Course Outline through Week 11

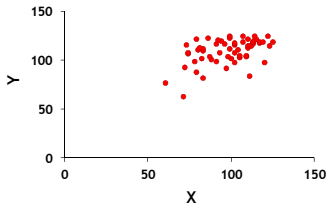
Between now and the 3rd exam we will focus on measuring the association between two variables, X & Y

1. When X is discrete and Y is continuous, we will use "analysis of variance" techniques (Done)
2. When X and Y are both discrete, we will use cross-tabular and χ^2 analyses (Done)
3. When X and Y are both continuous, we will use correlation & regression analyses (Today)

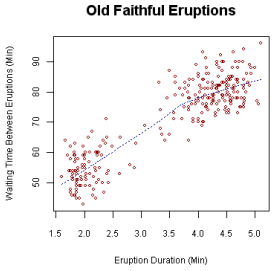
Scatterplots

Scatterplot

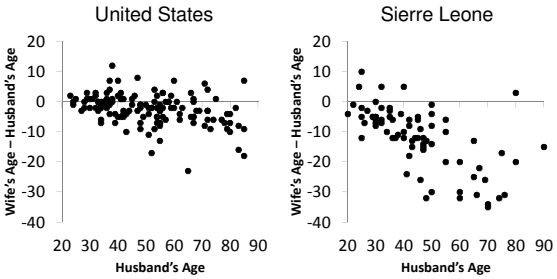
A diagram that displays the covariation of two continuous variables as a set of points on a Cartesian coordinate system



Scatterplots



Scatterplots



Scatterplots

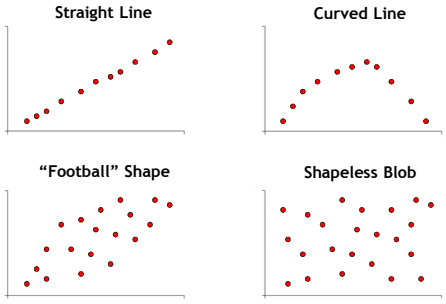
What is the basic shape of the relationship between the two variables? A straight line? A curve? A blob?

What is the direction of the relationship? Positive, negative, or uncertain?

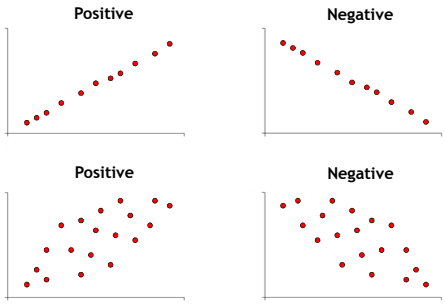
How much variability is there? How many points deviate from the basic pattern?

Are there outliers? Unusual observations?

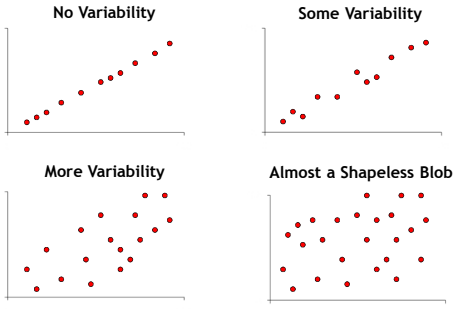
Scatterplots: Shape



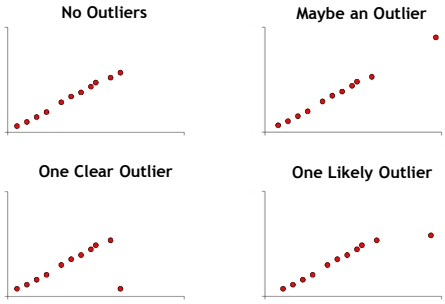
Scatterplots: Direction



Scatterplots: Variability

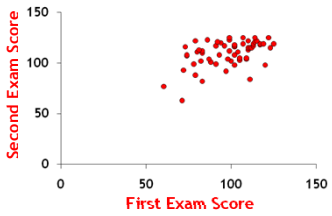


Scatterplots: Outliers



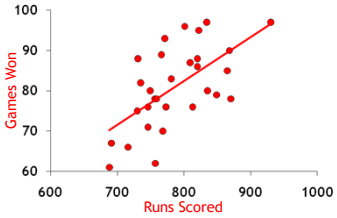
Worksheet

Describe the relationship depicted in this scatterplot



Bivariate Regression

Example: The scatterplot below relates the number of runs scored during the 2009 baseball season to the number of games won by MLB teams in 2009



Bivariate Regression

Conceptually speaking, bivariate regression involves drawing a line through the points on the scatterplot that comes closest to the points on the Y dimension

Regression analysis involves estimating an equation that…

- …describes how, on average, the response variable (Y) is related to the predictor variable (X)
- …allows us to make predictions about the value of the response variable (Y) given a specified value of the predictor variable (X)

When we "regress Y on X" we produce a model that predicts Y on the basis of X

Bivariate Regression

The algebraic equation for a line:

$$Y = a + bX$$

The prediction equation, which expresses the i^{th} individual's value of dependent variable Y as a function of predictor variable X, is:

$$\hat{Y}_i = a + b_{yx}X_i$$

The linear regression model recognizes deviations (or errors, e_i) from the prediction equation:

$$Y_i = a + b_{yx}X_i + e_i$$

Bivariate Regression

Given the prediction equation:

$$\hat{Y}_i = a + b_{yx}X_i$$

and the linear regression model:

$$Y_i = a + b_{yx}X_i + e_i$$

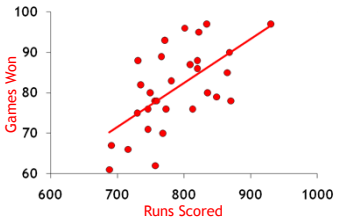
we see that the error term e_i (also know as the residual) can be expressed as the difference between the observed value of Y (from the linear regression model) and the predicted value of Y (from the prediction equation):

$$Y_i - \hat{Y}_i = [a + b_{yx}X_i + e_i] - [a + b_{yx}X_i] = e_i$$

Bivariate Regression

How do we know how to draw the regression line?

There are an infinite number of lines that one could draw through these points ... Why is the middle (red) line best?



Estimating a Regression Equation

The line that we draw … the values of intercept a and slope b_{YX} that we choose … maximizes our ability to predict the value of Y (and thus minimizes the prediction errors)

Mathematically, we choose the line for which

$$\sum_{i=1}^{N}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{N}e_i^2$$

is smallest

This is the "least squares error sum" criterion and produces ordinary least squares (OLS) estimates of intercept a and slope b_{YX}

Estimating a Regression Equation

The OLS estimate of slope b_{YX}:

$$b_{YX}=\frac{\sum_{i=1}^{N}(Y_i-\bar{Y})(X_i-\bar{X})}{\sum_{i=1}^{N}(X_i-\bar{X})^2}$$

$$\text{Covariance}_{YX}=s_{YX}=\frac{\sum_{i=1}^{N}(Y_i-\bar{Y})(X_i-\bar{X})}{n-1}$$

$$\text{Variance}_X=s_X^2=\frac{\sum_{i=1}^{N}(X_i-\bar{X})^2}{n-1}$$

so…

$$b_{YX}=\frac{\text{Covariance}_{XY}}{\text{Variance}_X}=\frac{s_{YX}}{s_X^2}$$

Estimating a Regression Equation

A computationally simpler formula for the OLS estimate of slope b_{XY}:

$$b_{XY}=\frac{n\sum_{i=1}^{n}Y_iX_i-\sum_{i=1}^{n}Y_i\sum_{i=1}^{n}X_i}{n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2}$$

The formula for the intercept a is:

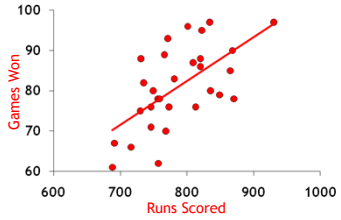
$$a=\bar{Y}-b\bar{X}$$

When slope b_{YX} equals zero, then the intercept equals the mean of Y and the predicted value of Y at all values of X is equal to the mean of Y

Estimating a Regression Equation

For the baseball example, the prediction equation is:

$\hat{Y}_i = -4.76 + 0.11X_i$



Interpreting the Regression Equation

$\hat{Y}_i = -4.76 + 0.11X_i$

How do we interpret this regression equation?

It says that for every one unit increase in X (runs) we should observe a 0.11 unit increase in Y (wins)

It also literally says that if a team were to score zero runs in a season … such that X=0 … we should observe that the team would win -4.76 game (more on this later)

Using the regression equation we can…

…estimate the average value of Y for a given value of X

…predict an individual's value of Y for a given value of X

Interpreting the Regression Equation

The equation $\hat{Y}_i = -4.76 + 0.11X_i$ means (in English) that

Expected Number of Wins = -4.76 + 0.11Runs

How many wins would we predict a team to win if they scored 801 runs?

Expected Number of Wins = -4.76 + 0.11(801) = 83.35

What is the average number of wins among teams that score 750 runs?

Expected Number of Wins = -4.76 + 0.11(750) = 77.74

How Well Does X Predict Y?

How well does a particular regression equation do in predicting values of the response variable Y?

- How strong the association is between X and Y
- If the association is extremely strong, then knowing X allows you to almost perfectly predict Y
- If the association is weak, then knowing X does nothing to allow you to predict the value of Y

As with ANOVA, we can ask how much of the variation in Y can be attributed to X and how much is random error

- That is, we can partition the variance in Y into the part attributable to X and the part attributable to error

How Well Does X Predict Y?

Start with the deviation:

$$Y_i - \bar{Y} = Y_i - \bar{Y}$$

Then add *and* subtract the predicted value from the right-hand side and rearrange the equation:

$$Y_i - \bar{Y} = Y_i - \bar{Y} + (\hat{Y}_i - \hat{Y}_i)$$

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

Deviations from the mean can be expressed as the sum of (1) deviations of the predicted value from the mean and (2) individual deviations from the predicted value

How Well Does X Predict Y?

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

Portion of the deviation from the mean that is attributable to X

Portion of the deviation from the mean that is attributable to random error

Deviations from the mean can be expressed as the sum of (1) deviations of the predicted value from the mean and (2) individual deviations from the predicted value

How Well Does X Predict Y?

If we square each side and then sum across cases we get:

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2+\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$$

Total Sum of Squares | Regression Sum of Squares | Error Sum of Squares

$$SS_{TOTAL} = SS_{REGRESSION} + SS_{ERROR}$$

If there is no association between Y and X, then knowing X does not help predict Y

In this case, our best guess about Y-hat is Y-bar; thus $SS_{REGRESSION}$ equals zero and $SS_{TOTAL}=SS_{ERROR}$

How Well Does X Predict Y?

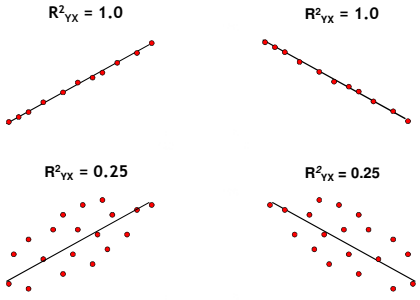
The coefficient of determination (R^2_{YX}) indicates the proportion of the total variation in Y that is determined by its linear relationship with X

$$R^2_{YX}=\frac{SS_{TOTAL}-SS_{ERROR}}{SS_{TOTAL}}=\frac{SS_{REGRESSION}}{SS_{TOTAL}}$$

If $R^2_{YX}=0$, then $SS_{REGRESSION}=0$, which suggests that there is no association between Y and X

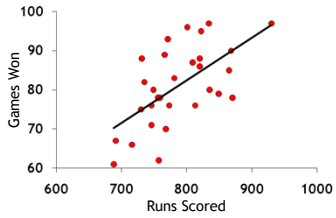
If $R^2_{YX}=1$, then $SS_{REGRESSION}=SS_{TOTAL}$, which suggests that there is no error variation and that we can perfectly predict Y based on X

How Well Does X Predict Y?



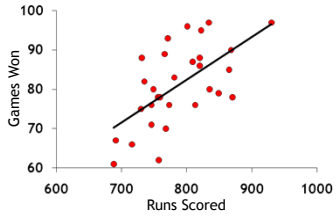
How Well Does X Predict Y?

For the baseball example: $R^2_{YX} = \frac{SS_{REGRESSION}}{SS_{TOTAL}} = \frac{1128.163}{2948.967} = 0.383$

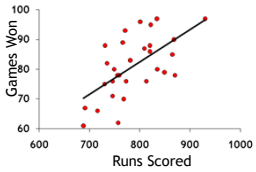


How Well Does X Predict Y?

An R^2_{YX} of 0.383 means that 38.3% of the variation in Y (Wins) is "explained" by X (Runs Scored)

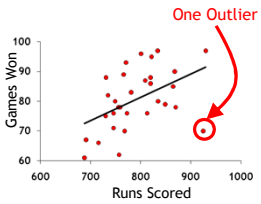


Caution I: Outliers?



$R^2_{YX} = 0.383$

$\hat{Y}_i = -4.764 + 0.109X_i$

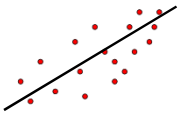


$R^2_{YX} = 0.227$

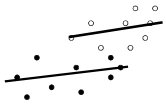
$\hat{Y}_i = 18.703 + 0.078X_i$

Caution II: One Population?

All Cases Included in One Group

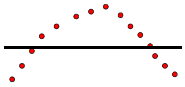


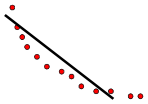
Cases Analyzed as Two Distinct Groups



Caution III: Linear Association?

Linear Regression Equations Fit to Curvilinear Data

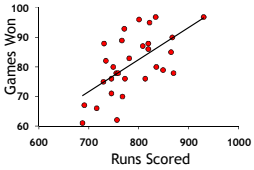




Caution IV: Predicting Beyond X

Literally, the equation to the right says that a team would win -4.764 games if they scored zero runs

Lesson: Don't make predictions beyond the observed range of X (~700 to ~900 runs in this case)



$R^2_{YX} = 0.383$

$\hat{Y}_i = -4.764 + 0.109X_i$

Caution V: Correlation is Not Causation

Regression and correlation are methods for describing the association between continuous variables

In order to make causal statements … for example, that X affects Y … several other things have to be true … more on all of this later

Just because X and Y are highly correlated does not necessarily mean that X causes Y …

- It may be that Y causes X instead
- Some third variable may completely account for the correlation
- Some third variable may partially account for the correlation

Correlation

The correlation coefficient (r_{YX}) summarize the strength and direction of the association between two continuous variables

Correlation equals the square root of R^2_{YX}, but it can also be computed as

$$r_{YX} = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_x}\right)\left(\frac{Y_i - \bar{Y}}{s_y}\right)$$

r_{YX} can also be expressed as the ratio of the covariance of Y and X to the product of the variances of Y and X

$$r_{YX} = s_{xy} / s_Y s_X$$

Correlation

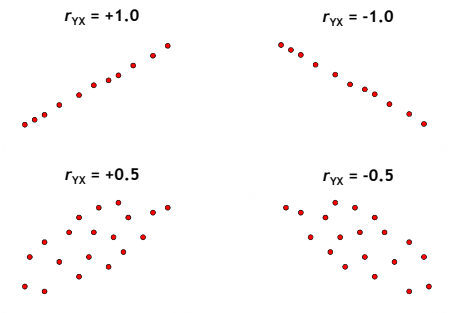
Correlation always ranges from -1 to +1

Correlations between 0 and +1 indicate a positive relationship; if r_{YX}=+1, then there is a perfect positive association

Correlations between -1 and 0 indicate a negative association; if r_{YX}=-1, then there is a perfect negative association

If r_{YX}=0, there is absolutely no association

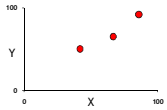
Correlation



Correlation: Example

Consider the following data on three students' scores on a mid-term exam and on a final exam:

Student:	1	2	3
X: Mid-term:	86	67	42
Y: Final Exam:	92	65	50



Here are the summary statistics for each variable:

	X	Y
Mean	65	69
Standard Deviation	22.1	21.3
N = 3		

Correlation: Example

The formula for correlation looks complicated, but it only involves the means and standard deviations of the two quantitative variables

In our example:

$$r_{YX} = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_x}\right)\left(\frac{Y_i - \bar{Y}}{s_y}\right)$$

$$r_{YX} = \left(\frac{1}{3-1}\right)\left[\left(\frac{86-65}{22.1}\right)\left(\frac{92-69}{21.3}\right) + \left(\frac{67-65}{22.1}\right)\left(\frac{65-69}{21.3}\right) + \left(\frac{42-65}{22.1}\right)\left(\frac{50-69}{21.3}\right)\right]$$

$$r_{YX} = \left(\frac{1}{2}\right)[(.95)(1.08) + (0.09)(-0.19) + (-1.04)(-.89)] = \left(\frac{1}{2}\right)(1.935)$$

$$r_{YX} = 0.97$$

Correlation and Slope

r_{YX} is not the same as the slope b_{YX}, but there is a close relationship between the two. Formally:

$$b_{YX} = \frac{s_{YX}}{s_X^2} \rightarrow s_{YX} = b_{YX} s_X^2$$

$$r_{YX} = \frac{s_{YX}}{s_Y s_X} \rightarrow s_{YX} = r_{YX} s_Y s_X$$

$$b_{YX} s_X^2 = r_{YX} s_Y s_X \rightarrow b_{YX} = \frac{r_{YX} s_Y s_X}{s_X^2} \rightarrow b_{YX} = r_{YX} \frac{s_Y}{s_X}$$

Recommended Formulas

The easiest way (and order in which) to compute the statistics we covered today:

Compute the mean and standard deviation of each variable

Compute r_{YX} as $r_{YX} = \left(\frac{1}{N-1}\right)\sum_{i=1}^{N}\left(\frac{X_i - \bar{X}}{s_x}\right)\left(\frac{Y_i - \bar{Y}}{s_y}\right)$

Compute the slope b_{YX} as $b_{YX} = r_{YX}\frac{s_Y}{s_X}$

Compute the intercept a as $a = \bar{Y} - b\bar{X}$

Compute R^2_{YX} as $R^2_{YX} = r^2_{YX}$

Worksheet

Here are values of 2 variables, X and Y, for n=4 people:

	X	Y	
Person #1	3	10	$\bar{X}$ = 4
Person #2	5	8	s_X = 1.826
Person #3	2	11	$\bar{Y}$ = 10
Person #4	6	11	s_Y = 1.414

Using these data and summary statistics:

1. Draw a scatterplot
2. Compute and interpret r_{YX}
3. Compute and interpret the least-squares regression line; draw it on your scatterplot
4. Compute and interpret R^2_{YX}

BREAK

Correlation and Regression: Review

Correlation (r_{YX}) measures the strength and direction of the association between continuous variables Y and X

Bivariate regression involves drawing a line through the points on the scatterplot such that the sum of the squared prediction errors equals zero

Regression analysis allows us to…

- …quantify the degree to which variability in Y is "explained by" variability in X (using R^2_{YX});
- …describe how, on average, the response variable (Y) is related to the predictor variable (X) (using b_{YX}); and
- …make predictions about the value of the response variable (Y) given a specified value of the predictor variable (X)

Inferences About Associations

When we talked about describing associations between continuous variables, we were using sample data

We would like to be able to make inferences about associations between continuous variables in the population from which the sample were drawn

For example:

- Is the relationship observed in the sample data strong enough to confidently conclude that there is a relationship in the population?
- What can we infer about the likely values of the slope in the population based on the slope in the sample?

Inferences About Associations

Hypothesis Tests About ρ^2_{YX}

- R^2_{YX} is a sample estimate of population parameter ρ^2_{YX}
- If ρ^2_{YX} equals zero, then X does nothing to explain variability in Y

Hypothesis Tests About ρ_{YX}

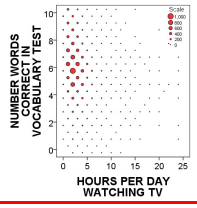
- r_{YX} is a sample estimate of population parameter ρ_{YX}
- If ρ_{YX} equals zero, then there is no correlation between X and Y

Hypothesis Tests About Slope β_{YX}

- b_{YX} is a sample estimate of population parameter β_{YX}
- If β_{YX} equals zero, then the regression of Y on X has a zero slope

Example

The scatterplot below relates GSS respondents' hours per week watching TV (X) & their vocabulary test score (Y)



$\bar{Y} = 6.01$ $\bar{X} = 2.95$

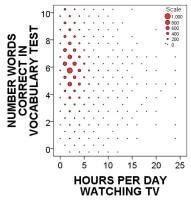
$s_y = 2.13$ $s_x = 2.31$

$r_{YX} = -0.193$ $n = 15{,}357$

Note: These are the only statistics required to estimate the least squares regression equation and to perform the tests described today

Example

The scatterplot below relates GSS respondents' hours per week watching TV (X) & their vocabulary test score (Y)



$$b_{YX} = r_{YX}\frac{s_Y}{s_X}$$

$$b_{YX} = -0.193\frac{2.13}{2.31} = -0.178$$

$$a = \bar{Y} - b\bar{X}$$

$$a = 6.01 + 0.178(2.95)$$

$$a = 6.535$$

$$R^2_{YX} = r^2_{YX} = 0.193^2 = 0.037$$

Assumptions of the Regression Model

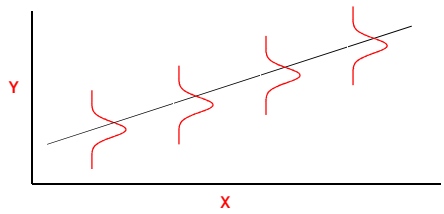
Assumptions of the Regression Model

Assumptions we make when using sample-based regression models to make inferences about associations in the population:

1. The functional form of the relationship between X and Y is appropriately specified; usually this means checking for linearity
2. There are no extreme outliers
3. The variability of the prediction errors is constant across the observed values of X (assumption of homoskedasticity)
4. The values of Y are normally distributed at each value of X (assumption of normality)
5. The observations are independent

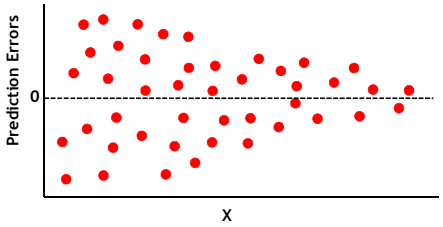
Assumptions of the Regression Model

Graphical summary of the first four assumptions:



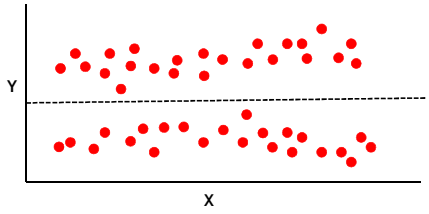
Assumptions of the Regression Model

What does it look like if variability of the prediction errors are not constant across all values of X? (*Note*: This is just one possibility)



Assumptions of the Regression Model

What does it look like if the values of Y are not normally distributed at all values of X in the population? (*Note*: This is just one possibility)



Assumptions of the Regression Model

The first three assumptions…

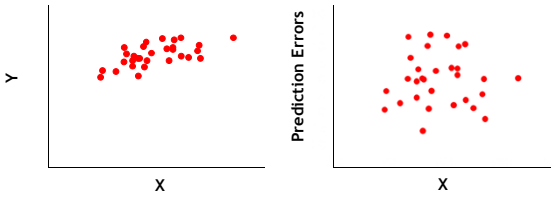
1. The functional form of the relationship between X and Y is appropriately specified; usually this means checking for linearity
2. There are no extreme outliers
3. The variability of the prediction errors is constant across the observed values of X

…can be checked using two plots

First, create a scatterplot with X on the horizontal axis and Y on the vertical axis

Second, create a scatterplot with X on the horizontal axis and the prediction errors on the vertical axis

Assumptions of the Regression Model



1. Is the association between X and Y plausibly linear?
2. Are there extreme outliers?
3. Is the variability of the prediction errors (the residuals) constant across the range of X?

Assumptions of the Regression Model

The fourth assumption…

4. The values of Y are normally distributed at each value of X

…can be checked by examining a histogram of the prediction errors (or residuals)

The fifth assumption…

5. The observations are independent

…is a matter of appropriate research design

Assumptions of the Regression Model

In later courses you may learn about more sophisticated techniques for diagnosing violations of the assumptions of the linear regression model

For now, just be aware of these assumptions

If the assumptions are not met, then hypothesis tests about ρ^2_{YX}, ρ_{YX}, and β_{YX} are generally invalid

Inferences About Associations

Hypothesis Testing in 6 Steps

1. State the null (H_0) and alternative (H_1) hypotheses
2. Check that the sample data conform to basic assumptions; if they do not, then do not go any further
3. Choose an α probability level … that is, a probability associated with incorrectly rejecting the null hypothesis
4. Determine the "critical value" … that is, how large the test statistic must be in order to reject the null hypothesis at the given α level
5. Calculate the test statistic … F or Z or t, depending
6. Compare the test statistic to the critical value

Inferences About ρ^2_{YX}

Inferences About ρ^2_{YX}

State the null (H_0) and alternative (H_1) hypotheses

H_0: $\rho^2_{YX} = 0$
H_1: $\rho^2_{YX} > 0$

This is a one-sided test (with no <) because ρ^2_{YX} cannot possibly be less than zero

Failing to reject the null means failing to reject the hypothesis that X explains none of the variation in Y

Inferences About ρ^2_{YX}

Check that the sample data conform to basic assumptions; if they do not, then do not go any further

The assumptions of the regression model described earlier must hold for hypothesis tests about ρ^2_{YX} to be valid

Inferences About ρ^2_{YX}

Choose an α probability level … that is, a probability associated with incorrectly rejecting the null hypothesis

Let's choose α=0.01

Inferences About ρ^2_{YX}

Determine the "critical value" … that is, how large the test statistic must be in order to reject the null hypothesis at the given α level

The hypothesis test for ρ^2_{YX} is (as described below) an F test with df_{NUM}=1 and df_{DENOM}=n-1

In our example, we want $F_{1,15356}$ for α=0.01 which is 6.63

We will thus reject H_0 if our F statistic exceeds 6.63

Inferences About ρ^2_{YX}

Calculate the test statistic

Remember from before…

$$\sum_{i=1}^{N}(Y_i-\bar{Y})^2=\sum_{i=1}^{N}(\hat{Y}_i-\bar{Y})^2+\sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2$$

Total Sum of Squares | Regression Sum of Squares | Error Sum of Squares

$SS_{TOTAL} = SS_{REGRESSION} + SS_{ERROR}$

We defined $R^2_{YX}=\frac{SS_{TOTAL}-SS_{ERROR}}{SS_{TOTAL}}=\frac{SS_{REGRESSION}}{SS_{TOTAL}}$

Inferences About ρ^2_{YX}

Calculate the test statistic

The F statistic here is

$$F_{1,N-2}=\frac{SS_{REGRESSION}/1}{SS_{ERROR}/n-2}=\frac{MS_{REGRESSION}}{MS_{ERROR}}$$

There is an analogy between this and ANOVA … in both cases we are asking whether variation in Y can be attributed to individuals' values on X

If X and Y are associated, then $MS_{REGRESSION}$ (and thus F) will be larger

Inferences About ρ^2_{YX}

Calculate the test statistic

Computationally:

$$SS_{TOTAL}=(s_Y^2)(n-1)$$
$$SS_{REGRESSION}=(R^2_{YX})(SS_{TOTAL})$$
$$SS_{ERROR}=SS_{TOTAL}-SS_{REGRESSION}$$

In our example:

$$SS_{TOTAL}=(2.13^2)(15{,}357-1)=69{,}668.6$$
$$SS_{REGRESSION}=(0.193^2)(69{,}668.6)=2{,}595.1$$
$$SS_{ERROR}=69{,}668.6-2{,}595.1=67{,}073.5$$

$$F_{1,N-2}=\frac{SS_{REGRESSION}/1}{SS_{ERROR}/n-2}=\frac{2{,}595.1/1}{67{,}073.5/15{,}355}=594.1$$

Inferences About ρ^2_{YX}

Compare the test statistic to the critical value

- If the test statistic is larger than the critical value, then reject H_0
- If the test statistic is less than or equal to the critical value, then do not reject H_0

We can restate the hypotheses:

H_0: $\rho^2_{YX} = 0$ ➔ Fail to reject H_0 if $F \leq 6.63$

H_1: $\rho^2_{YX} > 0$ ➔ Reject H_0 if $F > 6.63$

Since F=594.1, we reject H_0 … so it appears that in the population X and Y are associated, such that X accounts for some of the variability in Y

Worksheet

Mean of X:	6.50	SD of X:	2.95
Mean of Y:	7.15	SD of Y:	1.46
r_{XY}:	0.51		
n:	20		

Test the hypothesis that ρ^2_{YX} --- the population proportion of variation in Y explained by X --- is zero in the population; use α=0.05

Inferences About Correlation (ρ_{YX})

Inferences About Correlation (ρ_{YX})

Since r_{YX} is just the square root of R^2_{YX} (in the case of bivariate regression), a hypothesis test about ρ_{YX} will yield the same result as a hypothesis test about ρ^2_{YX}

Another way to directly test hypotheses about the correlation coefficient ρ_{YX} is to utilize the r-to-Z transformation:

$$Z_r = \left(\frac{1}{2}\right) \ln\left(\frac{1+r_{yx}}{1-r_{yx}}\right)$$

The following test statistic has a standard normal distribution:

$$Z = \frac{Z_r - 0}{\sqrt{1/n-3}}$$

Inferences About Correlation (ρ_{YX})

State the null (H_0) and alternative (H_1) hypotheses

H_0: $\rho_{YX} = 0$
H_1: $\rho_{YX} \neq 0$

This is a two-sided test since ρ_{YX} can range from -1 to +1

Failing to reject the null means failing to reject the hypothesis that X and Y are uncorrelated in the population

Inferences About Correlation (ρ_{YX})

Check that the sample data conform to basic assumptions; if they do not, then do not go any further

The assumptions of the regression model described earlier must hold for hypothesis tests about ρ_{YX} to be valid

Inferences About Correlation (ρ_{YX})

Choose an α probability level … that is, a probability associated with incorrectly rejecting the null hypothesis

Let's choose α=0.05

Inferences About Correlation (ρ_{YX})

Determine the "critical value" … that is, how large the test statistic must be in order to reject the null hypothesis at the given α level

Given α=0.05 for this two-sided test, the critical value Z* equals 1.96

Inferences About Correlation (ρ_{YX})

Calculate the test statistic

$$Z_r = \left(\frac{1}{2}\right)\ln\left(\frac{1+r_{yx}}{1-r_{yx}}\right) = \left(\frac{1}{2}\right)\ln\left(\frac{1-0.193}{1+0.193}\right) = -0.195$$

$$Z = \frac{Z_r - 0}{\sqrt{1/n-3}} = \frac{-0.195-0}{\sqrt{1/15{,}354}} = -24.2$$

Inferences About Correlation (ρ_{YX})

Compare the test statistic to the critical value

- If the test statistic is larger than the critical value, then reject H_0
- If the test statistic is less than or equal to the critical value, then do no reject H_0

We can restate the hypotheses:

H_0: $\rho_{YX} = 0$ ➔ Fail to reject H_0 if $|Z| \leq 1.96$

H_1: $\rho_{YX} \neq 0$ ➔ Reject H_0 if $|Z| > 1.96$

Since Z=-24.2, we reject H_0 … so it appears that in the population X and Y are correlated

Worksheet

Mean of X:	6.50	SD of X:	2.95
Mean of Y:	7.15	SD of Y:	1.46
r_{XY}:	0.51		
n:	20		

Test the hypothesis that ρ_{YX} --- the population correlation between X and Y --- is zero in the population; use α=0.05

Inferences About Slope (β_{YX})

Inferences About Slope (β_{YX})

The formula for the sample prediction equation is:

$$\hat{Y}_i = a + b_{YX} X_i$$

The formula for the population prediction equation is:

$$\hat{Y}_i = \alpha + \beta_{YX} X_i \quad \text{(sometimes written as } E(Y_i) = \alpha + \beta_{YX} X_i \text{)}$$

We use b_{YX} as an estimate of β_{YX}

Because b_{YX} is a sample estimate, we know that they would vary from sample to sample if we were to take repeated samples of the same size from the population

Inferences About Slope (β_{YX})

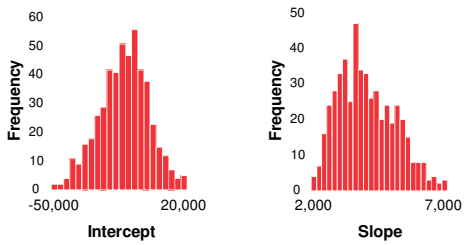
Example: In a population that includes more than 52,000 individuals, the population regression of income (Y) on years of education (X) yields prediction equation:

$$\hat{Y}_i = -18{,}402 + 3{,}892 X_i$$

I drew 523 random samples, each of size n=100, from this population and estimated 523 separate regression models using these sample data

Obviously not all 523 of the intercepts equal -18,402 and not all of the 523 slopes equal 3,892 … sampling variability produces distributions of both

Inferences About Slope (β_{YX})



Inferences About Slope (β_{YX})

The sampling distribution of b_{YX} is normally distributed, centered over β_{YX}, with variance: $\sigma_b^2 = \frac{\sigma_e^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$

where σ^2_e is the population variance of the prediction errors

We can use MS_{ERROR} as a sample estimate of σ^2_e and the denominator can be re-expressed as $(s^2_X)(n-1)$, so the standard error of the sampling distribution of b_{YX} is

$$s_b^2 = \frac{MS_{ERROR}}{(s_x^2)(n-1)}$$

Inferences About Slope (β_{YX})

State the null (H_0) and alternative (H_1) hypotheses

H_0: $\beta_{YX} = 0$
H_1: $\beta_{YX} \neq 0$

This is normally a two-sided test, although it needn't be

Inferences About Slope (β_{YX})

Check that the sample data conform to basic assumptions; if they do not, then do not go any further

The assumptions of the regression model described earlier must hold for hypothesis tests about β_{YX} to be valid

Inferences About Slope (β_{YX})

Choose an α probability level … that is, a probability associated with incorrectly rejecting the null hypothesis

Let's go with α=0.05

Inferences About Slope (β_{YX})

Determine the "critical value" … that is, how large the test statistic must be in order to reject the null hypothesis at the given α level

Because we are using sample-based estimates of the variability in the sampling distribution of b_{YX}, we will conduct a t (instead of a Z) test

Because MS_{ERROR} has n-2 degrees of freedom, we will select a critical value of t with df=n-2

For a two-sided test with α=0.05 and n-2=15,355 degrees of freedom, the critical value t*=1.96

Inferences About Slope (β_{YX})

Calculate the test statistic

The test statistic t with df=N-2 equals:

$$t_{n-2} = \frac{b_{YX} - 0}{s_b} = \frac{b_{YX} - 0}{\sqrt{\frac{MS_{ERROR}}{(s_X^2)(n-1)}}}$$

In our example:

$$t_{n-2} = \frac{-0.178 - 0}{\sqrt{\frac{4.368}{(2.31^2)(15{,}356)}}} = -24.379$$

Inferences About Slope (β_{YX})

Compare the test statistic to the critical value

- If the test statistic is larger than the critical value, then reject H_0
- If the test statistic is less than or equal to the critical value, then do no reject H_0

We can restate the hypotheses:

H_0: $\beta_{YX} = 0$ ➔ Fail to reject H_0 if $|t| \leq 1.96$

H_1: $\beta_{YX} \neq 0$ ➔ Reject H_0 if $|t| > 1.96$

Since t = -24.379, we reject H_0

Worksheet

Mean of X:	6.50	SD of X:	2.95
Mean of Y:	7.15	SD of Y:	1.46
r_{XY}:	0.51		
n:	20		

Test the hypothesis that β_{YX} --- the population slope relating Y to X --- is zero in the population; use α=0.05

Bonus: Confidence Intervals for β_{YX}

Inferences About Slope (β_{YX})

Using the sample estimate (b_1) of β_{XY} and the standard error of the sampling distribution of β_{YX} (above), we can compute confidence intervals for β_{YX}

The standard error is:

$$s_b = \sqrt{\frac{MS_{ERROR}}{(s_x^2)(n-1)}}$$

So the confidence interval can be expressed as:

$$C.I. = b_1 \pm t^* \sqrt{\frac{MS_{ERROR}}{(s_x^2)(n-1)}}$$

Inferences About Slope (β_{YX})

For our example, a 95% confidence interval would be:

$$C.I. = b_1 \pm 1.96 \sqrt{\frac{MS_{ERROR}}{(s_x^2)(n-1)}} = -0.178 \pm 1.96 \sqrt{\frac{4.368}{(2.31^2)(15{,}356)}}$$

$$C.I. = -0.178 \pm 1.96(0.0074)$$

$$C.I. = -0.178 \pm 0.014$$

…so we are 95% certain that β_{YX} falls in within the interval -0.192 to -0.164

Worksheet

Mean of X:	6.50	SD of X:	2.95
Mean of Y:	7.15	SD of Y:	1.46
r_{XY}:	0.51		
n:	20		

Construct a 95% confidence interval for β_{YX} --- the population slope relating Y to X

Standardized Regression Coefficients

Standardized Coefficients (β^*_{YX})

For a variety of reasons researchers often like to express the slope of the regression line in standardized terms

This is useful when:

- The metric of X is in an arbitrary scale, or a scale that is not intrinsically meaningful
- We want to better understand the magnitude of the association between X and Y

Instead of asking…

- "How many units does Y change as a result of a one unit change in X?"

we might ask,

- "How many standard deviations does Y change as a result of a one standard deviation change in X?"

Standardized Coefficients (β^*_{YX})

The standardized slope, or beta coefficient (or beta weight) is expressed as

$$\beta^*_{YX} = (b_{YX})\left(\frac{s_X}{s_Y}\right)$$

In bivariate regression, the standardized slope thus equals the correlation, r_{YX}

In our example:

$$\beta^*_{YX} = (-0.178)\left(\frac{2.31}{2.13}\right) = -0.193$$

Want More?

David Lane's Books
- http://onlinestatbook.com/2/regression/regression.html
- http://davidmlane.com/hyperstat/prediction.html

Stat Trek
- http://stattrek.com/regression/linear-regression.aspx

Lowry's Book (Chapter 3)
- http://vassarstats.net/textbook/

Dallal's Book (see "Simple Linear Regression" section)
- http://www.jerrydallal.com/LHSP/LHSP.htm