Data Science - Regression Models - Quiz 3 - Coursera

Quiz 3

This is Quiz 3 from Coursera's Regression Models class within the Data Science Specialization. This publication is intended as a learning resource, all answers are documented and explained.

---

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.

- **Answer:** -6.071

Explanation:

R assumes the first level of the factor is the reference level (4 cylinder). The coefficients give the betas for each factor. Changing from a 4 cyclinder engine to an 8 cyclinder loses 6 mpg holding weight fixed.

```
#Loading and examining the Data
data(mtcars)
head(mtcars)
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
#Fitting model
fit <- lm(mpg ~ factor(cyl) + wt,mtcars)
summary(fit)$coef
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  33.990794  1.8877934 18.005569 6.257246e-17
## factor(cyl)6 -4.255582  1.3860728 -3.070244 4.717834e-03
## factor(cyl)8 -6.070860  1.6522878 -3.674214 9.991893e-04
```

```
## wt            -3.205613  0.7538957 -4.252065 2.130435e-04
```

```
#Selecting coefficient
summary(fit)$coef[3,1]
```

```
## [1] -6.07086
```

---

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?.

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

Explanation:

The unadjusted beta values are higher. Weight is confounding significantly.

```
fit <- lm(mpg ~factor(cyl), mtcars)
afit <- lm(mpg~factor(cyl) + wt,mtcars)


summary(fit)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   26.663636  0.9718008 27.437347 2.688358e-22
## factor(cyl)6  -6.920779  1.5583482 -4.441099 1.194696e-04
## factor(cyl)8 -11.563636  1.2986235 -8.904534 8.568209e-10
```

```
summary(afit)$coef
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  33.990794  1.8877934 18.005569 6.257246e-17
## factor(cyl)6 -4.255582  1.3860728 -3.070244 4.717834e-03
## factor(cyl)8 -6.070860  1.6522878 -3.674214 9.991893e-04
## wt           -3.205613  0.7538957 -4.252065 2.130435e-04
```

---

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.

- **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.**

```
fit <- lm(mpg ~factor(cyl)+wt, mtcars)
Ifit <- lm(mpg~factor(cyl)*wt,mtcars)
anova(fit,Ifit)
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(cyl) + wt
## Model 2: mpg ~ factor(cyl) * wt
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 183.06
## 2     26 155.89  2     27.17 2.2658 0.1239
```

---

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:

```
fit4 <- lm(mpg ~ I(wt * 0.5) + factor(cyl), data = mtcars)
```

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

Explanation:

Mtcars reports the weight in units of 1000 lbs. Using I(wt*.5) doubles the weight coefficient from the previous model. This reflects a 2000 lbs (1 ton) increase holding the factor variable fixed.

```
summary(fit)
##
## 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(fit4)
```

```
## 
## 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
```

---

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)
```

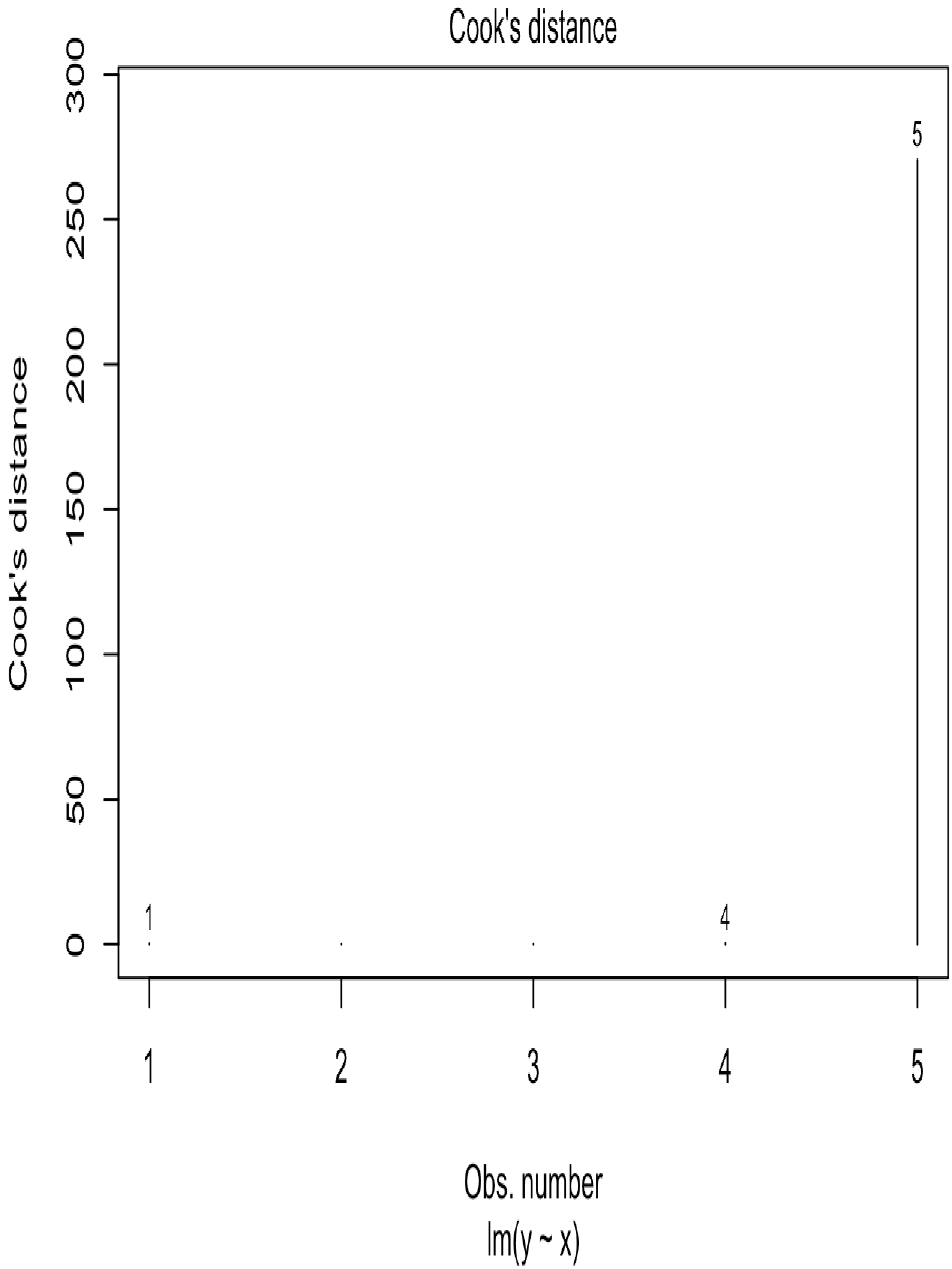
Give the hat diagonal for the most influential point

- **0.9946**

Explanation:

Generate linear model, use R to compute hat values. Cook's distance shows point of interest.

```
fit5 <- lm(y~x)
hatvalues(fit5)
##         1         2         3         4         5
## 0.2286650 0.2438146 0.2525027 0.2804443 0.9945734
plot(fit5, which = 4)
```


Cook's distance
Cook's distance
0
50
100
150
200
250
300
1
2
3
4
5
Obs. number
lm(y ~ x)

---

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

- **-134**

Explanation:

Generate linear model, use R to compute dfbeta values.

```
fit6 <- lm(y~x)
dfbetas(fit6)
##   (Intercept)            x
## 1  1.06212391   -0.37811633
## 2  0.06748037   -0.02861769
## 3 -0.01735756    0.00791512
## 4 -1.24958248    0.67253246
## 5  0.20432010 -133.82261293
```

---

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.

- **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.**

Explanation:

This is an example of Simpsons paradox and the importance of model selection. Below is an example from the swiss dataset which shows the Beta value flipping when all variables are included. Agriculture is highly correlated with education. If you take out this correlation effect the coefficient flips.

```
data(swiss)
```

```
summary(lm(Fertility~Agriculture,data=swiss))$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 60.3043752 4.25125562 14.185074 3.216304e-18
## Agriculture  0.1942017 0.07671176  2.531577 1.491720e-02
```

```
summary(lm(Fertility~., swiss))
```

```
## 
## Call:
## lm(formula = Fertility ~ ., data = swiss)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -15.2743  -5.2617   0.5032   4.1198  15.3213 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
## Agriculture      -0.17211    0.07030  -2.448  0.01873 *  
## Examination      -0.25801    0.25388  -1.016  0.31546    
## Education        -0.87094    0.18303  -4.758 2.43e-05 ***
## Catholic          0.10412    0.03526   2.953  0.00519 ** 
## Infant.Mortality  1.07705    0.38172   2.822  0.00734 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671 
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```