Simple Linear Regression:
One Qualitative IV

Simple linear regression with one qualitative IV variable is essentially identical to linear regression with quantitative variables. The primary difference between the two is how one interprets the regression coefficients.

The Regression Equation for One Dichotomous IV

Suppose we are interested in determining whether student achievement in high school biology is affected by the use of cooperative learning in the classroom. As an experiment, two biology classes taught by the same instructor are used. In the first class, students are taught to use the jigsaw method of cooperative learning. The second class serves as the control group, and these students do not participate in any form of cooperative activity. The outcome of interest is end-of-year achievement scores as measured by a final examination administered to both classes. The data are presented in Table 1.

In the regression equation of this data, achievement is the DV, denoted Y, and use of cooperative learning strategy is the IV, denoted, X. A given student's examination grade, Y, may be expressed as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{1}$$

where

Y_i signifies the i'th student's score, β_1 is the population regression coefficient expressing the relationship between X and Y,

β_0 is the population intercept for the equation, and

ε_i is the a random error term.

Table 1
Student Ratings and Course Grades Data

Achieve	Instruction
75.00	0.00
89.00	0.00
72.00	0.00
81.00	0.00
82.00	0.00
87.00	1.00
95.00	1.00
89.00	1.00
91.00	1.00
82.00	1.00

Note. For instruction, 0 indicates the control group, and 1 denotes the experimental group. The mean for the control group is 79.8 (SD = 6.61), and for the experimental group is 88.8 (SD = 4.816).

Note that both β_0 and β_1 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:

$$Y_i = b_0 + b_1 X_i + e_i \tag{2}$$

where

b_0 is the sample intercept, b_1 is the sample regression coefficient, and e is the sample error term in the model. 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. As with simple regression with one quantitative IV, the most importance goal is to obtain the best possible mathematical description of the relationship between Y and X.

Ordinary Least Squares
As before, OLS is used to obtain estimates of population regression coefficients.

Dummy Coding and Coefficient Interpretation
When the IV is dichotomous, a simple procedure for coding the IV is <u>dummy</u> coding. With dummy coding, one group takes the value 0, and the other group the value 1. The data in Table 1 for group membership is dummy coded: the control group equals 0, and the experimental group 1.

Dummy variable coefficients typically represent group means or deviations from means. For example, b_0 represents the mean for the group whose value is 0, and the equation $b_0 + b_1$ represents the mean for the group whose value is 1.

For the current example, the regression equation is:

$$Y_i = b_0 + b_1 X_i + e_i$$

$$Y_i = 79.8 + 9.00 (X_i) + e_i$$

Note that the control group, whose dummy value is 0.00, has a mean of 79.8, which is replicated by the following regression equation:

Control Mean = $b_0 + b_1 X$
Control Mean = $b_0 + b_1 (0)$
Control Mean = b_0
Control Mean = $79.8 + 9.00 (0)$
Control Mean = 79.8

The mean of the experimental group is:

Experimental Mean = $b_0 + b_1 X$
Experimental Mean = $b_0 + b_1 (1)$
Experimental Mean = $b_0 + b_1$
Experimental Mean = $79.8 + 9.00 (1)$
Experimental Mean = $79.8 + 9.00$
Experimental Mean = 88.8.

As outlined above, one may interpret b_0 as group 0's mean, and b_1 is the deviation or difference between the two groups' means, which in this example is 9.00 points.

Inferential Procedures
Once sample coefficients are obtained from OLS, one must next determine whether the sample estimates are likely to be due to chance fluctuations in the data. As before, each coefficient estimate is divided by its standard error, and the product forms a t ratio.

The hypothesis of interest when a dummy variable is present is:

H_0: $\beta_1 = 0.00$.

Recall that β_1 represents the mean difference between the two groups. If H_0 is not rejected, then one may conclude that the groups do not differ statistically. If, however, β_1 is rejected, then one may conclude that the mean difference between the groups is statistically significant.

The non-directional alternative hypothesis states that a difference does exist:

H_1: $\beta_1 \neq 0.00$.

Note that H_0 listed above is identical to the null hypothesis for the two-independent samples t-test:

H_0: $\mu_1 = \mu_2$

or

H_0: $\mu_1 - \mu_2 = 0.00$,

that is, the difference between the two groups is 0.00.

Also note that the null hypothesis for ANOVA with only two groups is identical to the two-independent samples t-test and regression with a dummy IV. The null in ANOVA for two groups is:

H_0: $\mu_1 = \mu_2$.

As these hypotheses reveal, regression with a dummy variable is identical to the two-independent samples t-test and to ANOVA with two groups. Output from all three will be shown shortly to confirm this.

The procedure for testing H_0 for dummy IVs is identical to the procedure outlined for regression with one quantitative IV. Since the ratio of the coefficient to its standard error forms a t-ratio, one may choose to use critical t-values for determining statistical significance. But with statistical software, p-values are usually reported, so one may simply employ the usual decision rule to test H_0:

If p ≤ α, reject H_0, otherwise fail to reject H_0.

Example Output
Results obtained from the regression equation stated above for the data in Table 1 follows:

```
        * * * *   M U L T I P L E   R E G R E S S I O N   * * * *
--------------------- Variables in the Equation ----------------------
Variable              B        SE B      95% Confdnce Intrvl B      Beta
GROUP          9.000000    3.657868       .564949    17.435051    .656323
(Constant)    79.800000    2.586503     73.835519    85.764481


----------- in ------------
Variable           T  Sig T
GROUP          2.460  .0393
(Constant)    30.852  .0000
```

Substituting the above estimates into the regression equation yields:

$$Y' = b_0 + b_1 X_i$$

$$Y' = 79.8 + 9.00 \, (X_i)$$

The ratio of b_1 to its standard error, SE_{b1}, results in a t-ratio of:

$$t_{b1} = b_1 / SE_{b1} = 9.00 / 3.658 = 2.46.$$

     The corresponding p-value is .039. Since this p-value is less than .05, the null is rejected, so we conclude that a statistically significant difference exists between the two groups.
     Output for a t-test of the data in Table 1 is presented below.

t-tests for Independent Samples of GROUP

```
                                Number
Variable                       of Cases      Mean        SD     SE of Mean
-----------------------------------------------------------------------
ACHIEVE

contro                             5        79.8000      6.611      2.956
experi                             5        88.8000      4.817      2.154
-----------------------------------------------------------------------

        Mean Difference = -9.0000

        Levene's Test for Equality of Variances: F= .629   P= .451

      t-test for Equality of Means                                    95%
Variances  t-value    df    2-Tail Sig    SE of Diff        CI for Diff
-----------------------------------------------------------------------
Equal        -2.46     8        .039          3.658      (-17.435, -.565)
Unequal      -2.46   7.31       .042          3.658      (-17.575, -.425)
-----------------------------------------------------------------------
```

Note that the t-ratio obtained from the t-test, 2.46, and the corresponding p-value, .039, are identical to the results found with regression.

     ANOVA output is also listed below.

```
- - - - - O N E W A Y - - - - -

      Variable  ACHIEVE
   By Variable  GROUP
                              Analysis of Variance

                                 Sum of        Mean          F      F
          Source          D.F.   Squares      Squares      Ratio   Prob.

Between Groups            1      202.5000     202.5000     6.0538  .0393
Within Groups            8      267.6000      33.4500
Total                    9      470.1000
```

The p-value for the ANOVA is identical to the p-value from the regression. In addition, the relationship between the t-ratio and the F-ratio reported in ANOVA is

$$F = t^2$$

when df_1 is equal to 1. Thus,

$$F = t^2$$

$$6.05 = (2.46)^2 = 6.05.$$

As these numbers reveal, regression with a dummy variable is identical to the two-independent samples t-test and the one-way ANOVA with two groups.

<u>Interval Estimation: Confidence Intervals (CI)</u>   Recall that a CI represents an upper and lower bound to the <u>point estimate</u> of the regression coefficient. Thus, CIs indicate the precision of the particular estimate.
        The CI for b_1 is:

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

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_{b1} is the standard error of b_1.
        For the current example, the 95% confidence interval (.95CI) for b_1 is

.95CI: $b_1 \pm t_{(\alpha/2, df)} SE_{b1}$
.95CI: $9.00 \pm (2.306)(3.658)$
.95CI: 9.00 ± 8.435
.95CI: $(17.435, 0.565)$.

Such a CI enables the researcher to state that one may be 95% confident that the true population difference between the two groups in achievement may be as high as 17.44 points, or as low as only 0.57 points. As this example illustrates, CIs provide a sense of precision that point estimates do not. Due to the .95CI, we can note that the benefit derived from cooperative learning may be only as large as half a point in achievement! Clearly the results obtained indicate the benefit estimated is not very precise; a larger sample, such as n = 100, would provide a much smaller CI, and thus more precise estimate of the benefit of cooperative learning.

Model Fit

Interpretations for Multiple R, R^2, adj. R^2, SEE, and MSE provided in notes on regression with one quantitative IV continue to hold here. The linkage between regression and ANOVA also holds. The ANOVA summary table for the current example is provided below.

Source	D.F.	Sum of Squares	Mean Squares	F Ratio	F Prob.
Between Groups	1	202.5000	202.5000	6.0538	.0393
Within Groups	8	267.6000	33.4500		
Total	9	470.1000			

The regression and residual sums of squares reported in regression are also provided below.

```
Analysis of Variance
                DF        Sum of Squares      Mean Square
Regression       1           202.50000        202.50000
Residual         8           267.60000         33.45000
F =      6.05381     Signif F =  .0393
```

Note that the model fit information for regression, in terms of MSE (33.45) and F ratios are identical to the ANOVA results. The R^2 may be calculated for both ANOVA and regression using

$$R^2 = SSR/SST$$
$$R^2 = 202.5/470.1$$
$$R^2 = 0.43.$$

About 43% of the variance in Y is explained or accounted for with X. The adj. R^2 is .36.

Overall Model Fit and Statistical Inference

With any regression or ANOVA model, the initial question of interest is whether the model, i.e., the combination of Xs, explains more variation in Y than would be expected by chance. This hypothesis, the hypothesis of overall model fit, is denoted

$$H_0: R^2 = 0$$

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 by the Xs), 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 hypothesis is 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).

As before, the overall F test is used to test H_0. F is calculated, like ANOVA, using any of the following formulae:

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

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

$$F = \frac{R^2/k}{(1 - R^2)/(n-k-1)} = \frac{.431/1}{(1 - .431)/(10 - 1 - 1)} = 6.05$$

or

$$F = \frac{SSR/df_r}{SSE/df_e} = \frac{202.5/1}{267.6/8} = 6.05.$$

With $df_1 = 1$ and $df_2 = 8$, the .05 level critical F value is

$$_{.05}F_{1,8} = 5.32.$$

Since the calculated F ratio is larger than the critical, the 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 .03, which is clearly smaller than .05.

A final model fit statistic is adj. R^2. 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

$$\text{adj. } R^2 = 1 - \frac{\text{MSE}}{\text{VAR(Y)}} = 1 - \frac{33.45}{52.23} = 1 - .6404 = .3596.$$

Assessment of Assumption Violations

Residuals are calculated as explained earlier. A residual is the difference between the predicted and observed Y:

$$e_i = Y - Y'.$$

For example, the predicted achievement level for students in the experimental group is:

$$Y' = b_0 + b_1(1)$$
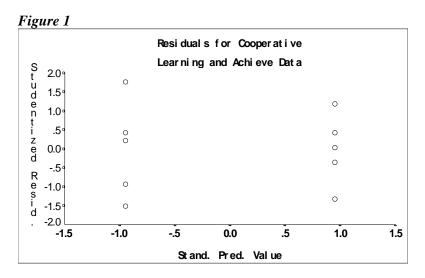$$Y' = 79.8 + 9.00 (1)$$
$$Y' = 88.8$$

If a student in this group had an observed achievement score of 95, then the residual is:

$$e_i = Y - Y'$$
$$e_i = 95 - 88.8$$
$$e_i = 6.20.$$

The predicted score for this individual was 6.20 point below the student's observed score.

To assess violations, one may simply plot the Studentized residuals against standardized predicted values, as is done below in Figure 1. While scatterplots of data with categorical IVs will appear different from scatterplots with quantitative IVs, assessment of violations will remain essentially unchanged. Primary emphasis should be on determining whether too much variation exists across groups (heterogeneity of variance). Notice from the graph below that both groups have scatter, or spread, that is essentially identical; thus, there is no apparent violation of assumptions with this data.

Figure 1

Residuals for Cooperative Learning and Achieve Data

Factors Affecting Power and Precision

The factors outlined in simple regression with one quantitative IV continue to hold here. Precision typically refers to the width of the CI, although it may also refer to the size of MSE or SEE.

Standardized Regression Equation

When categorical (i.e., qualitative) IVs are included in the regression equation, one should not use or report the standardized regression equation.

Obtaining and Reporting Regression Results
Regression results can be obtained from SPSS using the command below:

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

```
          * * * *   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
ACHIEVE   84.300   7.227
GROUP       .500    .527


N of Cases =     10


Correlation, 1-tailed Sig:


            ACHIEVE       GROUP
ACHIEVE      1.000         .656
               .            .020


GROUP         .656        1.000
              .020          .


Variable(s) Entered on Step Number
   1..    GROUP


Multiple R           .65632
R Square             .43076
Adjusted R Square    .35960
Standard Error      5.78360


Analysis of Variance
                  DF       Sum of Squares      Mean Square
Regression         1          202.50000        202.50000
Residual           8          267.60000         33.45000


F =        6.05381       Signif F =  .0393


--------------------- Variables in the Equation ----------------------
Variable             B        SE B      95% Confdnce Intrvl B     Beta
GROUP         9.000000   3.657868      .564949    17.435051   .656323
(Constant)   79.800000   2.586503    73.835519    85.764481


----------- in ------------
Variable          T  Sig T
GROUP         2.460  .0393
(Constant)   30.852  .0000
```

The IV is GROUP. Note all components previously discussed, such as R, R^2, and adj. R^2; the regression coefficients b_0 (denoted as constant), and b_1 (B for GROUP); the ANOVA summary table with SSR, SSE, and MSE; and the inferential tests—overall F test (F = 6.05, p = .039), 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. For the current example, the table would appear as:

Table 1
Descriptive Statistics for Biology Achievement by Type of Instruction

Instruction	Mean	SD	N
Coop. Learning	88.80	4.82	5
Control	79.80	6.61	5

Table 2
Summary of Regression Analysis for Type of Instruction Predicting Biology Achievement

Variable	b	se b	95%CI	t
Cooperative Learning	9.00	3.658	.56, 17.44	2.46*
Intercept	79.8	2.586	73.84, 85.76	30.85*

Note. R^2 = .43, adj. R^2 = .36, p < .05, F = 6.05, MSE = 33.45, df = 1,8, n = 10.
*p < .05.

This table does not contain standardized coefficient estimates or estimates of ΔR^2 estimates. For simple regression with one qualitative variable, nether of these estimates are applicable.

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

> Students exposed to cooperative learning average nine points higher on achievement than their control counterparts. This difference is statistically significant at the .05 level. Approximately 43% of variance in achievement in this sample is statistically related to the type of instruction used.

An alternative presentation of results, without tables, is:

> Students exposed to cooperative learning average nine points higher on achievement than their control counterparts. This difference is statistically significant ($F_{1,8}$ = 6.05, p = .039, MSE = 33.45) at the .05 level. Approximately 43% of variance in achievement in this sample is statistically related to the type of instruction used, with an adj. R^2 of .36.

Exercises

(1) 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. In some classes, students are exposed to a test every week. In other classes, students are tested three times during the quarter. Is there evidence that testing frequency is related to average achievement?

Quarter	Testing Frequency During Quarter	Overall Class Achievement on Final Exam
Fall 1991	3 times	85.5
Winter 1992	3 times	86.5
Spring 1992	3 times	88.9
Summer 1992	weekly	89.1
Fall 1992	3 times	87.2
Winter 1993	weekly	90.5
Spring 1993	weekly	89.8
Summer 1993	weekly	92.5
Fall 1994	weekly	89.3
Winter 1994	3 times	90.1

(2) Two classes of educational research were taught with two different methods of instruction, teacher guided (TG) and self paced (SP). Which had the better student achievement at the end of the quarter?

TG scores: 95, 93, 87, 88, 82, 92
SP scores: 78, 89, 83, 90, 78, 86

(3) A researcher wishes to know whether home study is related to general achievement amongst high school students. Students were surveyed, and all students who indicated that they routinely studied were coded 1, others were coded 0.

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

(4) Determine whether a statistical difference exists between men and women in weight:
Men: 156, 158, 175, 203, 252, 195
Women: 149, 119, 168, 123, 155, 126

The Regression Equation for One Categorical IV

Often qualitative variables take more than two categories. For each distinct group or category in the IV, c - 1 dummy variables are needed to represent the variable in the regression equation, where c is the number of categories. For example, suppose one wishes to model attitude toward increased millage rates by race of local residents. Three groups are of interest, Black, Hispanic, and White residents. Since there are three groups, (c - 1 = 3 - 1 = 2) two dummy variables will be needed to represent this variable in the regression equation. Or if, for instance, a categorical IV had 9 groups, 9 - 1 = 8 dummy variables would be needed. Dummy variable representation of the qualitative variable race is illustrated below in Table 2.

Table 2

Y = attitude	race	B = Black	H = Hispanic
7	B	1	0
9	B	1	0
6	B	1	0
8	B	1	0
9	H	0	1
8	H	0	1
8	H	0	1
6	H	0	1
6	W	0	0
5	W	0	0
3	W	0	0
3	W	0	0

Note. Means on Y by group: Black = 7.5, Hispanic = 7.75, and White = 4.25.

Each column or dummy variable is referred to as a <u>vector</u>. When calculating the degrees of freedom for the overall F test, to be discussed later, each vector is considered a variable, that is:

$$df_e = df_2 = n - k - 1$$

where n is the overall sample size, and k is the number of variables (or vectors) in the model. For the data in Table 1, the error degrees of freedom for the overall F test are:

$$df_2 = 12 - 2 - 1 = 9.$$

The sample regression equation for this data is

$$Y_i = b_0 + b_1 B + b_2 H + e_i$$

where B represents the dummy variable for Black residents and H is the dummy variable for Hispanic residents.

Coefficient Interpretation

As with one dichotomous dummy variable, the intercept, b_0, represents the mean for the omitted group, which in this example is White residents. For example, when both B and H, the dummy variables, equal zero, then:

$$Y'_{White} = b_0 + b_1 B + b_2 H$$

$Y'_{White} = b_0 + b_1(0) + b_2(0)$
$Y'_{White} = b_0$

where the predicted mean for White residents is symbolized as Y'_{White}. The predicted mean for Black residents is:

$Y'_{Black} = b_0 + b_1B + b_2H$
$Y'_{Black} = b_0 + b_1(1) + b_2(0)$
$Y'_{Black} = b_0 + b_1(1)$.

And the predicted mean for Hispanic residents is:

$Y'_{Hispanic} = b_0 + b_1B + b_2H$
$Y'_{Hispanic} = b_0 + b_1(0) + b_2(1)$
$Y'_{Hispanic} = b_0 \qquad + b_2(1)$.

Note that the omitted group, White residents in this example, serves as the comparison or reference group. By comparison, what is meant is that the other groups are compared to the omitted group. As a result, the regression coefficients, b_1 and b_2, in this example, represent the mean deviation of the Black and Hispanic residents' responses, respectively, from the White residents' mean. For example, White respondents' mean on Y is 4.25 and Black respondents' mean is 7.5. The difference between these two groups is

7.5 - 4.25 = 3.25.

This difference of 3.25 points will be reflected in the regression equation by the regression coefficient associated with the dummy for Black residents—b_1. Similarly, the difference between Hispanic and White residents is

7.75 - 4.25 = 3.5,

and this difference is represented in the regression equation by b_2.

The regression equation for the data in Table 2 is

$Y' = b_0 + b_1B + b_2H$
$Y' = 4.25 + 3.25(B) + 3.50(H)$.

Note that the regression coefficient for Black respondents is $b_1 = 3.25$, which is the exact mean difference between White and Black respondents in terms of Y. Also, the mean difference between White and Hispanic respondents is 3.50, which is reflected by b_2 in the equation above.

Inferential Procedures for Coefficients and Pairwise Comparisons
The null hypothesis for each regression coefficient is expressed, symbolically, as previously outlined:

$H_0: \beta_0 = 0.00$,

H_0: $\beta_1 = 0.00$, and

H_0: $\beta_2 = 0.00$.

Since b_1 represents a <u>pairwise</u> comparison of White to Black respondents, an equivalent null for H_0: $\beta_1 = 0.00$ is

H_0: $\mu_{White} = \mu_{Black}$

and for H_0: $\beta_2 = 0.00$, the pairwise comparison of White to Hispanic respondents, the equivalent null is:

H_0: $\mu_{White} = \mu_{Hispanic}$

   As mentioned, b_1 represents the mean difference between White and Black respondents. When one simply compares one group to another, such contrasts are often referred to as pairwise comparisons. The ratio of b_1 to its standard error, se_{b1}, forms the usual t-ratio for the comparison. Also, p-values are presented for H_0: $\beta_1 = 0.00$. Similarly, b_2 is the mean difference between White and Hispanic respondents. As with b_1, ratio b_2 / se_{b2} forms a t-ratio that may be used to test the tenability of H_0.
   The procedure for testing H_0 for dummy IVs is identical to the procedure outlined for regression with one quantitative IV. Since the ratio of the coefficient to its standard error forms a t-ratio, one may choose to use critical t-values for determining statistical significance. But with statistical software, p-values are usually reported, so one may simply employ the usual decision rule to test H_0:

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

For example, the se of b_1 is 0.957, so the ratio is

$t_{b1} = b_1$ / $se_{b1} = 3.25$ / $0.957 = 3.39603$

with a corresponding p-value of .0079. The null hypothesis of no difference, i.e.,

H_0: $\beta_1 = 0.00$ or H_0: $\mu_{White} = \mu_{Black}$

is rejected since .0079 is less than .05.
   Similarly, the p-value for b_2 is .0053, which is smaller than .05, so the null is rejected, i.e., reject

H_0: $\beta_2 = 0.00$, or H_0: $\mu_{White} = \mu_{Hispanic}$.

   There is one remaining comparison, that of Black to Hispanic respondents. There are two ways this comparison may be assessed. First, the null hypothesis of no mean difference between Black and Hispanic respondents is

H_0: $\beta_1 - \beta_2 = 0.00$,

which is equivalent to

H_0: $\mu_{Black} = \mu_{Hispanics}$.

To test this hypothesis, one may use either (a) a hand-calculation procedure, or (b) a computer-calculation procedure. Both are illustrated below.

(a) Hand-calculation Procedure for Testing H_0: $\beta_1 - \beta_2 = 0.00$

To test the null of no difference, one must first compute the estimated difference between the two groups. The formula is not complex; the difference is simply:

$b_1 - b_2$ = difference between Black and Hispanic respondents.

To form the corresponding t-ratio requires the calculation of the standard error of the difference $b_1 - b_2$, i.e., calculating se_{b1-b2}. The formula for a given difference is:

$$se_{bi-bj} = \sqrt{SE_{bi}^2 + SE_{bj}^2 - 2COV_{bibj}}$$

where

se_{bi}^2 = the variance error of b_i is simply se_{bi} squared;

se_{bj}^2 = the variance error of b_j is simply se_{bj} squared; and

COV_{bibj} = the covariance between b_i and b_j., and this covariance is
multiplied by 2.

The covariance between the two regression coefficients may be obtained from SPSS by specifying, under the STATISTICS command in LINEAR REGRESSION, the output of COVARIANCE MATRIX.

To illustrate, hand computation of the t-ratio for the difference between Blacks and Hispanics is performed below. To assist in the computation, selected output from the regression analysis is presented.

```
          * * * *   M U L T I P L E   R E G R E S S I O N   * * * *
Var-Covar Matrix of Regression Coefficients (B)
Below Diagonal:  Covariance    Above:  Correlation


               H            B
H           .91667       .50000
B           .45833       .91667
```

Note above the variance-covariance matrix of the regression coefficients. The covariance between the coefficients is located on the lower diagonal of the matrix. The covariance between the coefficients for dummy variables B and H is $COV_{b1b2} = 0.458$.

The regression coefficients and their standard errors are reported below.

```
--------------------- Variables in the Equation ----------------------
Variable            B        SE B     95% Confdnce Intrvl B      Beta
B             3.250000    .957427    1.084151    5.415849      .774139
H             3.500000    .957427    1.334151    5.665849      .833688
(Constant)    4.250000    .677003    2.718513    5.781487

----------- in ------------
```

```
Variable           T  Sig T
B              3.395  .0079
H              3.656  .0053
(Constant)     6.278  .0001
```

The standard error of the coefficient for the dummy variable representing Black respondents is $se_{b1} = 0.957$, and the se for the dummy representing Hispanic respondents is $se_{b2} = 0.957$.

The estimated mean difference between the two group of respondents is:

$b_1 - b_2 = 3.25 - 3.50 = -0.25$.

The standard error of this difference is:

$$SE_{bi-bj} = \sqrt{SE_{bi}^2 + SE_{bj}^2 - 2COV_{bibj}}$$
$$SE_{b1-b2} = \sqrt{SE_{b1}^2 + SE_{b2}^2 - 2COV_{b1b2}}$$
$$SE_{b1-b2} = \sqrt{0.957^2 + 0.957^2 - (2)0.458}$$
$$= \sqrt{0.9158 + 0.9158 - 0.916}$$
$$= \sqrt{0.9156}$$
$$= 0.9569 \cong 0.957.$$

The calculated t-ratio resulting from this comparison is:

$$t_{b1-b2} = b_1 - b_2 / SE_{b1-b2}$$
$$t_{b1-b2} = (3.25 - 3.50) / 0.957$$
$$= (3.25 - 3.50) / 0.957$$
$$= -0.25 / 0.957 = -0.261.$$

This t-ratio must be compared against a critical t to determine statistical significance. A critical t, t_{crit}, is determined using the following degrees of freedom:

$df = n - k - 1$.

Note that this is the same df as that found for df_2 for the overall F test, discussed above. Remember that k represents the number of variables (and/or vectors) in the data. Thus, the df for the critical t value is:

$df = 12 - 2 - 1 = 9$.

The table of critical t-values indicates that $_{.05}t_{crit}$ is ± 2.262. Since the calculated t-ratio is smaller, in absolute value, than the critical value, the null is not rejected and one may conclude that Black and Hispanic respondents have, on average, essentially the same attitude toward millage rate increases.

(b) Computer-calculation Procedure for Testing $H_0: \beta_1 - \beta_2 = 0.00$

While the hand-computation procedure is necessary in some situations, an easier procedure does exist. The goal is to obtain the t-ratio for the mean difference between Black and Hispanic respondents. To do this, simply create another dummy variable to represent White respondents. The resulting data would appear as that in Table 3.

Table 3

Y = attitude	Race	B = Black	H = Hispanic	W = White
7	B	1	0	0
9	B	1	0	0
6	B	1	0	0
8	B	1	0	0
9	H	0	1	0
8	H	0	1	0
8	H	0	1	0
6	H	0	1	0
6	W	0	0	1
5	W	0	0	1
3	W	0	0	1
3	W	0	0	1

Note. Means on Y by group: Black = 7.5, Hispanic = 7.75, and White = 4.25.

To obtain the mean difference between Black and Hispanic respondents, simply re-run the regression using the following equation:

$$Y_i = b_0 + b_1 B + b_2 W + e_i$$

where B represents the dummy variable for Black respondents and W is the dummy variable for White respondents. This regression equation treats Hispanics as the omitted or reference group. As a result, the new b_1 coefficient will represent the mean difference between Black and Hispanic respondents, therefore, the resulting t-ratio and p-value can be used to assess whether a statistically significant difference exists between these two groups.

Results of the new regression are as follows:

```
--------------------- Variables in the Equation ----------------------
Variable               B       SE B     95% Confdnce Intrvl B      Beta
B              -.250000    .957427   -2.415849     1.915849    -.059549
W             -3.500000    .957427   -5.665849    -1.334151    -.833688
(Constant)     7.750000    .677003    6.218513     9.281487

----------- in ------------
Variable          T   Sig T
B             -.261   .7999
W            -3.656   .0053
(Constant)   11.448   .0000
```

Note that the new b_1 is equal to -.25, which is the mean difference obtained between Black and Hispanic respondents in the hand-computation section above. Also note that the obtained t-ratio of -.261 is identical to the hand-calculated t-ratio presented earlier. Finally, note the p-value associated with b_1. Since this p-value is larger than .05, .79 > .05, we may conclude that Black and Hispanic respondents do not differ in their attitude toward millage rate increases.

Note that when making comparisons using this method, the overall model fit statistics (e.g., R^2, adj. R^2, etc.) will remain exactly the same whether one uses Black, Hispanic, or White respondents as the reference group, so it is only necessary to report results for the initial overall model fit statistics from the fist model.

Interval Estimation: Confidence Intervals (CI)

A CI may be calculated for each comparison. For example, the CI for b_1 is:

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

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_{b1} is the standard error of b_1.

For the current example, in which the regression model is

$$Y_i = b_0 + b_1B + b_2H + e_i$$

the 95% confidence interval (.95CI) for b_1 is

.95CI: $b_1 \pm t_{(\alpha/2,df)}SE_{b1}$
.95CI: $3.25 \pm (2.262)(0.957)$
.95CI: 3.25 ± 2.165
.95CI: (5.415, 1.085).

Note that this CI may be obtained from the regression output presented below.

```
--------------------- Variables in the Equation ----------------------
Variable              B        SE B    95% Confdnce Intrvl B      Beta
B               3.250000    .957427    1.084151    5.415849    .774139
H               3.500000    .957427    1.334151    5.665849    .833688
(Constant)      4.250000    .677003    2.718513    5.781487

----------- in ------------
Variable          T   Sig T
B               3.395  .0079
H               3.656  .0053
(Constant)      6.278  .0001
```

The .95CI for the comparison between Black and Hispanic respondents, i.e.,

$$b_1 - b_2,$$

can be obtained either through the hand calculation methods or the computer calculation method. Recall that the difference was

$$b_1 - b_2 = 3.25 - 3.50 = -0.25$$

with a standard error of

$$SE_{b1-b2} = 0.957.$$

The .95CI for this comparison is

.95CI: $b_1-b_2 \pm t_{(\alpha/2,df)}SE_{b1-b2}$

.95CI: -0.25 ± (2.262)(0.957)
.95CI: -0.25 ± 2.165
.95CI: (1.915, -2.415).

Model Fit

Interpretations for Multiple R, R^2, adj. R^2, SEE, and MSE provided in notes on regression with one quantitative IV continue to hold here. The linkage between regression and ANOVA also holds. In fact, the regression of Y on a qualitative IV is exactly equal to one-way ANOVA, as will be illustrated below.

The model fit statistics for the model

$$Y_i = b_0 + b_1 B + b_2 H + e_i$$

are:

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

Variable(s) Entered on Step Number
   1..    H
   2..    B

Multiple R            .80557
R Square              .64894
Adjusted R Square     .57092
Standard Error       1.35401

Analysis of Variance
                    DF        Sum of Squares       Mean Square
Regression           2             30.50000          15.25000
Residual             9             16.50000           1.83333

F =       8.31818     Signif F =   .0090
```

Note that the model fit information for regression, in terms of MSE (33.45) and F ratios are identical to ANOVA results, presented below, for data in Table 2.

```
                 - - - - -  O N E W A Y  - - - - -
        Variable  Y
     By Variable  RACE
                              Analysis of Variance
                              Sum of        Mean          F      F
          Source       D.F.   Squares       Squares     Ratio  Prob.
Between Groups           2     30.5000       15.2500    8.3182  .0090
Within Groups            9     16.5000        1.8333
Total                   11     47.0000

                            Standard   Standard
Group    Count     Mean     Deviation   Error    95 Pct Conf Int for Mean
Grp 1       4     7.5000      1.2910     .6455      5.4458   TO      9.5542
Grp 2       4     7.7500      1.2583     .6292      5.7478   TO      9.7522
Grp 3       4     4.2500      1.5000     .7500      1.8632   TO      6.6368

Total      12     6.5000      2.0671     .5967      5.1867   TO      7.8133
```

Note that the regression sums of squares, SSR, is identical to the ANOVA SS_b, both equal to 30.50. Similarly, all the other components are identical, including the overall model F ratio, 8.31, and its associated p-value, .009. In short, regression with a qualitative IV is identical to ANOVA.

The R^2 may be calculated for both ANOVA and regression using

R^2 = SSR/SST
R^2 = 30.5/47
R^2 = 0.649.

The adj. R^2 is .57.

Overall Model Fit and Statistical Inference

Remember that with any regression or ANOVA model, the initial question of interest is whether the model, i.e., the combination of Xs or <u>vectors</u>, explains more variation in Y than would be expected by chance. You should test the overall hypothesis <u>before</u> testing hypotheses for individual coefficients. The overall hypothesis in regression is denoted

H_0: R^2 = 0.00

or

H_0: β_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 by the Xs), 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 hypothesis is 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). If the overall hypothesis is rejected, the next step in the analysis is to examine and test the individual regression coefficients.

Also note that the overall fit hypothesis for regression tests the same thing as the overall ANOVA hypothesis of

H_0: $\mu_1 = \mu_2 = \mu_3$,

or

H_0: $\mu_{Blacks} = \mu_{Hispanics} = \mu_{Whites}$.

As before, the overall F test is used to test the H_0. F is calculated, like ANOVA, using any of the following formulae:

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

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 (or vectors) in the model and n is the sample size or total number of observations in the sample.

For the example data, the model F is

$$F = \frac{R^2/k}{(1 - R^2)/(n-k-1)} = \frac{.6489/2}{(1 - .6489)/(12 - 2 - 1)} = 8.31$$

or

$$F = \frac{SSR/df_r}{SSE/df_e} = \frac{30.5/2}{16.5/9} = 8.31.$$

With $df_1 = 2$ and $df_2 = 9$, the .05 level critical F value is

$$_{.05}F_{2,9} = 4.26.$$

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

The corresponding p-value for F is .009. Since this value is less than .05, the null is rejected.

Example Output
Results obtained from the regression equation stated above the data in Table 2 follows:

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

           * * * *   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
Y        6.500    2.067
B         .333     .492
H         .333     .492
N of Cases =    12


Correlation, 1-tailed Sig:
               Y          B          H
Y          1.000       .357       .447
              .         .127       .073

B           .357      1.000      -.500
            .127         .         .049

H           .447      -.500      1.000
            .073       .049         .


Variable(s) Entered on Step Number
   1..    H
   2..    B


Multiple R          .80557
R Square            .64894
Adjusted R Square   .57092
Standard Error     1.35401


Analysis of Variance
                 DF       Sum of Squares       Mean Square
Regression        2           30.50000          15.25000
Residual          9           16.50000           1.83333

F =      8.31818       Signif F =  .0090


Var-Covar Matrix of Regression Coefficients (B)
Below Diagonal: Covariance    Above:  Correlation


              H           B
H          .91667      .50000
B          .45833      .91667


-------------------- Variables in the Equation ----------------------
Variable            B        SE B     95% Confdnce Intrvl B       Beta
B            3.250000     .957427     1.084151     5.415849     .774139
H            3.500000     .957427     1.334151     5.665849     .833688
(Constant)   4.250000     .677003     2.718513     5.781487
```

```
----------- in ------------
Variable          T  Sig T
B             3.395  .0079
H             3.656  .0053
(Constant)    6.278  .0001
```

Results for this analysis could be presented as follows.

Table 1
Descriptive Statistics for Attitude to Millage Increases by Race

Race	Mean	SD	N
Black	7.50	1.29	4
Hispanic	7.75	1.26	4
White	4.25	1.50	4

Table 2
Regression Analysis of Millage Rate Attitude by Race

Variable	b	se b	95% CI	t
Black	3.25	0.957	1.08, 5.42	3.395*
Hispanic	3.50	0.957	1.33, 5.67	3.656*
Intercept	4.25	0.677	2.72, 5.78	6.278*

Note. $R^2 = .65$, adj. $R^2 = .57$, $p < .05$, n = 12, $F_{2,9} = 8.31$, $MSE = 1.83$, n = 12.
*$p < .05$.

Table 3
Comparisons in Mean Millage Attitudes by Race

Contrast	Estimated Mean Difference	Standard Error of Difference	95% CI of Difference
Black vs. White	3.25*	0.957	1.08, 5.42
Hispanic vs. White	3.50*	0.957	1.33, 5.67
Black vs. Hispanic	-0.25	0.957	-2.42, 1.92

*$p < .05$.

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

> Analysis of the data indicated that differences in attitudes toward millage rates varied by race. All pairwise comparisons between the three groups of respondents were performed, and results are presented in Table 3. Note that, in general, both Black and Hispanic respondents hold statistically more positive attitudes to rate increases than do White respondents, and there does not appear to be a substantial, or statistically significant, attitude difference between Black and Hispanic respondents.

While one may report results without tables, I recommend that tables be used to report descriptive statistics, regression results, and pairwise comparisons when three or more groups are involved.

Exercises

(1) According to the leadership literature, there are a number of different leadership styles. Listed below are scores obtained from an instrument designed to measure a particular leadership style, which will be referred to as style X. Of interest is whether X differs by school district type in terms of urbanity. A stratified random sample of school principals were selected from three district types (mostly urban, mostly suburban, and mostly rural).

The scores on style X range from 100 to 0. The closely the score to 100, the more the respondent conforms to style X, while the closer the score to 0, the less the respondent conforms to style X.

Is there any evidence that X differs among the three district types?

District Type	Style X
Urban	85
Urban	98
Urban	75
Urban	63
Urban	91
Urban	49
Urban	62
Suburban	49
Suburban	48
Suburban	56
Suburban	78
Suburban	35
Suburban	50
Rural	33
Rural	95
Rural	26
Rural	11
Rural	33
Rural	25
Rural	65

(2) Which of the following small car makers, if any, have the better fuel economy in terms of miles per gallon (MPG)? For this study, a small car is defined as any vehicle with a 2.2 litre (or less), 4-cylinder engine with less than 130 horsepower.

Maker	MPG	Maker	MPG	Maker	MPG	Maker	MPG
Honda	43	Toyota	37	GM	33	Subaru	33
Honda	37	Toyota	38	GM	31	Subaru	36
Honda	29	Toyota	41	GM	29	Subaru	34
Honda	38	Toyota	36	GM	33	Subaru	33
Honda	46	Toyota	33	GM	40	Subaru	36
Honda	35	Toyota	37	GM	33	Subaru	32
Honda	34	Toyota	34	GM	37	Subaru	37
Honda	37	Toyota	33	GM	26	Subaru	37

(3) A researcher is interested in learning whether frequency of reading at home to elementary-aged children produces differential effects on reading achievement. After obtaining information from a randomly selected sample of parents about this behavior, the following classifications and standardized achievement scores were recorded. (Note: frequency classifications as follows: A = less than once per month, B = once to three times per month, C = more than three times per month.)

Freq. of Reading	Achievement
A	48
A	37
A	47
A	65
B	57
B	39
B	49
B	45
C	61
C	55
C	51
C	30

Is frequency of reading at home related to student reading achievement?

(4) Does a difference in salary for beginning assistant professors exist at GSU by college? Which colleges appear to have difference salary levels from the other colleges?

College	Salary	College	Salary
Education	33500	Education	34000
Business	32900	Business	43000
Health	30000	Health	29000
Arts & Science	26200	Arts & Science	24000
Technology	27250	Technology	29000
Education	35000	Education	32500
Business	37000	Business	44500
Health	49000	Health	31000
Arts & Science	29000	Arts & Science	23500
Technology	27500	Technology	30000

Answers to Exercises

NOTE – these tables have not been altered to agree in format with tables presented above. Use the tables presented below only for checking your results, not for creating APA style tables.

(1) According to the leadership literature, there are a number of different leadership styles. Listed below are scores obtained from an instrument designed to measure a particular leadership style, which will be referred to as style X. Of interest is whether X differs by school district type in terms of urbanity. A stratified random sample of school principals were selected from three district types (mostly urban, mostly suburban, and mostly rural).

The scores on style X range from 100 to 0. The closer the score to 100, the more the respondent conforms to style X, while the closer the score to 0, the less the respondent conforms to style X. Is there any evidence that X differs among the three district types?

Answer

Table 1.X
Descriptive Statistics for Style by Location

Location	Mean	SD	N
Urban	74.71	17.67	7
Suburban	52.67	14.19	6
Rural	41.14	28.87	7

Table 1.Y
Regression of Style on Location

Variable	B	SE B	t
Rural	-33.57	11.51	-2.92*
Suburban	-22.05	11.98	-1.84
Intercept	74.71	8.14	9.18*

Note. $R^2 = .34$, adj. $R^2 = .26$, $p < .05$, n = 20, $F_{2,17} = 4.38$, MSE = 463.62.
* $p < .05$.

Table 1.Z
Contrasts in Mean Levels of Style X by Location of School

Contrast	Estimated Mean Difference	Standard Error of Difference	95% Confidence Interval of Difference
Rural vs. Urban	-33.57*	11.51	-57.85, -9.29
Suburban vs. Urban	-22.05	11.98	-47.32, 3.23
Suburban vs. Rural	11.52	11.98	-13.75, 36.80

* $p < .05$.

Analysis of the data (see Tables 1.Y and 1.Z) indicates that differences in Style X varies by location of school. All pairwise comparisons between the three locations were performed, and results are presented in Table 1.Z. In general, principals in schools located in mostly urban areas tend to hold the highest level of Style X, and this level is statistically greater than the level exhibited by principals located in mostly rural areas. There is no statistical difference between levels of Style X found in suburban areas and the levels observed in either urban or rural areas.

(2) Which of the following small car makers, if any, have the better fuel economy in terms of miles per gallon (MPG)? For this study, a small car is defined as any vehicle with a 2.2 litre (or less), 4-cylinder engine with less than 130 horsepower.

Answer

Table 2.X
Descriptive Statistics for MPG by Auto Maker

Maker	Mean	SD	N
Honda	37.38	5.26	8
Toyota	36.13	2.75	8
GM	33.75	4.37	8
Subaru	34.75	1.98	8

Table 2.Y
Regression of MPG on Maker

Variable	B	SE B	t
Honda	4.63	1.91	2.42
Subaru	2.00	1.91	1.05
Toyota	3.38	1.91	1.77
Intercept	32.75	1.35	24.27

Note. $R^2 = .19$, adj. $R^2 = .10$, $p = .12$, n = 32, $F_{3,28} = 2.16$, MSE = 14.56.* $p < .05$.

Table 2.Z
Contrasts in Mean Levels of MPG by Auto Maker

Contrast	Estimated Mean Difference	Standard Error of Difference	95% Confidence Interval of Difference
Honda vs. GM	4.63*	1.91	0.72, 8.53
Subaru vs. GM	2.00	1.91	-1.91, 5.91
Toyota vs. GM	3.38	1.91	-0.53, 7.28
Honda vs. Toyota	1.25	1.91	-2.66, 5.16
Subaru vs. Toyota	-1.38	1.91	-5.28, 2.53
Honda vs. Subaru	2.63	1.91	-1.28, 6.53

* $p < .05$.

Analysis of the data (see Table 2.Y) indicates that no statistical differences in average MPG were observed among the four automobile makers compared.

(EDUR 8132 Class: Note that while the overall model was not statistically significant, there was one statistically significant contrast—Honda vs. GM. Odds are this one rejection of the null is a Type 1 error. If multiple comparison procedures, such as Scheffé or Bonferroni, were used to control for inflation of the Type 1 error rate, then it is doubtful that Honda would be different from GM.)

(3) A researcher is interested in learning whether frequency of reading at home to elementary-aged children produces differential effects on reading achievement. After obtaining information from a randomly selected sample of parents about this behavior, the following classifications and standardized achievement scores were recorded. (Note: frequency classifications as follows: a = less than once per month, b = once to three times per month, c = more than three times per month.)

Is frequency of reading at home related to student reading achievement?

Answer

Table 3.X
Descriptive Statistics for Achievement by Frequency of Reading

Frequency	Mean	SD	N
Less than Once per Month	49.25	11.62	4
One to three times per Month	47.50	7.55	4
More than three times per Month	49.25	13.48	4

Table 3.Y
Regression of Achievement on Frequency of Reading at Home

Variable	B	SE B	t
One to three times	-1.75	7.89	-0.22
Three times or more	-0.00	7.89	-0.00
Intercept	49.20	5.58	8.83

Note. $R^2 = .007$, adj. $R^2 = -.21$, $p = .97$, n = 12, $F_{2,9} = 0.03$, MSE = 124.50.
* $p < .05$.

Table 3.Z
Contrasts in Mean Levels of Achievement by Frequency of Reading at Home

Contrast	Estimated Mean Difference	Standard Error of Difference	95% Confidence Interval of Difference
One to Three vs. None	-1.75	7.89	-19.60, 16.10
Three + vs. None	-0.00	7.89	-17.85, 17.85
One to Three vs. Three +	-1.75	7.89	-19.60, 16.10

* $p < .05$.

The analysis of the data (see Table 3.Y) indicates that no statistical differences in average levels of achievement exists among students exposed to different levels of reading at home.

(4) Does a difference in salary for beginning assistant professors exist at GSU by college? Which colleges appear to have difference salary levels from the other colleges?

Answer

Table 4.X
Descriptive Statistics for Salary by College

College	Mean	SD	N
Education	33750.00	1040.83	4
Business	39350.00	5384.24	4
Health	34750.00	9535.02	4
Arts & Science	25675.00	2507.82	4
Technology	28437.50	1297.03	4

Table 3.Y
Regression of Achievement on Frequency of Reading at Home

Variable	B	SE B	t
Health	1000.00	3591.12	0.28
Technology	-5312.50	3591.12	-1.48
Arts	-8075.00*	3591.12	-2.25
Business	5600.00	3591.12	1.56
Intercept	33750.00	2539.31	13.29

Note. $R^2 = .55$, adj. $R^2 = .43$, $p = .013$, $n = 20$, $F_{4,15} = 4.52$, MSE = 25792291.66.
* $p < .05$.

Table 3.Z
Contrasts in Mean Levels of Salary by College

Contrast	Estimated Mean Difference	Standard Error of Difference	95% Confidence Interval of Difference
Arts vs. Education	-8075.00	3591.12	-15729, -420
Business vs. Technology	10912.50	3591.12	3258, 18566
Arts vs. Business	-13675.00	3591.12	-21329, -6020
Arts vs. Health	-9075.00	3591.12	-16729, -1420

Note. All possible pairwise comparisons were performed; only those that were statistically significant at the .05 level were reported.

The data analysis (see Tables 4.Y and 4.Z) indicates that differences in salary do appear to exist among the various colleges at Georgia Southern. In general, the College of Arts and Sciences has the lowest average salary, and this salary is statistically less than the average salaries for faculty in the Colleges of Education, Business, and Health. While the College of Business has the highest average salary, this salary is only statistically higher than the colleges of Arts and Sciences, and Technology. No other statistically significant differences were observed.