1. Consider the given data set that describes the relationship between “velocity” of an enzymatic reaction \( V \) and the substrate concentration \( C \). You are asked to investigate whether this linear transformation results in a good or poor fit by doing the following steps:

   a. Generate a scatterplot of \( V \) vs \( C \). Comment on the shape.

   There is a nonlinear relationship between \( V \) and \( C \). The spread in the observations at each level of \( C \) appears reasonably constant.

   ![Scatterplot of V vs C](image1)

   b. Define new variables for \( 1/V \) and \( 1/C \) in SAS and generate a scatterplot. Does the fit appear linear? Does there appear to be any violation of assumptions?

   ![Scatterplot of 1/V vs 1/C](image2)

   While there does appear to be a linear relationship, the spread in the observations at each level of \( 1/C \) is no longer constant. Also, the small values of \( C \) are now extremely large values of \( 1/C \). Combine this with a larger variance and they may have a strong influence on the fit.
c. Is the distribution of $1/C$ different than $C$? Are there any points that may be more influential in determining the fit?

See description above. The linear regression may try to better fit those large values of $1/C$ at the expense of those observations with smaller $1/C$ values.

d. Determine the least squares regression line for $1/V$ vs $1/C$. Save the residuals and predicted values. Does the residual plot suggest any problems?

Yes. The residual plot suggests nonconstant variance. One can also see that the residuals for small values of $1/C$ are either both positive or negative suggesting a poor fit.

e. Convert this regression line back into the original nonlinear model and plot the predicted curve on a scatterplot of $V$ vs $C$. Comment on the fit. To generate the predicted curve, simply take the predicted values from the regression model and “re-inverse” them. For example, consider new1 to be the data set containing the residuals and predicted values.

The curve appears to overpredict for low concentrations and underpredict at large values of $C$. Other estimation approaches, like nonlinear regression would be better to use here.
2. KNNL Problem 4.12. This can be done in SAS or by hand. To fit a model with no intercept in SAS, you need to add the \texttt{NOINT} option on \texttt{MODEL} statement line.

Using $b_1 = \sum XY / \sum X^2$ we get $\hat{Y} = 18.0283X$.

From the scatterplot, this appears to provide a very good fit. Management is interested in testing whether $H_0 : \beta_1 = 17.5$ vs the alternative $H_a : \beta_1 \neq 17.5$. In order to compute the test statistic, we need the $s(b_1)$. You can show that $\text{MSE} = 20.3113$ and $s(b_1) = .07948$. Thus the test statistic is $t^* = (18.0283 - 17.5)/.07948 = 6.65$. Since $t(.99, 11) = 2.718$, we reject and conclude an adjustment needs to be made. Since $\hat{Y} = 18.0283X + e$, the standard error of a predicted value at $X = 10$ is $\sqrt{(10 \times .07948)^2 + 20.3113} = 4.576$. Thus the prediction interval is $180.283 \pm 2.718(4.576) = (167.845, 192.721)$.

3. KNNL Problem 4.21

The covariance matrix of $b$ is $(X'X)^{-1} \sigma^2$. Since $\bar{X}$ is 0, the off-diagonal elements of this matrix will be zero. The parameter estimates are uncorrelated. Because $b$ is bivariate normal this implies the estimates are independent. Independence of the confidence interval implies that there is no correlation between the lower endpoints. If the variance were known
so the amount added and subtracted from the mean were constant, then the intervals would be independent. In this case, the variance is estimated from the data so there is a slight amount of correlation.

4. KNNL Problem 4.25
In this case $\hat{Y}_h = b_1 X_h$ so the variance will be $X_h^2 s^2(b_1) = X_h^2 \frac{MSE}{\sum X_i^2}$.

5. KNNL Problem 6.2

\[ a) X = \begin{bmatrix} X_{11} & X_{12} & X_{11}^2 \\ X_{21} & X_{22} & X_{21}^2 \\ X_{31} & X_{32} & X_{31}^2 \\ X_{41} & X_{42} & X_{41}^2 \\ X_{51} & X_{52} & X_{51}^2 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \]

\[ b) X = \begin{bmatrix} 1 & X_{11} & \log X_{12} \\ 1 & X_{21} & \log X_{22} \\ 1 & X_{31} & \log X_{32} \\ 1 & X_{41} & \log X_{42} \\ 1 & X_{51} & \log X_{52} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \]

6. (SAS Exercise) Use the brand preference data described in KNNL Problem 6.5. Run the linear regression with moisture and sweetness of the product as the explanatory variables and degree of liking as the response variable.

a. Summarize the regression results by giving the fitted regression equation, the value of $R^2$.

The output is shown below. The estimated regression equation is $\hat{Y} = 37.65 + 4.425 X_1 + 4.375 X_2$. The $R^2 = .9521$.

b. State the results of the significance test for the null hypothesis that the two regression coefficients for the explanatory variables are all zero (give null and alternative hypotheses, test statistic with degrees of freedom, $p$-value, and a brief conclusion in words).

The $p$-value of the F-test is < .0001. This suggests that the combination of moisture and sweetness is helpful in predicting the brandness liking. This $p$-value is associated with the null hypothesis that $\beta_1 = \beta_2 = 0$ versus the alternative that at least one of them is not equal to zero. The test statistic is 129.08 and the degrees of freedom are 2 and 13.

c. Describe the results of the hypothesis tests for the individual regression coefficients (give null and alternative hypotheses, test statistic with degrees of freedom, $p$-value, and a brief conclusion in words).

The $p$-values are shown below. Both $p$-values are less than 0.0001, this implies that we would reject the hypothesis that $\beta_i = 0$ for the alternative $\beta_i \neq 0$. The test statistics are 14.70 and 6.50, the degrees of freedom are 13. The conclusion would be
that either variable after fitting/adjusting for the other variable significantly helps predict brandness liking.

d. Give separate 95% confidence intervals for the regression coefficients of sweetness and moisture. What is the relationship between these confidence intervals and the above hypothesis results?

The 95% CIs are shown below. Because 0 is not included in either confidence interval, this implies that we would reject the hypothesis that $\beta_i = 0$ for the alternative $\beta_i \neq 0$. This can also be seen from the P-values.

Dependent Variable: liking

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>1872.70000</td>
<td>936.35000</td>
<td>129.08</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>13</td>
<td>94.30000</td>
<td>7.25385</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>15</td>
<td>1967.00000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE: 2.69330
Dependent Mean: 81.75000
Coeff Var: 3.29455

Parameter Estimates

| Variable   | DF | Estimate | Error | t Value | Pr > |t| |
|------------|----|----------|-------|---------|------|------|
| Intercept  | 1  | 37.65000 | 2.99610| 12.57   | <.0001|
| moisture   | 1  | 4.42500  | 0.30112| 14.70   | <.0001|
| sweetness  | 1  | 4.37500  | 0.67332| 6.50    | <.0001|

Parameter Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>31.17731 44.12269</td>
</tr>
<tr>
<td>moisture</td>
<td>1</td>
<td>3.77447  5.07553</td>
</tr>
<tr>
<td>sweetness</td>
<td>1</td>
<td>2.92037  5.82963</td>
</tr>
</tbody>
</table>

NOTE: If you look at the residual plots, to make sure the assumptions are reasonable, there is some evidence of model misfit (a curvilinear relationship between the residuals and predicted values). This may indicate there is an interaction. I fit the model with interaction and the interaction was not significant.

/* SAS Codes for HW4Q1 */
data hw4q1;
  input conc vel;
datalines;
  0.02 76
  0.02 47
  0.06 97
  0.06 107
  0.11 123
  0.11 139
  0.22 159
  0.22 152
*/
options nocenter l=75;
symbol v=circle i=sm70 c=black;
proc gplot data=hw4q1;
    plot vel*conc;
run;

data hw4q1; set hw4q1;
    cinv = 1/conc; vinv = 1/vel;
proc gplot data=hw4q1;
    plot vinv*cinv;
run;

proc reg data=hw4q1;
    model vinv = cinv;
    output out=q1out r=resid p=pred;
run;
symbol v=circle i=none c=black;
proc gplot data=q1out;
    plot resid*cinv/vref=0;
run;

data q1out; set q1out;
    pinv = 1/pred;
symbol1 v=circle i=none c=black;
symbol2 v=plus i=sm5 c=black;
proc gplot data=q1out;
    plot vel*conc=1 pinv*conc=2/vref=0 overlay;
run; quit;

/* SAS Codes for HW4Q2 */
data hw4q2;
    infile 'd:\nobackup\tmp\CH04PR12.txt';
    input typo galley;
proc reg data=hw4q2;
    model typo = galley/noint;
    output out=q2out p=pred;
run;
symbol1 v=circle i=none c=black;
symbol2 v=none i=join c=black;
proc gplot data=q2out;
    plot typo*galley=1 pred*galley=2/overlay;
run; quit;

/* SAS Codes for HW4Q6 */
data hw4q6;
    infile 'd:\nobackup\tmp\CH06PR05.txt';
    input liking moisture sweetness;
proc reg data=hw4q6;
    model liking = moisture sweetness/clb;
run; quit;