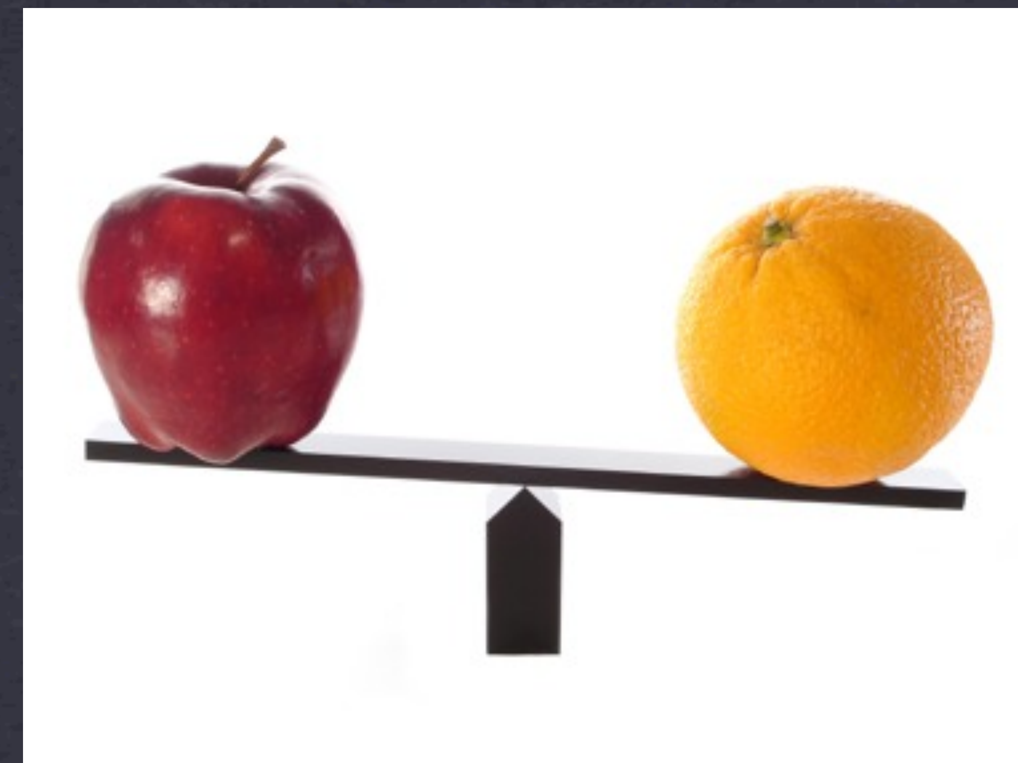
HOW TO RELATE TWO VARIABLES



PROJECT

CHAPTER 11 - REGRESSION

CARTOON GUIDE TO STATISTICS CHAPTER SUMMARY

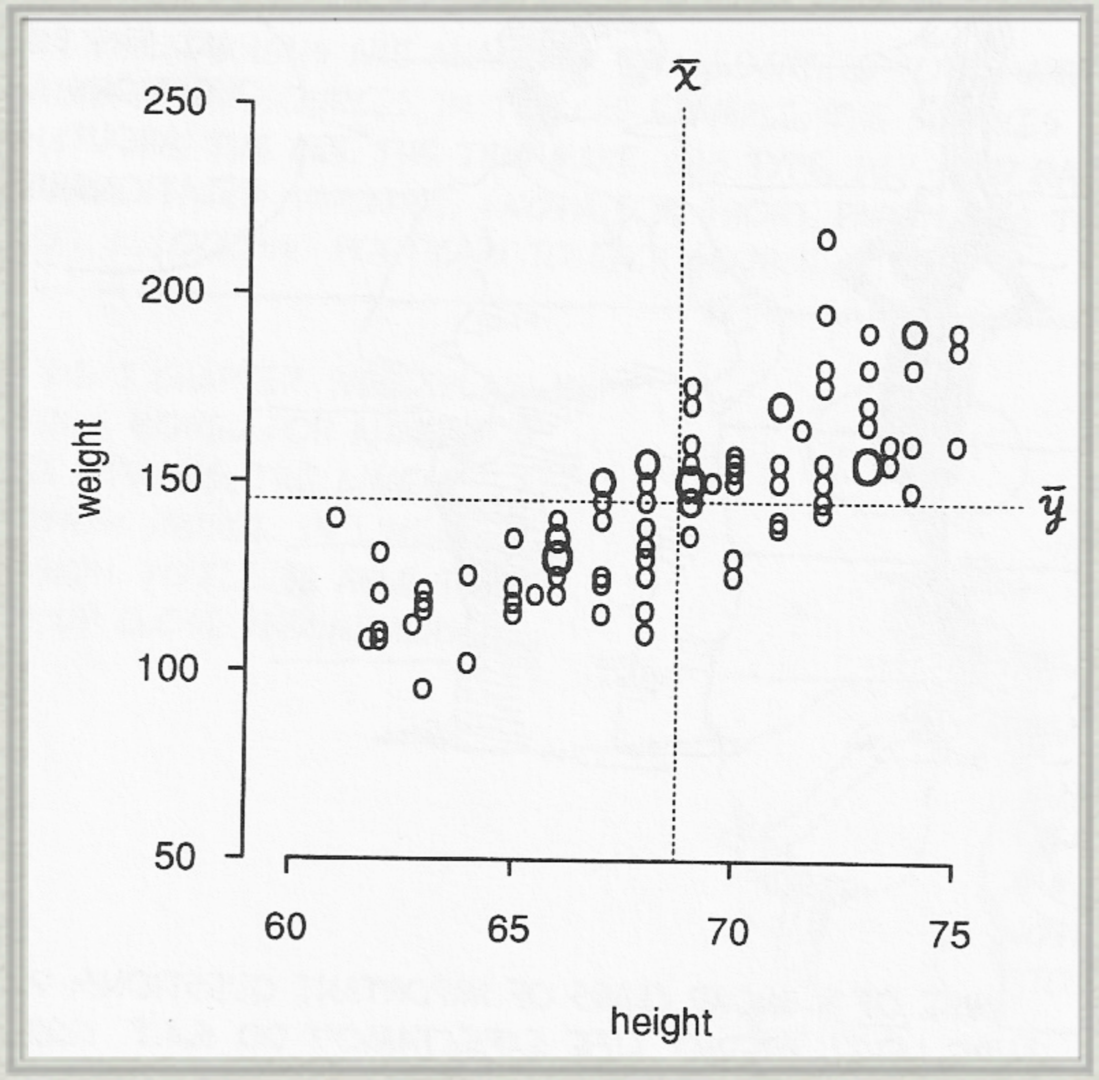
DATE 2010 - 02 - 21

BY JIM CASEY

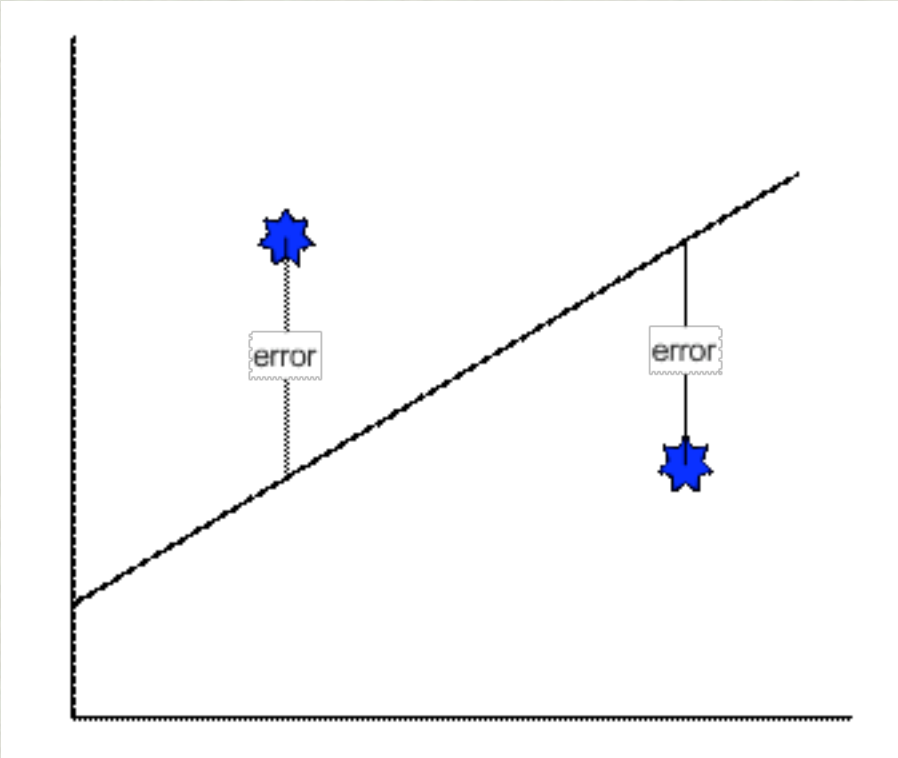
CLASS GEOG 3000

We often use a graph to show how two variables relate

- This is a scatterplot of two variables
- Height and weight are two such variables
- Does one influence the other?
- What about other factors such as *random* variation?



The Regression Line, also called the Least Squares Line

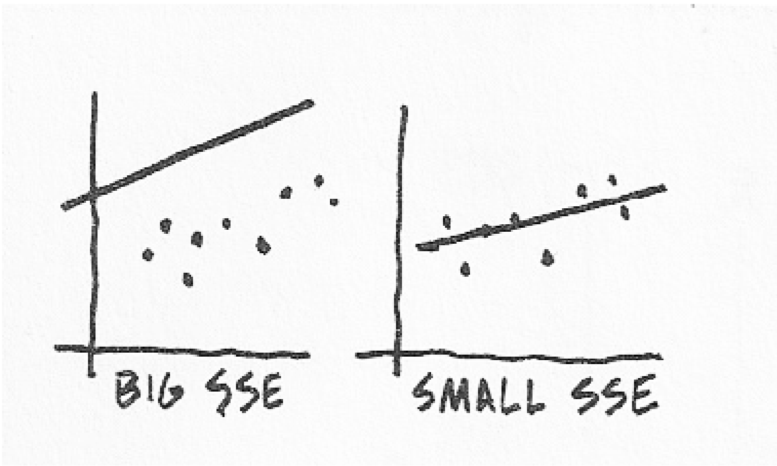


- This line is positioned to minimize the distance between the line and *all y* values
- The distance between any *y* value and the regression line is called ‘error’
- We calculate error values to measure how much the predicted values differ from actual values

The Sum of Squared Errors - SSE

- The regression line is the line with the smallest SSE value

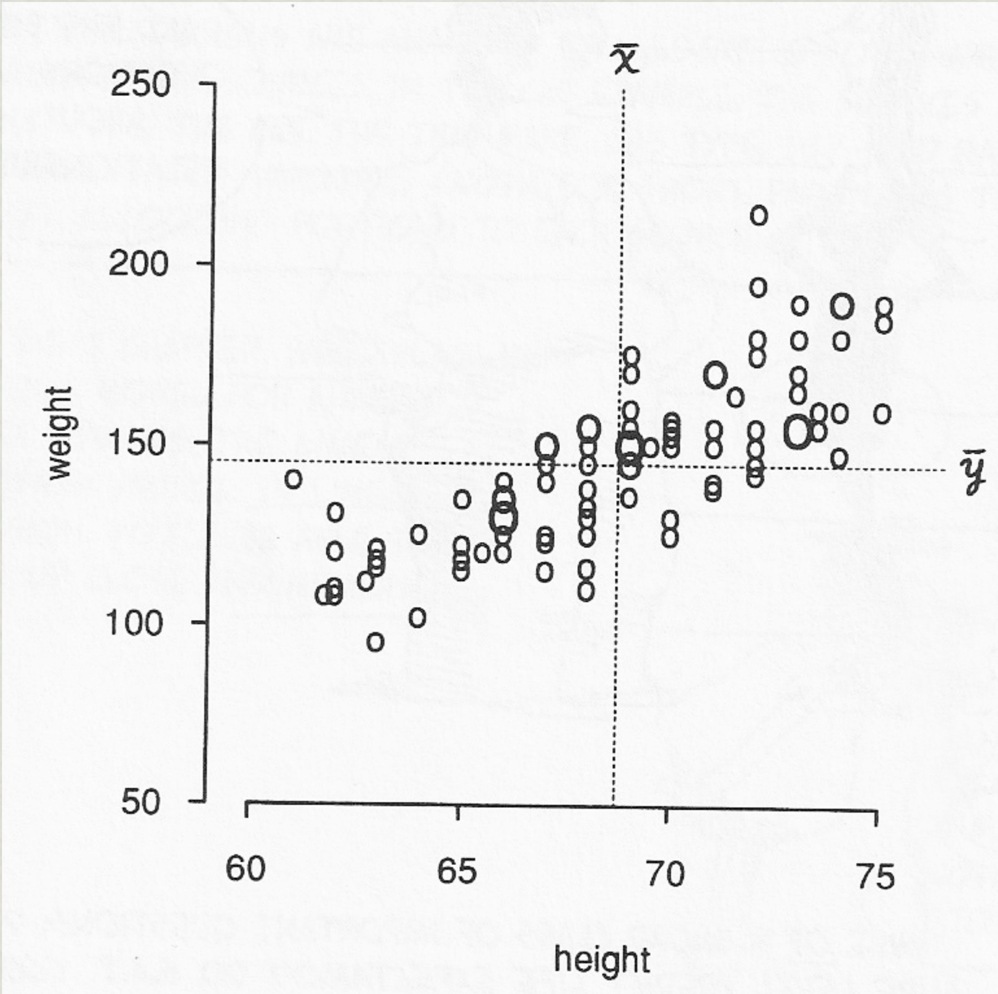
- SSE is found by calculating the sum of squared errors



$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Regression Analysis

- Now, how do we find where the line goes, again?
- We use a formula to find *y* from *a, b and x*
- the *x* axis is the **independent** variable
- the *y* axis is the **dependent** variable



The Regression Line Formula

- We can calculate *y* after finding the value of *b* and *a*

$$y = a + bx$$

WHERE

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

AND

$$a = \bar{y} - b\bar{x}$$

(HERE $\bar{x}$ AND $\bar{y}$ ARE THE MEANS OF $\{x_i\}$ AND $\{y_i\}$ RESPECTIVELY.)

* The sum of squares for *x* and *y* measure the spread of *x* and *y* around the mean

* The combined sum of squares is used to find the variable *b* as well

* We abbreviate these as shown below

$$ss_{xx} = \sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$ss_{yy} = \sum_{i=1}^{n}(y_i-\bar{y})^2$$

$$ss_{xy} = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

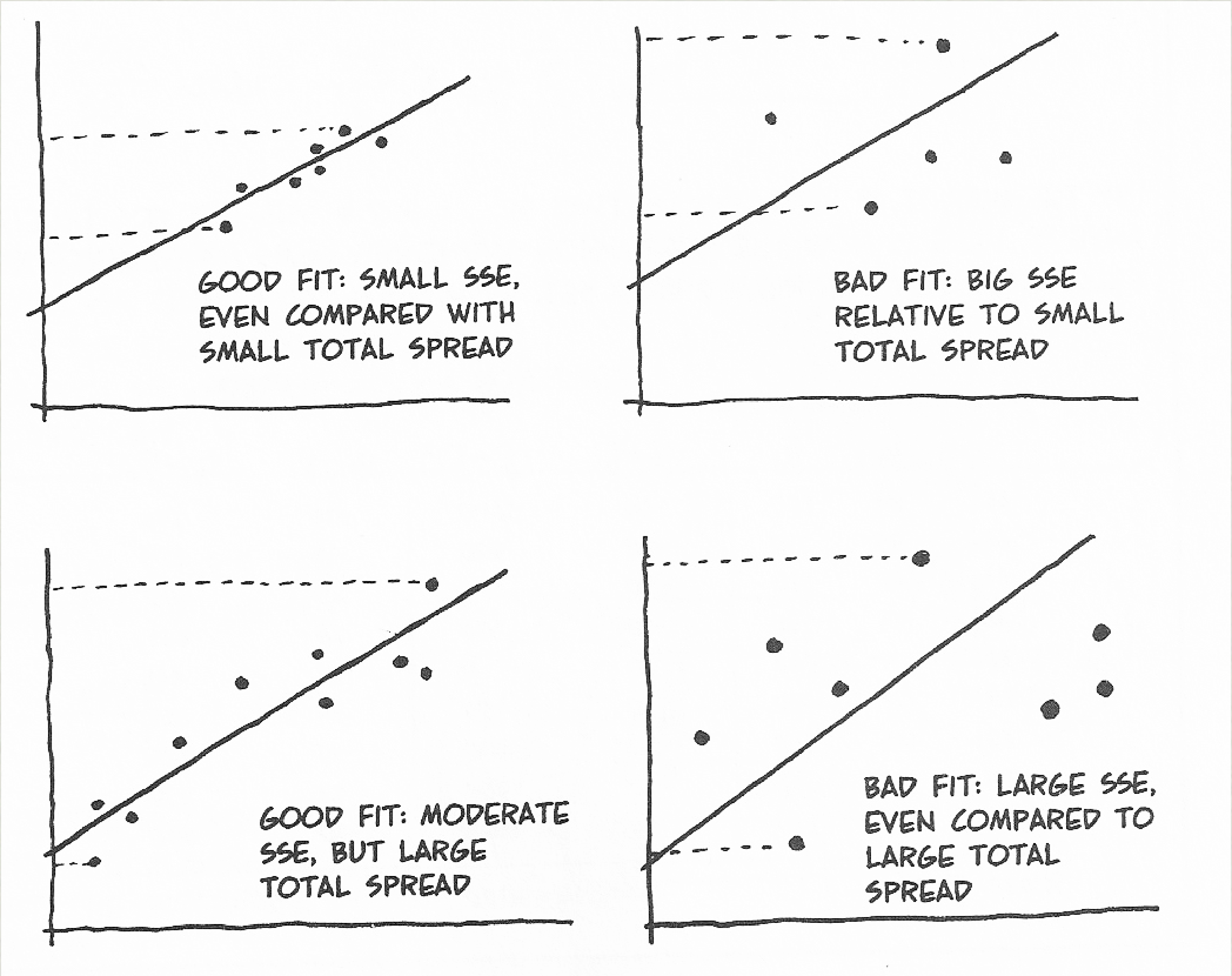
ANOVA - Analysis of Variance

Measuring Goodness of Fit

How well does the line fit the data?

- Some regression lines fit very closely with their data - representing less error
- Some regression lines fit, but have lots of points far away from the line - representing more error
- More error looks like ‘noise’ on the scatterplot
- The ‘fat pencil test’ shows a tight fit, because most of the data can be covered by laying a (fat) pencil over the line

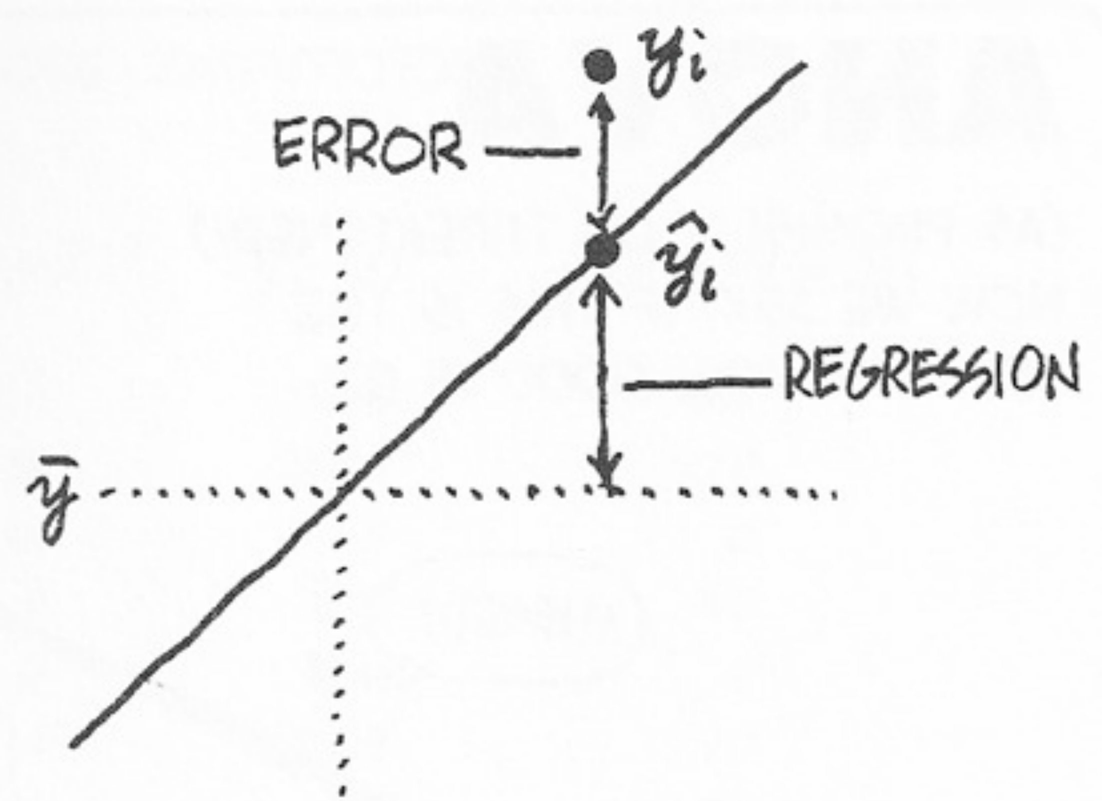
Good Fit vs Bad Fit



Measuring the Variability in *y*

- We use y^{hat} to represent predicted values as determined by the regression line

$$\hat{y}_i = a + b x_i$$



- Thus we can quantify the sources of variability with the sum of squares as shown below

- SSR measures the predicted values of *y*

SOURCE OF VARIABILITY	SUM OF SQUARES
REGRESSION	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
ERROR	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
TOTAL	$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

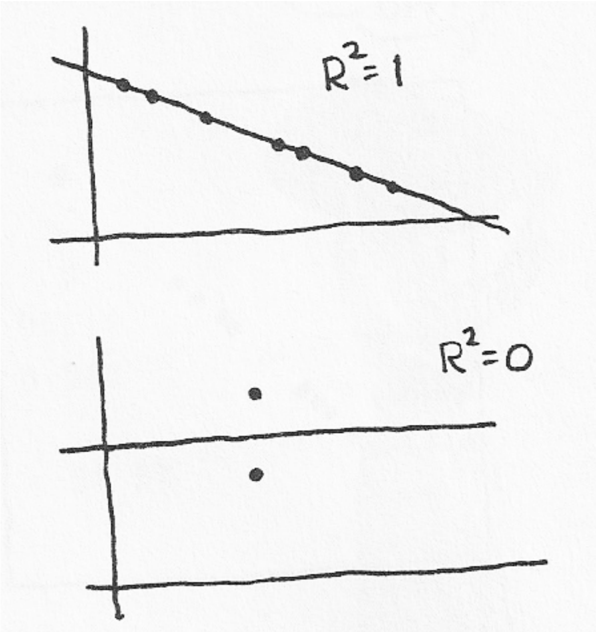
How do the predicted values and actual values relate?

- For this, we need R^2
- R^2 is the correlation coefficient between the actual and predicted values
- This helps us know how well the regression line approximates the actual data - how they are *associated*
- R^2 values can range from -1 to +1
- R^2 is shown with it's own line on a graph

Interpreting R^2 Values

- When $R^2 = 1$, there is a perfect relationship
- When $R^2 = 0$, there is no relationship
- We can know from this whether *y* will go up or down when *x* is increased

$$R^2 = \frac{SSR}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$



Statistical Inference

What is this telling us?

A Regression *Model* for the Entire Population

- To produce a linear regression of the entire population, this formula can be used

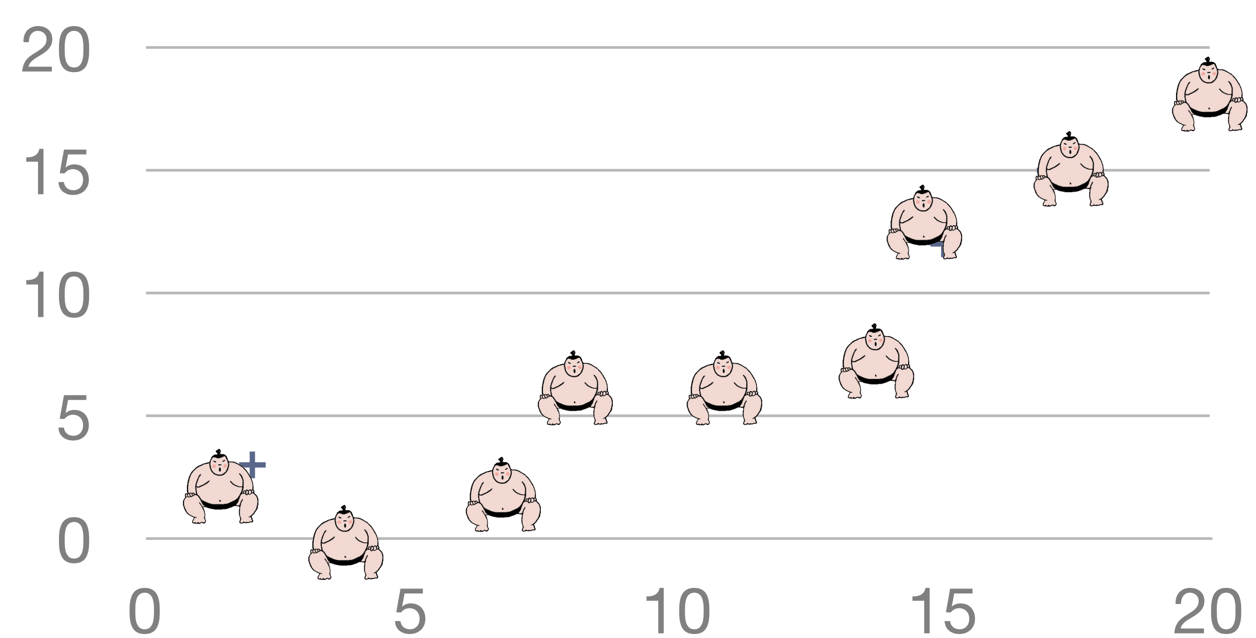
$$Y = \alpha + \beta x + \epsilon$$

Y IS THE DEPENDENT RANDOM VARIABLE; x IS THE INDEPENDENT VARIABLE (WHICH MAY OR MAY NOT BE RANDOM); α AND β ARE THE UNKNOWN PARAMETERS WE SEEK TO ESTIMATE; AND ϵ REPRESENTS RANDOM ERROR FLUCTUATIONS.

Estimating α and β Using Samples

- Using the formula, we can compare values reached by using the least squares method

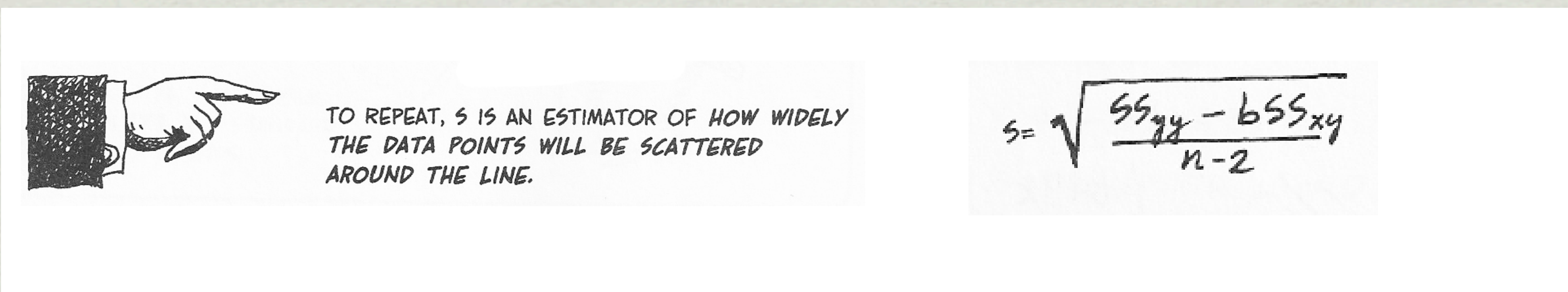
- Using different samples, we have *a* and *b* to compare with β and α



$$Y = \alpha + \beta x + \epsilon$$

Calculating an Estimator

- From samples, we can calculate *s*



Confidence Intervals

Feeling about 95% confident in this one...

How Reliable are our Estimates?

- To measure a 95% confidence interval, use the following formula

$$\beta = b \pm t_{.025} SE(b)$$

$$\alpha = a \pm t_{.025} SE(a)$$

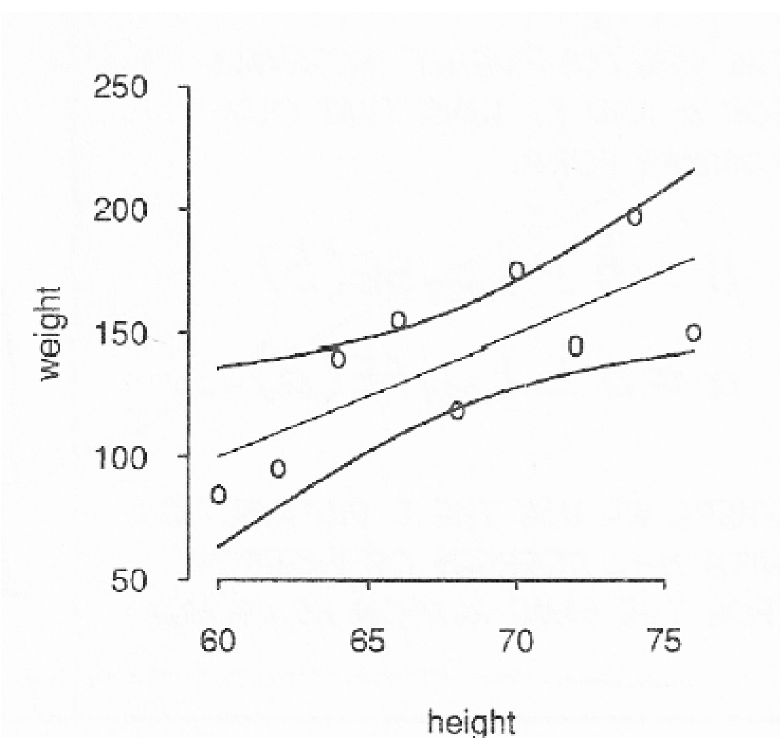
HELPFUL LINK:

http://en.wikipedia.org/wiki/Confidence_interval

Predicting Mean Response

- Mean response is an estimate of an expected value of *y* at a known value of *x*

- The prediction interval is the distribution area between the curved lines



FOR $Y = \alpha + \beta x_0$ IS

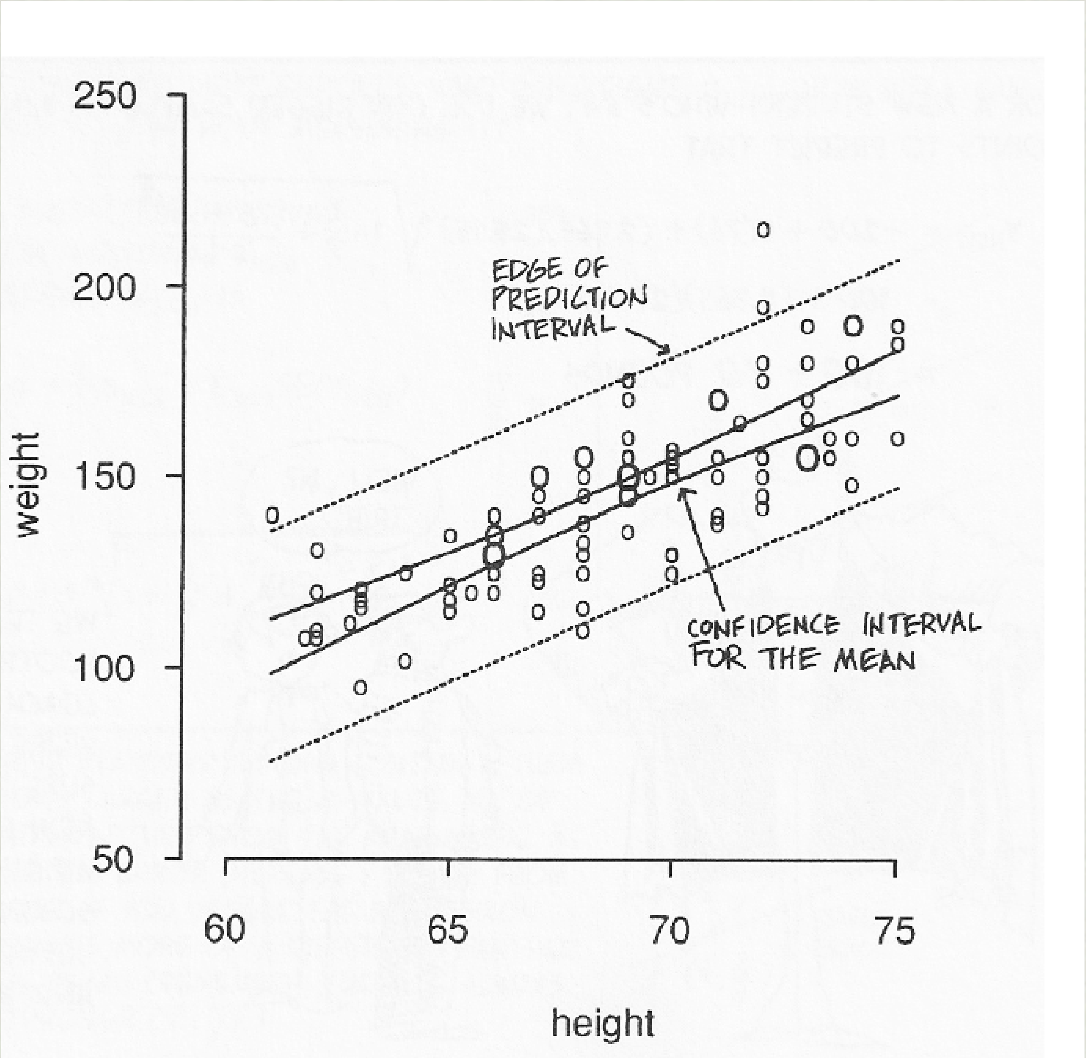
$$\alpha + \beta x_0 = a + b x_0 \pm t_{.025} SE(\hat{y})$$

WHERE

$$SE(\hat{y}) = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

What makes a Good Predictor?

- Many carefully chosen samples will provide more data points for analysis, and better predictions
- Consider more variables than initially considered - it may not be the factor that you anticipated



Hypothesis Testing

- Are these two related?
- The null hypothesis assumes there is no relationship or $H_0 : \beta=0$
- This assumes that *x* does not affect *y* at all
- To test this, we use the *t* test statistic

- We have choices for an alternate hypothesis

- Using H_a : $\beta > 0$ will reject the null hypothesis at $\alpha = .05$ significance level and conclude that there is a relationship

$$t = \frac{b}{SE(b)}$$

$t > t_\alpha$ FOR $H_a : \beta > 0$

$t < t_\alpha$ FOR $H_a : \beta < 0$

$|t| > |t_{\alpha/2}|$ FOR $H_a : \beta \neq 0$

Multiple Linear Regression

- Analyzes one dependent variable and *multiple* independent variables
- This is similar to the other linear regression we have done, but uses matrix algebra - best done on a computer!

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_n x_n + \epsilon$$

Nonlinear Regression

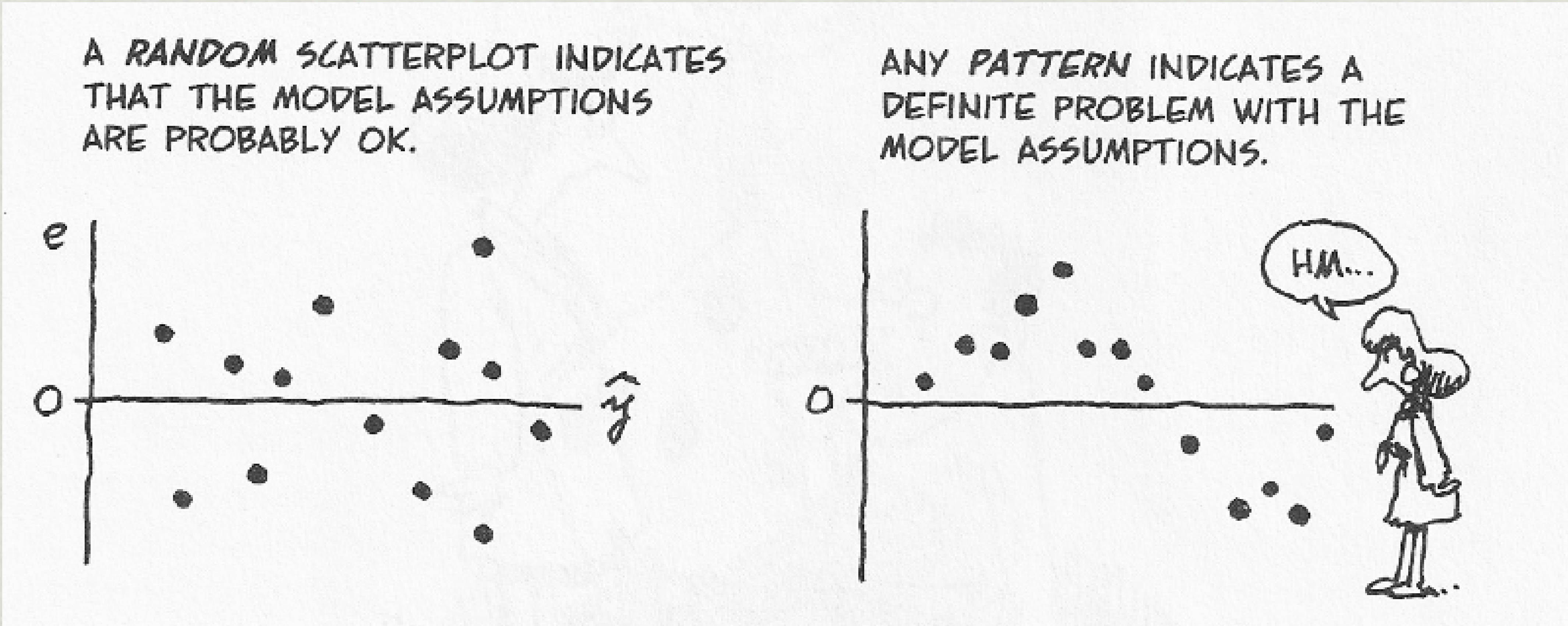
- Not all regressions follow a line
- For data that follows a nonlinear curve, linear regression can be used, such as with the following formula

$$Y = \alpha + \beta_1 x + \beta_2 x^2 + \epsilon$$

AND TREAT x AND x^2 AS INDEPENDENT VARIABLES IN A LINEAR MODEL.

Regression Diagnostics

- Plot the residual e_i against the predictor y_i
- This will create a scatterplot that helps identify any patterns not yet detected



That's All!

- We have covered the basics of regression analysis
- More resources are available at the following links

http://mathworld.wolfram.com/

http://en.wikipedia.org/wiki/Regression_analysis