1. A SAS data set named fuel.dat was created using the following input statement:

```sas
input State $ Pop Tax Numeric Income Roads Fuel;
```

Answer parts (a) to (e) below. (The answers to these parts must not be SAS/GRAPH procedures e.g., gplot or gchart)

(a) (2) Give the name of a SAS procedure that produces one-way or two-way tables of frequency counts and \( \chi^2 \) tests of independence for category variables LicGrp and IncomeGrp. \texttt{freq}\$

(b) (2) Give the name of a SAS procedure that enables you to create a variety of tables of statistics computed for variables such as Fuel, classified by variables such as LicGrp and IncomeGrp, that may be highly customized. \texttt{tabulate}\$

(c) (2) Give the name of a SAS procedure to produce horizontal barcharts of fuel use for three income groups in separate panels showing data for States in two fuel tax groups. \texttt{sgpanel}\$

(d) (2) Give the name of a SAS procedure that you may use to obtain a histogram of a continuous-valued variable such as Roads in ODS graphics. \texttt{univariate, sgplot}\$

(e) (2) Give the name of a SAS procedure that you may use to obtain a scatterplot matrix of variables Pop, Tax, Numeric, Income, Roads, and Fuel using ODS statistical graphics. \texttt{sgscatter}\$

2. (a) (4) Examine the proc step

```sas
proc format;
  value ms 1 = 'Single'
          2 = 'Married'
          3 = 'Divorced'
          4 = 'Widowed' ;
```

Suppose that the variable MaritalStatus has values of 1, 2, 3 or 4. Write a SAS statement would make SAS print the formatted values of this variable in any proc step.

```sas
format MaritalStatus ms. ;
```

(b) (4) To examine whether a Poisson distribution provides a reasonable model for the number of cell clumps per algae species in a lake the following statement is included in a proc freq step:

```sas
tables Clumps/ nocum testp=(3.7 12.2 20.1 22.1 18.2 12.0 6.6 5.1);
```

Explain as much as possible the details of the statistical test this statement will produce.

Calculates a \( \chi^2 \) goodness-of-fit test for testing the hypothesis that number of clumps follow a Poisson distribution

(c) (4) The following statement is included in a proc sgplot step using the fuel data set as input, where LicGrp and IncomeGrp are group variables each with 3 categories:

```sas
hbar LicGrp/response=Fuel stat=mean group=IncomeGrp;
```

Explain as much as possible the SAS output this statement will produce.

A horizontal bar chart of mean fuel consumption for each of 3 LicGrps, divided into income groups.
3. The graph below of side-by-side boxplots summarize birth rates of 60 countries according to the category variable level of technology.

(a) (4) Compare the shapes of the distributions across the three categories of level of technology. Estimate largest birth rate for all countries? The smallest?

The shapes of the 3 distributions are different: skewed to the left, symmetric and skewed to the right as level of Technology increases.

54, 10

(b) (4) Compare the location of the distributions across the three categories of level of technology. Which category appears to have more of the countries with lower birth rates? With higher birth rates?

Location decreases from about 45 to about 15 as measured by the median as level of Technology increases.

High Level of Technology category; Low group as 50% of

(c) (4) Compare the spread (or dispersion) of the distributions for the three categories of level of technology. Which category of countries (with respect to level of technology) have more countries with a largest range of birth rates?

The spread increases and then decreases as Technology level increases

Moderate Category

(d) (4) Name 3 features in the boxplots below you could use to verify that model assumptions needed for analyzing this data using ANOVA methods are plausible.

Shapes: do they indicate normal samples?

Spread: size of the boxes (IQR): are they similar?

Outliers: are there extreme values that need to be examined to see if they are outliers?

Box Plots of Birth Rate by Level of Technology
4. Examine the proc step

```
proc univariate data=mylib.fueldata cibasic normal mu0=4 5;
  var Income Roads;
  title 'Use of Proc Univariate to Examine Distributions';
run;
```

Explain statistics produced by the proc univariate step above for each of the two variables given below:

(a) (3) Income

- Moments, Basic Statistics, Quantiles of Income Variable
- 95% CI for \( \mu, \sigma \) of Income Population
- Test of \( H_0: \mu = 4 \) vs. \( H_a: \mu \neq 4 \)

(b) (3) Roads

- Same statistics as above for Roads Population except
- Test of \( H_0: \mu = 5 \) vs. \( H_a: \mu \neq 5 \) for Roads Variability

5. Below is a Q-Q plot of measurements of tire wear on two different brands of tires:

```
5 -Q-Q plot of Tire Wear Data
```

(a) (2) Using the above graph, compare the shapes of the distributions of tire wear measurements of the two brands of tires.

The two populations have the same shape as the respective quantiles lie on a straight line approximately.

(b) (2) Using the above graph, compare the median tire wear of the two brands of tires.

\[
\text{Median} = \frac{(\text{Brand} \#1 \text{ Median}) + (\text{Brand} \#2 \text{ Median})}{2} \quad \text{Brand} \#1 \text{ Median } \times 16 \quad \text{Brand} \#2 \text{ Median } \times 12
\]

(c) (2) Using the above graph, compare the variability of tire wear of the two brands of tires.

\[
\text{Approximate Slope} = \frac{25-13}{20-16} = 1.2
\]

So \( \text{Var (Brand} \#1) \approx 1.2^2 \), i.e. ~1.5 times \( \text{Var (Brand} \#2) \)
6. Consider following data:

<table>
<thead>
<tr>
<th>x</th>
<th>22</th>
<th>19</th>
<th>24</th>
<th>28</th>
<th>18</th>
<th>21</th>
<th>26</th>
<th>15</th>
<th>14</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>11.2</td>
<td>12.4</td>
<td>14.7</td>
<td>18.5</td>
<td>10.8</td>
<td>12.3</td>
<td>13.1</td>
<td>11.4</td>
<td>8.7</td>
<td>11.9</td>
<td>12.4</td>
</tr>
<tr>
<td>x(contd.)</td>
<td>15</td>
<td>25</td>
<td>13</td>
<td>17</td>
<td>8</td>
<td>6</td>
<td>7</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y(contd.)</td>
<td>7.1</td>
<td>11.1</td>
<td>7.2</td>
<td>8.9</td>
<td>6.7</td>
<td>4.6</td>
<td>8.3</td>
<td>9.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The procedure reg in SAS was used to perform analysis of this set of data using the model

\[ y = \beta_0 + \beta_1 x_1 + \epsilon \]

where \( \epsilon \) are assumed to be independently distributed as \( N(0, \sigma^2) \) variables. Answer the following questions based on the results appearing in the output attached to the end of the question.

i) (2) What is the value of \( R^2 \)? What does this say generally about the fit of this model?

\[ R^2 = 67.8\% \]

Indicates moderate fit with 67.8% of the variability in \( y \) being explained by a straight line model with \( x_1 \) as the explanatory variable.

ii) (2) Give the residual sum of squares and its degrees of freedom. What is the estimate \( \hat{\sigma}^2 \)?

\[ SSE = 60.33 \quad d.f. = 18 \]

\[ \hat{\sigma}^2 = 3.35 \]

iii) (2) What is the F statistic for testing \( H_0 : \beta_1 = 0 \) vs. \( H_a : \beta_1 \neq 0 \)? Make a decision based on the p-value.

\[ F = 37.97 \quad p < .0001 \]

Since \( p < .05 \), we reject \( H_0 \) at \( \alpha = .05 \). Thus a simple linear regression model with \( x_1 \) as the explanatory variable is feasible.

iv) (4) Using the estimate of \( \beta_1 \) and its standard error, compute the t-statistic for testing \( H_0 : \beta_1 = 0 \) vs. \( H_a : \beta_1 \neq 0 \).

\[ t = \frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)} = \frac{.41992}{.06798} = 6.16 \]

v) (4) Using the estimate of \( \beta_1 \) and its standard error, compute a 95% confidence interval for \( \beta_1 \).

\[ \hat{\beta}_1 \pm t_{.025, 18} \cdot \hat{\sigma}(\hat{\beta}_1) = .41992 \pm 2.101 \cdot .06798 \]

\[ = (0.2761, 0.5617) \]

vi) (4) Use the value of \( h_{44} \) to compute the standard error of the residual for observation 4.

\[ \text{S.E.}(e_4) = \sqrt{1 - h_{44}} = 1.83079 \sqrt{1 - 0.2108} \]

\[ = 1.6264 \]

vii) (4) Use the value of \( h_{44} \) to compute the standard error of \( \hat{\beta}_4 \)

\[ \text{S.E.}(\hat{\beta}_4) = \sqrt{h_{44}} = 1.83079 \sqrt{0.2108} \]

\[ = 0.8406 \]
viii) (4) Use the residual for observation 4 and its standard error to calculate the corresponding studentized residual.

\[ r_{4f} = \frac{e_4}{\sqrt{1-h_{44}}} = \frac{e_4}{\text{r.e.}\,(e_4)} = \frac{3.4751}{1.626} = 2.137 \]

ix) (4) Use an appropriate plot to determine possible y-outliers. Name this plot, give the case number(s) of the suspected y-outliers and explain why you think each is a y-outlier.

Plot B or C. Case D is outside ±2 range for R(Student), indicating possible y-outlier. Need to perform a test for y-outlier.

x) (4) Find any cases, if any, that may be x-outliers using a suitable cut-off value. Name the plot that you may use to determine if x-outliers are present. Does this plot indicate any x-outliers?

Cut-off for hat i's \( \frac{2\hat{e}(e+1)}{n} \) = 4 = .2

Cases D & R have hat values >.2 (in Table)

Plot C indicates this clearly (see leverage axis).

xi) (4) Find any cases from the output statistics (give case number), if any, that may be influential explaining why you selected these. Use a plot to determine an influential case, name this plot and say why you selected this case.

Case D has a large Cook's D (in Table)

Plot E clearly pinpoints this case (#4).

xii) (4) If you find any case to be influential, explain why or why not this case should be examined carefully. Use other related case statistics and plots in your explanation.

Case D is clearly influential and is a possible y-outlier. Thus we need to verify if this is a correct observation (Plots C & E).

xiii) (4) Examine the residuals vs. predicted values plot. State what you think this plot indicates about the fit of the data to a straight line model. Is the straight line model adequate?

This plot (Plot A) clearly shows random scatter around the zero reference line and does not show a pattern (and even spread) as y ↑

xiv) (4) Examine the normal probability plot of the residuals. Name this plot. What assumption about the model is checked using this plot? Explain whether the plot provides evidence to support this assumption.

The normal probability plot (Plot D) shows no marked deviation (or pattern) away from a straight line implying that this assumption is plausible.