Chapter 9 – Correlation and Regression

Using your College Student data do the following problems. Print your outputs after typing your interpretations on them. Please circle the key parts of the output that you discuss.

9.1. What is the correlation between students' height and parent's height? Also produce a scatterplot. Interpret the results, including statistical significance, direction, and effect size.

Selection of the Statistic

The variables of interest for Problem 9.1 are height and parent's height. Both of these variables are scale. (We found in Chapter 5 that they were normally distributed.) The best choice for a statistic when you are interested in investigating the relationship between two scale variables is Pearson correlation. Pearson correlation produces an *r*. The closer the *r* value is to +1.0, the stronger the positive the relationship that exists between the two variables. If the *r* value is negative, this indicates a negative relationship. For example, hours spent watching TV is negatively related to hours spent studying (see Problem 9.2). This indicates that students who watch a lot of TV tend to study less.

Assumptions of Correlation **(Pearson *r*)**

1. **The two variables are linearly related (Pearson *r* will not detect a curvilinear relationship)**
2. **Scores on one variable are normally distributed for each value of the other variable and vice versa. (If degrees of freedom are greater than 25, failure to meet this assumption has little consequence.)**

Assumptions of Spearman Rho **(correlation based on ranked data)**

- Data on both variables are at least ordinal

How to Produce the Selected SPSS Output

To answer Problem 9.1 with Windows:

- Click on Analyze ⇒ Correlate ⇒ Bivariate. This will open the Bivariate Correlations window
- Highlight student height in inches and same sex parent's height
- Click on the arrow to move the variables into the Variable(s) box
- Click on Options – This will open the Bivariate Correlations: Options window
- Click on Means and Standard deviations and Exclude cases listwise
- Click on Continue and O.K.

To answer Problem 9.1 with syntax:

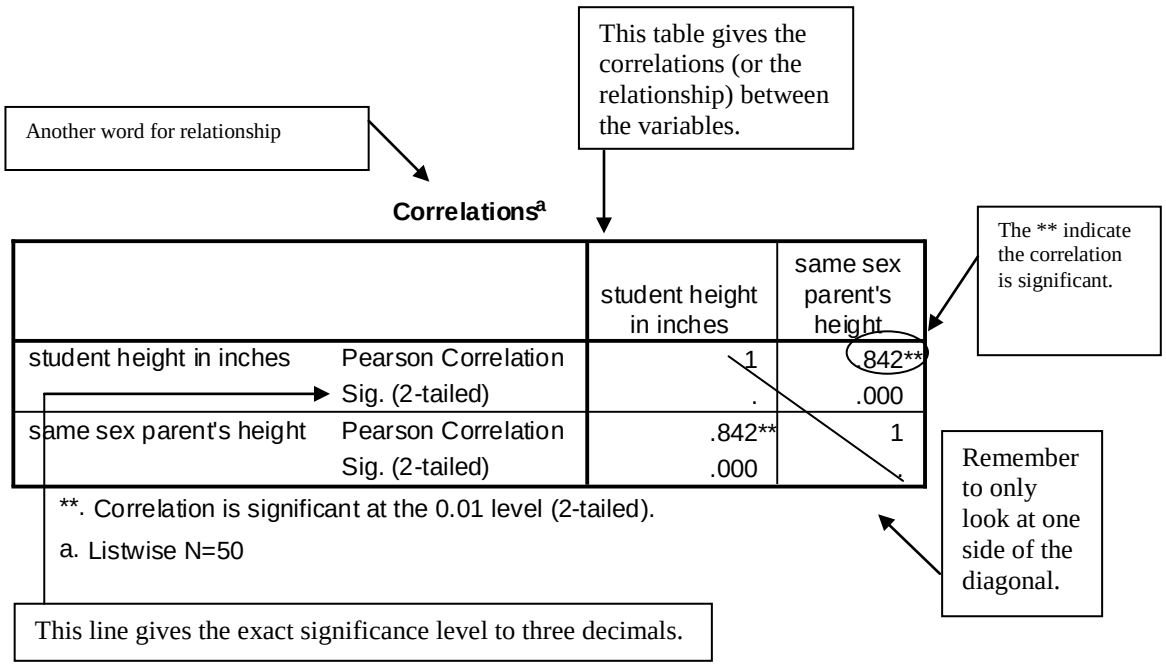
```
CORRELATIONS
  /VARIABLES=height pheight
  /PRINT=TWOTAIL NOSIG
  /STATISTICS DESCRIPTIVES
  /MISSING=LISTWISE .
```

SPSS Output for Problem 9.1

Correlations

Descriptive Statistics

	Mean	Std. Deviation	N
student height in inches	67.3000	3.9396	50
same sex parent's height	66.7800	5.1042	50



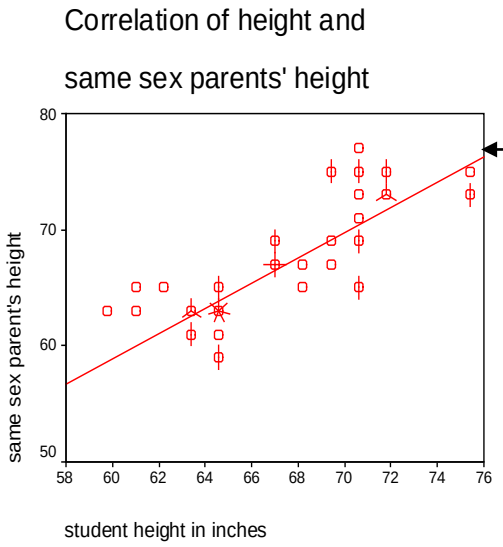
To create the graph for Problem 9.1 with Windows:

- Click on Graphs ⇒ Scatter… This will open the Scatterplot window
- Click on Simple and then Define
- Highlight same sex parent's height
- Click on the arrow to move it into the Y Axis box
- Highlight student height in inches
- Click on the arrow to move it into the X Axis box
- Click on Titles. This will open the Title box
- Type "Correlation of height and " in Line 1 under Title
- Type "same sex parents' height" in Line 2 under Title
- Click on Continue and O.K.
- To get the regression line – Double click on the Scatterplot. This will make the SPSS Chart Editor appear
- Click on Chart ⇒ Options
- Click on Total under Fit
- Click on Continue and O.K.
- Close the SPSS Chart Editor box

To answer the graph for Problem 9.1 with syntax:

```
GRAPH
  /SCATTERPLOT(BIVAR)=height WITH pheight
  /MISSING=LISTWISE .
```

Graph



It is helpful to look at the regression line and which way it goes. The line in this graph shows a positive correlation because its slope is from lower left to the upper right of the plot.

Rsq = 0.7084

This is the *r*-squared value. It tells us that 71 percent of the variance in parent's height can be predicted from student height.

Description of Output 9.1

The first table shows the mean, standard deviation and *N* for each variable. The second table presents the correlations. Remember to only look on one side of the diagonal, as the information is redundant. The top number in the correlation matrix (in this case it is .842) is the value of *r*. The second number is the *p* value (in this case .000), and the number at the bottom is the *N*, which was included in the analysis. The graph shows the relationship between the two variables. *R* squared indicates the amount of variance in same sex parent's height that can be predicted from student height in inches.

Example of APA Results and Discussion for Problem 9.1

Results

There was a statistically significant correlation between student height in inches and same sex parent's height, r (48) = .842, $p < .001$. The direction of the correlation was positive, which means that tall parents tend to have tall children, and vice versa. Using Cohen's (1988) guidelines, the r indicates a very large effect size. The r squared indicates that 71% of the variance in student height in inches can be predicted from the same sex parent's height.

9.2. Write a question that can be answered via correlational analysis with two approximately normal or *scale* variables. Run the appropriate statistics to answer the question. Interpret the results.

Selection of the Procedure

There are several variables you can choose to answer this problem. Any two variables that are scale would be appropriate. If you are unsure what the measurement is for a variable, you can check the codebook in Appendix B, or look on the Variable View Screen (the last column will list the measurement type). If one or both variables in a correlation is ordered but not scale (ranks or severely non normal), you can use the Spearman correlation (rho).

How to Produce the Selected SPSS Output

(See Problem 9.1)

SPSS Output for Problem 9.2

Correlations

Descriptive Statistics

	Mean	Std. Deviation	N
hours of study per week	15.62	8.31	50
amount of tv watched per week	11.98	6.10	50

Correlations

		hours of study per week	amount of tv watched per week
hours of study per week	Pearson Correlation	1.000	-.316*
	Sig. (2-tailed)	.	.025
	N	50	50
amount of tv watched per week	Pearson Correlation	-.316*	1.000
	Sig. (2-tailed)	.025	.
	N	50	50

*. Correlation is significant at the 0.05 level (2-tailed).

Notice this significance level. Two asterisks (**) means $p<.01$ and one (*) means $p<.05$.

To help yourself to remember to look at each relationship only once, it can be helpful to draw a line similar to this on your output. Note that the correlation (-.32) is the same above and below the line.

Description of Output 9.2

When viewing an output of correlations, first check to see if the *N*'s are the same for each variable; only persons with data for both variables will be included in the correlation. Note that correlations assess the strength of the association between two variables. Then check the correlations table. Remember to only look at one side of the diagonal since the information is repetitive. The 1.000's are correlations of the variable with itself.

Example of How to Write About Problem 9.2

Results

There was a statistically significant negative correlation between hours of study per week and amount of TV watched per week, r (48) = -.32, p = .025. The direction of the correlation was negative, indicating that students who study a lot tend not to watch a lot of TV. Using Cohen's guidelines, the r = -.32 indicates a medium effect size.

9.3. Make a correlation matrix using at least four appropriate variables. Identify, using the variable names, the two strongest and two weakest correlations. What were the *r* and *p* values for each correlation?

Selection of the Statistic

For Pearson correlations to be appropriate both variables should be scale level (ordered and normally distributed). Thus, you should select four scale variables to use when answering this problem.

How to Produce the Selected SPSS Output

(See Problem 9.1)

SPSS Output for Problem 9.3

Correlations

Descriptive Statistics

	Mean	Std. Deviation	N
amount of tv watched per week	11.98	6.10	50
hours of study per week	15.62	8.31	50
student's current gpa	3.172	.391	50
hours per week spent working	26.12	14.86	49

Exact significance level, *p*

Correlations

		amount of tv watched per week	hours of study per week	student's current gpa	hours per week spent working
amount of tv watched per week	Pearson Correlation	1.000	-.316*	-.253	-.541**
	Sig. (2-tailed)	.	.025	.076	.000
	N	50	50	50	49
hours of study per week	Pearson Correlation	-.316*	1.000	.109	.219
	Sig. (2-tailed)	.025	.	.453	.131
	N	50	50	50	49
student's current gpa	Pearson Correlation	-.253	.109	1.000	.303*
	Sig. (2-tailed)	.076	.453	.	.034
	N	50	50	50	49
hours per week spent working	Pearson Correlation	-.541**	.219	.303*	1.000
	Sig. (2-tailed)	.000	.131	.034	.
	N	49	49	49	49

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Notice that two levels of significance ($p<.01$ and $p<.05$) are marked with asterisks.

Description of Output 9.3

The Descriptive Statistics table indicates that hours per week spent working has only 49 observations. The other three variables have 50.

In the second table, called a correlation matrix, each of the four variables is correlated with the other three. Notice, in the first column, that the amount of TV watched per week is significantly *negatively* correlated with hours of study per week (r = -.316, p = .025) and with hours per week spent working (r = -.541, p = .001), which is the strongest correlation in the matrix. However, amount of TV is not significantly correlated with students GPA (r = -.253, p = .076) because the p or Sig. is greater than .05.

The boxes below the diagonal in the hours of study per week column indicate that it is *not* significantly correlated with student's GPA (r = .109, p = .453), the weakest in the matrix, or with hours per week spent working (r = .219, p = .131). In addition, student's current GPA is significantly positively correlated with hours per week spent working (r =

.303, p = .034). The number of subjects with data for each pair of variables (49 or 50, in this matrix) is indicated by the *N*s. The degrees of freedom (*df*) are *N*-2.

Example of How to Write About Problem 9.3

Results

Table 9.3 shows that three of the six pairs of variables were significantly correlated. The strongest correlation, with a large effect size, was between amount of TV watched per week and hours per week spent working, r (47) = -.541, p < .001. This means that students who had a relatively high number of hours watching TV were likely to spend relatively few hours working. Likewise, TV watching was negatively correlated with studying, but amount of work and GPA were positively correlated. As shown in Table 9.3 both of these correlations were of a medium size according to Cohen (1988).

Table 9.3

Intercorrelations, Means and Standard Deviations for Scale Variables

	Variable	1	2	3	4	*M*	*SD*
1.	TV watched	--	-.32*	-.54*	-.25	11.98	6.10
2.	Study	--	--	.22	.11	15.62	8.31
3.	Work	--	--	--	.30*	26.12	14.86
4.	GPA	--	--	--	--	3.17	.39

* p < .05

9.4. Is there a combination of gender and same sex parent's height that predicts student's height better than either one of these variables alone?

Selection of the Statistic

Problem 9.4 is asking a prediction question; "What variables significantly predict student's height?" When asking this type of question, multiple regression is an appropriate choice for a statistic. Multiple regression is also used to answer other questions when there are multiple independent variables that are scale or dichotomous and one scale dependent variable.

Key elements needed:

- Two or more predictor variables (another name for independent variables) that are scale (in this case, same sex parent's height) or dichotomous (in this example, gender).
- A response variable (another name for dependent or outcome variable) which is scale (in this case student's height)
- A linear (approximately straight line) relationship needs to exist between the predictors and response/outcome

Assumptions of Regression

1. **The relationship between the independent variables and the dependent variable is linear (in other words, the regression of the DV on the combination of IVs is linear).**
2. There are several other more complex assumptions not discussed here.

How to Produce the Selected SPSS Output

To answer Problem 9.4 with Windows:

- Click on Analyze ⇒ Regression ⇒ Linear. This will open the Linear Regression window.
- Highlight student height in inches and move it into the Dependent box
- Highlight same sex parent's height and gender and move them into the Independent(s) box
- Click on Statistics. This will open the Linear Regression: Statistics window
- Select Descriptives
- Click on Continue and O.K.

How to answer Problem 9.4 with syntax:

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

SPSS Output for Problem 9.4

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
student height in inches	67.3000	3.9396	50
same sex parent's height	66.7800	5.1042	50
gender of student	1.48	.50	50

See Chapter

Correlations

		student height in inches	same sex parent's height	gender of student
Pearson Correlation	student height in inches	1.000	.842	-.782
	same sex parent's height	.842	1.000	-.782
	gender of student	-.782	-.782	1.000
Sig. (1-tailed)	student height in inches	.	.000	.000
	same sex parent's height	.000	.	.000
	gender of student	.000	.000	.
N	student height in inches	50	50	50
	same sex parent's height	50	50	50
	gender of student	50	50	50

These are the correlations of the dependent variable with each of the predictors.

High correlations among predictors might affect the results.

These are the predictor variables.

Indicates what method was used to enter the predictor variables. There are many methods that can be used. We will only use the "Enter" or simultaneous method.

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	gender of student, same sex parent's height[a]	.	Enter

All predictors simultaneously entered in the analysis.

a. All requested variables entered.

b. Dependent Variable: student height in inches

The dependent variable

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.865[a]	.748	.737	2.0196

a. Predictors: (Constant), gender of student, same sex parent's height

Indicates the *F* statistic's significance level. If significant (<.05), then the predictors are significantly predicting the dependent variable.

The ANOVA or *F* statistic

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	568.794	2	284.397	69.725	.000[a]
	Residual	191.706	47	4.079		
	Total	760.500	49			

a. Predictors: (Constant), gender of student, same sex parent's height

b. Dependent Variable: student height in inches

Significance of each predictor and the constant.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	40.466	7.176		5.639	.000
	same sex parent's height	.457	.091	.592	5.038	.000
	gender of student	-2.491	.917	-.319	-2.715	.009

a. Dependent Variable: student height in inches

Description of Output 9.4

The table of Descriptives shows the number (*N)* for each variable. In multiple regression, SPSS uses only the participants with no missing data on the selected variables. In this example there is no missing data, which is indicated by the *N*'s being equal to 50.

After the table of Descriptive statistics, there is table of correlations which shows how each variable is related to all the others. The first column (or top row) shows the correlations of each predictor with the dependent/outcome variable. It is desirable to have high correlations between each predictor and the outcome variable. In this example student height in inches is highly correlated with same sex parent's height (r = .842) and with gender of student (r = -.782). This is good. It is not good to have high correlations between predictor variables because this might indicate the presence of multicollinearity. If there are high correlations among the predictor variables, then the VIF or Tolerance scores should be checked. In this example, there is a high correlation between the two predictors, same sex parent's height and gender of student, (r = -.782). Checking the VIF score (not shown here) indicated that multicollinearity was not a problem (see Mertler and Vannatta, 2001, for more on how to do this).

The third table shows which variables are predictors and dependent and the method used to run the multiple regressions. In this case, the method used was "Enter" which enters all the variables at once into the equation. Next the Model Summary table provides

information on how much variance in the dependent variable can be predicted from the combination of the predictors (see adjusted R^2). In this example, adjusted $R^2 = .737$ indicates that 74% of the variance in student height in inches can be predicted from the combination of gender of student and same sex parent's height. The ANOVA table provides an overall indication of whether this combination is statistically significant, in this example $p < .001$ which tells us that the model significantly predicts the dependent variable.

Finally, the Coefficients table is a key table. The unstandardized beta (*B*) and standard error would be used if you want to write out the model or equation to predict an individual student's height given his or her gender, same sex parent's height, and the constant. In this example the model would be written as student height = 40.47 + .46*p height –2.47*gender. Remember to multiply variables before adding or subtracting. The value for a person's gender (1 = male or 2 for female) should be multiplied by 2.47 and then subtracted from 40.47 + .46 times that person's parent's height in inches. The standardized betas (β) are similar to correlation coefficients. The *t* and Sig. values indicate whether the betas are significant predictors, assuming that the other variables are in the model. In this case both betas are statistically significant which indicates that both variables significantly contribute to the model. If a beta (β) is not significant, you might want to consider not including it in the model, although doing so might cause other currently significant predictors to become non significant. An indication of the effect size can be obtained from the multiple correlation coeffiecients: they can vary from +1.0 to –1.0. *R* equals 87, which according to Cohen's (1988) guidelines is a large effect. The effect sizes for multiple regression are similar to those for *r* and are shown on page 21.

Example of How to Write About Problem 9.4

Results

Multiple regression was conducted to investigate the best predictors of student height. The assumptions of multiple regression were checked and none were violated. When the combination of variables to predict student height included same sex parent's height and gender, $F(2, 49) = 69.73$, $p < .001$, student height = 40.47 + .46* *p* height – 2.49* gender. For example, this indicates that if *p* height equals 70 (5'10") the son's height would be predicted to be 70.18 (40.47 + .46 *70 – 2.49*1). The adjusted *R* squared value was .737. This indicates that 74% of the variance in student height was explained by the model. According to Cohen (1988) this is a large effect.

Table 9.4a

Means, Standard Deviations, and Intercorrelations for Student Height in Inches and Predictors Variables (N=50)

Variable	*M*	*SD*	Parent's height	Gender
Student Height	67.3	3.94	.842*	-.782*
Predictor variable				
1. Same sex parent's height	66.8	5.10	--	-.782*
2. Gender	1.48	.50		--

*$p < .001$.

Table 9.4b

Simultaneous Multiple Regression Analysis Summary for Gender and Same Sex Parent's Height Predicting Student's Height in Inches (N = 50)

Variable	*B*	*SEB*	β
Same sex parent's height	.46	.09	.59*
Gender	-2.49	.92	-.32*
Constant	40.47	7.18	

Note. $R^2 = .75$; $F(2,47) = 69.73$, $p < .001$.
*$p < .01$.

Discussion

From the data in this sample, it appears that one can predict student height from the same sex parent's height and gender very well. This study indicates that once a pregnant woman knows the sex of her child, the height of the child can be predicted from the child's gender and same sex parent's height. Jones (1998) has demonstrated this in previous studies.

9.5. Is there a combination of hours of TV watching, hours of studying, and hours of work that predicts current GPA?