hw1

Name: Yuhao Wang, UNI: yw3204

9/15/2018

1.19

a)

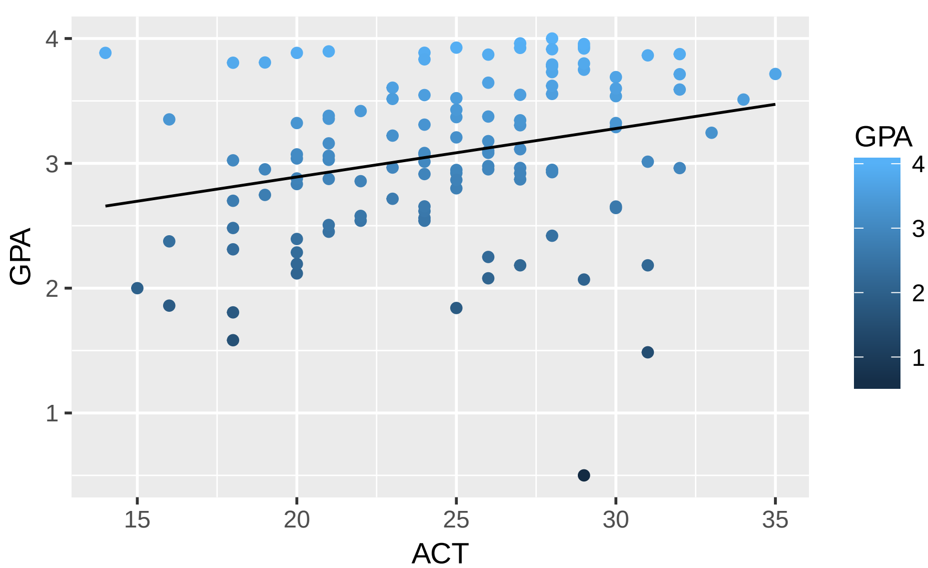
```
gpa <- read.table("CH01PR19.txt", header = FALSE)
names(gpa) <- c("Y", "X")
reg <- lm(Y ~ X, data = gpa)
reg
```

```
##
## Call:
## lm(formula = Y ~ X, data = gpa)
##
## Coefficients:
## (Intercept)            X
##     2.11405      0.03883
```

According to the result, $\hat{\beta}_0 = 18.98$, $\hat{\beta}_1 = 1.87$, and then the estimated regression line is: $Y = 1.87X + 18.98$.

b)

```
library(ggplot2)
pred <- predict(reg)
ggplot(gpa, aes(x = X, y = Y)) + geom_point(aes(color = Y)) +
scale_x_continuous(name = "ACT") + scale_y_continuous(name = "GPA") +
geom_line(aes(y = pred)) + scale_color_continuous(name = "GPA")
```



The function appears to be fit well.

c)

The point estimate is :$\hat{Y} = 1.87 * 30 + 18.98 = 75.08$.

d)

The change is: 1.87.

1.29

It means the expected value of Y is 0 when X is 0. The line fitted will pass through the original point.

1.30

It means variable X has no impact on Y however X changes. The line fitted will be horizontal.

1.33

Our goal is to minimize the following n squared deviations:

$Q = \Sigma_{i=1}^{n}(Y_i - \beta_0)^2$

Taking derivative in terms of β_0, we have:

$\frac{dQ}{d\beta_0} = \Sigma_{i=1}^{n} 2 * (\beta_0 - Y_i)$

Setting the derivative to 0, we find the estimator:

$\hat{\beta}_0 = \frac{\Sigma_{i=1}^{n} Y_i}{n}$.

1.34

By definition, we check the equation below:

$\mathbb{E}(\hat{\beta}_0) = \beta_0$.

Plugging in the result we got in 1.33 to the left hand side, we have:

$\mathbb{E}(\hat{\beta}_0)$

$= \mathbb{E}(\frac{\Sigma_{i=1}^{n} Y_i}{n})$

$= \frac{\Sigma_{i=1}^{n} \mathbb{E}(Y_i)}{n}$

$= \frac{\Sigma_{i=1}^{n} (\beta_0 + \mathbb{E}(\epsilon_i))}{n}$

$= \beta_0$

Thus, we proved that the estimator is unbiased.

1.39

a)

Denote the data as: (X_1, Y_{11}), (X_1, Y_{12}), (X_2, Y_{21}), (X_2, Y_{22}), (X_3, Y_{31}), (X_3, Y_{32}).

And further denote: $\bar{Y_1} = \frac{Y_{11}+Y_{12}}{2}$, $\bar{Y_2} = \frac{Y_{21}+Y_{22}}{2}$, $\bar{Y_3} = \frac{Y_{31}+Y_{32}}{2}$, $\bar{X} = \frac{X_1+X_2+X_3}{3}$, $\bar{Y} = \frac{\Sigma_{i=1}^3 \Sigma_{j=1}^3 Y_{ij}}{6}$.

We first fit the regression line on the original data. According to the formula, we have:

$\hat{\beta}_1 = \frac{\Sigma_{i=1}^3 \Sigma_{j=1}^3 (Y_{ij}-\bar{Y})(X_i-\bar{X})}{2*\Sigma_{i=1}^3 (X_i-\bar{X})^2}$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 * \bar{X}$

Then, we fit it on the following data: $(X_1, \bar{Y_1})$, $(X_2, \bar{Y_2})$, $(X_3, \bar{Y_3})$. The estimate is:

$\hat{\beta}_1{}^* = \frac{\Sigma_{i=1}^3 (X_i-\bar{X})(\bar{Y_i}-\bar{Y})}{\Sigma_{i=1}^3 (X_i-\bar{X})^2}$

$\hat{\beta}_0{}^* = \bar{Y} - \hat{\beta}_1{}^* * \bar{X}$

To prove the two regression lines are identical, we only need to prove:

$\hat{\beta}_1 = \hat{\beta}_1{}^*$

We start from $\hat{\beta}_1$.

$\hat{\beta}_1$

$= \frac{\Sigma_{i=1}^3 (X_i-\bar{X})\Sigma_{j=1}^2 (Y_{ij}-\bar{Y})}{2*\Sigma_{i=1}^3 (X_i-\bar{X})^2}$

$= \frac{\Sigma_{i=1}^3 (X_i-\bar{X})*2*(\bar{Y_i}-\bar{Y})}{2*\Sigma_{i=1}^3 (X_i-\bar{X})^2}$

$= \frac{\Sigma_{i=1}^3 (X_i-\bar{X})(\bar{Y_i}-\bar{Y})}{\Sigma_{i=1}^3 (X_i-\bar{X})^2}$

$= \hat{\beta}_1{}^*$

Thus, we proved they are identical.

b)

Because of the special condition given, we can first simplify the form of $\hat{\beta}_1$ and $\hat{\beta}_0$.

$\hat{\beta}_1$

$= \frac{\Sigma_{i=1}^3 (X_i-\bar{X})(\bar{Y_i}-\bar{Y})}{\Sigma_{i=1}^3 (X_i-\bar{X})^2}$

$= \frac{5(\bar{Y_3}-\bar{Y})-5(\bar{Y_1}-\bar{Y})}{50}$

$= \frac{\bar{Y_3}-\bar{Y_1}}{10}$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 * \bar{X} = \bar{Y} - (\bar{Y_3} - \bar{Y_1})$

We then calculate the estimate of σ^2 based on the original data.

$\hat{\sigma^2}$

$= \frac{1}{4}\Sigma_{i=1}^3 \Sigma_{j=2}^2 (Y_{ij} - \hat{Y_i})^2$

$= \frac{1}{4}\Sigma_{i=1}^3 \Sigma_{j=1}^2 (Y_{ij} - \bar{Y} - (\bar{Y_3} - \bar{Y_1}) * \frac{i-2}{2})^2$

Thus, we can estimate the $\hat{\sigma^2}$ without fitting a regression line just based on the formula above.

1.41

a)

To find the linear square estimator, we need to minimize the following function:

$Q = \Sigma_{i=1}^{n}(Y_i - \beta_1 * X_i)^2$

By taking derivative of Q, we get:

$\frac{dQ}{d\beta_1} = \Sigma_{i=1}^{n} 2 * (\beta_1 * X_i - Y_i)$

Let the derivative be 0. We find the estimator of β_1:

$\hat{\beta}_1 = \frac{\Sigma_{i=1}^{n} Y_i}{\Sigma_{i=1}^{n} X_i}$

b)

For $i = 1, .., n$, the p.d.f of Y_i is:

$f(Y_i) = \frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{Y_i - \beta_1 X_i}{2\sigma^2}\}$

Then, the likelihood function can be written as follow:

$l(\beta_1) = \Pi_{i=1}^{n} f(Y_i) = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{(Y_i - \beta_1 X_i)^2}{2\sigma^2}\} = (\frac{1}{\sqrt{2\pi}\sigma})^n * exp\{-\Sigma_{i=1}^{n} \frac{(Y_i - \beta_1 X_i)^2}{2\sigma^2}\}$

To maximize the likelihood function $l(\beta_1)$ is equivalent to minimize:

$\Sigma_{i=1}^{n}(Y_i - \beta_1 * X_i)^2$.

So the MLE is the same of the solution of question a, which is:

$\hat{\beta}_1 = \frac{\Sigma_{i=1}^{n} Y_i}{\Sigma_{i=1}^{n} X_i}$

c)

We compute the expectation of $\hat{\beta}_1$:

$\mathbb{E}(\hat{\beta}_1)$

$= \mathbb{E}(\frac{\Sigma_{i=1}^{n} Y_i}{\Sigma_{i=1}^{n} X_i})$

$= \frac{\Sigma_{i=1}^{n} \mathbb{E}(Y_i)}{\Sigma_{i=1}^{n} X_i}$

$= \frac{\Sigma_{i=1}^{n} (\mathbb{E}(\epsilon_i) + \beta_1 * X_i)}{\Sigma_{i=1}^{n} X_i}$

$= \frac{\Sigma_{i=1}^{n} (\beta_1 * X_i)}{\Sigma_{i=1}^{n} X_i}$

$= \beta_1$

So by definition, the estimator is unbiased.

1.43

a)

```r
CDI <- read.table("APPENC02.txt", header = FALSE)
reg_1 <- lm(V8 ~ V5, CDI)
reg_2 <- lm(V8 ~ V9, CDI)
reg_3 <- lm(V8 ~ V16, CDI)
reg_1
```

```
##
## Call:
## lm(formula = V8 ~ V5, data = CDI)
##
## Coefficients:
## (Intercept)           V5
##  -1.106e+02    2.795e-03
```

```r
reg_2
```

```
##
## Call:
## lm(formula = V8 ~ V9, data = CDI)
##
## Coefficients:
## (Intercept)           V9
##    -95.9322       0.7431
```

```r
reg_3
```

```
##
## Call:
## lm(formula = V8 ~ V16, data = CDI)
##
## Coefficients:
## (Intercept)          V16
##    -48.3948       0.1317
```

The estimated regression functions are:

$Y = 0.002795 * X_1 - 0.01106$

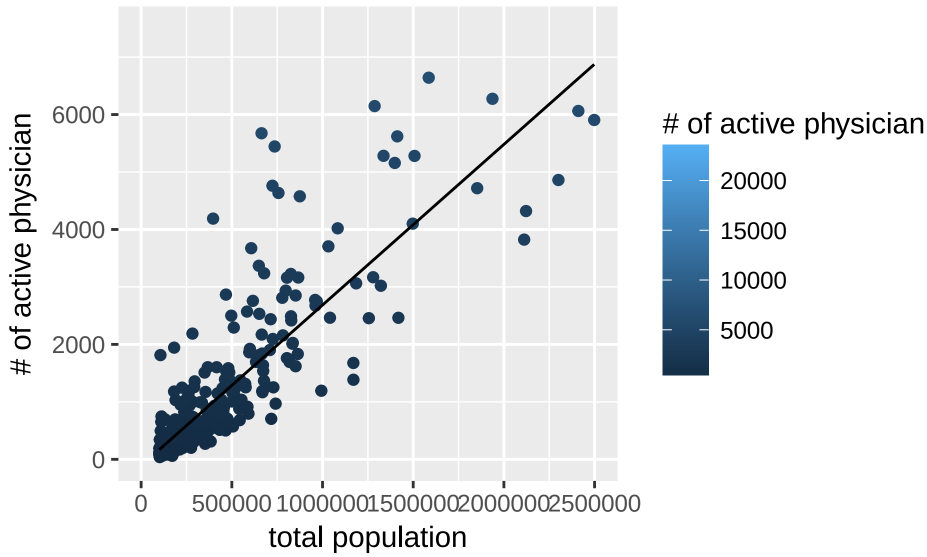
$Y = 0.7431 * X_2 - 95.9322$

$Y = 0.1317 * X_3 - 48.3948$

b)

```r
pred1 <- predict(reg_1)
pred2 <- predict(reg_2)
pred3 <- predict(reg_3)

ggplot(CDI, aes(x = V5, y = V8)) + geom_point(aes(color = V8)) +
scale_x_continuous(name = "total population", limits = c(0, 2500000)) +
scale_y_continuous(name = "# of active physician", limits = c(0, 7500)) +
geom_line(aes(y = pred1)) + scale_color_continuous(name = "# of active physician")
```

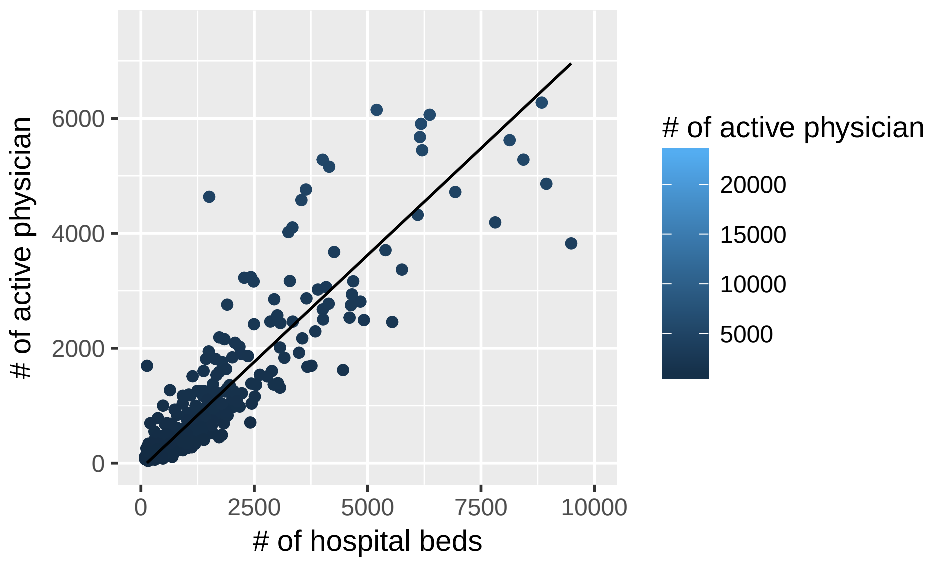
```
## Warning: Removed 3 rows containing missing values (geom_point).

## Warning: Removed 3 rows containing missing values (geom_path).
```




```r
ggplot(CDI, aes(x = V9, y = V8)) + geom_point(aes(color = V8)) +
scale_x_continuous(name = "# of hospital beds", limits = c(0, 10000)) +
scale_y_continuous(name = "# of active physician", limits = c(0, 7500)) +
geom_line(aes(y = pred2)) + scale_color_continuous(name = "# of active physician")
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

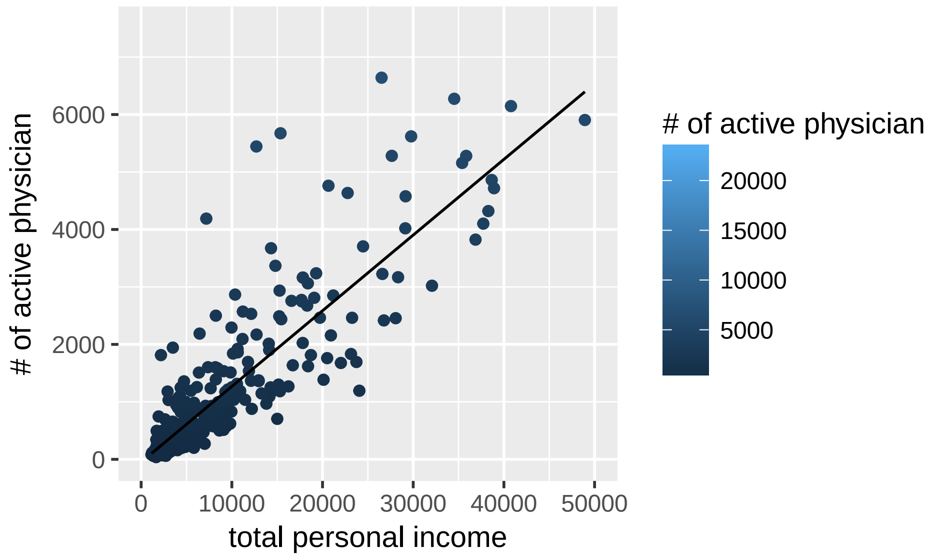
```
## Warning: Removed 10 rows containing missing values (geom_path).
```




```r
ggplot(CDI, aes(x = V16, y = V8)) + geom_point(aes(color = V8)) +
scale_x_continuous(name = "total personal income", limits = c(0, 50000)) +
scale_y_continuous(name = "# of active physician", limits = c(0, 7500)) +
geom_line(aes(y = pred3)) + scale_color_continuous(name = "# of active physician")
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

```
## Warning: Removed 4 rows containing missing values (geom_path).
```



It seems linear regression provides a good fit for each of the three variables.

c)

```
MSE1 = sum((CDI$V8-pred1)^2) / (nrow(CDI)-2)
MSE2 = sum((CDI$V8-pred2)^2) / (nrow(CDI)-2)
MSE3 = sum((CDI$V8-pred3)^2) / (nrow(CDI)-2)
MSE1
```

```
## [1] 372203.5
```

```
MSE2
```

```
## [1] 310191.9
```

```
MSE3
```

```
## [1] 324539.4
```

MSE is calculated above and variable number of hospital beds leads to the smallest variability.

1.44

a)

```
lms <- list()
inds <- list()

for(i in 1:4) {
  ind = CDI$V17 == i;
  lm <- lm(CDI[ind, ]$V15 ~ CDI[ind,]$V12);
  lms[[i]] <- lm
  inds[[i]] <- ind
}

lms[[1]]
```

```
##
## Call:
## lm(formula = CDI[ind, ]$V15 ~ CDI[ind, ]$V12)
##
## Coefficients:
##    (Intercept)  CDI[ind, ]$V12
##         9223.8           522.2
```

```
lms[[2]]
```

```
##
## Call:
## lm(formula = CDI[ind, ]$V15 ~ CDI[ind, ]$V12)
##
## Coefficients:
##    (Intercept)  CDI[ind, ]$V12
##        13581.4           238.7
```

```
lms[[3]]
```

```
##
## Call:
## lm(formula = CDI[ind, ]$V15 ~ CDI[ind, ]$V12)
##
## Coefficients:
##    (Intercept)  CDI[ind, ]$V12
##        10529.8           330.6
```

```
lms[[4]]
```

```
##
## Call:
## lm(formula = CDI[ind, ]$V15 ~ CDI[ind, ]$V12)
##
## Coefficients:
##    (Intercept)  CDI[ind, ]$V12
##         8615.1           440.3
```

The estimated function are listed below:

$Y = 522.2 * X + 9223.8$

$Y = 238.7 * X + 13581.4$

$Y = 330.6 * X + 10529.8$

$Y = 440.3 * X + 8615.1$

b)

Apparently, they are quite different from each other. It might be that there are other factors influencing Y(per capita income).

c)

```
mse = c()
for(i in 1:4) {
```

```r
  pred <- predict(lms[[i]]);
  mse[i] = sum((CDI[inds[[i]], ]$V15-pred)^2) / (sum(inds[[i]])-2)
}
mse
```

```
## [1] 7335008 4411341 7474349 8214318
```

The MSE is provided above. Not all the variability is the same. It might be in some region the variable X(the percentage of individuals having at least a bachelor's degree) contributes more to the response Y(per capita income), in others less.