Multiple Linear Regression

1. Purpose—To Model Dependent Variables

Purpose of multiple and simple regression is the same, to model a DV using one or more predictors (IVs) and perhaps also to obtain a prediction equation.

2. Regression Equations

Primary difference in equations for multiple regression compared with simple regression is addition of IVs (Xs) which also leads to slight difference in literal interpretation (focus now on partial effects interpretation).

Population

Multiple regression (one DV and multiple IVs)

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \ldots + \beta_p X_{ip} + \varepsilon_i$

where

Y_i = represents individual scores on the DV per each ith person
β_0 = the intercept of the equation (predicted value of Y' when all IVs = 0.00)
β_p = the slope relating the pth IV (X_i) to the DV (Y_i)
X_{ip} = represents individual scores on the pth IV per each ith person
ε_i = residual or error term; defined as the deviation between observed Y and predicted Y'

Note that each symbol has same meaning as in simple regression except that now several IVs (the Xs) exist which means there are also several slopes, one for each IV.

Prediction equation

$Y' = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \ldots + \beta_p X_{ip}$

Where Y' is the predicted value of the DV in the population given the partial effects of each X. Note absence of ε; since means are predicted based upon the equation, individual score deviations from the prediction (Y-Y') are not included.

Sample

Multiple regression (one DV and one IV)

$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i3} + b_4 X_{i4} + \ldots + b_p X_{ip} + e_i$

Prediction equation

$Y' = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i3} + b_4 X_{i4} + \ldots + b_p X_{ip}$

Literal interpretation

Literal interpretation of regression coefficients b_0, b_1, and b_2 (these interpretations also apply to population parameters shown above):

b_0 = predicted value of Y, Y', when each of the predicts, the Xs, all equal 0
b_1 = expected change in predicted Y' for a one unit change in X_1 controlling for other Xs (holding other predictors at a constant value, such as their respective mean scores).
b_2 = expected change in predicted Y' for a one unit change in X_2 controlling for other Xs (holding other predictors at a constant value, such as their respective mean scores).
b_p = expected change in Y' for a one unit change in X_p, holding constant the other predictors.

3. Controlling Effects; Partial Effects

Multiple regression offers a much more realistic modeling opportunity compared with simple regression. Usually outcomes of interest, DVs, are functions of multiple causes and predictors simultaneously, and multiple regression can help show and model those factors.

Partial effects illustrated

Graphical display

To illustrate the partial effects of multiple regression, consider the following fictional data that includes mathematics scores, student height, student sex (0 = female, 1 = male), and a second set of mathematics scores.

Table 1: Fictional Mathematics Scores, Height, Sex, and Other Mathematic Scores

Math Scores	Height	Sex	Other Math	Math Scores	Height	Sex	Other Math
9	11	1	.	3	5	0	.
8	10	1	.	2	4	0	.
9	10	1	10	3	4	0	3
10	10	1	11	4	4	0	2
7	9	1	.	1	3	0	.
8	9	1	12	2	3	0	4
9	9	1	.	3	3	0	3
10	9	1	11	4	3	0	.
11	9	1	.	5	3	0	.
8	8	1	.	2	2	0	5
9	8	1	12	3	2	0	4
10	8	1	13	4	2	0	.
9	7	1	.	3	1	0	.

Note: For sex, 0 = female, 1 = male

These data are plotted in Figure 1 below. Examine Figure 1 and determine whether a relationship exists between mathematics scores and height.

Figure 1

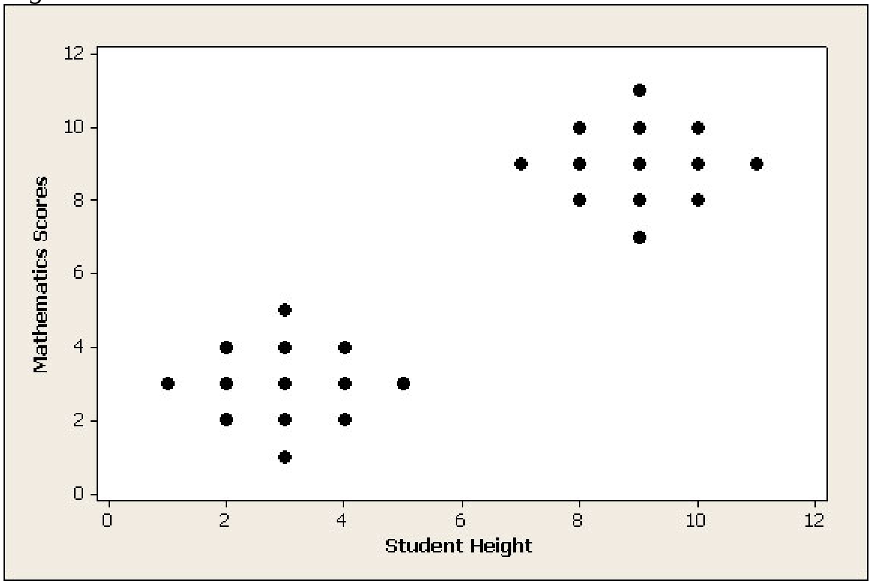


Figure 2 below shows how regression will model these data if sex is not controlled in the regression model. Note the positive relation between height and mathematics scores.

Figure 2

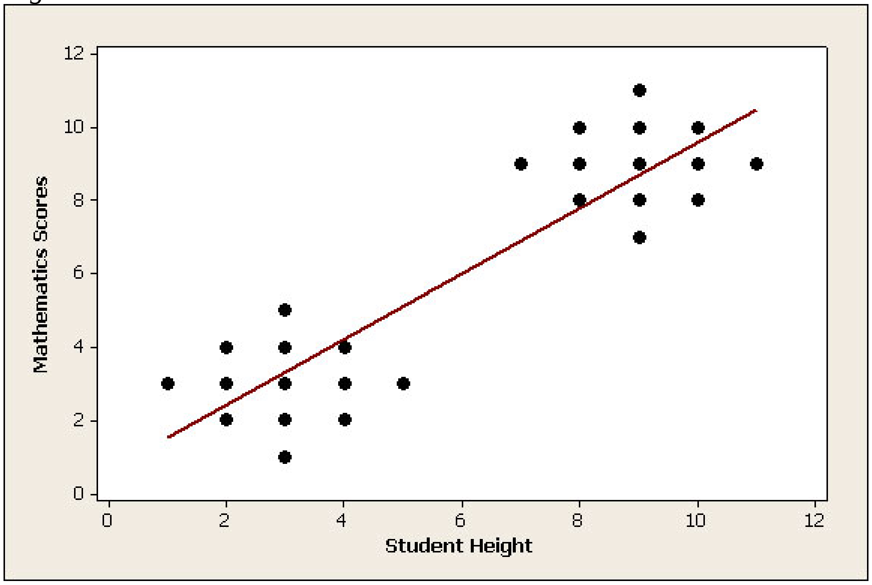
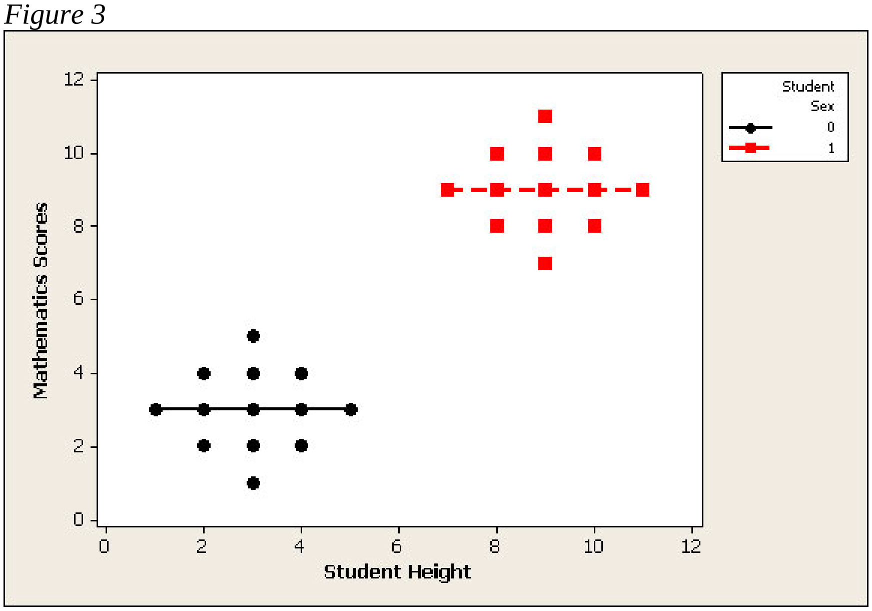


Figure 3 below identifies scores by sex and shows the regression equation between height and mathematics scores computed separately by sex. Note the slope of each line. If these data are modeled by multiple regression, then a more realistic analysis should arise.

Figure 3



Regression analysis

Model these data using regression first without sex as a second variable, then model again with sex. How does adding sex into the regression equation change the slope estimates?

Model 1 (simple regression with only height as predictor):

$Y_i = b_0 + b_1 Height_{i1} + e_i$

What is the literal interpretation for these coefficients?
What is the R^2 value for this model?

Model 2 (regression with both height and sex as predictors):

$Y_i = b_0 + b_1 Height_{i1} + b_2 Sex_{i2} + e_i$

What is the literal interpretation for these coefficients?
What is the R^2 value for this model? How did the R^2 value change from Model 1?

Partial effects illustrated, again

Graphical display

To show partial effects again, consider sex, height, and "other mathematics" scores provided in Table 1 above. These data are plotted in Figure 4 below. Examine Figure 4 and determine whether a relationship exists between mathematics scores and height.

Figure 4

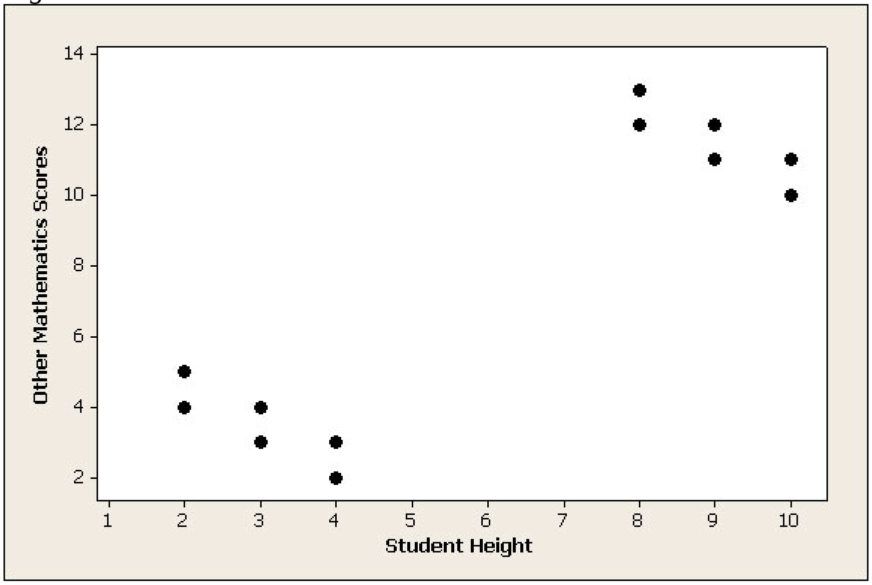


Figure 5 shows how regression will model these data if sex is not controlled in the regression model. Note again the positive relation between height and mathematics scores.

Figure 5

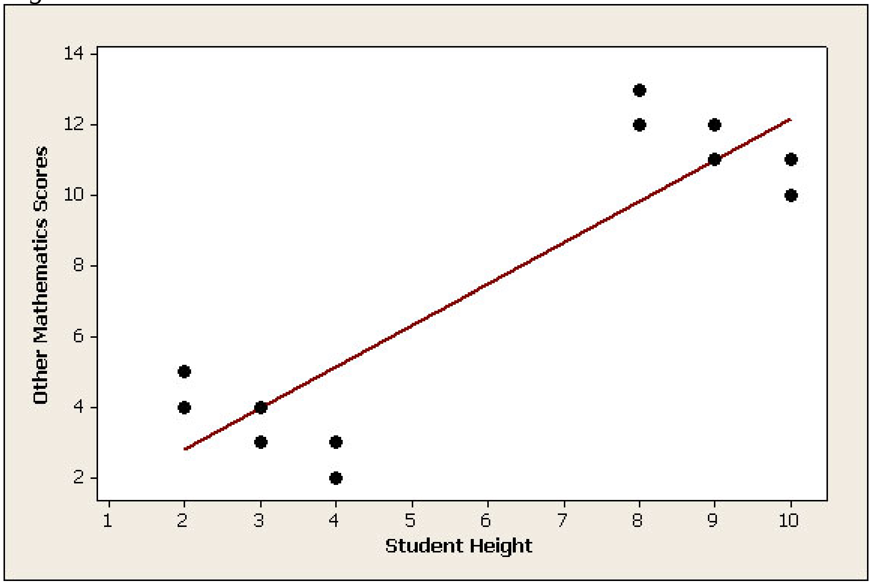
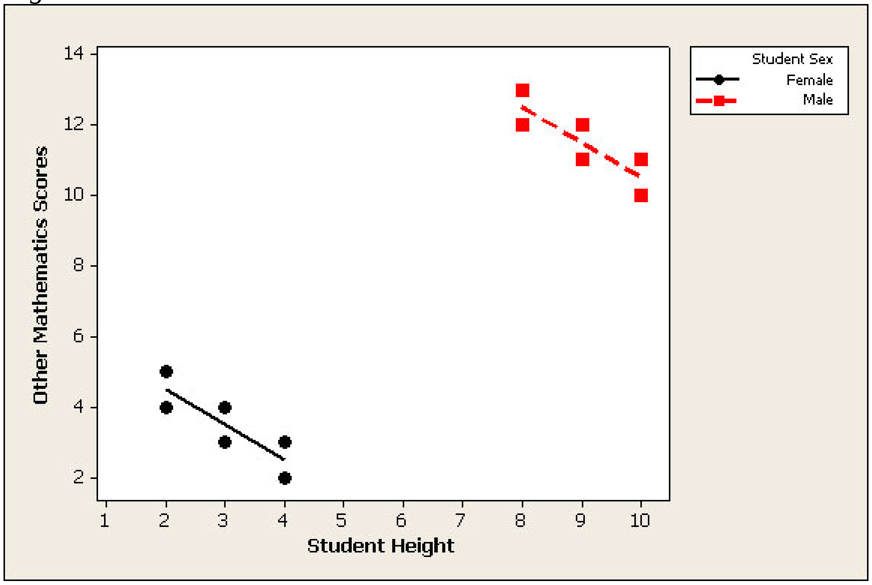


Figure 6 below shows how these data should be modeled if modeled correctly in regression.

Figure 6



Regression Analysis

Model these data using regression first without sex as a second variable, then model again with sex. How does adding sex into the regression equation change the slope estimates?

Model 1 (simple regression with only height as predictor):

$Y_i = b_0 + b_1 \text{Height}_{i1} + e_i$

What is the literal interpretation for these coefficients?
What is the R^2 value for this model?

Model 2 (regression with both height and sex as predictors):

$Y_i = b_0 + b_1 \text{Height}_{i1} + b_2 \text{Sex}_{i2} + e_i$

What is the literal interpretation for these coefficients?
What is the R^2 value for this model? How did the R^2 value change from Model 1?

3. Estimation of the Regression Equation and Residuals

Estimation for multiple regression relies on ordinary least squares, same as with simple regression. In addition, residuals are calculated in the same way—obtain predicted value of Y' and e = Y–Y'.

Note: Illustrate residuals with mathematics scores and height data.

4. Literal Interpretation of Coefficients

Example 1: Mathematics scores, height, and sex data

Using the first set of mathematics scores from Table 1, the following is obtained through regression:

$Y' = b_0 + b_1 Height_{i1} + b_2 Sex_{i2}$

$$\begin{aligned} Y' &= b_0 + b_1 Height_{i1} + b_2 Sex_{i2} \\ &= 3.00 + 0.00(Height) + 6.00(Sex) \end{aligned}$$

What is literal interpretation of b_0, b_1, and b_2?

b_0 = the predicted mean mathematics score for students with height = 0.00 and sex = 0.00 is 3.00 (i.e., females with height of 0.00 are predicted to have mathematics score of 3.00)

b_1 = mathematics scores are expected to increase by 0.00 for a 1 point increase in height controlling for sex

b_2 = mathematics scores are expected to increase by 6.00 for a 1 point increase in sex controlling for height

Recall that to "control for" means to hold at a constant value the other predictors in the model.

Example 2: Ice cream sales by price, income, and temperature

The data listed below were reported by Kadiyala, Koteswara Rao (1970). "Testing for the Independence of Regression Disturbances." Econometrics, 38(1), 97—117.

The data consists of ice cream sales over a 30 week period taken over several years from March 1950 to July 1953. The variables include the following:

Sales (consumption)	=	Measured in pints per capita.
Price	=	Price of ice cream in dollars.
Income	=	Weekly family income in dollars.
Temperature	=	Mean temperature in degrees Fahrenheit.

The data are located here in an Excel, Mintab, and SPSS file:

Excel: http://www.bwgriffin.com/gsu/courses/edur8131/data/ice-cream.xls
SPSS: http://www.bwgriffin.com/gsu/courses/edur8131/data/ice-cream.sav
Minitab: http://www.bwgriffin.com/gsu/courses/edur8131/data/ice-cream2.MTW

Using these ice cream sales data, estimate the following model:

$\text{Sales'} = b_0 + b_1\text{Price}_{i1} + b_2\text{Income}_{i2} + b_3\text{Temperature}_{i3}$

Address the following:

(a) Provide literal interpretation for each of the four regression coefficients.
(b) If temperature drops by 10 degrees, what is the expected change in sales?
(c) If ice cream prices increase by 25 cents, what is the expected change in sales?
(d) If family income drops $10 per week, what is the expected change in sales?
(e) What is the predicted ice cream sales if temperature = 70, family income = 70, and price = 28 cents?

Example 3: House prices in Albuquerque 1993

The data are a random sample of records of re-sales of homes from Feb 15 to Apr 30, 1993 from the files maintained by the Albuquerque Board of Realtors.

Price = Prices in thousands of dollars.
Square Feet = Size of house in square feet living space.
Age = Age of house in years.
Features = Number out of 11 features (dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access)
Tax = Annual taxes in dollars

The data are located here in an Excel, Mintab, and SPSS file:

Excel: http://www.bwgriffin.com/gsu/courses/edur8131/data/house-prices.xls
SPSS: http://www.bwgriffin.com/gsu/courses/edur8131/data/house-prices.sav
Minitab: http://www.bwgriffin.com/gsu/courses/edur8131/data/house-prices.MTW

Using these housing data, estimate the following model:

$\text{Price'} = b_0 + b_1\text{Square Feet}_{i1} + b_2\text{Age}_{i2} + b_3\text{Features}_{i3}$

Address the following:

(a) Provide literal interpretation for each of the four regression coefficients.
(b) If house size drops by 100, what is the expected change in price?
(c) If age of house increases by 25 years, what is the expected change in price?
(d) If number of features increases by 2, what is the expected change in price?
(e) What is the predicted sales price for a house if square feet = 3000, age = 25, and features = 6?

5. Model Fit

As with simple regression, model fit (how well model reproduces observed Y) is assess via the following indices:

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

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

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

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

6. ΔR^2 = Part Correlation—Increment in R2 due to an IV

To be added.

7. Inference in Regression

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

Overall Model Fit

Null

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

or equivalently

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

Alternative

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

Meaning

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

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

Test of H_0: $R^2 = 0$

Same as with simple regression, use overall F test (and corresponding p-value) to assess significance of complete model in predicting Y.

Coefficient Inference

Null and Alternative

H_0: $\beta_p = 0.00$

The null states that relation between X_1 and Y is zero controlling for other IV; no relation between X_1 and Y controlling for other IVs.

H_1: $\beta_p \neq 0.00$.

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

As with simple regression, H_0: $\beta_p = 0.00$ may be tested with a t-ratio:

$t = b_1 / SE_{b1}$

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

The decision rule for t-test:

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

The decision rule for p-values:

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

All of the above is the same for simple regression.

Confidence Interval for b_1: Inference and Estimation

Formula for CI about regression coefficient:

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

where

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

Interpretation: one may be 95% confidence that the true population coefficient may be as large as [upper bound] or as small as [lower bound].

8. Reporting Regression Results

Like with simple regression, two tables are often used, one for descriptive information and one for regression results. Using the ice cream data, below are sample tables and written results.

Table 2: Descriptive Statistics and Correlations for Ice Cream Sales Data

	Correlations			
Variable	1	2	3	4
1. Sales	---			
2. Price	-.26	---		
3. Income	.05	-.11	---	
4. Temperature	.78*	-.11	-.33	---
Mean	0.36	0.28	84.60	49.10
SD	0.07	0.01	6.25	16.42

Note: n = 30
* p < .05

Table 3: Regression of Ice Cream Sales on Price, Income, and Temperature

Variable	b	se b	95% CI	t
Price	-1.04	0.83	-2.76. 0.67	-1.25
Income	0.003	0.001	0.001, 0.006	2.82*
Temperature	0.003	0.001	0.003, 0.004	7.76*
Intercept	0.20	0.27	-0.36, 0.75	0.73

Note: R^2 = .72, adj. $\underline{R}^2$ = .69, F = 22.18*, df = 3,26; n = 30
*p < .05.

> Results of the regression analysis show that both temperature and weekly family income are positively and statistically associated with ice cream sales. Ice cream price is not a statistically significant predictor of sales in this analysis. The greater the family income, and the greater the temperature, the higher will be predicted sales of ice cream.

9. Exercises

(1) A researcher wishes to know whether number of hours studied at home is related to general achievement amongst high school students. Student ability should be controlled to assess better the effects of studying.

High School GPA	IQ	Time Spent Studying Per Week (in Hours)
3.33	117	3
1.79	90	5
2.21	101	12
3.54	121	9
2.89	105	11
2.54	110	1
2.66	112	0
1.10	85	3
3.67	128	2

(2) Does SAT adequately predict college success, once rank is controlled?

Freshmen Collegiate GPA	HS Rank*	SAT Scores
3.33	52	1000
1.79	233	750
2.21	150	890
3.54	43	1100
2.89	95	900
2.54	43	860
2.66	120	1010
1.10	280	640
3.67	33	1240

*Out of 300 students.

(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	Average IQ in Class	Overall Class Ach. on Final Exam
Fall 1991	1	105	85.5
Winter 1992	2	108	86.5
Spring 1992	3	108	88.9
Summer 1992	4	109	89.1
Fall 1992	5	107	87.2
Winter 1993	6	110	90.5
Spring 1993	7	108	89.8
Summer 1993	8	114	92.5
Fall 1994	9	110	89.3
Winter 1994	10	112	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. The principal also has standardized ITBS test scores for each student.

Student	GPA	ITBS	Tardies/Absences
Bill	3.33	65	2
Bob	1.79	40	10
Stewart	2.21	50	5
Linda	3.54	70	6
Lisa	2.89	49	3
Ann	2.54	55	4
Fred	2.66	58	6
Carter	1.10	37	12
Bill	3.10	55	3
Sue	2.10	45	8
Loser	2.31	51	6
Kathy	3.67	63	2

Exercises Answers

(1) A researcher wishes to know whether number of hours studied at home is related to general achievement amongst high school students. Student ability should be controlled to assess better the effects of studying.

Table 1

Descriptive Statistics and Correlations among GPA, Time Spent Studying, and IQ

Variable		Correlations	
	GPA	Time	IQ
GPA	---		
Time	.033	---	
IQ	.963*	-.149	---
Mean	2.637	5.11	107.67
SD	.844	4.46	14.053

n = 9

* p < .05

Table 2

Regression of GPA on Time Spent Studying and IQ

Variable	b	se	95%CI	t
Time	.034	.015	-.005, .07	2.16
IQ	.059	.005	.047, .072	11.83*
Intercept	-3.93	.56	-5.32, -2.56	-7.00*

Note. R^2 = .96, adj. R^2 = .95, $F_{2,6}$ = 70.09, n = 9.

*p < .05.

Regression results show that time spent studying is not statistically related to GPA once students' IQ scores are taken into account. The relationship between IQ and GPA is statistically significant and positive. The greater one's IQ, the higher, on average, is one's GPA. Time spent studying does not appear to predict one's GPA.

(2) Does SAT adequately predict college success, once rank is controlled?

Table 1
Descriptive Statistics and Correlations among GPA, HS Rank, and SAT

Variable	Correlations		
	GPA	Rank	SAT
GPA	---		
Rank	-.93*	---	
SAT	.94*	-.83*	---
Mean	2.64	116.56	932.22
SD	0.84	89.33	180.33

n = 9
*p < .05.

Table 2
Regression of GPA on Rank and SAT

Variable	b	se	95%CI	t
Rank	-.005	.002	-.008, -.001	-3.03*
SAT	.002	.001	.001, .004	3.28*
Intercept	.864	.862	-1.24, 2.97	1.00

Note. $R^2 = .95$, adj. $R^2 = .93$, $F_{2,6} = 57.43$*, n = 9.
*p < .05.

Both HS rank and SAT scores are statistically related to student GPA. Results show that as one’s rank increases, predicted GPA declines; however, as SAT scores increase, GPA tends to increase.

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

Table 1
Descriptive Statistics and Correlations between Testing Frequency, Achievement, and IQ

Variable	Correlations		
	Final	Testing	IQ
Final	---		
Testing	.75*	---	
IQ	.89*	.77*	---
Mean	88.94	5.50	109.10
SD	2.06	3.03	2.56

n = 10
* p < .05

Table 2
Regression of Achievement on Testing Frequency and IQ

Variable	b	se	95%CI	t
Testing	.12	.18	-0.31, 0.54	0.66
IQ	.61	.21	0.10, 1.11	2.86*
Intercept	22.06	22.45	-31.02, 75.14	0.98

Note. R^2 = .80, adj. R^2 = .74, $F_{2,7}$ = 13.94*, n = 10.
*p < .05.

Correlations show that testing frequency and IQ are both related to student achievement, but regression results show that once IQ is taken into account, testing frequency no longer predicts student achievement. Students with higher IQs tend to obtain higher achievement scores; testing frequency does not seem to relate to achievement.

(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. The principal also has standardized ITBS test scores for each student.

Table 1

Descriptive Statistics and Correlations among GPA, ITBS, and Absences

Variable	Correlations		
	GPA	ITBS	Absences
GPA	---		
ITBS	.92*	---	
Absences	-.85*	-.72*	---
Mean	2.60	53.17	5.58
SD	.76	9.93	3.15

n = 12

*p < .05.

Table 2

Regression of GPA on Absences and ITBS scores

Variable	b	se	95%CI	t
ITBS	.048	.01	.03, .07	4.62*
Absences	-.096	.03	-.17, -.02	-2.92*
Intercept	.58	.70	-1.00, 2.17	0.83

Note. $R^2 = .92$, adj. $R^2 = .90$, $F_{2,9} = 50.47$*, n = 12.

*p < .05.

Regression results show that both ITBS scores and number of absences are statistically related to student GPA. As the number of absences increase, GPA tends to decline; as ITBS scores increase, GPA tends to increase.