R Notebook

Hyunjoon Rhee

2/5/2021

```
library(faraway)
data(prostate)
```

Problem 1 a

```
library(faraway)
data(prostate)
summary(prostate)
```

```
##      lcavol           lweight           age             lbph        
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863  
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863  
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001  
##  Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004  
##  3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581  
##  Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263  
##       svi              lcp             gleason          pgg45       
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00  
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00  
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00  
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38  
##  3rd Qu.:0.0000   3rd Qu.: 1.1786   3rd Qu.:7.000   3rd Qu.: 40.00  
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00  
##       lpsa        
##  Min.   :-0.4308  
##  1st Qu.: 1.7317  
##  Median : 2.5915  
##  Mean   : 2.4784  
##  3rd Qu.: 3.0564  
##  Max.   : 5.5829
```

Above result are the minimum, 1st quartile, median, mean, 3rd quartile, maximum of the data. As it is shown in the summary, svi and gleason shows a categorical data. Whereas other continuous variables show good range of distribution of the data. Age and pgg45 could be categorized as discrete variable.

Problem 1 b

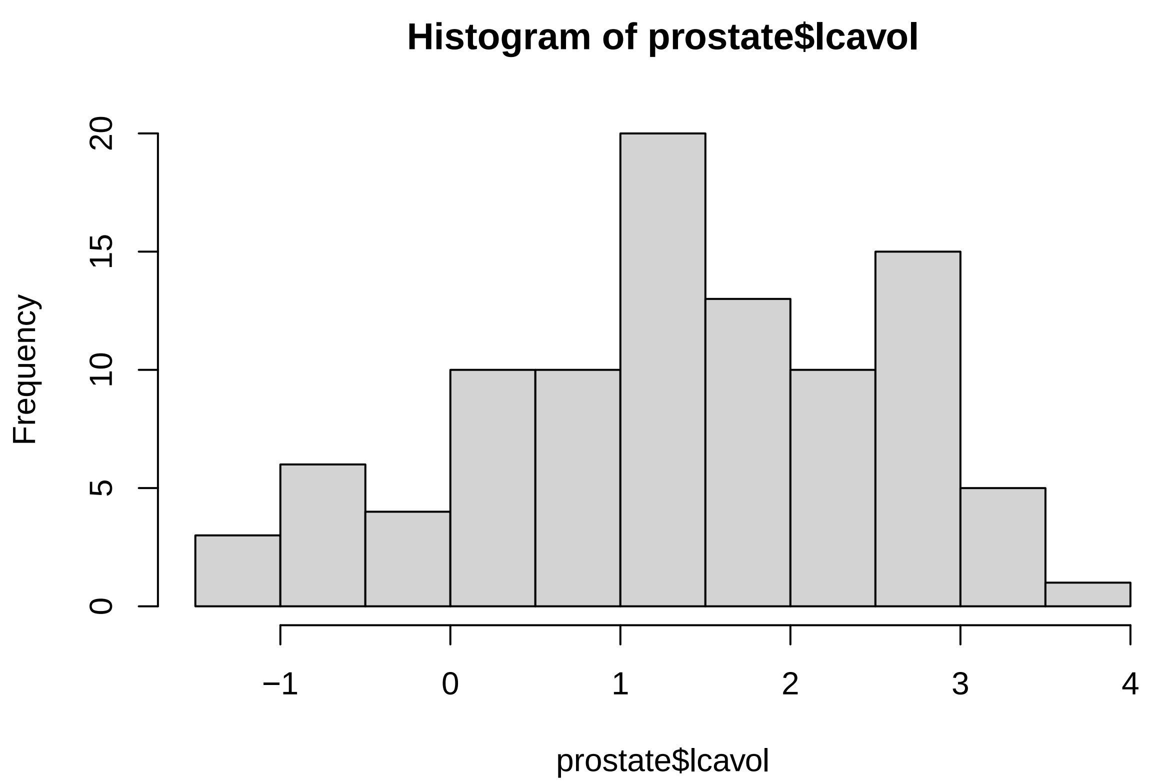
```
prostate$gleason <- factor(prostate$gleason)
prostate$svi <- factor(prostate$svi)
summary(prostate)
```

```
##      lcavol           lweight          age             lbph          svi
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863   0:76
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863   1:21
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001
##  Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004
##  3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581
##  Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263
##       lcp            gleason     pgg45             lpsa
##  Min.   :-1.3863   6:35    Min.   :  0.00   Min.   :-0.4308
##  1st Qu.:-1.3863   7:56    1st Qu.:  0.00   1st Qu.: 1.7317
##  Median :-0.7985   8: 1    Median : 15.00   Median : 2.5915
##  Mean   :-0.1794   9: 5    Mean   : 24.38   Mean   : 2.4784
##  3rd Qu.: 1.1786           3rd Qu.: 40.00   3rd Qu.: 3.0564
##  Max.   : 2.9042           Max.   :100.00   Max.   : 5.5829
```

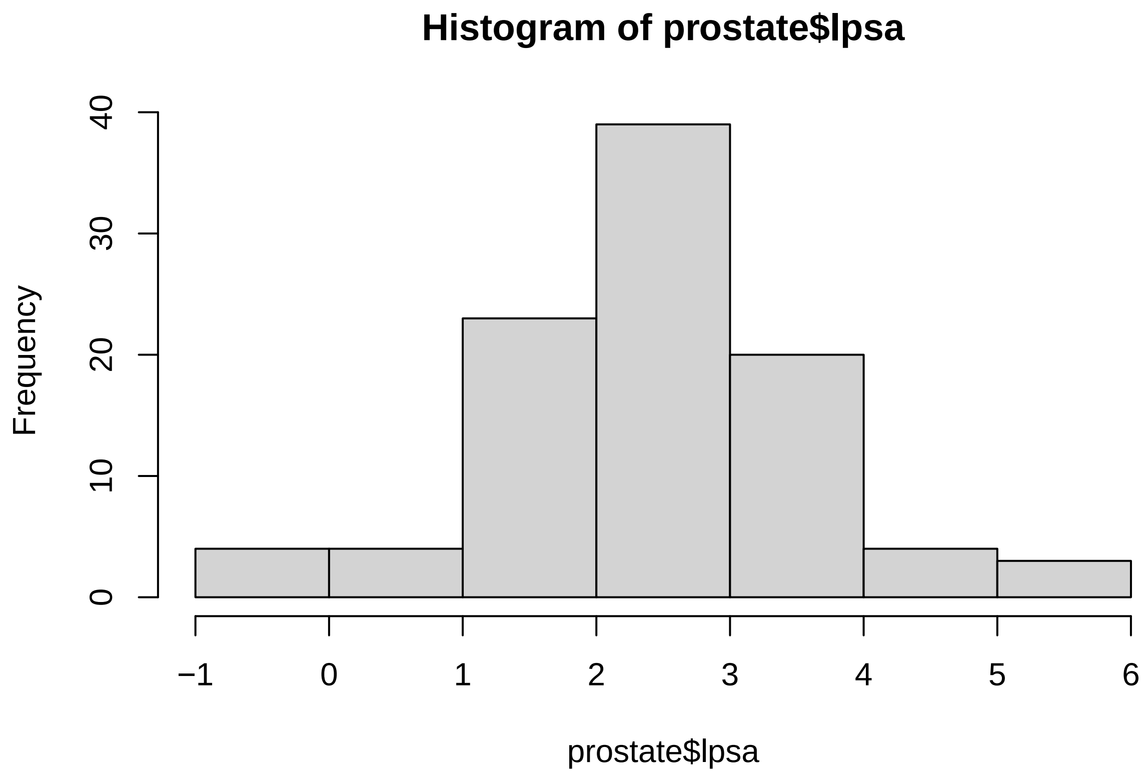
By using factor it became clearer for svi and gleason to be categorized into number of occurence in the data.

Problem 1 c

```r
hist(prostate$lcavol)
```




```r
hist(prostate$lpsa)
```



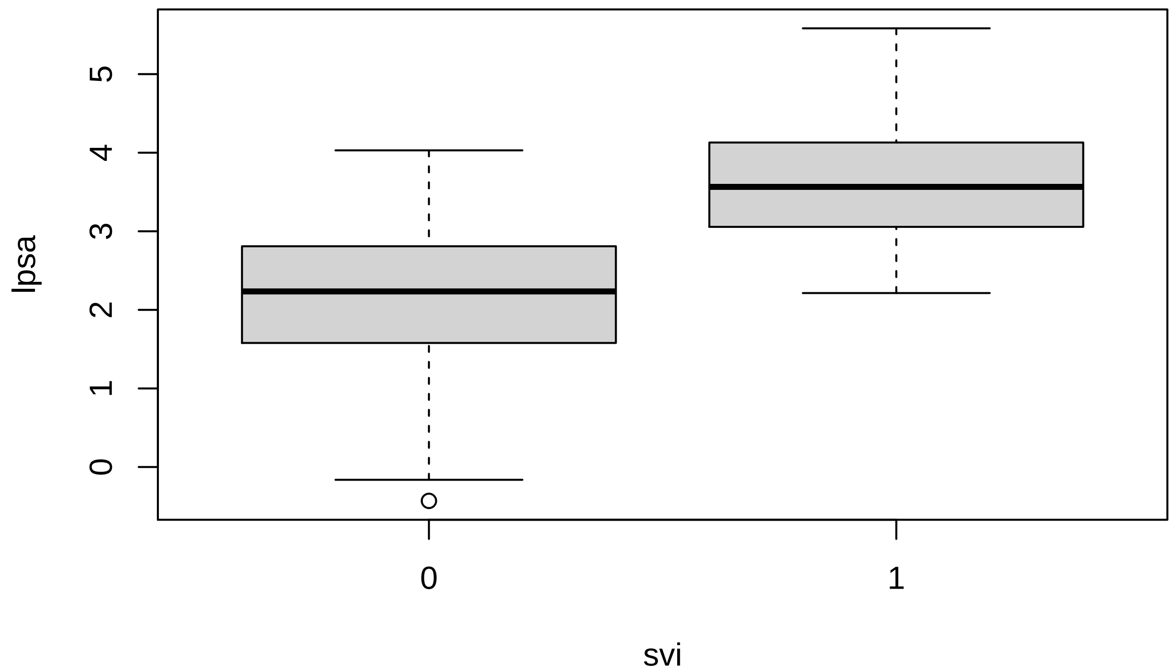

```
hist(prostate$lbph)
```


```r
hist(prostate$lcp)
```

Histogram for lcavol and lpsa seems to have a normal distribution although futher analysis is required. The remaining three are skewed to the left.

Problem 1 d

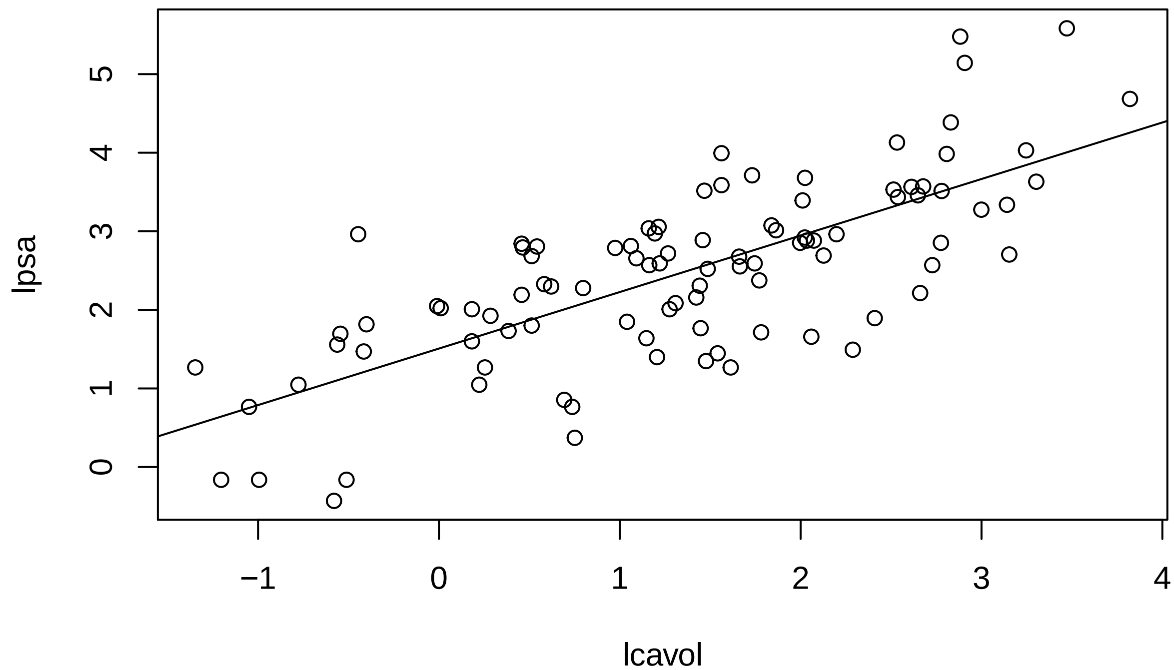
```
plot(lpsa ~ svi, prostate)
```



Because the plot is with categorical variable the plot looks abnormal and hard to analyze. But there seems to be a difference in the mean of lpsa with depending on svi.

Problem 1 e

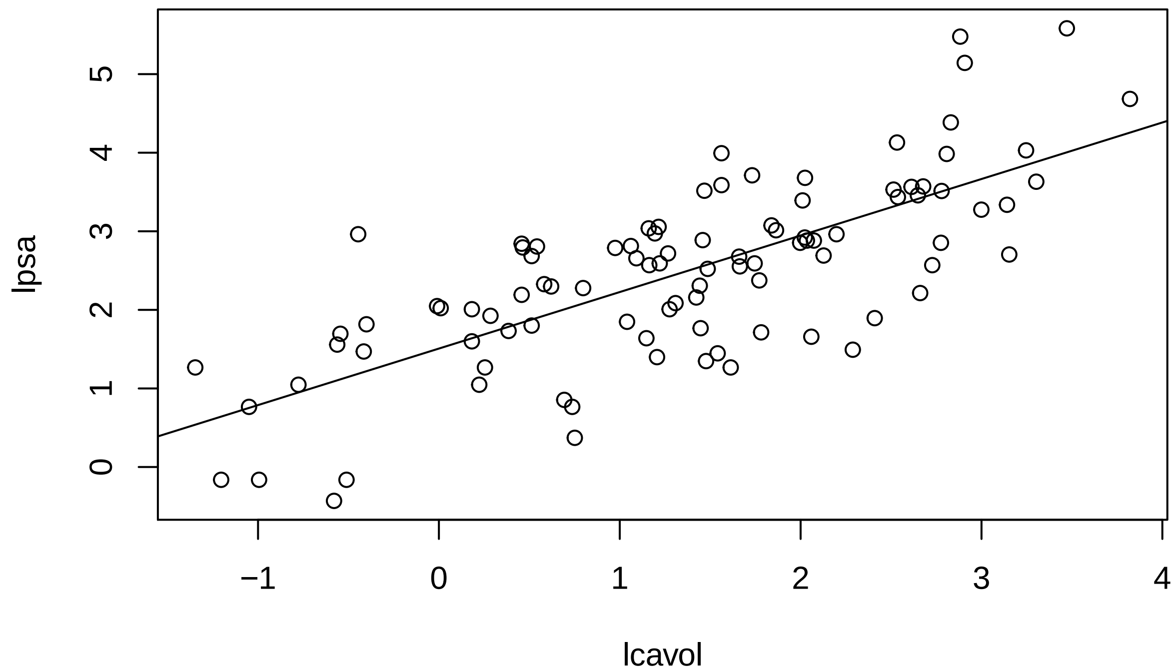
```
plot(lpsa ~ lcavol, prostate)
abline(lm(lpsa ~ lcavol, prostate))
```



This plot seems to show that there might be linear relationship between lpsa and lcavol. Further analysis is required to check it.

Problem 2 a

```
a = lm(lpsa ~ lcavol, prostate)
plot(lpsa ~ lcavol, prostate)
abline(lm(lpsa ~ lcavol, prostate))
```




```r
summary(a)$coef
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 1.5072979 0.12193682 12.36130 1.722234e-21
## lcavol      0.7193201 0.06819288 10.54832 1.118616e-17
```

The slope is 0.7193 and the intercept is 1.50729. This regression shows that there is a linear relationship between the two.

Problem 2 b

```r
a = lm(lpsa ~ lcavol, prostate)
summary(a)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67625 -0.41648  0.09859  0.50709  1.89673
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.50730    0.12194   12.36   <2e-16 ***
```

```
## lcavol       0.71932    0.06819   10.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7875 on 95 degrees of freedom
## Multiple R-squared:  0.5394, Adjusted R-squared:  0.5346
## F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

The residual standard error is 0.7875 with R squared value of 0.5394. The R squared value shows that it shows 53 percentage variation.

Problem 2 c

```r
mean(resid(a))
```

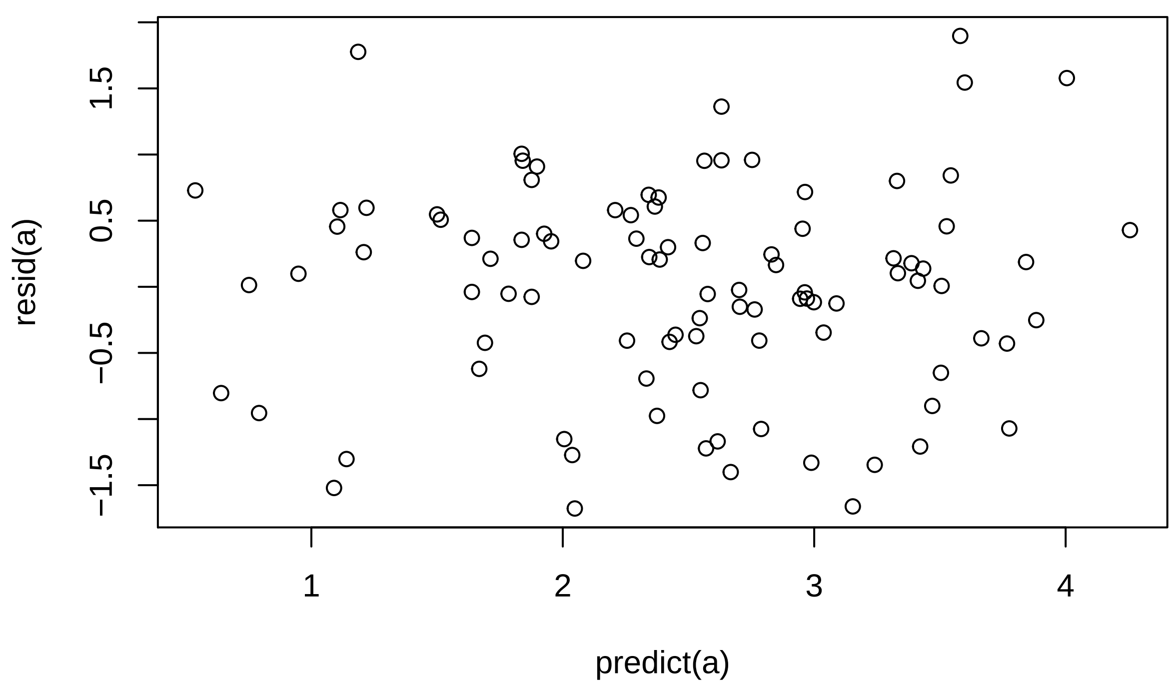
```
## [1] -7.886956e-17
```

```r
median(resid(a))
```

```
## [1] 0.09859487
```

Problem 2 d

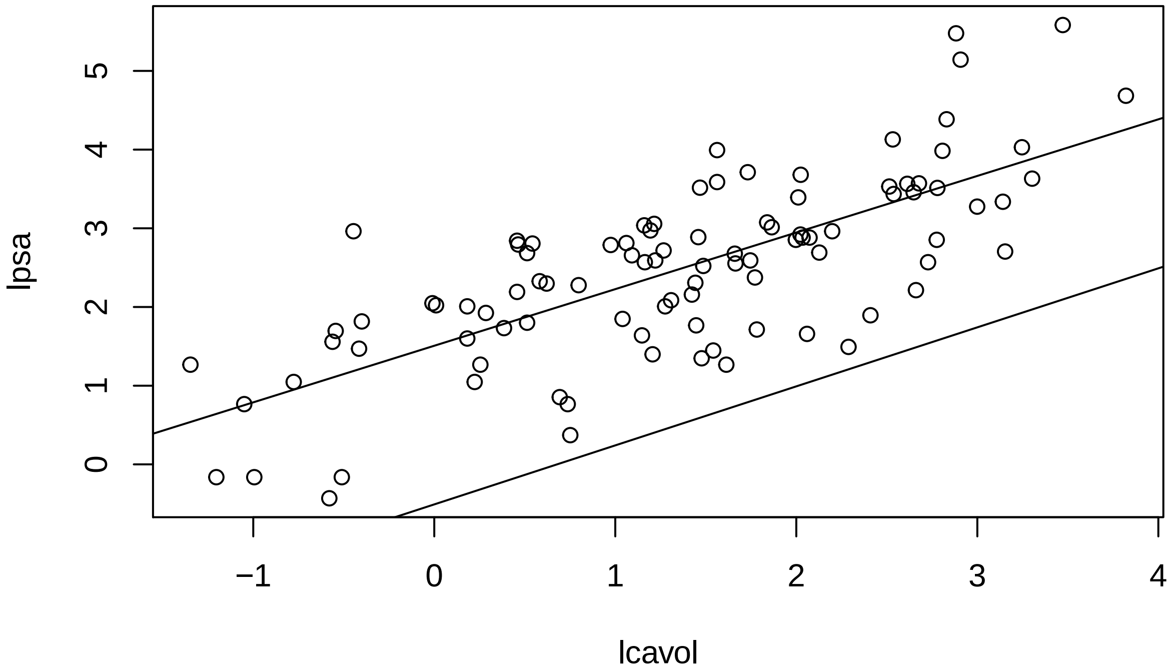
```r
plot(predict(a),resid(a))
```



The fitted vs residual plot shows that there is a constant variance in the data.

Problem 2 e

```
plot(lpsa ~ lcavol, prostate)
abline(lm(lpsa ~ lcavol, prostate))
abline(lm(lcavol~lpsa,prostate))
```



The two lines do not intersect.

3\. a) $Y_i = \beta x_i + e_i$

$$e_i = y_i - \beta x_i$$

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

$$\frac{\partial s}{\partial \beta} = 0$$

$$2 \sum_{i=1}^{n} (y_i - \beta x_i)(-x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta} \sum_{i=1}^{n} x_i^2 = 0$$

$$\hat{\beta} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

b) i) $E(\hat{\beta} / x) = \beta$

$$E(\hat{\beta} / x) = E\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \middle| x \right)$$

$$= \frac{\sum x_i \, E(y_i)}{\sum_{i=1}^{n} x_i^2}$$ x is given

$$= \frac{\sum_{i=1}^{n} x_i E(\beta x_i + e_i)}{\sum_{i=1}^{n} x_i^2}$$

$$= \frac{\sum_{i=1}^{n} x_i E(\beta x_i + 0)}{\sum_{i=1}^{n} x_i^2}$$

$$= \frac{\beta \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} = \beta$$

ii) $var(\hat{\beta} / x) = \dfrac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$

$$= var\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} / x \right)$$

$$\frac{1}{\sum_{i=1}^{n} (x_i^2)^2} var\left(\sum_{i=1}^{n} x_i y_i / x \right)$$

$$= \frac{1}{(\sum_{i=1}^{n} x_i^2)^2} \sum_{i=1}^{n} x_i^2 \, v(y_i | x)$$

$$= \frac{1}{(\sum_{i=1}^{n} x_i^2)^2} (\sum_{i=1}^{n} x_i^2) \, v(\beta x_i + e_i)$$

$$= \frac{1}{\sum_{i=1}^{n} x_i^2} v(e_i) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

iii) $\hat{\beta} \mid X \sim N\left(\beta, \dfrac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right)$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$= \sum_{i=1}^{n} c_i y_i \quad \cdots \quad c_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2} \Rightarrow \text{constant}$$

$$y_i \sim N(\beta x_i, \sigma^2)$$

$$E(\hat{\beta}) = \beta, \quad Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

$$\hat{\beta} \mid x \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right)$$

4 a) $(y_i-\hat{y}_i) = (y_i-\bar{y}) - \hat{\beta}_1(x_i-\bar{x})$

$$\hat{y}_i = \hat{\beta}_1 x_i + \beta_0$$

$$\bar{y} = E[y_i] = \hat{\beta}_1\bar{x} + \beta_0$$

$$y_i - (\hat{\beta}_1 x_i + \beta_0) = y_i - (\hat{\beta}_1\bar{x} + \beta_0) - (\hat{\beta}_1 x_i + \beta_0) + (\hat{\beta}_1\bar{x} + \beta_0)$$

$$= y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})$$

b) $\hat{y}_i - \bar{y} = \hat{\beta}_1 x_i + \beta_0 - \hat{\beta}_1 \bar{x} + \beta_0)$

$\therefore\ \hat{y}_i - y = \beta_1(x_i - x)$

c) $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$

$$\sum_{i=1}^{n}(y_i - y_i^{\wedge})(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}[(y_i-\bar{y}) - \hat{\beta}_1(x_i-\bar{x})][\hat{\beta}_1(x_i-\bar{x})]$$

$$= \hat{\beta}_1\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x}) - \hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$= \hat{\beta}_1 S_{xy} - \hat{\beta}_1^2 S_{xx}$$

$$= \frac{S_{xy}}{S_{xx}} S_{xy} - \frac{S_{xy}^2}{S_{xx}^2} S_{xx}$$

$$= \frac{S_{xy}^2}{S_{xx}} - \frac{S_{xy}^2}{S_{xx}} = 0$$

d) $\sum_{i=1}^{n} (y_i - \bar{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$

$$= \sum_{i=1}^{n} \left[(y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + \underbrace{2 (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{\to 0} \right]$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$Tss \quad = \quad Fss \quad + \quad Rss$$