Simple Linear Regression: One Qualitative IV

1. Purpose

As noted before regression is used both to explain and predict variation in DVs, and adding to the equation categorical variables extends regression flexibility and enables one to perform group contrasts in a way that is mathematically identical to ANOVA. 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.

2. Regression Equations

Population

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, (1)

where

Y_i = the i'th student's score,
β_1 = population regression coefficient expressing the relationship between X and Y,
β_0 = population intercept for the equation, and
ε_i = random error term.

Population 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

$Y_i = b_0 + b_1 X_i + e_i$, (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.

Sample Prediction Equation

$Y' = b_0 + b_1 X_i$

Note that the above equations are exactly the same as found when the IV is quantitative (refer back to notes on regression with one quantitative predictor).

2. Fictional Data for Group Comparison

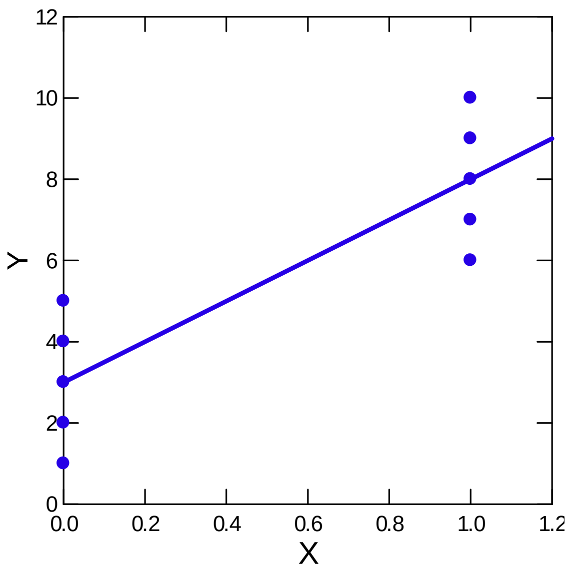
Table 1: Fictional Data by Sex (X)

Y	X	X = Sex
1	0	M
2	0	M
3	0	M
4	0	M
5	0	M
6	1	F
7	1	F
8	1	F
9	1	F
10	1	F

Note there are two groups X = 0 and X = 1 (i.e., X = 0 = Male, X = 1 = Female).

What are Y means on for X = 0? X = 1?

Slope for Fictional Data



(a) Given this regression equation, $Y' = b_0 + b_1X_i$, what would be the value of b_0?
(b) What would be the value of b_1?
(c) How to interpret this regression line?
(d) Find the coefficient estimates for the fictional data in SPSS.
(e) What is the literal interpretation for b_0?
(f) What is the literal interpretation for b_1?

3. Dummy Variable Coding

One method for representing categorical IV in regression is dummy coding. One group is coded with 0, the other with 1. The variable X in Table 1 above represents dummy coding for sex.

A dummy variable in regression represents the mean difference between the groups coded with 0 and 1. In the current example, the regression equation is:

$Y_i = b_0 + b_1X_i + e_i$,
$Y_i = 3.0 + 5.0(X_i) + e_i$,

Males are coded as X = 0, females coded as X = 1, so:

Predicted Y' mean for males = $Y_i = b_0 + b_1X_i$
Predicted Y' mean for males = $Y_i = b_0 + b_1\times 0$
Predicted Y' mean for males = $Y_i = b_0$

Predicted Y' mean for males = $Y_i = 3.0 + 5.0(X_i)$
Predicted Y' mean for males = $Y_i = 3.0 + 5.0(0)$
Predicted Y' mean for males = $Y_i = 3.0$

and the predicted Y' mean for females

Predicted Y' mean for females = $Y_i = b_0 + b_1X_i$
Predicted Y' mean for females = $Y_i = b_0 + b_1\times 1$
Predicted Y' mean for females = $Y_i = b_0 + b_1$

Predicted Y' mean for females = $Y_i = 3.0 + 5.0(X_i)$
Predicted Y' mean for females = $Y_i = 3.0 + 5.0(1)$
Predicted Y' mean for females = $Y_i = 3.0 + 5.0$
Predicted Y' mean for females = $Y_i = 8.0$

Given the above note the following:

b_0 = predicted Y' mean for group coded with dummy = 0;
$b_0 + b_1$ = predicted Y' mean for group coded with dummy = 1;
b_1 = predicted mean difference in Y between the two groups.

4. Fictional Data Example #2

Using the data below, create a dummy variable to represent student classification level.

Answer the following questions:

(a) What is the coefficient estimates for this model, that is, what are the values for b_0 and b_1?
(b) What is the literal interpretation for b_0?
(c) What is the literal interpretation for b_1?
(d) What is the predicted mean level of attitude for graduate students?
(e) What is the predicted mean level of attitude for undergraduate students?

Table 2: Fictional Data of Attitude Toward Recreation Activity Center and Student Classification Level

Attitude toward RAC at GSU	Student Level
2	Graduate
3	Graduate
5	Graduate
1	Graduate
2	Graduate
4	Undergraduate
5	Undergraduate
3	Undergraduate
5	Undergraduate
4	Undergraduate

Note: Attitude toward RAC scoring: 1 = very unhappy, 5 = very happy

5. Inferential Procedures for Regression Coefficient

Hypothesis testing is performed in exactly the same method as discussed previously with regression:

b_1/se = t-ratio

and

H_0: $\beta_1 = 0.00$
H_1: $\beta_1 \neq 0.00$

- Recall that β_1 is 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, H_0 is rejected, then one may conclude that the mean difference between the groups is statistically significant.

Dummy Variable for Two Groups: t-test and ANOVA Linkage

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.

Show result similarity with fictional data using t-test and ANOVA.
Note when $df_1 = 1$, then $F = t^2$

6. Confidence Intervals

CI represents an upper and lower bound to the point estimate of the regression coefficient. Thus, CIs indicate the precision of the particular estimate. The CI for b is:

$b \pm$ critical $t_{(\alpha/2,df)} \times SE_b$

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 is the standard error for the regression coefficient.

Illustrate 95%CI for intercept and slope for current example.
Calculate 99% CI for b_1.

Also, as before, if 0.00 lies within the CI, one will fail to reject H_0.

7. Model Fit

As before, model fit remains the same: R^2, adjusted R^2, SEE, and MSE. For example:

$$\text{adj. } R^2 = 1 - \frac{MSE}{VAR(Y)} = \textit{calculate for current example}$$

Similarly, errors are calculated in the same way:

$e = Y - Y'$

Calculate e for first observation in sample data.
Calculate e for last observation in sample data.

8. Overall Model Fit and Statistical Inference

To test the regression model as a whole—to learn whether any of the predictors are related to Y, or to learn whether the combination of predictors predict more variation in Y than one would expect by chance—one may test whether R^2 differs from 0.00:

H_0: $R^2 = 0$

or

H_0: $\beta_j = 0.00$ (all regression slopes equal 0.00)

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

$$F = \frac{SSR/df_1}{SSE/df_2} = \frac{SSR/k}{SEE/(n-k-1)} = \frac{MS_R}{MSE}$$

where;

SSR	=	regression sums of squares;
SSE	=	residual sums of squares;
df_1	=	regression degrees of freedom;
df_2	=	residual degrees of freedom;
k	=	number of independent variables (vectors) in the model;
n	=	sample size (or number of observations in sample);

MS_R = mean square (same as ANOVA) due to regression (e.g., between);
MSE = mean square error (same as ANOVA mean square within).

As before there are two sources for degrees of freedom:

df_1 = k (number of predictors in regression equation)

and

df_2 = n - k - 1

For the example data, the model F is

Illustrate calculation of F ratio with current data, find df and critical F – do we reject H_0?

9. Reporting Regression Results

Below is an example of APA styled reporting of results using the first fictional data given in Table 1:

Table 3: Descriptive Statistics for Y and X

Variable	Correlation	
	Y	X
Y	---	
X (sex)	.87*	---
Mean	5.50	0.50
SD	3.03	0.53

Note. n = 10, Sex coded 0 = male and 1 = female
* p < .05

Table 4: Summary of Regression of Y on X (Sex)

Variable	b	se b	95%CI	t
X (Sex)	5.00	1.00	2.68, 7.31	5.00*
Intercept	3.00	0.71	1.37, 4.63	4.24*

Note. R^2 = .76, adj. R^2 = .73, F = 25.00, MSE = 2.50, df = 1,8, n = 10; Sex coded 0 = male and 1 = female
*p < .05.

> Regression results show that females have, at the .05 level of significance, a statistically higher mean on Y than do males.

Note: The above regression 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.

10. 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	Regularly Study 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

11. Answers to Exercises

(1)

Table 1a: Descriptive Statistics and Correlations between Testing Frequency and Algebra Achievement

Variable	Correlations	
	Algebra Final	Testing Dummy
Algebra Final	---	
Testing Dummy	.67*	---
Mean	88.94	0.50
SD	2.06	0.53

Note. n = 10; Testing Dummy 1 = weekly tests, 0 = three tests per semester.
* $p < .05$

Table 1b: Regression of Algebra Achievement on Testing Frequency

Variable	b	se	95%CI	t
Testing Dummy	2.60	1.03	0.22, 4.98	2.52*
Intercept	87.64	0.73	85.96, 89.32	120.20*

Note. $R^2 = .44$, adj. $R^2 = .37$, F = 6.36*, df = 1,8; n = 10;
Testing Dummy 1 = weekly tests, 0 = three tests per semester.
*$p < .05$.

Regression results show that testing frequency appears to be related, at the .05 level of significance, to algebra achievement. Students who tested weekly scored about 2.60 points higher in algebra, on average, than did students who were test three times during the semester.

(2)

Table 2a: Descriptive Statistics and Correlations between Type of Instruction and Educational Research Achievement

Variable	Correlations	
	Ed. Research Scores	Instruction Dummy
Ed. Research Scores	---	
Instruction Dummy	.52	---
Mean	86.75	0.50
SD	5.58	0.52

Note. n = 10; Instruction Dummy 1 = teacher guided, 0 = self-paced.
* $p < .05$

Table 2b: Regression of Type of Instruction on Educational Research Scores

Variable	b	se	95%CI	t
Testing Dummy	5.50	2.90	-0.95, 11.95	1.90
Intercept	84.00	2.05	79.44, 88.56	41.03*

Note. $R^2 = .27$, adj. $R^2 = .19$, F = 3.61, df = 1,10; n = 12;
Instruction Dummy 1 = teacher guided, 0 = self-paced.
*$p < .05$.

Regression results show that there is not a statistical difference in mean educational research scores between those in teacher guided instruction and those in self-paced instruction. Students appear to perform similarly whether in teacher guided or self-paced instruction.

(3)

Table 3a: Descriptive Statistics and Correlations between GPA and Home Study

Variable	Correlations	
	GPA	Home Study
GPA	---	
Home Study Dummy	.59	---
Mean	2.64	0.56
SD	0.84	0.53

Note. n = 9; Home Study Dummy 1 = regularly studies at home, 0 = does not regularly study at home.
* $p < .05$

Table 3b: Regression of GPA on Home Study

Variable	b	se	95%CI	t
Home Study Dummy	0.95	0.49	-0.21, 2.10	1.94
Intercept	2.11	0.36	1.25, 2.97	5.80*

Note. $R^2 = .35$, adj. $R^2 = .26$, F = 3.78, df = 1,7; n = 9;
Home Study Dummy 1 = regularly studies at home, 0 = does not regularly study at home.
*$p < .05$.

Regression results show that there is not a statistical difference in mean GPA between students who regularly study at home and those who do not regularly study at home. GPA appears to be similar for both those who do and do not regularly study at home.

(4)

Table 4a: Descriptive Statistics and Correlations between Sex and Weight

Variable	Correlations	
	Sex Dummy	Weight
Sex Dummy	---	
GPA	.68*	---
Mean	0.50	164.92
SD	0.52	38.03

Note. n = 12; Sex Dummy 1 = males, 0 = females.
* $p < .05$

Table 4b: Regression of Weight on Sex

Variable	b	se	95%CI	t
Sex Dummy	49.83	16.79	12.42, 87.25	2.97*
Intercept	140.00	11.87	113.54, 166.46	11.79*

Note. $R^2 = .47$, adj. $R^2 = .42$, F = 8.81*, df = 1,10; n = 12;
Sex Dummy 1 = males, 0 = females.
*$p < .05$.

Regression results show that there is a statistically significant difference in mean weight between male and female participants—males, on average, weigh about 50 pounds more than females.