Math 327 Chapter 5 Homework

David Oniani

October 28, 2019

The data are from Montgomery, Peck, and Vining, Chapter 5, Exercise 5. A glass bottle manufacturing company recorded the average number of defects per 10,000 bottles due to stones (small pieces of rock embedded in the bottle wall) and the number of weeks since the last furnace overhaul. Number of defects will be modeled as a function of time (weeks since last overhaul).

```
# Load the data file, data_prob_5_5.Rdata

load ("./data_prob_5_5.Rdata")
```

Fit a linear model to the data and examine the residuals

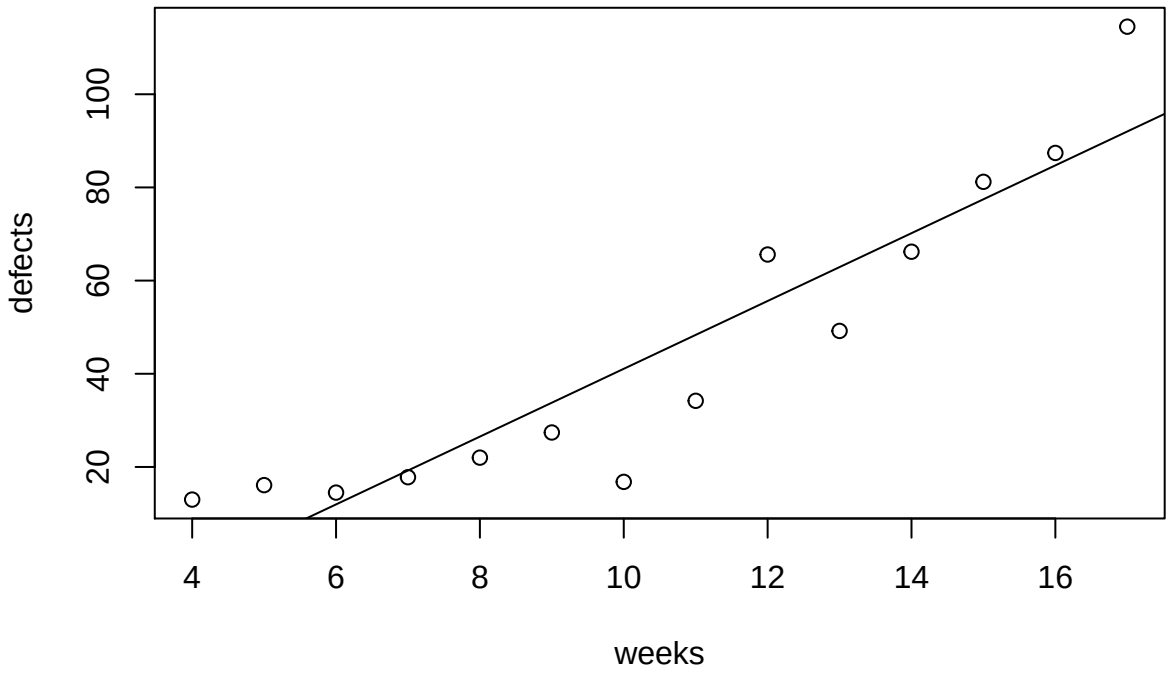
```
plot (defects ~ weeks, data=data_prob_5_5)
fit1 = lm (defects ~ weeks, data=data_prob_5_5)
summary (fit1)
```

```
##
## Call:
## lm(formula = defects ~ weeks, data = data_prob_5_5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.2688  -5.9229   0.5497   8.4203  22.4943
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -31.6982     9.7758  -3.243  0.00705 **
## weeks         7.2767     0.8692   8.372 2.35e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.11 on 12 degrees of freedom
## Multiple R-squared:  0.8538, Adjusted R-squared:  0.8416
## F-statistic: 70.09 on 1 and 12 DF,  p-value: 2.354e-06
```

```
confint (fit1)
```

```
##                   2.5 %     97.5 %
## (Intercept) -52.997842 -10.398642
## weeks         5.382938   9.170469
```

```
abline (fit1)
```



Q1: Describe the results of the fitted model.

The mean response value when weeks is 0 equals -31.698 defects and is between -52.998 defects and -10.399 defects with 95% confidence.

The estimated mean response value is 7.277 defects per weeks which changes between 5.383 defects per weeks and 9.170 defects per weeks, for any 1-unit increase in the predictor, with 95% confidence.

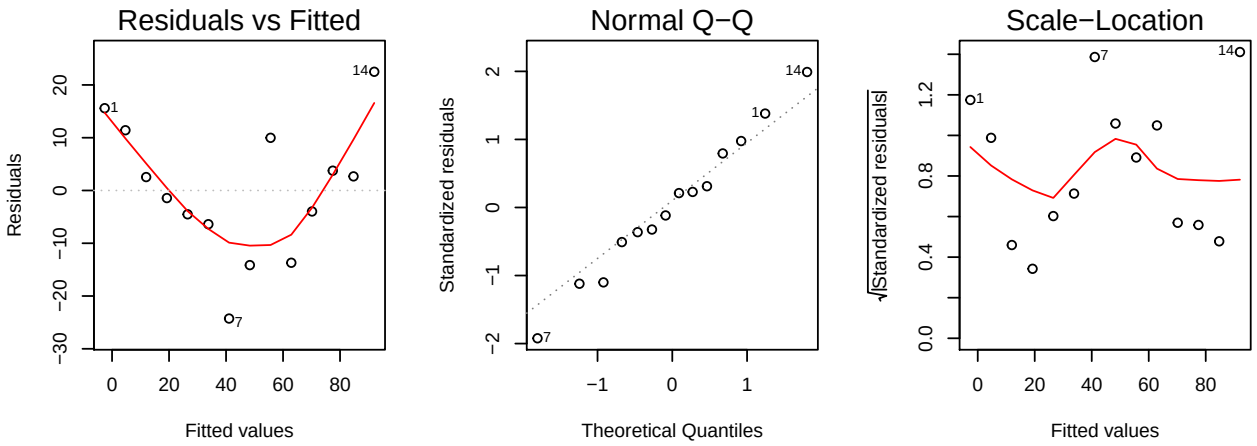
It should be noted that both intercept and slope are significant estimates with p-values of 0.007 and 2.350×10^{-6} respectively.

The Multiple R-squared value is 0.8538 which tells us that 85.38% of total variation in defects is explained by weeks.

Residual standard error is 13.11 meaning that on average, predictions of the model are 13.11 defects off from the actual value.

Examine the residuals from the linear fit

```
par (mfrow=c(1,3))
plot (fit1, which=1:3)
```



Q2: Interpret the residual plots.

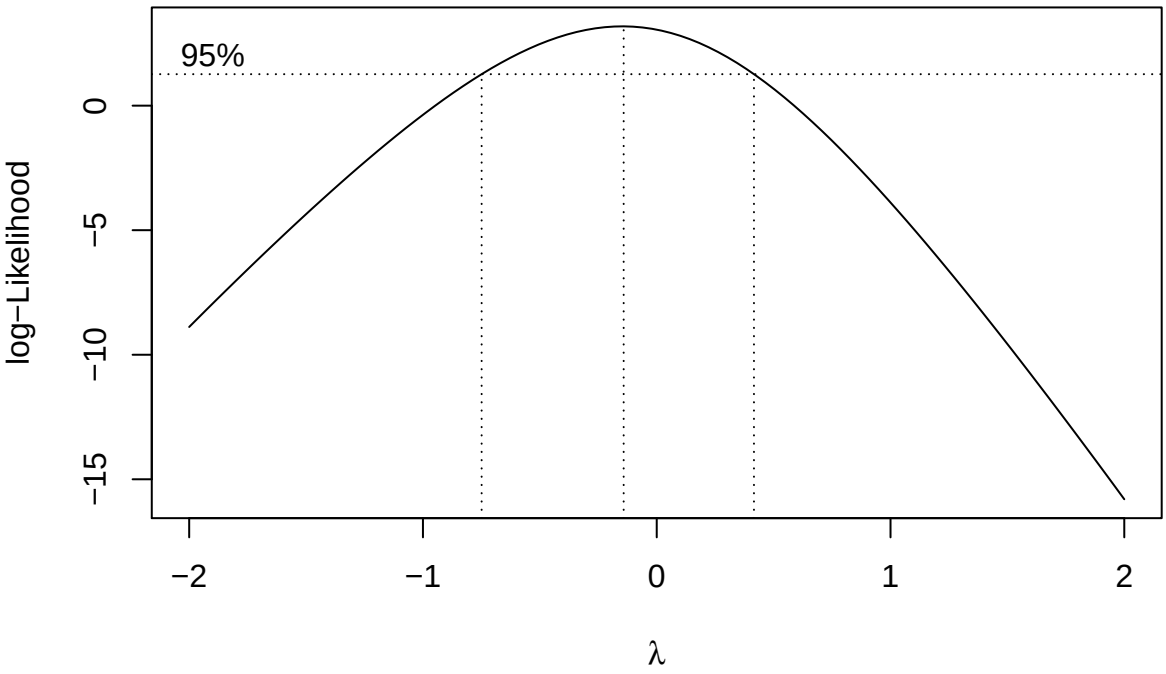
The first plot (Residuals VS Fitted Values) shows a non-constant spread. There is a strong evidence of heteroscedasticity.

The second plot (Standardized Residuals VS Theoretical Quantiles) shows a weak evidence that both sets of quantiles are coming from normal distributions. This is the case since the points are, to some degree, aligned on the line, but again, this visual check is not a strong evidence. Besides, the alignment across the line is irregular and most of the points are missed.

The third plot ($\sqrt{\text{Standardized Residuals}}$ VS Fitted values) shows us a set of lines which are very far from forming a straight line. This plot shows that the residuals are not spread equally along the ranges of predictors and that the variance is not constant.

Do a Box-Cox analysis

```
# NOTE: the MASS package must be installed.
MASS::boxcox (fit1)
```



Q3: Interpret the Box-Cox plot.

Box-Cox plot has the form of a flipped parabola ($f(x) = -x^2$). The λ (lambda) value is approximately -0.1 and is between -0.8 and 0.4 with 95% confidence. These observations suggest the **nonnormality of errors** in a linear model. Therefore, it is reasonable to do a log transformation of the response variable (defects). This would help us normalize the errors as well as address the non-linearity of the distribution.

Fit Log(Defects) vs. Weeks

```
fit2 = lm (log(defects) ~ weeks, data=data_prob_5_5)
summary (fit2)
```

```
##
## Call:
## lm(formula = log(defects) ~ weeks, data = data_prob_5_5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62990 -0.06982  0.00977  0.07727  0.38529
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.71622    0.17311   9.914 3.93e-07 ***
## weeks        0.17351    0.01539  11.273 9.68e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2322 on 12 degrees of freedom
## Multiple R-squared:  0.9137, Adjusted R-squared:  0.9065
## F-statistic: 127.1 on 1 and 12 DF,  p-value: 9.676e-08
```

```
confint (fit2)
```

```
##                 2.5 %    97.5 %
## (Intercept) 1.3390378 2.0934079
## weeks       0.1399697 0.2070414
```

Q4: Interpret the results of fit2.

The mean response value when weeks is 0 equals 1.716 `log`(defects) and is between 1.339 `log`(defects) and 2.093 `log`(defects) with 95% confidence.

The estimated mean response value is 0.174 `log`(defects) per weeks which changes between 0.140 `log`(defects) per weeks and 0.207 `log`(defects) per weeks, for any 1-unit increase in the predictor, with 95% confidence.

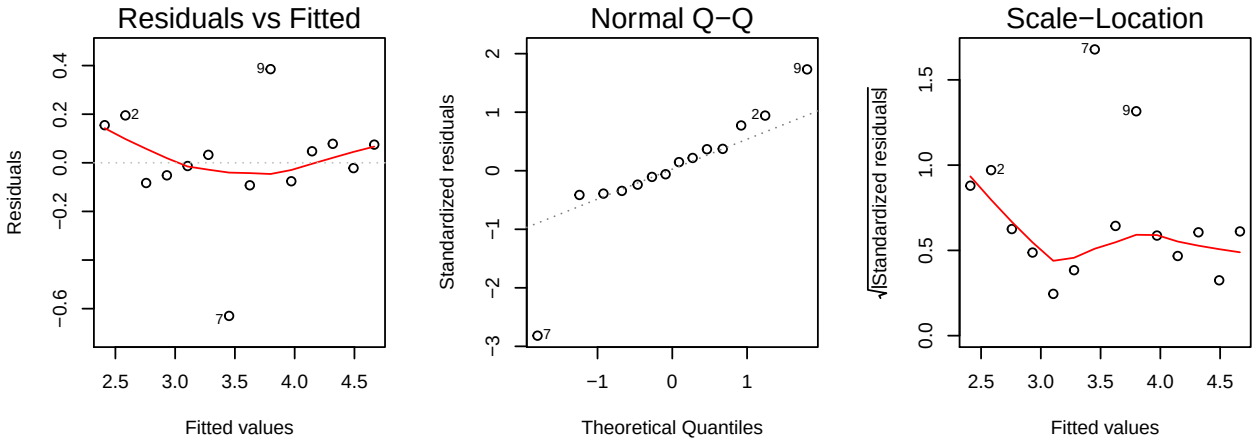
It should be noted that both intercept and slope are significant estimates with p-values of 3.930×10^{-7} and 9.680×10^{-8} respectively.

The Multiple R-squared value is 0.9137 which tells us that 91.37% of total variation in defects is explained by weeks.

Residual standard error is 0.2322 meaning that on average, predictions of the model are 0.2322 `defects` off from the actual value.

Residual analysis of fit2

```
par (mfrow=c(1,3))
plot (fit2, which=1:3)
```



Q5: Interpret the residual plots for fit2.

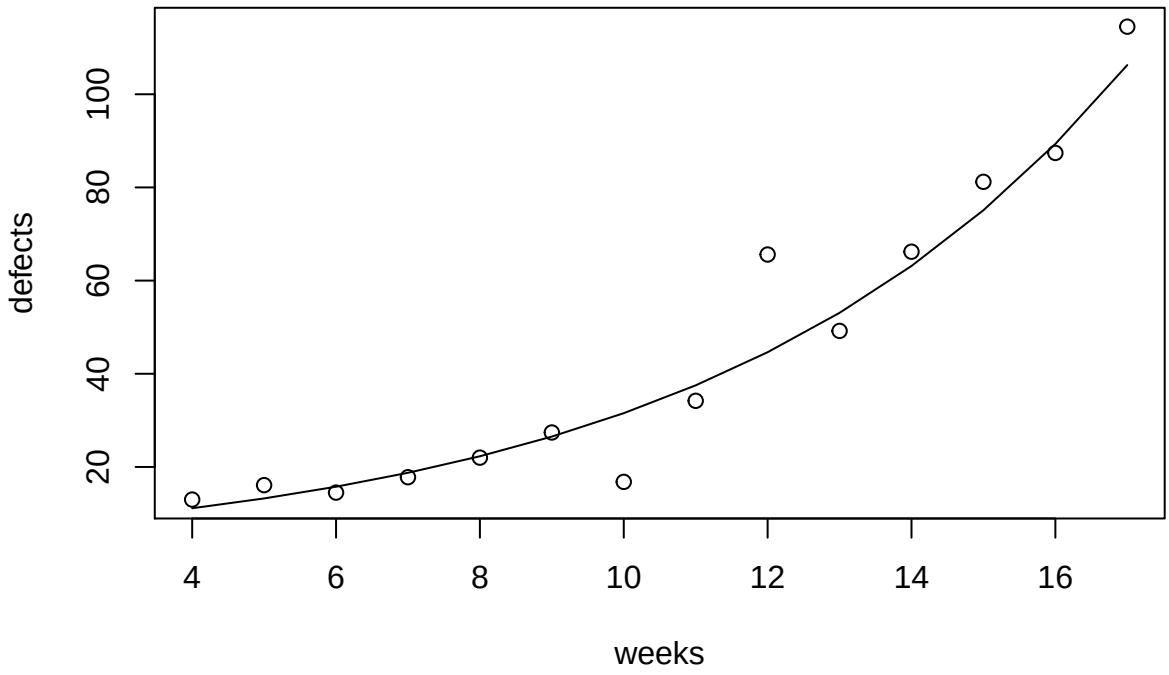
The first plot (Residuals VS Fitted Values) shows a non-constant spread. There is a weak evidence of heteroscedasticity. This model does seem to be better than the previous model.

The second plot (Standardized Residuals VS Theoretical Quantiles) shows a rather strong evidence that both sets of quantiles are coming from normal distributions. This is the case since the points are, to some degree, aligned on the line. Overall, this does seem to be better than the previous model as most points are aligned on the line.

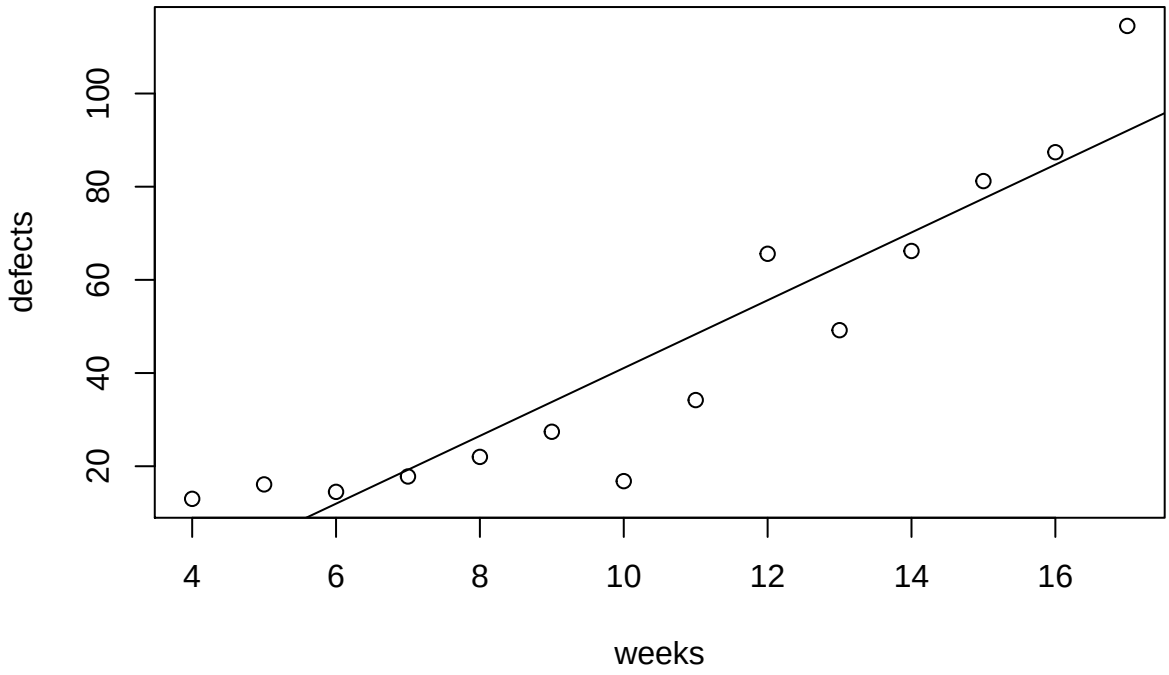
The third plot ($\sqrt{\text{Standardized Residuals}}$ VS Fitted values) shows us a set of lines which are very far from forming a straight line. That said, the plot shows that the residuals are spread equally along the ranges of predictors and that the variance is nearly constant. Besides, the square root of Standardized Residuals show the decreasing trend. This does seem to be better than the previous model.

Plot fit2 on the original scale of defects vs weeks

```
plot (defects ~ weeks, data=data_prob_5_5)
pred.logscale = predict (fit2)
pred.origscale = exp (pred.logscale)
ord.weeks = order (data_prob_5_5$weeks)
lines (data_prob_5_5$weeks [ord.weeks], pred.origscale[ord.weeks])
```




```
# The code chunk was not here initially
# It was added for the sake of comparing new and initial models
plot(defects ~ weeks, data=data_prob_5_5)
abline(lm(defects ~ weeks, data=data_prob_5_5))
```



Q6: Comment on the plot above. What do you see? What do you wonder?

There certainly is either quadratic or, more likely, a logarithmic relationship between defects and weeks. There seem to be a couple outliers. Introducing either quadratic term or doing a log-transform would probably not be a bad idea. I wonder why the relationship has such shape. Besides, this is not a linear regression as we fitted a curve instead of the line. That said the new fit is a lot better than the initial model.