

Simple Linear Regression: One Quantitative IV

Linear regression is frequently used to explain variation observed in a dependent variable (DV) with theoretically linked independent variables (IV). For example, one may wish to explain 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. With simple linear regression, the relationship between one IV and one DV is examined. The nature of the relationship between two these quantitative variables is expressed in a fashion similar to Pearson's product moment correlation coefficient, r . In fact, the relationship expressed is identical to r in some circumstances.

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.

The Regression Equation

Suppose we are interested in determining whether student evaluation of instructors is related to grades given in course. Specifically, are students' ratings of the instructor in the domain of the instructor's ability to evaluate student performance affected by actual grades given in the course?

Assume that after each course taught by the instructor, an evaluation form is administered to all students. Among the several items for evaluation, students are asked to rate the instructor's ability to evaluate student performance—the actual item reads as follows:

The instructor demonstrated skill in evaluating student performance.

Students have the option of providing a rating that ranges from 1 (strongly disagree) to 5 (strongly agree).

Some professors argue that students cannot objectively rate instructors in this category since students are affected by the grades they will receive. Therefore, the question of interest is whether grades given in the class affect student ratings of the instructor's ability to evaluate student performance. The data for this problem are listed below in Table 1.

In the regression equation of this data, student ratings of the instructor's ability to evaluate student performance is the DV, denoted Y , and the percentage of A's given in the course by the instructor is the IV, denoted X . A given course rating, Y , may be expressed as:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \tag{1}$$

or

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where

Y_i signifies the i^{th} rating received by the instructor for a given course,

β ($= \beta_1$) is the population regression coefficient expressing the relationship between X and Y ,

α ($= \beta_0$) is the population intercept for the equation, and

ε_i is, supposedly, a random error.

Table 1
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
EDR852	WINT	1993	3.40	999.00
EDR751	WINT	1993	4.20	53.00
EDR751	FALL	1992	3.50	41.00
EDR751	FALL	1992	3.80	47.00

Note. 999 represents missing data.

Note that both α and β are devoid of the subscript i . This implies that these two values are constants—they do not change from subject to subject or from observation to observation as do Y , X , and ε .

A corresponding sample regression equation also exists:

$$Y_i = a + bX_i + e_i,$$

or

$$Y_i = b_0 + b_1X_i + e_i,$$

or

(2)

$$Y_i = \hat{\alpha} + \hat{\beta}X_i + \hat{\varepsilon}_i,$$

or

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1X_i + \hat{\varepsilon}_i,$$

where

$a (= b_0)$ is the sample intercept,

$b (= b_1)$ is the sample regression coefficient, and

e is the sample error term in the model.

The hat or caret, $\hat{}$, over the population coefficient symbols simply denotes that α , β , and ε are sample estimates of the population parameters.

The goal of regression is to find a mathematical solution for b_0 and b_1 that will best fit the data reflected in Y and X . Several things are meant by best fit. Of most importance is the goal to obtain the best possible mathematical description of the relationship between Y and X , and this is primarily what we are interested in finding.

Ordinary Least Squares (or simply Least Squares)

The mathematical procedure used to obtain the best fitting estimates of the regression coefficients is called (ordinary) least squares (OLS or LS). The goal of OLS is to minimize the sum of squared residuals or errors, i.e., minimize $\sum e^2$, hence the term least squares.

What is a *residual*? To determine this, one first needs to understand how to obtain predicted values for the DV. The sample regression equation can be used to obtain predicted values for Y . The equation is:

$$Y' = b_0 + b_1X$$

or (3)

$$\hat{Y} = b_0 + b_1X$$

where \hat{Y} and Y' both represent the predicted value of the DV for a given level of the IV, X .

For example, suppose the following coefficient estimates were obtained:

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

In this equation the intercept is equal to 2.47, and the relationship between X and Y is estimated to be 0.034.

To obtain a predicted value of Y , simply substitute into X a given value of the IV. For example, the estimated effect of grades given in a course upon student ratings is $b_1 = 0.034$. If the instructor were to give, for example, 50% of the student in a class A's, then the predicted rating for that instructor would be:

$$\begin{aligned} Y' &= b_0 + b_1X \\ Y' &= 2.47 + 0.034(X) \\ 4.17 &= 2.47 + 0.034(50). \end{aligned}$$

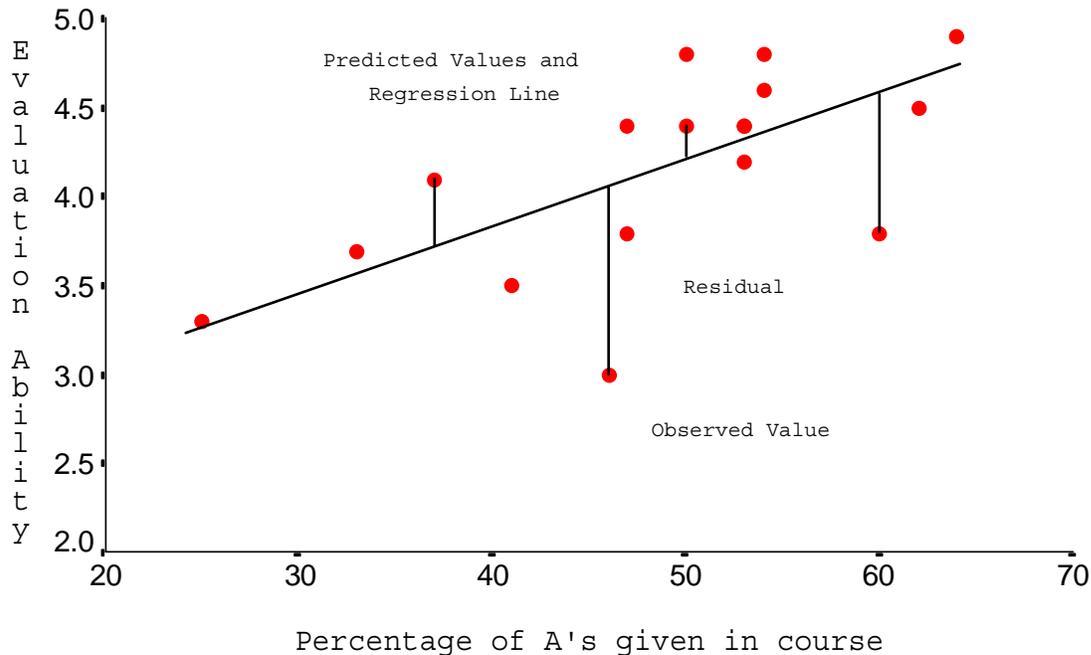
Now suppose we know that the actual student rating is 4.35 when 50% of the students get A's within a class. The residual or error in prediction for this rating is:

$$\begin{aligned} e_i &= Y - Y' \\ e_i &= 4.35 - 4.17 \\ e_i &= 0.18. \end{aligned}$$

The residual for this rating is 0.18, which indicates that the OLS estimate under-predicted this particular observation.

As previously noted, the goal of OLS is to obtain estimates for the regression coefficients that will provide the smallest possible residuals for all observations in the sample, and when certain assumptions are met, OLS estimates do provide the smallest possible residuals. In short, OLS attempts to find the regression line that passes through all observations and that provides the smallest set of squared residuals, $\sum e^2$.

Residuals can be depicted graphically as follows:



Regression Coefficient Interpretation

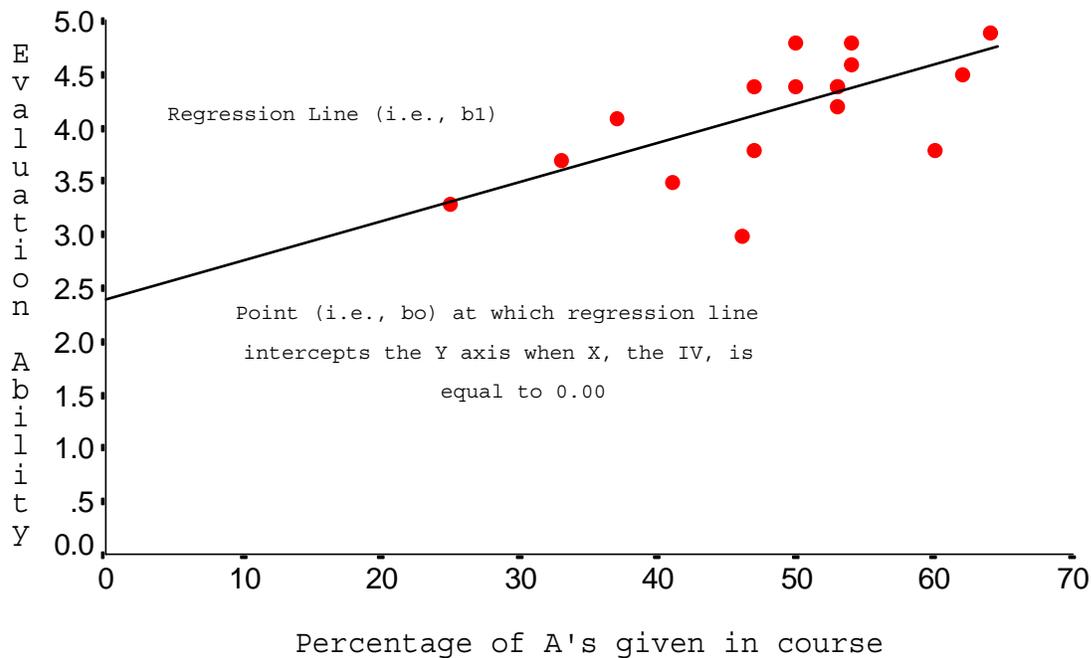
The coefficients obtained from OLS indicate the nature of the relationship between two variables. For example, $b_1 = 0.034$, and this indicates that

a one unit increase in X changes Y by b_1 units. In terms of the example data, a one percentage point increase in the number of A's given in a course is expected to increase student ratings by 0.034 points. (Note that as X increases, Y also is expected to increase—this indicates a positive relationship.)

So, for each 1 percentage point increase in A's given in a course, the predicted student rating will increase by .034. Note that coefficients may be interpreted in a fashion similar to Pearson's r . A positive b_1 indicates that a positive relationship exists between X and Y (assuming both variables have scales in which larger numbers represent more positive results), and a negative b_1 indicates an inverse relationship. Note that a positive relationship, as indicated by $b_1 = .034$, is depicted in the graph above.

Sometimes it is easier to understand or conceptualize such a change by considering large unit increases in the IV, X. For example, how would ratings be expected to change if one were to give 60% A's rather than 50% A's? Note that this is a change of 10 percentage points ($60 - 50 = 10$), so to determine how much increase is likely to result, multiple 10 by b_1 . The instructor would expect to gain about $0.034 (10) = 0.34$ points in ratings.

The other regression coefficient, b_0 , represents the intercept or constant. Graphically speaking, b_0 is the point at which the regression line, b_1 , crosses the Y axis when the IV is equal to 0.00 ($X = 0.00$). See the illustration below. In most cases, but not all, the exact value of b_0 will be of little theoretical interest since many IVs in behavioral research lack meaning when their value is 0.00.



Often b_0 is important because it is necessary for calculating predicted values, Y' , of the DV. As noted above, to calculate a predicted value of the DV, the regression equation is needed. Recall the earlier example:

$$Y' = b_0 + b_1X$$

$$Y' = 2.47 + 0.034(X)$$

$$4.17 = 2.47 + 0.034(50).$$

Recall the discussion of interpretation of b_1 . This coefficient denotes the change in predicted Y for a unit change in X . Also recall that for the current example, a 10 point change in percentage of students receiving an A, say from 50 to 60%, in the course corresponds to a gain in rating of about $0.034(10) = 0.34$ points.

A more informative presentation of this information would provide the estimated (or predicted) values in ratings for a change in percentage A's from 50 to 60. Thus, the predicted student rating for the instruction for 50% A's is:

$$Y' = b_0 + b_1X$$

$$Y' = 2.47 + 0.034(X)$$

$$4.17 = 2.47 + 0.034(50),$$

and the predicted rating for the instructor should 60% of the students receive an A is $4.17 + 0.34 = 4.51$, or

$$Y' = b_0 + b_1X$$

$$Y' = 2.47 + 0.034(X)$$

$$4.51 = 2.47 + 0.034(60).$$

Thus, the instructor should expect an average rating of 4.17 should half the class receive A's, and a rating of 4.51 should 60% receive A's!

Inferential Procedures

Once sample coefficients have been estimated, one must next determine whether the sample estimates are likely to be due to chance fluctuations in the data. Recall that the two-group t-test uses a calculated t ratio to determine whether a statistically significant relationship (or difference) exists. Significance tests in regression uses the same t ratios for significance tests.

In regression, the null hypothesis states that no relationship is expected between the IV and DV, that is,

$$H_0: \beta_1 = 0.00.$$

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:

$$H_1: \beta_1 \neq 0.00.$$

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. As with the two-group t-test, this calculated t ratio may be compared against a critical t ratio found in t tables to determine statistical significance. Or, one may simply use p-values obtained for each coefficient to determine statistical significance.

For example, the SE for b_1 is $SE_{b_1} = .011$, so the t ratio for b_1 is:

$$\begin{aligned} t &= b_1 / SE_{b_1} \\ &= .034 / .011 \\ &= 3.091. \end{aligned}$$

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

If $t \geq |t_{crit}|$ (in absolute value) reject H_0 , otherwise fail to reject H_0 .

If $3.09 \geq 2.131$ reject H_0 , otherwise fail to reject H_0 .

In terms of p-values, the p-value for the obtained t ratio of 3.091 is $p = .006$. The usual decision rule, for non-directional tests, applies:

If $p \leq \alpha$, reject H_0 , otherwise fail to reject H_0 ,

so, since the obtained p is less than alpha, H_0 is rejected:

If $.006 \leq .05$, reject H_0 , otherwise fail to reject H_0 ,

and one may conclude that a positive relationship exists between X and Y since b_1 is positive.

If necessary, one may also test b_0 in a similar fashion.

Interval Estimation: Confidence Intervals (CI)

A confidence interval represents an 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 . The problem with the point estimate is that it does not give a description of how precise the sample estimate is of the population coefficient. A confidence interval represents, with a set level of precision, a range of possible values for b_1 . Thus, the

confidence interval provides some indication of the exactness or precision with which sample estimates are derived from the sample data.

A confidence interval for b_1 may be formed as:

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

where t is the critical t value obtained from a table of t values representing a two-tailed alpha (α) level (such as .05) with degrees of freedom equal to $n-k-1$, and SE_{b_1} is the standard error of b_1 described above.

For the current example, the 95% confidence interval (.95CI) for b_1 is

$$.95CI: b_1 \pm t_{(\alpha/2, df)} SE_{b_1}$$

$$.95CI: 0.034 \pm (2.131)(0.011)$$

$$.95CI: 0.034 \pm 0.023$$

$$.95CI: (0.057, 0.011).$$

Such a CI enables the researcher to state that one may be 95% confidence that the true population coefficient may be as high as 0.057 or as low as 0.011. As this illustrates, CIs provide a sense of precision that point estimates do not.

Model Fit

When models are fit to data, a question of concern is how well the model explains or predicts the data. It is possible for one to obtain a model in which one or more IVs are statistically related to the DV (i.e., $H_0: \beta_j = 0.00$ is rejected, where j represents any coefficient, such as 0, 1, etc.), but that little variation in the DV can be accounted for—that is, little of the changes in Y can be explained or predicted by the X 's. In short, noting that some X 's are significantly related to Y does not mean that one has produced a good model of Y . By good model, I mean that we understand what causes, explains, or predicts variation in Y .

For linear regression, several indices of overall model fit are available. The most commonly used indicators of model fit are *Multiple R* and *Multiple R²*. Multiple R is the multivariate equivalent of Pearson's r . R simply represents the correlation between predicted Y , Y' , and observed Y ; as such, Multiple R is sometimes referred to as the *Coefficient of Multiple Correlation*. Multiple R^2 is referred to as the *coefficient of determination*, and R^2 roughly represents the proportion of variation in Y that can be accounted for (or predicted) by X (or multiple X 's). Multiple R^2 is also known as a measure of *proportional reduction in error* (PRE) that results from the model when attempting to predict Y .

Both R and R^2 range between 0.00 and 1.00. The larger the value of R^2 , the better the predictive power of the X 's. Since R is the correlation between Y' and Y , it makes sense that as Y is better predicted, using various X 's, the match between Y' and Y will be stronger. As the match between Y' and Y becomes stronger, the difference between Y' and Y will be smaller—that is, residuals will be smaller. The smaller the residuals, the larger R and R^2 . In short, the smaller the residuals between Y' and Y , the better the model will fit the data, thus R^2 will be closer to 1.00.

For the current example, the $R = .6365$, and $R^2 = .405$.

Note that for regression, it is also possible to obtain sums of squares, SS , like those obtained from ANOVA. Usually for regression the SS_{total} (or SST) are broken into two components, a SS due to the regression—explained or regression sums of squares (SSR)—and a SS due to the residuals—unexplained or residual sums of squares (SSE).

R^2 can be calculated from SSR and SST as follows:

$$R^2 = SSR/SST.$$

For the current example,

$$\begin{aligned}R^2 &= SSR/SST; \\ &= 1.986/4.902. \\ &= 0.405.\end{aligned}$$

Thus, for the current example, approximately 40.5% of the variance in student ratings and be accounted for by knowledge of the grades given in the class.

Overall Model Fit and Statistical Inference

With a given sample of data it is possible to obtain regression coefficients, R , and R^2 that suggest some relationship in the population between X and Y . To determine, however, if the model of Y with various X 's is explaining variation in Y above what one would expect by chance alone requires the use of statistical significance tests.

The question of interest is whether the combination of the X 's (for multiple regression) explains any variation in Y beyond what could be due to chance. The hypothesis of interest is:

$$H_0: R^2 = 0$$

or

$$H_0: \beta_j = 0.00;$$

note that both hypotheses are equivalent.

The alternative hypotheses would simply indicate that R^2 is not equal to 0.00 (thus some variation in Y is being explained or predicted), or that at least one of the regression coefficients is not likely to equal zero (for the multiple regression interpretation).

In short, if the null hypotheses are rejected, then one may conclude that some aspect of the model used, i.e., the IVs selected, is statistically related to Y (or at least predicts Y).

To test the null H_0 , one must perform an *overall F test* of R^2 . This overall F test is simply the F test learned in one-way ANOVA. F is calculated, like ANOVA, using the following formula:

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

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:

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

The 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 . The more commonly used symbols for any F test are df_1 and df_2 . The formulas for F dfs are:

$$df_1 = k$$

and

$$df_2 = n - k - 1$$

where k is the number of IVs in the model (or number of vectors or columns) and n is the sample size or number of observations in the sample.

For the example 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

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

Since the calculated F ratio is larger than the critical, then null hypothesis of no explained or predicted variation is rejected. In short, the model appears to provide some explanatory power for Y.

As you might expect, p-values are provided for the F test, with the usual decision rule. The p-value for the obtained F ratio is .006, which is clearly smaller than .05.

A final model fit statistic is the adjusted R^2 , adj. R^2 . With sample data there is a tendency in regression analysis to overstate or overestimate the model explained variance in Y—that is, R^2 is typically too large relative to the population parameter; R^2 is biased upwards. A correction for this bias results in the adj. R^2 measure. Adj. R^2 will always be smaller than R^2 , with the degree of difference less in larger samples. The formula for adj. R^2 is:

$$\text{adj. } R^2 = 1 - \frac{\text{MSE}}{\text{VAR}(Y)}$$

where MSE is the variance error of the estimate, and VAR(Y) is the variance of Y. For example, the variance of Y is 0.3069, and MSE = 0.1944, so

$$\text{adj. } R^2 = 1 - \frac{\text{MSE}}{\text{VAR}(Y)} = 1 - \frac{0.1944}{0.3069} = 1 - .6334 = \underline{.3666}.$$

Obtaining and Reporting Regression Results

Regression results can be obtained from SPSS using the following command, which is found under the regression menu (results reported below command):

```
REGRESSION /DESCRIPTIVES MEAN STDDEV CORR SIG N /MISSING LISTWISE
/STATISTICS COEFF OUTS CI R ANOVA /CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN /DEPENDENT perform /METHOD=ENTER per_as.
```

* * * * M U L T I P L E R E G R E S S I O N * * * *

Listwise Deletion of Missing Data

	Mean	Std Dev	Label
PERFORM	4.153	.554	Evaluation Ability
PER_AS	48.765	10.220	Percentage of A's given in course

N of Cases = 17

Correlation, 1-tailed Sig:

	PERFORM	PER_AS
PERFORM	1.000	.637
	.	.003
PER_AS	.637	1.000
	.003	.

Multiple R	.63652
R Square	.40516
Adjusted R Square	.36550
Standard Error	.44092

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	1	1.98623	1.98623
Residual	15	2.91612	.19441

F = 10.21684 Signif F = .0060

----- Variables in the Equation -----

Variable	B	SE B	95% Confdnce Intrvl B	Beta
PER_AS	.034476	.010786	.011486 .057466	.636521
(Constant)	2.471719	.536738	1.327690 3.615749	

----- in -----

Variable	T	Sig T
PER_AS	3.196	.0060
(Constant)	4.605	.0003

The IV is PER_AS (percentage A's given in the course). Note all components previously discussed, such as R, R², and adj. R²; the regression coefficients b₀ (denoted as constant), and b₁ (B for PER_AS); the ANOVA summary table with SSR, SSE, and MSE; and the inferential tests—overall F test ($F = 10.21$, $p = .006$), and t tests for the coefficients.

Reporting results may take several forms. Most common is a tabular display, although for models with few IVs, reporting results within the text of your manuscript may be feasible.

APA tabular style appears on page 132 of the 4th edition manual. For the current example, the table could appear as:

Table X
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. $R^2 = .36$, $F = 10.22^*$, $df = 1,15$; $n = 17$

* $p < .05$.

When tables are used, one may report results (and interpretations) as:

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.

Exercises

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

Student	High School GPA	Estimated Number of Hours of Study Per Week at Home
Bill	3.33	3
Bob	1.79	5
Stewart	2.21	12
Linda	3.54	9
Lisa	2.89	11
Ann	2.54	1
Fred	2.66	0
Carter	1.10	3
Kathy	3.67	2

(2) Does SAT adequately predict college success?

Student	Freshmen Collegiate GPA	SAT Scores
Bill	3.33	1000
Bob	1.79	750
Stewart	2.21	890
Linda	3.54	1100
Lisa	2.89	900
Ann	2.54	860
Fred	2.66	1010
Carter	1.10	640
Kathy	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.

Student	GPA	Tardies/Absences
Bill	3.33	2
Bob	1.79	10
Stewart	2.21	5
Linda	3.54	6
Lisa	2.89	3
Ann	2.54	4
Fred	2.66	6
Carter	1.10	12
Bill	3.10	3
Sue	2.10	8
Loser	2.31	6
Kathy	3.67	2