APSTA-GE 2003: Intermediate Quantitative Methods

Sample Solution - Assignment 3

Created on: 11/10/2020

Modified on: 11/10/2020

Part 1

Question 1

The regression output contains the results of T-test about the regression coefficients.

For coefficient of `mheight`, what are the null and alternative hypotheses tested here?

Below we use β_1 to represent the regression coefficient for `mheight`.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 47.56386    0.77925   61.04   <2e-16 ***
mheight      0.32566    0.01197   27.20   <2e-16 ***
```

A. H_0: β_1 = 0 versus H_1: β_1 = 1.

B. H_0: β_1 = 0 versus H_1: β_1 not equal to 0.

C. H_0: β_1 = 0 versus H_1: β_1 > 0

D. H_0: β_1 = 1 versus H_1: β_1 = 0

```r
# Load the dataset `parent_son.csv` to R using read.csv(),
# define the dataset as `dat`.
dat <- read.csv("../data/parent_son.csv")

# Check the structure of `dat`
str(dat)
```

```
## 'data.frame':    1078 obs. of  4 variables:
##  $ id     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fheight: num  65 63.3 65 65.8 61.1 ...
##  $ sheight: num  59.8 63.2 63.3 62.8 64.3 ...
##  $ mheight: num  44.7 53.3 59.3 48 61.5 ...
```

Answer: Q1

H_0: β_1 = 0.

H_1: β_1 not equal to 0.

Question 2

Based on the regression result, we can conclude that, at significance level 5%, son's height is statistically significantly related to mother's height.

Answer: Q2

This is true because p-value is smaller than 0.05.

Question 3

Now let's consider a different hypothesis testing problem.

Hypothetically suppose that, on average, sons perfectly inherit mother's height gene (but not their father's); that the average height of sons born to mother's of the same height is expected to be the same.

In this case, we can hypothesize the slope coefficient to be 1. The hypotheses to be tested will then be:

H_0: β_1 = 1; H_a: β_1 not equal to 1.

Calculate the test statistic (T-test) that tests such hypotheses using this formula:

$$T = \frac{\hat{\beta}_1 - \beta_1(\text{hypothesize value})}{SE(\hat{\beta}_1)}$$

Answer: Q3

```
beta1 <- 0.32566
beta1_hypo <- 1
se_beta1 <- 0.01197
(T_Q3 <- (beta1 - beta1_hypo) / se_beta1)
```

```
## [1] -56.33584
```

Test statistic is -56.34.

Question 4

What's the p-value for this test? (Sample size n = 1058)

Hint: you can compute the probability under the curve in R using `pt(T, n - 2)`. This function gives you the probability that P(t<T) under t distribution with degrees of freedom $n - 2$.

To get the correct p-value for a two-sided test, you need to twice this probability.

The p-value for this test is: ____ (round to three digits after decimal point (1 thousandth).

If the answer is less than 0.001, put 0)

Answer: Q4

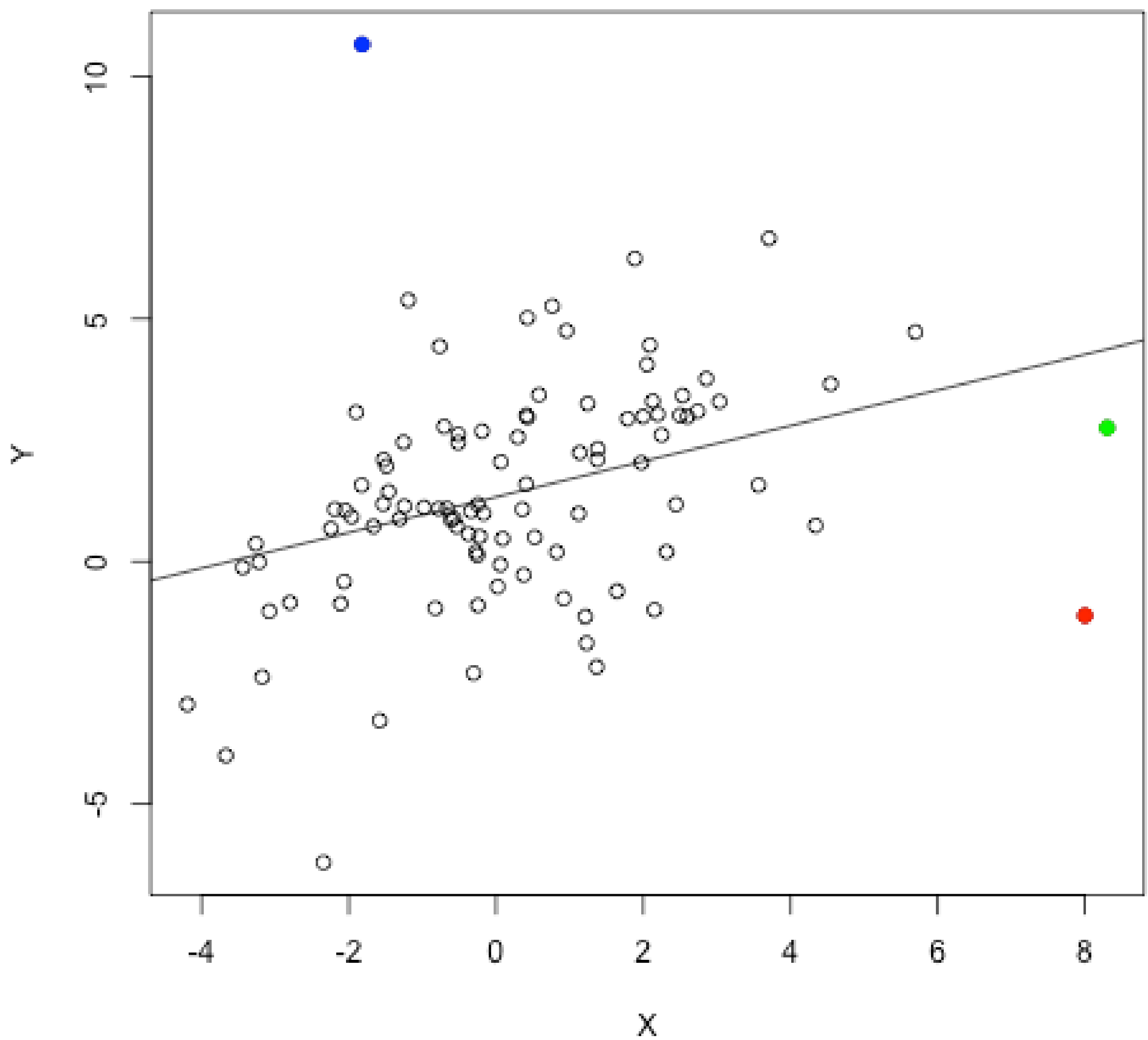
```
n <- 1058
pt(T_Q3, n - 2) * 2
```

```
## [1] 1.794446e-320
```

The p-value is 0.

Part 2 - Regression Diagnostics

Question 1



The scatter plot between X and Y is shown in this figure, together with a fitted regression line. Examine the three colored dots (red, green and blue dots) and match them with their characteristics.

A. an outlier along the Y but not along the X

B. an outlier along the X but not along the Y

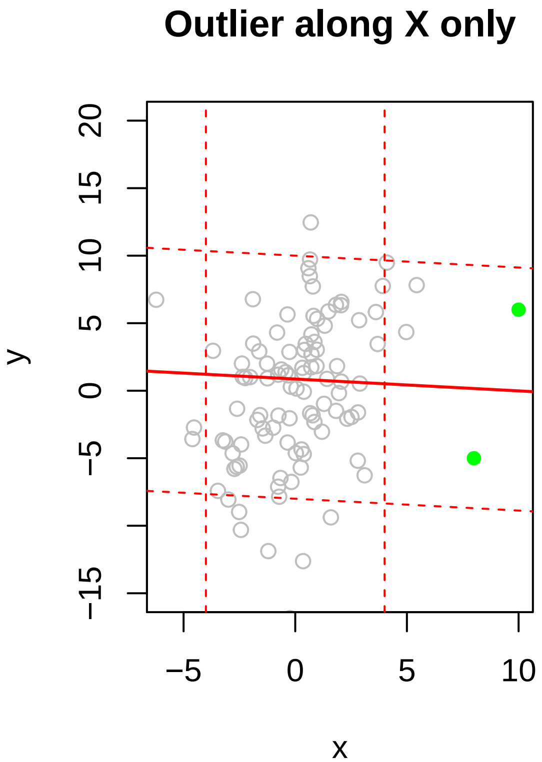
C. an outlier along both the Y and the X

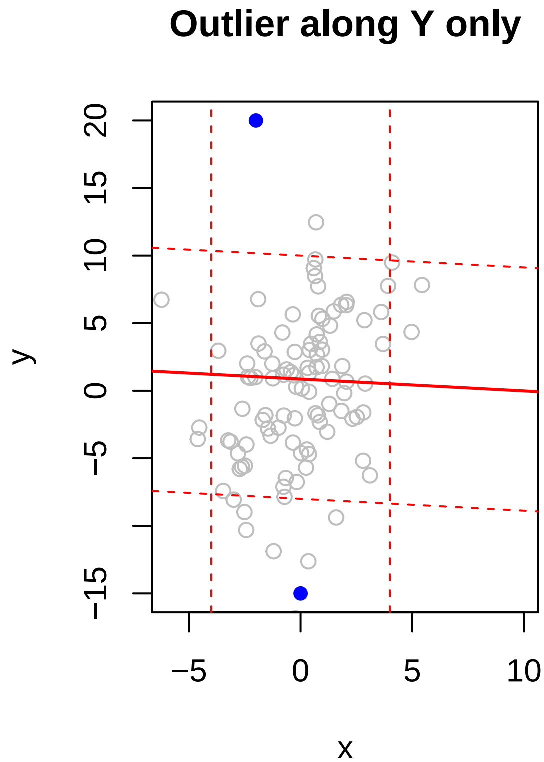
Answer: P2Q1

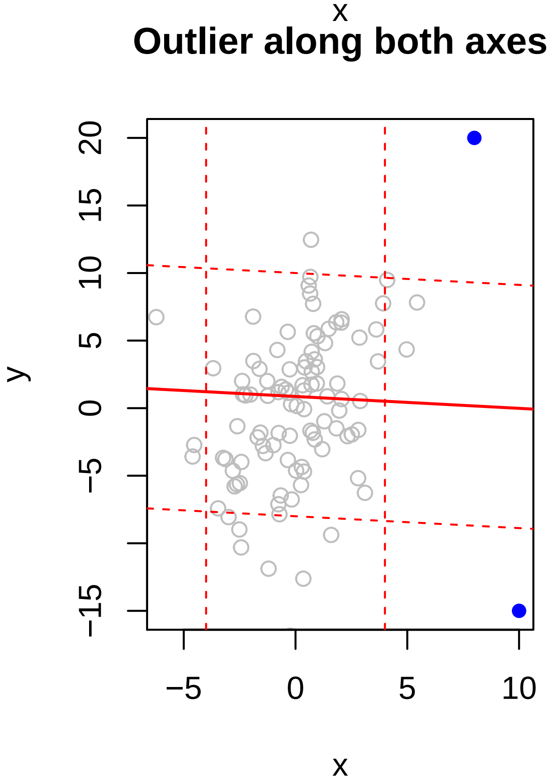
Outlier along the x axis only: green dot

Outlier along the y axis only: blue dot

Outlier along both axes: red dot







Question 2

The dataset of the figure in Part2: Q1 is given in this question. Furthermore, the blue dot is data entry 98, the green dot is data entry 99 and the red dot is data entry 100. Answer the following True/False questions through experiments (namely, examining the regression results dropping one colored dot at a time.)

Dropping a data point that is only an outlier along the X (high leverage only) will not likely change the regression results very much.

Answer: P2Q2

```
dat <- read.csv("../data/hwk3_part2.csv")
tail(dat)
```

```
##             X         Y
## 95  -2.189718  1.074797
## 96  -1.825038  1.571325
## 97   2.049429  4.056204
## 98  -1.817718 10.652952
## 99   8.300000  2.752453
## 100  8.000000 -1.116266
```

```
# Drop data entry 99: high leverage point
dat_no99 <- dat[-99, ]

(mod_P2Q2_all <- lm(Y ~ X, data = dat))
```

```
##
## Call:
## lm(formula = Y ~ X, data = dat)
##
## Coefficients:
## (Intercept)            X
##      1.3277       0.3671
```

```
(mod_P2Q2_no99 <- lm(Y ~ X, data = dat_no99))
```

```
##
## Call:
## lm(formula = Y ~ X, data = dat_no99)
##
## Coefficients:
## (Intercept)            X
##      1.3384       0.3972
```

This is true.

Question 3

Dropping a data point that is both an outlier along the X and along the Y will have a large impact on the regression coefficient estimates.

Hint: you can try this out using the dataset given from last question.

Answer: P2Q3

```r
dat_no100 <- dat[-100, ]

print(mod_P2Q2_all)
```

```
##
## Call:
## lm(formula = Y ~ X, data = dat)
##
## Coefficients:
## (Intercept)            X
##      1.3277       0.3671
```

```r
(mod_P2Q3_no100 <- lm(Y ~ X, data = dat_no100))
```

```
##
## Call:
## lm(formula = Y ~ X, data = dat_no100)
##
## Coefficients:
## (Intercept)            X
##      1.3636       0.4622
```

This is true.

Question 4

Dropping a data point that is a large outlier along the Y axis will decrease the Standard Error (SE) of the regression coefficients when the sample size is small or moderate (~n<=100).

Answer: Q8

```r
dat_no98 <- dat[-98, ]

mod_P2Q4_no98 <- lm(Y ~ X, data = dat_no98)

summary(mod_P2Q2_all)$coefficients
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 1.3277262  0.2258410 5.879031 5.714131e-08
## X           0.3671333  0.1000508 3.669468 3.955052e-04
```

```r
summary(mod_P2Q4_no98)$coefficients
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 1.2143713 0.20345554 5.968731 3.913949e-08
## X           0.4095113 0.08997414 4.551433 1.545665e-05
```

The standard error of X decreases from 0.1 to 0.09. This is true.

END: Sample Solution - Assignment 3