Linear Regression

1. Purpose—To Model Dependent Variables

Linear regression is used model variation observed in a dependent variable (DV) with theoretically linked independent variables (IV). For example, one may wish to model why students obtain different scores on achievement tests. Possible reasons for these differences include intelligence, ability, or teaching strategies. Linear regression enables researchers to determine if any or all of these IVs are related (and therefore possibly explain) variation observed in achievement.

A second, and less common, reason researchers use linear regression is to obtain a prediction equation. The goal in prediction is to find highly related IVs that may be used to predict a subject's outcome, like probability of dropping out, low achievement, etc. Thus, prediction equations are usually for determining which students, for example, may benefit most from specialized programs, etc.

2. Regression Equations

Population

Simple regression (one DV and one IV)

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (or sometimes $Y_i = \alpha + \beta_1 X_i + \varepsilon_i$)

where

Y_i = represents individual scores on the DV per each i^{th} person
β_0 = the intercept of the equation (predicted value of Y' when IV = 0.00)
β_1 = the slope relating IV (X_i) to the DV (Y_i)
X_i = represents individual scores on the IV per each i^{th} person
ε_i = residual or error term; defined as the deviation between observed Y and predicted Y'

Prediction equation

$Y' = \beta_0 + \beta_1 X_i$.

Where Y' is the predicted value of the DV in the population. Note absence of ε; since means are predicted based upon the equation, individual score deviations from the prediction are not included.

Sample

Simple regression (one DV and one IV)

$Y_i = b_0 + b_1 X_i + e_i$ (or sometimes $Y_i = a + b_1 X_i + e_i$)

where

Y_i = represents individual scores on the DV per each i^{th} person
b_0 = the intercept of the equation (predicted value of Y' when IV = 0.00)
b_1 = the slope relating IV (X_i) to the DV (Y_i)
X_i = represents individual scores on the IV per each i^{th} person
e_i = residual or error term; defined as the deviation between observed Y and predicted Y'

Prediction equation

$Y' = b_0 + b_1X_i$.

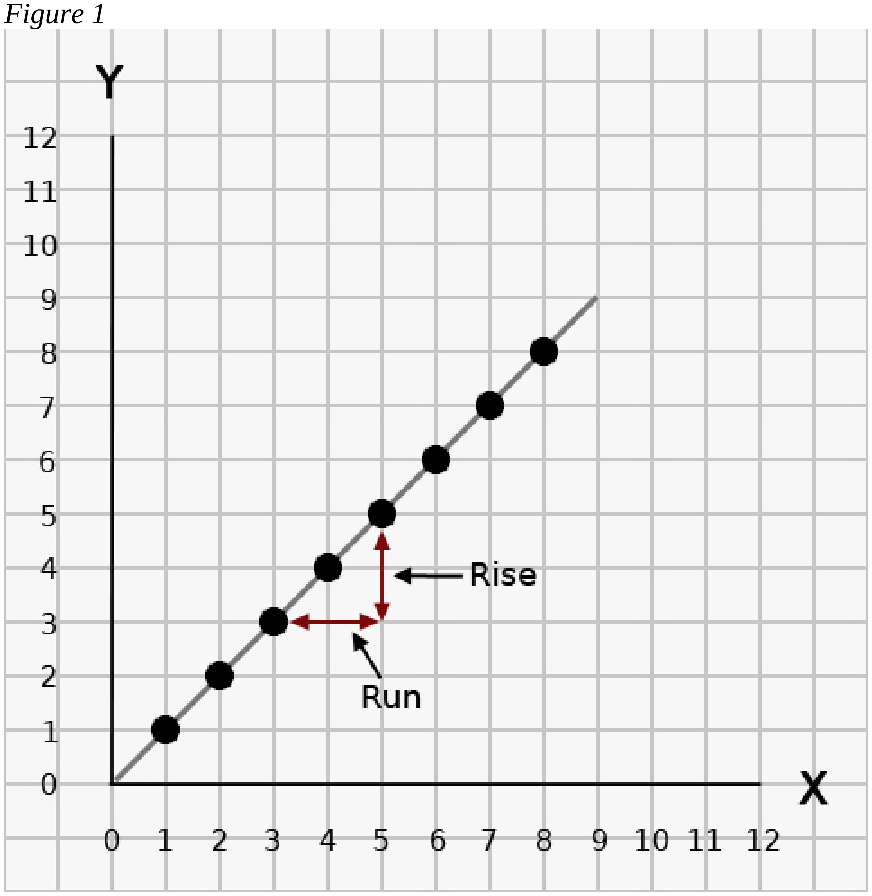
Where Y' is the predicted value of the DV in the sample. Note absence of e; since means are predicted based upon the equation, individual score deviations from the prediction are not included.

Literal interpretation of regression coefficients $b_0 + b_1$ (these interpretations also apply to population parameters shown above):

b_0 = predicted value of DV, Y', when X = 0
b_1 = expected change in predicted Y' for a one unit increase in X

Intercept and Slope of a Line Illustrated

Figure 1



b_0 = regression line cross the Y axis; it is also the predicted value of Y when X = 0.00.
b_1 = for simple regression it is the rise in Y divided by the run in X, b_1 = rise/run.

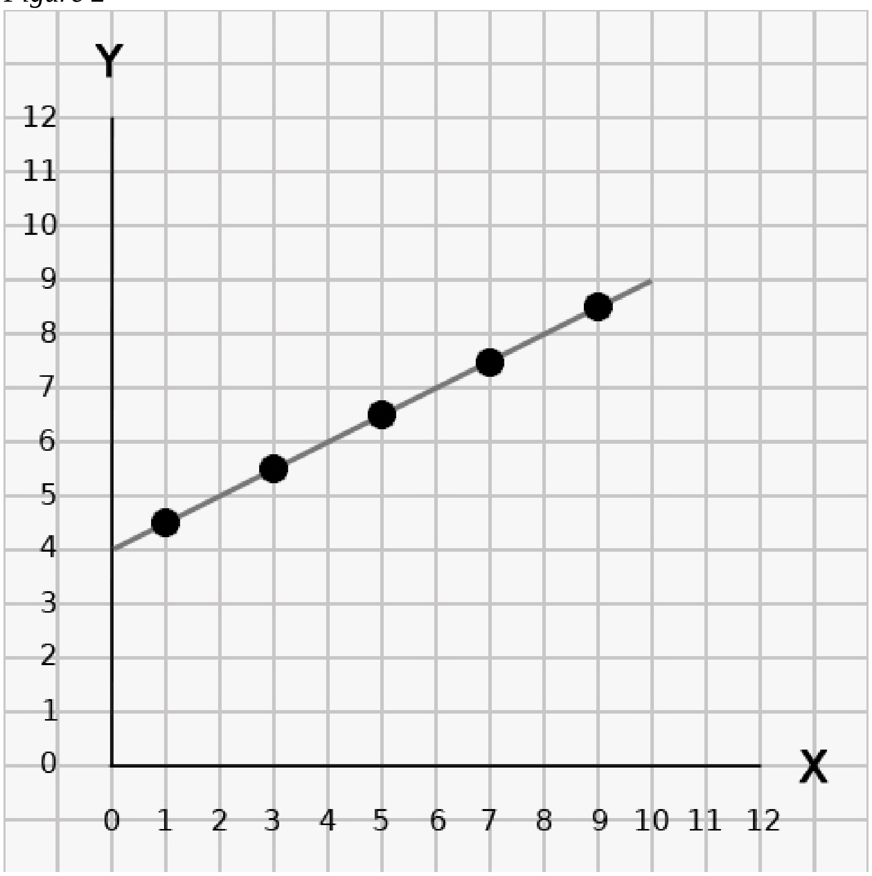
The sample prediction equation for the above line is $Y' = b_0 + b_1X_i$. Determine the following:

(a) b_0 =
(b) b_1 =

(c) Find Y' when X = 5 using the prediction equation
(d) Find Y' when X = 2 using the prediction equation
(e) Find Y' when X = 12 using the prediction equation
(f) Find Y' when X = 0 using the prediction equation
(g) What is literal interpretation of b_0?
(h) What is literal interpretation of b_1?

Note: Illustrate with statistical software.

Figure 2

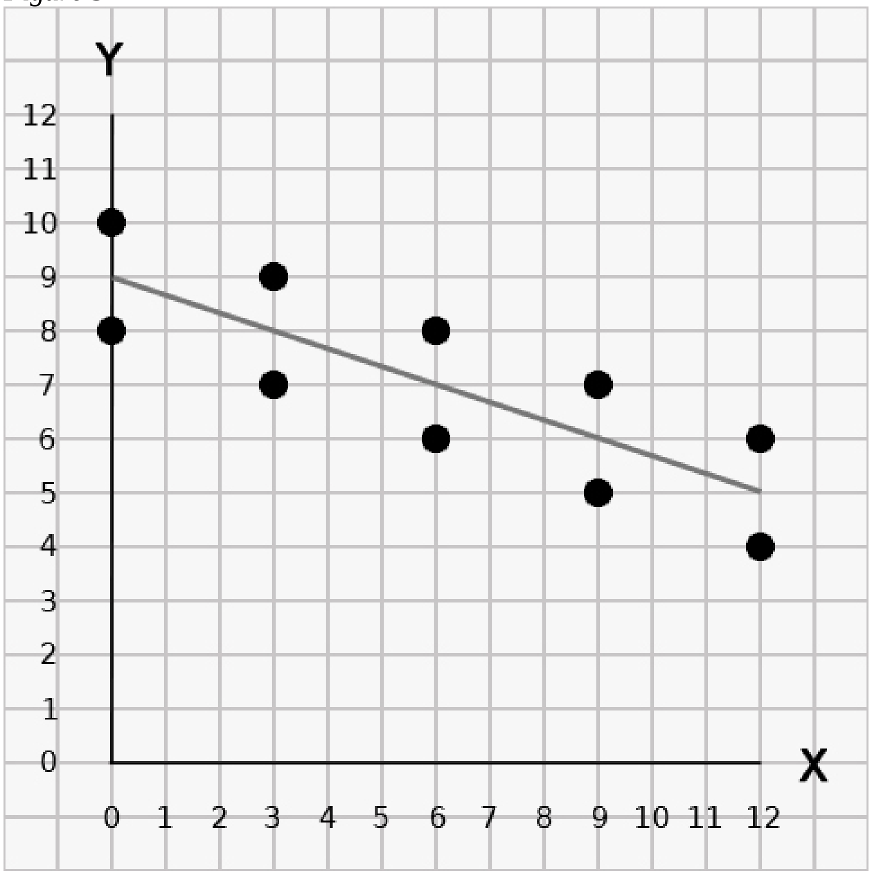


The sample prediction equation for the above line is $Y' = b_0 + b_1X_i$. Determine the following:

(a) b_0 =
(b) b_1 =
(c) Find Y' when X = 7 using the prediction equation
(d) Find Y' when X = 2 using the prediction equation
(e) Find Y' when X = 0 using the prediction equation
(f) What is literal interpretation of b_0?
(g) What is literal interpretation of b_1?

Note: Illustrate with statistical software.

Figure 3



The sample prediction equation for the above line is $Y' = b_0 + b_1X_i$. Determine the following:

(a) b_0 =
(b) b_1 =
(c) Find Y' when X = 4.5 using the prediction equation
(d) Find Y' when X = 9 using the prediction equation
(e) Find Y' when X = 0 using the prediction equation
(f) What is literal interpretation of b_0?
(g) What is literal interpretation of b_1?

Note: Illustrate with statistical software. Data for Figure 3 below.

Table 1: Data for Figure 3

Y	X	Y	X
10	0	6	6
8	0	7	9
9	3	5	9
7	3	6	12
8	6	4	12

3. Estimation of the Regression Equation and Residuals

Ordinary Least Squares (OLS or LS)

Method of finding best fitting regression line by minimizing the sum of squared residuals or errors, i.e., minimize $\sum e^2$, hence the term least squares. Formula for OLS not covered in this course.

Residual

Any discrepancy between observed Y and predicted Y', e = Y–Y'

Residuals Illustrated

The following data will be used to illustrate residuals, e.

Table 2: Student Ratings and Course Grades Data

Course	Quarter	Year	Student Ratings (mean ratings for course)	Percent A's
EDR852	FALL	1994	3.00	46.00
EDR761	FALL	1994	4.40	47.00
EDR761	FALL	1993	4.40	53.00
EDR751	SUMM	1994	4.50	62.00
EDR751	SUMM	1994	4.90	64.00
EDR761	SPRI	1994	4.40	50.00
EDR751	SPRI	1994	3.70	33.00
EDR751	WINT	1994	3.30	25.00
EDR751	WINT	1994	4.40	53.00
EDR751	FALL	1993	4.80	50.00
EDR751	SUMM	1993	4.80	54.00
EDR751	SUMM	1993	3.80	60.00
EDR751	SPRI	1993	4.60	54.00
EDR761	SPRI	1993	4.10	37.00
EDR751	WINT	1993	4.20	53.00
EDR751	FALL	1992	3.50	41.00
EDR751	FALL	1992	3.80	47.00

Obtained prediction equation will be

$$
\begin{aligned}
Y' &= b_0 + b_1X \\
&= 2.47 + 0.034(X).
\end{aligned}
$$

Predicted Y' when X = 50:

$$
\begin{aligned}
Y' &= 2.47 + 0.034(X) \\
&= 2.47 + 0.034\,(50). \\
4.17 &= 2.47 + 1.70.
\end{aligned}
$$

Obtaining residuals -- In Table 1 Y = 4.40 when X = 50 (Y also is 4.80 when X = 50).

e_i = Y - Y'
e_i = 4.40 - 4.17
e_i = 0.23.

also

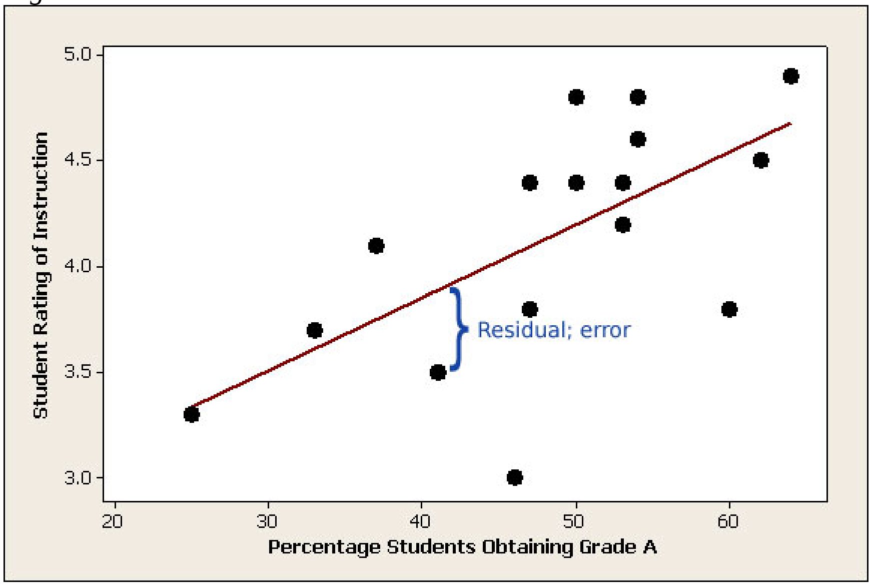
e_i = Y - Y'
e_i = 4.80 - 4.17
e_i = 0.63.

Note that the OLS coefficient estimates under-predicted this particular observation.

OLS attempts to find the regression line that passes through all observations and that provides the smallest set of squared residuals, $\sum e^2$. See Figure 4 below.

Based upon Figure 4, what is likely to be the intercept?

Figure 4



See Figure 5 below for different scaling of X and Y.

Note: Illustrate calculating residuals for data above with statistical software.

4. Literal Interpretation of Coefficients

Using current example data, the estimated regression equation is:

$$Y' = b_0 + b_1X$$
$$= 2.47 + 0.034(X).$$

What is literal interpretation of b_0 and b_1?

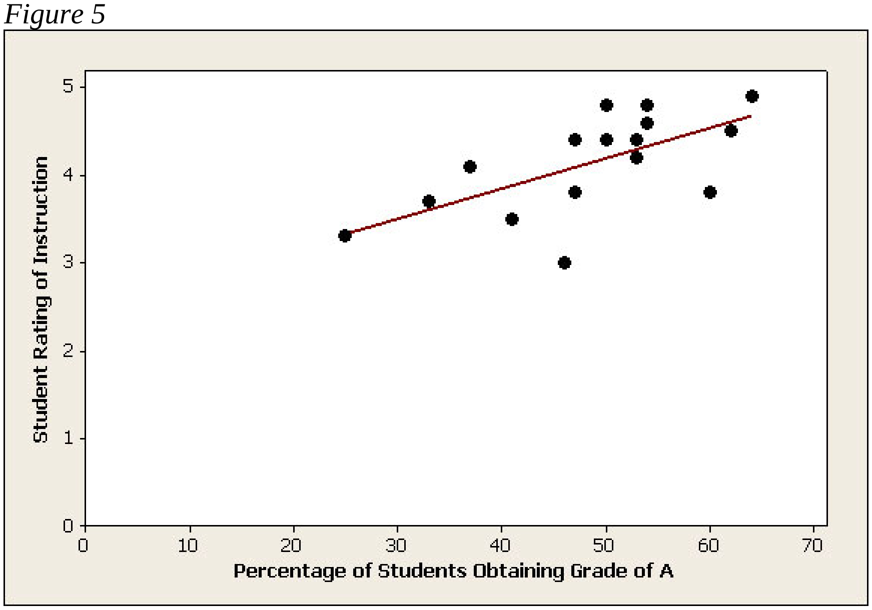
b_0 = predicted mean student rating of instruction is 2.47 for those classes in which 0.00% of students obtained the grade of A

b_1 = for each 1% increase in number of students who obtain the grade of A in a class, the mean student rating of instruction is predicted to increase by 0.034

Find the following:

(a) If X (percentage of students with the grade of A) increases by 10, what is the amount of change expected in student ratings?
(b) If X decreases by 5, what is the amount of change expected in student ratings?
(c) If X = 25, what is the predicted mean student rating?
(d) If X = 75, what is the predicted mean student rating?

Figure 5



Additional Examples for Interpretation

Example 1*: What is the boiling point of water in degrees Fahrenheit—212°F, correct? Yes, at barometric pressure of about 29.92 inches of mercury (in/Hg), which is sea level.*

Table 3: Boiling Point of Water and Corresponding Barometric Pressure in in/Hg

BPt	Pressure	BPt	Pressure	BPt	Pressure
194.5	20.79	200.9	23.89	209.5	28.49
194.3	20.79	201.1	23.99	208.6	27.76
197.9	22.4	201.4	24.02	210.7	29.04
198.4	22.67	201.3	24.01	211.9	29.88
199.4	23.15	203.6	25.14	212.2	30.06
199.9	23.35	204.6	26.57	209.5	28.49

Example 2*: Does cotton yield, in pounds per acre, vary by amount of irrigation water applied in feet per acre?*

Table 4: Cotton Yield in Pounds per Acre and Irrigation Amount in Feet per Acre

Irrigation	Yield		Irrigation	Yield
1.8	260		1.5	280
1.9	370		1.5	230
2.5	450		1.2	180
1.4	160		1.3	220
1.3	90		1.8	180
2.1	440		3.5	400
2.3	380		3.5	650

Example 3*: Does weight of car correlate with gasoline consumption?*

Table 5: Fuel Consumption and Weight for Several Cars

Car	Weight (in 1,000s of pounds)	Gallons per 100 miles	Miles per gallon
AMC Concord	3.4	5.5	18.2
Chevy Caprice	3.8	5.9	16.9
Ford Wagon	4.1	6.5	15.4
Chevette	2.2	3.3	30.3
Corona	2.6	3.6	27.8
Mustang	2.9	4.6	21.7
Mazda GLC	2.0	2.9	34.5
AMC Sprint	2.7	3.6	27.8
VW Rabbit	1.9	3.1	32.3
Buick Century	3.4	4.9	20.4

5. Model Fit

How well does the regression model fit the observed Y scores? Does the regression model reproduce Y well?

Multiple R = Pearson product moment correlation, r, between Y and Y', denoted simply as R. Sometimes this is referenced as the *Coefficient of Multiple Correlation*.

Multiple R^2 = Value of R squared, sometimes referenced as the *coefficient of determination*. R^2 represents a measure known as the *proportional reduction in error* (PRE) that results from the model when attempting to predict Y.

Adjusted R^2 = Similar in interpretation to R^2 above, but calculated differently. Adjusted R^2 is proportional reduction in error variance of Y, i.e., adj. R^2 = [var(Y)-var(e)] / var(Y) where var(e) is variance of residuals calculated as $n – df_1 – 1$ (not $n – 1$ as is typical of variance formula for sample). Another formula: adj. $R^2 = 1 – [MSE/var(Y)]$

- Both R^2 adjusted R^2 range between 0.00 and 1.00, and R ranges between -1.00 and 1.00 although R should not be negative.
- Closer R to 1.00, the better the model fits Y.
- Closer R to 1.00, the smaller will be residuals, why?

Student ratings and course grades data (Table 2): R = .6365, R^2 = .405, and adj. R^2 = .366 (about 40% of variance in student ratings can be predicted by knowledge of grades given in class; reduce prediction error of Y by 40% knowing student grades).

What are model fit statistics for the three data examples listed above?

6. Inference in Regression

Two types of inferential tests are common in regression, a test of overall model fit and tests of regression coefficients.

Overall Model Fit

Null

H_0: $R^2 = 0$ (tests whether model R^2 differs from 0.00; if R^2 = 0.00, then no reduction in prediction error)

or equivalently

H_0: $\beta_p = 0.00$ (states that slopes of predictors all equal 0.00, so none of the Xs have association with Y)

Alternative

H_1: $R^2 \neq 0$; or
H_0: at least one β_p is not 0.00

Meaning

H_0: $R^2 = 0$: The regression model does not predict Y; model does not reduce error in prediction.

H_1: $R^2 \neq 0$: Regression does predict some variability in Y; model does reduce some error in prediction of Y. Some aspect of the model used, i.e., the IVs selected, is statistically related to Y (or at least predicts Y).

Test of H_0: $R^2 = 0$

Overall F test is used to test the model null hypothesis. This overall F test is the same F test learned in one-way ANOVA.

$$F = \frac{SSR/df_r}{SSE/df_e} = \frac{SSR/k}{SSE/(n - k - 1)} = \frac{MS_{reg}}{MSE}$$

where;

SSR = regression sums of squares;
SSE = residual sums of squares;
df_r = regression degrees of freedom;
df_e = residual degrees of freedom;
k = number of independent variables (vectors) in the model;
n = sample size (or number of observations in sample);
MS_{reg} = mean square (same as ANOVA) due to regression (e.g., between);
MSE = mean square error (same as ANOVA mean square within).

The overall F test may also be calculated using R^2 as the basis:

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

Calculated and critical F value

F test has two degrees of freedom, one due to regression (explained variation) denoted df_r or df_1, and one due to residuals or error which is denoted as df_e or df_2.

$df_1 = k$ (where k = number of IVs/predictors in model)

and

$df_2 = n - k – 1$ (where n = sample size)

Degrees of freedom are used to find the critical F ratio.

For the student ratings of instruction and percentage of students with grade A data, the model F is

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{.405/1}{(1 - .405)/(17 - 1 - 1)} = 10.21$$

or

$$F = \frac{SSR/df_r}{SSE/df_e} = \frac{1.986/1}{2.916/15} = 10.21.$$

With $df_1 = 1$ and $df_2 = 15$, the .05 level critical F value is

${}_{.05}F_{1,15} = 4.54$.

Since the calculated F ratio is larger than the critical, then null hypothesis is rejected. We conclude that student grades are related to student evaluations of instruction.

Calculated F and p-values

If using software for the F test, likely a p-value will be reported. Compare p against α, and if $p \le \alpha$ reject model H_0.

Using the student rating data, p = .006, since .006 < .05, reject H_0.

Coefficient Inference

The inferential test of the overall model provides evidence of whether the model, as a whole, predicts Y, but to identify which X is related to Y, individual inferential tests of coefficients are needed. (This statement holds for multiple regression, for simple regression testing H_0: $R^2 = 0$ is equivalent to testing H_0: $\beta_1 = 0.00$).

Null and Alternative

H_0: $\beta_1 = 0.00$ (States that relation between X_1 and Y is zero; no relation between X_1 and Y)

If the null is rejected, then one may conclude that a relationship does exist between IV and DV. A non-directional alternative hypothesis states that a relationship does exist, thus:

H_1: $\beta_1 \neq 0.00$.

Testing H_0: $\beta_1 = 0.00$ with t-test

For each regression coefficient estimated, a corresponding standard error (SE) is also estimated. The ratio of the coefficient to its SE provides a t ratio.

For the student ratings and percentage of grade A data, the t-ratio for X is

$t = b_1 / SE_{b1}$
$= .034 / .011$
$= 3.091.$

Degrees of freedom for t-test

Degrees of freedom (df) for this t-test is defined a n – k – 1 where k is the number of predictors in the regression equation.

With the student rating data there n = 17 and there is one predictor (percentage of students in class with grade of A), so

$df = n – k – 1 = 17 – 1 – 1 = 15$

The critical t value, using an α of .05, is ${}_{.05}t_{15} = 2.131$, so the null hypothesis is rejected:

If $t \leq -t_{crit}$ or $t \geq t_{crit}$ reject H_0, otherwise fail to reject H_0.

Since $3.09 \geq 2.131$ *reject* H_0.

Calculated t and p-values

Statistical software will usually report p-values for t-tests and if the p-value is less than α, reject H_0 for that particular regression coefficient.

For the current example, the p-value for a t-ratio of 3.091 is p = .006. The usual decision rule, for non-directional tests, applies, so reject H_0.

Confidence Interval for b_1: Inference and Estimation

- CI is the upper and lower bound to the *point estimate* of regression coefficients.
- A point estimate is the single best estimate of the population coefficient denoting the relationship between X and Y, and for simple regression is b_1.
- A confidence interval represents, with a set level of precision, a range of possible values for b_1.

Formula for CI about regression coefficient:

$b_1 \pm t_{(\alpha/2,df)} SE_{b1}$

where

t is the critical t value, and SE_{b1} is the standard error of b_1.

Using the student ratings data, the 95% confidence interval (.95CI) for b_1 is

.95CI: $b_1 \pm t_{(\alpha/2,df)} SE_{b1}$
.95CI: $0.034 \pm (2.131)(0.011)$
.95CI: 0.034 ± 0.023
.95CI: 0.011, 0.057.

Interpretation: one may be 95% confidence that the true population coefficient may be as large as 0.057 or as small as 0.011.

Since 0.00 does not lie within this interval, H_0 will be rejected since 0.00, which is specified in the null hypothesis, is not one of the likely population values for the regression coefficient.

7. Reporting Regression Results

Often regression results will be presented in two tables, one of correlations and descriptive statistics and a second with regression estimates.

Table 6: Correlations and Descriptive Statistics for Student Ratings and Percentage A's Given in Class

	1	2
1. Student Ratings	---	
2. Percent Grade A	.64*	---
Mean	4.15	48.77
SD	0.55	10.22
Scale Min/Max Values	1 to 5	0 to 100

Note. n = 17
* p < .05

Table 7: Regression of Student Ratings on Percentage A's Given in Class

Variable	b	se b	95% CI	t
Percent A's	0.03	0.01	.01, .06	3.20*
Intercept	2.47	0.54	1.33, 3.62	4.61*

Note. R^2 = .41, adj. R^2 = .36, F = 10.22*, df = 1,15; n = 17
*p < .05.

> There is a positive and statistically significant relationship between student ratings of the instructor's ability to evaluate student performance and the percentage of students in the class who received high grades. In those classes where a high percentage of students received a grade of A, student ratings of the instructor were also high; in those classes where fewer students received a grade of A, the instructor was rated lower.

8. Exercises

(1) A researcher wishes to know whether number of hours studied at home is related to general achievement amongst high school students.

High School GPA	Estimated Number of Hours of Study Per Week at Home
3.33	3
1.79	5
2.21	12
3.54	9
2.89	11
2.54	1
2.66	0
1.10	3
3.67	2

(2) Does SAT adequately predict college success?

Freshmen Collegiate GPA	SAT Scores
3.33	1000
1.79	750
2.21	890
3.54	1100
2.89	900
2.54	860
2.66	1010
1.10	640
3.67	1240

(3) A teacher is convinced that frequency of testing within her classroom increases student achievement. She runs an experiment for several years in her algebra class. The frequency in which she presents tests to the class varies across quarters. For example, one quarter students are tested only once during the term, while in another quarter students are tested once every week. Is there evidence that testing frequency is related to average achievement?

Quarter	Testing Frequency During Quarter	Overall Class Achievement on Final Exam
Fall 1991	1	85.5
Winter 1992	2	86.5
Spring 1992	3	88.9
Summer 1992	4	89.1
Fall 1992	5	87.2
Winter 1993	6	90.5
Spring 1993	7	89.8
Summer 1993	8	92.5
Fall 1994	9	89.3
Winter 1994	10	90.1

(4) An administrator wishes to know whether a relationship exists between the number of tardies or absences a student records during the year and that student's end-of-year achievement as measured by GPA. The administrator randomly selects 12 students and collects the appropriate data.

GPA	Tardies/Absences
3.33	2
1.79	10
2.21	5
3.54	6
2.89	3
2.54	4
2.66	6
1.10	12
3.10	3
2.10	8
2.31	6
3.67	2

Exercise Answers

(1) A researcher wishes to know whether number of hours studied at home is related to general achievement amongst high school students.

Table 1
Descriptive Statistics and Correlation between High School GPA and Number of Hours Studied

Variable	Correlations	
	GPA	Hours Studied
GPA	---	
Hours Studied	.03	---
Mean	2.64	5.11
SD	0.84	4.46

Note. n = 9
* p < .05

Table 2
Regression of HS GPA on Hours Studied

Variable	b	se	95%CI	t
Hours Studied	0.006	0.07	-0.16, 0.18	0.09
Intercept	2.60	0.47	1.49, 3.72	5.50*

Note. R^2 = .001, adj. R^2 = -.14, F =0.008., df = 1,7; n = 9.
*p < .05.

Results of both correlation and regression show that hours studied is not statistically related to high school GPA at the .05 level. These results indicate that the amount of time one spends studying in high school does not covary, associate, or predict one's high school GPA.

(2) Does SAT adequately predict college success?

Table 1
Descriptive Statistics and Correlation between College GPA and SAT scores

Variable	Correlations	
	GPA	SAT
GPA	---	
SAT	.94*	---
Mean	2.64	932.22
SD	0.84	180.33

Note. n = 9
* $p < .05$

Table 2
Regression of College GPA on SAT scores

Variable	b	se	95%CI	t
SAT	0.004	0.001	0.003, 0.006	6.98*
Intercept	-1.44	0.59	-2.85, -0.04	-2.43*

Note. $R^2 = .87$, adj. $R^2 = .86$, F =48.72*, df = 1,7; n = 9.
*$p < .05$.

There is a statistically significant association between SAT scores and college GPA. Regression results show that students with higher SAT scores also tend to have higher college GPAs.

(3) A teacher is convinced that frequency of testing within her classroom increases student achievement. She runs an experiment for several years in her algebra class. The frequency in which she presents tests to the class varies across quarters. For example, one quarter students are tested only once during the term, while in another quarter students are tested once every week. Is there evidence that testing frequency is related to average achievement?

Table 1
Descriptive Statistics and Correlation between Testing Frequency and Student Achievement

Variable	Correlations	
	Testing Freq.	Achievement
Testing Freq.	---	
Achievement	.75*	---
Mean	5.50	88.94
SD	3.03	2.06

Note. n = 10
* $p < .05$

Table 2
Regression of Achievement on Testing Frequency

Variable	b	se	95%CI	t
Testing Freq.	0.51	0.16	0.15, 0.88	3.23*
Intercept	86.13	0.98	83.86, 88.39	87.58*

Note. R^2 = .57, adj. R^2 = .51, F =10.42*, df = 1,8; n = 10.
*$p < .05$.

There is a statistically significant relationship between testing frequency and mean classroom achievement. Results show that as testing frequency within the classroom increases, mean performance on tests also increases.

(4) An administrator wishes to know whether a relationship exists between the number of tardies or absences a student records during the year and that student's end-of-year achievement as measured by GPA. The administrator randomly selects 12 students and collects the appropriate data.

Table 1
Descriptive Statistics and Correlation between Student Absences and GPA

Variable	Correlations	
	Absences	GPA
Absences	---	
GPA	-.85*	---
Mean	5.58	2.60
SD	3.15	0.76

Note. n = 12
* $p < .05$

Table 2
Regression of GPA on Number of Absences

Variable	b	se	95%CI	t
Num. Absences	-0.21	0.04	-0.29, -0.12	-5.13*
Intercept	3.75	0.25	3.18, 4.31	14.79*

Note. $R^2 = .72$, adj. $R^2 = .70$, $F = 26.26$*, df = 1,10; n = 12.
*$p < .05$.

Analysis of number of absences and student GPA shows a statistically significant and negative association. Students with more absences throughout the school year tend to obtain lower GPAs.