Regression Models: Quiz 3

Marcos Gestal

Question 1

Consider the `mtcars` data set. Fit a model with mpg as the outcome that includes number of cylinders as a factor variable and weight as confounder. Give the adjusted estimate for the expected change in mpg comparing 8 cylinders to 4.

```r
data(mtcars)

model <- lm(mpg~factor(cyl) + wt, data=mtcars)

summary(model)
```

Solution

```
##
## Call:
## lm(formula = mpg ~ factor(cyl) + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5890 -1.2357 -0.5159  1.3845  5.7915
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.9908     1.8878  18.006  < 2e-16 ***
## factor(cyl)6  -4.2556     1.3861  -3.070 0.004718 **
## factor(cyl)8  -6.0709     1.6523  -3.674 0.000999 ***
## wt            -3.2056     0.7539  -4.252 0.000213 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.557 on 28 degrees of freedom
## Multiple R-squared:  0.8374, Adjusted R-squared:   0.82
## F-statistic: 48.08 on 3 and 28 DF,  p-value: 3.594e-11
```

```r
# Answer: -6.071
```

Question 2

Consider the `mtcars` data set. Fit a model with mpg as the outcome that includes number of cylinders as a factor variable and weight as a possible confounding variable. Compare the effect of 8 versus 4 cylinders on mpg for the adjusted and unadjusted by weight models. Here, adjusted means including the weight variable as a term in the regression model and unadjusted means the model without weight included. What can be said about the effect comparing 8 and 4 cylinders after looking at models with and without weight included?.

Solution

- Holding weight constant, cylinder appears to have more of an impact on mpg than if weight is disregarded.
- Within a given weight, 8 cylinder vehicles have an expected 12 mpg drop in fuel efficiency.

- Including or excluding weight does not appear to change anything regarding the estimated impact of number of cylinders on mpg.

- ***Holding weight constant, cylinder appears to have less of an impact on mpg than if weight is disregarded.***

```
data(mtcars)

model1 <- lm(mpg ~ factor(cyl) + wt, data=mtcars)
model2 <- lm(mpg ~ factor(cyl), data=mtcars)

summary(model1)
```

```
##
## Call:
## lm(formula = mpg ~ factor(cyl) + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5890 -1.2357 -0.5159  1.3845  5.7915
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.9908     1.8878  18.006  < 2e-16 ***
## factor(cyl)6  -4.2556     1.3861  -3.070 0.004718 **
## factor(cyl)8  -6.0709     1.6523  -3.674 0.000999 ***
## wt            -3.2056     0.7539  -4.252 0.000213 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.557 on 28 degrees of freedom
## Multiple R-squared:  0.8374, Adjusted R-squared:   0.82
## F-statistic: 48.08 on 3 and 28 DF,  p-value: 3.594e-11
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = mpg ~ factor(cyl), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2636 -1.8357  0.0286  1.3893  7.2364
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   26.6636     0.9718  27.437  < 2e-16 ***
```

```
## factor(cyl)6  -6.9208     1.5583  -4.441 0.000119 ***
## factor(cyl)8 -11.5636     1.2986  -8.905 8.57e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.223 on 29 degrees of freedom
## Multiple R-squared:  0.7325, Adjusted R-squared:  0.714
## F-statistic:  39.7 on 2 and 29 DF,  p-value: 4.979e-09
```

Question 3

Consider the `mtcars` data set. Fit a model with mpg as the outcome that considers number of cylinders as a factor variable and weight as confounder. Now fit a second model with mpg as the outcome model that considers the interaction between number of cylinders (as a factor variable) and weight. Give the P-value for the likelihood ratio test comparing the two models and suggest a model using 0.05 as a type I error rate significance benchmark.

Solution

- The P-value is small (less than 0.05). So, according to our criterion, we reject, which suggests that the interaction term is not necessary.

- The P-value is larger than 0.05. So, according to our criterion, we would fail to reject, which suggests that the interaction terms is necessary.

- ***The P-value is larger than 0.05. So, according to our criterion, we would fail to reject, which suggests that the interaction terms may not be necessary.***
- The P-value is small (less than 0.05). So, according to our criterion, we reject, which suggests that the interaction term is necessary

- The P-value is small (less than 0.05). Thus it is surely true that there is no interaction term in the true model.

- The P-value is small (less than 0.05). Thus it is surely true that there is an interaction term in the true model.

```r
data(mtcars)

model1 <- lm(mpg ~ factor(cyl) + wt, data=mtcars)
model2 <- update(model1, mpg ~ factor(cyl) + wt + wt*factor(cyl))

summary(model1)
```

```
##
## Call:
## lm(formula = mpg ~ factor(cyl) + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5890 -1.2357 -0.5159  1.3845  5.7915
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.9908     1.8878  18.006  < 2e-16 ***
## factor(cyl)6  -4.2556     1.3861  -3.070 0.004718 **
## factor(cyl)8  -6.0709     1.6523  -3.674 0.000999 ***
## wt            -3.2056     0.7539  -4.252 0.000213 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.557 on 28 degrees of freedom
## Multiple R-squared:  0.8374, Adjusted R-squared:   0.82
## F-statistic: 48.08 on 3 and 28 DF,  p-value: 3.594e-11
```

```r
summary(model2)
```

```
##
## Call:
## lm(formula = mpg ~ factor(cyl) + wt + factor(cyl):wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1513 -1.3798 -0.6389  1.4938  5.2523
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      39.571      3.194  12.389 2.06e-12 ***
## factor(cyl)6    -11.162      9.355  -1.193 0.243584
## factor(cyl)8    -15.703      4.839  -3.245 0.003223 **
## wt               -5.647      1.359  -4.154 0.000313 ***
## factor(cyl)6:wt   2.867      3.117   0.920 0.366199
## factor(cyl)8:wt   3.455      1.627   2.123 0.043440 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.449 on 26 degrees of freedom
## Multiple R-squared:  0.8616, Adjusted R-squared:  0.8349
## F-statistic: 32.36 on 5 and 26 DF,  p-value: 2.258e-10
```

Question 4

Consider the `mtcars` data set. Fit a model with mpg as the outcome that includes number of cylinders as a factor variable and weight inlcuded in the model as

```
lm(mpg ~ I(wt * 0.5) + factor(cyl), data = mtcars)
```

How is the wt coefficient interpretted?

Solution

- ***The estimated expected change in MPG per one ton increase in weight for a specific number of cylinders (4, 6, 8).***
- The estimated expected change in MPG per one ton increase in weight.

- The estimated expected change in MPG per half ton increase in weight for the average number of cylinders.

- The estimated expected change in MPG per half ton increase in weight for for a specific number of cylinders (4, 6, 8).

- The estimated expected change in MPG per half ton increase in weight.

```
data(mtcars)

model <- lm(mpg ~ I(wt * 0.5) + factor(cyl), data = mtcars)

summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ I(wt * 0.5) + factor(cyl), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5890 -1.2357 -0.5159  1.3845  5.7915
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     33.991      1.888  18.006  < 2e-16 ***
## I(wt * 0.5)     -6.411      1.508  -4.252 0.000213 ***
## factor(cyl)6    -4.256      1.386  -3.070 0.004718 **
## factor(cyl)8    -6.071      1.652  -3.674 0.000999 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.557 on 28 degrees of freedom
## Multiple R-squared:  0.8374, Adjusted R-squared:   0.82
## F-statistic: 48.08 on 3 and 28 DF,  p-value: 3.594e-11
```

Question 5

Consider the following data set

```
x <- c(0.586, 0.166, -0.042, -0.614, 11.72)
y <- c(0.549, -0.026, -0.127, -0.751, 1.344)
```

Give the hat diagonal for the most influential point

```


model <- lm(y ~ x)
influence.measures(model)
```

Solution

```
## Influence measures of
##   lm(formula = y ~ x) :
##
##    dfb.1_     dfb.x     dffit cov.r   cook.d   hat inf
## 1  1.0621 -3.78e-01    1.0679 0.341 2.93e-01 0.229   *
## 2  0.0675 -2.86e-02    0.0675 2.934 3.39e-03 0.244
## 3 -0.0174  7.92e-03   -0.0174 3.007 2.26e-04 0.253   *
## 4 -1.2496  6.73e-01   -1.2557 0.342 3.91e-01 0.280   *
## 5  0.2043 -1.34e+02 -149.7204 0.107 2.70e+02 0.995   *
```

```
# Answer: 0.9946
```

Question 6

Consider the following data set

```r
x <- c(0.586, 0.166, -0.042, -0.614, 11.72)
y <- c(0.549, -0.026, -0.127, -0.751, 1.344)
```

Give the slope dfbeta for the point with the highest hat value.

```r
model <- lm(y ~ x)
influence.measures(model)
```

Solution


```
## Influence measures of
##   lm(formula = y ~ x) :
##
##    dfb.1_     dfb.x     dffit cov.r   cook.d   hat inf
## 1  1.0621 -3.78e-01    1.0679 0.341 2.93e-01 0.229   *
## 2  0.0675 -2.86e-02    0.0675 2.934 3.39e-03 0.244
## 3 -0.0174  7.92e-03   -0.0174 3.007 2.26e-04 0.253   *
## 4 -1.2496  6.73e-01   -1.2557 0.342 3.91e-01 0.280   *
## 5  0.2043 -1.34e+02 -149.7204 0.107 2.70e+02 0.995   *
```


```
# Answer: -134
```

Question 7

Consider a regression relationship between Y and X with and without adjustment for a third variable Z. Which of the following is true about comparing the regression coefficient between Y and X with and without adjustment for Z.

Solution

- The coefficient can't change sign after adjustment, except for slight numerical pathological cases.

- ***It is possible for the coefficient to reverse sign after adjustment. For example, it can be strongly significant and positive before adjustment and strongly significant and negative after adjustment.***
- For the the coefficient to change sign, there must be a significant interaction term.
- Adjusting for another variable can only attenuate the coefficient toward zero. It can't materially change sign.