Business Statistics Made Easy in SAS®

Gregory Lee

Practice Exercises with Solutions

support.sas.com/bookstore

Business Statistics Made Easy in SAS

By Prof Gregory John Lee

Practice Exercises (v.1, incl. answers)

Table of Contents

Preface: How to Use these Practice Questions

Welcome to the general-access practice questions for *Business Statistics Made Easy in SAS* by Prof Gregory John Lee (1st edition, 2015). These questions are designed for general access and use by all readers. This preface explains various aspects of this document.

TYPES OF QUESTIONS

There are various types of exam, test and/or general practise questions in this document for different potential uses:

- There are general theory / concept questions in various formats from multiple choice to essays. Some of these are pure theory, some hypothetical implementation.
- There are practical implementation questions where you are given data and asked to analyse it in SAS and answer questions.
- There are some questions where you are given pre-generated SAS outputs or other statistical findings / tables / graphs, and asked to answer questions on these.

The applicability of any given question to your needs is therefore debatable. You may be doing a basic course where only multiple choice is used as an examination technique: in that case using the other forms of questions will help your understanding but obviously will not apply to your course's examination setting. Another example is questions based on SAS code: if your course or needs does not require learning of the code, those questions will not apply to you.

DATASETS & PRINTOUTS FOR QUESTIONS

Any datasets or printouts attached to a question will be contained in the folder this document came in, available at https://support.sas.com/publishing/authors/lee.html.

QUESTION SETS WITH AND WITHOUT ANSWERS

There are two versions of this document in the folder, one without the answers embedded and another containing most of the answers. Please use your own discretion in which you would prefer to use, obviously the most success in practising is obtained by not having the answers immediately at hand.

LIMITED-ACCESS INSTRUCTOR QUESTIONS

There are several sets of instructor questions available for test/exam setting and the like. Please apply using the form at https://support.sas.com/publishing/authors/lee.html to get access to these. Obviously only genuine, confirmed instructors may access these sets, and we ask instructors not to disseminate the instructor sets.

COMPLEMENTARY PRACTICE OPPORTUNITIES

I encourage readers who wish to practise to make use of many other opportunities and sources other than the book. There are hundreds of statistics textbooks and online resources that can give alternate explanations of the book's general statistical principles as well as questions. When it comes to practising SAS implementation specifically or the interpretation of outputs, the first stop may be the various SAS User Guides (such as the SAS/STAT user guides). All these are available online, and provide many data examples. A similar resource is the many SAS papers that can be found online, many of which contain easy-to-follow examples.

1. Chapter 2 Introduction to the Statistical Process

1) (Open-book question): Imagine you are a marketing researcher seeking to measure the average response of customers to a new advertisment. Describe the overall statistical process you would undertake – using the overall statistics process described in Chapter Two of the book – for this situation, including the challenges you will face. Specifically relate each part of the answer to the specific marketing research example. (10 marks).

Answer: Answer is simple run-through of the Ch 2 statistical process but with the specific advertising research focus. In Step 1 the problem is explained specifically in terms of the advertising research (2 marks). In Step 2 data is gathered: student should specifically address the fact that data are gathered from customers, prfereably from a big enough and representative sample, and that some measure of response to the advertisement is needed, the more specificity the better (3 marks). In Step 3 the generation of statistics is covered. The average has specifically been mentioned as the focus. Depending on how much farther in the book the class has reached, the student may consider alternate measures of central tendency (such as the median), spread, and statistical significance. Issues of accuracy and size as formost (3 marks). Finally potential implications for the business should be mentioned depending on the outcome (2 marks).

2) (Closed-book question): Describe the overall statistical process as described to you in Chapter 2 of the book. (10 marks).

Answer is simple repeat of Ch 2 model of statistical process.

2. Chapters 3 & 4 Introduction to Data Exercises

2.1. TYPE OF VARIABLE

3) What type of variable is "*amount of leave time taken by employees"*, measured in hours?
   a) Ratio *(variable has a natural 0 and a range over a smooth continuum)*
   b) Interval
   c) Ordinal
   d) Categorical.

4) Which of the following types of data would "Industry" be (i.e. ascetaining what industry a company is in)? (1 mark)
   a) Ratio
   b) Interval
   c) Ordinal
   d) Categorical *(variable values reflect membership in a group only)*

2.2. MISSING DATA

5) Say you have three variables. One is Gender, the second is Age measured in days since birth, and the third is composed of seven sub-questions each of which measures an aspect of Employee Stress (i.e. a multi-item scale) each on a seven-point Likert scale. In each of these variables there is missing data. What are **all** your options for dealing with the missing data? **(7 marks)**

For Gender:

- *Imputation with mode but poor option*
- *Multiple imputation*
- *Leave as is if not too many*

Age:

- *Imputation e.g. with average*
- *Multiple imputation*
- *Bootstrapping methods*
- *Leave as is if not too many*

Stress

- *If not too many in row average into composite stress score and ignore missingness*
- *Imputation with median on individual items*
- *Multiple imputation*
- *Bootstrapping methods*

2.3. MULTI-ITEM SCALES

6) Say that you wish to measure employee commitment using a multi-item scale with 10 sub-questions in a survey, each of which is assessed on a 7-point Likert scale. What steps do you need to go through to use this multi-item scale? **(10 marks)**

Reverse the scoring on reverse-worded items

Deal with missing data – as per missing data question above

Internal reliability – Cronbach alpha

Aggregate – either average, factor analysis, or sum (if no missing items)

3. Chapter 5: Introduction to SAS

I do not specifically suggest testing this chapter outside of practical implementation in other chapters. Therefore there are no questions for Chapter 5.

4. Chapter 6: SAS Programs, Data Manipulation, Analysis & Reporting

4.1. SAS PROGRAMMING THEORY

7) The following SAS code contains three mistakes that would stop it from working. Identify each of the mistakes. (7).

```
DATA=Production;
  SET Factory.Production
  Total = Sum(of Daily1-Daily365);
  IF Month = January THEN Current = Yes;
RUN
```

Answer. Should be "DATA Production", i.e. no equals sign (2). There should be a semicolon at the end of the SET and RUN lines / commands (3). In the "IF-THEN" line, Jan and Yes should be in inverted commas, i.e. **IF Month = "January" THEN Current = "Yes";** (2)

4.2. COMBINING DATASETS

8) In SAS, say you have an employee information database called "HR.Employees" and a different sheet of employee performance data called HR.Performance. Each database identifies employees using a unique employee number (a variable called "employee_num"). (Further details: each employee appears on only one line in each dataset, and both datasets are sorted by employee number). Answer the following questions.
   a) What does the "HR" refer to in the dataset name? (2)
   b) Say you want to create a new dataset in the SAS 'Work' location called "Combined", which would combine the two datasets so that every employee's information and performance would be aligned on one line. Provide the SAS code you would use to achieve this merging. (8)

Answer. The "HR" refers to a permanent library location in SAS in which the "Employees" and "Performance" datasets are located (2). The code to achieve the merging of the two into the new dataset is (8):

```
DATA Combined;
  MERGE hr.employees hr.performance;
  BY employee_num;
RUN;
```

4.3. SAS OUTPUT / REPORTING

9) Explain how you can create a PDF file from SAS 9 output (5).

Answer. Student should discuss ODS functionality in general and for a complete answer possible give and example of the applicable code.

5. Chapters 7 & 8: Descriptive & Associational Statistics

5.1. THEORY

5.1.1. Measures of Centrality & Spread

10) Which of the following is NOT true of the *mean/average*:

a) The average is a measure of central tendency
b) The average is appropriate for continuous variables
c) The average is appropriate for ordinal variables *(averages are more appropriate for continuous data)*
d) The average is the sum of the variable values divided by the number of values.

11) Which of the following is NOT true of the *median*:

a) The median is a measure of central tendency

b) The median is the number that half of the variable values are less than and half the variable values are greater than

c) The median should only be used for ordinal data *(is is possible and desirable to look at medians of continuous variables too)*

d) An appropriate measure of spread when using the median is the inter-quartile range.

12) Which of the following is suitable as a measure of spread for ordinal variables:

a) The standard deviation

b) The inter-quartile range *(values 25% and 75% along the variable range)*

c) The mode

d) The spread of lowest to highest

e) The variance.

13) If you have a mean of 5 and a standard deviation of 2 then which of the following is true assuming the variable has the appropriate distribution:

a) The variance of the variable is 8

b) The variable runs from a low of 3 to a high of 7

c) Approximately two-thirds of the population is expected to lie between 3 and 7 *(since about 2/3 of the population is expected to lie between the average – the standard deviation and the average + the standard deviation)*

d) Approximately two-thirds of the population is expected to lie between 5 and 7

14) If you have a variable with the values 1,4,6,8,10,12,14,16,18,20 then the median is:

a) 10

b) 10.5

c) 11 *(the variable has an even number of values, so we take the average between the middle two. The middle values are 10 and 12, the average between them is 11)*

d) 11.5

e) 12

f) 12.5

g) 6.5.

15) If you have a variable with the data 2,4,6,8,10,12,14,16,18,20,30 what is the inter-quartile range? (1 mark)

a) 8
b) 6
c) 2 to 18
d) 6 to 18 *(6 is 25% the way up the values and 18 is 75% of the way up. Another way to see this: 12 is the median, and 6, 12 and 18 each have and equal number of values on either side of them, i.e. 2 values)*
e) 8 to 16
f) 10 to 16

16) If you have a variable with a mean of 22.34 and a median of 2.59, what would you conclude?

Answer: the mean and median are very different. There are probably some extremely large outliers pulling the average up, whereas these outliers will not affect the median so long as there are not too many of them.

5.1.2. Correlations & Causation

17) If I told you there was a correlation of .34 between employee satisfaction and productivity, what would you understand from this? **(3 marks)**

Answer: students should reference that this reflects a moderate positive association without assumption of causality

5.2. APPLIED SIMPLE DATA ANALYSIS IN SAS

18) Consulting company descriptive and associative statistics

Based on Dataset: "Dataset_2_Consulting"

In this dataset you are a consulting company interested in whether customer ratings of consultant service are affected by the knowledge and/or extroversion of the individual consultant. We have data on the following variables: a) "Tenure" (no. months in the company), b) "Female" referring to gender, where 1 = female 0 = male, c) "Department", referring to which department the respondent is in, d) "Age" in years, e) "Extrovert1-Extrovert6" a multi-item scale referring to each consultant's score on an extroversion personality scale, f) "Knowledge1-Knowledge5" a multi-item scale referring to supervisor's ratings of consultants on their knowledge, g) "CustServ1-CustServ7" a multi item scale of customer ratings of the consultant (this is the core variable). The multi-item scales are all rated on 1-4 point answer scales. Save this dataset to file, import into the software package, and answer the following questions.

a) Calculate the mean of "Tenure" BEFORE replacing missing data. Answer to nearest 2nd decimal (1 mark)
b) Calculate the standard deviation of "Tenure" BEFORE replacing missing data. Answer to the nearest second decimal place (1 mark)
c) Replace missing "Tenure" data with the mean of the variable, and calculate the standard deviation of "Tenure" after replacing missing data. Answer to the nearest second decimal place. (2 marks)
d) Has the standard deviation of Tenure changed after replacing missing data compared to before replacement? (1 mark)

e) Justify / explain why you think your answer in the previous question has happened (3 marks)
f) What would the appropriate choice of central tendency measure be for the variable "Age"? (1 mark)
g) What would the appropriate choice of central tendency measure be for the variable "CustServ1"? (1 mark)
h) What would the appropriate choice of central tendency measure be for the variable "Department"? (1 mark)

Arranging and associating your data. Do the following to the assignment dataset. First, replace all age and tenure missing values with the averages of those variables. Second, create summary Extroversion, Knowledge and Customer Service variables using the method suggested in Readings 1 and 2 (i.e. not replacing individual item missing data, and averaging the items to make aggregate variables). Third, create a correlation matrix of the variables Tenure, Age, Extroversion, Knowledge and Customer Service. Now answer the following with regard to these variables and the theory.

i) Which of the following is true about about correlations? (1 mark)
    i) Correlations run between a low of 0 and a high of 1
    ii) A high correlation shows that one variable causes another
    iii) Correlations close to -1 indicate weak linear relationships
    iv) A correlation of -.89 would indicate that when one variable is high the other tends to be low
j) Report the correlation between tenure and knowledge, rounded to 2 decimals (1 mark)
k) Report the p-value of the correlation between tenure and knowledge, to 2 decimals. Answer to the nearest second decimal. (1 mark)

l) Look at the correlation between tenure and customer service. With regard to this correlation, which of the following is true (1 mark):
   i) This is a large and statistically significant correlation
   ii) This is a small correlation which is significant at the 1% level
   iii) This is a small correlation which is significant at the 5% level
   iv) This correlation is both small and statistically non-significant

m) With regard to the correlation between extroversion and customer service which of the following is true (1 mark):
   i) This is a very strong and statistically significant correlation
   ii) This is a very strong but statistically not significant correlation
   iii) This is a weak to moderate correlation, and is statistically significant
   iv) This is a weak to moderate correlation, but is statistically not significant

n) Explain/justify your answer in the previous question (3 marks)

Estimate the associations between the variable Department and the variables Female, Extroversion, and Customer Service. Answer the following.

o) Which department has the most extroverted consultants? (1 mark)
   i) Department 1
   ii) Department 2
   iii) Department 3
   iv) Department 4

p) What is the average extroversion score for the highest-extroversion department? (2 marks)

q) Which department would you categorise as the most problematic? (2 marks)
   i) Department 1
   ii) Department 2
   iii) Department 3
   iv) Department 4

r) Justify your answer in the previous question (3 marks)

5.3. APPLIED DATA ANALYSIS WITH PRE-GENERATED OUTPUTS

5.3.1. Correlation Analysis

19) Refer to the SAS output *Appendix A* (which should open in your web browser). This output is a correlation analysis of the four variables in "*Dataset_1_Satisfaction*". Assess the correlations between the variables**. (7 marks).**

Answer.

The Pearson Correlation Coefficients gives the correlations. The biggest correlation is between Sat_M and Sat_W (r = .41) ***[1]****, which is a moderate* ***[1]****, positive* ***[1]*** *correlation. Other moderate correlations are between Sat_K and Sat_W (r = .32)* ***[1]*** *and Sat_M and Sat_K (r = .29)* ***[1]****. Correlations between Sat_C and other variables are all negative* ***[1]*** *but small* ***[1]*** *in size (in the r = .10 to .20 range).*

6. Chapter 9: Using Basic Statistics To Check & Fix Data

6.1. APPLIED DATA ANALYSIS WITH PRE-GENERATED OUTPUTS

6.1.1. Missing Data Analysis

20) Refer to the SAS output *Appendix B*, which contains a missing data analysis of the data from "*Dataset_1_Satisfaction*" (using "Code09a Gregs missing data analysis" given with this book). Assess missingness of data.

Answer.

As seen in the first table of the output ("Missing data by observations") there are many observations that are missing 100% of the data, as well as four observations missing 75% and so on. You would probably delete those with 100% missing data, and would need to consider other observations with high degrees of missing data depending on your needs. The second table ("Missing data by variables") shows that Sat_M has the highest proportion of missingness (19%), Sat_W has 7% missing, and so on. However, you should probably delete

observations with 100% missing data (which are meaningless anyway) before looking at final numbers for variables.

6.1.2. Reliability Analysis

21) Refer again the SAS output *Appendix A*, which also contains a Cronbach alpha analysis of the four variables in "*Dataset_1_Satisfaction*". It contains a reliability analysis of four survey items that are designed to form a multi-item scale. Explain in as much detail as possible the following: a) what this analysis is trying to assess / achieve, b) what you would infer from the statistical output, c) all the steps you could engage in after this analysis to deal with and use the scale in a multivariate statistical analysis. **(10 marks)**

Answer

It is an analysis of reliability that seeks to test the following:

- *The extent to which answers given the these multi-item scale items are **consistent** **(1)***
- *This is done by looking at pooled inter-item **correlations** (Cronbach alpha) **(1)***
- *Generally we hope to achieve a Cronbach alpha of .65/.7 or above **(1)***

In the current output:

- *Initial Cronbach Alpha is very poor - .343 **(1)***
- *Analysis of the Cronbach Coefficient Alpha with Deleted Variable table indicates that the item that really does not work with the others is Sat C **(1)***
    - *it has negative correlation with the other variables of -.21 **(1)***
    - *if deleted the alpha would rise dramatically to .62) **(1)***
- *Sat C might either not fit **(1)** or may be a reverse item that the researcher forgot to deal with **(1)***

To use in further analysis:

- *We must try deal with the poor reliability. Options are:*
    - *Check if Sat C is reversed item. If so, reverse and re-check reliability* ***(1)***
    - *If Sat_C is not a reverse item, perhaps remove it and analyse it separately* ***(1)***
    - *However, if we remove Sat C alpha is only .62. You might argue that this is OK, poor or that 3 items are not enough to be sure. Either aggregate all 3 or leave everything as seperate items in final analysis* ***(1)***
- *If we are aggregating items decide on aggragation strategy:*
    - *Sum (if no missing data)* ***(1)***
    - *Average* ***(1)***
    - *Factor analysis can aggregate* ***(1)***

7. Chapter 10: Graphing in SAS

22) Refer to the dataset "Dataset_7_Customer_Gain. This dataset refers to a retail study in which:

- The variable "Customer gain" is a retail company's data on the average proportional growth or loss in customers gained per week by each of its 350 stores (i.e. .15 would indicate a 15% growth rate in customers).
- The main question that the company wants answered is whether one of two new store looks (the variable "**Look**") has materially affected customer gain. The variable store is categorical and has the three values 'Old look' and 'Rebrand 1' and 'Rebrand 2' where 'rebrand' refers to one of two new store looks. The company started implementing rebranding without testing, but store managers report that the new branding appears to be turning customers away or failing to attract walk-ins as before. Is it?
- There are some other variables:
    - ✓ The **Location** of the store ("Mall", "Non-Mall") indicating whether the shop is in a shopping mall or not.
    - ✓ Average **Customer Satisfaction** for each store, and

- ✓ Each store's score on a **Mystery Shopper** rating undertaken by the company

a) Import this data and graph the following in SAS:
   i) A scatter plot of satisfaction versus customer gain;
   ii) The same scatter plot of satisfaction versus customer gain but differentiated (grouped) by store location;
   iii) A bar graph of average customer gain across the three store looks;
   iv) A box-and-whisker graph of customer gain across the three store looks;
b) Interpret each of these graphs for business implications.

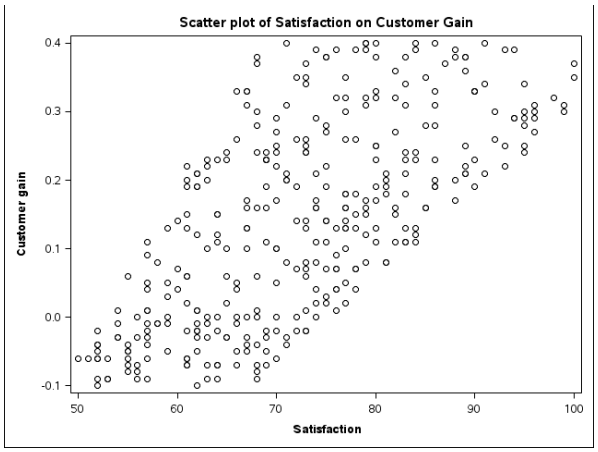
[Marks would depend on whether this was closed or open book]

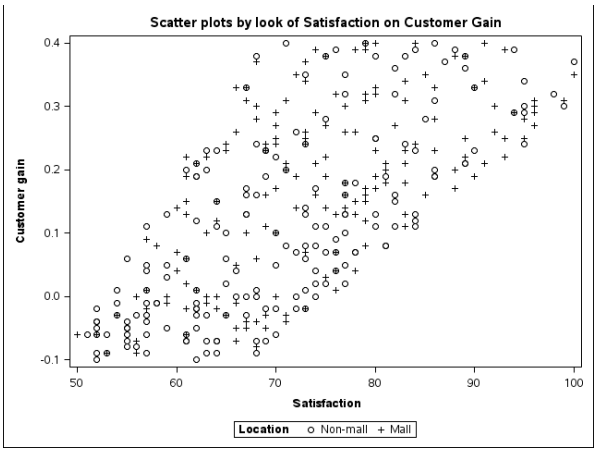
Answers. SAS code for this could be something like the following (note I have made the practice folder the location for the "Practice" library – you need to create your own library. Note also you can make the graphs much 'prettier' see the textbook and SAS helpfiles for how):

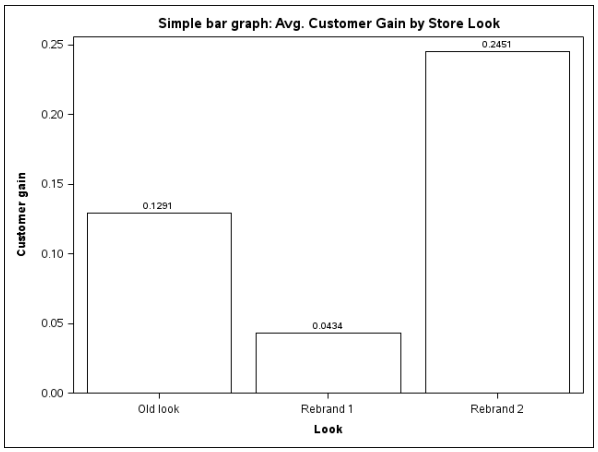
```
/*Scatter graphs*/
title'Scatter plot of Satisfaction on Customer Gain';
proc sgplot data=Practice.Dataset_7_Customer_Gain;
  Scatter y = Customer_Gain x = Satisfaction;
run;
/*Grouped scatter graph*/
title'Scatter plots by location of Satisfaction on Customer Gain';
proc sgplot data=Practice.Dataset_7_Customer_Gain;
  Scatter y = Customer_Gain x = Satisfaction/ group=Location;
run;
/*Simple bar graph*/
Title 'Simple bar graph: Avg. Customer Gain by Store Look';
```

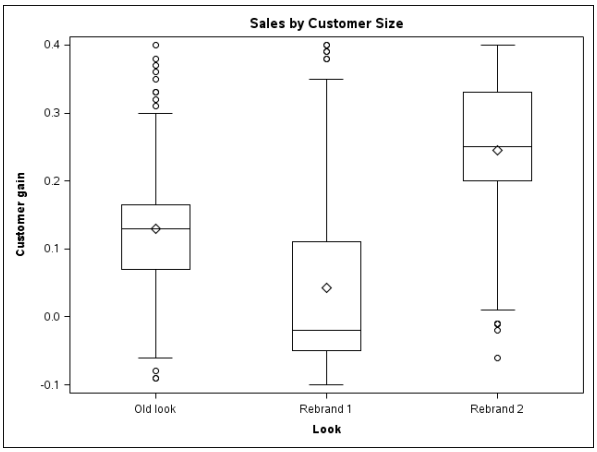
```
proc sgplot data=Practice.Dataset_7_Customer_Gain;
   vbar Look / stat=mean response=Customer_Gain datalabel = Customer_Gain;
run;
/*Box-and-Whisker plot*/
proc sgplot data=Practice.Dataset_7_Customer_Gain;
   title "Sales by Customer Size";
   vbox Customer_gain / category=Look;
run;
```

You should get graphs like the following:









For part b, the first scatter plot tells us there seems to be a fairly strong positive linear association between satisfaction and customer gain (correlation analysis through PROC CORR will reveal a correlation of r = .68, p < .001). Satisfaction surveys of customers may be

a good leading indicator of store success in this case. The grouped scatter plot seems to indicate this relationship holds equally true for both locations (mall and non-mall). The simple bar graph tells us that customer gain is highest for Rebrand 2 (24.5%), then Old Look (12.9%) and lowest for Rebrand 1 (4.3%). This initial evidence perhaps (tentatively!) suggests Rebrand 2 should be chosen for all stores. Finally, the box-and-whisker plot gives us the same averge splits as the bar graph but expresses the median and the spread as well. For instance, Rebrand 1 has several large outliers on the upper end of the scale. Rebrand 2 has a markedly different average and median, it is likely the outliers causing a long upper whisker seen are causing this. Perhaps medians should be used instead?

8. Chapter 11: Fitting Models to Data

23) Imagine a telecommunications company with a database of millions of customers' spending patterns, call times, data downloads, demographics, and the like. They come to you and ask if you can find any useful patterns in this data. What approach to statistics is this called?

Answer: Data mining

9. Chapter 12 Size vs. Accuracy

Accuracy of statistics

24) Say I calculate a variable's mean and also get a 95% confidence interval of .13 to .98. What does this tell you? (1 mark)
   a) That this variable runs from a low of .13 to .98
   b) That two-thirds of the variable lay between .13 and .98
   c) That the variable is definitiely significantly larger than zero
   d) That the variable mean is significantly larger than zero with 95% confidence

25) A statistic that is very statistically significant (1 mark):

a) Is accurate but not necessarily large
b) Is accurate and needs to be large
c) Is a large statistic but can be inaccurate
d) Has an important impact on the world around it

26) A confidence interval of 99% will be (1 mark):
a) Less accurate than a confidence interval of 95%
b) Wider than a confidence interval of 95%
c) Narrower than a confidence interval of 95%
d) The same range as a a confidence interval of 95% but more accurate

27) Bootstrapping is (1 mark):
a) A good way to estimate averages of variables
b) An old-fashioned method of estimating p-values, that has been replaced by technology
c) Often a superior method of estimating confidence intervals

28) Which of the following is true about statistical power (1 mark):
a) Power of .80 tells you there's an 80% chance of a false positive, i.e. that your test finds an effect that does not exist
b) Power of .80 tells you that there is an 80% chance of not finding an effect that does exist
c) Power assesses the ability of a statistical test to find an effect that exists
d) Power assesses the chance that your sample is too small

29) If you have a statistical finding that indicates an impact on profitability with a confidence interval that includes zero and a power value of .04 what would this tell you (2 marks)?
a) The statistic has a significant and powerful impact on profitability

b) The statistic has a non-significant and weak impact on profitability

c) The statistic may have a significant impact on profitability but we can't tell

d) The statistic has a significant but small impact on profitability

30) Say that you have a correlation of .36 that you think is reasonably moderate in size. It has a 95% bootstrapped confidence interval from -.07 to .71 and a p-value = .23. Your post-hoc power for this test is .43. What would you conclude based on these tests, and how would you proceed if this statistic is important to you? **(8 marks)**

Answer. Statistic is moderate in size but non-significant. Power is low so you cannot be sure that the significance should be taken as an indication of non-importance, could just be sample size effect. You could try following:

- *Increase sample size*
- *Lower your test alpha if avoiding Type I error is not as important to you as originally stated*
- *Try changing the variance involved – add covariates not previously used (in this case partial correlations), improve model structure (are you using right kind of correlation?) (.5), re-check measurement of variables to decrease measurement error (.5)*

31) What is a-priori power used for? (2 marks)

Answer. A priori power is used in advance of gathering data to determine necessary characteristics of the study to obtain a feasible result: usually the sample size necessary to find a result of the magnitudes required or the magnitude required given a known sample size.

10. Chapter 13 Regression

10.1. REGRESSION THEORY QUESTIONS

32) When thinking about model structure, which of the following would NOT stop you from using the usual multiple linear regression taught to you on the course? **(1 mark)**

a) The independent variable cause each other
b) There is feedback from the dependent variable to the independent variables
c) The independent variables are independent of each other *(the independent variables are supposed to be independent of each other)*
d) The dependent variable is categorical

33) Which of the following is true of the studentised residual? (1 mark)

a) The studentized residual measures whether an outlier is an influential point *(outliers are not necessarily influential so this is debatable)*
b) A studentized residual score close to zero indicates a poor fitting regression *(opposite is true)*
c) A studentized residual score bigger than 3 or less than -3 is a potentially big outlier
d) The studentized residual statistic is a measure of how much error exists in the whole model *(residuals relate to a single observation only)*

34) Which of the following is not true of heteroskedasticity: (1 mark)

a) It indicates the extent to which the whole regression line fits equally well along the level of the predictors
b) It is a major problem in regression as it 'swings' the slopes *(as seen in the notes, heteroscedasticity does not on average affect slopes dramatically)*
c) It only really affects the confidence intervals
d) It is assessed through the residual plots
e) It can be ameliorated through bootstrapping

35) Autocorrelation can be assessed through: (1 mark)
a) Residual diagnostics like Cooks D
b) The Durbin-Watson statistics
c) Explaining the correlations between variables
d) The R^2 statistics

36) The adjusted R square statistic: (1 mark)
a) Penalises the R square for the addition of more independent variables that do not add much value
b) Picks up non-linearity as opposed to the R^2 which expresses linearity
c) Is usually higher than the raw R square as it is adjusted for error
d) Is a measure of how strongly your independent variable affect each other

37) The ANOVA F statistic in regression: (1 mark)

a) Is an indication of how strongly the independent variables affect the dependent variable *(partly true, but since sample size also affects the p-value here you will often get significant ANOVA F p-values with poor predictoir variables because of a large sample)*
b) Is an indication of whether there are "Far away" outliers
c) Indicates good fit if the p-value is big
d) Is a measure of accuracy for the R square

38) In a regression, what does two VIF scores of 14 and 15 respectively and a condition number of 103 mean, and if you would respond to it what are all the possible responses? (5 marks)

You have likely multicollinearity. Remove one of the variables. Combine the variables if they are similar constructs. Ridge regression. Leave alone as not always problem.

39) In regression, if you have an R^2 = .03 and an ANOVA F p-value of .00, what would you conclude about the fit of the regression, and the relationship between these two statistics? (3 marks)

The regression does not fit – R^2 of .03 indicates that the model only accounts for 3% of the dependent variable interest. The high level of statistical significance is almost certainly driven by excessively large sample size, and simply in this context suggests high accuracy to the low R^2. Assume the regression does not fit.

40) Say I have a regression built from prior years of data that estimates % of staff turnover as a factor of engagement and measured stress levels as follows:

% staff turnover = .13 + .012*Engagement + .0054*Stress

Now, in the current year say your average Engagement score in your unit is 12 and your Stress score is 54. What is your expected level of % of turnover?

Expected % staff turnover = .13 + .012*12 + .0054*54

= 57%

41) In regression, what would a Durbin Watson statistic of .12 infer? (1 mark)

Autocorrelation: correlated residuals

42) In regression what would it mean if you had an R^2 of .32 but an adjusted R^2 of .23? (2 marks)

Since adjusted R^2 includes an adjustment for number of variables and punishes many variables that contribute little this would probably suggest there are too many low-contributing variables.

43) Briefly define what a variance is and why it is so important in regression. (2 marks)

A variance expresses the spread of a variable away from its central point, it is the standard deviation squared. The variance is important in regression because in regression we try to explain the variance of the dependent variable.

44) Briefly explain the concept of endogeneity in regression, how it can be tested, and what methods exist to deal with it **(3 marks)**

Independent variables inferred to cause each other in the process of causing dependent variable ***(1)***

Theory will be main indicator ***(.5)****, can test via structural equation modelling or reasonable correlation sizes* ***(.5)***

Solutions are structural equation modelling or techniques like 2-stage least squares ***(1)***

45) Say that in a regression the original un-bootstrapped 95% confidence interval for a slope is -.88 to .45, and the bootstrapped 95% confidence interval is .19 to 1.03. What do the differences in the two confidence intervals mean, why might this have occurred, and what might this then tell you? (6 marks)

Answers. The un-bootstrapped interval includes zero and therefore suggests that the slope is not statistically significantly different from zero. The bootstrapped interval is substantially different, and notably does not include zero, it suggests a positive slope statistically higher than zero. This difference suggests that the data have certain problematic properties, possibly including issues such as malformed residuals or outliers. This would suggest that the original slopes are suspect, you should find out the root problem and address it.

46) Sometimes regressions are not linear. Write an essay in which you achieve the following. First, imagine and explain in writing and diagrams a tenable regression example in business, that is *not* already given in your notes, in which a specific non-linear pattern might be present. Second, explain in as much detail as possible how you might have identified this non-linearity. Third, explain in as much detail as possible how you would model this non-linearity using a regression. (12 marks).

See the textbook on this topic.

47) What is the null hypothesis test when assessing the p-value of a regression slope? **(2 marks)**

Answer. That the statistic = 0.

48) Explain the Leverage statistic and explain what it should be used for **(1 mark).**

Answer. It assesses the extent to which outliers exist in the independent variables. It should be used as part of an outlier analysis, especially to estimate the location of the outlier effect.

49) For each of the following regression situations explain in as much detail as possible the nature of the problem, and all the major steps you might undertake to fix the problem:

a) More variables than observations (3 marks)
b) A lot of missing data (4 marks)
c) The possibility that the independent variables cause each other (4 marks)
d) A big difference between the raw and adjusted R^2 scores (4 marks)
e) A good R^2 and an ANOVA F with p = .21 (4 marks)
f) A partial residual plot where the residuals are above zero at low and high values of an independent variable and below zero for middle values of the independent variable (4 marks)
g) A residual plot where the residuals are diamond shaped (4 marks)
h) A decent R^2 where none of the independent variable betas are high (3 marks)

(30 marks)

See the textbook for this.

10.2. APPLIED REGRESSION QUESTIONS IN SAS

50) Open the dataset "Dataset_3_Sales"and import it into SAS. This sample dataset has a set of employees, and contains their City (Boston, Miami and New York), Tenure (in months), Age (in years), IQ and average sales figures per week. There is also a variable "Employee" just for identification, do not use it. Run an initial regression (if you are using code from the textbook folder, use "Code13a Multiple regression" or a variant thereof) in which the variable "Sales" is the dependent variable and the independent variables are Boston and Miami (dummy variables for city, so that New York is the missing reference category), Tenure, Age and IQ. With regard to the regression, answer the following:

a) Give the highest variable correlation. (1 mark) *Answer: r = .915 between IQ and Sales.*
b) Give the value of the highest VIF. (1 mark) *Answer: VIF = 1.93 for IQ*
c) In your opinion, is multicollinearity a possible problem in this regression? Explain why it was/was not a problem, mentioning all relevant tests (1 mark) *Answer: No – VIFs are low (< 10), no condition index is > 100, no correlations between independent variables exceeds .90*
d) In your opinion, is non-linearity a possible issue in this regression? Explain briefly why you thought non-linearity was/was not an issue, giving all relevant tests. If you think non-linearity might be an issue, suggest possible solutions without going into too much detail (5 marks)

Answer

Maybe, the main residual plot appears to possibly have a positive curvilinear shape. The partial residual plots show some systematic shapes. Tenure may have the curvilinear shape. IQ also seems curvilinear in the residuals.

Solutions for non-linearity are to use the correct mathematical shape that expresses the particular line, in this case, a mathematical shape for a parabola expresses a curvilinear relationship. Fitting this to the data may give a better fit.

e) What is the value of the highest Cook's D value? Do you think that the highest Cook's D value indicates that the observation is influential? Explain why you think the highest Cook's D is/is not indicative of an influential outlier. (3 marks) *Answer: Highest Cook's D = .31. This is far higher than the next biggest one – more than twice the size. This suggests an influential observation.*

f) For the observation with the highest Cook's D, can you locate in which variables the effect may be operating? Explain which specific statistics and numerical values you used to come to your coinclusion (5 marks) *Answer: The Hats score of the most influential point is high, suggesting the effect is at least partly in the indepenedent variables. Examination of the independent variables for this top row shows that employee 60 has perhaps an unusually combination of high age and tenure (around 60 for each), and a very low IQ (101) and low Sales.*

g) Do you believe that there may be heteroskedasticity in the regression? Explain your answer (why you do/do not believe that heteroskedasticity exists in the regression or why you're not sure), and what, if any, your repsonse might be? (3 marks) *Answer: It's actually hard to tell at this stage since the residuals already show non-linearity. The non-linearity would need to be dealt with and then the residuals re-assessed for heteroskedascticity.*

h) Are the residuals normally distributed? Explain briefly why you do/do not believe the residuals to be normally distributed (1 mark) *Answer: The residuals are mostly normally distributed, since the normality plot of residuals roughly follows the perfect black line of normality, and in the Normal P-P plot the residuals follow the*

upwards sloping diagonal. Residual normality appears fine. However, if you are transforming for non-linearity then you would also need to check after the transformation.

i) Based on all the regression assumption checks, would you do anything to adjust the basic regression? Explain your answer with specific reference to any changes or reasons for not changing. NOTE THAT YOU SHOULD NOT ACTUALLY FINALLY IMPLEMENT ANY CHANGES TO YOUR INITIAL REGRESSION, LEAVE IT AS IT IS. (5 marks)

Answer: Model structure appears appropriate – there is no reason to believe indepednent variables will cause each other here, nor that Sales will have feedback effects on other variables (except possibly for City: is it possible the company might move the best salespeople to a specific location?). Fixing the non-linearity is the major step here, and if done would change all other tests. However, if you believe non-linear analysis is not necessary then running robust regression or deleting or weighting outliers with very high Cooks Ds here may be a possibility. Other tests appear OK as discussed above.

In this section, assume that the initial regression you ran on the data is assumed to fit (i.e. make NO changes to the regression like deleting observations and changing the equation). Answer the following.

j) What is the R square statistic for the equation? (1 mark) *Answer: R^2 = .90*

k) In your opinion, considering the context, do the R square statistics suggest that this regression equation fits? Justify why you do/do not think the R square statistics indicate a good fit (2 marks) *Answer: R^2 = .90 and adjusted R^2 = .89. These are big by almost any standard, suggesting this equation explains about 90% of the variance of Sales. Comparison to other research on Sales would agree.*

l) Give the p-value of the ANOVA F statistic. Does the ANOVA F statistic indicate good fit for the regression, and explain the reason for your answer? (3

marks). *Answer: $p < .0001$. This indicates good fit since it is a low p-value suggesting the R^2 is significantly > 0.*

m) Which independent variables do you consider to have potentially meaningful impacts on Sales? You can pick more than one option. Explain your answer, including a comparison between the slopes. (6 marks). *Answer: I would base this on a combination of effect sizes – especially standardised slopes – and accuracy. IQ is clearly the most important predictor at $\beta = .73$ ($p < .001$). The only other predictor worth mentioning is probably Tenure at $\beta = -.28$ ($p < .001$). Other predictors have small and non-significant betas. The true practical importance can also be assessed from unstandardised slopes, but these need more business context to be judged.*

n) Compare the slopes of Tenure and IQ. Which of the following is true: (1 mark)
   i) IQ has approximately 3 times the impact on the dependent variable that Tenure has (*$\beta = .73$ is just under three times the size of $\beta = -.28$).*
   ii) IQ has approximately 2 times the impact on the dependent variable that Tenure has
   iii) Tenure confidence interval is negative which is a sign it does not fit, unlike IQ
   iv) Tenure's confidence interval is negative which is why it has less influence than IQ

o) What does the unstandardised slope of "Tenure" mean? (1 mark)
   i) When tenure increases by one month, salespeople tend to sell $605.76 more per week
   ii) When tenure increases by one standard deviation, salespeople tend to sell $605.76 more per week
   iii) When tenure increases by one month, salespeople tend to sell $605.76 less per week
   iv) When tenure increases by one standard deviation, salespeople tend to sell $605.76 less per week

p) What does the standardised slope of "Tenure" mean? (1 mark)
   i) When tenure increases by one month, salespeople tend to sell 28% more per week
   ii) When tenure increases by one standard deviation, salespeople tend to sell about $2800 more per week
   iii) When tenure increases by one month, salespeople tend to sell 28% less per week
   iv) When tenure increases by one standard deviation, average sales drops by .28 standard deviations

q) What does the unstandardised slope of "Miami" mean? (1 mark)
   i) Every time another salesperson moves to Miami, sales decrease by $3787
   ii) Being a Miami salesperson means you sell $3787 less than the intercept value
   iii) Miami salespeople sell, on average, $3787 less than those from New York *(slopes of dummy variables compare average levels of the Dependent variable for the category of that dummy variable versus the missing reference category)*
   iv) Miami salespeople sell, on average, $3787 less than those from Boston

r) Which of the following is NOT true of the bootstrapped confidence intervals: (1 mark)
   i) Both the IQ and Tenure slopes are significant at the 95% level
   ii) The IQ and Tenure slopes are have statistically significant confidence intervals at 95% and 99% confidence levels *(you can't know this from running only the 95% intervals – you would at least have to run the 99% intervals as well)*
   iii) Boston's slope is not statistically significant
   iv) IQ and Tenure are the only statistically significant independent variables

s) What do you conclude from the confidence intervals about the influence of the variable "Age" (1 mark):

i) Because the confidence interval includes zero, Age definitely has no impact on the dependent variable *(you don't currently know the 99% interval and, anyway, see point iii below)*

ii) The confidence interval is narrow enough to believe Age is accurate

iii) Age probably has almost no impact on Sales. However, the small sample size might make it non-significant due to power issues

iv) Age is statistically significant at the 90% level

t) [**This question applies financial extrapolation to a regression question. You may wish to read the Extrapolating Statistics to Business Outcomes chapter 17 before answering**].

We estimate that if in the future we use selection procedures based on IQ tests we could increase average new salesperson IQ by 20 points. If this regression is correct, what would we expect the change in ANNUAL sales to be (note that the dependent variable is weekly sales), assuming that weekly gains refer to a 52 week calendar year. Estimate this in various ways:

i) Per-employee annual average sales improvement (4 marks)

ii) Low and high estimates of per-employee annual improvements. Use the non-bootstrapped confidence intervals for consistency (4 marks)

iii) Assuming that we are looking to hire 15 new salespeople, and that in other respects the new hires are identical in city and other variable distributions other than IQ to old employees (in other worsds you should only focus on IQ), what is the total sales improvement? (3 marks)

Answers:

The slope of IQ on sales (holding other variables constant) is $1,078.14, which is a per-week per-employee estimate. The slope has a confidence interval of $899.71 to $1256.57.

*i) Assuming that this regression is stable, an increase in average IQ of 20 would lead to a Sales improvement of $1,078.14*20 per week = $21,562 per week per employee = $1,121,265.60 per annum (i.e. $21.562 * 52).*

*ii) Low and high per-annum gains can be used by using the bootstrapped confidence intervals in place of the slope above (e.g. the low estimate is $899.71 slope * 20 IQ points improvement * 52 weeks in year). The table below shows the estimates.*

Slope level	*Low*	*Medium*	*High*
Slope sizes	*$899.71*	*$1,078.14*	*$1,256.57*
Gain p week p employee	*$ 17,994*	*$ 21,563*	*$ 25,131*
*Gain p year p employee (prev row * 52)*	*$ 935,698*	*$ 1,121,266*	*$ 1,306,833*
*Annual gain over new staff (prev row * 15)*	*$ 14,035,476*	*$ 16,818,984*	*$ 19,602,492*

iii) The final row in the table above simply multiplies the per-employee gains by 15 to reflect the size of the new workforce. Total gains to the company are on average $16,818 million with a 95% confidence interval of $14.035 – $19.602 million.

10.3. APPLIED REGRESSION QUESTIONS WITH GENERATED PRINTOUTS

51) Loss to Competitors Regression

This question uses the SAS printout (which should open in your webpage browser) "Appendix C Loss to competitors regression". This question refers to a regression study in which:

- The dependent variable is "Loss to competitors", and is managers' estimated percentage loss of staff to other organizations.
- There are 6 independent variables, namely:
    - the type of organization ("Type"), separated into Services, Manufacturing, and Government
    - the extent to which the company gives good *job aspects* (entitled "Job")
    - the extent to which the company gives good *compensation* ("Pay")
    - the extent to which the company gives good *development opportunities* ("Development")
    - the extent to which the company has good *environmental aspects* ("Environment")
    - the extent to which outside factors are adverse to retention and replacement of staff ("*External Factors*" – a higher score in this column means that retention and replacement of staff is harder).

In Appendix C you will see a SAS output of this regression. Answer the following questions.

a) Evaluate - to the greatest extent possible within the time given - the suitability of the regression for the various data assumptions. In doing so, refer **explicitly** to parts of the output that you are referencing for your analysis, and where necessary to specific statistical values (e.g. "as seen in the *ABC* table, the largest *XXX* value is only *XYZ* which suggests *ZZZ* is not problematic" or the like)

(15 marks)

Answers:

- *Check <u>model structure</u>. Theory may possibly support inter-predictor causation, for instance higher development might increase pay as more skilled staff are presumably paid more, and good job aspects and environment might seem to reinforce each other.*

The sizes of some independent variable correlations may bear this out, for instance the two former relationships are correlated .65 and .77. This may infer endogeneity. There may also be feedback: loss to competitors may affect subsequent HR decisions such as how much to pay staff (2 marks).

- *Check multicollinearity, which is not a problem here because no independent variable is correlated more than .90, no VIF is > 10 or relatively high to others, no condition index is >100 (3 marks).*
- *Assess nonlinearity, no residual plot has an obvious nonlinear pattern*
- *Assess outliers. Cook Ds in Column C11: higher ones are not significantly higher than those below, no unlikely that there's overly influential outliers overall. Top row's RStudent is 3.27 which is >3 therefore high but seemingly not influential. Some of this row's DFBetas are high compared to others, indicating possible influence in some of the independent variables*
- *Assess heteroskedasticity. None of the residual plots are significantly uneven from top to bottom so regression is relatively homoscedastic, although not perfectly so.*
- *Assess residual normality. The histogram of residuals does not follow a perfect normal line and the normality plot does not quite follow the 45 degree line upwards, which may indicate some non-normality which can be expected to affect accuracy of confidence intervals.*
- *Assess autocorrelation. The Durbin Watson stat is between 2 and 4 therefore within acceptable bounds*

b) If you feel that any given data assumption is a problem, then suggest possible remedies

(7 marks)

Answers. Potential endogeneity or feedback dealt with through structural equation modeling. Possible non-normality and/or mild heteroskedasticity can be addressed through transformations, bootstrapping or weighted regression, although bootstrapping is probably

sufficient here. Outliers addressed by robust regression, possible comparison of regression with and without biggest outlier, bootstrapping.

c) Assuming data assumptions in this output were found to be satisfied, use the rest of the output to evaluate the global fit of the regression, i.e. fit that is evaluated after data assumptions have been assessed and sorted out. Again refer to specific parts of the output and specific statistics. In your opinion does the regression fit?

(8 marks)

Answers. R^2 = .07 and adjusted R^2 = .03, indicating that only about 7% of the variance of the dependent variable is explained by the model (3% adjusting for number of independent variables indicating a large no. of poor quality predictors). The ANOVA F has a p-value >.10, suggesting poor fit (R^2 not significantly different from 0), although this may be due to small sample it's more a confirmation of small effect. Regression does not fit.

d) Assuming data assumptions and other fit statistics in this output were found to be sufficient for fit, evaluate the actual regression equation, notably suggesting which independent variables are more or less strongly associated with the outcome variable and in what way they are associated. Again refer to specific parts of the output and specific statistics.

(10 marks)

Answer. None of the independent variables has a particularly strong effect on Loss. The highest is Job with a standardized coefficient of .28 (at best a weakly moderate effect), which is double the effect of the next highest. Job's slope is the only one significant at a 95% confidence level according to the bootstrap confidence intervals, its unstandardised effect indicates that for every 1 unit increase in the Job variable there is a 7.58% decrease in Loss to Competitors. One may also consider External factors (β = .15, B = 8.95, p < 10), but this is weak.

e) Say hypothetically that you decide the regression fits and is suitable for prediction. Using the current equation, therefore, if you have a manufacturing company with scores of Job = 5.5, Pay = 6, Development = 7, Environment = 5, and External Factors = 4, what would you predict the Loss to Competitors factor to be for this company?

(5 marks)

Answer.

*Forecast = 36.6870 -0.8950*7 + 4.4739*5 + 8.9485*4 -7.5807*5.5 + 2.6287*6 -11.0206(1) = 51.64.*

[TOTAL FOR QUESTION = 45 MARKS]

11. Chapter 14 Categories Explaining a Continuous Variable

11.1. THEORY QUESTIONS

52) What are the two main assumptions of a t-test? (2 marks)

Answer. Normality of data, equality of variances.

53) Explain the importance and use of the Equality of Variances test in a T-Test **(3 marks).**

The equality of variances tests is one of the tests that can be used to assess the appropriateness of a parametric T-Test, which assume equal variances. Failing equal variances, adjusted T-Tests or non-parametric tests may be more appropriate.

54) When comparing means across groups, as we do in t-tests when comparing two groups, what are the four main remedies for non-normal data or data with unequal variances? (8 marks)

Answer. Transform the dependent variable to normality. Bootstrap. Use alternate, corrected versions of the test. Use non-parametric tests. (2 marks each)

55) Briefly explain how some non-parametric tests use ranks to solve data issues. (5 marks)

Answer. If we use the ranks of a set of data, the distribution becomes near-perfect (e.g. if there are no ties the distribution is perfectly uniform, ties add just a little more complexity). Distances between data points that could create outliers, for instance, become irrelevant, as do other distributional issues. This creates data that is easy to compare.

11.2. APPLIED DATA ANALYSIS IN SAS

56) Access the dataset "Dataset_7_customer_gain" in the practice folder in SAS. See question 22) on page 17 for the explanation of the variables. Run a parametric and non-parametric t-test to check whether mystery shopper scores differ significantly between mall and non-mall locations. Comment on all aspects of the analysis, including assumptions, findings and weaknesses of this approach. (15 marks)

Answers. First, the SAS code may be something like (assuming you linked the folder to a library called "Practice"):

```
ods graphics on;
Proc TTest data=Practice.Dataset_7_customer_gain;
   Class Location;
   Var Customer_gain;
Run;
proc npar1way data=Practice.Dataset_7_customer_gain wilcoxon HL;
```

```
   Class Location;
   Var Customer_gain;
run;
ods graphics off;
```

First, we assess assumptions.

- *The groups are roughly equal in size, which is a good start.*
- *Analysis of the plots may indicate non-normality.*
- *The equality of variances test is non-significant (p = .3156) so we do not reject the hypothesis that the variances are equal, which is good. The actual standard deviations are .135 vs. .1457 which seem close together but this could be debatable, and graphs seem to have roughly equally distributed data.*

We could transform the variables to deal with the non-normality, but perhaps let us proceed for now without that. Then, the t-test p-values are significant, suggesting a statistically significant difference, and the confidence interval for the pooled difference between the means is .0358 to .095.

Location	Method	Mean	95% CL Mean	
Mall		0.1752	0.1547	0.1956
Non-mall		0.1098	0.0883	0.1312
Diff (1-2)	Pooled	0.0654	0.0358	0.095
Diff (1-2)	Satterthwaite	0.0654	0.0359	0.0949

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	348	4.35	<.0001
Satterthwaite	Unequal	347.87	4.36	<.0001

The actual means are 17.52% growth for malls versus 10.98% growth in non-mall locations, so this test seems to suggest that malls have statistically significantly higher growth than non-mall locations.

The non-parametric test seems to agree, the Wilcoxon test is significant at p < .001 (so we reject a zero different between the groups) and the Hodges-Lehmann confidence interval is also above zero suggesting significant differences.

This test has a substantial weakness (among others): there are no controls for the many, many other other variables that could affect customer gain, so this is possibly an unrealistic result unless it perseveres when these other variables are accounted for at the same time.

12. Chapter 15: Categorical Associations

12.1. APPLIED DATA ANALYSIS IN SAS

57) Access the dataset "Dataset_7_customer_gain" in the practice folder in SAS. See question 22) on page 17 for the explanation of the variables. Focussing on the categorical data, answer the following questions:

a) The industry norm for mall versus non-mall locations is 60% 30%.

i) How does our retail organization compare – are we statistically significantly different in store distribution from the norm?

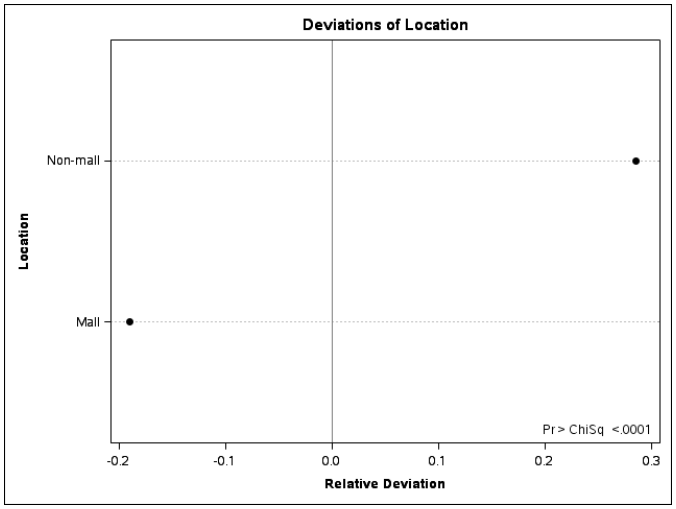
ii) Say that the strategy actually believes malls will fall out of consumer favour – is your finding desirable or undesirable as an indicator of competitive advantage given that assumption?

Answer. This is a classic one-way chi-square test. The SAS code might be something like the following, see the book:

```
proc freq data=Practice.Dataset_7_customer_gain;
   Tables Location/out=Locations;
run;
proc sort data=Locations;
  by Location;
run;
ods graphics on;
```

```
proc freq data=Locations order=data;
   tables Location / nocum chisq testp=(60 40)
            plots(only)=deviationplot(type=dotplot);
   weight Count;
   title 'Locations of malls';
run;
ods graphics off;
```

In our preliminary results, we see that 48.57% of our stores are currently in malls and the rest not. The chi-square statistic is significant, suggesting that our proportions are significantly different from those of the industry. The relative deviation plot is an elegant way to see how big the deviations are: our mall percentage is about 20% different from the benchmark of 60% (we have 48%, benchmark is 60%, deviation is [48-60]/60 = -12/60 = 20) and our non-mall percentage is about 30% different.



This is competitively good supposing the strategy is correct that consumers will start trending away from malls – we are less vested in malls now than our competitors. However, previously we saw previous customer gain in malls to be higher: we'd need to ponder the truth of the assumption!

b) Have the looks (rebrands and retention of old looks) been applied equally across the mall versus non-mall locations?

Answer. This is a classic two-way chi-square test. The SAS code might be something like the following, see the book:

```
proc freq data=Practice.Dataset_7_customer_gain noprint;
   Tables Location*Look/out=Combinations;
run;
proc sort data=Combinations;
  by Location Look;
run;
ods graphics on;
proc freq data=Combinations;
  tables Location*Look / Chisq;
  weight Count;
  title 'Location &Look proportions';
run;
ods graphics off;
```

In this case, the Chi-Square, Likelihood Ratio Chi-Square and Mantel-Haenszel Chi-Square all agree (because of high p-values) that there does not seem to be any association between look and location.

13. Chapter 16: Business Reporting with SAS

There are no questions on this chapter as it draws on a combination of previous chapters.

14. Chapter 17: Extrapolating Stats to Business Outcomes

58) Say you wish to evaluate the effect of training spend on the sales levels of your salesforce. Including various other control variables, you do a regression that

estimates an unstandardized slope for the independent variable "Training spend" (annual Dollars spent on employees on training) on the dependent variable "Sales" (annual sales of the employees) of B = 1.454, p < .001. However, training also has indirect costs like the cost of having employees away from their jobs – for every Dollar spent on training you estimate that it costs another $.19 in indirect costs. You spent $2.67 million training your sales force last year, and the salesforce generated sales revenues worth $41,997,252. This year, you wish to argue for an increase in the training budget to $40 million. Answer the following:

a) What was the profitability of training last year assuming the regression and other calculations are correct? **(4 marks)**
b) What is the ROI of last year's training? **(2 marks)**
c) If the company's required cost of capital on such projects is 15%, should the company increase the training budget to $4 million? Argue both using profitability and ROI. **(5 marks)**
d) What would be the problems with applying this model to forecasting future profitability of training budget increases?**(3 marks)**

Answers:

(a)

Profitability of last year's training assuming regression model is true is:

$2.67million(1.454 – 1.19) = $704,880*

[The reason for the 1.19 is that every Dollar of direct training spend is an expense as well as an extra $1.90 so $1+$1.90 = a multiplier of 1.19]

(b)

*The ROI is Profit/Cost so $704,880/1.19*2.67million = $0.22 = 22%*

(c)

The ROI of 22% exceeds the 15% cost of capital suggesting profitability to training. So long as more training makes the same returns – which is debatable – increasing investment to $4m

makes sense. Note that the reference to total sales levels does not really have a role to play here.

(d)

If the regression model and slopes are not stable over time then it will not hold. Also, the effect of training may not have an ongoing, linear effect in the same person (i.e. every Rand spent garners an extra $12.54 in sales no matter how much training has come before for that person) – it is more plausible that the effect of training spend on a given person depends on factors like prior experience and training.

15. Chapter 18: Miscellaneous Business Statistics Topics

15.1.1. Big data

59) What are the characteristics of big data, and some solutions for dealing with big data that you have learned about in the book?

See the textbook for the basic theory on this topic.

15.1.2. Data warehousing

60) Describe how traditional data warehouses work.

See the textbook for the basic theory on this topic.

61) (For practicing managers / organizational members): If your organization has a data warehouse and you have contact with it comment on whether it fulfills the needs of a good warehouse in terms of this book, and whether you believe it needs changes (explain your answer either way – for or against change). If you do not have a warehouse, do you believe in terms of this book's description of them that one would suit your organization? Be careful to reference the specific nature of your organization in answering this. (If you have no contact with a warehouse that is there, answer the second part pretending you had none).

Answers can vary.

15.1.3. Simulation

62) Give a practical business example of mathematical simulation *not* covered in your course reading or the videos shown to you in class, if possible from your personal working environment or experience. Go into as much detail as possible about how the simulation might work, including if possible numerical examples or figures such as flow charts, and including as much discussion on the principles of simulation and the OR method as possible. Note, however, that you do not have longer than about 18-20 minutes and you are not expected to produce more than could be expected in such a time.

[TOTAL FOR QUESTION = 10 MARKS]

See the simulation section.