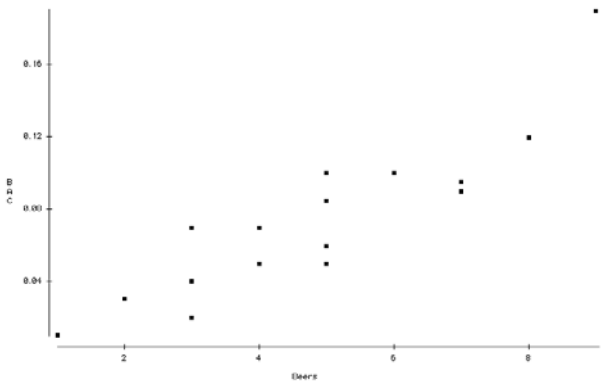
ST517 Note Outline 9: Introduction to Regression

Lecture 9.1: The Regression Equation

Example: In February 1986 sixteen introductory statistics students participated in an experiment to determine how beer consumption influenced blood alcohol content. Each of the students was randomly assigned a number of 12 ounce beers to consume between 1 and 9. After drinking these, the students had their blood alcohol recorded using a breathalyzer. The students were also subjected to a field sobriety (scored 1 to 10) test by a university police officer.



beers	BAC
5	0.1
2	0.03
9	0.19
8	0.12
3	0.04
7	0.095
3	0.07
5	0.06
3	0.02
5	0.05
4	0.07
6	0.1
5	0.085
7	0.09
1	0.01
4	0.05

Remember the equation of a line from math class?

Notation

y

y_i

$\hat{y}_i$

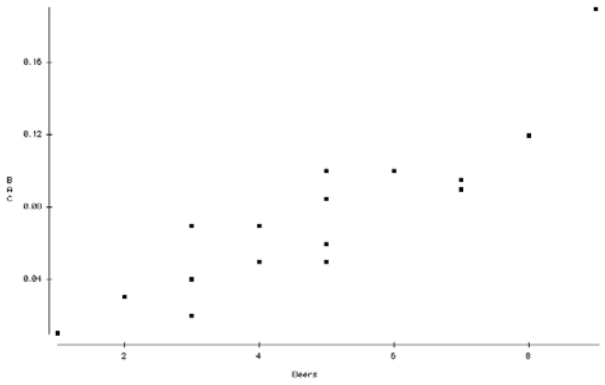
x

x_i

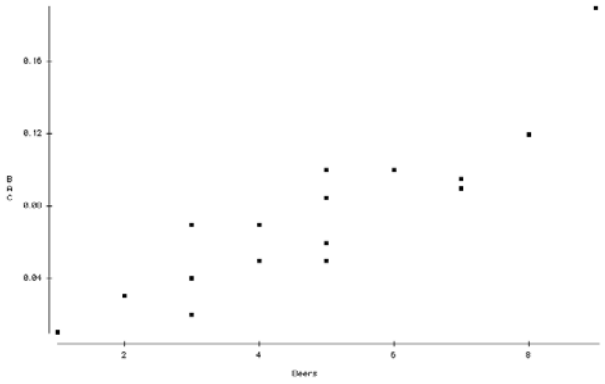
$\hat{\beta}_0$ (or b_0)

$\hat{\beta}_1$ (or b_1)

What is the line?



Another option



Which is better?

- **<u>Residuals</u>** -

$$\text{Residual} = y_i - \hat{y}_i$$

- Also called prediction errors -

Sum of Squares

$$SSE = \sum_i \left(y_i - \hat{y}_i \right)^2$$

Least Squares Line – Line that relates x and y, that minimizes the sum of squared errors

- Form:

Coefficient estimates

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1}^{n}(x_1 - \bar{x})(y_1 - \bar{y})}{\sum_{i=1}^{n}(x_1 - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: BAC

Number of Observations Read	16
Number of Observations Used	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.02338	0.02338	55.94	<.0001
Error	14	0.00585	0.00041783		
Corrected Total	15	0.02922			

Root MSE	0.02044	R-Square	0.7998
Dependent Mean	0.07375	Adj R-Sq	0.7855
Coeff Var	27.71654		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.01270	0.01264	-1.00	0.3320
beers	1	0.01796	0.00240	7.48	<.0001

Interpretation

- Slope -

- Intercept -

Example: What BAC would we predict for someone who drank 3 beers?

Other Output

Minitab:

Regression Analysis: BAC versus beers

The regression equation is
BAC = - 0.0127 + 0.0180 beers

Predictor	Coef	SE Coef	T	P
Constant	-0.01270	0.01264	-1.00	0.332
beers	0.017964	0.002402	7.48	0.000

S = 0.0204410 R-Sq = 80.0% R-Sq(adj) = 78.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.023375	0.023375	55.94	0.000
Residual Error	14	0.005850	0.000418		
Total	15	0.029225			

JMP:

Linear Fit

BAC = -0.012701 + 0.0179638*beers

Summary of Fit

RSquare	0.799841
RSquare Adj	0.785544
Root Mean Square Error	0.020441
Mean of Response	0.07375
Observations (or Sum Wgts)	16

Lack Of Fit

Analysis of Variance

Parameter Estimates

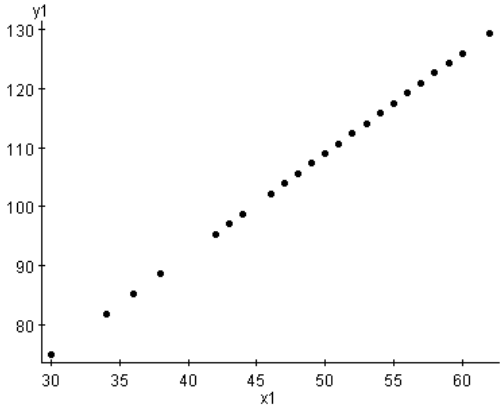
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.012701	0.012638	-1.00	0.3320
beers	0.0179638	0.002402	7.48	<.0001*

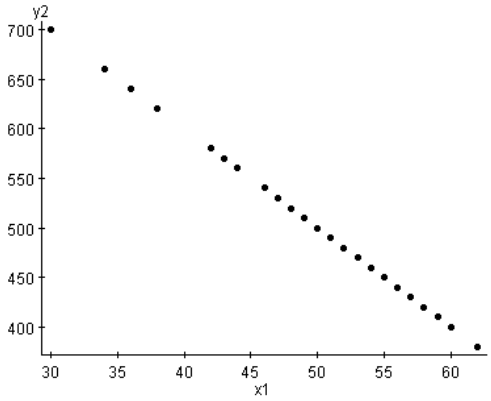
Lecture 9.2: Summaries of the Relationship

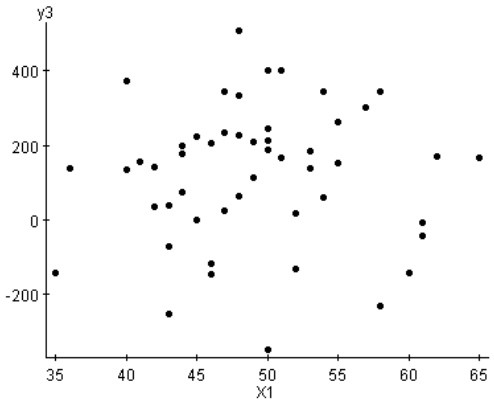
Correlation

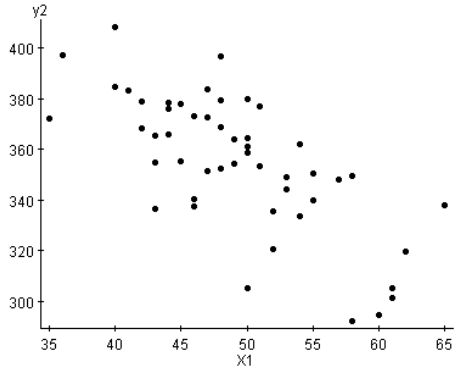
- Notation: r
- Ranges from -1 to +1

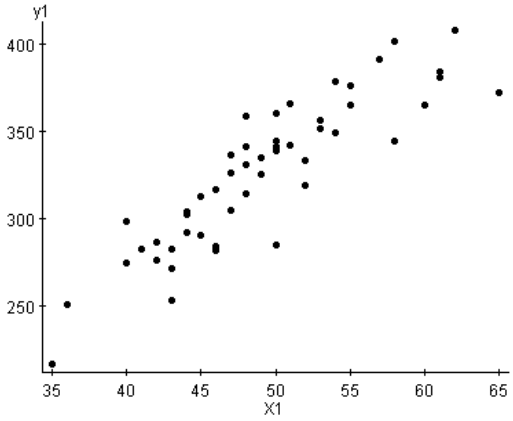
- Hypothetical Examples:

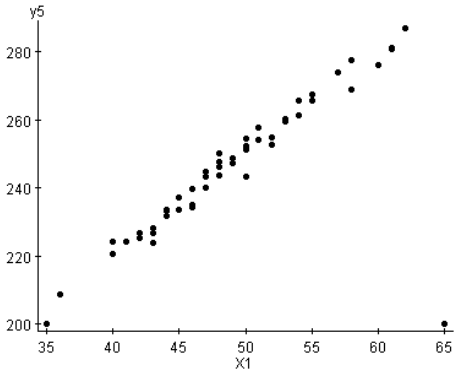


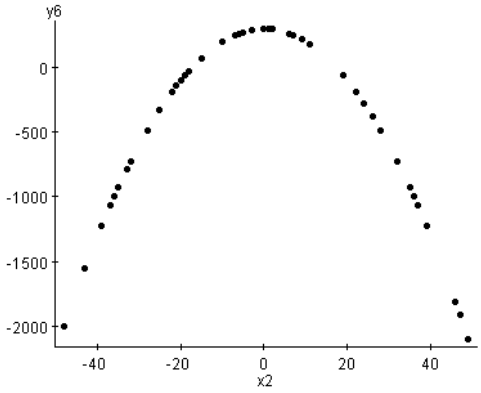












Pearson's Correlation Coefficient

$$\rho = \frac{(X, Y)}{\sigma_X \sigma_Y}$$

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}} = \frac{\sum_{i=1}^{n}(x_1 - \bar{x})(y_1 - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_1 - \bar{x})^2 \sum_{i=1}^{n}(y_1 - \bar{y})^2}} = \hat{\beta}_1 \frac{s_x}{s_y}$$

Coefficient of Determination

- R^2 or r^2 – the square of the correlation

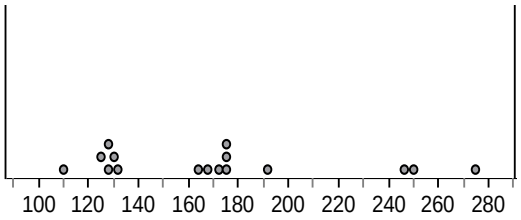
- Interpretation: R^2 represents percentage of variability in *y* that is accounted for by straight line relationship with *x*

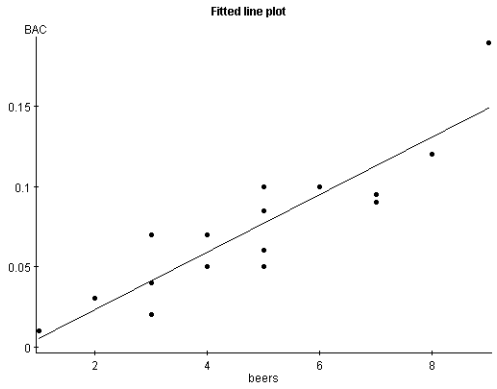
The Standard Deviation of the Points around the Regression Line

Recall: $SSE = \sum_i \left(y_i - \hat{y}_i \right)^2$

Note: Sample standard Deviation

$$s = \sqrt{\frac{\sum \left(y_i - \bar{y} \right)^2}{n-1}}$$





Mean Square Error: Indicates how spread out points are around predictions

$$s^2 = \frac{SSE}{n-2}$$

- Standard deviation: $s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum \left(y_i - \hat{y}_i \right)^2}{n-2}}$

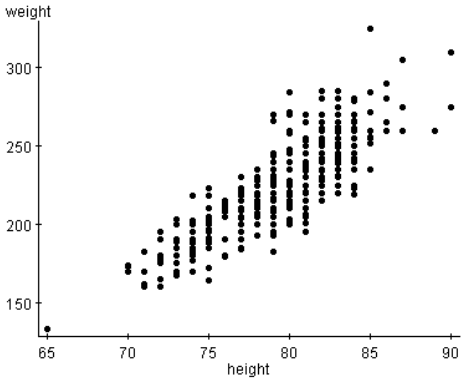
Lecture 9.3: Notes about regression and correlation

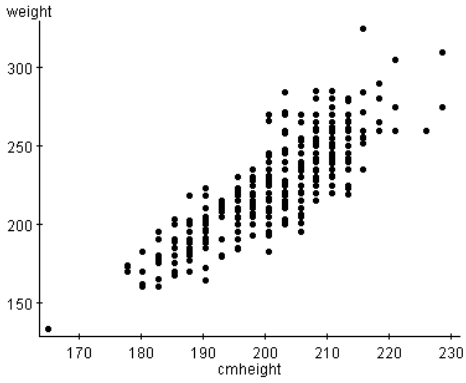
- Straight line relationships only

- Beware outliers

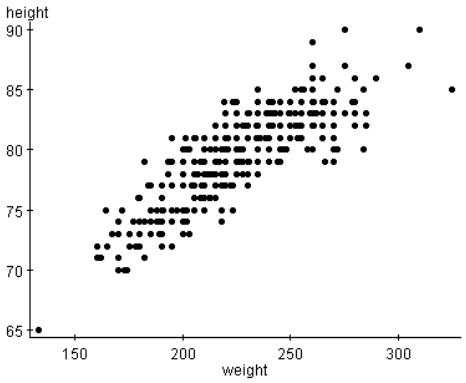
- Correlation does not depend on the units of measure, but the regression does

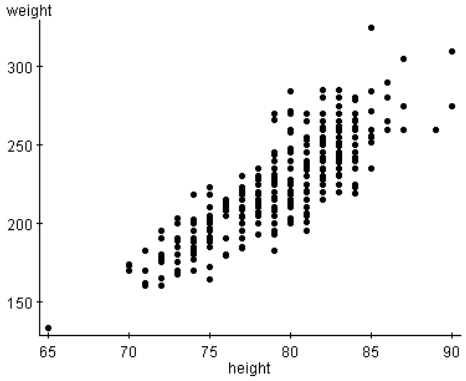
EX: Height and weight of NBA players





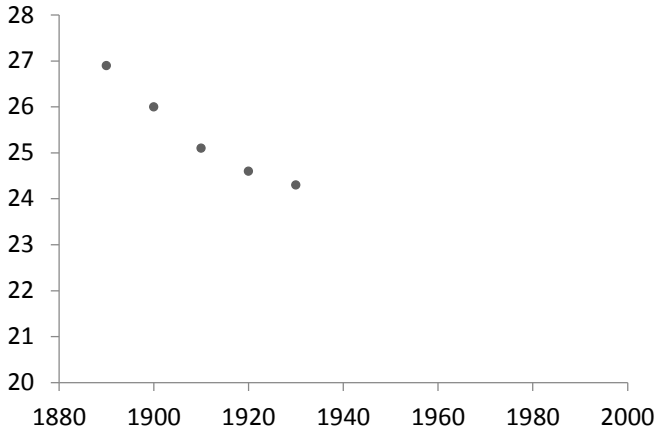
- For regression is important which variable is X and which is Y, for correlation it is not





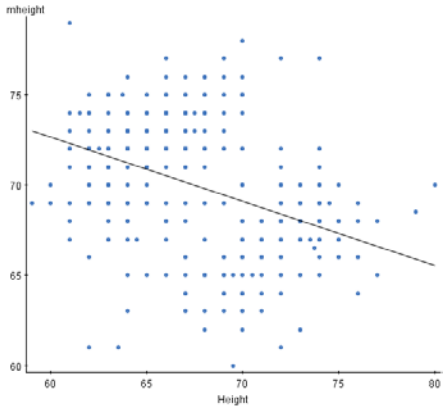
- **Extrapolation:** Predicting beyond the range of the data observed

EX: Predicting age at first marriage



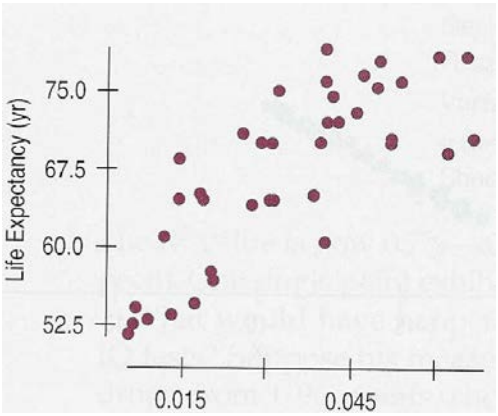
- **Simpson's Paradox:** When the direction of a relationship is surprising

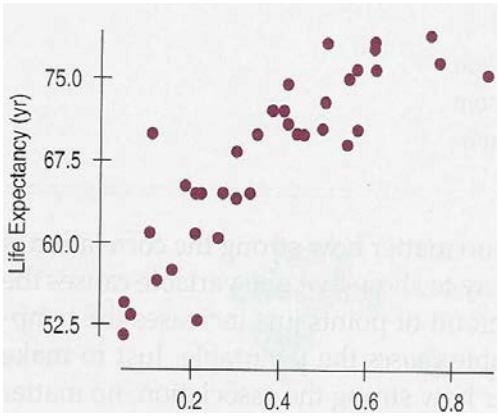
EX: Preferred height



- Correlation does not imply causation

EX: How to improve life expectancy for developing countries?





Lecture 9.4: Inference for the Slope

Recall: The regression equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

- This line is estimated based on sample data
- It is estimating the true form of the relationship for the population, the true regression line:

$$\hat{y} = \beta_0 + \beta_1 x$$

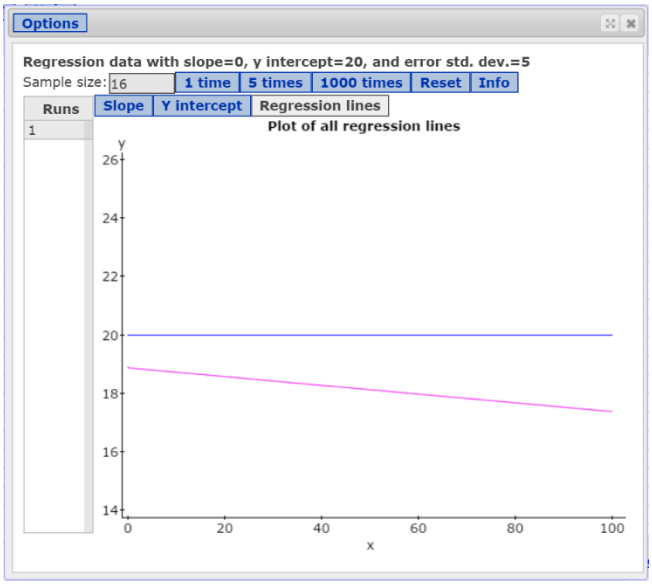
- Written another way: $E(Y) = \beta_0 + \beta_1 X$

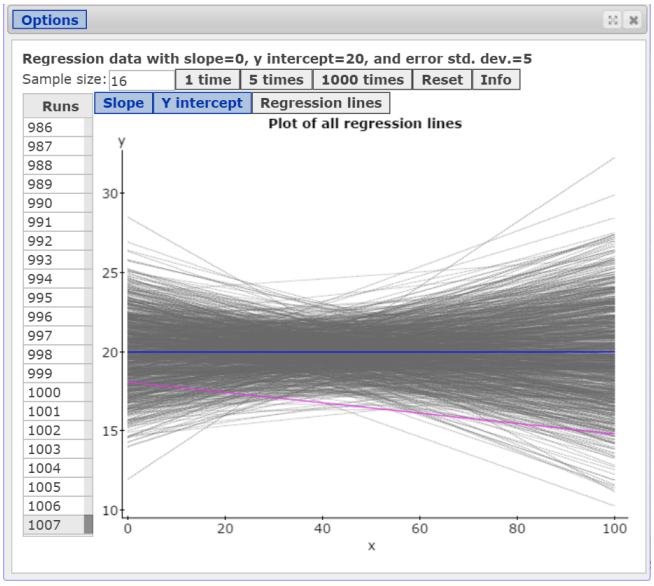
Inference for the slope:

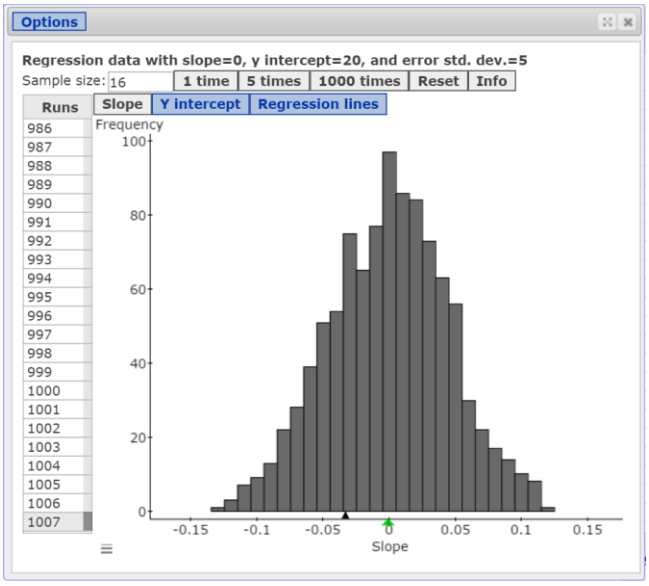
- Why?

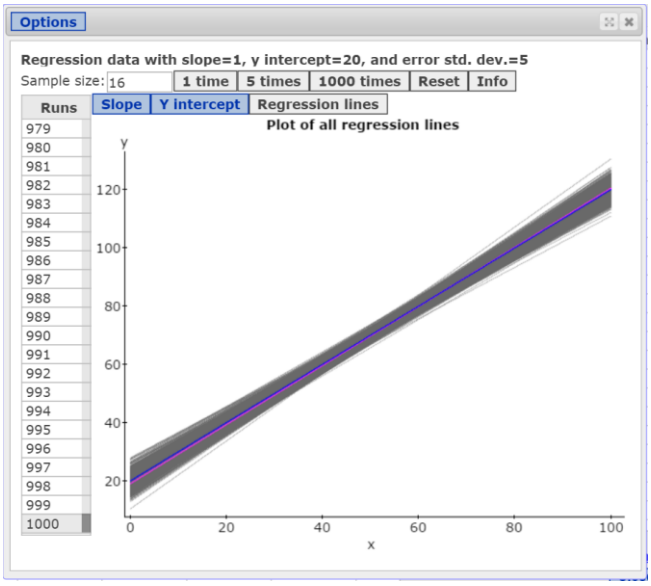
Applet to simulate sampling variability for regression

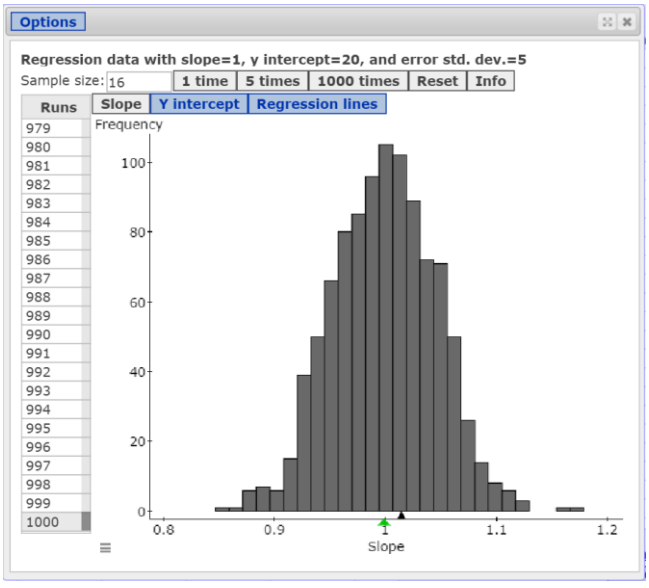
- statcrunch.stat.ncsu.edu
- Applets > Regression > Simulation




Options
Regression data with slope=0, y intercept=20, and error std. dev.=5
Sample size: 16
1 time
5 times
1000 times
Reset
Info
Runs
Slope
Y intercept
Regression lines
Plot of all regression lines
y
x


Options
Regression data with slope=0, y intercept=20, and error std. dev.=5
Sample size: 16
1 time
5 times
1000 times
Reset
Info
Runs
Slope
Y intercept
Regression lines
Frequency
Slope


Options
Regression data with slope=1, y intercept=20, and error std. dev.=5
Sample size: 16
1 time
5 times
1000 times
Reset
Info
Runs
Slope
Y intercept
Regression lines
Plot of all regression lines
y
x


Options
Regression data with slope=1, y intercept=20, and error std. dev.=5
Sample size: 16
1 time
5 times
1000 times
Reset
Info
Runs
Slope
Y intercept
Regression lines
Frequency
Slope

Summary: Distribution of the sample slope

- Shape:
- Centered at:
- Standard error (estimated standard deviation) of the sample slope:

$$s.e.(\hat{\beta}_1) = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Hypothesis Test for the Slope

- Steps 1, 6, and 7 are the same as always
- Hypotheses:

- Type of test:
- Conditions:
- Test Statistic: $t = \frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)}$
- Null distribution:
- p-value found under Ho in direction of Ha
    - Notes about how to find p-value from Note Outline 6 apply

Note: The test for the slope of a simple linear regression is equivalent to a test about the correlation coefficient.

Confidence Interval for the Slope: $\hat{\beta}_1 \pm t^*_{\alpha/2,\nu} s.e.(\hat{\beta}_1)$

- Multiplier found using:

Example: Is there a significant relationship between a person's height and the number of credit hours they are taking? The results of a survey taken by a random sample of 417 students were used to explore this question. Use the following computer output to carry out a test of hypothesis to see if there is a significant relationship between these.

Root MSE	4.19753	R-Square	0.0006
Dependent Mean	67.36451	Adj R-Sq	-0.0018
Coeff Var	6.23108		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	68.08513	1.42662	47.72	<.0001
credits	1	-0.04668	0.09145	-0.51	0.6100

BAC Example: Is there evidence of a significant positive relationship between BAC and the number of beers consumed?

Root MSE	0.02044	R-Square	0.7998
Dependent Mean	0.07375	Adj R-Sq	0.7855
Coeff Var	27.71654		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.01270	0.01264	-1.00	0.3320
beers	1	0.01796	0.00240	7.48	<.0001

Lecture 9.5: Assumptions for the Model

Recall: All statistical inference (intervals, tests) is based on assumptions

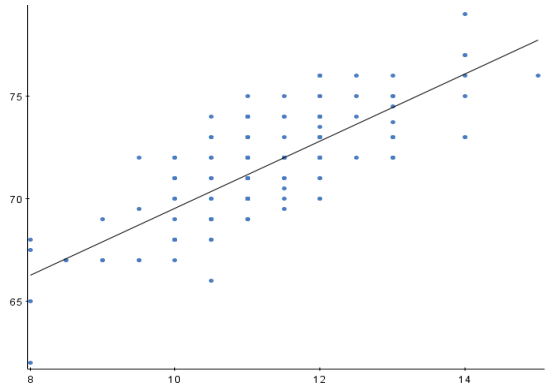
- E.g. Random samples, normal populations, etc.
- Must be true to know the distribution of the statistic

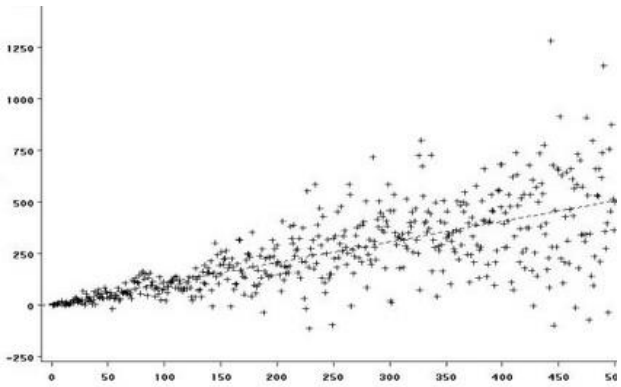
Note: Conditions must be met for the model, and any inference about model, to be valid

Assumptions

1. A straight line is the correct model for the data.

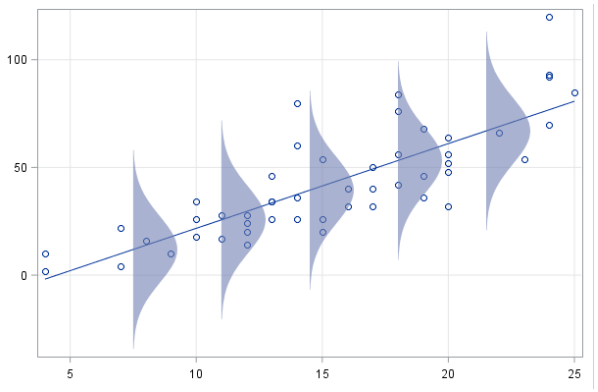
2. The spread of the points around the line have the same standard deviation _____ for all x.

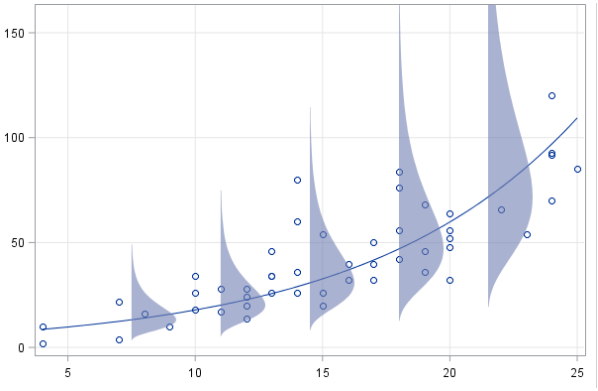




3. The random errors are independent of each other.

4. The points are normally distributed around the line.





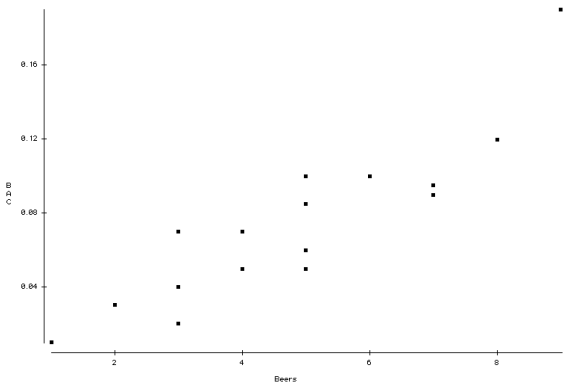
How do we evaluate assumptions?

- Think about how data was collected

- Look at the scatterplot

- Estimate the random errors with the residuals

BAC Example: Do the assumptions for the regression, inference appear to be met?



Lecture 9.6: Confidence and Prediction Intervals for Y

Confidence interval for average at a particular value of *x*

$$\hat{y} \pm t^*_{\alpha/2,\nu} \cdot s \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Example: Average BAC for someone who drank 7 beers

Root MSE	0.02044	R-Square	0.7998
Dependent Mean	0.07375	Adj R-Sq	0.7855
Coeff Var	27.71654		

Using output from proc means:

Variable	Mean	Std Dev	Variance	Corrected SS
beers	4.8125000	2.1975365	4.8291667	72.4375000
BAC	0.0737500	0.0441399	0.0019483	0.0292250

Recall: We assume that the individuals are spread out around the true line

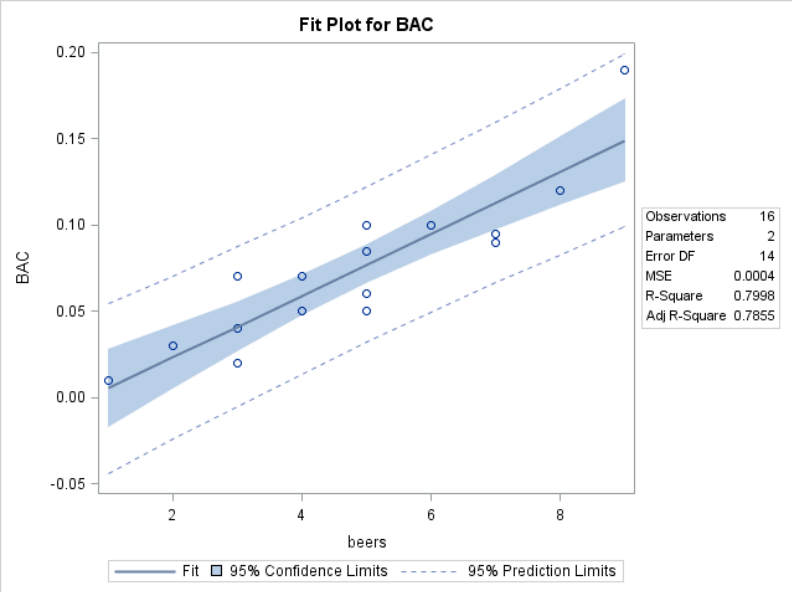
Predicting Individuals = Additional uncertainty

Prediction interval for an individual

$$\hat{y} \pm t^*_{\alpha/2,\nu} \cdot s \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Example: Predicting BAC for an individual who drank 7 beers

Visually:



Lecture 9.7: Additional Examples

Use the following to answer questions 1 to 10: Is there a relationship between the weight of a car and its fuel efficiency? To investigate this, a consumer organization has collects data for 50 randomly selected car models. They use the weight of the car (in thousands of pounds) to predict the fuel efficiency (in miles per gallon or MPG). SAS Output from their analysis is below and on the next page; use this to answer the questions that follow.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	865.41049	865.41049	148.58	<.0001
Error	48	279.56951	5.82436		
Corrected Total	49	1144.98000			

Root MSE	2.41337	**R-Square**	0.7558
Dependent Mean	25.02000	**Adj R-Sq**	0.7507
Coeff Var	9.64577		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	48.73931	1.97558	24.67	<.0001
weight	**1**	-8.21362	0.67383	-12.19	<.0001

1. On the next page is a panel of plots produced by PROC REG. The plots outlined are the residual plot and qq-plot; based on these, do the conditions for a linear regression analysis using this data appear to be valid?
2. Is there a positive or negative relationship between weight and fuel efficiency?
3. What percent of the variability in fuel efficiency is explained by its linear relationship with weight?
4. Write the estimated least squares regression equation (i.e. write the estimated model).
5. On average, how far are the points from the estimated equation? (In other words: what is the average distance between the points and the line?)
6. Use the estimated model to predict fuel efficiency for a car that weighs 4 thousand pounds.
7. One car that weighs 4 thousand pounds had an actual fuel efficiency of 15 MPG. What is the residual for this particular car?
8. What is the value of the correlation between weight and fuel efficiency?

9. Is there a significant linear relationship between weight and fuel efficiency? Conduct the appropriate hypothesis test.
10. What parts of the test in question (9) would change if we were interested in a *negative* relationship between weight and fuel efficiency?