LINEAR REGRESSION MODELS W4315

HOMEWORK 4 ANSWERS

October 19, 2009

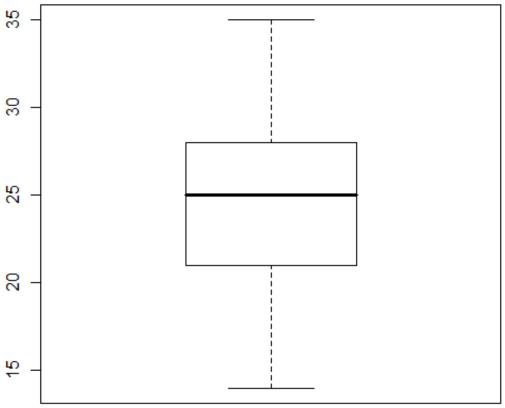
Instructor: Frank Wood (10:35-11:50)

1. (50 points) Problem 3.3 in the book.
N.B. 1. If you need any software for this problem, do not use the embedded linear regression commands, say, 'lm' in R is not allowed. 2. If you are using software, please attach the code to your handed-in homework. Results alone without code will not be accepted.

Answer:

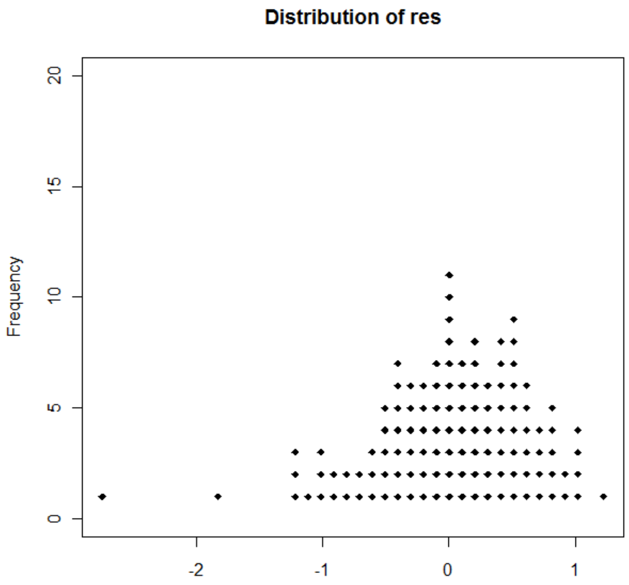
(a) Continuing the code from last homework with the same notation, the R code to draw the plot is: boxplot(x), which is illustrated in figure 1. There doesn't seem to be anything noteworthy about the ACT scores other than it seems rather symmetric with no apparent outliers. The median seems to lie exactly in the middle of the range and the IQR. Because the data size is 120, we can be fairly certain that the distribution is symmetrical.

Figure 1: 3.3a box plot of the ACT scores



(b) The R code is: dotchart(y-y.hat) which gives the dot plot of residuals: It shows roughly the distribution of the error terms. It appears to be centered around 0 and does look normal.

Figure 2: 3.3b dot plot of the residuals



However, there are a couple of outliers.
(c) The R code is:

plot(y.hat,y-y.hat,xlab="fitted Y values",ylab="residuals")

which give the residual plot as: There are a few residuals that depart quite distantly from the regression model as seen in the plot. But in general, the residuals are scattered evenly about the model without a pattern, which is a good thing for setting up the linear regression model.
(d) The R code to draw a qq-plot is "qqnorm" or "qqplot", who have different parameters. You are suggested to go check the detail using R built-in help system. Here we can use:
$qqnorm(res)$
$qqline(res, col = 2)$
directly to draw the qqplot as follows: As for the correlation between the sorted residuals and expected value under normality assumption, we can use the following R code:
$sort\_res < -sort(res, index.return = TRUE)\$x$
$index < -sort(res, index.return = TRUE)\$ix$
$expected\_under\_normal < -sqrt(MSE) * pnorm((c(1 : n) - .375)/(n + .25))$
$cor(sort\_res, expected\_under\_normal)$ → this is the test statistics r

Figure 3: 3.3c residual plot

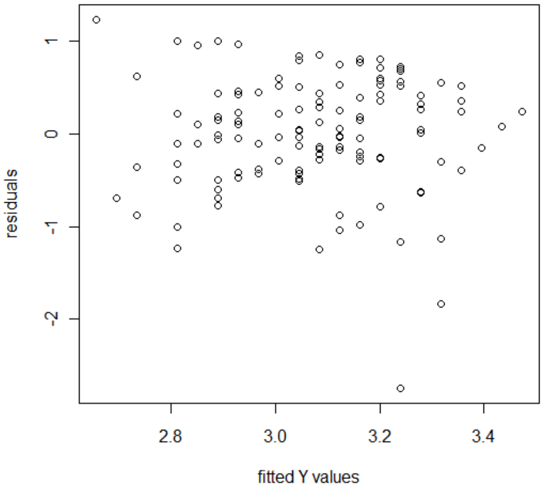
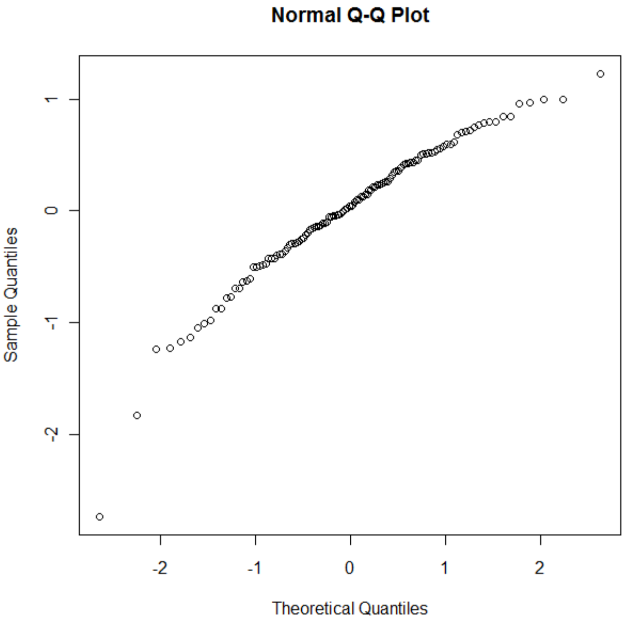


Figure 4: 3.3d qqplot to test normality assumption of the model



In this specific setting, the formulation of the hypothesis is:
$H_0 : Normalityassumptionholds \leftrightarrow H_1 : Normalityassumptionfails$
Then we get the correlation between the two is around .97. Refer to Table B.6 and we can find that $.97 < .98$, so we reject H_0 which means that the normality assumption doesn't hold here. This keeps sync with conclusion that we can get from the qqplot, since from it,

we can see that the dots do not all align up on one straight line. But also notice that the skewness is not severe and the test statistics is not far smaller than the criteria. So normality assumption may not hold here, but the violation is not very significant.
(d)Conduct the Brown-Forsythe test:
$res.1 < -res[x < 26]$
$res.2 < -res[x >= 26]$
$n1 < -length(res.1)$
$n2 < -length(res.2)$
$med.1 < -median(res.1)$
$med.2 < -median(res.2)$
$d1 < -abs(res.1 - med.1)$
$d2 < -abs(res.2 - med.2)$
$s < -sqrt((sum((d1 - mean(d1))^2) + sum((d2 - mean(d2))^2))/(n - 2))$
$t.BF < -(mean(d1) - mean(d2))/s/(sqrt(1/n1 + 1/n2))$
$qt(.995, n - 2)$
The decision rule is that if $|t.BF| < t(.995, 118)$, then we accept the H_0, otherwise we reject it. Here we have t.BF to be -0.896, so we accept the assumption that error terms have equal variances, or at least the error variance does not vary with the level of X. It does not contradict the conclusion we arrived at (c).
(f) The R code is:
$x2 < -data[, 3]$
$x3 < -data[, 4]$
$plot(x2, res, xlab = "intelligencetestscore", ylab = "residuals")$
$plot(x3, res, xlab = "highschoolclassrankpercentile", ylab = "residuals")$
The residual plots are:(Please refer to the last page.) From the graphs we can see there is a transparent linear trend between the residuals and X2, but no obvious trend between residuals and X3. So we should include X2 into our model, and proceed to further investigation into the case concerning X3.

2. (25 points) Problem 3.19 in the book.

Answer:

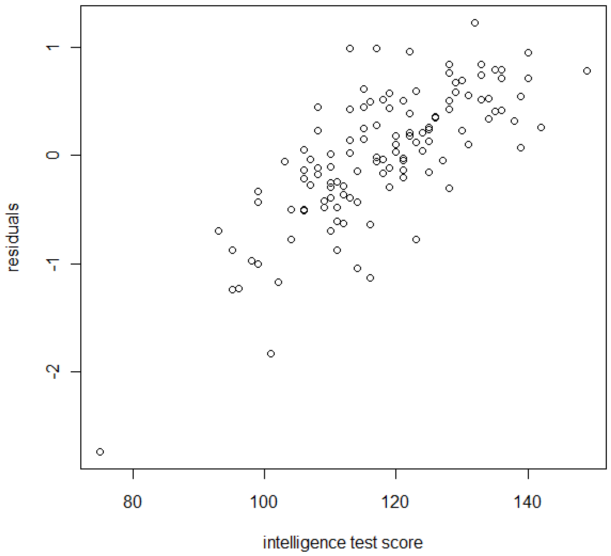
In either cases, the residuals represent the distance from the Y's to the fitted Y's. If there is no violation of the assumptions, there should be apparent pattern when plotting residuals again the fitted Y values, which is more meaningful when carrying out residual analysis. As for plotting residuals 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 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 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.

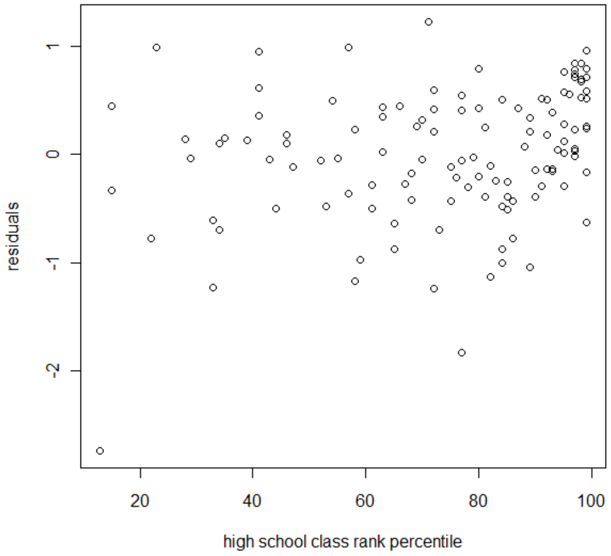
3. (25 points) Do problem 3.20 in the book.

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.



(a) 3.3f plot of residuals against X_2



(b) 3.3f plot of residuals against X_3