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:

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

Then, the tutuorial will ask what to call you so enter your name and next type `{r eval=FALSE} 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 describles 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("~/Google Drive Swat/Swat docs/Stat 21/Homework/")
car_data <- read_csv("mileage.csv", skip=2, col_names=FALSE)
```

```
## Parsed with column specification:
## cols(
##   X1 = col_character(),
##   X2 = col_double(),
##   X3 = col_double(),
##   X4 = col_double(),
##   X5 = col_character()
## )
```

```
colnames(car_data) <- c("car", "mpg", "displacement", "weight", "transmission")
#head(car_data) ## note: transmission is not yet a factor in R
```

```r
car_data2 <- car_data %>% mutate(transmission_cat = fct_infreq(transmission))
#head(car_data2)
```

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)

Solution:

```r
MLR_car1 <- lm(mpg ~ displacement + transmission_cat,
               data=car_data2)
summary(MLR_car1)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + transmission_cat, data = car_data2)
##
## 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_catM  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
```

Note we have fit the following regression model:

$$\hat{y} = \hat{\beta}_0 + (\hat{\beta}_1 \times displacement) + (\hat{\beta}_2 \times type)$$

where $type$ is a dummy variable with $type = 0$ indicating an automatic and $type = 1$ indicating a manual.

Based on the individual t-test for the signifigance of the difference in average mileage between manual and automatic cars in this model, no transmission does not have a significant effect. (p-value of 0.818 which is much larger than any typical α)

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)

Solution:

```r
MLR_car1_interaction <- lm(mpg ~ displacement + transmission_cat +
                           displacement*transmission_cat,
                           data=car_data2)
summary(MLR_car1_interaction)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + transmission_cat + displacement *
##     transmission_cat, data = car_data2)
```

```
##
## 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_catM             13.483040   3.846945   3.505 0.001557 **
## displacement:transmission_catM -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
```

Note we have fit the following regression model:

$$\hat{y} = \hat{\beta}_0 + (\hat{\beta}_1 \times displacement) + (\hat{\beta}_2 \times type) + (\hat{\beta}_3 \times displacement \times type)$$

where *type* is a dummy variable with $type = 0$ indicating an automatic and $type = 1$ indicating a manual.

There is a significant interaction between engine displacement and the type of transmission. For an automatic transmission type, the predicted mileage per gallon is $(29.436591) - (0.035116 \times displacement)$ which indicates that, *on average*, for every increase of one cubic inch in displacement, miles per gallon decreases by 0.035. For an manual transmission type, the predicted mileage per gallon is $(29.436591 + 13.483040) + (-0.035116 - 0.081659)x_1 = 42.91963 - (0.116775 \times displacement)$ which indicates that, *on average*, for every increase of one cubic inch in displacement, miles per gallon decreases by 0.117.

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)

Solution:

```
MLR_car2 <- lm(mpg ~ weight + transmission_cat,
               data=car_data2)
summary(MLR_car2)
```

```
##
## Call:
## lm(formula = mpg ~ weight + transmission_cat, data = car_data2)
##
## 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_catM  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
```

Note we have fit the following regression model:

$$\hat{y} = \hat{\beta}_0 + (\hat{\beta}_1 \times weight) + (\hat{\beta}_2 \times type)$$

where *type* is a dummy variable with $type = 0$ indicating an automatic and $type = 1$ indicating a manual.

Let $\alpha = 0.05$. Based on the individual t-test for the signifigance of the difference in average mileage between manual and automatic cars in this model, no transmission does not have a significant effect since p-value$= 0.071 > \alpha$.

d) Modify the model developed in part c 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)

Solution:

```
MLR_car2_interaction <- lm(mpg ~ weight + transmission_cat +
                             weight*transmission_cat, data=car_data2)
summary(MLR_car2_interaction)
```

```
##
## Call:
## lm(formula = mpg ~ weight + transmission_cat + weight * transmission_cat,
##     data = car_data2)
##
## 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_catM        28.6553504  6.2299643   4.600 8.28e-05 ***
## weight:transmission_catM -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
```

Note we have fit the following regression model:

$$\hat{y} = \hat{\beta}_0 + (\hat{\beta}_1 \times weight) + (\hat{\beta}_2 \times type) + (\hat{\beta}_3 \times weight \times type)$$

where *type* is a dummy variable with $type = 0$ indicating an automatic and $type = 1$ indicating a manual.

There is a significant interaction between engine displacement and the type of transmission. For an automatic transmission type, the predicted mileage per gallon is $29.4530696 - (0.0030367 \times weight)$ which indicates that, *on average*, for every increase of one cubic inch in displacement, miles per gallon decreases by 0.003. For an manual transmission type, the predicted mileage per gallon is $(29.4530696 + 28.6553504) + ((-0.0030367 - 0.0094807) \times weight) = 58.10842 - (0.0125174 \times weight)$ which indicates that, *on average*, for every increase of one cubic inch in displacement, miles per gallon decreases by 0.013.

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)

Solution:

Although the main effects of the categorical variable for transmission type are not significant (at the $\alpha = 0.05$ level) in either of the non-interaction term models, in both cases when we account for an interaction effect between transmission type and weight of a vehicle and transmission type and displacement of a vehicle, we see that all interaction and main effects terms are significant. Note that the adjusted R-squared value for the interaction model with car displacement (0.8348) is higher than the R-squared value for the interaction model with car weight (0.8189). So if were were to choose which model to use out of the four models above, we'd want to go with

$$\hat{y} = \hat{\beta}_0 + (\hat{\beta}_1 \times displacement) + (\hat{\beta}_2 \times type) + (\hat{\beta}_3 \times displacement \times type).$$

We may also consider a model with all three variables, displacement, weight, and transmission type and the interactions of transmission type. For this model we get

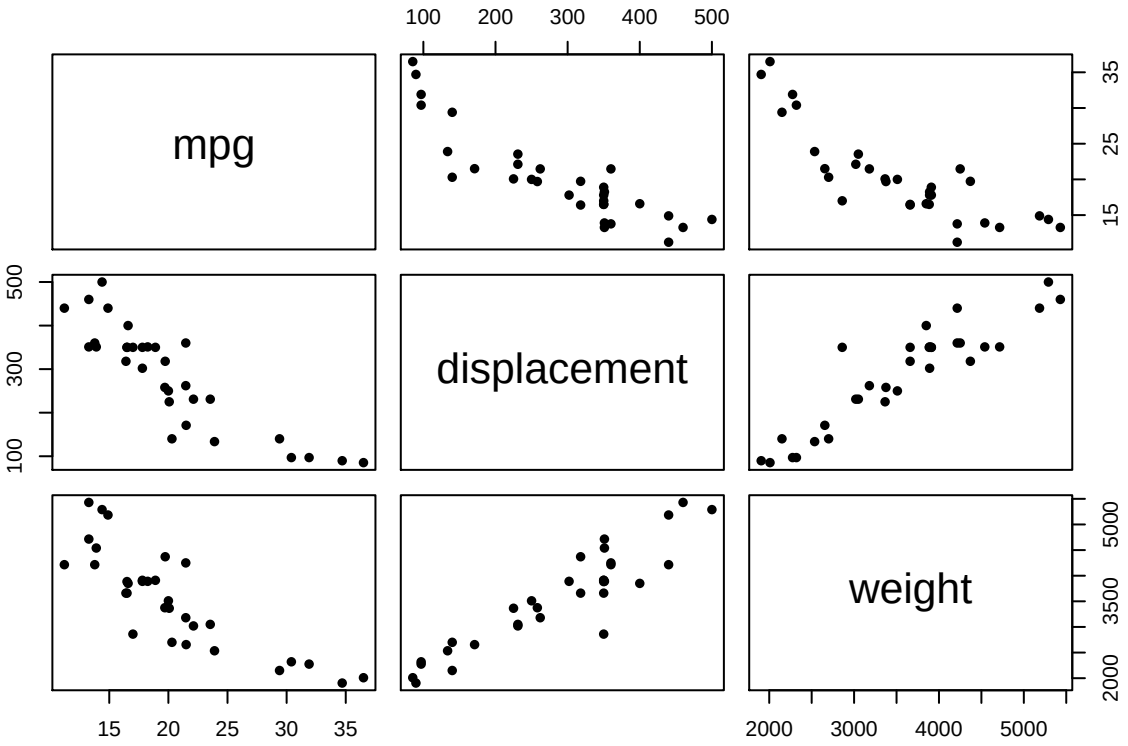
```
cor(car_data2$displacement, car_data2$weight)
```

```
## [1] 0.9210804
```

```
MLR_car3_interaction <- lm(mpg ~ displacement + weight + transmission_cat +
                           displacement*transmission_cat +
                           weight*transmission_cat,
                           data=car_data2)
summary(MLR_car3_interaction)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + weight + transmission_cat +
##     displacement * transmission_cat + weight * transmission_cat,
##     data = car_data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5978 -1.1567  0.3685  1.2488  4.7509
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    30.1364066  3.0504414   9.879 2.73e-10 ***
## displacement                   -0.0306983  0.0123978  -2.476  0.02011 *
## weight                         -0.0005567  0.0012510  -0.445  0.65999
## transmission_catM              24.3705102  7.2248129   3.373  0.00234 **
## displacement:transmission_catM -0.0067241  0.0469619  -0.143  0.88725
## weight:transmission_catM       -0.0084692  0.0048310  -1.753  0.09137 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.484 on 26 degrees of freedom
## Multiple R-squared:  0.8704, Adjusted R-squared:  0.8455
## F-statistic: 34.93 on 5 and 26 DF,  p-value: 9.7e-11
```

```
car_data2 %>% select(-c(car,transmission_cat,transmission)) %>% pairs(pch=16)
```



Based on this model output, the effect of weight on mileage is no longer significant; although, the adjusted R-squared value does increase a bit to 0.8455. Upon further investigation however (plotting the numerical variables and looking for collinearity), we see that displacement and weight are highly correlated. This is also reflected in the results of the individual t-tests in this model summary. So it makes more sense to just include one of these terms and based on the above output, it looks like the model in part (b) preforms the best (with respect to adjusted R-squared, the overall F-test, and the significance of the individual predictors.) Regardless, the main message is we **do** want to choose a model that accounts for the interaction of transmission type!.