

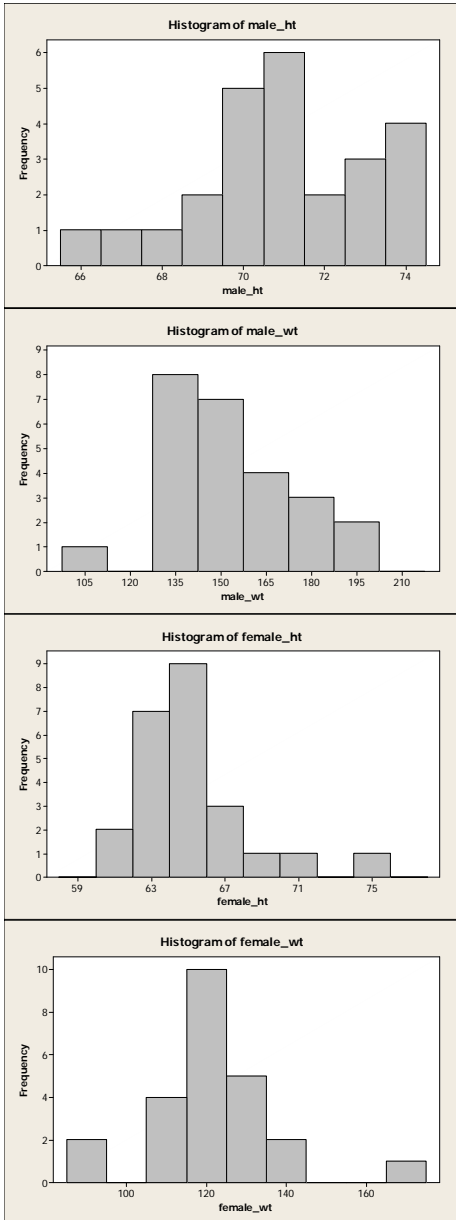
Chapter 9: Linear Regression

Data Set:

- Heights and Weights of Male and Female High School Students
- <http://www.canyons.edu/faculty/morrowa/140/datasets/>

1. Normally Distributed?

a) Create histograms for each variable. Describe the histograms.



Male heights appear to be bimodal. There is a skew to the left. There do not appear to be any outliers or gaps.

I adjusted this histogram to have 8 bins. It's okay if you left it as is, but your description needs to match your graph. Male weights are unimodal and skew to the right. There is a gap separating a potential outlier on the low end.

Note: I adjusted so that there were 10 bins. Female heights are slightly skewed to the right. There is a gap on the high end, separating a potential outlier at 75. Data is unimodal.

Female weights are unimodal and nearly symmetric. There are gaps and outliers on the high and low ends.

b) Calculate summary statistics.

Variable	Mean	StDev	Median	IQR
male_ht	70.960	2.169	71.000	3.000
male_wt	153.56	21.01	150.00	29.00
female_ht	64.729	3.210	64.000	2.875
female_wt	120.58	15.93	120.00	17.00

c) Is it reasonable to say that these data are drawn from a populations that are normally distributed?
Females: Yes... The histograms are unimodal and roughly symmetric.
Males: It's possible, but not as clear.

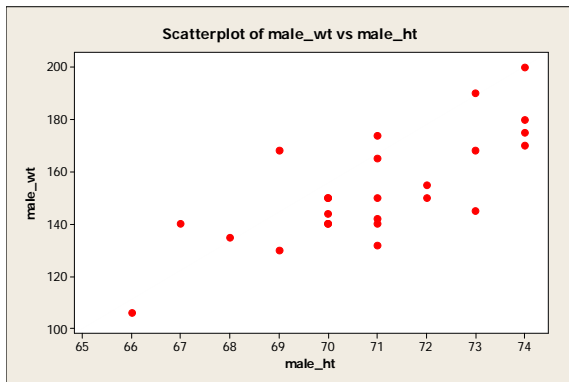
2. Correlation

- a) Make a scatterplot for each gender. Describe the scatterplots. *Note: There is a best choice for the explanatory and response variables. Check your answer before moving on.*

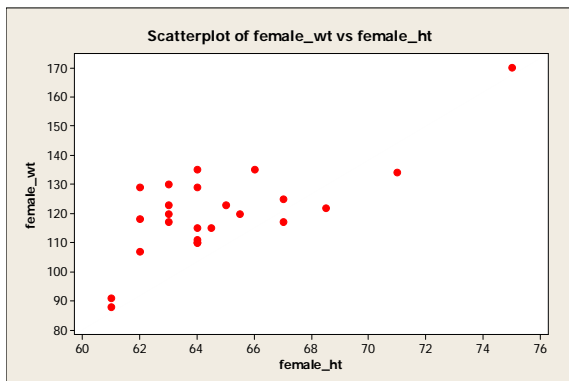
EXPLANATORY: HEIGHT

RESPONSE: WEIGHT

Height determines weight. Also, people are less likely to lie about their heights, and so we might use their heights to estimate weights.



The scatterplot for male heights and weights shows a moderate positive linear relationship. Taller males are generally heavier. There do not appear to be any outliers or non-constant variance.



Likewise, there is a moderate, positive linear relationship between female heights and female weights. Taller females are generally heavier. There is one female who is much taller than the others (a potential outlier), but this female's weight fit the overall pattern.

Note: The female outlier is called a point of high leverage, but does not appear to be an influential point.

- b) If it is appropriate to do so, calculate and interpret r for each gender.

Since we suspect a linear relationship, it is appropriate.

Pearson correlation of male_ht and male_wt = 0.754

Pearson correlation of female_ht and female_wt = 0.738

- c) What lurking variables might influence one's weight?

Some possibilities: Eating disorders, genetics, eating habits, poverty level, ...

Note: Anything you list here that might affect one's weight is fine.

3. Regression

- a) Find the least squares line of best fit for each gender.

MALES:

$$\text{male_wt} = -364 + 7.30 \text{ male_ht}$$

FEMALES:

$$\text{female_wt} = -117 + 3.66 \text{ female_ht}$$

- b) Explain what the slope means in the context of this problem.

MALES: For each inch taller a male is, our model predicts that he will be 7.30 pounds heavier.

FEMALES: For each inch taller a female is, our model predicts that she will be 3.66 pounds heavier.

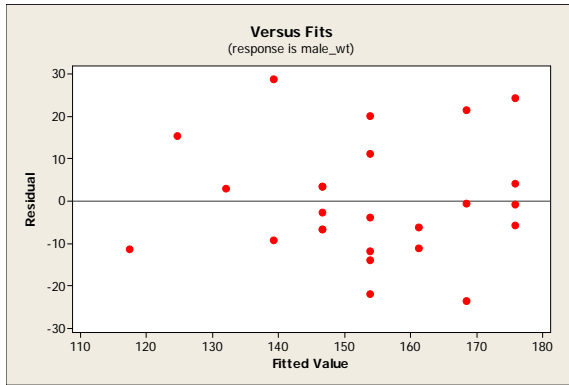
- c) Explain what the intercept means in the context of this problem.

MALES/FEMALES: There is no interpretation possible because it is not possible to have a male/female who is 0" tall.

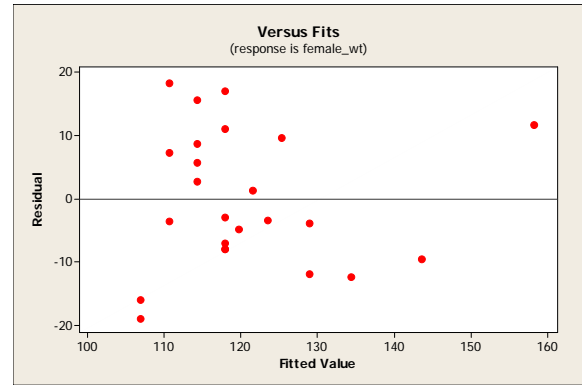
4. Residuals

a) Provide residual plots for each gender.

MALE



FEMALE



b) Do you think the line is a good model for each? Why or why not?

Both residual plots appear reasonably scattered (there is no significant pattern in either). The linear model appears appropriate for both.

It is important to note that we might improve the female model by removing the outlier (the heaviest female).

c) What does a positive residual mean?

A positive residual means that the individual weighs more than was predicted by the model, given their height.

d) What is the residual for the female who was 71" tall in the study?

Looking back at the data, we can see that she actually weighed 134 pounds. To determine her residual, we first must calculate her predicted weight: 143.56 pounds. The residual is calculated as (observed – expected) = $134 - 143.56 = -9.56$ pounds.

5. Predictions. Predict weights for each of the following:

- a) A 60" Male: 60" is outside the range of given male heights, so our prediction may not be accurate. If we were to calculate it, it would be 73.56 pounds.
- b) A 70" Male: 146.55 pounds
- c) A 7'2" Male: 86" is outside the range of given male heights, so our prediction may not be accurate. If we were to calculate it, it would be 263.35 pounds
- d) 20" Newborn baby boy: Prediction is not possible. Our model was not created using baby data.

6. R^2

a) Calculate R^2 for each gender.

MALE: 56.8%

FEMALE: 54.5%

b) Provide an interpretation of R^2 in the context of this problem.

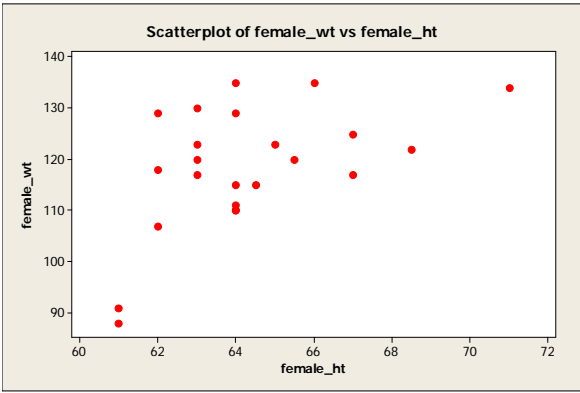
For males, 56.8% of the variation in weights is explained by their linear relationship with height. For females, 54.5% of the variation in weights is explained by their linear relationship with height.

c) Using R^2 , comment on the strength of the linear relationship for each gender.

R^2 for each model is above 50%. This indicates that the strength of the linear model is weak to moderate.

7. Outliers

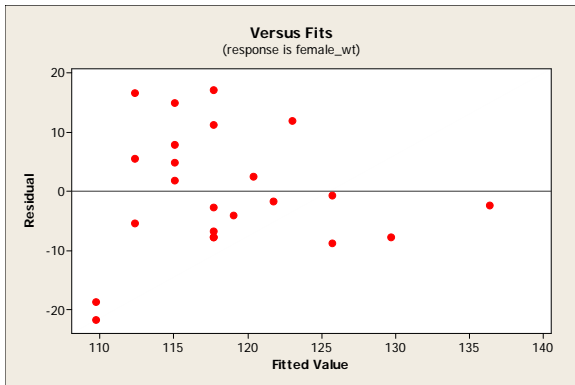
a) The 75" female is a potential outlier. Remove the result and redo #2-4. What changes?



The linear pattern in the scatterplot appears weaker. If we imagine the lowest 2 weights and the highest weight did not exist, there might not be much of a linear pattern.

The value of r reflects this as well, as it has decreased from the original value: 0.524.

The regression equation is
 $female_wt = - 53.1 + 2.67 female_ht$



The residual plot does not show any clear pattern, which would indicate that the linear model is appropriate.

There is cause for concern, however, as fitted values above 125 appear to only have negative residuals.