Homework 6 Solutions

Due: Thursday 4/02/20 by 8:30am

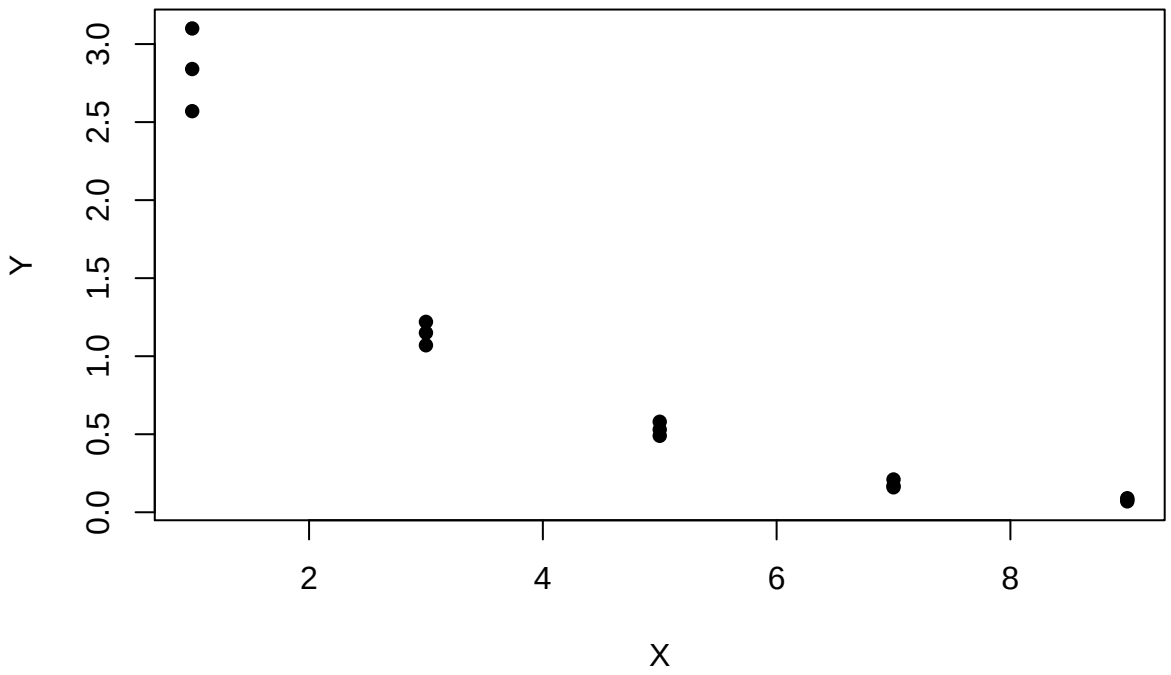
Rubric:

- Maximum of 2 points each for 5. and 6., determined as follows:
  - 0 points for no solutions whatsoever or incomplete solutions;
  - 1 point for solutions provided for each part, but at least one incorrect solution;
  - 2 points for correct solutions to each part;
- Maximum of 3 points for 1., 2., 3., and 4., determined as follows:
  - 0 points for no solutions whatsoever or `R` output only;
  - 1 point for an honest effort but very few correct answers or `R` output only plus a figure;
  - 2 points for mostly correct answers but at least one substantial issue;
  - 3 points for nearly/exactly correct.

1. Problem 3.16 from the `.pdf` version of the textbook, parts (a), (c)-(f). Requires use of the `solution_concentration` data that has been posted on the Homework page.

(a)

```
link <- url("http://maryclare.github.io/stat525/content/homework/solution_concentration.RData")
load(link)
close(link)
Y <- data$Y
X <- data$X
plot(X, Y, pch = 16)
```



I might try using a log transformation of Y.

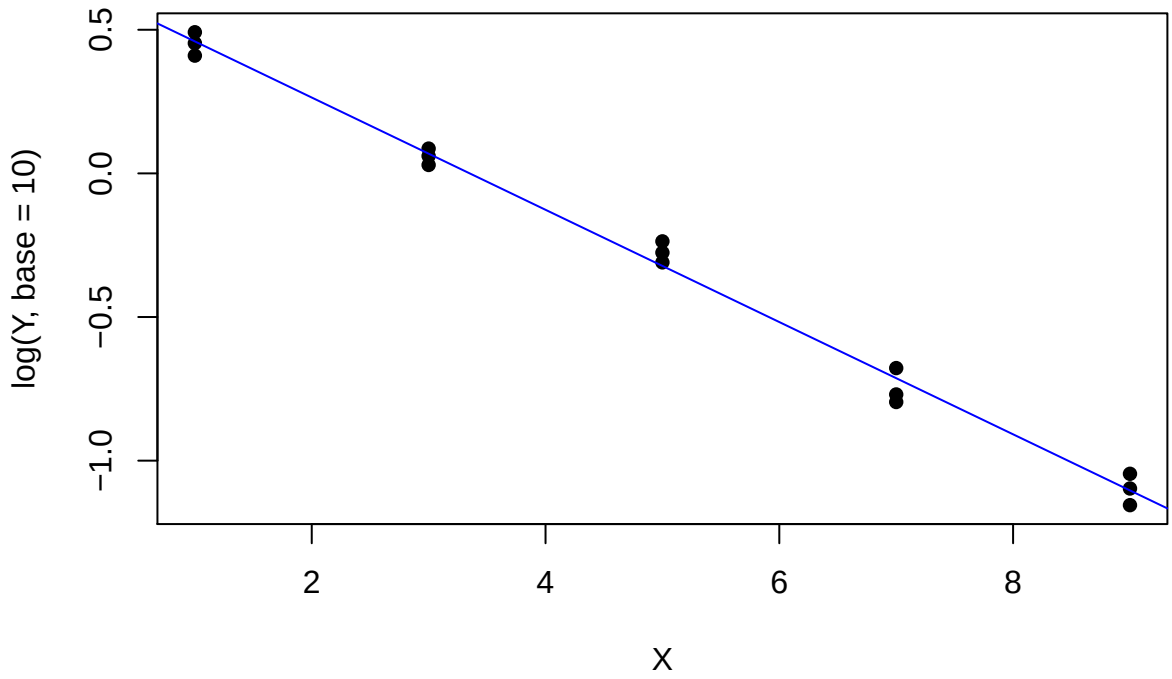
(c)

```
linmod <- lm(I(log(Y, base = 10)~X))
b0 <- linmod$coef[1]
b1 <- linmod$coef[2]
```

The estimated linear regression function is $0.65 + -0.2X_i$.

(d)

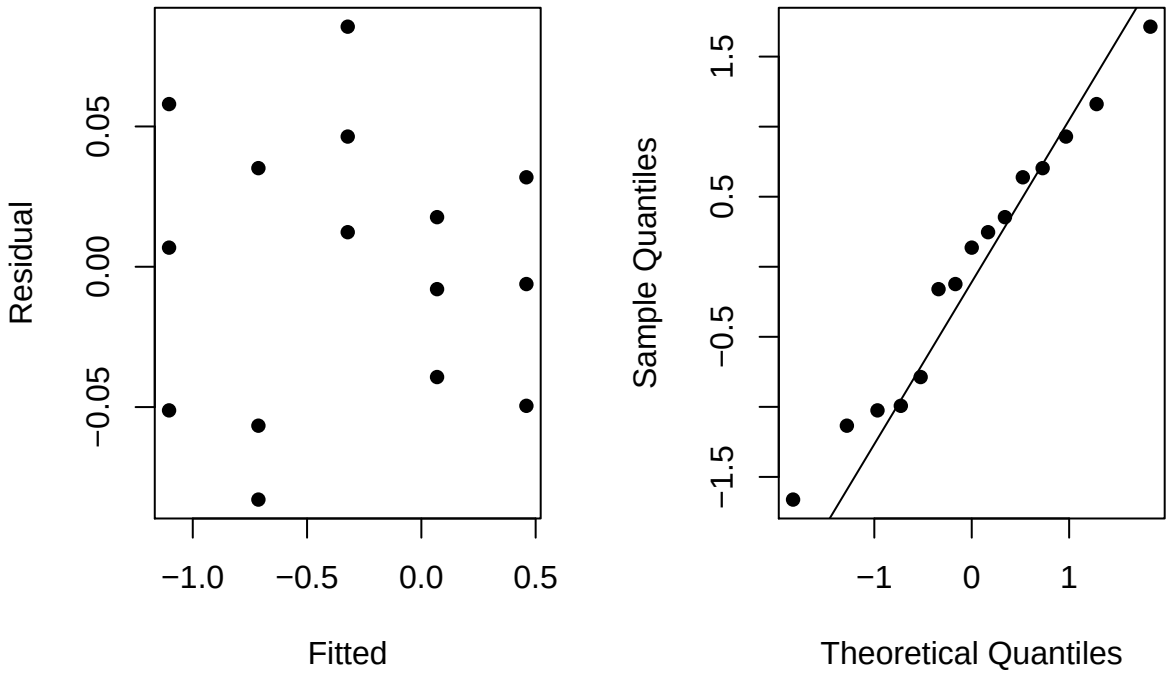
```
plot(X, log(Y, base = 10),
     pch = 16)
abline(a = b0, b = b1, col = "blue")
```



The regression line appears to be a very good fit to the transformed data.

(e)

```
par(mfrow = c(1, 2))
plot(linmod$fitted.values, linmod$residuals,
     pch = 16,
     xlab = "Fitted", ylab = "Residual")
qqnorm(linmod$residuals/summary(linmod)$sigma,
       pch = 16, main = "")
qqline(linmod$residuals/summary(linmod)$sigma)
```



The plot of the residuals against the fitted values indicates that the transformation yields residuals with approximately constant variance. The normal probability of plot of the residuals indicates thta most of the residuals are consistent with what we would expect if the model were true.

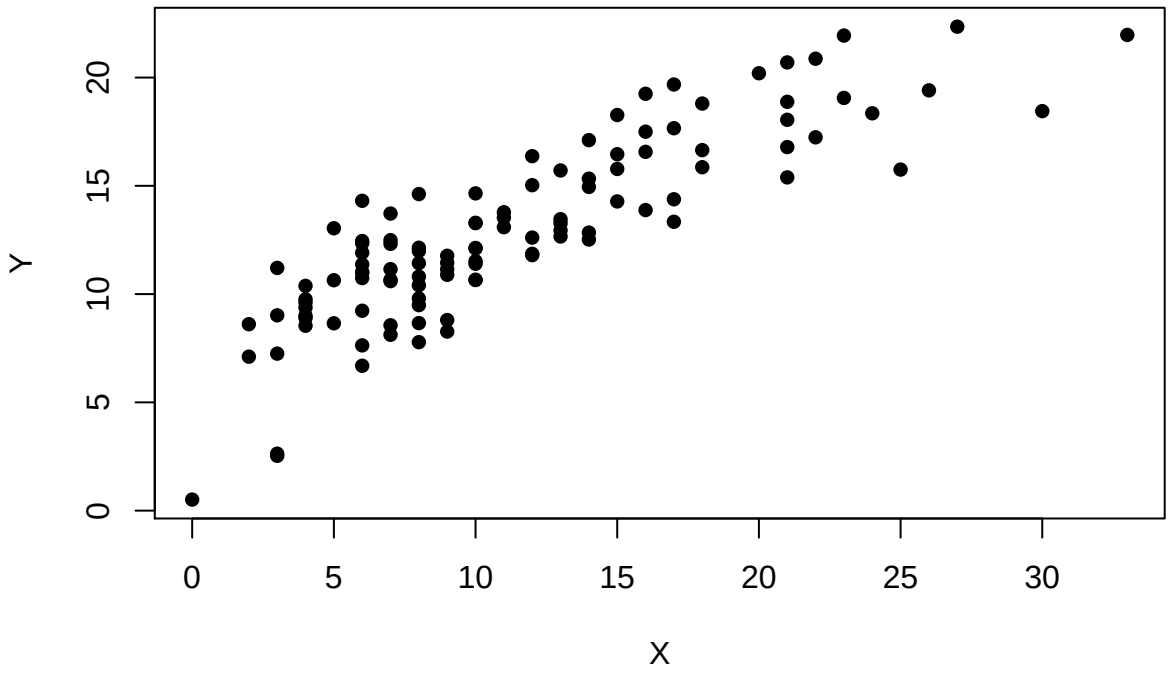
(f)

The estimated regression function in the original units is $4.52 \times 10^{-0.2X_i}$.

2. Problem 3.18 from the `.pdf` version of the textbook. Requires use of the `production_time` data that has been posted on the Homework page.

(a)

```
link <- url("http://maryclare.github.io/stat525/content/homework/production_time.RData")
load(link)
close(link)
Y <- data$Y
X <- data$X
plot(X, Y, pch = 16)
```



A linear relation does not appear adequate here. Rather, the relation between X and Y appears to be curvilinear. Because there is little evidence of variance nonstationarity, it would make more sense to transform X as opposed to Y.

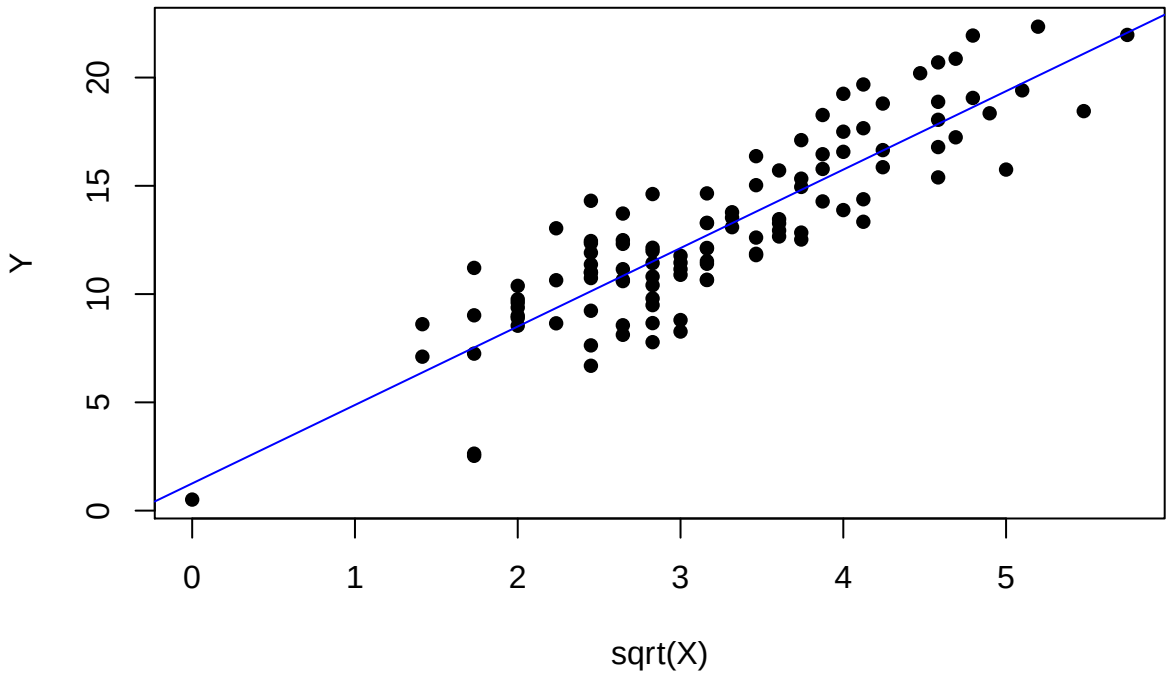
(b)

```
linmod <- lm(Y~I(sqrt(X)))
b0 <- linmod$coef[1]
b1 <- linmod$coef[2]
```

The estimated linear regression function is $1.25 + 3.62\sqrt{X_i}$.

(c)

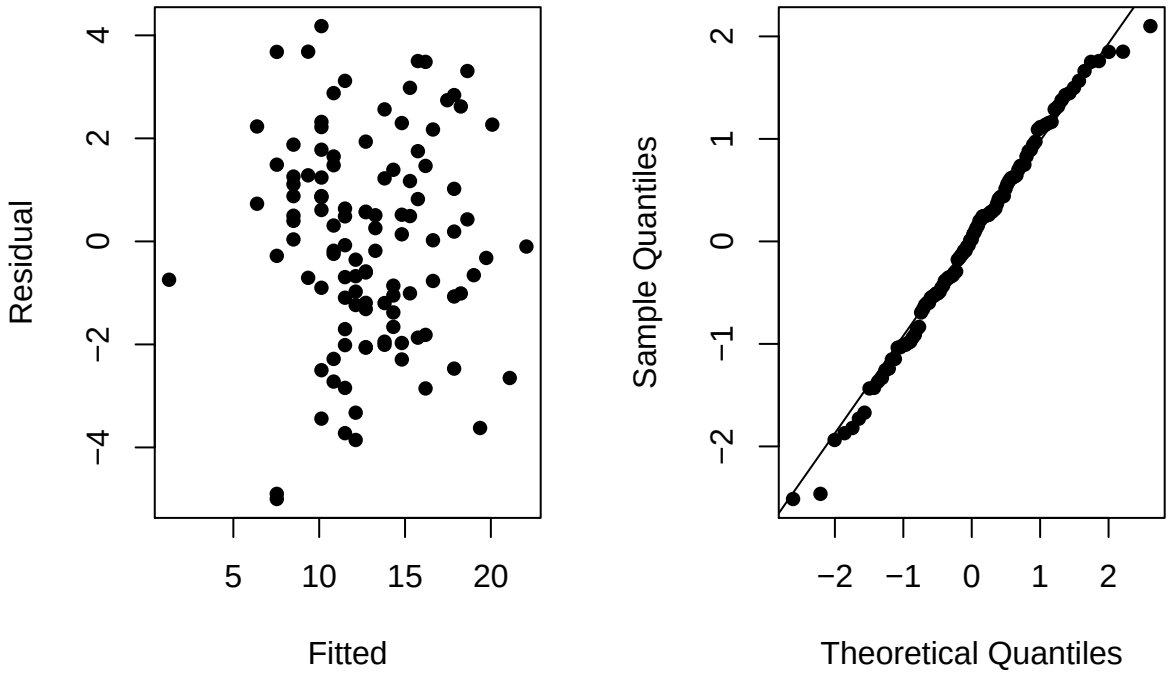
```
plot(sqrt(X), Y,
     pch = 16)
abline(a = b0, b = b1, col = "blue")
```



The regression line appears to be a very good fit to the transformed data.

(d)

```
par(mfrow = c(1, 2))
plot(linmod$fitted.values, linmod$residuals,
     pch = 16,
     xlab = "Fitted", ylab = "Residual")
qqnorm(linmod$residuals/summary(linmod)$sigma,
       pch = 16, main = "")
qqline(linmod$residuals/summary(linmod)$sigma)
```



The plot of the residuals against the fitted values indicates that the transformation yields residuals with approximately constant variance. The normal probability of plot of the residuals indicates thta most of the residuals are consistent with what we would expect if the model were true.

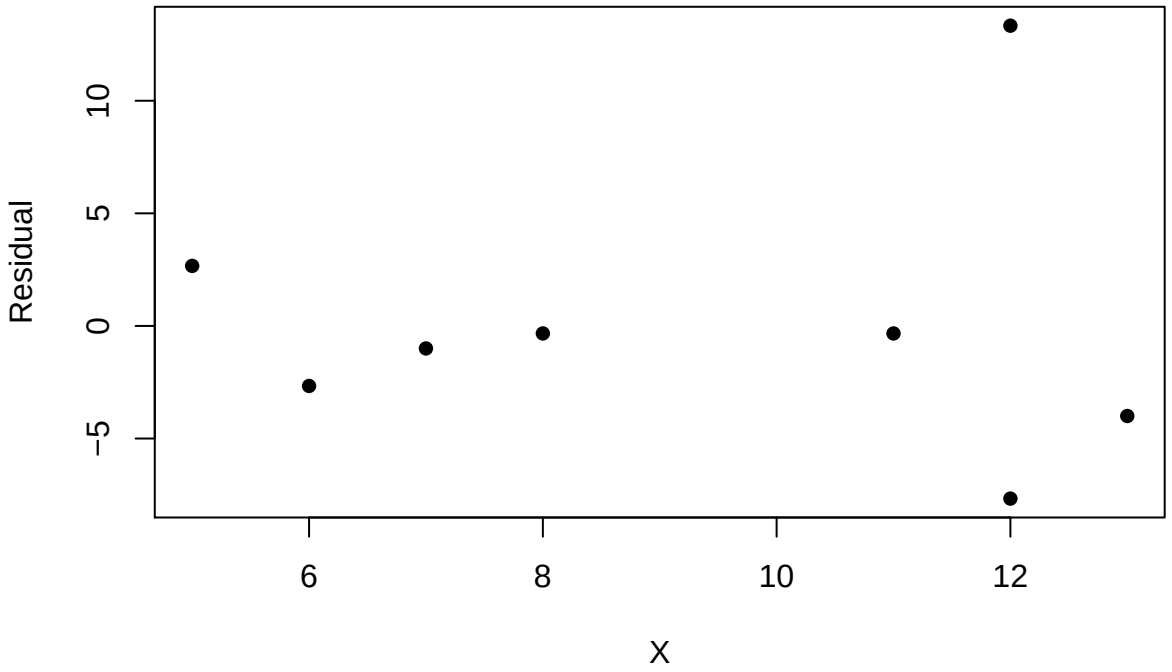
(e)

The estimated regression function in the original units is the same, because we did not transform the response: $1.25 + 3.62\sqrt{X_i}$.

3. Problem 3.24 from the `.pdf` version of the textbook. Requires use of the `blood_pressure` data that has been posted on the Homework page.

(a)

```
link <- url("http://maryclare.github.io/stat525/content/homework/blood_pressure.RData")
load(link)
close(link)
Y <- data$Y
X <- data$X
n <- length(Y)
linmod <- lm(Y~X)
b0 <- linmod$coefficients[1]
b1 <- linmod$coefficients[2]
plot(X, linmod$residuals, pch = 16,
     ylab = "Residual")
```



We obtain the estimated regression function $48.67 + 2.33X_i$.

(b)

```
linmod.no7 <- lm(Y[-7]~X[-7])
b0 <- linmod.no7$coefficients[1]
b1 <- linmod.no7$coefficients[2]
```

We obtain the estimated regression function $53.07 + 1.62X_i$.

(c)

```
s <- summary(linmod.no7)$sigma
s.pred <- s*sqrt(1 + 1/(n - 1) + (12 - mean(X[-7]))^2/(sum((X[-7] - mean(X[-7]))^2)))
hat <- b0 + b1*12
alpha <- 0.01
low <- hat + qt(alpha/2, (n - 1) - 2)*s.pred
hig <- hat + qt(1 - alpha/2, (n - 1) - 2)*s.pred
# Note - this could also be obtained with
# predict(linmod.no7, data.frame("X" = 12), interval = "prediction", level = 1 - alpha)
```

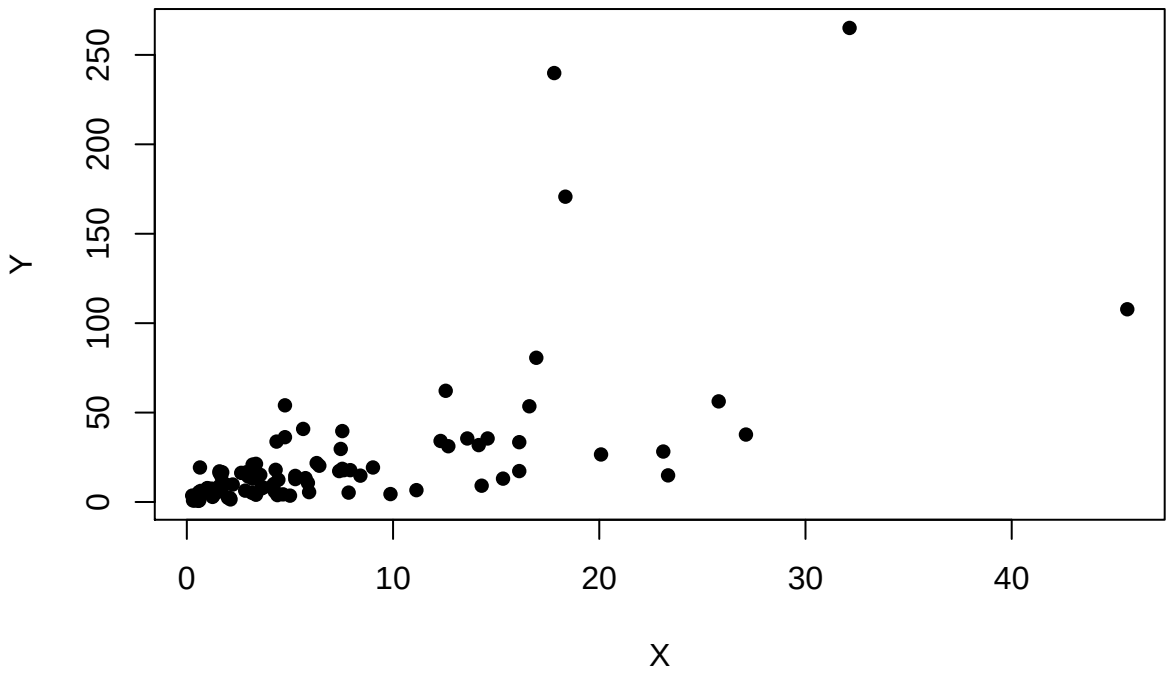
A 99 percent prediction interval for a new Y value at $X = 12$ is given by $(60.31, 84.74)$. The observed value $Y_7 = 90$ falls outside of this prediction interval. This suggests that Y_7 is an outlier that is not well explained by the same model as the rest of the observed data and requires further investigation.,

4. Problem 3.32 from the `.pdf` version of the textbook. When the question asks you to "include and justify appropriate remedial measures," explain what modifications you might make to address assumptions that may not be met, e.g. transformations of X, transformations of Y, or exclusion of one or more outliers. Requires use of the `prostate_cancer` data that has been posted on the Homework page.

```
link <- url("http://maryclare.github.io/stat525/content/homework/prostate_cancer.RData")
load(link)
close(link)
Y <- data$Y
X <- data$X
n <- length(Y)
```

First, we make a scatter plot of the observed data.

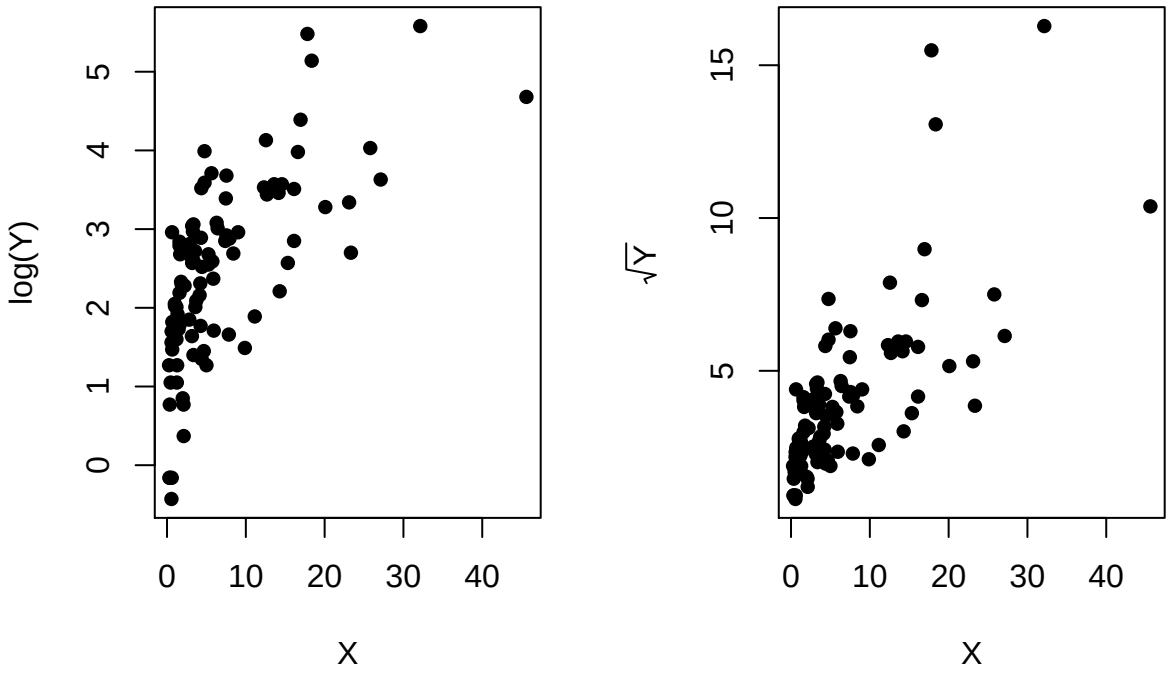
```
plot(X, Y, pch = 16)
```



It is clear that there is evidence of nonconstant variance, so we will consider several transformations of the Y to see if the variance can be stabilized with a simple transformation. We consider taking the logarithm of the response Y and the square root of the response Y.

```
par(mfrow = c(1, 2))
plot(X, log(Y), pch = 16)
```

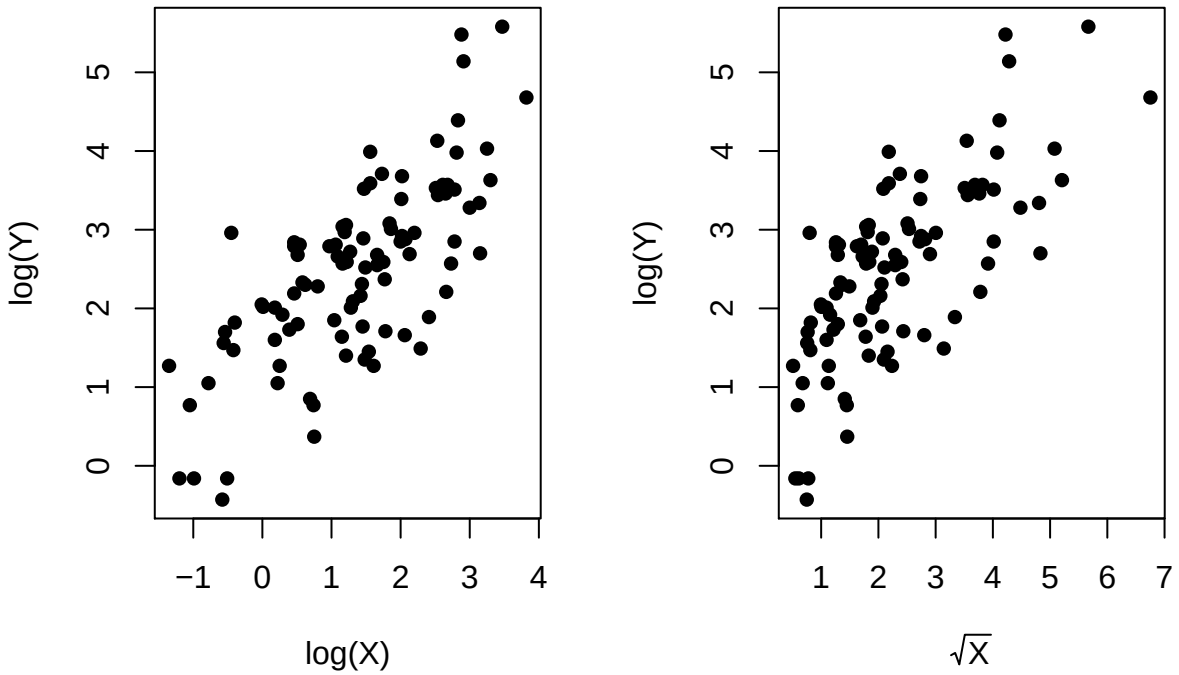
```
plot(X, sqrt(Y), pch = 16,
     ylab = expression(sqrt(Y)))
```



Based on these plots, it appears that taking the logarithm of the response $\log(Y)$ provides more variance stabilization than taking the square root of the response $\sqrt{Y}$, so we will consider the logarithm of the response going forward. In the scatter plot of the covariate X against the logarithm of the response $\log(Y)$, we see evidence of a nonlinear relationship.

To address this, we consider transforming the covariate X. Specifically, we make scatterplots of the logarithm of X against the response Y and the square root of X against the response Y.

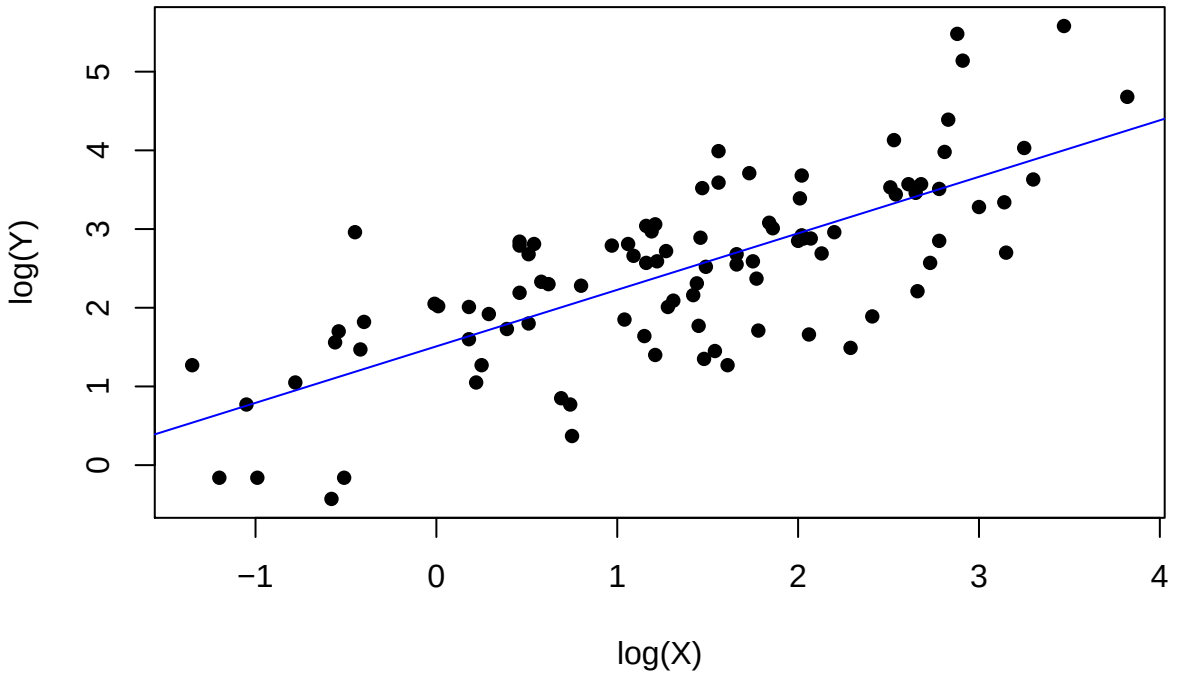
```
par(mfrow = c(1, 2))
plot(log(X), log(Y), pch = 16)
plot(sqrt(X), log(Y),
     xlab = expression(sqrt(X)), pch = 16)
```



Based on these plots, it appears that taking the logarithm of the covariate X provides a more linear relationship between the transformed covariate and response $\log(Y)$.

We fit a simple linear regression model to the logarithm of the response $\log(Y)$ and the logarithm of the covariate $\log(X)$. A scatterplot of the data with the estimated regression function is provided.

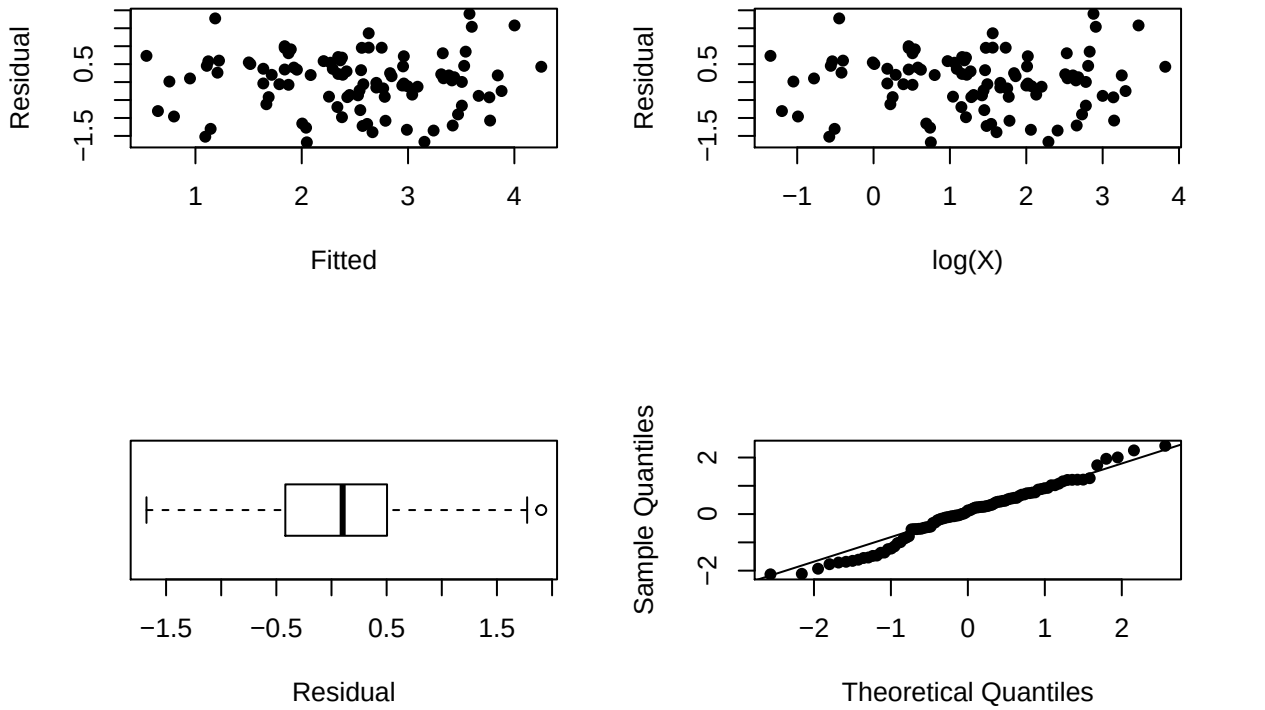
```
linmod <- lm(I(log(Y))~I(log(X)))
plot(log(X), log(Y), pch = 16)
abline(a = linmod$coefficients[1],
       b = linmod$coefficients[2], col = "blue")
```



Diagnostic scatter plots of the residuals from the final model against the corresponding fitted values and the

logarithm of the covariate $\log(X)$ indicate that the main strengths of this model; there is no strong evidence of remaining variance non-stationarity or nonlinearity. However, the final model does have some weaknesses. A box plot of the residuals indicates that the residuals are roughly symmetric about zero, with slightly more negative residuals than we might expect if the errors were normally distributed. This is more clearly evident in the normal probability plot of the residuals; residuals greater than zero are generally consistent with what we would expect if the errors were normally distributed, but residuals smaller than zero tend to be smaller than we would expect if the errors were normally distributed. Altogether, this suggests that a simple linear regression model to the logarithm of the response $\log(Y)$ and the logarithm of the covariate $\log(X)$ is a reasonable choice, although there is some slight evidence of violations of the normality assumption.

```
par(mfrow = c(2, 2))
plot(linmod$fitted.values, linmod$residuals,
     xlab = "Fitted",
     ylab = "Residual",
     pch = 16)
plot(log(X), linmod$residuals,
     ylab = "Residual",
     pch = 16)
boxplot(linmod$residuals, horizontal = TRUE,
        xlab = "Residual")
qqnorm(linmod$residuals/(summary(linmod)$sigma),
       main = "", pch = 16)
qqline(c(linmod$residuals/(summary(linmod)$sigma)))
```




```
fit.ci <- predict(linmod, data.frame("X" = 20),
                  interval = "confidence", level = 1 - 0.05)
orig.scale.fit.ci <- exp(fit.ci)
```

We obtain an estimate of the mean PSA level for a patient whose cancer volume is 20 cc of approxiamtely 38.9, and a corresponding approximate 95% confidence interval of $(29.58, 51.15)$.

5. Problem 6.1 from the `.pdf` version of the textbook.

(a)

$$\begin{pmatrix} 1 & X_{11} & X_{11}X_{12} \\ 1 & X_{21} & X_{21}X_{22} \\ 1 & X_{31} & X_{31}X_{32} \\ 1 & X_{41} & X_{41}X_{42} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

(b)

$$\begin{pmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ 1 & X_{31} & X_{32} \\ 1 & X_{41} & X_{42} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

6. Problem 6.2 from the `.pdf` version of the textbook.

(a)

There are two possible answers to this problem. One is

$$\begin{pmatrix} 1 & X_{11} & X_{12} & X_{11}^2 \\ 1 & X_{21} & X_{22} & X_{21}^2 \\ 1 & X_{31} & X_{32} & X_{31}^2 \\ 1 & X_{41} & X_{42} & X_{41}^2 \\ 1 & X_{51} & X_{52} & X_{51}^2 \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} 0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Another is

$$\begin{pmatrix} X_{11} & X_{12} & X_{11}^2 \\ X_{21} & X_{22} & X_{21}^2 \\ X_{31} & X_{32} & X_{31}^2 \\ X_{41} & X_{42} & X_{41}^2 \\ X_{51} & X_{52} & X_{51}^2 \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

(b)

$$\begin{pmatrix} 1 & X_{11} & \log_{10}(X_{12}) \\ 1 & X_{21} & \log_{10}(X_{22}) \\ 1 & X_{31} & \log_{10}(X_{32}) \\ 1 & X_{41} & \log_{10}(X_{42}) \\ 1 & X_{51} & \log_{10}(X_{52}) \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$