The Simple Linear Regression Equation

Shane Hutton, Ph.D.
Vanderbilt University

Simple Linear Regression

- Simple linear regression (SLR) is a statistical technique for determining the best-fitting line for a set of bivariate data.
- The best-fitting straight line is called the regression line.

Simple Linear Regression

- The regression line is a mathematical model of the relationship between X and Y.
- It can be used to:
  - Explain the relationship between X and Y
  - Predict the value of Y from a known value of X

Simple Linear Regression

- We can decompose a Y score into:

$$Y = a + bX + e$$

Y = known value of Y; known score on DV

X = known value of X; known score on IV

a = intercept

b = slope

e = random error term (or residual)

Regression Line

- The regression line is represented by the equation:

$$\hat{Y} = a + bX$$

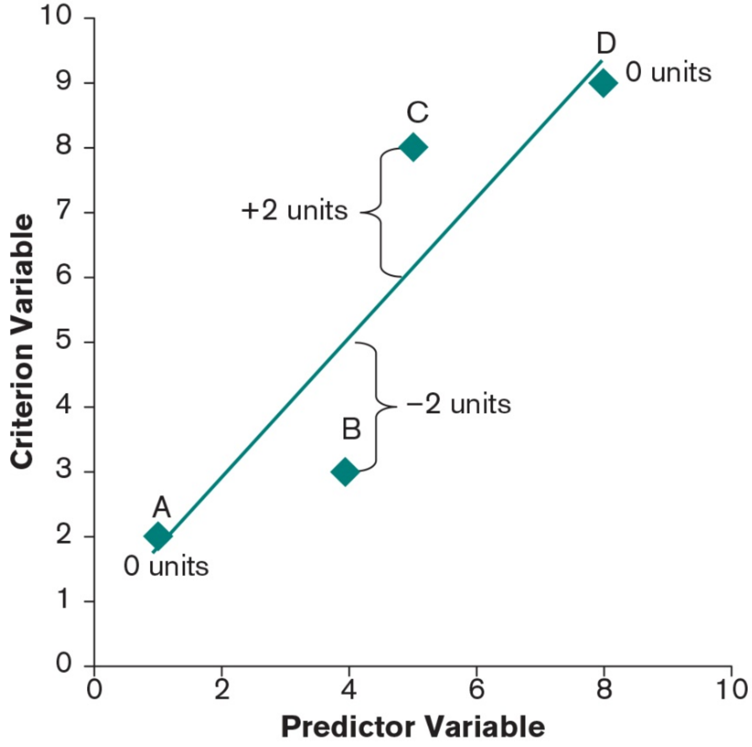
- $\hat{Y}$= predicted value of Y; predicted score on DV
- a = intercept; predicted value of Y when $X = 0$
- b = slope; the predicted change in Y for each unit of increase in X
- Notice:

$$Y = \hat{Y} + e$$

Simple Linear Regression

FIGURE 16.2 A Table and Scatter Plot of Four Hypothetical Data Points

Predictor Variable	Criterion Variable
X	***Y***
1	2
4	3
5	8
8	9



The regression line and the distances, in units, between the data points and the regression line are shown. Both the table and the scatter plot show the same data.

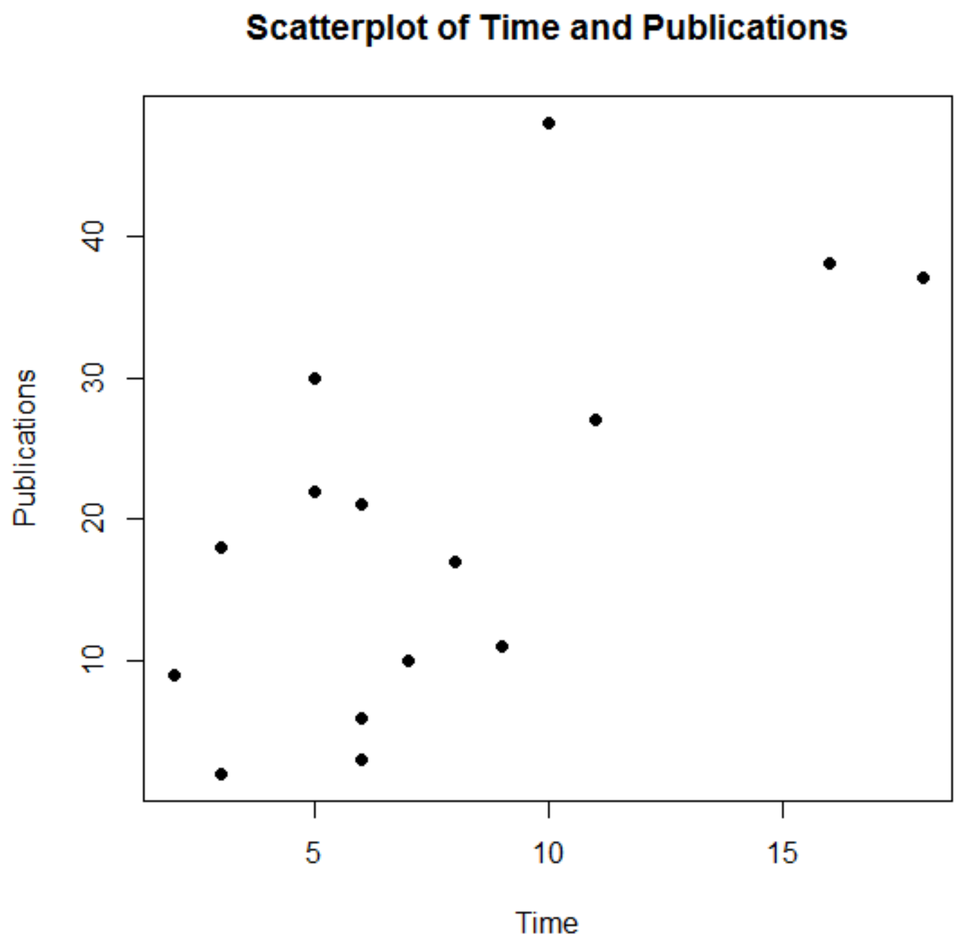
VANDERBILT
PEABODY COLLEGE

Ordinary Least Squares (OLS) Estimation

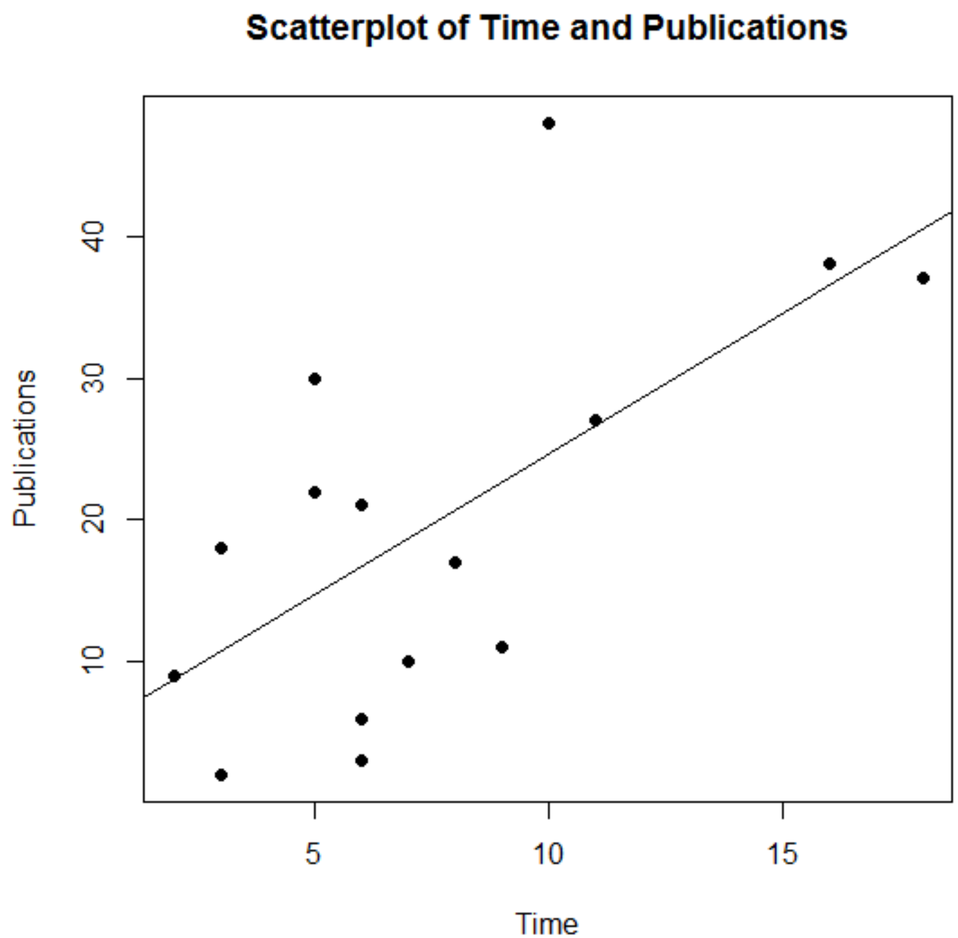
Shane Hutton, Ph.D.

Vanderbilt University

OLS Estimation



OLS Estimation



OLS Estimation

- Ordinary Least Squares (OLS) estimation minimizes the sum of squared errors (SSE), that is, we want to minimize

$$SSE = \sum e^2 = \sum (Y - a - bX)^2$$

OLS Estimation

- For those who have taken calculus, this is a calculus optimization problem
- Minimize:

$$\frac{\partial SSE}{\partial a} = \frac{\partial}{\partial a}\left[\sum (Y - a - bX)^2\right]$$

$$\frac{\partial SSE}{\partial b} = \frac{\partial}{\partial b}\left[\sum (Y - a - bX)^2\right]$$

OLS Estimation

- Using calculus, we obtain

$$b = \frac{\sum[(X - M_X)(Y - M_Y)]}{\sum(X - M_X)^2} = \frac{SS_{XY}}{SS_X}$$

$$a = M_Y - bM_X$$

OLS Estimation

- There is also a mathematical relationship between the correlation and slope for simple linear regression

$$b = r\left(\frac{s_Y}{s_X}\right)$$

- where
  - s_Y is the standard derivation for Y and s_X is the standard deviation for X

VANDERBILT
PEABODY COLLEGE

VANDERBILT
PEABODY COLLEGE

Computing the Regression Line

Shane Hutton, Ph.D.
Vanderbilt University

Computing the Regression Line

- Equations to find the intercept and slope:

$$b = \frac{\sum[(X - M_X)(Y - M_Y)]}{\sum(X - M_X)^2}$$

$$a = M_Y - bM_X$$

- For the example: $M_{time} = 7.67, M_{pubs} = 19.93$

Computing the Regression Line

Professor	Time	Publications	$X - M_x$	$Y - M_y$	$(X - M_x)(Y - M_y)$	$(X - M_x)^2$
1	3	18	-4.67	-1.93	9.02	21.78
2	6	3	-1.67	-16.93	28.22	2.78
3	3	2	-4.67	-17.93	83.69	21.78
4	8	17	0.33	-2.93	-0.98	0.11
5	9	11	1.33	-8.93	-11.91	1.78
6	6	6	-1.67	-13.93	23.22	2.78
7	16	38	8.33	18.07	150.56	69.44
8	10	48	2.33	28.07	65.49	5.44
9	2	9	-5.67	-10.93	61.96	32.11
10	5	22	-2.67	2.07	-5.51	7.11
11	5	30	-2.67	10.07	-26.84	7.11
12	6	21	-1.67	1.07	-1.78	2.78
13	7	10	-0.67	-9.93	6.62	0.44
14	11	27	3.33	7.07	23.56	11.11
15	18	37	10.33	17.07	176.36	106.78

Computing the Regression Line

- Summing up the last two columns
  - $\sum[(X - M_X)(Y - M_Y)] = 581.67$
  - $\sum(X - M_X)^2 = 293.33$
- Plugging into the slope and intercept equations
  - $b = \frac{581.67}{293.33} = 1.98$
  - $a = 19.93 - 1.98(7.67) = 4.74$

Computing the Regression Line

- The regression equation

$$\widehat{pubs} = 4.74 + 1.98time$$

- The predicted number of publications for a professor who just received their Ph.D. is 4.74.
- For each additional year after Ph.D. completion, the predicted number of publications will increase by 1.98.

VANDERBILT
PEABODY COLLEGE

VANDERBILT
PEABODY COLLEGE

Predicted Values, Residuals and the Standardized SLR Model

Shane Hutton, Ph.D.

Vanderbilt University

Predicted Values

- The regression line can be used to make predictions about Y at observed or unobserved values of the independent variable X.
- The predictions are denoted $\hat{Y}$ but the actual observed values are Y.
- For a given level of X there will be one $\hat{Y}$, even though we could have observed multiple Y values.
- It is important to use caution when making predictions outside the range of X.

Predicted Values

- Predict the number of publications for a professor who earned their Ph.D. 6 years ago.

$$\widehat{pubs} = 4.74 + 1.98(6) = 16.62$$

- Thus, we predict a professor will have 16.62 publications 6 years after earning a Ph.D.

Predicted Values

- We actually observed three professors (professors 2, 6 and 12) who had earned their Ph.D. 6 years ago.
  - We observed they had 3, 6, and 21 publications,
  - We predicted 16.62 publications.
- Each professor will have an associated residual.

Residuals

- Recall the equation:

$$Y = \hat{Y} + e$$

- Solving for the residual (or error), we get:

$$e = Y - \hat{Y}$$

Residuals

- We can compute a residual for each professor:
  - $e_2 = 3 - 16.62 = -13.62$
  - $e_6 = 6 - 16.62 = -10.62$
  - $e_{12} = 21 - 16.62 = 4.38$
- We can find predicted values and residuals for each professor in the dataset.

Predicted Values and Residuals

Professor	Time (X)	Pubs (Y)	$\hat{Y}$	$Y - \hat{Y}$
1	3	18	10.68	7.32
2	6	3	16.62	-13.62
3	3	2	10.68	-8.68
4	8	17	20.58	-3.58
5	9	11	22.56	-11.56
6	6	6	16.62	-10.62
7	16	38	36.42	1.58
8	10	48	24.54	23.46
9	2	9	8.70	0.30
10	5	22	14.64	7.36
11	5	30	14.64	15.36
12	6	21	16.62	4.38
13	7	10	18.60	-8.60
14	11	27	26.52	0.48
15	18	37	40.38	-3.38

The Standardized Model

- Consider standardizing the X and Y.

$$z_X = \frac{X - M_X}{s_X} \qquad z_Y = \frac{Y - M_Y}{s_Y}$$

- The equation for the simple linear regression line with standardized variables is:

$$\hat{z}_Y = \beta z_X$$

The Standardized Model

- The textbook chooses to represent a standardized regression coefficient with β.
  - This is not always the case.
- The equation does not have an intercept because the intercept is 0 when X and Y are standardized.
- For simple linear regression, $r = \beta$

The Standardized Model

- The standardized simple linear regression model for the example is:

$$\hat{z}_{pubs} = .656 z_{time}$$

- Recall: $r = .656$

VANDERBILT
PEABODY COLLEGE

The Standard Error of the Estimate and the Coefficient of Determination in SLR

Shane Hutton, Ph.D.

Vanderbilt University

Standard Error of the Estimate

- The standard error of the estimate (s_e) is a measure of the accuracy of the predicted values (i.e., points on the regression line).
- It measures the variability of Y values around the regression line.

Standard Error of the Estimate

- The standard error of the estimate is defined as:

$$s_e = \sqrt{\frac{SSE}{df}}$$

  - Where $df = n - 2$ for simple linear regression

Standard Error of the Estimate

Professor	Time (X)	Pubs (Y)	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
1	3	18	10.68	7.32	53.58
2	6	3	16.62	-13.62	185.50
3	3	2	10.68	-8.68	75.34
4	8	17	20.58	-3.58	12.82
5	9	11	22.56	-11.56	133.63
6	6	6	16.62	-10.62	112.78
7	16	38	36.42	1.58	2.50
8	10	48	24.54	23.46	550.37
9	2	9	8.70	0.30	0.09
10	5	22	14.64	7.36	54.17
11	5	30	14.64	15.36	235.93
12	6	21	16.62	4.38	19.18
13	7	10	18.60	-8.60	73.96
14	11	27	26.52	0.48	0.23
15	18	37	40.38	-3.38	11.42

Standard Error of the Estimate

- Example:

$$SSE = \sum(Y - \hat{Y})^2 = 1521.52$$

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1521.52}{15-2}} = 10.82$$

Standard Error of the Estimate

- The larger the value of s_e, the more variability of Y around the regression line.
  - This means less accurate predictions.
- It does not tell you exactly about the accuracy of a single prediction (i.e., how much error there will be in any single prediction).
  - s_e represents the average of all the errors.

Standard Error of the Estimate

- The interpretation has two parts:
  - State that this is the average error in predictions of the dependent variable
  - Evaluate that amount of error relative to the range of the predicted variable using terms such as small, moderate and large

Standard Error of the Estimate

- Interpretation
  - The average error in predictions of ___ (DV) is ___ points. Given the range of ___ (DV) scores, this seems like a ___ (small, moderate, larger) amount of error.

Standard Error of the Estimate

- Publication scores ranged from 2 to 48.
- Interpretation:
  - The average error in predictions of publications is 10.82 points. Given the range of publication scores, this seems like a moderate amount of error.

Coefficient of Determination

- Coefficient of determination: r^2 or R^2
  - Proportion of variance in Y that is accounted for by X
- Example: Number of years since Ph.D. and number of publications ($\mathrm{r} = .656$)
- Coefficient of determination
  - $R^2 = (.656)^2 = .43$
- 43% of the variability in publications is explained by time since Ph.D.

VANDERBILT
PEABODY COLLEGE

VANDERBILT PEABODY COLLEGE

Question Review: Predicted Values and Residuals

Shane Hutton, Ph.D.

Vanderbilt University

Question Review

- A teacher asked her students to report how many hours they had spent studying for the last midterm during the two days prior to the midterm. After collecting the data, she found the following regression equation:

$$\widehat{grade} = 70 + 3hour$$

Question Review

- Predict the grade for a student who studied for 5 hours during the two days prior to the midterm:

$$\widehat{grade} = 70 + 3(5) = 85$$

- Suppose we actually observed a student who studied 5 hours and her grade was an 89. What is that student's residual value?

$$e = 89 - 85 = 4$$

VANDERBILT
PEABODY COLLEGE

VANDERBILT
PEABODY COLLEGE

Inference in SLR

Shane Hutton, Ph.D.
Vanderbilt University

Hypothesis Testing

- Sample data are used to determine the a and b for the OLS regression line:

$$\hat{Y} = a + bX$$

- But we are truly interested in the population model:

$$\hat{Y} = \alpha^* + \beta^* X$$

Hypothesis Testing

- α^* and β^* are population parameters, which we don't know.
  - α^* is the population intercept
  - β^* is the population slope
- a and b are parameter estimates, which we estimate from sample data (a random sample from the population).
  - a is the estimate of α^*
  - b is the estimate of β^*
- This is the same idea as using r to estimate ρ or M to estimate μ.

Hypothesis Testing

- Likewise, r^2 is an estimate of ρ^2.
  - ρ^2 is the population coefficient of determination.
  - r^2 is the sample coefficient of determination.
- We can formulate hypothesis tests for α^*, β^* and ρ^2
  - t-tests are used for α^* and β^*.
  - The F-test is used for ρ^2.

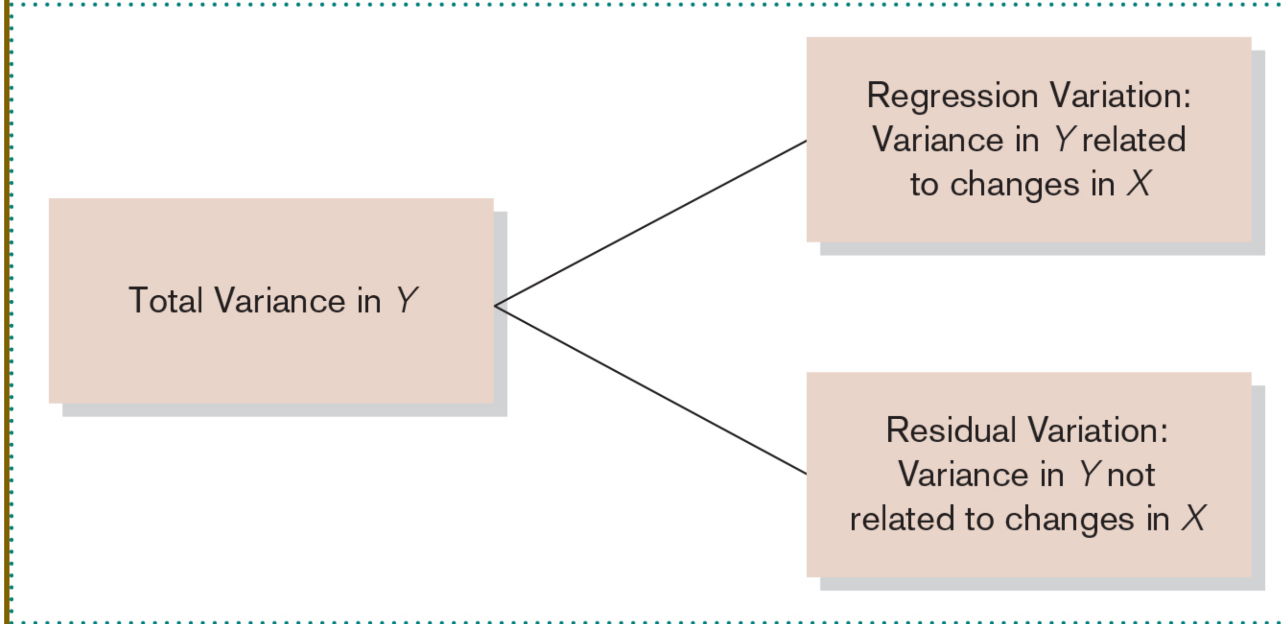
Hypothesis Test for ρ^2

- **Step 1:** State the Hypotheses
  - The null hypothesis is that X does not explain significant variability in Y.
    - H_0: $\rho^2 = 0$
  - The alternative hypothesis is that X does explain significant variability in Y,
    - H_1: $\rho^2 > 0$
- **Step 2:** Select the Statistical Test and the Significance Level

Hypothesis Test for ρ^2

- **Step 3:** Calculate the Test Statistic
  - This is the F-statistic

FIGURE 16.6 An Analysis of Regression Measures the Variance in *Y*



Some of the variance is attributed to changes in *X*; the remaining variance is not attributed to changes in *X*.

Hypothesis Test for ρ^2

- **Step 3:** Calculate the Test Statistic
  - This is the F-statistic:

TABLE 16.4 The *F* Table for an Analysis of Regression

Formulas for Completing the Analysis of Regression				
Source of Variation	***SS***	***df***	***MS***	F_{obt}
Regression	$r^2 SS_Y$	1	$\frac{SS_{regression}}{df_{regression}}$	$\frac{MS_{regression}}{MS_{residual}}$
Residual (error)	$(1 - r^2)SS_Y$	$n - 2$	$\frac{SS_{residual}}{df_{residual}}$	
Total	$SS_{regression} + SS_{residual}$	$n - 1$		

Hypothesis Test for ρ^2

- **Step 4:** Make a Decision
  - Find the rejection region in the F distribution
  - Determine the p-value

Hypothesis Test for α^* and β^*

- **Step 1:** State the Hypotheses
  - For the intercept:
    - The null hypothesis is that the intercept is equal to 0 in the population.
    - H_0: $\alpha^* = 0$
    - The alternative hypothesis the intercept is not equal to 0 in the population.
    - H_1: $\alpha^* \neq 0$

Hypothesis Test for α^* and β^*

- For the slope:
  - The null hypothesis is that the slope is equal to 0 in the population (or there is no linear relationship between X and Y).
  - H_0: $\beta^* = 0$
  - The alternative hypothesis the slope is not equal to 0 in the population (or there is a linear relationship between X and Y).
  - H_1: $\beta^* \neq 0$

- **Step 2:** Select the Statistical Test and the Significance Level

Hypothesis Test for α^* and β^*

- **Step 3:** Calculate the Test Statistic
  - This is the t-statistic:

$$t = \frac{a}{s_a} \qquad\qquad t = \frac{b}{s_b}$$

  - Where
    - a is the parameter estimate for the intercept and s_a is the estimated standard error for the slope
    - b is the parameter estimate for the intercept and s_b is the estimated standard error for the slope

Hypothesis Test for α^* and β^*

- **Step 4:** Make a Decision
  - Find the rejection region in the *t*-distribution
    - $df = n - 2$
  - Determine the p-value

Hypothesis Testing

- For simple linear regression, the F-test for $\rho^2 = 0$ is equivalent to testing if the slope $\beta^* = 0$.
  - If $\rho^2 = 0$, then there is no variability in Y being explained by X
  - Thus, $\beta^* = 0$
- There is a mathematical relationship between the t-statistic used in the t-test for the slope and the F-statistic used in the F-test for the coefficient of determination.

$$F = t^2$$

Confidence Intervals

- Confidence intervals can be computed for the parameter estimates.
- The are computed for the intercept and slope as follows:

$$a \pm t_{cv}(s_a)$$

$$b \pm t_{cv}(s_b)$$

- If 0 is contained in the CI, then fail to reject the null hypothesis.

VANDERBILT
PEABODY COLLEGE

Example: Inference in SLR

Shane Hutton, Ph.D.

Vanderbilt University

Hypothesis Test for ρ^2

- **Step 1:** State the Hypotheses
  - The null hypothesis is that time does not explain significant variability in publications.
  - H_0: $\rho^2 = 0$
  - The alternative hypothesis is that time does explain significant variability in publications.
  - H_1: $\rho^2 > 0$
- **Step 2:** Select the Statistical Test and the Significance Level
  - F-test, $\alpha = .05$

Hypothesis Test for ρ^2

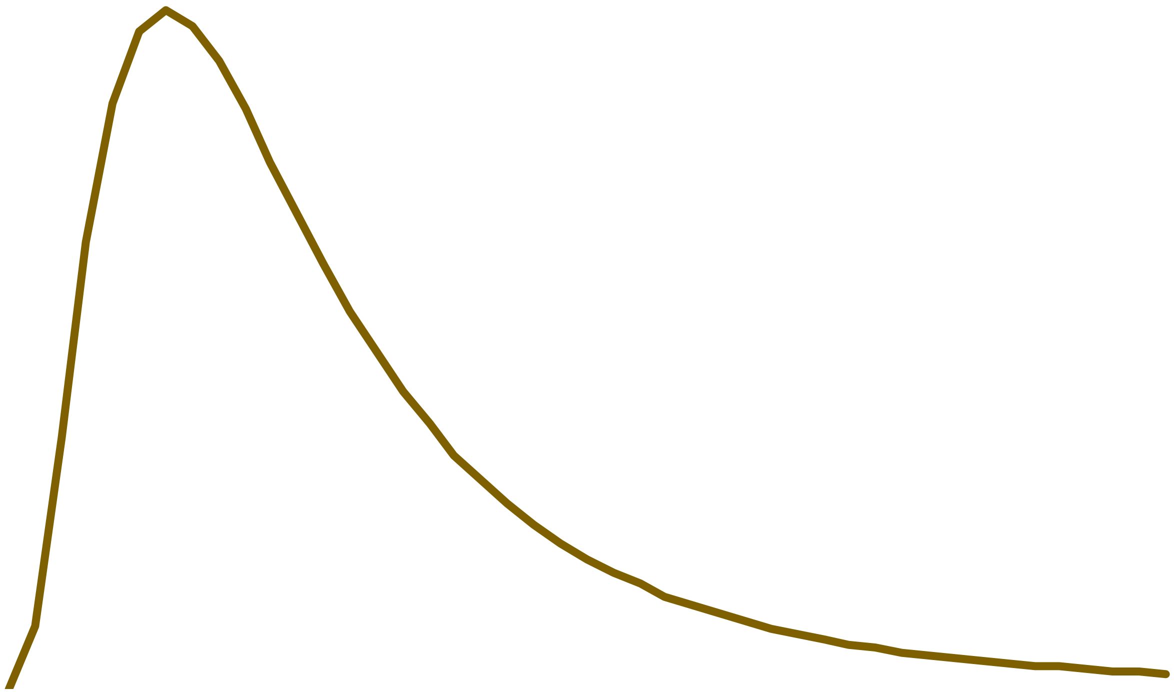
- Step 3: Calculate the Test Statistic
  - The F-statistic

Source	SS	df	MS	F
Regression	1153.42	1	1153.42	9.85
Residual	1521.52	15-2=13	117.04	
Total	2674.94	15-1=14		

Hypothesis Test for ρ^2

- **Step 4:** Make a Decision

$df_n = 1$
$df_d = 13$
$f_{cv} = 4.67$



Hypothesis Test for α^* and β^*

- **Step 1:** State the Hypotheses
  - For the intercept:
    - The null hypothesis is that the intercept is equal to 0 in the population.
    - H_0: $\alpha^* = 0$
    - The alternative hypothesis the intercept is not equal to 0 in the population.
    - H_1: $\alpha^* \neq 0$

Hypothesis Test for α^* and β^*

- For the slope:
  - The null hypothesis is that the slope is equal to 0 in the population (or there is no linear relationship between X and Y).
  - H_0: $\beta^* = 0$
  - The alternative hypothesis the slope is not equal to 0 in the population (or there is a linear relationship between X and Y).
  - H_1: $\beta^* \neq 0$

Hypothesis Test for α^* and β^*

- **Step 2:** Select the Statistical Test and the Significance Level
  - t-test, $\alpha = .05$

Hypothesis Test for α^* and β^*

- **Step 3:** Calculate the Test Statistic
  - The t-statistic
    - Intercept:

$$t = \frac{4.73}{5.59} = 0.85$$

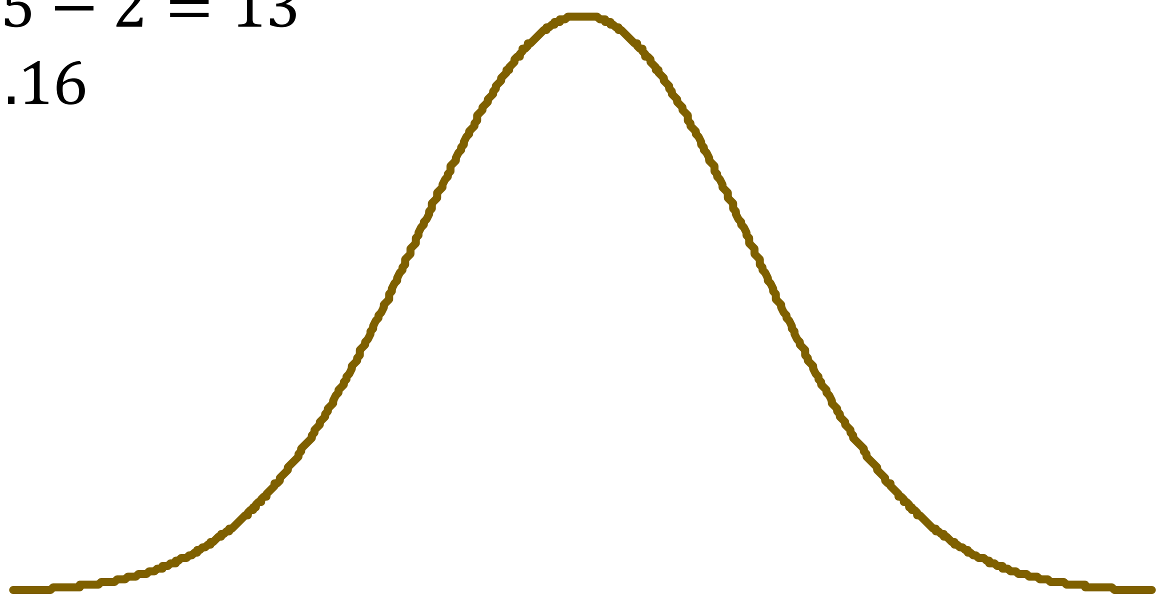
    - Slope:

$$t = \frac{1.98}{0.63} = 3.14$$

Hypothesis Test for α^* and β^*

- **Step 4:** Make a Decision

$df = 15 - 2 = 13$
$t_{cv} = 2.16$



Hypothesis Testing

- Notice for the slope

$$9.85 = 3.14^2$$

Confidence Intervals

- Compute 95% confidence intervals:

$$4.73 \pm 2.16(5.59) \quad \Rightarrow \quad [-7.34, 16.80]$$

$$1.98 \pm 2.16(.63) \quad \Rightarrow \quad [0.62, 3.34]$$

- The 95% CI for the intercept contains 0 so we fail to reject the null hypothesis for the intercept but the 95% CI for the slope does not contain 0 so we reject the null hypothesis for the slope.

VANDERBILT
PEABODY COLLEGE

VANDERBILT
PEABODY COLLEGE

APA Reporting in SLR

Shane Hutton, Ph.D.
Vanderbilt University

APA Reporting

- Report the test statistic, *df*, and *p*-value.
- Report effect size if significant .
- The data points are often summarized in a scatter plot displaying the regression line.
- The regression line equation can be reported in the scatter plot.

APA Reporting

- **Conclusion:** A simple linear regression analysis showed that the number of publications can be significantly predicted from time since Ph.D., $F(1, 13) = 9.85$, $p < .05$, $R^2 = .43$.

VANDERBILT
PEABODY COLLEGE