39. The data is to be used in a study concerning the prediction of yearly income from possible predictors sex, area of residence, number of children, age, job satisfaction score, weekly TV hours, and weekly radio hours. A 0.05 significance level is chosen for all hypothesis testing, and the updated SPSS data file survey created in Exercise #20 will be used.

(a) Complete the following statements concerning the potential independent variables that are qualitative in the SPSS data file:

The variable sex is can be treated as an indicator (dummy) variable, since male is coded with 0 and female is coded with 1; that is, we may write

\[
\text{sex} = \begin{cases} 
0 & \text{for male} \\
1 & \text{for female} 
\end{cases}
\]

The variable area of residence cannot be treated as an indicator (dummy) variable, since rural is coded with 1, suburban is coded with 2, and urban is coded with 3. We can define the following three dummy variables:

\[
R_1 = \begin{cases} 
1 & \text{for rural} \\
0 & \text{otherwise} 
\end{cases} \quad R_2 = \begin{cases} 
1 & \text{for suburban} \\
0 & \text{otherwise} 
\end{cases} \quad R_3 = \begin{cases} 
1 & \text{for urban} \\
0 & \text{otherwise} 
\end{cases}
\]

Any 2 of these dummy variables is sufficient to represent the variable area of residence.

(c) Go to the document titled Using SPSS for Windows (which can be accessed from the appropriate link on the course syllabus web page), go to the section titled Data Entry and Manipulation, and use the steps in the subsection titled Creating New Variables by Recoding Existing Variables as a guide to recode the variable area of residence into the indicator (dummy) variable \( R_1 \) in part (b); then repeat this for the dummy variable(s) \( R_2 \) and \( R_3 \) in part (b).

(d) Go to the document titled Using SPSS for Windows (which can be accessed from the appropriate link on the course syllabus web page), go to the section titled Hypothesis Tests Involving Two or More Variables, and use the first 6 steps in the subsection titled Performing a Multiple Linear Regression with Checks for Multicollinearity and of Linearity, Homoscedasticity, and Normality Assumptions as a guide to creating two scatter plots, each one with the dependent variable on the vertical axis and one of the potential quantitative independent variables on the horizontal axis.
After noting that Step 7 is has already been done in part (c), use the remaining steps beginning with Step 8 as a guide to do a multiple regression for the prediction of drying time from all of the potential predictors and to obtain output useful in checking for multicollinearity and in verifying linearity, homoscedasticity, and normality assumptions. Once you have successfully generated SPSS output, add a title to the top of the output in the following format:

YOUR NAME - Survey Data Exercise 39(d)

Verify that your SPSS output contains all of the following items:

1. two scatterplots, each displaying the least squares line;
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.

(e) Complete the following sentences concerning the assumptions required to proceed with a multiple linear regression analysis:

The points look like they vary around the graph of a straight line with positive slope on the scatterplot corresponding to **age and TV hrs**. The points look like they vary around the graph of a straight line with negative slope on the scatterplot corresponding to **children, job satisfaction, radio hrs**. The points on the residual plot look **randomly** distributed above and below the horizontal line at zero. Consequently, it appears that the **linearity** assumption and the **homoscedasticity** assumption are both satisfied.

The points on the normal probability plot look **reasonably close to** the diagonal line. Consequently, it appears that the **normality** assumption is satisfied.
(f) Use SPSS to do a stepwise regression for the prediction of wheat yield from the potential predictors 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 or More Variables, and using the steps in the subsection titled Performing a Regression (or Related Procedure) to Build a Model as a guide.

Once you have successfully generated SPSS output, add a title to the top of the output in the following format:

YOUR NAME - General Exercise 39(f)

Verify that your SPSS output contains all of 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.

(g) Complete the following sentences concerning multicollinearity:

The correlation matrix contains \textbf{no} at least one correlation with an absolute value greater than 0.8 between any pair of independent variables. Also, we see that \textbf{no} at least one independent variable shows a tolerance less than 0.10 (i.e., VIF > 10). Consequently, we believe that multicollinearity will /will not be a problem.

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

There \textbf{three steps} in the stepwise regression performed at the 0.05 level.

In the first step, the predictor variable entered into the regression equation was \textbf{weekly TV hours} ($t_{28} = 3.214, t_{28; 0.025} = 2.048, p = 0.003$).

This variable accounted for \textbf{27.0\%} of the variance in yearly income. The correlation between yearly income and weekly TV hours is $r_{in, tv} = +0.519$ ($n = 30$).
Complete the following sentences concerning the stepwise regression analysis:

In the second step, the predictor variable entered into the regression equation was age ($t_{27} = 4.276$, $t_{27; 0.025} = 2.052$, $p < 0.001$).

The two predictor variables weekly TV hours and age accounted for 56.4% of the variance in yearly income. The partial correlation between yearly income and age given weekly TV hours is $r_{in, age | tv} = +0.635$ ($n = 30$).

In the third step, the predictor variable entered into the regression equation was the indicator variable for the rural area of residence ($t_{26} = -5.683$, $t_{26; 0.025} = 2.056$, $p < 0.001$). The three predictor variables accounted for 80.6% of the variance in yearly income. The partial correlation between yearly income and the indicator variable for the rural area given weekly TV hours and age is $r_{in, R1 | tv, age} = -0.744$ ($n = 30$).

\[
\hat{\text{income}} = -4.825 + 0.855(\text{tv}) + 1.060(\text{age}) - 21.183(\text{R}_1)
\]

(i) For each of the estimated regression coefficients in the least squares regression equation, write a one-sentence interpretation describing precisely what that coefficient is estimating.

For each increase of one hour in weekly TV hours, yearly income increases on average by about 0.855 thousand dollars.

For each increase of one year in age, yearly income increases on average by about 1.060 thousand dollars.

On average, yearly income is about 21.183 thousand dollars smaller in the rural area than in other areas.
(j) Complete the following sentences concerning the least squares regression equation:

The fact that the indicator variable for a rural area of residence \( (R_1) \) is in the least squares regression equation suggests a statistically significant difference between the rural area and the other two areas but no statistically significant difference between the other two areas (i.e., between the suburban area and urban area).

We can write two separate least squares regression equations. The equation for the rural area \( (R_1 = 1) \) is

\[
\hat{\text{income}} = -26.008 + 0.855(tv) + 1.060(age)
\]

and the equation for the other two areas \( (R_1 = 0) \)

is

\[
\hat{\text{income}} = -4.825 + 0.855(tv) + 1.060(age)
\]