HOMEWORK 4 SOLUTION

3.7

a. Stem-and-Leaf Display: x

Stem-and-leaf of x  N = 60
Leaf Unit = 1.0

8   4  1122334
15  4  567788
21  5  123344
29  5  56777999
(8)  6  00013334
23  6  5556889
16  7  001223

I agree that the plot is consistent with the random selection of women from each 10-year age group for 40-60. Women in their mid to late 70’s are not represented.

b. The residuals are scattered around zero, with what appears to be a symmetric distribution. Checking for normality would be difficult from this plot.

c. The two graphs provide the same information. No clear departure from the regression line is evident.
Residuals Versus the Fitted Values
(response is y)

Residuals Versus x
(response is y)

d.
$H_0$: Normal
$H_a$: not normal
$r = 0.9897$
Decision rule: If $r \geq 0.984$ conclude $H_0$, otherwise $H_a$.
Conclude $H_0$.

e. $SSR^* = 31, 833.4, SSE = 3, 874.45, \chi^2_{BP} = (31, 833.4/2) \div (3, 874.45/60)^2 = 3.817116, \chi^2(.99; 1) = 6.63$. If $\chi^2_{BP} \leq 6.63$ conclude error variance constant, otherwise error variance not constant. Conclude error variance constant. The same results as part c.

3.10
a. There is an outlier in the observations – since it is at the largest predicted value it can be influential and pull the regression line towards it changing the slope significantly.
b. There are 4 points fall outside 1 standard deviation. If the normal error model is appropriate, there should be about 3.8 or 4 points fall outside.

3.15
a. **Regression Analysis: y15 versus x15**

The regression equation is
\[ y_{15} = 2.58 - 0.324x_{15} \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2.5753</td>
<td>0.2487</td>
<td>10.35</td>
<td>0.000</td>
</tr>
<tr>
<td>x15</td>
<td>-0.32400</td>
<td>0.04330</td>
<td>-7.48</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ S = 0.474314 \quad \text{R-Sq} = 81.2\% \quad \text{R-Sq(adj)} = 79.7\% \]

b. \( H_0: E(Y) = \beta_0 + \beta_1X \),
\[ Ha: E(Y) \text{ not equal } \beta_0 + \beta_1X \]
\[ F^* = 58.60 \]
Decision Rule: \( F(0.975; 3, 10) = 4.82562 \). If \( F^* \leq 4.82562 \) conclude \( H_0 \), otherwise \( Ha \).
Conclude \( Ha \).

Or you can use the fact that the P-value is below 0.025 which means \( H_0 \) would be rejected.
Regression       1  12.597  12.597  55.99  0.000
Residual Error  13   2.925   0.225
Lack of Fit    3   2.767   0.922  58.60  0.000
Pure Error    10   0.157   0.016
Total           14  15.522

c. The lack of fit only indicates that a linear regression is inappropriate, but does not indicate what regression function to try next. A study of the residuals might help.

3.16
a. By examining the scatter plot, I might try to transform Y to logY to attain a constant variance and linearity. There are problems with both – log should improve the constant variance – if linearity is still a problem you might transform x as well.

c. The regression equation is
logY15 = 0.655 - 0.195 x15

d. The regression line appears to be a good fit for the transformed data and the constant variance assumption seems satisfied on the new scale.
e. The following graph shows a good linear regression fit for the transformed data.
f. The transformed regression equation is
\[ Y_{15} = \frac{4.518559}{10^{0.195x}} \]

3.15
a. I'd try to transform x first, since the variances seem to be constant.
b. Regression Analysis: $y_{18}$ versus $\sqrt{x_{18}}$

The regression equation is

$y_{18} = 1.25 + 3.62 \sqrt{x_{18}}$

c. The graph shows a better fit after the transformation.
d. The graphs show the linear regression model is a good fit after transformation.
e. The question is ambiguous – when transforming x the y’s are still in the original units, although x is not.

\[(y - 1.25)^2 = 13.1044 \times \]

3.20 The error terms are still independent normal after the transformation. But if the same transformation is used on Y, the error terms no longer have normal distribution.

3.23

\[H_0: E\{Y\} = \beta_1 X,\]

\[H_a: E\{Y\} \text{ not equal } \beta_1 X\]

The degrees of freedom are full = 10 and reduced = 19.

3.23

a. There seems an outlier at x=12.

**Regression Analysis: y24 versus x24**

The regression equation is

\[y24 = 48.7 + 2.33 \times x24\]
b. Case 7 pulls the regression line toward the count-clockwise direction. The regression equation is
\[ y_{24} = 53.1 + 1.62 \times x_{24} \]

\[ b. \text{ Case 7 pulls the regression line toward the count-clockwise direction.} \]
\[ \text{The regression equation is} \]
\[ y_{24} = 53.1 + 1.62 \times x_{24} \]

\[ \text{c. Case 7 falls out of the 99\% prediction interval. That means case 7 doesn’t seem to follow} \]
\[ \text{the pattern of the rest of the data.} \]

\[ \begin{array}{cccc}
\text{New} & \text{Obs} & \text{Fit} & \text{SE Fit} & \text{99\% CI} & \text{99\% PI} \\
& 1 & 72.52 & 1.47 & (66.58, 78.47) & (60.31, 84.74) \\
\end{array} \]

\[ 3.32 \]

Based on the original data, the linear regression model is fitted. And the residual seems to increase as \( X \) increase. Although the normal probability plot suggests the normal assumption for the error term isn't satisfied, this plot is not appropriate when the variances are unequal.

**Regression Analysis: PSA level versus cancel volume**

The regression equation is
\[ \text{PSA level} = 1.12 + 3.23 \text{ cancel volume} \]
Since the residuals increase as the cancel volume increases and the increasing rate of PSA level with the increase of cancel volume is decreasing, so I transform PSA level to log(PSA)
level). The fitted line is as below. Under this model, the assumptions fit much better. The transformation of the response variable makes the explanation a little harder.

**Regression Analysis: log(y) versus x**

The regression equation is

\[
\log(y) = 1.81 + 0.0962 \times x
\]

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95% CI</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.7292</td>
<td>0.1719</td>
<td>(3.3880, 4.0705)</td>
<td>(1.9604, 5.4980)</td>
</tr>
</tbody>
</table>

At the cancel volume 20 cc, the estimated mean for PSA level is 41.64578.
Residuals Versus the Fitted Values
(response is log(y))

Residuals Versus x
(response is log(y))