HW4

2025-02-16

Homework 4: Multiple Linear Regression

Setup 1

This question involves the use of multiple linear regression on the Auto data set. The data set is provided in a CSV file. Check the UCI repository for the details of the data set, link: https://archive.ics.uci.edu/ml/datasets/Auto+MPG.

Question set 1

a. Load data in R and see the summary statistics. Mention if you need any pre-processing of the data. Hints: missing data is coded as a '?' symbol. Origin is a categorical variable. The name of a car should not be used to model mpg. Finally, you may delete or impute missing values (see question c).

```
data = read.csv("/Users/atanugiri/OneDrive - University of Texas at El Paso/Class Documents/Data Mining/
head(data)
```

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                        name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

```
summary(data)
```

```
##       mpg          cylinders      displacement     horsepower        weight    
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613  
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225  
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804  
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978  
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615  
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140  
##   acceleration        year           origin          name          
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   Length:392        
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   Class :character  
##  Median :15.50   Median :76.00   Median :1.000   Mode  :character  
##  Mean   :15.54   Mean   :75.98   Mean   :1.577                     
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000                     
##  Max.   :24.80   Max.   :82.00   Max.   :3.000
```

```
data$name = as.factor(data$name)
data$origin = as.factor(data$origin)
summary(data)
```

```
##       mpg          cylinders      displacement     horsepower        weight    
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613  
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225  
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804  
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978  
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615  
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140  
##                                                                              
##   acceleration        year       origin                 name    
##  Min.   : 8.00   Min.   :70.00   1:245   amc matador       :  5  
##  1st Qu.:13.78   1st Qu.:73.00   2: 68   ford pinto        :  5  
##  Median :15.50   Median :76.00   3: 79   toyota corolla    :  5  
##  Mean   :15.54   Mean   :75.98           amc gremlin       :  4  
##  3rd Qu.:17.02   3rd Qu.:79.00           amc hornet        :  4  
##  Max.   :24.80   Max.   :82.00           chevrolet chevette:  4  
##                                          (Other)           :365
```

b. Produce a pair plot that includes all relevant variables in the data set. You may use the 'ggpairs' function from the GGally package.

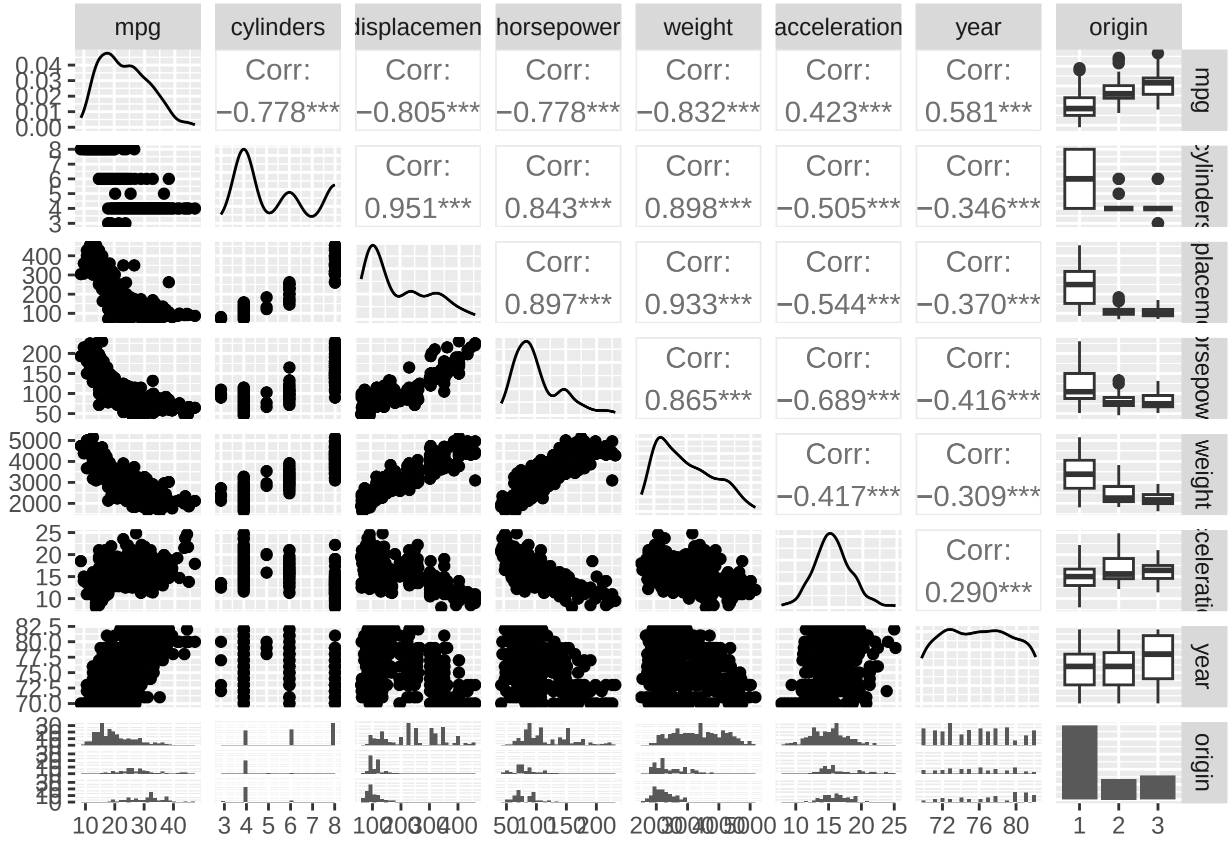
```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from   
##   +.gg   ggplot2
```

```
ggpairs(data[-9])
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



c. Compute the matrix of correlations between the variables. You will need to exclude the name. Also, do not use the origin, as it is a categorical variable. However, the origin should be used for modeling in the next question. Hints: you may use the cor() function from base R. The corrplot package gives a nice visualization of the correlation matrix.

```
(cor(data[-c(8,9)]))
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
##              acceleration       year
## mpg             0.4233285  0.5805410
## cylinders      -0.5046834 -0.3456474
## displacement   -0.5438005 -0.3698552
## horsepower     -0.6891955 -0.4163615
## weight         -0.4168392 -0.3091199
## acceleration    1.0000000  0.2903161
## year            0.2903161  1.0000000
```

d. Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output. For instance:

```
lm_fit = lm(mpg ~ . - name, data = data)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0095 -2.0785 -0.0982  1.9856 13.3608
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.795e+01  4.677e+00  -3.839 0.000145 ***
## cylinders    -4.897e-01  3.212e-01  -1.524 0.128215
## displacement  2.398e-02  7.653e-03   3.133 0.001863 **
## horsepower   -1.818e-02  1.371e-02  -1.326 0.185488
## weight       -6.710e-03  6.551e-04 -10.243  < 2e-16 ***
## acceleration  7.910e-02  9.822e-02   0.805 0.421101
## year          7.770e-01  5.178e-02  15.005  < 2e-16 ***
## origin2       2.630e+00  5.664e-01   4.643 4.72e-06 ***
## origin3       2.853e+00  5.527e-01   5.162 3.93e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.307 on 383 degrees of freedom
## Multiple R-squared:  0.8242, Adjusted R-squared:  0.8205
## F-statistic: 224.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

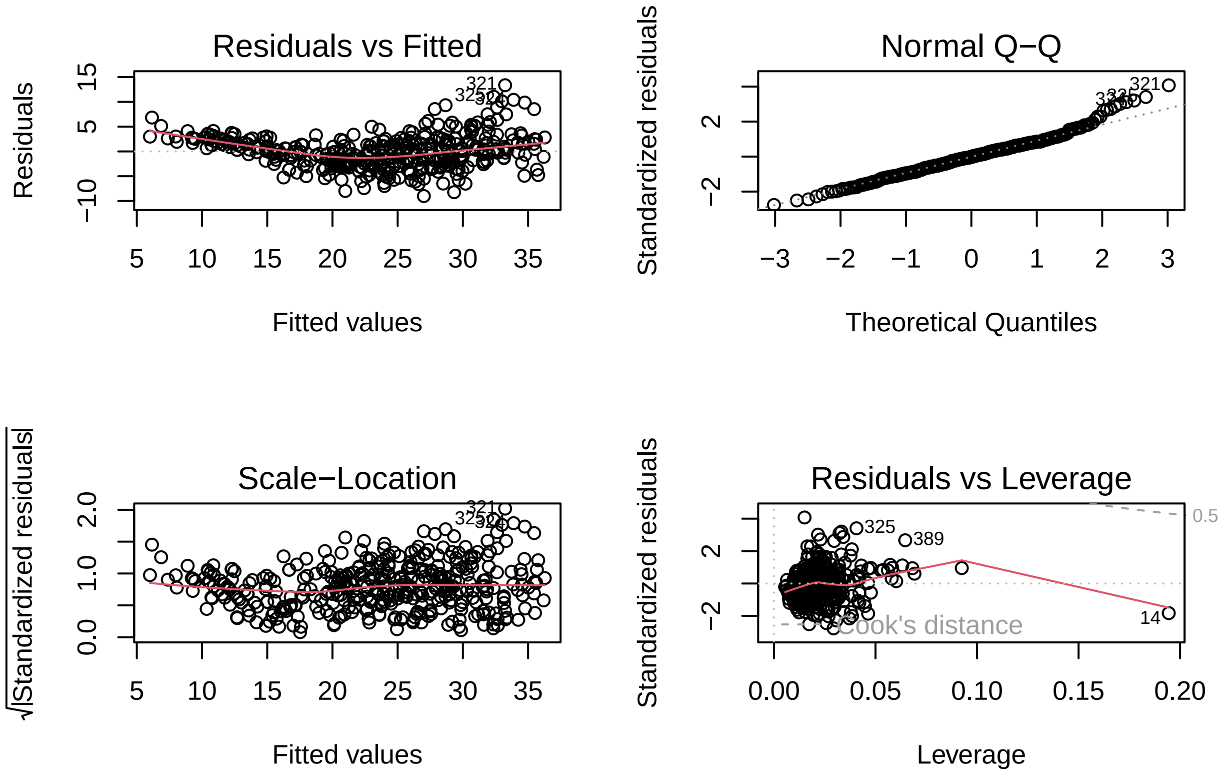
i) Is there a relationship between the predictors and the response? i) Yes, there is a relationship between the predictors and the response as p $< 2.2e^{-16}$

ii) Which predictors appear to have a statistically significant relationship to the response? ii) displacement, weight, year, and origin.

iii) What does the coefficient for the year variable suggest? iii) For each unit of increase in displacement, weight, year, and origin the mpg increases by 0.02, - 0.006, 0.75, and 1.43 units, respectively if all other variables remain fixed.

e. Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

```
par(mfrow = c(2,2))
plot(lm_fit)
```



The residuals are not distributed uniformly on both sides of the line line at y = 0. Higher values in Q-Q plot shows residuals does not follow normal distribution.

f. Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant? Formula $y \sim .^2$ is used to include all interaction terms.

```
data1 = data
data1$name = NULL
lm_fit_2 = lm(mpg ~ .^2, data = data1)
summary(lm_fit_2)
```

```
##
## Call:
## lm(formula = mpg ~ .^2, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6008 -1.2863  0.0813  1.2082 12.0382
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.401e+01  5.147e+01   0.855 0.393048
## cylinders                   3.302e+00  8.187e+00   0.403 0.686976
## displacement               -3.529e-01  1.974e-01  -1.788 0.074638 .
## horsepower                  5.312e-01  3.390e-01   1.567 0.117970
## weight                     -3.259e-03  1.820e-02  -0.179 0.857980
## acceleration              -6.048e+00  2.147e+00  -2.818 0.005109 **
## year                       4.833e-01  5.923e-01   0.816 0.415119
## origin2                   -3.517e+01  1.260e+01  -2.790 0.005547 **
## origin3                   -3.765e+01  1.426e+01  -2.640 0.008661 **
## cylinders:displacement    -6.316e-03  7.106e-03  -0.889 0.374707
## cylinders:horsepower       1.452e-02  2.457e-02   0.591 0.555109
## cylinders:weight           5.703e-04  9.044e-04   0.631 0.528709
## cylinders:acceleration     3.658e-01  1.671e-01   2.189 0.029261 *
## cylinders:year            -1.447e-01  9.652e-02  -1.499 0.134846
## cylinders:origin2         -7.210e-01  1.088e+00  -0.662 0.508100
## cylinders:origin3          1.226e+00  1.007e+00   1.217 0.224379
## displacement:horsepower   -5.407e-05  2.861e-04  -0.189 0.850212
## displacement:weight        2.659e-05  1.455e-05   1.828 0.068435 .
## displacement:acceleration -2.547e-03  3.356e-03  -0.759 0.448415
## displacement:year          4.547e-03  2.446e-03   1.859 0.063842 .
## displacement:origin2      -3.364e-02  4.220e-02  -0.797 0.425902
## displacement:origin3       5.375e-02  4.145e-02   1.297 0.195527
## horsepower:weight         -3.407e-05  2.955e-05  -1.153 0.249743
## horsepower:acceleration   -3.445e-03  3.937e-03  -0.875 0.382122
## horsepower:year           -6.427e-03  3.891e-03  -1.652 0.099487 .
## horsepower:origin2        -4.869e-03  5.061e-02  -0.096 0.923408
## horsepower:origin3         2.289e-02  6.252e-02   0.366 0.714533
## weight:acceleration       -6.851e-05  2.385e-04  -0.287 0.774061
## weight:year               -8.065e-05  2.184e-04  -0.369 0.712223
## weight:origin2             2.277e-03  2.685e-03   0.848 0.397037
## weight:origin3            -4.498e-03  3.481e-03  -1.292 0.197101
## acceleration:year          6.141e-02  2.547e-02   2.412 0.016390 *
## acceleration:origin2       9.234e-01  2.641e-01   3.496 0.000531 ***
## acceleration:origin3       7.159e-01  3.258e-01   2.198 0.028614 *
## year:origin2               2.932e-01  1.444e-01   2.031 0.043005 *
## year:origin3               3.139e-01  1.483e-01   2.116 0.035034 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.628 on 356 degrees of freedom
## Multiple R-squared:  0.8967, Adjusted R-squared:  0.8866
## F-statistic: 88.34 on 35 and 356 DF,  p-value: < 2.2e-16
```

cylinders:acceleration, acceleration:year, acceleration:origin, and year:origin.

g. Try a few different transformations of the variables, such as $log(X)$, $\sqrt{X}$, X^2. Note that X^2 transformation needs $I()$ in the formula: $y \sim I(X^2)$. For $\sqrt(X)$ and X^2, you can simply use $log(X)$ and $\sqrt{X}$, respectively. Comment on your findings. You may also consider transforming the response variable. The goal is to be familiar with some variable transformations, although they may not be the optimum ones.

```
lm_fit_trans = lm(mpg ~ . + log(displacement) + sqrt(horsepower) + year^2, data = data1)
summary(lm_fit_trans)
```

```
##
## Call:
## lm(formula = mpg ~ . + log(displacement) + sqrt(horsepower) +
##     year^2, data = data1)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -9.8186 -1.5724 -0.0212  1.4778 11.7555
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       55.0369549  8.5393929   6.445 3.50e-10 ***
## cylinders          0.1358752  0.2916219   0.466  0.64153
## displacement       0.0266484  0.0128895   2.067  0.03937 *
## horsepower         0.2988372  0.0610741   4.893 1.47e-06 ***
## weight            -0.0033028  0.0006727  -4.910 1.36e-06 ***
## acceleration      -0.2691184  0.0985969  -2.729  0.00664 **
## year               0.7615771  0.0463315  16.438  < 2e-16 ***
## origin2            0.8508379  0.5373529   1.583  0.11416
## origin3            1.2401229  0.5414392   2.290  0.02254 *
## log(displacement) -6.7092007  2.2755252  -2.948  0.00339 **
## sqrt(horsepower)  -7.8076253  1.3624471  -5.731 2.04e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.933 on 381 degrees of freedom
## Multiple R-squared:  0.8624, Adjusted R-squared:  0.8587
## F-statistic: 238.7 on 10 and 381 DF,  p-value: < 2.2e-16
```

R^2 has slightly decreased.

Setup 2

This question should be answered using the Carseats data set. The data set is provided in a CSV file.

Question set 2

a. Fit a multiple regression model to predict Sales using Price, Urban, and ShelveLoc.

```
data = read.csv("/Users/atanugiri/OneDrive - University of Texas at El Paso/Class Documents/Data Mining
summary(data)
```

```
##      Sales          CompPrice       Income        Advertising
##  Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
##  Median : 7.490   Median :125   Median : 69.00   Median : 5.000
##  Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635
##  3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
##  Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000
##    Population        Price        ShelveLoc              Age
##  Min.   : 10.0   Min.   : 24.0   Length:400         Min.   :25.00
##  1st Qu.:139.0   1st Qu.:100.0   Class :character   1st Qu.:39.75
##  Median :272.0   Median :117.0   Mode  :character   Median :54.50
##  Mean   :264.8   Mean   :115.8                      Mean   :53.32
##  3rd Qu.:398.5   3rd Qu.:131.0                      3rd Qu.:66.00
##  Max.   :509.0   Max.   :191.0                      Max.   :80.00
##    Education       Urban                US
##  Min.   :10.0   Length:400         Length:400
##  1st Qu.:12.0   Class :character   Class :character
##  Median :14.0   Mode  :character   Mode  :character
##  Mean   :13.9
##  3rd Qu.:16.0
##  Max.   :18.0
```

```
data$ShelveLoc = as.factor(data$ShelveLoc)
data$Urban = as.factor(data$Urban)
data$US = as.factor(data$US)
summary(data)
```

```
##      Sales          CompPrice       Income        Advertising
##  Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
##  Median : 7.490   Median :125   Median : 69.00   Median : 5.000
##  Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635
##  3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
##  Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000
##    Population        Price          ShelveLoc         Age          Education
##  Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0
##  1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0
##  Median :272.0   Median :117.0   Medium:219   Median :54.50   Median :14.0
##  Mean   :264.8   Mean   :115.8                Mean   :53.32   Mean   :13.9
##  3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0
##  Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.0
##  Urban       US
##  No :118   No :142
##  Yes:282   Yes:258
##
##
##
##
```

```
lm_fit = lm(Sales ~ Price + Urban + ShelveLoc, data = data)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + ShelveLoc, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.614 -1.321 -0.004  1.360  5.001
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     11.80818    0.52172  22.633  < 2e-16 ***
## Price           -0.05699    0.00406 -14.036  < 2e-16 ***
## UrbanYes         0.29375    0.21095   1.392    0.165
## ShelveLocGood    4.92633    0.28642  17.200  < 2e-16 ***
## ShelveLocMedium  1.88631    0.23512   8.023 1.19e-14 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.915 on 395 degrees of freedom
## Multiple R-squared:  0.5448, Adjusted R-squared:  0.5402
## F-statistic: 118.2 on 4 and 395 DF,  p-value: < 2.2e-16
```

b. Provide an interpretation of each coefficient in the model. Be careful – some of the variables in the model are qualitative!

```
contrasts(data$Urban)
```

```
##     Yes
## No    0
## Yes   1
```

```
contrasts(data$ShelveLoc)
```

```
##        Good Medium
## Bad       0      0
## Good      1      0
## Medium    0      1
```

For Urban ‘No’ is baseline. For ShelveLoc ‘Bad’ is baseline.

For each unit of increase in Price, Urban, ShelveLocGood, and ShelveLocMedium the Sales increases by -0.06, 4.93, and 1.89 units, respectively if all other variables remain fixed.

c. Write out the model in equation form, being careful to handle the qualitative variables properly. $Sales = 11.80818 - 0.05699 * Price + 0.29375 * UrbanYes + 4.92633 * ShelveLocGood + 1.88631 * ShelveLocMedium$

d. Add the interaction between Urban and Price in the model. Interpret the fitted coefficients.

```
lm_fit_2 = lm(Sales ~ Price*Urban + ShelveLoc, data = data)
summary(lm_fit_2)
```

```
##
## Call:
## lm(formula = Sales ~ Price * Urban + ShelveLoc, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3565 -1.2763 -0.0569  1.3895  4.8002
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       12.987446   0.904196  14.364  < 2e-16 ***
## Price             -0.067390   0.007678  -8.777  < 2e-16 ***
## UrbanYes          -1.358875   1.057116  -1.285    0.199
## ShelveLocGood      4.934686   0.285909  17.260  < 2e-16 ***
## ShelveLocMedium    1.896242   0.234742   8.078 8.12e-15 ***
## Price:UrbanYes     0.014408   0.009031   1.595    0.111
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 1.911 on 394 degrees of freedom
## Multiple R-squared:  0.5477, Adjusted R-squared:  0.542
## F-statistic: 95.43 on 5 and 394 DF,  p-value: < 2.2e-16
```

The model is written as:
$Sales = \beta_0 + \beta_1 * Price + \beta_2 * UrbanYes + \beta_{31} * ShelveLocGood + \beta_{32} * ShelveLocMedium + \beta_5 * Price * UrbanYes$

If Urban = No
$Sales = \beta_0 + \beta_1 * Price + \beta_{31} * ShelveLocGood + \beta_{32} * ShelveLocMedium$
If Urban = Yes
$Sales = \beta_0 + \beta_1 * Price + \beta_2 + \beta_{31} * ShelveLocGood + \beta_{32} * ShelveLocMedium + \beta_5 * Price$
$= (\beta_0 + \beta_2) + (\beta_1 + \beta_5) * Price + \beta_{31} * ShelveLocGood + \beta_{32} * ShelveLocMedium$

e. For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$? For Price, ShelveLoc we can reject null hypothesis.

f. Now fit a multiple linear model for Sales using all variables provided in the data set (intercept and main effects only). Comment on the model fitting.

```r
lm_fit = lm(Sales ~ ., data = data)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = Sales ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8692 -0.6908  0.0211  0.6636  3.4115
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***
## CompPrice        0.0928153  0.0041477  22.378  < 2e-16 ***
## Income           0.0158028  0.0018451   8.565 2.58e-16 ***
## Advertising      0.1230951  0.0111237  11.066  < 2e-16 ***
## Population       0.0002079  0.0003705   0.561    0.575
## Price           -0.0953579  0.0026711 -35.700  < 2e-16 ***
## ShelveLocGood    4.8501827  0.1531100  31.678  < 2e-16 ***
## ShelveLocMedium  1.9567148  0.1261056  15.516  < 2e-16 ***
## Age             -0.0460452  0.0031817 -14.472  < 2e-16 ***
## Education       -0.0211018  0.0197205  -1.070    0.285
## UrbanYes         0.1228864  0.1129761   1.088    0.277
## USYes           -0.1840928  0.1498423  -1.229    0.220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.019 on 388 degrees of freedom
## Multiple R-squared:  0.8734, Adjusted R-squared:  0.8698
## F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```

Adjusted R-squared has increased.

g. Fit a smaller model that only uses the predictors for which there is evidence of association with the outcome. We will discuss variable selection in a later chapter, but for this question, select variables with significant p-values.

```
lm_fit2 = lm(Sales ~ . - Population - Education - Urban - US, data = data)
summary(lm_fit2)
```

```
## 
## Call:
## lm(formula = Sales ~ . - Population - Education - Urban - US, 
##     data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.7728 -0.6954  0.0282  0.6732  3.3292 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      5.475226   0.505005   10.84   <2e-16 ***
## CompPrice        0.092571   0.004123   22.45   <2e-16 ***
## Income           0.015785   0.001838    8.59   <2e-16 ***
## Advertising      0.115903   0.007724   15.01   <2e-16 ***
## Price           -0.095319   0.002670  -35.70   <2e-16 ***
## ShelveLocGood    4.835675   0.152499   31.71   <2e-16 ***
## ShelveLocMedium  1.951993   0.125375   15.57   <2e-16 ***
## Age             -0.046128   0.003177  -14.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.019 on 392 degrees of freedom
## Multiple R-squared:  0.872,  Adjusted R-squared:  0.8697 
## F-statistic: 381.4 on 7 and 392 DF,  p-value: < 2.2e-16
```

h. How well do the models in (f) and (g) fit the data? You may use anova() function to compare to models. The Adjusted R-squared values very similar. However, the second model has less variables. So, we should prefer 2nd model.

```
anova(lm_fit, lm_fit2)
```

```
## Analysis of Variance Table
## 
## Model 1: Sales ~ CompPrice + Income + Advertising + Population + Price + 
##     ShelveLoc + Age + Education + Urban + US
## Model 2: Sales ~ (CompPrice + Income + Advertising + Population + Price + 
##     ShelveLoc + Age + Education + Urban + US) - Population - 
##     Education - Urban - US
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    388 402.83                           
## 2    392 407.39 -4   -4.5533 1.0964  0.358
```

p = 0.358. So the models are equivalent.