Unit 2 - Forecasting Elantra Sales

Problem 1

```
elantra <- read.csv("elantra.csv")
train <- subset(elantra, Year <= 2012)
test <- subset(elantra, Year > 2012)
```

Problem 2.1

```
model <- lm(ElantraSales ~ Unemployment + CPI_all + CPI_energy + Queries,
            train)
summary(model)
```

```
##
## Call:
## lm(formula = ElantraSales ~ Unemployment + CPI_all + CPI_energy +
##     Queries, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6785.2 -2101.8  -562.5  2901.7  7021.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   95385.36  170663.81   0.559    0.580
## Unemployment  -3179.90    3610.26  -0.881    0.385
## CPI_all        -297.65     704.84  -0.422    0.676
## CPI_energy       38.51     109.60   0.351    0.728
## Queries          19.03      11.26   1.690    0.101
##
## Residual standard error: 3295 on 31 degrees of freedom
## Multiple R-squared:  0.4282, Adjusted R-squared:  0.3544
## F-statistic: 5.803 on 4 and 31 DF,  p-value: 0.00132
```

0.4282

Problem 2.2

0 variables

Problem 2.3

-3179.90

Problem 2.4

For an increase of 1 in Unemployment, the prediction of Elantra sales decreases by approximately 3000.

Problem 3.1

```
model <- lm(ElantraSales ~ Month + Unemployment + CPI_all + CPI_energy + Queries,
            train)
summary(model)
```

```
##
## Call:
## lm(formula = ElantraSales ~ Month + Unemployment + CPI_all +
##     CPI_energy + Queries, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6416.6 -2068.7  -597.1  2616.3  7183.2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  148330.49  195373.51   0.759   0.4536
## Month           110.69     191.66   0.578   0.5679
## Unemployment  -4137.28    4008.56  -1.032   0.3103
## CPI_all        -517.99     808.26  -0.641   0.5265
## CPI_energy       54.18     114.08   0.475   0.6382
## Queries          21.19      11.98   1.769   0.0871 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3331 on 30 degrees of freedom
## Multiple R-squared:  0.4344, Adjusted R-squared:  0.3402
## F-statistic: 4.609 on 5 and 30 DF,  p-value: 0.003078
```

0.4344

Problem 3.2

The model is not better because the adjusted R-squared has gone down and none of the variables (including the new one) are very significant.
The first option is incorrect because (ordinary) R-Squared always increases (or at least stays the same) when you add new variables. This does not make the model better, and in fact, may hurt the ability of the model to generalize to new, unseen data (overfitting).
The second option is correct: the adjusted R-Squared is the R-Squared but adjusted to take into account the number of variables. If the adjusted R-Squared is lower, then this indicates that our model is not better and in fact may be worse. Furthermore, if none of the variables have become significant, then this also indicates that the model is not better.

Problem 3.3

110.69 * (3 - 1) = 221.38 110.69 * (5 - 1) = 442.76

Problem 3.4

By modeling Month as a factor variable, the effect of each calendar month is not restricted to be linear in the numerical coding of the month.

Problem 4.1

```
train$Month_Factor <- as.factor(train$Month)
test$Month_Factor <- as.factor(test$Month)
model <- lm(ElantraSales ~ Month_Factor + Unemployment + CPI_all + CPI_energy +
              Queries, train)
summary(model)
```

```
##
## Call:
## lm(formula = ElantraSales ~ Month_Factor + Unemployment + CPI_all +
##     CPI_energy + Queries, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3865.1 -1211.7   -77.1  1207.5  3562.2
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    312509.280 144061.867   2.169 0.042288 *
## Month_Factor2    2254.998   1943.249   1.160 0.259540
## Month_Factor3    6696.557   1991.635   3.362 0.003099 **
## Month_Factor4    7556.607   2038.022   3.708 0.001392 **
## Month_Factor5    7420.249   1950.139   3.805 0.001110 **
## Month_Factor6    9215.833   1995.230   4.619 0.000166 ***
## Month_Factor7    9929.464   2238.800   4.435 0.000254 ***
## Month_Factor8    7939.447   2064.629   3.845 0.001010 **
## Month_Factor9    5013.287   2010.745   2.493 0.021542 *
## Month_Factor10   2500.184   2084.057   1.200 0.244286
## Month_Factor11   3238.932   2397.231   1.351 0.191747
## Month_Factor12   5293.911   2228.310   2.376 0.027621 *
## Unemployment    -7739.381   2968.747  -2.607 0.016871 *
## CPI_all         -1343.307    592.919  -2.266 0.034732 *
## CPI_energy        288.631     97.974   2.946 0.007988 **
## Queries            -4.764     12.938  -0.368 0.716598
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2306 on 20 degrees of freedom
## Multiple R-squared:  0.8193, Adjusted R-squared:  0.6837
## F-statistic: 6.044 on 15 and 20 DF,  p-value: 0.0001469
```

0.8193

Problem 4.2

Month, CPI_all, CPI_energy, Unemployment

Problem 5.1

```
cor(train[c("Unemployment","Month","Queries","CPI_energy","CPI_all")])
```

```
##              Unemployment      Month    Queries CPI_energy    CPI_all
## Unemployment    1.0000000 -0.2036029 -0.6411093 -0.8007188 -0.9562123
## Month          -0.2036029  1.0000000  0.0158443  0.1760198  0.2667883
## Queries        -0.6411093  0.0158443  1.0000000  0.8328381  0.7536732
## CPI_energy     -0.8007188  0.1760198  0.8328381  1.0000000  0.9132259
## CPI_all        -0.9562123  0.2667883  0.7536732  0.9132259  1.0000000
```

Unemployment, Queries, CPI_all

Problem 5.2

```
cor(train[c("Unemployment","Month","Queries","CPI_energy","CPI_all")])
```

```
##              Unemployment      Month    Queries CPI_energy    CPI_all
## Unemployment    1.0000000 -0.2036029 -0.6411093 -0.8007188 -0.9562123
## Month          -0.2036029  1.0000000  0.0158443  0.1760198  0.2667883
## Queries        -0.6411093  0.0158443  1.0000000  0.8328381  0.7536732
## CPI_energy     -0.8007188  0.1760198  0.8328381  1.0000000  0.9132259
## CPI_all        -0.9562123  0.2667883  0.7536732  0.9132259  1.0000000
```

Unemployment, CPI_energy, CPI_all

Problem 6.1

```
model <- lm(ElantraSales ~ Month_Factor + Unemployment + CPI_all + CPI_energy,
            train)
summary(model)
```

```
##
## Call:
## lm(formula = ElantraSales ~ Month_Factor + Unemployment + CPI_all +
##     CPI_energy, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3866.0 -1283.3  -107.2  1098.3  3650.1
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    325709.15  136627.85   2.384 0.026644 *
## Month_Factor2    2410.91    1857.10   1.298 0.208292
## Month_Factor3    6880.09    1888.15   3.644 0.001517 **
## Month_Factor4    7697.36    1960.21   3.927 0.000774 ***
```

```
## Month_Factor5    7444.64    1908.48   3.901 0.000823 ***
## Month_Factor6    9223.13    1953.64   4.721 0.000116 ***
## Month_Factor7    9602.72    2012.66   4.771 0.000103 ***
## Month_Factor8    7919.50    2020.99   3.919 0.000789 ***
## Month_Factor9    5074.29    1962.23   2.586 0.017237 *
## Month_Factor10   2724.24    1951.78   1.396 0.177366
## Month_Factor11   3665.08    2055.66   1.783 0.089062 .
## Month_Factor12   5643.19    1974.36   2.858 0.009413 **
## Unemployment    -7971.34    2840.79  -2.806 0.010586 *
## CPI_all         -1377.58     573.39  -2.403 0.025610 *
## CPI_energy        268.03      78.75   3.403 0.002676 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2258 on 21 degrees of freedom
## Multiple R-squared:  0.818,  Adjusted R-squared:  0.6967
## F-statistic: 6.744 on 14 and 21 DF,  p-value: 5.73e-05
```

Queries

Problem 6.2

```
pred <- predict(model, test)
SSE <- sum((pred - test$ElantraSales)^2)
SSE
```

```
## [1] 190757747
```

Problem 6.3

```
mean(train$ElantraSales)
```

```
## [1] 14462.25
```

Problem 6.4

```
SST <- sum((mean(train$ElantraSales) - test$ElantraSales)^2)
1 - SSE/SST
```

```
## [1] 0.7280232
```

Problem 6.5

```
max(abs(pred - test$ElantraSales))
```

```
## [1] 7491.488
```

Problem 6.6

```
test[which.max(abs(pred - test$ElantraSales)), ]
```

```
##    Month Year ElantraSales Unemployment Queries CPI_energy CPI_all Month_Factor
## 14     3 2013        26153          7.5     313    244.598 232.075            3
```

03/2013