Chapter 6

Bivariate Correlation & Regression

6.1 Scatterplots and Regression Lines

6.2 Estimating a Linear Regression Equation

6.3 R-Square and Correlation

6.4 Significance Tests for Regression Parameters

Scatterplot: a positive relation

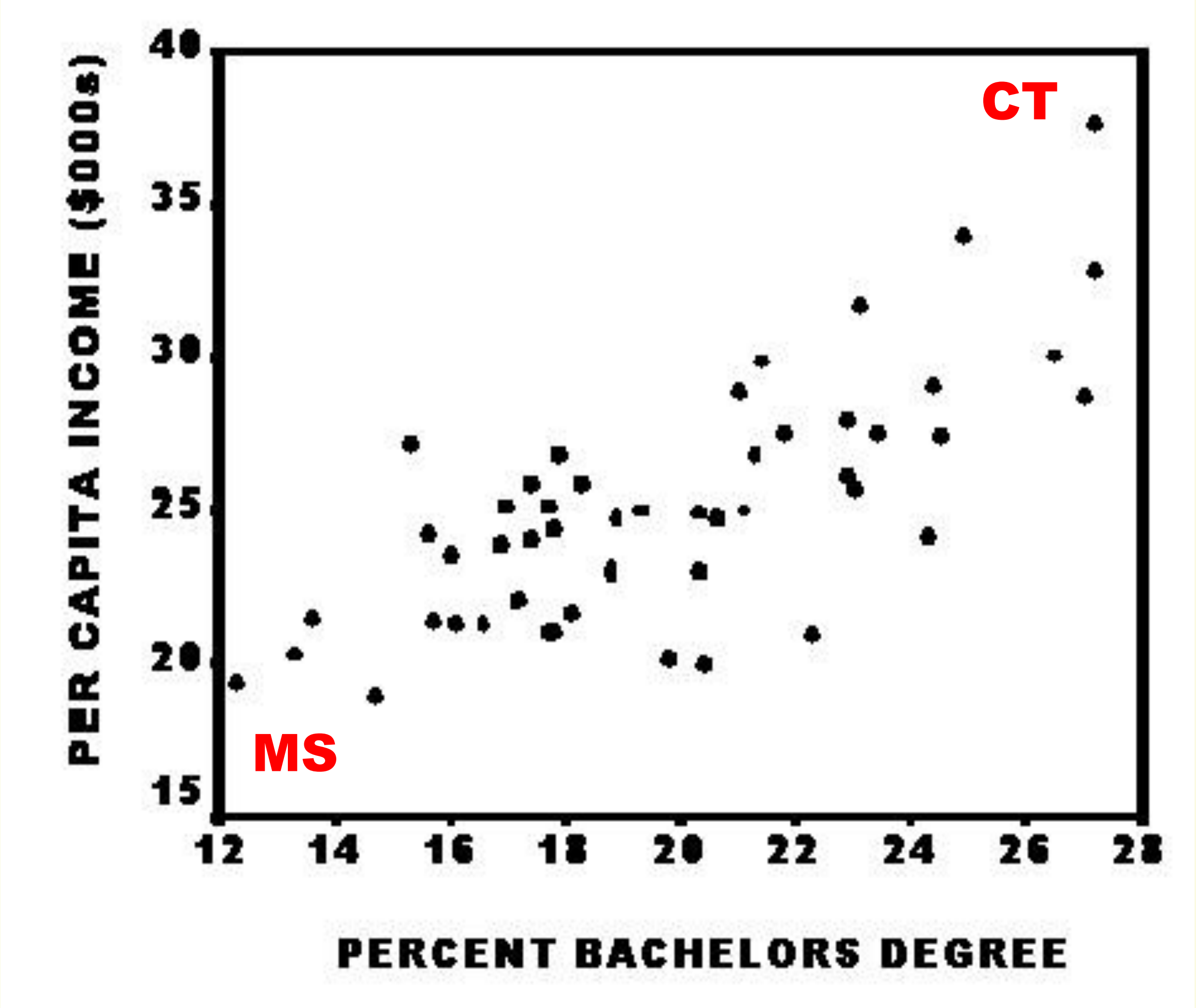
Visually display relation of two variables on X-Y coordinates

<u>50 U.S. States</u>

Y = per capita income

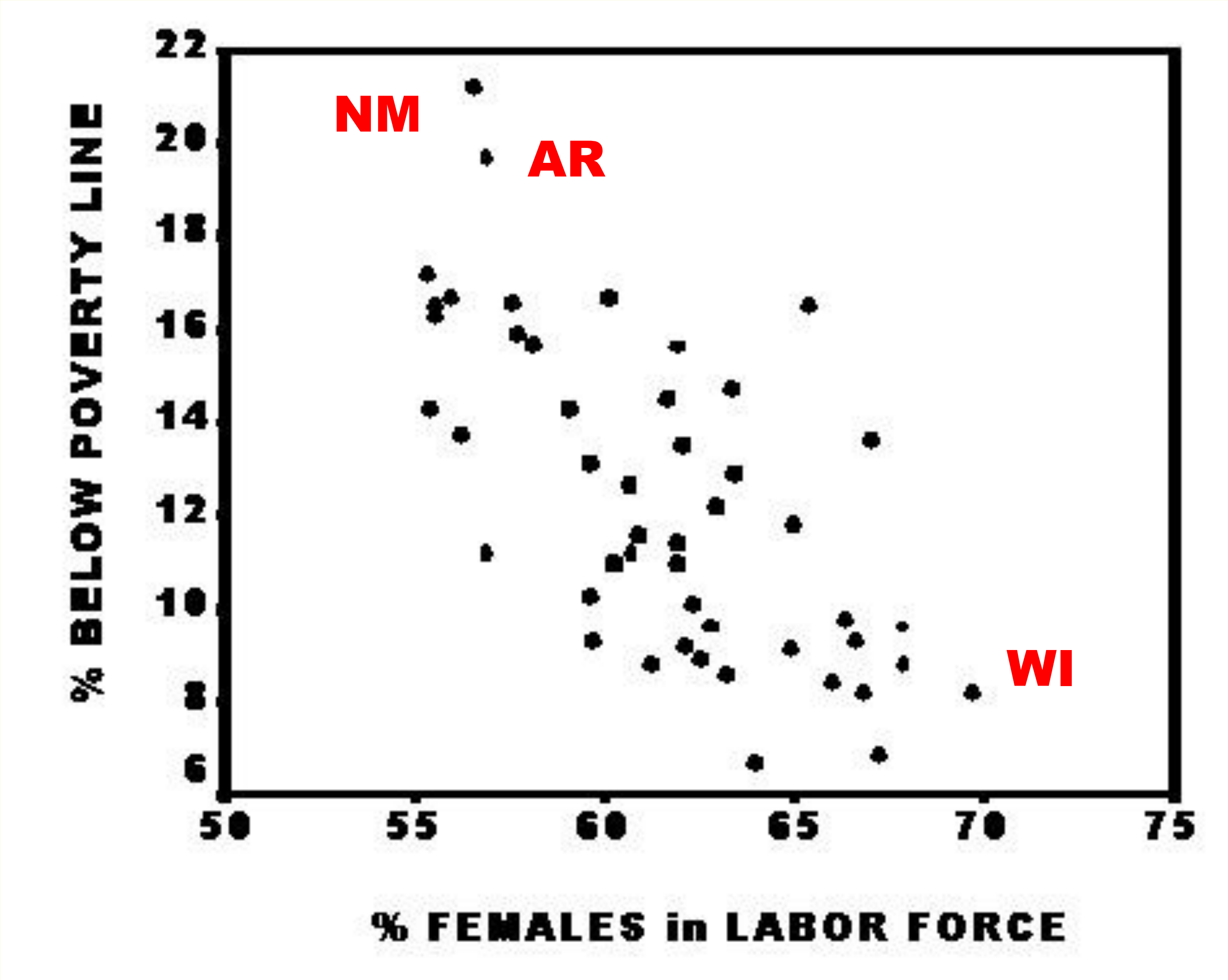
X = % adults with BA degree

Positive relation: increasing X related to higher values of Y



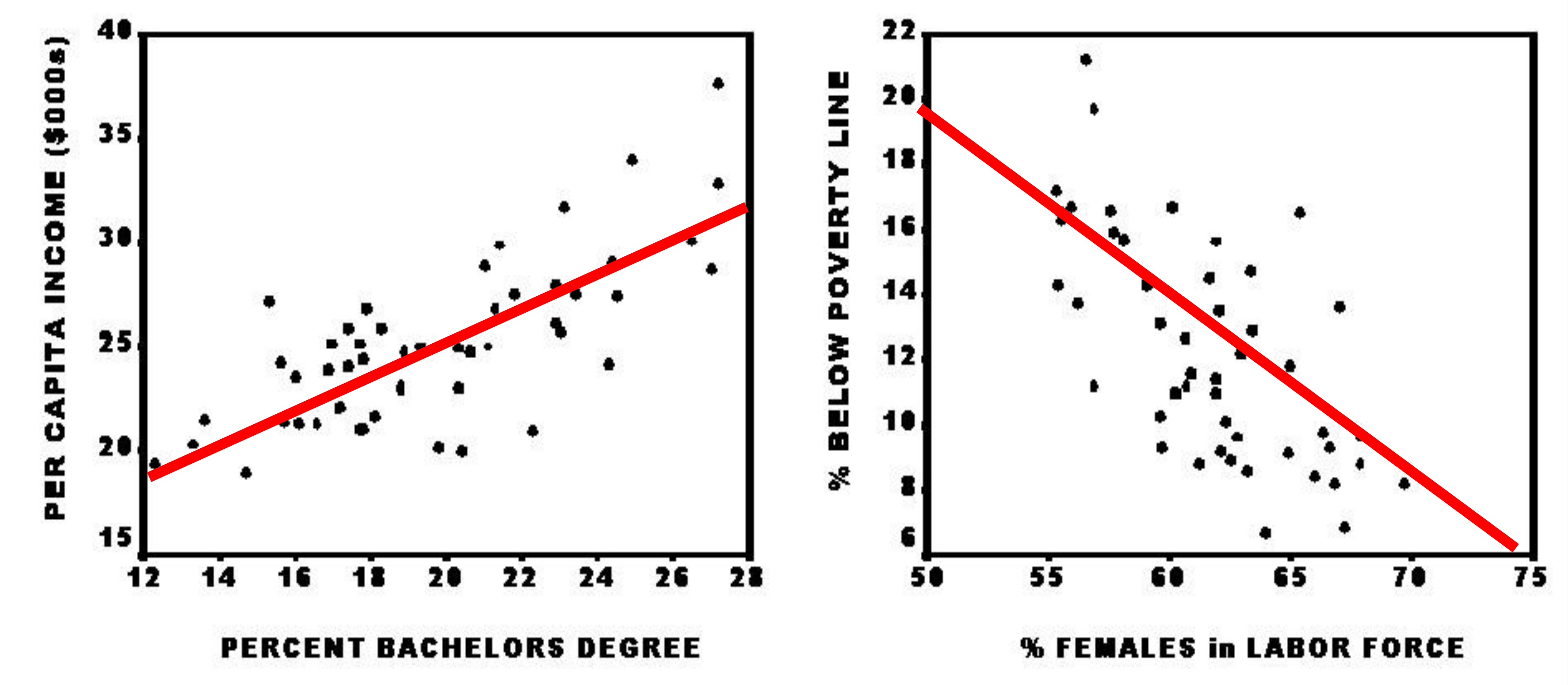
Scatterplot: a negative relation

Y = % in poverty X = % females in labor force



Summarize scatter by regression line

Use linear regression to estimate "best-fit" line thru points:



How can we use sample data on the Y & X variables to estimate population parameters for the best-fitting line?

Slopes and intercepts

We learned in algebra that a line is uniquely located in a coordinate system by specifying: (1) its slope ("rise over run"); and (2) its intercept (where it crosses the Y-axis)

Equation has a bivariate linear relationship:

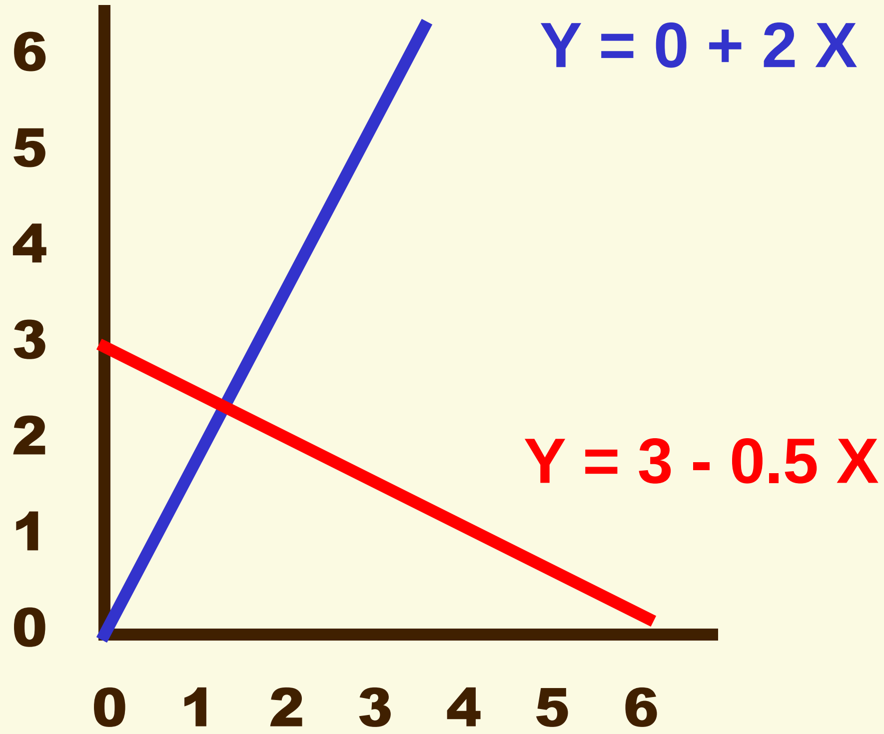
Y = a + bX

where:

b is slope

a is intercept

DRAW THESE 2 LINES:



Prediction equation vs. regression model

In prediction equation, caret over Y_i indicates <u>predicted</u> ("expected") score of *i*th case for independent value X_i :

$$\hat{\mathbf{Y}}_{\mathbf{i}} = \mathbf{a} + \mathbf{b}_{\mathbf{YX}}\mathbf{X}_{\mathbf{i}}$$

But we can never perfectly predict social relationships!

Regression model's error term indicates <u>how discrepant</u> is the predicted score from observed value of the *i*th case:

$$\mathbf{Y}_{\mathbf{i}} = \mathbf{a} + \mathbf{b}_{\mathbf{YX}}\mathbf{X}_{\mathbf{i}} + \mathbf{e}_{\mathbf{i}}$$

Calculate the magnitude and sign of the i^{th} case's error by subtracting 1^{st} equation from 2^{nd} equation (see next slide):

$$\mathbf{Y}_{\mathbf{i}} - \hat{\mathbf{Y}}_{\mathbf{i}} = \mathbf{e}_{\mathbf{i}}$$

Regression error

The regression error, or residual, for the ith case is the difference between the value of the dependent variable predicted by a regression equation and the observed value of that case.

Subtract the prediction equation from the linear regression model to identify the ith case's error term

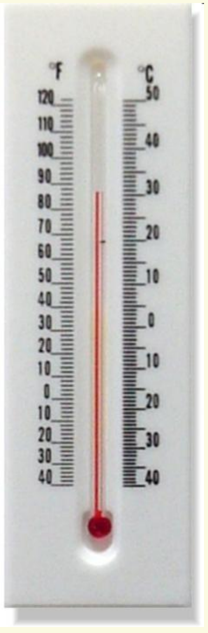
$$Y_i = a + b_{YX}X_i + e_i$$

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

$$Y_i - \hat{Y}_i = e_i$$

An analogy: In weather forecasting, an error is the difference between the weatherperson's predicted high temperature for today and the actual high temperature observed today:

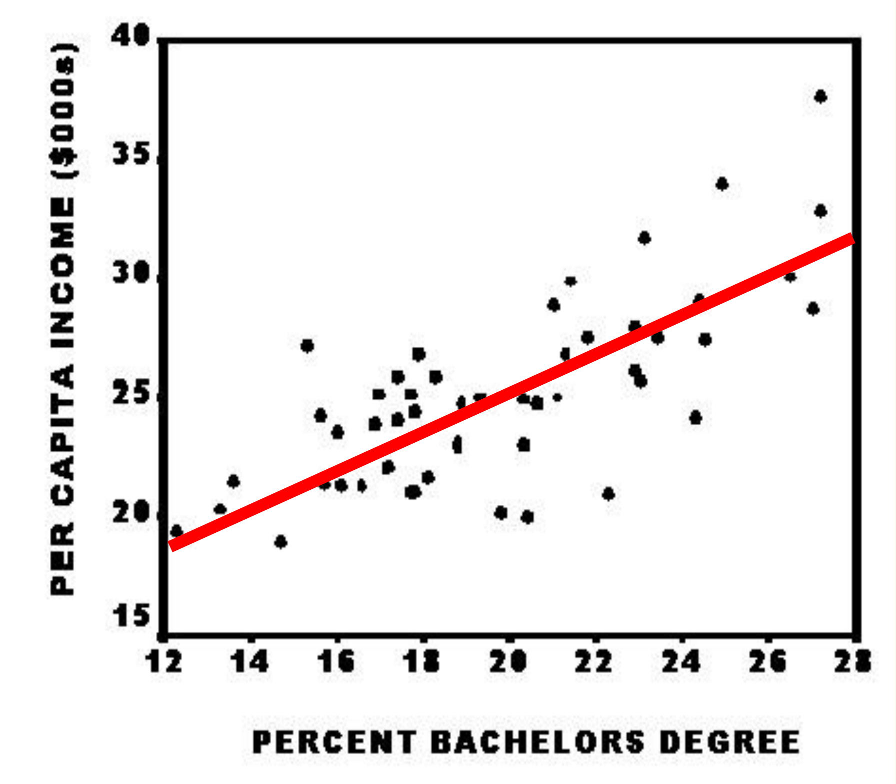
Observed temp 86º - Predicted temp 91º = Error -5º



The Least Squares criterion

Scatterplot for state Income & Education has a positive slope

To plot the regression line, we apply a criterion yielding the “best fit” of a line through the cloud of points



Ordinary least squares (OLS) a method for estimating regression equation coefficients -- intercept (a) and slope (b) -- that minimize the sum of squared errors

OLS estimator of the slope, *b*

Because the sum of errors is always 0, we want parameter estimators that will minimize the sum of squared errors:

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

Fortunately, both OLS estimators have this desired property

Bivariate regression coefficient:

$$b_{YX} = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$$

Numerator is sum of product of deviations around means; when divided by *N* – 1 it's called the covariance of Y and X.

If we also divide the denominator by *N* – 1, the result is the now-familiar variance of X.

Thus,

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

OLS estimator of the intercept, *a*

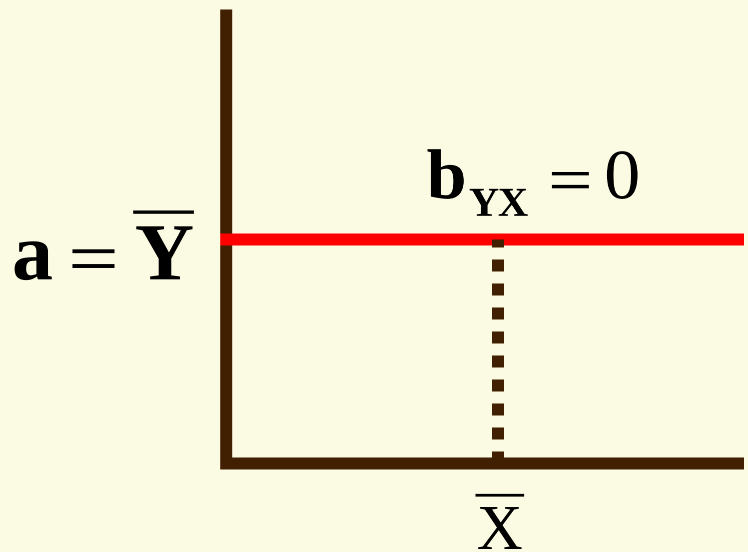
The OLS estimator for the intercept (a) simply changes the mean of Y (the dependent variable) by an amount equaling the regression slope's effect for the mean of X:

Two important facts arise from this relation:

$$\mathbf{a} = \overline{\mathbf{Y}} - \mathbf{b}\overline{\mathbf{X}}$$

(1) The regression line always goes through the point of both variables' means!

(2) When the regression slope is zero, for every X we only predict that Y equals the intercept *a*, which is also the mean of the dependent variable!



Use these two bivariate regression equations, estimated from the 50 States data, to calculate some predicted values:

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

1. Regress income on bachelor's degree:

$$\mathbf{\hat{Y}_i} = \$9.9 + 0.77\ \mathbf{X_i}$$

What predicted incomes for:

X_i = 12%: Y=_______________

X_i = 28%: Y=_______________

2. Regress poverty percent on female labor force pct:

$$\mathbf{\hat{Y}_i} = 45.2\% - 0.53\ \mathbf{X_i}$$

What predicted poverty % for:

X_i = 55%: Y=_______________

X_i = 70%: Y=_______________

Use these two bivariate regression equations, estimated from the 2008 GSS data, to calculate some predicted values:

$$\hat{\mathbf{Y}}_{\mathbf{i}} = \mathbf{a} + \mathbf{b}_{\mathbf{YX}} \mathbf{X}_{\mathbf{i}}$$

3. Regress church attendance per year on age (N=2,005)

$$\hat{Y}_i = 8.34 + 0.28\ X_i$$

What predicted attendance for:

X_i = 18 years: Y=____________

X_i = 89 years: Y=____________

4. Regress sex frequency per year on age (N=1,680)

$$\hat{Y}_i = 121.44 - 1.46\ X_i$$

What predicted activity for:

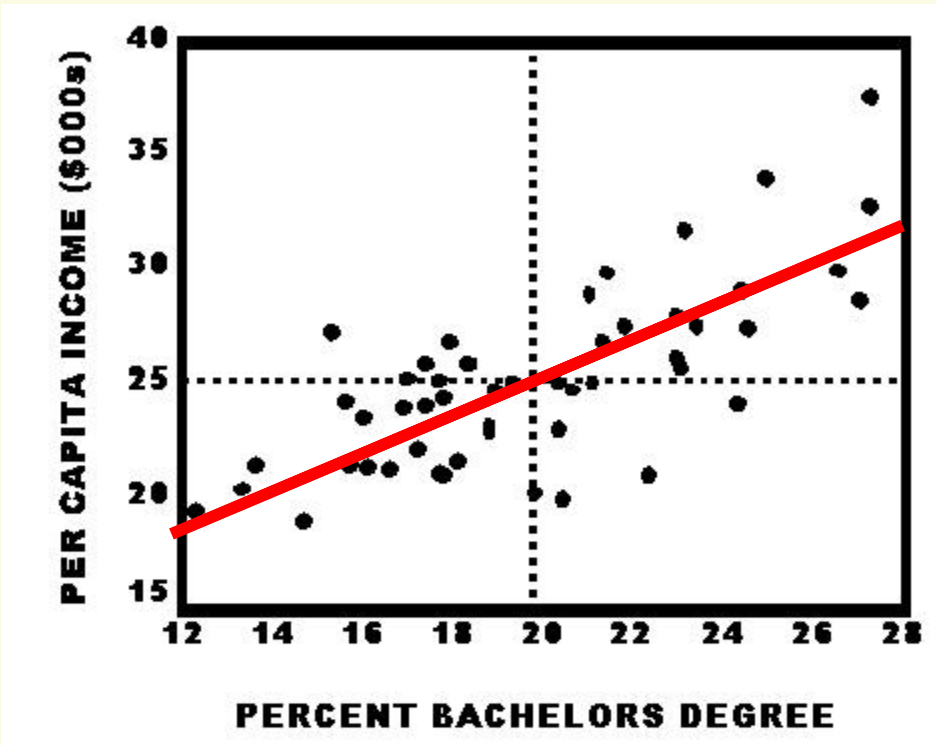
X_i = 18 years: Y=____________

X_i = 89 years: Y=____________

Linearity is not always a reasonable, realistic assumption to make about social behaviors!

Errors in regression prediction

Every regression line through a scatterplot also passes through the <u>means</u> of both variables; i.e., point $(\bar{\mathbf{Y}}, \bar{\mathbf{X}})$



We can use this relationship to divide the <u>variance of Y</u> into a double deviation from:

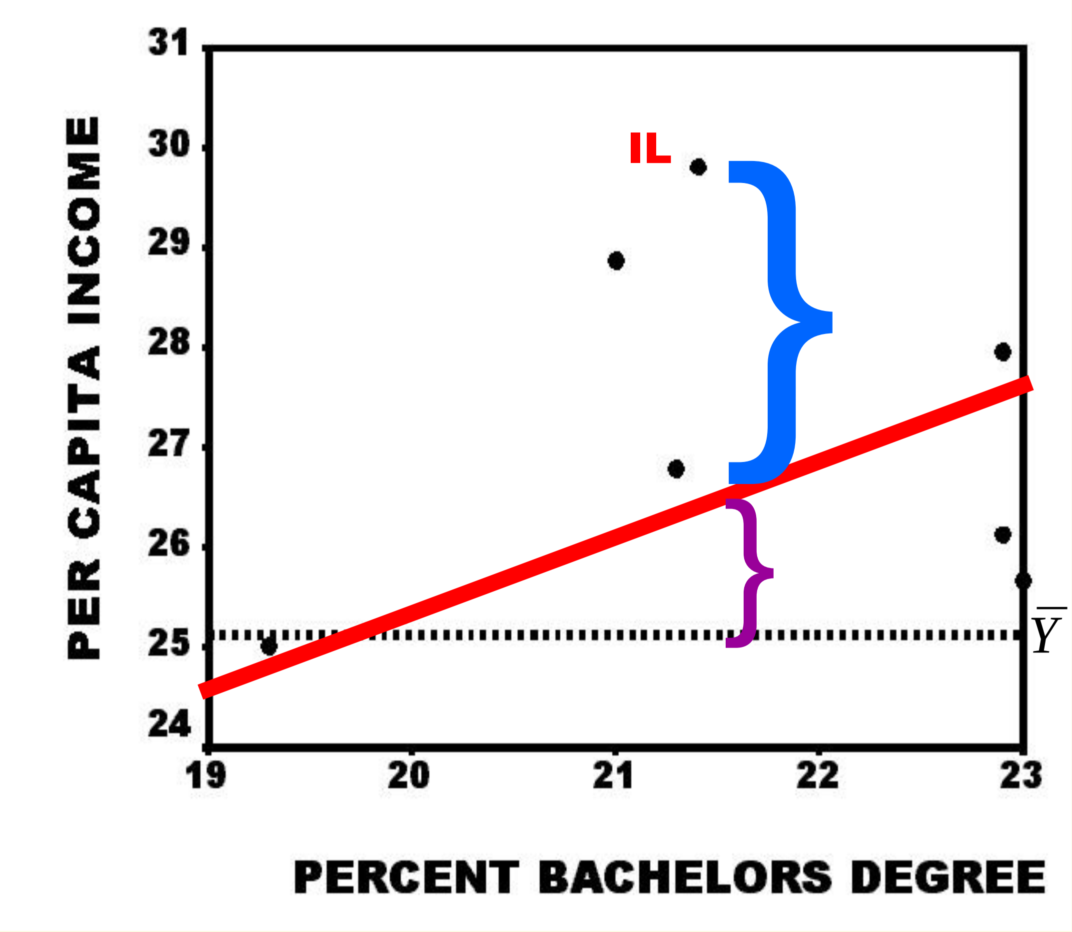
(1) the regression line

(2) the Y-mean line

Then calculate a sum of squares that reveals <u>how strongly</u> Y is predicted by X.

Illinois double deviation

In Income-Education scatterplot, show the difference between the mean and Illinois' Y-score as the sum of two deviations:



Error deviation of observed and predicted scores

$$= \mathbf{Y_i} - \mathbf{\hat{Y}_i}$$

Regression deviation of predicted score from the mean

$$= \mathbf{\hat{Y}_i} - \mathbf{\bar{Y}}$$

Partitioning the sum of squares

Now generalize this procedure to all *N* observations

1. Subtract the mean of Y from the *i*th observed score (= case i's deviation score): $\mathbf{Y_i} \quad - \mathbf{\overline{Y}}$

2. Simultaneously subtract *and* add *i*th predicted score (leaves the deviation unchanged): $Y_i \quad - \hat{Y}_i + \hat{Y}_i \quad - \overline{Y}$

3. Group these four elements into two terms: $(Y_i \quad - \hat{Y}_i) + (\hat{Y}_i \quad - \overline{Y})$

4. Square both grouped terms: $(Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \overline{Y})^2$

5. Sum the squares across all *N* cases: $\sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \overline{Y})^2$

6. Step #5 equals the sum of the squared deviations in step #1 (which is also the numerator of the variance of Y): $\sum (Y_i - \overline{Y})^2$

Therefore: $$\sum (Y_i - \overline{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \overline{Y})^2$$

Naming the sums of squares

Each result of the preceding partition has a name:

$$\sum(\mathbf{Y_i} - \overline{\mathbf{Y}})^2 = \sum(\mathbf{Y_i} - \hat{\mathbf{Y}}_\mathbf{i})^2 + \sum(\hat{\mathbf{Y}}_\mathbf{i} - \overline{\mathbf{Y}})^2$$

TOTAL sum of squares

ERROR sum of squares

REGRESSION sum of squares

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

The relative proportions of the two terms on the right indicate how well or poorly we can predict the variance in Y from its linear relationship with X

The SS_{TOTAL} should be familiar to you – it's the numerator of the variance of Y (see the Notes for Chapter 2). When we partition the sum of squares into the two components, we're analyzing the variance of the dependent variable in a regression equation.

Hence, this method is called **the analysis of variance** or **ANOVA**.

Coefficient of Determination

If we had no knowledge about the regression slope (i.e., $b_{YX} = 0$ and thus $SS_{REGRESSION} = 0$), then our only prediction is that the score of Y <u>for every case</u> equals the mean (which also equals the equation's intercept **a**; see slide #10 above).

$$\hat{\mathbf{Y}}_{\mathbf{i}} = \mathbf{a} + \mathbf{b}_{\mathbf{YX}}\mathbf{X}_{\mathbf{i}}$$

$$\hat{\mathbf{Y}}_{\mathbf{i}} = \mathbf{a} + 0\mathbf{X}_{\mathbf{i}}$$

$$\hat{\mathbf{Y}}_{\mathbf{i}} = \mathbf{a}$$

But, if $b_{YX} \neq 0$, then we can use information about the *i*th case's score on X to improve our predicted Y for case *i*. We'll still make errors, but the stronger the Y-X linear relationship, the more accurate our predictions will be.

R^2 as a PRE measure of prediction

Use information from the sums of squares to construct a standardized proportional reduction in error (PRE) measure of prediction success for a regression equation

This PRE statistic, the coefficient of determination, is the proportion of the variance in Y "explained" statistically by Y's linear relationship with X.

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

$$R^2_{YX} = \frac{\sum (Y_i - \bar{Y})^2 - \sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

The range of R-square is from 0.00 to 1.00, that is, from no predictability to "perfect" prediction.

Find the R^2 for these 50-States bivariate regression equations

1. R-square for regression of income on education

$SS_{REGRESSION}$ = 409.3

SS_{ERROR} = 342.2

SS_{TOTAL} = 751.5

R^2 = ___________

2. R-square for poverty-female labor force equation

$SS_{REGRESSION}$ = _____

SS_{ERROR} = 321.6

SS_{TOTAL} = 576.6

R^2 = ___________

Here are some R^2 problems from the 2008 GSS

3. R-square for church attendance regressed on age

$SS_{REGRESSION}$ = 67,123

SS_{ERROR} = 2,861,928

SS_{TOTAL} = ________

R^2 = ____________

4. R-square for sex frequency-age equation

$SS_{REGRESSION}$ = 1,511,622

SS_{ERROR} = ____________

SS_{TOTAL} = 10,502,532

R^2 = ____________

The correlation coefficient, r

Correlation coefficient is a measure of the direction and strength of the linear relationship of two variables

Attach the sign of regression slope to square root of R^2:

$$\mathbf{r_{YX}} = \mathbf{r_{XY}} = \sqrt{\mathbf{R}^2_{\mathbf{YX}}}$$

Or, in terms of covariances and standard deviations:

$$\mathbf{r_{YX}} = \frac{\mathbf{s_{YX}}}{\mathbf{s_Y s_X}} = \frac{\mathbf{s_{XY}}}{\mathbf{s_X s_Y}} = \mathbf{r_{XY}}$$

Calculate the correlation coefficients for these pairs:

Regression Eqs.	R^2	b_{YX}	r_{YX}
Income-Education	0.55	+0.77	
Poverty-labor force	0.44	-0.53	
Church attend-age	0.018	+0.19	
Sex frequency-age	0.136	-1.52	

Comparison of r and R^2

This table summarizes differences between the correlation coefficient and coefficient of determination for two variables.

	Correlation Coefficient	Coefficient of Determination
Sample statistic	r	R^2
Population parameter	ρ	ρ^2
Relationship	$r^2 = R^2$	$R^2 = r^2$
Test statistic	*t* test	*F* test

Sample and population

Regression equations estimated with sample data can be used to test hypotheses about each of the three corresponding population parameters

Sample equation: $\hat{Y}_i = a + b_{YX} X_i$ R^2_{YX}

Population equation: $\hat{Y}_i = \alpha + \beta_{YX} X_i$ ρ^2_{YX}

Each pair of null and alternative (research) hypotheses are statements about a population parameter. Performing a significance test requires using sample statistics to estimate a standard error or a pair of mean squares.

Hypotheses about slope, β

A typical null hypothesis about the population regression slope is that the independent variable (X) has <u>no linear relation</u> with the dependent variable (Y).

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

Its paired research hypothesis is nondirectional (a two-tailed test):

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

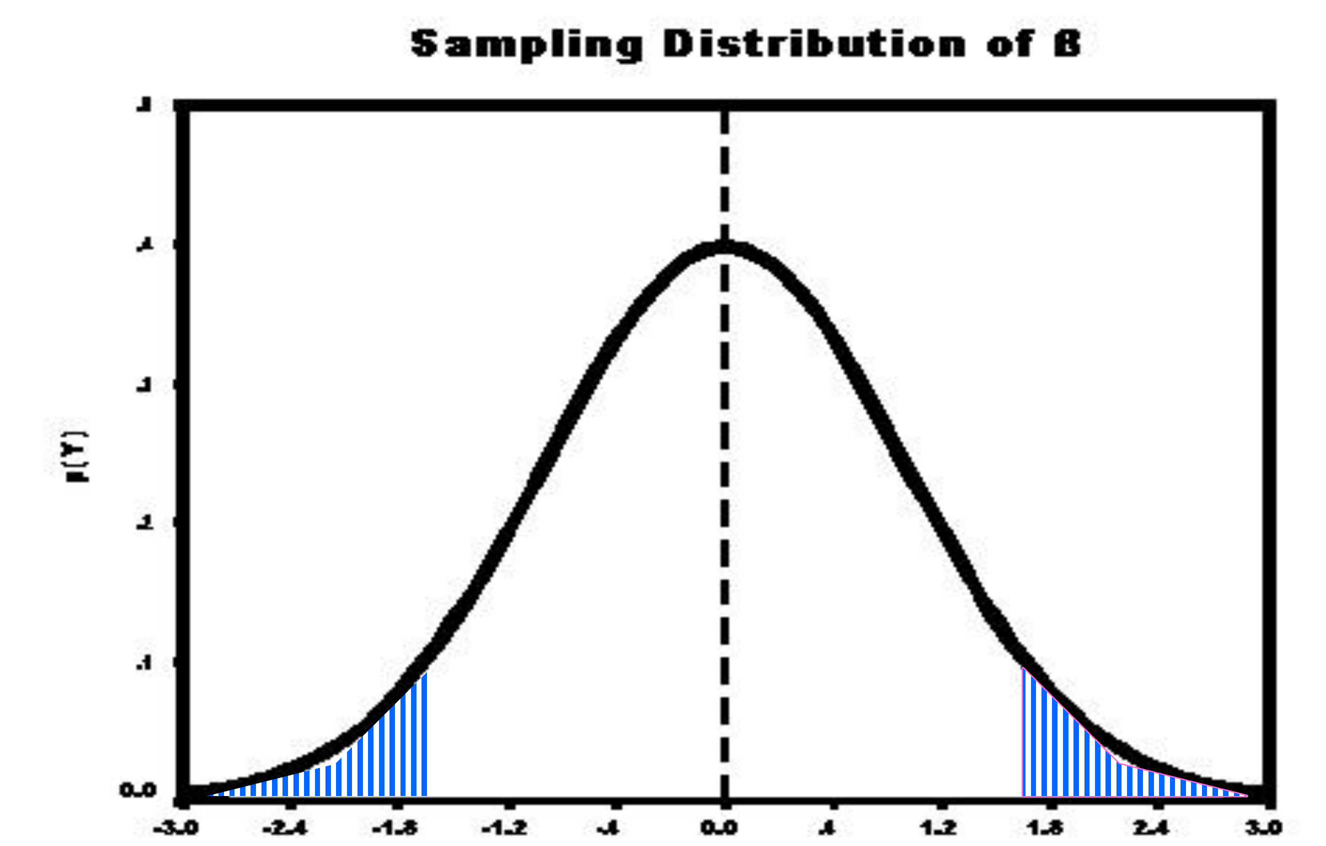
Other hypothesis pairs are directional (one-tailed tests):

$$H_0: \beta_{YX} \leq 0 \qquad \text{or} \qquad H_0: \beta_{YX} \geq 0$$

$$H_1: \beta_{YX} > 0 \qquad\qquad\quad H_1: \beta_{YX} < 0$$

Sampling Distribution of β

The Central Limit Theorem, which let us analyze the sampling distribution of large-sample means as a normal curve, also treats the sampling distribution of β as normal, with mean β = 0 and standard error σ_β. Hypothesis tests may be one- or two-tailed.



$\beta = 0$

The t-test for β

To test whether a large sample's regression slope (b_{YX}) has a low probability of being drawn from a sampling distribution with a hypothesized population parameter of zero ($\beta_{YX} = 0$), apply a t-test (same as Z-test for large *N*).

$$t = \frac{b_{YX} - \beta_{YX}}{s_b}$$

where s_b is the sample estimate of the standard error of the regression slope.

SSDA#4 (pp. 192) shows how to calculate this estimate with sample data. But, in this course we will rely on SPSS to estimate the standard error.

Here is a research hypothesis: The greater the percentage of college degrees, the higher a state's per capita income.

1. Estimate the regression equation (s_b in parens):

$$\hat{Y}_i = \$9.9 + 0.77\ X_i$$
$$(2.1) \quad (0.10)$$

2. Calculate the test statistic:

$$t = \frac{b_{YX} - \beta_{YX}}{s_b} = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

3. Decide about the null hypothesis (one-tailed test):

____

α	1-tail	2-tail
.05	**1.65**	**±1.96**
.01	**2.33**	**±2.58**
.001	**3.10**	**±3.30**

4. Probability of Type I error:

____

5. Conclusion: ____________________________________

For this research hypothesis, use the 2008 GSS (*N*=1,919):
The more siblings respondents have, the lower their occupational prestige scores.

1. Estimate the regression equation (s_b in parentheses):

$$\hat{Y}_i = 46.87 - 0.85\,X_i$$
$$(0.47) \quad (0.10)$$

2. Calculate the test statistic:

$$t = \frac{b_{YX} - \beta_{YX}}{s_b} = \underline{\hspace{10cm}}$$

3. Decide about the null

hypothesis (one-tailed test): ____________________________

4. Probability of Type I error: ____________________________

5. Conclusion:

____

Research hypothesis: The number of hours people work per week is unrelated to number of siblings they have.

1. Estimate the regression equation (s_b in parentheses):

$$\hat{Y}_i = 41.73 + 0.08\,X_i$$
$$(0.65) \quad (0.14)$$

2. Calculate the test statistic:

$$t = \frac{b_{YX} - \beta_{YX}}{s_b} = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

3. Decide about the null

hypothesis (two-tailed test): ________________________

4. Probability of Type I error: ________________________

5. Conclusion:

____

Hypothesis about the intercept, α

Researchers rarely have any hypothesis about the population intercept (the dependent variable's predicted score when the independent variable = 0).

Use SPSS's standard error for a t-test of this hypothesis pair:

$$H_0 : \alpha = 0$$

$$H_1 : \alpha \neq 0$$

$$t = \frac{a - \alpha}{s_a}$$

Test this null hypothesis: the intercept in the state income-education regression equation is zero.

$$t = \frac{a - \alpha}{s_a} = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

Decision about H_0 (two-tailed): ____________________

Probability of Type I error: ________________________

Conclusion: ____

Chapter 3

3.11 The Chi-Square and F Distributions

Chi-Square

Two useful families of theoretical statistical distributions, both based on the Normal distribution:

Chi-square and *F* distributions

The Chi-square (χ^2) family: for ν normally distributed random variables, square and add each Z-score

ν (Greek nu) is the degrees of freedom (df) for a specific χ^2 family member

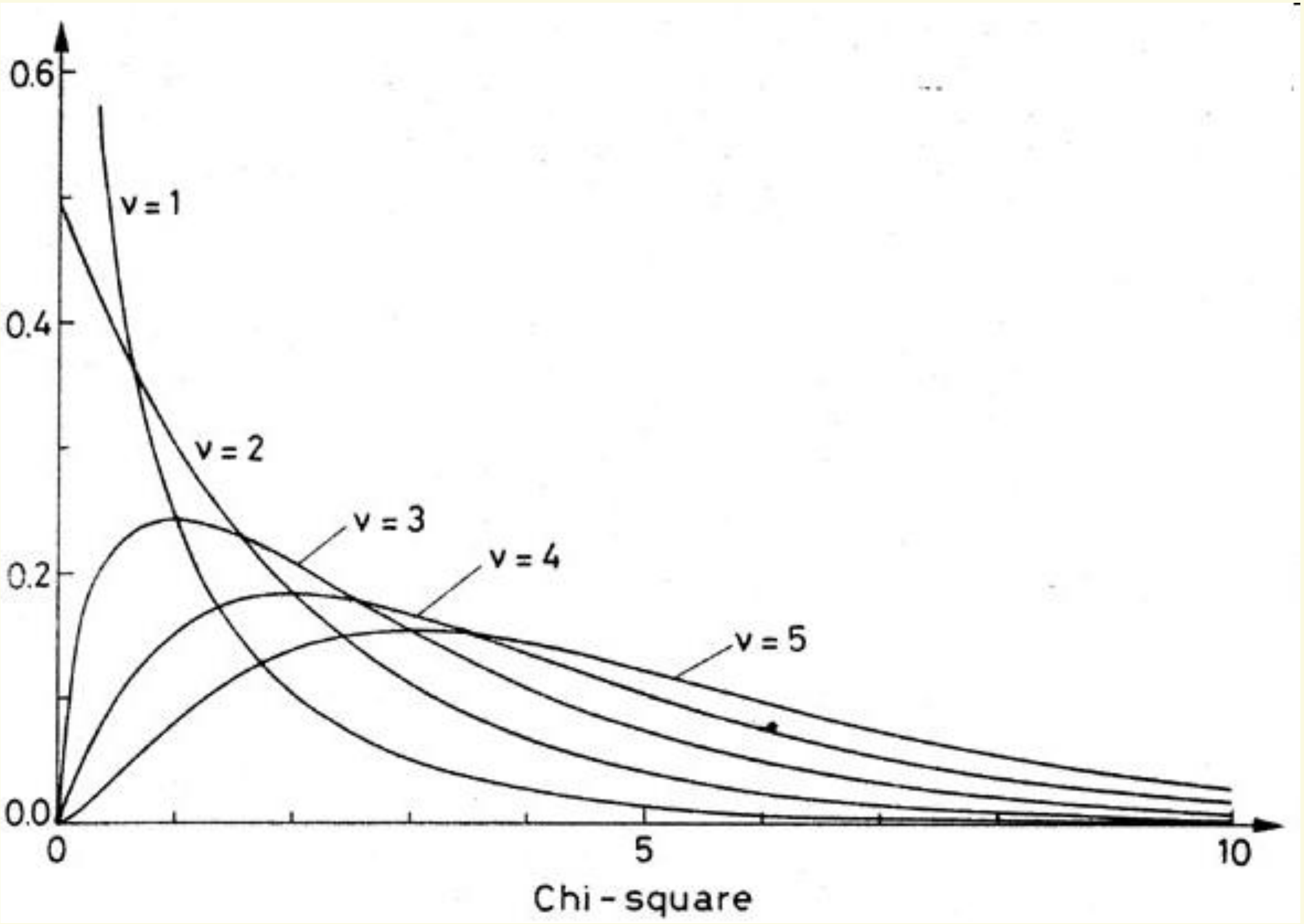
For $\nu = 2$:

$$\mathbf{Z}_1^2 = \frac{(\mathbf{Y}_1 - \boldsymbol{\mu}_{\mathbf{Y}})^2}{\boldsymbol{\sigma}_{\mathbf{Y}}^2} \qquad \mathbf{Z}_2^2 = \frac{(\mathbf{Y}_2 - \boldsymbol{\mu}_{\mathbf{Y}})^2}{\boldsymbol{\sigma}_{\mathbf{Y}}^2}$$

$$\boldsymbol{\chi}_{\boldsymbol{\nu}=2}^2 = \mathbf{Z}_1^2 + \mathbf{Z}_2^2$$

Shapes of Chi-Square

Mean for each $\chi^2 = \nu$ and variance = 2ν. With larger *df*, plots show increasing symmetry but each is positively skewed:



Areas under a curve can be treated as probabilities

The F Distribution

The *F* distribution family: formed as the ratio of two independent chi-square random variables.

Ronald Fischer, a British statistician, first described the distribution 1922. In 1934, George Snedecor tabulated the family's values and called it the F distribution in honor of Fischer.

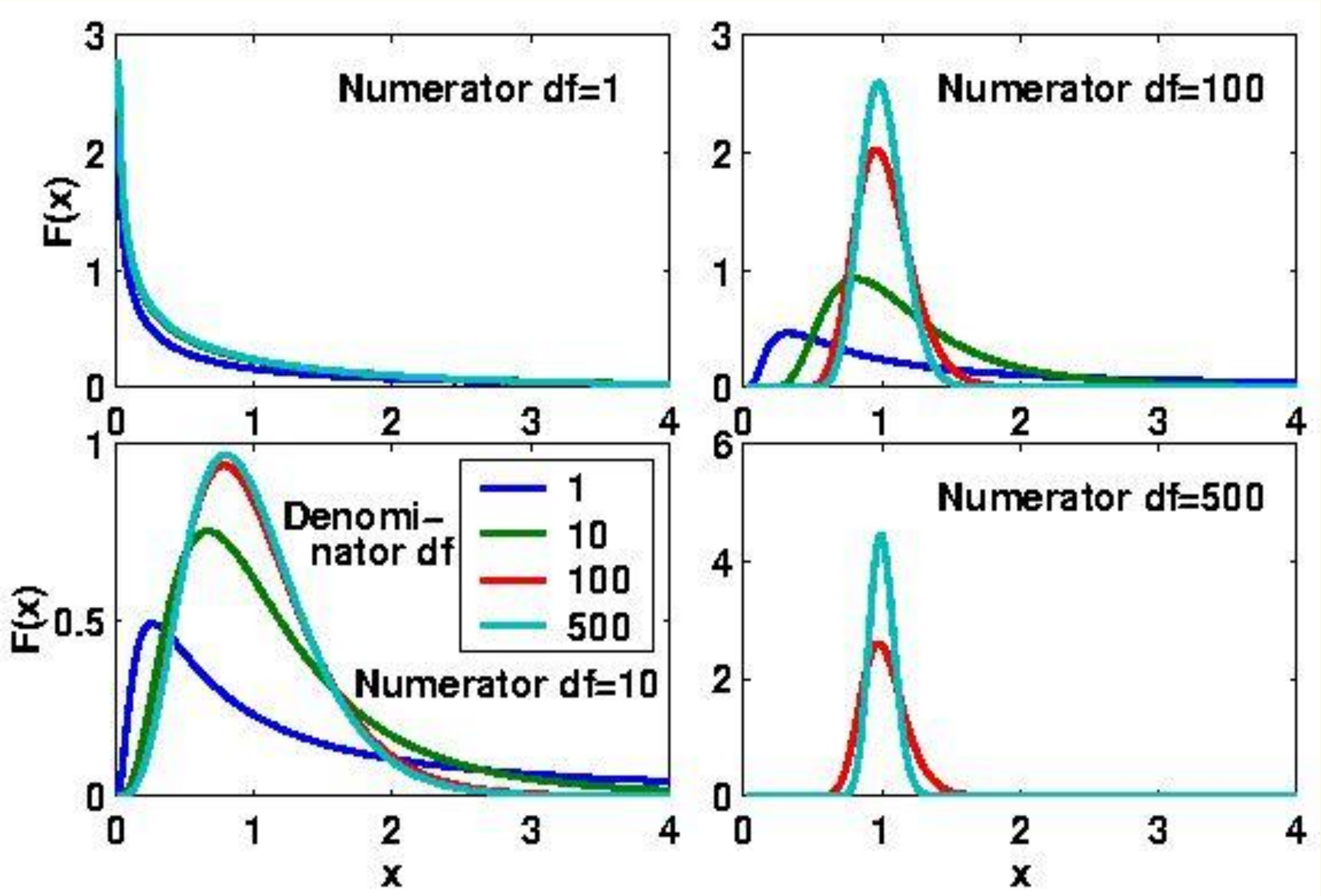
Every member of the *F* family has two degrees of freedom, one for the chi-square in the numerator and one for the chi-square in the denominator:

$$\mathbf{F} = \frac{\chi^2_{\nu_1} / \nu_1}{\chi^2_{\nu_2} / \nu_2}$$

F is used to test hypotheses about whether the variances of two or more populations are equal (analysis of variance = ANOVA)

F is also used in tests of "explained variance" in multiple regression equations (also called ANOVA)

Each member of the F distribution family takes a different shape, varying with the numerator and denominator *df*s:



Chapter 6

Return to hypothesis testing for regression

Hypothesis about ρ^2

A null hypothesis about the population coefficient of determination (Rho-square) is that none of the dependent variable (Y) variation is due to its linear relation with the independent variable (X):

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

The only research hypothesis is that Rho-square in the population is greater than zero:

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

Why is H_1 never written with a <u>negative</u> Rho-square (i.e., $\rho^2 < 0$)?

To test the null hypothesis about ρ^2, use the **F** distribution, a ratio of two chi-squares each divided by their degrees of freedom:

Degree of freedom: the number of values free to vary when computing a statistic

Calculating degrees of freedom

The concept of degrees of freedom (*df*) is probably better understood by an example than by a definition.

Suppose a sample of N =4 cases has a mean of 6.

I tell you that $Y_1 = 8$ and $Y_2 = 5$; what are Y_3 and Y_4?

Those two scores can take <u>many</u> values that would yield a mean of 6 ($Y_3 = 5$ & $Y_4 = 6$; or $Y_3 = 9$ & $Y_4 = 2$)

But, if I now tell you that $Y_3 = 4$, what <u>must</u> Y_4 = _____

Once the mean and *N*-1 other scores are fixed, the *N*th score has <u>no freedom to vary</u>.

The three sums of squares in regression analysis "cost" differing degrees of freedom, which must be "paid" when testing a hypothesis about ρ^2.

df for the 3 Sums of Squares

1. SS_{TOTAL} has $df = N - 1$, because for a fixed total all scores except the final score are free to vary

2. Because the $SS_{REGRESSION}$ is estimated from one regression slope (b_{YX}), it "costs" 1 *df*

3. Calculate the *df* for SS_{ERROR} as the difference:

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

$$N - 1 \quad = \quad 1 \quad + df_{ERROR}$$

Therefore: $df_{ERROR} = N-2$

Mean Squares

To standardize F for different size samples, calculate mean (average) sums of squares per degree of freedom, for the three components

$$\frac{SS_{TOTAL}}{df_{TOTAL}} = \frac{SS_{REGRESSION}}{df_{REGRESSION}} + \frac{SS_{ERROR}}{df_{ERROR}} \Rightarrow \frac{SS_{TOTAL}}{N-1} = \frac{SS_{REGRESSION}}{1} + \frac{SS_{ERROR}}{N-2}$$

Label the two terms on the right side as Mean Squares:

$$SS_{REGRESSION}/1 = MS_{REGRESSION}$$

$$SS_{ERROR}/(N-2) = MS_{ERROR}$$

The **F statistic** is thus a ratio of the two Mean Squares:

$$F = \frac{MS_{REGRESSION}}{MS_{ERROR}}$$

SS_{TOTAL} / df_{TOTAL} = the variance of Y (see the Notes for Chapter 2), further indication we're conducting an analysis of variance (ANOVA).

Analysis of Variance Table

One more time: The F test for 50 State Income-Education

Calculate and fill in the two MS in this summary ANOVA table, and then compute the F-ratio:

Source	SS	df	MS	F
Regression	409.3			
Error	342.2			
Total	751.5		-------------------	

A decision about H_0 requires the critical values for F, whose distributions involve the two degrees of freedom associated with the two Mean Squares

Critical values for F

In a population, if ρ^2 is greater than zero (the H_1), then the $MS_{REGRESSION}$ will be significantly larger than MS_{ERROR}, as revealed by the *F* test statistic.

An F statistic to test a null hypothesis is a ratio two Mean Squares. Each MSs has a different degrees of freedom (*df* = 1 in the numerator, *df* = *N*-2 in the denominator).

For large samples, use this table of critical values for the three conventional alpha levels:

α	df_R, df_E	c.v.
.05	1, ∞	3.84
.01	1, ∞	6.63
.001	1, ∞	10.83

Why are the c.v. for F always positive?

Test the hypothesis about ρ^2 for the occupational prestige-siblings regression, where sample R^2 = 0.038.

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

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

Source	SS	df	MS	F
Regression	14,220			
Error	355,775			
Total	369,995		--------------------	

Decide about null hypothesis: ______________________________

Probability of Type I error: ______________________________

Conclusion: ______________________________

Test the hypothesis about ρ^2 for the hours worked-siblings regression, where sample $R^2 = 0.00027$.

Source	SS	df	MS	F
Regression	68			
Error	251,628			
Total	251,696		--------------------	

Decide about null hypothesis: ______________________

Probability of Type I error: ______________________

Conclusion: ______________________

Will you make <u>always</u> the same or different decisions if you test hypotheses about <u>both</u> β_{YX} and ρ^2 for the same bivariate regression equation? Why or why not?