Homework 2
36-401
Due Friday, September 17 at 3:00

Searching for solutions on the web or anywhere else is forbidden.

1. Suppose that
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n$$
where $\mathbb{E}[\epsilon_i|X_i] = 0$ and $\text{Var}[\epsilon_i|X_i] = \sigma^2$. Let $\widehat{\beta}_0$ and $\widehat{\beta}_1$ be the least squares estimates.
(a) Show that $\mathbb{E}[\epsilon_i] = 0$ and $\text{Var}[\epsilon_i] = \sigma^2$.
(b) Show that
$$\widehat{\beta}_1 = \frac{\sum_i (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_i (X_i - \overline{X})^2}$$
and
$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}.$$
(c) Find $\mathbb{E}[\widehat{\beta}_0|X_1, \ldots, X_n]$ and $\text{Var}[\widehat{\beta}_0|X_1, \ldots, X_n]$.
(d) Recall that the observed residuals are defined by $\widehat{\epsilon}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)$. Show that $\sum_i \widehat{\epsilon}_i = 0$.

2. Suppose that
$$Y_i = \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n$$
where $\mathbb{E}[\epsilon_i] = 0$ and $\text{Var}[\epsilon_i] = \sigma^2$. In other words, there is no intercept in this model. This is called regression through the origin.
(a) Find the least squares estimate $\widehat{\beta}_1$ for this model.
(b) Find $\mathbb{E}[\widehat{\beta}_1|X_1, \ldots, X_n]$ and $\text{Var}[\widehat{\beta}_1|X_1, \ldots, X_n]$.
(c) Generate data as follows:

```
n = 100
X = rnorm(n)
epsilon = rnorm(n)
Y = 2*X + epsilon
```

Fit a regression through the origin. By default, R will always include an intercept. Use the help command

help(lm)

to see how to force R to omit the intercept. Give a plot of your data and report the estimated slope and its standard error.

Now fit the data but include an intercept. Report the estimated slope and its standard error.

3. Simulation problem.
(a) Generate $n = 100$ data points as follows. Take $X_i \sim \text{Uniform}(-2, 2)$. Then set
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n$$
where $\beta_0 = 15$, $\beta_1 = -2$ and $\epsilon_i \sim N(0, 1)$. Here is the R code:

```
n = 100
x = runif(n,-2,2)
eps = rnorm(n,0,1)
y = 15 - 2*x + eps
```

Plot the data. Fit the regression line. Add the fitted line to the plot. Give the standard diagnostic plots and comment on them.

(b) Repeat the experiment in part (a) 1,000 times. You will get a different value of $\widehat{\beta}_1$ each time. Denote these by $\widehat{\beta}_1^{(1)}, \ldots, \widehat{\beta}_1^{(1000)}$. What is the mean of these values? What do you expect the mean to be? Plot a histogram of $\widehat{\beta}_1^{(1)}, \ldots, \widehat{\beta}_1^{(1000)}$. Here is the R code: (make sure you understand the code):

```
nsim = 1000
beta1 = rep(0,1000)
n = 100
for(i in 1:nsim){
     x = runif(n,-2,2)
     eps = rnorm(n,0,1)
     y = 15 - 2*x + eps
     beta1[i] = lm(y ~ x)$coef[2]
     }
```

4. Load the data on housing prices:

```
library(wooldridge)
data(hprice2)
str(hprice2)
names(hprice2)
help(hprice2)
```

The data are from *Hedonic Housing Prices and the Demand for Clean Air*, by Harrison, D. and D.L. Rubinfeld, Journal of Environmental Economics and Management 5, 81-102. We are interested in these variables:

price: median housing price.

nox: Nitrous Oxide concentration; parts per million.

crime: number of reported crimes per capita.

rooms: average number of rooms in houses in the community.

dist: weighted distance of the community to 5 employment centers.

stratio: average student-teacher ratio of schools in the community.

(a) Let's create a data frame with the variables of interest then do some plots:

```
df = data.frame(price,nox,crime,rooms,dist,stratio)
pairs(df)
summary(df)
```

Briefly comment on the plots. Do you see any relationships between the variables?

(b) Crime seems to be bunched up at low levels making it hard to see what is going on. Let's replace crime with log(crime). We will also replace price with log(price):

```
logcrime = log(crime)
logprice = log(price)
df = data.frame(logprice,nox,logcrime,rooms,dist,stratio)
pairs(df)
```

(c) Fit the regression model:

$$\text{logprice} = \beta_0 + \beta_1 \text{nox} + \beta_2 \text{logcrime} + \beta_3 \text{rooms} + \beta_4 \text{dist} + \beta_5 \text{stratio}$$

What are the estimated values of $\widehat{\beta}_0, \ldots, \widehat{\beta}_5$? What is $\widehat{\sigma}^2$?

(d) Plot the residuals. Comment on the plots. Which point has the largest residual?

(e) Plot the Cook's distance. Which point has the largest influence?

(f) Give 90 percent confidence intervals for the coefficients. Make sure to use the `sandwich` package. What is the effect of pollution `nox` on our prediction of price?

(g) Suppose there is a new house with these values:

nox = 8, logcrime = 4, rooms = 6, dist = 3, stratio 20.

Give a 90 percent prediction interval for its price.

5. Recall that the hat matrix is $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

(a) Show that $\mathbf{H}$ is idempotent, that is, $\mathbf{H} = \mathbf{H}^2$.

(b) Show that the trace of $\mathbf{H}$ is $d + 1$.

(c) Suppose there is only one covariate X. In that case

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}.$$

Find an explicit expression for $\mathbf{H}$.