SMML Homework 2

Kevin Linares and Jamila Sani

9/17/2024

1. Using the Auto dataset in the ISLR2 package fit a regression model of miles per gallon (mpg) as a function of acceleration and save as an object called "mod1".

```
library(ISLR2)
?Auto
head(Auto)
```

```
##   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
```

```
dim(Auto)
```

```
## [1] 392   9
```

```
names(Auto)
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"
## [6] "acceleration" "year"         "origin"       "name"
```

```
mod1<-lm(mpg~acceleration,Auto)
summary(mod1)
```

```
##
## Call:
## lm(formula = mpg ~ acceleration, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.989  -5.616  -1.199   4.801  23.239
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.8332     2.0485   2.359   0.0188 *
## acceleration   1.1976     0.1298   9.228   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.08 on 390 degrees of freedom
## Multiple R-squared:  0.1792, Adjusted R-squared:  0.1771
## F-statistic: 85.15 on 1 and 390 DF,  p-value: < 2.2e-16
```

```
coef(mod1)
```

```
##  (Intercept) acceleration
##     4.833250     1.197624
```

```
coef(mod1)[1]
```

```
## (Intercept)
##     4.83325
```

```
coef(mod1)[2]
```

```
## acceleration
##     1.197624
```

2. How would you interpret the intercept and slope from #1?

- Intercept => when acceleration=0, the mean expected mpg is 4.83
- Slope => For every second increase in acceleration from 0 to 60 miles per hour (mph), the mean expected miles per gallon (mpg) increases by 1.2.

3. Does the intercept make sense? Show how you would refit the model to get an interpretable intercept and explain whether this is necessary. Interpret the intercept from the refit model.

- The intercept is not meaningful => when acceleration=0, the mean expected mpg is 4.83 (which is not possible without any distance covered) thus the need to refit the model by centering the predictor variable (acceleration).
- Refitted model intercept => the expected mpg for mean accelaton is 23.45

```
Auto$acc_cen <- Auto$acceleration - mean(Auto$acceleration) # centering to refit model
summary(lm(mpg ~ acc_cen, Auto))                            # refitted model
```

```
##
## Call:
## lm(formula = mpg ~ acc_cen, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.989  -5.616  -1.199   4.801  23.239
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.4459     0.3576  65.564   <2e-16 ***
## acc_cen       1.1976     0.1298   9.228   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.08 on 390 degrees of freedom
## Multiple R-squared:  0.1792, Adjusted R-squared:  0.1771
## F-statistic: 85.15 on 1 and 390 DF,  p-value: < 2.2e-16
```

4. What is the null hypothesis being tested via the t statistic for acceleration? Does this hypothesis change if the predictor is centered?

- H_0: Beta_1 = 0 (the slope or true mean of acceleration is equal to zero => no linear relationship) - Wald Test.

- Having centered the predictor, the coefficient for acceleration is 1.1976 with a standard error of 0.1298, thus t-stat = coeff/se = 9.23
- No, the hypothesis does not change if centered.

5. What are some of the reasons you might choose a more flexible modeling approach? What are some of the tradeoffs of choosing a more flexible approach?

- Flexible models are more adaptable to data, particularly with complicated non-linear relationships thus addressing real-life problems simple models fall short.
- Flexible models also minimize residuals or errors, suggestive of decreases in bias and difference between average prediction values of the model and observed data.
- However, tradeoffs in using flexible models include:
  - lack of interpretability or the degree to which we can comprehend modeling decisions and making predictions.
  - they can easily overfit observed data making the model less generalizable to new data. This is called high variance i.e. when small changes in training data result in large changes in predictions or inferences.
- It is vital to understand the variance-bias tradeoffs resulting from fitting flexible vs. simple models to data and finding a balance that is dependent on research goals.