STAT 500 Homework 9

Vu Thi-Hong-Ha; NetID: 851924086

November 2, 2020

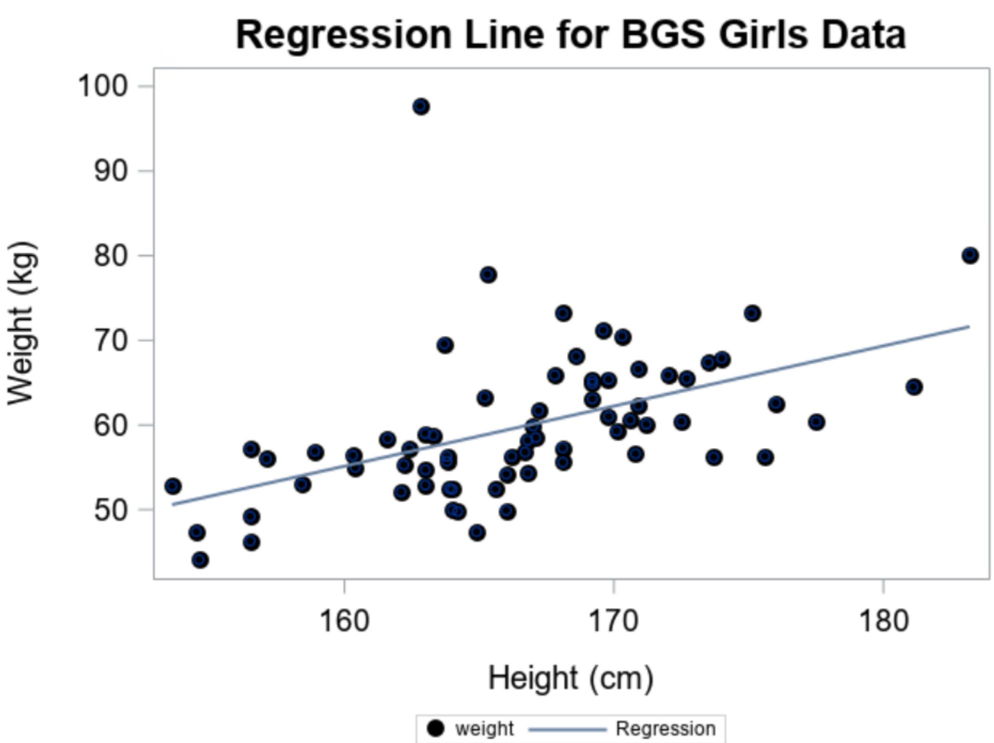
1 Question 1

(a)

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-58.48504	24.99519	-2.34	0.0222
height	1	0.71014	0.14998	4.73	<.0001

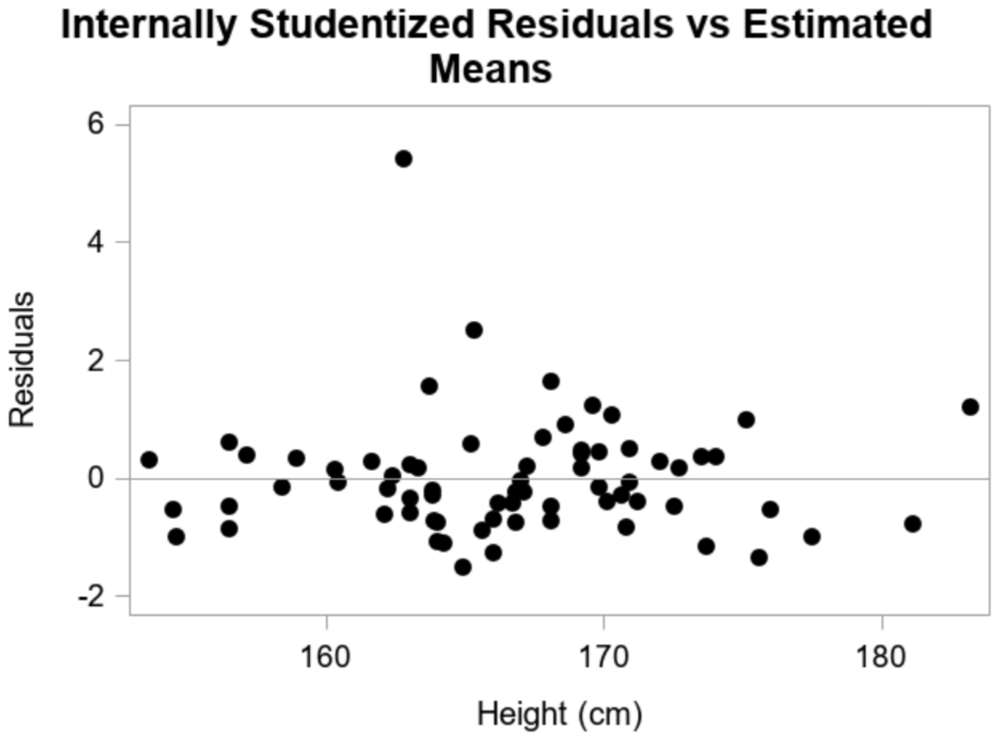
The estimate for β_0 is -58.48504 with standard error = 24.99519, and the estimate for β_1 is 0.71014 with standard error = 0.14998.

(b)



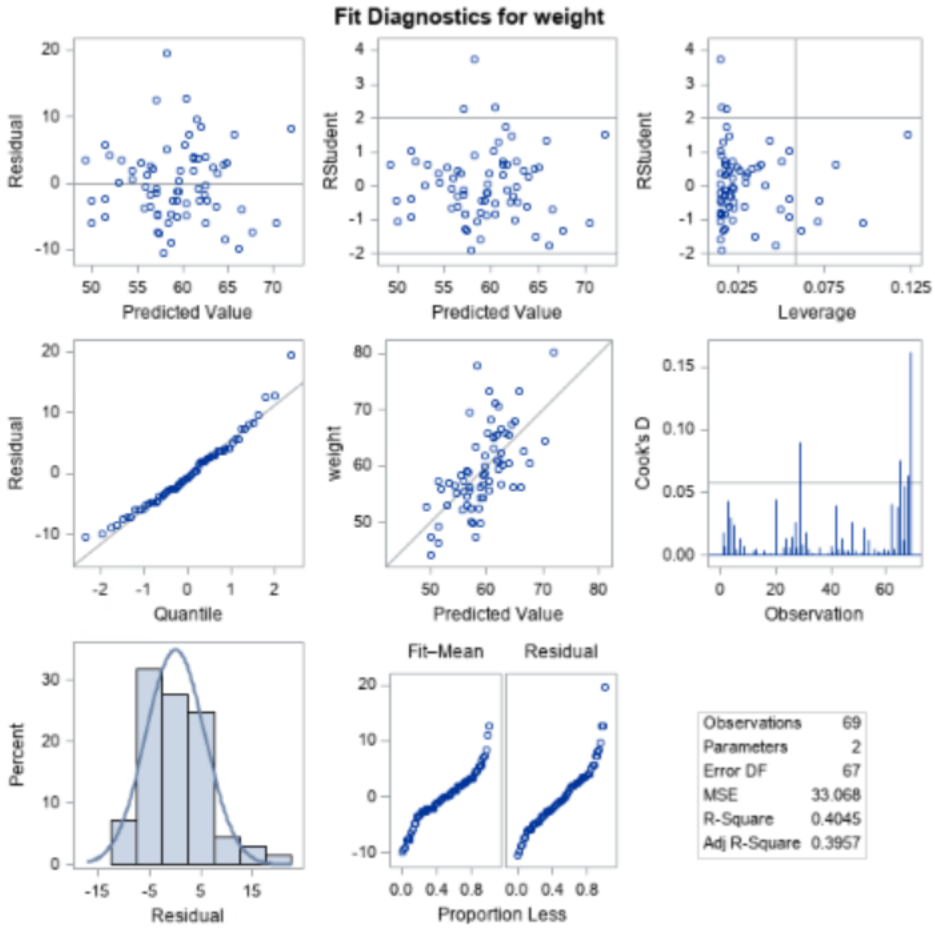
The plot suggests that there might be a linear relationship between weight and height.

(c)



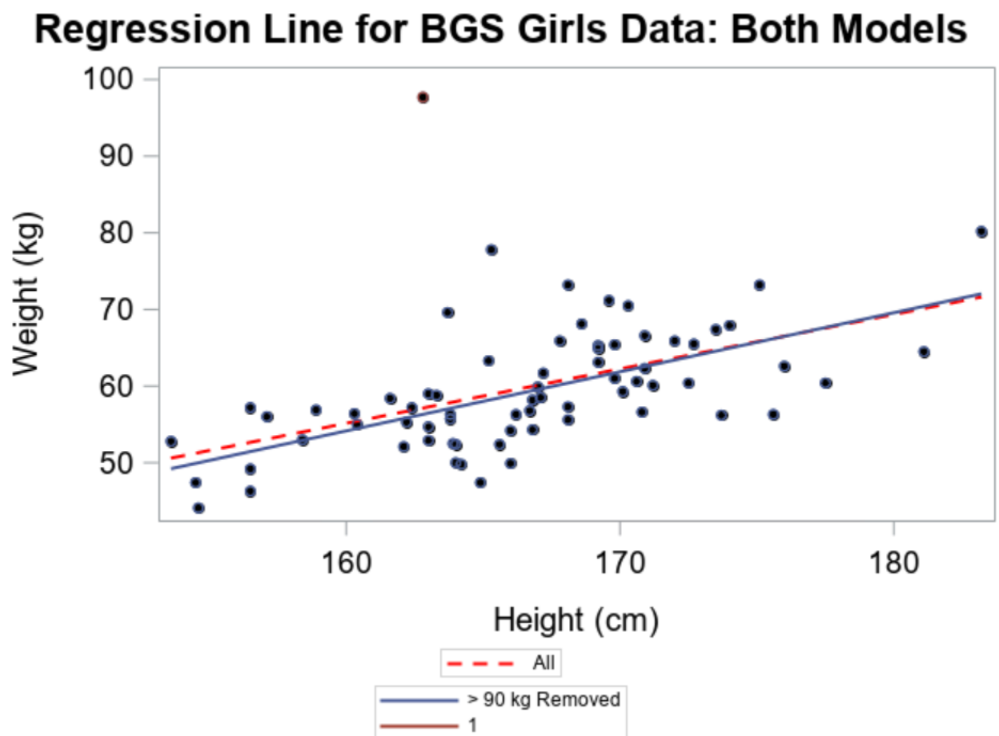
There is one data point that has absolute residual value greater than 2 and smaller than 3, so it potentially is an outlier. There is one data point that has absolute residual value greater than 4, which clearly is an outlier.

(d)



From the Leverage vs RStudent plot, we can see there are several plots that have higher leverage than the threshold. There are also three points that have studentized residual greater than 3, one of which has studentized residual greater than 3. These indicate that there are still outliers in the dataset. From the q-q plot, we can see the data seems to be right skewed. Also, with the Predicted Value vs. weight plot, it does not seem that the regression line is well fitted by the data.

(e)



The new estimates when we remove the data point greater than 90 kg is:

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-69.21339	19.05090	-3.63	0.0005	-107.23915	-31.18763
height	1	0.77100	0.11428	6.75	<.0001	0.54291	0.99910

The estimate for the intercept is affected the most if we remove the apparent outlier; it changes from -58.4850 to -69.21339.

(f)

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-145.23598	363.04838	-0.40	0.6904
height	1	1.68313	4.35132	0.39	0.7001
height2	1	-0.00273	0.01303	-0.21	0.8346

$a_0 = -145.23598$, $S_{a_0} = 363.04838$, $t = -0.40$, $p - value = 0.6904$

$a_1 = 1.68313$, $S_{a_1} = 4.35132$, $t = 0.39$, $p - value = 0.7001$

$a_2 = -0.00273$, $S_{a_2} = 0.01303$, $t = -0.21$, $p - value = 0.8346$

At a significance level of 0.05, we fail to reject the null that each regression parameter is zero. Thus, none of these tests support putting the quadratic term ino the model.

2 Question 2

(a)

Let $x_1 = 150$, $x_2 = 200$, $x_3 = 250$, $x_4 = 300$. Then $\bar{x} = 225$. Also, $\bar{Y} = \dfrac{66*6+81*6+89*6+92*6}{6*4} = 82$.

$$b_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{(x_1-\bar{x})\sum_{j=1}^{6}(Y_{1j}-\bar{Y}) + \cdots + (x_4-\bar{x})\sum_{j=1}^{6}(Y_{4j}-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{(x_1-\bar{x})\sum_{j=1}^{6}Y_{1j} - (x_1-\bar{x})*6*\bar{Y} + \cdots + (x_4-\bar{x})\sum_{j=1}^{6}Y_{4j} - (x_4-\bar{x})*6*\bar{Y}}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = 0.172$$

$b_0 = \bar{Y} - b_1\bar{x} = 43.3$

$\hat{Y_{1j}} = 43.3 + 0.172*150 = 69.1$, $\hat{Y_{2j}} = 43.3+0.172*200 = 77.7$, $\hat{Y_{3j}} = 43.3+0.172*250 = 86.3$, $\hat{Y_{4j}} = 43.3+0.172*300 = 94.9$ for $j = 1, 2, \ldots, 6$.

$SS_{error} = \sum_{j=1}^{6}(Y_{1j}-\hat{Y_{1j}})^2 + \cdots \sum_{j=1}^{6}(Y_{4j}-\hat{Y_{4j}})^2$.

Consider $\sum_{j=1}^{6}(Y_{1j}-\hat{Y_{1j}})^2$:

$\sum_{j=1}^{6}(Y_{1j}-\hat{Y_{1j}})^2 = \sum_{j=1}^{6}Y_{1j}^2 - 2\hat{Y_{1j}}\sum_{j=1}^{6}Y_{1j} + 6\hat{Y_{1j}}^2$, in which

$\sum_{j=1}^{6}Y_{1j}^2 = (n_1-1)Var_1 - n_1\bar{Y_1}^2 + 2\bar{Y_1}\sum_{j=1}^{6}Y_{1j} = (6-1)*1.15 - 6*66^2 + 2*66*(66*6) = 26141.75$

$2\hat{Y_{1j}}\sum_{j=1}^{6}Y_{1j} = 2*69.1*(66*6) = 54727.2$

$6\hat{Y_{1j}}^2 = 6*69.1^2 = 28648.86$

Then, $\sum_{j=1}^{6}(Y_{1j}-\hat{Y_{1j}})^2 = 26141.75 - 54727.2 + 28648.86 = 63.41$

Thus, $SS_{error} = 63.41 + 70.34 + 50.49 + 54.96 = 239.2$.

Thus, $MSE_{error} = \hat{\sigma}^2 = \dfrac{239.2}{24-2} = 10.87273$

Hence,

$$Var(b_0) = \hat{\sigma}^2 * \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) = 10.87273 * \left(\frac{1}{24} + \frac{225^2}{75000}\right) = 7.792124$$

$$Var(b_1) = \hat{\sigma}^2 * \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{10.87273}{75000} = 0.0001449697$$

$SE(b_0) = \sqrt{7.792124} = 2.791438$, and $SE(b_1) = \sqrt{0.0001449697} = 0.01204034$

(b)

$$SS_{regression} = \sum_{i=1}^{r=4} n_i(\hat{Y_i} - \bar{Y_{..}})^2 = 6*(69.1-82)^2 + 6*(77.7-82)^2 + 6*(86.3-82)^2 + 6*(94.9-82)^2 = 2218.8$$

$$SS_{lack-of-fit} = \sum_{i=1}^{r=4} n_i(\hat{Y_i} - Y_{i.})^2 = 6*(69.1-66)^2 + 6*(77.7-81)^2 + 6*(86.3-89)^2 + 6*(94.9-92)^2 = 217.2$$

$$SS_{pure-error} = \sum_{i=1}^{4}\sum_{j=1}^{6}(Y_{ij} - \bar{Y_{i.}})^2 = \sum_{j=1}^{6}(Y_{1j} - \bar{Y_{1j}})^2 + \cdots \sum_{j=1}^{6}(Y_{4j} - \bar{Y_{4j}})^2 = (n_1 - 1)Var_1 + \ldots (n_4 - 1)Var_4 = 22$$

Source of variation	D.f	SS	MS
Regression on X	1	2218.8	2218.8
Lack of fit	$r - 2 = 2$	217.2	108.6
Pure error	$n - r = 20$	22	1.1
Corrected Total	$n - 1 = 23$	2458	

(d) Lack of fit test:

Test statistic $F = \dfrac{108.6}{1.1} = 98.72727$

$F_{(2,20),0.95} = 3.49283$

$p - value < .0001$

At a significance level of 0.05, we reject the null that $E(Y_{ij}|x_i) = \beta_0 + \beta_1 x_i$.

(d) We can use quadratic regression instead of simple linear regression.