**Student Data**

For a sample of high school students, data on area of residence, hours spent watching TV per week, hours spent studying per week, and high school grade point average are recorded. This data is stored in the SPSS data file student, where the variable area of residence is named residence, the variable weekly TV hours is named tvhrs, the variable weekly study hours is named stdhrs, and the variable grade point average is named gpa.

1. A simple linear regression is of interest to predict high school grade point average from weekly study hours is of interest. The students used to collect the data will be treated as a random sample of high school students.

   (a) Complete the following sentences:
   The dependent (response) variable is \( Y = \) “high school gpa”, and the independent (explanatory) variable is \( X = \) “weekly study hours”.
   This data is observational, since all of the variables must be random.

   (b) In order to use SPSS to do the calculations needed for a simple linear regression, first, go to the section titled Hypothesis Tests Involving Two Variables, and use the first five steps in the subsection titled Performing a Simple Linear Regression with Checks of Linearity, Homoscedasticity, and Normality Assumptions as a guide to use SPSS to create a graph of the least squares line on a scatterplot. Once you have successfully generated SPSS output, add a title to the top of the output in the following format:

   **YOUR NAME – Student Data Exercise 1(b)**

   Next, use the remaining steps to do calculations and create other graphs needed for a simple linear regression. Verify that your SPSS output contains all of the following items:

   (1) a graph of the least squares line on a scatterplot;

   (2) tables titled Descriptive Statistics, Correlations, Model Summary, ANOVA, and Coefficients;

   (3) a normal curve super-imposed onto a histogram, a normal probability plot, and a residual plot.

   Submit this copy of the SPSS output with this assignment.
(c) Complete the following sentences concerning the assumptions required to proceed with a simple linear regression analysis:

Since the points on the scatterplot look like they vary around the graph of a **straight line** curve instead of a line with a **positive** slope, and the points on the residual plot look **randomly**/**non-randomly** distributed above and below the horizontal line at zero, then it appears that the **linearity** assumption is satisfied.

Since the points on both the scatterplot and the residual plot seem to show roughly the same variation everywhere, it appears that the uniform variance assumption is satisfied.

Since the points on the normal probability plot appear reasonably close to the diagonal line, it appears that the normality assumption is satisfied.

(d) Complete the following sentences concerning the simple linear regression analysis:

Using the $f$ test in the analysis of variance for the simple linear regression, we find that the linear regression to predict **high school gpa** from **weekly study hours** is statistically significant at the $0.05$ level ($f_{1,25} = 32.780$, $f_{1,25; 0.05} = 4.24$, $p < 0.001$), that is, that the linear relationship between **high school gpa** and **weekly study hours** is statistically significant. **The data suggest that the linear relationship is positive.**

About **56.7%** of the variation in **high school gpa** is explained by **weekly study hours**, and the correlation is $r = +0.753$. 
(e) Complete the following sentences concerning the least squares line in the simple linear regression analysis:

\[ \hat{y} = 1.965 + 0.038x \quad \text{OR} \quad gpa = 1.965 + 0.038(\text{std}) \]

The least squares regression equation is

With each increase of one hour in weekly study time, there is an average increase of 0.038 in high school gpa.

A 95% confidence interval for the slope is of interest, since the linear regression is statistically significant.

We can 95% confident that the slope in the regression to predict high school gpa from weekly study time is between 0.024 and 0.052.

(f) Use the SPSS output to find each of the following: \( n = 27 \)

\[ \bar{x} = 24.00 \quad \bar{y} = 2.8807 \quad s_x = 10.583 \quad s_y = 0.53592 \]

standard error of estimate = \( s = 0.35950 \quad \sum (x - \bar{x})^2 = (26)(10.583)^2 = 2887.286 \)

(g) First, find the \( t \) test statistic in a hypothesis test to see if there is any evidence that the slope in the simple linear regression is different from zero; then, complete the following sentences:

\[ t_{25} = 5.725 \]

The \( f \) test statistic \( f_{1,n-2} \) in the analysis of variance for the simple linear regression and the \( t \) test statistic with a two-sided test comparing the slope to 0 are really testing the same hypothesis, since \( f_{1,n-2} = f^2_{n-2} \). For this data we have \( 32.780 = (5.725)^2 \).
(h) The 0.05 significance level is to be used for a hypothesis test to see if there is any evidence that the slope in the simple linear regression is different from 0.05. Complete the following statement of results for this hypothesis test:

Using the appropriate t test, we find that the difference between the estimated slope (0.038) and the hypothesized slope (0.05) is not statistically significant at the 0.05 level (\( t_{25} = -1.780, t_{25;0.025} = 2.060, 0.05 < p < 0.10, \) two-sided).

We conclude that the slope in the regression to predict high school gpa from weekly study hours is not different from 0.05.

A 95% confidence interval for the slope is not of interest, since the null hypothesis is not rejected.

(i) The 0.05 significance level is to be used for a hypothesis test to see if there is any evidence that the mean high school gpa with 10 hours of study per week is larger than 2.00. Complete the following statement of results for this hypothesis test:

Using the appropriate t test, we find that the difference between the mean grip strength for 20 year old right-handed males estimated from the least squares line (1.965 + 0.038(10) = 2.347) and the hypothesized mean (2.00) is statistically significant at the 0.05 level (\( t_{25} = 2.986, t_{25;0.05} = 1.708, 0.001 < p < 0.01, \) one-sided).

We conclude that the mean high school gpa with 10 hours of study per week is larger than 2.00.

A 95% confidence interval for the mean high school gpa with 10 hours of study per week is of interest, since the null hypothesis is rejected.

We can 95% confident that the mean high school gpa with 10 hours of study per week is between 2.108 and 2.586.
(j) Find and interpret a 95% prediction interval for the gpa of a high school student with 10 hours of study per week.

We are 95% confident that the gpa for a randomly selected high school student with 10 hours of study per week will be between 1.569 and 3.125.

OR

At least 95% of high school students with 10 hours of study per week have a gpa between 1.569 and 3.125.

(k) For what weekly study time will the confidence interval for mean high school gpa and the prediction interval for a particular high school student’s gpa both have the smallest length?

24 study hours per week
2. A simple linear regression is of interest to predict high school grade point average from weekly TV hours is of interest. The students used to collect the data will be treated as a random sample of high school students.
(a) Complete the following sentences:
   The dependent (response) variable is \( Y = \) “high school gpa” \( \), and
   the independent (explanatory) variable is \( X = \) “weekly TV hours” \( \).
   This data is \underline{observationally experimental}, since \underline{all of the variables must be random}.

(b) In order to use SPSS to do the calculations needed for a simple linear regression, first, go to the section titled Hypothesis Tests Involving Two Variables, and use the first five steps in the subsection titled Performing a Simple Linear Regression with Checks of Linearity, Homoscedasticity, and Normality Assumptions as a guide to use SPSS to create a graph of the least squares line on a scatterplot. Once you have successfully generated SPSS output, add a title to the top of the output in the following format:

   YOUR NAME – Student Data Exercise 1(b)

Next, use the remaining steps to do calculations and create other graphs needed for a simple linear regression. Verify that your SPSS output contains all of the following items:

1. a graph of the least squares line on a scatterplot;
2. tables titled Descriptive Statistics, Correlations, Model Summary, ANOVA, and Coefficients;
3. a normal curve super-imposed onto a histogram, a normal probability plot, and a residual plot.

Submit this copy of the SPSS output with this assignment.
(c) Complete the following sentences concerning the assumptions required to proceed with a simple linear regression analysis:

Since the points on the scatterplot look like they vary around the graph of a \underline{straight line} curve instead of a line with a \underline{negative} slope, and the points on the residual plot look \underline{randomly} distributed above and below the horizontal line at zero, then it appears that the \underline{linearity} assumption is satisfied.

Since the points on both the scatterplot and the residual plot seem to show roughly the same variation everywhere, it appears that the uniform variance assumption is satisfied.

Since the points on the normal probability plot appear reasonably close to the diagonal line, it appears that the normality assumption is satisfied.

(d) Complete the following sentences concerning the simple linear regression analysis:

Using the \( f \) test in the analysis of variance for the simple linear regression, we find that the linear regression to predict \underline{high school gpa} from \underline{weekly TV hours} is statistically significant at the \( 0.05 \) level (\( f_{1,25} = 38.223, f_{1,25}^{0.05} = 4.24, p < 0.001 \)), that is, that the linear relationship between \underline{high school gpa} and \underline{weekly TV hours} is statistically significant. The data suggest that the linear relationship is positive.

About \( 60.5\% \) of the variation in \underline{high school gpa} is explained by \underline{weekly TV hours}, and the correlation is \( r = -0.778 \).

Student Data Exercises
(e) Complete the following sentences concerning the least squares line in the simple linear regression analysis:

\[ \hat{y} = 4.081 - 0.043x \quad \text{OR} \quad gpa = 4.081 - 0.043(tvh) \]

The least squares regression equation is \[ \hat{y} = 4.081 - 0.043x \quad \text{OR} \quad gpa = 4.081 - 0.043(tvh) \].

With each increase of one \[ \text{hour in weekly TV time} \], there is an average \[ \text{decrease} \] of \[ 0.043 \text{ in high school gpa} \].

A \[ 95\% \] confidence interval for the slope is \[ \text{of interest, since the linear regression is statistically significant} \].

We can \[ 95\% \] confident that the slope in the regression to predict high school gpa from weekly TV time is between \(-0.057\) and \(-0.028\).

---

(f) Use the SPSS output to find each of the following: \[ n = 27 \]

\[ \bar{x} = 28.22 \quad \bar{y} = 2.8807 \quad s_x = 9.795 \quad s_y = 0.53592 \]

standard error of estimate \[ s = 0.35950 \quad \sum (x - \bar{x})^2 = 2494.493 \]

(g) First, find the \( t \) test statistic in a hypothesis test to see if there is any evidence that the slope in the simple linear regression is different from zero; then, complete the following sentences:

\[ t_{25} = -6.182 \]

The \( f \) test statistic \( f_{1, n-2} \) in the analysis of variance for the simple linear regression and the \( t \) test statistic with a two-sided test comparing the slope to 0 are really testing the same hypothesis, since \[ f_{1, n-2} = f_{n-2}^2 \]. For this data we have \[ 38.223 = (-6.182)^2 \].
(h) The 0.05 significance level is to be used for a hypothesis test to see if there is any evidence that the slope in the simple linear regression is different from −0.06. Complete the following statement of results for this hypothesis test:

Using the appropriate t test, we find that the difference between the estimated slope (−0.043) and the hypothesized slope (−0.06) is not statistically significant at the 0.05 level ($t_{25} = -2.537$, $t_{25;0.025} = 2.060$, $0.01 < p < 0.025$, two-sided).

We conclude that the slope in the regression to predict high school GPA from weekly TV hours is not different from −0.06.

A 95% confidence interval for the slope is of interest, since the null hypothesis is not rejected.

This confidence interval was obtained in part (e).

(i) The 0.05 significance level is to be used for a hypothesis test to see if there is any evidence that the mean high school GPA with 30 hours of TV per week is less than 3.00. Complete the following statement of results for this hypothesis test:

Using the appropriate t test, we find that the difference between the mean grip strength for 20 year old right-handed males estimated from the least squares line (4.081 − 0.043(30) = 2.805) and the hypothesized mean (3.00) is statistically significant at the 0.05 level ($t_{25} = -2.897$, $t_{25;0.05} = 1.708$, $0.001 < p < 0.01$, one-sided).

We conclude that the mean high school GPA with 30 hours of TV per week is less than 3.00.

A 95% confidence interval for the mean high school GPA with 30 hours of TV per week is of interest, since the null hypothesis is rejected.

We can 95% confident that the mean high school GPA with 30 hours of TV per week is between 2.667 and 2.944.
(j) Find and interpret a 95% prediction interval for the gpa of a high school student with 10 hours of study per week.

We are 95% confident that the gpa for a randomly selected high school student with 10 hours of study per week will be between 2.084 and 3.526.

OR

At least 95% of high school students with 10 hours of study per week have a gpa between 2.084 and 3.526.

(k) For what weekly study time will the confidence interval for mean high school gpa and the prediction interval for a particular high school student’s gpa both have the smallest length?

28.22 study hours per week
3. It is of interest to see if a multiple linear regression to predict high school grade point average from both predictor variables weekly study hours and weekly TV hours is an improvement over the prediction using just one of the two predictor variables. A 0.05 significance level is chosen for all hypothesis testing.

(a) Use SPSS to obtain a correlation matrix for all the quantitative variables by going to the document titled Using SPSS for Windows (which can be accessed from the appropriate link on the course syllabus web page), going to the section titled Hypothesis Tests Involving Two Variables, and using the steps in the subsection titled Generating a Correlation Matrix with p-values as a guide; there is no need to use the steps in this subsection to create scatter plots, since this was done already in previous exercises. Once you have successfully generated SPSS output, add a title to the top of the output in the following format:

YOUR NAME – Student Data Exercise 3(b)

(b) Use the fact that the $t$ test in an analysis of variance for a simple linear regression is equivalent to the two-sided test of statistical significance of the Pearson correlation between the dependent variable and the independent variable to complete the following sentences concerning each simple linear regression to predict high school grade point average:

Since the Pearson correlation between gpa and weekly study hours is not statistically significant at the 0.05 level ($r = 0.753, n = 27$, $p < 0.001$, two-sided), then the linear regression to predict gpa from weekly study hours is not statistically significant at the 0.05 level.

Since the Pearson correlation between gpa and weekly TV hours is not statistically significant at the 0.05 level ($r = -0.778, n = 27$, $p < 0.001$, two-sided), then the linear regression to predict gpa from weekly TV hours is not statistically significant at the 0.05 level.
Among the two possible independent variables, the one that would be chosen for a simple linear regression to predict $\text{gpa}$ is $\text{weekly TV hours}$, since this is the independent variable which has the strongest statistically significant Pearson correlation with $\text{gpa}$, and therefore will result in the largest statistically significant $t$ test statistic (i.e., the one with the smallest corresponding $p$-value).

(c) Use SPSS to obtain the ANOVA table for the regression to predict high school gpa from weekly study hours and weekly TV hours by doing the following:

Select the Analyze > Regression > Linear options to display the Linear Regression dialog box.

From the list of variables on the left, select the dependent (response) variable gpa and click on the arrow pointing toward the Dependent slot.

From the list of variables on the left, select each independent (explanatory or predictor) variable weekly study hours and weekly TV hours, and click on the arrow pointing toward the Independent(s) section (where selection of more than one variable is permitted).

Click on the OK button, after which the SPSS output will be generated.

Verify that your SPSS output contains a table titled Model Summary and a table titled ANOVA.
Calculate the $f$ statistic used to test whether or not the addition of weekly study hours after weekly TV hours is statistically significant by performing the following steps:

We shall let $SSR(X_1, X_2) = \text{the regression sum of squares from the ANOVA table with both variables } X_1 \text{ and } X_2 \text{ in the model}$, let $SSR(X_1) = \text{the regression sum of squares from the ANOVA table with only } X_1 \text{ in the model}$, and let $MSR(X_1, X_2) = \text{the error mean square from the ANOVA table with both } X_1 \text{ and } X_2 \text{ in the model}$. Calculate the $f$ statistic to decide if the addition of $X_2$ to the model after $X_1$ is statistically significant as follows:

$$\frac{SSR(X_1, X_2) - SSR(X_1)}{MSE(X_1, X_2)} = \frac{4.614 - 4.515}{0.119} = 0.832$$

The numerator degrees of freedom for this $f$ statistic is 1 (the number of variables being added to the model), and the denominator degrees of freedom is the same as the degrees of freedom associated with $MSE(X_1, X_2)$.

Complete the following statement concerning this $f$ statistic:

The $f$ test concerning the addition of __________ weekly study hours __________ after __________ weekly TV hours __________ is not statistically significant at the __________ 0.05 __________ level ($f_{1,24} = 0.832, f_{1,24; 0.05} = 4.26, 0.10 < p$).

We conclude that the addition of weekly study hours after weekly TV hours in the regression equation does not improve the prediction of gpa.

(d) Use SPSS to do a stepwise regression for the prediction of high school gpa from weekly study hours and weekly TV hours by doing the following:

Select the Analyze > Regression > Linear options to display the Linear Regression dialog box.

From the list of variables on the left, select the dependent (response) variable “gpa” and click on the arrow pointing toward the Dependent slot.

From the list of variables on the left, select all of the independent (explanatory or predictor) variables “weekly study hours” and “weekly TV hours”, and click on the arrow pointing toward the Independent(s) section (where selection of more than one variable is permitted).

In the Method slot, select the Stepwise option.

Click on the OK button, after which the SPSS output will be generated.

Verify that your SPSS output contains the following items:
(1) a table titled **Variables Entered/Removed**;
(2) a table titled **Model Summary**;
(3) a table titled **ANOVA**;
(4) a table titled **Coefficients**;
(5) a table titled **Excluded Variables**.

(e) Complete the following sentences concerning the stepwise regression analysis:

There ___ was one ___ step in the stepwise regression performed at the ___ 0.05 ___ level. The predictor variable ___ weekly TV hours ___ was entered in the first step (___ $t_{25} = -6.182$, $t_{25;0.025} = 2.060$, $p < 0.001$ ___) and accounted for ___ 60.5% ___ of the variance in ___ gpa ___.

The correlation between ___ gpa ___ and ___ weekly TV hours ___ is ___ $r_{gpa,TV} = -0.778$, $n = 27$ ___.

The predictor variable weekly study hours was not entered into the regression equation.

The least squares regression equation from the model at the final step is ___ $\hat{gpa} = 4.081 - 0.043(tvh)$ ___.

---

Student Data Exercises  14