APSTA-GE 2003: Intermediate Quantitative Methods

Sample Solution - Assignment 5

Created on: 11/09/2020

Modified on: 11/10/2020

Instructions

In this part, you will analyze dataset `toy_example3.csv` (in the `data` folder).

```
# Load the dataset `toy_example3.csv` to R using read.csv(),
# define the dataset as `dat`.
dat <- read.csv("../data/toy_example3.csv")

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

```
## 'data.frame':    500 obs. of  6 variables:
##  $ X          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ id         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Sex        : Factor w/ 2 levels "F","M": 1 2 1 1 1 2 2 2 1 1 ...
##  $ Birth.Order: int  1 1 1 2 1 1 1 3 2 2 ...
##  $ Height     : num  72.3 68 60 67.1 68.4 ...
##  $ Weight     : num  168 129 127 124 144 ...
```

```
# Number of rows: 500
nrow(dat)
```

```
## [1] 500
```

```
# Numer of columns: 6
ncol(dat)
```

```
## [1] 6
```

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

`dat` has 500 rows and 6 columns:

- `X`: Index, same as `id`
- `id`: Index, same as `X`
- `Sex`: Biological sex, factor with 2 levels, "F" as *female*, "M" as *male*
- `Birth.Order`: Birth order, integer
- `Height`: Height, numeric
- `Weight`: Weight, numeric

Question 1

First, fit a regression model using `height` and biological sex (`Male`) (i.e., Female is the reference group) as predictors for `weight`, assuming an **additive relationship**.

Then, fit the regression again assuming an **interaction relationship** between `height` and biological sex (`Male`). Match the assumptions/interpretations with the appropriate model.

Hint: make sure you code variable "sex" as instructed.

1. The regression coefficient of `height` is the same for male and for female.
2. The male and female difference in `weight` is the same at any `height`.
3. The regression coefficient of `height` is more pronounced among males than among females, i.e. greater positive slope among males.
4. The male and female difference in `weight` is greater among taller people.

Answer: Q1

To fit a multiple regression model, we need to determine variables as components first.

- Dependent variable: `Weight`
- Independent variables:
  - `Height`
  - `Male`

Since we don't have a `Male` variable in our data frame, we need to create one. We know that `Male` is based on `Sex`. Therefore, we can use dummy coding to create a new variable, `Male`, so that `Male` is 1 if `Sex` = 1; 0 if `Sex` = 0.

```
# Create a new variable, Male, based on Sex
# For each row in `dat`:
#   if `Sex` is "M", then `Male` is 1;
#   if `Sex` is not "M", then `Male` is 0.
dat$Male <- ifelse(dat$Sex == "M", yes = 1, no = 0)
```

Then, we're ready to fit a regression model. First, create one with additive relationship. We **assume** that, in a regression model with additive effects only, each independent variable is independent with others. There is no relationship among independent variables. In this case, we assume that there is no relationship between `Height` and `Sex` (probably not true in reality but let's assume in this way in our model). We assume that, on average, `Weight` is determined by adding up `Height` with `Sex` indicator:

$$\text{Weight} = \beta_0 + \beta_1 \cdot \text{Height} + \beta_2 \cdot \text{Male} + \varepsilon$$

where β_0 is the intercept, β_1 is the correlation coefficient of `Height`, β_2 is the coefficient of `Male`, and ε is the error.

To clarify terms in this case:

- The intercept is the value of `Weight` when all independen variables are zero: $\beta_0 = \text{Weight}_{x=0}$.
- β_1 is the average (expected) influence of one additional unit increase of `Height` on `Weight`, controlling the `Male` status.
- β_2 is the average (expected) difference of `Weight` for male and female with same `Height`.
- ε is the data variance that cannot be explained by our model.

```
# Fit a multiple regression model with additive effects,
# check results using summary().
```

```r
mod_Q1_add <- lm(Weight ~ Height + Male, data = dat)
summary(mod_Q1_add)
```

```
##
## Call:
## lm(formula = Weight ~ Height + Male, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -70.368 -13.384  -0.107  14.263  62.384
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.1872    15.2535   2.176     0.03 *
## Height        1.4185     0.2301   6.164 1.46e-09 ***
## Male          9.7232     2.0477   4.748 2.69e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.24 on 497 degrees of freedom
## Multiple R-squared:  0.1844, Adjusted R-squared:  0.1811
## F-statistic: 56.17 on 2 and 497 DF,  p-value: < 2.2e-16
```

To clarify terms in the summary:

- The relationship between `Weight` and `Height`, and `Weight` and `Male` are both statistically significant.
- On average, for same biological sex, one additional unit increase of `Height` leads to 1.4185 increase of `Weight`.
- On average, for same `Height`, male is expected to be 9.7232 units heavier than female.
- Confidence interval is: Estimate $\pm$ 1.96 $\cdot$ Standard Error We can get confidence interval for each coefficient estimate.
- t-value is the difference to the null hypothesis. It measures the distance to the null value ($\beta = 0$). For example, the t-value of `Height` in our case is 6.164. This means that, in a t-distribution (a bell-shaped curve similar to normal distribution but with thicker tails), our t-value is 6.164 units away from null. If we perform an one-tail t-test, we can get the p-value of `Height` by calculating the area under the curve (AUC) of the t-distribution that is higher than 6.164. This is the probability that null hypothesis (no difference of `Weight` for one additional `Height` increase) holds. If we perform a two-tail t-test, then the p-value is the AUC of the t-distribution that is higher than 6.164 or lower than -6.164. In our case, the p-value is very close to 0 (definitely lower than 0.05). This indicates that, with no more than 5% of Type I error, we reject the null hypothesis. it is extremely rare that `Weight` will be the same for one additional `Height` increase.
- R-squared measures how well can the model explain the variance of the data. It is monotonic increasing as we sequentially including more predictiors. As we including more independent variables, as long as they are correlated with the dependent variable, our model tends to perform better.
- The adjusted R-squared is designed to panelize overfitting. With large number of independent variables, we can refer to this statistic for an accurate estimation of model explanation ability. While R-squared is monotonic, adjusted R-squared is not. It decreases if overfit accumulates.
- F-statistic measures the distance to the null hypothesis ($MSS = 0$). Similar to t-score, it measures the distance between the F-value of this model and the null, which is zero. Then, by calculating the AUC that is higher than the F-value (one-tail F-test), we can get the probability of getting the null, which is the p-value. In this case, p-value is very close to zero. With more than 95% of confidence, we reject the null hypothesis. The sample on which our model is based can well represent the variance of the population.

Now, we can write down our model estimation:

$$\text{Weight} = 33.1872 + 1.4185 \cdot \text{Height} + 9.7232 \cdot \text{Male} + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, 20.2^2)$.

To predict, for a male that is 70 inches high, we predict the weight to be 142.2054 lbs:

$$\text{Weight'} = 33.1872 + 1.4185 \cdot 70 + 9.7232 \cdot 1 = 142.2054$$

Then, let's fit a model with interactive relationship:

$$\text{Weight} = \beta_0 + \beta_1 \cdot \text{Height} + \beta_2 \cdot \text{Male} + \beta_3 \cdot \text{Height} \cdot \text{Male} + \varepsilon$$

where β_0 is the intercept, β_1 is the correlation coefficient of `Height`, β_2 is the coefficient of `Male`, β_3 is the coefficient of the interactive relationship, and ε is the error.

```
mod_Q1_inter <- lm(Weight ~ Height * Male, data = dat)
summary(mod_Q1_inter)
```

```
##
## Call:
## lm(formula = Weight ~ Height * Male, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -68.90 -13.00  -0.50  13.52  61.24
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  68.9328    20.9023   3.298  0.00104 **
## Height        0.8774     0.3158   2.778  0.00568 **
## Male        -68.0695    31.3503  -2.171  0.03038 *
## Height:Male   1.1400     0.4584   2.487  0.01322 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.14 on 496 degrees of freedom
## Multiple R-squared:  0.1944, Adjusted R-squared:  0.1895
## F-statistic:  39.9 on 3 and 496 DF,  p-value: < 2.2e-16
```

We can write down our model estimation:

$$\begin{aligned}
\text{Weight} &= 68.9328 + 0.8774 \cdot \text{Height} - 68.0695 \cdot \text{Male} + 1.14 \cdot \text{Height} \cdot \text{Male} + \varepsilon \\
&= 68.9328 + (0.8774 + 1.14 \cdot \text{Male}) \textbf{ Height} - 68.0695 \cdot \text{Male} + \varepsilon \\
&= 68.9328 + (-68.0695 + 1.14\text{Height}) \textbf{ Male} + 0.8774 \cdot \text{Height} + \varepsilon
\end{aligned}$$

where $\varepsilon \sim \mathcal{N}(0, 20.07^2)$. The equation can also be written in form of conditional variable: holding `Male` as constant (either 1 for male or 0 for female) and `Height` as variable or holding `Height` as constant and `Male` as variable.

To interpret, For male, **an additional unit increase** in `Height` leads to a `Weight` increase of 2.8807 lbs. For female, **an additional unit increase** in `Height` leads to a `Weight` increase of 0.8774 lbs.

$$\text{Weight"}_{male} = 0.8774 + 1.14 \cdot 1 = 2.8807$$
$$\text{Weight"}_{female} = 0.8774 + 1.14 \cdot 0 = 0.8774$$

To predict, for a male that is 70 inches high, we predict the weight to be 142.0813 lbs:

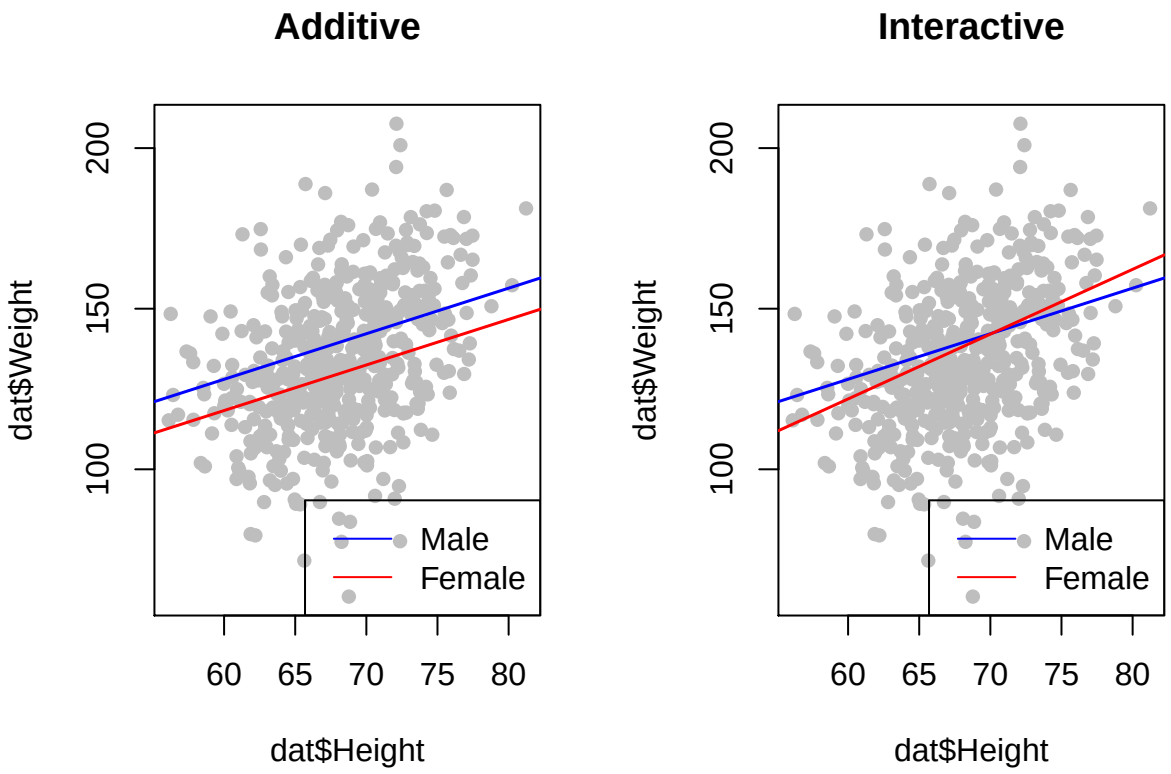
$$\text{Weight"} = 68.9328 + 0.8774 \cdot 70 - 68.0695 \cdot 1 + 1.14 \cdot 70 \cdot 1 = 142.0813$$

Finally, we can visualize these two models and evaluate question statements.

The regression coefficient of `height` is the same for male and for female.

```r
par(mfrow = c(1, 2)) # Create a 1-by-2 grid canvas
# First, create a scatter plot with addtive regression lines
plot(dat$Height, dat$Weight, pch = 16, col = "gray") # pch: point character
title(main = "Additive")
abline(a = 33.1872 + 9.7232, b = 1.4185, lwd = 1.5, col = "blue") # lwd = line width
abline(a = 33.1872,          b = 1.4185, lwd = 1.5, col = "red")
legend("bottomright", legend = c("Male", "Female"),
       lty = c("solid", "solid"), col = c("blue", "red"))

# Second, a scatter plot with interactive regression lines
plot(dat$Height, dat$Weight, pch = 16, col = "gray")
title(main = "Interactive")
abline(a = 33.1872 + 9.7232, b = 1.4185, lwd = 1.5, col = "blue") # lwd = line width
abline(a = 68.9328 - 68.0695, b = 0.8774 + 1.14, lwd = 1.5, col = "red")
legend("bottomright", legend = c("Male", "Female"),
       lty = c("solid", "solid"), col = c("blue", "red"))
```



1. The regression coefficient of `height` is the same for male and for female.

- **Additive model**

2. The male and female difference in `weight` is the same at any `height`.

- **Additive model**

3. The regression coefficient of `height` is more pronounced among males than among females, i.e. greater positive slope among males.

- **Interaction model**

4. The male and female difference in `weight` is greater among taller people.

- **Interaction model**

Question 2

In order to answer the following questions, you will need to create a new variable with "centered" height. The mean height in this data set is 67.6 inches. Generate a new variable called "`height0`" using the code below (assumes you've named the dataset "`dat`"). Round to two decimal places and do not use scientific notation.

Report the mean value of `height0`.

Answer: Q2

Centerization measures the distance of each `Height` value to the average of `Height`. Therefore, the average of centerized `Height` should be zero.

For example, vector A has three elements: 3, 5, 10. The average is 6. Their distance to the average is -3, -1, 4. The average of their distances is 0.

```
dat$height0 <- dat$Height - mean(dat$Height)
round(mean(dat$height0), digits = 2)
```

```
## [1] 0
```

Question 3

Run a new regression of `weight` (DV) regressed on an interaction between `height0` and `male` (IVs), namely: weight = height0 + male + height0*male.

Compare this model to the model with the original height variable interacted with male (weight = height + male + height*male). Which of the following coefficients are different in these two models (check all that apply)?

A. Intercept

B. Coefficients of `height` and `height0`

C. Coefficient of `male`

D. Coefficient of the interaction term

Answer: Q3

```
mod_Q3 <- lm(Weight ~ height0 * Male, data = dat)
coefficients(mod_Q3)
```

```
##  (Intercept)      height0         Male height0:Male
##  128.6649813    0.8774105    9.5370690    1.1399694
```

```
coefficients(mod_Q1_inter)
```

```
## (Intercept)      Height        Male Height:Male
##  68.9328135   0.8774105 -68.0695310   1.1399694
```

A. Intercept - **Check**. The intercept in the new model is 128.665. Originally, it was 68.9328.

B. Coefficients of `height` and `height0` - They are both 0.8774.

C. Coefficient of `male` - **Check**. The coefficient of `Male` is now 9.5371.

D. Coefficient of the interaction term - They are both 1.14.

Question 4

In the model when regressing `weight` on `height0` interacted with `male`:

Which of the following is a correct interpretation of the **intercept**?

A. Expected average weight of males who have height=0.

B. Expected average weight of males who are average height.

C. Expected average weight of females who have height=0.

D. Expected average weight of females who are average height.

Answer: Q4

```
mod_Q4 <- mod_Q3
coefficients(mod_Q4)
```

```
##  (Intercept)      height0         Male height0:Male
##  128.6649813    0.8774105    9.5370690    1.1399694
```

The intercept is the theoritical `Weight` of a person with the following attributes: - `Male` = 0 - `height0` = 0 (average)

$$\text{Weight}_{intercept} = 128.665 + 0.8774 \cdot 0 + 9.5371 \cdot 0 + 1.14 \cdot 0 \cdot 0 = 128.665$$

Therefore, it is the expected average weight of females (`Male` = 0) who are at average height.

Question 5

In the model when regressing `weight` on `height0` interacted with `male`:

Which of the following is a correct interpretation of the slope coefficient on `height0`?

A. Average difference in weight for males whose heights differ by one inch.

B. Average difference in weight for females whose heights differ by one inch.

C. Average difference in weight for females whose heights differ by one standard deviation.

D. Average difference in weight for males whose heights differ by the average height in the dataset.

Answer: Q5

```
mod_Q5 <- mod_Q3
coefficients(mod_Q3)
```

```
##  (Intercept)      height0         Male height0:Male
##  128.6649813    0.8774105    9.5370690    1.1399694
```

The slope coefficient of `height0` is the estimated `Weight` increase of an additional unit increase in `height0` for a person with the following attributes: - `Male` = 0 - `height0` = 0 (average)

$$\text{Weight}_{height0} = 0 + 0.8774 \cdot 1 + 9.5371 \cdot 0 + 1.14 \cdot 0 \cdot 0 = 0.8774$$

Therefore, it is the average difference in weight for females (`Male` = 0) whose heights differe by one inch.

Question 6

In the model when regressing `weight` on `height0` interacted with `male`:

Which of the following is a correct interpretation of the slope coefficient on `male`?

A. Average difference in weight between males and females when height=0

B. Average difference in weight for males whose heights differ by one inch

C. Average difference in weight between males and females who are at the average height

D. Average difference in height of males and females who are at average weight

Answer: Q6

```
mod_Q6 <- mod_Q3
coefficients(mod_Q6)
```

```
##  (Intercept)      height0         Male height0:Male
##  128.6649813    0.8774105    9.5370690    1.1399694
```

The coefficient of `Male` is the estimated `Weight` difference between male and female when they are at the average height (`height0` = 0).

$$\text{Weight}_{male} = 0 + 0.8774 \cdot 0 + 9.5371 \cdot A + 1.14 \cdot 0 \cdot 0 = B$$

where $A \in \{0, 1\}, B \in \{0, 9.5371\}$.

Therefore, it is the average difference in weight between males and females who are at the average height.

Question 7

In the model when regressing `weight` on `height0` interacted with `male`:

What is the expected difference in weight for males whose heights differ by one inch? (round to 2 decimals)

Answer: Q7

```
mod_Q7 <- mod_Q3
coefficients(mod_Q7)
```

```
##  (Intercept)      height0         Male height0:Male
##  128.6649813    0.8774105    9.5370690    1.1399694
```

Since `Male` = 1, we can get:

$$\text{Weight}_{height+1} = 0.8774 + 1.14 = 2.0174$$

Question 8

When replacing `height` with `height0` in these models, the R squares of the models are not affected.

Answer: Q8

Centerization is just shifting the axis. Before centerization, the data is centered at the average. After centerization, the data is centered at zero. It does not improve model's interpretation ability.

```
# Validate
mod_Q8 <- mod_Q3

summary(mod_Q1_inter)$r.square
```

```
## [1] 0.1943972
```

```
summary(mod_Q8)$r.square
```

```
## [1] 0.1943972
```

Question 9

Run two separate models regressing `weight` on centered height (`height0`), one for male and one for female (in other words, split the data into male and female samples separately, then run the regression model `Weight ~ height0` for each sample separately). This approach is sometimes called to stratify the sample based on a covariate (in this case, "`male`").

Examine the results of these two male and female specific models, and compare the results with the model you ran in previous questions (`Weight ~ height0 + male + height0*male`). Choose all the answers that are correct.

Hint: `lm(weight ~ height0, data = data[data$Male == 0, ]), lm(weight ~ height0, data = data[data$Male == 1, ])`

A. The intercept and slope of the male regression line between weight and height(centered) are the same between two approaches.

B. The intercept and slope of the female regression line between weight and height(centered) are the same between two approaches.

C. Whenever comparable, the SEs are greater in the interaction model than in the sex-stratified model(s).

D. The model sum of squares of the interaction model equals to the model sum of squares of the female model plus the model sum of squares of the male model.

E. The model degrees of freedom of the interaction model equals to the model degrees of freedom of male model plus the female model.

F. The residual degrees of freedom of the interaction model equals to the residual degrees of freedom of male model plus the female model.

G. One difference between the two approaches is that in the interaction model we only estimate one model variance, while in the gender stratified approach, we estimate two separate model variances.

H. One difference between the two approaches is that we now have smaller number of observations in each of the two gender specific models than in the one combined model.

Answer: Q9

```
# Select the `dat` by row where `Male` = 1; define the new dataset as `dat_male`
dat_male <- dat[dat$Male == 1, ]
# Repeat for `dat_female`
dat_female <- dat[dat$Male == 0, ]

mod_Q9_male <- lm(Weight ~ height0, data = dat_male)
mod_Q9_female <- lm(Weight ~ height0, data = dat_female)
mod_Q9_all <- mod_Q3
```

A. The intercept and slope of the male regression line between weight and height(centered) are the same between two approaches. **Check**

```
coefficients(mod_Q9_male)
```

```
## (Intercept)     height0
##   138.20205     2.01738
```

```
coefficients(mod_Q9_all)
```

```
##  (Intercept)      height0         Male height0:Male
##  128.6649813    0.8774105    9.5370690    1.1399694
```

For the stratified approach:

- Intercept: 138.202
- Slope (`height0`): 2.017

For the interaction approach: (`Male` = 1)

- Intercept: $128.665 + 9.537 = 138.202$
- Slope (`height0`): $0.877 + 1.14 = 2.017$

B. The intercept and slope of the female regression line between weight and height(centered) are the same between two approaches. **Check**

```
coefficients(mod_Q9_female)
```

```
## (Intercept)     height0
## 128.6649813   0.8774105
```

```
coefficients(mod_Q9_all)
```

```
##  (Intercept)      height0         Male height0:Male
##  128.6649813    0.8774105    9.5370690    1.1399694
```

For the stratified approach:

- Intercept: 128.665
- Slope (`height0`): 0.877

For the interaction approach: (`Male` = 0)

- Intercept: 128.665
- Slope (`height0`): 0.877

C. Whenever comparable, the SEs are greater in the interaction model than in the sex-stratified model(s).

```
summary(mod_Q9_all)$coefficients
```

```
##                 Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  128.6649813  1.4084865 91.349817 2.070838e-312
## height0        0.8774105  0.3158443  2.777984  5.677015e-03
## Male           9.5370690  2.0384364  4.678620  3.730198e-06
## height0:Male   1.1399694  0.4584334  2.486663  1.322259e-02
```

```
summary(mod_Q9_male)$coefficients
```

```
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 138.20205  1.4977297 92.274362 5.062254e-190
## height0       2.01738  0.3377194  5.973538  8.280203e-09
```

Here, the standard error of `height0` in the interaction approach is smaller than the one in the stratified approach. The stratified approach splits the sample into two. Therefore, each stratified model is based on a smaller sample and, therefore, has less references when performing least-square calculation. This generates larger errors and, thus, higher standard error value.

D. The model sum of squares of the interaction model equals to the model sum of squares of the female model plus the model sum of squares of the male model.

```
# MSS for the interaction model
sum(anova(mod_Q9_all)[1:3, "Sum Sq"])
```

```
## [1] 48524.02
```

```r
# MSS for the stratified model
anova(mod_Q9_male)["height0", "Sum Sq"] + anova(mod_Q9_female)["height0", "Sum Sq"]
```

```
## [1] 18073.81
```

Model sum of squares (MSS) is the data variance that can be explained by the model. Since stratified models are based on smaller samples, the explaination ability of these models are weaker. Therefore, they have smaller sum of MSS values than the interaction model's.

E. The model degrees of freedom of the interaction model equals to the model degrees of freedom of male model plus the female model.

```r
# Model DF for the interaction model
sum(anova(mod_Q9_all)[1:3, "Df"])
```

```
## [1] 3
```

```r
# Model DF for the stratified model
anova(mod_Q9_male)["height0", "Df"] + anova(mod_Q9_female)["height0", "Df"]
```

```
## [1] 2
```

The reason why the degrees of freedom between these two approaches is different is because the interaction model has an interaction term, `height0:Male`, which takes an additional mean calculation.

F. The residual degrees of freedom of the interaction model equals to the residual degrees of freedom of male model plus the female model. **Check**

```r
# Residual DF for the interaction model
sum(anova(mod_Q9_all)["Residuals", "Df"])
```

```
## [1] 496
```

```r
# Residual DF for the stratified model
anova(mod_Q9_male)["Residuals", "Df"] + anova(mod_Q9_female)["Residuals", "Df"]
```

```
## [1] 496
```

The stratification takes $n - k - 1$ to calculate the residual degress of freedom, where k is the number of variables.

G. One difference between the two approaches is that in the interaction model we only estimate one model variance, while in the gender stratified approach, we estimate two separate model variances. **Check**

This is true.

H. One difference between the two approaches is that we now have smaller number of observations in each of the two gender specific models than in the one combined model. **Check**

This is also true.

Question 10

In the following questions, you will fit a regression model treating `birthweight` as the dependent variable and the variable that records mother's race/ethnicity as covariate using `toy_ex3.csv` (in the data folder).

Fit a regression of `birth weight` on race/ethnicity using "`Hispanic`" as reference group, and match the regression coefficients below.

Hint: You can either generate dummy variables to represent variable "`Race`" (code below) or re-level the "`Race`" variable so that the reference group is `Hispanic`:

```
dat$White <- as.numeric(dat$Race=="W")

dat$Black <- as.numeric(dat$Race=="B")

dat$Hispanic <- as.numeric(dat$Race=="H")

dat$Other <- as.numeric(dat$Race=="O")
```

Answer: Q10

```
dat <- read.csv("../data/toy_ex3.csv")
str(dat)
```

```
## 'data.frame':    1078 obs. of  2 variables:
##  $ BirthWeight: num  3769 4137 3847 3566 3309 ...
##  $ Race       : Factor w/ 4 levels "B","H","O","W": 4 4 4 4 4 4 4 4 4 4 ...
```

```
# Option 1: dummy coding
dat1 <- dat
dat1$White <- as.numeric(dat1$Race=="W")
dat1$Black <- as.numeric(dat1$Race=="B")
dat1$Hispanic <- as.numeric(dat1$Race=="H")
dat1$Other <- as.numeric(dat1$Race=="O")

# Option 2: relevel
dat2 <- dat
ref_Q10 <- c("H")
dat2$Race <- relevel(dat2$Race, ref = ref_Q10)

mod_Q10_1 <- lm(BirthWeight ~ Black + Other + White, data = dat1)
mod_Q10_2 <- lm(BirthWeight ~ Race, data = dat2)

coefficients(mod_Q10_1)
```

```
## (Intercept)       Black       Other       White
##  3277.04273   -42.31621   154.78760   120.29266
```

```
coefficients(mod_Q10_2)
```

```
## (Intercept)       RaceB       RaceO       RaceW
##  3277.04273   -42.31621   154.78760   120.29266
```

Coefficients:

White: 120.3, Black: -42.3, Other: 154.8

Question 11

Based on the regression you just ran: on average, babies born to black mothers are expected to weigh ____ grams less than Hispanic babies.

Answer: Q11

```
mod_Q11 <- mod_Q10_1
coefficients(mod_Q11)
```

```
## (Intercept)       Black       Other       White
##  3277.04273   -42.31621   154.78760   120.29266
```

Because `Hispanic` is the reference group, we set hispanic as the baseline and assume that the average weight of babies born to hispanic mothers is the intercept, 3277.04. Therefore, on average, babies born to black mothers are expected to weigh 42.31621 grams less than Hispanic babies.

Question 12

Based on the same regression result, at level 5%, there is a statistically significant difference in birth weight between Black and Hispanic babies.

Answer: Q12

```
mod_Q12 <- mod_Q10_1
summary(mod_Q12)
```

```
##
## Call:
## lm(formula = BirthWeight ~ Black + Other + White, data = dat1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1476.20  -308.72    -4.06   294.11  1457.78
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3277.04      28.80 113.795  < 2e-16 ***
## Black         -42.32      42.32  -1.000  0.31752
## Other         154.79      39.17   3.952 8.26e-05 ***
## White         120.29      39.08   3.078  0.00214 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 459.9 on 1074 degrees of freedom
## Multiple R-squared:  0.02967,    Adjusted R-squared:  0.02696
## F-statistic: 10.95 on 3 and 1074 DF,  p-value: 4.391e-07
```

```
# Check the p-value
summary(mod_Q12)$coefficients["Black", "Pr(>|t|)"]
```

```
## [1] 0.3175239
```

The p-value of `Black` in our model is higher than 0.05. Fail to reject the null at level 5%.

Question 13

On average, the birth weight of a White baby is ____ grams higher than that of a Black baby (round to one decimal place).

Answer: Q13

We can just calculate the difference between the coefficient of `White` and `Black`.

$120.293 - (-42.316) = 162.609$

END: Sample Solution - Assignment 5