STATS 413 Hw1

Shu Zhou

2020/9/14

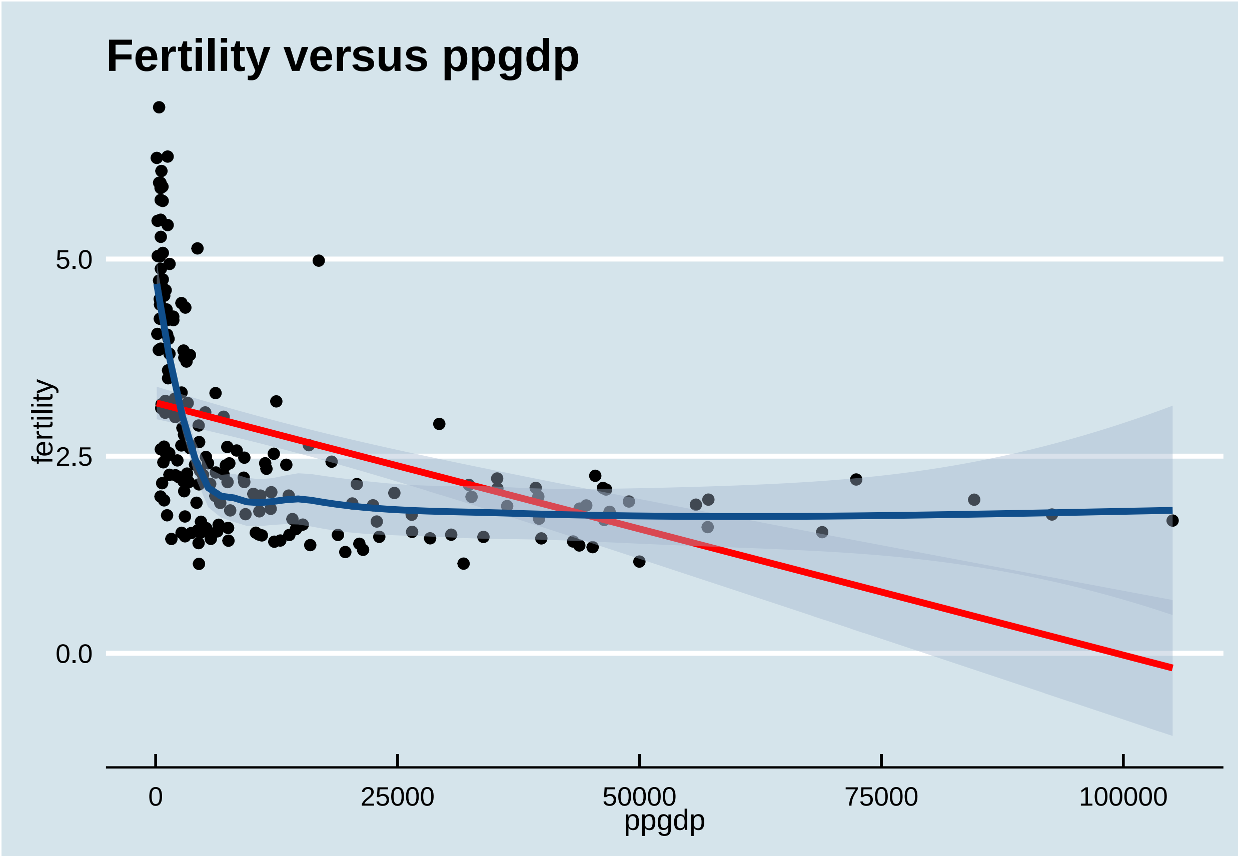
```r
##This is the Assignment 1 of STATS 413
##Author: Shu Zhou
##UMID: 19342932
```

1. United Nations Data

```r
##(a)
##The predictor variable is ppgdp and the response variable is fertility.

##(b)
ggplot(UN11,  aes (y=fertility, x=ppgdp ) )+ geom_point()+
  scale_fill_brewer(palette = "OrRd")+
  geom_smooth(method = "lm", col = "red", fill = "lightsteelblue3", size = 1.2)+
 geom_smooth(method = "loess", col = "dodgerblue4", fill = "lightsteelblue3",
             size = 1.2)+
  ggtitle( "Fertility versus ppgdp")+
  labs(x = "ppgdp", y = "fertility")+
  theme_economist()+
  theme(axis.text.x = element_text(size=10),
        axis.text.y = element_text(size=10), legend.position = "right")
```

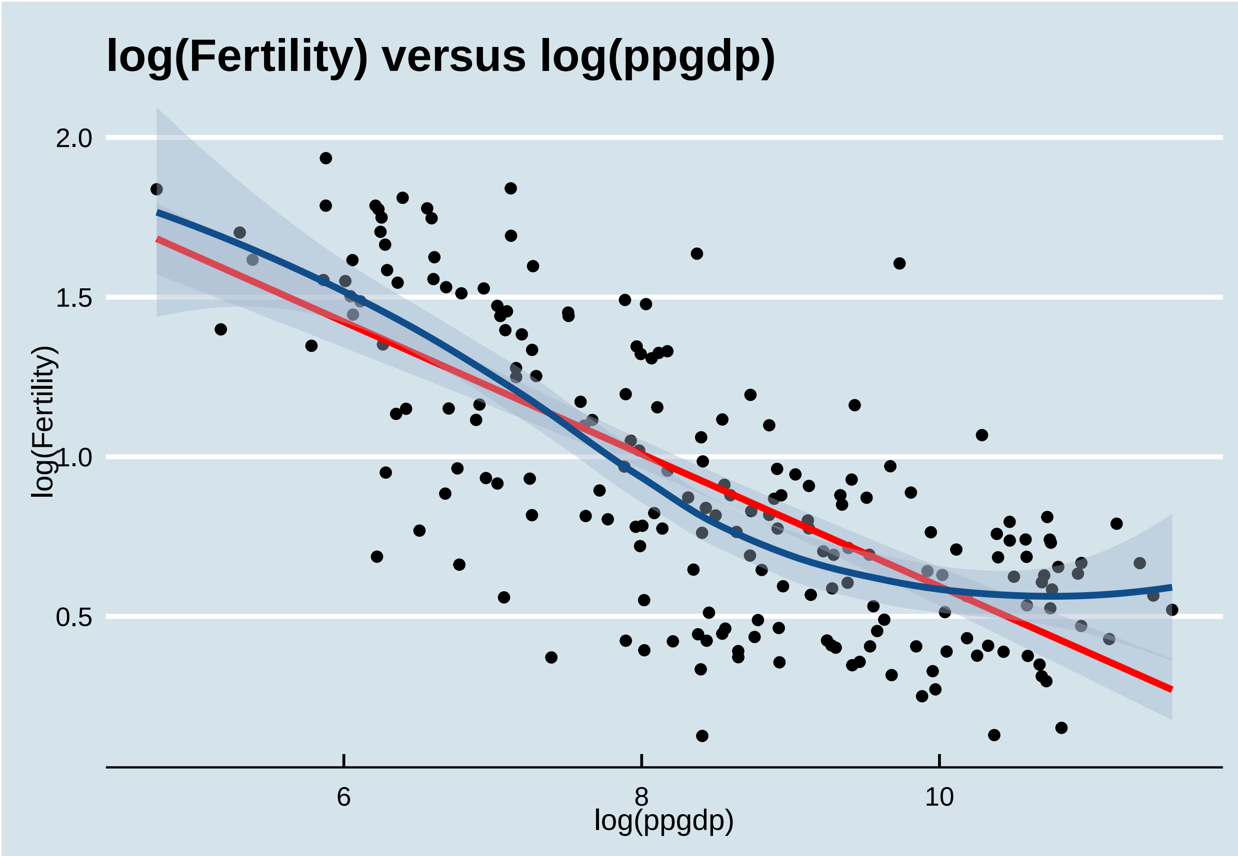
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```




```
##A straight line is not plausible for a summary to this relationship

##c
UN11$logPpgdp = log(UN11$ppgdp)
UN11$logFert = log(UN11$fertility)
ggplot(UN11,  aes (y=logFert, x=logPpgdp ) )+ geom_point()+
  scale_fill_brewer(palette = "OrRd")+
  geom_smooth(method = "lm", col = "red", fill = "lightsteelblue3", size = 1.2)+
 geom_smooth(method = "loess", col = "dodgerblue4", fill = "lightsteelblue3",
             size = 1.2)+
  ggtitle( "log(Fertility) versus log(ppgdp)")+
  labs(x = "log(ppgdp)", y = "log(Fertility)")+
  theme_economist()+
  theme(axis.text.x = element_text(size=10),
        axis.text.y = element_text(size=10), legend.position = "right")
```

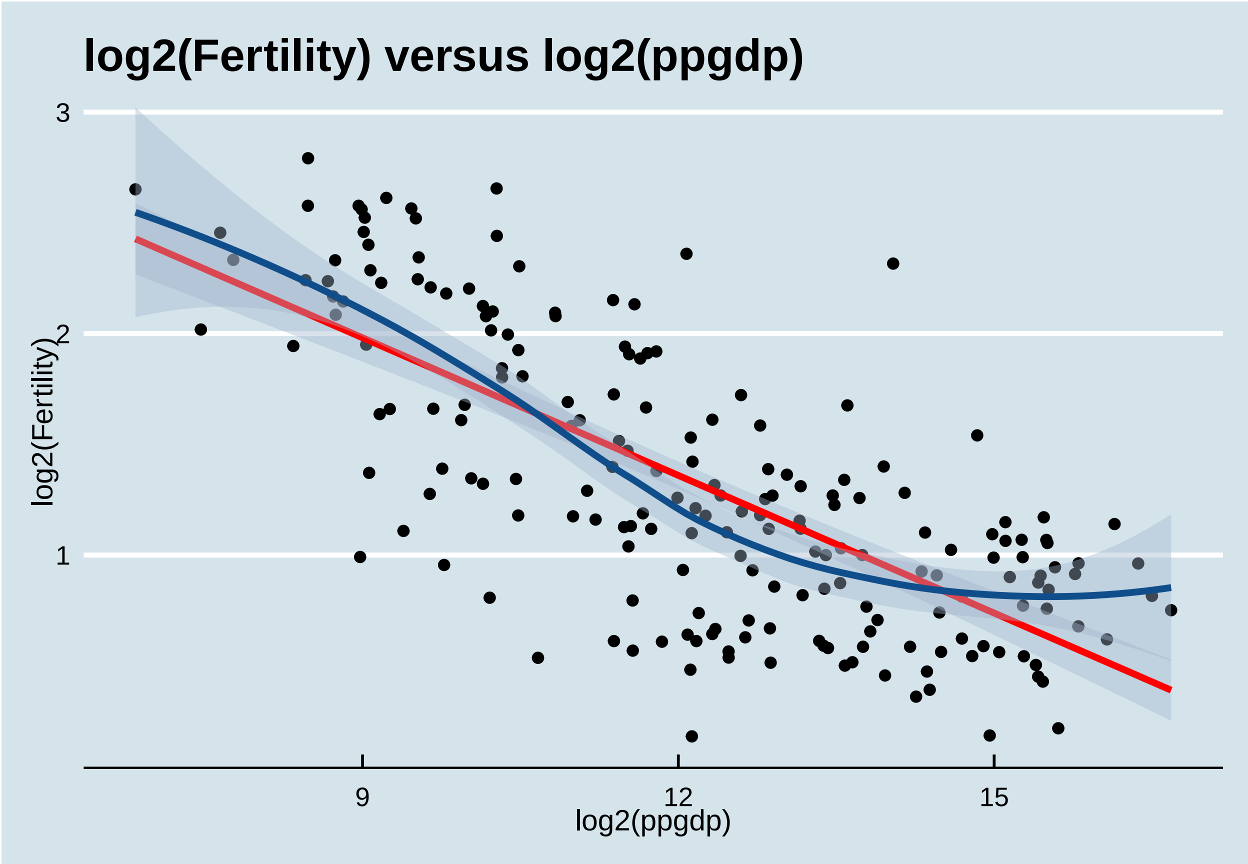
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```




```r
##The simple linear regression mode is much more plausible for a summary to
##this relationship, which shows that the higher log(PPgdp), the lower
## log(Fertility)

UN11$log2Ppgdp = log(UN11$ppgdp,base = 2)
UN11$log2Fert = log(UN11$fertility, base = 2 )
ggplot(UN11,  aes (y=log2Fert, x=log2Ppgdp ) )+ geom_point()+
  scale_fill_brewer(palette = "OrRd")+
  geom_smooth(method = "lm", col = "red", fill = "lightsteelblue3", size = 1.2)+
 geom_smooth(method = "loess", col = "dodgerblue4", fill = "lightsteelblue3",
             size = 1.2)+
  ggtitle( "log2(Fertility) versus log2(ppgdp)")+
  labs(x = "log2(ppgdp)", y = "log2(Fertility)")+
  theme_economist()+
  theme(axis.text.x = element_text(size=10),
        axis.text.y = element_text(size=10), legend.position = "right")
```

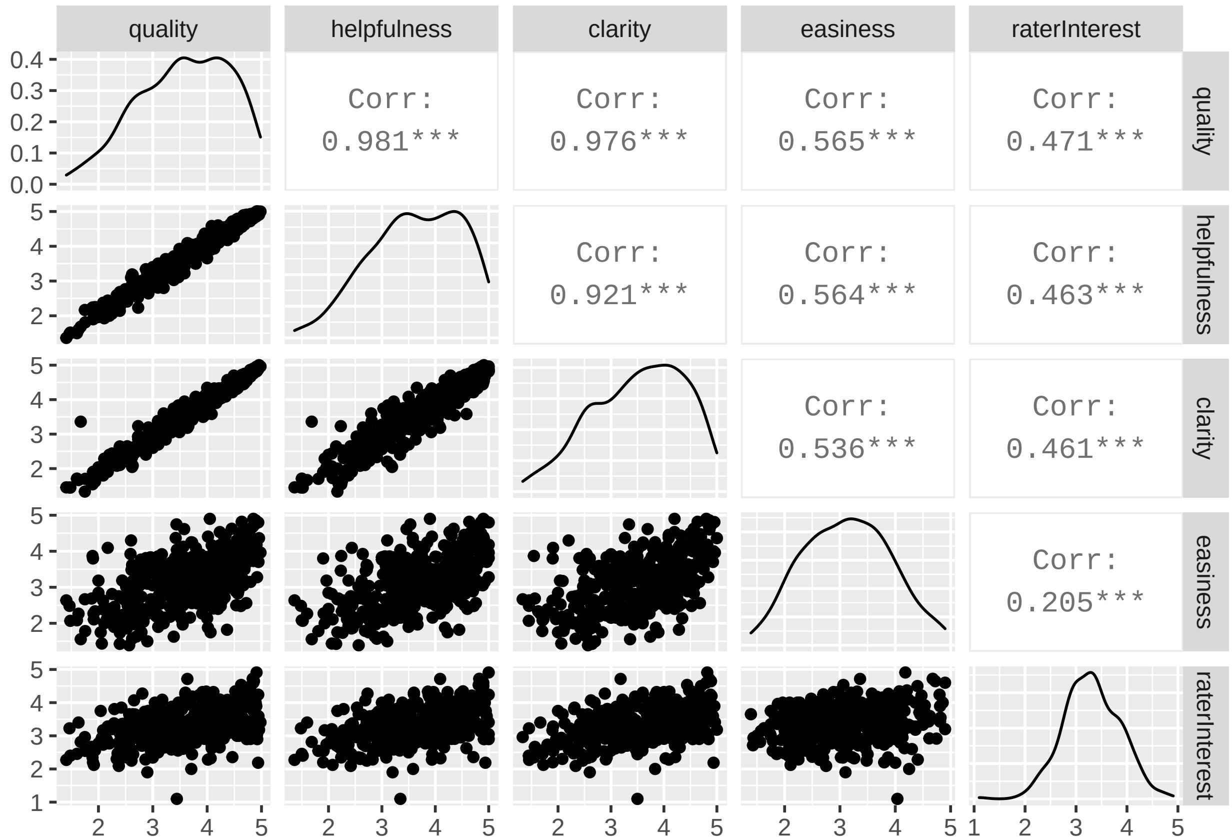
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```




```
##Hence,the shape of the graph won't change, but the values on the axes will.
```

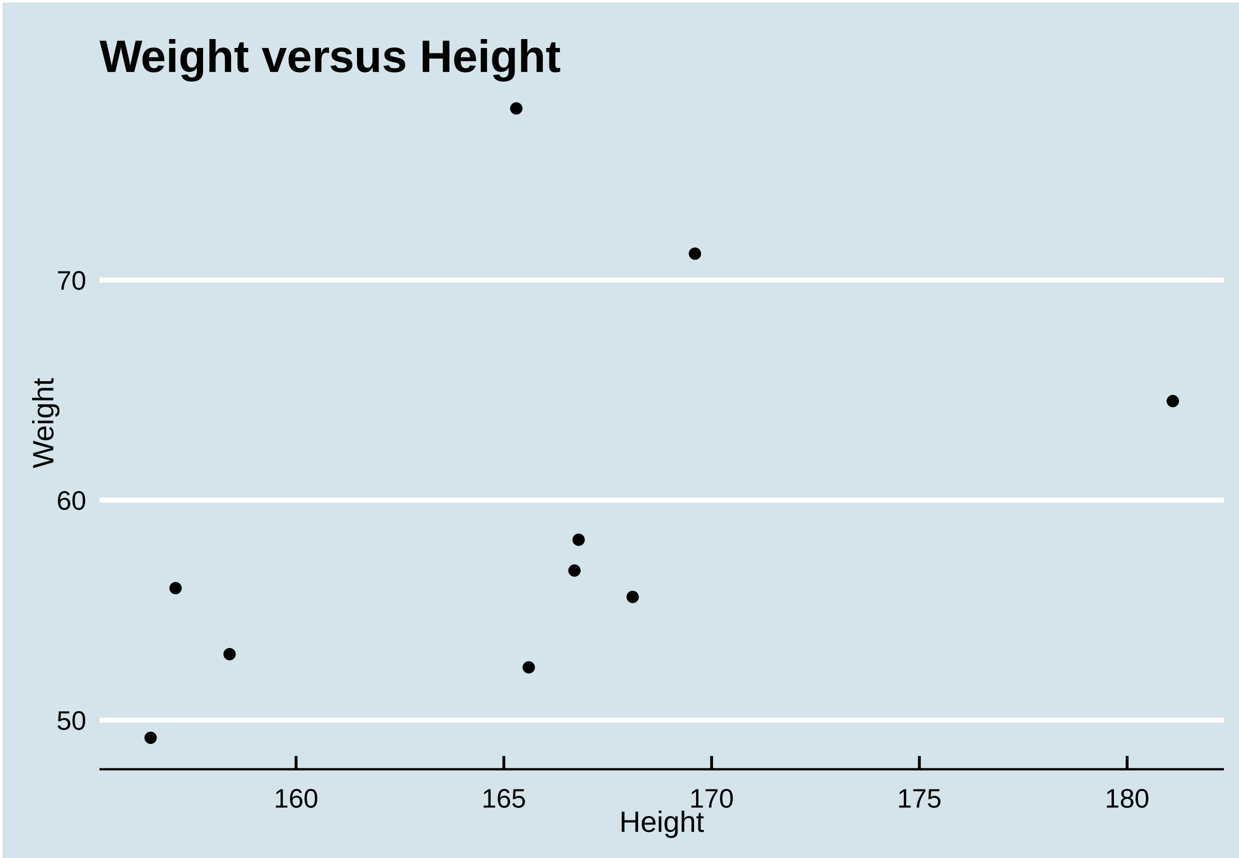
2.Professor ratings Data

```
ggpairs (Rateprof[,9: 13])
```



3.Height and Weight data

```
##(a)
ggplot(Htwt,  aes (y=wt, x=ht ) )+ geom_point()+
  scale_fill_brewer(palette = "OrRd")+
  ggtitle( "Weight versus Height")+
  labs(x = "Height", y = "Weight")+
  theme_economist()+
  theme(axis.text.x = element_text(size=10),
        axis.text.y = element_text(size=10), legend.position = "right")
```



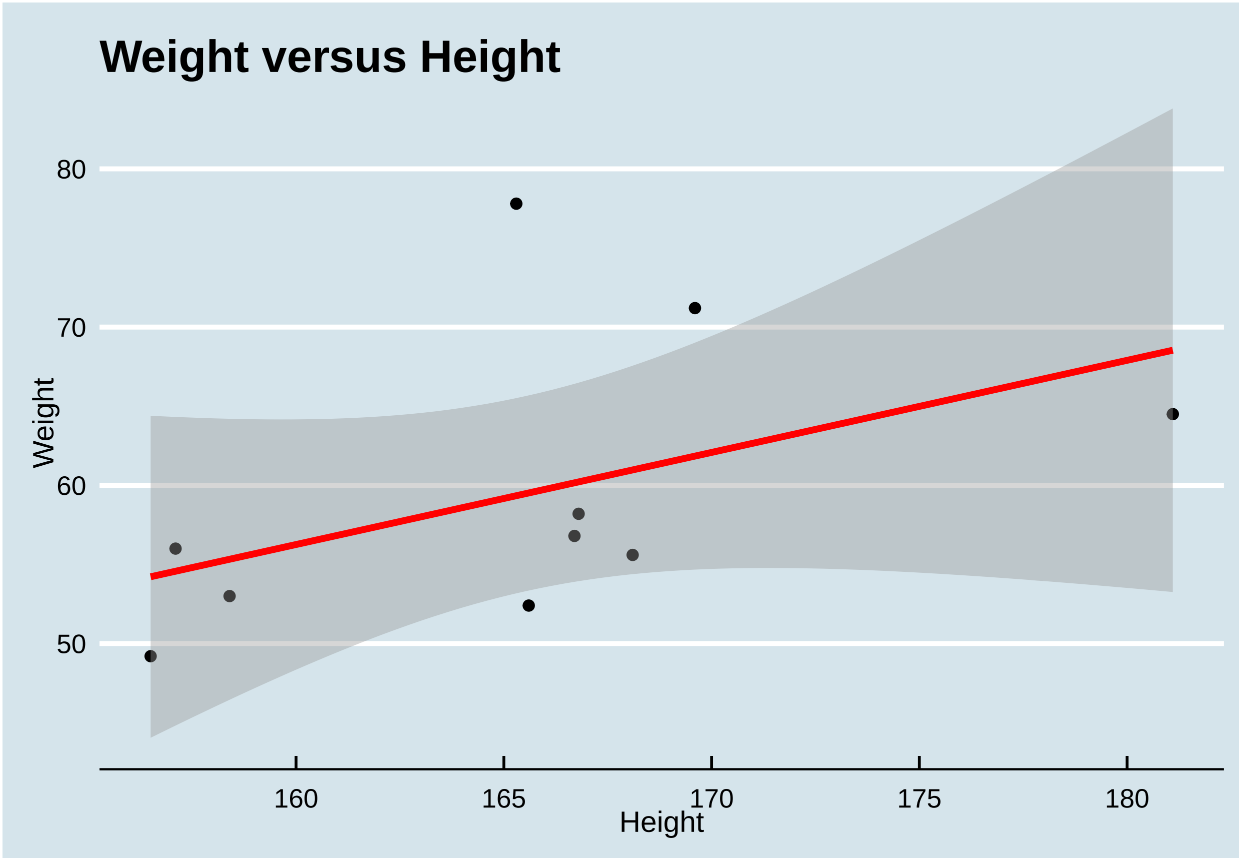

```
##A straight line is not plausible for a summary to this relationship, since
## there is no clear linear pattern between this two variables

##(b)
lm(Htwt$wt ~ Htwt$ht)
```

```
##
## Call:
## lm(formula = Htwt$wt ~ Htwt$ht)
##
## Coefficients:
## (Intercept)      Htwt$ht
##    -36.8759       0.5821
```

```
ggplot(Htwt,  aes (y=wt, x=ht ) )+ geom_point()+
  scale_fill_brewer(palette = "OrRd")+
  geom_smooth(method = "lm", col = "red", size = 1.2)+
  ggtitle( "Weight versus Height")+
  labs(x = "Height", y = "Weight")+
  theme_economist()+
  theme(axis.text.x = element_text(size=10),
        axis.text.y = element_text(size=10), legend.position = "right")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```r
##(c)
fm<-lm(Htwt$wt ~ Htwt$ht)
summary(fm)
```

```
##
## Call:
## lm(formula = Htwt$wt ~ Htwt$ht)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1166 -4.7744 -2.8412  0.5696 18.4581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.8759    64.4728  -0.572    0.583
## Htwt$ht       0.5821     0.3892   1.496    0.173
##
## Residual standard error: 8.456 on 8 degrees of freedom
## Multiple R-squared:  0.2185, Adjusted R-squared:  0.1208
## F-statistic: 2.237 on 1 and 8 DF,  p-value: 0.1731
```

```r
summary(fm)$sigma^2          ##Estimate of Sigma^2
```

```
## [1] 71.5017
```

```r
summary(fm)$coefficient[1,2] ##Estimated Std Error of Intercept
```

```
## [1] 64.4728
```

```r
summary(fm)$coefficient[2,2] ##Estimated Std Error of Slope
```

```
## [1] 0.3891815
```

4. Simple Linear Regression

According to Cramer's rule, the function we calculate $\hat{\beta}_0$ and $\hat{\beta}_1$

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i \tag{1}$$

We divide n on both sides of the equation, hence it becomes

$$\hat{\beta}_0 + \hat{\beta}_1 \frac{\sum x_i}{n} = \frac{\sum y_i}{n} \tag{2}$$

Since $\frac{\sum x_i}{n} = \overline{x_i}$ and $\frac{\sum y_i}{n} = \overline{y_i}$
Hence

$$\hat{\beta}_0 + \hat{\beta}_1 \overline{x_i} = \overline{x_i} \tag{3}$$

Which shows that the least squared line always passes through the mean point

5. Multi-task Regression

a

We first calculate the $\hat{\beta}$ that minimizes the RSS

$$y = X\hat{\beta} \tag{4}$$

$$(X^T X)\hat{\beta} = X^T y \tag{5}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{6}$$

Hence RSS is calculated by

$$RSS = \hat{e}^T \hat{e} = (y - X\hat{\beta})^T (y - X\hat{\beta}) = y^T y - 2\hat{\beta}^T X^T y + \hat{\beta}^T X^T X \hat{\beta} \tag{7}$$

Hence

$$RSS = y^T [1 - X(X^T X)^{-1} X^T] y \tag{8}$$

b

We have already calculated the $\hat{\beta}$ in (a) eq. (4) - eq. (6).

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{9}$$

c

The regression coefficients from the m separate regressions will be different from the matrix of regression coefficients that minimizes the RSS.

When we derive a multiple regression, any element in the vector $\hat{\beta}$ minimizes the the correlation between any dependent variable y and the regressor.

So, in this problem, when we try to calculate the regression coefficients separately, each coefficient was not significant enough.

As a result, the RSS calculated form the regression coefficients in this problem will be much larger.=