LINEAR REGRESSION MODELS W4315

HOMEWORK 4 ANSWERS
February 27, 2010

Instructor: Frank Wood

1. (50 points) [1] Refer to **Copier maintenance** Problem 1.20.

a. Prepare a dot plot for the number of copiers serviced X_i. What information is provided by this plot? Are there any outlying cases with respect to this variable?

b. The cases are given in time order. Prepare a time plot for the number of copiers serviced. What does your plot show?

d. Prepare residual plots of e_i versus $\hat{Y}_i$ and e_i versus X_i on separate graphs. Do these plots provide the same information? What departures from regression model (2.1) can be studied from these plots? State your findings. And the model (2.1) is as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where:

$$\beta_0 \text{ and } \beta_1 \text{ are parameters}$$
$$X_i \text{ are known constants}$$
$$\epsilon_i \text{ are independent } N(0, \sigma^2)$$

e. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. DOes the normality assumption appear to be tenable here? Use table B.6 and $\alpha = .10$.

f. Prepare a time plot of the residuals to ascertain whether the error terms are correlated over time. What is your conclusion?

h. Information is given below on two variables not included in the regression model, namely, mean operational age of copiers serviced on the call(X_2, in months) and years of experience of the service person making the call(X_3). Plot the residuals against X_2 and X_3 on separate

---

[1]This is problem 3.4 in 'Applied Linear Regression Models(4th edition)' by Kutner etc.

graphs to ascertain whether the model can be improved by including either or both of these variables. What do you conclude?

N.B. 1. If you need any software for this problem, do not use the embedded linear regression commands, say, 'regress' in MATLAB is not allowed. 2. If you are using software, please attach the code at the back of your handed-in homework instead of mixing codes with the results. 3. You don't have to hand in part c and h of this problem.

Answer:

(a) The dot plot for the number of copiers serviced X_i was shown below. You can use the 'dotchart' function in R to draw the dot plot.
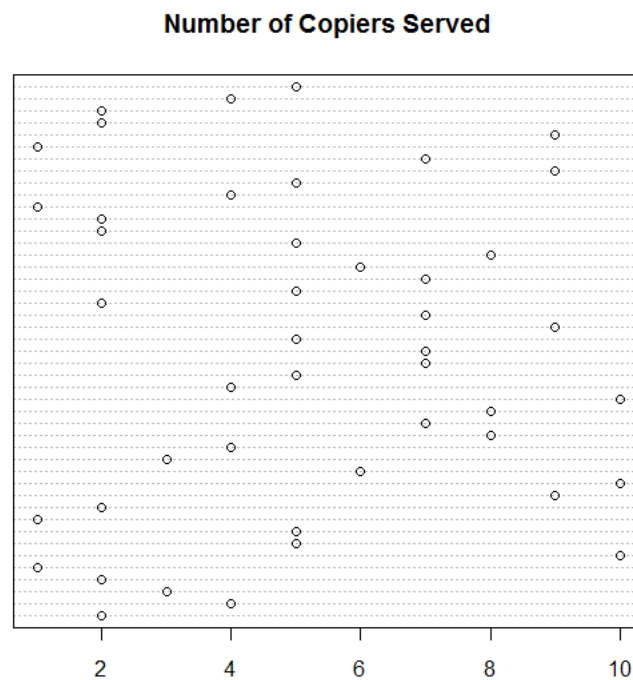

Figure 1: Dot Plot for X_i

The range for X_i was between 1 to 10, they were evenly distributed and no outlying cases were observed with respect to X_i.

R Code:

hw4=read.table("CH01PR20.txt")

dotchart(hw4[,2],main="Number of Copiers Served")

2

(b)Figure 2 was the plot of the number of copiers serviced versus time. The plot did not show any special pattern. The number of copiers serviced was not correlated over time.
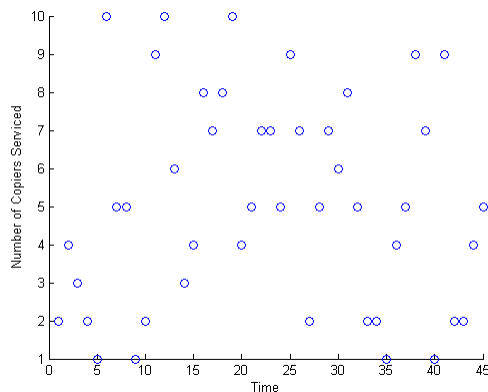

Figure 2: The time plot for X_i

Matlab Code:
pr1=textread('CH03PR04.txt');
X=pr1(:,2);
Y=pr1(:,1);
n=size(X);
n=n(1);
time=1:n;
scatter(time,X)
xlabel('Time')
ylabel('Number of Copiers Serviced')

(d) The residual versus X and $\hat{Y}$ plots were shown in Figure 3 and Figure 4. The variance of residuals seemed to be constant with the change of X and $\hat{Y}$. The constant variance assumption of the simple linear regression model was not violated in this case. Both of plots indicated the presence of potential outliers. As we can see, most of the residuals were in the range of -15 and 15, however, there were two residuals much smaller than -15 which raised a flag for further analysis. The two plots provided the same information.
 Matlab Code:
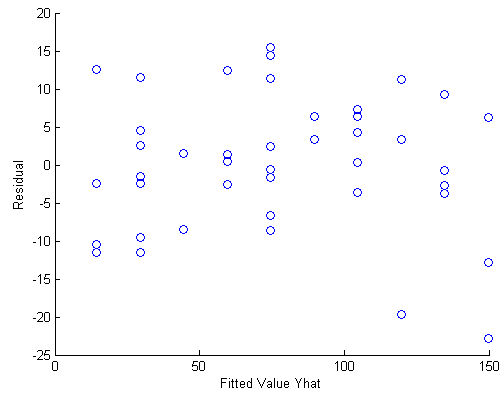avgX = mean(X);
avgY = mean(Y);
$SXX = sum((X - avgX).^2);$

3

Figure 3: Residual versus Fitted Value Plot

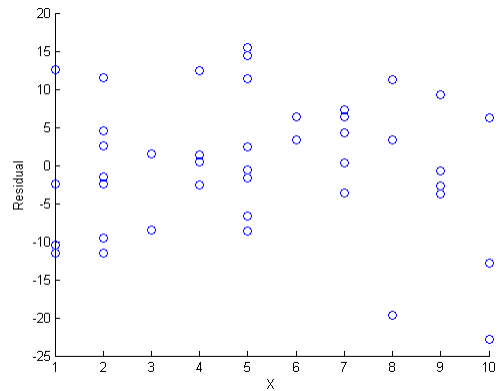

Figure 4: Residual versus X Plot

$SXY = sum((X - avgX). * (Y - avgY));$
b1 = SXY/SXX;
b0 = avgY- b1*avgX;
resid=Y-b0-b1*X
Yhat=b0+b1*X
scatter(Yhat,resid)
xlabel('Fitted Value Yhat')
ylabel('Residual')
scatter(X,resid)
xlabel('X')
ylabel('Residual')

(e)A normal probability plot of the residuals was shown in Figure 5. The coefficient of cor-

4

relation between the ordered residuals and their expected values under normality is 0.9891. With n=45, from Table B.6, the critical value for the coefficient of correlation between the ordered residuals and the expected values under normality when the distribution of error terms is normal using a 0.01 significance level is 0.979. Since 0.9891 > 0.979, the assumption of normality appeared reasonable.
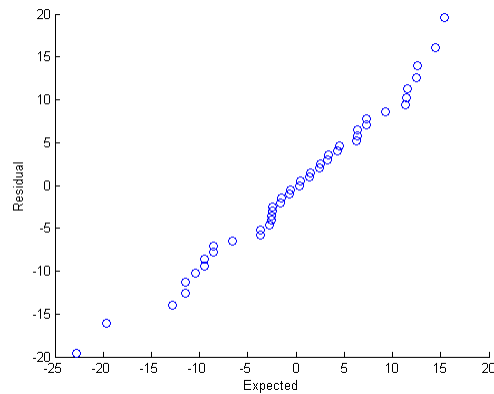

Figure 5: Normal Probability Plot of the Residuals

Matlab Code:
newresid=sort(resid)
k=1:n
z=norminv((k-0.375)/(n+0.25))
$mse = sum(resid.^2)/(n-2)$
expected=z*sqrt(mse)
scatter(newresid,expected)
xlabel('Expected')
ylabel('Residual')
sxy=(expected-mean(expected))*(newresid-mean(newresid))
$sxx = sum((expected - mean(expected)).^2)$
$syy = sum((newresid - mean(newresid)).^2)$
r=sxy/sqrt(sxx*syy)

(f) The residual versus time plot did not show any evidence that the error terms were correlated over time.
Matlab Code:
scatter(time,resid)
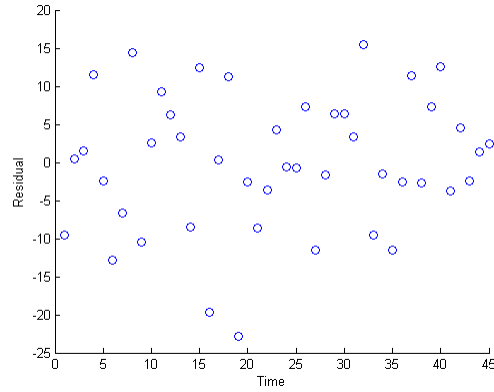
5

xlabel('Time')
ylabel('Residual')

Figure 6: Residual versus Time Plot

(h) The residual versus X_2 plot showed a positively correlated linear trend between the residual and the mean operational age of copiers serviced(X_2) which indicated that the model might be improved by including X_2. The residual versusX_3 plot did not show any special trend or pattern which indicated that adding X_3 into the model may not bring any improvement in terms of model fitting.
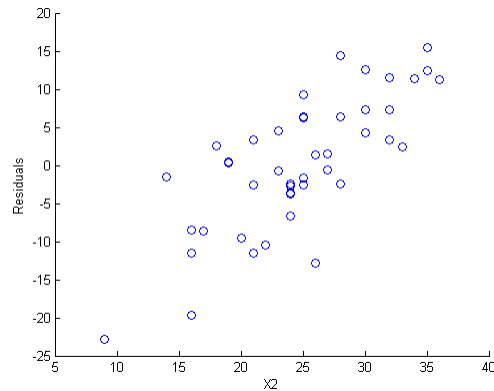

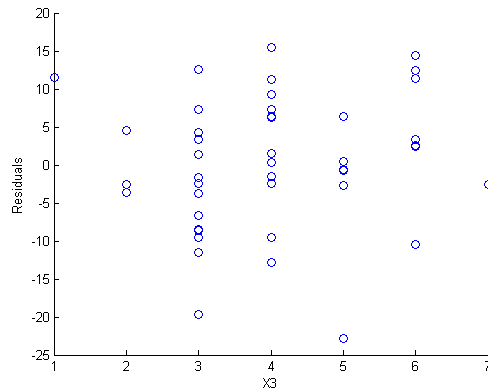Figure 7: Residual versus X_2 Plot

Matlab Code:
X2=pr1(:,3);

Figure 8: Residual versus X_3 Plot

X3=pr1(:,4);
scatter(X2,resid)
xlabel('X2')
ylabel('Residuals')
scatter(X3,resid)
xlabel('X3')
ylabel('Residuals')

2. (25 points) [2] A student fitted a linear regression function for a class assignment. The student plotted the residuals e_i against Y_i and found a positive relation. When the residuals were plotted against the fitted values $\hat{Y}_i$, the student found no relation. How could this difference arise? Which is the more meaningful plot?

Answer:

In either cases, the residuals represent the distance from the Y's to the fitted Y's. If there is no violation of the assumptions, there should be apparent pattern when plotting residuals again the fitted Y values, which is more meaningful when carrying out residual analysis. As for plotting residuals against observed Y's, it should always manifest positive relations between the two, so it's not sensible. You can explain this from two perspectives. One is by intuition: for large Y's, it is more likely that the values depart from the regression line(especially if the line is pretty flat, and the data are more spread out), and for small Y values, it's more likely that they're somewhere around the regression line, thus have a less significant residual. Another way to explain the positive correlation comes from the

---

[2]This is problem 3.19 in 'Applied Linear Regression Models(4th edition)' by Kutner etc.

decomposition of the Y:

$\hat{e} = Y - \hat{Y}$, so we have $Y = \hat{e} + \hat{Y}$. Since $\hat{(Y)}$ and $\hat{(e)}$ are independent, we have that the covariance between $\hat{(e)}$ and Y is always positive, even when the assumption is not violated. In this sense, the residual plot against fitted Y is more meaningful, actually it is the most classic residual plot that we usually use.

3. (25 points) [3] If the error terms in a regression model are independent $N(0, \sigma^2)$, what can be said about the error terms after transformation $X' = 1/X$ is used? Is the situation the same after transformation $Y' = 1/Y$ is used?

Answer:

If the transformation is exerted only on X, then the error terms remain normally distributed, however, the mean will shift–this is because X is not taken as random variables in the regression setting, so if only X is transformed, the model's distribution of course is not going to change.

Nevertheless, if Y is reversed, then the error term's distribution will change in the way that it will not even follow a normal distribution. This is simply because the reverse of a normally distributed random variable will not still follow normal distribution. So when taking transformations, we should be careful about the inference made after fitting the model, since the distribution may be different than which we carry forward the classic inference.

---

[3]This is problem 3.20 in 'Applied Linear Regression Models(4th edition)' by Kutner etc.