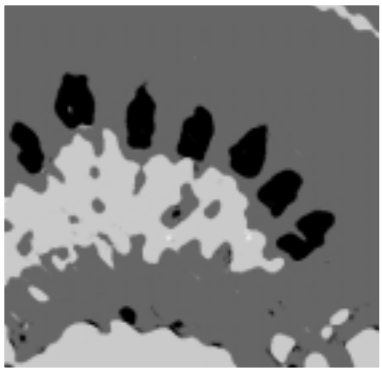


INTERDISCIPLINARY LIVELY APPLICATIONS PROJECT

MATERIALS

1. Problem Statement (four parts, including one optional part)
2. Solutions
3. Notes for the Instructor

Computing Requirements: Scientific Calculator (optional)

Keeping the Drinking Water Supply Safe

Interdisciplinary Lively Applications Project

Title: Keeping the Drinking Water Supply Safe

Authors: Laura J. Steinberg, Civil and Environmental Engineering,
Tulane University
John R. Liukkonen, Mathematics, Tulane University

Editors: David C. Arney
William Fox
Brian Winkel

Math Subjects: Elementary linear algebra, Calculus of
several variables

Prerequisite Skills:
Solving systems of lenear equations
Elementary linear algebra
Plotting and analyzing plots

Engineering Concepts Examined:
Sorption of contaminants
Suspended solids transport in natural waters
Stage-discharge curves
Unit operations for water treatment
Drinking water quality standards

Computing Requirements:
A computer package such as *Matlab, Maple, Derive,* or *S-Plus*
which easily carries out:
- elementary linear algebra operations
- plots one vector against another

Interdisciplinary Lively Applications Project is funded
By the National Science Foundation, Directorate of
Education and Human Resources Division of
Undergraduate Education, NSF Grant #9455980

NSF Initiative:
Mathematics Sciences and Their Applications Throughout the Curriculum
(CCD-MATH)

REGULATORY BACKGROUND

In order to protect public health, the United States Environmental Protection Agency (USEPA) regulates the concentrations of many contaminants in drinking water. For example, turbidity (visible material in suspension) in the water supply is limited to 1 Turbidity Unit. Turbidity is caused by particles or organic matter which run off into the water from adjacent land, reside in the bottom of rivers, or are formed from dead or live microorganisms in the river. Not only are the particles unhealthy themselves, but they also carry many organic contaminants attached to them. Concentrations of these organic compounds are also regulated. For example, polychlorinated biphenyl (PCB) cannot exceed 0.5 micrograms/liter in drinking water. In order to ensure that contaminated drinking water is not consumed by the public, health officials test the water at regular intervals and whenever there is a suspicion of contamination.

THE SITUATION

The towns of Freeburg and Cajunville use the Bluewater River as a source of drinking water. Freeburg, the county seat, has a water treatment plant which removes particles from the water using chemical treatment followed by sedimentation. Cajunville, being a poor farming community, only settles its water and can produce potable water only when the concentration of suspended particles is less than 40 mg/l.

Elevated levels of polychlorinated biphenyl (PCB), a cancer-causing compound, have recently been discovered in the Bluewater River just upstream of the Freeburg water treatment plant intake. It is believed that they were discharged into the river by CARCINOGENS-R-US Inc. The residents of Freeburg are very concerned about drinking this water.

State health officials studied the problem for four years. They found that Freeburg can remove all of PCB which is *attached* to particles in the water by their sedimentation basin. However, it cannot remove any PCB which is *dissolved* in the water. Cajunville, being upstream of the discharge, is unaffected by the PCB but cannot provide potable water when particle concentrations exceed 40 mg/l. We know from physical chemistry that the percentage of PCB which is dissolved in the water is directly related to the particle concentration in the water. So, when the particle concentration is high, the percentage of PCB which is dissolved is low. In this case, Freeburg can remove most of the contaminant in its sedimentation basins. However, at the same time, Cajunville cannot drink the water because it cannot remove the particles.

We will investigate methods of statistical analysis which will permit us to advise Freeburg and Cajunville about the potability of their water under varying hydrologic conditions. In Part 1, we introduce Linear Regression Models, Part 2 discusses Transformations to improve the estimates made with naive linear regression models, and in Part 3 we apply these techniques to the water quality problems of Freeburg and Cajunville. Finally, in Part 4 we provide additional explanation about transformations.

PART 1: LINEAR REGRESSION

Suppose we have observations on N cases involving a predictor variable x and a response variable y. These observations take the form $(x_1,y_1),\ldots,(x_N,y_N)$. If we believe that the variable y is roughly a linear function of the variable x, then we may form the model

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ &\vdots \\ y_N &= \beta_0 + \beta_1 x_N + \varepsilon_N \end{aligned} \tag{1}$$

where the x_i's and y_i's are the observed predictors and responses, β_0 and β_1 are unknown parameters we wish to choose well, and the ε_is are unobserved

We may save space and effort by rewriting this in vector form as

$$\vec{y} = \beta_0 \vec{1} + \beta_1 \vec{x} + \vec{\varepsilon} \tag{2}$$

Here $\vec{y}$, $\vec{x}$, and $\vec{\varepsilon}$ are the column N-vectors formed from the corresponding entities in (1) and $\vec{1}$ is the column N-vector of 1's.When we choose β_0 and β_1 we choose a fitted linear summary $\hat{y} = \beta_0 \vec{1} + \beta_1 \vec{x}$ for $\vec{y}$; $\vec{\varepsilon}$ then represents the vector of "errors" in our summary.

The question is, how to choose the parameters β_0 and β_1. The most popular method is to use the **least squares criterion**: i.e., choose β_0 and β_1 to minimize the distance squared $\left\| \vec{y} - \left(\beta_0 \vec{1} + \beta_1 \vec{x} \right) \right\|^2$. Geometrically, this situation can be represented by **Figure 1**. The figure is drawn in $\mathbf{R}^3$ but represents a situation in $\mathbf{R}^N$.

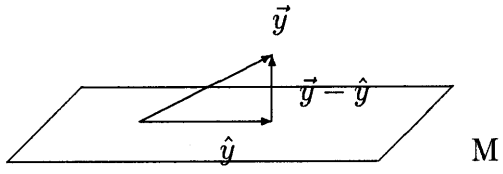


Figure 1.

Here the set of possible fitted linear summaries $\hat{y} = \beta_0 \vec{1} + \beta_1 \vec{x}$ is the plane M, and the problem is to minimize the distance from the vector $\vec{y}$ to this plane. The picture makes clear that to choose the "best" model in the least squares sense, we need to choose β_0 and β_1 so that $\hat{y} = \beta_0 \vec{1} + \beta_1 \vec{x}$ is the orthogonal projection of $\vec{y}$ onto M; i.e., so that

$\vec{y} - \hat{y}$ is perpendicular to *M*. The $\vec{y}$ depicted in the figure is in fact this orthogonal projection. For $\hat{y}$ to be this projection we need the dot product of $\vec{y} - \hat{y}$ with both $\vec{1}$ and $\vec{x}$ to be 0; i.e.,

$$\begin{aligned} \vec{1} \cdot \left(\vec{y} - \left(\beta_0 \vec{1} + \beta_1 \vec{x} \right) \right) &= 0 \\ \vec{x} \cdot \left(\vec{y} - \left(\beta_0 \vec{1} + \beta_1 \vec{x} \right) \right) &= 0 \end{aligned} \tag{3}$$

or in coordinate terms

$$\begin{aligned} n\beta_0 + \left(\Sigma x \right) \beta_1 &= \Sigma y \\ \left(\Sigma x \right) \beta_0 + \left(\Sigma x^2 \right) \beta_1 &= \Sigma xy \end{aligned} \tag{4}$$

These last equations are called the normal equations, and their solutions provide the "least squares coefficients" $\hat{\beta}_0$ and $\hat{\beta}_1$. From now on we will let $\hat{y} = \hat{\beta}_0 \vec{1} + \hat{\beta}_1 \vec{x}$ be the optimal linear summary, or "fitted model", and the error vector, or residual vector will then be $e = \vec{y} - \hat{y}$.

The length of the residual vector has an interesting use. If we let

$$s = \frac{\|e\|}{\sqrt{N-2}} = \sqrt{\frac{\Sigma_i e_i^2}{N-2}} \tag{5}$$

then *s* is a standard measure of the spread of *y* about the least squares line. For large well behaved data sets, 95% of the *y* values should fall within 2*s* of the least squares line. For more detail on this you will need to study a course in statistics covering linear regression.

REQUIREMENT 1.

Consider the data in **Table 1** for the the variables *x* and *y*. Write down the normal equations for this data set and solve them. Find $\hat{y}$ and the residual vector *e*. What is *s*? Plot *e* against $\hat{y}$ and *e* against *x*. What observations can you make about these plots?

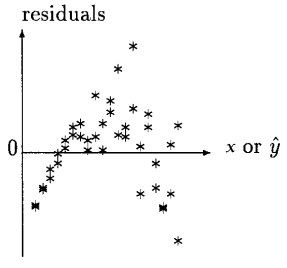
Table 1.

x	1	2	4	5	7	8	9	10
y	–1.3	1.7	6.7	10.1	8.8	11.0	16.1	12.8

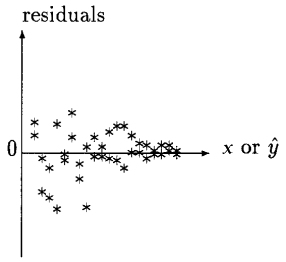
PART 2: TRANSFORMATIONS

The above methods work very well when we simply want a linear summary model. But when we assume that the true state of nature consists of y being a function of x with added noise, the situation changes somewhat, even in the case that y is a linear function of x with added noise. When we formally assume that the ε_i terms are random noise, then classical theory (the Gauss Markov Theorem) tells us that the least squares method described above is the optimal procedure when y is a linear function of x **plus evenly scattered noise**. But if y is a linear function of x plus noise **the magnitude of whose scatter changes** in some systematic fashion, then the normal equations yield **suboptimal** estimates. Thus, when we determine that the magnitude of the scatter varies in some systematic fashion, we need to transform y or x to bring the residual terms into closer conformity with the Gauss-Markov assumptions of random evenly scattered noise.

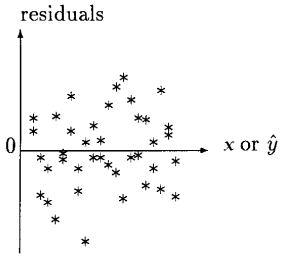
We need tools to tell us when such model change is necessary, and what change to make in such a case. In practice we proceed graphically: we look at a scatterplot of the residuals versus x or $\hat{y}$. We judge from the plot whether the error terms have an even random scatter about 0, or whether there is some clear dependence of that scatter on x or $\hat{y}$. From the same plot we can also check whether there is remaining higher order, curved dependence of the residuals on x or $\hat{y}$. There will be no further **linear** dependence of the actual residuals (as opposed to the magnitude of scatter) on x or $\hat{y}$ since such linear dependence has been aborbed into the model. Below are examples of such scatterplots. All the depicted plots except the one labeled "Even Scatter" indicate a need to transform.



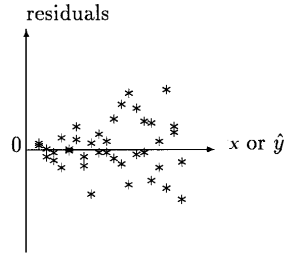
Curvilinear Right Fan



Left Fan



Even Scatter



Right Fan

In case we do need to transform, how do we do it? We usually choose a transformation from the following Box-Cox family of power transformations. That is for the response variable will replace y by y^{λ} where λ is some real number. (If some of the y values are negative and λ is fractional we will translate all y values by a common constant to make them all positive.)

We need to be aware of several additional things here. First, in this context the transformation y^0 is interpreted to be $\ln(y)$. Second, in the interests of getting a model which has some chance of interpretation, we will usually choose λ to be a half-integer. Also, we choose λ in the interval [-2,2], as more extreme powers result in unstable transformations. Finally, we may need to try more than one half integer before we arrive at the best looking residual plot. If we are transforming y we correct a right fan by taking a larger positive power of y than we have, and a left fan by a smaller power (or negative power) of y than we have. If we are transforming x we do just the opposite. To remember this the following "ladder of transformations" is helpful.

Ladder of Transformations

$$y^{-2} \quad y^{-1.5} \; y^{-1} \quad y^{-0.5} \ln(y) \quad y^{+0.5} \; y^{+1} \quad y^{+1.5} \; y^{+2}$$
$$x^{+2} \quad x^{+1.5} \; x^{+1} \quad x^{+0.5} \ln(x) \quad x^{-0.5} \; x^{-1} \quad x^{-1.5} \; x^{-2}$$

← To correct a right fan move in this direction

To correct a left fan move in this direction →

Requirement 1.

Consider the data in **Table 2** for the variables x and y. Carry out a regression of y on x, and plot the residuals versus $\hat{y}$. Is a transformation indicated? If so, select a transformation y^{λ} bringing the residual plot into decent shape.

(As always, consider for λ only half-integers in the interval [-2,2].)

Table 2.

x	5.7	5.8	5.8	5.8	6.4	7.3	8.1	8.8	9.4	9.5
y	5.1	3.4	9.8	6.5	6.8	7.9	7.1	8.5	9.2	11.3
x	11.4	11.5	11.8	13.7	14.5	14.6	14.6	15.0	15.9	16.4
y	9.9	9.7	8.3	13.9	11.0	7.8	15.7	9.4	13.2	13.4
x	17.7	18.9	19.7	19.7	20.2	21.8	23.1	23.6	24.8	26.4
y	13.2	18.7	4.8	16.7	19.2	29.4	6.7	18.8	31.6	15.2
x	26.6	27.3	29.6	30.2	31.8	32.0	32.1	33.3	33.7	34.2
y	19.9	11.4	27.3	10.1	22.2	30.5	32.6	29.9	32.4	16.9

PART 3: CAJUNVILLE AND FREEBURG: A CASE OF THREATENED WATER SUPPLIES?

Now, let's apply our knowledge of regression to the drinking water problems of Freeburg and Cajunville. We want to know if the water is safe for the townspeople to drink. We will make this determination under two conditions: first during a period of little rain when the river is low, and, second, soon after rainshowers from a major storm have created flood conditions in the Bluewater River Valley. We want to determine the concentration of suspended solids in Bluewater River during each of these periods. Using this information, we will be able to advise the townspeople when their drinking water should be tested for contaminants.

We have collected two data sets to help us answer this question.

We will make use of United States Geological Survey (USGS) data which have been collected for the Bluewater River. The USGS has recorded flow rate and suspended solids concentration on 90 occasions. These data are reproduced in **Table 3**.

We also have collected USGS data with simultaneous observations of flow rate and river depth (**Table 4**).

From state health officials we have learned that the total concentration of PCB (the dissolved plus the attached phase of PCB) in Bluewater River is 0.75 μg/1. Also, we have located an equation from physical chemistry which enables us to compute the fraction of total PCB concentration which is in the dissolved form. Using this equation, we can determine how much PCB is dissolved in water at different levels of suspended particle concentrations.

$$\text{fraction dissolved} = 1 - \frac{0.16 \times \text{ suspended solids conc}}{1 + 0.16 \times \text{ suspended solids conc}} \qquad (6)$$

Recall from The Situation section that PCB in the *dissolved* phase cannot be removed from Freeburg's drinking water but that Freeburg's sedimentation basin can remove PCB *attached* to suspended particles.

REQUIREMENT 1.

Using the data in Table 3, solve the normal equations for a linear regression of the response variable, suspended solids concentration, on the predictor variable, flow rate. What are the coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$?
Plot the residuals vs. $\hat{y}$. Discuss the pattern, if any, in this plot.
Does there seem to be non-constant variance, or non-linearity in the residuals? If so, find an appropriate transformation for *y* and/or *x*, and solve the normal equations on the transformed data set. What are the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ for the transformed data set ? Find *s* (see Equation 5 in Part 1).

REQUIREMENT 2.

Actually, flow is not typically measured by the USGS. Instead, the depth of the water (called the "stage" by hydrologists) is measured and regression equations are developed to predict flow as a function of stage. From the data given in Table 4, find a linear regression model and the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ which we can use to predict flow at Freeburg and Cajunville as a function of river stage.

REQUIREMENT 3

a. We suppose the water depth is measured at 14.5 feet during a dry period. Use the two regression equations you developed in Requirements 1 and 2 to estimate the concentration of suspended solids in the water.

b. Your model giving flow as a function of depth is very accurate; however, your model giving sediment concentration as a function of flow is subject to considerable uncertainty. We can be 90% sure that the true sediment concentration is no more than the estimated concentration plus 1.3*s* and also 90% sure that the true concentration is no less than the estimate minus 1.3*s*. Compute these upper and lower bounds of the suspended solids concentrations.

REQUIREMENT 4

a. You are needed to give some advice to public health officials in Cajunville. Compare the upper bound you found for suspended particle concentrations with the 40 mg/l maximum that the Cajunville's water treatment plant can remove. Is your confidence bound greater? If so, what recommendation would you make to the public health officials with regard to testing Cajunville's water when the river depth is 14.5 feet?

b. Suppose that an unusually wet spring creates high water in the river. The depth is measured at 25 feet. What recommendation would you make to the public health officials now?

c. Find the median value of the measured flow from Table 4. Using this as the "normal" flow rate in the river, verify that under normal conditions, Cajunville's water supply is safe.

REQUIREMENT 5

a. Now, consider the case of Freeburg. When the water depth is 14.5 feet, use the *lower* bound on the suspended sediment concentration in Equation 6 to determine the fraction of PCB which is dissolved. Note that we use the lower bound here instead of upper bound because this is the worst case scenario; the lower the suspended solids concentration, the greater the percentage of PCB which is dissolved and untreatable by Freeburg.

b. Use the value that you computed in part a to determine the worst case value for the PCB concentration when the river depth is 14.5 feet. What recommendation would you make to public health officials regarding testing the potable water supply of Freeburg?

c. What recommendation would you make when the river depth reaches 25 feet?

d. Verify that under normal flow conditions, Freeburg's drinking water meets the USEPA's drinking water standards.

REQUIREMENT 6.

Discuss the difference in the hydrologic threats to Cajunville's and Freeburg's water supply.

suspended part.(mg/l)	flow rate (ft^3/sec)	suspended part.(mg/l)	flow rate (ft^3/sec)	suspended part.(mg/l)	flow rate (ft^3/sec)
96	17400	12	25500	8	20000
97	26200	8	23000	12	19400
105	27600	5	21100	8	18900
47	23500	5	18800	8	18100
28	18700	3	16800	7	12800
23	16600	5	15700	9	10700
26	16500	4	15300	7	9980
11	14300	5	15500	5	9260
11	11600	3	14300	7	9050
27	11600	5	13400	38	8500
7	11300	4	13700	19	6400
6	10200	3	14700	5	7070
5	8840	6	14600	6	7130
5	8360	7	13900	5	6880
5	7760	5	14000	5	6520
6	7570	6	15100	5	6520
5	7760	6	15200	6	5100
4	7830	6	14900	6	4040
3	10000	7	14900	5	4890
14	12900	6	15800	5	5000
19	17600	5	16300	7	5320
77	28700	7	17700	21	8290
44	25900	23	18200	11	11100
16	19900	40	19600	11	9690
9	18900	50	28900	8	8030
7	18300	107	36600	14	9400
8	17700	40	31500	27	11800
9	19600	16	27200	13	12100
9	21600	9	23800	9	11600
9	24500	9	21200	7	10300

Table 3

flow rate (ft^3/sec)	river depth (ft)	flow rate (ft^3/sec)	river depth (ft)	flow rate (ft^3/sec)	river depth (ft)
4065	11	31732	24	11095	17
36039	26	28775	23	13081	17
11016	16	36724	26	32229	24
2866	11	7765	13	3398	11
19151	20	34357	25	8624	15
17803	19	11320	16	42663	27
27282	23	15499	18	11066	16
38134	26	44922	28	2515	10
17691	19	33268	25	5083	12
14793	18	42254	27	34201	25
9229	15	18065	20	14809	18
14841	18	4845	12	44810	28
8133	15	17803	19		
17439	19	6579	13		

Table 4

PART 4 (OPTIONAL): WHY WE TRANSFORM—BEST LINEAR UNBIASED ESTIMATORS

The mathematical rationale for our pursuit of residual plots with even scatter is the Gauss-Markov Theorem. To bring this to bear on our situation, we now assume that in model (1) or (2) above the error terms ε_i are independent **random variables**. We assume that for each i the expectation $E(\varepsilon_i)=0$, and we denote the variance $v(\varepsilon_i)$ by v_i The mathematical formulation of **evenly scattered noise** is that **all the v_i's are the same**.

A **linear estimator** of β_1 is any linear combination $\tilde{\beta}_1=\Sigma_i c_i y_i$ of the observed y_i's. The c_i's are allowed to depend on the x_i's, since, as is customary, we take the y_i's as random and the x_i's as "known", not random. The formulas for $E(\tilde{\beta}_1)$ and $V(\tilde{\beta}_1)$, easily derived from basic principles of probability:

$$E(\tilde{\beta}_1)=\Sigma_i c_i(\beta_0+\beta_1 x_i)=(\Sigma_i c_i)\beta_0+(\Sigma_i c_i x_i)\beta_1 \tag{5}$$

$$V(\tilde{\beta}_1)=\Sigma_i c_i^2 v_i \tag{6}$$

We call $\tilde{\beta}_1$ **unbiased** if we can show that $E(\tilde{\beta}_1)=\beta_1$ no matter what the true value of β_1. From (5) we see that this is so, exactly when $\Sigma_i c_i = 0$ and $\Sigma_i c_i x_i= 1$. In the unbiased case the expected squared error of $\tilde{\beta}$ (from its target) is the same as the variance of $\tilde{\beta}_1$. So from among unbiased linear estimators we seek the one with minimum variance. That is, to choose the best linear unbiased estimator, we want to choose the c_i's to minimize $\Sigma_i c_i^2 v_i$ subject to the conditions

$$\Sigma_i c_i=0 \text{ and } \Sigma_i c_i x_i=1. \tag{7}$$

The Gauss-Markov Theorem says that **when all the v_i's are the same**, then the least squares estimators are also the best linear unbiased estimators (i.e., minimum variance linear unbiased.) We also know that when the v_i's vary, then the least squares estimators **are not** the minimum variance linear unbiased estimators. You will be asked below to work through two examples—in the first the error variances are all the same, and in the second the error variances vary by case. You will see that the least squares estimate agrees with the optimal estimate in the first case but not the second.

REQUIREMENT 1.

Consider the data set in Table 1 of Part 1 above. Assuming that all variances $v_i = v$ are the same, then to minimize $v(\tilde{\beta}_1)$ we want to minimize $\Sigma_i c_i^2$. Use Lagrange multipliers to determine the c_i's which minimize $\Sigma_i c_i^2$ subject to conditions (7) above. Compare the unbiased linear estimator $\Sigma_i c_i y_i$ with the least squares estimator $\hat{\beta}_1$.

REQUIREMENT 2.

Now consider the data set in Table 1 of Part 1 above. Assuming that for some unknown positive constant k we have $v_i = kx_i$ for each i (so that the v_i's are varying), we now want to minimize $\Sigma_i c_i^2 x_i$. Determine the c_i's which minimize $\Sigma_i c_i^2 x_i$ subject to conditions (7) above. Compare $\Sigma_i c_i y_i$ with the least squares estimator $\hat{\beta}_1$.

SOLUTIONS

PART 1.

Requirement 1.

n=8, Σx =46, Σx^2 = 340, Σy=65.9, $\Sigma\, xy$=501.9. These are enough to write down the normal equations. We get $\hat{\beta}_0$ =–1.1281, $\hat{\beta}_1$ =1.6288.

We calculate $\hat{y}$ and e and get s=2.2292. **Figure 1** shows a plot of the residuals e versus x.

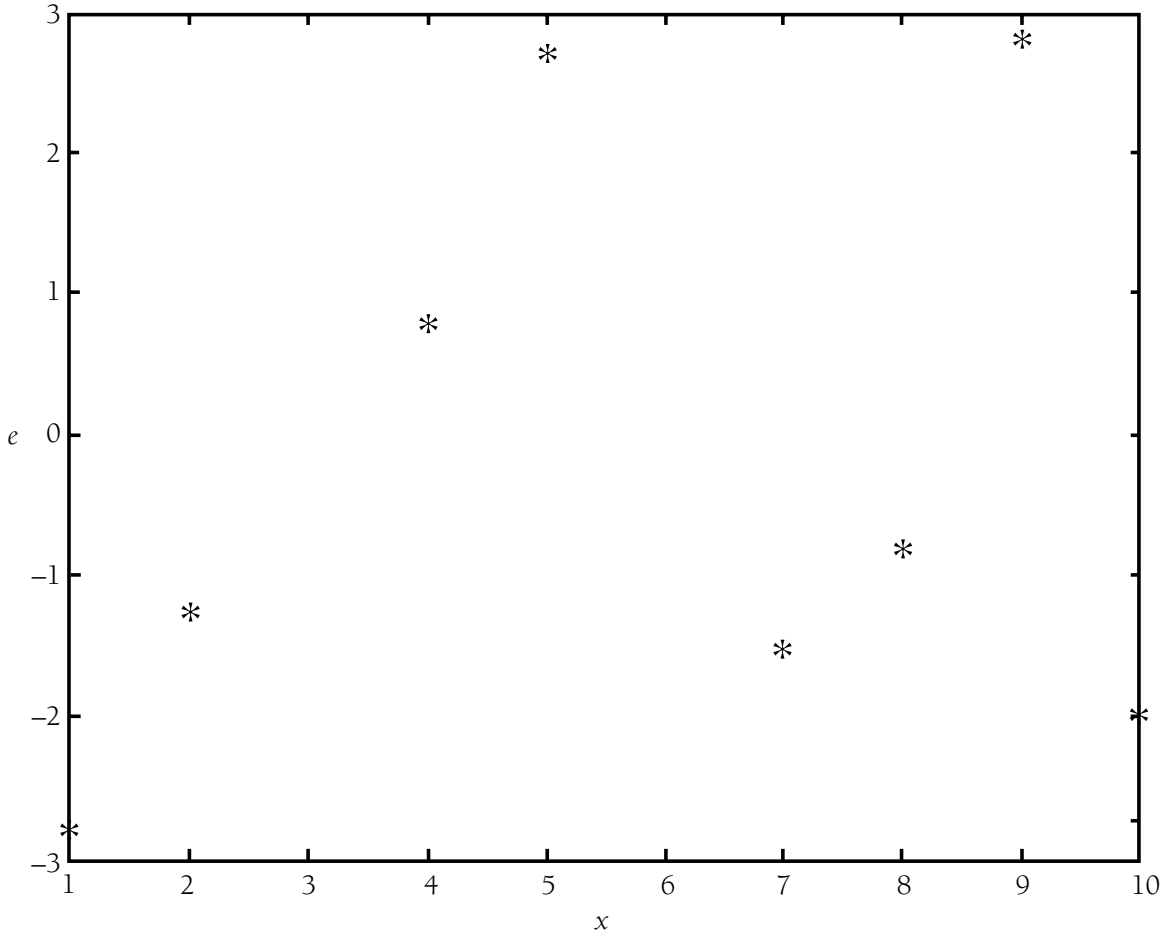


Figure 1.

PART 2.

Requirement 1.

For y untransformed, we get $\hat{\beta}_0$ =-1.968,

$\hat{\beta}_1$ =0.69946 For response variable $(y+15)^{.5}$ we get

$\hat{\beta}_0$ =1.8681, $\hat{\beta}_1$ =0.4666. See **Figure 2.** for residual plot, which shows a right fan. Either $y^{0.5}$ or ln y represents an acceptable transformation.

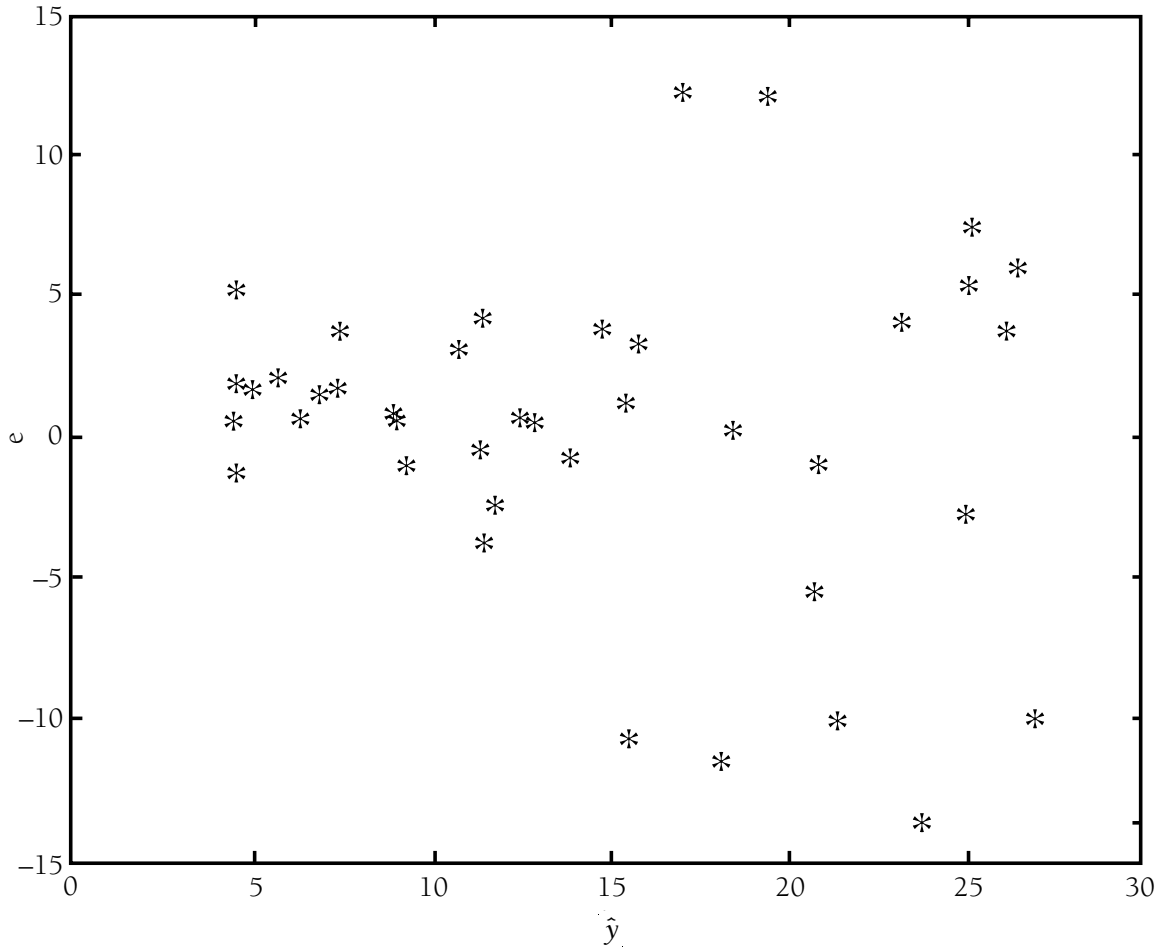


Figure 2.

PART 3

Requirement 1.

Using ln y as the dependent variable and $1000x$

as the independent variable, we get the prediction equation:

$$\ln y = 1.2061 + 0.0727x \qquad (1)$$

with $s = 0.7178$.

Requirement 2.

Using ln y as the dependent variable and x is the independent variable, we get the prediction equation:

$$\ln y = 6.83 + 0.146x \qquad (2)$$

Requirement 3.

The estimated concentration is 8.6mg/l. The upper and lower bounds are 14.9mg/l and 2.3mg/l.

Requirement 4.

a. No need to test the water.

b. At 25 feet the upper and lower suspended sediment concentrations are 113.0mg/l and 17.5mg/l, respectively. Therefore, under high stage conditions, the water should be tested at Cajunville.

c. The median flow value is 16290 cfs. The upper bound on the suspended particle concentration is 27.8mg/l —therefore under normal flow Cajunville's water supply is potable.

Requirement 5.

a. At 2.3mg/l 73% of the PCB is dissolved.

b. The worst case occurs if we have 2.3mg/l giving us 73% of the PCB in the dissolved phase. This yields a dissolved concentration of 0.55 μgrams/l. Since this value is greater than 0.5 μgrams/l, the water should be tested in Freeburg.

c. At 25 feet, the worst case is 0.75 μgrams/l at 0.26% dissolved, or 0.20μgrams/l dissolved. At this level, Freeburg need not test the water.

d. Under normal flow conditions, the worst case concentration is 0.44μgrams per liter so the water is potable for Freeburg.

PART 4.

Requirement 1.

Let $\vec{c}$ be the optimal vector we seek. From the theory of Lagrange multipliers, $\vec{c} = a\vec{1} + b\vec{1}$ for a,b to be determined. Substituting this into the side conditions $\vec{c} \cdot \vec{1} = 0$ and $\vec{c} \cdot \vec{x} = 1$ and we get the equations $na + \Sigma xb = 0$ and $\Sigma xa + \Sigma x^2 b = 1$. Solving for a,b we then get $\vec{c}$, and we check that $\vec{c} \cdot \vec{y}$ is the same as the least squares estimate $\hat{\beta}_1$, which we got in Part 1.

Requirement 2.

For each i let $d_i = c_i x_i^{.5}$, so that $c_i = d_i x_i^{-.5}$. Also let $u_i = x_i^{-.5}$ and $V_i = x_i^{-.5}$ for each i. We want to minimize $\Sigma c_i^2 x_i$ subject to $\vec{c} \cdot \vec{1} = 0$ and $\vec{c} \cdot \vec{x} = 1$. But this translates to minimize $\Sigma\, d_i^2$ subject to $\vec{d} \cdot \vec{u} = 0$ and $\vec{d} \cdot \vec{v} = 1$. Reasoning as in Part 4 Requirement 1 we see that $\vec{d} = a\vec{u} + b\vec{v}$ for a, b undetermined. Substituting this into the side conditions we are led to $\vec{u} \cdot \vec{u}a + \vec{u} \cdot \vec{v}b = 0$ and $\vec{v} \cdot \vec{u}a + \vec{v} \cdot \vec{v}b = 1$. Solving for a, b we get $\vec{d}$ and then $\vec{c}$. Then we get $\vec{c} \cdot \vec{y} = 1.8541$ as our best linear unbiased estimate for β_1, and this is of course different from the least squares estimate.