AP Statistics

LINEAR REGRESSION
CHAP 7

The object [of statistics] is to discover methods of condensing information concerning large groups of allied facts into brief and compendious expressions suitable for discussion.

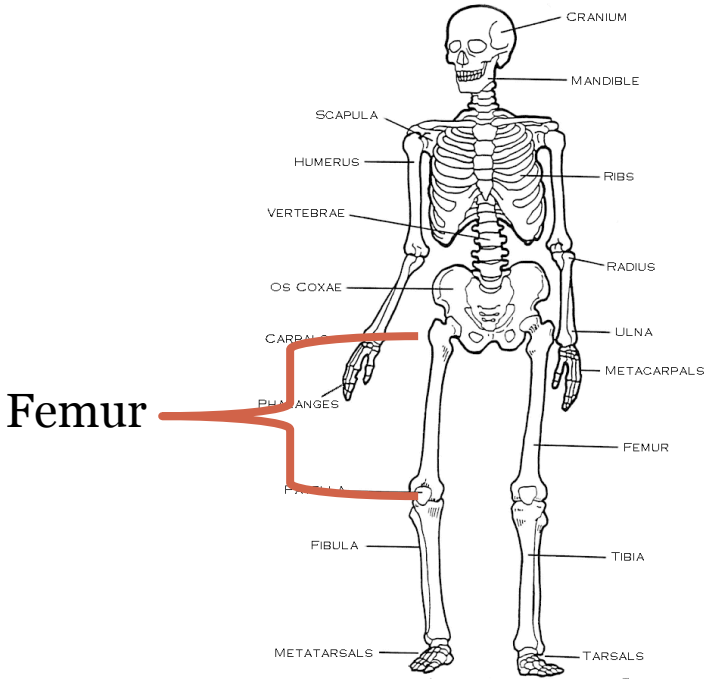
Francis Galton (1822-1911)

Relationship between 2 Quantitative Variables

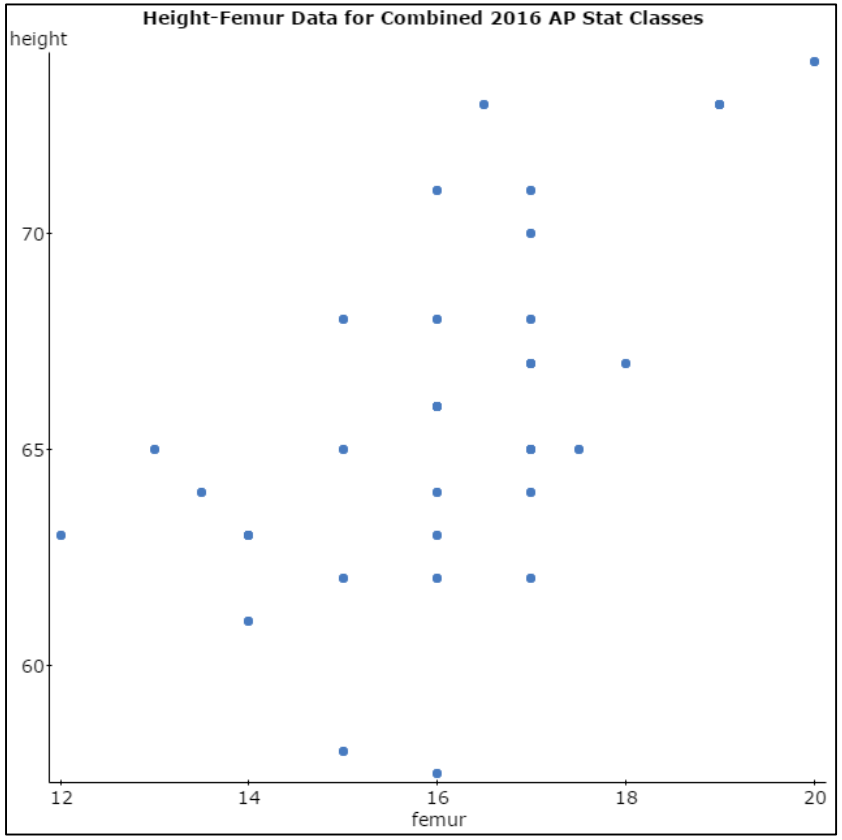
In the mid-20th century Dr. Mildred Trotter, a forensic anthropologist, determined relationships between dimensions of various bones and a person's height. The relationships she found are still in use today in the effort to identify missing persons based on skeletal remains (think C.S.I) .

One relationship compares the length of the femur to the person's height.

Measure the length of your femur along the outside of your leg (in inches). Also measure your height (also in inches) if you don't already know it. Record both at the front of the room.



Scatterplot of Height vs. Femur



Femur length is the **predictor** variable and height is the **response** variable.

Association between variables:

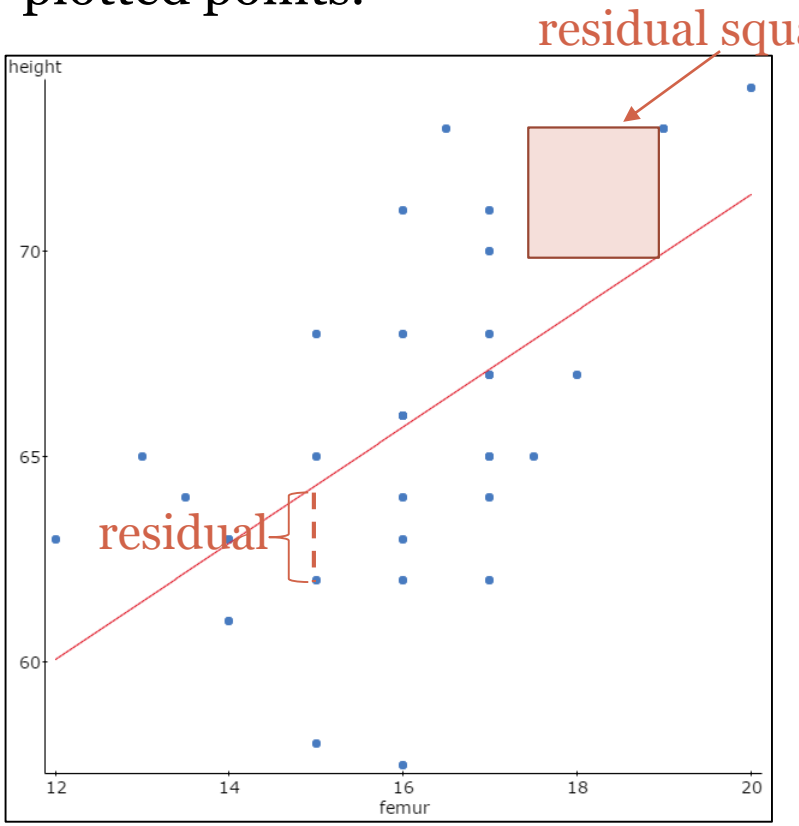
Direction – positive

Form – linear

Strength – moderate (correlation = 0.591)

Linear Model

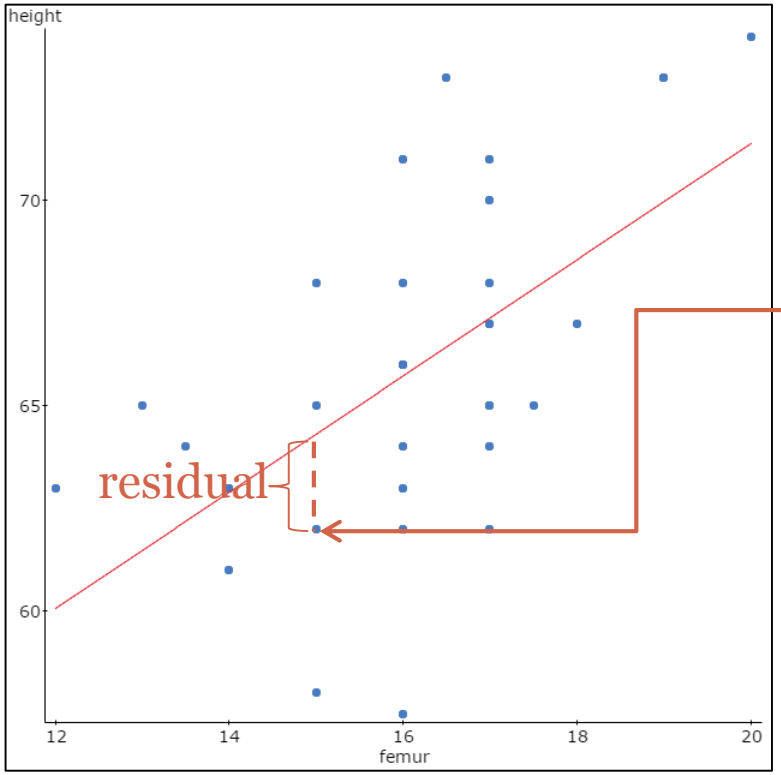
A **linear model** is a line that passes "through the middle" of the plotted points.



The line we will create is called the **"Least Squares Regression Line."** It minimizes something called the **squared residuals** (**residuals** are also called **errors**).

Linear Model

A **residual** is the difference between the observed response and what the model predicts for the response:



$$\text{residual} = \text{observed } y - \text{predicted } y$$

$$\text{residual} = y - \hat{y}$$ ← "y-hat"

Example: The indicated point is for a femur length of 15 in. The **observed height** is 62 in., and the model **predicts a height** of 64.4 in. The residual is:

$$\text{residual} = 62 - 64.4$$

$$= -2.4$$

The model over-predicts height for this data point by 2.4 in.

Least Squares Regression Equation

The form of the **linear model** is:

$$\hat{y} = b_0 + b_1 x$$

$$b_0 = \bar{y} - b_1 \bar{x} \qquad b_1 = \frac{r s_y}{s_x}$$

Note 1 – These formulas are on your AP Formula Sheet

Note 2 – See optional Math Box on p. 180 to "look under the hood"

Least Squares Regression Equation

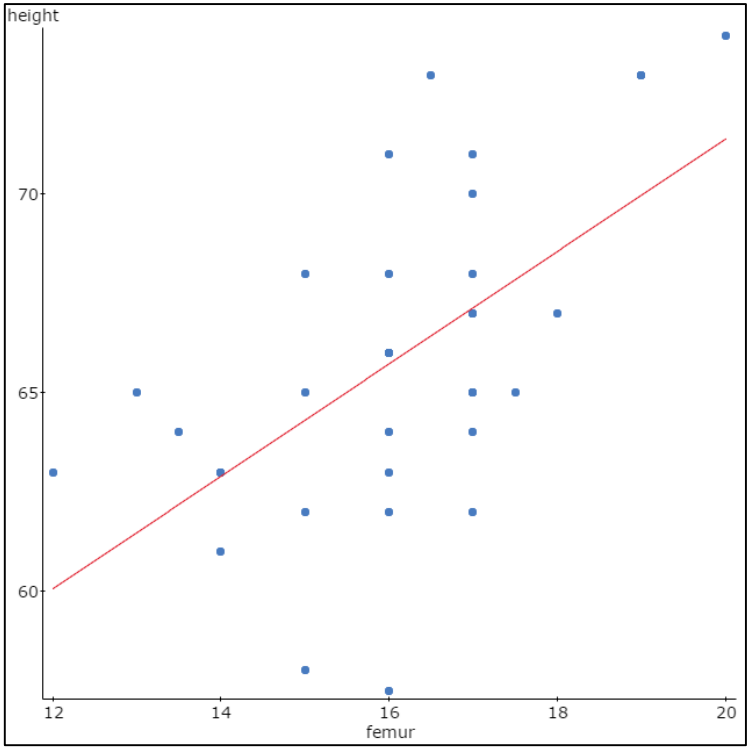
If we standardized the data, the standard deviation s would be 1:

$$b_1 = \frac{rs_y}{s_x} = r\frac{1}{1} = r$$

Since r can never be larger than 1, the numerator indicates that the predicted y tends to be closer to the mean (in standard deviations) the its corresponding x. This property is called **regression to the mean**.

So, when x increases by 1SD, y increases by only r times 1SD. This is called *regression to the mean*.

Back to Our Data …



Variable	Mean	Std. dev.
femur	16.10	1.71
height	65.87	4.10

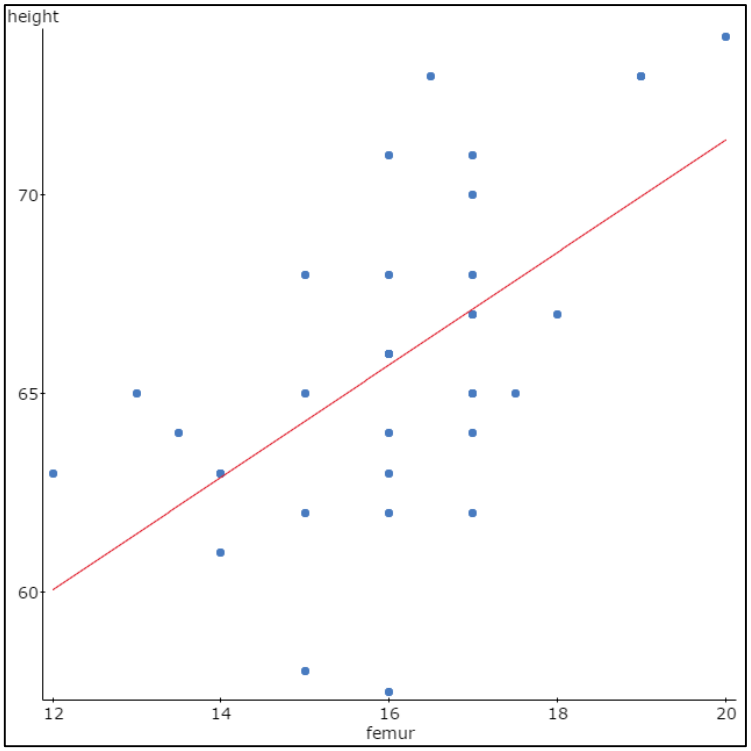
$r = 0.591$

Compute the Least Squares Regression Line Model:

$$b_1 = \frac{rs_y}{s_x} \qquad b_0 = \bar{y} - b_1\bar{x}$$

$$\hat{y} = b_0 + b_1 x$$

Back to Our Data …



Calculate the Least Squares Regression Model using your calculator (see p. 193 for TI Tips).

Regression on the Computer

Simple linear regression results:
Dependent Variable: height
Independent Variable: femur
height = 43.06999 + 1.4157449 femur
Sample size: 34
R (correlation coefficient) = 0.59071902
R-sq = 0.34894896
Estimate of error standard deviation: 3.3562315

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	43.06999	5.5348135	≠ 0	32	7.7816516	<0.0001
Slope	1.4157449	0.3418507	≠ 0	32	4.1414119	0.0002

Interpret the Slope and Intercept

$$\text{height} = 43.07 + 1.42(\text{femur})$$

Predicted height if femur length is 0 inches!

For every 1-inch increase in femur length we expect an increase of about 1.42 inches in height.

Make Predictions

$$\text{height} = 43.07 + 1.42(\text{femur})$$

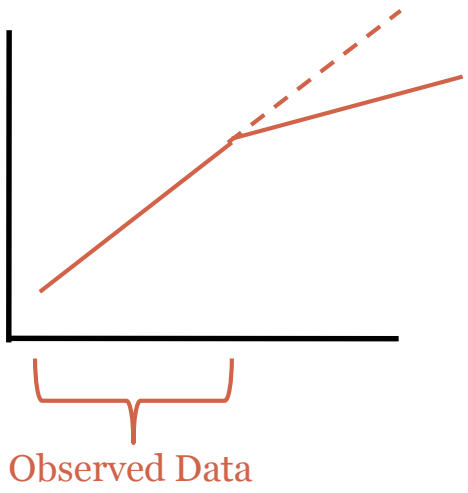
Predict a person's height if the femur length is:

1. 18.5 inches
2. 25 inches

Make Predictions

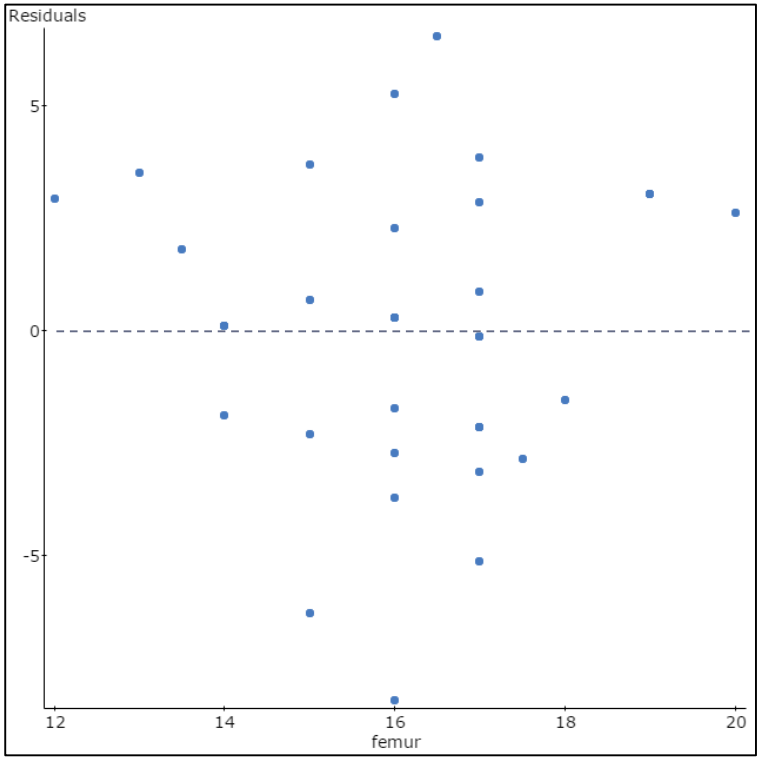
A prediction made outside the range of the predictor (explanatory) data is called **extrapolation**.

Proceed with caution. The model may not hold true beyond the observed data.



How Good Is The Model?

Make a **residual plot**. If there is no apparent pattern to the residuals, then the Least Squares Regression (LSR) Model is reasonable.

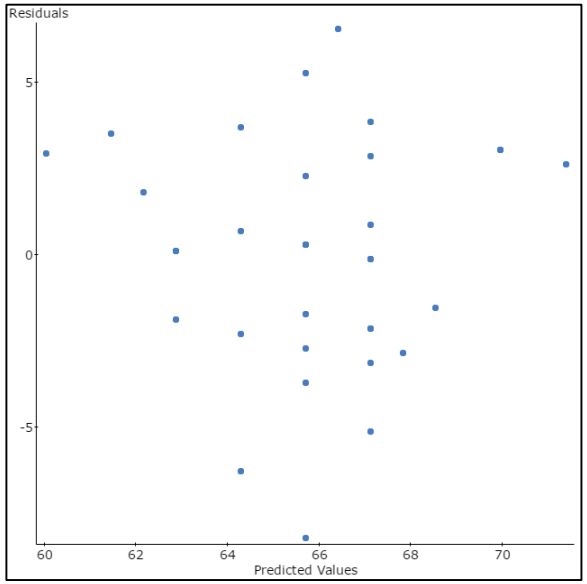


No pattern means the model captured all of the "interesting" relationship between the variables.

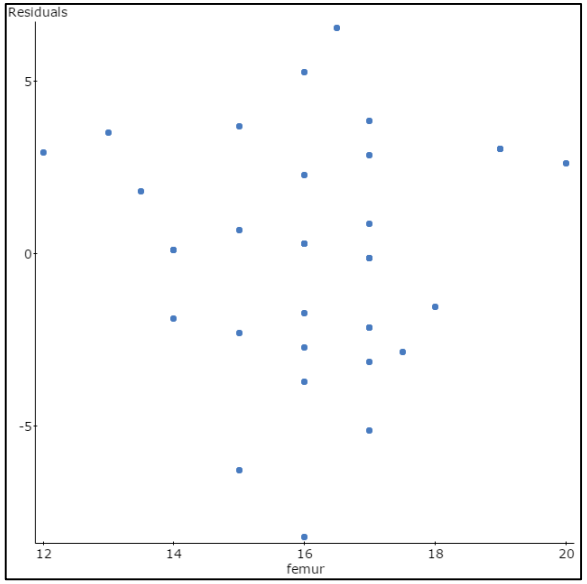
The mean of the residuals will always be 0, and the smaller the residuals from the model, the closer the model fits the data.

How Good Is The Model?

Software usually plots the residuals against the predicted values (heights in this case). When there is only one explanatory variable, the result is the same when the residuals are plotted against the explanatory variable (other than scale).



Residuals vs. Predicted Values



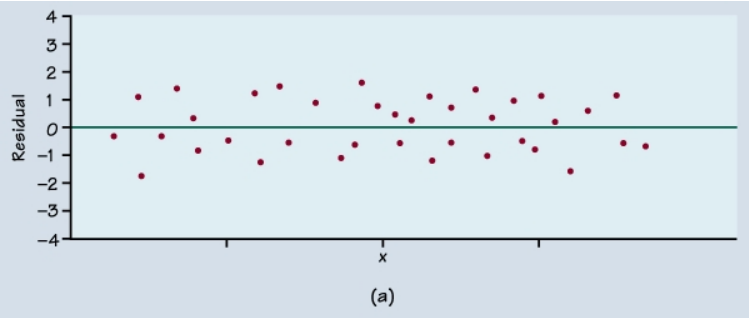
Residuals vs. Explanatory Variable

How Good Is The Model?

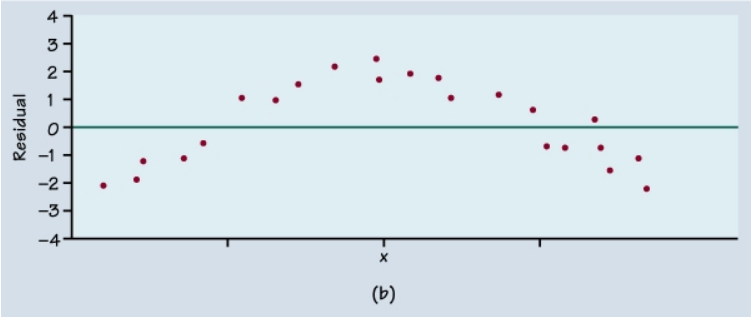
See TI Tips on p. 193 for instructions on creating a residual plot on your calculator.

Note – Every time you have your calculator create a LSR Model, a residuals list (RESID) is created. ***Every time****. The most common mistake made by AP Stat students with regards to this is to forget to rerun the LSR command when working with a new set of data.*

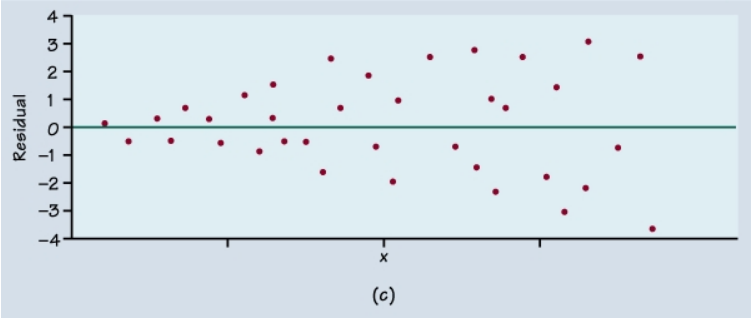
How Good Is The Model?



No left-over underlying pattern, so the linear model is reasonable.



Curved pattern left-over after regression, so linear model did not capture entire variable relationship.



Residuals increase in size with an increase in *x*, so linear model did not capture entire variable relationship.

After Deciding On The Model …

A statistic we will call “R-squared*” tells us something about how well the model captures the relationship in the variability present in the association between the predictor and response variables.

It goes something like this …

*It has a technical name, but no one really ever uses it.

R-squared

What if I told you a new student was joining the class next week, and I want you to predict his/her height? Further, the only model you have is based on the mean height of students in the AP Stats class (y-bar = 65.9 in.).

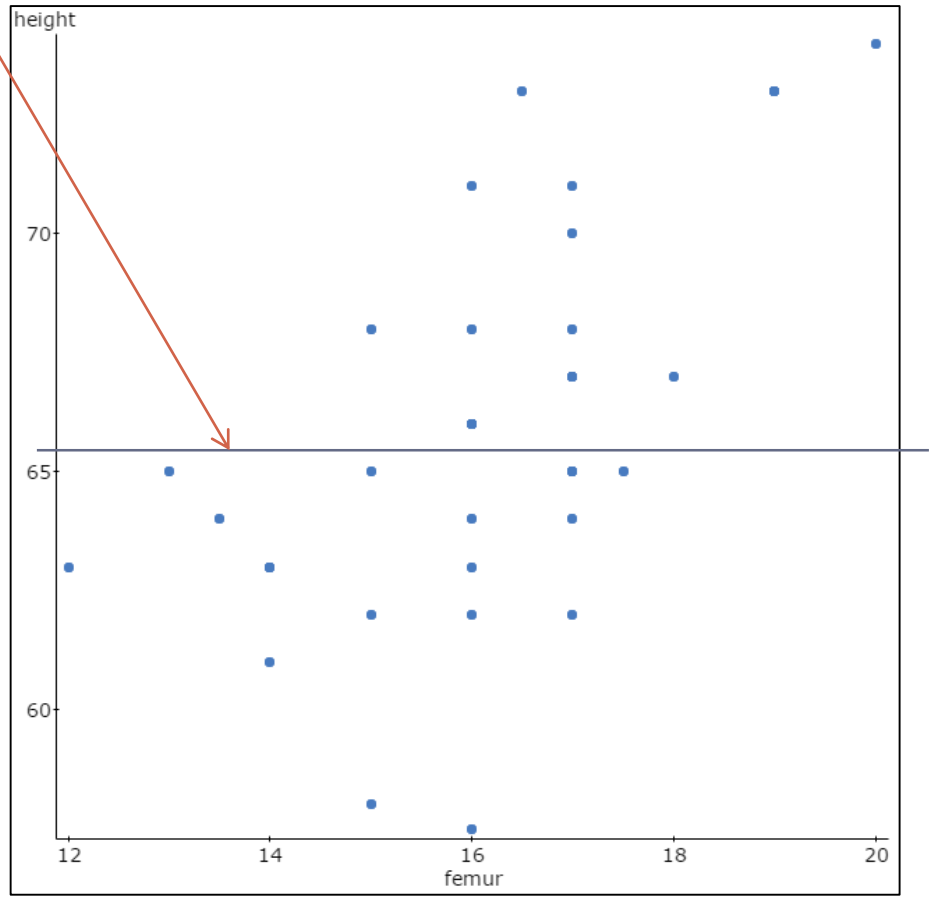
What height is your best guess?

That's right, the mean height of 65.9 in.

R-squared

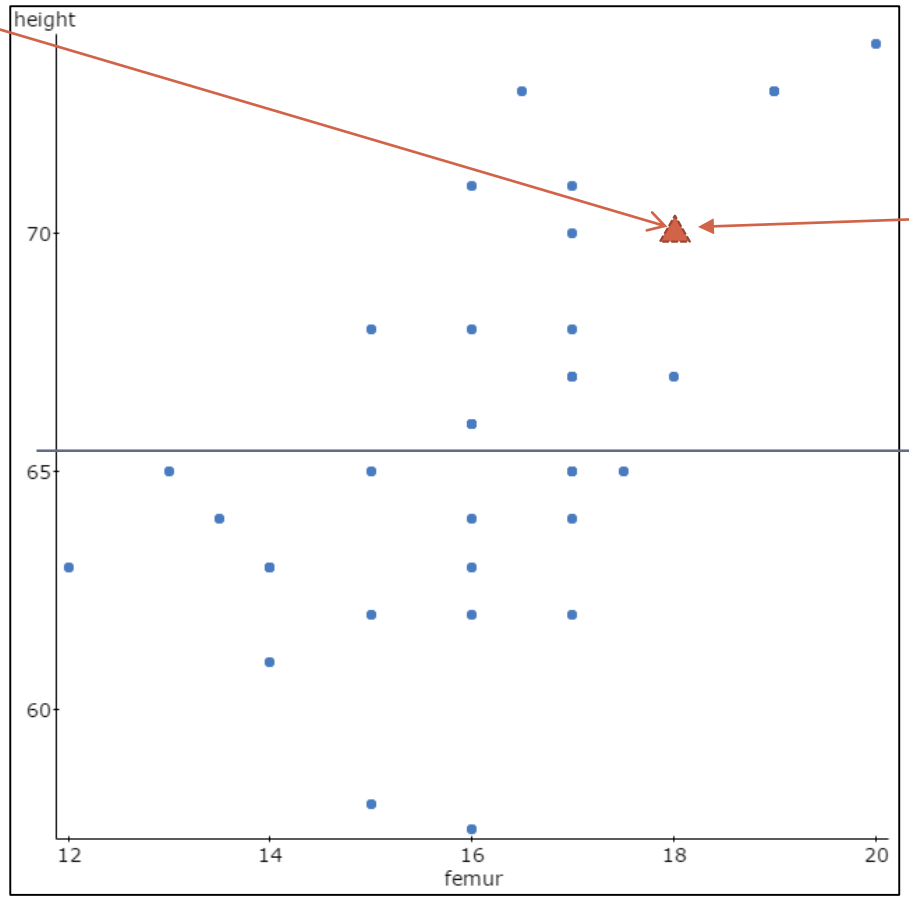
The “mean height” model:

$\hat{y} = 65.9$



R-squared

Here's the new student …

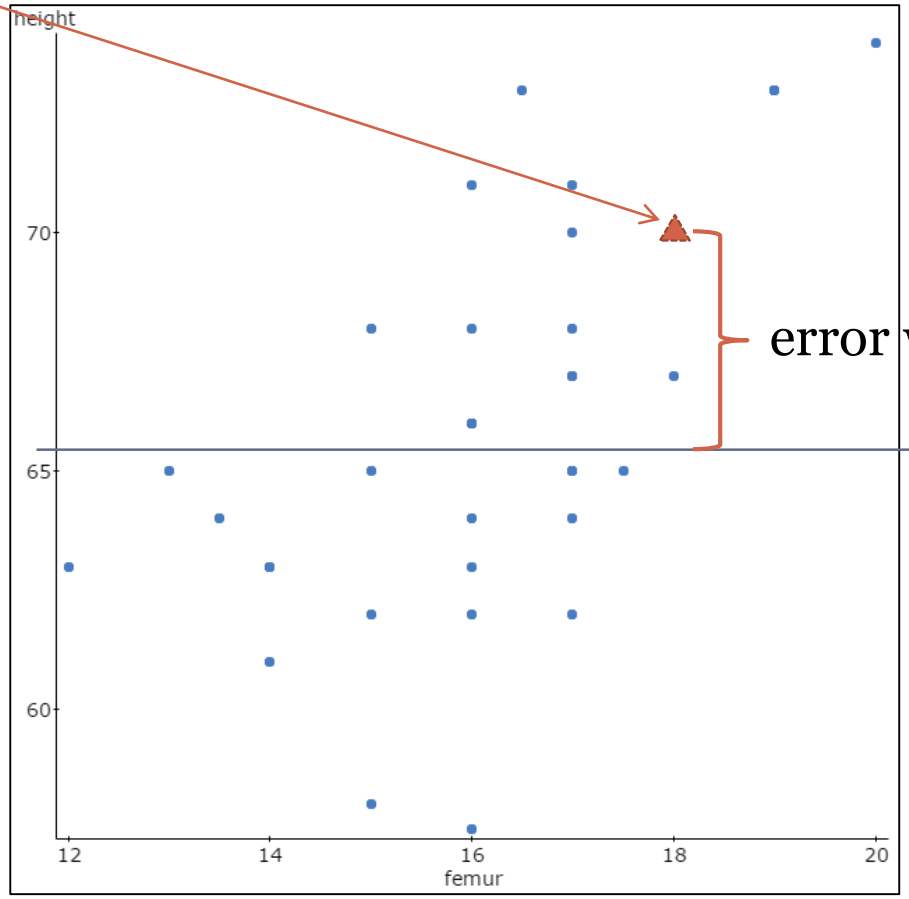


Femur = 18 in
Height = 70in

$\hat{y} = 65.9$

R-squared

Here's the new student …



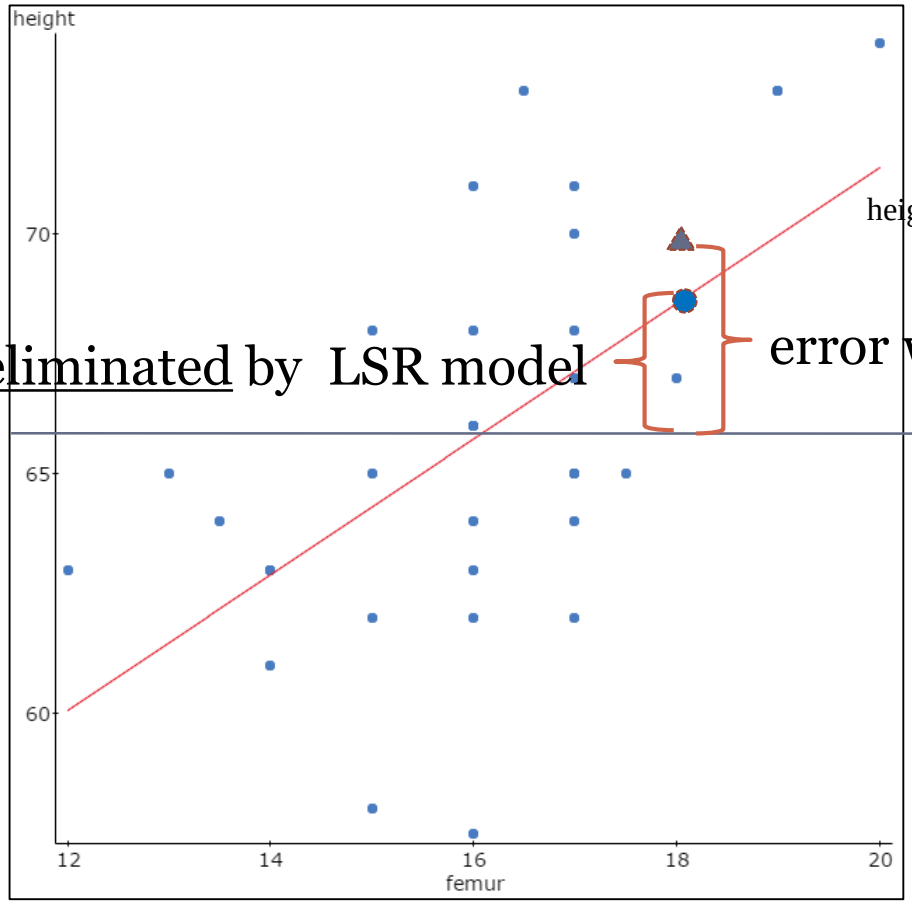
$\hat{y} = 65.9$

error wrt mean model

The mean height is not a satisfactory guess … we need more information.

R-squared

Now we know the student's femur length to be 18 in.



error eliminated by LSR model

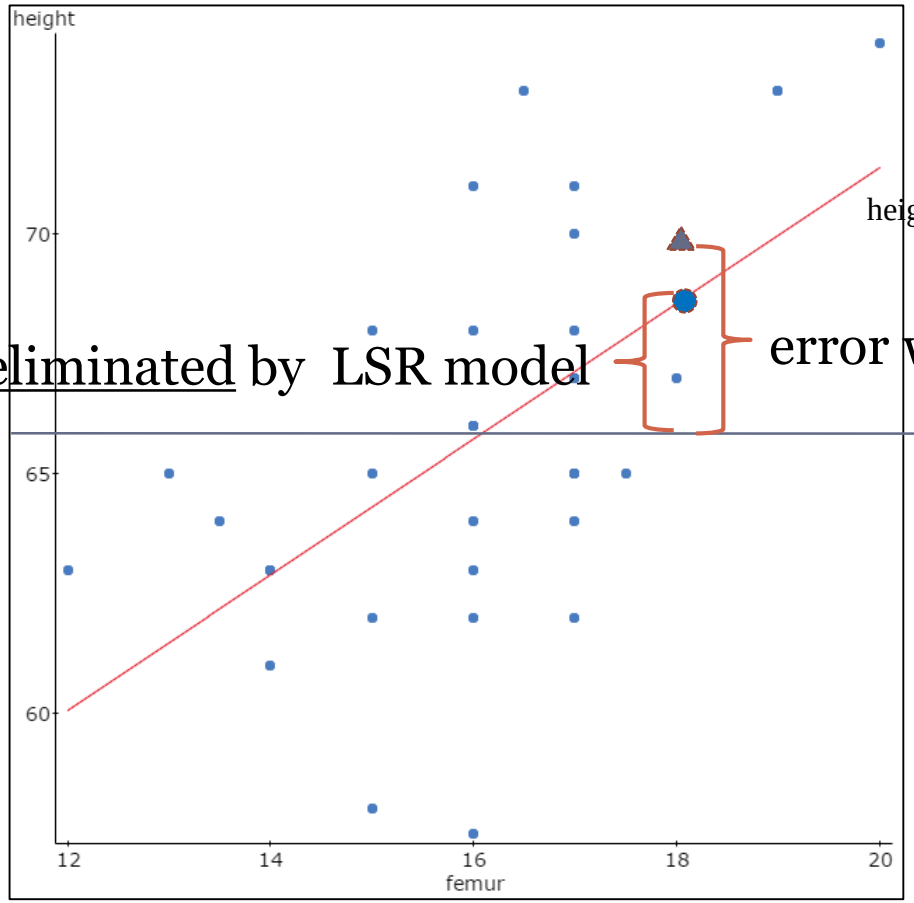
error wrt mean model

$\hat{y} = 65.9$

A better estimate is based on the LSR Model … 68.63 in.

R-squared

The ratios of these two errors (summed for all data points) is *r*-squared.



R-squared

R-squared is *the fraction of variation in height accounted for by the regression model with femur length*. Specifically …

… About 34.9% of the variation in height can be attributed to the linear regression model with femur length.

R-squared

R-squared will always be between 0% and 100%.

R-squared describes the percent of variation in the response variable that can be explained (or attributed to) the regression with the explanatory variable.

In the case of heights predicted from femur lengths, if only 34.9% of the observed variation of heights can be attributed to differences in femur length, there must be other factors that contribute to differences in heights.

What might some of these factors be?

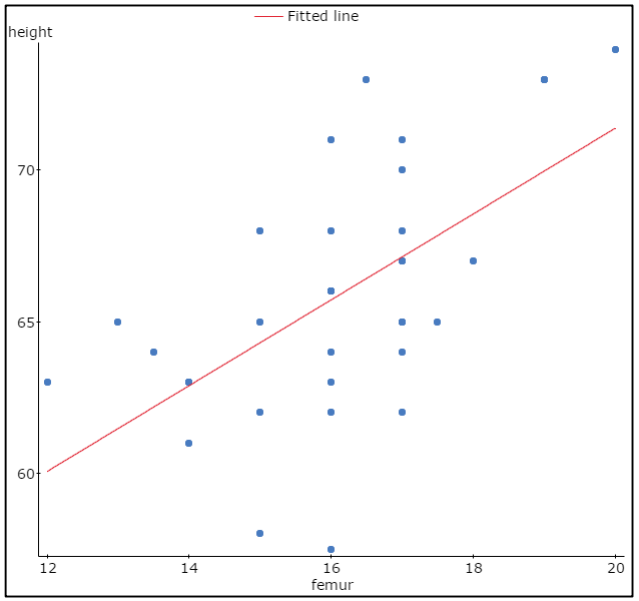
r vs. R-squared

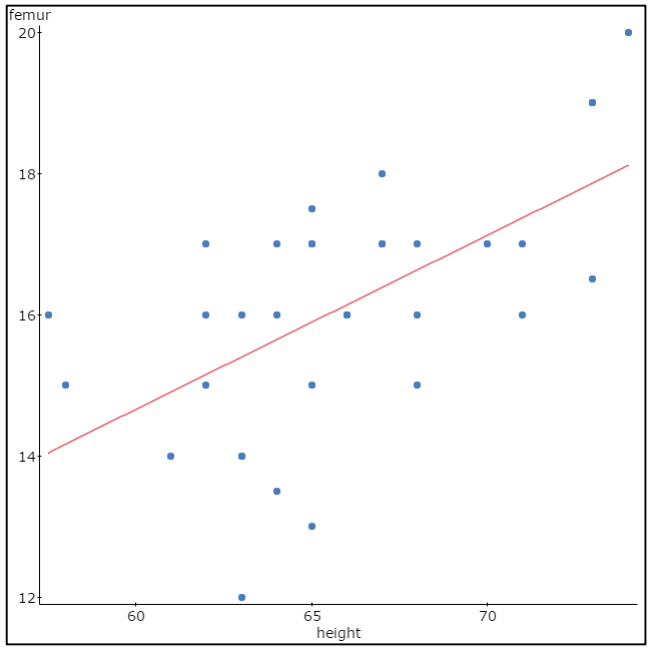
Correlation *r* describes the strength and direction of a linear association between two quantitative variables.

R-squared describes the amount of variation observed in the response variable accounted for through the linear model with the predictor variable.

What if the Variables are Reversed?

Because the Least Squares Regression Line minimizes the vertical distances to the observed data, you can't ever (no, not ever … don't even think it!) use a regression equation to predict the explanatory variable from the response variable.





$$\text{height} = 43.07 + 1.42(\text{femur})$$

$$\text{femur} = -0.13 + 0.25(\text{height})$$

Summary

- If the scatterplot doesn't look straight, don't fit a line!
- Watch out for outliers (in all directions).
- Do use a residual plot to see if the linear model looks reasonable.
- Don't use R-squared to determine if a linear model is reasonable. R-squared only tells you something about the amount of variation in the response variable that can be traced back to the a particular predictor variable.
- Association does not imply causation. A regression model with high r and R-squared in no way means the predictor variable causes the response variable.
- Extrapolate with caution.

Assignment

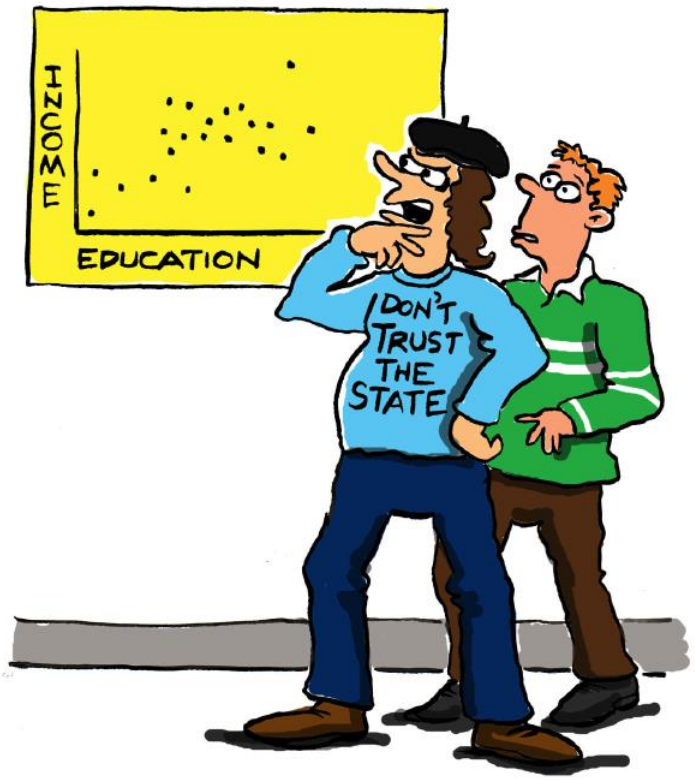
Read Chapter 7

Assignment 1:

Do Ch 7 exercises #1-11 odd, 17,

Assignment 2:

Do Ch 7 exercises # 19-31 odd, 37-41 odd, 45, 53, 57, 59



"I think the Government is behind this plot.

www.causeweb.org
John Landers