Stat 021 Homework 6

Suzanne Thornton

Due: Friday, Nov. 8, 12:00pm

Instructions: A **pdf** version of your homework must be submitted to Gradescope by **noon** on the due date. The course passcode is **MPKJ4Z**. If you are having trouble getting your *.Rmd* file to compile, you need to get help with this **before** the due date.

You are allowed to hand in **only one** late homework assignment throughout the semester. If you need to hand in this particular assignment late, you must let me know via email by noon on the due date.

You are encouraged to study with your peers to help complete the homework assignments but no copying is allowed. If I see that two or more homework assignments are copied, all students involved will receive a grade of 0 on that assignment and will forfeit (perhaps retroactively) the opportunity to hand in a late homework.

Q0) To help you with some programming tools you will need on your final project, please complete the R *swirl* tutorial on writing functions. You can access this tutorial by typing the following commands in the R console window:

```r
install.packages("swirl")
library("swirl")
swirl()
```

Then, the tutuorial will ask what to call you so enter your name and next type

```r
main()
```

Make the following sequence of selections:

- 1: R Programming: The basics of programming in R
- 1: R Programming
- 9: Functions

Please complete this tutorial up until you get to the part about binary operators (this occurs at about 94% of the way through). Although you don't get points for this problem, it will dramatically help you with your final project and the material in this tutorial is fair game for future homework assignments.

Q1) Read the data uploaded to Moodle called "mileage.csv". This data describes the gasoline mileage performance for 32 automobiles. Use this data to answer the following questions. Before fitting any models make sure the data is being correctly read into R.

```r
mile <- read.csv("mileage.csv")
mile <- mile[-c(1), ]

mile$transmission_type <- as.factor(mile$transmission_type)
mile$mpg <- as.numeric(as.character(mile$mpg))
mile$displacement <- as.numeric(as.character(mile$displacement))
mile$weight <- as.numeric(as.character(mile$weight))
sapply(mile, class)
```

```
##               Car               mpg      displacement            weight
##          "factor"         "numeric"         "numeric"         "numeric"
## transmission_type
##          "factor"
```

```
# mile_standard <- mile %>% mutate_at(vars("mpg", "displacement", "weight"), funs(scale))
mile$transmission_type <- relevel(mile$transmission_type, "A")

# a)
mlr_mile <- lm(formula = mpg ~ displacement + transmission_type, data = mile)
summary(mlr_mile)
```

```
## 
## Call:
## lm(formula = mpg ~ displacement + transmission_type, data = mile)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.9153 -1.8679  0.1302  1.7907  6.7826 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        33.101927   3.068171  10.789 1.15e-11 ***
## displacement       -0.045742   0.008694  -5.262 1.23e-05 ***
## transmission_typeM  0.517276   2.227587   0.232    0.818    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.117 on 29 degrees of freedom
## Multiple R-squared:  0.7724, Adjusted R-squared:  0.7567 
## F-statistic:  49.2 on 2 and 29 DF,  p-value: 4.779e-10
```

```
# b)
mlr2_mile <- lm(formula = mpg ~ displacement + transmission_type +
                  displacement*transmission_type , data = mile)
summary(mlr2_mile)
```

```
## 
## Call:
## lm(formula = mpg ~ displacement + transmission_type + displacement * 
##     transmission_type, data = mile)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.2712 -1.2660  0.1412  1.5336  4.6750 
## 
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     29.436591   2.702869  10.891 1.42e-11 ***
## displacement                    -0.035116   0.007681  -4.572 8.94e-05 ***
## transmission_typeM              13.483040   3.846945   3.505 0.001557 ** 
## displacement:transmission_typeM -0.081659   0.021292  -3.835 0.000653 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.568 on 28 degrees of freedom
## Multiple R-squared:  0.8508, Adjusted R-squared:  0.8348 
## F-statistic: 53.21 on 3 and 28 DF,  p-value: 1.092e-11
```

```
# c)
mlr3_mile <- lm(formula = mpg ~ weight + transmission_type, data = mile)
summary(mlr3_mile)
```

```
## 
## Call:
## lm(formula = mpg ~ weight + transmission_type, data = mile)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.2095 -2.2586  0.3033  2.2403  7.0699 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        34.2056710  3.8447835   8.897 8.73e-10 ***
## weight             -0.0042267  0.0009466  -4.465 0.000112 ***
## transmission_typeM  3.7157618  1.9791784   1.877 0.070552 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.354 on 29 degrees of freedom
## Multiple R-squared:  0.7364, Adjusted R-squared:  0.7182 
## F-statistic:  40.5 on 2 and 29 DF,  p-value: 4.025e-09
```

```
# d)
mlr4_mile <- lm(formula = mpg ~ weight + transmission_type +
                  weight*transmission_type, data = mile)
summary(mlr4_mile)
```

```
## 
## Call:
## lm(formula = mpg ~ weight + transmission_type + weight * transmission_type, 
##     data = mile)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.4534 -1.8453  0.3717  1.4173  4.9229 
## 
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               29.4530696  3.2887177   8.956 1.04e-09 ***
## weight                    -0.0030367  0.0008114  -3.743 0.000834 ***
## transmission_typeM        28.6553504  6.2299643   4.600 8.28e-05 ***
## weight:transmission_typeM -0.0094807  0.0022902  -4.140 0.000289 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.689 on 28 degrees of freedom
## Multiple R-squared:  0.8365, Adjusted R-squared:  0.8189 
## F-statistic: 47.73 on 3 and 28 DF,  p-value: 3.908e-11
```

```
# e)
mlr5_mile <- lm(formula = mpg ~ displacement + transmission_type + weight*transmission_type, data = mil
summary(mlr5_mile)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + transmission_type + weight *
##     transmission_type, data = mile)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4998 -1.1688  0.4337  1.2783  4.7483
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               30.1468382  2.9937445  10.070 1.22e-10 ***
## displacement              -0.0311669  0.0117390  -2.655 0.013139 *
## transmission_typeM        24.9621003  5.8182776   4.290 0.000205 ***
## weight                    -0.0005188  0.0012003  -0.432 0.668992
## transmission_typeM:weight -0.0090907  0.0020820  -4.366 0.000167 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.438 on 27 degrees of freedom
## Multiple R-squared:  0.8703, Adjusted R-squared:  0.8511
## F-statistic:  45.3 on 4 and 27 DF,  p-value: 1.348e-11
```

a) Build a linear regression model relating gasoline mileage, y to engine displacement x_1 and the type of transmission, x_2. (Note that transmission type is a binary categorical variable.) Does the type of transmission significantly affect the mileage performance? Justify your answer. (4 points)

Transmission type does not significantly effect mileage between types A and M. The coefficent suggests that on average manual transmission performs slightly better in relation to fuel efficency (33.101927 for automatic versus 33.6192 for manual type) but because the p-value for $\hat{\beta}_2$ is not significant at alpha level 0.05 we can not reject the null hypothesis.

b) Modify the model developed in part a to include an interaction between engine displacement and the type of transmission. What is the average effect on gasoline mileage when the engine is automatic? What is the average effect on gasoline mileage when the engine is manual? (4 points)

Using dummy variables you can look at the average effects of automatic ($\chi_2 = 0$) versus manual ($\chi_2 = 1$) engines. The average effect of automatic engines is $\hat{\beta}_0 + \hat{\beta}_1\chi_1$ or $29.436591 - 0.035116\chi_1$. The average effect of manual engines is $\hat{\beta}_0 + \hat{\beta}_2 + \chi_1(\hat{\beta}_1 + \hat{\beta}_3)$ or $29.436591 + 13.483040 + \chi_1(-0.035116 - 0.081659) = 42.91963 + \chi_1(-0.116775)$. The difference between the two transmission types on average is $\hat{\beta}_2 + \hat{\beta}_3\chi_1 = 13.483040 - 0.081659\chi_1$.

c) Build a linear regression model relating gasoline mileage, y, to vehicle weight x_3 and the type of transmission x_2 . Does the type of transmission significantly affect the mileage performance? Justify your answer. (4 points)

Again, the difference between M and A transmission type does not significantly effect mileage in this model. Although the coefficent suggests that the manual transmissions performs better in relation to fuel efficency (manual transmission on average increases mpg) the p-value for the test on $\hat{\beta}_2$ suggests that it is not significant at alpha level 0.05. However, the p-value is much closer to being significant compared to part (a) suggesting that there may be a relationship and more data should be collected. Again, although the model is good at accounting for the variance in y it does not appear to be disproportionally because of the transmission type variable

d) Modify the model developed in part a to include an interaction between vehicle weight and the type of transmission. What is the average effect on gasoline mileage when the transmission is automatic? What is the average effect on gasoline mileage when the transmission is manual? (4 points)

Using dummy variables you can look at the average effects of automatic ($\chi_2 = 0$) versus manual ($\chi_2 = 1$) engines. The average effect of automatic engines is $\hat{\beta}_0 + \hat{\beta}_1\chi_1$ or $29.4530696 - 0.0030367\chi_1$. The average effect of manual engines is $\hat{\beta}_0 + \hat{\beta}_2 + \chi_1(\hat{\beta}_1 + \hat{\beta}_3)$ or $29.4530696 + 28.6553504 + \chi_1(-0.0030367 - 0.0094807) = 58.10842 + \chi_1(-0.0125174)$. The average difference between type a and type m is $\hat{\beta}_2 + \hat{\beta}_3\chi_1 = 28.6553504 - 0.0094807\chi_1$.

e) Based off of the results for parts (a)-(d), what terms do you think should be included in the final regression model and why? (4 points)

Based off only the results from parts (a-d), the value with the highest adjusted R^2 value would be best because it explains the most variance in the response variable. Part b is the best model according to the adjusted R^2 value (0.8348).

If you wanted a more experimental method, looking at the largest t-value estimate values from each part (a-d) and putting the variables/interactions with the largest magnitude from each regression model would be another way to choose which variables should be included in the final regression model. For example in my final regression (mpg ~ displacement + transmission_type + weight + weight:transmission_type) includes displacement because it had the greatest magnitude t-value from parts (a-b), weight because part (c), and transmission type with weight:transmission_type because part (d) had the same magnitude if you round to the first decimal. This method provides a slightly better adjusted R^2 value (0.8511), but given the fact that the method essentially relys that you have an intuitive sense of how t-values relate to variable significane it would be better to go with the first option because it is less likely to be problematic. Additionally, the improvment in the adjusted R^2 value does not seem to justify favoring the second method over the first.