

Business Statistics Made Easy in SAS®

Gregory Lee

Practice Exercises



support.sas.com/bookstore

This set of Practice Exercises is a companion piece to the following SAS Press book: Lee, Gregory. Business Statistics Made Easy in SAS®. Copyright © 2015, SAS Institute Inc., Cary, NC, USA. ALL RIGHTS RESERVED.

Business Statistics Made Easy in SAS

By Prof Gregory John Lee

Practice Exercises (v.1, no answers)

Table of Contents

| | |
|---|-----------|
| PREFACE: HOW TO USE THESE PRACTICE QUESTIONS..... | 3 |
| 1. CHAPTER 2 INTRODUCTION TO THE STATISTICAL PROCESS..... | 5 |
| 2. CHAPTERS 3 & 4 INTRODUCTION TO DATA EXERCISES..... | 5 |
| 2.1. Type of Variable..... | 5 |
| 2.2. Multi-Item Scales..... | 6 |
| 3. CHAPTER 5: INTRODUCTION TO SAS..... | 6 |
| 4. CHAPTER 6: SAS PROGRAMS, DATA MANIPULATION, ANALYSIS & REPORTING..... | 6 |
| 4.1. SAS Programming Theory..... | 6 |
| 4.2. Combining Datasets..... | 7 |
| 4.3. SAS Output / Reporting..... | 7 |
| 5. CHAPTERS 7 & 8: DESCRIPTIVE & ASSOCIATIONAL STATISTICS..... | 7 |
| 5.1. Theory..... | 7 |
| 5.1.1. Measures of Centrality & Spread..... | 7 |
| 5.1.2. Correlations & Causation..... | 9 |
| 5.2. Applied Simple Data Analysis in SAS..... | 9 |
| 5.3. Applied Data Analysis with Pre-Generated Outputs..... | 12 |
| 5.3.1. Correlation Analysis..... | 12 |
| 6. CHAPTER 9: USING BASIC STATISTICS TO CHECK & FIX DATA..... | 12 |
| 6.1. Applied Data Analysis with Pre-Generated Outputs..... | 12 |
| 6.1.1. Missing Data Analysis..... | 12 |

| | | |
|---------|---|----|
| 6.1.2. | <i>Reliability Analysis</i> | 13 |
| 7. | CHAPTER 10: GRAPHING IN SAS | 13 |
| 8. | CHAPTER 11: FITTING MODELS TO DATA | 14 |
| 9. | CHAPTER 12 SIZE VS. ACCURACY | 14 |
| 10. | CHAPTER 13 REGRESSION | 16 |
| 10.1. | Regression Theory Questions | 16 |
| 10.2. | Applied Regression Questions in SAS | 20 |
| 10.3. | Applied Regression Questions with Generated Printouts | 24 |
| 11. | CHAPTER 14 CATEGORIES EXPLAINING A CONTINUOUS VARIABLE 26 | |
| 11.1. | Theory Questions | 26 |
| 11.2. | Applied Data Analysis in SAS | 26 |
| 12. | CHAPTER 15: CATEGORICAL ASSOCIATIONS | 27 |
| 12.1. | Applied Data Analysis in SAS | 27 |
| 13. | CHAPTER 16: BUSINESS REPORTING WITH SAS | 27 |
| 14. | CHAPTER 17: EXTRAPOLATING STATS TO BUSINESS OUTCOMES | 27 |
| 15. | CHAPTER 18: MISCELLANEOUS BUSINESS STATISTICS TOPICS | 28 |
| 15.1.1. | <i>Big data</i> | 28 |
| 15.1.2. | <i>Data warehousing</i> | 28 |
| 15.1.3. | <i>Simulation</i> | 29 |

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

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
 - b) Interval
 - c) Ordinal
 - d) Categorical.
- 4) Which of the following types of data would “Industry” be (i.e. ascertaining what industry a company is in)? (1 mark)
 - a) Ratio
 - b) Interval
 - c) Ordinal
 - d) Categorical
 - e) 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)

2.2. 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)

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
```

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.
- What does the "HR" refer to in the dataset name? (2)
 - 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)

4.3. SAS OUTPUT / REPORTING

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

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*:
- The average is a measure of central tendency
 - The average is appropriate for continuous variables
 - The average is appropriate for ordinal variables
 - 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*:
- 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
- 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
- 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
- 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
- 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
- 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?

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)

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).

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.

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)

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]

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?

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

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
 - 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
- b) A studentized residual score close to zero indicates a poor fitting regression
- 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

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
- c) It usually 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^2 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 always an indication of how strongly the independent variables affect the dependent variable
- 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)

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)

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:

$$\% \text{ staff turnover} = .13 + .012 * \text{Engagement} + .0054 * \text{Stress}$$

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?

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

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

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

- 44) Briefly explain the concept of endogeneity in regression, how it can be tested, and what methods exist to deal with it **(3 marks)**
- 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)
- 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).
- 47) What is the null hypothesis test when assessing the p-value of a regression slope? **(2 marks)**
- 48) Explain the Leverage statistic and explain what it should be used for **(1 mark)**.
- 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:
- More variables than observations (3 marks)
 - A lot of missing data (4 marks)
 - 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)

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)
- b) Give the value of the highest VIF. (1 mark)
- 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)
- 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)

- 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)
- 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 conclusion (5 marks)
- 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 response might be? (3 marks)
- h) Are the residuals normally distributed? Explain briefly why you do/do not believe the residuals to be normally distributed (1 mark)
- 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)

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)
- 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)

- 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).
- 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).
- 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
 - 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
 - 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
 - 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
 - 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 words you should only focus on IQ), what is the total sales improvement? (3 marks)

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)

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

(7 marks)

- 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)

- 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)

- 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)

[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)
- 53) Explain the importance and use of the Equality of Variances test in a T-Test (3 marks).
- 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)
- 55) Briefly explain how some non-parametric tests use ranks to solve data issues. (5 marks)

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 13 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)

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 13 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?
- b) Have the looks (rebrands and retention of old looks) been applied equally across the mall versus non-mall locations?

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)**

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?

15.1.2. Data warehousing

- 60) Describe how traditional data warehouses work.
- 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).

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]