



Practical Data Analysis with JMP[®]

Student Solutions

Robert Carver

**THE
POWER
TO KNOW[®]**

This set of Student Solutions is a companion piece to the following SAS Press book: Carver, Robert. Practical Data Analysis with JMP®. Copyright © 2010, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

Practical Data Analysis with JMP®

Copyright © 2010, SAS Institute Inc., Cary, NC, USA

ISBN 978-1-60764-475-0

ISBN 978-1-60764-487-3 (electronic book)

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.

U.S. Government Restricted Rights Notice: Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227-19, Commercial Computer Software-Restricted Rights (June 1987).

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

1st printing, July 2010

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at support.sas.com/publishing 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 registered trademarks or trademarks of their respective companies.

Student Solutions

Chapter 2

Scenario 1

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

Scenario 2

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

Scenario 3

Student answers will vary. Excel sheet should be imported into JMP.

Scenario 4

This data table contains significant 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 5

The sampling weight is equal across the stratified sample at approximately 1.6, reflecting the fact that there are approximately equal numbers of countries within each of the four quartiles.

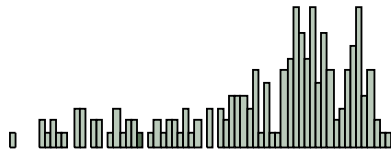
2 *Practical Data Analysis with JMP*

Student Solutions

Chapter 3

Scenario 1

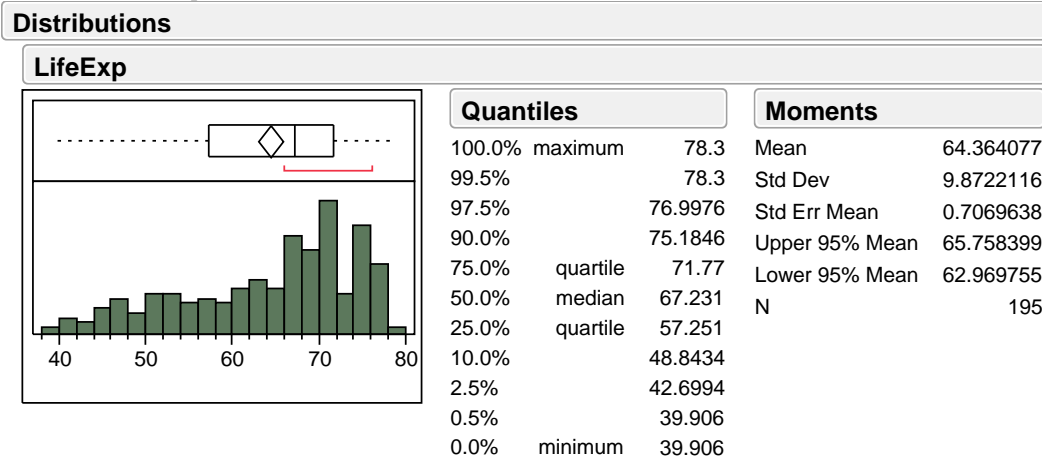
- a. Using the grabber tool, click and drag upwards to increase the number of bars in the histogram. A second peak near 80 appears when as the number of bars increases, while the peak at 75 remains.



- 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

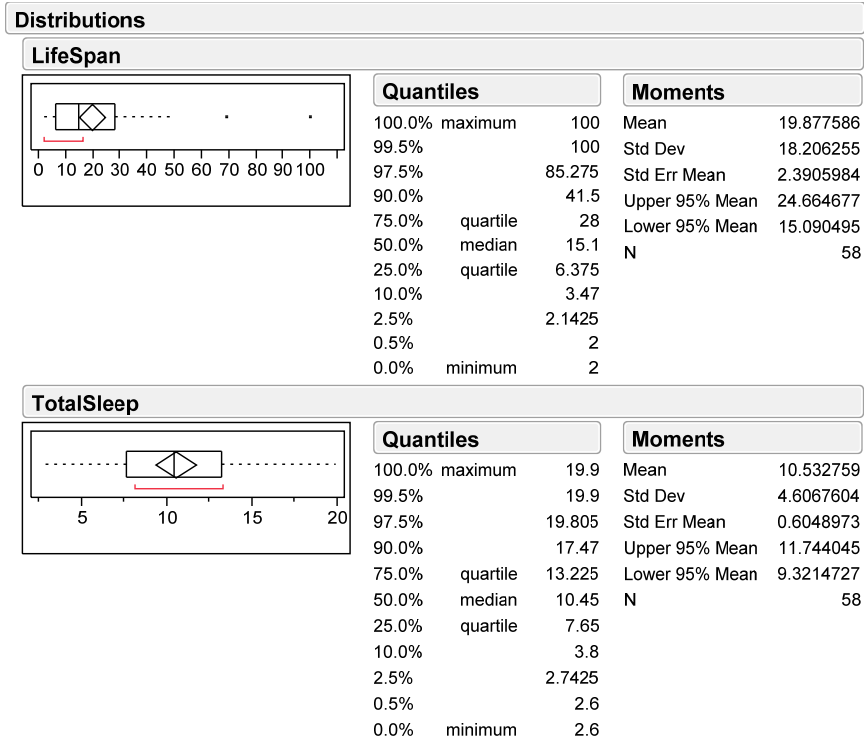


- c. The standard deviation is 9.87 in the 1985 data compared to 10.4 in the 2010 data.

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.

2 Practical Data Analysis with JMP



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.

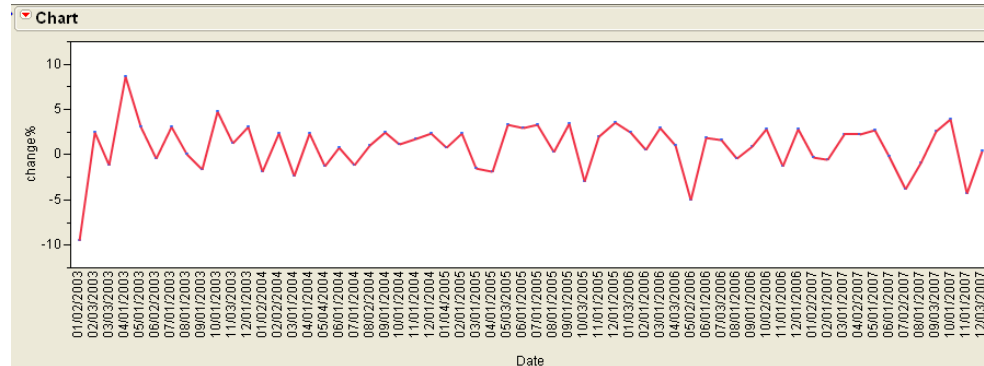
g. The animals that sleep in the most exposed locations are also the largest in terms of body weight. This may be because larger animals cannot hide as easily, or due to sheer size, they can sleep in exposed locations safely.

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.

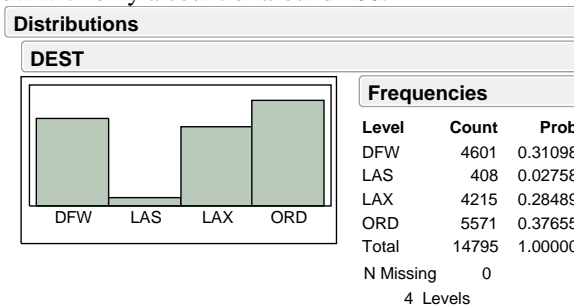
c. The FTSE declines approximately 25% of the time.

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.



- 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.

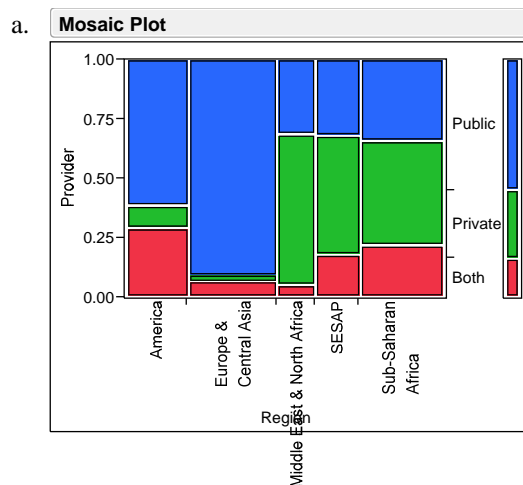
Scenario 6

- 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.

Student Solutions

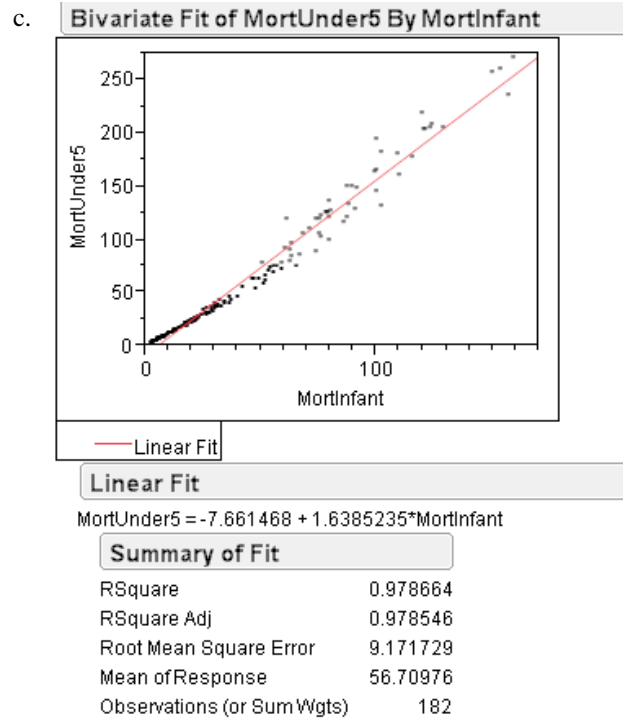
Chapter 4

Scenario 1



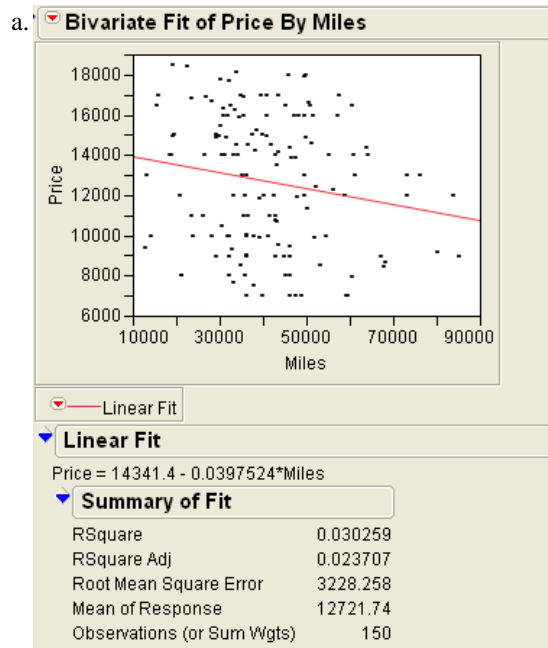
Public provision is most common by far in the Americas and Europe & Central Asia. Private provision seems to be the norm in the rest of the world. Most areas have relatively few countries with both public and private, though such arrangements are fairly common (more than 25% of countries) in the Americas.

2 Practical Data Analysis with JMP



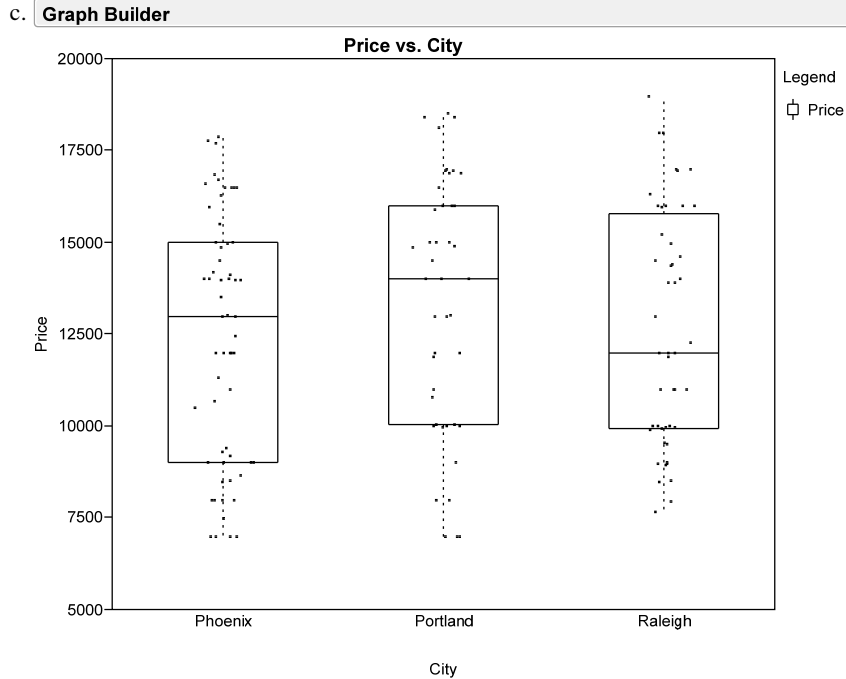
The fitted line and RSquare value are shown above. There is a strong, positive linear relationship between the two variables.

Scenario 2



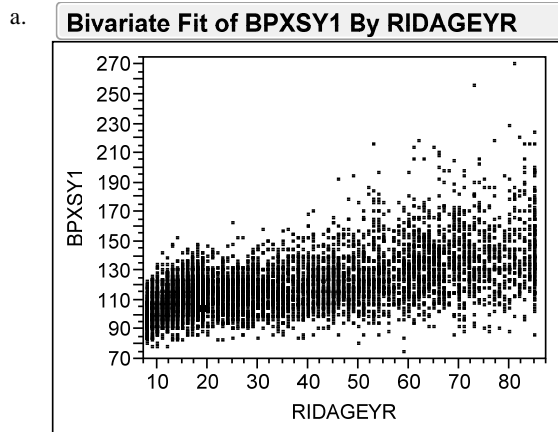
The plot, equation and Rsquare are shown above. The correlation coefficient is 0.17395. There is a weak negative relationship between mileage and price: the higher the mileage, the lower the price.

4 Practical Data Analysis with JMP



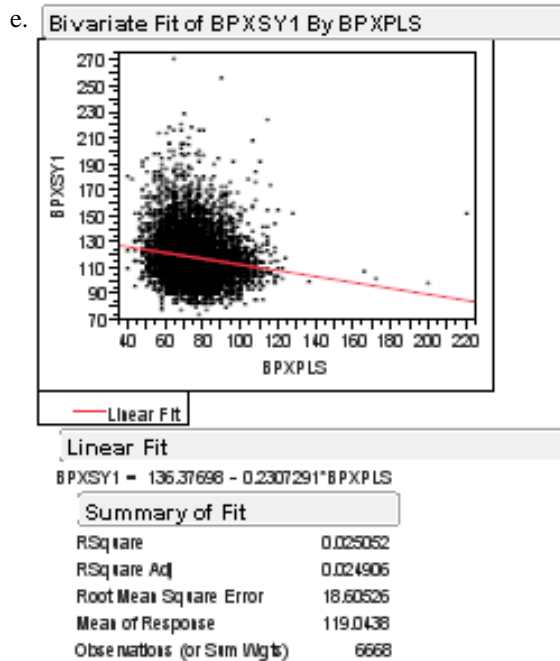
The distribution of price across the three cities seems to be fairly uniform. The box plot shows similar middle 50% with varying means. They also have very similar spreads.

Scenario 3



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.



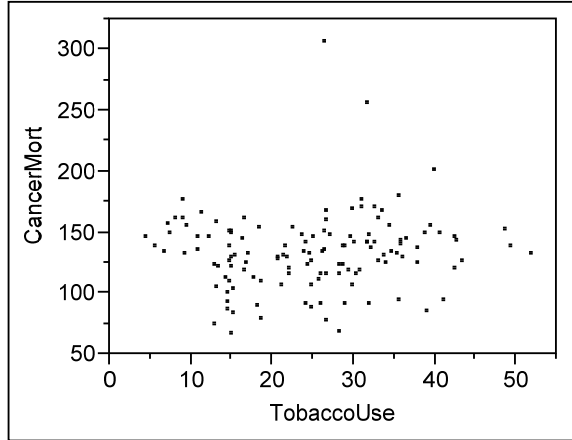
It appears there is little evidence of a relationship between pulse and blood pressure, as the r squared statistic is .02, which is very low.

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.

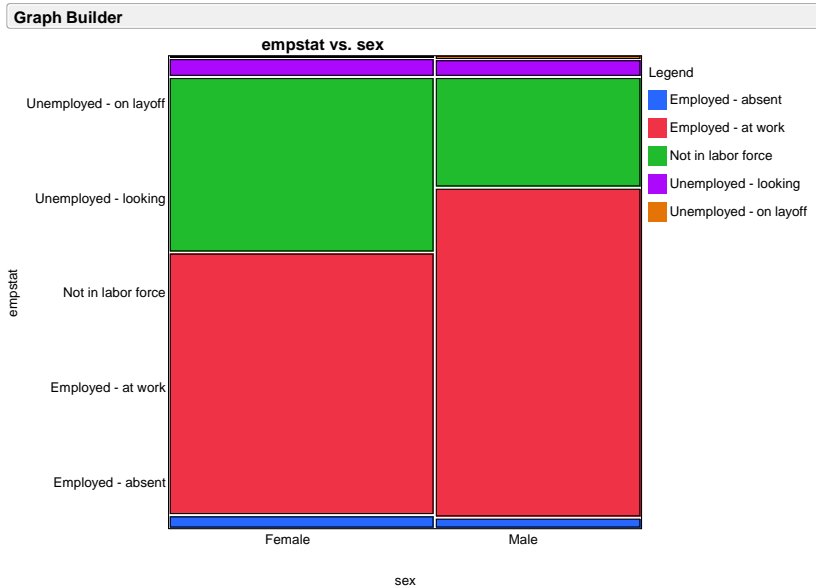
6 Practical Data Analysis with JMP

c. **Bivariate Fit of CancerMort By TobaccoUse**



Here again, we find scant evidence of a relationship.

Scenario 5



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.

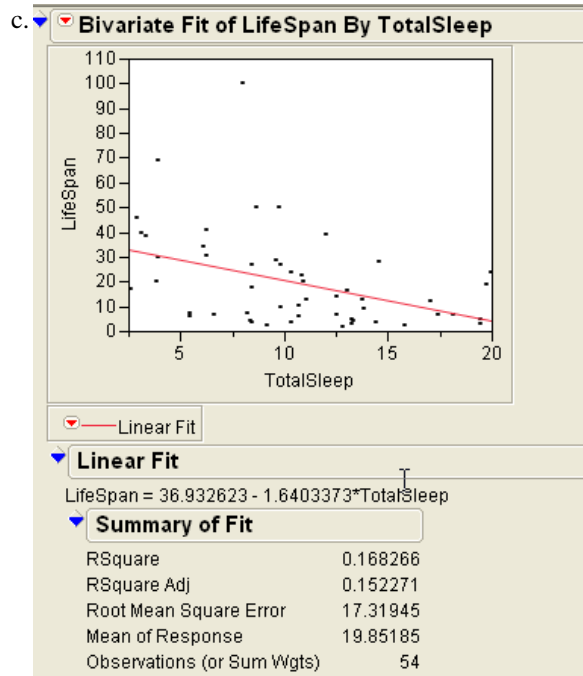
Scenario 6

a. **Contingency Table**

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

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

8 Practical Data Analysis with JMP



There seems to be little evidence of a relationship between lifespan and total sleep as the Rsquare statistic is only 0.168.

Student Solutions

Chapter 5

Scenario 1

NOTE: parts a through d use a contingency table like this one:

		Region					
		America	Europe & Central Asia	Middle East & North Africa	SESAAP	Sub-Saharan Africa	
MatLeave90+	Count						
	Total %						
	Col %						
	Row %						
	No	15	0	12	12	21	60
		9.26	0.00	7.41	7.41	12.96	37.04
		48.39	0.00	63.16	48.00	48.84	
		25.00	0.00	20.00	20.00	35.00	
	Yes	16	44	7	13	22	102
		9.88	27.16	4.32	8.02	13.58	62.96
	51.61	100.00	36.84	52.00	51.16		
	15.69	43.14	6.86	12.75	21.57		
	31	44	19	25	43	162	
	19.14	27.16	11.73	15.43	26.54		

a. $Pr(\text{Sub-Saharan Africa}) = 0.2654$

c. $Pr(\text{Longer than 90 and Sub-Saharan Africa}) = 0.1358$

NOTE: for the remaining parts of this problem, use a contingency table like this one:

		Region					
		America	Europe & Central Asia	Middle East & North Africa	SESAAP	Sub-Saharan Africa	
Provider	Count						
	Total %						
	Col %						
	Row %						
	Both	9	3	1	4	9	26
		5.73	1.91	0.64	2.55	5.73	16.56
		29.03	6.82	5.26	18.18	21.95	
		34.62	11.54	3.85	15.38	34.62	
	Private	3	1	12	11	18	45
		1.91	0.64	7.64	7.01	11.46	28.66
	9.68	2.27	63.16	50.00	43.90		
	6.67	2.22	26.67	24.44	40.00		
Public	19	40	6	7	14	86	
	12.10	25.48	3.82	4.46	8.92	54.78	
	61.29	90.91	31.58	31.82	34.15		
	22.09	46.51	6.98	8.14	16.28		
	31	44	19	22	41	157	
	19.75	28.03	12.10	14.01	26.11		

2 Practical Data Analysis with JMP

$$e.Pr(Both) = 0.1656$$

$$g.Pr(Both | SESAP) = 0.1818.$$

Scenario 2

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

		DMDMARTL								
Count	Married	Widowed	Divorced	Separated	Never Married	Living with Partner	Refused	Don't Know		
Total %										
Col %										
Row %										
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 d 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 3

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

		Binge Freq				Total %
		At least once a week	At least once a month	At least once a year	Never	
Accident	No	415	557	1071	1545	3588
	Yes	70	31	56	56	213
		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	
		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. $Pr(\text{Binge at least once a week}) = 0.1276$.

c. $Pr(\text{Accident}) = 0.0560$.

e. $Pr(\text{Accident} \mid \text{binge at least once a week}) = 0.1443$.

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 4

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

a. $Pr(\text{Not in labor force}) = 0.3122$

Parts c and d rely on this table:

		fullpart			Total %
		Full time	NIU (Not in universes)	Part time	
sex	Female	7700	7739	3179	18618
	Male	9083	3981	1286	14350
		23.36	23.47	9.64	56.47
		45.88	66.03	71.20	
		41.36	41.57	17.07	
		27.55	12.08	3.90	43.53
		54.12	33.97	28.80	
		63.30	27.74	8.96	
		16783	11720	4465	32968
		50.91	35.55	13.54	

c. $Pr(\text{Part-time or female}) = Pr(\text{part-time}) + Pr(\text{female}) - Pr(\text{part-time and female}) = 0.1354 + 0.5647 - 0.0964 = 0.6037$

e. **Frequencies**

Level	Count	Prob
Divorced	4193	0.12718
Married - spouse absent	393	0.01192
Married - spouse present	17088	0.51832
Never married	7721	0.23420
Separated	843	0.02557
Widowed	2730	0.08281
Total	32968	1.00000
N Missing	0	

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:

$$Pr(\text{Not Married}) = 0.12718 + 0.23420 + 0.08281 = 0.44419.$$

Scenario 5

a. $Pr(\text{Central}) = 0.2863$

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 $Pr(\text{Evacuation}) = 189/468 = 0.4038$.

e.

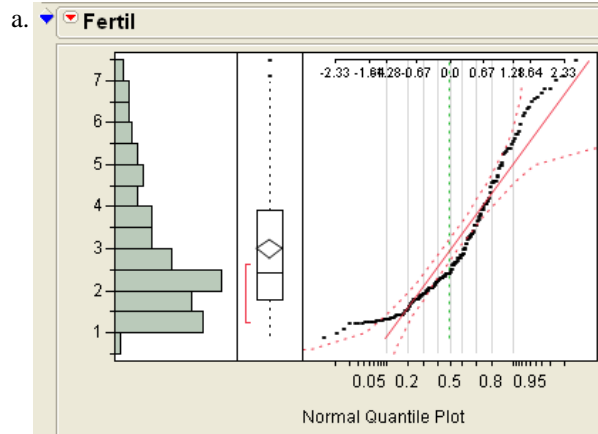
		LRTYPE_TEXT				
		LEAK	N/A	OTHER	RUPTURE	
EXPLO	Count					
	Total %					
	Col %					
	Row %					
	No	42	18	139	59	258
	Yes	41	7	70	32	150
		10.29	4.41	34.07	14.46	63.24
		50.60	72.00	66.51	64.84	
		16.28	6.98	53.88	22.87	
		10.05	1.72	17.16	7.84	36.76
		49.40	28.00	33.49	35.16	
		27.33	4.67	46.67	21.33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

$$Pr(\text{Rupture or Explosion}) = Pr(\text{Rupture}) + Pr(\text{Explosion}) - Pr(\text{Rupture and Explosion}) = 0.2230 + 0.3676 - 0.0784 = 0.5122.$$

Student Solutions

Chapter 6

Scenario 1

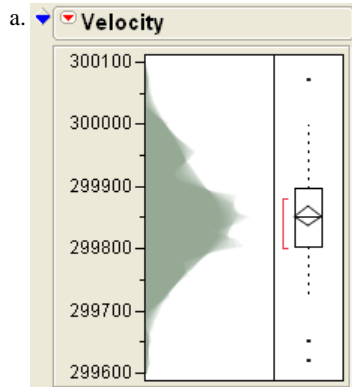


The normal quantile plot appears to the left. The distribution is strongly skewed positively, and therefore the normal model is not suitable for this variable.

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.

2 Practical Data Analysis with JMP

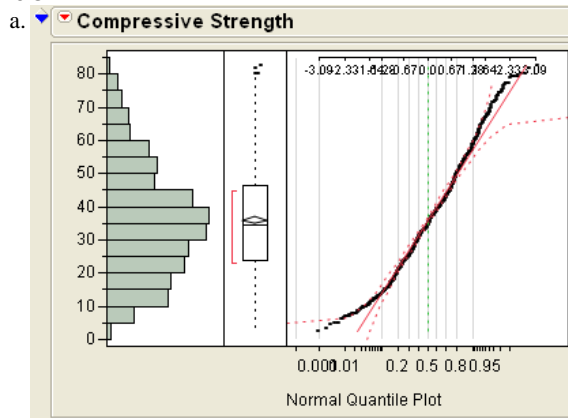
Scenario 2



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



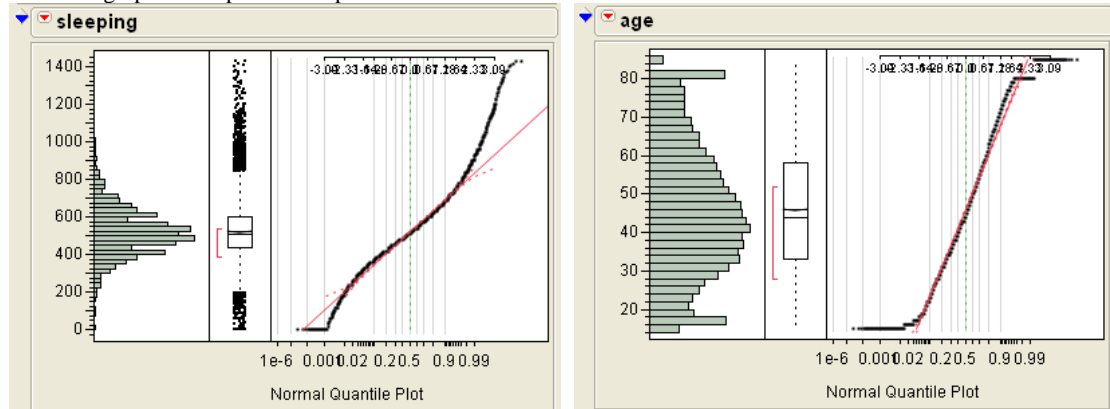
This distribution shows mild skewness. The lower tail is truncated and therefore shorter and thicker than a normal distribution would be.

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

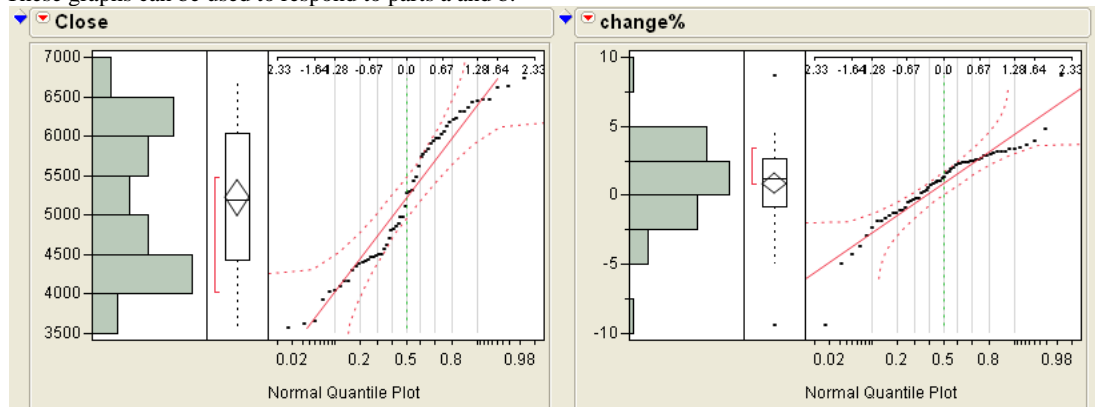
Use these graphs to respond to all parts:



- 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 than 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

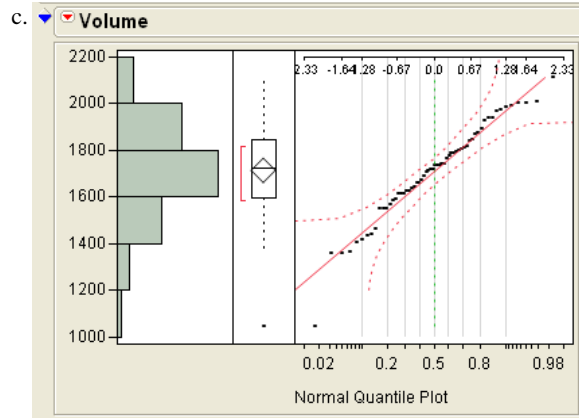
These graphs can be used to respond to parts a and b.



- 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%.

4 Practical Data Analysis with JMP



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

Student Solutions

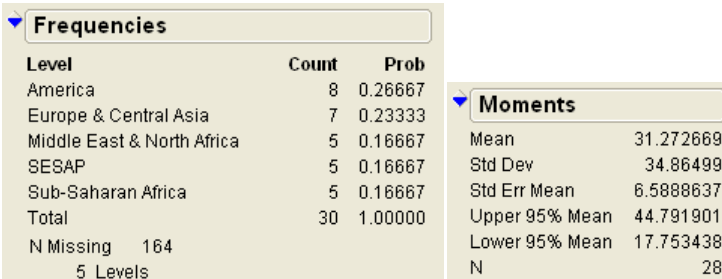
Chapter 7

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. 

Level	Count	Prob
America	8	0.26667
Europe & Central Asia	7	0.23333
Middle East & North Africa	5	0.16667
SESAP	5	0.16667
Sub-Saharan Africa	5	0.16667
Total	30	1.00000
N Missing	164	
5 Levels		

Moments	
Mean	31.272669
Std Dev	34.86499
Std Err Mean	6.5888637
Upper 95% Mean	44.791901
Lower 95% Mean	17.753438
N	28

Student answers will vary. 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 31.27 (note that in this sample only 28 of 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.

2 *Practical Data Analysis with JMP*

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.

Student Solutions

Chapter 8

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

		Hypoth	
Binomial Test	Level Tested	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. The sample does not 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.

2 Practical Data Analysis with JMP

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

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.

Student Solutions

Chapter 9

Scenario 1

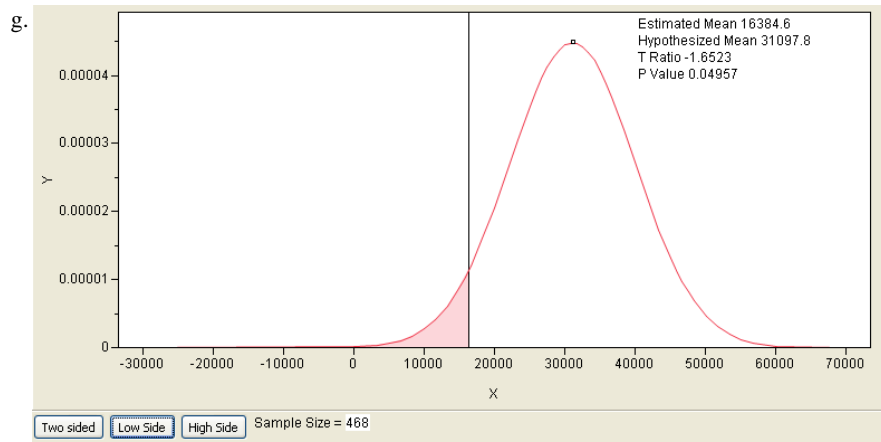
a. Probably. These columns contain continuous data, and though both distributions are strongly right-skewed, both have a sufficiently large number of observations to rely on the Central Limit Theorem. The critical question is whether we can view this particular time period as representative of the overall 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

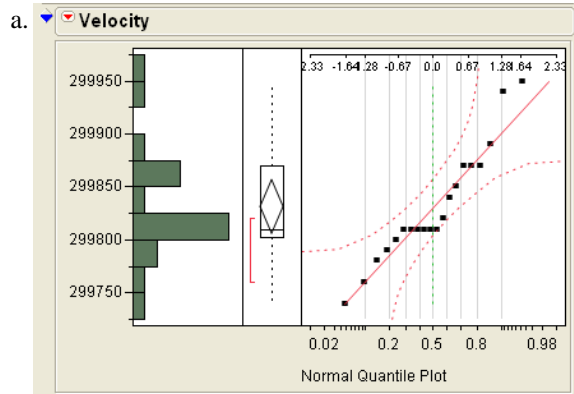
We can be 99% confident that the mean dollar cost of lost natural gas is between $-\$6646.48$ and $\$39,415.62$.



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.

2 Practical Data Analysis with JMP

Scenario 2



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 ($n = 6774$) we can rely on the Central Limit Theorem to proceed.

b. **Confidence Intervals**

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	11.37924	10.46774	12.29075	0.950
Std Dev	38.26994	37.63623	38.92551	0.950

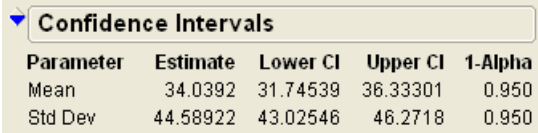
We can be 95% confident that the mean flight delay was somewhere between 10.47 and 12.29 minutes.

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. The screenshot shows a software interface with a dropdown menu labeled 'Confidence Intervals' and a table below it. The table has five columns: Parameter, Estimate, Lower CI, Upper CI, and 1-Alpha. The rows are Mean and Std Dev.

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	34.0392	31.74539	36.33301	0.950
Std Dev	44.58922	43.02546	46.2718	0.950

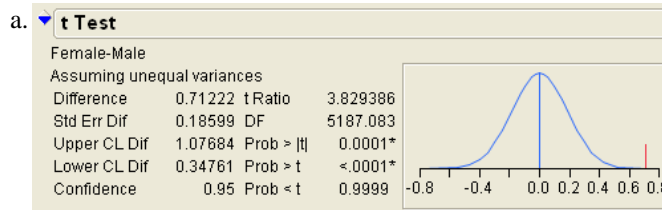
We can be 95% confident that the mean time from scheduled departure to wheels off is between 31.75 and 36.33 minutes.

Student Solutions

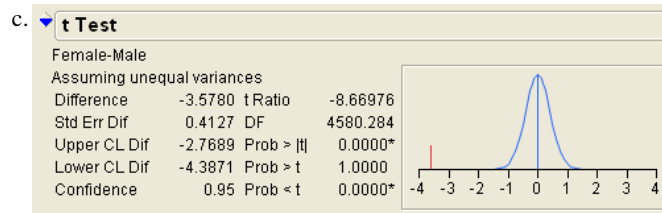
Chapter 10

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.



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



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

- 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 $+\$32,748,087$.
- We should first note that we have modest sample sizes ($n=35$ and $n=43$) from strongly skewed distributions. Therefore, we should use the Wilcoxon test rather than the t (results below). The relevant P-value is 0.3018, which is insufficient to reject the null hypothesis.

2 Practical Data Analysis with JMP

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)				
Level	Count	Score Sum	Score Mean	(Mean-Mean0)/Std0
SOUTHERN	35	1282.50	36.6429	-1.033
SOUTHWEST	43	1798.50	41.8256	1.033

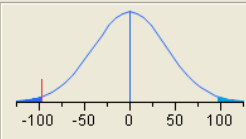
2-Sample Test, Normal Approximation			
S	Z	Prob> Z	
1282.5	-1.03259	0.3018	

1-way Test, ChiSquare Approximation		
ChiSquare	DF	Prob>ChiSq
1.0770	1	0.2994

Scenario 3

a. **t Test**

RUPTURE-LEAK			
Assuming unequal variances			
Difference	-97.68	tRatio	-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. **Test**

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	Score Mean	(Mean-Mean0)/Std0
0	8	72.500	9.0625	-2.569
1	24	455.500	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.

Student Solutions

Chapter 11

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	0.0039

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).

Scenario 2

a.

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

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*

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

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

2 Practical Data Analysis with JMP

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	0.2407
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**

		Binge Freq				
		At least once a week	At least once a month	At least once a y ear	Never	
Count						
Total %						
Col %						
Row %						
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	

Tests

N	DF	-LogLike	RSquare (U)
3801	3	33.676445	0.0069

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 bingeing 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*

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, making the Chi-Square test unreliable. That said, the test results point in the direction of rejecting the null hypothesis.

4 Practical Data Analysis with JMP

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*

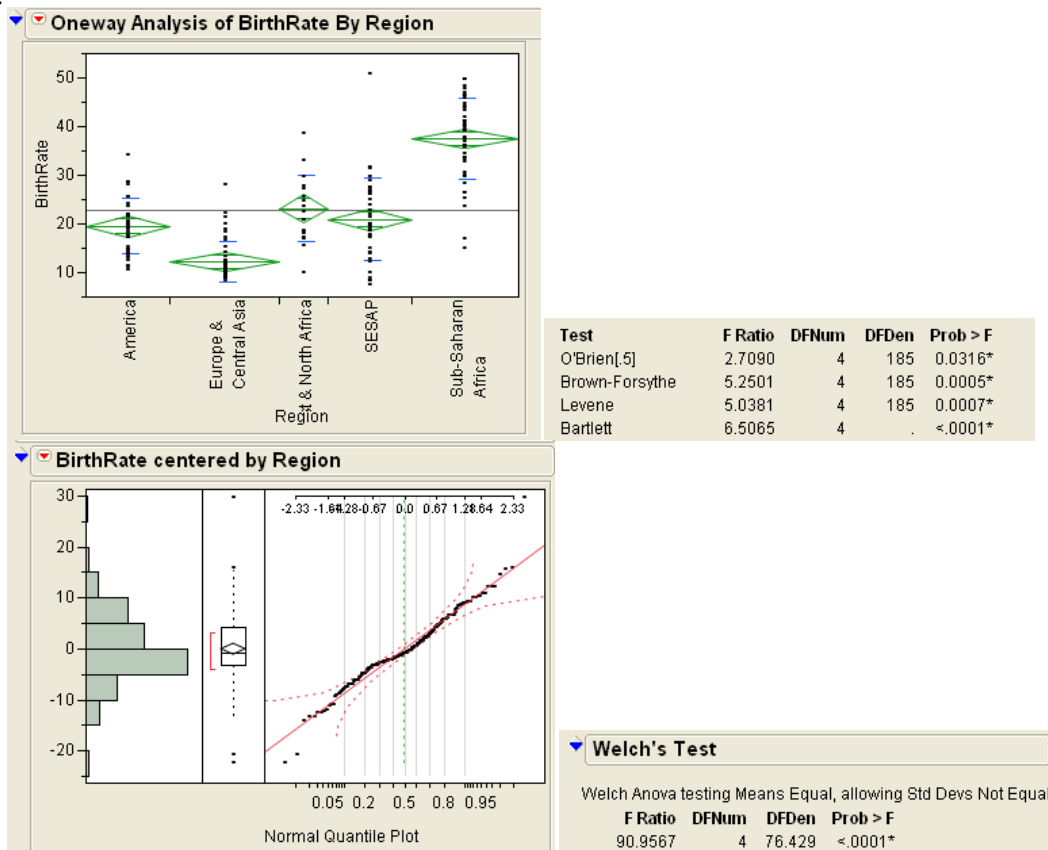
According to the Chi-Square test the two variables are not independent. There is sufficient evidence to reject a null hypothesis that they are independent.

Student Solutions

Chapter 12

Scenario 1

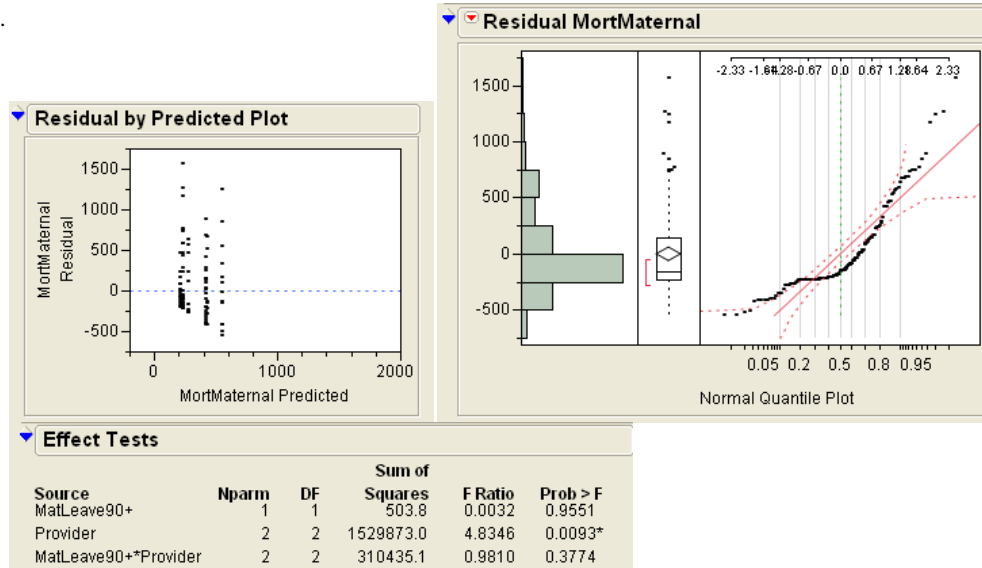
a.



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 Sub-Saharan birth rates are unusually high, and that birth rates in Europe and Central Asia are unusually low.

2 Practical Data Analysis with JMP

c.

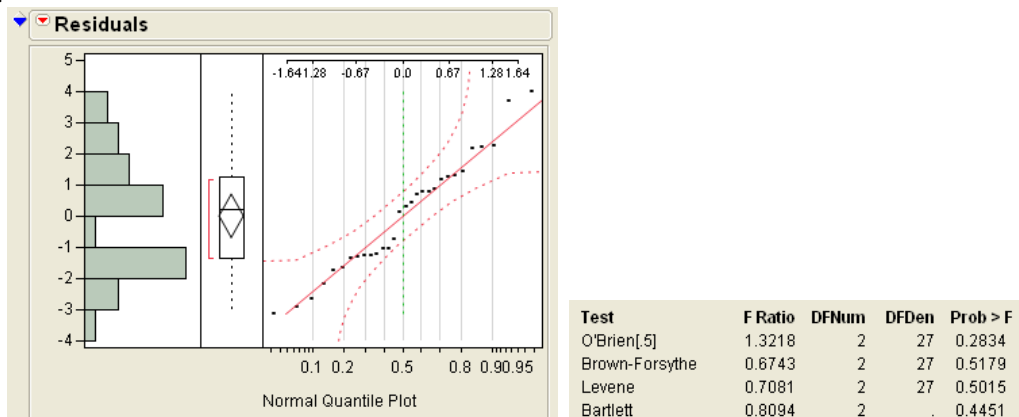


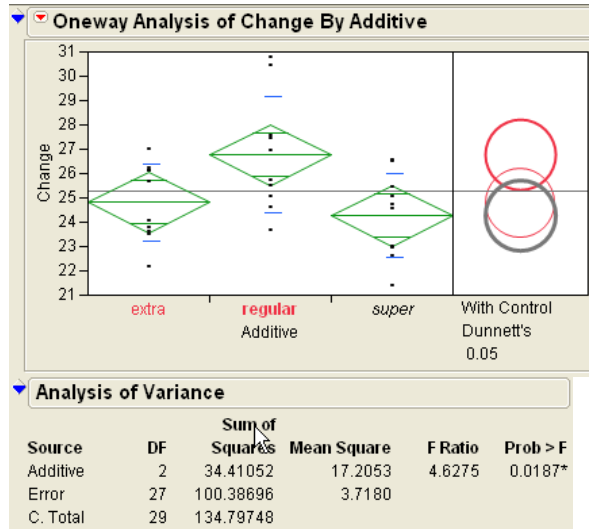
We start by evaluating conditions. The Residual by Predicted Plot raises some question about the equality of variances, but it is not definitive. The residuals do not appear to be normally distributed, but we have reasonably large samples and can rely on the Central Limit Theorem.

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.



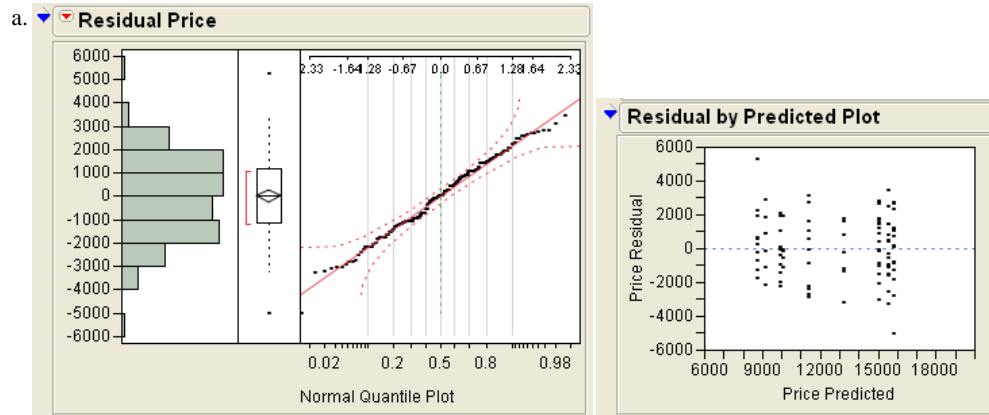


We see no evidence that the ANOVA assumptions have been violated; variances across the three groups appear to be equal and residuals are approximately normal. The F Ratio of 4.6275 and corresponding P-value of 0.0187 indicate that we should reject the null hypothesis of equal means; there is compelling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

- 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.

4 Practical Data Analysis with JMP

Scenario 3



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*

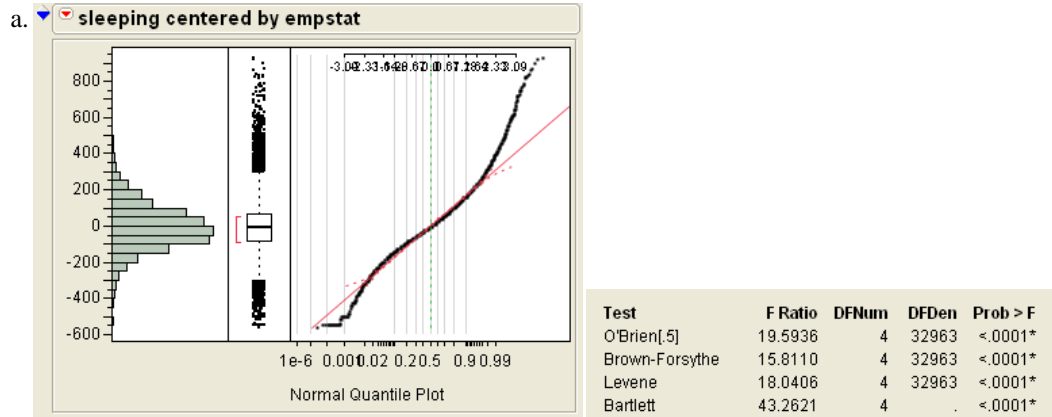
A review of the Effect Tests shows that we have a significant interaction effect as well as significant main effects. This tells us that prices vary by city and by model, and what's more the impact of model varies across the cities.

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

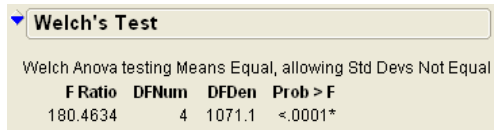
Levels not connected by same letter are significantly different.

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

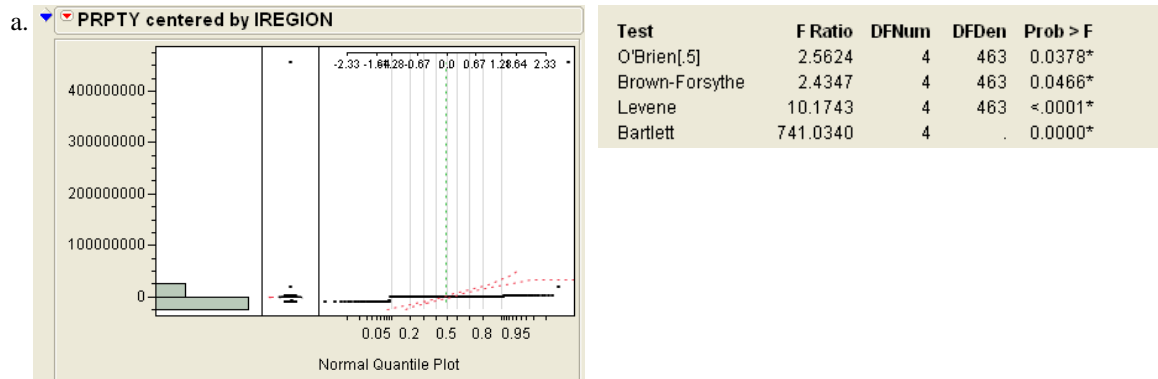


As usual we start by evaluating assumptions. We have a very large sample, so the Central Limit Theorem applies and we need not be concerned with normality (above we see the residuals are unimodal and symmetric, but depart from the normal model in the tails). We also see evidence that the variances are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



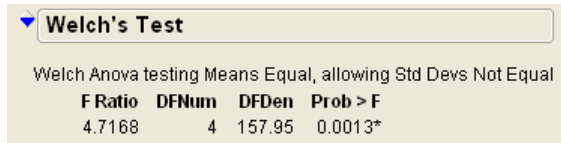
Both Welch's test and the standard ANOVA results strongly indicate that there are significant differences in group means. There is no control group here. Tukey's HSD indicates that employed people at work get the least sleep and unemployed people who are looking report the most. All others are significantly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

Scenario 5



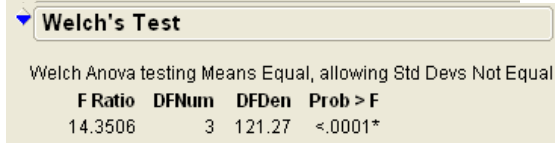
6 Practical Data Analysis with JMP

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.

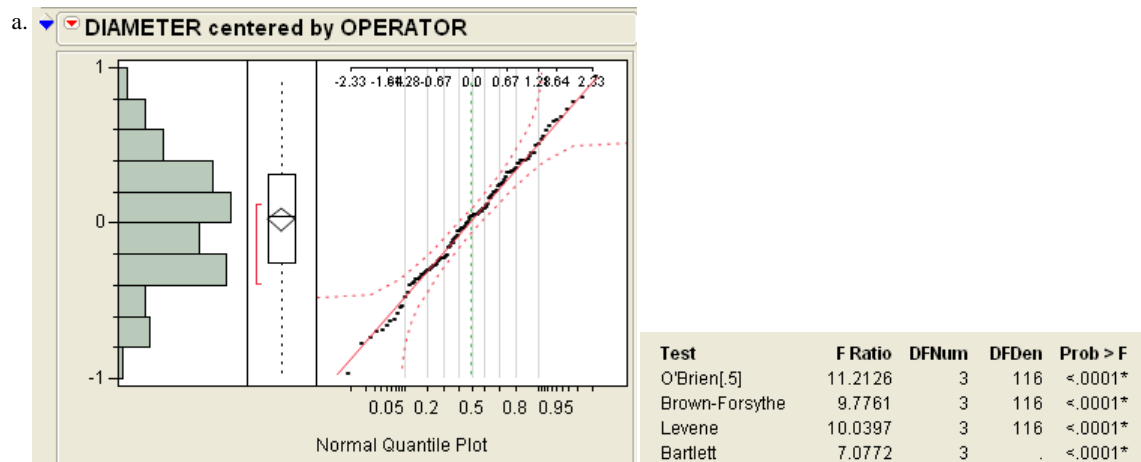


c. The distribution of residuals (not shown here) raises questions about normality and the usual tests indicate that the variances of the different disruption-type subgroups are unequal. According to Welch's test, there is at least one disruption type that differs from the others in terms of time required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

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*

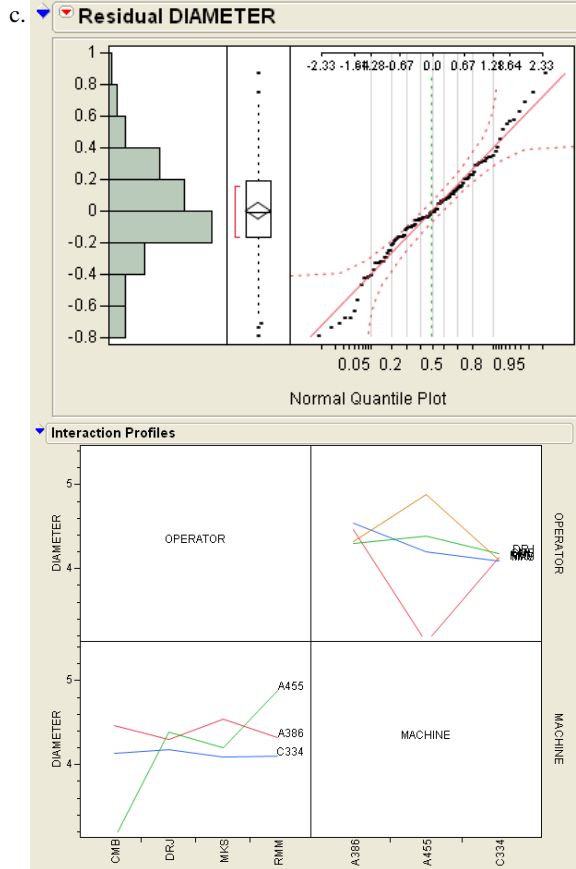


Scenario 6



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.



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.

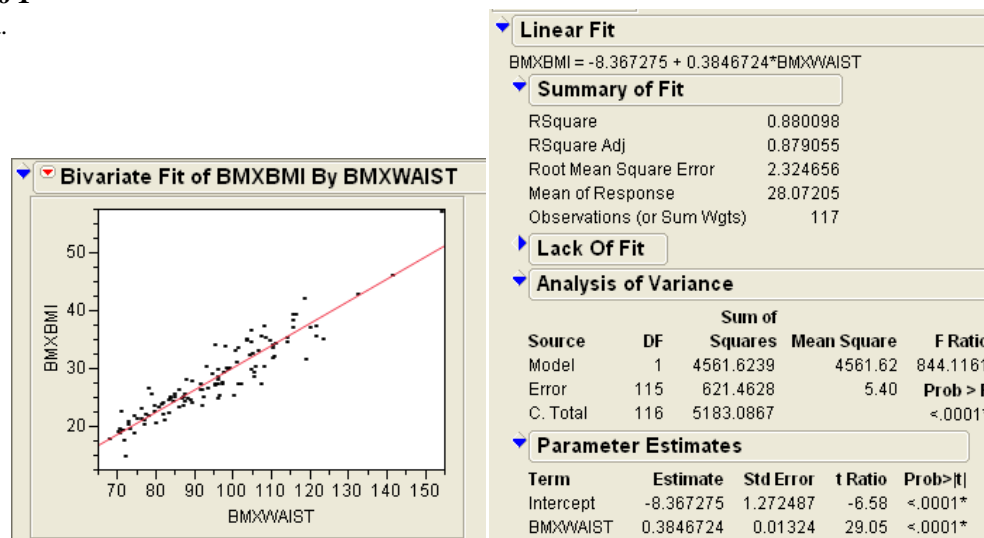
The interaction plots indicate interaction effects between operator and machine, making it difficult to interpret the main effects of machine and operator separately.

Student Solutions

Chapter 13

Scenario 1

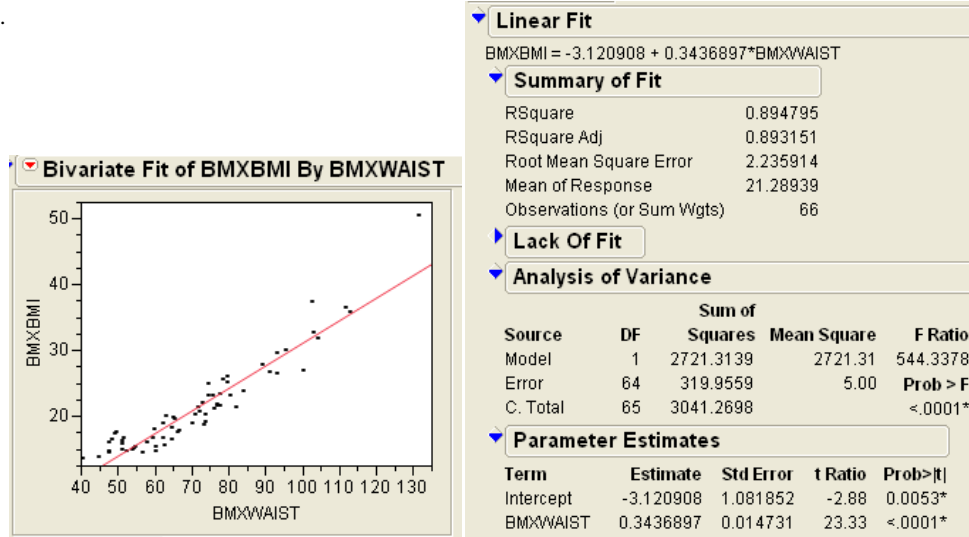
a.



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.

2 Practical Data Analysis with JMP

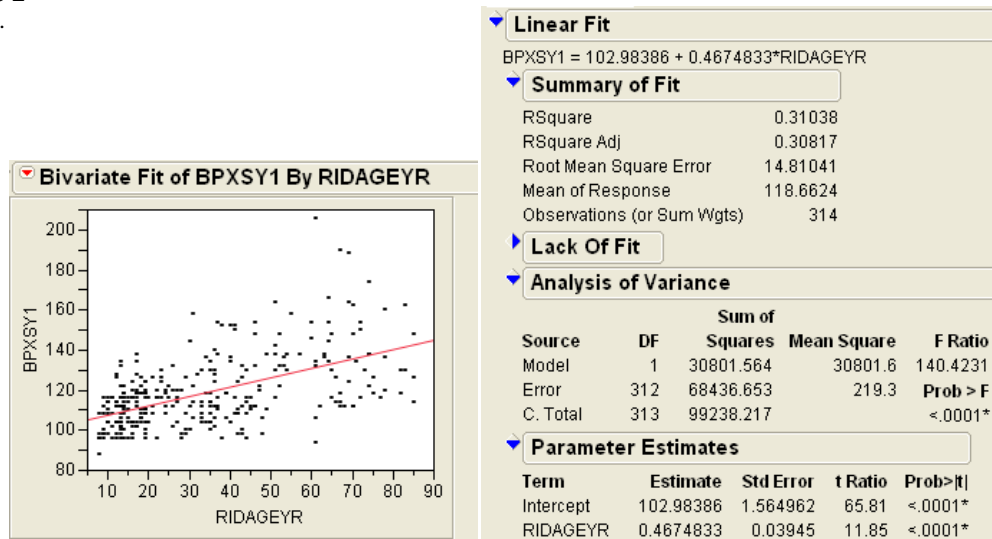
c.



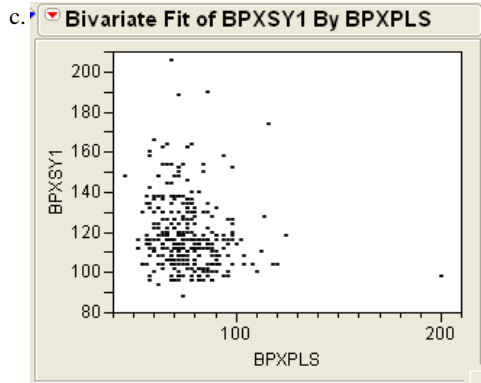
If we restrict the analysis to females under the age of 17 we find a slightly stronger relationship between Waist and BMI. The estimated slope is slightly smaller than before (0.344 vs. 0.385) but otherwise the regression models are very similar.

Scenario 2

a.

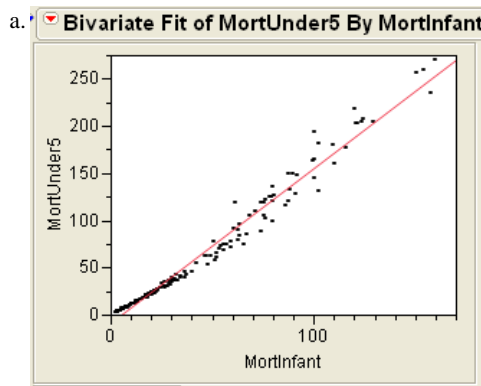


In this regression we find a weak ($R^2 = 0.31$) but highly significant positive relationship. Subjects who differ in age by 1 year tend to have, on average, systolic BP that is approximately 0.47 points higher per year. This is not a strong relationship because age accounts for less than one-third of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



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



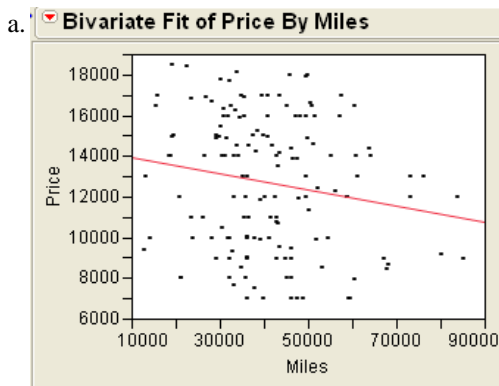
The estimated equation is:

$$\text{MortUnder5} = -7.661468 + 1.6385235 * \text{MortInfant} \quad \text{and } R^2 = 0.979$$

– 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



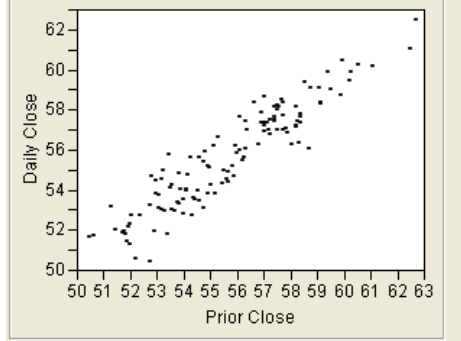
$$\text{Price} = 14341.4 - 0.0397524 * \text{Miles} \quad R^2 = 0.03.$$

4 Practical Data Analysis with JMP

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. **Bivariate Fit of Daily Close By Prior Close**



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

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*

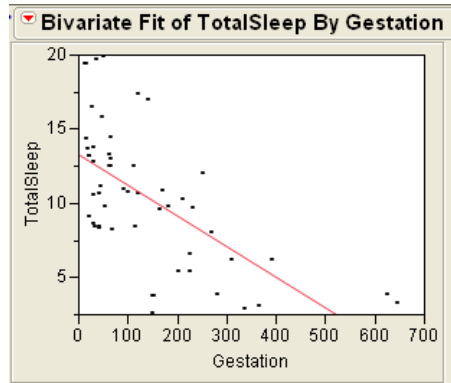
Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	0.1785088	2.225508	0.08	0.9363
Partb	0.6099486	0.027365	22.29	<.0001*

Custom Test	
Golden Mean	
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

Source	DF	Sum of 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*

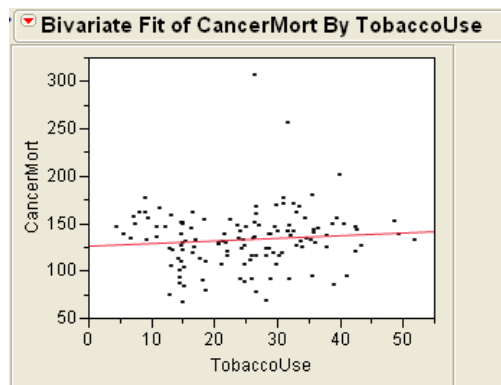
Parameter Estimates

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

Source	DF	Sum of 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

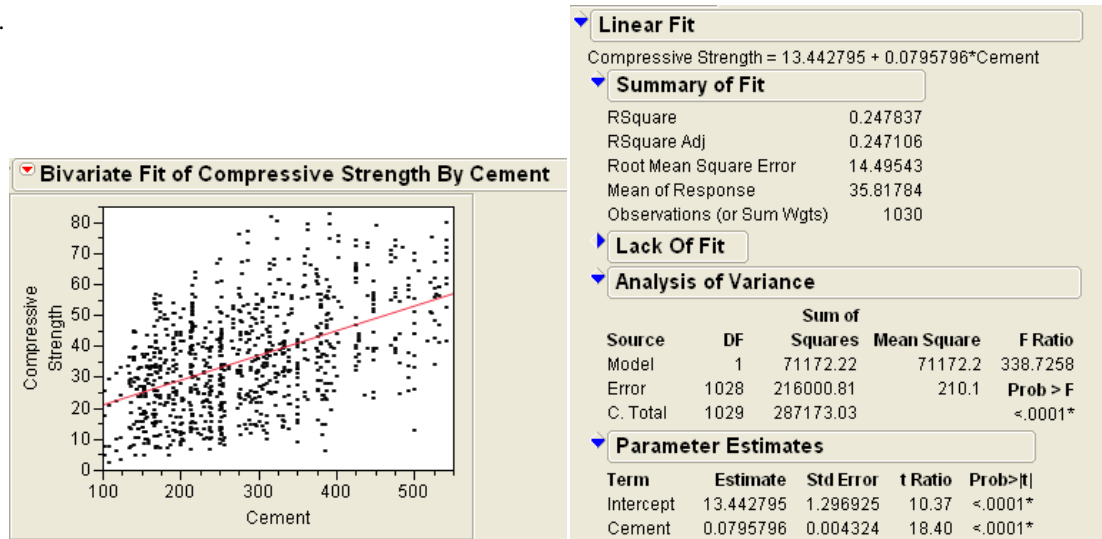
We find a non-significant relationship here – Tobacco Use is not a useful predictor of cancer deaths in a country.

6 Practical Data Analysis with JMP

- 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.



This is a highly significant, but weak, positive relationship. For each additional kg of cement in the mixture, compressive strength increases on average by 0.08 megapascals.

Scenario 10

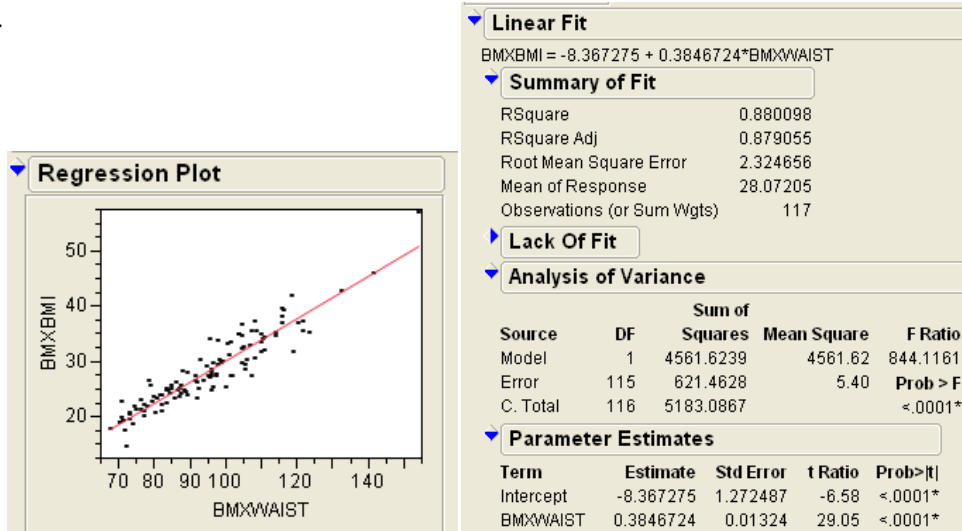
- a. There are slight differences, but when we round the major statistics we find that all four models are nearly identical: $Y_i = 3 + 0.5 X_i$. All R^2 (0.66) and p-values (0.0022 for the slope) are the same.
- c. In the other three graphs, the points do not fall in a linear pattern at all. This illustrates 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).

Student Solutions

Chapter 14

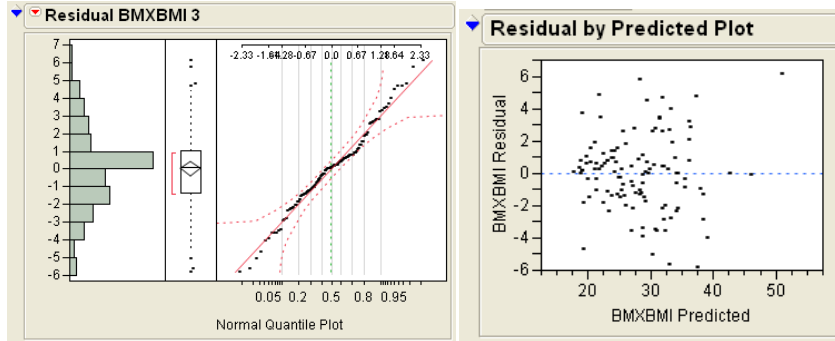
Scenario 1

a.



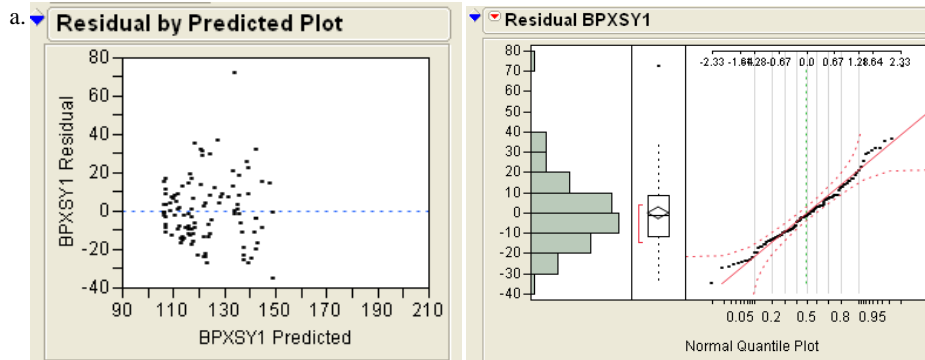
We first performed this regression in Chapter 13. Above are the regression results for adult females. We find a significant relationship between waist circumference and BMI, with the waist measurement accounting for about 88% of the variation in BMI. Each addition centimeter of waist circumference is 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.

2 Practical Data Analysis with JMP

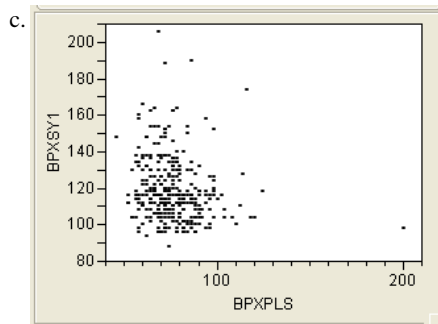


- c. Looking at the fitted line graph, it appears that the mean BMI for women with 68 cm. waists is approximately 18.

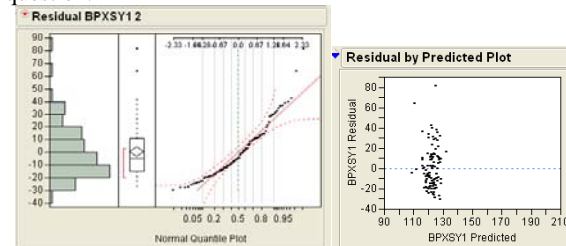
Scenario 2



Once again we see the suggestion of heteroskedasticity on the left side of the graph. The residuals are largely normal in shape, except for a single right-side outlier. We can probably use the model safely.

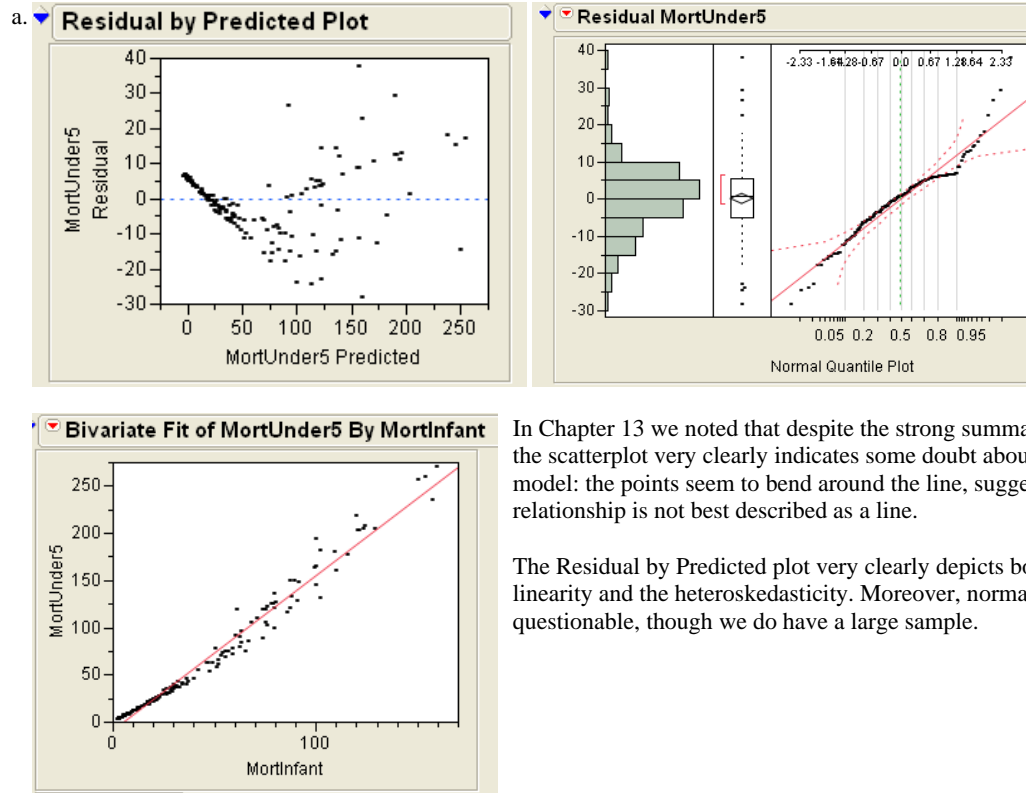


The scatterplot to the left shows little or no relationship between pulse and systolic BP. If anything, there may be a very weak negative relationship here, contrary to the suspicion expressed in the question.



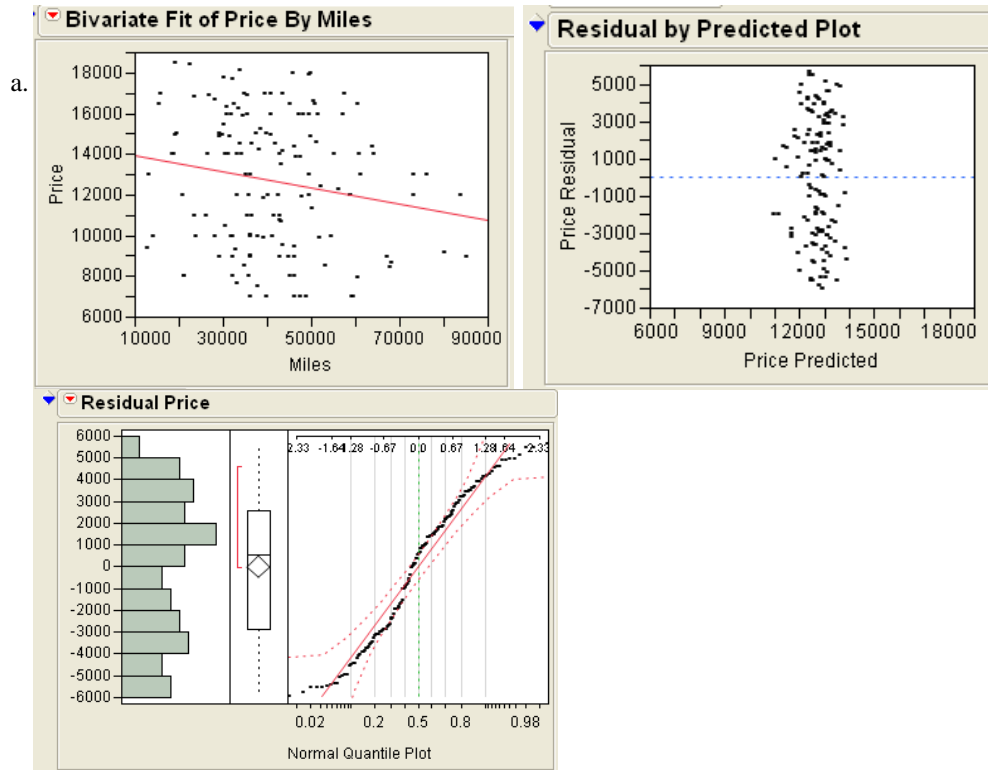
The residuals graphs cast doubt on normality (though the Central Limit Theorem applies); there does not seem to be a problem with constant variance.

Scenario 3



4 Practical Data Analysis with JMP

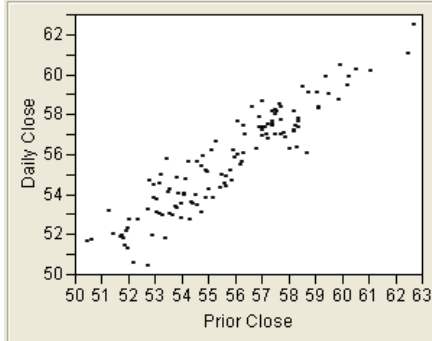
Scenario 4



The residuals are clearly not normally distributed, though the sample is reasonably large. There is no obvious problem with constant variance.

- 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 \$6,200 to \$19,500.

Scenario 5

a. **Bivariate Fit of Daily Close By Prior Close**

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

c.

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

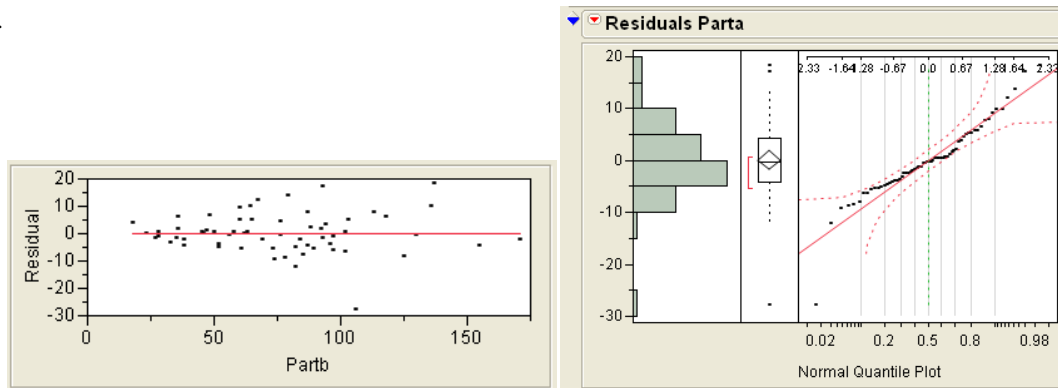
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	5.1911903	1.775282	2.92	0.0041*
Prior Close	0.9060521	0.031839	28.46	<.0001*

The Random Walk model specifies a slope of 1 and intercept of 0. In the table of parameter estimates, we see that we reject the null that the intercept = 0. Moreover, in a custom test comparing the estimated slope to a hypothetical parameter of 1.0, we reject the null; that is, we find that 0.906 is significantly different from 1. Therefore, the Random Walk model does not suit this set of data.

6 Practical Data Analysis with JMP

Scenario 6

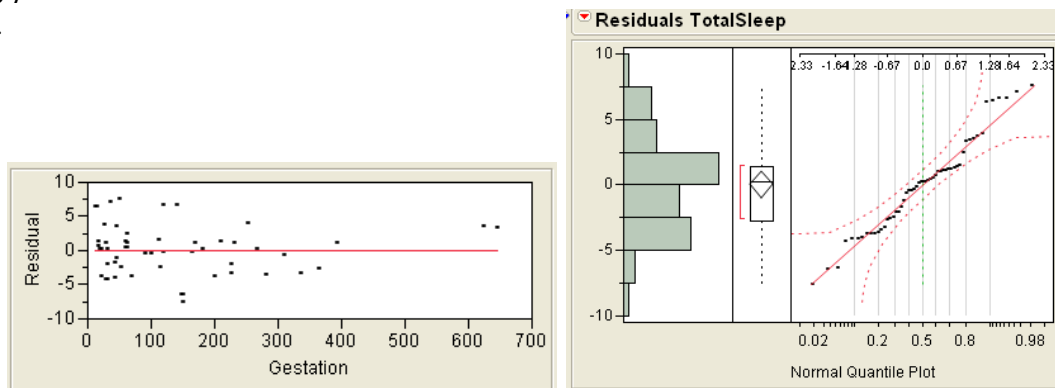
a.



With the Haydn data, in the Residual vs. Partb plot we find a heteroskedastic pattern; the residuals do deviate from normality, but the distribution is single peaked, moderately and we have a large sample.

Scenario 7

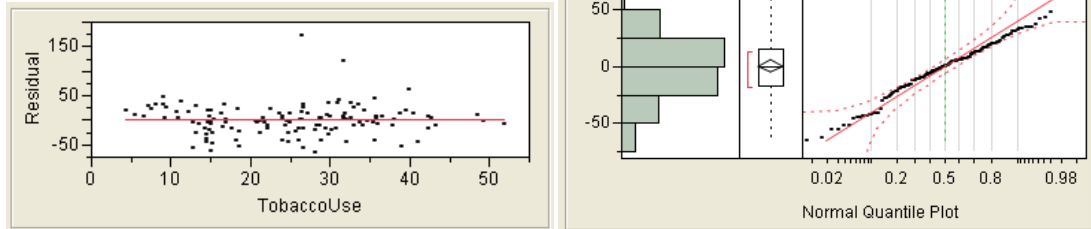
a.



Here 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.

Scenario 8

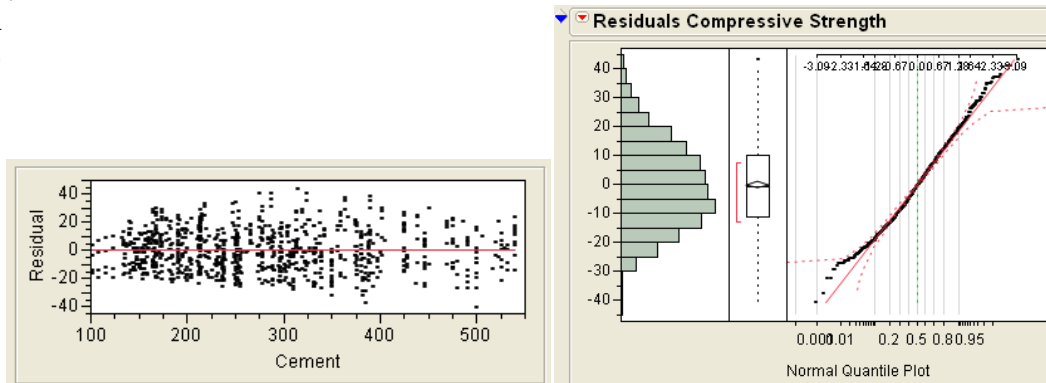
a.



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

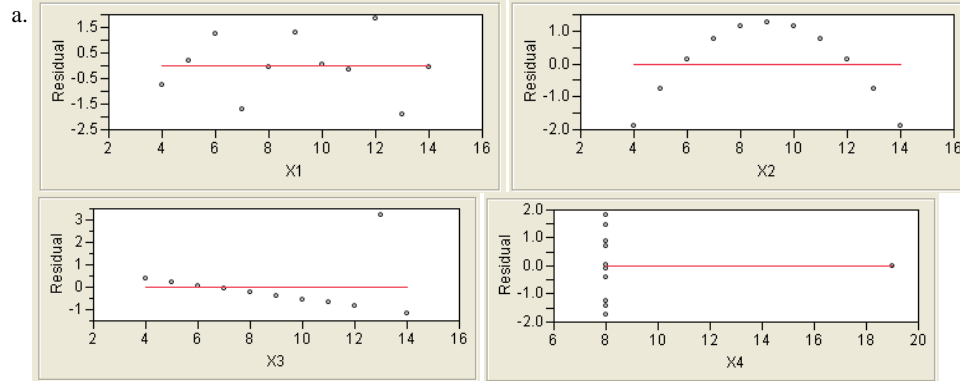


These residuals look good... the Residual vs. Cement plot shows an even scatter above and below the 0-line and the normal quantile plot shows that the residuals follow a nearly normal distribution except for the lower tail. In any case, we have a very large sample, so the CLT applies. We can safely interpret 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 0.08 megapascals.

8 Practical Data Analysis with JMP

Scenario 10



Above are the four plots of residuals vs. X; normality plots are not shown here. 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.

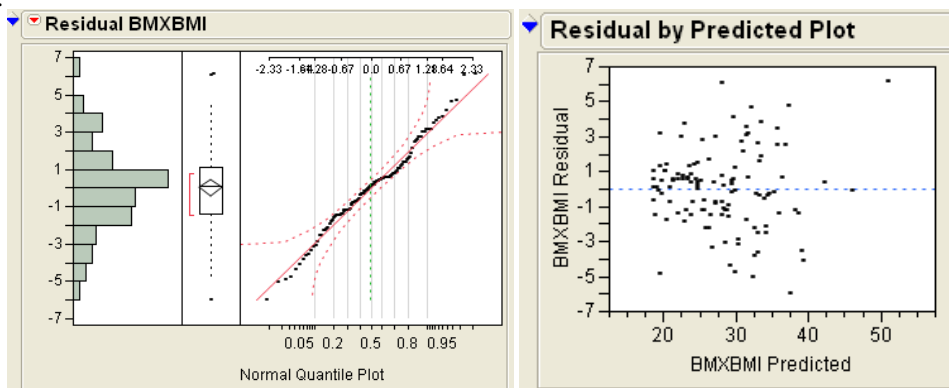
b. The four residual vs. X plots indicate that only the first model is suitable for interpretation and use.

Student Solutions

Chapter 15

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 here:

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*

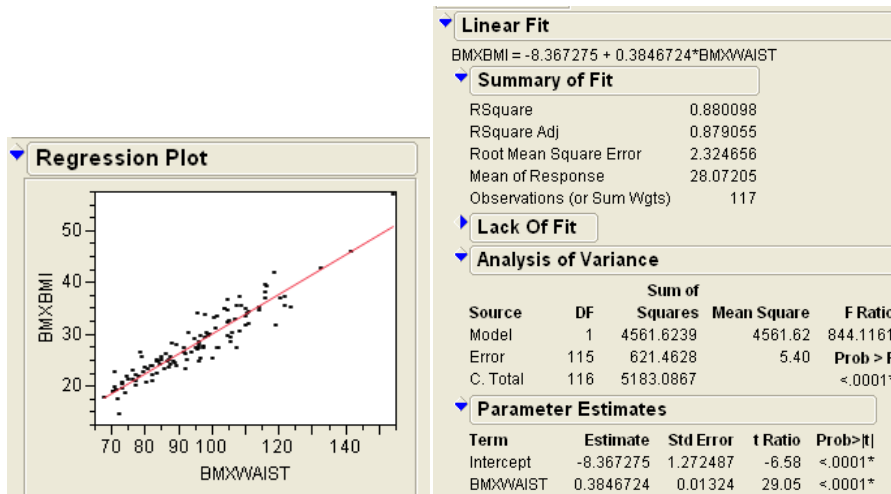
Parameter Estimates

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

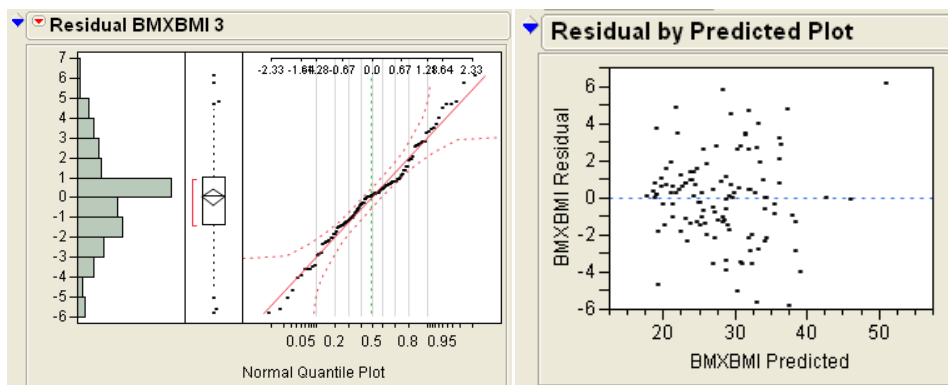
2 Practical Data Analysis with JMP

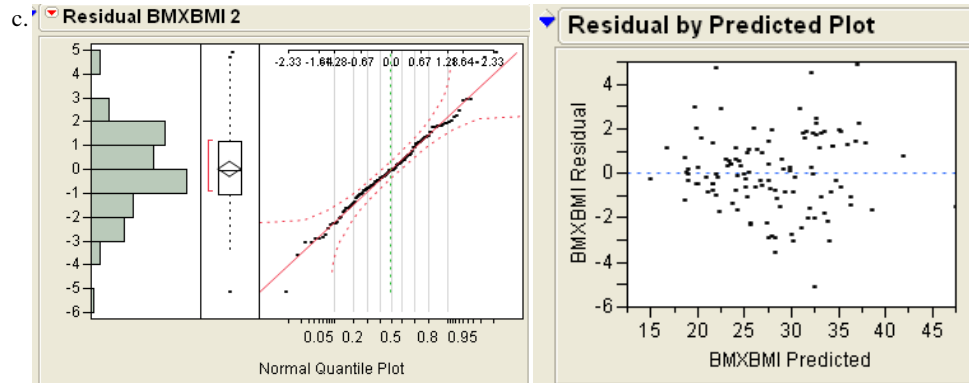
We find a strong relationship between BMI and the model, but this model is not much of an improvement over the previous model (shown again below). The intercept has changed dramatically, though in this model the intercept does not have much meaning. The effect size for the Waist measurement is almost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

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



We first performed this regression in Chapter 13. 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.





In the model using waist and thigh circumference (note typographical error in early printings of the book that this is referred to as wrist circumference), we find residuals that are approximately 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 significant 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*

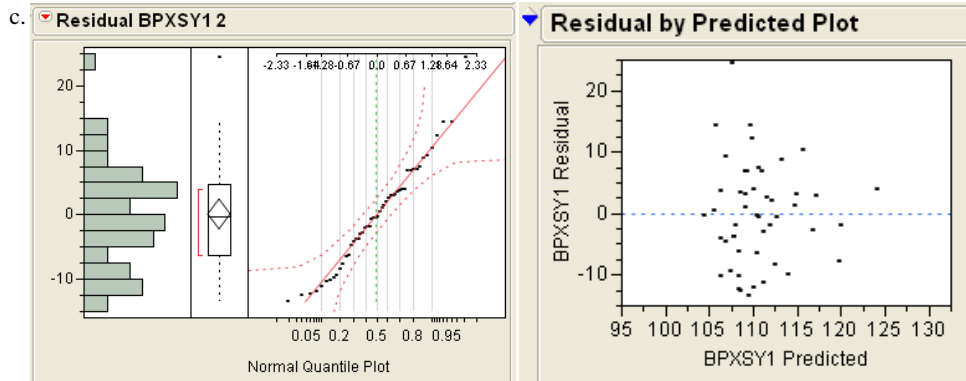
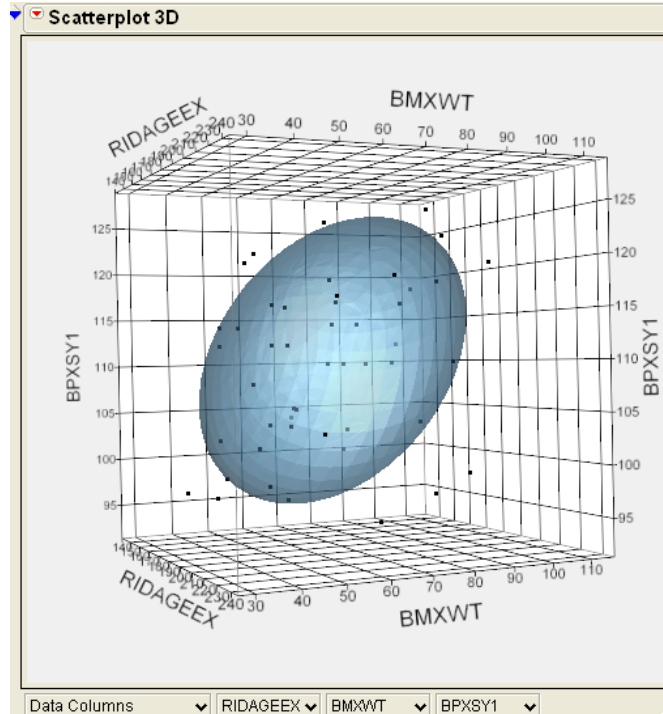
Parameter Estimates

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.

4 Practical Data Analysis with JMP



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

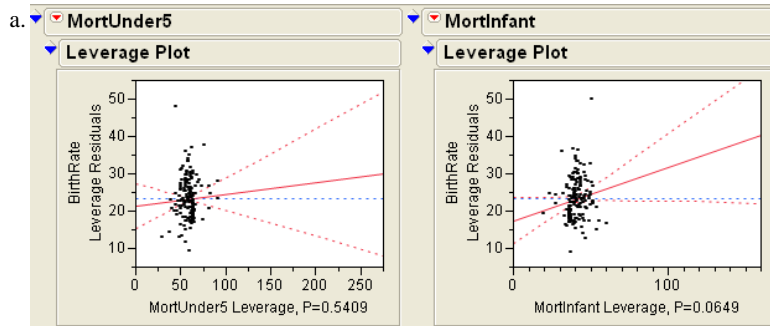
Analysis of Variance

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*
BPXD11	0.149581	0.142639	1.05	0.2998

Scenario 3



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

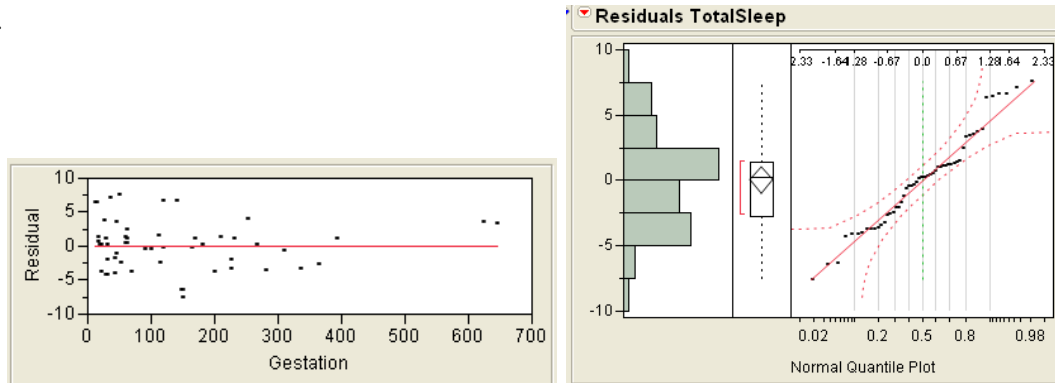
Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	13.286898	0.702092	18.92	<.0001*	.
MortMaternal	0.0074981	0.002851	2.63	0.0093*	7.4325523
MortUnder5	0.0316232	0.051616	0.61	0.5409	61.071947
MortInfant	0.1418103	0.076308	1.86	0.0649	48.623078

This model should not be used or interpreted.

6 Practical Data Analysis with JMP

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

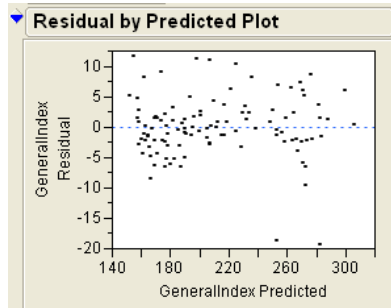
This model is not an improvement over the prior two. We still see heteroskedasticity in the plot of residuals vs. fitted values (not shown here). We see evidence of collinearity in the large VIF for BrainWt, and only the Gestation variable is statistically significant.

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:

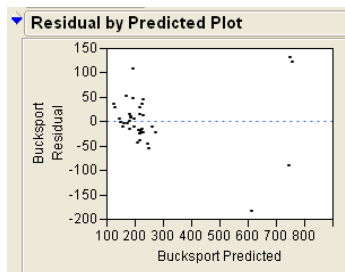
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-45.42024	2.919964	-15.56	<.0001*
BasicGoods	1.3979391	0.015697	89.06	<.0001*

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:



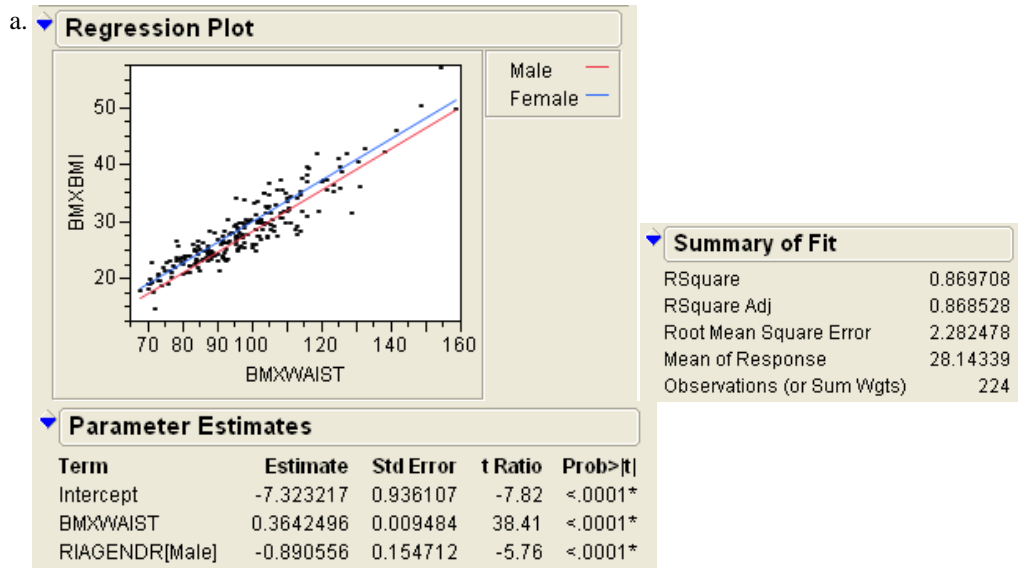
The residuals appear to have a non-constant variance, which raises a problem with using this model for prediction or estimation. The model adjusted R^2 is approximately 0.9 which indicates a very good fit. Both variables are statistically significant and we see no real evidence of collinearity.

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

Student Solutions

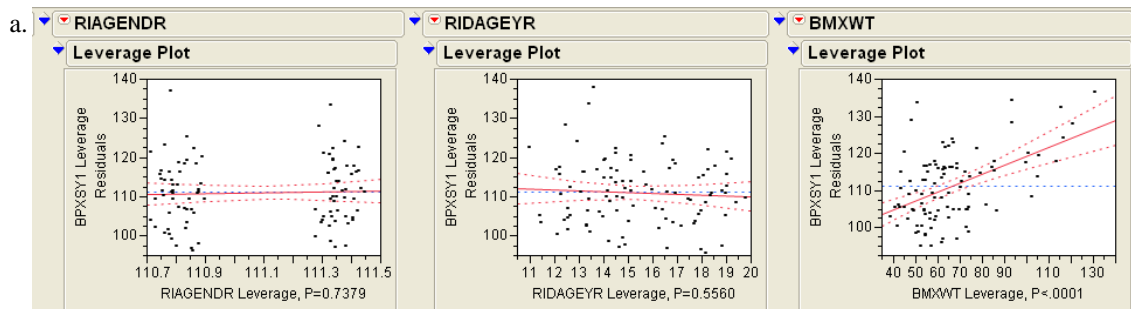
Chapter 16

Scenario 1



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



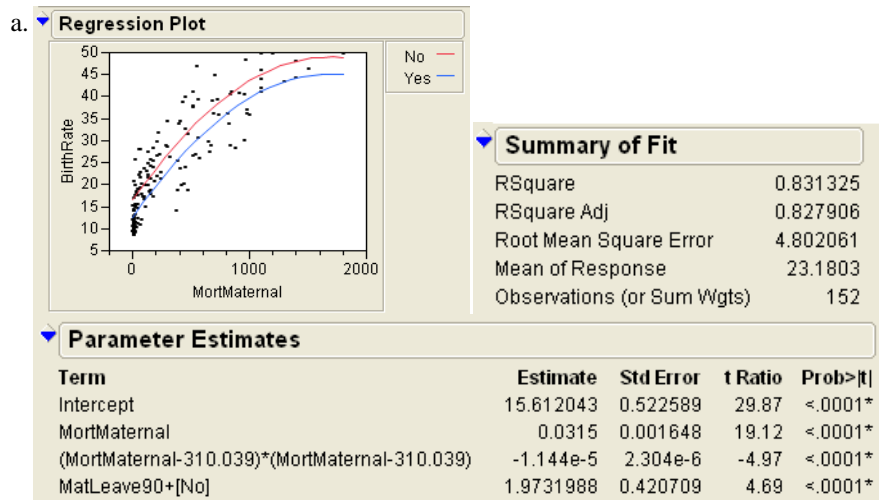
2 Practical Data Analysis with JMP

Summary of Fit	
RSquare	0.258143
RSquare Adj	0.23496
Root Mean Square Error	8.262977
Mean of Response	111.08
Observations (or Sum Wgts)	100

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	98.460896	5.397262	18.24	<.0001*
RIAGENDR[Male]	0.2812469	0.838053	0.34	0.7379
RIDAGEYR	-0.214684	0.36334	-0.59	0.5560
BMXWT	0.2416745	0.04334	5.58	<.0001*

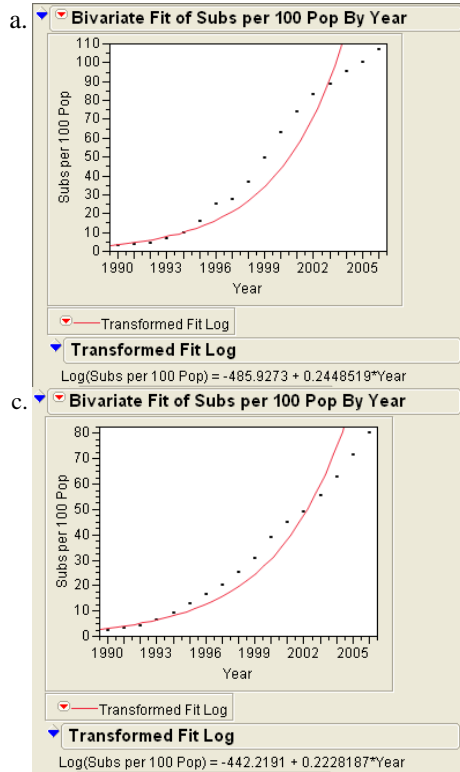
The leverage plots indicate collinearity problems, which are borne out by the parameter estimates. We see that the model has rather poor fit, and only the Weight variable is statistically significant.

Scenario 3



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



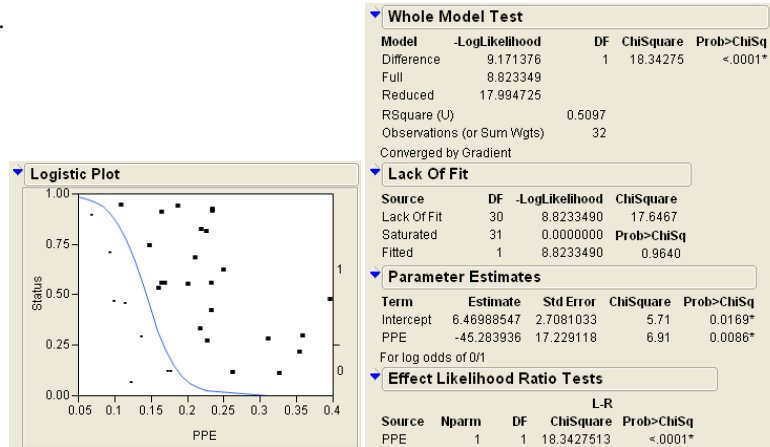
For Denmark, the annual growth rate is $e^{0.2448519} - 1 = 0.277$ or 27.7% per year.

For the U.S., the annual growth rate is $e^{0.2228187} - 1 = 0.249$ or 24.9% per year.

4 Practical Data Analysis with JMP

Scenario 5

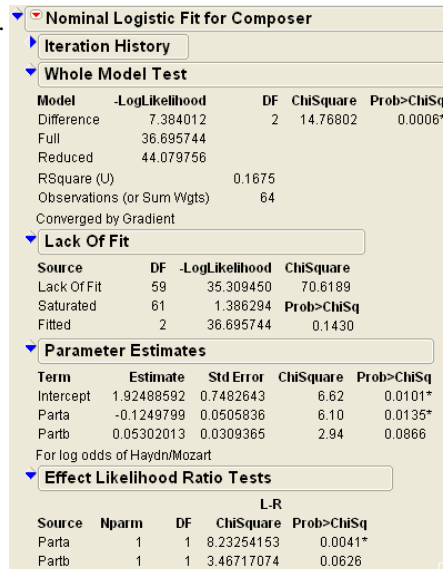
a.



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.

Scenario 6

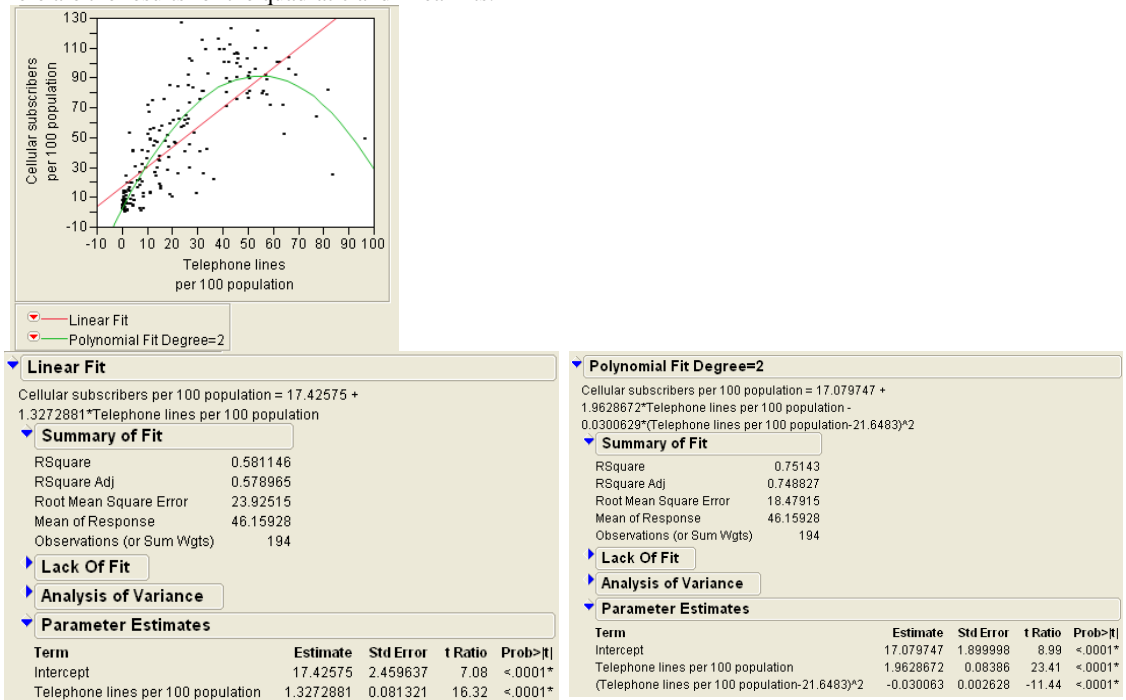
a.



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:



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

Student Solutions

Chapter 17

Scenario 1

a. **Model Comparison**

Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	AIC Rank	SBC Rank	MAPE	MAE
Winters Method (Additive)	104	50.177156	735.50239	743.52087	0.940	729.50239	2	2	3.315726	6.109066
Winters Method (Additive)	101	45.377639	711.39900	719.33217	0.942	705.399	1	1	3.132245	5.846902

As shown above, using a 6-month season 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**

Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	AIC Rank	SBC Rank	MAPE	MAE
AR(1)	109	99.891165	830.57245	835.99151	0.878	826.57245	3	3	4.290917	7.913439
ARI(1, 1)	108	46.320348	736.88434	742.28530	0.949	732.88434	2	2	2.877901	5.367753
ARI(2, 1)	107	42.046266	727.42716	735.52860	0.954	721.42716	1	1	2.825473	5.258672

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

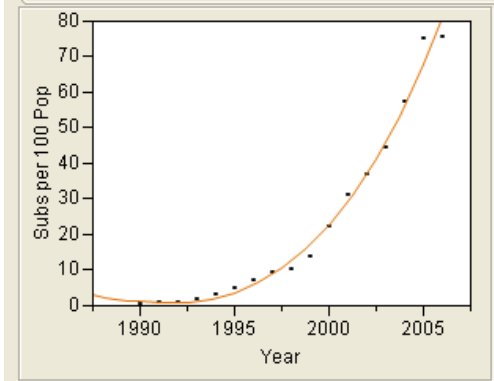
Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	AIC Rank	SBC Rank	MAPE	MAE
Linear (Holt) Exponential Smoothing	107	519.71465	1000.9686	1006.3513	0.875	996.96864	2	2	5.741653	17.872988
Winters Method (Additive)	104	528.09261	989.97456	997.99305	0.873	983.97456	1	1	5.876557	18.367321
AR(1)	109	689.3022	1044.5854	1050.0045	0.832	1040.5854	4	4	6.783594	21.001116
ARI(1, 1)	108	597.81201	1017.5667	1022.9677	0.870	1013.5667	3	3	5.840147	18.695112

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:

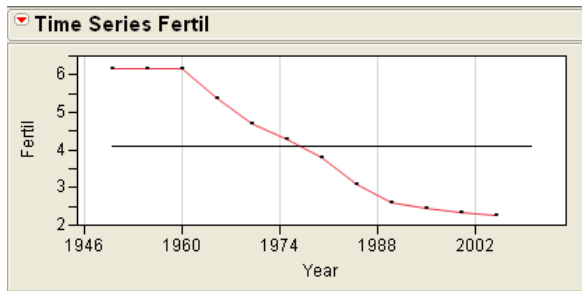
2 Practical Data Analysis with JMP



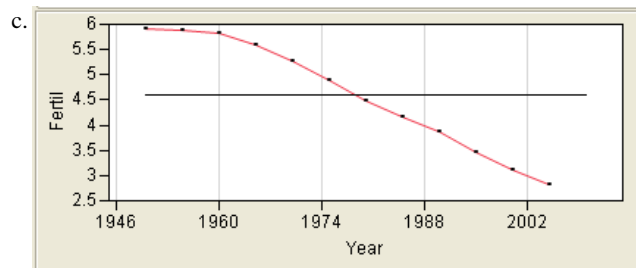
e. These countries are all best approximated by different models. Effective time-series modeling requires the use of a variety of approaches.

Scenario 4

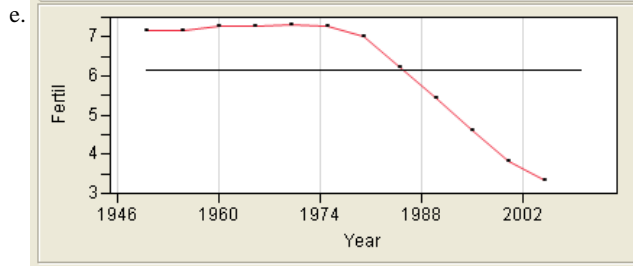
a. The fertility rate in Brazil has declined following an S-shaped curve:



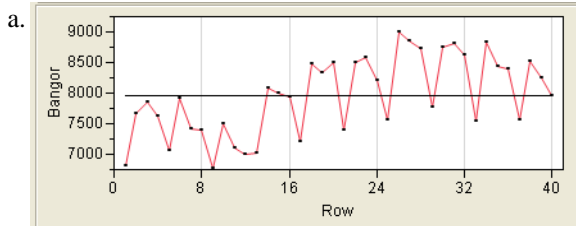
An AR(1,1) model fits moderately well, with relatively high RSquare (0.969), low variance (0.077) and MAPE and MAE of 5.35% and 0.20 respectively.



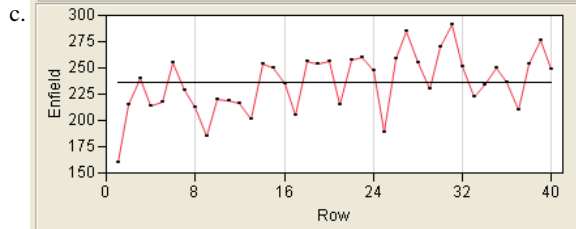
India's decline is very regular, especially since 1960. Linear Exponential Smoothing (Holt's method) and AR(1,1) models both fit extremely well.



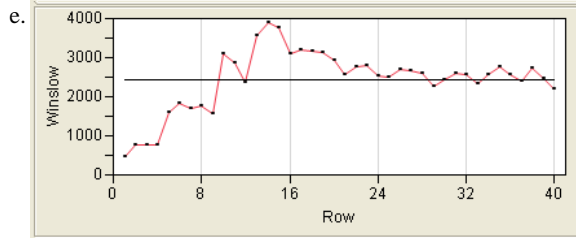
Scenario 5



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.

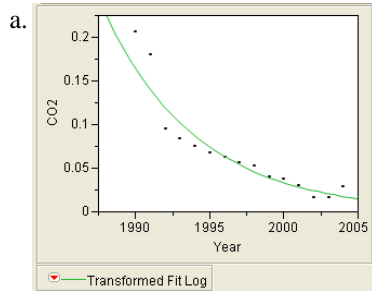


Enfield: This pattern is much like the one in Bangor; Once again an AR(4,1) model fits well.



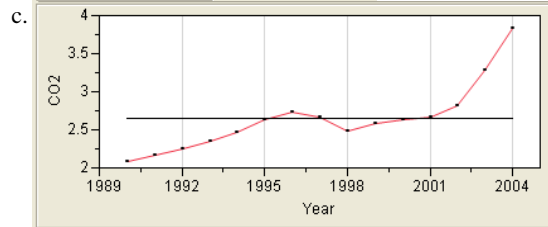
Winslow: Here we see the dramatic change occurring roughly half-way through the time series. Simple exponential smoothing provides a reasonably good model.

Scenario 6

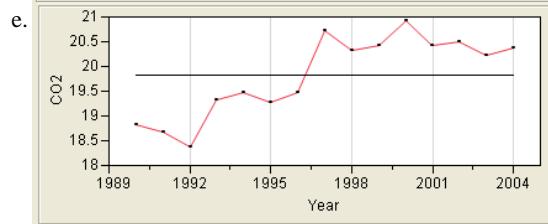


CO2 emissions in Afghanistan have fallen since the series began, and have leveled off (with minor increases) in most recent years.

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

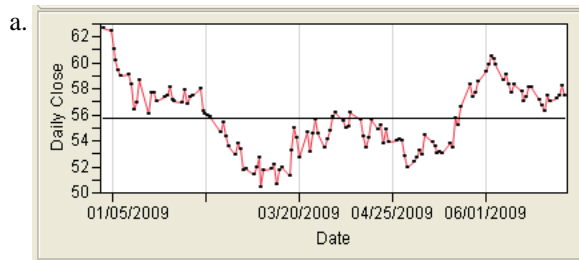


In sharp contrast to the prior two graphs, China's CO2 emissions have been rapidly rising. A 3rd-degree polynomial (cubic) provides a moderately good fit, as does AR(1,1).



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



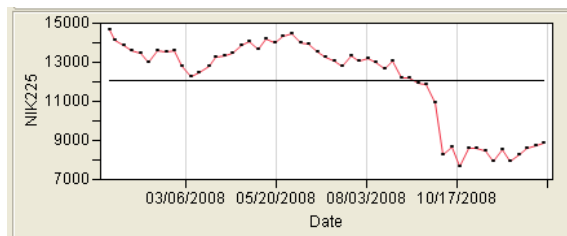
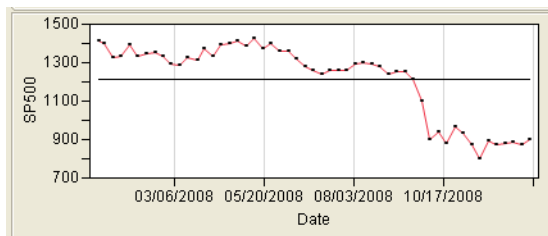
The series to the left would be poorly described with any type of linear trend model because it exhibits several changes of direction. Because we have just 6 months of data, we should not use Winter's method which accounts for seasonal variation.

Scenario 8

a. **Correlations**

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

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



For the S&P no model is perfect, AR(2,1) provides a comparably low variance, MAE, MAPE, and high RSqr.

Much like the S&P series, the Nikkei is well-modeled with an AR(2,1) model.

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

Student Solutions

Chapter 18

Scenario 1

a.

Pattern
++-2
--+1
+--1
--+2
--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.

2 Practical Data Analysis with JMP

Scenario 3

a.

	Pattern	Gender	TestCondition	Sleep
1	122	Female	SleepFirst	Interrupted
2	212	Male	AwakeFirst	Interrupted
3	222	Male	SleepFirst	Interrupted
4	221	Male	SleepFirst	FullNight
5	121	Female	SleepFirst	FullNight

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, amount of money offered.

- 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 $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 2^7 = 128$ runs.

Scenario 5

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

	Pattern	Impact Modifier	Thermal Stabilizer	Anti-UV	Charpy Index
1	223	CPE	Pb	10	▪
2	342	ABS	Sn	5	▪
3	212	CPE	PdBaCd	5	▪
4	413	MBS	PdBaCd	10	▪
5	113	ACR	PdBaCd	10	▪

- c. The full-factorial design has 480 runs and the response-surface custom design has 640. 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 but the intermediate 6.5 level is only tested in 5% of the experimental runs.

Scenario 6

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

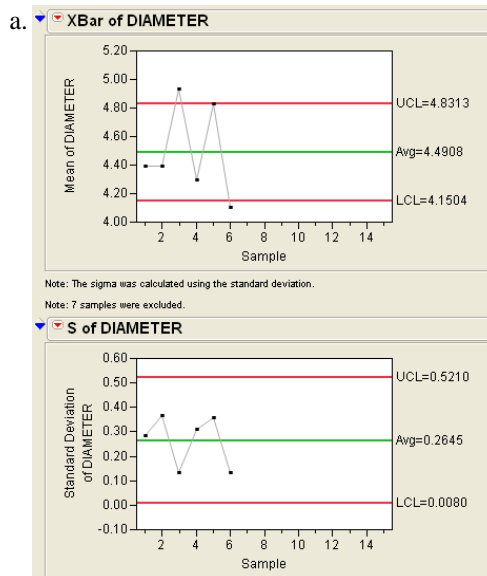
	Pattern	SubjectLine	Salutation	CaltoAction	Promotion	Closing	Response
1	11131	Crayola	Hi	As Crayola	Amazon	Crayola	▪
2	22132	Help	Greetings	As Crayola	Amazon	Education	▪
3	22122	Help	Greetings	As Crayola	Product	Education	▪
4	21221	Help	Hi	Because	Product	Crayola	▪
5	21222	Help	Hi	Because	Product	Education	▪

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

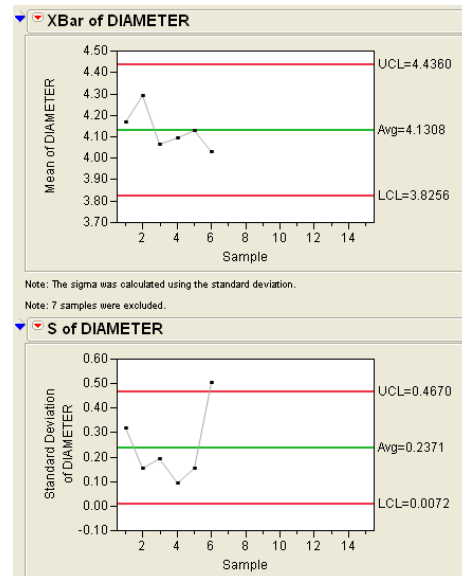
Student Solutions

Chapter 19

Scenario 1

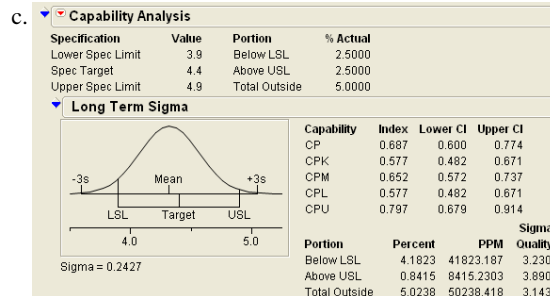


Machine A455



Machine C334

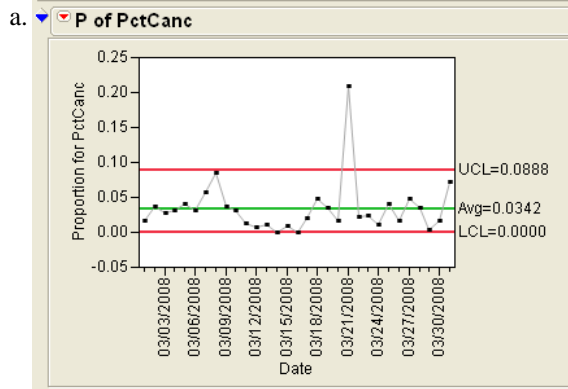
As we can see in the graphs above, Machine C335 may have an unstable standard deviation and machine A455 shows two sample means beyond the control limits. These machines should be inspected closely for possible adjustment.



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 4.5.

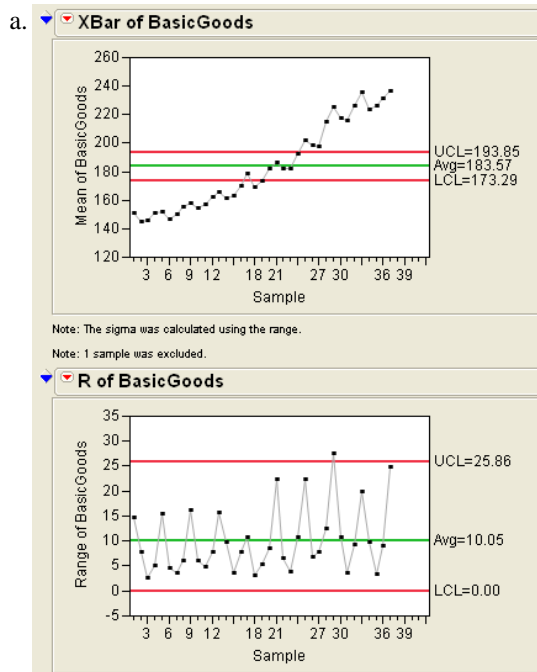
2 Practical Data Analysis with JMP

Scenario 2

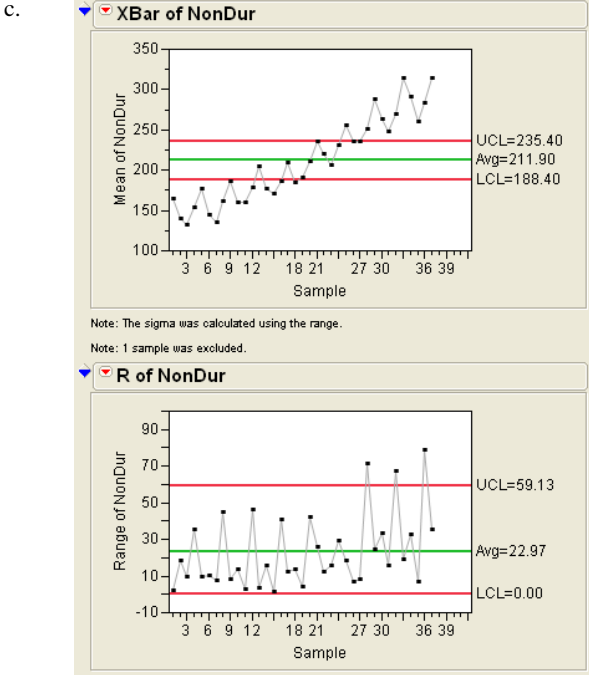


This process is out of control at one point. Because a day with 0 cancellations is desirable, we should not be concerned about dates with values below the LCL. However, the chart shows 1 date well above the UCL.

Scenario 3



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.



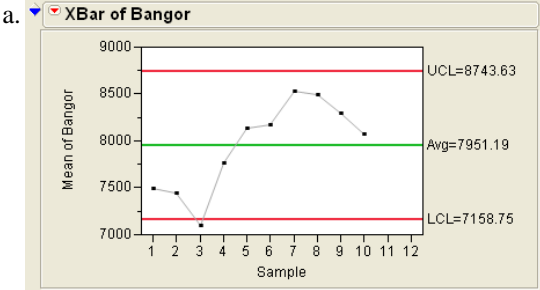
Once again we see a steady pattern of growth, with clear seasonal variation. In contrast to the control chart for Basic Goods, the one for NonDurables may exhibit a more linear upward trend, and substantial growth in variability (the R Chart) in the most recent years.

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

Scenario 4

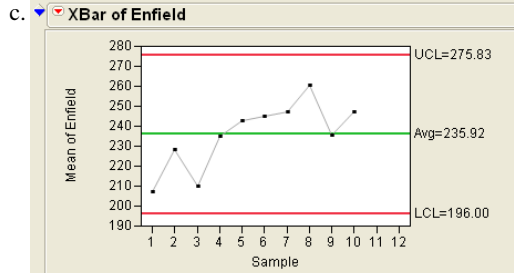
- a. In most regions except for the Southwest the standard deviations are sufficiently unstable that we should not interpret the Xbar charts. In the Southwest, the standard deviations have been steadily increasing but the limited data (only five sample mean) indicates increasing mean times to restore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

Scenario 5

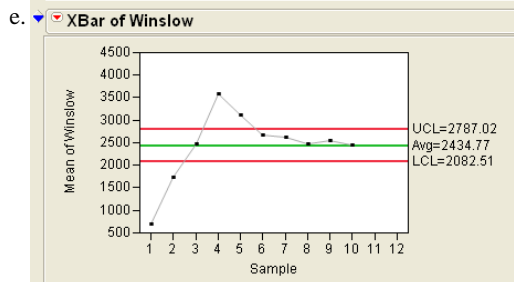


Bangor: The S chart is stable; early in the study period there was one year below the LCL. Otherwise Bangor has remained within limits, though the 6 most recent years have been above average.

4 Practical Data Analysis with JMP

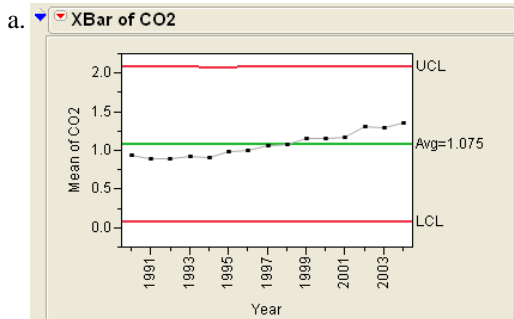


Enfield: This pattern is much like the one in Bangor. The S chart is stable throughout. Otherwise Enfield has remained within limits, though the 5 of the 6 most recent years have been above average

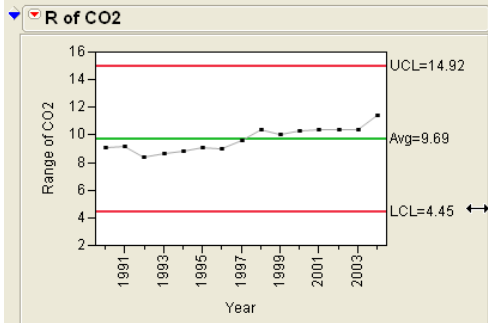


Winslow: In year 3 the S chart (not shown) shows the sample standard deviation above the UCL; otherwise the standard deviations are moderately stable. The Xbar chart shows a process out of control until year 6, after which the process seems to be in control.

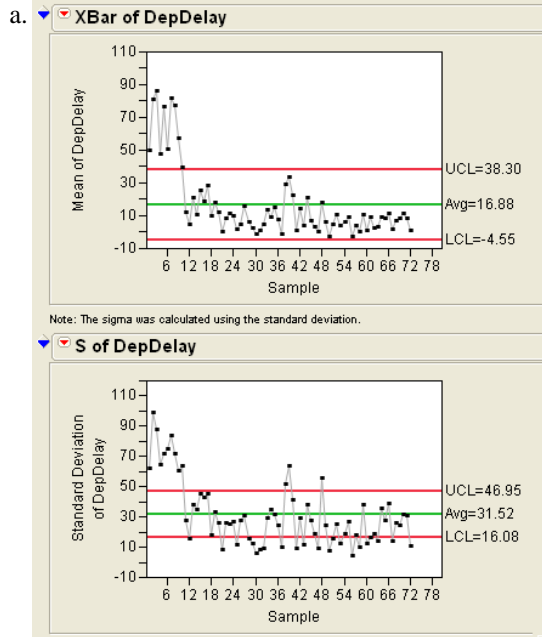
Scenario 6



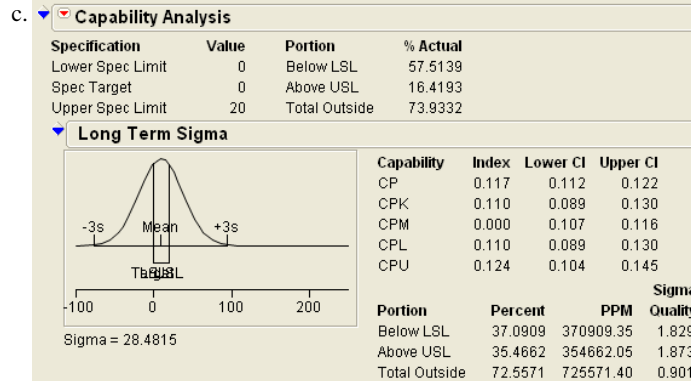
Emissions in most regions are relatively stable In Africa (shown to the left), both the ranges and means have been steadily rising over the 15-year period.



Scenario 7



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.

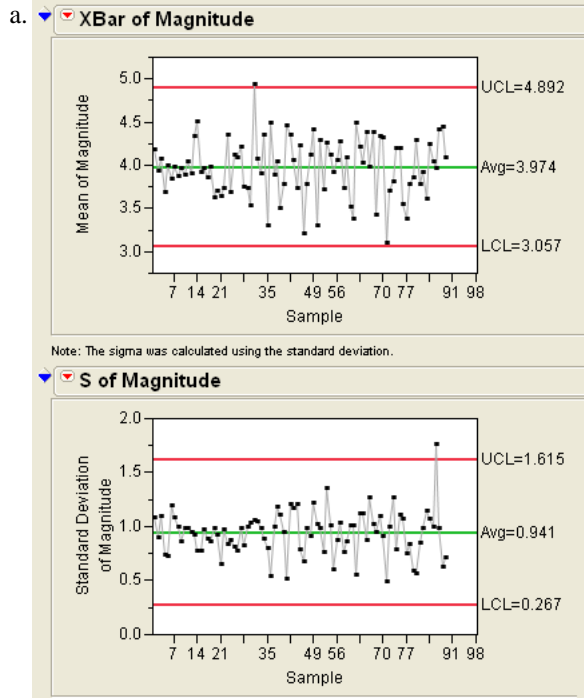


We need to select the weekday flights; because the target involves individual flights, we make a Run Chart. The critical capability limit here is the USL, which we set at 20 minutes; the other values may be set to zero

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.

6 Practical Data Analysis with JMP

Scenario 8



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.