Homework 5 Solutions

Due: Thursday 3/26/20 by 8:30am

Rubric:

- Maximum of 2 points each for 1., 2., and 5., 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 3., 4., 6., and 7., 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 2.10 from the `.pdf` version of the textbook.

(a) A **prediction** interval is appropriate, because we are interested in the value of a new observation.
(b) A **confidence** interval is appropriate, because we are interested in the average value of new observations with the same disposable income of $23,500.
(c) A **prediction** interval is appropriate, because we are interested in the value of a new observation.

2. Problem 2.11 from the `.pdf` version of the textbook.

There is a difference - we generally have more uncertainty about one or a finite number of new observations at $X = X_h$ than about the average across infinitely many observations at $X = X_h$, which would be the mean response at $X = X_h$.

3. Problem 2.28 from the `.pdf` version of the textbook, parts (a)-(b).

(a)

```
link <- url("http://maryclare.github.io/stat525/content/homework/muscle.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]
X.h <- 60
Y.hat.h <- b0 + b1*X.h
s.sq <- summary(linmod)$sigma^2

# Construct interval by hand
s.sq.Y.hat.h <- s.sq*(1/n + (X.h - mean(X))^2/(sum((X - mean(X))^2)))
int <- c(Y.hat.h + qt(0.025, n - 2)*sqrt(s.sq.Y.hat.h),
  Y.hat.h + qt(0.975, n - 2)*sqrt(s.sq.Y.hat.h))
# Note: This gives the same result as
```

```
int <- predict(linmod, newdata = data.frame("X" = 60),
               level = 0.95, interval = "confidence")[1, 2:3]
```

We obtain a 95 percent confidence interval of $(82.8347139, 87.0589529)$.

(b)

```
# Construct interval by hand
s.sq.Y.hat.h <- s.sq*(1/n + (X.h - mean(X))^2/(sum((X - mean(X))^2)))
int <- c(Y.hat.h + qt(0.025, n - 2)*sqrt(s.sq + s.sq.Y.hat.h),
  Y.hat.h + qt(0.975, n - 2)*sqrt(s.sq + s.sq.Y.hat.h))
# Note: This gives the same result as
int <- predict(linmod, newdata = data.frame("X" = 60),
               level = 0.95, interval = "prediction")[1, 2:3]
```

We obtain a 95 percent prediction interval of $(68.450669, 101.4429978)$.

4. Problem 2.31 from the `.pdf` version of the textbook, parts (a)-(b).

(a)

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

n <- length(Y)

linmod <- lm(Y~X)

ssto <- sum((Y - mean(Y))^2)
sse <- sum(linmod$residuals^2)
ssr <- sum((linmod$fitted.values - mean(Y))^2)

msto <- ssto/(n - 1)
mse <- sse/(n - 2)
msr <- ssr/1
```

Source of Variation	SS	df	MS
Regression	9.3462942×10^7	1	9.3462942×10^7
Error	4.5527317×10^8	82	5.5521118×10^6
Total	5.4873611×10^8	83	6.6112784×10^6

(b)

```
F.star <- msr/mse
F.quantile.01 <- qf(0.99, 1, n  - 2)
p.val <- pf(F.star, 1, n - 2, lower.tail = FALSE)
```

To decide whether or not there is a linear association between amount of muscle mass and age, we could conduct an F test of the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_a : \beta_1 \neq 0$. The decision rule for a level-$\alpha = 0.01$ test based on the test statistic $F^* = MSR/MSE$ would be:

- If $F^* \leq F(0.99; 1; 82)$, conclude H_0
- If $F^* > F(0.99; 1; 82)$, conclude H_a.

Because $F^* = 16.8337645$ and $F(0.99; 1; 82) = 6.9544199$, we would reject H_0 and conclude H_a.

The p-value of the test is $P(F > F^*) = 9.5713958 \times 10^{-5}$, the probability that a F random variable with 1 and 82 degrees of freedom is greater than F^*. This is identical to the p-value of the t test we performed in last week's homework! The decision rules are also equivalent. The decision rule for the t test was:

- If $t_{82}(0.005) \leq t^* \leq t_{82}(0.995)$, conclude H_0
- If $t^* < t_{82}(0.005)$ or $t^* > t_{82}(0.995)$, conclude H_a.

This can be rewritten as:

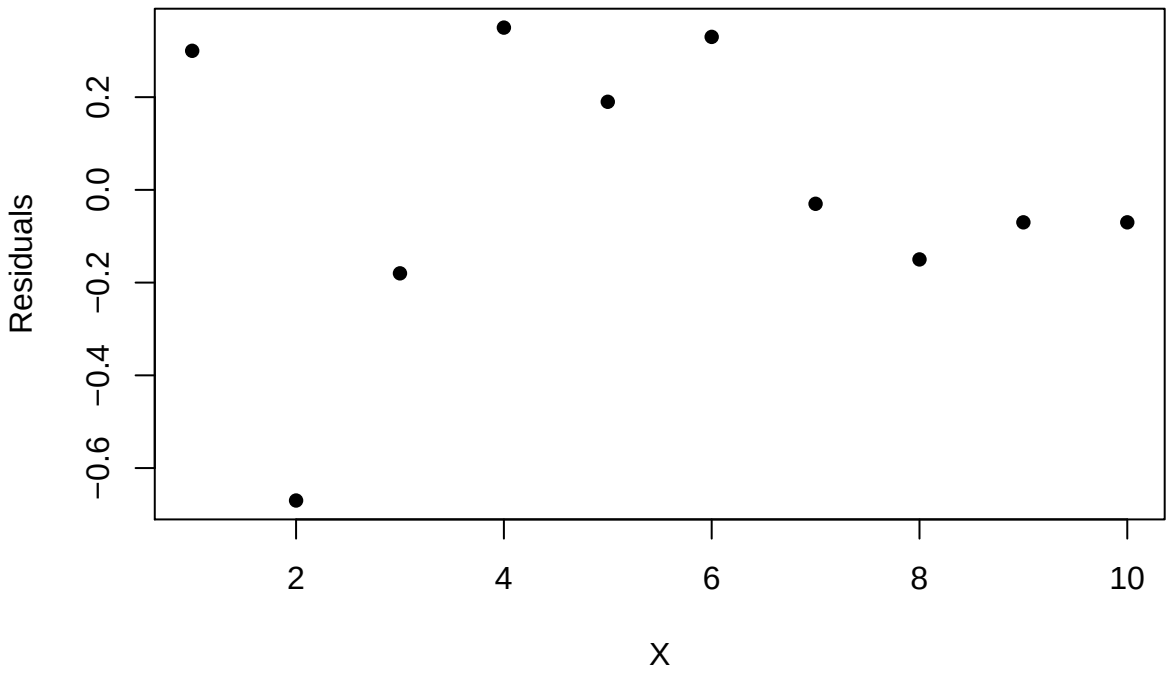
- If $(t^*)^2 \leq t_{82}(0.995)^2$, conclude H_0
- If $(t^*)^2 > t_{82}(0.995)^2$, conclude H_a.

From last time, we had $t^* = -4.1028971$, and $t_{82}(0.995) = -2.6371234$. We can see that $(t^*)^2 = 16.8337645 = F^*$, and $t_{82}(0.995)^2 = 6.9544199 = F^*$, i.e. the test statistics and decision rules are numerically equivalent!

5. Problem 3.2 from the `.pdf` version of the textbook. The idea is to make an example plot of the residuals $e_1, \ldots, e_n$ against the predictors $X_1, \ldots, X_n$, that depicts what you might expect to see under the circumstances described by each part of the problem.

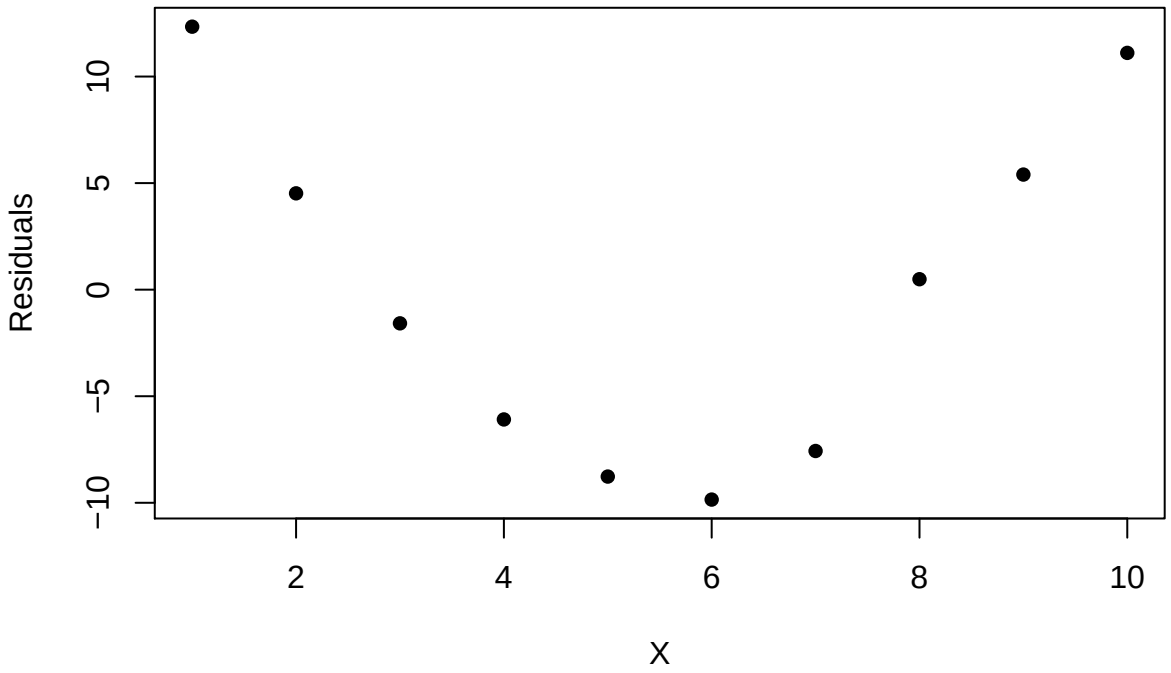
(a)

```
X <- 1:10
e <- c(0.30, -0.67, -0.18,  0.35,  0.19,
       0.33, -0.03, -0.15, -0.07, -0.07)
plot(X, e, ylab = "Residuals", pch = 16)
```



(b)

```
X <- 1:10
e <- c(12.34,  4.52, -1.58, -6.09, -8.77,
      -9.85, -7.57,  0.49,  5.40, 11.11 )
plot(X, e, ylab = "Residuals", pch = 16)
```



6. Problem 3.7 from the `.pdf` version of the textbook, parts (a)-(d). You only need to do the first part of (d), i.e. construct a normal probability plot of the residuals. Note that this will require a bit of material that we will cover Tuesday 3/10.

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

n <- length(Y)
```

We can use `stem(X, scale = 3)` to create a stem-and-leaf plot.

Stem	Leaf
40	000
42	0000
44	00
46	0000
48	00
50	0
52	000
54	000
56	0000
58	000
60	0000
62	000
64	0000
66	0
68	000
70	000
72	000
74	0

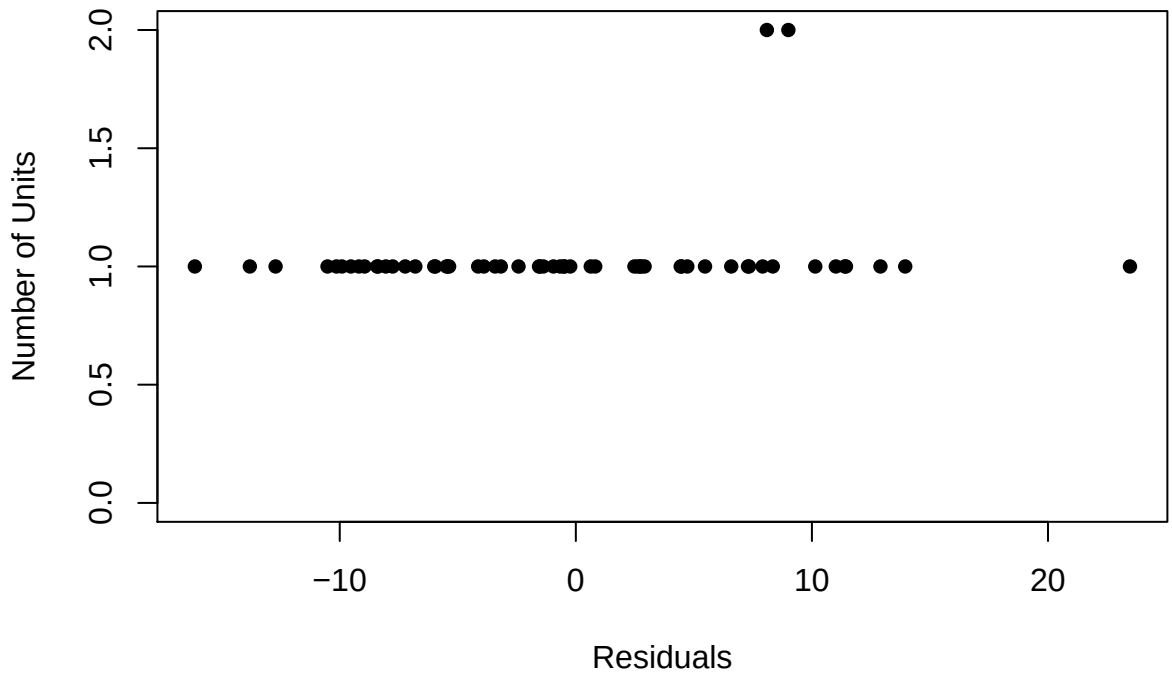
Stem	Leaf
76	0000
78	00000

This plot appears to be consistent with the random selection of women from each 10-year age group, because there are no clear patterns of some ages being better represented in the data than others.

(b)

A dot plot is given below. If you used a bar chart like we did in class, that is ok and will be given full credit, however I think this is clearer.

```
linmod <- lm(Y~X)
e <- linmod$residuals
tab <- table(e)
plot(e, e, xlim = range(e), ylim = c(0, 2),
     xlab = "Residuals",
     ylab = "Number of Units", type = "n")
points(as.numeric(names(tab)),
       tab, pch = 16)
```

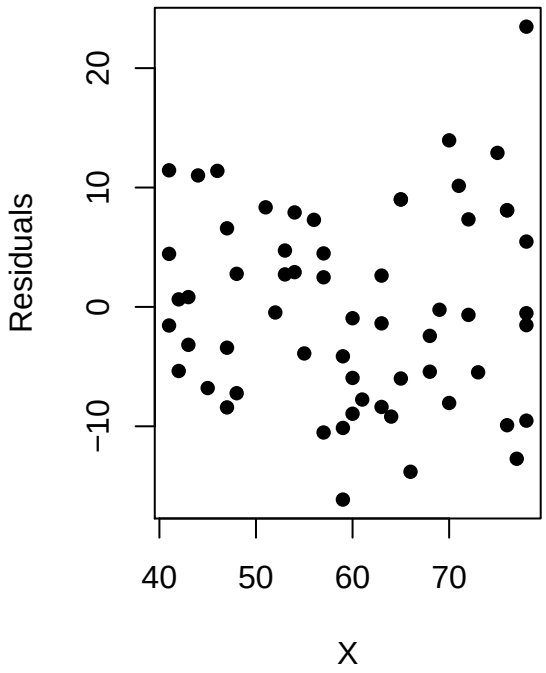


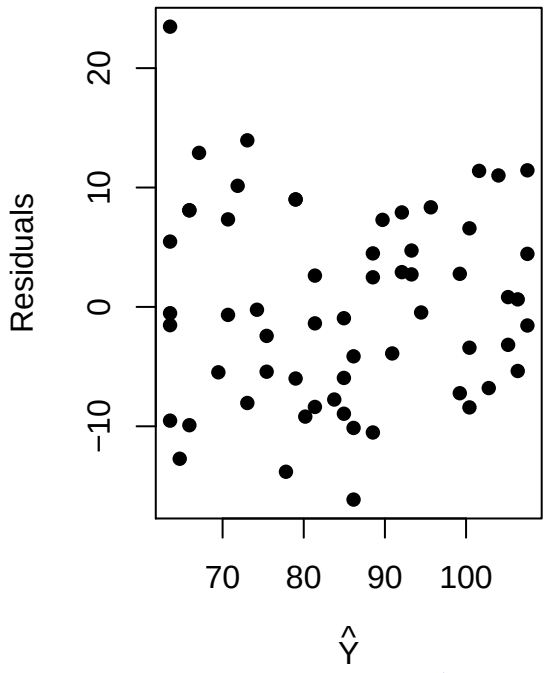
The residuals seem evenly distributed about 0, but it is difficult to discern much from a dot plot.

(c)

```
Y.hat <- linmod$fitted.values

par(mfrow = c(1, 2))
plot(X, e, ylab = "Residuals", pch = 16)
plot(Y.hat, e, xlab = expression(hat(Y)),
     ylab = "Residuals", pch = 16)
```

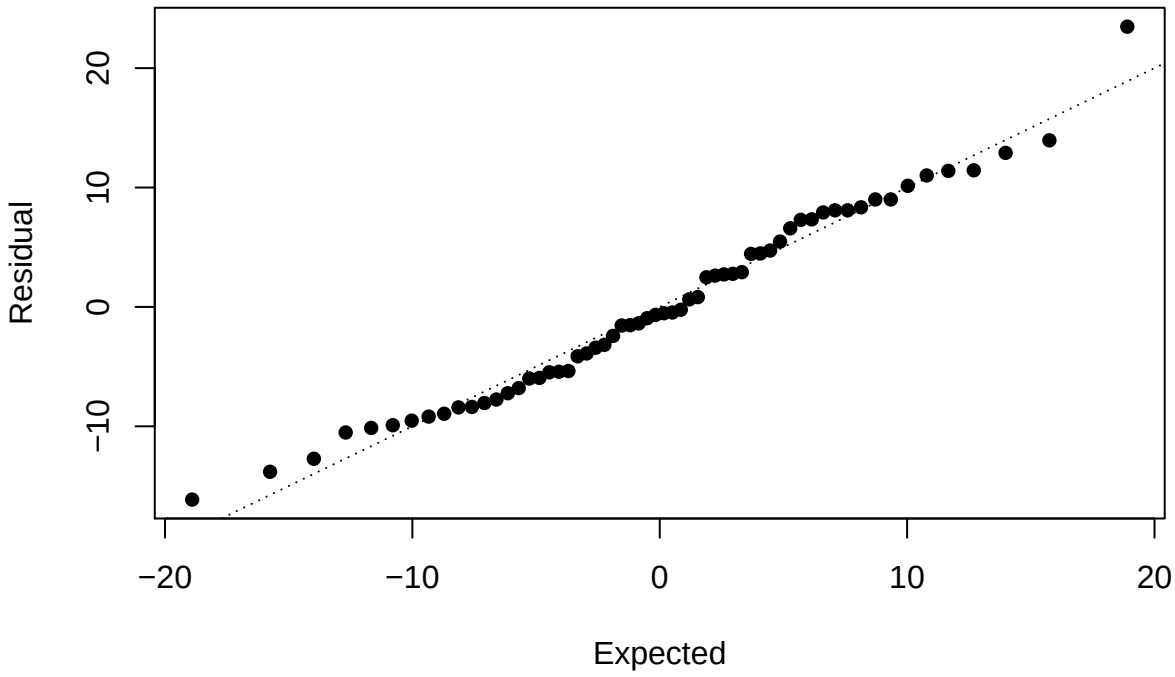




There are no striking relationships between the residuals and the covariate X or the fitted values $\hat{Y}$. There may be slight evidence of increasing variance as $\hat{Y}$ increases, and of a slight quadratic trend in X, but these seem minor.

(d)

```
MSE <- sum(e^2)/(n - 2)
Ee <- sqrt(MSE)*qnorm((1:n - 0.375)/(n + 0.25),
                      mean = 0, sd = 1)
plot(Ee, sort(e),
     pch = 16,
     xlab = "Expected",
     ylab = "Residual")
abline(a = 0, b = 1, lty = 3)
```



With the exception of the most extreme residuals, the observed residuals resemble what we would expect to observe if the errors were normally distributed.

7. Problem 3.8 from the `.pdf` version of the textbook, parts (a)-(d). You only need to do the first part of (d), i.e. construct a normal probability plot of the residuals. Note that this will require a bit of material that we will cover Tuesday 3/10.

(a)

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

n <- length(Y)
```

We can use `stem(X, scale = 3)` to create a stem-and-leaf plot.

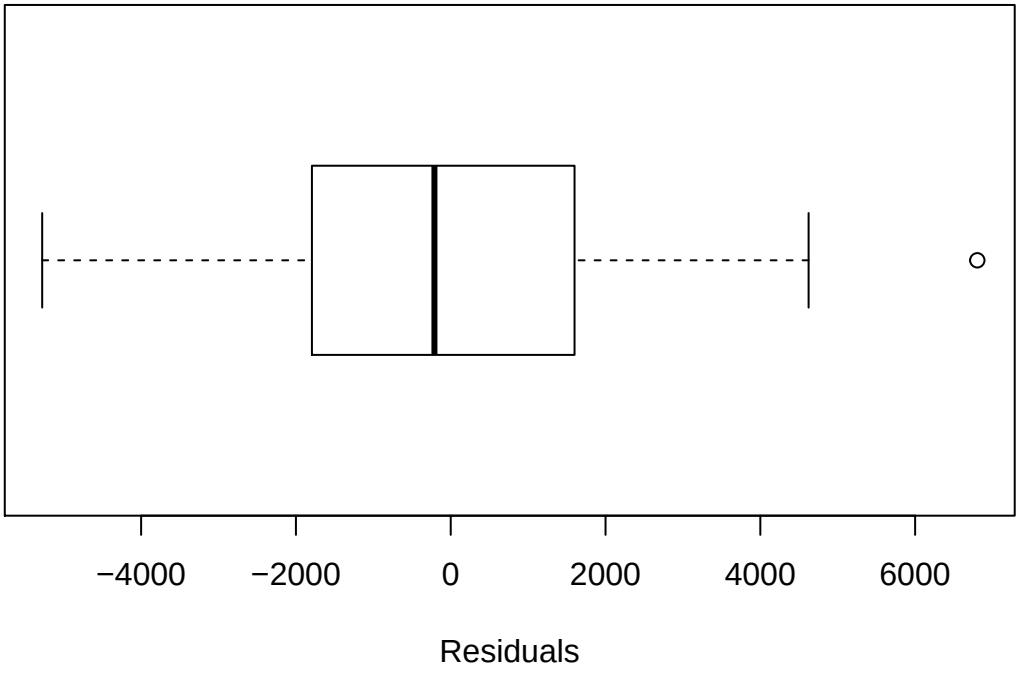
Stem	Leaf
60	0
62	
64	0000
66	00
68	0
70	00
72	00
74	00000000
76	00000000000
78	0000000000000
80	00000000000
82	000000000000000
84	00000000
86	0

Stem	Leaf
88	0000
90	00

The plot indicates that most of the counties represented in the data have high school diploma per capita rates of around 75-80 percent. In every county, at least 60 percent but no more than 90 percent of the residents have a high school diploma.

(b)

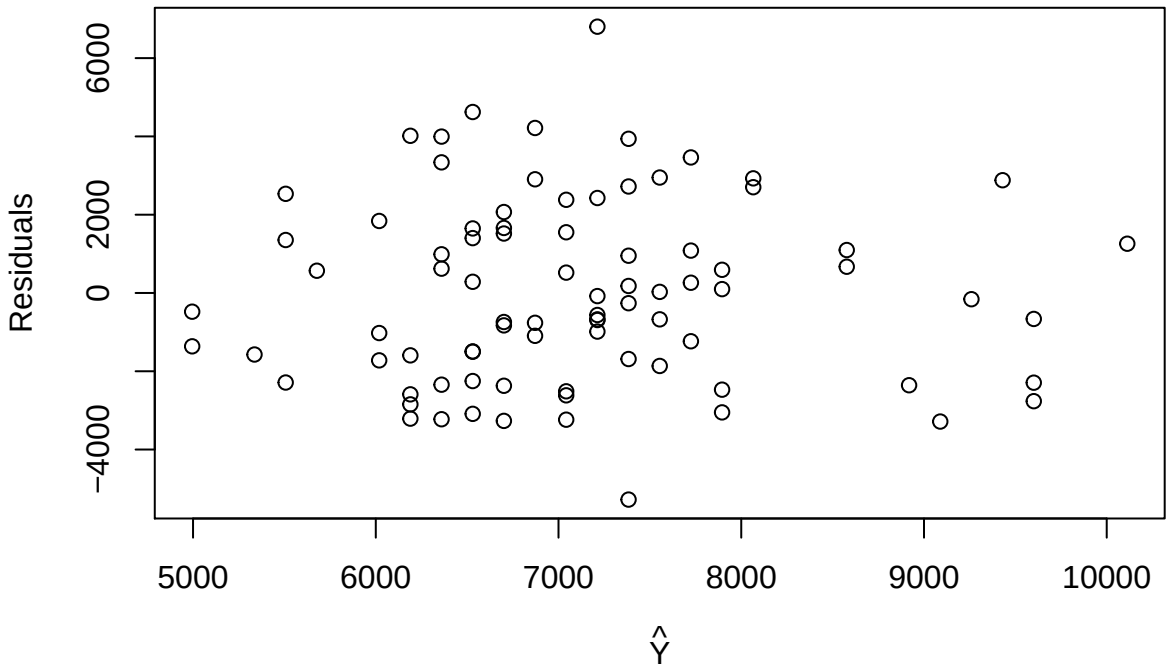
```
linmod <- lm(Y~X)
e <- linmod$residuals
boxplot(e,
        horizontal = TRUE,
        xlab = "Residuals")
```



The residuals appear to be slighty skewed to the right, with several very large residuals and generally more positive residuals than negative residuals.

(c)

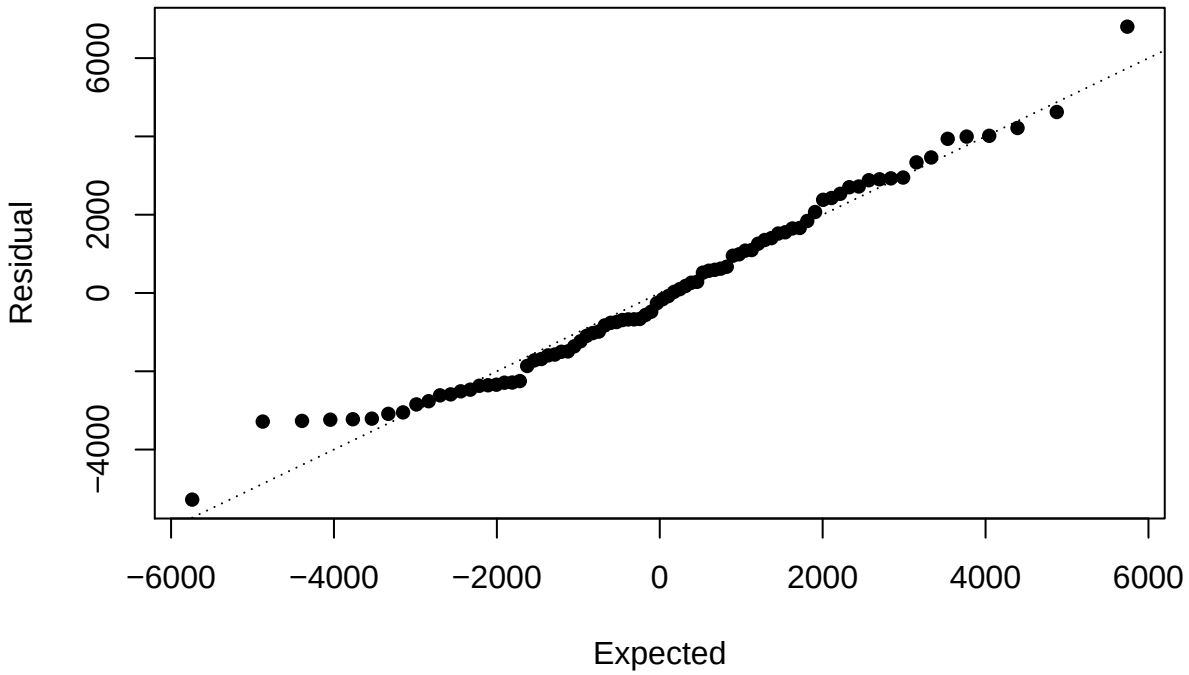
```
Y.hat <- linmod$fitted.values
plot(Y.hat, e, xlab = expression(hat(Y)),
     ylab = "Residuals")
```



The residuals tend to be larger and have greater variance for nearly average values of $\hat{Y}$, which suggests the presence of some slight nonconstant variance or nonlinearity. However overall, there are no striking trends.

(d)

```
MSE <- sum(e^2)/(n - 2)
Ee <- sqrt(MSE)*qnorm((1:n - 0.375)/(n + 0.25),
                      mean = 0, sd = 1)
plot(Ee, sort(e),
     pch = 16,
     xlab = "Expected",
     ylab = "Residual")
abline(a = 0, b = 1, lty = 3)
```



Again, with the exception of the most extreme residuals, the residuals are consistent with what we would expect if the errors were indeed normally distributed.