Linear Regression Exercises

Ethan Marcano

2/1/2022

Predicting Life Expectancy in the United States

A) We want to explore the data of different factors within the United States.

First, we want to import StateData.csv.

```
StateData <- read_csv("StateData.csv")
```

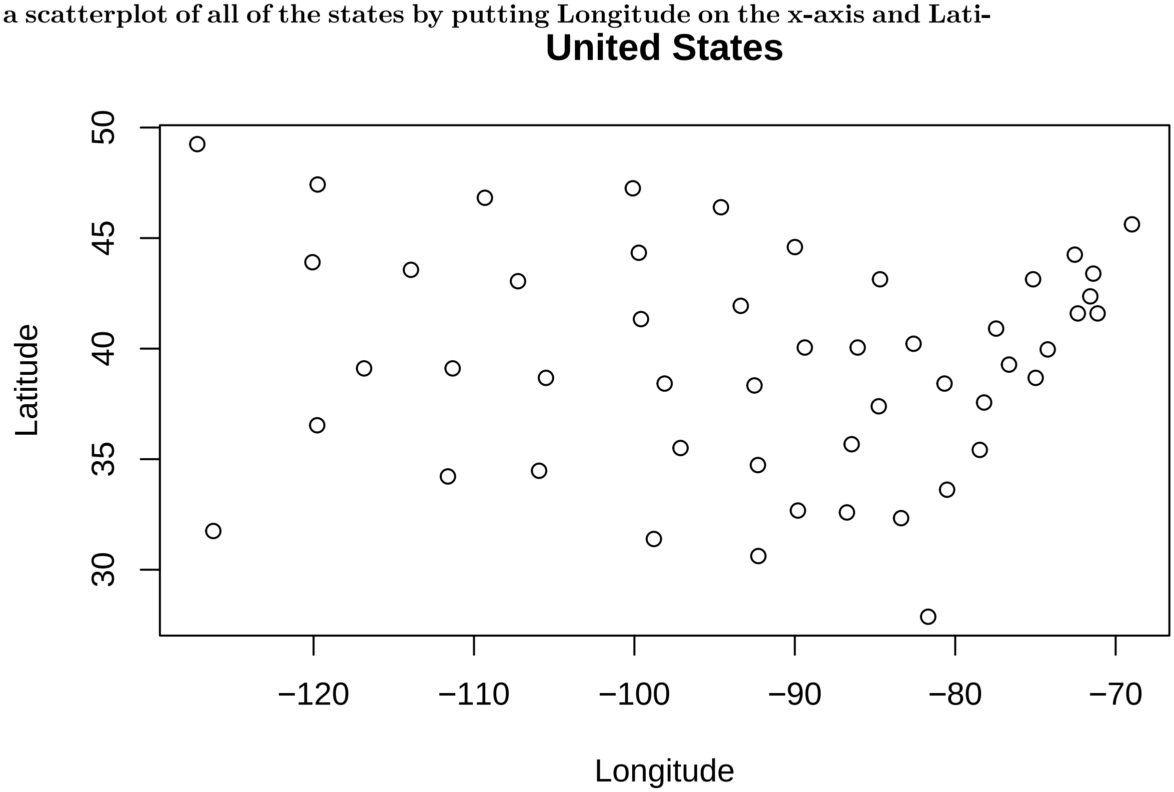
```
## Rows: 50 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Region
## dbl (10): Population, Income, Illiteracy, LifeExp, Murder, HighSchoolGrad, F...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
summary(StateData)
```

```
##    Population        Income       Illiteracy       LifeExp
##  Min.   :  365   Min.   :3098   Min.   :0.500   Min.   :67.96
##  1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12
##  Median : 2838   Median :4519   Median :0.950   Median :70.67
##  Mean   : 4246   Mean   :4436   Mean   :1.170   Mean   :70.88
##  3rd Qu.: 4968   3rd Qu.:4814   3rd Qu.:1.575   3rd Qu.:71.89
##  Max.   :21198   Max.   :6315   Max.   :2.800   Max.   :73.60
##      Murder       HighSchoolGrad      Frost             Area
##  Min.   : 1.400   Min.   :37.80   Min.   :  0.00   Min.   :  1049
##  1st Qu.: 4.350   1st Qu.:48.05   1st Qu.: 66.25   1st Qu.: 36985
##  Median : 6.850   Median :53.25   Median :114.50   Median : 54277
##  Mean   : 7.378   Mean   :53.11   Mean   :104.46   Mean   : 70736
##  3rd Qu.:10.675   3rd Qu.:59.15   3rd Qu.:139.75   3rd Qu.: 81162
##  Max.   :15.100   Max.   :67.30   Max.   :188.00   Max.   :566432
##    Longitude          Latitude        Region
##  Min.   :-127.25   Min.   :27.87   Length:50
##  1st Qu.:-104.16   1st Qu.:35.55   Class :character
##  Median : -89.90   Median :39.62   Mode  :character
##  Mean   : -92.46   Mean   :39.41
##  3rd Qu.: -78.98   3rd Qu.:43.14
##  Max.   : -68.98   Max.   :49.25
```

```
plot(StateData$Longitude, StateData$Latitude, main="United States", xlab = "Longitude", ylab = "Latitud
```

i) First, let's create a scatterplot of all of the states by putting Longitude on the x-axis and Latitude on the y-axis.



This scatterplot was generated via a built-in R function, and used factors in the StateData dataset.

```
regiongrad <-  split(StateData$HighSchoolGrad, StateData$Region)

sapply(regiongrad, mean)
```

ii) We want to see which region of the United States (West, North Central, South, Northeast) has the highest average high graduation rate.

```
## North Central     Northeast         South          West
##      54.51667      53.96667      44.34375      62.00000
```

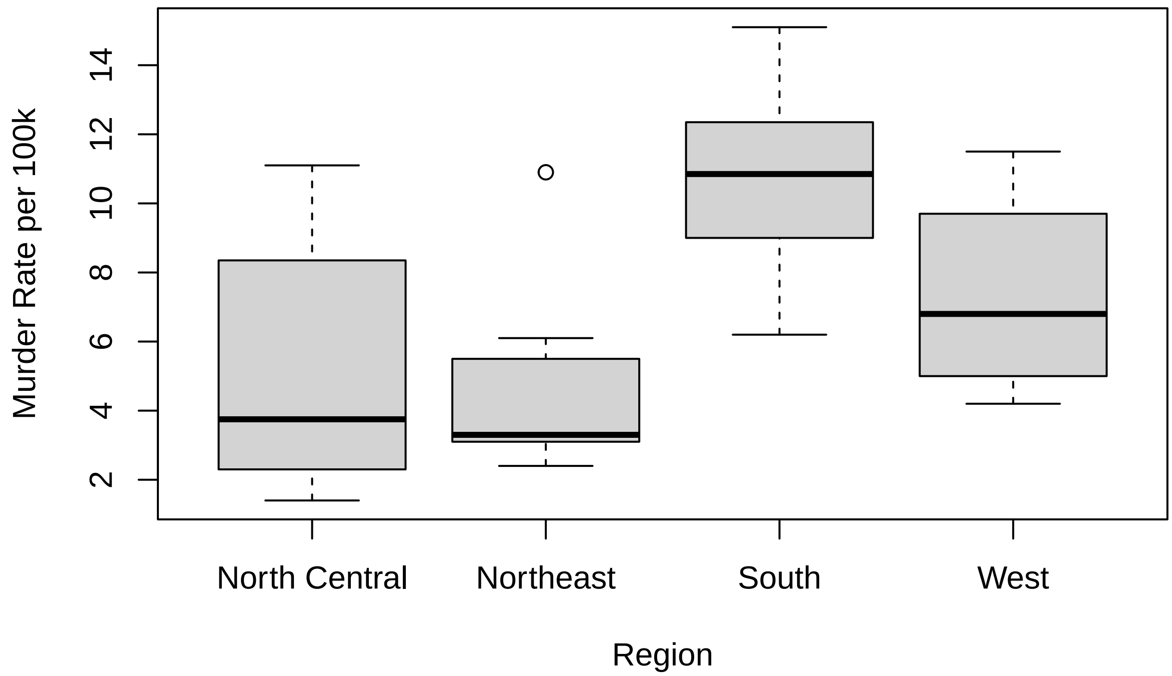
With this in mind, the highest average graduation rate in 1970 is 62% in the West.

iii) Create a box plot of the variable Murder for each Region (four box plots total)

1. Describe the statistical distribution of the murder rate for each region.
2. Which region has the highest median murder rate?

3.The largest range of values?

```
regionmurder <- split(StateData$Murder, StateData$Region)
boxplot(regionmurder, xlab = "Region", ylab = "Murder Rate per 100k")
```



1. Statistical Distribution:
   - North Central has a wide range in Murder Rate, with the median low at 4. IQR is wide and corresponds to range.
   - Northeast has a small range and a low Murder Rate, with the median below North Central's 4. Interestingly, there is an outlier at 11. IQR is narrow.
   - South has a wide range and higher median Murder Rate than the other regions, at ~11. IQR is relatively narrow, but it has a wide range between Max and Min.
   - West has a medium range with a median murder rate of ~7 per 100k. IQR is somewhat narrow, and Max and Min are close to their quartiles.
2. The south has the highest median at ~11 murders per 100k.
3. North Central has the largest range with the min at 1 and the max at 11 (10 units).

B) Build a linear regression model to predict life expectancy (LifeExp) using the following variables as independent variables: Population, Income, Illiteracy, Murder, HighSchoolGrad, Frost, and Area.

```
LifeExpPredict = lm(LifeExp ~ Population + Income + Illiteracy + Murder + HighSchoolGrad + Frost + Area

summary(LifeExpPredict)
```

```
##
## Call:
## lm(formula = LifeExp ~ Population + Income + Illiteracy + Murder +
##     HighSchoolGrad + Frost + Area, data = StateData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48895 -0.51232 -0.02747  0.57002  1.49447
##
```

```
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.094e+01  1.748e+00  40.586  < 2e-16 ***
## Population      5.180e-05  2.919e-05   1.775   0.0832 .
## Income         -2.180e-05  2.444e-04  -0.089   0.9293
## Illiteracy      3.382e-02  3.663e-01   0.092   0.9269
## Murder         -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
## HighSchoolGrad  4.893e-02  2.332e-02   2.098   0.0420 *
## Frost          -5.735e-03  3.143e-03  -1.825   0.0752 .
## Area           -7.383e-08  1.668e-06  -0.044   0.9649
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7448 on 42 degrees of freedom
## Multiple R-squared:  0.7362, Adjusted R-squared:  0.6922
## F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

```
summary(LifeExpPredict)$coefficient
```

```
##                     Estimate   Std. Error     t value     Pr(>|t|)
## (Intercept)     7.094322e+01 1.747975e+00 40.58594017 2.510609e-35
## Population      5.180036e-05 2.918703e-05  1.77477309 8.318351e-02
## Income         -2.180424e-05 2.444256e-04 -0.08920603 9.293422e-01
## Illiteracy      3.382032e-02 3.662799e-01  0.09233464 9.268712e-01
## Murder         -3.011232e-01 4.662073e-02 -6.45899735 8.679582e-08
## HighSchoolGrad  4.892948e-02 2.332328e-02  2.09788176 4.197175e-02
## Frost          -5.735001e-03 3.143230e-03 -1.82455682 7.518682e-02
## Area           -7.383166e-08 1.668163e-06 -0.04425927 9.649075e-01
```

i) What is the regression equation produced by your model? Include all of the coefficients and independent variables they correspond to.

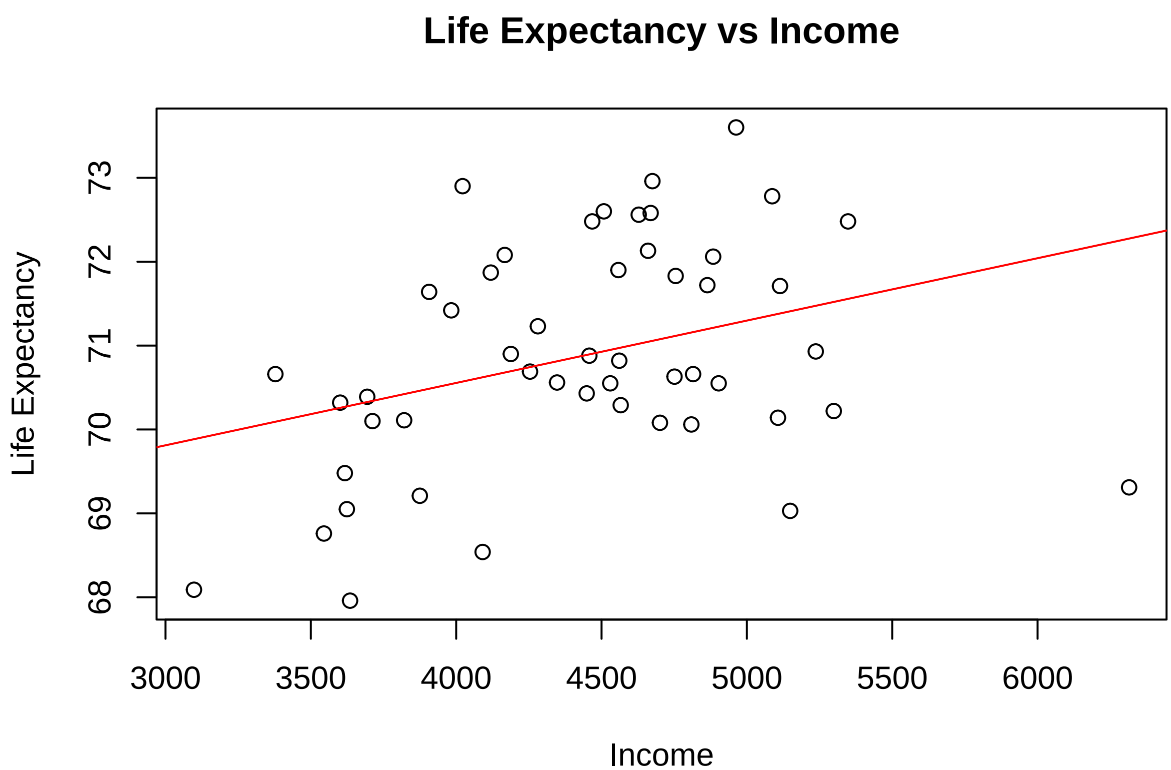
- $y = 70.94 + .00005X_1 + -.000022X_2 + .0338X_3 + -.3011X_4 + .0489X_5 + -.0057X_6 + -.0000007X_7$

ii) What is the interpretation of the coefficient for Income?

- For an increase in income, life expectancy decreases slightly. Income is not a statistically significant determinant for life expectancy.

```
plot(StateData$Income, StateData$LifeExp, main = "Life Expectancy vs Income", xlab = "Income", ylab = "I
abline(lm(StateData$LifeExp ~ StateData$Income), col="red")
```

iii) Create a scatterplot with Income on the x-axis, and LifeExp on the y-axis. Does this relationship agree with the coefficient for Income in your linear regression model? Why or why



not?

- It generally agrees with the coefficient, with the relationship between the two variables being weak.

C) Rebuild the linear regression model, using the set of independent variables you think is the best for predicting LifeExp. This means any subset of the 7 independent variables previously used. Use the significance of the coefficients, the R^2 of the model, and the interpretability of the model when selecting the final set of variables.

```
revismodel = lm(StateData$LifeExp ~ StateData$Murder + StateData$HighSchoolGrad + StateData$Population +
summary(revismodel)
```

```
##
## Call:
## lm(formula = StateData$LifeExp ~ StateData$Murder + StateData$HighSchoolGrad +
##     StateData$Population + StateData$Frost)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47095 -0.53464 -0.03701  0.57621  1.50683
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              7.103e+01  9.529e-01  74.542  < 2e-16 ***
## StateData$Murder        -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
## StateData$HighSchoolGrad  4.658e-02  1.483e-02   3.142  0.00297 **
## StateData$Population     5.014e-05  2.512e-05   1.996  0.05201 .
## StateData$Frost         -5.943e-03  2.421e-03  -2.455  0.01802 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7197 on 45 degrees of freedom
```

```
## Multiple R-squared:  0.736,  Adjusted R-squared:  0.7126
## F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

```
summary(revismodel)$coefficients
```

```
##                                Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)                7.102713e+01 9.528530e-01 74.541541 8.612596e-49
## StateData$Murder          -3.001488e-01 3.660946e-02 -8.198669 1.774520e-10
## StateData$HighSchoolGrad  4.658225e-02 1.482706e-02  3.141704 2.968091e-03
## StateData$Population      5.013998e-05 2.512002e-05  1.996017 5.200514e-02
## StateData$Frost          -5.943290e-03 2.420875e-03 -2.455017 1.801778e-02
```

i) What is your new linear regression equation?

- $y = 71.02 + -.3001X_1 + .0466X_2 + .0005X_3 + -.0059X_4$

ii) Compare and contrast this model to the original model, paying special attention to the R^2 of the model and significance of the coefficients.

- The multiple R^2 of the model is slightly worse than the original. However, all of the coefficients here are statistically significant.

```
predict_vector <- predict(revismodel)

vector_frame <- data.frame(predict_vector)

coordinates <- data.frame(StateData$Latitude, StateData$Longitude)

est_lifexp <- cbind(coordinates, vector_frame)

print(est_lifexp %>% arrange(desc(vector_frame)))
```

iii) Using your simplified model, create a vector of predictions for the dataset StateData.

```
##    StateData.Latitude StateData.Longitude predict_vector
## 47            47.4231           -119.7460       72.68272
## 21            42.3645            -71.5800       72.44105
## 37            43.9078           -120.0680       72.41445
## 15            41.9358            -93.3714       72.39653
## 23            46.3943            -94.6043       72.26560
## 27            41.3356            -99.5898       72.17032
## 11            31.7500           -126.2500       72.09317
## 44            39.1063           -111.3300       72.05753
## 7             41.5928            -72.3573       72.03459
## 41            44.3365            -99.7238       72.01161
## 49            44.5937            -89.9941       72.00996
## 16            38.4204            -98.1156       71.90352
## 34            47.2517           -100.0990       71.87649
## 19            45.6226            -68.9801       71.86095
## 5             36.5341           -119.7730       71.79565
## 39            41.5928            -71.1244       71.76007
## 29            43.3934            -71.3924       71.72636
## 30            39.9637            -74.2336       71.59612
## 12            43.5648           -113.9300       71.49989
## 3             34.2192           -111.6250       71.41416
```

```
## 26             46.8230            -109.3200        71.40025
## 38             40.9069             -77.4500        71.38046
## 36             35.5053             -97.1239        71.15860
## 8              38.6777             -74.9841        71.12647
## 6              38.6777            -105.5130        71.10354
## 35             40.2210             -82.5963        71.08549
## 45             44.2508             -72.5450        71.06135
## 14             40.0495             -86.0808        70.90159
## 50             43.0504            -107.2560        70.87679
## 32             43.1361             -75.1449        70.62937
## 9              27.8744             -81.6850        70.61539
## 20             39.2778             -76.6459        70.51852
## 48             38.4204             -80.6665        70.44983
## 13             40.0495             -89.3776        70.19244
## 46             37.5630             -78.2005        70.14691
## 25             38.3347             -92.5137        70.10610
## 31             34.4764            -105.9420        70.03119
## 43             31.3897             -98.7857        69.97886
## 22             43.1361             -84.6870        69.86893
## 2              49.2500            -127.2500        69.85740
## 4              34.7336             -92.2992        69.57374
## 28             39.1063            -116.8510        69.52482
## 42             35.6767             -86.4560        69.46583
## 33             35.4195             -78.4686        69.28624
## 17             37.3915             -84.7674        69.24418
## 18             30.6181             -92.2724        69.15045
## 40             33.6190             -80.5056        69.06109
## 24             32.6758             -89.8065        69.00535
## 10             32.3329             -83.3736        68.63694
## 1              32.5901             -86.7509        68.48112
```

- Which state does your model predict to have the lowest life expectancy? **Alabama**
- Which state actually has the lowest life expectancy? **Mississippi**
- Which state does your model predict to have the highest life expectancy? **Washington**
- Which state actually has the highest life expectancy? **Hawaii**

Climate Change

Studying the relationship between average global temperature and several other factors.

A) Start by splitting the dataset into a training set (observations =< 2006) and a testing set (observations > 2006). This will build the model and evaluate the predictive ability of the model. Build a linear regression model to predict Temp using all of the other variables as independent variables, using the training set.

```
climate <- read_csv("ClimateChange.csv")
```

```
## Rows: 308 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (11): Year, Month, MEI, CO2, CH4, N2O, CFC.11, CFC.12, TSI, Aerosols, Temp
##
## i Use `spec()` to retrieve the full column specification for this data.
```

```
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
train_data <- subset(climate, Year <= 2006)
test_data <- subset(climate, Year > 2006)

climatemodel = lm(Temp ~ CFC.11 + CFC.12 + CO2 + N2O + CH4 + Aerosols + TSI + MEI, data = train_data)

summary(climatemodel)
```

```
##
## Call:
## lm(formula = Temp ~ CFC.11 + CFC.12 + CO2 + N2O + CH4 + Aerosols +
##     TSI + MEI, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25888 -0.05913 -0.00082  0.05649  0.32433
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.246e+02  1.989e+01  -6.265 1.43e-09 ***
## CFC.11      -6.631e-03  1.626e-03  -4.078 5.96e-05 ***
## CFC.12       3.808e-03  1.014e-03   3.757  0.00021 ***
## CO2          6.457e-03  2.285e-03   2.826  0.00505 **
## N2O         -1.653e-02  8.565e-03  -1.930  0.05467 .
## CH4          1.240e-04  5.158e-04   0.240  0.81015
## Aerosols    -1.538e+00  2.133e-01  -7.210 5.41e-12 ***
## TSI          9.314e-02  1.475e-02   6.313 1.10e-09 ***
## MEI          6.421e-02  6.470e-03   9.923  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09171 on 275 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7436
## F-statistic: 103.6 on 8 and 275 DF,  p-value: < 2.2e-16
```

```r
summary(climatemodel)$coefficients
```

```
##                  Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) -1.245943e+02 1.988680e+01 -6.2651739 1.431046e-09
## CFC.11      -6.630489e-03 1.625983e-03 -4.0778339 5.957288e-05
## CFC.12       3.808103e-03 1.013523e-03  3.7572927 2.097199e-04
## CO2          6.457359e-03 2.284643e-03  2.8264197 5.052521e-03
## N2O         -1.652800e-02 8.564948e-03 -1.9297260 5.466931e-02
## CH4          1.240419e-04 5.158324e-04  0.2404694 8.101456e-01
## Aerosols    -1.537613e+00 2.132523e-01 -7.2103008 5.411273e-12
## TSI          9.314108e-02 1.475488e-02  6.3125609 1.095945e-09
## MEI          6.420531e-02 6.470206e-03  9.9232260 4.898887e-20
```

i) What is the linear regression equation produced by your model?

- $y = -124.6 + 0.006X_1 + .0038X_2 + .0064X_3 + -.0165X_4 + .00012X_5 + -1.537X_6 + .0931X_7 + .0642X_8$

ii) Evaulate the quality of the model. What is the R^2 value? Which independent variables are significant?

- The model does a good job, with most independent variables related significantly to Temp. The multiple R-squared value is 0.7509. The significant independent variables are: CFC.11, CFC.12, C02, Aerosols, TSI, and MEI.

iii) What is the simplest explanation for this contradiction (N20 and CFC-11 associated with high temperatures, but not clear in model)

- The model as a whole reflects recent industrialization, and while there is a negative correlation for the two variables, it does not reflect real world values.

```
cor(train_data)
```

iv) Compute the correlations between all independent variables in the training set. Which independent variables is N20 highly correlated with (>0.7)? Which independent variables is CFC.11 high correlated with (>0.7)?

```
##                  Year         Month           MEI         CO2         CH4
## Year       1.00000000 -0.0279419602 -0.0369876842  0.98274939  0.91565945
## Month     -0.02794196  1.0000000000  0.0008846905 -0.10673246  0.01856866
## MEI       -0.03698768  0.0008846905  1.0000000000 -0.04114717 -0.03341930
## CO2        0.98274939 -0.1067324607 -0.0411471651  1.00000000  0.87727963
## CH4        0.91565945  0.0185686624 -0.0334193014  0.87727963  1.00000000
## N2O        0.99384523  0.0136315303 -0.0508197755  0.97671982  0.89983864
## CFC.11     0.56910643 -0.0131112236  0.0690004387  0.51405975  0.77990402
## CFC.12     0.89701166  0.0006751102  0.0082855443  0.85268963  0.96361625
## TSI        0.17030201 -0.0346061935 -0.1544919227  0.17742893  0.24552844
## Aerosols -0.34524670  0.0148895406  0.3402377871 -0.35615480 -0.26780919
## Temp       0.78679714 -0.0998567411  0.1724707512  0.78852921  0.70325502
##                   N2O      CFC.11        CFC.12         TSI    Aerosols
## Year       0.99384523  0.56910643  0.8970116635  0.17030201 -0.34524670
## Month      0.01363153 -0.01311122  0.0006751102 -0.03460619  0.01488954
## MEI       -0.05081978  0.06900044  0.0082855443 -0.15449192  0.34023779
## CO2        0.97671982  0.51405975  0.8526896272  0.17742893 -0.35615480
## CH4        0.89983864  0.77990402  0.9636162478  0.24552844 -0.26780919
## N2O        1.00000000  0.52247732  0.8679307757  0.19975668 -0.33705457
## CFC.11     0.52247732  1.00000000  0.8689851828  0.27204596 -0.04392120
## CFC.12     0.86793078  0.86898518  1.0000000000  0.25530281 -0.22513124
## TSI        0.19975668  0.27204596  0.2553028138  1.00000000  0.05211651
## Aerosols -0.33705457 -0.04392120 -0.2251312440  0.05211651  1.00000000
## Temp       0.77863893  0.40771029  0.6875575483  0.24338269 -0.38491375
##                  Temp
## Year       0.78679714
## Month     -0.09985674
## MEI        0.17247075
## CO2        0.78852921
## CH4        0.70325502
## N2O        0.77863893
## CFC.11     0.40771029
## CFC.12     0.68755755
## TSI        0.24338269
## Aerosols -0.38491375
## Temp       1.00000000
```

- N20 correlations: Year, C02, CH4, CFC.12
- CFC.11 correlations: CH4, CFC.12

B) Build a new linear regression model, this time only using MEI, TSI, Aerosols, and N20 as the independent variables. Use the training data set.

```r
revised_climate = lm(Temp ~ N2O + MEI + TSI + Aerosols, data = train_data)
summary(revised_climate)
```

```
##
## Call:
## lm(formula = Temp ~ N2O + MEI + TSI + Aerosols, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27916 -0.05975 -0.00595  0.05672  0.34195
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.162e+02  2.022e+01  -5.747 2.37e-08 ***
## N2O          2.532e-02  1.311e-03  19.307  < 2e-16 ***
## MEI          6.419e-02  6.652e-03   9.649  < 2e-16 ***
## TSI          7.949e-02  1.487e-02   5.344 1.89e-07 ***
## Aerosols    -1.702e+00  2.180e-01  -7.806 1.19e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09547 on 279 degrees of freedom
## Multiple R-squared:  0.7261, Adjusted R-squared:  0.7222
## F-statistic: 184.9 on 4 and 279 DF,  p-value: < 2.2e-16
```

```r
summary(revised_climate)$coefficients
```

```
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -116.22685815 20.223028005 -5.747253 2.373584e-08
## N2O            0.02531975  0.001311434 19.306911 2.487588e-53
## MEI            0.06418576  0.006651795  9.649389 3.373572e-19
## TSI            0.07949028  0.014875381  5.343747 1.893732e-07
## Aerosols      -1.70173707  0.217995842 -7.806282 1.193197e-13
```

i) How does the coefficient for N20 in this model compare to the coefficient in the previous model?

- The N20 coefficient in this model is positively correlated with Temp, as opposed to negatively in the previous model.

ii) How does the coefficient of this model compare to the previous one? Consider the R^2 value and the signficance of the independent variables when answering this question.

- The coefficient of the model is similar, but the original model has a slightly higher R^2 value. The independent variables are all highly related to each other.

C) Using the simplified model you created in part (B), calculate predictions for the testing dataset. What is the R^2 on the test set? What does this tell you about the model?

```r
test_climate = lm(Temp ~ N2O + MEI + TSI + Aerosols, data = test_data)

summary(test_climate)
```

```
##
## Call:
## lm(formula = Temp ~ N2O + MEI + TSI + Aerosols, data = test_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21741 -0.02439  0.01930  0.03430  0.17768
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1334.70893  951.60350   1.403   0.1769
## N2O           -0.05695    0.04289  -1.328   0.2000
## MEI            0.06019    0.03111   1.934   0.0681 .
## TSI           -0.96384    0.69064  -1.396   0.1789
## Aerosols      71.32377   30.89366   2.309   0.0324 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0821 on 19 degrees of freedom
## Multiple R-squared:  0.5212, Adjusted R-squared:  0.4204
## F-statistic: 5.171 on 4 and 19 DF,  p-value: 0.005444
```

```
predict(test_climate)
```

```
##         1         2         3         4         5         6         7         8
## 0.5677143 0.5205199 0.4143372 0.4391000 0.4572208 0.3580366 0.3717633 0.3571366
##         9        10        11        12        13        14        15        16
## 0.3516708 0.3101729 0.3716495 0.3252922 0.2914086 0.2330769 0.2693219 0.2600574
##        17        18        19        20        21        22        23        24
## 0.2745682 0.3449913 0.3569679 0.3747531 0.3875059 0.3746566 0.3566392 0.3434387
```

- The Multiple R^2 is 0.5212.
- The model has a low R^2, suggesting that the independent variables do not significantly explain Temperature variance.

Hyundai Elantra

Forecasting Hyundai Elantra sales.

A) Split the dataset into training (2010, 2011, 2012) and testing (2013, 2014). Build a linear regression model to predict monthly Elantra sales (ElantraSales) using Unemployment, Queries, CPI.Energy, and CPI.All. Use the training set to build the model.

```
elantra <- read_csv("Elantra.csv")
```

```
## Rows: 50 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (7): Month, Year, ElantraSales, Unemployment, Queries, CPI.Energy, CPI.All
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
etrain_data <- subset(elantra, Year <= 2012)
etest_data <- subset(elantra, Year > 2012)
```

```r
elantramodel = lm(ElantraSales ~ Unemployment + Queries + CPI.Energy + CPI.All, data = etrain_data)

summary(elantramodel)
```

```
##
## Call:
## lm(formula = ElantraSales ~ Unemployment + Queries + CPI.Energy +
##     CPI.All, data = etrain_data)
##
## 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
## Queries          19.03      11.26   1.690    0.101
## CPI.Energy       38.51     109.60   0.351    0.728
## CPI.All        -297.65     704.84  -0.422    0.676
##
## 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
```

```r
summary(elantramodel)$coefficients
```

```
##                 Estimate   Std. Error    t value  Pr(>|t|)
## (Intercept)  95385.36360 170663.81417  0.5589080 0.5802400
## Unemployment -3179.89957   3610.26225 -0.8807946 0.3852069
## Queries         19.02968     11.25896  1.6901807 0.1010267
## CPI.Energy      38.50604    109.60117  0.3513287 0.7277185
## CPI.All       -297.64563    704.83667 -0.4222902 0.6757278
```

i) What is the linear regression equation produced by your model? Make sure to give the coefficients for each of the independent variables.

- $y = 95385.4 + -3179.9X_1 + 19.02X_2 + 38.51X_3 + -297.6X_4$

ii) What is the $\mathbf{R^2}$ of the model?

- The multiple R^2 is .4282.

iii) Which variables are signficant? What does this tell you about the model?

- None of the variables are statistically significant. This model shows that those independent variables do not significantly explain variance in elantra sales.

B) We want to incorporate seasonality into our model by using the Month variable. Build a new linear regression model, this time using the Month variable as an additional independent variable, using the training data.

```r
monthmodel = lm(ElantraSales ~ Month + Unemployment + Queries + CPI.Energy + CPI.All, data = etrain_dat

summary(monthmodel)
```

```
##
## Call:
## lm(formula = ElantraSales ~ Month + Unemployment + Queries +
##     CPI.Energy + CPI.All, data = etrain_data)
##
## 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
## Queries          21.19      11.98   1.769   0.0871 .
## CPI.Energy       54.18     114.08   0.475   0.6382
## CPI.All        -517.99     808.26  -0.641   0.5265
## ---
## 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
```

```
summary(monthmodel)$coefficients
```

```
##                  Estimate   Std. Error    t value   Pr(>|t|)
## (Intercept)  148330.48770 195373.50659  0.7592150 0.45364852
## Month           110.68527    191.65738  0.5775163 0.56790018
## Unemployment  -4137.28256   4008.55786 -1.0321125 0.31026872
## Queries          21.18552     11.97849  1.7686295 0.08712393
## CPI.Energy       54.18332    114.07565  0.4749770 0.63824315
## CPI.All        -517.99104    808.25901 -0.6408726 0.52647121
```

i) Describe your new model. What is the regression equation? What is the $\mathbf{R^2}$? Which variables are signficant?

- $y = 148330.5 + 110.69X_1 + -4137.3X_2 + 21.19X_3 + 54.18X_4 + -518X_5$
- The multiple R^2 is .4344.
- The queries variable is statistically significant.

ii) We are currently modeling Month as a numeric variable. This causes our model to see Feburary as "larger" than January and so on. Is this the right way to model this variable? What if we made Month a categorical variable instead?

- This is the wrong way to model the variable, as "time" is not increasing over itself.
- Making month a categorical variable would be the correct way to model sales over time.

C) Create a new linear regression model, this time with Month model as a categorical variable. You can manually change the values, or in R, convert Month to a factor variable.

```
etrain_data$factormonth <- as.factor(etrain_data$Month)

emonthmodel = lm(ElantraSales ~ factormonth + Unemployment + Queries + CPI.Energy + CPI.All, data = e
```

```r
summary(emonthmodel)
```

```
##
## Call:
## lm(formula = ElantraSales ~ factormonth + Unemployment + Queries +
##     CPI.Energy + CPI.All, data = etrain_data)
##
## 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 *
## factormonth2    2254.998   1943.249   1.160 0.259540
## factormonth3    6696.557   1991.635   3.362 0.003099 **
## factormonth4    7556.607   2038.022   3.708 0.001392 **
## factormonth5    7420.249   1950.139   3.805 0.001110 **
## factormonth6    9215.833   1995.230   4.619 0.000166 ***
## factormonth7    9929.464   2238.800   4.435 0.000254 ***
## factormonth8    7939.447   2064.629   3.845 0.001010 **
## factormonth9    5013.287   2010.745   2.493 0.021542 *
## factormonth10   2500.184   2084.057   1.200 0.244286
## factormonth11   3238.932   2397.231   1.351 0.191747
## factormonth12   5293.911   2228.310   2.376 0.027621 *
## Unemployment   -7739.381   2968.747  -2.607 0.016871 *
## Queries           -4.764     12.938  -0.368 0.716598
## CPI.Energy       288.631     97.974   2.946 0.007988 **
## CPI.All        -1343.307    592.919  -2.266 0.034732 *
## ---
## 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
```

```r
summary(emonthmodel)$coefficients
```

```
##                    Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)   312509.280182 144061.86707  2.1692713 0.0422884369
## factormonth2    2254.997812   1943.24856  1.1604269 0.2595399946
## factormonth3    6696.556764   1991.63473  3.3623418 0.0030989082
## factormonth4    7556.607380   2038.02192  3.7078146 0.0013916585
## factormonth5    7420.248994   1950.13889  3.8049849 0.0011095428
## factormonth6    9215.832605   1995.22974  4.6189331 0.0001658816
## factormonth7    9929.464426   2238.80038  4.4351718 0.0002544591
## factormonth8    7939.447434   2064.62932  3.8454590 0.0010095185
## factormonth9    5013.286649   2010.74490  2.4932485 0.0215417274
## factormonth10   2500.183753   2084.05722  1.1996714 0.2442864246
## factormonth11   3238.931505   2397.23116  1.3511136 0.1917468055
## factormonth12   5293.910735   2228.30966  2.3757518 0.0276210171
## Unemployment   -7739.381433   2968.74725 -2.6069520 0.0168712350
## Queries           -4.763646     12.93793 -0.3681922 0.7165981623
## CPI.Energy       288.631413     97.97365  2.9460108 0.0079881486
```

```
## CPI.All        -1343.306829    592.91880 -2.2655831 0.0347321946
```

i) Describe your new model. What is the regression equation? What is the R^2? Which variables are signficant?

- $y = 312509 + 2255X_1 + 6697X_2 + 7557X_3 + 7420X_4 + 9216X_5 + 9930X_6 + 7940X_7 + 5013X_8 + 2500X_9 + 3239X_{10} + 5294X_{11} + -7739X_{12} + -4.764X_{13} + 228.6X_{14} + -1343X_{15}$
- The multiple R^2 is 0.8193.
- The significant variables are factormonths3-9 (Spring and Summer), Unemployment, and both CPI stats.

```
etest2013 <- subset(etest_data, Year < 2014)

testsales = lm(ElantraSales ~ Unemployment + Queries + CPI.Energy + CPI.All + as.factor(Month), data = e

testsales2013 = lm(ElantraSales ~ Unemployment + Queries + CPI.Energy + CPI.All + as.factor(Month), data

summary(testsales2013)
```

ii) Using this model, make predictions on the test set. Remember to convert the Month variable to a categorical variable in the test set before making predictions. What is the R^2 of the model on the test set?

```
##
## Call:
## lm(formula = ElantraSales ~ Unemployment + Queries + CPI.Energy +
##     CPI.All + as.factor(Month), data = etest2013)
##
## Residuals:
## ALL 12 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (4 not defined because of singularities)
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        305214.197        NaN     NaN      NaN
## Unemployment        -8203.217        NaN     NaN      NaN
## Queries                63.439        NaN     NaN      NaN
## CPI.Energy              7.237        NaN     NaN      NaN
## CPI.All             -1057.323        NaN     NaN      NaN
## as.factor(Month)2    3111.729        NaN     NaN      NaN
## as.factor(Month)3    6214.775        NaN     NaN      NaN
## as.factor(Month)4    8282.719        NaN     NaN      NaN
## as.factor(Month)5    9099.584        NaN     NaN      NaN
## as.factor(Month)6    2604.812        NaN     NaN      NaN
## as.factor(Month)7    6088.100        NaN     NaN      NaN
## as.factor(Month)8    6398.771        NaN     NaN      NaN
## as.factor(Month)9          NA         NA      NA       NA
## as.factor(Month)10         NA         NA      NA       NA
## as.factor(Month)11         NA         NA      NA       NA
## as.factor(Month)12         NA         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 11 and 0 DF,  p-value: NA
```

```
summary(testsales)
```

```
##
## Call:
## lm(formula = ElantraSales ~ Unemployment + Queries + CPI.Energy +
##     CPI.All + as.factor(Month), data = etest_data)
##
## Residuals:
## ALL 14 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (2 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.860e+06        NaN     NaN      NaN
## Unemployment         2.910e+04        NaN     NaN      NaN
## Queries              1.544e+02        NaN     NaN      NaN
## CPI.Energy          -1.834e+01        NaN     NaN      NaN
## CPI.All              1.129e+04        NaN     NaN      NaN
## as.factor(Month)2  -5.764e+03        NaN     NaN      NaN
## as.factor(Month)3   4.332e+03        NaN     NaN      NaN
## as.factor(Month)4   1.671e+04        NaN     NaN      NaN
## as.factor(Month)5   1.207e+04        NaN     NaN      NaN
## as.factor(Month)6  -9.584e+03        NaN     NaN      NaN
## as.factor(Month)7   7.188e+02        NaN     NaN      NaN
## as.factor(Month)8   2.773e+03        NaN     NaN      NaN
## as.factor(Month)9  -9.892e+03        NaN     NaN      NaN
## as.factor(Month)10 -3.608e+03        NaN     NaN      NaN
## as.factor(Month)11         NA         NA      NA       NA
## as.factor(Month)12         NA         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 13 and 0 DF,  p-value: NA
```

```
predict(testsales)
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## 12174 15326 16219 16393 26153 24445 25090 22163 23958 24700 19691 14876 16751
##    14
## 21692
```

- Ran into an error regarding the model and month factorization and got a multiple R-squared of 1.
- The predictions are still interesting regarding their showing of seasonality, even with the error.

D) From what you saw in the problem, what can you conclude about predicting Hyundai Elantra sales? Do you think these conclusions generalize to predicting sales for other products?

- Many of the independent variables are not useful on their own, but introducing months shows that there is a seasonality associated with Hyundai Elantra sales. You could probably predict sales for many products based on cultural patterns.

E) If you could collect additional independent variables for this problem which variables do you think would be useful for predicting sales?

- I would collect region and inflation rate. It would be interesting to see which regions have higher sales, as well as if inflation has an effect on sales overall (both in car price and loan).