Stat 615/415 (Regression) HW #5 – Solutions

Regression Diagnostics (chap. 3 [3.1-3.7, 3.11]).

1. (**3.12**) A student does not understand why the sum of squares SSPE is called a *pure error sum of squares* "since the formula looks like the one for an ordinary sum of squares". Explain.

   SOLUTION. *The pure error sum of squares measures variation that can be caused only by the randomness of responses. It cannot be caused by an omitted nonlinear trend because the most general nonlinear model has already been fitted, and this accounts for all nonlinear trends.*

2. (**3.19**) A student fitted a linear regression function for a class assignment. The student plotted the residuals e_i against responses Y_i and found positive relation. When the residuals were plotted against the fitted values $\hat{Y}_i$, the student found no relation.

   (a) How could the differences arise? Which is the more meaningful plot?

   SOLUTION.

   (a) *There will always be no linear relation between residuals e_i and fitted values $\hat{Y}_i = b_0 + b_1 X_i$, because we already know that there is no linear relation between e_i and X_i.*
   There should still be a relation between residuals e_i and actual responses Y_i. Responses above the linear regression line have positive residuals, and the ones below the regression line have negative residuals. This explains the positive correlation which is what the student observed.
   The plot against fitted values $\hat{Y}_i$ is more meaningful because it lets one see any omitted nonlinear terms.

3. (Computer project, **3.3**) Refer to the GPA data from the previous h/w assignments.

   (a) Plot residuals e_i against the fitted values $\widehat{Y}_i$. What departures from the standard regression assumptions can be detected from this plot?

   (b) Prepare a Normal Q-Q plot of the residuals and use it to comment on whether the data passes or fails the assumption of normality. Conduct the Shapiro-Wilk test for normality.

   (c) Test whether residuals in this regression analysis have the same variance.

   (d) Conduct the lack-of-fit test and state your conclusion.

   SOLUTION.

   (a) *Very minor nonlinear trend can be seen, so the linear model is probably appropriate. The variance seems to be increasing, so it may be non-constant. No evidence of any correlations among regression residuals.*

(b) *The Q-Q plot is linear in the middle but nonlinear at the ends, with several clearly extreme points. Shapiro-Wilk test rejects the assumption of normality with a very low p-value.*

(c) *With a high p-value of 0.42397, there is no evidence of non-constant variance.*

(d) *With a high p-value of 0.6324, there is no evidence of a lack of fit.*

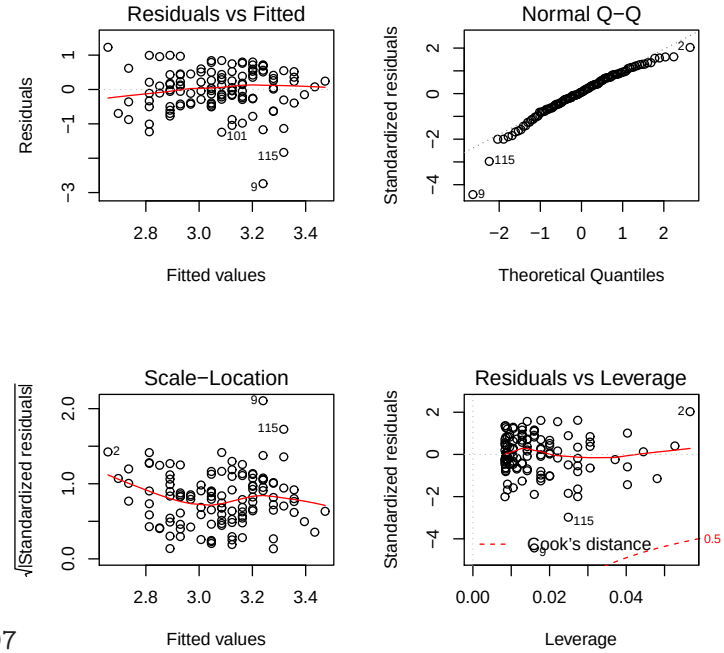
R code:

```
> X = V2; Y = V1;
> par(mfrow=c(2,2))
> reg = lm(Y~X)
> plot(reg)
> shapiro.test(rstudent(reg))

Shapiro-Wilk normality test

data:  rstudent(reg)
W = 0.94176, p-value = 5.638e-05

> install.packages("car"); library(car);
> ncvTest(reg)
Non-constant Variance Score Test
Variance formula:     fitted.values
Chisquare = 0.6392758, Df = 1, p = 0.42397
> anova(reg,full)
Analysis of Variance Table
Model 1:  Y   X
Model 2:  Y   as.factor(X)
Res.Df RSS Df Sum of Sq F Pr(>F)
1 118 45.818
2 99 39.332 19 6.4857 0.8592 0.6324
```



4. (Computer project,) **Crime rate** data set is available on our Blackboard site.

A criminologist studies the relationship between level of education and crime rate in medium-sized U.S. counties. She collected data from a random sample of 84 counties; X is the percentage of individuals in the county having at least a high-school dipoma, and Y is the crime rate (crimes reported per 100,000 residents) last year.

i	1	2	3	4	5	6	7	8	$\cdots$
Y_i	8487	8179	8362	8220	6246	9100	6561	5873	$\cdots$
X_i	74	82	81	81	87	66	68	81	$\cdots$

A linear regression of Y on X is then fit to these data. Test:

(a) normal distribution of residuals;

(b) constant variance of residuals;

(c) presence of outliers;

(d) lack of fit.

SOLUTION.

(a) *Shapiro-Wilk test fails to reject Normality. No evidence that responses don't follow a Normal distribution. The p-value is 0.1559.*

(b) *The Breusch-Pagan test fails to reject the assumption of a constant variance. No evidence of heteroscedasticity. The p-value is 0.94338.*

(c) *At the individual level of $\alpha = 0.05$, there is one outlier, observation #27 with the studentized residual $t = 3.048$. However, keeping the familywise error rate at the level of 0.05, there is no evidence of any outliers.*

(d) *There is no evidence of lack of fit. The p-value is 0.8066.*

R code:

```
> X = V2; Y = V1;
> reg = lm(Y∼X)
> t = rstudent(reg)
> n = length(X)
> shapiro.test(t)
> ncvTest(reg)
> t[ abs(t) > 1-qt(0.05/2/n,n-2) ]
named numeric(0)
> t[ abs(t) > 1-qt(0.05/2,n-2) ]
> full = lm(Y   as.factor(X))
> anova(reg,full)
```

5. For the "toy" example, consider a small data set

X	0	0	1	2
Y	0	2	2	3

Try to do as much as you can by hand, without the use of a computer. The numbers are quite simple!

(a) Plot these data and draw the least squares regression line, which has the expression $y = 1 + x$.

(b) Compute all the residuals.

(c) Compute all sums of squares by hand, from their definitions:

$$\begin{aligned}
\text{SSTot} &= \sum_i (Y_i - \bar{Y})^2 \\
\text{SSReg} &= \sum_i (\hat{Y}_i - \bar{Y})^2 \\
\text{SSErr} &= \text{SSTot} - \text{SSReg} = \sum_i (Y_i - \hat{Y}_i)^2 \\
\text{SS}_{\text{PE}} &= \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2 \\
\text{SS}_{\text{LOF}} &= \text{SSErr} - \text{SS}_{\text{PE}} = \sum_j \sum_i (\bar{Y}_j - \hat{Y}_j)^2
\end{aligned}$$

Then conduct the lack-of-fit test. Explain the result.

SOLUTION.

X_i	0	0	1	2
Y_i	0	2	2	3
$\bar{Y}$	1.75	1.75	1.75	1.75
$Y_i - \bar{Y}$	−1.75	.25	.25	1.25
$\hat{Y}_i$	1	1	2	3
e_i	−1	1	0	0
$\hat{Y} - \bar{Y}$	−.75	−.75	.25	1.25
$\bar{Y}_j$	1	1	2	3
$\hat{Y}_{ij} - \bar{Y}_j$	0	0	0	0
$Y_{ij} - \hat{Y}_{ij}$	−1	1	0	0

Then

$$\begin{aligned}
SSTot &= \sum_i (Y_i - \bar{Y})^2 = 4.75 \\
SSReg &= \sum_i (\hat{Y}_i - \bar{Y})^2 = 2.75 \\
SSErr &= SSTot - SSReg = \sum_i (Y_i - \hat{Y}_i)^2 = 2 \\
SS_{PE} &= \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2 = 2 \\
SS_{LOF} &= SSErr - SS_{PE} = \sum_j \sum_i (\bar{Y}_j - \hat{Y}_j)^2 = 0
\end{aligned}$$

The F-statistic is $F = 0$, and the p-value is $p = 1$, so there is no evidence of a lack of fit.

There cannot be any evidence of a lack of fit because both the full and the reduced models produced precisely the same fitted values $\hat{Y}_{ij} = \bar{Y}_j$. The two models predict responses in the same way! Certainly, there is no evidence that the full model is any better.