SCM 651 Assignment 2

Pan Chen; Yifan Liu; Siyao Xu

1. **Full Model**

```
> summary(LinearModel.7)
Call:
lm(formula = logmove ~ logprice + BRAND + Season + BRAND * logprice +
   Feat + AGE9 + AGE60 + EDUC + ETHNIC + INCOME + NOCAR + SINGLE +
   POVERTY, data = Dataset)
Residuals:
   Min      1Q  Median      3Q     Max
-4.5807 -0.5451 -0.0097  0.5168  3.7773
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -4.55233    1.24249  -3.664  0.00025 ***
logprice                  -2.86563    0.07170 -39.965  < 2e-16 ***
BRAND[T.MINMAID]           0.31546    0.07512   4.200 2.69e-05 ***
BRAND[T.TROPICANA]         1.71789    0.09256  18.560  < 2e-16 ***
Season[T.Spring]           0.09822    0.02296   4.278 1.90e-05 ***
Season[T.Summer]          -0.05587    0.02374  -2.354  0.01861 *
Season[T.Winter]           0.10012    0.02279   4.394 1.12e-05 ***
Feat                       0.52766    0.01873  28.166  < 2e-16 ***
AGE9                       1.16234    0.99554   1.168  0.24301
AGE60                      3.02475    0.38929   7.770 8.49e-15 ***
EDUC                       1.00126    0.14936   6.704 2.12e-11 ***
ETHNIC                     0.09843    0.10530   0.935  0.34993
INCOME                     0.74235    0.10968   6.768 1.37e-11 ***
NOCAR                      1.33548    0.27994   4.771 1.86e-06 ***
SINGLE                     0.98018    0.50306   1.948  0.05138 .
POVERTY                    1.71746    0.99036   1.734  0.08291 .
logprice:BRAND[T.MINMAID]   -0.03421    0.09721  -0.352  0.72494
logprice:BRAND[T.TROPICANA]  0.59291    0.10391   5.706 1.19e-08 ***
```

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

Residual standard error: 0.8899 on 11982 degrees of freedom

Multiple R-squared: 0.5584, Adjusted R-squared: 0.5577

F-statistic: 891.1 on 17 and 11982 DF, p-value: < 2.2e-16

(a) Price Elasticity of each brand:

FG: -2.86563

Minute Maid: -2.86563-0.03421 = -2.89984

Tropicana: -2.86563+0.59291 = -2.27272

(b)

i. Demographic variables that are not significant at a 90% level of confidence: **AGE9 and ETHNIC.**

<u>First Method:</u>

<u>Restricted Model</u>

```
> rmodel <- lm(logmove ~ logprice + BRAND + Season + BRAND * logprice + Feat
+   + AGE60 + EDUC + INCOME + NOCAR + SINGLE + POVERTY, data=Dataset)
> summary(rmodel)
Call:
lm(formula = logmove ~ logprice + BRAND + Season + BRAND * logprice +
   Feat + AGE60 + EDUC + INCOME + NOCAR + SINGLE + POVERTY,
   data = Dataset)
Residuals:
   Min      1Q  Median      3Q     Max
-4.5729 -0.5484 -0.0116  0.5134  3.7802
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -4.46740    1.18323  -3.776  0.00016 ***
logprice            -2.86257    0.07169 -39.932  < 2e-16 ***
BRAND[T.MINMAID]     0.31621    0.07512   4.210 2.58e-05 ***
BRAND[T.TROPICANA]   1.71791    0.09253  18.567  < 2e-16 ***
```

```
Season[T.Spring]           0.09872    0.02296   4.300 1.72e-05 ***
Season[T.Summer]          -0.05535    0.02374  -2.332  0.01974 *
Season[T.Winter]           0.10083    0.02278   4.425 9.71e-06 ***
Feat                       0.52775    0.01873  28.180  < 2e-16 ***
AGE60                      2.56335    0.16250  15.774  < 2e-16 ***
EDUC                       1.00556    0.14928   6.736 1.70e-11 ***
INCOME                     0.76101    0.10837   7.022 2.30e-12 ***
NOCAR                      1.18116    0.24770   4.768 1.88e-06 ***
SINGLE                     0.68034    0.31651   2.150  0.03161 *
POVERTY                    2.93967    0.66960   4.390 1.14e-05 ***
logprice:BRAND[T.MINMAID]   -0.03522    0.09721  -0.362  0.71714
logprice:BRAND[T.TROPICANA]  0.59211    0.10388   5.700 1.23e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.89 on 11984 degrees of freedom
Multiple R-squared:  0.5583,      Adjusted R-squared:  0.5577
F-statistic:  1010 on 15 and 11984 DF,  p-value: < 2.2e-16
> anova(rmodel, LinearModel.7)
Analysis of Variance Table
Model 1: logmove ~ logprice + BRAND + Season + BRAND * logprice + Feat +
   AGE60 + EDUC + INCOME + NOCAR + SINGLE + POVERTY
Model 2: logmove ~ logprice + BRAND + Season + BRAND * logprice + Feat +
   AGE9 + AGE60 + EDUC + ETHNIC + INCOME + NOCAR + SINGLE +
   POVERTY
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1  11984 9491.6
2  11982 9489.1  2    2.5444 1.6064 0.2007
```

<u>Second Method:</u>

```
> local({
+   .Hypothesis <- matrix(c(0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
+   0,0,0,0,1,0,0,0,0,0,0,0), 2, 18, byrow=TRUE)
```

```
+   .RHS <- c(0,0)
+   linearHypothesis(LinearModel.1, .Hypothesis, rhs=.RHS)
+ })
Linear hypothesis test

Hypothesis:
AGE9 = 0
ETHNIC = 0

Model 1: restricted model
Model 2: logmove ~ logprice + BRAND + Season + BRAND * logprice + Feat +
    AGE9 + AGE60 + EDUC + ETHNIC + INCOME + NOCAR + SINGLE +
    POVERTY

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1  11984 9491.6
2  11982 9489.1  2    2.5444 1.6064 0.2007
```

Because in both method, Pr(>F) is greater than 0.01, we accept the null hypothesis that the coefficients of AGE9 and ETHNIC are all zeros at a 99% level of confidence.

ii.

```
> local({
+   .Hypothesis <- matrix(c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,
+   0,0,0,0,0,0,0,0,0,0,1), 2, 18, byrow=TRUE)
+   .RHS <- c(0,0)
+   linearHypothesis(LinearModel.10, .Hypothesis, rhs=.RHS)
+ })
Linear hypothesis test

Hypothesis:
logprice:BRAND[T.MINMAID] = 0
logprice:BRAND[T.TROPICANA] = 0
```

Model 1: restricted model

Model 2: logmove ~ logprice + BRAND + Season + BRAND * logprice + Feat +
    AGE60 + EDUC + INCOME + NOCAR + SINGLE + POVERTY + AGE9 +
    ETHNIC

```
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1  11984 9524.7
2  11982 9489.1  2    35.673 22.523 1.725e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because P value (1.725e-10) is less than 0.01, reject the null hypothesis that price elasticity of demand is same for all three brands at a 99% level of confidence.

iii.

```
> local({
+   .Hypothesis <- matrix(c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0), 1, 18,
+   byrow=TRUE)
+   .RHS <- c(0)
+   linearHypothesis(LinearModel.10, .Hypothesis, rhs=.RHS)
+ })
Linear hypothesis test

Hypothesis:
logprice:BRAND[T.MINMAID] = 0

Model 1: restricted model
Model 2: logmove ~ logprice + BRAND + Season + BRAND * logprice + Feat +
    AGE60 + EDUC + INCOME + NOCAR + SINGLE + POVERTY + AGE9 +
    ETHNIC
```

```
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1  11983 9489.2
2  11982 9489.1  1  0.098051 0.1238 0.7249
```

Since P value (0.7249) > 0.01, accept the null hypothesis that price elasticity of demand is same for FG and Minute Maid at a 99% level of confidence.

(c)

```
> GLM.11 <- glm(Feat ~ BRAND + Season, family=binomial(logit), data=Dataset)

> summary(GLM.11)

Call:
glm(formula = Feat ~ BRAND + Season, family = binomial(logit),
    data = Dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0255  -0.9359  -0.8668   1.3945   1.5695

Coefficients:
                   Estimate Std. Error z value   Pr(>|z|)
(Intercept)        -0.44539    0.04655  -9.568    < 2e-16 ***
BRAND[T.MINMAID]   -0.23037    0.04714  -4.887 0.00000102 ***
BRAND[T.TROPICANA] -0.12890    0.04673  -2.758   0.005808 **
Season[T.Spring]   -0.17982    0.05423  -3.316   0.000913 ***
Season[T.Summer]   -0.21097    0.05645  -3.737   0.000186 ***
Season[T.Winter]    0.07705    0.05282   1.459   0.144605
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 15484  on 11999  degrees of freedom
Residual deviance: 15419  on 11994  degrees of freedom
AIC: 15431

Number of Fisher Scoring iterations: 4



> exp(coef(GLM.11))  # Exponentiated coefficients ("odds ratios")
      (Intercept)   BRAND[T.MINMAID] BRAND[T.TROPICANA]   Season[T.Spring]
        0.6405752          0.7942411          0.8790605          0.8354217
  Season[T.Summer]   Season[T.Winter]
        0.8098016          1.0800956
```

I = -0.44539 – 0.23037*Minute Maid – 0.12890*Tropicana – 0.17982*Spring - 0.21097*Summer + 0.07705*Winter

Interpretation:

<u>Brand in Fall:</u>

FG Fall: I = -0.44539 ← Highest likely to be on sale in Fall

MM Fall: I = -0.44539 – 0.23037 ← Lowest likely to be on sale in Fall

TRO Fall: I = -0.44539 – 0.12890 ← Medium likely to be on sale in Fall

<u>Brand: Florida Gold in all seasons</u>

FG Spring: I = -0.44539 – 0.17982

FG Summer: I = -0.44539 – 0.21097

FG Fall: I = -0.44539

FG Winter: I = -0.44539 + 0.07705 ← FG is most likely to be on sale in Winter, compared with other three seasons.

<u>Brand: Minute Maid in all seasons</u>

MM Spring: I = -0.44539 – 0.23037 – 0.17982

MM Summer: I = -0.44539 – 0.23037 – 0.21097

MM Fall: I = -0.44539 – 0.23037

MM Winter: I = -0.44539 – 0.23037 + 0.07705 ← Minute Maid is most likely to be on sale in Winter, compared with other three seasons.

Brand: Tropicana in all seasons

TRO Spring: I = -0.44539 – 0.12890 – 0.17982

TRO Summer: I = -0.44539 – 0.12890 – 0.21097

TRO Fall: I = -0.44539 – 0.12890

TRO Winter: I = -0.44539 – 0.12890 + 0.07705 ← Tropicana is most likely to be on sale in Winter, compared with other three seasons.

To sum up, all three brands are most likely to be on sale in winter, compared with other three seasons. For a given season, Florida Gold is most likely to be on sale, compared with the other two brands.

(d)

(i)

```
> local({
+   .Hypothesis <- matrix(c(0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,1), 3, 6,
+   byrow=TRUE)
+   .RHS <- c(0,0,0)
+   linearHypothesis(GLM.11, .Hypothesis, rhs=.RHS, test="Chisq")
+ })
Linear hypothesis test


Hypothesis:
Season[T.Spring] = 0
Season[T.Summer] = 0
Season[T.Winter] = 0
Model 1: restricted model
Model 2: Feat ~ BRAND + Season
  Res.Df Df  Chisq    Pr(>Chisq)
1  11997
2  11994  3 39.524 0.00000001344 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since Pr(>chi) (0.00000001344) is less than 0.01, reject the null hypothesis that a brand is equally likely to be on sale (Feat=1) in all four seasons at a 99% level of confidence.

(ii)

```
> local({
+   .Hypothesis <- matrix(c(0,1,-1,1,0,0), 1, 6, byrow=TRUE)
+   .RHS <- c(0)
+   linearHypothesis(GLM.11, .Hypothesis, rhs=.RHS, test="Chisq")
+ })
Linear hypothesis test

Hypothesis:
BRAND[T.MINMAID] - BRAND[T.TROPICANA]  + Season[T.Spring] = 0

Model 1: restricted model
Model 2: Feat ~ BRAND + Season

  Res.Df Df  Chisq Pr(>Chisq)
1  11995
2  11994  1 15.089  0.0001025 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since Pr(>chi) (0.0001025) is less than 0.01, reject the null hypothesis season being same, Minute maid and Tropicana are equally likely to be sale at a 99% level of confidence.

2.

```
> orangejuice <- lm(logmove ~ BRAND + Feat + logprice, data=Dataset)

> summary(orangejuice)
Call:
```

```
lm(formula = logmove ~ BRAND + Feat + logprice, data = Dataset)
Residuals:
    Min      1Q  Median      3Q     Max
-4.6036 -0.5760 -0.0087  0.5556  4.0641
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.55428    0.03942  115.53   <2e-16 ***
BRAND[T.MINMAID]     0.27105    0.02090   12.97   <2e-16 ***
BRAND[T.TROPICANA]   2.21267    0.02383   92.85   <2e-16 ***
Feat                 0.56347    0.01937   29.10   <2e-16 ***
logprice            -2.52854    0.04635  -54.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.9276 on 11995 degrees of freedom
Multiple R-squared:  0.5196, Adjusted R-squared:  0.5195
F-statistic:  3244 on 4 and 11995 DF,  p-value: < 2.2e-16
```

Prediction Intervals

```
>  predict(orangejuice, interval="prediction",level=0.95,newdata=newdata)
```

	fit	lwr	upr
1	**3.985557**	**2.1670171**	**5.804097**
2	**5.307591**	**3.4889478**	**7.126234**
3	**2.549643**	**0.7311073**	**4.368178**
4	**3.871676**	**2.0530234**	**5.690329**
5	**2.657873**	**0.8393360**	**4.476411**
6	**3.853480**	**2.0348311**	**5.672129**

3.

```
> GLM.13 <- glm(delaynew ~ d1 + d2 + d3 + d4 + d5, family=binomial(logit),
+   data=flightdelay)

> summary(GLM.13)
```

```
Call:
glm(formula = delaynew ~ d1 + d2 + d3 + d4 + d5, family = binomial(logit),
   data = flightdelay)
Deviance Residuals:
   Min       1Q   Median       3Q      Max
-0.9521  -0.7202  -0.5404  -0.5404   1.9981
Coefficients:
           Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4155     0.1328 -10.654  < 2e-16 ***
d1            0.0330     0.2163   0.153 0.878717
d2           -0.4348     0.1272  -3.418 0.000632 ***
d3            0.3397     0.1372   2.475 0.013330 *
d4            0.1985     0.1591   1.247 0.212259
d5            0.5196     0.1366   3.804 0.000142 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
   Null deviance: 2168.5  on 2200  degrees of freedom
Residual deviance: 2123.5  on 2195  degrees of freedom
AIC: 2135.5
Number of Fisher Scoring iterations: 4
> exp(coef(GLM.13))  # Exponentiated coefficients ("odds ratios")
(Intercept)          d1          d2          d3          d4          d5
  0.2428120   1.0335522   0.6473794   1.4044612   1.2195530   1.6813316
```

l = -1.4155 + 0.0330*d1 – 0.4348*d2 + 0.3397*d3 + 0.1985*d4 + 0.5196*d5

Hypothesis tests 95%:

(1) B1=B2=0 Given destination and time of departure, flights from all three origins (BWI, DCA and IAD) are equally likely to be delayed.

```
> local({
+   .Hypothesis <- matrix(c(0,1,0,0,0,0,0,0,1,0,0,0), 2, 6, byrow=TRUE)
+   .RHS <- c(0,0)
```

```
+   linearHypothesis(GLM.13, .Hypothesis, rhs=.RHS, test="Chisq")
+ })
Linear hypothesis test

Hypothesis:
d1 = 0
d2 = 0

Model 1: restricted model
Model 2: delaynew ~ d1 + d2 + d3 + d4 + d5

  Res.Df Df  Chisq Pr(>Chisq)
1   2197
2   2195  2 12.604   0.001833 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since Pr(>chisq) (0.001833) is less than 0.05, reject the null hypothesis.

(2) B3=B4=0 Given origin and time of departure, flights to all three destinations (JFK, LGA and EWR) are equally likely to be delayed.

```
> local({
+   .Hypothesis <- matrix(c(0,0,0,1,0,0,0,0,0,0,1,0), 2, 6, byrow=TRUE)
+   .RHS <- c(0,0)
+   linearHypothesis(GLM.13, .Hypothesis, rhs=.RHS, test="Chisq")
+ })
Linear hypothesis test

Hypothesis:
d3 = 0
d4 = 0

Model 1: restricted model
```

Model 2: delaynew ~ d1 + d2 + d3 + d4 + d5

```
  Res.Df Df Chisq Pr(>Chisq)
1   2197
2   2195  2 6.142    0.04637 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since Pr(>chisq) (0.04637) is less than 0.05, reject the null hypothesis.

(3) B1=B4=0 .

B1=0. Given time of departure and destination, flights from IAD and BWI origins are equally likely to be delayed.

B4=0. Given time of departure and origins, flights to LGA and JFK destinations are equally likely to be delayed.

Combined: Given time of departure, for flights from either IAD or BWI (origin) and to either LGA or JFK (destination), they are equally likely to be delayed

```
> local({
+   .Hypothesis <- matrix(c(0,1,0,0,0,0,0,0,0,0,1,0), 2, 6, byrow=TRUE)
+   .RHS <- c(0,0)
+   linearHypothesis(GLM.13, .Hypothesis, rhs=.RHS, test="Chisq")
+ })
Linear hypothesis test

Hypothesis:
d1 = 0
d4 = 0

Model 1: restricted model
Model 2: delaynew ~ d1 + d2 + d3 + d4 + d5

  Res.Df Df  Chisq Pr(>Chisq)
1   2197
```

2   2195 2 1.5887   **0.4519**

Since Pr(>chisq) (0.4519) is greater than 0.05, accept the null hypothesis.

(4) **B2+B3=0 Given time of departure, flights are equally likely to be delayed for the following two combinations of origin and destination: 1) origin=IAD and destination=LGA; 2) origin=DCA and destination=EWR.**

```
> local({
+   .Hypothesis <- matrix(c(0,0,1,1,0,0), 1, 6, byrow=TRUE)
+   .RHS <- c(0)
+   linearHypothesis(GLM.13, .Hypothesis, rhs=.RHS, test="Chisq")
+ })
```

Linear hypothesis test

Hypothesis:

d2 + d3 = 0

Model 1: restricted model

Model 2: delaynew ~ d1 + d2 + d3 + d4 + d5

  Res.Df Df Chisq **Pr(>Chisq)**

1   2196

2   2195 1 0.1899   **0.663**

Since Pr(>chisq) (0.663) is greater than 0.05, accept the null hypothesis.

(5) **B0+B3+B5=0 A flight has a 50% likelihood to be delayed if the flight departures from IAD to EWR in the evening (6:00pm or later).**

```
> local({
+   .Hypothesis <- matrix(c(1,0,0,1,0,1), 1, 6, byrow=TRUE)
+   .RHS <- c(0)
+   linearHypothesis(GLM.13, .Hypothesis, rhs=.RHS, test="Chisq")
+ })
```

Linear hypothesis test

Hypothesis:

(Intercept) + d3 + d5 = 0

Model 1: restricted model

Model 2: delaynew ~ d1 + d2 + d3 + d4 + d5

	Res.Df	Df	Chisq	**Pr(>Chisq)**	
1	2196				
2	2195	1	12.689	**0.0003678**	***

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

Since Pr(>chisq) (0.0003678) is less than 0.05, reject the null hypothesis.