Stat 021 Homework 6

Kayonna L Tindle

Due: Saturday, Nov. 9, 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:

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

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

```
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.

```
setwd("~/Stat 21 HW")
library(readr)
mileage <- read_csv("mileage.csv")
```

```
## Parsed with column specification:
## cols(
##   Car = col_character(),
##   mpg = col_double(),
##   displacement = col_double(),
##   weight = col_double(),
##   transmission_type = col_character()
## )
```

```
mileage_standardized <- mileage %>% mutate_at(vars("displacement", "weight"), funs(scale))
```

```
## Warning: `lang()` is deprecated as of rlang 0.2.0.
## Please use `call2()` instead.
## This warning is displayed once per session.
```

```
## Warning: `new_overscope()` is deprecated as of rlang 0.2.0.
## Please use `new_data_mask()` instead.
## This warning is displayed once per session.
```

```
## Warning: `overscope_eval_next()` is deprecated as of rlang 0.2.0.
## Please use `eval_tidy()` with a data mask instead.
## This warning is displayed once per session.
```

```
## Warning: `is_lang()` is deprecated as of rlang 0.2.0.
## Please use `is_call()` instead.
## This warning is displayed once per session.
```

```
## Warning: `mut_node_car()` is deprecated as of rlang 0.2.0.
## This warning is displayed once per session.
```

```
## Warning: The `printer` argument is deprecated as of rlang 0.3.0.
## This warning is displayed once per session.
```

```
#While I presented the code for standardizing mileage, I left the data understandized since our interce
```

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)

Response to a) We can use two metrics, the p-value for significance and the adjusted R-squared value to observe the significance of the variable on the car mileage. When we observe the p-value of our beta_2 coefficient, relating to the transmission type for the manual transmission we will observe that this variable does not have a significant effect on our y variable as the p-value is greater than 0.05. However, we can observe that our r adjusted square value does increase in response to adding the interaction between the two predictor variables (see below).

```
MLR_mileage <- lm(mpg ~ displacement + transmission_type, data = mileage)
summary(MLR_mileage)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + transmission_type, data = mileage)
##
## 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) 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)

part b response) First we must rework our model from $y = beta_0 + beta_1$ times $x_1 + beta_2$ times $x_2 + beta_3$ times x_1 times x_2, to $beta_0 + beta_1$ times x_1 when we make x_2 our dummy variable representing tranmission type where 0 equals automatic and 1 equals manual. From there we can assign values to our coefficient to find the average effect on gasoline mileage when the engine is automatic being represented by the equation y = 29.44 + 29.41 times x_1.

Similarly we can rework our original model to assess the average effect when the engine is manual to get the equation $beta_0 + beta_2 + beta_1$ times x_1. From there we can assign values to our coefficient to find the average effect on gasoline mileage when the engine is manual being represented by the equation y equals (29.4 + 42.92) + 13.48 times x_2.

```
MLR_mileage_interaction <- lm(mpg ~ displacement + transmission_type + displacement*transmission_type, 
summary(MLR_mileage_interaction)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + transmission_type + displacement *
##     transmission_type, data = mileage)
##
## 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) 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)

part c response) Again using two metrics, the p-value for significance and the adjusted R-squared value to observe the significance of the variable on the car mileage... when we observe the p-value of our $beta_2$ coefficient, relating to the transmission type for the manual transmission we will observe that this variable does not have a significant effect on our y variable as the p-value is greater than 0.05. However, again we can observe that our r adjusted square value does increase in response to adding the interaction between the two predictor variables weight and transmission type (although slightly less than the increase in part a's interacting terms).

```
MLR_mileage2 <- lm(mpg ~ weight + transmission_type, data = mileage)
summary(MLR_mileage2)
```

```
##
```

```
## Call:
## lm(formula = mpg ~ weight + transmission_type, data = mileage)
##
## 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) 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)

Response to d) First we must rework our model from $y = beta_0 + beta_1$ times $x_1 + beta_2$ times $x_2 + beta_3$ times x_3, to $beta_0 + beta_1$ times x_1 when we make x_2 our dummy variable representing tranmission type where 0 equals automatic and 1 equals manual. From there we can assign values to our coefficient to find the average effect on gasoline mileage when the engine is automatic being represented by the equation $y = 29.45 + 29.45$ times x_1.
Note: we can use the same equation as in part b because we are not acually using the displacement again in the model in part d, some weight will replace displacement as the $beta_1$ coefficient. . . . so similarly we can rework our original model to assess the average effect when the engine is manual to get the equation $beta_0$ plus $beta_2$ plus $beta_1$ times x_1. . From there we can assign values to our coefficient to find the average effect on gasoline mileage when the engine is manual being represented by the equation y equals (29.45 plus 29.45) plus 28.66 times x_2.

```
MLR_mileage_interaction2 <- lm(mpg ~ weight + transmission_type + weight*transmission_type, data = mile
summary(MLR_mileage_interaction2)
```

```
##
## Call:
## lm(formula = mpg ~ weight + transmission_type + weight * transmission_type,
##     data = mileage)
##
## 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) 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)

Considering the results we obtained from the regression models, it seems that engine displacement had the greatest effect on the adjusted r-square value and showed significance (even outside the context of interaction terms). Weight also seems to be a useful (although potentially less useful) for the same reasons. Futhermore, there does seem to be an interaction between transmission type and engine displacement effect (but less so for tranmission type and weight). However, I definitely believe the fact holds true here that place/industry knowledge would be important to truly grasp which variables should be viewed in the context of others (as an interaction term) and generally what variables should be logically considered in the model. In addition to doing some research to gain subject knowledge, I would certainly focus building a model using engine displacement (and transmission type as an interacting term with engine displacement). Weight could also be useful, but I would not assign transmission type as an interacting term.