**SCM 651 Fall 2017 Data Analysis Exam Version 6**

**Total Points = 50**

**Due Date: Midnight, Tuesday, December 12, 2017**

**Pan Chen**

**Questions**

1(5 points) Please use the file Q1Q2V6.csv to answer this question. This data set provides:

* Sales (move) for three brands light beer: Amstel, Coors and Miller, at a store in a given week.
* The store, week, and Storeweek = 1000\*store + week
* Season (Fall, Winter, Spring, or Summer)
* Brand
* Feat (1 if product is on sale, 0 if not)
* Log of move
* Price
* Log of price
* Demographic variables at the store location: AGE9, AGE60, EDUC, ETHNIC, INCOME, NOCAR, SINGLE, POVERTY, RETIRED, UNEMP, NWHITE

Please answer the following questions using pivot tables. You may submit the answers either as an Excel worksheet, or in Word.

For each season (Fall, Winter, Spring, and Summer), what percentage of time each brand was on sale (Feat = 1)?

Fall:

|  |  |  |  |
| --- | --- | --- | --- |
| Season | Fall |  |  |
|  |  |  |  |
| **Count of Feat** | **Column Labels** |  |  |
| **Row Labels** | **0** | **1** | **Grand Total** |
| AMSTEL | 72.82% | 27.18% | 100.00% |
| COORS | 67.61% | 32.39% | 100.00% |
| MILLER | 38.31% | 61.69% | 100.00% |
| **Grand Total** | **57.14%** | **42.86%** | **100.00%** |

Winter:

|  |  |  |  |
| --- | --- | --- | --- |
| Season | Winter |  |  |
|  |  |  |  |
| **Count of Feat** | **Column Labels** |  |  |
| **Row Labels** | **0** | **1** | **Grand Total** |
| AMSTEL | 64.02% | 35.98% | 100.00% |
| COORS | 55.49% | 44.51% | 100.00% |
| MILLER | 42.34% | 57.66% | 100.00% |
| **Grand Total** | **53.56%** | **46.44%** | **100.00%** |

Spring:

|  |  |  |  |
| --- | --- | --- | --- |
| Season | Spring |  |  |
|  |  |  |  |
| **Count of Feat** | **Column Labels** |  |  |
| **Row Labels** | **0** | **1** | **Grand Total** |
| AMSTEL | 83.00% | 17.00% | 100.00% |
| COORS | 45.42% | 54.58% | 100.00% |
| MILLER | 41.59% | 58.41% | 100.00% |
| **Grand Total** | **57.58%** | **42.42%** | **100.00%** |

Summer:

|  |  |  |  |
| --- | --- | --- | --- |
| Season | Summer |  |  |
|  |  |  |  |
| **Count of Feat** | **Column Labels** |  |  |
| **Row Labels** | **0** | **1** | **Grand Total** |
| AMSTEL | 68.74% | 31.26% | 100.00% |
| COORS | 47.54% | 52.46% | 100.00% |
| MILLER | 52.31% | 47.69% | 100.00% |
| **Grand Total** | **56.28%** | **43.72%** | **100.00%** |

2(6+6+4+4 = 20 points) Use the package R and the file Q1Q2V6.csv to answer this question. Please prepare answers as a Word file where the relevant output from R is pasted.

2(a) Fit a regression model with dependent variable log of move and the following independent variables:

* Log of price
* Brand
* Season
* Interaction between Brand and log of price
* Feat
* Demographic variables given in the data set.

Based on this output, what is the price elasticity of demand of each brand?

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -9.698973 1.857127 -5.223 1.82e-07 \*\*\*

logprice -0.755361 0.183796 -4.110 4.01e-05 \*\*\*

BRAND[T.COORS] 6.956143 0.729096 9.541 < 2e-16 \*\*\*

BRAND[T.MILLER] 10.337733 0.653176 15.827 < 2e-16 \*\*\*

Season[T.Spring] -0.004622 0.029561 -0.156 0.875762

Season[T.Summer] 0.311733 0.027208 11.458 < 2e-16 \*\*\*

Season[T.Winter] 0.022173 0.028885 0.768 0.442735

Feat 0.237462 0.026642 8.913 < 2e-16 \*\*\*

AGE9 14.635849 1.552845 9.425 < 2e-16 \*\*\*

AGE60 7.403149 1.361445 5.438 5.61e-08 \*\*\*

EDUC 0.481296 0.212067 2.270 0.023270 \*

ETHNIC 0.470918 0.340365 1.384 0.166543

INCOME 0.609242 0.151662 4.017 5.96e-05 \*\*\*

NOCAR 1.744395 0.394366 4.423 9.89e-06 \*\*\*

NWHITE -1.960181 0.326047 -6.012 1.94e-09 \*\*\*

POVERTY -4.351303 1.337255 -3.254 0.001145 \*\*

RETIRED -5.161696 1.386072 -3.724 0.000198 \*\*\*

SINGLE 5.173396 0.677573 7.635 2.61e-14 \*\*\*

UNEMP 9.226259 1.225970 7.526 6.02e-14 \*\*\*

logprice:BRAND[T.COORS] -2.614258 0.316473 -8.261 < 2e-16 \*\*\*

logprice:BRAND[T.MILLER] -3.116827 0.285330 -10.924 < 2e-16 \*\*\*

Price elasticity for COORS: -0.755361-2.614258=-3.369619

Price elasticity for MILLER: -0.755361-3.116827=-3.872188

Price elasticity for AMSTEL: 0.755361

**2(b) To answer each part of 2(b), please start with the model and the results from 2(a).**

**2(b)(i) From the regression output in 2(a), identify the demographic variables that are not significant at a 90% level of confidence. At a 99% level of confidence, test the null hypothesis that the coefficients of these demographic variables (that is, variables you identified) are all zeros. Provide the relevant information from the output.**

The demographic variables that are not significant at a 90% level of confidence:

ETHNIC

**At a 99% level of confidence, test the null hypothesis that the coefficients of these demographic variables (that is, variables you identified) are all zeros.**

H0: the coefficient of ETHNIC is zero

HA: the coefficient of ETHNIC is not zero

Hypothesis:

ETHNIC = 0

Model 1: restricted model

Model 2: logmove ~ logprice + BRAND + Season + BRAND \* logprice + Feat +

AGE9 + AGE60 + EDUC + ETHNIC + INCOME + NOCAR + NWHITE +

POVERTY + RETIRED + SINGLE + UNEMP

Res.Df RSS Df Sum of Sq F Pr(>F)

1 5980 3587.0

2 5979 3585.8 1 1.1481 1.9143 0.1665

P-value=0.1665>0.01, failed to reject null hypothesis.

**2(b)(ii) Starting with the model in 2(a) (that is, using that model as the full model), test the null hypothesis that the price elasticity of demand is same for all three brands at a 99% level of confidence.**

Hypothesis:

logprice:BRAND[T.COORS] = 0

logprice:BRAND[T.MILLER] = 0

Model 1: restricted model

Model 2: logmove ~ logprice + BRAND + Season + BRAND \* logprice + Feat +

AGE9 + AGE60 + EDUC + ETHNIC + INCOME + NOCAR + NWHITE +

POVERTY + RETIRED + SINGLE + UNEMP

Res.Df RSS Df Sum of Sq F Pr(>F)

1 5981 3673.4

2 5979 3585.8 2 87.592 73.026 < 2.2e-16 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

P-Value=2.2e-16 <0.01, reject the null hypothesis

**2(b)(iii) Starting with the model in 2(a) (that is, using that model as the full model), test the null hypothesis that the price elasticity of demand is same for Coors and Miller.**

Hypothesis:

logprice:BRAND[T.COORS] - logprice:BRAND[T.MILLER] = 0

Model 1: restricted model

Model 2: logmove ~ logprice + BRAND + Season + BRAND \* logprice + Feat +

AGE9 + AGE60 + EDUC + ETHNIC + INCOME + NOCAR + NWHITE +

POVERTY + RETIRED + SINGLE + UNEMP

Res.Df RSS Df Sum of Sq F Pr(>F)

1 5980 3587.0

2 5979 3585.8 1 1.1986 1.9986 0.1575

P-Value=0.1575>0.01, failed to reject the null hypothesis

**2(c) Fit a logit model with Feat as the dependent variable, and Brand and Season as independent variables. Very briefly describe how the probability that the product is on sale (Feat = 1) depends on Brand and Season.**

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.01917 0.06889 -14.795 < 2e-16 \*\*\*

BRAND[T.COORS] 0.77493 0.07089 10.931 < 2e-16 \*\*\*

BRAND[T.MILLER] 1.19436 0.06456 18.501 < 2e-16 \*\*\*

Season[T.Spring] 0.02563 0.07795 0.329 0.74228

Season[T.Summer] 0.05489 0.07287 0.753 0.45129

Season[T.Winter] 0.20655 0.07720 2.675 0.00746 \*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 8226.3 on 5999 degrees of freedom

Residual deviance: 7856.7 on 5994 degrees of freedom

AIC: 7868.7

Number of Fisher Scoring iterations: 4

Rcmdr> exp(coef(GLM.3)) # Exponentiated coefficients ("odds ratios")

(Intercept) BRAND[T.COORS] BRAND[T.MILLER] Season[T.Spring]

0.3608958 2.1704365 3.3014309 1.0259645

Season[T.Summer] Season[T.Winter]

1.0564287 1.2294308

I=-1.01917+0.77493\*COORS+1.19436\*MILLER+0.02563\*Spring+0.05489\*Summer+0.20655\*Winter

COORS

Spring: I=-1.01917+0.77493+0.02563

Summer: I=-1.01917+0.77493+0.05489

Fall: I=-1.01917+0.77493 (COORS is least likely to be on sale in fall)

Winter: I=-1.01917+0.77493+0.20655 (COORS is most likely to be on sale in winter)

MILLER

Spring: I=-1.01917+1.19436+0.02563

Summer: I=-1.01917+1.19436+0.05489

Fall: I=-1.01917+1.19436 (MILLER is least likely to be on sale in fall)

Winter: I=-1.01917+1.19436+0.20655 (MILLER is most likely to be on sale in winter)

AMSTEL

Spring: I=-1.01917+0.02563

Summer: I=-1.01917+0.05489

Fall: I=-1.01917 (AMSTEL is least likely to be on sale in fall)

Winter: I=-1.01917+0.20655 (AMSTEL is most likely to be on sale in winter)

In summary, all of the three brands are most likely to be on sale in winter, and are least likely to be on sale in fall.

**2(d) Starting with the logit model in 2(c), test the following null hypotheses at a 99% level of confidence:**

**2(d)(i) A brand is equally likely to be on sale (Feat = 1) in Fall, Spring and Summer.**

Hypothesis:

Season[T.Spring] = 0

Season[T.Summer] = 0

Model 1: restricted model

Model 2: Feat ~ BRAND + Season

Res.Df Df Chisq Pr(>Chisq)

1 5996

2 5994 2 0.5704 0.7519

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

P-value=0.7519>0.05, failed toreject the null hypothesis

**2(d)(ii) Season being same, Amstel and Miller are equally likely to be on sale.**

Hypothesis:

BRAND[T.MILLER] = 0

Model 1: restricted model

Model 2: Feat ~ BRAND + Season

Res.Df Df Chisq Pr(>Chisq)

1 5995

2 5994 1 342.27 < 2.2e-16 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

P-value<2.2e-16, reject the null hypothesis

3.(4+6 = 10 points) Please use the file Q3V6 to answer this question. This file includes two worksheets, where each worksheet provides, for a brand of beer:

* Store, Week, and Storeweek = 1000\*store + week
* Season (Fall, Winter, Spring, or Summer)
* Brand
* Sales (move)
* Log of move
* Feat (1 if product is on sale, 0 if not)
* Price
* Log of price

3(a) Using Access and the join variable Storeweek, create a file of **same store sales** that gives, store, week, storeweek, season, and, for a given combination of store and week, the move, log of move, price, log of price, and Feat for both brands. (Only include cases where data for both brands are available.)

3(b) Export the data file created in 3(a), save it as csv, and use R to answer the following questions.

For each of the two brands, fit a regression model where the dependent variable is log of move for that brand, and the independent variables are log prices for both brands, Feat for both brands, and Season.

How does the demand of each brand depend on its own price, the price of the other brand, and whether **the other brand** is on sale (Feat =1)?

**Please paste the relevant parts of the R output in your answers and also attach the data file you created as answer to 3(a).**

**For Budweiser:**

Call:

lm(formula = Bud\_logmove ~ Bud\_logprice + Busch\_logprice + Bud\_Feat +

Busch\_Feat + Season, data = df2)

Residuals:

Min 1Q Median 3Q Max

-2.51939 -0.53259 0.09783 0.62415 2.20551

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.64937 1.07561 8.971 < 2e-16 \*\*\*

Bud\_logprice -4.89511 0.46945 -10.427 < 2e-16 \*\*\*

Busch\_logprice 1.79738 0.36421 4.935 0.00000090903 \*\*\*

Bud\_Feat 0.03517 0.05853 0.601 0.548

Busch\_Feat 0.05342 0.05917 0.903 0.367

Season[T.Spring] 0.06362 0.06850 0.929 0.353

Season[T.Summer] 0.39320 0.06434 6.111 0.00000000132 \*\*\*

Season[T.Winter] -0.07295 0.06694 -1.090 0.276

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

1) Keeping all of the other variables the same, the more expensive Budweiser is, the less units Budweiser will sell.

2) Keeping other variables the same, the more expensive Busch is, the more units Budweiser will sell.

3) Keeping other variables the same, if Busch is on sale, Budweiser will likely to sell more, however whether Busch is on sale or not might not be a very good predictor of the move of Budweiser due to its very high P-value.

**Busch:**

Call:

lm(formula = Busch\_logmove ~ Busch\_logprice + Bud\_logprice +

Busch\_Feat + Bud\_Feat + Season, data = df2)

Residuals:

Min 1Q Median 3Q Max

-1.6815 -0.6123 -0.0305 0.5885 2.6017

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.24717 1.08526 5.756 1.08e-08 \*\*\*

Busch\_logprice -2.88893 0.36748 -7.861 8.11e-15 \*\*\*

Bud\_logprice 0.37682 0.47367 0.796 0.4264

Busch\_Feat 0.04845 0.05970 0.811 0.4173

Bud\_Feat 0.05339 0.05905 0.904 0.3662

Season[T.Spring] 0.19221 0.06911 2.781 0.0055 \*\*

Season[T.Summer] 0.26687 0.06492 4.111 4.20e-05 \*\*\*

Season[T.Winter] -0.04292 0.06754 -0.635 0.5253

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8212 on 1259 degrees of freedom

Multiple R-squared: 0.09212, Adjusted R-squared: 0.08707

F-statistic: 18.25 on 7 and 1259 DF, p-value: < 2.2e-16

1) Keeping other variables the same, the more expensive Busch is, the less units Busch will sell.

2) Keeping other variables the same, the more expensive Budweiser is, the more units Busch will sell, however the logprice of Budweiser might not be a very good predictor of the move of Busch due to its very high P-value.

3) Keeping other variables the same, if Budweiser is on sale, Busch will likely to sell more, however whether Budweiser is on sale or not might not be a very good predictor of the move of Busch due to its very high P-value.

4.(10 points) Please use the file Q4V6 to answer this question. This file has three worksheets obtained from World Bank data:

* Country: A complete list of country names and country codes.
* Internet: Country code, and the average number of residents out of 100 that had access to the internet in 2015. (The worksheet only includes countries for which the data are available.)
* Service: Country code, and the value added by services as % of GDP in 2015. (The worksheet only includes countries for which the data are available.)

Using Access and country code as the join variable, prepare the data files listed below. Submit the worksheets as answers. (Please note that in some cases, it is possible that the answer is a blank table.)

4(a) A data file that gives country name, country code, internet and service for only the cases where internet is **not missing** and service is **not missing**.

4(b) A data file that gives country name, country code, internet and service for only the cases where internet is **not missing** and service is **missing**.

4(c) A data file that gives country name, country code, internet and service for only the cases where internet is **missing** and service is **not missing**.

4(d) A data file that gives country name, country code, internet and service for only the cases where internet is **missing** and service is **missing**.

5.(5 points) The Excel fine Q5V6 has 150 cases drawn from the Dominicks data base and includes seven fields, ID, Store, Week, Brand, Move, Price and Feat. This worksheet includes one or more duplicates of the same rows. Each combination of the three fields STORE, WEEK and BRAND should be unique. Import the worksheet to Access and copy the table to a new table with structure only. Following the method described in the handout on Access, open the copy in design view, click Indexes under Table Tools, and create a primary key that is the combination of Store, Week and Brand. Save the change. Append the original table to the copy to remove duplicates. Your answer to question 5 is the table you get after removing duplicates.