36-401 Modern Regression HW #2 Solutions

DUE: 9/15/2017

Problem 1 [36 points total]

(a) (12 pts.)

In Lecture Notes 4 we derived the following estimators for the simple linear regression model:

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$
$$\widehat{\beta}_1 = \frac{c_{XY}}{s_X^2},$$

where

$$c_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \quad \text{and} \quad s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Since the formula for $\widehat{\beta}_0$ depends on $\widehat{\beta}_1$ we will calculate $\text{Var}(\widehat{\beta}_1)$ first. Some simple algebra[1] shows we can rewrite $\widehat{\beta}_1$ as

$$\widehat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})\epsilon_i}{s_X^2}.$$

Now, treating the $X_i's$ as fixed, we have

$$\begin{aligned}
\text{Var}\big(\widehat{\beta}_1\big) &= \text{Var}\left(\beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})\epsilon_i}{s_X^2}\right) \\
&= \text{Var}\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})\epsilon_i}{s_X^2}\right) \\
&= \frac{\frac{1}{n^2}\sum_{i=1}^{n}(X_i - \overline{X})^2\text{Var}(\epsilon_i)}{s_X^4} \\
&= \frac{\frac{\sigma^2}{n^2}\sum_{i=1}^{n}(X_i - \overline{X})^2}{s_X^4} \\
&= \frac{\frac{\sigma^2}{n}s_X^2}{s_X^4} \\
&= \frac{\sigma^2}{n \cdot s_X^2}.
\end{aligned}$$

[1] See (16)-(22) of Lecture Notes 4

Thus, $\mathrm{Var}(\widehat{\beta}_0)$ is given by

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta}_0) &= \mathrm{Var}\big(\overline{Y} - \widehat{\beta}_1\overline{X}\big) \\
&= \mathrm{Var}(\overline{Y}) + \overline{X}^2\mathrm{Var}(\widehat{\beta}_1) - 2\overline{X}\mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^n Y_i, \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2}\right) \\
&= \frac{\sigma^2}{n} + \overline{X}^2\mathrm{Var}(\widehat{\beta}_1) - \frac{2\overline{X}}{n\sum_{i=1}^n (X_i - \overline{X})^2}\mathrm{Cov}\left(\sum_{i=1}^n Y_i, \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})\right) \\
&= \frac{\sigma^2}{n} + \overline{X}^2\mathrm{Var}(\widehat{\beta}_1) - \frac{2\overline{X}}{n\sum_{i=1}^n (X_i - \overline{X})^2}\sum_{i=1}^n (X_i - \overline{X})\mathrm{Cov}(Y_i, Y_i) \\
&= \frac{\sigma^2}{n} + \overline{X}^2\mathrm{Var}(\widehat{\beta}_1) - \frac{2\overline{X}\sigma^2}{n\sum_{i=1}^n (X_i - \overline{X})^2}\underbrace{\sum_{i=1}^n (X_i - \overline{X})}_{=0} \\
&= \frac{\sigma^2}{n} + \overline{X}^2\mathrm{Var}(\widehat{\beta}_1) \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2\overline{X}^2}{n \cdot s_X^2} \\
&= \frac{\sigma^2\big(s_X^2 + \overline{X}^2\big)}{n \cdot s_X^2} \\
&= \frac{\sigma^2\sum_{i=1}^n X_i^2}{n^2 \cdot s_X^2}.
\end{aligned}$$

(b) (6 pts.)

$$\begin{aligned}
\sum_{i=1}^n \widehat{\epsilon}_i &= \sum_{i=1}^n \big(Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)\big) \\
&= \sum_{i=1}^n \big(Y_i - (\overline{Y} - \widehat{\beta}_1\overline{X}) - \widehat{\beta}_1 X_i\big) \\
&= \sum_{i=1}^n (Y_i - \overline{Y}) + \sum_{i=1}^n (\widehat{\beta}_1\overline{X} - \widehat{\beta}_1 X_i) \\
&= (n\overline{Y} - n\overline{Y}) + (n\widehat{\beta}_1\overline{X} - n\widehat{\beta}_1\overline{X}) \\
&= 0 + 0 \\
&= 0
\end{aligned}$$

(c) (12 pts.)

$$
\begin{aligned}
\sum_{i=1}^{n} \widehat{Y}_i \widehat{\epsilon}_i &= \sum_{i=1}^{n} (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)\widehat{\epsilon}_i \\
&= \widehat{\beta}_0 \underbrace{\sum_{i=1}^{n} \widehat{\epsilon}_i}_{=0} + \widehat{\beta}_1 \sum_{i=1}^{n} X_i \widehat{\epsilon}_i \\
&= \widehat{\beta}_1 \sum_{i=1}^{n} X_i \widehat{\epsilon}_i \\
&= \widehat{\beta}_1 \sum_{i=1}^{n} X_i \widehat{\epsilon}_i - \widehat{\beta}_1 \overline{X} \underbrace{\sum_{i=1}^{n} \widehat{\epsilon}_i}_{=0} \\
&= \widehat{\beta}_1 \sum_{i=1}^{n} (X_i - \overline{X})\widehat{\epsilon}_i \\
&= \widehat{\beta}_1 \sum_{i=1}^{n} (X_i - \overline{X})\Big(Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)\Big) \\
&= \widehat{\beta}_1 \sum_{i=1}^{n} (X_i - \overline{X})\Big(Y_i - (\overline{Y} - \widehat{\beta}_1 \overline{X}) - \widehat{\beta}_1 X_i\Big) \\
&= \widehat{\beta}_1 \sum_{i=1}^{n} (X_i - \overline{X})\Big((Y_i - \overline{Y}) - \widehat{\beta}_1 (X_i - \overline{X})\Big) \\
&= \widehat{\beta}_1 \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) - \widehat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2 \\
&= \widehat{\beta}_1 \cdot n \cdot c_{XY} - \widehat{\beta}_1^2 \cdot n \cdot s_X^2 \\
&= \widehat{\beta}_1 \cdot n \cdot c_{XY} - \widehat{\beta}_1 \cdot \frac{c_{XY}}{s_X^2} \cdot n \cdot s_X^2 \\
&= 0
\end{aligned}
$$

Note: The above implies

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{Y}_i - \frac{1}{n}\sum_{j=1}^{n}\widehat{Y}_j\Big)\Big(\widehat{\epsilon}_i - \frac{1}{n}\sum_{j=1}^{n}\widehat{\epsilon}_i\Big) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{Y}_i - \frac{1}{n}\sum_{j=1}^{n}\widehat{Y}_j\Big)\cdot \widehat{\epsilon}_i \\
&= \frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_i\widehat{\epsilon}_i - \frac{1}{n^2}\sum_{j=1}^{n}\widehat{Y}_j\sum_{i=1}^{n}\widehat{\epsilon}_i \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_i\widehat{\epsilon}_i}_{=0} - \left(\frac{1}{n}\sum_{j=1}^{n}\widehat{Y}_j\right)\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}_i\right)}_{=0} \\
&= 0.
\end{aligned}
$$

Linear Algebra interpretation: The observed residuals are orthogonal to the fitted values.

Statistical interpretation: The observed residuals are linearly uncorrelated with the fitted values.

(d) (6 pts.)

From the result in part (c) we have $\widehat{\beta}_1 = 0$.

Substituting this into the equation for $\widehat{\beta}_0$, we obtain the intercept

$$\begin{aligned}\widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} \widehat{\epsilon}_i \\ &= \overline{Y} - 0 \cdot 0 \\ &= \overline{Y}.\end{aligned}$$

Problem 2 [24 points]

(a) (8 pts.)

We compute the least squares estimate $\widehat{\beta}_1$ by minimizing the empirical mean squared error via a 1st derivative test.

$$\begin{aligned}\frac{\partial}{\partial \beta_1}\widehat{MSE}(\beta_1) &= \frac{\partial}{\partial \beta_1}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_1 X_i)^2\right) \\ &= \frac{2}{n}\sum_{i=1}^{n}(Y_i - \beta_1 X_i)(-X_i)\end{aligned}$$

Setting the derivative equal to 0 yields

$$\begin{aligned}-\frac{2}{n}\sum_{i=1}^{n}(Y_i - \beta_1 X_i)(X_i) &= 0 \\ \sum_{i=1}^{n}(Y_i X_i - \beta_1 X_i^2) &= 0 \\ \sum_{i=1}^{n} Y_i X_i - \beta_1 \sum_{i=1}^{n} X_i^2 &= 0 \\ \implies \widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}.\end{aligned}$$

Furthermore,

$$\begin{aligned}\frac{\partial^2}{\partial \beta_1^2}\widehat{MSE}(\beta_1) &= \frac{\partial}{\partial \beta_1}\left(-\frac{2}{n}\sum_{i=1}^{n}(Y_i X_i - \beta_1 X_i^2)\right) \\ &= \frac{2}{n}\sum_{i=1}^{n} X_i^2 \\ &> 0\end{aligned},$$

so $\widehat{\beta}_1$ is indeed the *minimizer* of the empirical MSE.

(b) (8 pts.)

$$\begin{aligned}
\mathbb{E}\big[\widehat{\beta}_1\big] &= \mathbb{E}\left[\frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n X_i(\beta_1 X_i + \epsilon_i)}{\sum_{i=1}^n X_i^2}\right] \\
&= \mathbb{E}\left[\frac{\beta_1 \sum_{i=1}^n X_i^2 + \sum_{i=1}^n X_i\epsilon_i}{\sum_{i=1}^n X_i^2}\right] \\
&= \mathbb{E}\left[\beta_1 + \frac{\sum_{i=1}^n X_i\epsilon_i}{\sum_{i=1}^n X_i^2}\right] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^n X_i^2}\mathbb{E}\left[\sum_{i=1}^n X_i\epsilon_i\right] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^n X_i^2}\sum_{i=1}^n X_i \cdot \underbrace{\mathbb{E}[\epsilon_i]}_{=0} \\
&= \beta_1
\end{aligned}$$

Thus, if the true model is linear *and* through the origin, then $\widehat{\beta}_1$ is an unbiased estimator for β_1.

(c) (8 pts.)

If the true model is linear, but not necessarily through the origin, then the bias of the regression-through-the-origin estimator $\widehat{\beta}_1$ is

$$\begin{aligned}
\text{Bias}(\widehat{\beta}_1) &= \mathbb{E}\big[\widehat{\beta}_1\big] - \beta_1 \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2}\right] - \beta_1 \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n X_i(\beta_0 + \beta_1 X_i + \epsilon_i)}{\sum_{i=1}^n X_i^2}\right] - \beta_1 \\
&= \mathbb{E}\left[\frac{\beta_0 \sum_{i=1}^n X_i + \beta_1 \sum_{i=1}^n X_i^2 + \sum_{i=1}^n X_i\epsilon_i)}{\sum_{i=1}^n X_i^2}\right] - \beta_1 \\
&= \mathbb{E}\left[\beta_1 + \frac{\beta_0 \sum_{i=1}^n X_i + \sum_{i=1}^n X_i\epsilon_i}{\sum_{i=1}^n X_i^2}\right] - \beta_1 \\
&= \beta_1 + \frac{\beta_0 \sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2} + \frac{1}{\sum_{i=1}^n X_i^2}\sum_{i=1}^n X_i \cdot \underbrace{\mathbb{E}[\epsilon_i]}_{=0} - \beta_1 \\
&= \frac{\beta_0 \sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2}.
\end{aligned}$$

Problem 3 [20 points total]

(a) (5 pts.)

```
set.seed(1)
n <- 100
X <- runif(n, 0, 1)
Y <- 5 + 3 * X + rnorm(n, 0, 1)

plot(X,Y, cex = 0.75)
model <- lm(Y ~ X)
abline(model, lwd = 2, col = "red")
```

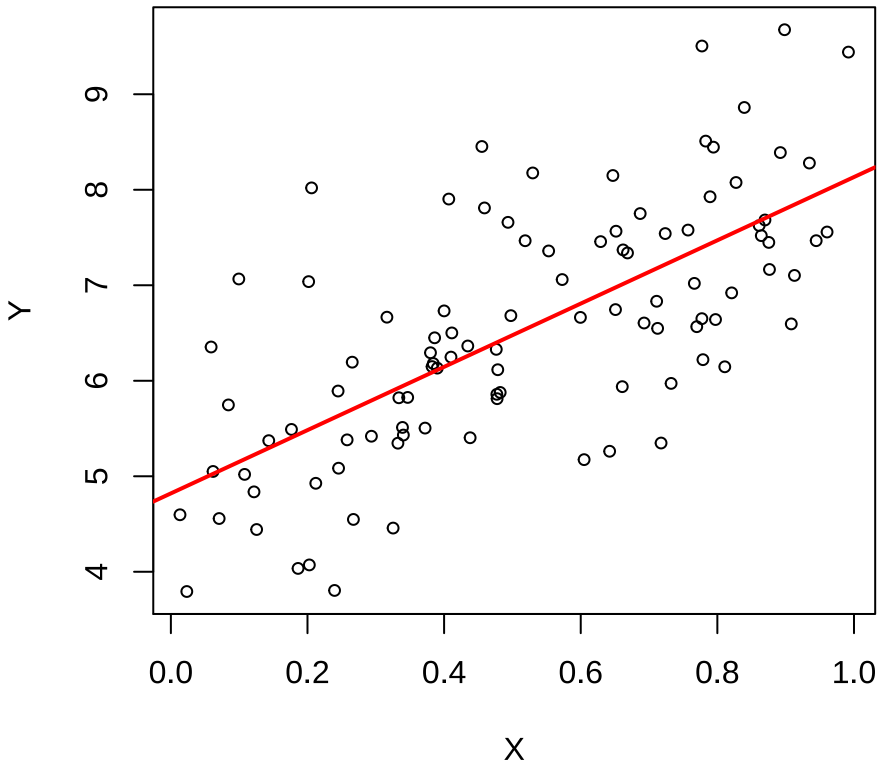

Figure 1: One hundred data points with the simple linear regression fit

(b) (5 pts.)

```
n <- 100
betas <- rep(NA,1,1000)

for (itr in 1:1000){
  X <- runif(n, 0, 1)
  Y <- 5 + 3 * X + rnorm(n, 0, 1)
  model <- lm(Y ~ X)
  betas[itr] <- model$coefficients[2]
}
```

```r
mean(betas)
```

```
## [1] 3.019629
```

Since 1000 is a reasonably large number of trials we expect the mean of $\beta_1^{(1)}, \ldots, \beta_1^{(1000)}$ to be close to

$$\begin{aligned}
\mathbb{E}[\widehat{\beta}_1] &= \mathbb{E}[\mathbb{E}[\widehat{\beta}_1 \mid X_1, \ldots, X_n]] \\
&= \mathbb{E}[\beta_1] \\
&= \mathbb{E}[3] \\
&= 3.
\end{aligned}$$

In the above experiment, we have

$$\frac{1}{1000} \sum_{i=1}^{1000} \beta_1^{(i)} = 3.019629.$$

```r
hist(betas, xlab = expression(hat(beta)[1]),  prob = FALSE, main = "", breaks = 50)
```

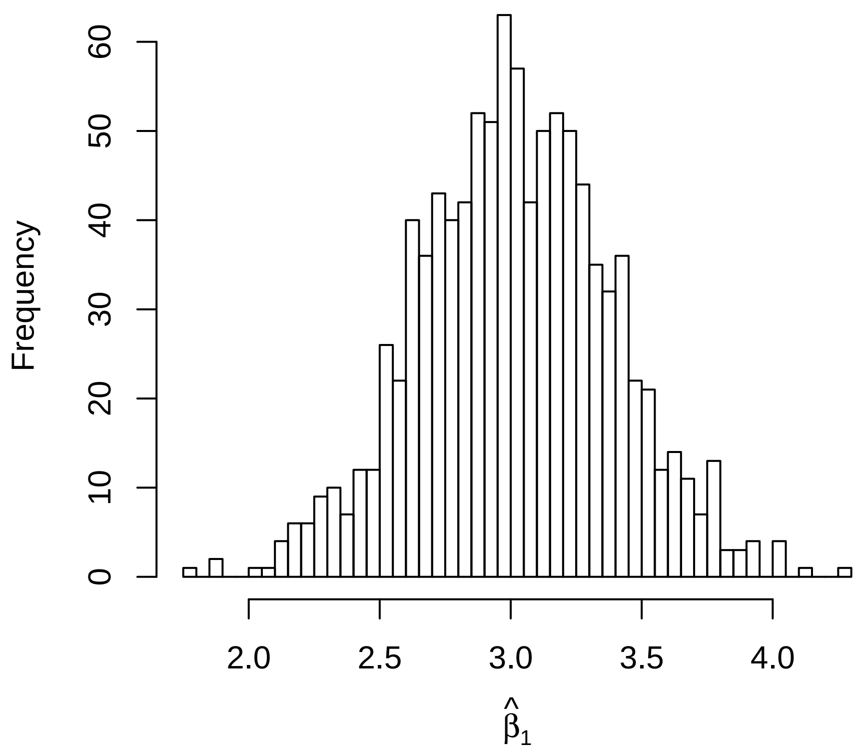

Figure 2: Histogram of linear regression slope parameters for Gaussian data

(c) (5 pts.)

```r
n <- 100
betas <- rep(NA,1,1000)

for (itr in 1:1000){
  X <- runif(n, 0, 1)
  Y <- 5 + 3 * X + rcauchy(n, 0, 1)
```

```
  model <- lm(Y ~ X)
  betas[itr] <- model$coefficients[2]
}

par(mfrow = c(1,2))
hist(betas, xlab = expression(hat(beta)[1]),  prob = FALSE, main = "", xlim = c(3-20,3+20),
     breaks = 750)
abline(v = 3, col = "red", lwd = 2)
hist(betas, xlab = expression(hat(beta)[1]),  prob = FALSE, main = "", breaks = 200)
```

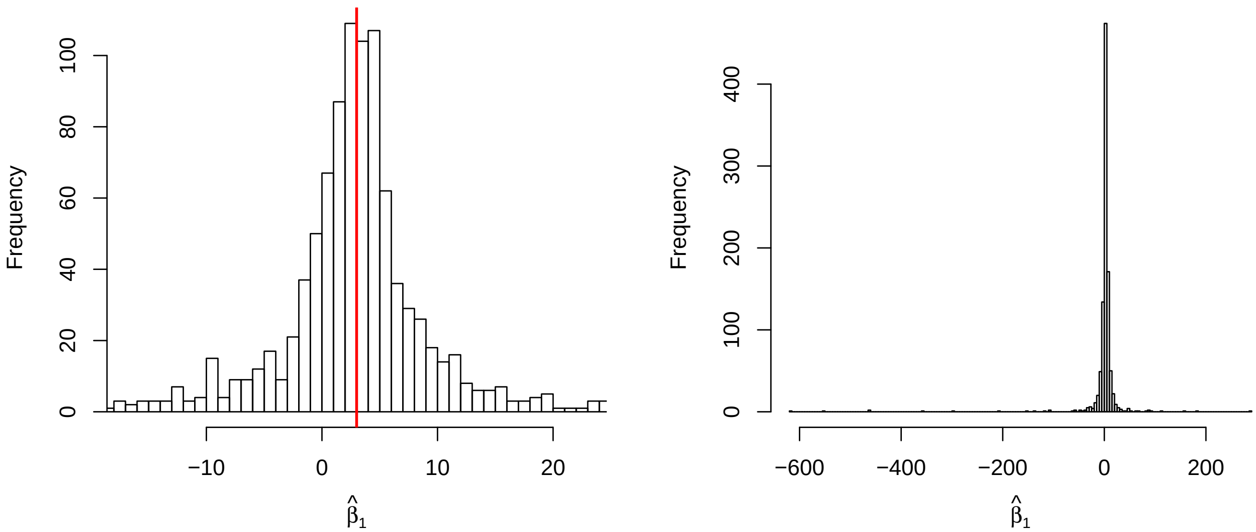

Figure 3: Histogram of linear regression slope parameters for Cauchy data (**Left**: restricted to the window (-17,23). **Right**: The full window.)

Notice that the distribution of $\beta_1^{(1)}, \ldots, \beta_1^{(1000)}$ still seems to be approximately centered around $\widehat{\beta}_1 = 3$, but the tails are now much fatter. In particular, from the plot on the right, we see that at least one trial of the experiment resulted in a value around $\widehat{\beta} \approx -600$.

(d) (5 pts.)

```
set.seed(1)
n <- 100
X <- runif(n, 0, 1)
W <- X + rnorm(n, 0, sqrt(2))
Y <- 5 + 3 * X + rnorm(n, 0, 1)

plot(X,Y, cex = 0.75)
model <- lm(Y ~ W)
abline(model, lwd = 2, col = "red")
```

```
n <- 100
betas <- rep(NA,1,1000)

for (itr in 1:1000){
  X <- runif(n, 0, 1)
  W <- X + rnorm(n, 0, sqrt(2))
```

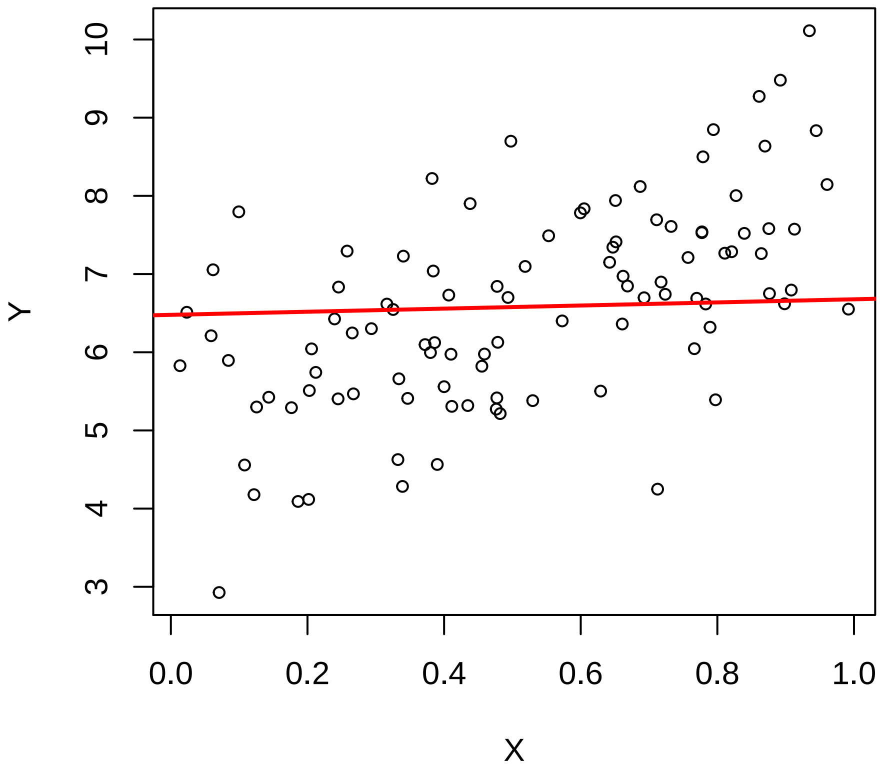

Figure 4: One hundred observations of Y vs. X with the simple linear regression fit of Y on W

```
  Y <- 5 + 3 * X + rnorm(n, 0, 1)
  model <- lm(Y ~ W)
  betas[itr] <- model$coefficients[2]
}

mean(betas)
```

```
## [1] 0.1198059
```

```
hist(betas, xlab = expression(hat(beta)[1]),  prob = FALSE, main = "", breaks = 50)
```

In the above experiment, we have

$$\frac{1}{1000}\sum_{i=1}^{1000}\beta_1^{(i)} = 0.06132475.$$

From this, and Figure 5, we conclude having errors on the X_i's biases $\widehat{\beta}_1$ downwards.

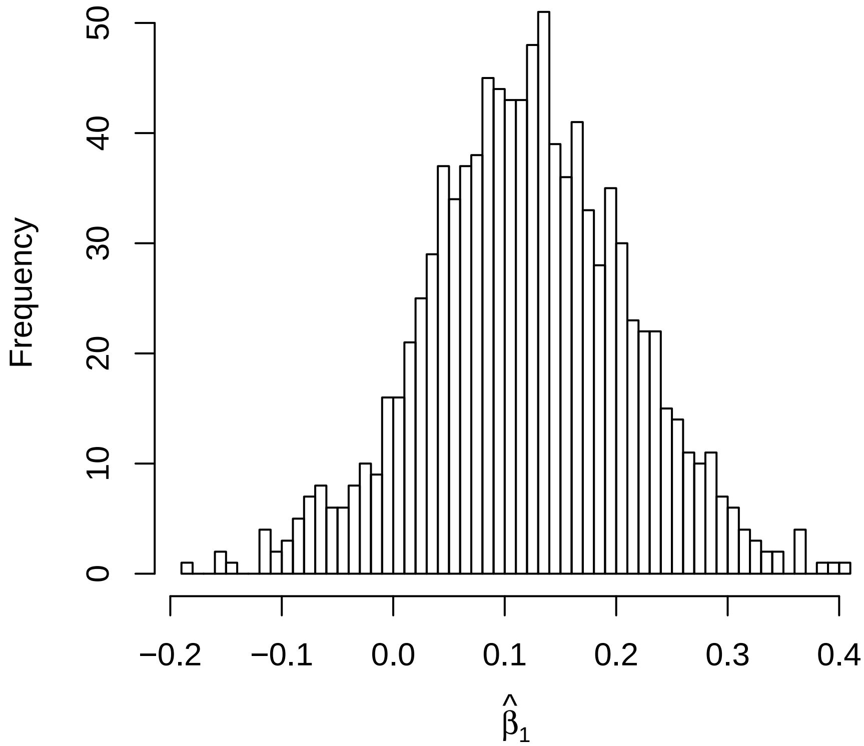

Figure 5: Histogram of linear regression slope parameters for data with errors on the X's

Problem 4 [20 points total]

```r
data(airquality)
```

(a) (5 pts.)

```r
summary(airquality)
```

```
##      Ozone           Solar.R           Wind             Temp      
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
##  NA's   :37       NA's   :7                                       
##      Month            Day      
##  Min.   :5.000   Min.   : 1.0  
##  1st Qu.:6.000   1st Qu.: 8.0  
##  Median :7.000   Median :16.0  
##  Mean   :6.993   Mean   :15.8  
##  3rd Qu.:8.000   3rd Qu.:23.0  
##  Max.   :9.000   Max.   :31.0  
## 
```

```r
pairs(airquality, cex = 0.5)
```

(b) (5 pts.)

```r
with(airquality, plot(Solar.R, Ozone, xlab = "Solar Radiation", ylab = "Ozone"))
model <- lm(Ozone ~ Solar.R, data = airquality)
abline(model, col = "red", lwd = 2)
```

Ozone and Solar Radiation appear to be positively correlated.

(c) (5 pts.)

```r
summary(model)$coefficients
```

```
##               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 18.5987278 6.74790416 2.756223 0.0068560215
## Solar.R      0.1271653 0.03277629 3.879795 0.0001793109
```

The intercept and slope of the least squares regression are

$$\widehat{\beta}_0 = 18.59873 \quad \text{and} \quad \widehat{\beta}_1 = 0.12717$$

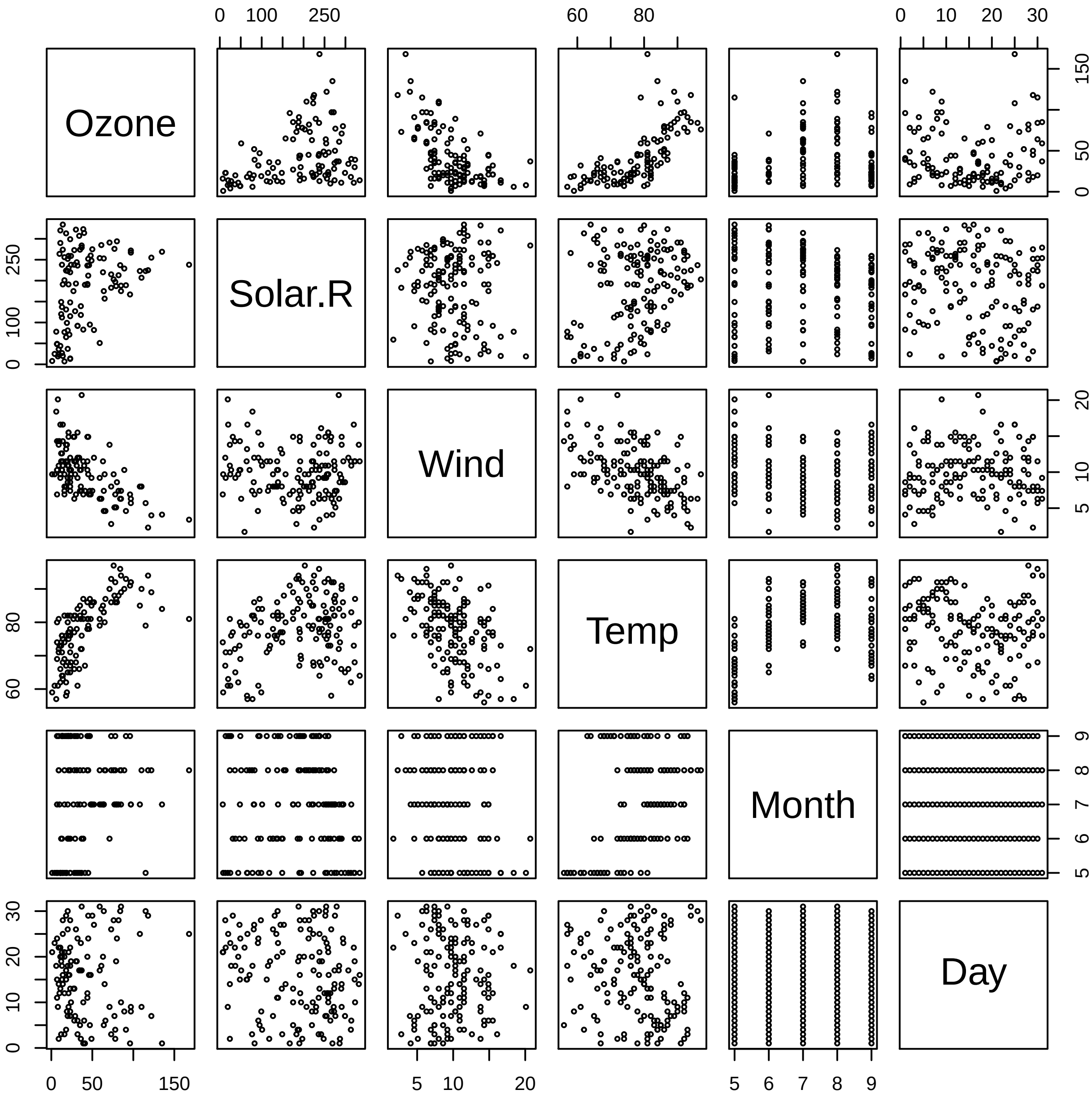

Figure 6: Pairwise relationships of variables in the **airquality** data set

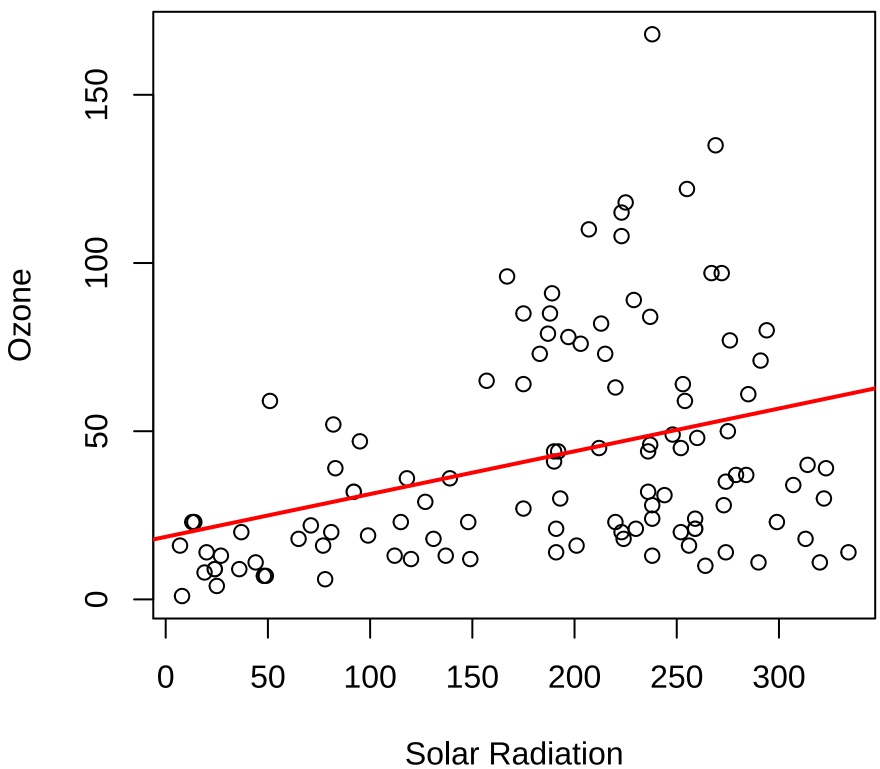

Figure 7: Ozone vs. solar radiation observations in the **airquality** data set

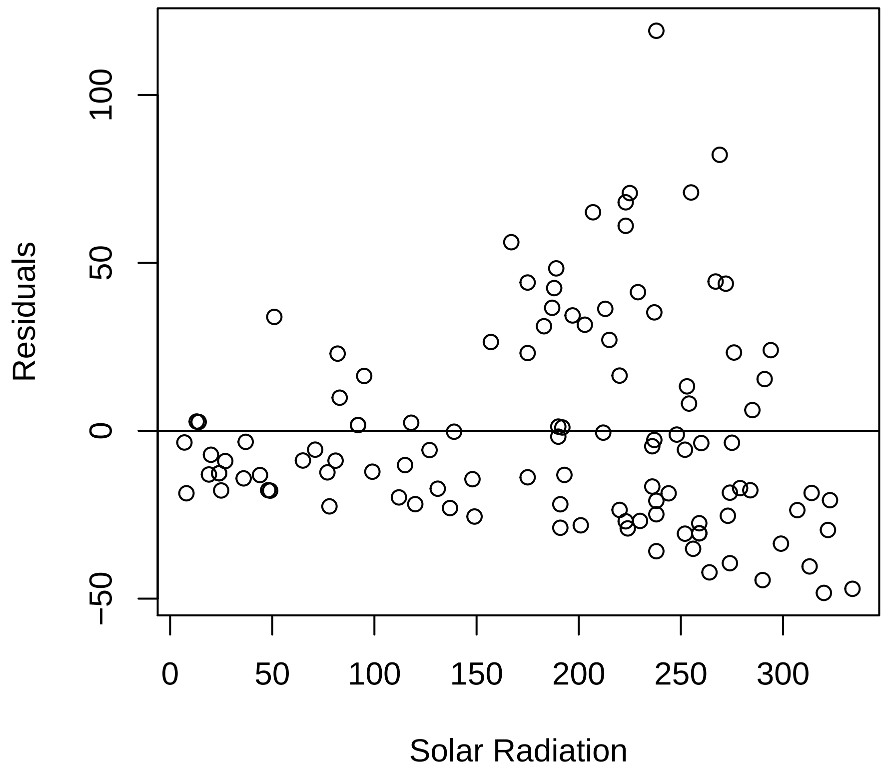

Figure 8: Linear regression residuals vs. solar radiation

(d) (5 pts.)

```
resids <- airquality$Ozone - predict(model, newdata = data.frame(Solar.R = airquality$Solar.R))
plot(airquality$Solar.R, resids, xlab = "Solar Radiation", ylab = "Residuals")
abline(h = 0)
```

No, the standard regression assumptions do not hold. The residuals are not symmetric about zero so the linear functional form assumption is not suitable. Furthermore, the residuals are highly heteroskedastic.