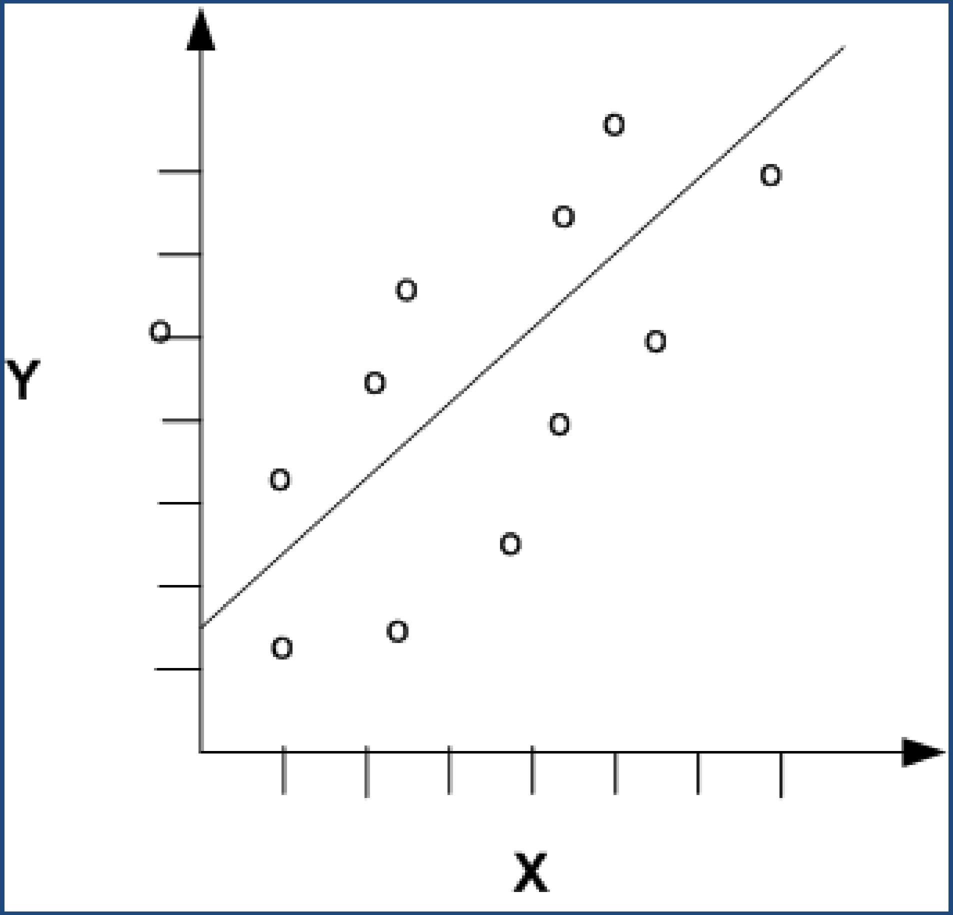
Part II – Exploring Relationships Between Variables

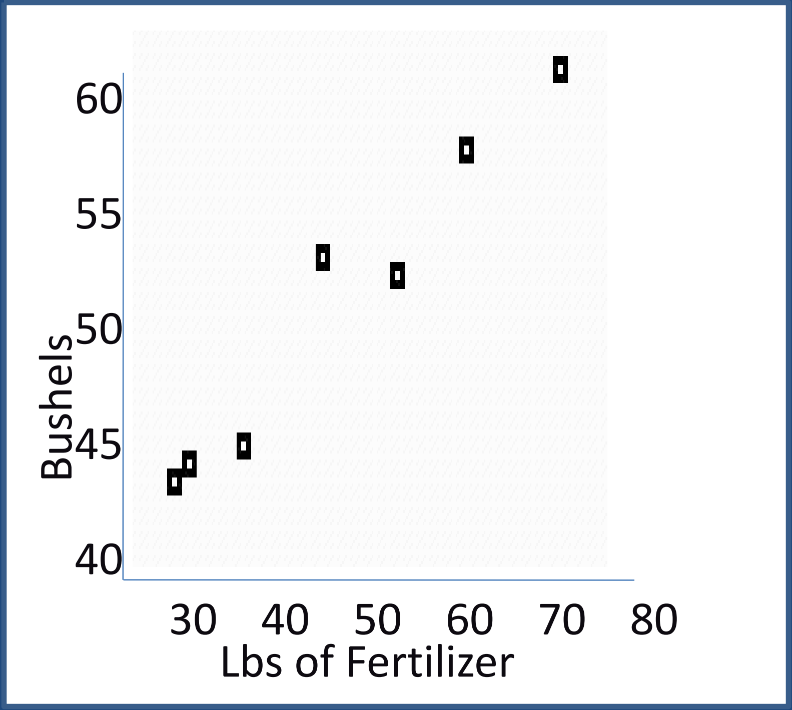
Ch. 8 – Linear Regression (Day 1)



The Linear Model

- In the last chapter, we learned to draw a picture (scatterplot) of the relationship between two variables, and to measure the strength and direction of that relationship (correlation)

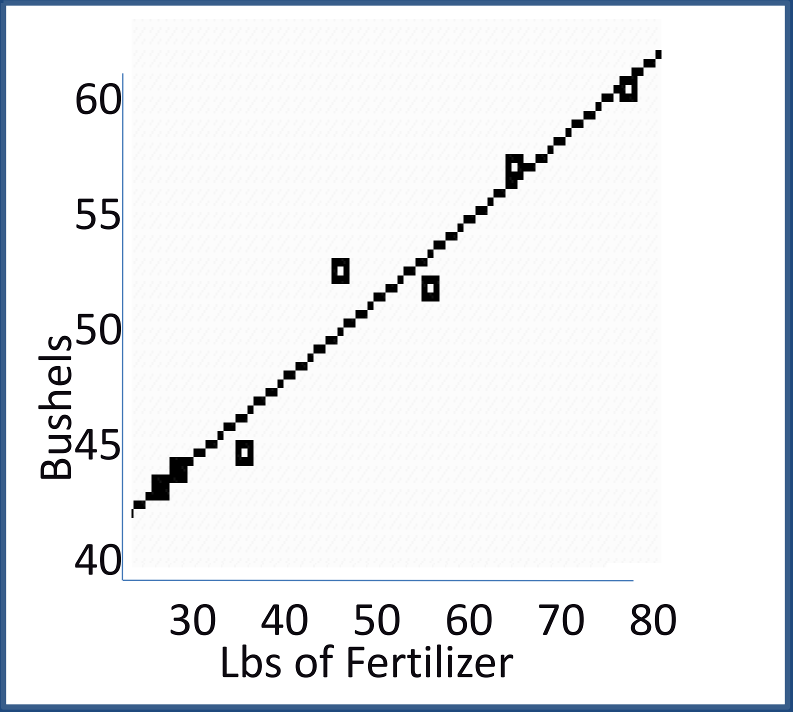
- The next step is to use what we learn from this analysis to make predictions about the variables



r = .9782 There is a strong, positive linear relationship between lbs of fertilizer and bushels of grain

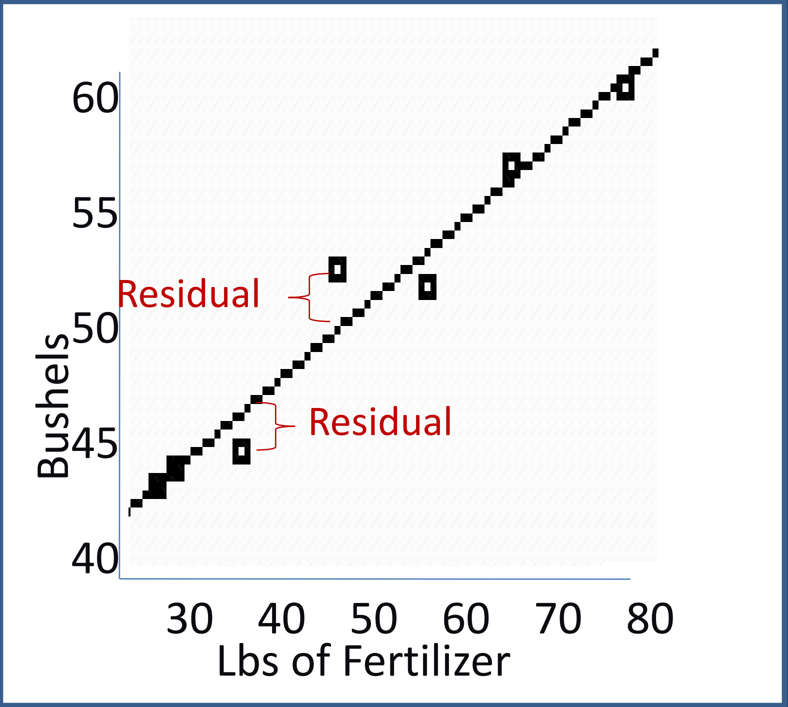
The Linear Model

- In Chapter 6, we used the Normal model to represent the distribution of a single quantitative variable
- In this chapter we will model the relationship between two quantitative variables using a linear model – the equation of a straight line through the data
- In models of the real world the line will not fit the data perfectly, but it will still be very useful in understanding how the two variables are related



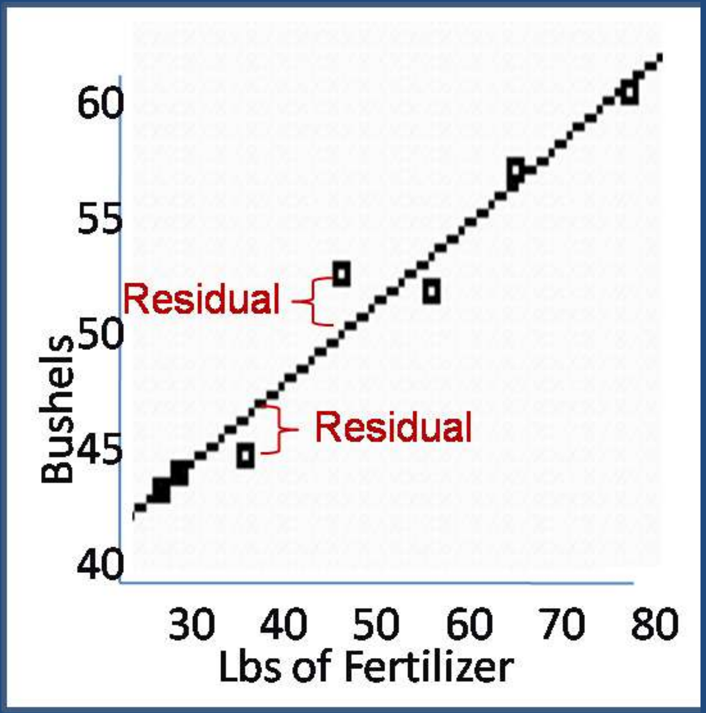
Residuals

- For most data sets, a straight line will not go through all of the data points – in fact, it might not go through any of them, but that doesn't mean it's a bad model
- We are looking for the line that will come *as close as possible* to all of the data points
- The distance between the actual data points and the line are called <u>residuals</u>



Residuals

- We will use the equation of the line to find <u>predicted values</u> for the response variable (in this case, bushels of grain)
- The symbol for the predicted value is $\hat{y}$ (pronounced y-hat)
- The line predicts that a field where 40 lbs of fertilizer is used will have a yield of 47 bushels of grain
- In our data set, the field where 40 lbs of fertilizer was used actually yielded 45 bushels

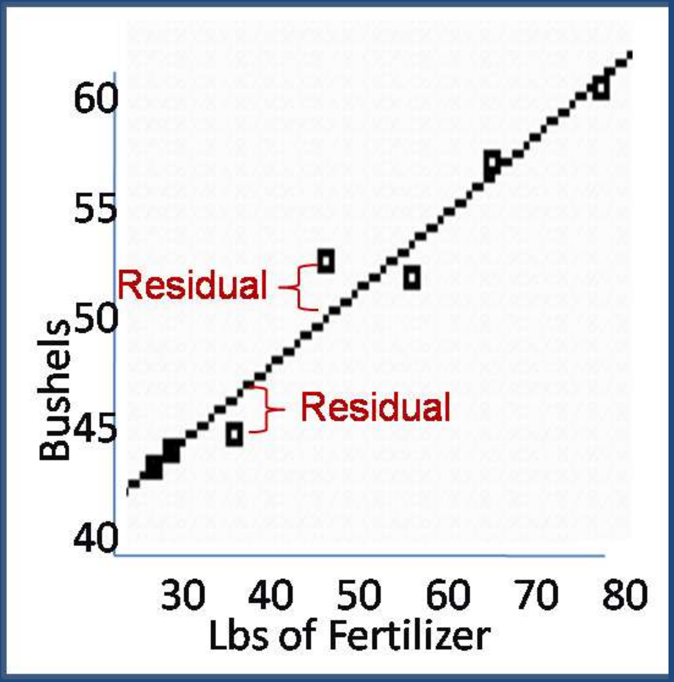


Residuals

- The <u>residual</u> for this point is the difference between what actually happened and what the line predicted

$$e = y - y$$

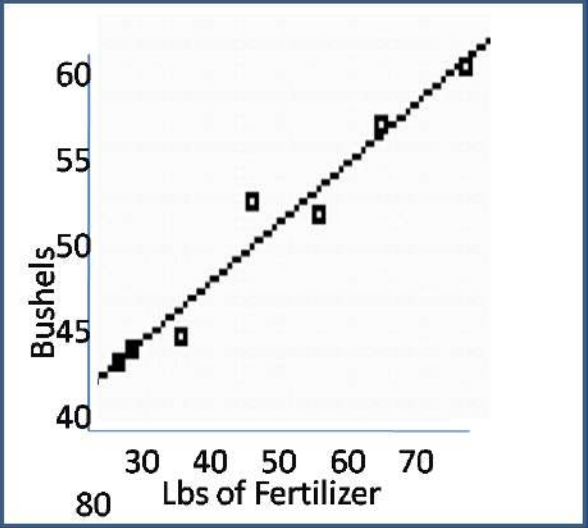
$$e = 45 - 47$$



- The negative residual for this point indicates that the actual value was below the prediction (the point was below the line)
- Points which fall above the line will have positive residuals

Line of Best Fit

- Remember that our model is the “line of best fit”, where the data points are as close as possible to the line
- In other words, we want the residuals to be as small as possible
- If we just found the sum of the residuals, the positive and negative residuals would cancel each other out and the sum would be zero
- Just as we did when we found standard deviation, we will use the squares of the residuals instead
- The linear model we will use is one which minimizes the squared residuals – it’s called the <u>least squares regression line</u>



Finding the Model

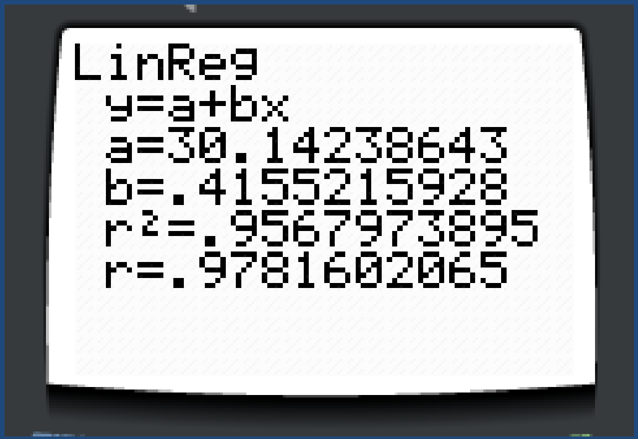
- The equation of the LSRL (Least Squares Regression Line) will be in the form

$$y = a + bx$$

- *a* is the intercept of the line, and *b* is the slope
- The formula for calculating these values can be found on your formula sheet and in your book
- We will find the line either with your calculator or with a computer printout – the main focus of our work will be on interpreting what the line tells us

Finding the Model

- You have already calculated the least squares regression line using your calculator – it's the same calculation you used to find r



- When you write the equation, you need to define the variables
- It should look like this:

$$y = 30.142 + .4155x$$

$\mathrm{x} = \text{lbs of fertilizer}$

$\mathrm{y} = \text{bushels of grain}$

or

$$\widehat{bushels} = 30.142 + .4155(fertilizer)$$

Interpreting the Model

$$\widehat{bushels} = 30.142 + .4155(\mathit{fertilizer})$$

The model has two parts:

- <u>**Slope**</u>: for every increase in x, this is the <u>predicted</u> or <u>average</u> amount we would expect y to change
- For this problem: For every additional pound of fertilizer, we predict about .42 more bushels of grain
- <u>**Intercept**</u>: Remember that this is the value of y when x is 0
- For this problem: If no fertilizer was used, the model predicts that about 30 bushels of grain will be produced.

Using the Model

- Use the line to predict the yield if 50 lbs of fertilizer were used

$$\widehat{bushels} = 30.142 + .4155(fertilizer)$$

$$\widehat{bushels} = 30.142 + .4155(50)$$

$$\widehat{bushels} = 50.917$$

- What is the residual for this point?

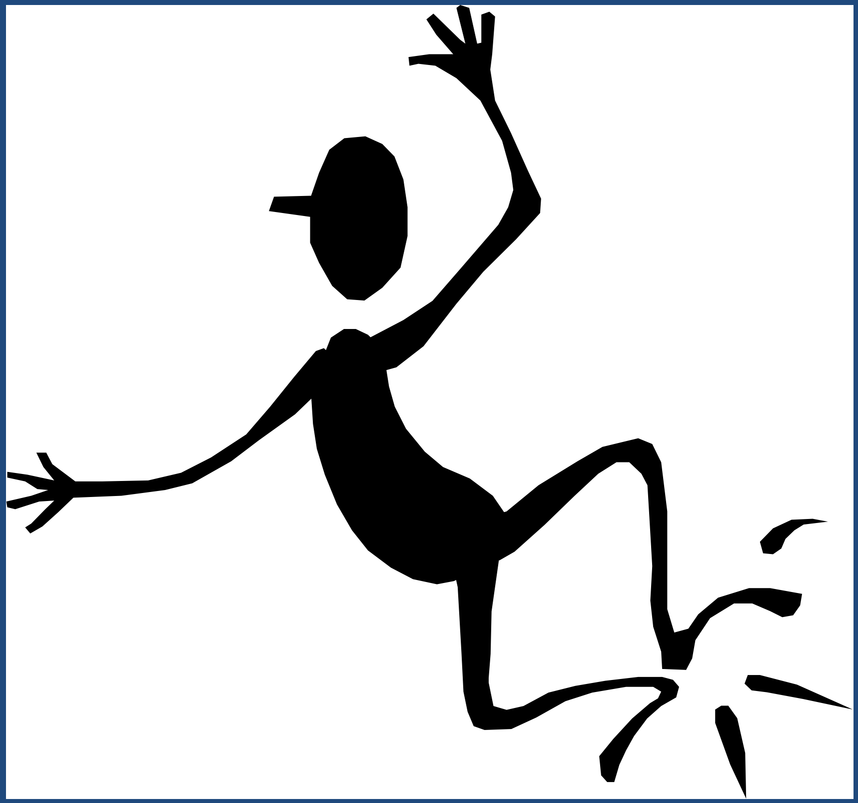
$$e = y - y$$

$$e =$$

Notes About the Regression Line

- Remember that predictions we make from the line are estimates – we don't expect them to match reality perfectly
- Since this model describes the average behavior of y in relation to x, the point $(\overline{x}, \overline{y})$ always falls on the regression line
- Be careful not to extrapolate too much – the model is only good for predictions within the range of the data we used to create it (or close to it)
- Some predictions will not make sense in the context of the problem (often including the intercept)

Very Short Homework!



- Assignment 8-1
- P. 192 #21, 22
- Watch video on how to do a residual plot on the calculator.