1. Use the patient satisfaction data described in KNNL Problem 6.15.

a) Compute the pairwise correlations between the X’s and between each X and Y. Which X variable appears to be the best individual predictor?

There do not appear to be any really strong correlations among the explanatory variables (all < 0.70). Age (x1) appears to be the best individual predictor since it has the largest correlation ($r = -0.79$).


<table>
<thead>
<tr>
<th></th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>1.00000</td>
<td>0.56795</td>
<td>0.56968</td>
<td>-0.78676</td>
</tr>
<tr>
<td>x2</td>
<td>0.56795</td>
<td>1.00000</td>
<td>0.67053</td>
<td>-0.60294</td>
</tr>
<tr>
<td>x3</td>
<td>0.56968</td>
<td>0.67053</td>
<td>1.00000</td>
<td>-0.64459</td>
</tr>
<tr>
<td>y</td>
<td>-0.78676</td>
<td>-0.60294</td>
<td>-0.64459</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

b) Run the linear regression with age, severity of illness and anxiety level as the explanatory variables and satisfaction as the response variable. Summarize the regression results.

The fitted regression line is $\hat{Y} = 158.49 - 1.14X_1 - 0.44X_2 - 13.47X_3$ where $X_1$ is age, $X_2$ is severity level, and $X_3$ is anxiety level. The coefficient of determination is $R^2 = 0.6822$ and the P-value of the $F$ test is < .0001 suggesting that this set of explanatory variables is helpful in explaining satisfaction. The test is testing the hypotheses

\[ H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \]
\[ H_a : \text{at least one } \beta_i \neq 0 \]

The test statistic is 30.05 and the degrees of freedom are 3 and 42.

Only one individual parameter t-tests is significant (Age) suggesting that perhaps not all variables are needed in predicting satisfaction. All of the coefficients are negative suggests an increase in any one of them results in an decrease in satisfaction.

c) Plot the residuals versus the predicted satisfaction and each of the explanatory variables. Are there any unusual patterns?

I do not see any unusual patterns in any of the residual plots.
d Examine the assumption of normality for the residuals using a qqplot or histogram. State your conclusions.

The data appears to be reasonably normal. No real deviations can be found in these plots.

![QQ Plot and Histogram](image)

e Predict the satisfaction for a 55 year old patient with illness severity 50 and anxiety level 2.8. Provide a 95% prediction interval with your prediction.

The following is from SAS using the CLI option.

<table>
<thead>
<tr>
<th>Obs</th>
<th>Variable</th>
<th>Value</th>
<th>Predicted Mean</th>
<th>Std Error</th>
<th>95% CL Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td></td>
<td>35.8859</td>
<td>4.6332</td>
<td>13.5381</td>
<td>58.2338</td>
</tr>
</tbody>
</table>

2. KNNL Problem 7.5

This can be done using Type I sum of squares. The following output summarizes the various extra sums of squares. As far as the test in part b, this is the same as the t-test since the \(X_3\) variable is fitted last. In this case, we would not reject and conclude that this variable could be dropped given the other two are already in the model. The test is testing the hypotheses

\[
H_0 : E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\
H_a : E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3
\]
### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>9120.46367</td>
<td>3040.15456</td>
<td>30.05</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>42</td>
<td>4248.84068</td>
<td>101.16287</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>45</td>
<td>13369</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Root MSE: 10.05798
- R-Square: 0.6822
- Dependent Mean: 61.56522
- Adj R-Sq: 0.6595
- Coeff Var: 16.33711

| Variable | DF  | Estimate  | Error  | t Value | Pr > |t| | Type I SS |
|----------|-----|-----------|--------|---------|-------|---|----------------|
| Intercept| 1   | 158.49125 | 18.1259| 8.74    | <.0001|   | 174353        |
| x2       | 1   | -0.44200  | 0.49197| -0.90   | 0.3741|   | 4860.26000    |
| x1       | 1   | -1.14161  | 0.21480| -5.31   | <.0001|   | 3896.04414    |
| x3       | 1   | -13.47016 | 7.09966| -1.90   | 0.0647|   | 364.15952     |

3. KNNL Problem 7.6

This can be done using the TEST option. The output below summarizes the test. Using $\alpha = .025$, we would reject the null hypothesis and assume that we cannot throw both the variables out of the model.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerator</td>
<td>2</td>
<td>422.53741</td>
<td>4.18</td>
<td>0.0222</td>
</tr>
<tr>
<td>Denominator</td>
<td>42</td>
<td>101.16287</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. KNNL Problem 7.14

These can be obtained using the pcorr1 pcorr2 options (interpretations of type I and type II are similar to ss1 and ss2, respectively).

| Variable | DF  | Parameter Estimate  | Standard Error  | t Value | Pr > |t| | Corr Type I | Corr Type II |
|----------|-----|---------------------|-----------------|---------|-------|---|----------------|-------------|
| Intercept| 1   | 158.49125           | 18.12589        | 8.74    | <.0001|   | 0.61898       | 0.40211     |
| x1       | 1   | -1.14161            | 0.21480         | -5.31   | <.0001|   | 0.94441       | 0.01886     |
| x2       | 1   | -0.44200            | 0.49197         | -0.90   | 0.3741|   | 0.07894       | 0.07894     |
| x3       | 1   | -13.47016           | 7.09966         | -1.90   | 0.0647|   | 0.07894       | 0.07894     |
Parameter Estimates

| Variable | DF | Estimate | Error   | t Value | Pr > |t| | Corr Type I | Corr Type II |
|----------|----|----------|---------|---------|-------|-----------------|-----------------|
| Intercept| 1  | 158.49125| 18.12589| 8.74    | <.0001| .               | .               |
| x2       | 1  | -0.44200 | 0.49197 | -0.90   | 0.3741| 0.36354         | 0.01886         |
| x1       | 1  | -1.14161 | 0.21480 | -5.31   | <.0001| 0.45787         | 0.40211         |
| x3       | 1  | -13.47016| 7.09966 | -1.90   | 0.0647| 0.07894         | 0.07894         |

The above outputs report these values:

- $r_{Y1}^2 = 0.61898$
- $r_{Y1|2}^2 = 0.45787$
- $r_{Y1|23}^2 = 0.40211$
- $r_{Y2}^2 = 0.36354$
- $r_{Y2|1}^2 = 0.09441$
- $r_{Y2|23}^2 = 0.01886$

Based on these, one can see that $x_1$ is strongly associated with $y$ and does not change all that much given other variables in the model. Variable $x_2$, however, changes dramatically depending on what is in the model.

5. KNNL Problem 7.18

It really does not matter whether you use the standardization transformation of the correlation transformation in this case. Thus, the stb option will compute the coefficients and give you the coefficients in the original units too.
The coefficients of determination among all pairs of predictor variables is simply the correlation squared (see the correlation coefficients in Q1.a). Since this is a linear transformation, the correlation between the untransformed and transformed variables are the same. This means that a similar type of interpretation can be made regarding each coefficient. The difference is now the change is in terms of standard deviation units.

6. Derive equation 7.56 (HINT: Recall that \( b_1 \) can be obtained by regressing the residuals of \( Y|X_2 \) vs \( X_1 \mid X_2 \))

The first relationship below uses the fact that \( b_1' \) can be obtained by regressing the residuals of \( Y|X_2 \) vs \( X_1 \mid X_2 \). The second plugs in the fitted values using \( b \) and \( b^* \) for the slope estimates. The remainder uses the relationship between the Pearson’s correlation coefficient and the slope.

\[
\begin{align*}
b_1 &= \frac{\sum (X_{i1} - \bar{X}_{i1})(Y_i - \bar{Y})}{\sum (X_{i1} - \bar{X}_{i1})^2} \\
&= \frac{\sum (X_{i1} - \bar{X}_1 - b(X_{i2} - \bar{X}_2))(Y_i - \bar{Y} - b^*(X_{i2} - \bar{X}_2))}{\sum (X_{i1} - \bar{X}_1 - b(X_{i2} - \bar{X}_2))^2} \\
&= \frac{\sum (X_{i1} - \bar{X}_1)(Y_i - \bar{Y}) - 2b \sum (X_{i2} - \bar{X}_2)(Y_i - \bar{Y}) + b^* \sum (X_{i2} - \bar{X}_2)^2}{\sum (X_{i1} - \bar{X}_1)^2 - b^2 \sum (X_{i2} - \bar{X}_2)^2} \\
&= \frac{\sum (X_{i1} - \bar{X}_1)(Y_i - \bar{Y}) - 2r_{12}r_{Y2}(n-1)s_1s_Y + r_{12}r_{Y2}(n-1)s_1s_y}{\sum (X_{i1} - \bar{X}_1)^2 - b^2 \sum (X_{i2} - \bar{X}_2)^2} \\
&= \frac{b'_1 - r_{12}r_{Y2}s_Y}{s_1} \\
&= \frac{b'_1 - r_{12}r_{Y2}s_Y}{1 - r_{12}^2}
\end{align*}
\]

You can also start from equations in 6.77 (page 241): eliminate \( b_0 \) by replacing \( Y_i \) with \( Y_i - \bar{Y} \), \( X_{i1} \) with \( X_{i1} - \bar{X}_1 \), \( X_{i2} - \bar{X}_2 \).

7. KNNL Problem 8.6

Based on the plot below, a quadratic relationship appears to be a good fit.

The regression output is also shown. The overall model is significant. This is testing \( H_0 : \beta_1 = \beta_2 = 0 \) versus the alternative that at least one of them is not equal to zero. Since the P-value is less than .01, we reject the null. Looking at the \( t \)-tests, both regression coefficients are significantly different from zero. The \( R^2 \) here is equal to 81.43%.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Sum of Mean</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Analysis of Variance
The Bonferroni approach was used to obtain simultaneous CI's. For an overall 99% level, we’d use 99.67% level for each of the three CI’s.

The prediction interval for age=15 is (10.9734, 29.3024). This means that with 99% confidence, we believe the observed response will be between these two numbers.

The individual $t$-test can be used to assess whether the quadratic term can be dropped. In this case, it cannot (see $t$-test result above). Note that $sx = \frac{X - 15.7778}{5.5006}$. The model, rewritten in terms of $X$ is $E(Y) = -26.32541 + 4.87357X - 0.1184X^2$.

8. KNNL Problem 8.15

It is assumed that there is the same rate of change between the number of copiers serviced and the number of minutes spent on the service call (slope is the same). The parameter
on $X_2$ represents the difference in slopes between the small and large copier models. The regression model is shown below

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>76966</td>
<td>38483</td>
<td>473.94</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>42</td>
<td>3410.32825</td>
<td>81.19829</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>44</td>
<td>80377</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 9.01101

Dependent Mean 76.26667

Adj R-Sq 0.9556

Coef Var 11.81513

Parameter Estimates

| Variable | DF | Estimate | Standard Error | t Value | Pr > |t| 95% Confidence Limits |
|----------|----|----------|----------------|---------|------|-----------------------|
| Intercept| 1  | -0.92247 | 3.09969        | -0.30   | 0.7675| -7.17789              |
| x1       | 1  | 15.04614 | 0.49000        | 30.71   | <.0001| 14.05728              |
| x2       | 1  | 0.75872  | 2.77986        | 0.27    | 0.7862| -4.85125              |

The stated model is $E(Y) = -0.92247 + 15.05X_1 + .75872X_2$. This implies that for the same number of copiers, the service call lasts .75872 minutes longer for the small compared to the large copiers. It should also be noted that as the number of copiers serviced increases, the longer the number of minutes spent on the service call. By including this in the model, we reduce the unexplained variation and thus make our comparison between large and small copiers more precise. It also allows us to make the large versus small comparison at the same level of copiers serviced.

The plot below shows the residuals versus $X_1X_2$. There does not appear to be any linear trend and the residuals are scattered randomly. I would not expect this to be a helpful predictor.

/* SAS Codes for Q1 */
data hw5q1;
    infile 'd:\nobackup\tmp\ch06pr15.txt';
    input y x1 x2 x3;
    proc corr data=hw5q1;
        var x1 x2 x3 y;
    run;

data newobs;
    x1=55; x2=50; x3=2.8;
    output;
run;

data thw5q1;
    set hw5q1 newobs;
run;

proc reg data=thw5q1;
    model y=x1 x2 x3/cli;
    output out=q1out r=resid p=pred;
run;

symbol v=circle i=none c=black;
proc gplot data=q1out;
    plot resid*(pred x1 x2 x3);
run;

proc univariate data=q1out;
    histogram resid / normal kernel (L=2);
    qqplot resid / normal (L=1 mu=est sigma=est);
run; quit;

/* SAS Codes for Q2 & Q3 */
proc reg data=hw5q1;
    model y=x2 x1 x3/ss1;
    Q3: test x2, x3;
run; quit;

/* SAS Codes for Q4 */
proc reg data=hw5q1;
    model y=x1 x2 x3/pcorr1 pcorr2;
run;

proc reg data=hw5q1;
    model y=x2 x1 x3/pcorr1 pcorr2;
run; quit;
/* SAS Codes for Q5 */
proc reg data=hw5q1;
   model y=x1 x2 x3/stb;
run; quit;

/* SAS Codes for Q7 */
goptions colors=(none);
data hw5q7;
infile 'D:\nobackup\tmp\ch08pr06.txt';
input y x;
sx = x;
data newobs;
input x y;
sx = x;
cards;
  10 .
  15 .
  20 .;

data thw5q7;
   set hw5q7 newobs;
   proc standard data=thw5q7 out=nhw5q7 mean=0 std=1;
      var sx;
data nhw5q7;
   set nhw5q7;
   sx2 = sx*sx;
   proc reg data=nhw5q7;
      model y=sx sx2 / clm alpha=.00333333333333333333;
      id x y;
      output out=q7out p=pred;
run;
   symbol1 v=circle i=none;
   symbol2 v=square i=join;
   proc sort data=q7out; by sx;
   proc gplot data=q7out;
      plot y*sx pred*sx / overlay;
   run;
   proc reg data=nhw5q7;
      model y=sx sx2/ cli alpha=.01;
id x;
Q7e: test sx2;
run;

proc univariate data=hw5q7;
  var x;
  output out=xstat mean=mean std=std;
run; quit;

/* SAS Codes for Q8 */
goptions colors=(none);
data hw5q8a;
  infile 'D:\nobackup\tmp\ch01pr20.txt';
  input y x1;
data hw5q8;
  set hw5q8a;
  infile 'D:\nobackup\tmp\ch08pr15.txt';
  input x2;
  x12 = x1*x2;

proc reg data=hw5q8;
  model y=x1 x2 / clb alpha=.05;
  output out=q8out r=resid;
run;

proc sort data=q8out; by x12;
symbol1 v=circle i=none;
proc gplot data=q8out;
  plot resid*x12/vref=0;
run;quit;