Practical Data Analysis with JMP®

Second Edition

Robert H. Carver

Student Solutions

support.sas.com/bookstore

This set of Solutions for Students is a companion piece to the following SAS Press book: Carver, Robert. *Practical Data Analysis with JMP®, Second Edition*. Copyright © 2014, SAS Institute Inc., Cary, NC, USA. ALL RIGHTS RESERVED.

Practical Data Analysis with JMP®, Second Edition

Copyright © 2014, SAS Institute Inc., Cary, NC, USA

ISBN 978-1-61290-823-6

All rights reserved. Produced in the United States of America.

For a hard-copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a) and DFAR 227.7202-4 and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513-2414.

July 2014

SAS provides a complete selection of books and electronic products to help customers use SAS® software to its fullest potential. For more information about our offerings, visit **support.sas.com/bookstore** or call 1-800-727-3228.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

2

Student Solutions to Application Scenarios

Scenario 1

Student answers will vary. Answers will depend on data set student selects to input into a new JMP data table

Scenario 2

Quantity of cement (component 1), expressed as kg in a m^3 mixture.
Quantity of Superplasticizer (component 5), expressed as kg in a m^3 mixture.
Quantity of Fine Aggregate (component 7), expressed as kg in a m^3 mixture.

Scenario 3

Columns that need to be corrected: DMDMARTL, RIDEXPRG, BPQ150A

Scenario 4

NHANES does not contain experimental data because the experimenters are not manipulating any of the variables. The data was not obtained through a designed experiment but through observation.

Scenario 5

Open the **Military** table, and select **Rows ▶Row Selection ▶Select Randomly** and specify a sample size of **500**. Then choose **Tables ▶Subset**.

Scenario 6

This data table contains monthly stock values and volume from the FTSE 100 index, from1 January 2003 through 1 December 2007. Data were collected by observation on the first day of each month. The date column is ordinal because it is a chronological variable. Open, High, Low, Close, Volume, and change% are all Continuous columns containing numeric measurements. Open is the FTSE 100 index's opening price. High represents the high price for the day. Low is the low price for that day. Close is the closing price for that day. Volume is the number of shares exchanged during the day. change% is how much the index changed from open to close.

Scenario 7

This data table contains statistics from earthquakes recorded worldwide between August 20, 2009 and September 19, 2009. Data was collected by observation on the first day of each month. The date column is ordinal because it is a chronological variable. Latitude is a continuous variable indicating the latitudinal coordinate of where the earthquake took place. Longitude is also a continuous variable indicating the longitudinal coordinate of where the earthquake took place. Magnitude is a continuous measurement of how strong the earthquake was, while depth is a continuous variable describing how far from the surface the epicenter was. Time is an ordinal column describing when the earthquake took place. This data was found by observation.

Scenario 8

This table contains observational data from the WHO regarding tobacco use, cardiovascular disease and cancer rates. Code is a nominal variable uniquely identifying each nation. Country is a nominal variable that provides the name of the country relating to the data. Region is also a nominal variable indicating the region where the country is located in. TobaccoUse is a continuous variable observed describing the prevalence of tobacco use in that country. Female and Male are both continuous variables that were found observationally which describe the prevalence of tobacco use for both genders. CVmort is the mortality rate from cardiovascular disease for this country and CancerMort is the cancer mortality rate for this country. Both are continuous.

Scenario 10

The variables are Activity (travel, feed, social), Period (morning, noon, afternoon, evening), and Groups (numeric). The observational units were groups of dolphins. Activity and Period are nominal and Groups (# dolphins in each group) is continuous.

Scenario 11

The columns are as follows:

marst : marital status (nominal). Respondent's marital status, one of six levels

empstat: employment status (nominal). Respondent's employment status, one of five levels.

sleeping minutes spent sleeping each day (continuous).

telff minutes spent on the telephone with family and friends each day (continuous).

Scenario 11

This data table appears to contain demographic, economic, crime and other statistics for the 50 US states and the District of Columbia. The three specific variables are all continuous, and represent the following:

smoke is the percentage of the state population that smokes.

fed_spend is the per capita amount of federal spending in the state (dollars)

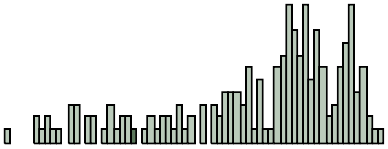
nuclear is the percentage of power coming from nuclear sources

3

Student Solutions to Application Scenarios

Scenario 1

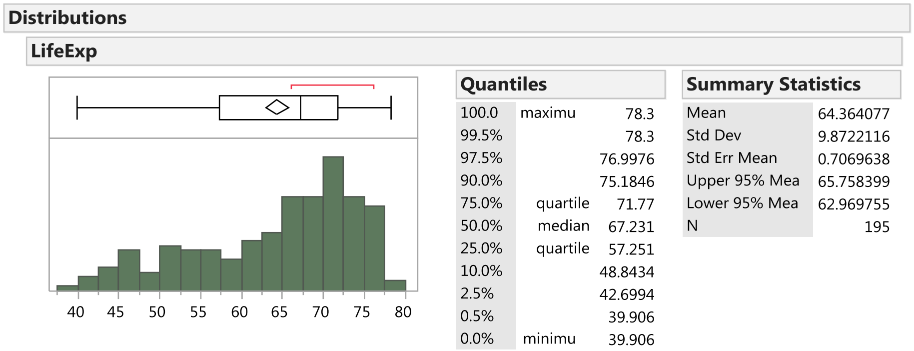
a. Using the grabber tool, click and drag upwards to increase the number of bars in the histogram. A second peak near 80 appears when as the number of bars increases, while the peak at 75 remains.



c. Scale can be manipulated in order to change the center, shape, and spread of a histogram, so it is important to carefully analyze and think critically about the choice of scale on an axis.

Scenario 2

a. This histogram has a shape that is skewed to the left, has a mean of about 70, and a spread described by a range from 35 to 80. It has one peak



Distributions

LifeExp

Quantiles

100.0	maximu	78.3
99.5%		78.3
97.5%		76.9976
90.0%		75.1846
75.0%	quartile	71.77
50.0%	median	67.231
25.0%	quartile	57.251
10.0%		48.8434
2.5%		42.6994
0.5%		39.906
0.0%	minimu	39.906

Summary Statistics

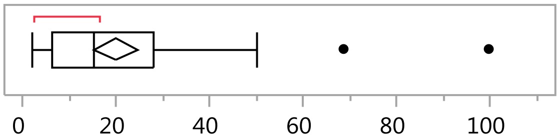
Mean	64.364077
Std Dev	9.8722116
Std Err Mean	0.7069638
Upper 95% Mea	65.758399
Lower 95% Mea	62.969755
N	195

c. The standard deviation is 9.87 in the 1985 data compared to 10.4 in the 2010 data.
.

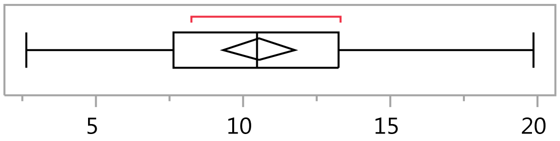
Scenario 3

a. The points furthest to the left and right indicate the minimum and maximum respectively. In each boxplot, the ends of the box represent the first and third quartiles, and the line within the box represents the median. The diamond shows the location of the mean. We see a handful of outlying points in the LifeSpan boxplot, but not in the TotalSleep plot.

LifeSpan



TotalSleep



c. 99.5% of the species have a life span less than 100 years.

e. The animals that get the most sleep tend to be relatively small animals and have low predation, exposure, and danger values.

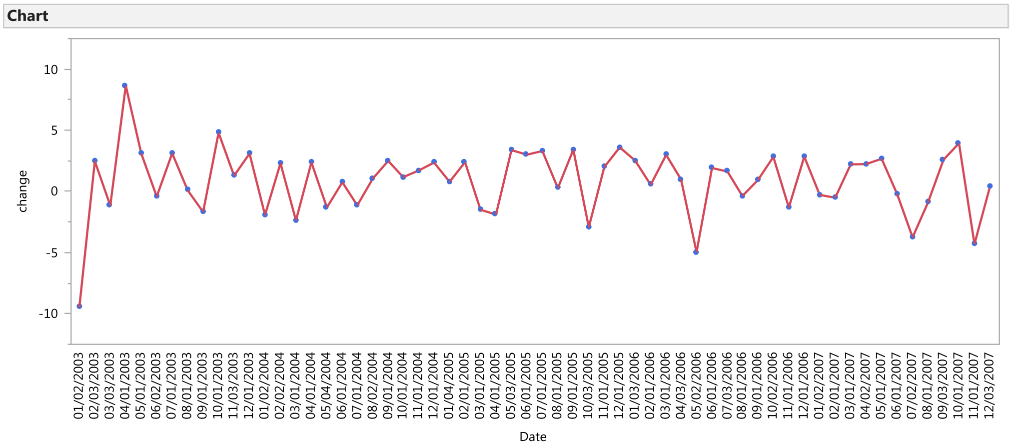
g. The animals that sleep in the most exposed locations are also the largest in terms of body weight. This may be because larger animals cannot hide as easily, or due to sheer size, they can sleep in exposed locations safely.

Scenario 4

a.... Volume has a nearly symmetrical and normal distribution. It ranges from 1043.49 to 2115.33 with a median of 1726.22 and a mean of 1710.49.

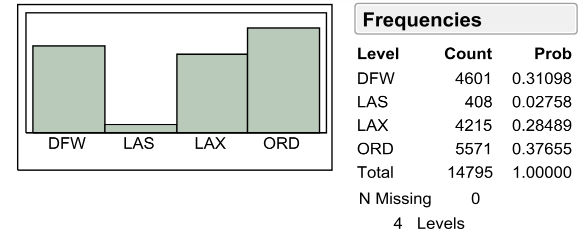
c. The FTSE declines approximately 25% of the time.

e. This line graph shows fluctuation without any obvious pattern. The monthly percentage change seems to vary at random from month to month, typically remaining approximately between –3% and +3%. There is no obvious growth over the five years, in contrast to the closing index value.



Scenario 5

a. The histogram has four bars. DFW, LAX, and ORD all have high with counts of over 4000 while LAS is low with only a count of around 400.



Distributions

DEST

Frequencies

Level	Count	Prob
DFW	4601	0.31098
LAS	408	0.02758
LAX	4215	0.28489
ORD	5571	0.37655
Total	14795	1.00000

N Missing 0
4 Levels

c. Because airlines attempt to schedule arrivals accurately, it is unlikely that very many flights would be extraordinarily early. However, given the many possible reasons for delays and the nature of travel, some flights can be exceptionally late. The practical minimum sets a lower bound for this variable, but there isn’t a comparable upper bound. As such, a few flights with very long delays will tend to skew the data.

Scenario 6

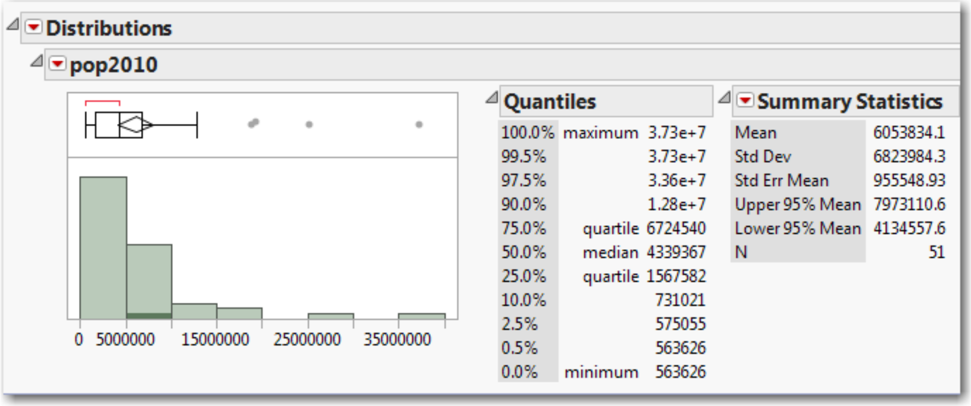
a. TobaccoUse is somewhat symmetrical with a mean of 24.77 and median of 25.6. It ranges from 4.3 to 51.8.

c.CVMort has two peaks at around 150 and 400. It is skewed to the right. It has a mean of 355.5 and a median of 375. It ranges from 106 to 713.

e.Europe & Central Asia and Sub-Saharan Africa have the highest count of countries in this data table. South Asia has the lowest count and America, East Asia & Pacific and Middle East & North Africa all fall in the middle.

Scenario 7

a.



Distributions

pop2010

Quantiles

100.0%	maximum	3.73e+7
99.5%		3.73e+7
97.5%		3.36e+7
90.0%		1.28e+7
75.0%	quartile	6724540
50.0%	median	4339367
25.0%	quartile	1567582
10.0%		731021
2.5%		575055
0.5%		563626
0.0%	minimum	563626

Summary Statistics

Mean	6053834.1
Std Dev	6823984.3
Std Err Mean	955548.93
Upper 95% Mean	7973110.6
Lower 95% Mean	4134557.6
N	51

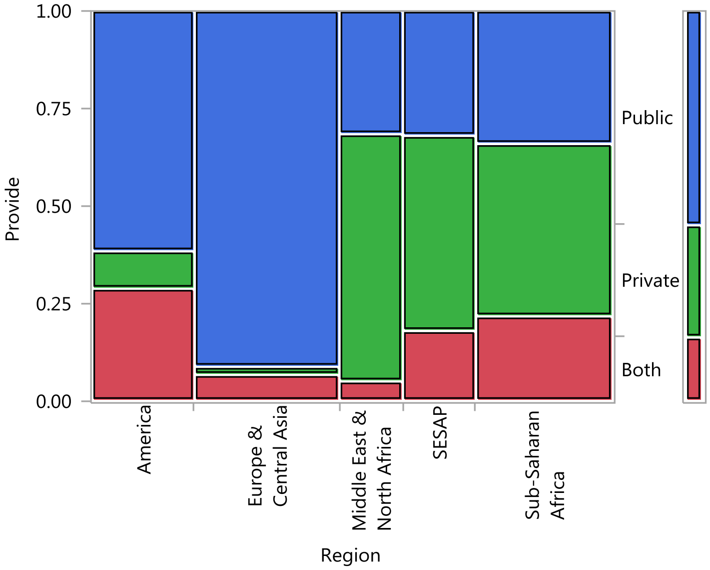
This is a strongly right-skewed distribution with 4 outliers (California, Texas, New York, and Florida). The mean population was 6,053,834 people and the median was just 4,339,367. States range from approximately 563,000 people in Wyoming to more than 37 million in California. The largest number of states have fewer than 5 million residents.

4

Student Solutions to Application Scenarios

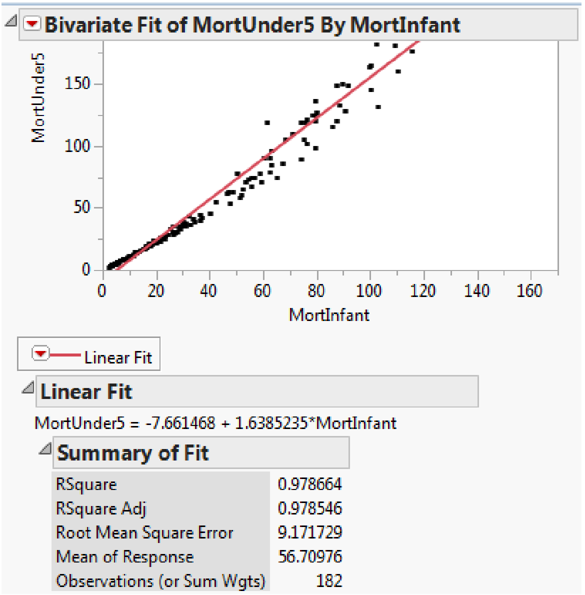
Scenario 1

a. **Mosaic Plot**



Public provision is most common by far in the Americas and Europe & Central Asia. Provide provision seems to be the norm in the rest of the world. Most areas have relatively few countries with both public and private, though such arrangements are fairly common (more than 25% of countries) in the Americas.

c.



Bivariate Fit of MortUnder5 By MortInfant

Linear Fit

Linear Fit

MortUnder5 = -7.661468 + 1.6385235*MortInfant

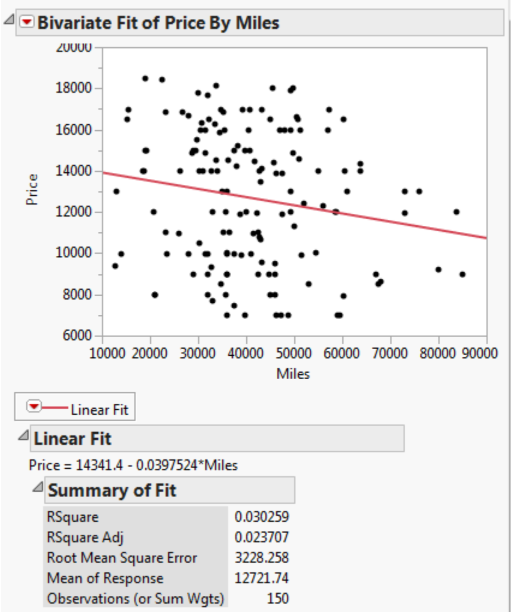
Summary of Fit

RSquare	0.978664
RSquare Adj	0.978546
Root Mean Square Error	9.171729
Mean of Response	56.70976
Observations (or Sum Wgts)	182

The fitted line and RSquare value are shown above. There is a strong, positive linear relationship between the two variables.

Scenario 2

a.



Bivariate Fit of Price By Miles

Linear Fit

Linear Fit

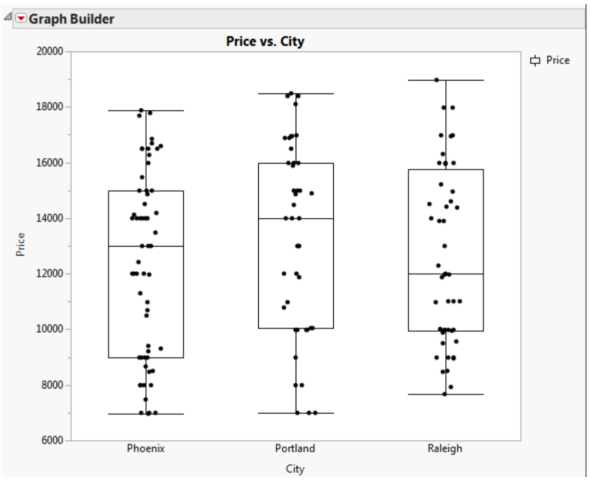
Price = 14341.4 - 0.0397524*Miles

Summary of Fit

RSquare	0.030259
RSquare Adj	0.023707
Root Mean Square Error	3228.258
Mean of Response	12721.74
Observations (or Sum Wgts)	150

The plot, equation and Rsquare are shown above. The correlation coefficient is 0.17395. There is a weak negative relationship between mileage and price: the higher the mileage, the lower the price.

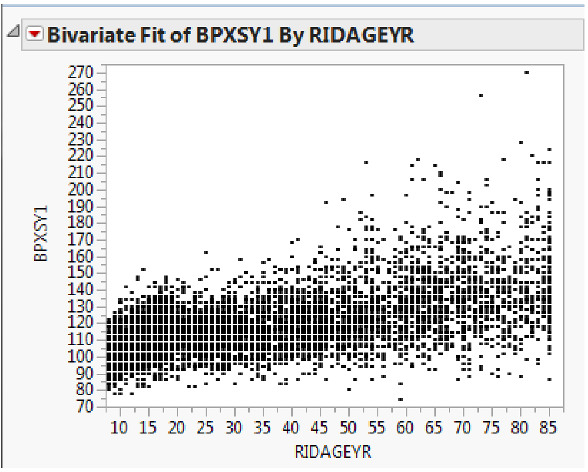
c.



The distribution of price across the three cities seems to be fairly uniform. The box plot shows similar middle 50% with varying means. They also have very similar spreads.

Scenario 3

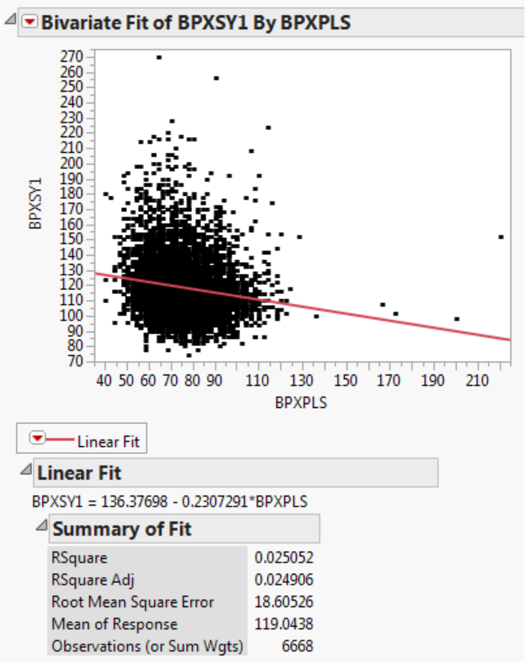
a.



As individuals get older, blood pressure increases.

c. Men have a higher average systolic blood pressure. Both genders have similar shape, being skewed to the right. Women have a far greater range, spanning from 70 to 270 while men have readings from 80 to about 210.

e.



Bivariate Fit of BPXSY1 By BPXPLS

Linear Fit

BPXSY1 = 136.37698 - 0.2307291*BPXPLS

Summary of Fit

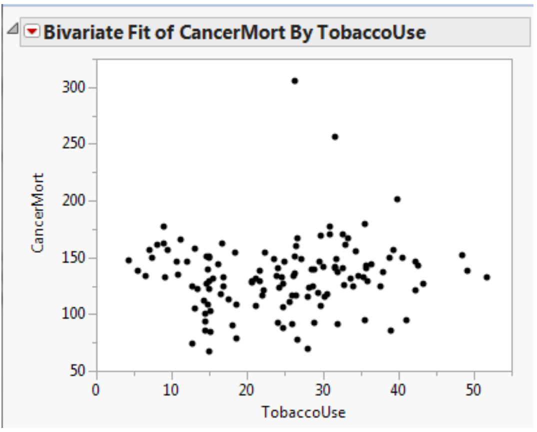
RSquare	0.025052
RSquare Adj	0.024906
Root Mean Square Error	18.60526
Mean of Response	119.0438
Observations (or Sum Wgts)	6668

It appears there is little evidence of a relationship between pulse and blood pressure, as the r squared statistic is .2, which is very low.

Scenario 4

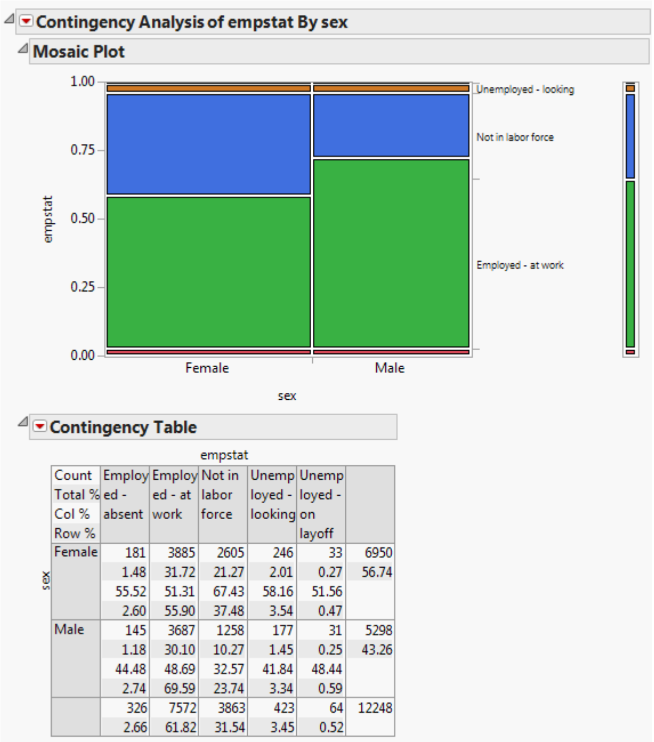
a. Tobacco is most heavily used in Europe and Central Asia and to a lesser extent in East Asia and the Pacific. There is a moderate use in the Middle East and North Africa as well as the Americas while South Asia and Sub-Saharan Africa has the lowest tobacco use.

c.



Here again, we find scant evidence of a relationship.

Scenario 5



empstat

Count Total % Col % Row %	Employed - absent	Employed - at work	Not in labor force	Unemployed - looking	Unemployed - on layoff	
Female	181	3885	2605	246	33	6950
	1.48	31.72	21.27	2.01	0.27	56.74
	55.52	51.31	67.43	58.16	51.56	
	2.60	55.90	37.48	3.54	0.47	
Male	145	3687	1258	177	31	5298
	1.18	30.10	10.27	1.45	0.25	43.26
	44.48	48.69	32.57	41.84	48.44	
	2.74	69.59	23.74	3.34	0.59	
	326	7572	3863	423	64	12248
	2.66	61.82	31.54	3.45	0.52	

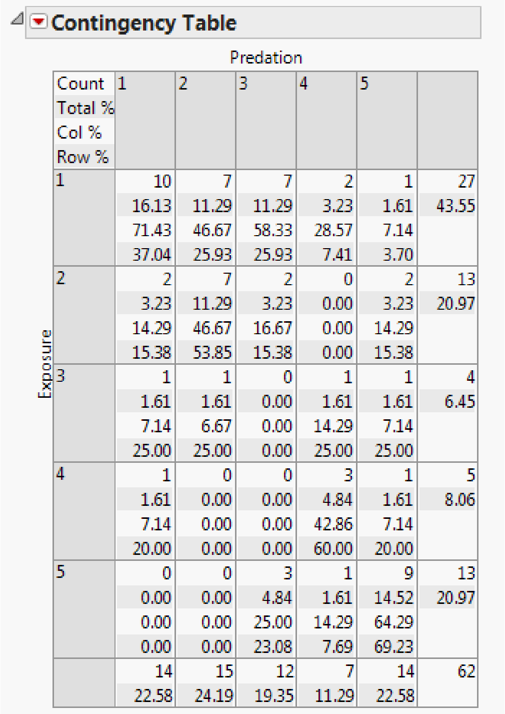
sex

More males were employed (at work) than females while more females were not in the labor force. About the same amount of males and females were unemployed and looking or employed and absent.

c.People employed had the lowest mean time spent sleeping. All employment statuses had nearly normal distributions with some like employed at work being more skewed to the right than others. Nearly all the spreads of employment categories ranged across the same amount of time.

Scenario 6

a.

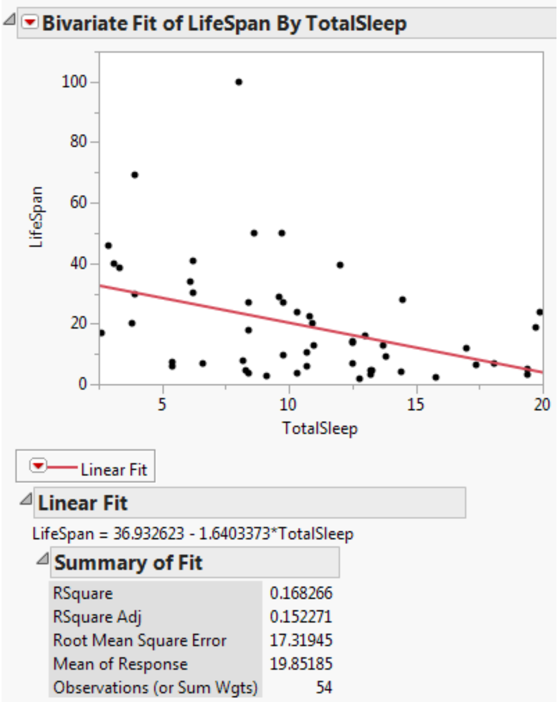


Contingency Table

		Predation					
Exposure	Count Total % Col % Row %	1	2	3	4	5	
	1	10 16.13 71.43 37.04	7 11.29 46.67 25.93	7 11.29 58.33 25.93	2 3.23 28.57 7.41	1 1.61 7.14 3.70	27 43.55
	2	2 3.23 14.29 15.38	7 11.29 46.67 53.85	2 3.23 16.67 15.38	0 0.00 0.00 0.00	2 3.23 14.29 15.38	13 20.97
	3	1 1.61 7.14 25.00	1 1.61 6.67 25.00	0 0.00 0.00 0.00	1 1.61 14.29 25.00	1 1.61 7.14 25.00	4 6.45
	4	1 1.61 7.14 20.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	3 4.84 42.86 60.00	1 1.61 7.14 20.00	5 8.06
	5	0 0.00 0.00 0.00	0 0.00 0.00 0.00	3 4.84 25.00 23.08	1 1.61 14.29 7.69	9 14.52 64.29 69.23	13 20.97
		14 22.58	15 24.19	12 19.35	7 11.29	14 22.58	62

Animals with lower exposure values seem to have lower predation ratings. Conversely, creatures with higher exposure values also had higher predation ratings.

c.



Bivariate Fit of LifeSpan By TotalSleep

Linear Fit

LifeSpan = 36.932623 - 1.6403373*TotalSleep

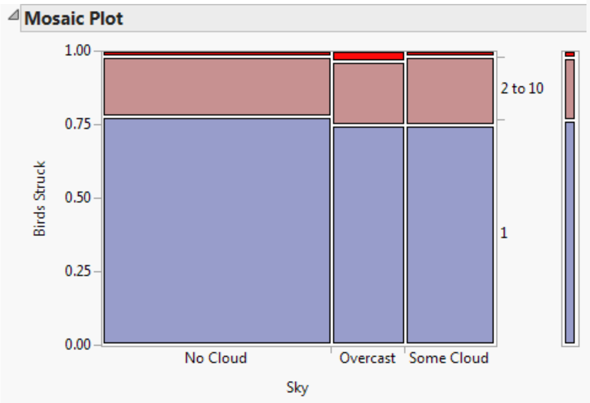
Summary of Fit

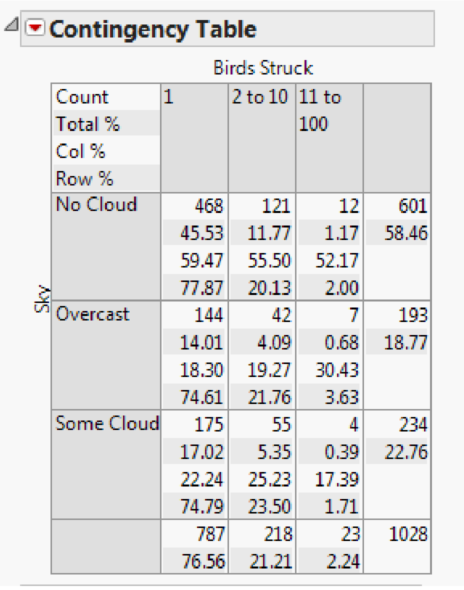
RSquare	0.168266
RSquare Adj	0.152271
Root Mean Square Error	17.31945
Mean of Response	19.85185
Observations (or Sum Wgts)	54

There seems to be evidence of a weak negative relationship between lifespan and total sleep. The Rsquare statistic is only 0.168.

Scenario 7

a.





Contingency Table

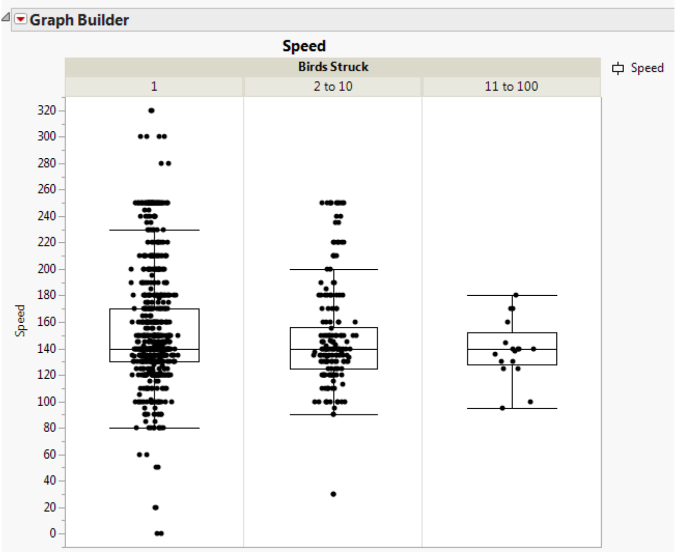
Birds Struck

Sky

Count Total % Col % Row %	1	2 to 10	11 to 100	
No Cloud	468 45.53 59.47 77.87	121 11.77 55.50 20.13	12 1.17 52.17 2.00	601 58.46
Overcast	144 14.01 18.30 74.61	42 4.09 19.27 21.76	7 0.68 30.43 3.63	193 18.77
Some Cloud	175 17.02 22.24 74.79	55 5.35 25.23 23.50	4 0.39 17.39 1.71	234 22.76
	787 76.56	218 21.21	23 2.24	1028

Neither the mosaic plot nor the contingency table show much evidence of large differences in number of birds struck across different sky conditions. Regardless of conditions, for example, it appears that about 75% of incidents involve a single bird.

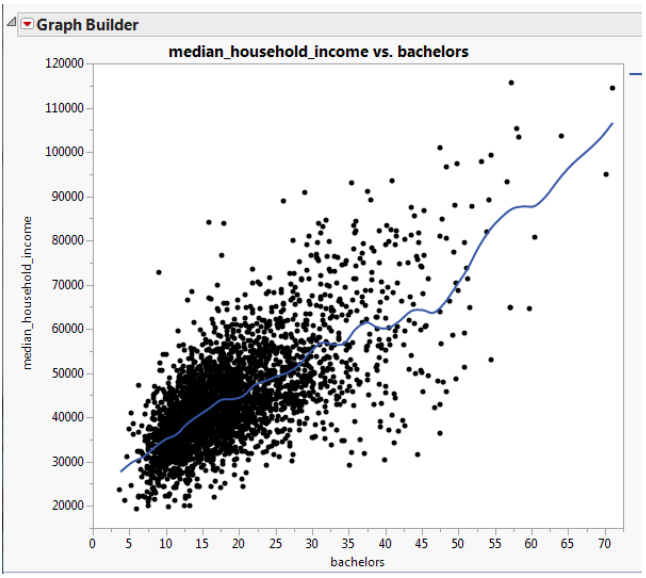
c.



There are various ways to approach this question. One simple way is to explore the relationship using Graph Builder. In this graph we see that median speed is approximately the same regardless of the number of birds struck. However, single-bird incidents occur at a wide variety of speed; as the number of birds involved increases, the variability of speed decreases.

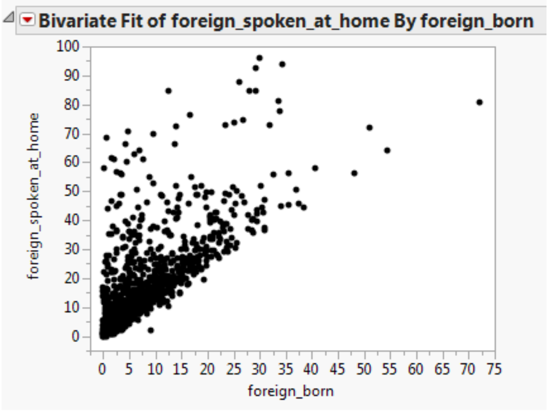
Scenario 8

a.



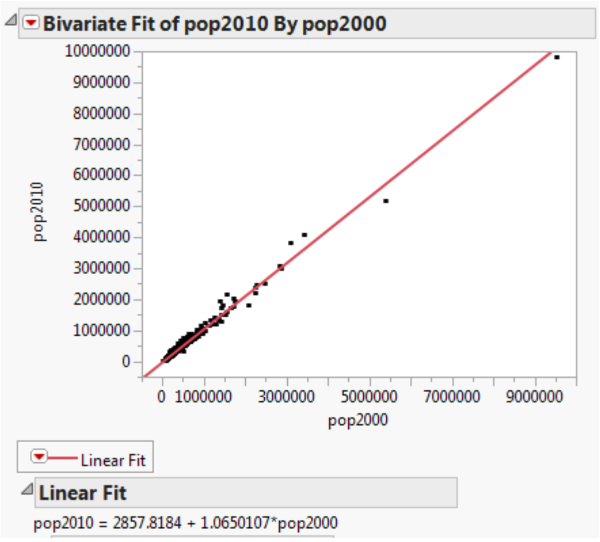
Using Graph Builder to investigate this relationship we find a positive but inconsistent relationship between income and percentage of population with a bachelor’s degree. There is a clear upward pattern with a lot of scatter, indicating that a relationship exists but it is not very strong.

c.



There are very few counties lying below the 45-degree diagonal line, indicating that the percentage of homes where a foreign language is spoken almost always exceeds the percentage of homes with a foreign-born member. This makes sense, assuming that homes with no foreign-born members would be less inclined to speak a foreign language.

e.



```
pop2010 = 2857.8184 + 1.0650107*pop2000
```

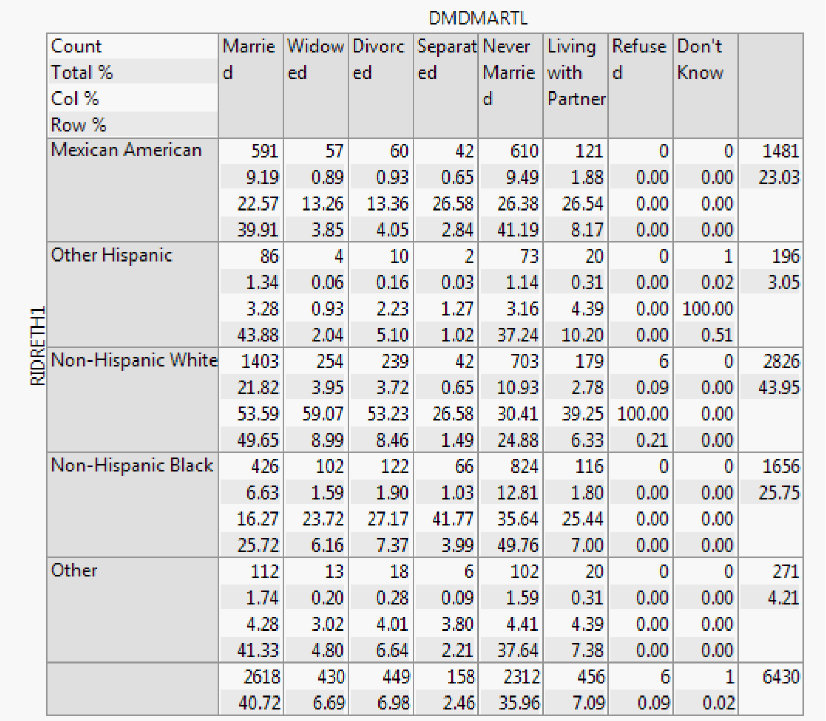
The slope of the line is approximately 1.065, indicating that on average, the population of US counties grew by 6.5% from 2000 to 2010.

6

Student Solutions to Application Scenarios

Scenario 1

NOTE: This contingency table provides the necessary information to respond to all parts:



DMDMARTL

	Count Total % Col % Row %	Married	Widowed	Divorced	Separated	Never Married	Living with Partner	Refused	Don't Know	
RIDRETH1	Mexican American	591	57	60	42	610	121	0	0	1481
		9.19	0.89	0.93	0.65	9.49	1.88	0.00	0.00	23.03
		22.57	13.26	13.36	26.58	26.38	26.54	0.00	0.00	
		39.91	3.85	4.05	2.84	41.19	8.17	0.00	0.00	
	Other Hispanic	86	4	10	2	73	20	0	1	196
		1.34	0.06	0.16	0.03	1.14	0.31	0.00	0.02	3.05
		3.28	0.93	2.23	1.27	3.16	4.39	0.00	100.00	
		43.88	2.04	5.10	1.02	37.24	10.20	0.00	0.51	
	Non-Hispanic White	1403	254	239	42	703	179	6	0	2826
		21.82	3.95	3.72	0.65	10.93	2.78	0.09	0.00	43.95
		53.59	59.07	53.23	26.58	30.41	39.25	100.00	0.00	
		49.65	8.99	8.46	1.49	24.88	6.33	0.21	0.00	
	Non-Hispanic Black	426	102	122	66	824	116	0	0	1656
		6.63	1.59	1.90	1.03	12.81	1.80	0.00	0.00	25.75
		16.27	23.72	27.17	41.77	35.64	25.44	0.00	0.00	
		25.72	6.16	7.37	3.99	49.76	7.00	0.00	0.00	
	Other	112	13	18	6	102	20	0	0	271
		1.74	0.20	0.28	0.09	1.59	0.31	0.00	0.00	4.21
		4.28	3.02	4.01	3.80	4.41	4.39	0.00	0.00	
		41.33	4.80	6.64	2.21	37.64	7.38	0.00	0.00	
		2618	430	449	158	2312	456	6	1	6430
		40.72	6.69	6.98	2.46	35.96	7.09	0.09	0.02	

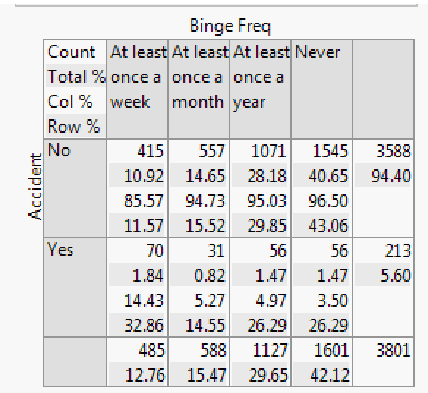
a. *Pr(Mexican American)* =0.2303

c. *Pr(Mexican American and Never Married)* = 0.0949.

e. No. In part e we found that *Pr(Never Married|Mexican American)*= 0.4114. The marginal probability *Pr(Never Married)* = 0.3596. Because the probabilities are unequal, we find that the events are not independent.

Scenario 2

For all of the questions that follow, we can use this contingency table:



Binge Freq

	Count Total % Col % Row %	At least once a week	At least once a month	At least once a year	Never	
Accident	No	415	557	1071	1545	3588
		10.92	14.65	28.18	40.65	94.40
		85.57	94.73	95.03	96.50	
		11.57	15.52	29.85	43.06	
	Yes	70	31	56	56	213
		1.84	0.82	1.47	1.47	5.60
		14.43	5.27	4.97	3.50	
		32.86	14.55	26.29	26.29	
		485	588	1127	1601	3801
		12.76	15.47	29.65	42.12	

a. *Pr(Binge at least once a week)* = 0.1276.

c. *Pr(Accident)* = 0.0560.

e. *Pr(Accident | binge at least once a week)* = 0.1443.

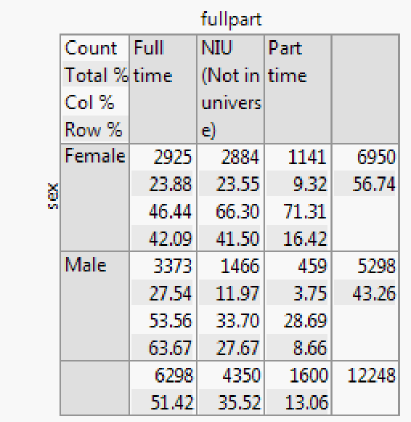
g. No. Comparing the results in parts a and f or parts c and e should lead to the conclusion that because the relevant marginal probabilities do not equal the corresponding conditionals, the events are not independent.

Scenario 3

NOTE: Different contingency tables are needed for different parts of this problem.

a. *Pr(Not in labor force) =0.3154*

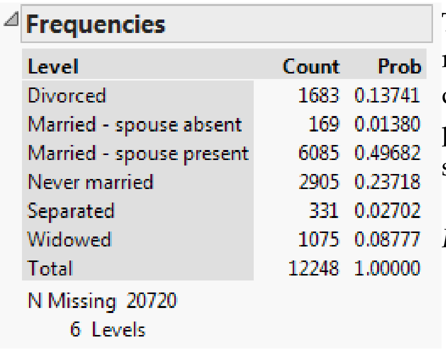
Parts c and d rely on this table:



		fullpart			
sex	Count Total % Col % Row %	Full time	NIU (Not in universe)	Part time	
	Female	2925	2884	1141	6950
		23.88	23.55	9.32	56.74
		46.44	66.30	71.31	
		42.09	41.50	16.42	
	Male	3373	1466	459	5298
		27.54	11.97	3.75	43.26
		53.56	33.70	28.69	
		63.67	27.67	8.66	
		6298	4350	1600	12248
		51.42	35.52	13.06	

c. *Pr(Part-time or female) = Pr(part-time)+ Pr(female) – Pr(part-time and female)* = 0.1306 + 0.5674 – 0.0932 = 0.6048

e.



Frequencies

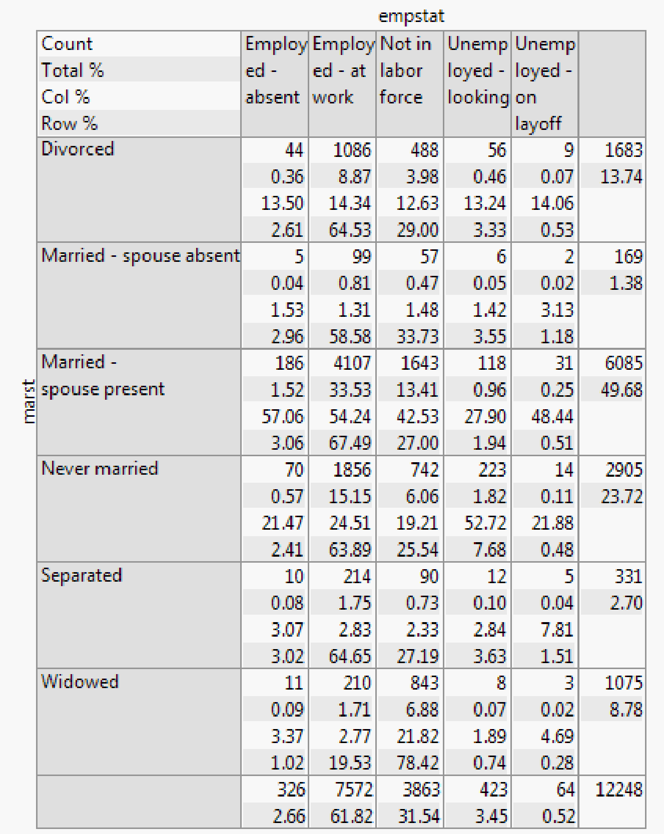
Level	Count	Prob
Divorced	1683	0.13741
Married - spouse absent	169	0.01380
Married - spouse present	6085	0.49682
Never married	2905	0.23718
Separated	331	0.02702
Widowed	1075	0.08777
Total	12248	1.00000

N Missing 20720
6 Levels

The marital status column identifies three types of respondents who are not married: those who are divorced, never married, or widowed. To find the probability of selecting a person who is not married, we sum the probabilities of these three categories:

Pr(Not Married) =0.13741 + 0.23718 + 0.08777 = 0.46236.

This table can be used for Part f:



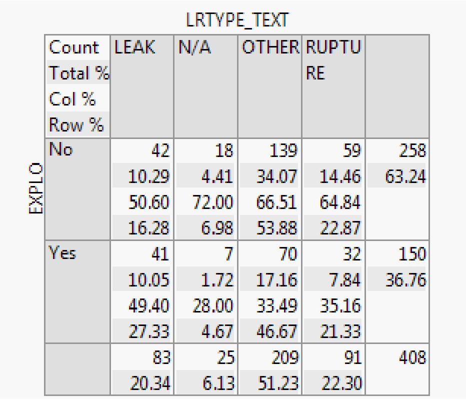
	empstat					
Count Total % Col % Row %	Employed - absent	Employed - at work	Not in labor force	Unemployed - looking	Unemployed - on layoff	
Divorced	44 0.36 13.50 2.61	1086 8.87 14.34 64.53	488 3.98 12.63 29.00	56 0.46 13.24 3.33	9 0.07 14.06 0.53	1683 13.74
Married - spouse absent	5 0.04 1.53 2.96	99 0.81 1.31 58.58	57 0.47 1.48 33.73	6 0.05 1.42 3.55	2 0.02 3.13 1.18	169 1.38
Married - spouse present	186 1.52 57.06 3.06	4107 33.53 54.24 67.49	1643 13.41 42.53 27.00	118 0.96 27.90 1.94	31 0.25 48.44 0.51	6085 49.68
Never married	70 0.57 21.47 2.41	1856 15.15 24.51 63.89	742 6.06 19.21 25.54	223 1.82 52.72 7.68	14 0.11 21.88 0.48	2905 23.72
Separated	10 0.08 3.07 3.02	214 1.75 2.83 64.65	90 0.73 2.33 27.19	12 0.10 2.84 3.63	5 0.04 7.81 1.51	331 2.70
Widowed	11 0.09 3.37 1.02	210 1.71 2.77 19.53	843 6.88 21.82 78.42	8 0.07 1.89 0.74	3 0.02 4.69 0.28	1075 8.78
	326 2.66	7572 61.82	3863 31.54	423 3.45	64 0.52	12248

Scenario 4

a. *Pr(Central)*= 0.2863

c. This problem is complicated by the fact that most cells in this column are blank and the remaining cells contain the label "Yes." There are 189 "Yes" values and 468 rows in all. Therefore *Pr(Evacuation)* = 189/468 = 0.4038.

e.



LRTYPE_TEXT

EXPLO

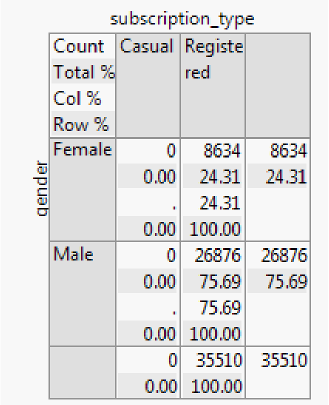
Count Total % Col % Row %	LEAK	N/A	OTHER	RUPTURE	
No	42 10.29 50.60 16.28	18 4.41 72.00 6.98	139 34.07 66.51 53.88	59 14.46 64.84 22.87	258 63.24
Yes	41 10.05 49.40 27.33	7 1.72 28.00 4.67	70 17.16 33.49 46.67	32 7.84 35.16 21.33	150 36.76
	83 20.34	25 6.13	209 51.23	91 22.30	408

Pr(Rupture or Explosion) = *Pr(Rupture)* + *Pr(Explosion)* – *Pr(Rupture and Explosion)* = 0.2230 + 0.3676 – 0.0784 = 0.5122.

Scenario 5

a.*Pr(Registered) = 0.6295*

c.This table applies to questions c and d:



		subscription_type		
	Count Total % Col % Row %	Casual	Registered	
gender	Female	0 0.00 . 0.00	8634 24.31 24.31 100.00	8634 24.31
	Male	0 0.00 . 0.00	26876 75.69 75.69 100.00	26876 75.69
		0 0.00	35510 100.00	35510

Pr(female who is registered) = 0.2431.

Note: The full contingency table is too large to reproduce effectively here.

e.*Pr(began at South Station)* = 0.0477 = 2,636/ 55,230 trips.

g.*Pr(start at South Station and end at Library)* = 0.0013 = 72/55,230 trips.

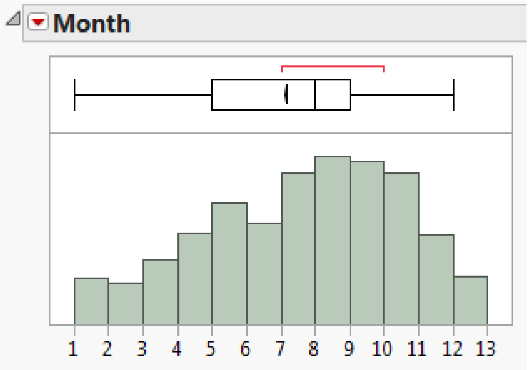
Scenario 6

a.A Fit Y by X contingency table shows that 19 of the 1,000 women were smokers with premature babies. Hence, the probability is 0.019.

c.A Fit Y by X contingency table shows that 11 of the 1,000 women were smokers and mature moms. Hence, the probability is 0.011.

Scenario 7

a.



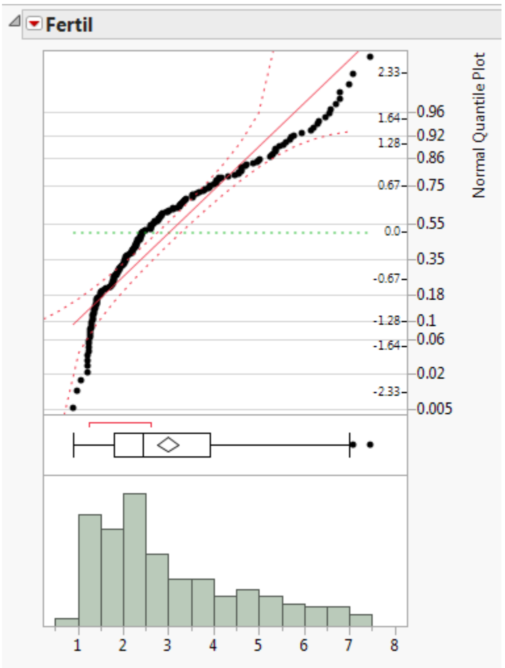
Birdstrikes occur most often July through October (months 7 through 10), and rather infrequently during December, January, and February. So, we would say that they do not occur with equal frequency through the year.

7

Student Solutions to Application Scenarios

Scenario 1

a.

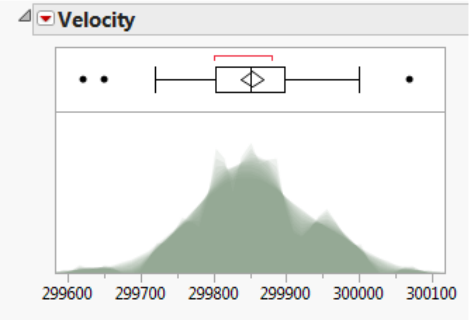


The normal quantile plot appears to the left. The distribution is strongly skewed positively, and therefore the normal model is not suitable for this variable.

c. Pr(X>5.5) = 1 – 0.9426 = 0.0574. In comparison, based on the reported quantiles, we find that more than 10% of the observed data lies above 5.5 children per woman.

Scenario 2

a.

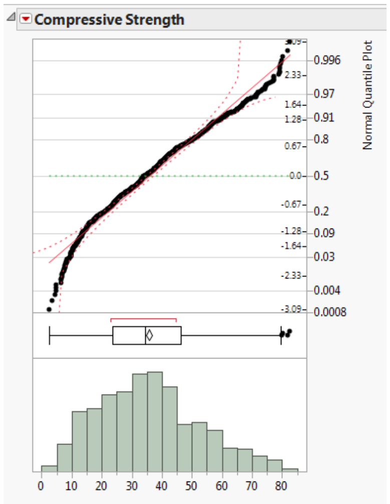


In the shadowgram to the left we see a generally symmetric distribution that seems to be mound-shaped. There may be some indication of a secondary peak at approximately 299,950 km/sec., but the overall impression is that the distribution might be well-described by the normal model.

c. The data set provides some support for the assumption. Michelson's various measurements of the speed of light seem to vary according to an approximate normal distribution.

Scenario 3

a.



This distribution shows mild skewness. The lower tail is truncated and therefore shorter and thicker than a normal distribution would be.

This set of Solutions for Students is a companion piece to the following SAS Press book: Carver, Robert. Practical Data Analysis with JMP®, Second Edition.

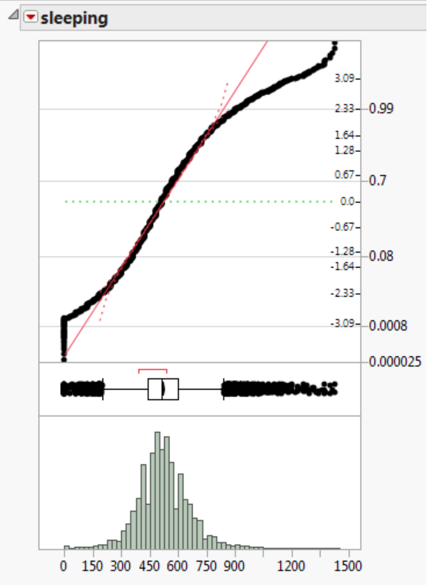
Scenario 4

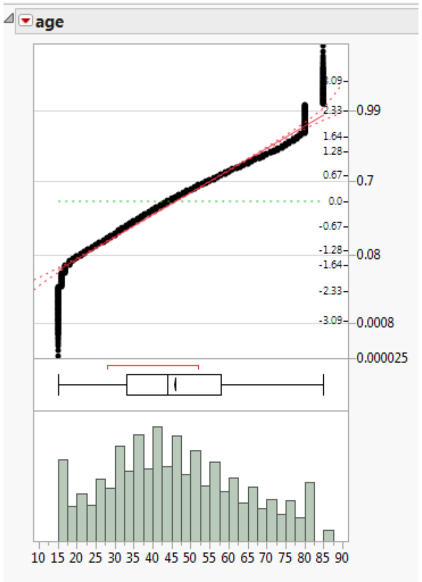
a. Student answers will vary. Most will likely choose the weekly change column corresponding to the Hang Seng market index, but others might select a different column (e.g. Tel Aviv or S&P). In these graphs, the points track most closely to the diagonal line.

c. The mean and standard deviation of the changes in Hang Seng for the weeks observed are –1.102065 and 5.242892. For a normal distribution with that mean and standard deviation, Pr (X <0) =0.5832, or approximately 0.58.

Scenario 5

Use these graphs to respond to all parts:



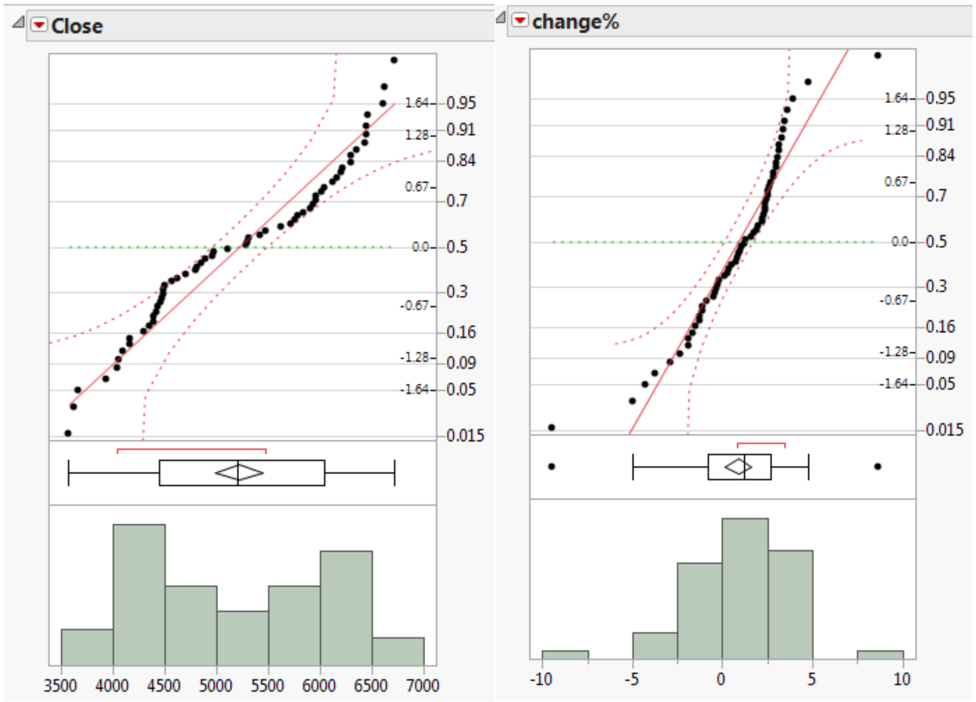


a. This histogram is mound-shaped with a single peak centered near 500 minutes. The large majority of respondents report between approximately 300 and 700 minutes of sleep per week.

c. The Age histogram is more skewed that the Sleeping histogram, with distinct secondary peaks in each tail. It appears to be centered near 40, but with the peaks in the tails it is difficult to generalize about the degree of dispersion. Again, the normal quantile plot casts doubts on using a normal model for this variable. The normal model seems to fit acceptably near the center of the distribution, but deviates quite dramatically in the tails.

Scenario 6

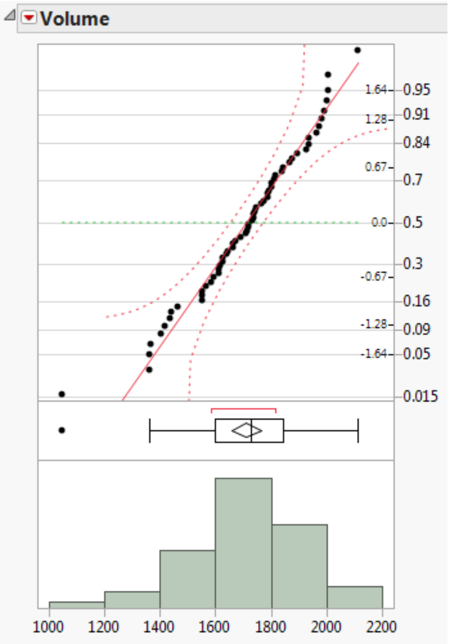
These graphs can be used to respond to parts a and b.



a. Closing values appear to be symmetric and bimodal, with peaks between 4000-5000 and 6000-6500. The center of the distribution is close to 5000 and it ranges from approximately 3500 to 7000.

In contrast, the %change column is moderately symmetric with a single peak just above 0. Most of the distribution lies between -5 % and +5 %.

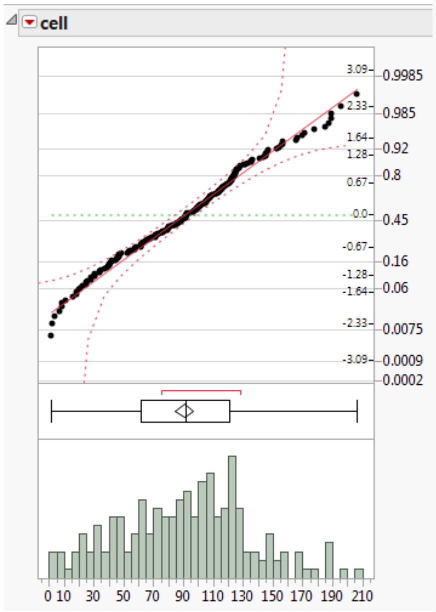
c.

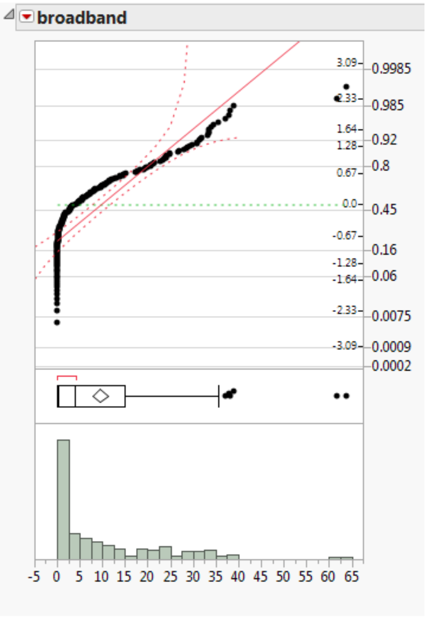


The volume column has a normal quantile plot that looks quite close to a normal distribution. It would be well described by a model ~N(1710.4911, 203.1369).

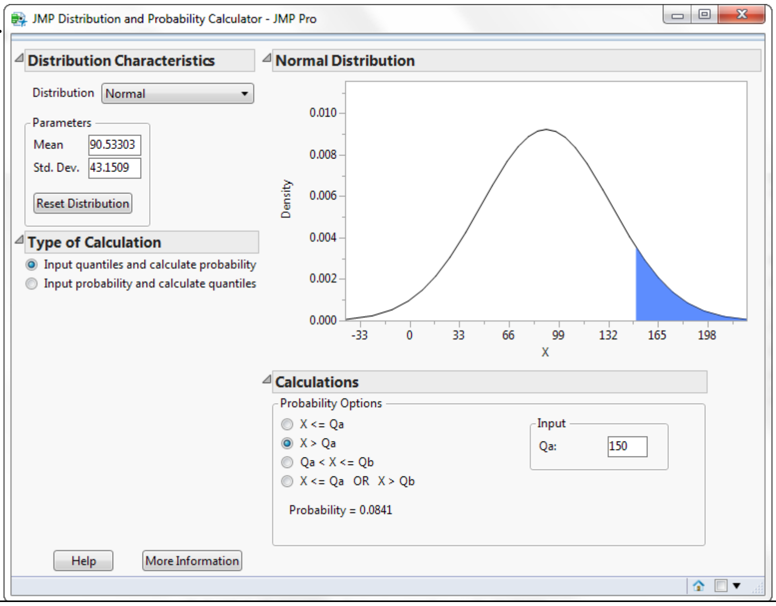
Scenario 7

a. Here are the graphs, which very clearly show that the cell column is better modeled as normal than the broadband data. The broadband histogram is strongly skewed to the right and its probability plot does not track the diagonal line at all.





c.



Distribution: Normal

Parameters
Mean: 90.53303
Std. Dev.: 43.1509

Reset Distribution

Type of Calculation

● Input quantiles and calculate probability
○ Input probability and calculate quantiles

Calculations

Probability Options
○ X <= Qa
● X > Qa
○ Qa < X <= Qb
○ X <= Qa OR X > Qb

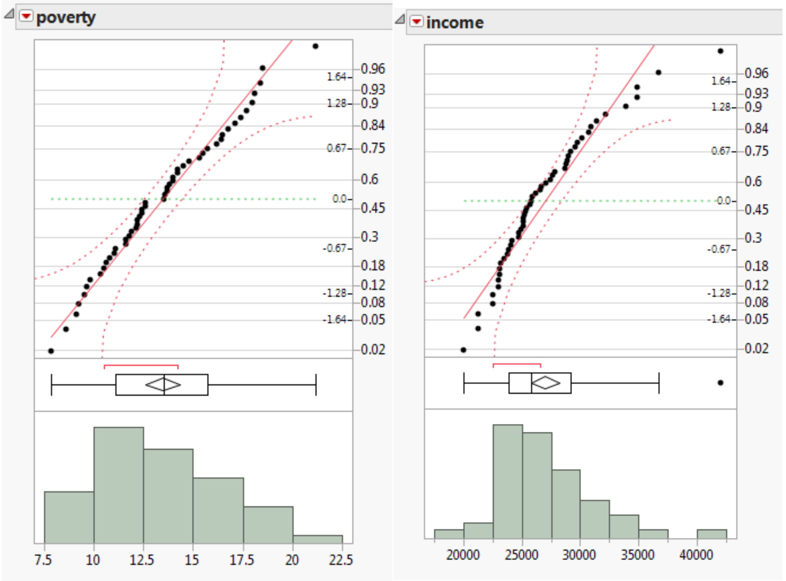
Input
Qa: 150

Probability = 0.0841

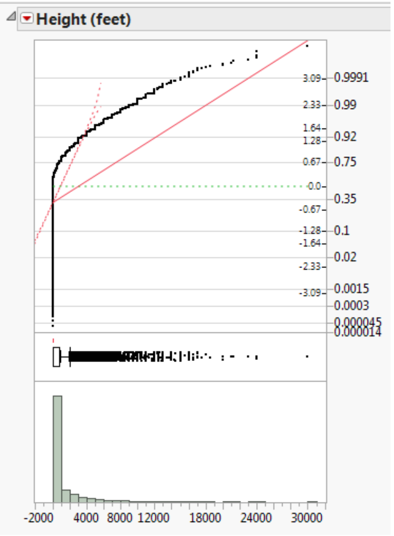
Using the normal model, we would estimate that approximately 8.4% of countries had 150 or more cell subscribers per 100 people.

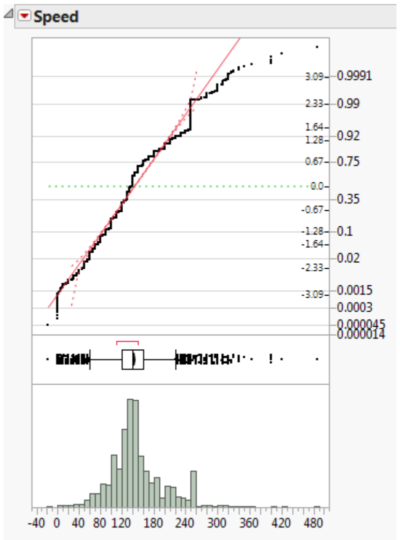
Scenario 8

a. Use Analyze > Distribution to obtain histograms, then red triangle Normal Quantile plot to produce these two plots:



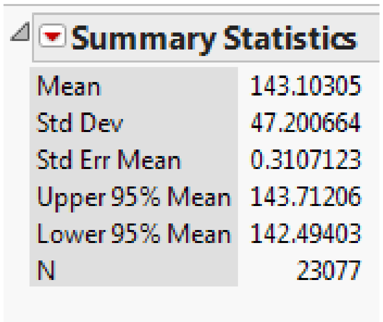
Scenario 9





a. The graphs above show that Height is very strongly skewed to the right with many outliers. Speed is more closely normal, through there is a second mode at approximately 250 mph.

c.



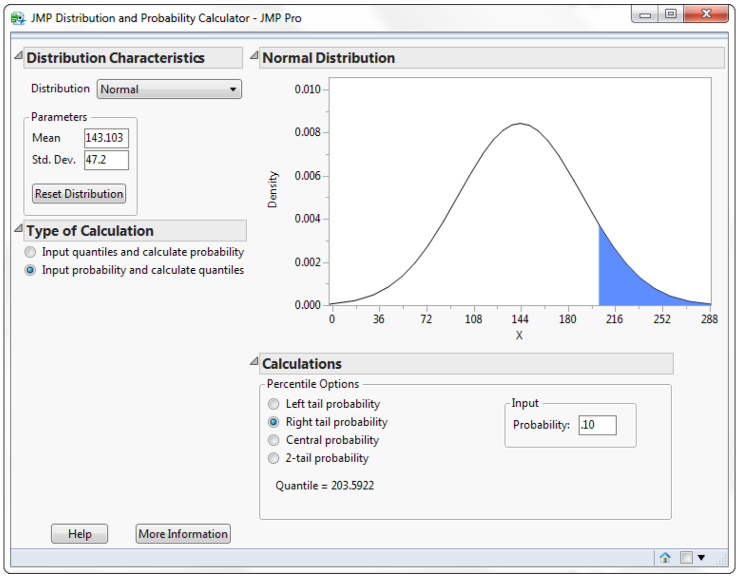
Summary Statistics

Mean	143.10305
Std Dev	47.200664
Std Err Mean	0.3107123
Upper 95% Mean	143.71206
Lower 95% Mean	142.49403
N	23077

From the distribution platform, we find that the mean is 143.103 mph and the standard deviation is 47.2.

Placing these values into the normal distribution calculator, we can approximate that the 90th percentile of the normal distribution is 203.6 mph.

In this instance, the normal approximation comes reasonably close to the observed data.



Distribution: Normal

Mean: 143.103

Std. Dev.: 47.2

○ Input quantiles and calculate probability
● Input probability and calculate quantiles

○ Left tail probability
● Right tail probability
○ Central probability
○ 2-tail probability

Probability: .10

Quantile = 203.5922

8

Student Solutions to Application Scenarios

Scenario 1

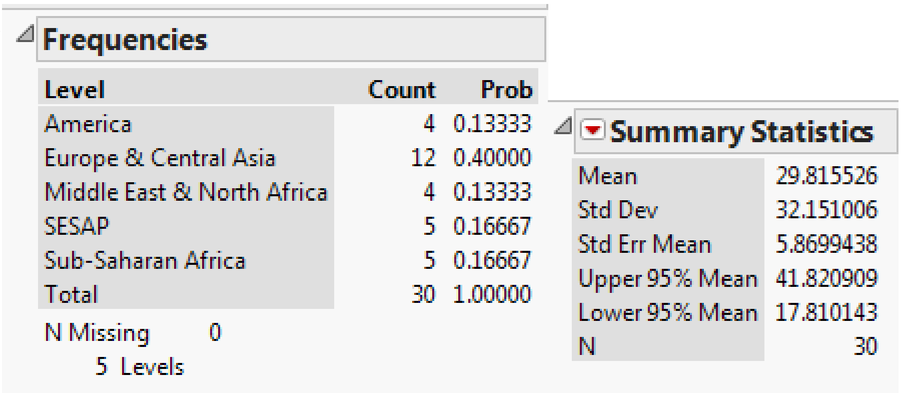
a.Student answers will vary due to the operation of the random number generator.

c.The probability that a SRS of 250 households would include 25 or fewer homes without Internet service is 0.00031368.

Scenario 2

a. The proportion of countries in Sub-Saharan Africa is 0.24227.

c.



Frequencies

Level	Count	Prob
America	4	0.13333
Europe & Central Asia	12	0.40000
Middle East & North Africa	4	0.13333
SESAP	5	0.16667
Sub-Saharan Africa	5	0.16667
Total	30	1.00000

N Missing 0

5 Levels

Summary Statistics

Mean	29.815526
Std Dev	32.151006
Std Err Mean	5.8699438
Upper 95% Mean	41.820909
Lower 95% Mean	17.810143
N	30

Student answers will vary due to random sampling. Above we find the results of one random sample—only 5 of the 30 countries are in Sub-Saharan Africa (16.7%). The mean mortality rate in the sample is 29.82 (note that in this sample all 30 countries reported an infant mortality rate). In general students' results will not match the population values shown in parts a & b due to sampling variation.

Scenario 3

a. Student answers will vary. In general, the sampling distribution will be bell-shaped and symmetrical, centered very near 0.40 and ranging from about 0.35 to 0.45.

c. Student answers will vary again. In general, the sampling distribution will be roughly bell-shaped and possibly a little left skewed, centered very near 0.95. Compare to the distribution in part c, this distribution will be steep and range only from about 0.90 to 1.00.

e. In part c we notice that the population with a proportion of .95 generates samples with comparatively small standard errors. The risks associated with sampling variation tend to be smaller in more uniform populations.

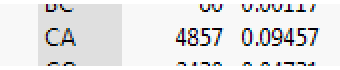
Scenario 4

a. Student responses will vary. In general, the sampling distribution will be bell-shaped and symmetrical, centered very near 15 with an overall standard error (std. deviation of the sample means) approximately equal to 0.10 and ranging from about 14.7 to 15.3.

c. Student responses will again vary. In general, the sampling distribution will be bell-shaped and symmetrical, centered very near 15 with an overall standard error (std. deviation of the sample means) approximately equal to 0.40 and ranging from about 13.8 to 16.2.

e. The results will be very similar to parts a and d though each student may have slightly different numerical results.

Scenario 5

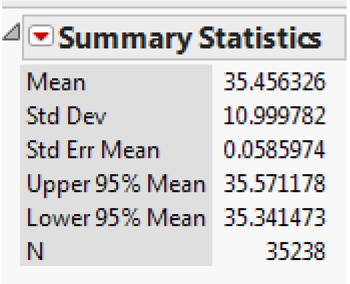
a. 

CA	4857	0.09457

4, 857 strikes occurred in California, for a proportion of 0.095, or about 9.5%.

c. Each student will obtain a different SRS, so these answers will vary. In general they will differ from the values in parts a and b due to the chance variation associated with random sampling.

Scenario 6

a. 

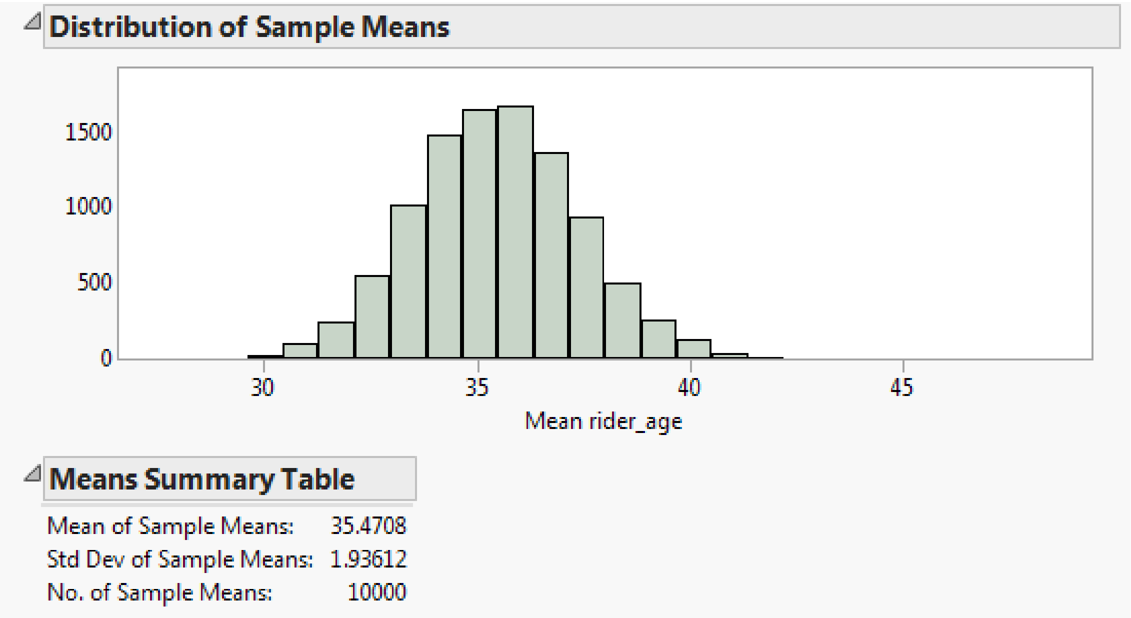
Summary Statistics

Mean	35.456326
Std Dev	10.999782
Std Err Mean	0.0585974
Upper 95% Mean	35.571178
Lower 95% Mean	35.341473
N	35238

The mean rider age is 35.46 years.

c. Using the CLT, we'd expect the sampling distribution of the sample mean to approach an approximately normal distribution as the sample size, *n*, grows large. The mean of the distribution should be 35.46 years with a standard error equal to approximately 11/(sqrt(n)).

e. Here are the results of **one** such simulation, rescaled for clarity:



Distribution of Sample Means

Means Summary Table

Mean of Sample Means:	35.4708
Std Dev of Sample Means:	1.93612
No. of Sample Means:	10000

The sampling distribution is symmetric and unimodal, with a mean at 35.47 years and a standard error of 1.936. Note that in part c the CLT would have predicted a mean of 35.46 and a standard error of $11/\sqrt{50} = 1.56$

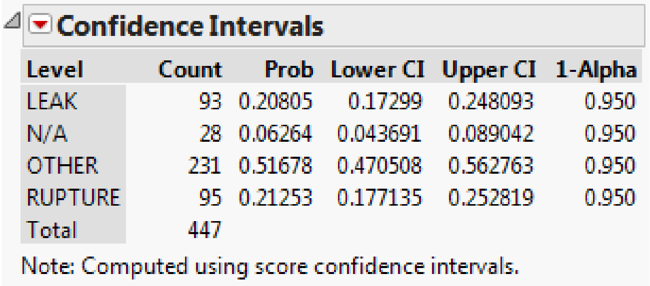
•

10

Student Solutions to Application Scenarios

Scenario 1

a.



Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
LEAK	93	0.20805	0.17299	0.248093	0.950
N/A	28	0.06264	0.043691	0.089042	0.950
OTHER	231	0.51678	0.470508	0.562763	0.950
RUPTURE	95	0.21253	0.177135	0.252819	0.950
Total	447				

Note: Computed using score confidence intervals.

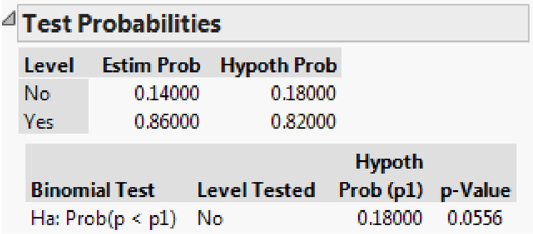
Based on the analysis shown to the left, 95 of 447 disruptions with known causes were ruptures. The estimated confidence interval is from 0.177 to 0.253. We can be 95% confident that the true population proportion is somewhere between 0.177 and 0.253.

c. When we lower the confidence level the interval becomes narrower.

Scenario 2

a. Yes. We have a random sample of sufficient size to invoke the Central Limit Theorem.

c.



Test Probabilities

Level	Estim Prob	Hypoth Prob
No	0.14000	0.18000
Yes	0.86000	0.82000

Binomial Test	Level Tested	Hypoth Prob (p1)	p-Value
Ha: Prob(p < p1)	No	0.18000	0.0556

With a p-Value of 0.0556, this sample falls just short of statistical significance. Assuming that we are using the standard 5% significance level, the sample does not quite provide sufficient evidence to conclude that the rate is currently below 18%.

e. A larger sample with the very same proportion provides more precision in the confidence interval (i.e. a narrower interval) and enhances the statistical significance of the test result.

Scenario 3

a. Yes. We have a random sample of sufficient size to invoke the Central Limit Theorem.

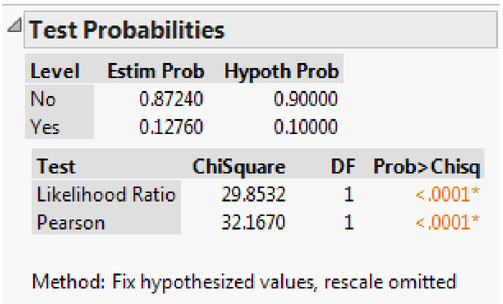
c. We can be 99% confident that the population proportion is between 0.071 and 0.085. Both intervals are centered at the same value, but the 99% interval is wider than the 95% interval.

e. The lower the confidence level, the narrower the interval.

Scenario 4

a. Yes. We have a random sample of sufficient size to invoke the Central Limit Theorem.

c.



Test Probabilities

Level	Estim Prob	Hypoth Prob
No	0.87240	0.90000
Yes	0.12760	0.10000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	29.8532	1	<.0001*
Pearson	32.1670	1	<.0001*

Method: Fix hypothesized values, rescale omitted

(For this question, it is simplest to create a small summary table). Create a Because of the question's wording, a two-tailed test is most appropriate here. Based on this random sample, we can confidently conclude that it is *not* credible to conclude that 10% of the population binge drinks at least once per week. If anything, this sample suggests a higher population proportion.

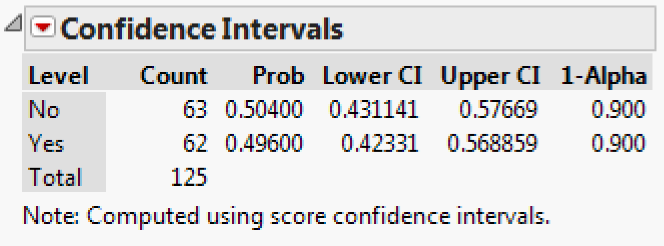
Scenario 5

a. It depends The total sample size is 189; because some events or combination of events are relatively rare, it may be the case that *np* <5, in which case we should not interpret the inferential results.

c. Although the observed relative frequency is 0.53, and thus greater than 0.5 the p-Value is 0.362 which is quite high enough that we can readily attribute the result to sampling error. In other words, a null hypothesis that the population proportion is 0.50 or less is still plausible, so we fail to reject the null.

Scenario 6

a.



Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
No	63	0.50400	0.431141	0.57669	0.900
Yes	62	0.49600	0.42331	0.568859	0.900
Total	125				

Note: Computed using score confidence intervals.

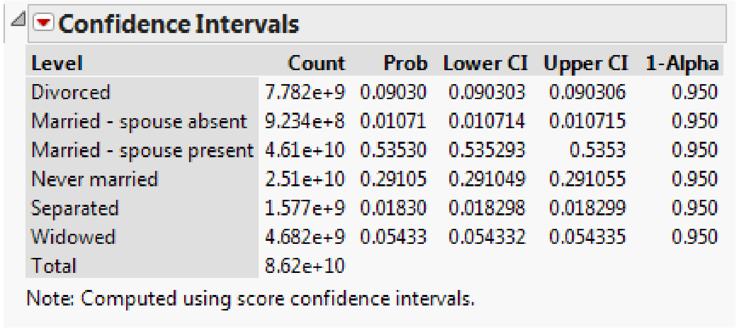
We can be 90% confident that the proportion of trading days on which McDonald's stock increases is somewhere between 0.423 and 0.569.

Scenario 7

a. Yes. We have very large samples, and can rely on the Central Limit Theorem.

c. The 99% confidence interval is (0.237 and 0..247). Like the prior interval, this interval is centered at 0.237, but is slightly wider.

e.



Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
Divorced	7.782e+9	0.09030	0.090303	0.090306	0.950
Married - spouse absent	9.234e+8	0.01071	0.010714	0.010715	0.950
Married - spouse present	4.61e+10	0.53530	0.535293	0.5353	0.950
Never married	2.51e+10	0.29105	0.291049	0.291055	0.950
Separated	1.577e+9	0.01830	0.018298	0.018299	0.950
Widowed	4.682e+9	0.05433	0.054332	0.054335	0.950
Total	8.62e+10				

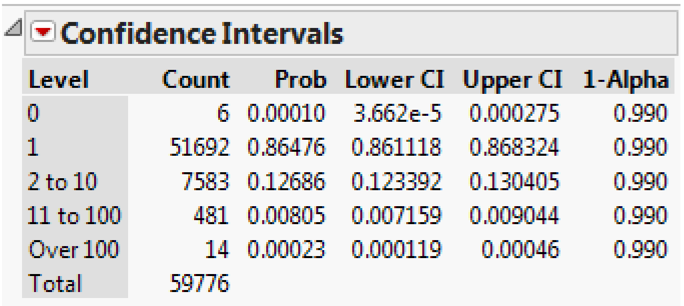
Note: Computed using score confidence intervals.

When we apply the sampling weights, the point estimate changes from 23.7% to 29.1%, and the 95% confidence interval is approximately 29.1049% to 29.1055% -- it shrinks dramatically in width, and is considerably higher than before.

Scenario 8

a. It depends on which variables we examine. We have a random sample of sufficient size to invoke the Central Limit Theorem, but there is a considerable amount of missing data.

c.



Confidence Intervals

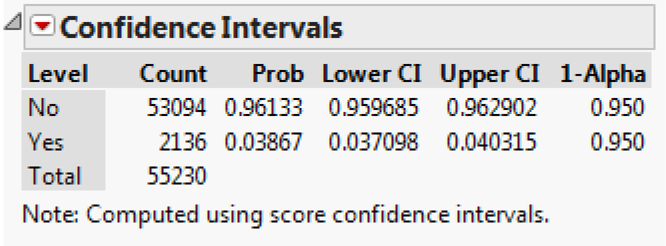
Level	Count	Prob	Lower CI	Upper CI	1-Alpha
0	6	0.00010	3.662e-5	0.000275	0.990
1	51692	0.86476	0.861118	0.868324	0.990
2 to 10	7583	0.12686	0.123392	0.130405	0.990
11 to 100	481	0.00805	0.007159	0.009044	0.990
Over 100	14	0.00023	0.000119	0.00046	0.990
Total	59776				

We can be 99% confident that, out of all instances where there is a bird strike, a single bird is struck somewhere between 86.1% and 86.8% of the time. The 99% CI is slightly wider than the 95% CI.

Scenario 9

a. Yes. We have a random sample of sufficient size to invoke the Central Limit Theorem, but there is a considerable amount of missing data.

c.This question is most easily done by creating a small summary table.



Level	Count	Prob	Lower CI	Upper CI	1-Alpha
No	53094	0.96133	0.959685	0.962902	0.950
Yes	2136	0.03867	0.037098	0.040315	0.950
Total	55230				

Note: Computed using score confidence intervals.

We can be 95% confident that between 3.9% and 4% start at the library.

e.The smaller sample makes the interval wider, but has no effect on the center of the interval.

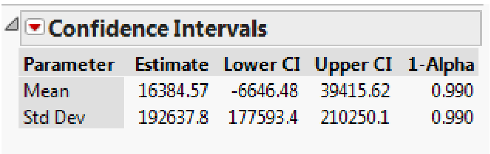
11

Student Solutions to Application Scenarios

Scenario 1

a. Probably. These columns contain continuous data, and though both distributions are strongly right-skewed, both have a sufficiently large number of observations to rely on the Central Limit Theorem. The critical question is whether we can view this particular time period as representative of the overall process of pipeline disruptions; if we can regard it as random, then we can proceed to make inferences.

c. The 90% interval is –$ 307,156 to $ 2,979,847. We can be 90% certain that the mean damage cost lies between these two values.

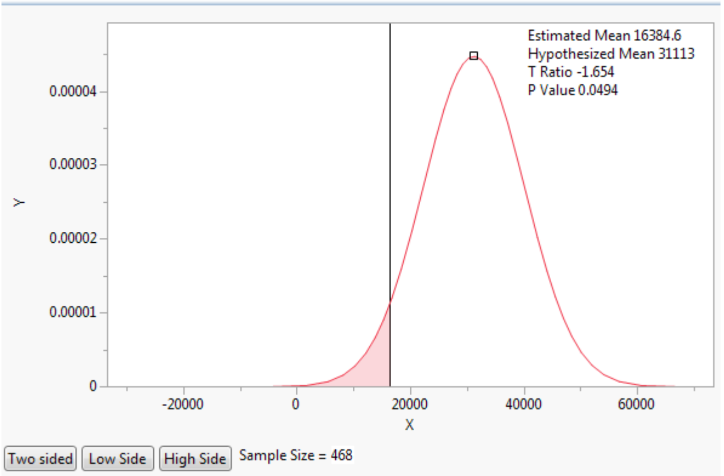
e. 

Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	16384.57	-6646.48	39415.62	0.990
Std Dev	192637.8	177593.4	210250.1	0.990

We can be 99% confident that the mean dollar cost of lost natural gas is between –$6646.48 and $39,415.62. NOTE: the distribution is so strongly right-skewed that we should be reluctant to draw conclusions from this sample, even with a sample size of 468.

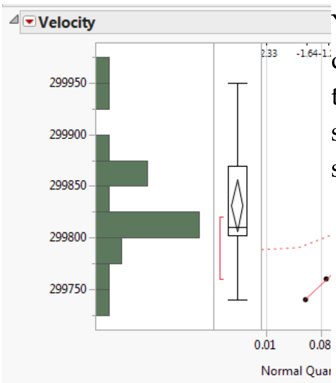
g.



Student answers will vary but should conclude that if the null hypothesis were that $\mu =$ approximately $ 31,100 then we would reject the null in favor of the one-sided alternative hypothesis.

Scenario 2

a.



Yes. We do not know the population σ so we will use the t-distribution. Because the sample is small (n = 20) we want to see if the sample data suggest that the population is roughly normal in shape. The histogram and normal quantile plots indicate mild skewness but no serious indication of non-normality.

c. From the confidence interval in part b we can see that Michelson would probably have (erroneously) concluded that the value 300,000 kps is not credible. The two-tailed hypothesis test yields a P-value < 0.0001 and a test statistic equal to –13.898; Michelson would have rejected a null hypothesis that the constant speed of light is 300,000 kps.

Scenario 3

a. Student answers will vary. On the one hand, because both measurements refer to the same child's height, we expect them to be quite similar. On the other hand, when a person stands the spine may compress slightly, so that standing height measurements may be less than reclining measurements.

Scenario 4

a. Yes. We do not know the population σ so we will use the t-distribution. Because the sample is so large (n = 1787) we can rely on the Central Limit Theorem to proceed.

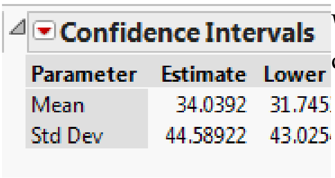
c. No. The interval is an estimate of the population *mean*, not the range of individual values. The interval provides an estimate of the location of the population mean acknowledging the uncertainty that arises from using a sample.

e. If the true population mean actually = 10 minutes the power of this test would be approximately 0.996. In other words, if the reality were that the mean flight is delayed 10 minutes, this test would detect that the mean is less than 12 minutes.

Scenario 5

a. Yes. We do not know the population σ so we will use the t-distribution. Because the sample is so large (n = 1455) we can rely on the Central Limit Theorem to proceed.

c.



Confidence Intervals

Parameter	Estimate	Lower
Mean	34.0392	31.745
Std Dev	44.58922	43.025

We can be 95% confident that the mean time from scheduled departure to wheels off is between 31.75 and 36.33 minutes.

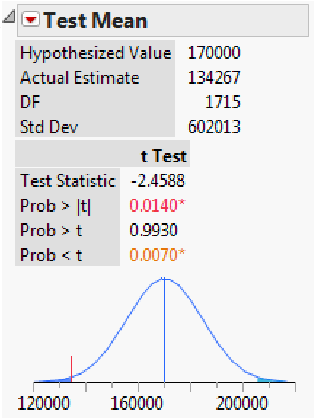
Scenario 6

a. The speed column does seem to satisfy the conditions: it is moderately symmetric and the sample is very large (n = 23,077) so we can rely on the Central Limit Theorem to proceed. We do not know the population σ so we will use the t-distribution.

The Cost of Repairs column is a smaller sample (n = 1716) and very strongly skewed. Even with the CLT, we should proceed with caution.

c. At the 99% confidence level, we can be 99% confident that the mean flight speed at impact is between 142.3 and 143.9 MPH.

e.



Test Mean

Hypothesized Value	170000
Actual Estimate	134267
DF	1715
Std Dev	602013

	t Test
Test Statistic	-2.4588
Prob > |t|	0.0140*
Prob > t	0.9930
Prob < t	0.0070*

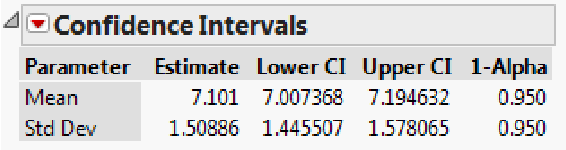
The test results indicate that the sample provides convincing evidence to reject the null hypothesis, yielding a very small P-value of just 0.007. The sample is, as noted, very right-skewed, but if anything that would overstate the population mean.

g. Student answers will vary, depending on which possible “True Mean” values they explore. It is useful to notice that the power of the test exceeds 90% for all true means below approximately $127,000.

Scenario 7

a. Yes. We do not know the population σ so we will use the t-distribution. The sample is large enough ($n = 1000$; some mothers’ gains are missing, $n = 973$) and the distributions are reasonably symmetric so we can rely on the Central Limit Theorem to proceed.

c.



Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	7.101	7.007368	7.194632	0.950
Std Dev	1.50886	1.445507	1.578065	0.950

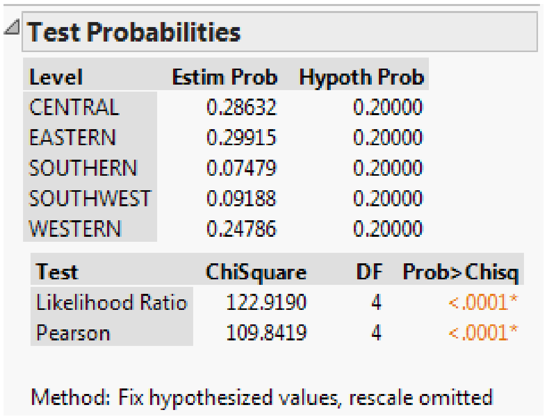
We can be 95% confident that the mean birthweight of infants in NC for the year 2004 was between 7 and 7.2 pounds.

12

Student Solutions to Application Scenarios

Scenario 1

a.



Test Probabilities

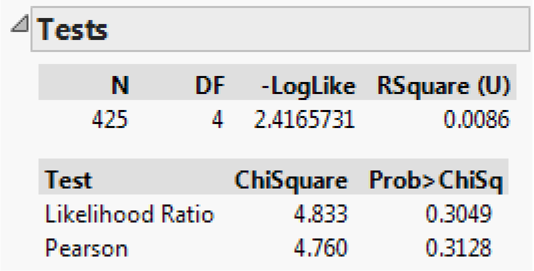
Level	Estim Prob	Hypoth Prob
CENTRAL	0.28632	0.20000
EASTERN	0.29915	0.20000
SOUTHERN	0.07479	0.20000
SOUTHWEST	0.09188	0.20000
WESTERN	0.24786	0.20000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	122.9190	4	<.0001*
Pearson	109.8419	4	<.0001*

Method: Fix hypothesized values, rescale omitted

No. At the 0.05 level of significance we reject the null hypothesis of equal probabilities.

c.



Tests

N	DF	-LogLike	RSquare (U)
425	4	2.4165731	0.0086

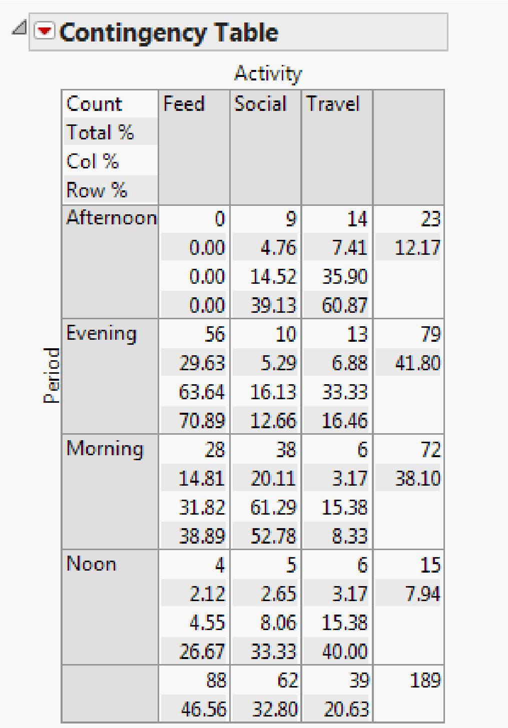
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	4.833	0.3049
Pearson	4.760	0.3128

Based on this sample, we would conclude that the variables are independent. We do not have sufficient evidence to conclude that the two variables are not independent (assuming a significance level of 0.05).

This set of Solutions for Students is a companion piece to the following SAS Press book: Carver, Robert. Practical Data Analysis with JMP®, Second Edition. Copyright © 2014, SAS Institute Inc., Cary, NC, USA. ALL RIGHTS RESERVED.

Scenario 2

a.

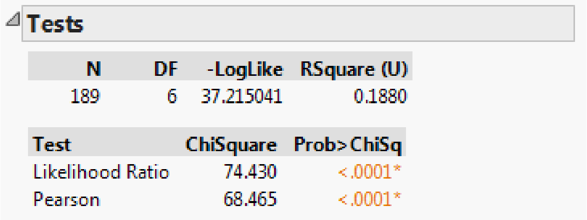


Contingency Table

Activity

Period

Count Total % Col % Row %	Feed	Social	Travel	
Afternoon	0 0.00 0.00 0.00	9 4.76 14.52 39.13	14 7.41 35.90 60.87	23 12.17
Evening	56 29.63 63.64 70.89	10 5.29 16.13 12.66	13 6.88 33.33 16.46	79 41.80
Morning	28 14.81 31.82 38.89	38 20.11 61.29 52.78	6 3.17 15.38 8.33	72 38.10
Noon	4 2.12 4.55 26.67	5 2.65 8.06 33.33	6 3.17 15.38 40.00	15 7.94
	88 46.56	62 32.80	39 20.63	189



Tests

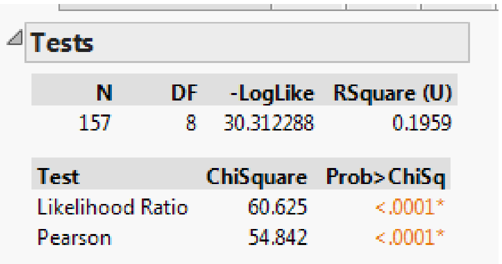
N	DF	-LogLike	RSquare (U)
189	6	37.215041	0.1880

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	74.430	<.0001*
Pearson	68.465	<.0001*

Because there are some cells with very small counts and expected counts, we should use caution making inferences from the ChiSquare test. However, we can note that the evidence points towards rejection of the null hypothesis of independence and we can also note (for example) that dolphins were regularly observed feeding in the morning and evening, but rarely if ever at other times.

Scenario 3

a.



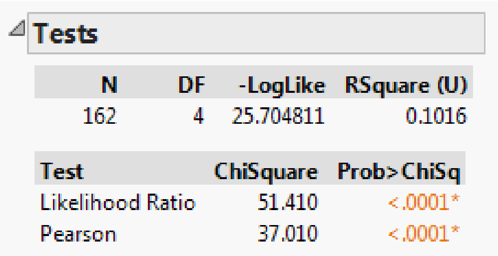
Tests

N	DF	-LogLike	RSquare (U)
157	8	30.312288	0.1959

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	60.625	<.0001*
Pearson	54.842	<.0001*

No. At the 0.05 level of significance we reject that null hypothesis that Provider and Region are independent.

c.



Tests

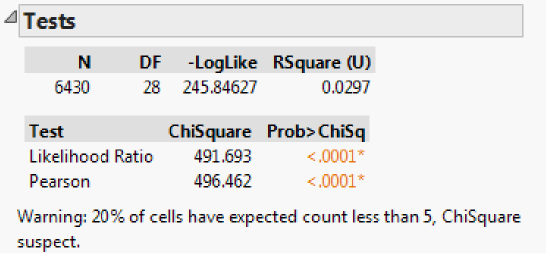
N	DF	-LogLike	RSquare (U)
162	4	25.704811	0.1016

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	51.410	<.0001*
Pearson	37.010	<.0001*

No. At the 0.05 level of significance we reject that null hypothesis that MatLeave90+ and Region are independent.

Scenario 4

a.



Tests

N	DF	-LogLike	RSquare (U)
6430	28	245.84627	0.0297

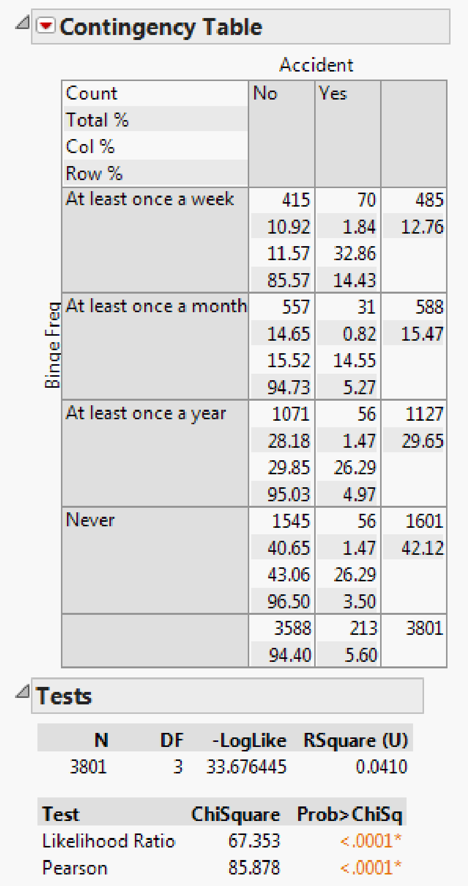
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	491.693	<.0001*
Pearson	496.462	<.0001*

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

Because there are a substantial proportion of cells with very small expected counts, we should use caution making inferences from the ChiSquare test. However, we can note that the evidence points toward rejecting the null hypothesis of independence. We might observe (for example) that married respondents were disproportionately non-Hispanic whites.

Scenario 5

a.



Contingency Table

		Accident		
	Count Total % Col % Row %	No	Yes	
Binge Freq	At least once a week	415 10.92 11.57 85.57	70 1.84 32.86 14.43	485 12.76
	At least once a month	557 14.65 15.52 94.73	31 0.82 14.55 5.27	588 15.47
	At least once a year	1071 28.18 29.85 95.03	56 1.47 26.29 4.97	1127 29.65
	Never	1545 40.65 43.06 96.50	56 1.47 26.29 3.50	1601 42.12
		3588 94.40	213 5.60	3801

Tests

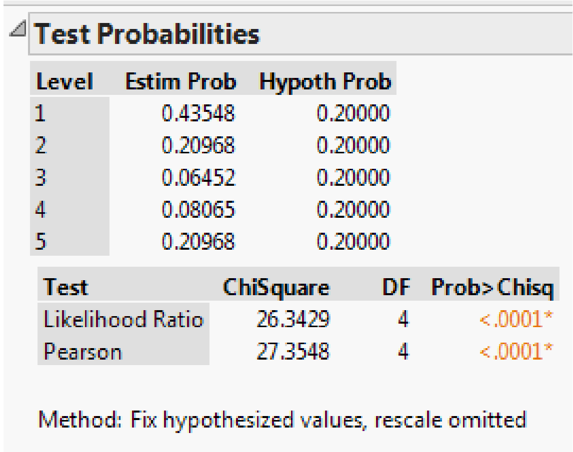
N	DF	-LogLike	RSquare (U)
3801	3	33.676445	0.0410

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	67.353	<.0001*
Pearson	85.878	<.0001*

No. At the 0.05 level of significance we reject that null hypothesis that binge drinking regularity and involvement in car accidents are independent. Students who report binging at least once a week are far more likely to have been involved in an accident than other students.

Scenario 6

a.



Test Probabilities

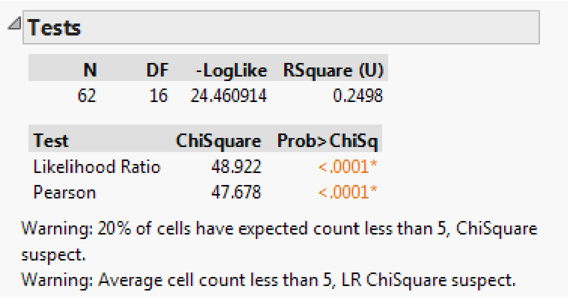
Level	Estim Prob	Hypoth Prob
1	0.43548	0.20000
2	0.20968	0.20000
3	0.06452	0.20000
4	0.08065	0.20000
5	0.20968	0.20000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	26.3429	4	<.0001*
Pearson	27.3548	4	<.0001*

Method: Fix hypothesized values, rescale omitted

The Chi-Square goodness-of-fit test indicates that the five categories are not equally distributed across mammalian species. We reject the null hypothesis that all proportions are equal at 0.20.

c.



Tests

N	DF	-LogLike	RSquare (U)
62	16	24.460914	0.2498

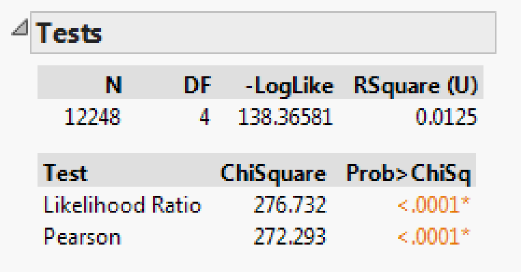
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	48.922	<.0001*
Pearson	47.678	<.0001*

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.
Warning: Average cell count less than 5, LR ChiSquare suspect.

The total sample size here leads to many cells with expected counts < 5, makng the Chi-Square test unreliable. That said, the test results point in the direction of rejecting the null hypothesis.

Scenario 7

a.



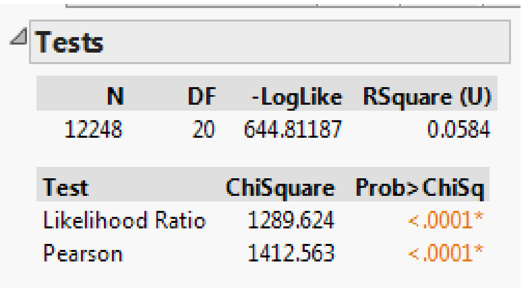
Tests

N	DF	-LogLike	RSquare (U)
12248	4	138.36581	0.0125

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	276.732	<.0001*
Pearson	272.293	<.0001*

According to the Chi-Square test the two variables are not independent. There is sufficient evidence to reject a null hypothesis that they are independent.

c.



Tests

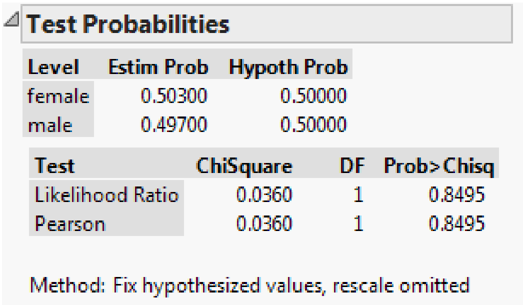
N	DF	-LogLike	RSquare (U)
12248	20	644.81187	0.0584

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	1289.624	<.0001*
Pearson	1412.563	<.0001*

According to the Chi-Square test the two variables are not independent. There is sufficient evidence to reject a null hypothesis that they are independent.

Scenario 8

a.



Test Probabilities

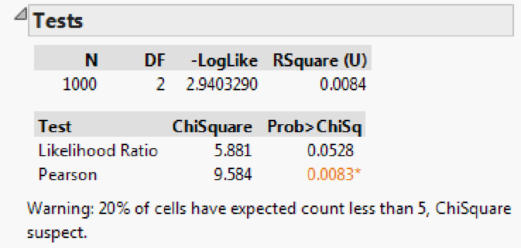
Level	Estim Prob	Hypoth Prob
female	0.50300	0.50000
male	0.49700	0.50000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	0.0360	1	0.8495
Pearson	0.0360	1	0.8495

Method: Fix hypothesized values, rescale omitted

According to the Chi-Square test there is not sufficient evidence to reject a null hypothesis that mothers are equally likely to give birth to a male as a female baby.

c.



Tests

N	DF	-LogLike	RSquare (U)
1000	2	2.9403290	0.0084

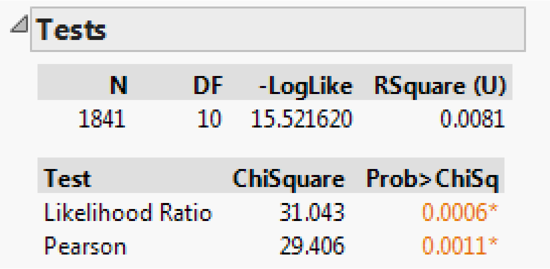
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	5.881	0.0528
Pearson	9.584	0.0083*

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

We should be reluctant to draw inferences about this question because of the high number of cells with counts less than 5. That said, Pearson's test does indicate sufficient evidence to reject a null hypothesis that they are independent. It would be wise to obtain a larger sample before drawing a conclusion.

Scenario 9

a.



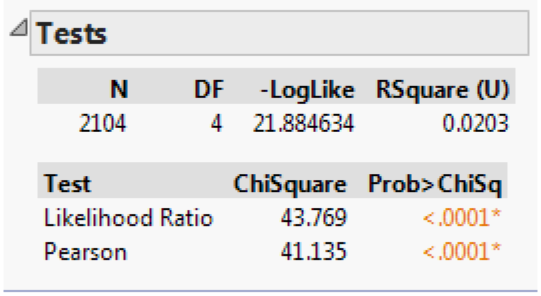
Tests

N	DF	-LogLike	RSquare (U)
1841	10	15.521620	0.0081

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	31.043	0.0006*
Pearson	29.406	0.0011*

According to the Chi-Square test the two variables are not independent. There is sufficient evidence to reject a null hypothesis that they are independent. The distribution of phase of flight is different at different airports.

c.



Tests

N	DF	-LogLike	RSquare (U)
2104	4	21.884634	0.0203

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	43.769	<.0001*
Pearson	41.135	<.0001*

According to the Chi-Square test the number of birds struck per incident does vary by airport. There is sufficient evidence to reject a null hypothesis that they are independent.

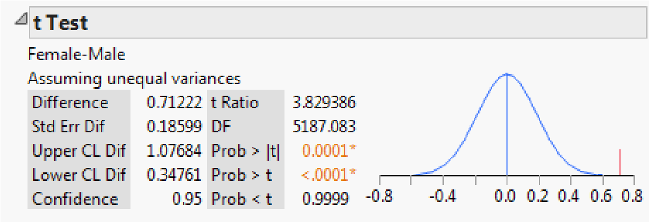
13

Student Solutions to Application Scenarios

Scenario 1

NOTE: Complete answers should note that we have continuous data, independent samples, and that the samples in each part of the question are large enough to rely on the Central Limit Theorem.

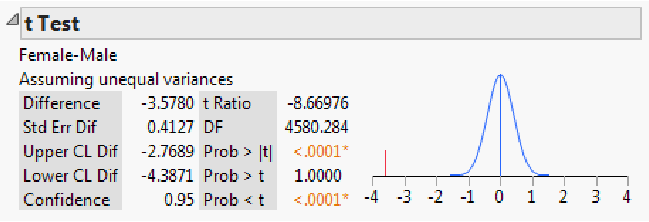
a.



```
t Test
Female-Male
Assuming unequal variances
Difference     0.71222  t Ratio    3.829386
Std Err Dif    0.18599  DF         5187.083
Upper CL Dif   1.07684  Prob > |t|  0.0001*
Lower CL Dif   0.34761  Prob > t    <.0001*
Confidence        0.95  Prob < t     0.9999
```

We can be 95% confident that the mean difference in Body Mass Index between men and women is between .34761 and 1.07684.

c.



```
t Test
Female-Male
Assuming unequal variances
Difference     -3.5780  t Ratio    -8.66976
Std Err Dif     0.4127  DF         4580.284
Upper CL Dif   -2.7689  Prob > |t|  <.0001*
Lower CL Dif   -4.3871  Prob > t     1.0000
Confidence        0.95  Prob < t    <.0001*
```

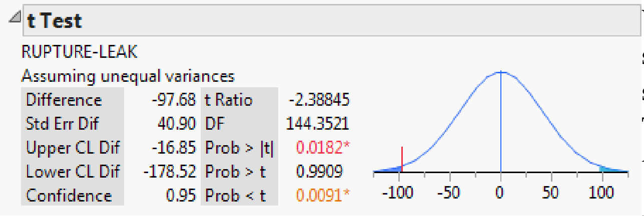
We can be 95% confident that the mean difference in Diastolic Blood Pressure between men and women is between –4.387 and –2.7689.

Scenario 2

a. We should first note that we have modest sample sizes (n= 35 and n=43) from strongly skewed distributions. Therefore, we should be reluctant to interpret the resulting interval at all. However, the reported 95% confidence interval is from –$11,026,606 to +$32,748,087.

Scenario 3

a.

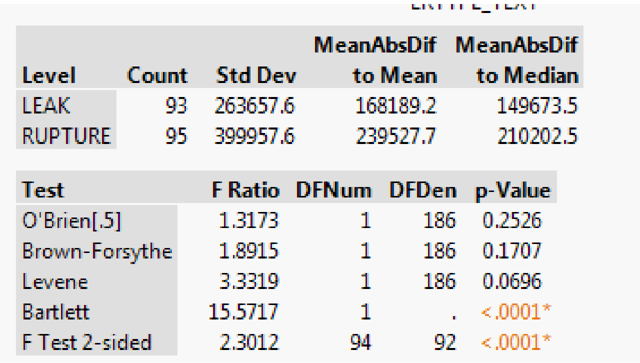


```
t Test
RUPTURE-LEAK
Assuming unequal variances
Difference      -97.68  t Ratio    -2.38845
Std Err Dif      40.90  DF         144.3521
Upper CL Dif    -16.85  Prob > |t|  0.0182*
Lower CL Dif   -178.52  Prob > t    0.9909
Confidence        0.95  Prob < t    0.0091*
```

We should first note that we have strongly skewed distributions but the sample sizes are reasonably large. Therefore, we can proceed to interpret the results of a t-test.

In this test, there is compelling evidence to suggest that it does not take longer to secure the area after a rupture than after a leak; to the contrary, leaks require more time.

c.



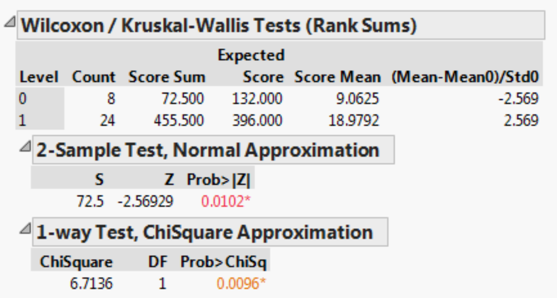
```
                               MeanAbsDif  MeanAbsDif
Level      Count   Std Dev       to Mean   to Median
LEAK          93  263657.6      168189.2    149673.5
RUPTURE       95  399957.6      239527.7    210202.5

Test              F Ratio  DFNum  DFDen  p-Value
O'Brien[.5]        1.3173      1    186   0.2526
Brown-Forsythe     1.8915      1    186   0.1707
Levene             3.3319      1    186   0.0696
Bartlett          15.5717      1      .   <.0001*
F Test 2-sided     2.3012     94     92   <.0001*
```

In this case the different tests of homogeneity of variance lead to different conclusions. Using Levene's test, we would fail to reject the null hypothesis of equal variances; with F Test 2-sided, we would reject the null and conclude that the variances are unequal. Given the ambiguity, it is safer to conclude that the variances are unequal when conducting the tests of means (above).

Scenario 4

a. Student answers will differ. We have only 8 individuals without PD, and for the baseline pitch and jitter, the distributions appear bimodal with few observations in the "center"; shimmer may be normally distributed for non-PD observations. Among individuals with PD ($n = 24$) the distributions tend to be skewed. As such, with non-normal distributions and small samples, this sample does not satisfy the conditions for the use of the t-test.

c.



Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
0	8	72.500	132.000	9.0625	-2.569
1	24	455.500	396.000	18.9792	2.569

2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
72.5	-2.56929	0.0102*

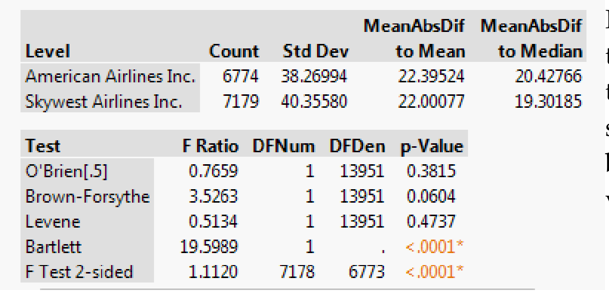
1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
6.7136	1	0.0096*

Based on the Wilcoxon test (assuming a significance level of $\alpha = 0.05$) we reject the null hypothesis that the mean jitter measurement is equal for both groups. There is a statistically significant difference in this sample data.

Scenario 5

a.



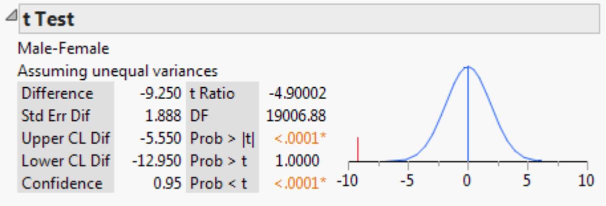
Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
American Airlines Inc.	6774	38.26994	22.39524	20.42766
Skywest Airlines Inc.	7179	40.35580	22.00077	19.30185

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	0.7659	1	13951	0.3815
Brown-Forsythe	3.5263	1	13951	0.0604
Levene	0.5134	1	13951	0.4737
Bartlett	19.5989	1	.	<.0001*
F Test 2-sided	1.1120	7178	6773	<.0001*

If we rely on Levene's test, we conclude that there is insufficient evidence to conclude that the variances are different; the F Test 2-sided leads to the opposite conclusion. To be safe we'll use the t-test assuming unequal variances for the next question.

Scenario 6

a.



t Test

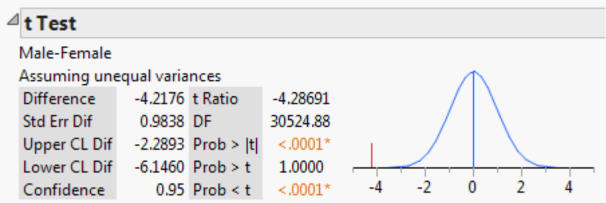
Male-Female

Assuming unequal variances

Difference	-9.250	t Ratio	-4.90002
Std Err Dif	1.888	DF	19006.88
Upper CL Dif	-5.550	Prob > |t|	<.0001*
Lower CL Dif	-12.950	Prob > t	1.0000
Confidence	0.95	Prob < t	<.0001*

Using just the 2003 data, we estimate with 95% confidence that females reported sleeping between 5.55 and 12.95 minutes more than males.

c.



t Test

Male-Female

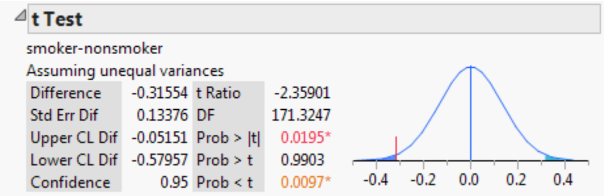
Assuming unequal variances

Difference	-4.2176	t Ratio	-4.28691
Std Err Dif	0.9838	DF	30524.88
Upper CL Dif	-2.2893	Prob > |t|	<.0001*
Lower CL Dif	-6.1460	Prob > t	1.0000
Confidence	0.95	Prob < t	<.0001*

Combining all of the data from both years, we can conclude with 95% confidence that men spend, on average, 2.3 to 6.1 fewer minutes per day socializing than do women.

Scenario 7

a.



t Test

smoker-nonsmoker

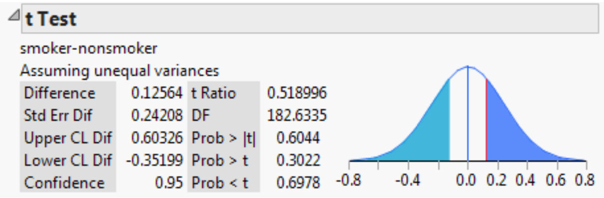
Assuming unequal variances

Difference	-0.31554	t Ratio	-2.35901
Std Err Dif	0.13376	DF	171.3247
Upper CL Dif	-0.05151	Prob > |t|	0.0195*
Lower CL Dif	-0.57957	Prob > t	0.9903
Confidence	0.95	Prob < t	0.0097*

Comment: Smoking status is missing (NA) for one respondent—filter out that case in order to compare the means of smokers and non-smokers.

The data provide sufficient evidence to reject the hypothesis of equal birth weights, and conclude that smokers have lower birthweight babies than non-smokers.

c.



t Test

smoker-nonsmoker

Assuming unequal variances

Difference	0.12564	t Ratio	0.518996
Std Err Dif	0.24208	DF	182.6335
Upper CL Dif	0.60326	Prob > |t|	0.6044
Lower CL Dif	-0.35199	Prob > t	0.3022
Confidence	0.95	Prob < t	0.6978

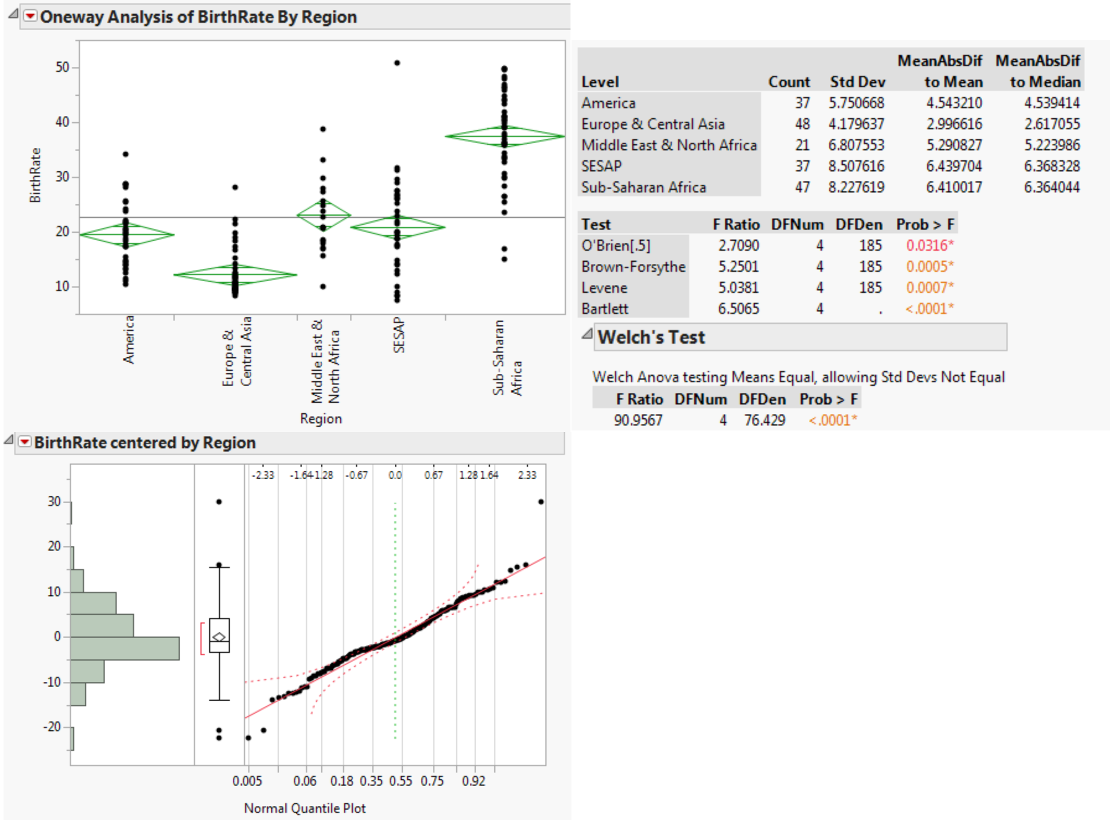
The data do not provide sufficient evidence to reject the hypothesis of equal number of weeks at delivery. We cannot conclude that there is any difference in the length of pregnancy between the two groups.

14

Student Solutions to Application Scenarios

Scenario 1

a.



Oneway Analysis of BirthRate By Region

Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
America	37	5.750668	4.543210	4.539414
Europe & Central Asia	48	4.179637	2.996616	2.617055
Middle East & North Africa	21	6.807553	5.290827	5.223986
SESAP	37	8.507616	6.439704	6.368328
Sub-Saharan Africa	47	8.227619	6.410017	6.364044

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	2.7090	4	185	0.0316*
Brown-Forsythe	5.2501	4	185	0.0005*
Levene	5.0381	4	185	0.0007*
Bartlett	6.5065	4	.	<.0001*

Welch's Test

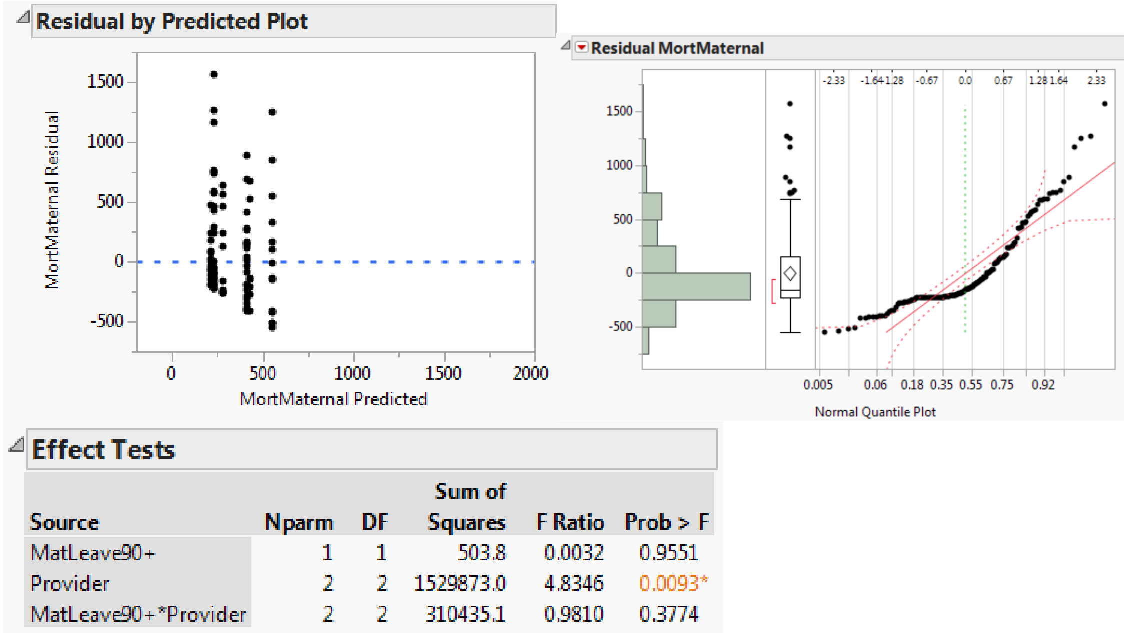
Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
90.9567	4	76.429	<.0001*

BirthRate centered by Region

In this case we find that the regional variances are not equal but the residuals do appear to be approximately normal. According to Welch's test, the mean birthrate is not equal across the regions of the world. Strictly speaking we cannot rely on a formal test to determine which regions differ. Visual inspection of the means diamonds in the Oneway graph suggests that SubSaharan birth rates are unusually high, and that birth rates in Europe and Central Asia are unusually low.

c.



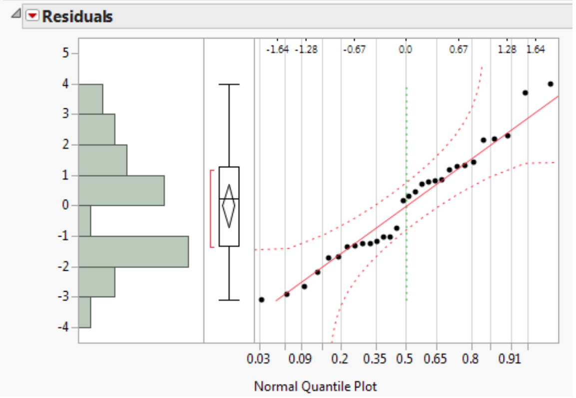
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
MatLeave90+	1	1	503.8	0.0032	0.9551
Provider	2	2	1529873.0	4.8346	0.0093*
MatLeave90+*Provider	2	2	310435.1	0.9810	0.3774

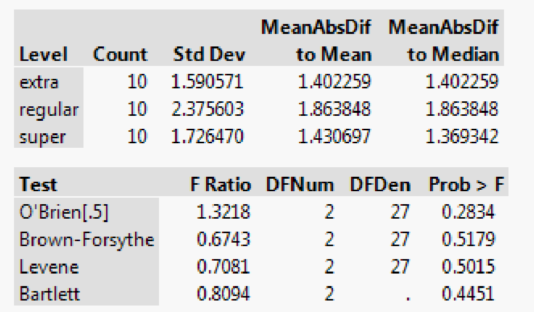
We start by evaluating conditions. The Residual by Predicted Plot raises some question about the equality of variances, but it is not definitive. The residuals do not appear to be normally distributed, but we have reasonably large samples and can rely on the Central Limit Theorem.

We find no significant interaction term, and we do find a significant main effect associated with the Provider of benefits. It appears that countries with private provision of maternity benefits have significantly higher rates of maternal mortality.

Scenario 2

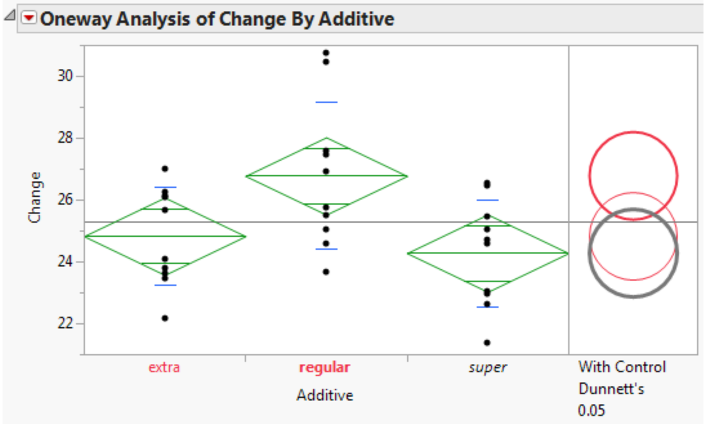
a. We see no evidence that the ANOVA assumptions have been violated; variances across the three groups appear to be equal and residuals are approximately normal. The F Ratio of 4.6275 and corresponding P-value of 0.0187 indicate that we should reject the null hypothesis of equal means; there is compelling evidence that the different additives lead to different mean changes.

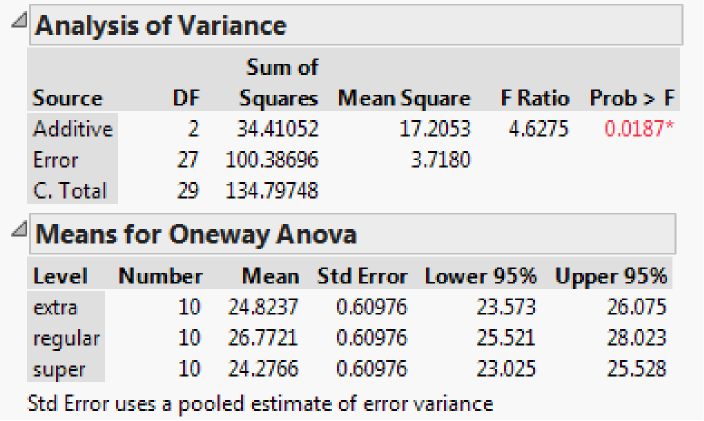




Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
extra	10	1.590571	1.402259	1.402259
regular	10	2.375603	1.863848	1.863848
super	10	1.726470	1.430697	1.369342

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	1.3218	2	27	0.2834
Brown-Forsythe	0.6743	2	27	0.5179
Levene	0.7081	2	27	0.5015
Bartlett	0.8094	2	.	0.4451





Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Additive	2	34.41052	17.2053	4.6275	0.0187*
Error	27	100.38696	3.7180		
C. Total	29	134.79748			

Means for Oneway Anova

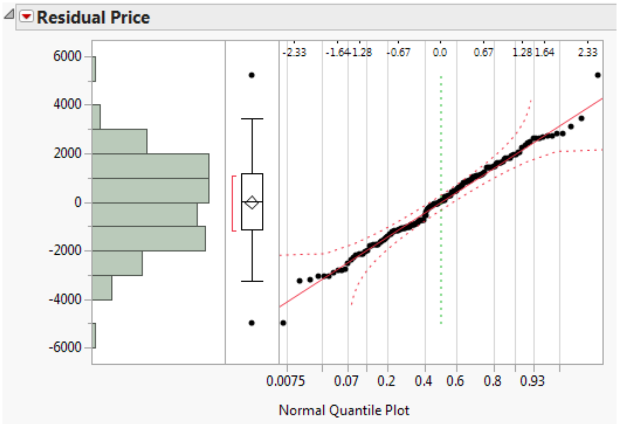
Level	Number	Mean	Std Error	Lower 95%	Upper 95%
extra	10	24.8237	0.60976	23.573	26.075
regular	10	26.7721	0.60976	25.521	28.023
super	10	24.2766	0.60976	23.025	25.528

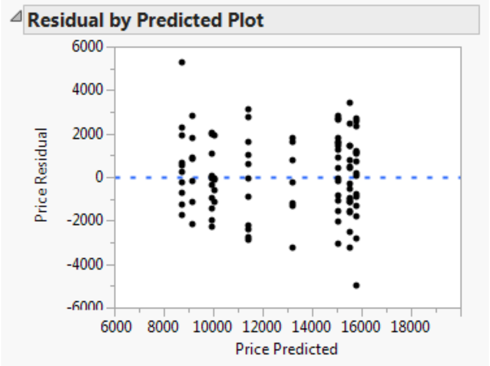
Std Error uses a pooled estimate of error variance

c. We find that there is a significant improvement in insulation with the "super" additive—the temperature change is smallest with that additive. The company should switch from regular to super.

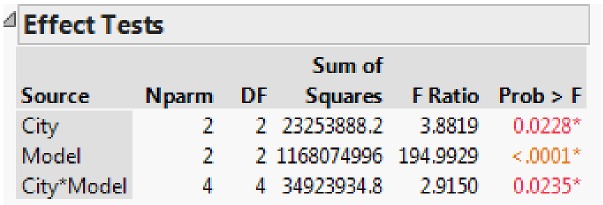
Scenario 3

a.





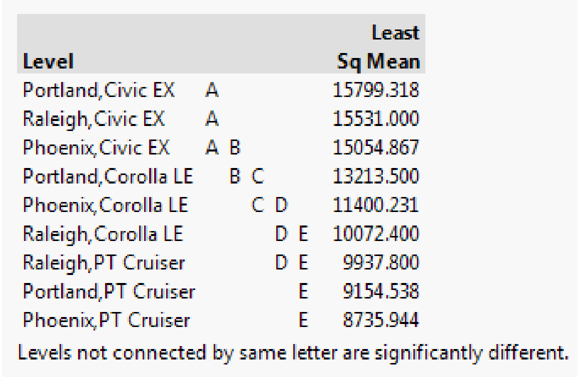
We start by evaluating conditions, and find no signs that the sample data violate the conditions for inference.



Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
City	2	2	23253888.2	3.8819	0.0228*
Model	2	2	1168074996	194.9929	<.0001*
City*Model	4	4	34923934.8	2.9150	0.0235*

A review of the Effect Tests shows that we have a significant interaction effect as well as significant main effects. This tells us that prices vary by city and by model, and what's more the impact of model varies across the cities.



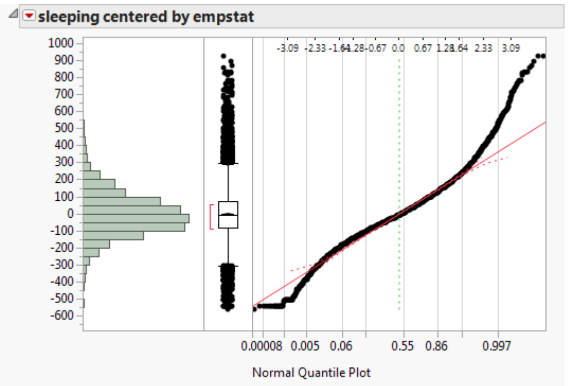
Level						Least Sq Mean
Portland,Civic EX	A					15799.318
Raleigh,Civic EX	A					15531.000
Phoenix,Civic EX	A	B				15054.867
Portland,Corolla LE		B	C			13213.500
Phoenix,Corolla LE			C	D		11400.231
Raleigh,Corolla LE				D	E	10072.400
Raleigh,PT Cruiser				D	E	9937.800
Portland,PT Cruiser					E	9154.538
Phoenix,PT Cruiser					E	8735.944

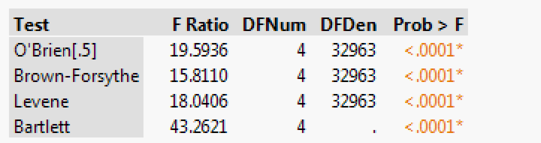
Levels not connected by same letter are significantly different.

When we apply Tukey's HSD (output not shown fully here) we see the complexity of the interactions; we should not make statements about main effects but can use the connecting letters report to identify differences among the model-city combinations.

Scenario 4

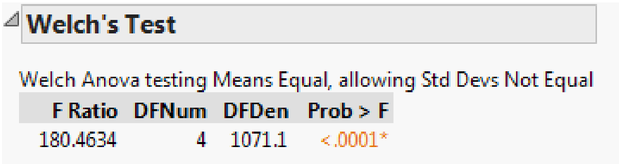
a.





Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	19.5936	4	32963	<.0001*
Brown-Forsythe	15.8110	4	32963	<.0001*
Levene	18.0406	4	32963	<.0001*
Bartlett	43.2621	4	.	<.0001*

As usual we start by evaluating assumptions. We have a very large sample, so the Central Limit Theorem applies and we need not be concerned with normality (above we see the residuals are unimodal and symmetric, but depart from the normal model in the tails). We also see evidence that the variances are unequal. In practice, because of the very large sample it is not surprising that we find significant differences.



Welch's Test

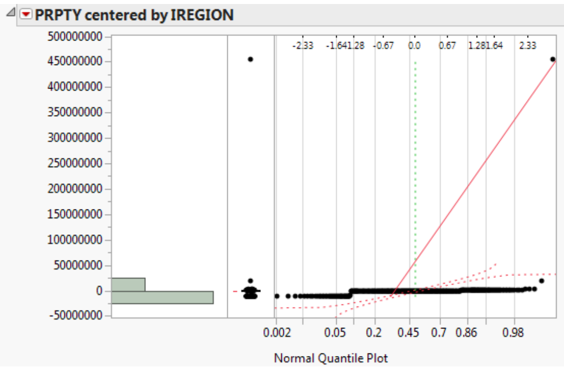
Welch Anova testing Means Equal, allowing Std Devs Not Equal

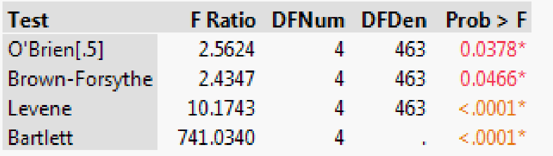
F Ratio	DFNum	DFDen	Prob > F
180.4634	4	1071.1	<.0001*

Both Welch's test and the standard ANOVA results strongly indicate that there are significant differences in group means. There is no control group here. Tukey's HSD indicates that employed people at work get the least sleep and unemployed people who are looking report the most. All others are significantly different from those two groups, but indistinguishable from one another.

Scenario 5

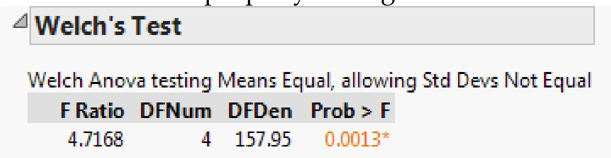
a.





Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	2.5624	4	463	0.0378*
Brown-Forsythe	2.4347	4	463	0.0466*
Levene	10.1743	4	463	<.0001*
Bartlett	741.0340	4	.	<.0001*

As we can see from the output, the sample data seem to violate the assumptions of normality and equal variance. Each of the regional subsamples is large enough to rely on the Central Limit Theorem with respect to normality. Using Welch's test (below) we would conclude that the mean costs of property damage are not identical across the regions.

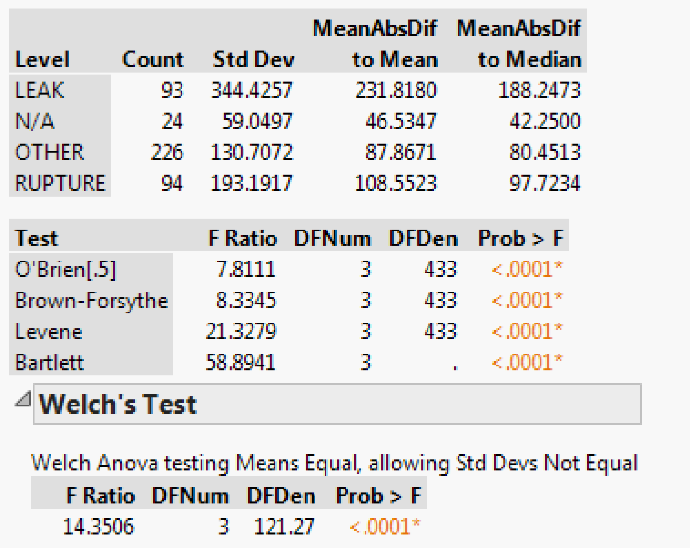


Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
4.7168	4	157.95	0.0013*

c. The distribution of residuals (not shown here) raises questions about normality and the usual tests indicate that the variances of the different disruption-type subgroups are unequal. According to Welch's test, there is at least one disruption type that differs from the others in terms of time required to make the area safe.



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
LEAK	93	344.4257	231.8180	188.2473
N/A	24	59.0497	46.5347	42.2500
OTHER	226	130.7072	87.8671	80.4513
RUPTURE	94	193.1917	108.5523	97.7234

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	7.8111	3	433	<.0001*
Brown-Forsythe	8.3345	3	433	<.0001*
Levene	21.3279	3	433	<.0001*
Bartlett	58.8941	3	.	<.0001*

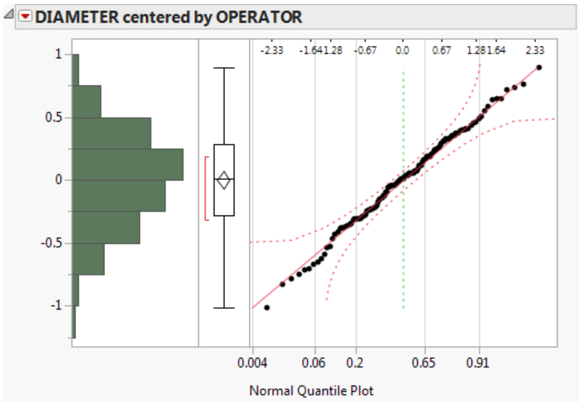
Welch's Test

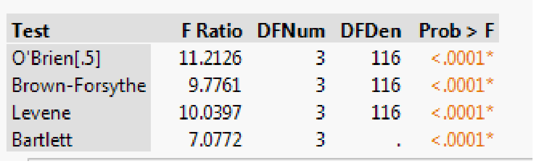
Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
14.3506	3	121.27	<.0001*

Scenario 6

a.



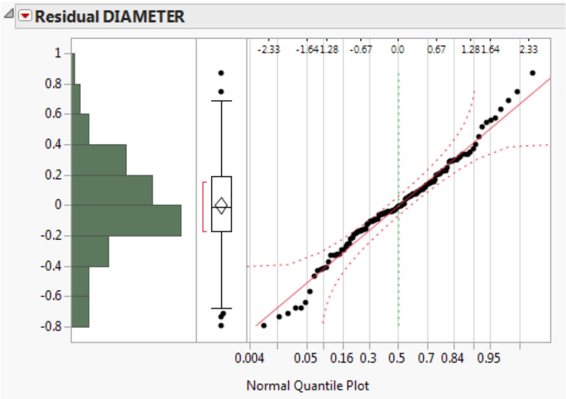


Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	11.2126	3	116	<.0001*
Brown-Forsythe	9.7761	3	116	<.0001*
Levene	10.0397	3	116	<.0001*
Bartlett	7.0772	3	.	<.0001*

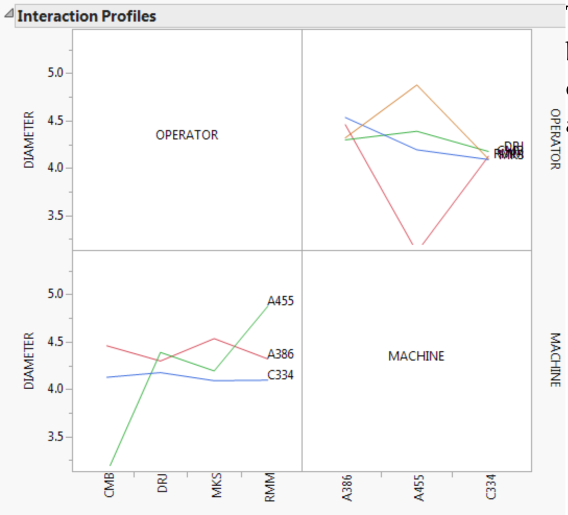
We start by examining assumptions. The residuals appear to be normally distributed (the sample sizes are large enough to rely on the Central Limit Theorem in this case), but the subsamples appear not to share a common variance.

Both Welch's test and the conventional ANOVA find no significant differences among group means.

c.



The assumption of normality does appear to be satisfied; visual inspection of residuals vs. predicted values does not reveal any obvious differences in group variances.



The interaction plots indicate interaction effects between operator and machine, making it difficult to interpret the main effects of machine and operator separately.

Scenario 7: NOTE-- Due to the large amount of output required in this problem, only a few selected results are shown.

a. We begin by checking the normality and equal variance assumptions. These are particularly important with such a small sample.

Among the three analyses, we find that there are unequal variances for the analysis of Yield by Popcorn type and Yield by batch.
The residuals in the analysis by Batch appear to be approximately normal, but the others do not. Due to the non-normality, we should use the Wilcoxon approach for the other two.

Yield vs. Batch satisfies normality, but not Equal Variance. Hence, we should consult Welch's test for Yield vs. batch.

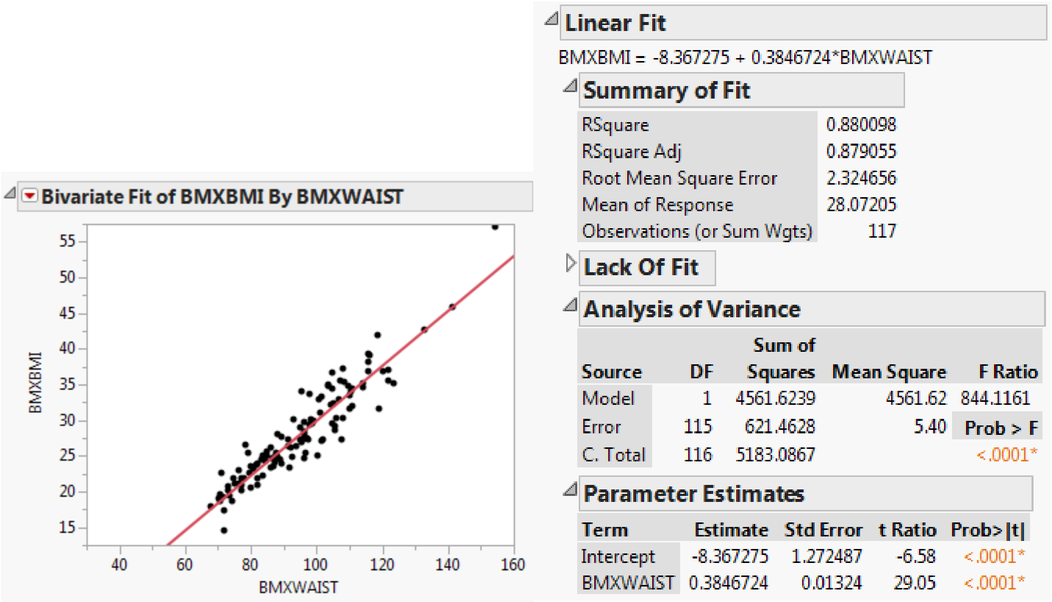
There are no significant main effects on yield for either Popcorn or Oil Amt.
However, the Welch's test result does indicate a significant effect of batch size (small batches improve yield).

15

Student Solutions to Application Scenarios

Scenario 1

a.



Linear Fit

BMXBMI = -8.367275 + 0.3846724*BMXWAIST

Summary of Fit

RSquare	0.880098
RSquare Adj	0.879055
Root Mean Square Error	2.324656
Mean of Response	28.07205
Observations (or Sum Wgts)	117

Lack Of Fit

Analysis of Variance

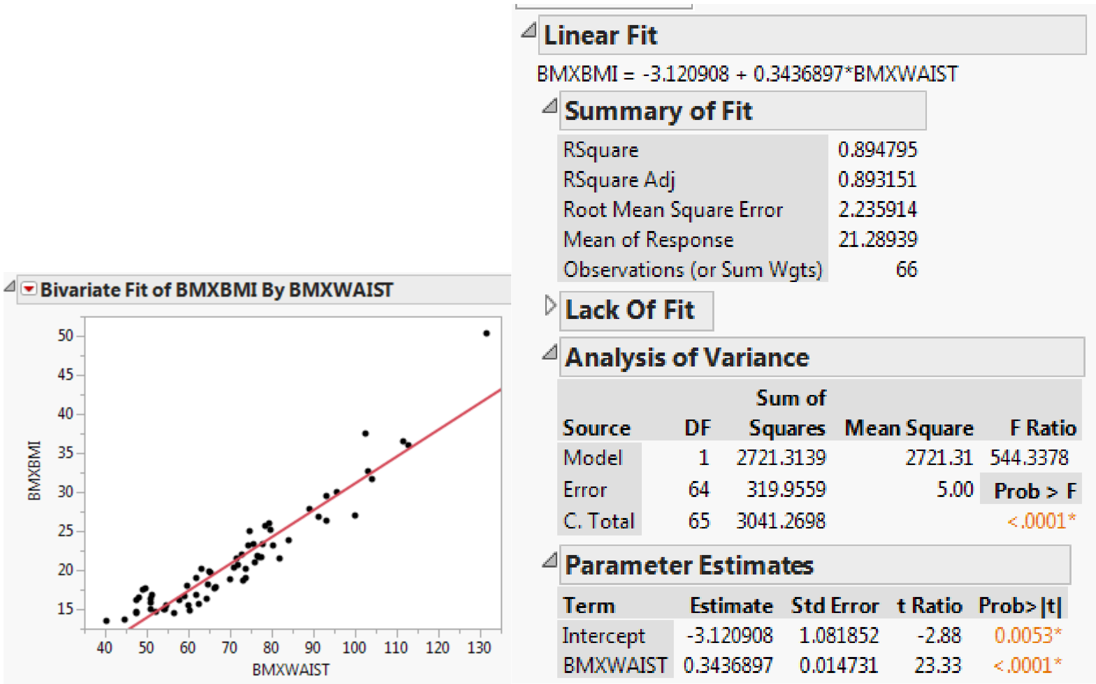
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	4561.6239	4561.62	844.1161
Error	115	621.4628	5.40	Prob > F
C. Total	116	5183.0867		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-8.367275	1.272487	-6.58	<.0001*
BMXWAIST	0.3846724	0.01324	29.05	<.0001*

Above are the regression results for adult females. We find a significant relationship between waist circumference and BMI, with the waist measurement accounting for about 88% of the variation in BMI. Each addition centimeter of waist circumference is associated with an increase of 0.3847 in BMI.

c.



Linear Fit

BMXBMI = -3.120908 + 0.3436897*BMXWAIST

Summary of Fit

RSquare	0.894795
RSquare Adj	0.893151
Root Mean Square Error	2.235914
Mean of Response	21.28939
Observations (or Sum Wgts)	66

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	2721.3139	2721.31	544.3378
Error	64	319.9559	5.00	Prob > F
C. Total	65	3041.2698		<.0001*

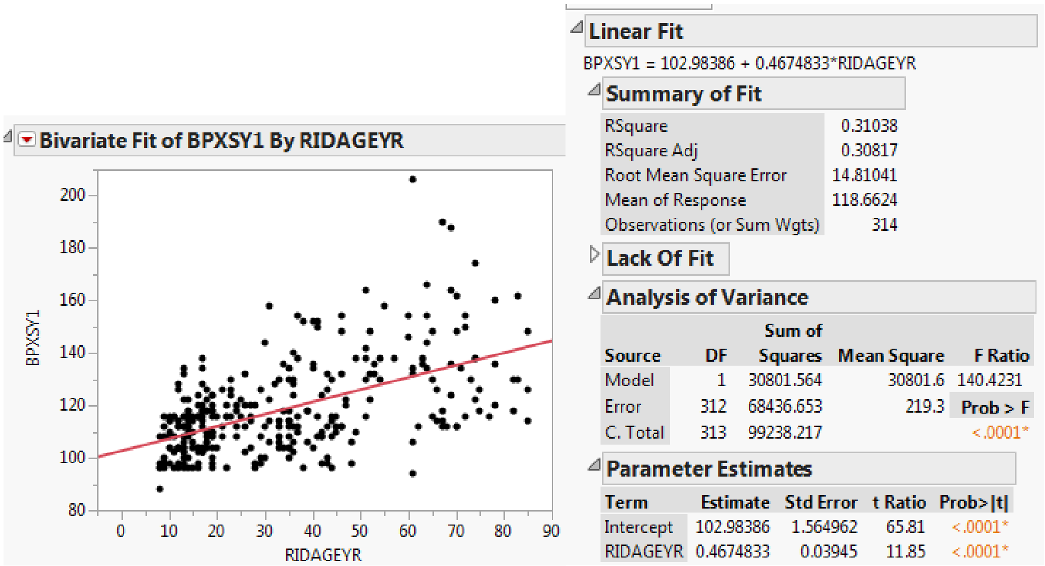
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-3.120908	1.081852	-2.88	0.0053*
BMXWAIST	0.3436897	0.014731	23.33	<.0001*

If we restrict the analysis to females under the age of 17 we find a slightly stronger relationship between Waist and BMI. The estimated slope is slightly smaller than before (0.344 vs. 0.385) but otherwise the regression models are very similar.

Scenario 2

a.



Linear Fit

BPXSY1 = 102.98386 + 0.4674833*RIDAGEYR

Summary of Fit

RSquare	0.31038
RSquare Adj	0.30817
Root Mean Square Error	14.81041
Mean of Response	118.6624
Observations (or Sum Wgts)	314

Lack Of Fit

Analysis of Variance

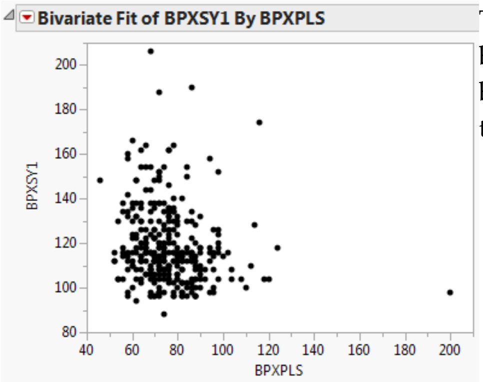
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	30801.564	30801.6	140.4231
Error	312	68436.653	219.3	Prob > F
C. Total	313	99238.217		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	102.98386	1.564962	65.81	<.0001*
RIDAGEYR	0.4674833	0.03945	11.85	<.0001*

In this regression we find a weak ($R^2 = 0.31$) but highly significant positive relationship. Subjects who differ in age by 1 year tend to have, on average, systolic BP that is approximately 0.47 points higher per year. This is not a strong relationship because age accounts for less than one-third of the variation in systolic BP.

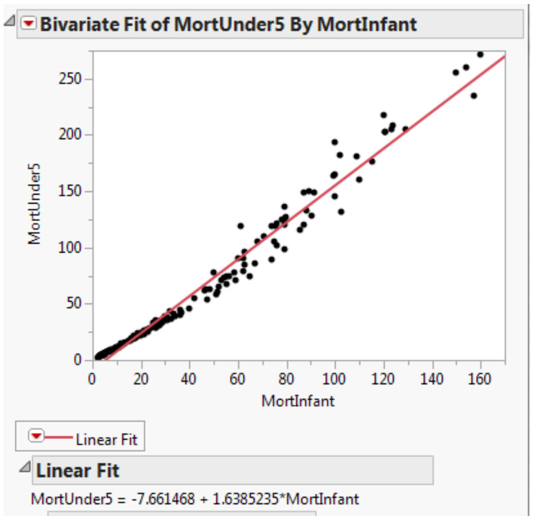
c.



The scatterplot to the left shows little or no relationship between pulse and systolic BP. If anything, there may be a very weak negative relationship here, contrary to the suspicion expressed in the question.

Scenario 3

a.



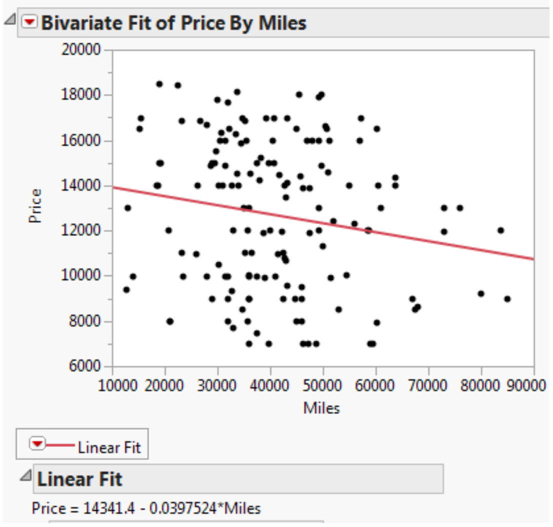
```
MortUnder5 = -7.661468 + 1.6385235*MortInfant
```

The estimated equation is appears beneath the graph, with $R^2 = 0.979$ – indicating a very strong relationship and excellent fit.

Despite the strong summary statistics, the scatterplot very clearly indicates some doubt about the linear model: the points seem to bend around the line, suggesting that the relationship is not best described as a line.

Scenario 4

a.



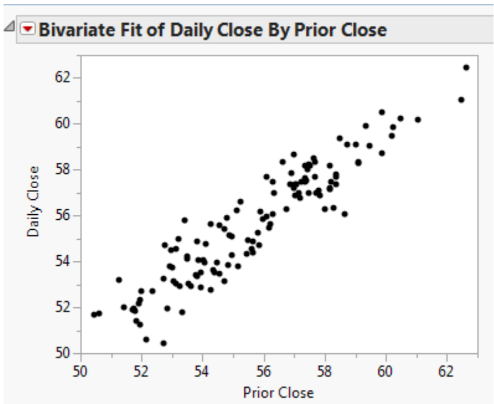
```
Price = 14341.4 - 0.0397524*Miles
```

The equation appears beneath the graph, and $R^2 = 0.03$.

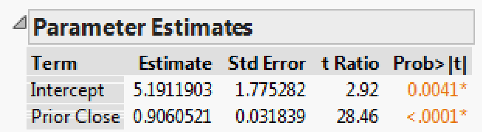
This regression shows there is a weak, significant negative relationship between mileage and price for used cars. The further a car has been driven, on average the lower the price (about 4 cents per mile, on average). However there is considerable scatter around the line.

Scenario 5

a.

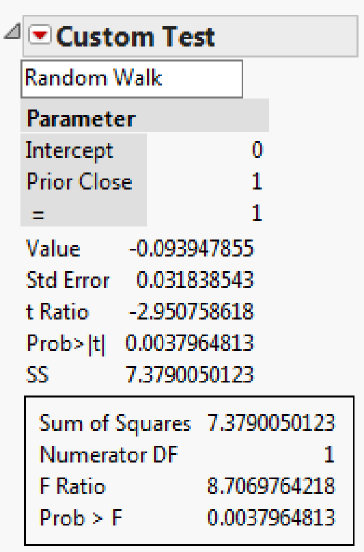


In the scatterplot we see a moderately strong positive linear association.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5.1911903	1.775282	2.92	0.0041*
Prior Close	0.9060521	0.031839	28.46	<.0001*



Custom Test

Random Walk

Parameter	
Intercept	0
Prior Close	1
=	1

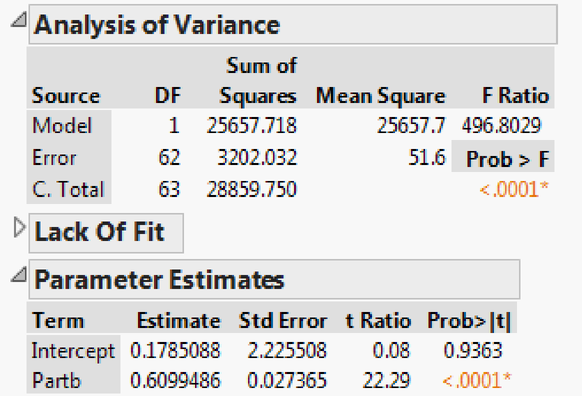
Value	-0.093947855
Std Error	0.031838543
t Ratio	-2.950758618
Prob>|t|	0.0037964813
SS	7.3790050123

Sum of Squares	7.3790050123
Numerator DF	1
F Ratio	8.7069764218
Prob > F	0.0037964813

Although the estimated slope of 0.906 might appear to be approximately 1, the custom test indicates a significant difference from 1 (p-value = .004). Moreover, we find that the Intercept is significantly different from 0.

Scenario 6

a.



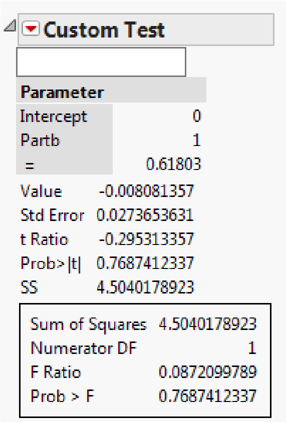
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	25657.718	25657.7	496.8029
Error	62	3202.032	51.6	Prob > F
C. Total	63	28859.750		<.0001*

Lack Of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.1785088	2.225508	0.08	0.9363
Partb	0.6099486	0.027365	22.29	<.0001*



Custom Test

Parameter	
Intercept	0
Partb	1
=	0.61803

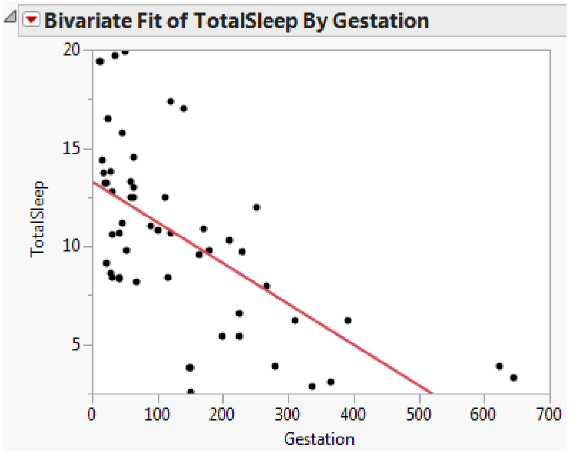
Value	-0.008081357
Std Error	0.0273653631
t Ratio	-0.295313357
Prob>|t|	0.7687412337
SS	4.5040178923

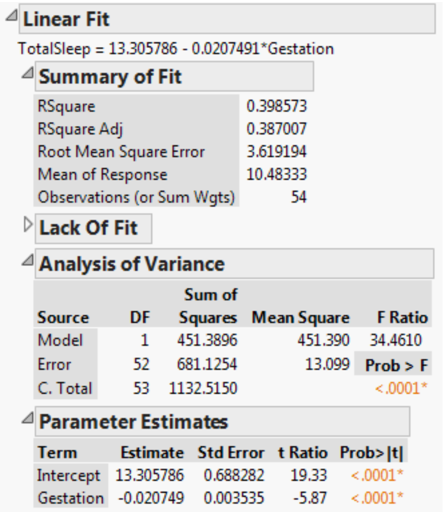
Sum of Squares	4.5040178923
Numerator DF	1
F Ratio	0.0872099789
Prob > F	0.7687412337

Using the Haydn data we find a similar story to the one we saw with Mozart. We again find the Golden Mean model plausible.

Scenario 7

a.



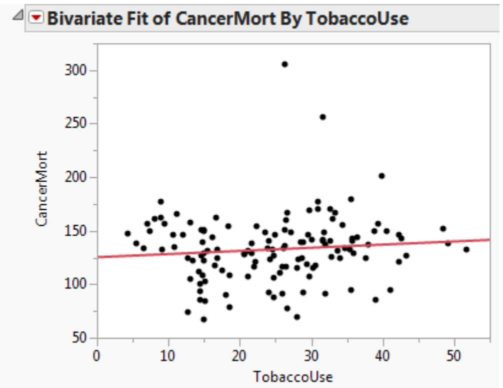


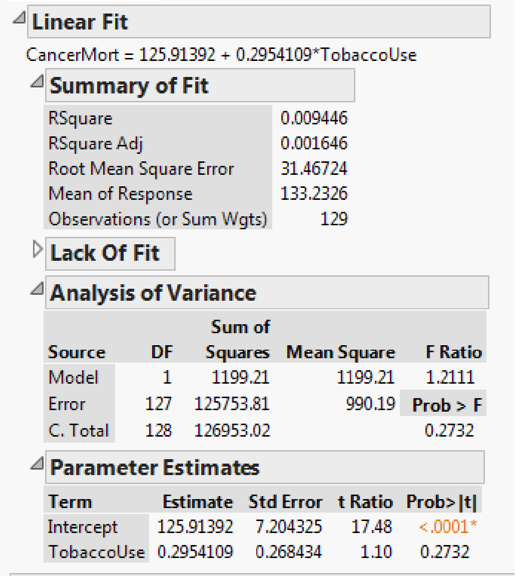
```
Linear Fit
TotalSleep = 13.305786 - 0.0207491*Gestation
Summary of Fit
RSquare                        0.398573
RSquare Adj                    0.387007
Root Mean Square Error         3.619194
Mean of Response               10.48333
Observations (or Sum Wgts)           54
Lack Of Fit
Analysis of Variance
                   Sum of
Source     DF     Squares  Mean Square   F Ratio
Model       1    451.3896      451.390   34.4610
Error      52    681.1254       13.099  Prob > F
C. Total   53   1132.5150                <.0001*
Parameter Estimates
Term        Estimate  Std Error  t Ratio  Prob>|t|
Intercept  13.305786   0.688282    19.33   <.0001*
Gestation  -0.020749   0.003535    -5.87   <.0001*
```

Here we find a significant, but weak, negative relationship. On average, each additional day of gestation is associated with a reduction of 0.02 hours of sleep per night. Gestation accounts for only about 40% of the variation in total sleep, so it is a fair predictor of sleep hours.

Scenario 8

a.





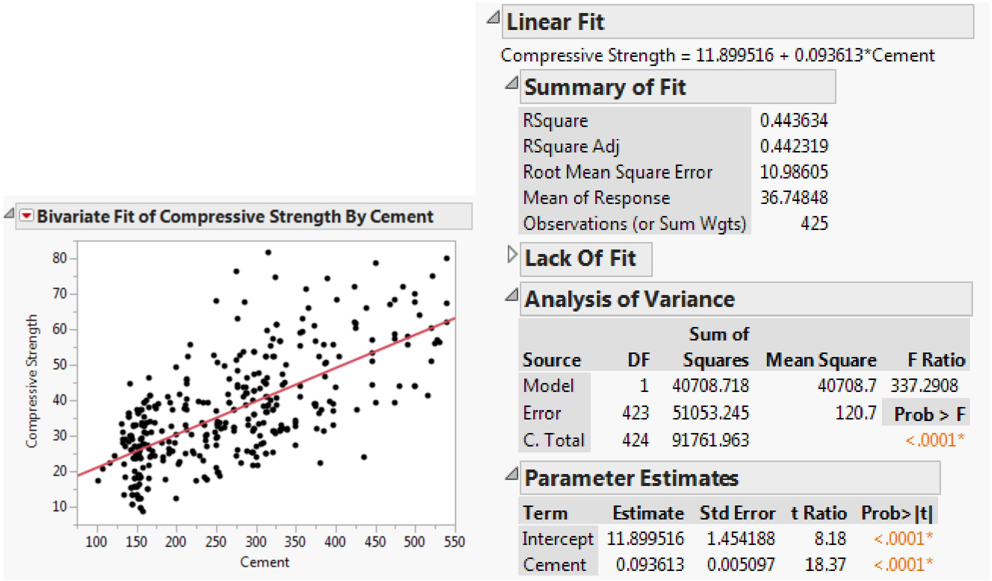
```
Linear Fit
CancerMort = 125.91392 + 0.2954109*TobaccoUse
Summary of Fit
RSquare                        0.009446
RSquare Adj                    0.001646
Root Mean Square Error         31.46724
Mean of Response               133.2326
Observations (or Sum Wgts)          129
Lack Of Fit
Analysis of Variance
                    Sum of
Source      DF     Squares  Mean Square   F Ratio
Model        1     1199.21      1199.21    1.2111
Error      127   125753.81       990.19  Prob > F
C. Total   128   126953.02                 0.2732
Parameter Estimates
Term          Estimate  Std Error  t Ratio  Prob>|t|
Intercept    125.91392   7.204325    17.48   <.0001*
TobaccoUse   0.2954109   0.268434     1.10    0.2732
```

We find a non-significant relationship here – Tobacco Use is not a useful predictor of cancer deaths in a country.

c. The aggregate prevalence of tobacco use obscures the fine distinctions in the amount and length of tobacco use in individuals. We'd really want to look at data at the individual level in order to determine the degree to which increased tobacco use influences the risks of death from cancer or from cardiovascular disease.

Scenario 9

a.



Linear Fit

Compressive Strength = 11.899516 + 0.093613*Cement

Summary of Fit

RSquare	0.443634
RSquare Adj	0.442319
Root Mean Square Error	10.98605
Mean of Response	36.74848
Observations (or Sum Wgts)	425

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	40708.718	40708.7	337.2908
Error	423	51053.245	120.7	Prob > F
C. Total	424	91761.963		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11.899516	1.454188	8.18	<.0001*
Cement	0.093613	0.005097	18.37	<.0001*

This is a highly significant, but weak, positive relationship. For each additional kg of cement in the mixture, compressive strength increases on average by 0.09 megapascals.

Scenario 10

a. There are slight differences, but when we round the major statistics we find that all four models are nearly identical: $Y_i = 3 + 0.5\ X_i$. All R^2 (0.66) and p-values (0.0022 for the slope) are the same.

c. In the other three graphs, the points do not fall in a linear pattern at all. This illustrates a substantial risk in running a linear regression without first examining the data visually. (In JMP we *always* see a scatterplot of the points either prior to fitting a model or in conjunction with fitting a model).

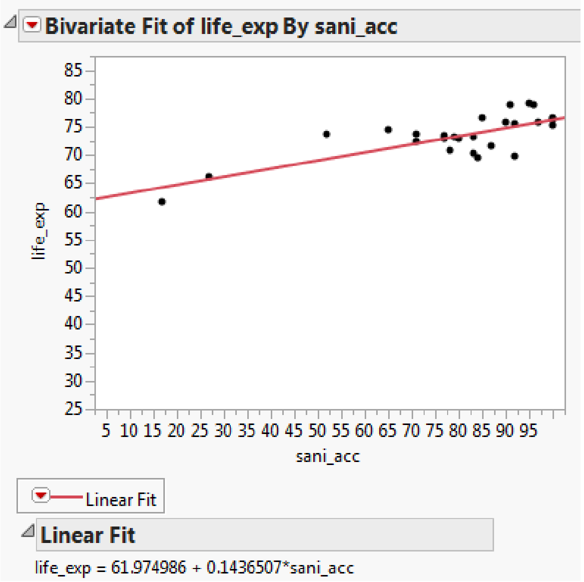
Scenario 11

a. The estimated equation is Price = 17625.688 - 0.054972*Miles. On average, the price declines approximately 5.5 cents per mile driven, and a car that had never been driving would have an asking price of $ 17,625.69.

c. The estimated equation is Price = 10659.169 - 0.0350164*Miles. Due to the large p-value for the slope, we cannot be confident that the true slope differs from 0, and hence should not venture an estimate of the price decline. The p-value for the intercept is significant , and we can estimate that a car that had never been driving would have an asking price of $ 10,659.17.

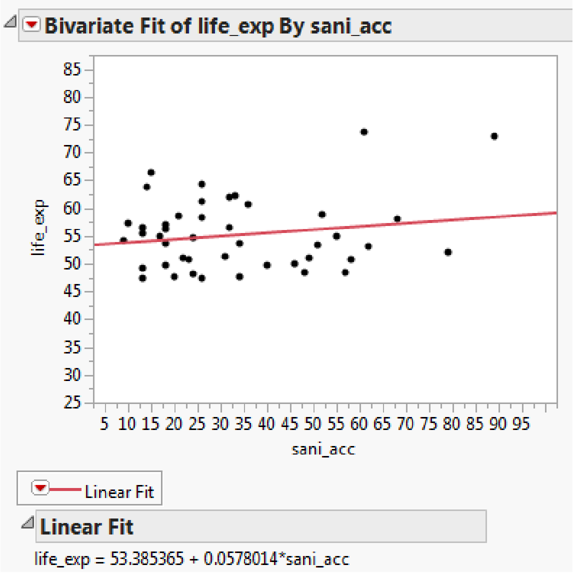
Scenario 12

a. Countries in which higher percentages of citizens have access to sanitation have greater life expectancies. The equation appears beneath the fitted line plot. The slope is significant at the 0.0001 level, and R^2 = 0.58.



life_exp = 61.974986 + 0.1436507*sani_acc

c.

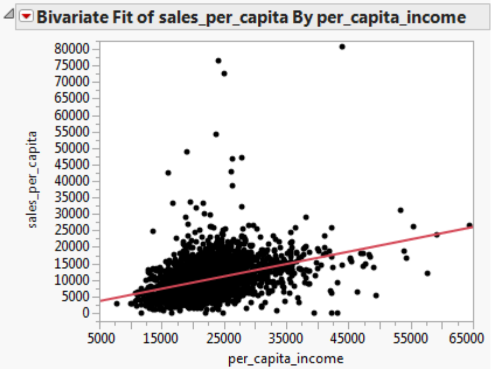


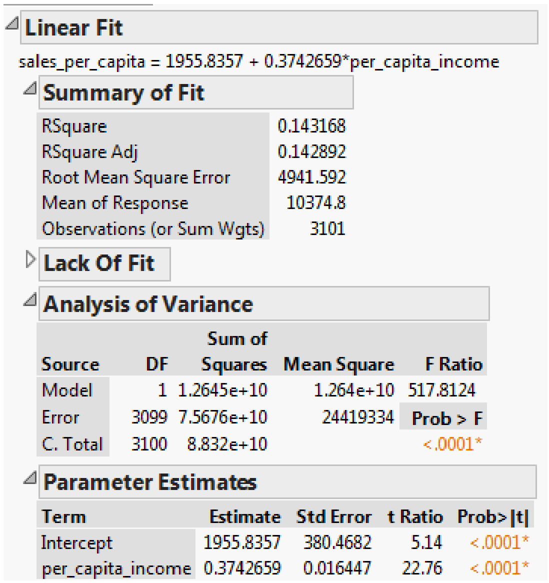
life_exp = 53.385365 + 0.0578014*sani_acc

The equation appears beneath the fitted line plot. In this case, the estimated slope is not significant (p-value = 0.253) and R² = 0..03.

Scenario 13

a.





sales_per_capita = 1955.8357 + 0.3742659*per_capita_income

Summary of Fit

RSquare	0.143168
RSquare Adj	0.142892
Root Mean Square Error	4941.592
Mean of Response	10374.8
Observations (or Sum Wgts)	3101

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1.2645e+10	1.264e+10	517.8124
Error	3099	7.5676e+10	24419334	Prob > F
C. Total	3100	8.832e+10		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1955.8357	380.4682	5.14	<.0001*
per_capita_income	0.3742659	0.016447	22.76	<.0001*

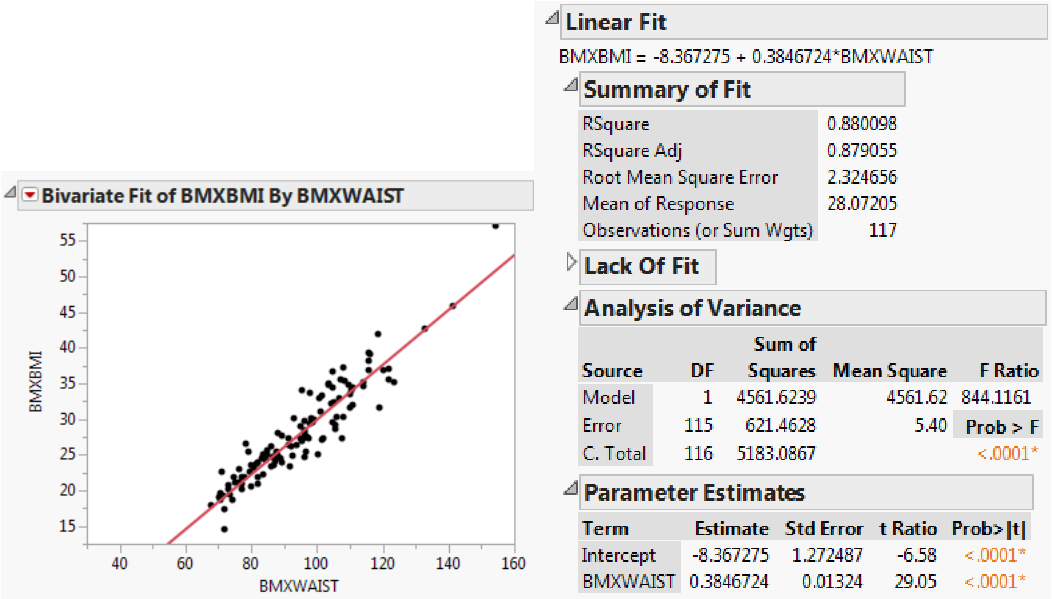
We first should note that the linear model is not particularly suitable for the cloud of points. There are a relatively small number of outlying points, but overall the trend is that higher per capita income is associated with higher retail spending. This makes logical sense because areas of higher incomes have residents who are in a position to spend more, other things being equal. On average, each additional dollar in per capita income is associated with an increase of approximately 37 cents in spending. The estimated slope is highly significant, but the relationship is weak, with $R^2 = 0.14$.

16

Student Solutions to Application Scenarios

Scenario 1

a.



Linear Fit

BMXBMI = -8.367275 + 0.3846724*BMXWAIST

Summary of Fit

RSquare	0.880098
RSquare Adj	0.879055
Root Mean Square Error	2.324656
Mean of Response	28.07205
Observations (or Sum Wgts)	117

Lack Of Fit

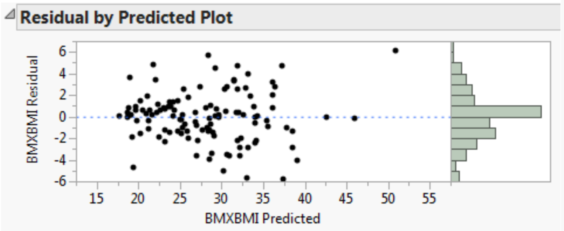
Analysis of Variance

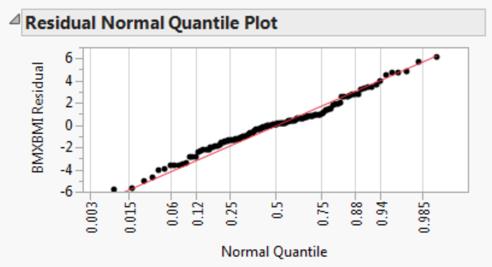
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	4561.6239	4561.62	844.1161
Error	115	621.4628	5.40	Prob > F
C. Total	116	5183.0867		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-8.367275	1.272487	-6.58	<.0001*
BMXWAIST	0.3846724	0.01324	29.05	<.0001*

We first performed this regression in the previous chapter. Above are the regression results for adult females. We find a significant relationship between waist circumference and BMI, with the waist measurement accounting for about 88% of the variation in BMI. Each addition centimeter of waist circumference is associated with an increase of 0.3847 in BMI. When we save the residuals and check their normality, we find the normality assumption seems to be reasonable. The graph of residuals vs. predicted values suggests that the dispersion of residuals increases as predicted values increase, though it is not an overly dramatic tendency. We can probably trust this model for predictions.

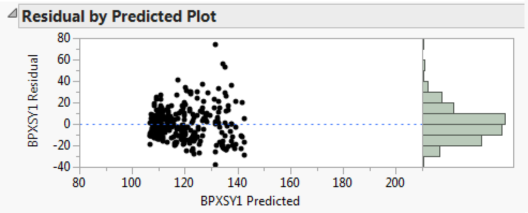


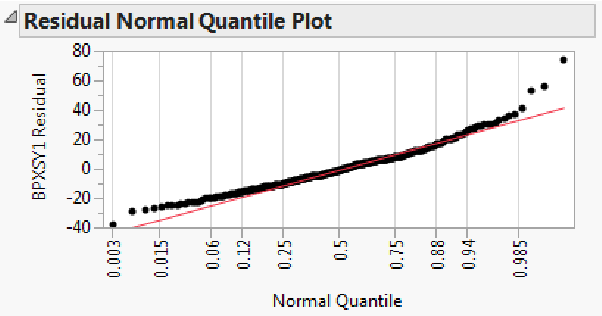


c. Looking at the fitted line graph , it appears that the mean BMI for women with 68 cm. waists is approximately18 .

Scenario 2

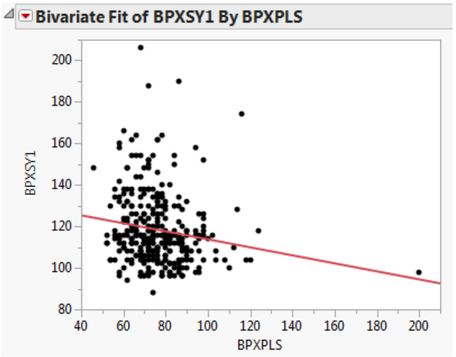
a.



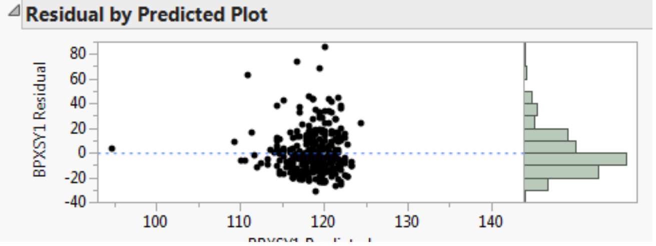


Once again we see the suggestion of heteroskedasticity on the left side of the graph. The residals are largely normal in shape, though somewhat right-skewed. We can probably use the model safely.

c.



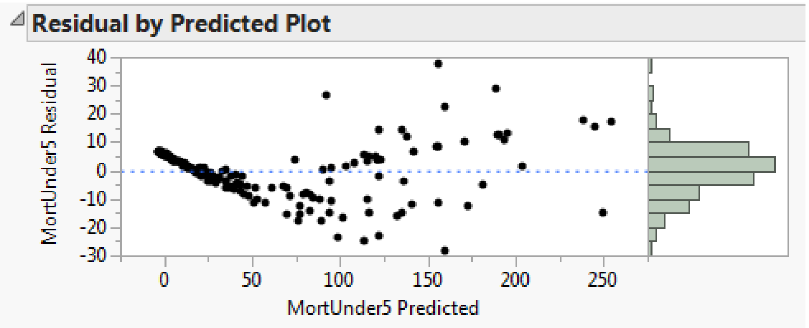
The scatterplot to the left shows little or no relationship between pulse and systolic BP. If anything, there may be a very weak negative relationship here, contrary to the suspicion expressed in the question.

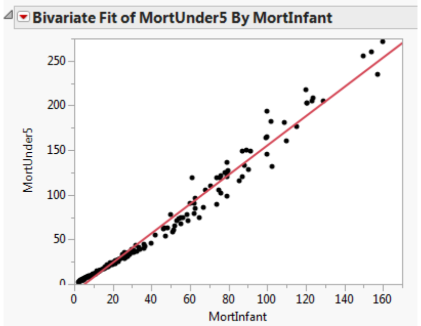


The residuals graphs cast doubt on both normality and constant variance.

Scenario 3

a.



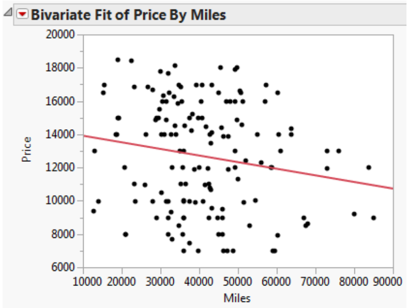


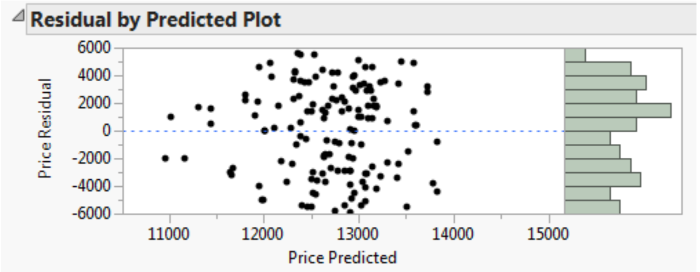
In Chapter 15 we noted that despite the strong summary statistics, the scatterplot very clearly indicates some doubt about the linear model: the points seem to bend around the line, suggesting that the relationship is not best described as a line.

The Residual by Predicted plot very clearly depicts both the non-linearity and the heteroskedasticity. Normality does not seem to present a serious problem.

Scenario 4

a.



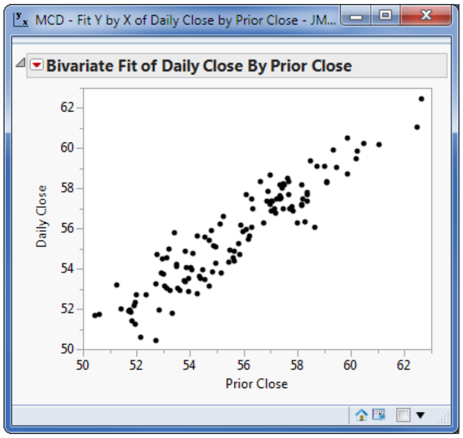


(Note: it is wise to adjust the horizontal axis on the residual by predicted plot to more clearly see the pattern.) The residuals are not normally distributed, there may be a problem with constant variance on the left side of the graph. The sample size may be large enough to rely on the Central Limit Theorem.

c. Student answers will vary. The prediction bands on this graph are quite wide, and even with rescaling the axes it is difficult to read predicted values of Y. A reasonable response would be that the price should fall between $6200 to $19,500.

Scenario 5

a.

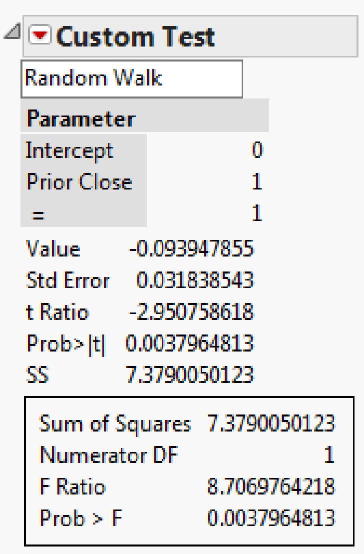


MCD - Fit Y by X of Daily Close by Prior Close - JM...

Bivariate Fit of Daily Close By Prior Close

In the scatterplot we see a moderately strong positive linear association.

b.



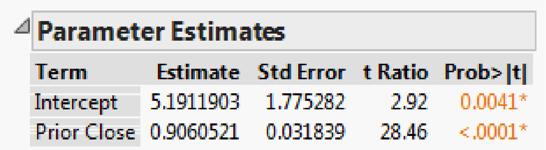
Custom Test

Random Walk

Parameter	
Intercept	0
Prior Close	1
=	1

Value	-0.093947855
Std Error	0.031838543
t Ratio	-2.950758618
Prob>|t|	0.0037964813
SS	7.3790050123

Sum of Squares	7.3790050123
Numerator DF	1
F Ratio	8.7069764218
Prob > F	0.0037964813



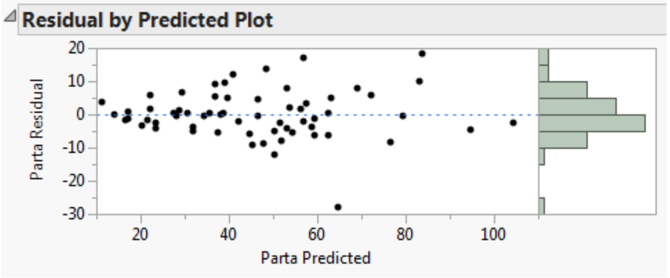
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5.1911903	1.775282	2.92	0.0041*
Prior Close	0.9060521	0.031839	28.46	<.0001*

Although the estimated slope of 0.906 might appear to be approximately 1, the custom test indicates a significant difference from 1 (p-value = .004). Moreover, we find that the Intercept is significantly different from 0. Therefore, the Random Walk model does not suit this set of data.

Scenario 6

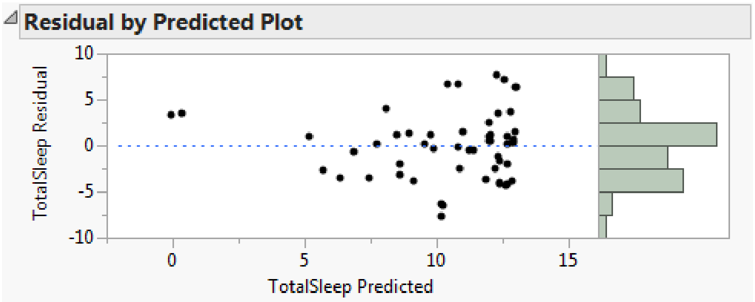
a.



With the Haydn data, in the Residual vs. Partb plot we find a heteroskedastic pattern; the residual do deviate slightly from normality, but the distribution is single peaked, so inference is probably appropriate.

Scenario 7

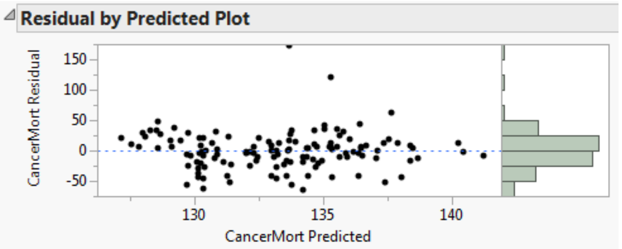
a.



Here we find a heteroskedastic pattern in which the variability of residuals increases as the Gestation period lengthens. Normality is not ideal, but the sample size may be enough to rely on the CLT. Given the non-constant variance, we should be reluctant to interpret or use the results of the regression.

Scenario 8

a.

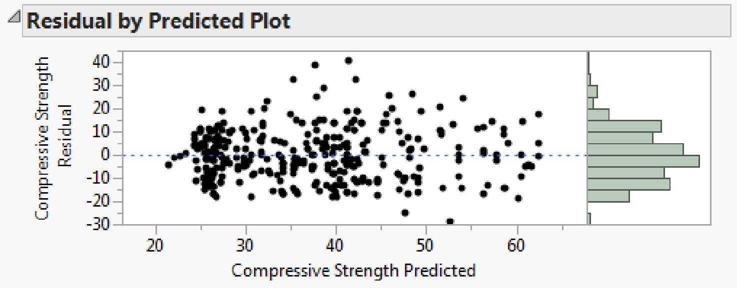


(Note: it is wise to adjust the horizontal axis on the residual by predicted plot to more clearly see the pattern.)

Recall that we find a non-significant relationship here – Tobacco Use is not a useful predictor of cancer deaths in a country. The residuals seem to show more variability in the middle range of tobacco use (non-constant variance), and residuals are nearly normal, with a long upper tail but large sample size. This model is not useful for inference.

Scenario 9

a.

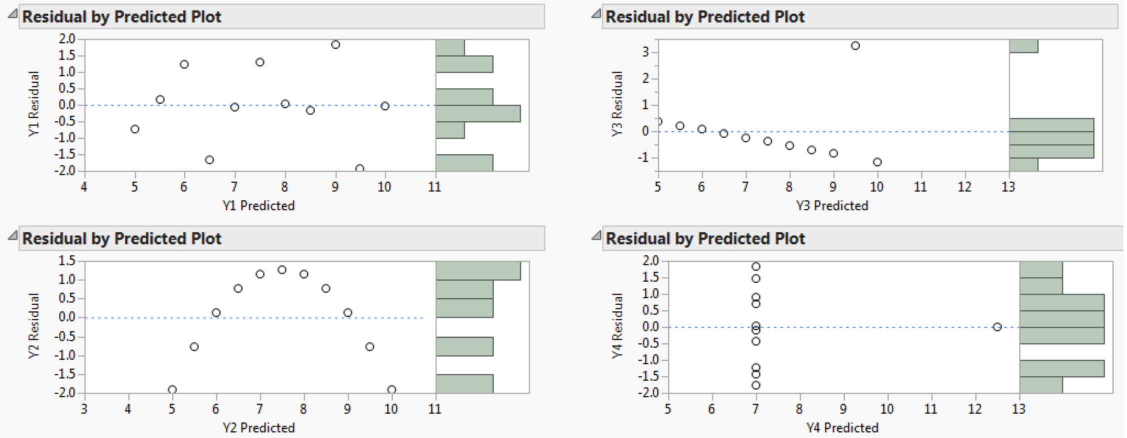


These residuals look good… the Residual vs. Cement plot shows an even scatter above and below the 0-line and the normal quantile plot shows that the residuals follow a nearly normal distribution except for the lower tail. In any case, we have a large sample, so the CLT applies. We can safely interpret the results.

This is a highly significant, but weak, positive relationship. For each additional kg of cement in the mixture, compressive strength increases on average by 0.005 megapascals.

Scenario 10

a.

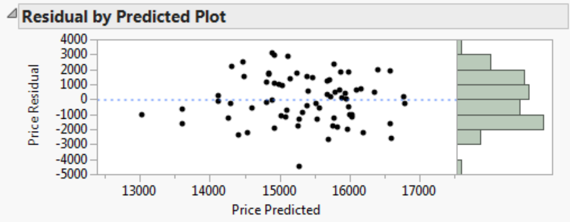


Above are the four plots of residuals vs. predicted. The residuals in the first regression are homoskedastic and approximately normal. The others indicate non-linearity and/or heteroskedasticity. Normality plots also indicate non-normal residuals in these small samples.

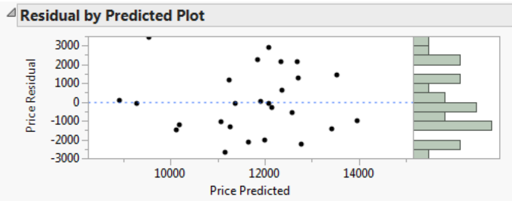
Scenario 11

a.

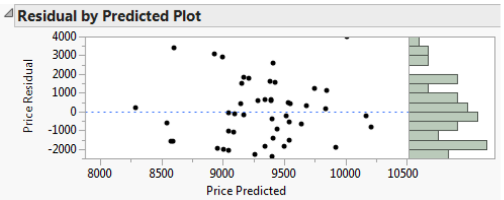
Civic:



Corolla:



Cruiser:

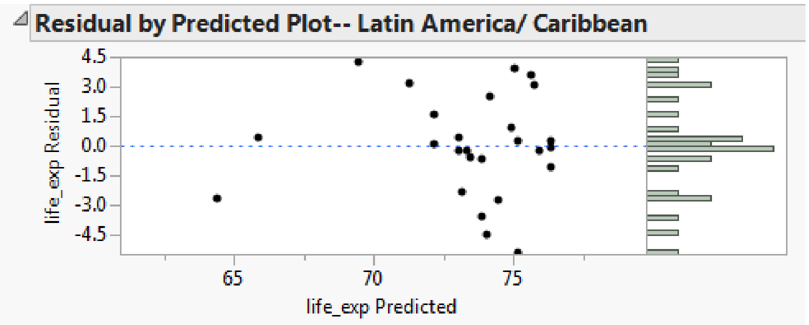


(Note: it is wise to adjust the horizontal axes on the residual by predicted plots to more clearly

see the pattern.)
Recall that we find a non-significant relationship for the Cruiser data. Each set of residuals would appear to have constant variance; the Civic data are most nearly normal, but normality is questionable for the others. Hence p-value estimates and confidence intervals may be inaccurate.

Scenario 12

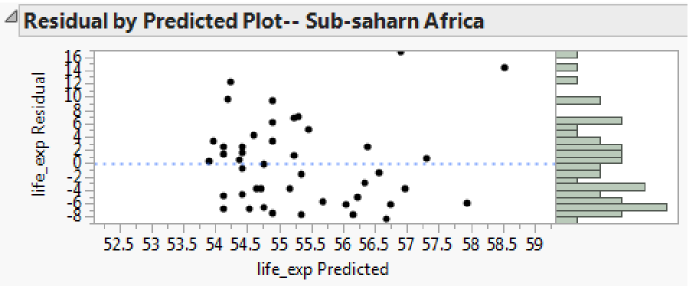
a.



Residual by Predicted Plot-- Latin America/ Caribbean

These residuals are unimodal and somewhat symmetric. With a small sample it is difficult to determine non-constant variance. No obvious violations, so inference is reasonable.

c.

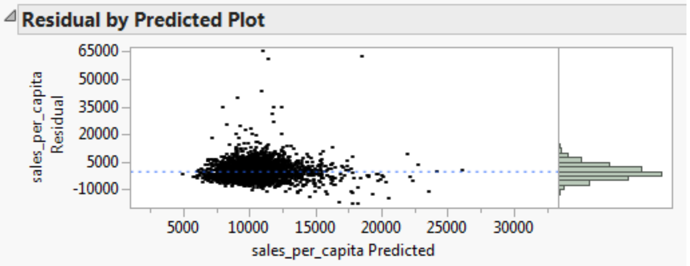


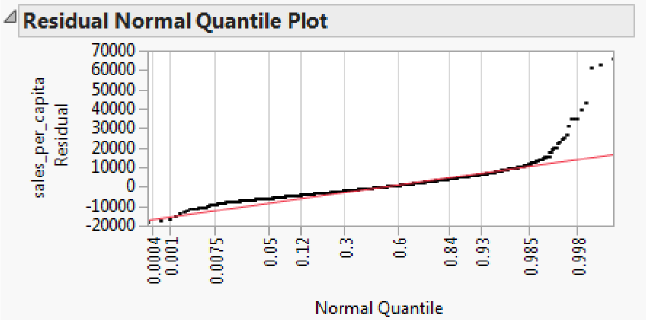
Residual by Predicted Plot-- Sub-saharn Africa

This is a non-significant relationship. The residuals seem to show constant variance, but a skewed and flat distribution. Inference is safe.

Scenario 13

a.





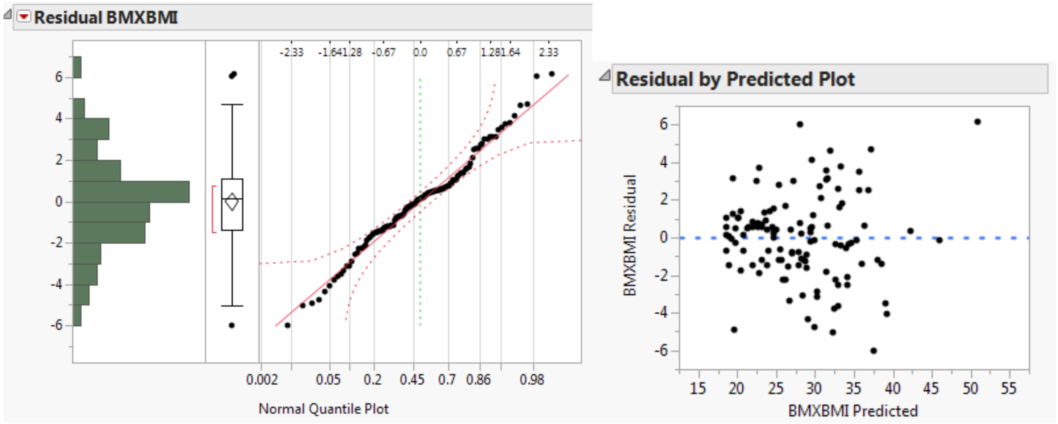
This is a very large sample, so the skewness to the right side may not be a major issue. The residuals do not appear to have a purely random pattern with constant variance, so judgments based on confidence intervals and p-values may be questionable.

18

Student Solutions to Application Scenarios

Scenario 1

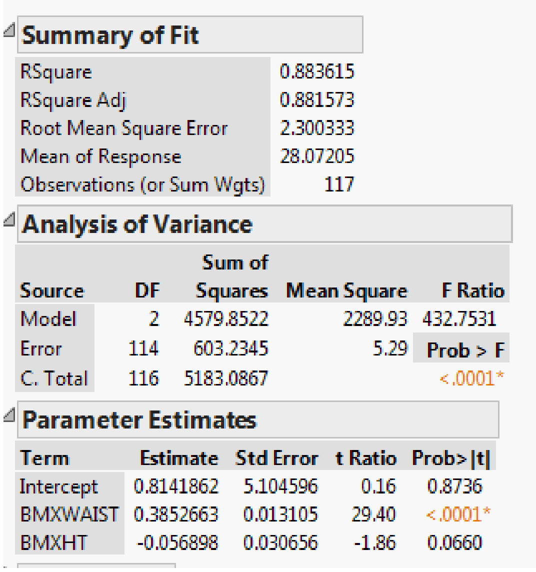
a.



The residual plots from this multiple regression model are very similar to those from the simple regression using Waist circumference as the only predictor (see those graphs below). We can use this set of data for estimation. The regression results themselves are shown below.

We find a strong relationship between BMI and the model, but this model is not much of an improvement over the previous model (shown again below). The intercept has changed dramatically, though in this model the intercept does not have much meaning. The effect size for the Waist measurement is almost equal to that of the single variable model, and the coefficient of height is not significant at the cusotmary .0.05 level. The height variable is not significant at the 0.05 level, though it is significant at the 0.10 level.

The two-variable model has a very small improvement in goodness of fit in comparison to the single-variable model.



Summary of Fit

RSquare	0.883615
RSquare Adj	0.881573
Root Mean Square Error	2.300333
Mean of Response	28.07205
Observations (or Sum Wgts)	117

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	4579.8522	2289.93	432.7531
Error	114	603.2345	5.29	Prob > F
C. Total	116	5183.0867		<.0001*

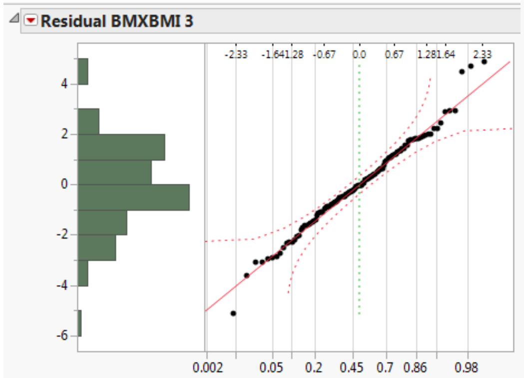
Parameter Estimates

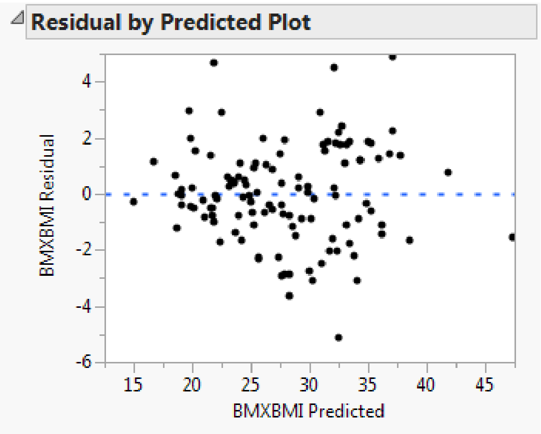
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.8141862	5.104596	0.16	0.8736
BMXWAIST	0.3852663	0.013105	29.40	<.0001*
BMXHT	-0.056898	0.030656	-1.86	0.0660

In short, the addition of the height data does not improve the model in any material way.

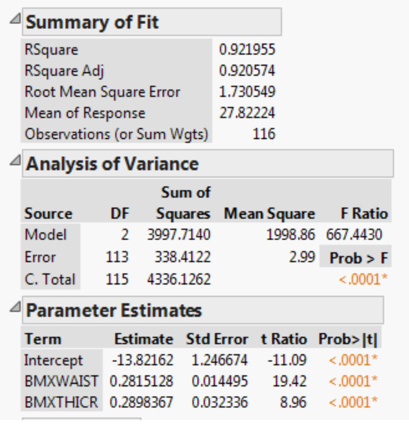
We first performed this regression in Chapter 15. At that time we found a significant relationship between waist circumference and BMI, with the waist measurement accounting for about 88% of the variation in BMI. Each addition centimeter of waist circumference is associated with an increase of 0.3847 in BMI. When we saved the residuals and check their normality, we find the normality assumption seems to be reasonable. The graph of residuals vs. predicted values suggested that the dispersion of residuals increases as predicted values increase, though it is not an overly dramatic tendency. We can probably trust this model for predictions.

c. NOTE: The scenario question mistakenly asks for you to use the Write Circumference as a predictor. The question should ask for <u>thigh</u> circumference.





In the model using waist and thigh circumference (note typographical error in early printings of the book that this is refrered to as wrist cicumference), we find residuals that are approximtely normal and more heteroskedastic than our prior models. In this sense, the model is less attractive than the earlier ones. On the other hand, the goodness of fit is improved (Adj. RSquare; see below) now equals 0.92 and both slopes are statistically signifiant and make logical sense.



Summary of Fit

RSquare	0.921955
RSquare Adj	0.920574
Root Mean Square Error	1.730549
Mean of Response	27.82224
Observations (or Sum Wgts)	116

Analysis of Variance

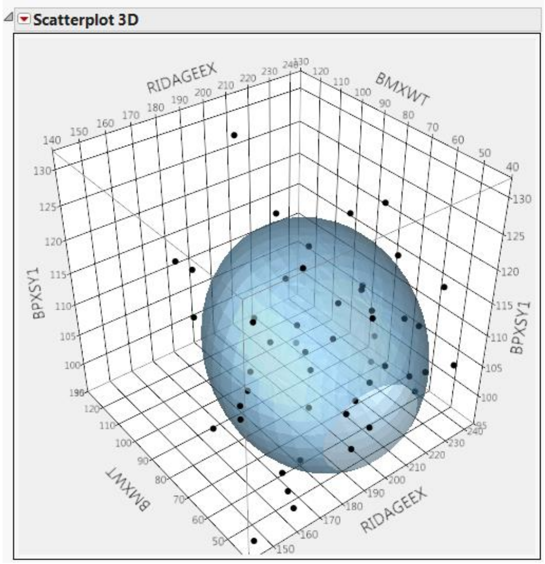
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	3997.7140	1998.86	667.4430
Error	113	338.4122	2.99	Prob > F
C. Total	115	4336.1262		<.0001*

Parameter Estimates

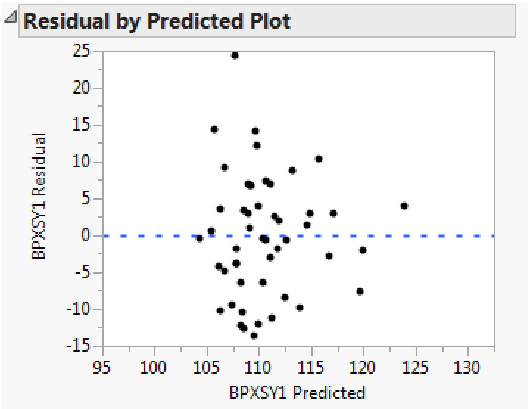
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-13.82162	1.246674	-11.09	<.0001*
BMXWAIST	0.2815128	0.014495	19.42	<.0001*
BMXTHICR	0.2898367	0.032336	8.96	<.0001*

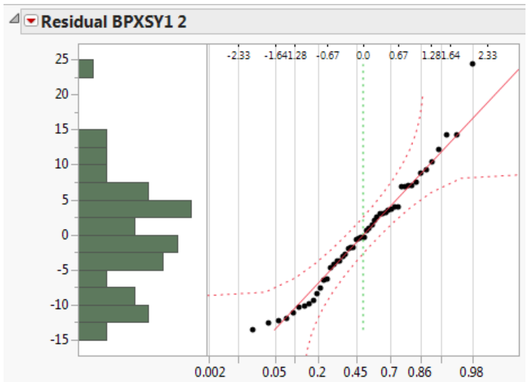
Scenario 2

a. Student answers will vary. One rotated scatterplot is shown here (including a density ellipsoid). We see a weak tendency for systolic BP to increase both as age and weight increase.

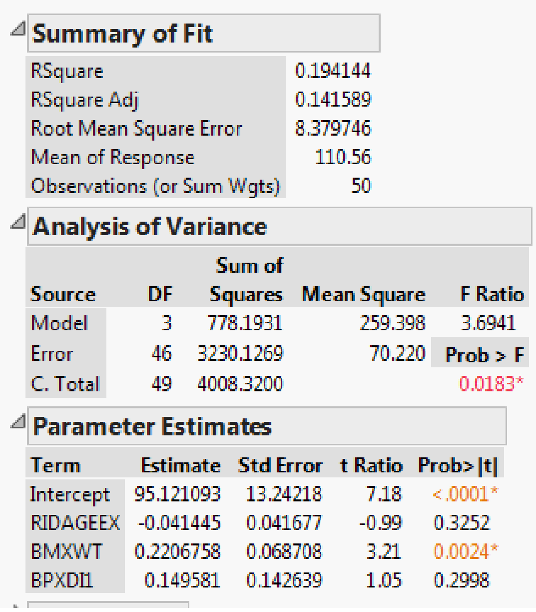


c.





Here again we find concerns about heteroskedasticity and normality; if we continue on to interpret the coefficient estimates, we see that the Diastolic BP adds little to the model. The estimated value is not significantly different from zero, and the adjusted R^2 is very nearly the same in the prior model using just 2 factors in the model. This model is no meaningful improvement over the prior one.



Summary of Fit

RSquare	0.194144
RSquare Adj	0.141589
Root Mean Square Error	8.379746
Mean of Response	110.56
Observations (or Sum Wgts)	50

Analysis of Variance

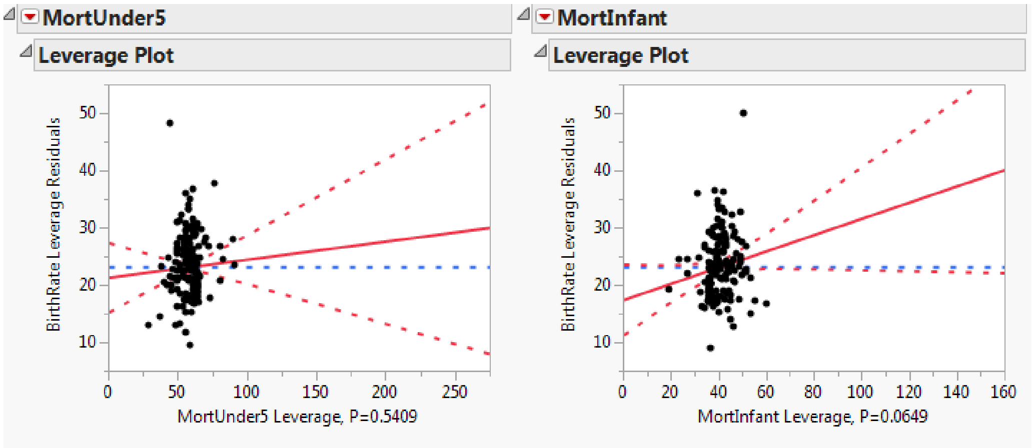
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	778.1931	259.398	3.6941
Error	46	3230.1269	70.220	Prob > F
C. Total	49	4008.3200		0.0183*

Parameter Estimates

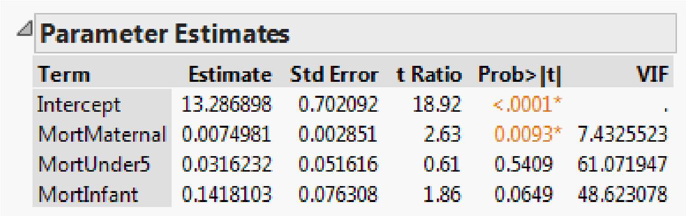
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	95.121093	13.24218	7.18	<.0001*
RIDAGEEX	-0.041445	0.041677	-0.99	0.3252
BMXWT	0.2206758	0.068708	3.21	0.0024*
BPXDI1	0.149581	0.142639	1.05	0.2998

Scenario 3

a.



The leverage plots immediately suggest a problem with collinearity, which is confirmed by the very high VIFs in the table of parameter estimates (below):



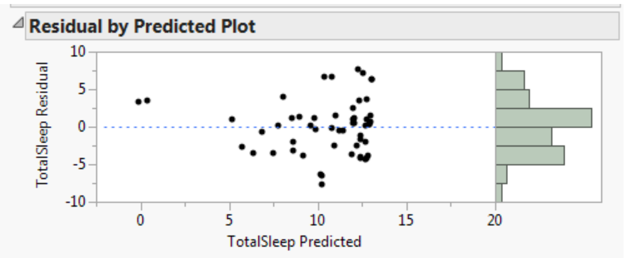
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	13.286898	0.702092	18.92	<.0001*	.
MortMaternal	0.0074981	0.002851	2.63	0.0093*	7.4325523
MortUnder5	0.0316232	0.051616	0.61	0.5409	61.071947
MortInfant	0.1418103	0.076308	1.86	0.0649	48.623078

This model should not be used or interpreted.

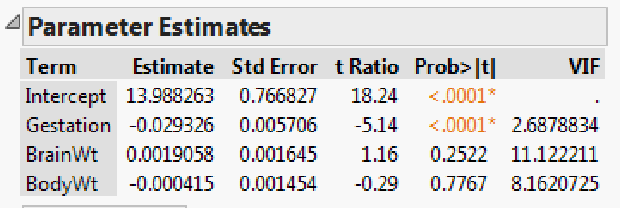
Scenario 4

a.



When we estimate a simple linear model using gestation as the factor, we find a heteroskedastic pattern in which the variability of residuals diminishes as the Gestation period lengthens. Normality is not ideal, but the sample size is large enough to rely on the CLT. Given the non-constant variance, we should be reluctant to interpret or use the results of the regression.

c.



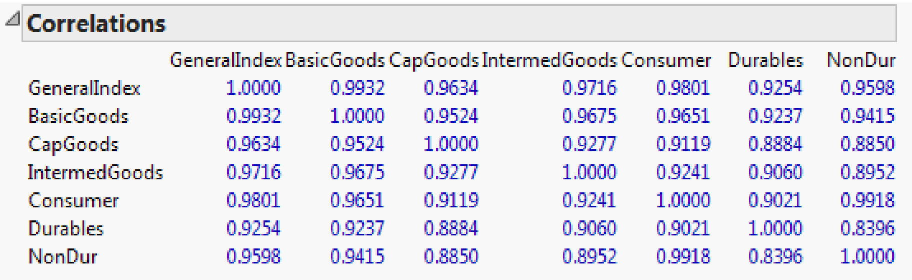
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	13.988263	0.766827	18.24	<.0001*	.
Gestation	-0.029326	0.005706	-5.14	<.0001*	2.6878834
BrainWt	0.0019058	0.001645	1.16	0.2522	11.122211
BodyWt	-0.000415	0.001454	-0.29	0.7767	8.1620725

This model is not an improvement over the prior two. We still see heteroskedasticity in the plot of residuals vs. fitted values (not shown here). We see evidence of collinearity in the large VIF for BrainWt, and only the Gestation variable is statistically significant.

Scenario 5

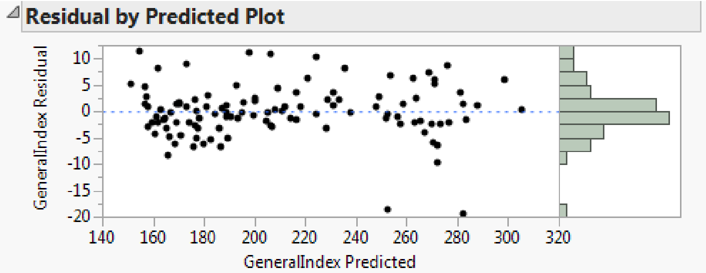
a.



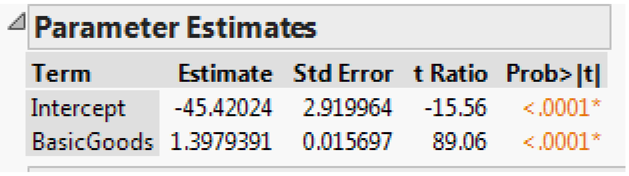
Correlations

	GeneralIndex	BasicGoods	CapGoods	IntermedGoods	Consumer	Durables	NonDur
GeneralIndex	1.0000	0.9932	0.9634	0.9716	0.9801	0.9254	0.9598
BasicGoods	0.9932	1.0000	0.9524	0.9675	0.9651	0.9237	0.9415
CapGoods	0.9634	0.9524	1.0000	0.9277	0.9119	0.8884	0.8850
IntermedGoods	0.9716	0.9675	0.9277	1.0000	0.9241	0.9060	0.8952
Consumer	0.9801	0.9651	0.9119	0.9241	1.0000	0.9021	0.9918
Durables	0.9254	0.9237	0.8884	0.9060	0.9021	1.0000	0.8396
NonDur	0.9598	0.9415	0.8850	0.8952	0.9918	0.8396	1.0000

In the correlation matrix we find that the Basic Goods index is most highly correlated with the General Index. The simple model that estimates monthly values of the General IIP from the Basic Goods IIP provides an excellent goodness of fit and the sample is large enough to invoke the CLT. However, we do see some evidence of non-linearity in the plot of residuals vs. fitted values (below):



Given the R^2 value of nearly 0.99, the non-linearity may not be a major problem. The estimation results are as follows:



Parameter Estimates

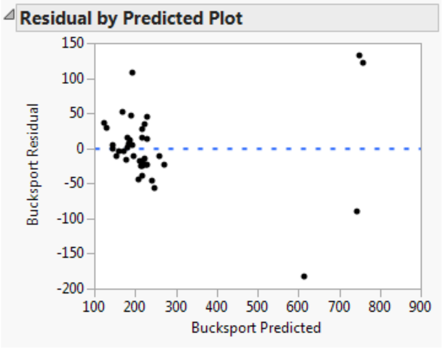
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-45.42024	2.919964	-15.56	<.0001*
BasicGoods	1.3979391	0.015697	89.06	<.0001*

An increase of 1 in the Basic Goods index will be accompanied on average by an increase of approximately 1.4 in the General Index.

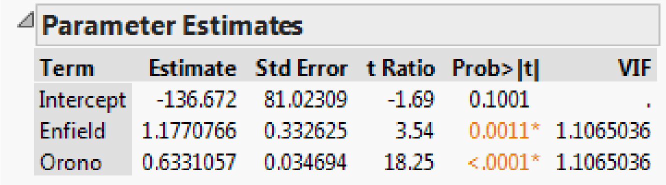
c. See discussion in part (b) above. It is not surprising that these index variables are all highly correlated because they all measure different aspects of the fundamental production activity within the Indian economy, and all reflect the general level of economic activity.

Scenario 6

a. Student models will vary. Here is one plausible result using the Enfield and Orono columns:



The residuals appear to have a non-constant variance, which raises a problem with using this model for prediction or estimation. The model adjusted R² is approximately 0.9 which indicates a very good fit. Both variables are statistically significant and we see no real evidence of collinearity.



Parameter Estimates

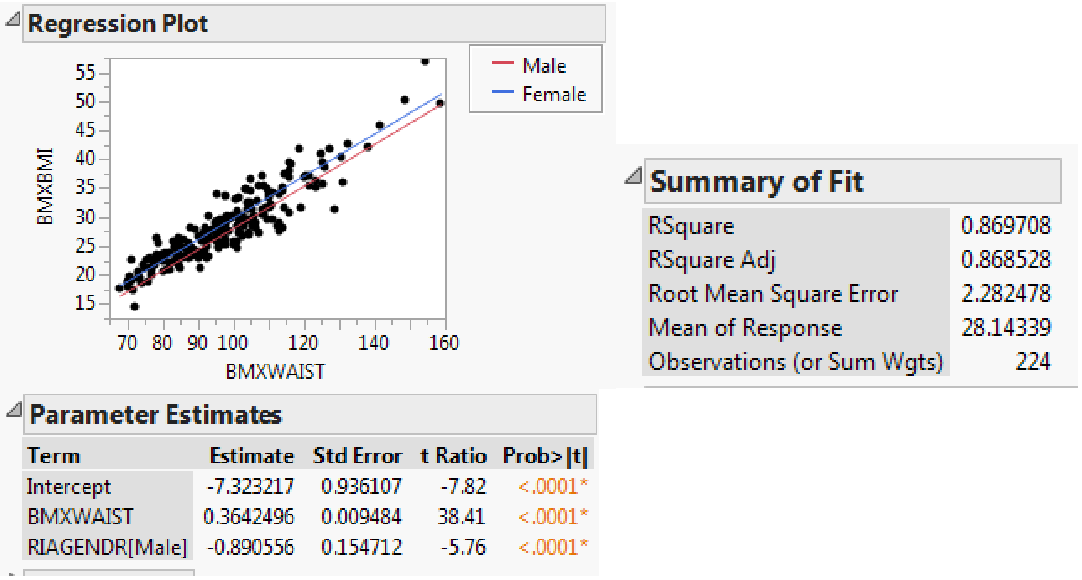
Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-136.672	81.02309	-1.69	0.1001	.
Enfield	1.1770766	0.332625	3.54	0.0011*	1.1065036
Orono	0.6331057	0.034694	18.25	<.0001*	1.1065036

19

Student Solutions to Application Scenarios

Scenario 1

a.



Summary of Fit

RSquare	0.869708
RSquare Adj	0.868528
Root Mean Square Error	2.282478
Mean of Response	28.14339
Observations (or Sum Wgts)	224

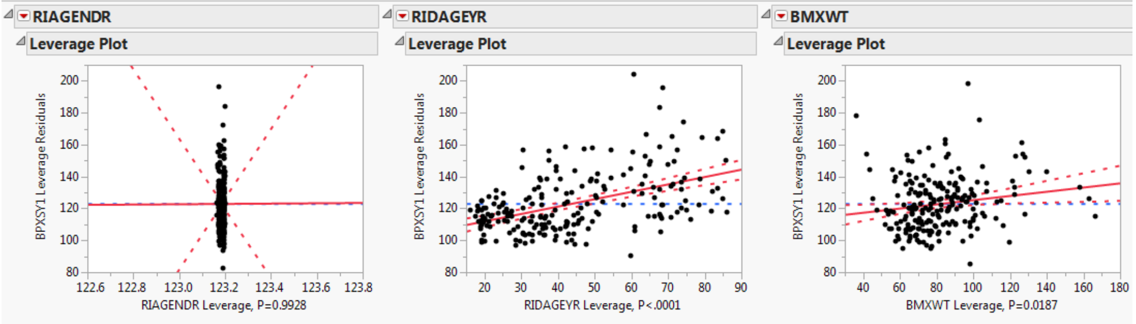
Parameter Estimates

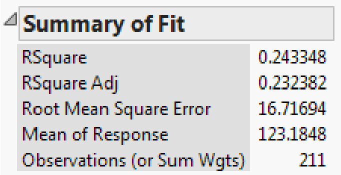
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-7.323217	0.936107	-7.82	<.0001*
BMXWAIST	0.3642496	0.009484	38.41	<.0001*
RIAGENDR[Male]	-0.890556	0.154712	-5.76	<.0001*

The key results are shown above. Compared to the model using waist circumference only, this model has a slightly higher adjusted RSquare and smaller Root Mean Square Error. Both variables are statistically significant. The residuals vs. fits graph is quite similar in both models, and this model makes logical sense.

Scenario 2

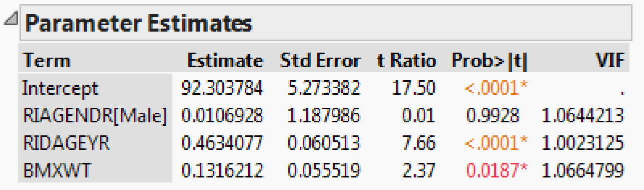
a.





Summary of Fit

RSquare	0.243348
RSquare Adj	0.232382
Root Mean Square Error	16.71694
Mean of Response	123.1848
Observations (or Sum Wgts)	211



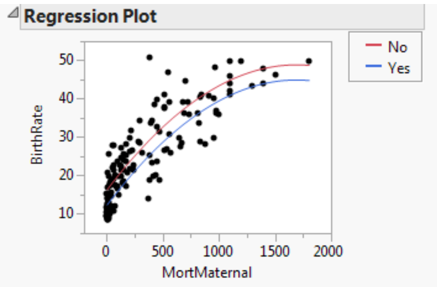
Parameter Estimates

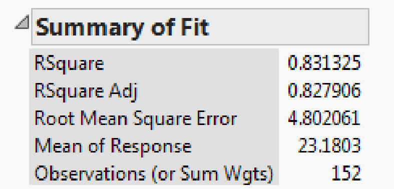
Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	92.303784	5.273382	17.50	<.0001*	.
RIAGENDR[Male]	0.0106928	1.187986	0.01	0.9928	1.0644213
RIDAGEYR	0.4634077	0.060513	7.66	<.0001*	1.0023125
BMXWT	0.1316212	0.055519	2.37	0.0187*	1.0664799

The leverage plots indicate collinearity problems. We see that the model has rather poor fit, and only the Gender variable is not statistically significant.

Scenario 3

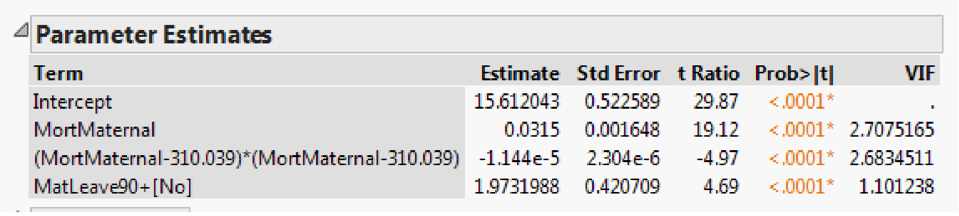
a.





Summary of Fit

RSquare	0.831325
RSquare Adj	0.827906
Root Mean Square Error	4.802061
Mean of Response	23.1803
Observations (or Sum Wgts)	152



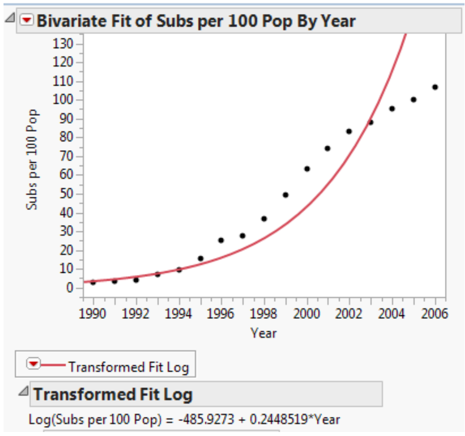
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	15.612043	0.522589	29.87	<.0001*	.
MortMaternal	0.0315	0.001648	19.12	<.0001*	2.7075165
(MortMaternal-310.039)*(MortMaternal-310.039)	-1.144e-5	2.304e-6	-4.97	<.0001*	2.6834511
MatLeave90+[No]	1.9731988	0.420709	4.69	<.0001*	1.101238

This model fits the data rather well, and all coefficients are significant. We find that other things equal higher rates of maternal mortality are associated with higher birthrates, and that after controlling for differences in maternal mortality, countries that do not offer lengthy maternity leaves have higher birthrates than countries with longer leaves. Residuals appear to be normally distributed with equal variances.

Scenario 4

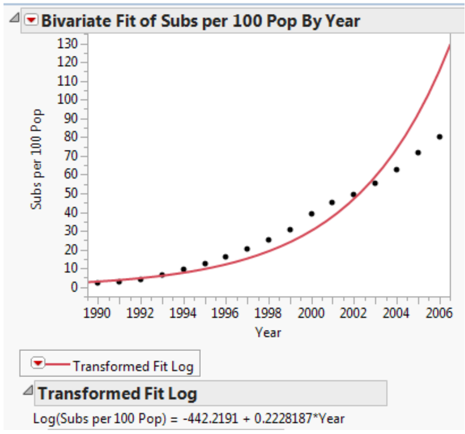
a.



Log(Subs per 100 Pop) = -485.9273 + 0.2448519*Year

For Denmark, the annual growth rate is $e^{0.2448519}$–1 = 0.277 or 27.7% per year.

c.

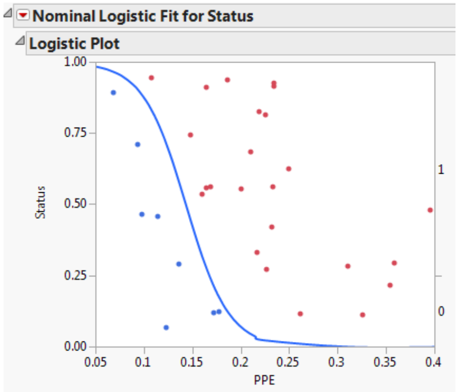


Log(Subs per 100 Pop) = -442.2191 + 0.2228187*Year

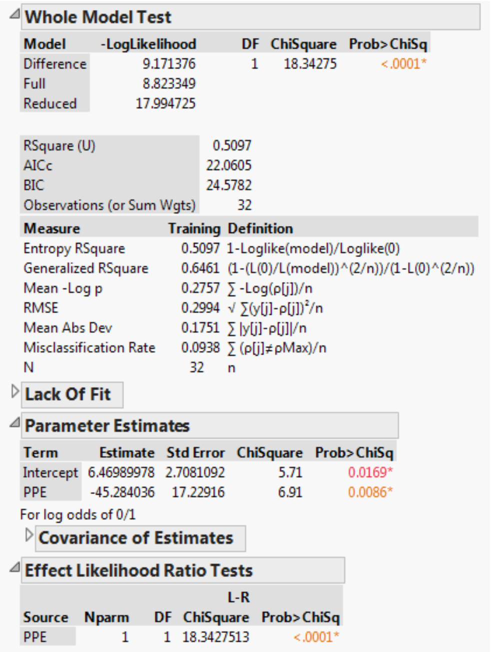
For the U.S., the annual growth rate is $e^{0. 2228187}$–1 = 0.249 or 24.9% per year.

Scenario 5

a.



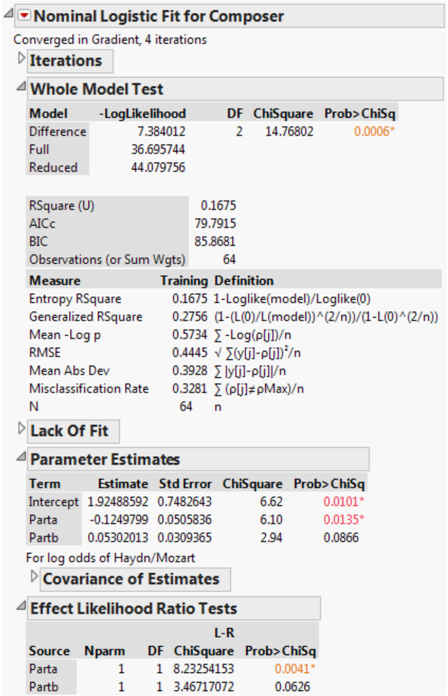
The logistic regression results appear to the left. The regressor, PPE, is statistically significant and we see that patients with Parkinson's Disease have significantly lower PPE values than patients without PD. In the Logistic Plot, the dark markers are patients with PD; we see that the estimated curve distinguishes between PD and non-PD patients.



```
Whole Model Test
Model        -LogLikelihood   DF  ChiSquare  Prob>ChiSq
Difference         9.171376    1   18.34275     <.0001*
Full               8.823349
Reduced           17.994725

RSquare (U)                   0.5097
AICc                         22.0605
BIC                          24.5782
Observations (or Sum Wgts)        32
Measure                  Training  Definition
Entropy RSquare            0.5097  1-Loglike(model)/Loglike(0)
Generalized RSquare        0.6461  (1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p                0.2757  ∑ -Log(ρ[j])/n
RMSE                       0.2994  √ ∑(y[j]-ρ[j])²/n
Mean Abs Dev               0.1751  ∑ |y[j]-ρ[j]|/n
Misclassification Rate     0.0938  ∑ (ρ[j]≠ρMax)/n
N                              32  n
Lack Of Fit
Parameter Estimates
Term          Estimate  Std Error  ChiSquare  Prob>ChiSq
Intercept   6.46989978  2.7081092       5.71     0.0169*
PPE         -45.284036   17.22916       6.91     0.0086*
For log odds of 0/1
  Covariance of Estimates
Effect Likelihood Ratio Tests
                              L-R
Source   Nparm   DF     ChiSquare  Prob>ChiSq
PPE          1    1    18.3427513     <.0001*
```

Scenario 6

a.



Nominal Logistic Fit for Composer

Converged in Gradient, 4 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	7.384012	2	14.76802	0.0006*
Full	36.695744			
Reduced	44.079756			

RSquare (U)	0.1675
AICc	79.7915
BIC	85.8681
Observations (or Sum Wgts)	64

Measure	Training	Definition
Entropy RSquare	0.1675	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.2756	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5734	∑ -Log(ρ[j])/n
RMSE	0.4445	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3928	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.3281	∑ (ρ[j]≠ρMax)/n
N	64	n

Lack Of Fit

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	1.92488592	0.7482643	6.62	0.0101*
Parta	-0.1249799	0.0505836	6.10	0.0135*
Partb	0.05302013	0.0309365	2.94	0.0866

For log odds of Haydn/Mozart

Covariance of Estimates

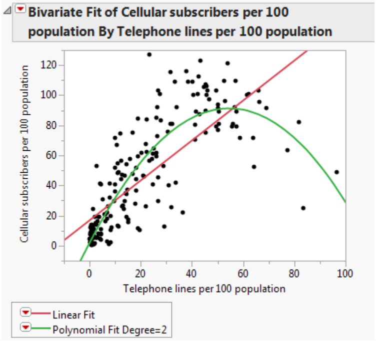
Effect Likelihood Ratio Tests

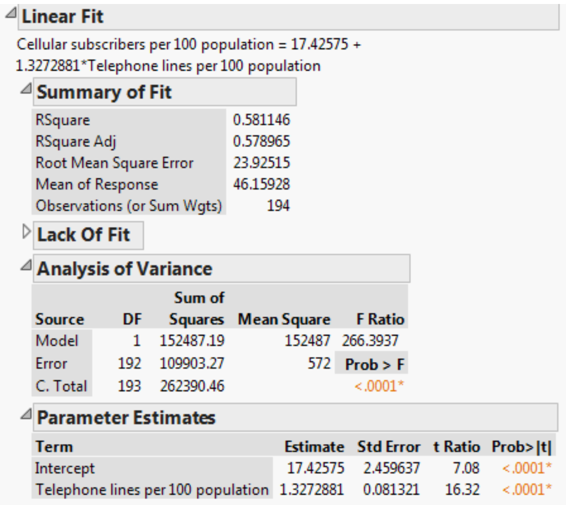
			L-R	
Source	Nparm	DF	ChiSquare	Prob>ChiSq
Parta	1	1	8.23254153	0.0041*
Partb	1	1	3.46717072	0.0626

The results are to the left. We find that the whole model is significant with a rather poor fit, as measured by U. Other things being equal, the longer Part a is the lower the odds that it was composed by Haydn. Conversely, the longer Part b is (holding Part a constant) the higher the odds that it was composed by Haydn.

Scenario 7

a. Here are the results for the quadratic and linear fits:





Linear Fit

Cellular subscribers per 100 population = 17.42575 + 1.3272881*Telephone lines per 100 population

Summary of Fit

RSquare	0.581146
RSquare Adj	0.578965
Root Mean Square Error	23.92515
Mean of Response	46.15928
Observations (or Sum Wgts)	194

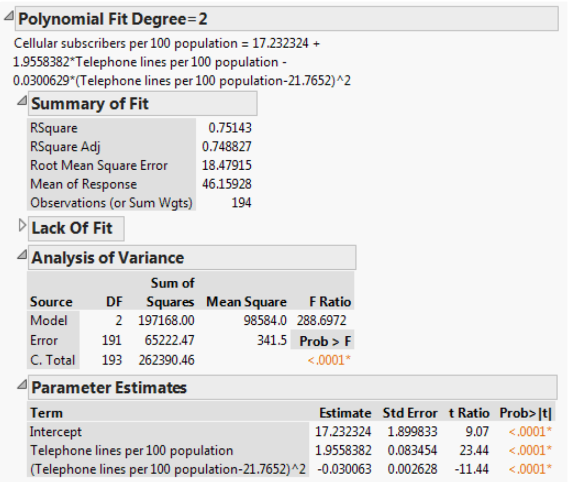
Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	152487.19	152487	266.3937
Error	192	109903.27	572	Prob > F
C. Total	193	262390.46		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	17.42575	2.459637	7.08	<.0001*
Telephone lines per 100 population	1.3272881	0.081321	16.32	<.0001*



Polynomial Fit Degree=2

Cellular subscribers per 100 population = 17.232324 + 1.9558382*Telephone lines per 100 population - 0.0300629*(Telephone lines per 100 population-21.7652)^2

Summary of Fit

RSquare	0.75143
RSquare Adj	0.748827
Root Mean Square Error	18.47915
Mean of Response	46.15928
Observations (or Sum Wgts)	194

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	197168.00	98584.0	288.6972
Error	191	65222.47	341.5	Prob > F
C. Total	193	262390.46		<.0001*

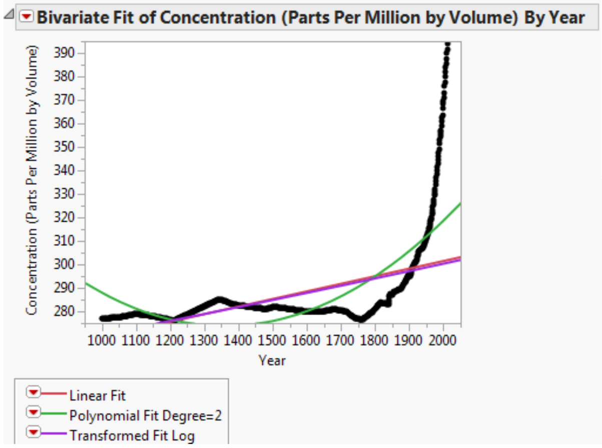
Parameter Estimates

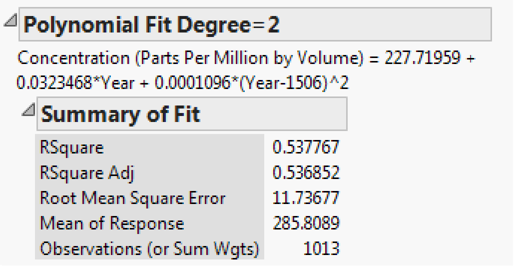
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	17.232324	1.899833	9.07	<.0001*
Telephone lines per 100 population	1.9558382	0.083454	23.44	<.0001*
(Telephone lines per 100 population-21.7652)^2	-0.030063	0.002628	-11.44	<.0001*

We can see that the quadratic model has better goodness of fit statistics, and graphically it is clear the parabolic model fits the observed points better than the linear model.

Scenario 8

a. Here are the results for the linear, quadratic and log- linear fits: The linear and log-linear are nearly indistinguishable. None of the models fit particularly well, which visual inspection makes clear. The quadratic model has the best fit of the three, but it is weak.





Polynomial Fit Degree=2

Concentration (Parts Per Million by Volume) = 227.71959 + 0.0323468*Year + 0.0001096*(Year-1506)^2

Summary of Fit

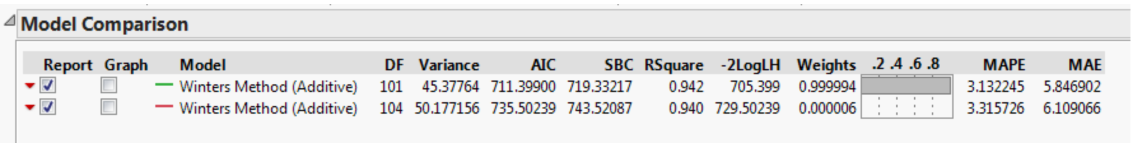
RSquare	0.537767
RSquare Adj	0.536852
Root Mean Square Error	11.73677
Mean of Response	285.8089
Observations (or Sum Wgts)	1013

20

Student Solutions to Application Scenarios

Scenario 1

a.

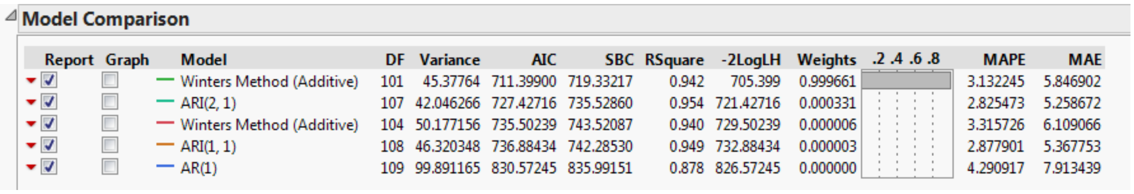


Model Comparison

Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	Weights	.2 .4 .6 .8	MAPE	MAE
☑	☐	Winters Method (Additive)	101	45.37764	711.39900	719.33217	0.942	705.399	0.999994		3.132245	5.846902
☑	☐	Winters Method (Additive)	104	50.177156	735.50239	743.52087	0.940	729.50239	0.000006		3.315726	6.109066

As shown above, using a 6-month season (top row) is a minor improvement over the 3-month season. The variance, MAPE, and MAE are smaller with this model than the earlier model, and RSquare is very slightly higher.

c.



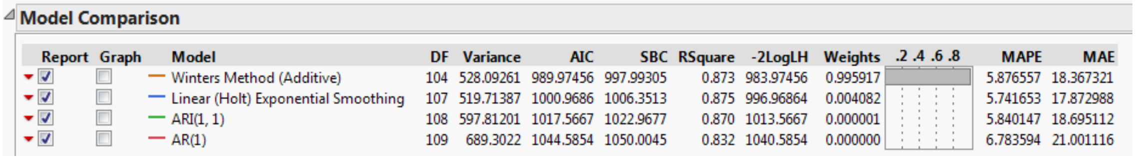
Model Comparison

Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	Weights	.2 .4 .6 .8	MAPE	MAE
☑	☐	Winters Method (Additive)	101	45.37764	711.39900	719.33217	0.942	705.399	0.999661		3.132245	5.846902
☑	☐	ARI(2, 1)	107	42.046266	727.42716	735.52860	0.954	721.42716	0.000331		2.825473	5.258672
☑	☐	Winters Method (Additive)	104	50.177156	735.50239	743.52087	0.940	729.50239	0.000006		3.315726	6.109066
☑	☐	ARI(1, 1)	108	46.320348	736.88434	742.28530	0.949	732.88434	0.000003		2.877901	5.367753
☑	☐	AR(1)	109	99.891165	830.57245	835.99151	0.878	826.57245	0.000000		4.290917	7.913439

As shown above, the AR(2,1) model is an improvement as indicated by all measures of fit.

Scenario 2

a. Student answers will vary. Responses should note that Durables show a marked upward trend with likely seasonal component. Below are summary results for several reasonable approaches. Among the methods available through the Time Series platform, Linear Exponential Smoothing outperforms the others according to the measures we have studied. The adjusted RSquare statistics for the regression-based models are inferior to all but the AR(1) model, as follows: Linear, (.854), Quadratic (.855), LogLinear (.867).



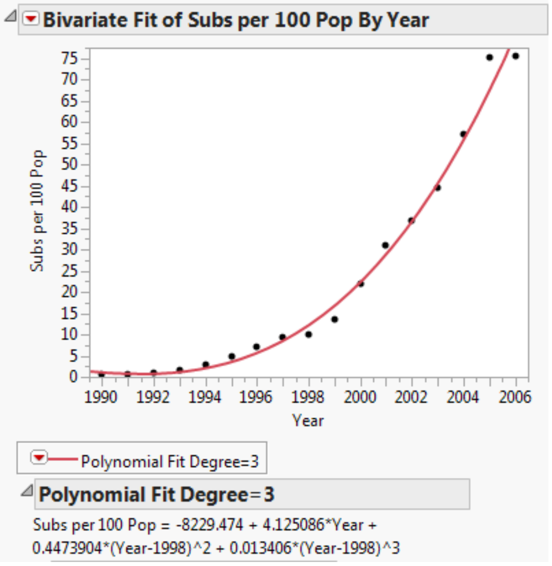
Model Comparison

Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	Weights	.2 .4 .6 .8	MAPE	MAE
☑	☐	Winters Method (Additive)	104	528.09261	989.97456	997.99305	0.873	983.97456	0.995917	▨	5.876557	18.367321
☑	☐	Linear (Holt) Exponential Smoothing	107	519.71387	1000.9686	1006.3513	0.875	996.96864	0.004082		5.741653	17.872988
☑	☐	ARI(1, 1)	108	597.81201	1017.5667	1022.9677	0.870	1013.5667	0.000001		5.840147	18.695112
☑	☐	AR(1)	109	689.3022	1044.5854	1050.0045	0.832	1040.5854	0.000000		6.783594	21.001116

Scenario 3

a. This is an annual series and therefore there can be no seasonal component.

c. Student answers will vary. For the Malaysia data, a 3rd degree polynomial (cubic) model provides a very good fit:

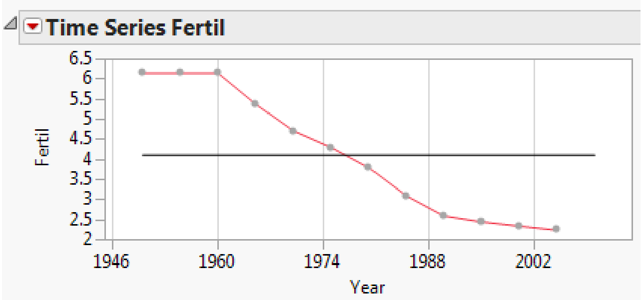


Subs per 100 Pop = -8229.474 + 4.125086*Year + 0.4473904*(Year-1998)^2 + 0.013406*(Year-1998)^3

e. These countries are all best approximated by different models. Effective time-series modeling requires the use of a variety of approaches.

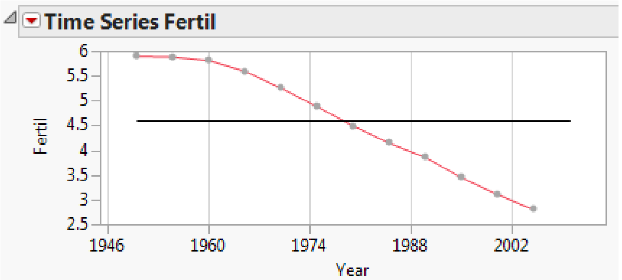
Scenario 4

a. The fertility rate in Brazil has declined following an S-shaped curve:



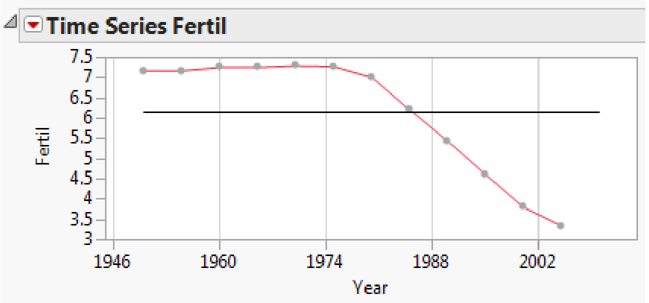
An AR(1,1) model fits modertately well, with relatively high RSquare (0.969), low variance (0.077) and MAPE and MAE of 5.35% and 0.20 respectively.

c.



India’s decline is very regular, especially since 1960. Linear Exponential Smoothing (Holt’s method) and AR(1,1) models both fit extremely well.

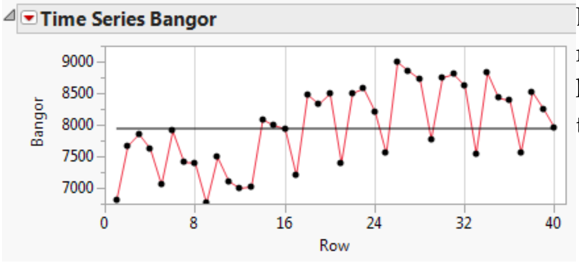
e.



Saudi Arabia’s decline is very regular, especially since 1980. Linear Exponential Smoothing (Holt’s method) and AR(1,1) models both fit extremely well.

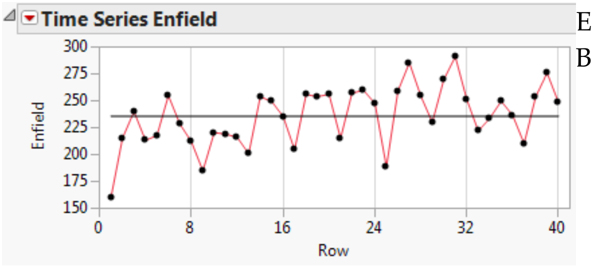
Scenario 5

a.



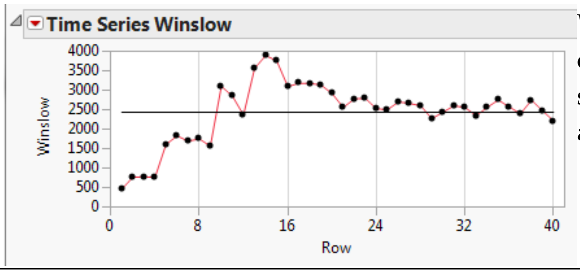
Bangor: For this series, an AR(4,1) works moderately well. The strong seasonal element here suggests that points are correlated with the observation 4 quarters earlier.

c.



Enfield: This pattern is much like the one in Bangor; Once again an AR(4,1) model fits well.

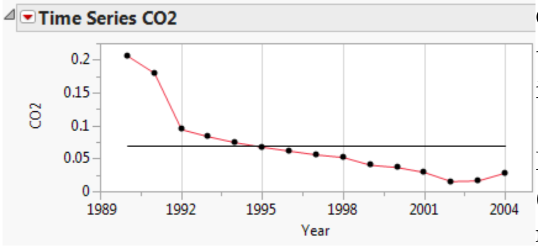
e.



Winslow: Here we see the dramatic change occurring roughly half-way through the time series. Simple exponential smoothing provides are reasonably good model.

Scenario 6

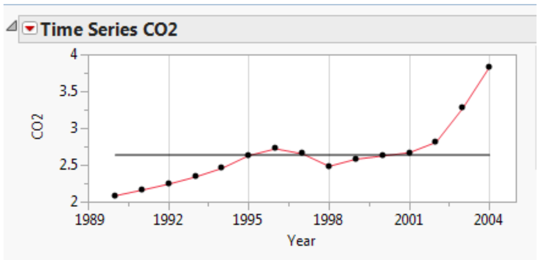
a.



CO2 emissions in Afghanistan have fallen since the series began, and have leveled off (with minor increases) in most recent years.

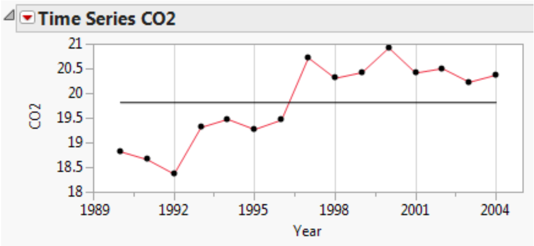
For this series, a log-linear model fits quite well (Rsqr =0.905). The other time series methods do not fit quite as well, though an AR(1,1) provides a good fit.

c.



In sharp contrast to the prior two graphs, China's CO2 emissions have been rapidly rising. A 3rd-degree polynomial (cubic) provides a moderately good fit, as does AR(1,1).

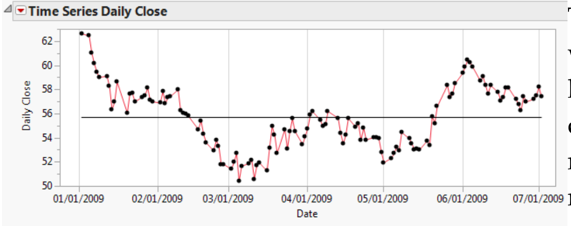
e.



CO2 emissions in the US rose for much of the period and seem to have leveled off, presenting a quite different pattern from the prior 4 nations. A 2nd degree polynomial fits best.

Scenario 7

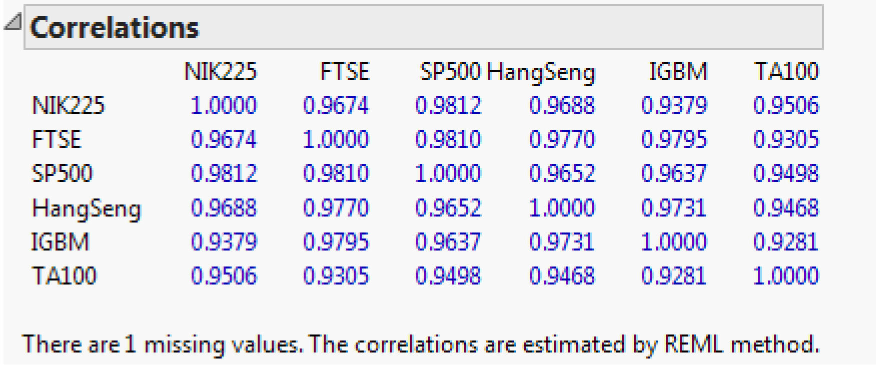
a.



The series to the left (expanded for clarity) would be poorly described with any type of linear trend model because it exhibits several changes of direction. Because we have just 6 months of data, we should not use Winter's method which accounts for seasonal variation.

Scenario 8

a.



Correlations

	NIK225	FTSE	SP500	HangSeng	IGBM	TA100
NIK225	1.0000	0.9674	0.9812	0.9688	0.9379	0.9506
FTSE	0.9674	1.0000	0.9810	0.9770	0.9795	0.9305
SP500	0.9812	0.9810	1.0000	0.9652	0.9637	0.9498
HangSeng	0.9688	0.9770	0.9652	1.0000	0.9731	0.9468
IGBM	0.9379	0.9795	0.9637	0.9731	1.0000	0.9281
TA100	0.9506	0.9305	0.9498	0.9468	0.9281	1.0000

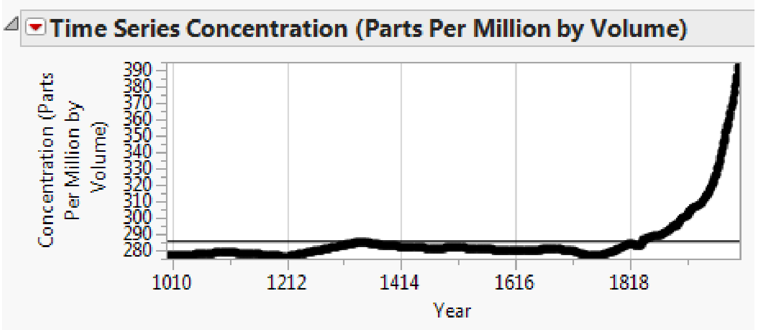
There are 1 missing values. The correlations are estimated by REML method.

The Nikkei225 has the highest correlation with the S&P500 (0.9812) and the FTSE100 is close behind with $r = 0.9810$)

c. Yes. Both markets are engaged in competition in the same global markets, and move very closely together as indicatd by their very high correlation.

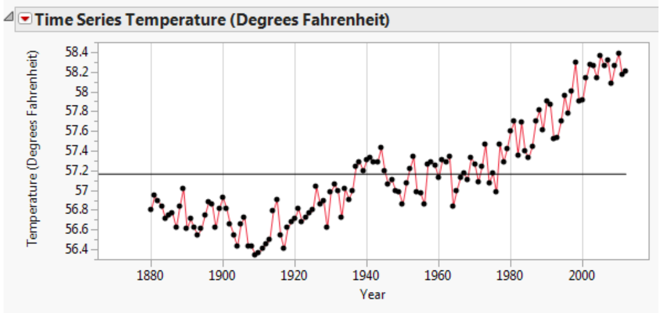
Scenario 9

a.



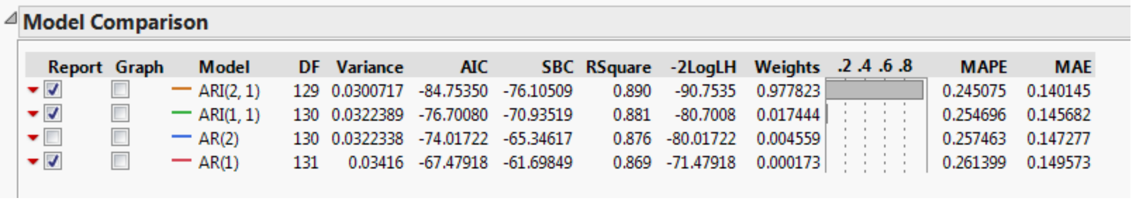
This is a non-stationary annual series with a curvilinear trend since approximately 1800. We saw in the previous chapter that a quadratic trend model fit to a degree, but this pattern is probably better modeled as an autoregressive process.

c.



This non-stationary annual series shows considerable variability. A trend model won't capture the year-to year oscillations, but an autoregressive model will.

Looking at the summary of 4 different AR models (below) it appears that the AR(2,1) model performs best. The estimates using that model are shown below the Model Comparison table.



Model Comparison

Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	Weights	.2 .4 .6 .8	MAPE	MAE
☑	☐	ARI(2, 1)	129	0.0300717	-84.75350	-76.10509	0.890	-90.7535	0.977823		0.245075	0.140145
☑	☐	ARI(1, 1)	130	0.0322389	-76.70080	-70.93519	0.881	-80.7008	0.017444		0.254696	0.145682
☐	☐	AR(2)	130	0.0322338	-74.01722	-65.34617	0.876	-80.01722	0.004559		0.257463	0.147277
☑	☐	AR(1)	131	0.03416	-67.47918	-61.69849	0.869	-71.47918	0.000173		0.261399	0.149573

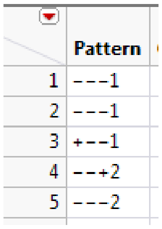
2013	58.272704408
2014	58.257001165
2015	58.262675668
2016	58.28208585
2017	58.290882083

21

Student Solutions to Application Scenarios

Scenario 1

a.



	Pattern
1	---1
2	---1
3	+--1
4	--+2
5	---2

The first 5 rows are shown to the left.

c. Assuming we follow the example presented in the chapter, we now have 50 experimental runs, the first 10 of which are assigned to team member #1. Each team member will perform 10 of the 16 possible runs, with each member having a slightly different pattern assigned randomly.

Scenario 2

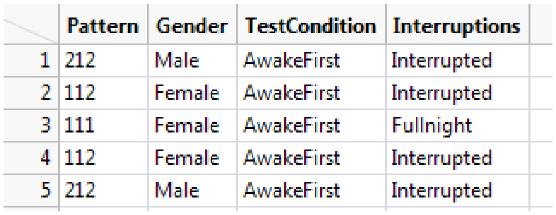
a. There will be 32 runs in a Resolution IV, full-factorial design.

c. [NOTE: The question should read: "Briefly explain what happens when we move from a **two**-factor screening design to a five-factor design."]

In a two-factor screening design there would be just four runs (2^2) and the five-factor model has $2^5 = 32$ runs.

Scenario 3

a.



	Pattern	Gender	TestCondition	Interruptions
1	212	Male	AwakeFirst	Interrupted
2	112	Female	AwakeFirst	Interrupted
3	111	Female	AwakeFirst	Fullnight
4	112	Female	AwakeFirst	Interrupted
5	212	Male	AwakeFirst	Interrupted

The first five rows of the data table, including Patterns, are shown above.

c. With 72 subjects, the prediction profiler shows that the variance ranges from approximately 0.042 to approximately 0.056. With 144 subjects, the corresponding variance range is reduced by half, ranging from approximately 0.021 to 0.028.

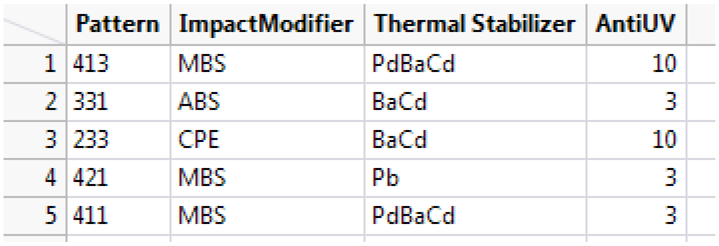
Scenario 4

a. Categorical factors: type of incentive, timing of incentive, survey mode, guarantee vs. lottery. Continuous factors: Duration of survey, number of contacts made, and amount of money offered.
[some students might classify "burden" of survey as categorical.]

c. Assuming that we use minimal number of factor levels described in b, and two factor levels for the continuous factors, we would have four dichotomous categorical factors and three continuous factors. This would, then, require 2 x 2 x 2 x 2 x 2 x 2 x 2 = 2^7 = 128 runs.

Scenario 5

a. Here are the first five rows of the table:

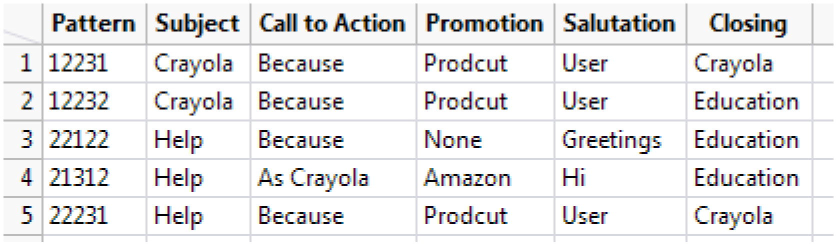


	Pattern	ImpactModifier	Thermal Stabilizer	AntiUV
1	413	MBS	PdBaCd	10
2	331	ABS	BaCd	3
3	233	CPE	BaCd	10
4	421	MBS	Pb	3
5	411	MBS	PdBaCd	3

c. The full-factorial design has 528 runs and the fractional custom design has 35. In the initial design, the Anti-UV additive is tested at levels of 3, 5 and 10 with each of the three tested in one-third of the runs. In the revised design, the levels are 3, 6.5, and 10. The profiler for the custom design reveals substantial interactions among the three factors; the fractional design can detect these, but the loss of resolution in this design could be costly.

Scenario 6

a. This table has 72,072 rows. Here are the first five:



	Pattern	Subject	Call to Action	Promotion	Salutation	Closing
1	12231	Crayola	Because	Prodcut	User	Crayola
2	12232	Crayola	Because	Prodcut	User	Education
3	22122	Help	Because	None	Greetings	Education
4	21312	Help	As Crayola	Amazon	Hi	Education
5	22231	Help	Because	Prodcut	User	Crayola

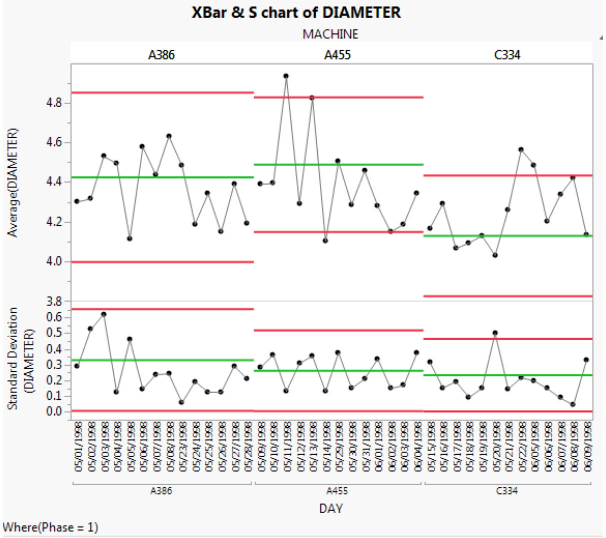
c. In the full factorial design, every combination of all levels the five factors (2 x 3 x 2 x 3 x 2 = 72) is tested whereas in the reduced custom design, far fewer are tested because interactions are limited to two factors at a time.

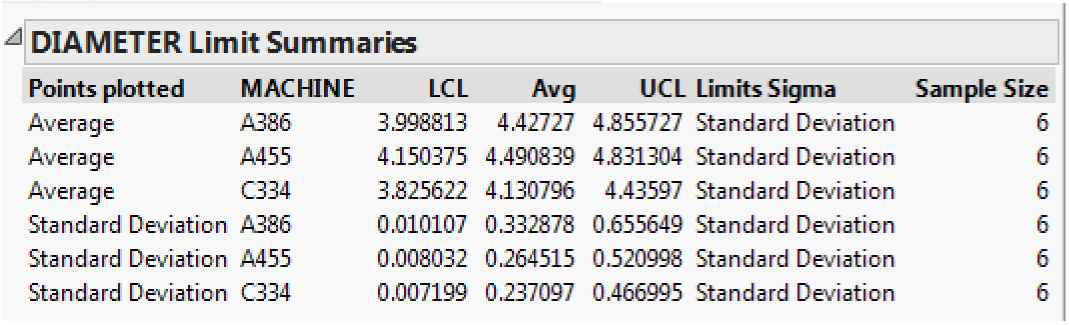
22

Student Solutions to Application Scenarios

Scenario 1

a.



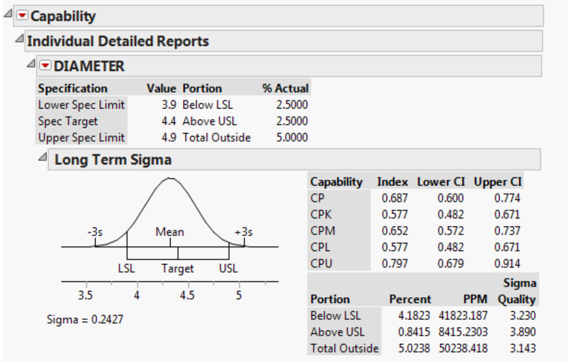


DIAMETER Limit Summaries

Points plotted	MACHINE	LCL	Avg	UCL	Limits Sigma	Sample Size
Average	A386	3.998813	4.42727	4.855727	Standard Deviation	6
Average	A455	4.150375	4.490839	4.831304	Standard Deviation	6
Average	C334	3.825622	4.130796	4.43597	Standard Deviation	6
Standard Deviation	A386	0.010107	0.332878	0.655649	Standard Deviation	6
Standard Deviation	A455	0.008032	0.264515	0.520998	Standard Deviation	6
Standard Deviation	C334	0.007199	0.237097	0.466995	Standard Deviation	6

As we can see in the graphs above, Machine C334 may have an unstable standard deviation and machine A455 shows two sample means beyond the control limits. These machines should be inspected closely for possible adjustment.

c.



Capability

Individual Detailed Reports

DIAMETER

Specification	Value	Portion	% Actual
Lower Spec Limit	3.9	Below LSL	2.5000
Spec Target	4.4	Above USL	2.5000
Upper Spec Limit	4.9	Total Outside	5.0000

Long Term Sigma

Sigma = 0.2427

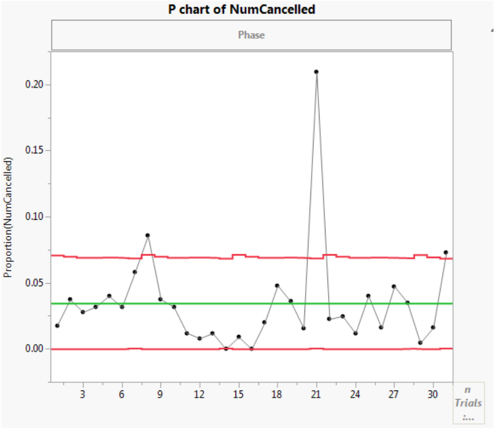
Capability	Index	Lower CI	Upper CI
CP	0.687	0.600	0.774
CPK	0.577	0.482	0.671
CPM	0.652	0.572	0.737
CPL	0.577	0.482	0.671
CPU	0.797	0.679	0.914

Portion	Percent	PPM	Sigma Quality
Below LSL	4.1823	41823.187	3.230
Above USL	0.8415	8415.2303	3.890
Total Outside	5.0238	50238.418	3.143

This capability analysis shows that 5% of the observations lie outside the capability limits, indicating that the process is capable of producing tubing that is within .5 mm of 4.4.

Scenario 2

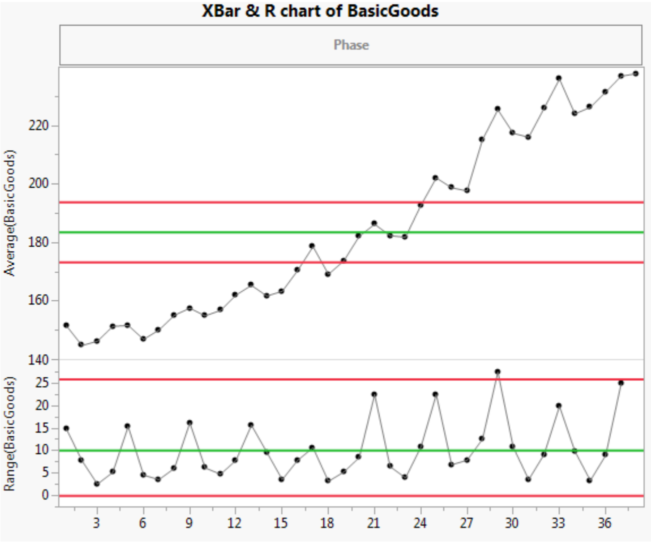
a.



This process is out of control at three points. Because a day with 0 cancellations is desirable, we should not be concerned about dates with values below the LCL. However, the chart shows 3 date well above the UCL.

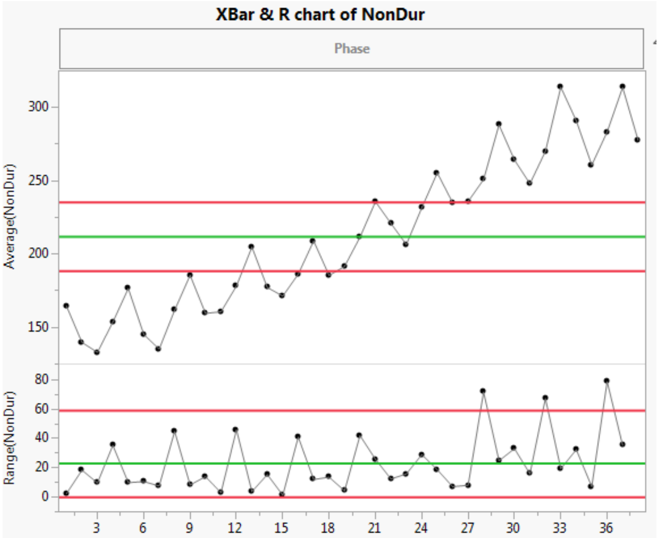
Scenario 3

a.



Production of basic goods has been rising steadily over time, which is a good thing. This is not a process designed for a constant target, but rather one of continuous growth.

c.



Once again we see a steady pattern of growth, with clear seasonal variation. In contrast to the control chart for Basic Goods, the one for NonDurables may exhibit a more linear upward trend, and substantial growth in variability (the R Chart) in the most recent years.

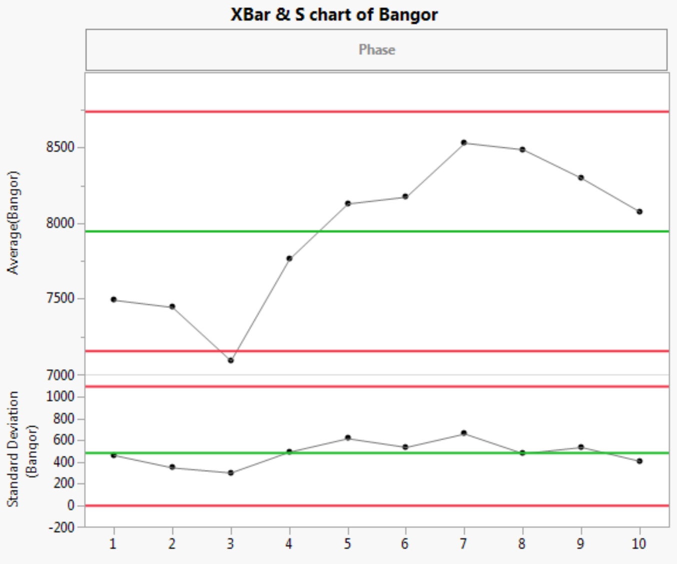
Because the need for basic goods probably follows the growth in population we might expect steady growth akin to population trends.

Scenario 4

a. In most regions except for the Southwest the standard deviations are sufficiently unstable that we should not interpret the Xbar charts. In the Southwest, the standard deviations have been steadily increasing but the limited data (only five sample mean) indicates increasing mean times to restore the area to safety, but still within control limits.

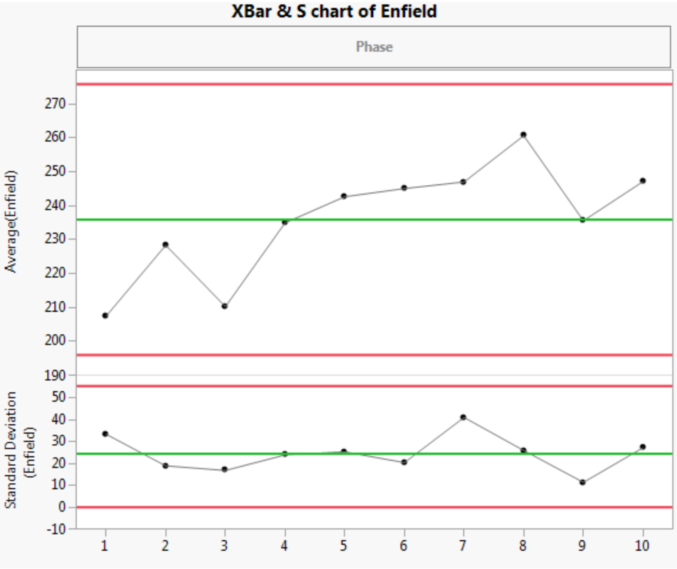
Scenario 5

a.



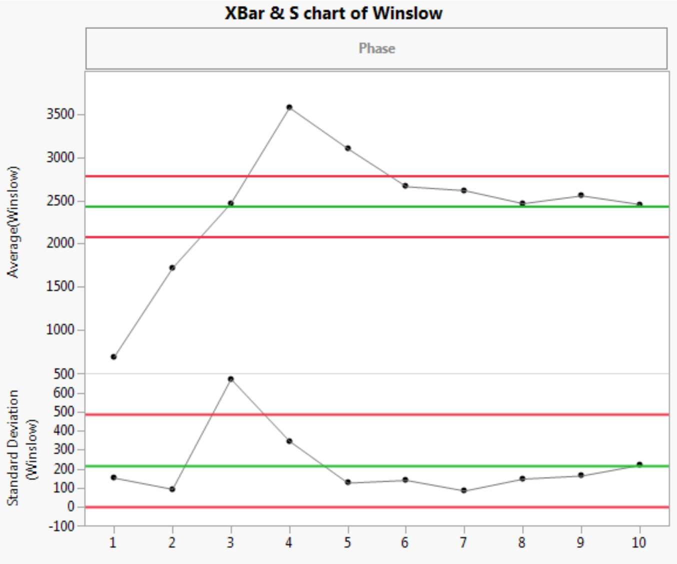
Bangor: The S chart is stable; early in the study period there was one year below the LCL. Otherwise Bangor has remained within limits, though the 6 most recent years have been above average.

c.



Enfield: This pattern is much like the one in Bangor. The S chart is stable throughout. Otherwise Enfield has remained within limits, though the 5 of the 6 most recent years have been above average

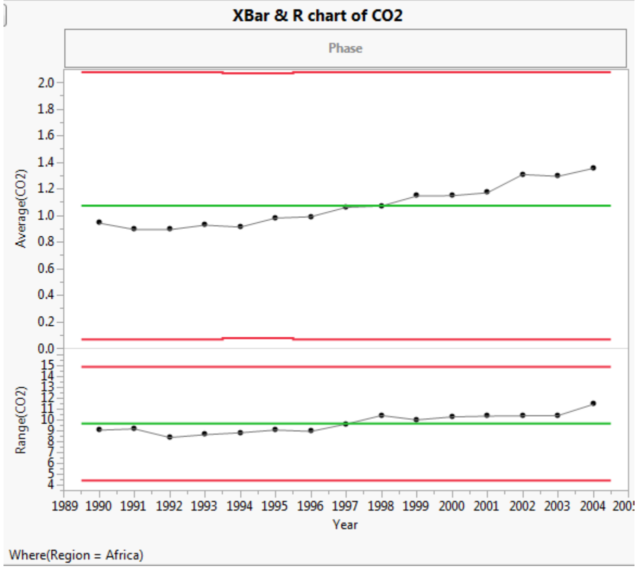
e.



Winslow: In year 3 the S chart (not shown) shows the sample standard deviation above the UCL; otherwise the standard deviations are moderately stable. The Xbar chart shows a process out of control until year 6, after which the process seems to be in control.

Scenario 6

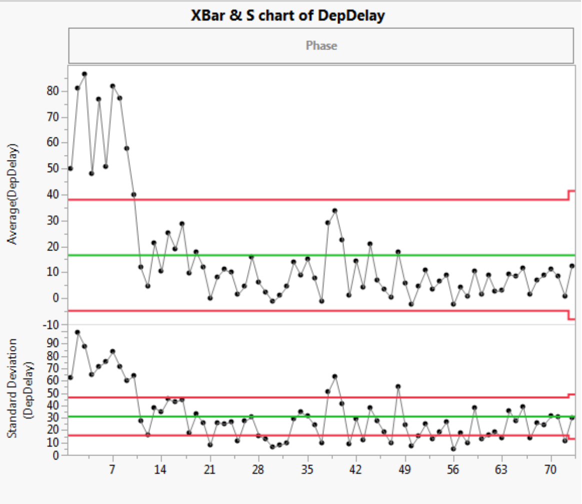
a.



Emissions in most regions are relatively stable In Africa (shown to the left), both the ranges and means have been steadily rising over the 15-year period.

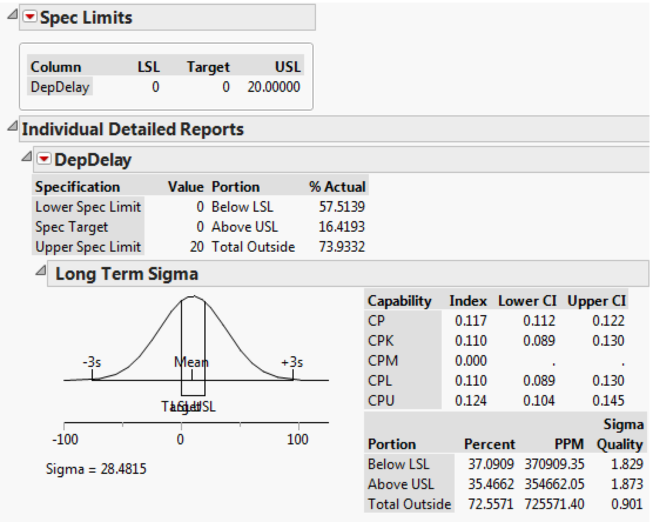
Scenario 7

a.



Given the instability in the standard deviations, we should be reluctant to interpret the Xbar chart. However, we might observe that for roughly the first 10 samples both the standard deviations and means tended to be substantially higher than for the remainder of the period. It would appear that there was a fundamental process change leading to shorter and more predictable departure delays sometime around the 10th sample.

c.



Spec Limits

Column	LSL	Target	USL
DepDelay	0	0	20.00000

Individual Detailed Reports

DepDelay

Specification	Value	Portion	% Actual
Lower Spec Limit	0	Below LSL	57.5139
Spec Target	0	Above USL	16.4193
Upper Spec Limit	20	Total Outside	73.9332

Long Term Sigma

Sigma = 28.4815

Capability	Index	Lower CI	Upper CI
CP	0.117	0.112	0.122
CPK	0.110	0.089	0.130
CPM	0.000	.	.
CPL	0.110	0.089	0.130
CPU	0.124	0.104	0.145

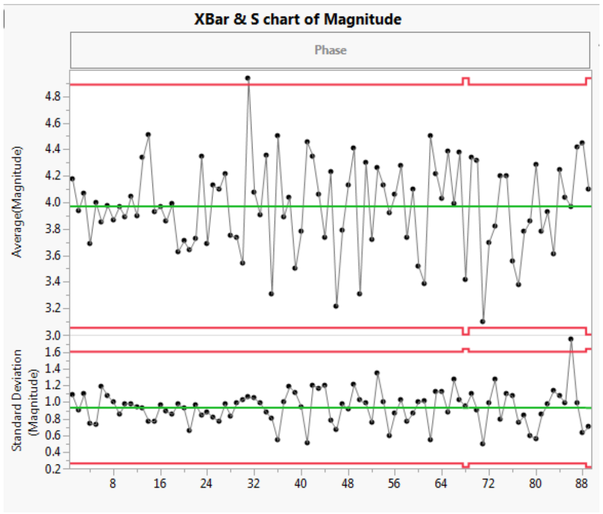
Portion	Percent	PPM	Sigma Quality
Below LSL	37.0909	370909.35	1.829
Above USL	35.4662	354662.05	1.873
Total Outside	72.5571	725571.40	0.901

We need to use the Data Filter to select only the weekday flights. In the Capability analysis, the critical capability limit here is the USL, which we set at 20 minutes; the other values may be set to zero

We see that 16% of the flights exceeded delays of more than 20 minutes. Therefore the current process is not capable of meeting the goal.

Scenario 8

a.



It appears that the variability of the process standard deviation has increased over time, with one recent S above the UCL. Nearly all of the sample means are within the control limits; early in the observation period (roughly the first 15 samples) the mean magnitudes remained quite close to 4.0. Since that time, the fluctuations in mean magnitude have increased even as the mean appears to have remained stable.