APSTA-GE 2003: Intermediate Quantitative Methods

Sample Solution - Assignment 2

Created on: 11/09/2020

Modified on: 11/10/2020

Instructions

In this section, you will use the data set `parent_son.csv` to conduct regression analysis and answer questions.

Throughout this homework (all parts), please round all numerical answers to the nearest 100th decimal point, unless otherwise instructed in the question.

```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 ...
```

First, load the dataset to R using `read.csv()` as shown above. Then, check dimensions and the structure:

`dat` has 1,078 rows and 4 columns:

- `id`: Index
- `fheight`: Father's height, numeric
- `sheight`: Son's height, numeric
- `mheight`: Mother's height, numeric

```r
# Create a function to round to the nearest 100th decimal point
rd <- function (n) {
  round(as.numeric(n), digits = 2)
}
```

Question 1

Estimate a simple linear regression using mother's height ("`mheight`") as independent variable X, and son's height ("`sheight`") as outcome variable Y.

Use R function: `lm(Y ~ X, data)`

The intercept is: ____

The slope is: ____

Answer: Q1

```
mod_Q1 <- lm(sheight ~ mheight, data = dat)
summary(mod_Q1)
```

```
##
## Call:
## lm(formula = sheight ~ mheight, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0013 -1.3998  0.0316  1.4004  6.7875
##
## 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 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.168 on 1076 degrees of freedom
## Multiple R-squared:  0.4075, Adjusted R-squared:  0.4069
## F-statistic: 739.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

```
intercept_mod_Q1 <- 47.56386
slope_mod_Q1 <- 0.32566
```

The intercept is 47.56.

The slope is 0.33.

Question 2

In your own language, explain what the intercept means (in the context of mother and son's height).

Answer: Q2

The intercept is the average height of sons whose mother's heights are 0 inch.

Question 3

In your own language, explain the meaning of the slope in the context of this analysis.

Answer: Q3

For two groups of mothers whose heights differ by one inch, the taller group will have sons who are 0.32 inches taller on average.

Question 4

Report the 95% confidence interval of the slope coefficient

lower bound: ____

upper bound: ____

Answer: Q4

```
confint(mod_Q1)
```

```
##                  2.5 %    97.5 %
## (Intercept) 46.0348472 49.092876
## mheight      0.3021704  0.349154
```

```
lwr_bound <- 0.3021704
upp_bound <- 0.349154
```

lower bound: 0.3.

upper bound: 0.35.

Question 5

What is the expected height of sons whose mothers' heights are 65 inches tall? (keep two decimal point)

Answer: Q5

We can manually predict by using the model equation:

$$\begin{aligned}
\text{Son's Height} &= 47.56386 + 0.32566 \cdot \text{Mother's Height} \\
\text{Mother's Height}_1 &= 65 \\
\text{Son's Height}_1 &= 47.56386 + 0.32566 \cdot 65 = 68.73
\end{aligned}$$

Alternatively, we can use `predict()`.

```
new_dat <- data.frame(mheight = 65)
predict(mod_Q1, newdata = new_dat)
```

```
##       1
## 68.7319
```

Question 6

Estimate the average height of sons whose mother's height is 60 inches and the confidence interval of this estimate.

Hint: you can use R command `predict()`. Use "?predict.lm" to learn about how "predict" works for "lm" objects.

Average height of sons whose mother's height is 60 inches: ____

The lower bound of the 95% CI: ____

The upper bound of the 95% CI: ____

Answer: Q6

```
new_dat <- data.frame(mheight = 60)
predict(mod_Q1, newdata = new_dat, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 67.10359 66.93103 67.27616
```

Average height of sons whose mother's height is 60 inches: 67.1.

The lower bound of the 95% CI: 66.93.

The upper bound of the 95% CI: 67.28.

Question 7

Maria is 60 inches tall, can you predict her son's adult height based on the regression model you estimated? Prescribe a 95% prediction interval of your prediction.

Predicted height: ____

Lower bound of the 95% prediction interval: ____

Upper bound of the 95% prediction interval: ____

Answer: Q7

```
predict(mod_Q1, newdata = new_dat, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 67.10359 62.84675 71.36044
```

Since we are referring to the same regression model, the predicted value will be the same. However, in this question, our prediction is based on the information of a single individual instead of the average, the prediction interval (individual-level) will be wider than the confidence interval (average-level) to reflect reduced accuracy.

Predicted height: 67.1.

Lower bound of the 95% prediction interval: 62.85.

Upper bound of the 95% prediction interval: 71.36.

Question 8

Please explain briefly why the prediction interval is considerably wider in Question 7 than the confidence interval in Question 6.

You can offer intuitions or use mathematical formula for the explanation.

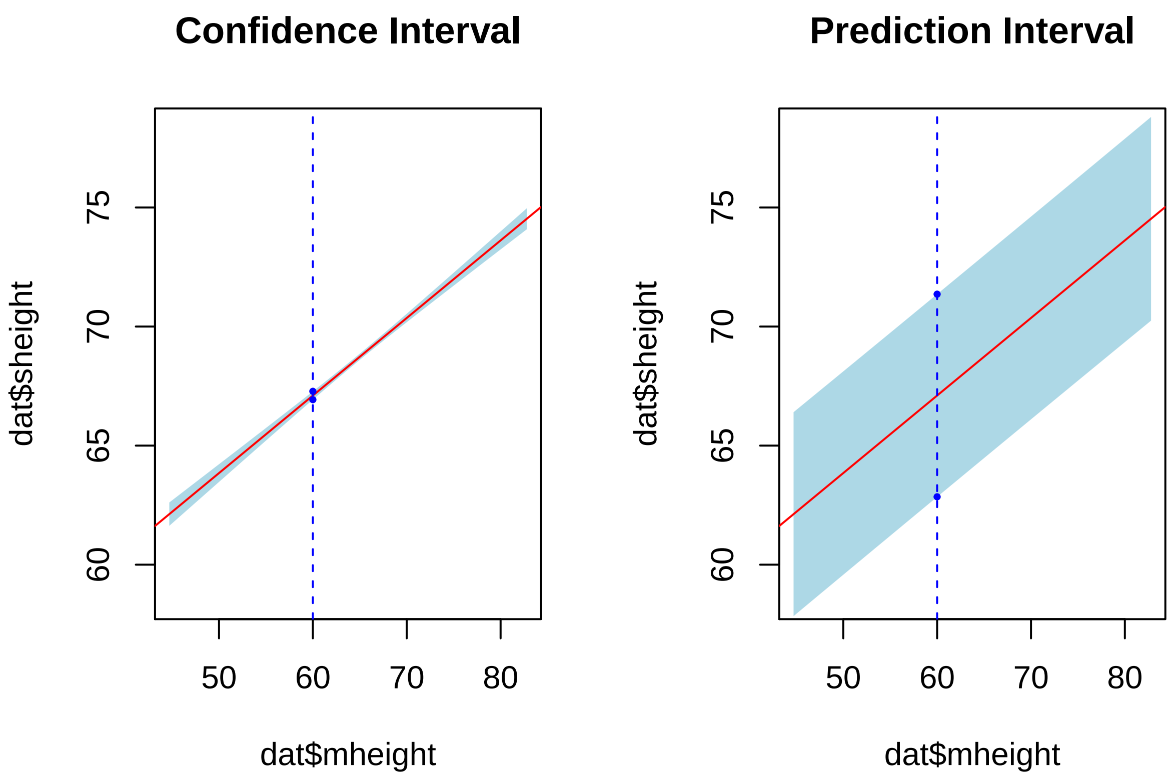
Answer: Q8

In Question 6, the confidence interval reflects the uncertainty around an estimated mean height of a group, while in Question 7, the prediction interval reflects the uncertainty around a predicted value of an individual. It is harder to predict individual's height than estimate the mean height of a group.

```
# Visualize the difference between individual and average
par(mfrow = c(1, 2)) # Create a 1-by-2 grid canvas

# Draw confidence interval
mheight <- sort(dat$mheight, decreasing = FALSE)
new_dat <- data.frame(mheight)
sheight <- predict(mod_Q1, newdata = new_dat, interval = "confidence")
plot(dat$mheight, dat$sheight, type = "n") # pch: point character
title(main = "Confidence Interval")
polygon(x = c(rev(mheight), mheight),
        y = c(rev(sheight[, "upr"]), sheight[, "lwr"]),
        col = "light blue", border = NA)
abline(a = 47.56386, b = 0.32566, col = "red")
abline(v = 60, lty = "dashed", col = "blue")
points(x = 60, y = 66.93, pch = 16, col = "blue", cex = .5)
points(x = 60, y = 67.28, pch = 16, col = "blue", cex = .5)

# Draw prediction interval
sheight <- predict(mod_Q1, newdata = new_dat, interval = "prediction")
plot(dat$mheight, dat$sheight, type = "n") # pch: point character
title(main = "Prediction Interval")
polygon(x = c(rev(mheight), mheight),
        y = c(rev(sheight[, "upr"]), sheight[, "lwr"]),
        col = "light blue", border = NA)
abline(a = 47.56386, b = 0.32566, col = "red")
abline(v = 60, lty = "dashed", col = "blue")
points(x = 60, y = 62.85, pch = 16, col = "blue", cex = .5)
points(x = 60, y = 71.36, pch = 16, col = "blue", cex = .5)
```



Question 9

The R squared is the percentage of the variability in Y explained by X. It can be calculated using the formula above.

Report the R squared: _____

Answer: Q9

```
mod_Q9 <- mod_Q1
summary(mod_Q9)$r.square
```

```
## [1] 0.4074578
```

The R squared is 0.41.

Question 10

Report the total sum of squares (TSS) for the dependent variable (son's height).

Total sum of square: _____

Answer: Q10

```
mod_Q10 <- mod_Q1
anova(mod_Q10)
```

```
## Analysis of Variance Table
##
```

```
## Response: sheight
##             Df Sum Sq Mean Sq F value    Pr(>F)
## mheight      1 3476.7  3476.7   739.9 < 2.2e-16 ***
## Residuals 1076 5055.9     4.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$TSS = MSS + RSS = 3476.7 + 5055.9 = 8532.6$$

Question 11

Report the model sum of squares (MSS).

Model Sum of Squares: ____

Answer: Q11

```r
mod_Q11 <- mod_Q1
anova(mod_Q11)
```

```
## Analysis of Variance Table
##
## Response: sheight
##             Df Sum Sq Mean Sq F value    Pr(>F)
## mheight      1 3476.7  3476.7   739.9 < 2.2e-16 ***
## Residuals 1076 5055.9     4.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$MSS = 3476.7$$

Question 12

Report the residual sum of squares (RSS).

Residual Sum of Squares: ____

Answer: Q12

```r
mod_Q12 <- mod_Q1
anova(mod_Q12)
```

```
## Analysis of Variance Table
##
## Response: sheight
##             Df Sum Sq Mean Sq F value    Pr(>F)
## mheight      1 3476.7  3476.7   739.9 < 2.2e-16 ***
## Residuals 1076 5055.9     4.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$RSS = 5055.9$$

Question 13

Report model sum of squares divided by the total sum of squares (MSS/TSS).

Answer: Q13

$$MSS/TSS = \frac{3476.7}{8532.6} = 0.4074608$$

The $\frac{MSS}{TSS}$ is 0.41.

Question 14

The model sum of squares captures the part of the variability in son's height (DV) that is due to their mother's height variability.

Answer: Q14

This is true.

END: Sample Solution - Assignment 2