STATS 415 Homework 2 Solutions

January 25, 2018

This exercise relates to the `Carseats` data in the `ISLR` package (or it can be found on the book's website). You may use `help(Carseats)` to learn more about the data set.

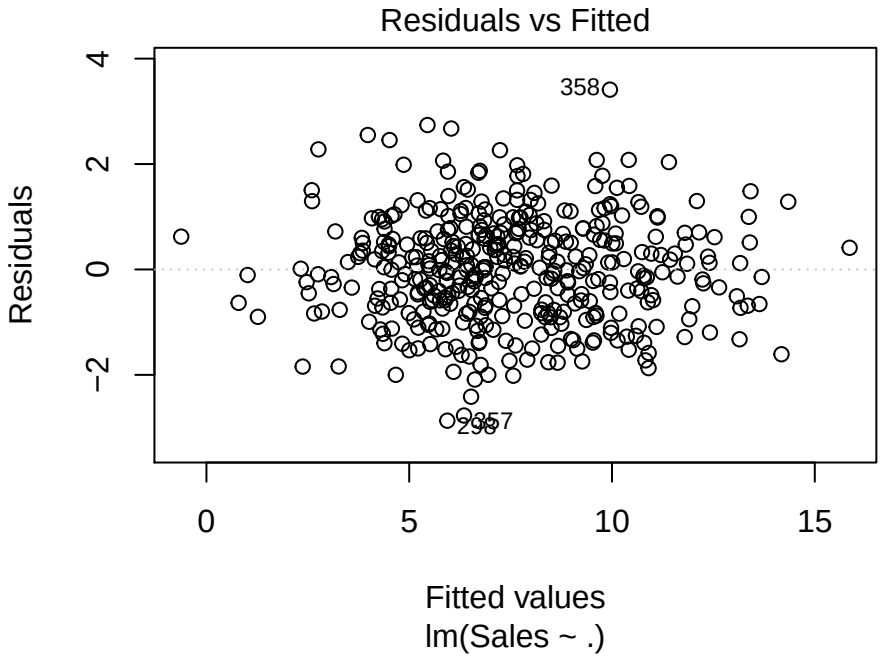
1. [5pts] Fit a multiple regression model to predict `Sales` using all other variables in the model. Report the values of the coefficients, and how well the model fits (using R^2). Include a plot of residuals and comment on any interesting features.

```
mod1 <- lm(Sales ~ ., data = Carseats)
summary(mod1)
```

```
##
## Call:
## lm(formula = Sales ~ ., data = Carseats)
##
## 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
```

The estimated coefficients are tabulated in the above summary. The adjusted R^2 value of 0.8698 indicates that the model is a good fit for the data: it can explain approximately 87% of the variability in `Sales`. A residual plot is below.

```r
plot(mod1, which = 1, add.smooth = F)
```



The plot shows no obvious pattern, and points appear to be scattered evenly around the x-axis. This suggests that the assumption that the errors are mean zero and have constant variance is not violated in this data set. (Note that it cannot tell us about independence of the errors or normality.)

2. [3pts] Which variables correspond to significant p-values? What is the null hypothesis the p-values are testing?

The variables `CompPrice`, `Income`, `Advertising`, `Price`, `ShelveLocGood`, `ShelveLocMedium`, and `Age` have significant p-values at the $\alpha = 0.05$ significance level. These p-values test the null hypothesis that the corresponding regression coefficients are equal to zero, e.g., $H_0 : \beta_{\texttt{CompPrice}} = 0$.

3. [3pts] Drop all the variables that are not significant in the previous model. (Note: this is not the best way to do model selection; we will study better ways later.) Fit the linear model with the remaining variables (but no interaction). It will include one categorical variable, `ShelveLoc`. Compare the fit of the model to the previous one using R^2.

```r
mod3 <- lm(Sales ~ . - Population - Education - Urban - US, data = Carseats)
summary(mod3)
```

```
##
## Call:
## lm(formula = Sales ~ . - Population - Education - Urban - US,
##     data = Carseats)
##
## 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
```

Adjusted R^2 is approximately the same in both models, showing that both are simiarly good fits for the data. Multiple R^2 has decreased slightly, but this is expected: multiple R^2 increases as more variables are added to the model.

4. [3pts] Use the `anova()` function to formally compare the two models and state your conclusion. Comment on the difference between their R^2 in light of your conclusion.

```
anova(mod3, mod1)
```

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

We fail to reject the null hypothesis that $\beta_{\texttt{Population}} = \beta_{\texttt{Education}} = \beta_{\texttt{Urban}} = \beta_{\texttt{US}} = 0$ at the $\alpha = 0.05$ significance level. There is not enough evidence to say that these models are different. In light of this, it makes sense that their adjusted R^2 values are almost identical, since the models differ only in that `mod1` contains variables that are not significantly associated with `Sales`.

5. [5pts] Write out the model from the previous question in equation form and interpret the coefficients. Be careful with the coefficients of the categorical variable.

The model is

$$\texttt{Sales}_i = \beta_0 + \beta_1 \texttt{CompPrice}_i + \beta_2 \texttt{Income}_i + \beta_3 \texttt{Advertising}_i + \beta_4 \texttt{Price}_i + \beta_5 \mathbb{1}\left(\texttt{ShelveLoc}_i = \texttt{Good}\right) + \beta_6 \mathbb{1}\left(\texttt{ShelveLoc}_i = \texttt{Medium}\right) + \beta_7 \texttt{Age}_i + \epsilon_i$$

We interpret the coefficients as follows:

- β_0 is the mean sales when $\texttt{CompPrice} = \texttt{Income} = \texttt{Advertising} = \texttt{Price} = \texttt{Age} = 0$ (i.e., competitors charge \$0, the community income is \$0, the local advertising budget is \$0, the company charges \$0 per carseat and the average age of the local population is 0), and when `ShelveLoc` is `Bad`.
- β_1 is the mean change in sales (in \$1000s) associated with a \$1 increase in the price charged by a competitor, holding all other variables constant.
- β_2 is the mean change in sales (in \$1000s) associated with a \$1000 increase in the community income level, holding all other variables constant.
- β_3 is the mean change in sales (in \$1000s) associated with a \$1000 increase in the local advertising budget, holding all other variables constant.
- β_4 is the mean change in sales (in \$1000s) associated with a \$1 increase in the carseat's price, holding all other variables constant.
- β_5 is the mean difference in sales (in \$1000s) between sites which give the carseat a `Good` shelving location versus a bad shelving location, holding all other variables constant.
- β_6 is the mean difference in sales (in \$1000s) between sites which give the carseat a `Medium` shelving location versus a bad shelving location, holding all other variables constant.
- β_7 is the mean change in sales (in \$1000s) associated with a one-year increase in the average age of the local population.

6. [4pts] Add an interaction term between the categorical variable `ShelveLoc` and the variable `Price`. Refit the model, report the estimated coefficients, and interpret the coefficients of the interaction term. Do the p-values associated with them suggest the interaction term is necessary?

```
mod6 <- lm(Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + Age +
             Price:ShelveLoc, data = Carseats)
summary(mod6)
```

```
##
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
##     ShelveLoc + Age + Price:ShelveLoc, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7984 -0.6896  0.0144  0.6743  3.3391
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.866758   0.696460   8.424 7.08e-16 ***
## CompPrice             0.092592   0.004159  22.262  < 2e-16 ***
## Income                0.015766   0.001849   8.528 3.32e-16 ***
```

```
## Advertising            0.116003   0.007746  14.975  < 2e-16 ***
## Price                 -0.098594   0.004677 -21.082  < 2e-16 ***
## ShelveLocGood          4.185088   0.747377   5.600 4.06e-08 ***
## ShelveLocMedium        1.535031   0.628915   2.441   0.0151 *
## Age                   -0.046494   0.003209 -14.490  < 2e-16 ***
## Price:ShelveLocGood    0.005619   0.006300   0.892   0.3730
## Price:ShelveLocMedium  0.003650   0.005386   0.678   0.4984
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.021 on 390 degrees of freedom
## Multiple R-squared:  0.8723, Adjusted R-squared:  0.8693
## F-statistic: 295.9 on 9 and 390 DF,  p-value: < 2.2e-16
```

We interpret the coefficients of the interaction term as follows:

- $\beta_{\texttt{Price:ShelveLocGood}}$ is the additional change in sales associated with a \$1 price increase at sites with a `Good` shelving location.
- $\beta_{\texttt{Price:ShelveLocMedium}}$ is the additional change in sales associated with a \$1 price increase at sites with a `Medium` shelving location.

The p-values for the individual interaction terms are both insignificant at the $\alpha = 0.05$ level, which suggests the interaction will not be significant overall.

7. [2pts] Use the `anova()` to formally compare model from Q3 to the model from Q6 and state your conclusion.

```
anova(mod3, mod6)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ (CompPrice + Income + Advertising + Population + Price +
##     ShelveLoc + Age + Education + Urban + US) - Population -
##     Education - Urban - US
## Model 2: Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age + Price:ShelveLoc
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    392 407.39
## 2    390 406.52  2   0.86946 0.4171 0.6593
```

Since the p-value of 0.66 is greater than 0.05, we fail to reject the null hypothesis that $\beta_{\texttt{Price:ShelveLocGood}} = \beta_{\texttt{Price:ShelveLocMedium}} = 0$ at the $\alpha = 0.05$ level. The association between `Price` and `Sales` does not change based on shelving location.