STA 242/ENV 255
M. Clyde
February 9, 1995.

Name ______________________________

Score ______________________________

First In–Class Exam and Solution

Notes:

1. This is an open book and open notes exam.

Problem 1. Prostate cancer is one of the most virulent forms of cancer. Generally, it has spread before being detected and is usually fatal. One dietary factor that has been studied for its relationship with prostate cancer is fat consumption. Data on fat consumption (in g/day) and prostate cancer death rates (per 100,000) for several countries were collected and a linear model was fit using Splus.

Refer to the Splus output at the end of the test to answer the following questions:

a) Does the value of R^2 suggest that fat consumption is a good predictor of the cancer mortality rate? Explain.

Yes - the value of R^2 is .7834, which means that 78% of the variation in mortality can be explained by the linear regression on Fat consumption. As the data indicates a linear relationship this is o.k.

b) For the model $\texttt{Mortality} = \beta_0 + \beta_1\texttt{Fat} + e$, test the hypothesis that $\beta_1 = 0$ versus the alternative hypothesis that $\beta_1 \neq 0$ at the $\alpha = .01$ level of significance. What is the meaning of b_1 in this problem? *(Don't forget about the units!)*

The p-value from the output is .0000 which is less than $\alpha = .01$ therefore, we would reject $H_o : \beta_1 = 0$ in favour of the alternative that it is nonzero.
The value of b_1 is .1134, which means that as a country's fat consumption increases by 1 gram/day there is an expected increase in .1134 deaths due to prostate cancer per 100,000 people.

c) Calculate a 95% confidence interval for β_1. What is the meaning of this interval?

The 95% CI is $.1134 \pm 2.048.0113 = (.009026, .1365)$. *Under repeated samples we would expect 95% of intervals calculated in the same manner to contain β_1. The 95% Bayesian probability interval using the prior distribution on the handout, is the same interval as the CI, but has the interpretation that we believe there is a .95 probability that β_1 given the existing data is in that interval.*

d) What is the estimated cancer mortality rate for the US? What is the prediction error for the US?

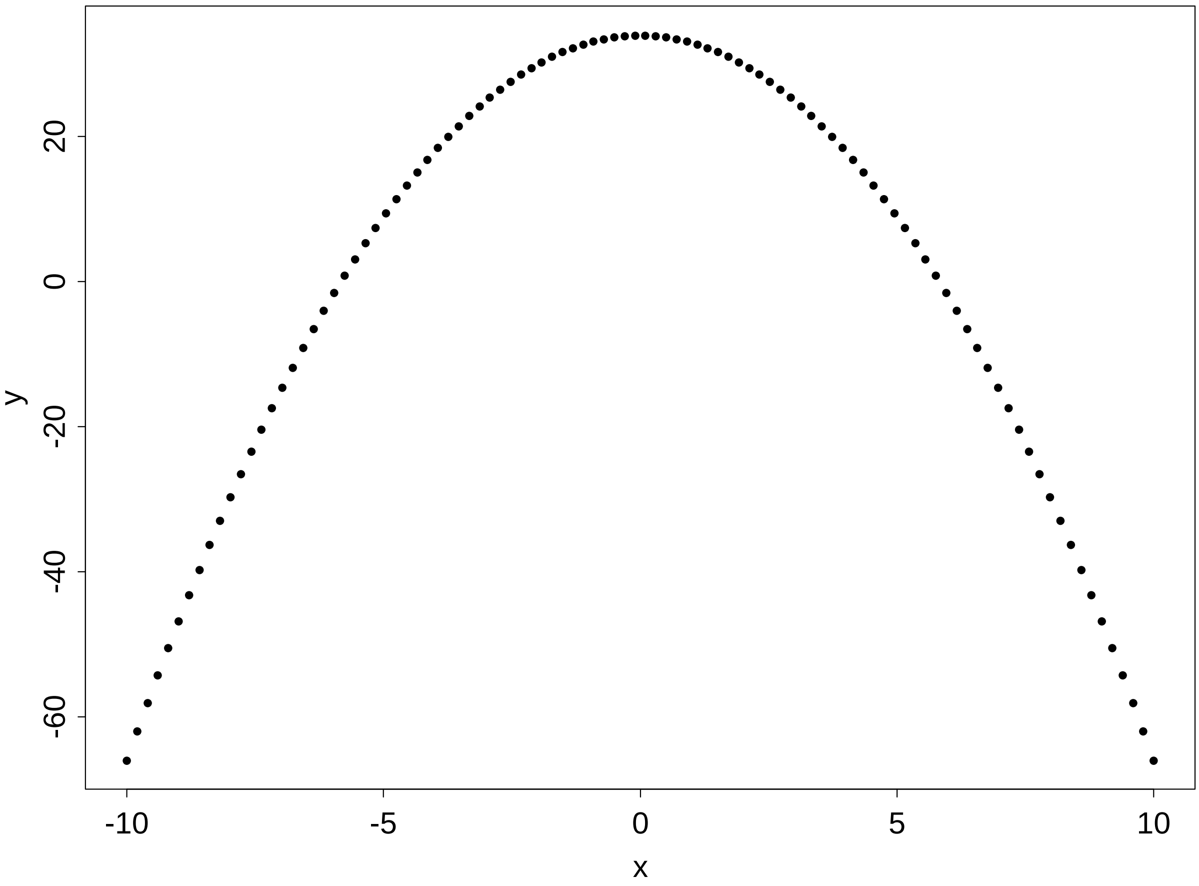
The estimated mortality for the US is $-1.0634 + .1134 * 147 = 15.61$ *deaths per 100,000 people. The prediction error is* $16.3 - 15.61 = .69$

e) Is there any evidence that the linear model is not appropriate? Explain.

Not really, the residual plot does not indicate any signs of nonconstant variance or lack of fit due to a nonlinear relationship. There are a couple of points with large residuals that we might want to examine closer.

f) Can we conclude from this study that eating a diet high in fat causes prostate cancer? Explain.

No. This is observation data and the observed strong relationship could be due to a third (unmeasured variable) that is correlated with both fat consumption and death rate due to prostate cancer. In fact, we do not know if the men with prostate cancer actually had a diet high in fat, since the measurements are aggregated over the country.



Problem 2.
The above data has an R^2 value of 0, yet it looks like there is a perfect relationship between Y and X. What is the explanation?

The R^2 value above is how much variation in Y is explained by the linear regression on X, $\hat{Y}_i = b_0 + b_1 X_i$. From the picture it is clear that there is not a linear relationship but a perfect curvilinear (in fact quadratic) relationship with X. This points out the danger of relying on computer output with out looking at residual plots or plots of the data in an analysis.

Problem 3. Fourteen indoor air samples in a house near a metal-pipe foundry in Phillsburg, New Jersey were collected. The investigators suspected that Benzo(a)pyrene (BaP) from the foundry was contaminating the household air. The measurements are concentrations of BaP-containing particles no larger than 10 micrograms. The measurements are in micrograms per cubic meter (μ/m^3). Refer to the figures at the end of the test.

a) Is it reasonable to summarize the distribution of BaP by a normal distribution using the mean and standard deviation? Explain.

No. The data are positively skewed as evidenced by the boxplot, histogram, symmetry plot, and normal quantile plot. So we cannot get a way with summarizing the data using just the mean and standard deviation as would be the case if the distribution were normal.

b) What is the median concentration (approximately)?

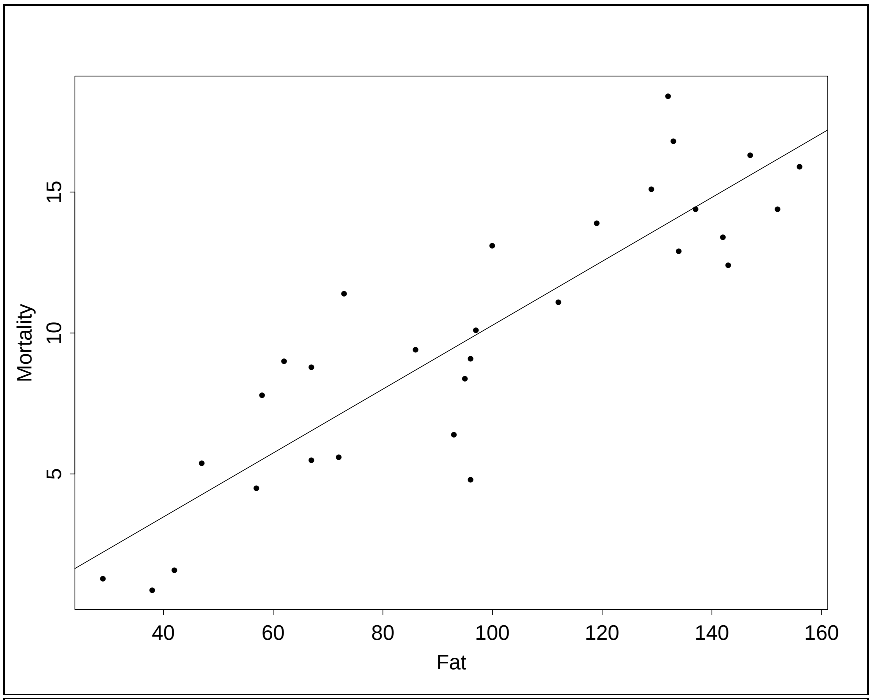
From the boxplot it looks like it is about 50 μ/m^3.

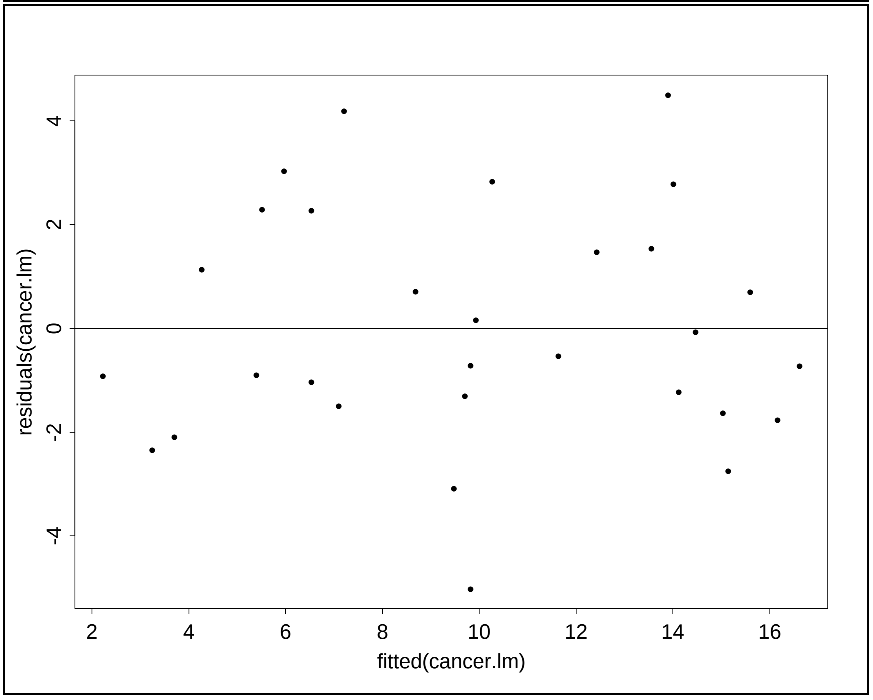
c) What types of transformations might reduce the skewness?

Because of the positive skew, values of $q < 1$ *would be appropriate. Air and water pollutant concentrations often are lognormally distributed, which means that the log(concentration) is normally distributed. So* $q = 0$ *might be a good candidate, but other values should be tried as well. Since there are no 0 or negative values we will not have to add a constant to the data.*

Output for Problem 1

```
                Fat Mortality
    ElSalvador  38        0.9
   Philippines  29        1.3
         Japan  42        1.6
        Mexico  57        4.5
        Greece  96        4.8
      Colombia  47        5.4
      Bulgaria  67        5.5
    Yugoslavia  72        5.6
        Poland  93        6.4
        Panama  58        7.8
        Israel  95        8.4
       Romania  67        8.8
     Venezuela  62        9.0
Czechoslovakia  96        9.1
         Italy  86        9.4
         Spain  97       10.1
      Portugal  73       11.4
       Finland 112       11.1
       Hungary 100       13.1
            UK 143       12.4
       Germany 134       12.9
        Canada 142       13.4
       Austria 119       13.9
        France 137       14.4
   Netherlands 152       14.4
     Australia 129       15.1
       Denmark 156       15.9
            US 147       16.3
        Norway 133       16.8
        Sweden 132       18.4

> cancer.lm <- lm(Mortality ~ Fat)
> summary(cancer.lm)

Call: lm(formula = Mortality ~ Fat)
Residuals:
   Min     1Q  Median    3Q   Max
 -5.02 -1.451 -0.6264 1.523 4.499

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  -1.0634   1.1705    -0.9085   0.3714
        Fat   0.1134   0.0113    10.0646   0.0000

Residual standard error: 2.295 on 28 degrees of freedom
Multiple R-Squared: 0.7834
F-statistic: 101.3 on 1 and 28 degrees of freedom, the p-value is
8.347e-11
```


Mortality
5
10
15
40
60
80
100
120
140
160
Fat


residuals(cancer.lm)
-4
-2
0
2
4
2
4
6
8
10
12
14
16
fitted(cancer.lm)

Output for Problem 3

```
> mean(indoor)
[1] 76.07143
> sqrt(var(indoor))
[1] 79.10992
```