Kristin Fridgeirsdottir
EMBALJ Spring 2014

Data Analytics for Management
Individual Assignment

Instructions:

Please read carefully and answer the questions. The data files that the questions refer to are posted on Portal under Session 6. Please communicate clearly your reasoning for the answers.

We will not read through submitted Excel files but you can copy tables and graphs from Excel to your report.

This is an individual assignment, which means that you should not consult with anyone.

The report is due on **Wednesday 14 May 2014 at midnight**. Please submit via Turnitin UK under subject: "DAM Individual Assignment".

This assignment counts for 50% of the final grade.

Question 1. Hotel Bookings (20%)

During high season, hotels often take on more bookings than the number of rooms they have. The reason for this is that often a certain portion of bookings get cancelled last minute. However, it can happen that everyone shows up and there are not enough rooms available. This overbooking situation can be costly for the hotel as they need to send their possibly unhappy guests to another hotel.

Hotel Reykjavik Natura is a newly renovated hotel next to Reykjavik's domestic airport and Reykjavik University. It has 220 rooms and is relatively conveniently located for downtown, only 5 min driving away.

The hotel is very busy in June, especially on the 17 June, Iceland's Independence Day. However, in the past they have had some guests cancelling last minute opting to travel out of town instead.

Hotel Reykjavik Natura is interested in understanding the implication of taking on more bookings than there are rooms. They estimate the cancelling probability on 17 June to be 4%.

The so-called binomial distribution can be used to describe uncertain situations of this kind where there are two possible outcomes; a guest shows up or not. Fortunately, the binomial distribution can be approximated by the normal distribution such that the number of guests that show up is normally distributed with a mean of $n \cdot p$ and a standard deviation of $\sqrt{n \cdot p \cdot (1-p)}$ where n is the number of rooms booked and p is the probability of showing up.

a) If the hotel books 228 rooms on 17 June what is the probability that there will be an overbooking situation?

b) If the hotel books 235 rooms on 17 June what is the probability of having empty rooms?

c) If the hotel wants only 2% probability of an overbooking situation on 17 June, how many rooms should it book?

d) You have been hired as a consultant for Hotel Reykjavik Natura. You have understood the probability structure of the overbooking problem and the next step is to advise them on how many rooms to book during peak season. Please make a list of the data you would need to get from the hotel in order to make your recommendation.

Question 2. Managing Subscriptions at London Today (20%)

London Today is a growing newspaper that has been giving out trial subscriptions to potential customers. To ensure that as many trial subscriptions as possible are converted to regular subscriptions, the London Today marketing department works closely with the distribution department to accomplish a smooth initial delivery process for the trial subscription customers. To assist in this effort, the marketing department needs to be able to better forecast the number of new regular subscriptions for the coming months.

A team consisting of managers from the marketing and distribution departments was convened to develop a better method of forecasting new subscriptions. Previously, after examining new subscription data for the previous 2 or 3 months, a group of three managers would develop a consensus on what the final forecast should be. Julia Gibson, who was recently hired by the company to provide special skills in quantitative forecasting methods, suggested that the department look for factors that might be helpful in predicting new subscriptions.

Members of the team noted that the forecasts in the last year had been particularly inaccurate because in some months much more time was spent on telemarketing than in other months. In particular, in the last month, only 1,055 hours were completed since callers were busy during the first week of the month attending training sessions on the personal but formal greeting style and a new standard presentation guide. Julia suggested that the data for the number of new subscriptions and the number of hours spent on telemarketing for new subscriptions for each month for the past 2 years be obtained from company records.

a) *What criticism can you make concerning the forecasting method the group of three managers used that involved taking the new subscriptions for the past 3 months as the basis for future projections?*

b) *What factors other than number of telemarketing hours spent might be useful in predicting the number of new subscriptions? Explain.*

c) *Analyze the data in the file,* Subscription data.xlsx, *and develop a statistical model to predict the average number of new subscriptions for a month based on the number of hours spent on telemarketing for new subscriptions. Interpret the model and comment on its quality.*

d) *If there are expected to be 1,200 hours spent on telemarketing in the coming month, estimate the average number of new subscriptions expected for the month. Indicate the assumptions upon which this prediction is based. Do you think these assumptions are valid? Explain.*

e) *Estimate the average number of new subscriptions for a month in which 2,000 hours were spent on telemarketing.*

Question 3. TV Movie Ratings (30%)

INN, RUV and S2, are the three main TV networks competing for the same audience. These networks rarely produce their own movies but rather contract with independent producers to have them made. When buying a movie from a producer, S2 is mainly concerned with attracting a large audience. A network's success in attracting a large audience is reflected in its Nielsen ratings (expressed as percentage of all households with televisions). S2 has collected data on ratings for 88 different movies that have been showed on the three main networks. They believe there are different factors that affect the ratings of a movie including whether the movie is based on true events and number of movie stars acting in it. They also believe that the rating for the program preceding the movie on the same network matters as well as average ratings of the programs on the two competing networks during the time the movie is broadcasted. The data they have collected can be found in Movie ratings.xlsx with the following variables:

Variable	**Description**
NETWORK	Broadcasting network (INN, RUV, S2)
RATING	Nielsen rating for movie
FACT	1 = based on true events, 0 = fictional
STARS	Number of famous actors or actresses
PREVIOUS RATING	Nielsen rating for program immediately preceding movie on same network
COMPETITION	Average of Nielsen ratings received by the two competing networks during the movie's broadcast

a) *Develop a regression model that would help to predict movie ratings.*

b) *Compare the movie ratings of the three networks INN, RUV, and S2.*

c) *The conventional industry wisdom is that fact-based movies have higher ratings than movies based on fictional stories. Do you agree with this view? Please explain.*

d) *On Sunday nights, S2 usually presents "Jeremy and Julia" at 8:00pm, followed by the Sunday night movie at 9:00pm. Typical ratings for "Jeremy and Julia" are 17.5. This week, S2 is considering replacing "Jeremy and Julia" with a live concert with the Icelandic group Sigurrós that is expected to garner a rating of 20 points.*

   Determine the expected change in ratings for the Sunday movie.

Question 4. Cash Withdrawal (30%)

A bank wants to analyse what variables have an effect on the amount of cash withdrawn from automatic teller machines (ATMs) located in residential neighbourhoods. A sample of total daily withdrawals from ATMs has been collected, together with information suspected to affect withdrawals. This information includes the median value of homes in the neighbourhood, the median family income in the neighbourhood, the average checking balance of customers in the neighbourhood, the distance to the next nearest ATM, and whether or not the withdrawals occurred on a weekend. A part of the data set is given in Table 1.

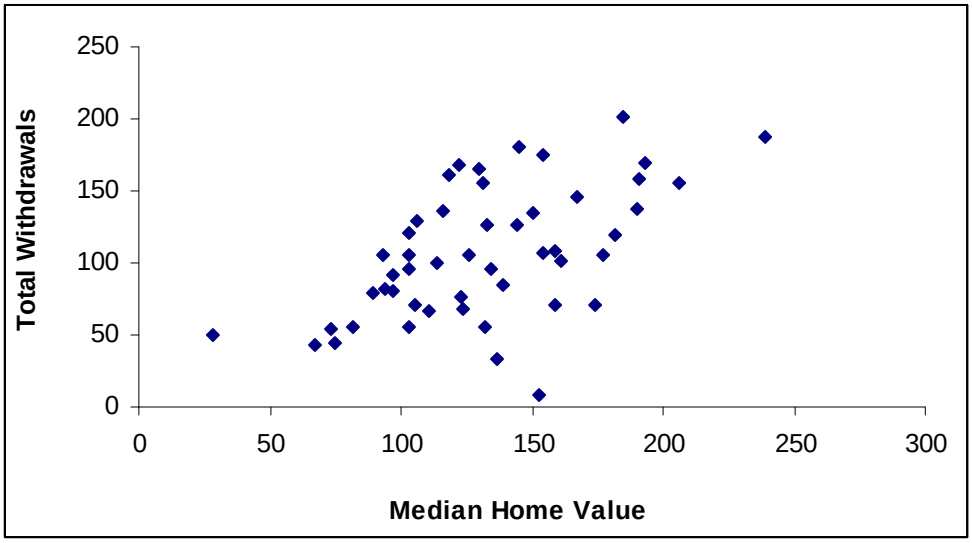
Table 1. Cash Withdrawal data

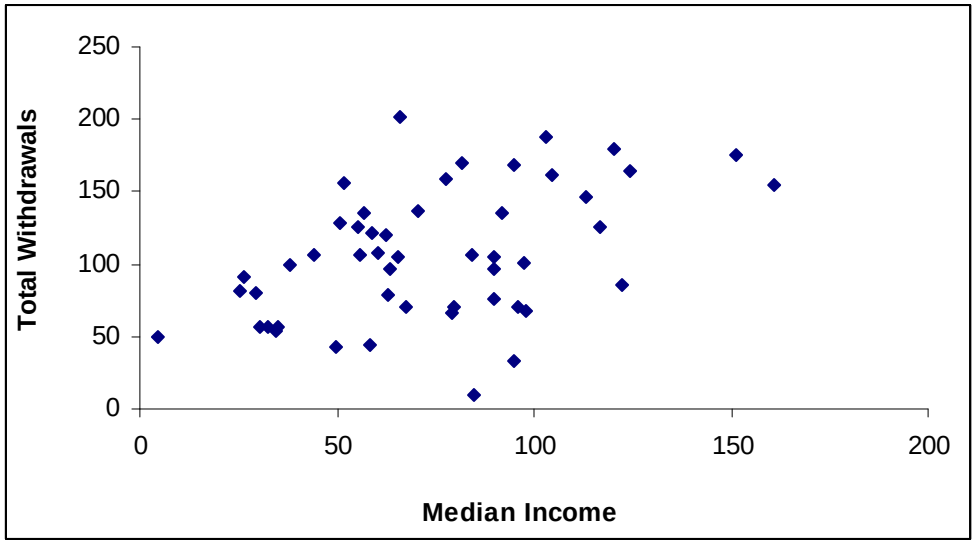
Total Daily Withdrawals (£1,000s)	Median Home Value (£1,000s)	Median Income (£1,000s)	Average Checking Balance (£1,000s)	Distance to Next ATM (miles)	Weekend
107	154	55.9	1690	4.6	No
56	82	30.6	891	4.4	No
50	28	4.5	60	4.1	Yes
135	150	56.7	1651	1.4	No
56	103	32.5	914	4.5	No
⋮	⋮	⋮	⋮	⋮	⋮
33	137	94.7	1164	7.4	No
105	103	90.1	1093	4.3	Yes
68	124	98	1269	6.1	No
126	133	116.6	1762	4.3	No
71	174	95.8	1483	5.8	No

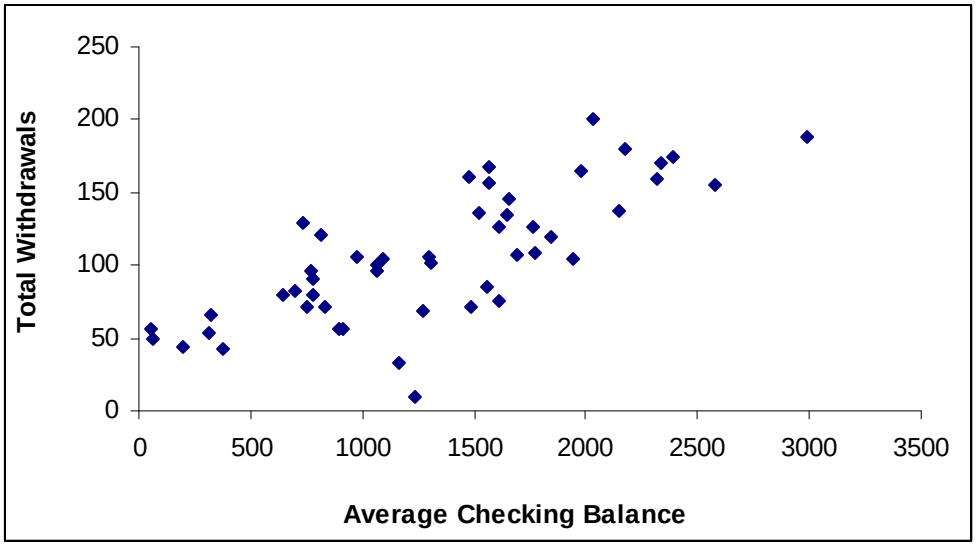
Table 2 contains the correlation coefficients between the different variables, where a dummy variable is used to indicate whether the withdrawal occurred on a weekend (1) or not (0). Also scatter plots between all the independent variables and total withdrawals are given.

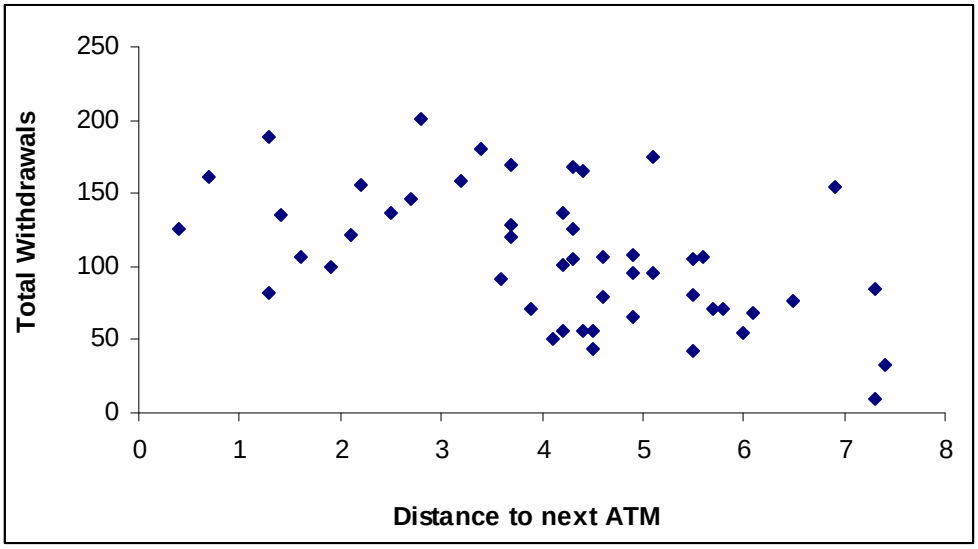
Table 2. Correlation matrix

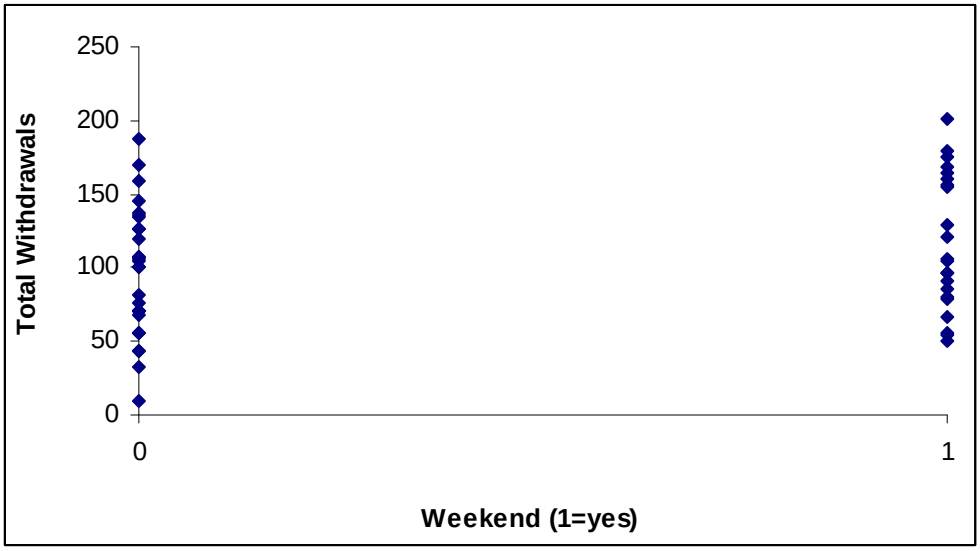
	Withdrawals	*Med. Home Value*	*Med. Income*	*Avg. Checking Balance*	*Distance Next ATM*	*Weekend*
Withdrawals	**1.00**					
Med. Home Value	**0.56**	**1.00**				
Med. Income	**0.45**	**0.52**	**1.00**			
Avg. Checking Balance	**0.76**	**0.84**	**0.63**	**1.00**		
Distance Next ATM	**-0.51**	**-0.06**	**0.26**	**-0.15**	**1.00**	
Weekend	**0.21**	**-0.30**	**0.09**	**-0.20**	**0.09**	**1.00**


Total Withdrawals
250
200
150
100
50
0
0
50
100
150
200
250
300
Median Home Value


Total Withdrawals
250
200
150
100
50
0
0
50
100
150
200
Median Income


Total Withdrawals
250
200
150
100
50
0
0
500
1000
1500
2000
2500
3000
3500
Average Checking Balance





a) Explain in practical terms the meaning of the correlation coefficient between the 'Distance to Next ATM' variable and the total withdrawals.

A multiple regression analysis was performed, with the results given in Table 3.

Table 3. Multiple Regression Results

Regression Statistics	
Multiple R	**0.95**
R Square	**0.91**
Adjusted R Square	**0.89**
Standard Error	**14.72**
Observations	**50**

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**69.527**	**10.118**	**6.87**	**0.00**	**49.135**	**89.919**
Median Home Value	**0.004**	**0.100**	**0.04**	**0.97**	**-0.198**	**0.206**
Median Income	**0.105**	**0.096**	**1.09**	**0.28**	**-0.089**	**0.298**
Average Checking Balance	**0.048**	**0.007**	**7.08**	**0.00**	**0.034**	**0.062**
Distance to Next ATM	**-12.076**	**1.400**	**-8.63**	**0.00**	**-14.898**	**-9.254**
Dummy Weekend	**35.163**	**4.631**	**7.59**	**0.00**	**25.829**	**44.497**

b) Do the results suggest that there is a relationship between the median family income in the neighbourhood and total ATM withdrawals in that neighbourhood?

Based on the results in Table 3, a new regression model was run. The results are given in Table 4.

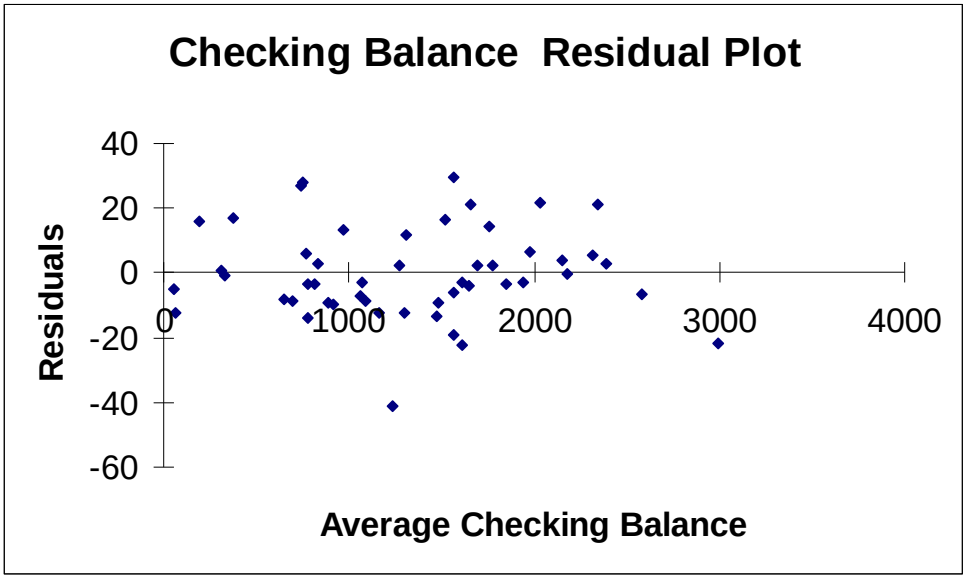
Table 4. Multiple Regression Results – Revised Model

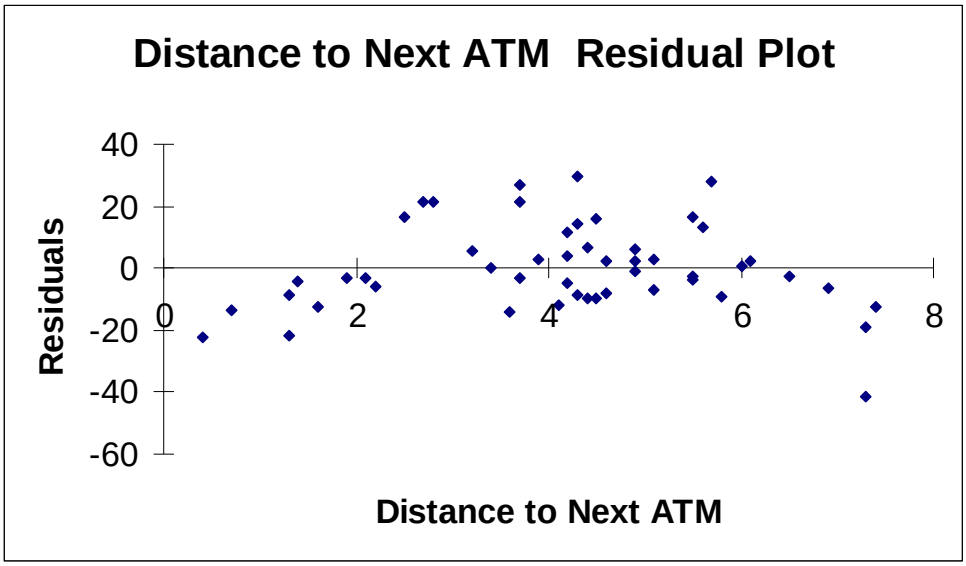
Regression Statistics	
Multiple R	**0.95**
R Square	**0.90**
Adjusted R Square	**0.90**
Standard Error	**14.59**
Observations	**50**

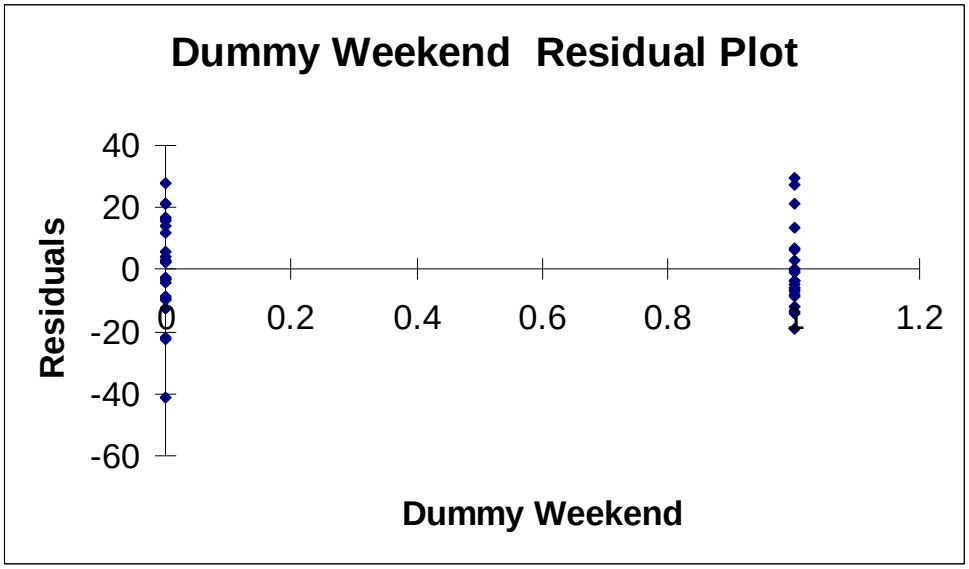
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**69.171**	**7.646**	**9.05**	**0.00**	**53.780**	**84.561**
Average Checking Balance	**0.052**	**0.003**	**16.50**	**0.00**	**0.046**	**0.058**
Distance to Next ATM	**-11.380**	**1.223**	**-9.31**	**0.00**	**-13.841**	**-8.919**
Dummy Weekend	**36.547**	**4.252**	**8.59**	**0.00**	**27.988**	**45.106**

c) For forecasting purposes, would you prefer the first model (with the results in Table 3), or the second model (with the results in Table 4)? Why?

Below, the three residual (error) plots for the revised model are given.







d) Do any of the residual plots highlight any problems with the regression analysis? If yes, interpret them and discuss how these could be resolved.

e) Provide a forecast for the total withdrawals on a weekday from an ATM located 5 miles from the nearest other ATM, in a neighbourhood where the average checking balance of the customers living in that neighbourhood is 1,000, the median value of the homes is 150 and the median family income is 100.