Solution to HW 4

Guanhua FANG

October 17, 2017

3.3

a. The ACT score are quite symmetric about its median and no extreme value exists.

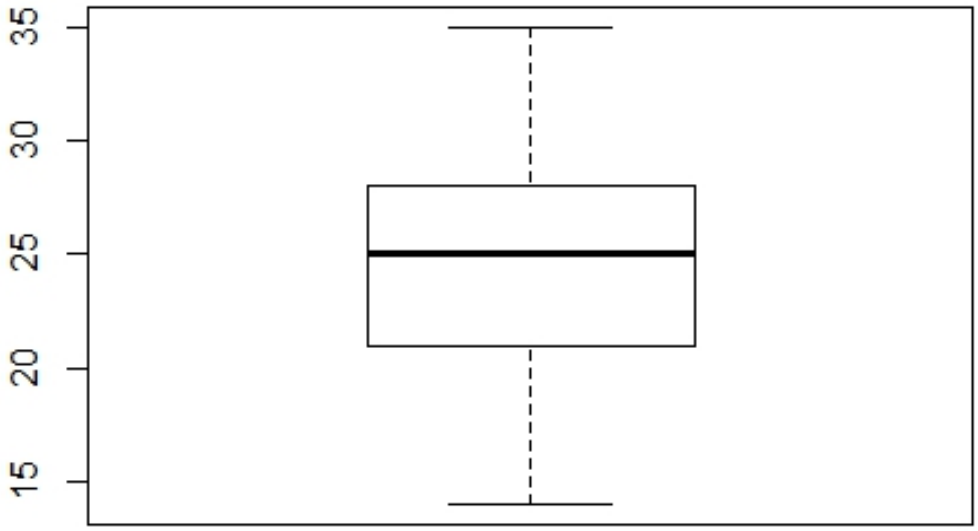


Figure 1: 3.3a

c. The plot shows that the variance of residuals are quite constant as fitted value $\hat{y}$ varies except one outlier.

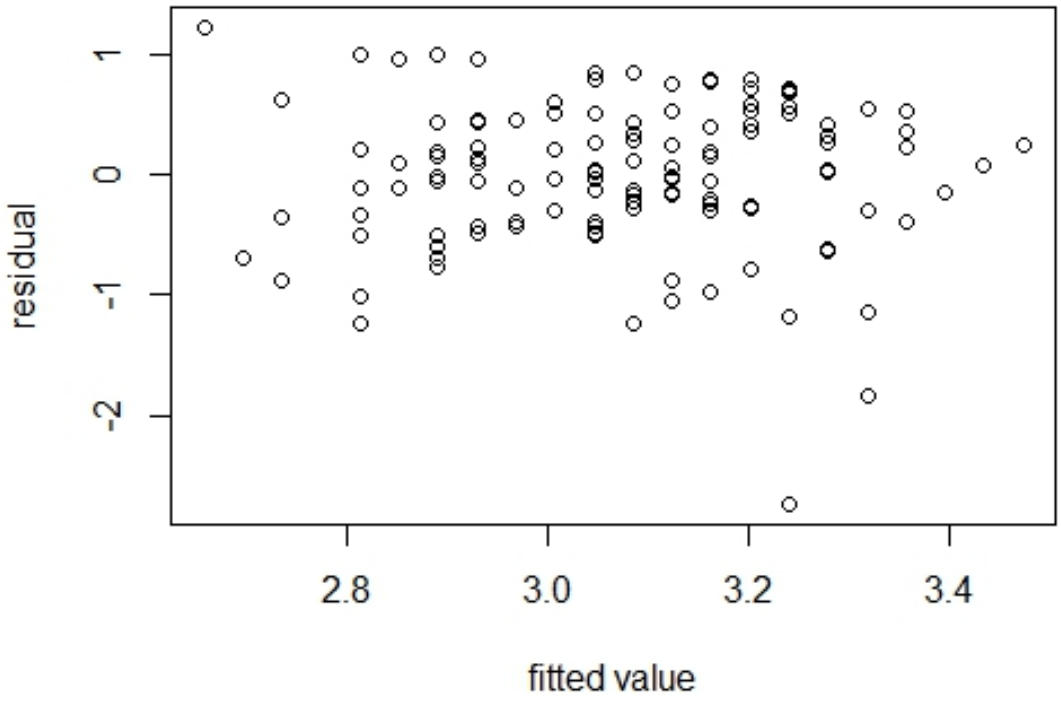


Figure 2: 3.3c

d. The correlation of coefficient is 0.974, which is below 0.987. So normality assumption is rejected.

e. The decision rule is whether $|t_{bf}| > qt(0.995, 118)$. We get $t_{bf} = -0.896$ and cut off point is 2.62. Then, null cannot be rejected. Yes, it support the finding in c.

f. Do two regression separately, we see that from residual plot. We cannot say too much about it, since there are too few points around two ends of intelligence score. For class rank, it seems a little bit uneven in terms of error variance of but not too much. In addition, both two qq plots show the approximate normality. If we do the regression using all three predictors, then "intelligence score" seems most significant. It should be included in the model.

3.9

From residual plot, we can find that the errors has the nonlinear relation with X. The assumption of linear model may be violated. A transformation like box-cox transformation may mitigate the problem, it depends on the original response value Y.

3.16

a. Use a log transformation on Y could achieve constant variance and linearity.

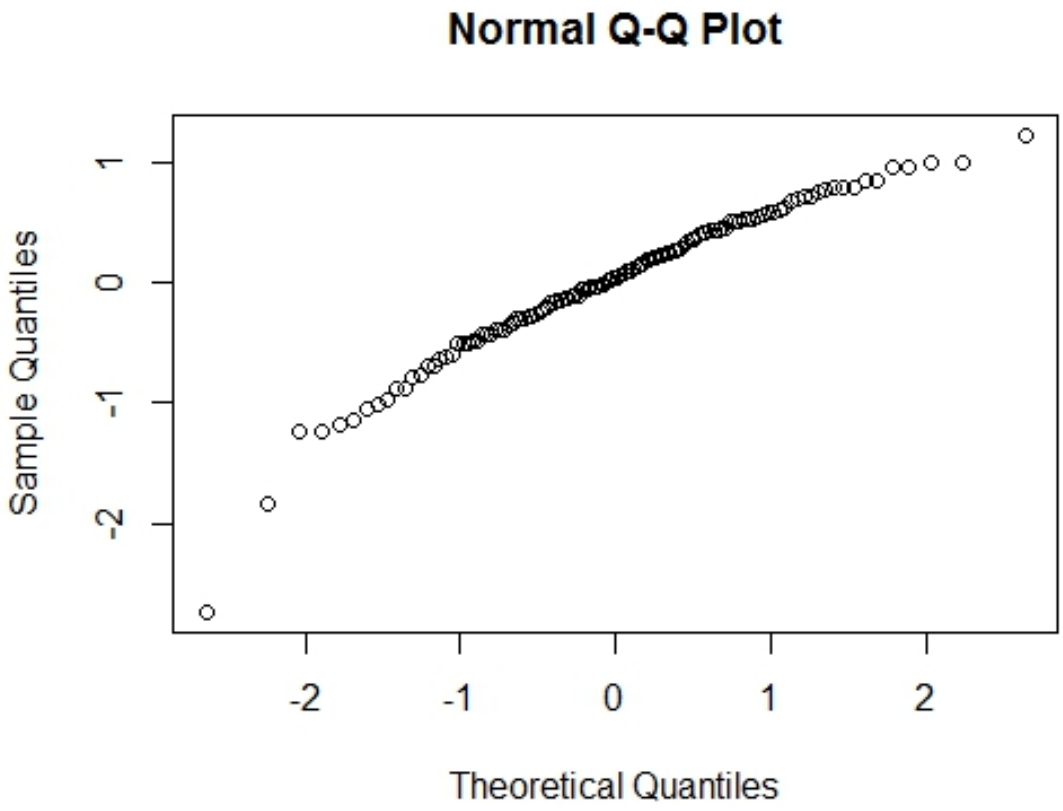


Figure 3: 3.3d

b. When $\lambda = .2$, the SSE is 0.0107. The standard deviation of transformed Y is 0.23. After eliminating the scale effect, then relative SSE will be 0.32. When $\lambda = .1$, the SSE is 0.0017. The standard deviation of transformed Y is 0.12. After eliminating the scale effect, then relative SSE will be 0.11. When $\lambda = 0$, the SSE is 0.17. The standard deviation of transformed Y is 1.32. After eliminating the scale effect, then relative SSE will be 0.096. When $\lambda = -.1$, the SSE is 0.0033. The standard deviation of transformed Y is 0.14. After eliminating the scale effect, then relative SSE will be 0.16. When $\lambda = -.2$, the SSE is 0.029. The standard deviation of transformed Y is 0.32. After eliminating the scale effect, then relative SSE will be 0.29. Hence, it suggest it is best when $\lambda = 0$.

c. The regression function is $y^{'} = -0.20x + 0.65$.

d. From the plot, the regression line seems good.

e. From the residual plot, it roughly shows the constant variance. The qq plot shows that the residuals and expected value are almost on the line. Fitting is not to bad.

f. The estimated function is $y = 10^{-0.2x+0.65}$.

3.23

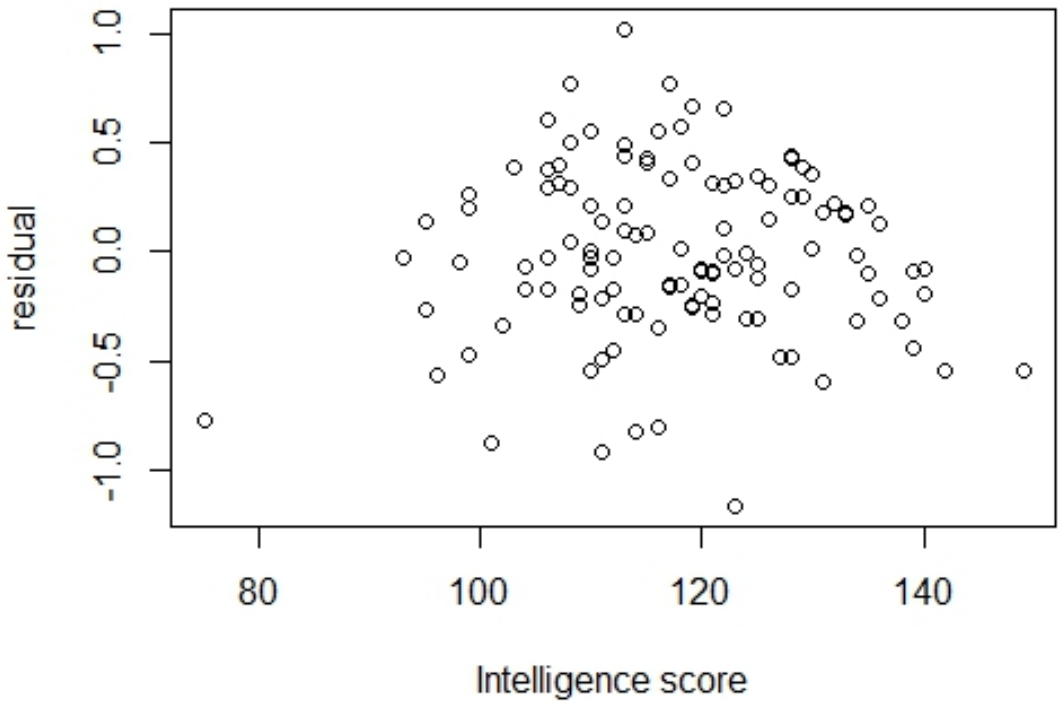


Figure 4: 3.3f

The model is $Y = \beta X + \epsilon$, then $EY = \beta X$. The full model is $Y_{ij} = \mu_j + \epsilon_{ij}$. The reduced model is $Y_{ij} = \beta_j X_j + \epsilon_{ij}$. Then, the degree of freedom for full model is $n - c = 10$ and the degree of freedom for reduced model is $n - 1 = 19$.

3.24

a. The regression function is $y = 48.67 + 2.33x$. From the residual plot, we can see there is a outlier while others are roughly around zero in reasonable distance.

b. After omitting the seventh case, we find the new regression function will be $y = 53.07 + 1.62x$. Case 7 is the outlier, the observation value is much higher than the expected value. Then, slope estimate is pulled up.

c. The y_h is $53.07 + 1.62 \times 12 = 72.51$. $\bar{x} = 8.86$. $SSX = 58.86$. Then, $s^2(y_h) = 2.645(1 + \frac{1}{7} + \frac{(12-8.86)^2}{58.86}) = 3.47$ Then, the confidence interval will be $(72.51 + \sqrt{3.47}qt(0.005, 5), 72.51 + \sqrt{3.47}qt(0.995, 5)) = (65.00, 82.02)$. The observation 7 falls outside the prediction interval. The significance level is around 2e-4.

3.31

The original scale seems not suitable to use linear model to fit. One therapy is use transformation

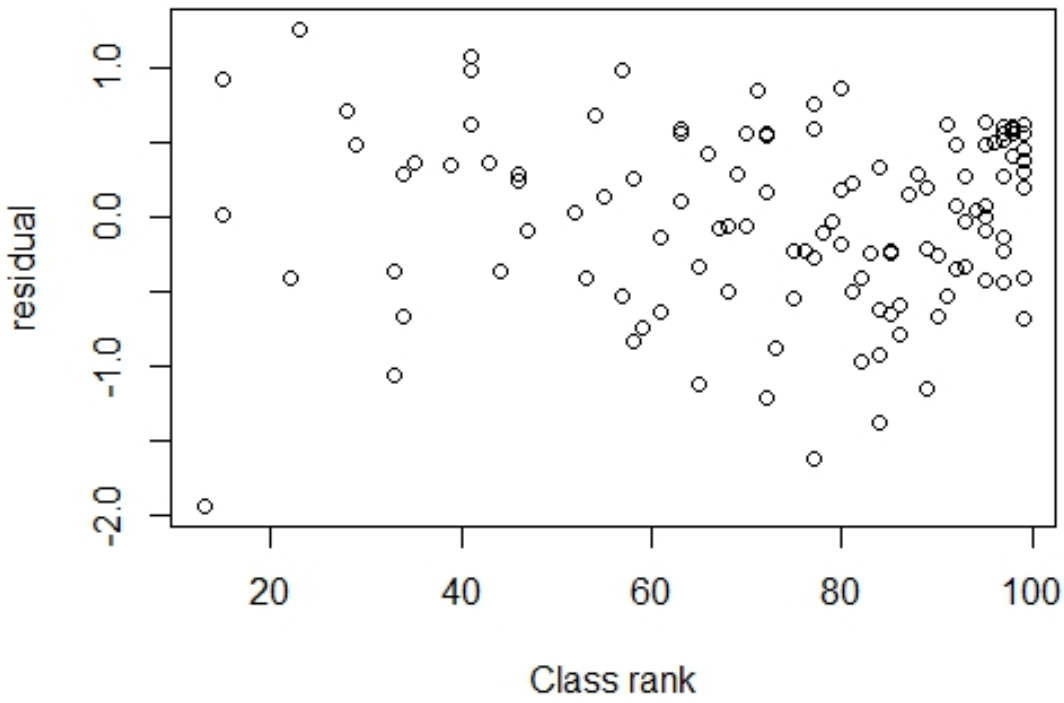


Figure 5: 3.3f

on Y and X. Based on my attempts, a better model candidate could be $y^{\lambda} = \beta_0 + \beta_1 \log x + \epsilon$. I use boxcox function in R to choose the best λ which is equal to -0.43. Then, I got the estimated function $y^{-0.43} = -0.047 \log x + 0.13$. Both coefficients are significant. (Notice: I use one unit of y is K dollars and unit of x is 1000 square feet.) Codes are available in appendix, the diagnosis plots is available after copying my codes in R. The estimated value at $X = 1100$ is 118K dollars and the estimated value at $X = 4900$ is 784K dollars. The final model is little bit complicated. The interpretation is not straightforward.

3.32

The original scale is not suitable. It is better to choose log-log scale. After doing log transformation. Then, the relationship is already almost linear. To make the model not over fitting, I would say it is OK to take this one. Couple of diagnosis plots are available in appendix after implementing code. The estimated function is $\log y = 0.72 \log x + 1.51$. The prediction at $X = 20$ is 38.93. The final model could be not that well-fitted, but simple to handle.

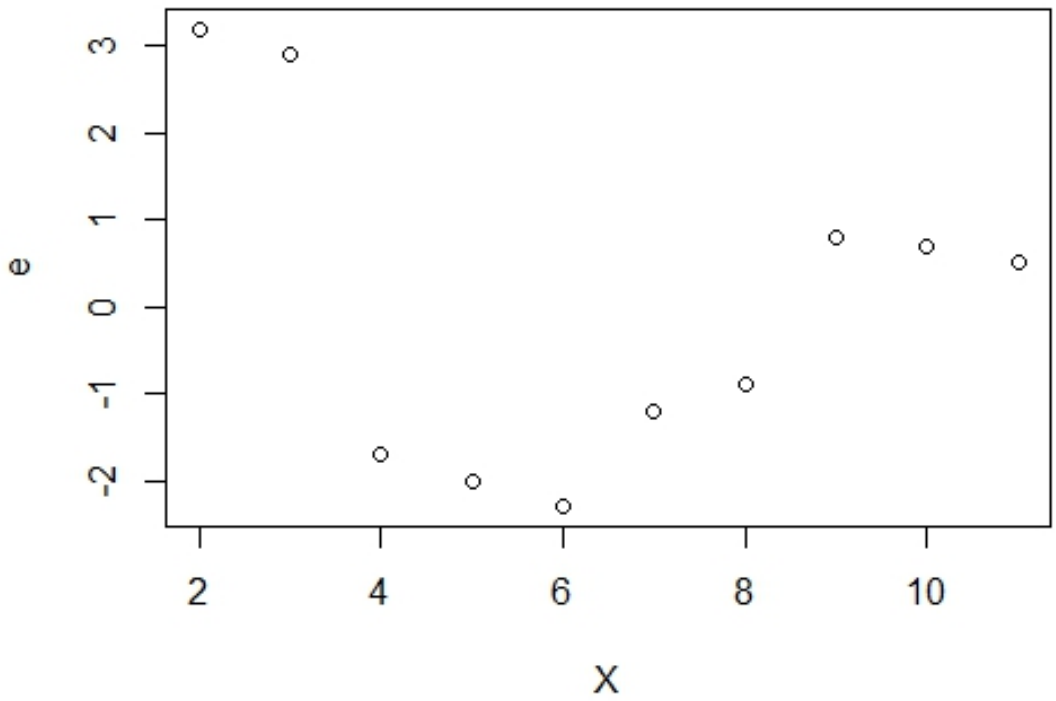


Figure 6: 3.9

A

```
Problem 3.31
# data called: data331
set.seed(5205);
randomset <- sample(1:522,200);
mydata <- data331[randomset,];
Price <- mydata$V2/1000;
Feet <- mydata$V3/1000;

# plot scatter plot
plot(Feet,Price, xlab = "Squared feet", ylab = "Sales price", main = "Scatter plot"
```

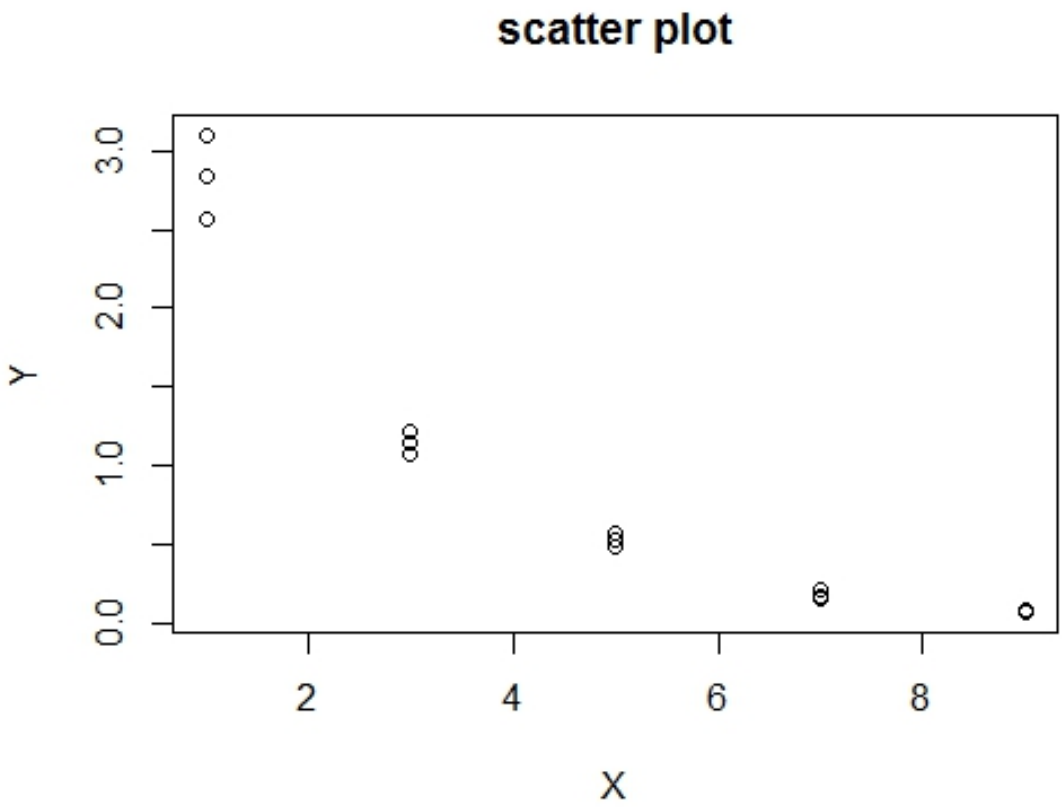


Figure 7: 3.16a

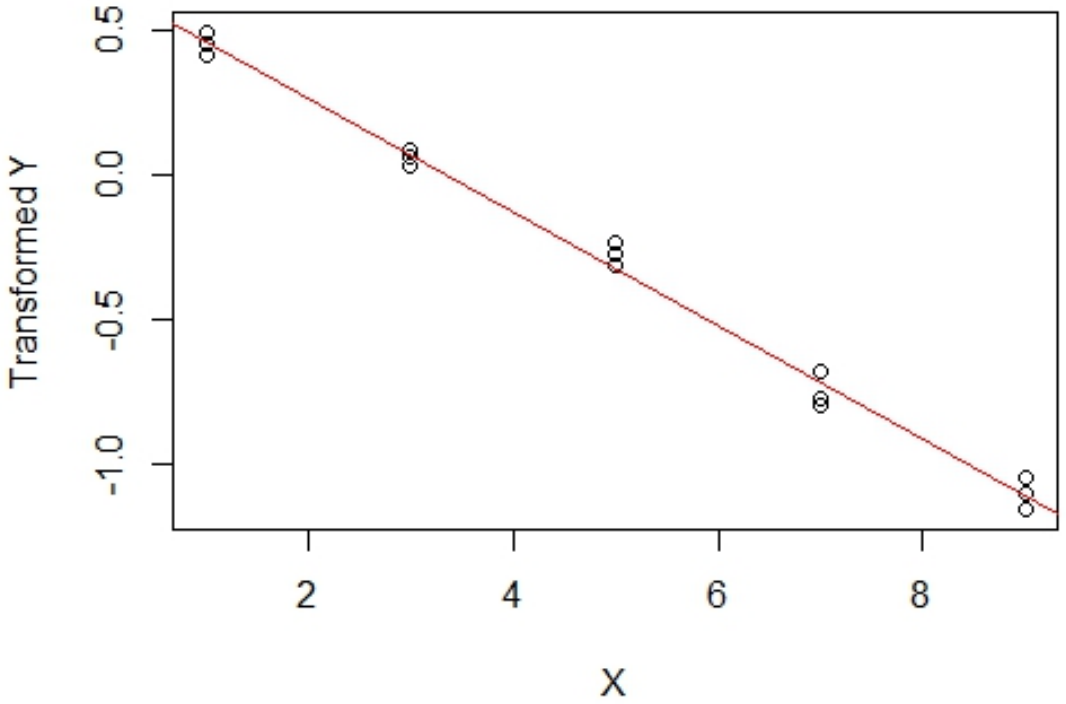
```
fit <- lm(Price~Feet)

# plot residual plot
plot(Feet, fit$residuals, xlab = "Squared feet", ylab = "residuals")

# qq plot
qqnorm(fit$residuals)

# use boxcox function
library(MASS)

# do not transform X
```




```
boxfit <- boxcox(Price ~ Feet, lambda = seq(-1,1, length = 10))
lambda.best <- boxfit$x[which.max(boxfit$y)]
# Seems not that good
plot(Feet, Price^lambda.best, xlab = "Squared Feet", ylab = "Transformed Sales")

# transform X
boxfit <- boxcox(Price ~ log(Feet), lambda = seq(-1,1, length = 10))
lambda.best <- boxfit$x[which.max(boxfit$y)]
plot(log(Feet), Price^lambda.best, xlab = "Squared Feet", ylab = "Transformed Sales

fit.best<-lm(Price^lambda.best~log(Feet))
# Acceptable
qqnorm(fit.best$residuals, main ="qqplot for transformed data")
```

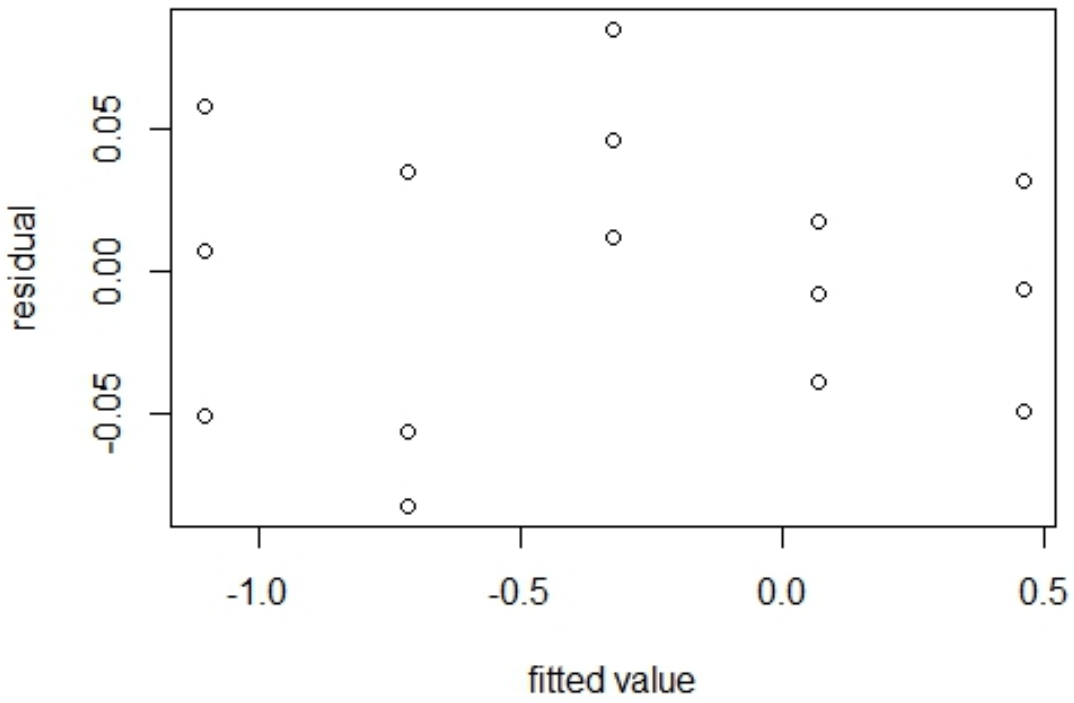


Figure 8: 3.16(e)

Problem 3.32

```
# data called: data332

PSA <- data332$V2
canvol <- data332$V3
plot(canvol, PSA, xlab = "cancer volume", ylab = "PSA  level")

# Apparently we need to do the log log transformation

plot(log(canvol), log(PSA), xlab = "cancer volume", ylab = "PSA  level")
```

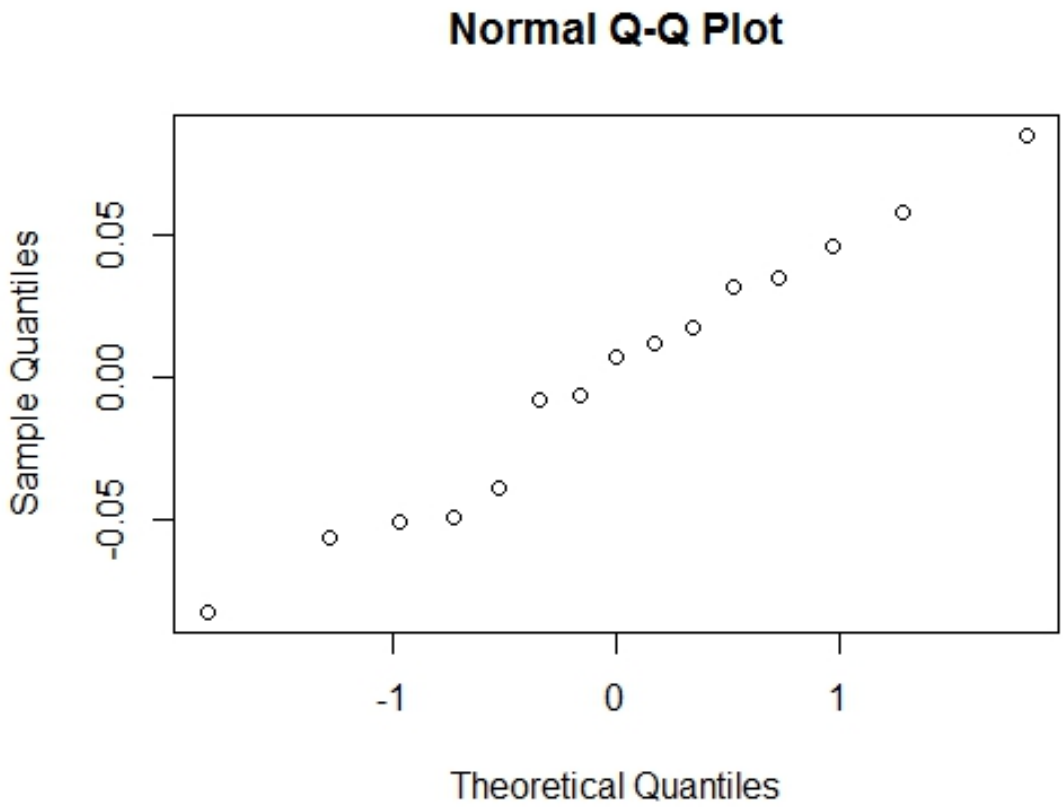


Figure 9: 3.16(e)

```
# It seems much better now


fit <- lm(log(PSA) ~ log(canvol))
plot(log(canvol), fit$residuals, xlab = "log of cancer volume", ylab = "residuals")
# residual plots seems good


qqnorm(fit$residuals)
# seems OK, not that bad
```

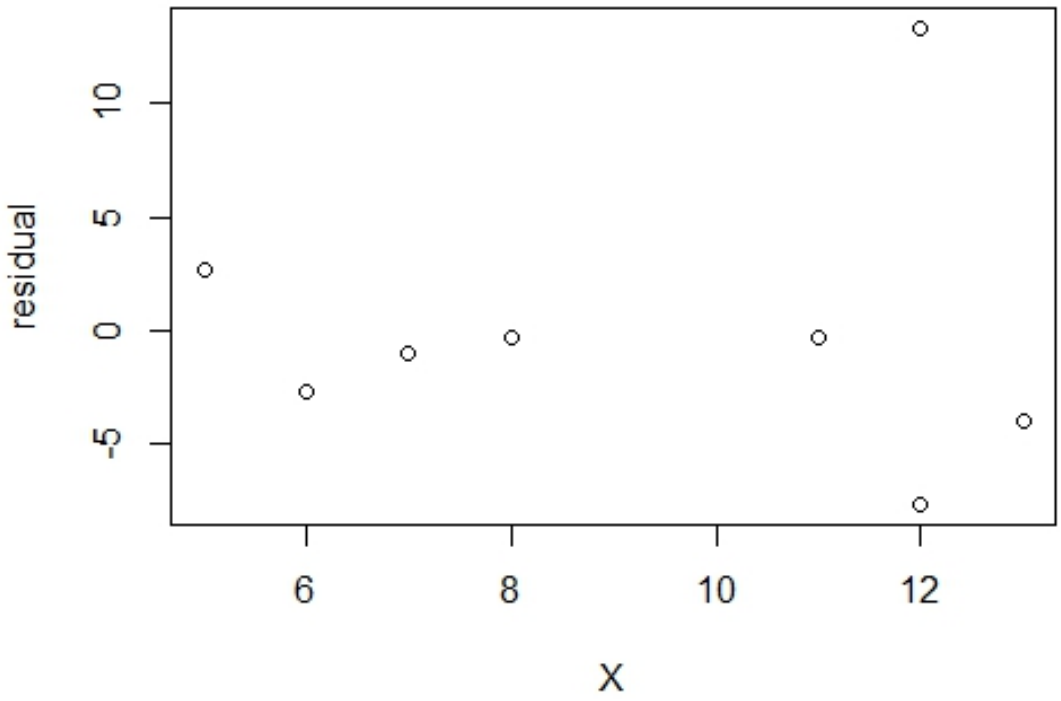


Figure 10: 3.16(a)