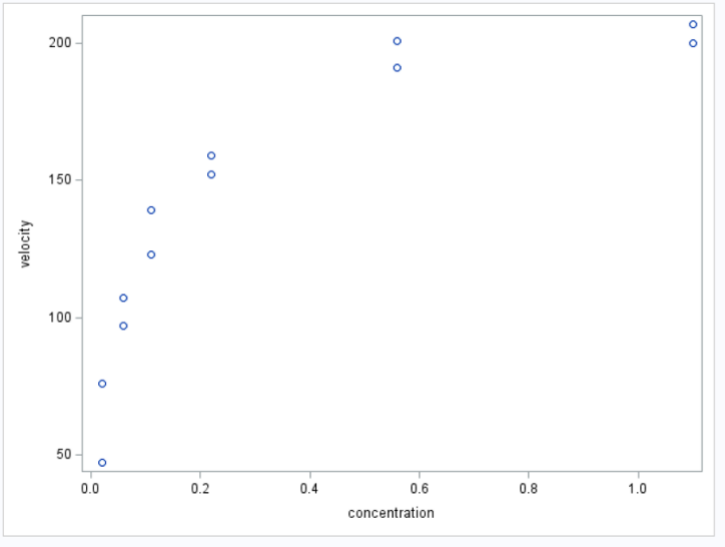
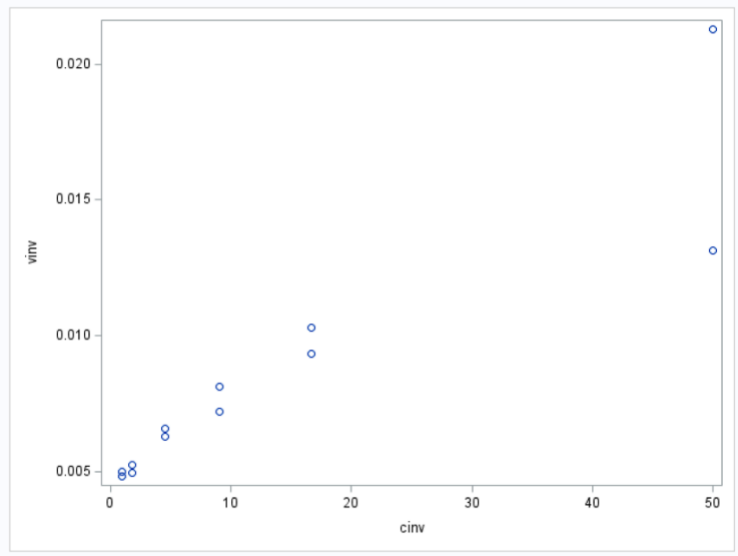
1. Consider the following data set that describes the relationship between "velocity" of an enzymatic reaction (V) and the substrate concentration (C). You are asked to investigate whether this linear transformation results in a good or poor fit by doing the following steps:

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



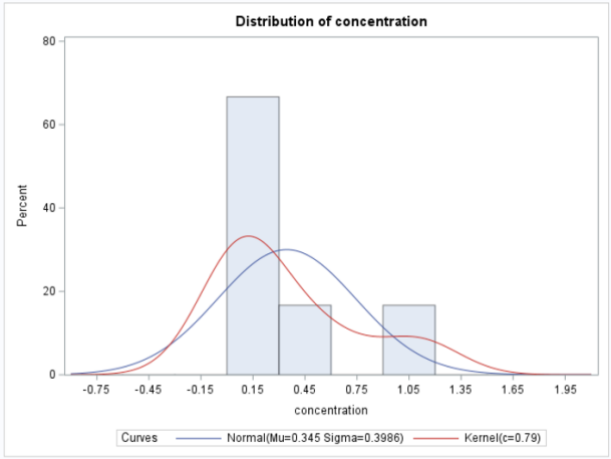
This plot doesn't pass the "eyeball test" for linear regression, as its clear that the data trend quickly departs from linear behavior around a concentration level of 0.2.

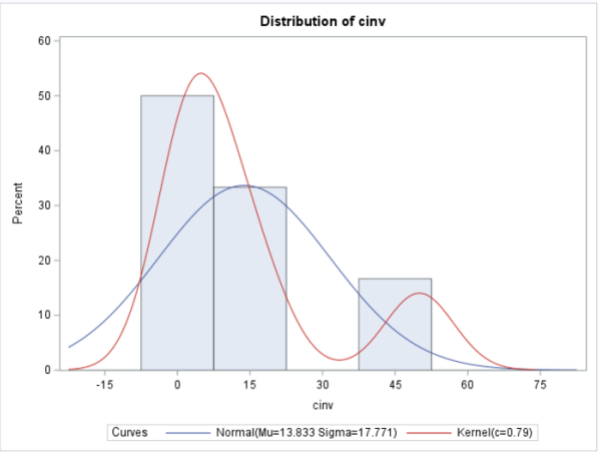
b. Define new variables for $\frac{1}{V}$ and $\frac{1}{C}$ in SAS and generate a scatterplot.



Now linear regression seems much more reasonable, though there are two significant outliers at $X = 50$ that could affect the appropriateness of the fit.

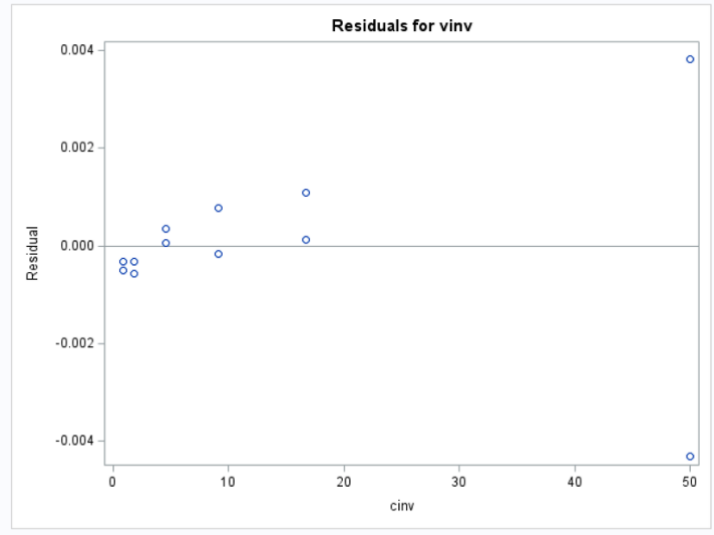
c. Is the distribution of $\frac{1}{C}$ different than C? Are there any points that may be more influential in determining the fit?





Their distributions appear similar, as evidenced by the histograms. Looking at the scatter plots above in parts a and b, it's evident that $C = 1.1$ and $\frac{1}{C} = 50$ are very influential points.

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



Number of Observations Read	12
Number of Observations Used	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.00021232	0.00021232	59.30	<.0001
Error	10	0.00003581	0.00000358		
Corrected Total	11	0.00024813			

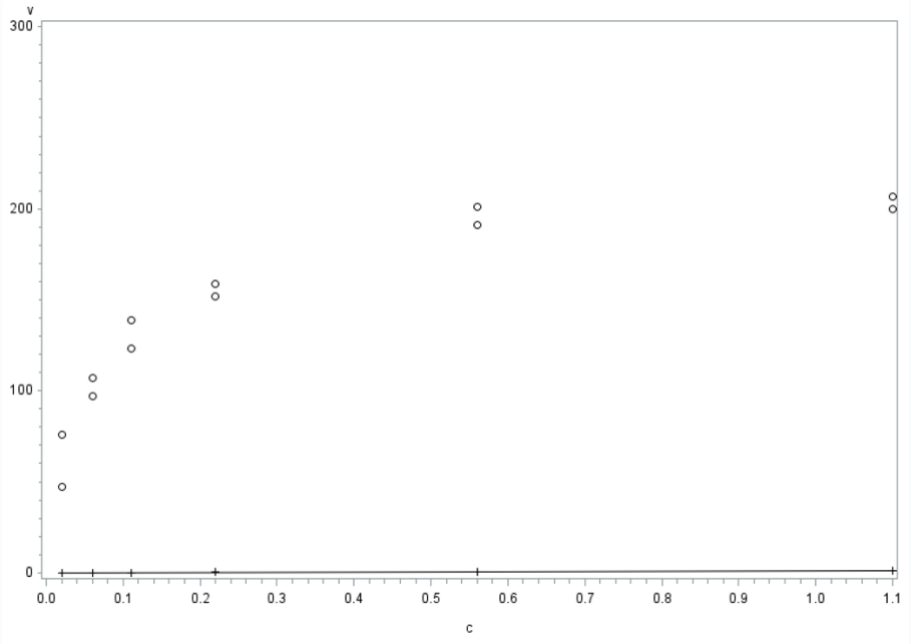
Root MSE	0.00189	R-Square	0.8557
Dependent Mean	0.00853	Adj R-Sq	0.8413
Coeff Var	22.19144		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.00511	0.00070400	7.25	<.0001
cinv	1	0.00024722	0.00003210	7.70	<.0001

The residuals increase in a conical fashion as $\frac{1}{C}$ increases, so it appears that the assumption of constant variance for linear regression is violated.

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.



A regression line seems appropriate for concentration levels up to 0.21, but quickly becomes nonviable for values beyond that threshold.

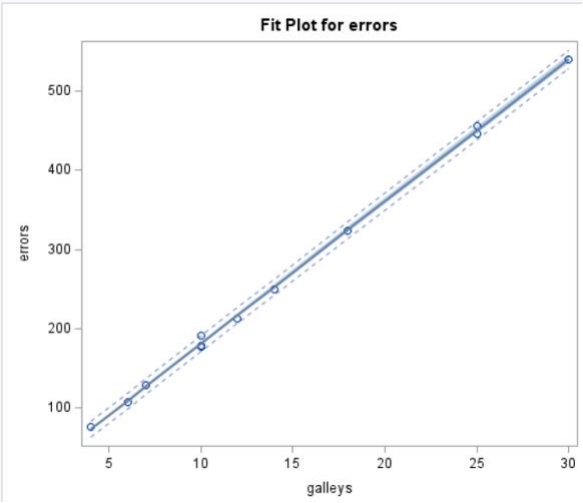
2. Typographical errors. Shown below are the number of galleys for a manuscript (X) and the total dollar cost of correcting typographical errors (Y) in a random sample of recent orders handled by a firm specializing in technical manuscripts. Since Y involves variable costs only, an analyst wished to determine whether regression-through-the-origin model (4.10) is appropriate for studying the relation between the two variables.

a. Fit regression model (4.10) and state the estimated regression function.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
galleys	1	18.02830	0.07948	226.82	<.0001

The estimated regression function is $\hat{Y}_i = 18.02830X_i$.

b. Plot the estimated regression function and the data. Does a linear regression function through the origin appear to provide a good fit here? Comment.



Yes, the data exhibits very linear behavior, as evidenced by the line-of-fit.

c. In estimating costs of handling prospective orders, management has used a standard of \$17.50 per galley for the cost of correcting typographical errors. Test whether or not this standard should be revised; use $\alpha = .02$. State the alternatives, decision rule, and conclusion.

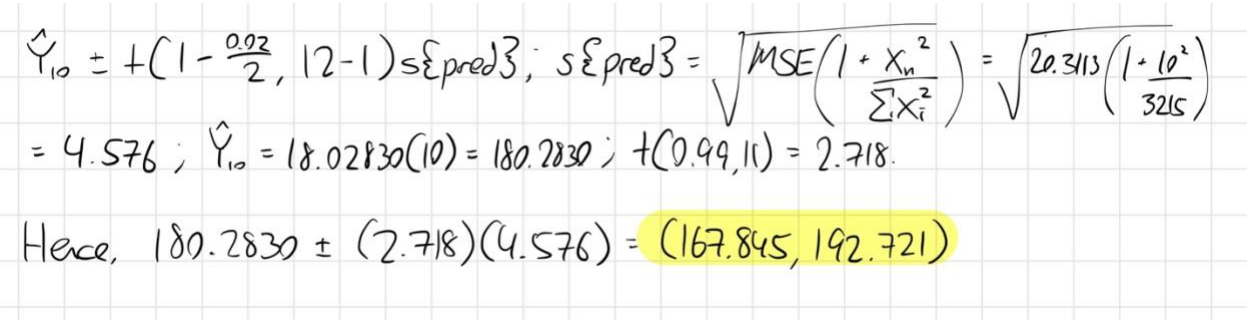
Let $H_0: \beta_1 = 17.50$ and $H_A: \beta_1 \neq 17.50$. Then $t^* = \frac{b_1 - \beta_1}{s\{b_1\}} = \frac{18.02830 - 17.5}{0.07948}$

$= 6.647$.

We reject H_0 if $|t^*| > t\left(1 - \frac{0.02}{2}, 12-1\right); t(0.99, 11) = 2.718$.

Since $6.647 > 2.718$, we conclude H_A; 17.50 per galley is not appropriate and should be revised.

d. Obtain a prediction interval for the correction cost on a forthcoming job involving 10 galleys. Use a confidence coefficient of 98 percent.



3. When the predictor variable is so coded that $\bar{X} = 0$ and the normal error regression model (2.1) applies, are b_0 and b_1 independent? Are the joint confidence intervals for β_0 and β_1 then independent?

When $\bar{X} = 0$, $b_0 = \bar{Y}$ and b_1's value is indeterminate without knowing any more information, but we know $\bar{Y}$ is needed to find b_1, so we know b_0 and b_1 are not independent because b_0 influences b_1.

Likewise, the joint confidence intervals for β_0 and β_1 require $s\{b_0\}$ and $s\{b_1\}$, respectively, which both require knowing the MSE. Thus these confidence intervals are not independent either.

4. Derive the formula for $s^2\{\hat{Y}_h\}$ given in Table 4.1 for linear regression through the origin.

We desire $s^2\{\hat{Y}_h\} = \frac{X_h^2 MSE}{\Sigma X_i^2}$ by Table 4.1. For regression through the origin, we know $\hat{Y}_h = b_1 X_h$. Then $\sigma^2\{\hat{Y}_h\} = \sigma^2\{b_1 X_h\}$. Because X_h is fixed, by definition of variance we have $\sigma^2\{b_1 X_h\} = X_h^2 \sigma^2\{b_1\}$.

Since $b_1 = \frac{\Sigma X_i Y_i}{\Sigma X_i^2}$ for regression through the origin, we know $\sigma^2\{b_1\} = \frac{\sigma^2}{\Sigma X_i^2}$. Because MSE is an unbiased estimator of σ^2, we also have $s^2\{b_1\} = \frac{MSE}{\Sigma X_i^2}$. So, we conclude,

$$s^2\{\hat{Y}_h\} = s^2\{b_1 X_h\} = X_h^2 s^2\{b_1\} = X_h^2 \frac{MSE}{\Sigma X_i^2}.$$

5. Set up the X matrix and β vector for each of the following regression models (assume $i = 1, \ldots, 5$.

a. $Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \epsilon_i$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & X_{13}^2 \\ X_{21} & X_{22} & X_{23}^2 \\ X_{31} & X_{32} & X_{33}^2 \\ X_{41} & X_{42} & X_{43}^2 \\ X_{51} & X_{52} & X_{53}^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix}$$

b. $\sqrt{Y_i} = \beta_0 + \beta_1 X_{i1} + \beta_2 \log_{10} X_{i2} + \epsilon_i$

$$\begin{bmatrix} \sqrt{Y_1} \\ \sqrt{Y_2} \\ \sqrt{Y_3} \\ \sqrt{Y_4} \\ \sqrt{Y_5} \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \log_{10} X_{12} \\ 1 & X_{21} & \log_{10} X_{22} \\ 1 & X_{31} & \log_{10} X_{23} \\ 1 & X_{41} & \log_{10} X_{42} \\ 1 & X_{51} & \log_{10} X_{52} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix}$$

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 and R^2.

Root MSE	2.69330	R-Square	0.9521
Dependent Mean	81.75000	Adj R-Sq	0.9447
Coeff Var	3.29455		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard 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

$R^2 = 0.9521$ and the fitted regression equation is $\hat{Y}_i = 37.65 + 4.425X_{i1} + 4.375X_{i2}$.

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).

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1872.70000	936.35000	129.08	<.0001
Error	13	94.30000	7.25385		
Corrected Total	15	1967.00000			

The null and alternative hypotheses are, $H_0: \beta_1 = \beta_2 = 0$; $H_A: \beta_1 \neq 0$ and $\beta_2 \neq 0$, respectively. The test stat is $F^* = 129.08$ with 13 degrees of freedom and the p-value is < 0.0001. Hence, we conclude H_A, that is, there is a relationship between liking the product and the two associated predictors: moisture and sweetness.

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).

Parameter Estimates

Variable	DF	Parameter Estimate	Standard 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

For β_0, $H_A: \beta_0 = 0$ and $H_A: \beta_0 \neq 0$. Test stat is $t^* = 12.57$ with 13 degrees of freedom and the p-value is < 0.0001. We conclude $H_A: \beta_0 \neq 0$, so there is statistical evidence to suggest that the intercept is not equal to 0.

For β_1, $H_A: \beta_1 = 0$ and $H_A: \beta_1 \neq 0$. Test stat is $t^* = 14.70$ with 13 degrees of freedom and the p-value is < 0.0001. We conclude $H_A: \beta_1 \neq 0$, so there is statistical evidence to suggest that there is a linear relationship between moisture and degree of liking.

For β_2, $H_A: \beta_2 = 0$ and $H_A: \beta_2 \neq 0$. Test stat is $t^* = 6.50$ with 13 degrees of freedom and the p-value is < 0.0001. We conclude $H_A: \beta_2 \neq 0$, so there is statistical evidence to suggest that there is a linear relationship between sweetness and degree of 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?

Parameter Estimates

Variable	DF	Parameter Estimate	Standard 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

For moisture, $b_1 \pm t\left(1 - \frac{0.05}{2}, 15 - 2\right) s\{b_1\}$. Then $b_1 = 4.425$, $t(0.975, 13) = 2.160$, and $s\{b_1\} = 0.30112$. So $4.425 \pm (2.160)(0.30112) =$ $(3.775, 5.075)$.

For sweetness, $b_2 \pm t\left(1 - \frac{0.05}{2}, 15 - 2\right) s\{b_2\}$. Then $b_2 = 4.375$, $t(0.975, 13) = 2.160$, and $s\{b_2\} = 0.67332$. So $4.375 \pm (2.160)(0.67332) =$ $(2.921, 5.829)$.

Neither 95% confidence interval contains 0, which supports the conclusion from the above hypothesis test that neither β_1 nor β_2 are equal to 0.