# Final, MA 331-B

Daniel Detore

May 15, 2025

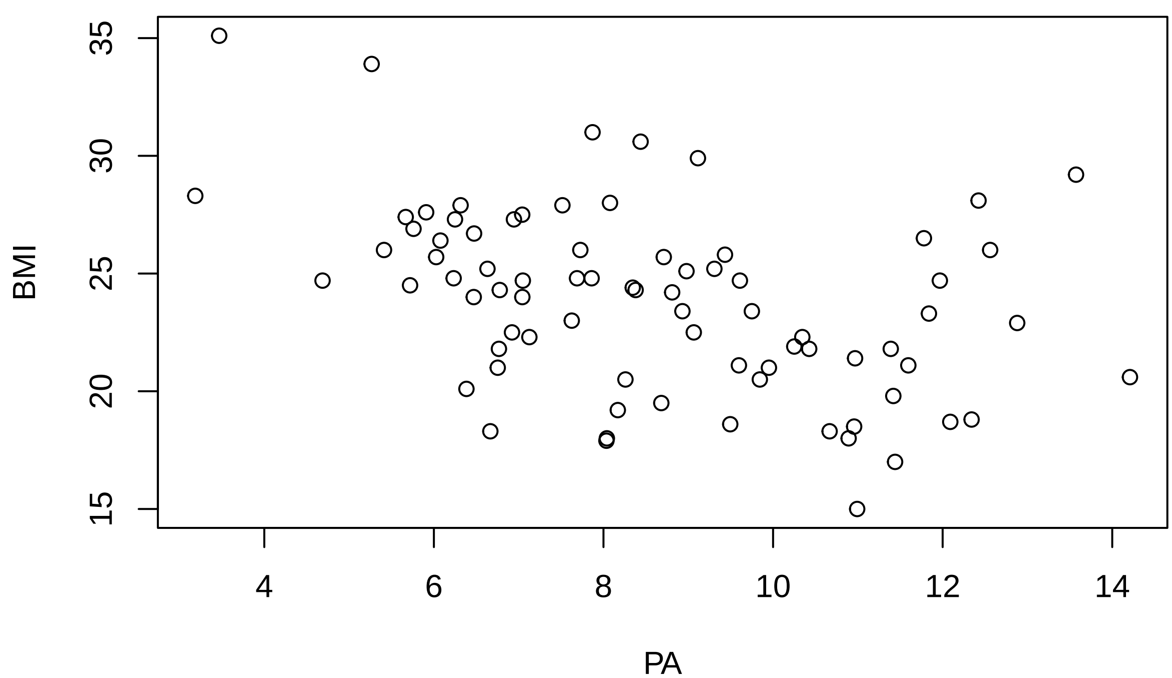
I pledge my honor that I have abided by the Stevens Honor System.

## Problem 1

```r
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.4.3
```

```r
pabmi <- read_excel("pabmi.xls")
plot(pabmi)
```




```r
PA <- pabmi$PA
BMI <- pabmi$BMI
n <- length(PA)
```

```r
cor(PA, BMI)
```

```
## [1] -0.428121
```

We can check for significance of association between Y and X by testing on the slope parameter.

We have $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$. Our testing statistic $t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}} \sim T_{n-2=77} = -4.16$.

```r
cor.test(BMI, PA)
```

```
##
##  Pearson's product-moment correlation
##
## data:  BMI and PA
## t = -4.157, df = 77, p-value = 8.291e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5930885 -0.2286556
## sample estimates:
##       cor
## -0.428121
```

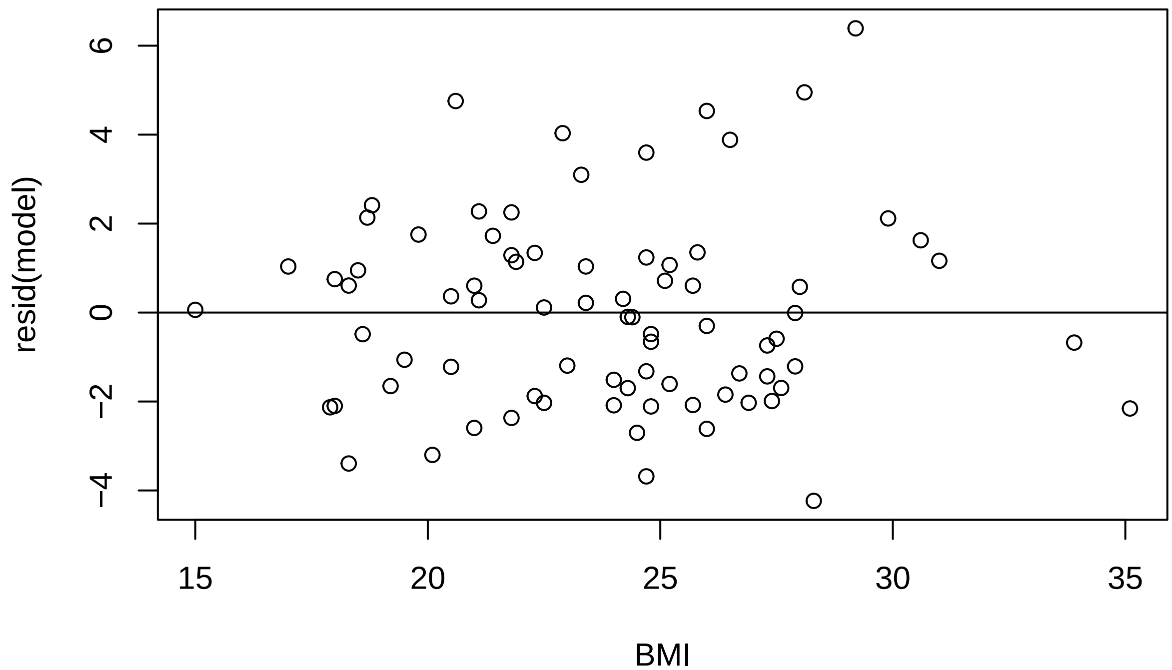
This p-value is very low so we may reject $H_0$ and assume $H_a$. Thus X and Y are significantly associated.

# Problem 2

```r
model <- lm(PA ~ BMI, pabmi);
sm <- summary(model); sm
```

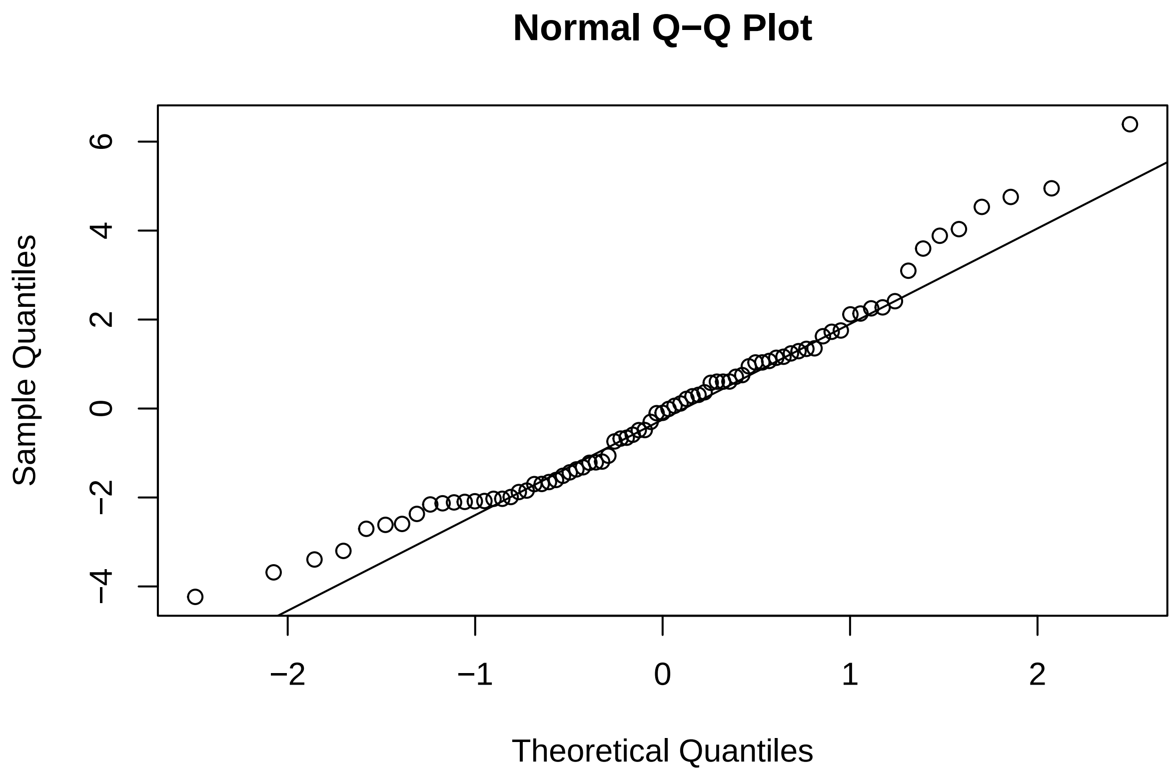
```
##
## Call:
## lm(formula = PA ~ BMI, data = pabmi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2341 -1.6975 -0.0971  1.2011  6.3905
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.89115    1.53333   9.712 5.14e-15 ***
## BMI         -0.26399    0.06351  -4.157 8.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.178 on 77 degrees of freedom
## Multiple R-squared:  0.1833, Adjusted R-squared:  0.1727
## F-statistic: 17.28 on 1 and 77 DF,  p-value: 8.291e-05
```

```r
b0 <- sm$coefficients[1, 1]
b1 <- sm$coefficients[2, 1]
plot(BMI, resid(model))
abline(h = 0)
```




The residuals have no relation pattern, which shows that this model is a good fit for the data.

```
qqnorm(resid(model))
qqline(resid(model))
```




Some of the QQ plot seems to somewhat follow a straight line, which means this model might not be a perfect fit for the data.

## Problem 3

```
b0
```

```
## [1] 14.89115
```

```
b1
```

```
## [1] -0.2639942
```

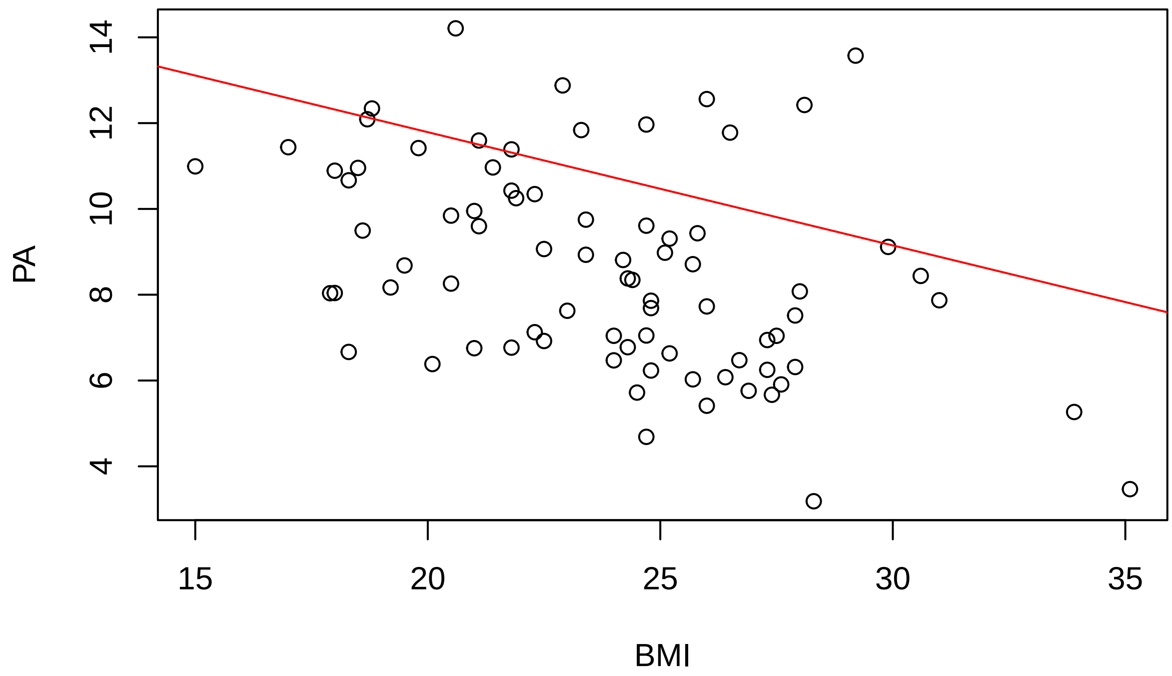
```
e <- sm$sigma; e
```

```
## [1] 2.177997
```

```
sprintf('Our regression equation is: Y =  %fX + %f + %f',
    b1, b0, e)
```

```
## [1] "Our regression equation is: Y =  -0.263994X + 14.891148 + 2.177997"
```

```
plot(BMI, PA)
abline(b0 + e, b1, col="red")
```




## Problem 4

The CI of the intercept parameter is $\hat{\beta}_0 \pm t_{1-\alpha/2}(n-2) \times SE_{\hat{\beta}_0}$.

```
cl <- .95
SEb0 <- sm$coef[1,2]
MOE <- pt(1 - (cl / 2), n-2) * SEb0; MOE
```

```
## [1] 1.072493
```

```
sprintf('Our confidence interval is: [%f, %f]',
    b0 - MOE,
    b0 + MOE)
```

```
## [1] "Our confidence interval is: [13.818655, 15.963641]"
```

The CI of the slope parameter is $\hat{\beta}_0 \pm t_{1-\alpha/2}(n-2) \times SE_{\hat{\beta}_0}$.

```
SEb1 <- sm$coef[2,2]
MOE <- pt(1 - (cl / 2), n-2) * SEb1; MOE
```

```
## [1] 0.04441977
```

```
sprintf('Our confidence interval is: [%f, %f]',
    b1 - MOE,
    b1 + MOE)
```

```
## [1] "Our confidence interval is: [-0.308414, -0.219574]"
```

## Problem 5

To test the significance of slope, we have $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$ and we observe test statistic $t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}} \sim T_{n-2} =$

```
t <- b1/SEb1; t
```

```
## [1] -4.156974
```

```
2 * (1 - pt(abs(t), n-2))
```

```
## [1] 8.290511e-05
```

Since this value is $< 0.01$ we can reject $H_0$ and assume $H_a$.

To test the significance of the intercept, we have $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$ and we observe test statistic $t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}} \sim T_{n-2} =$

```
t <- b0/SEb0; t
```

```
## [1] 9.71165
```

```
2 * (1 - pt(abs(t), n-2))
```

```
## [1] 5.107026e-15
```

Since this value is $< 0.01$ we can reject $H_0$ and assume $H_a$.

## Problem 6

Our coefficient of determination $R^2 =$

```
RR <- sm$r.squared; RR
```

```
## [1] 0.1832876
```

We know $RSE = S = \sqrt{\frac{SSE}{n-2}} \implies SSE = RSE^2 \times (n-2) =$

```
RSE <- sm$sigma
SSE <- (RSE^2) * (n-2); SSE
```

```
## [1] 365.2625
```

We also know $SST = SSM + SSE$ and $R^2 = \frac{SSM}{SST} \implies SSM = R^2 \times SST$ thus $SST = R^2 \times SST + SSE \implies SST - R^2 \times SST = SSE \implies SST(1 - R^2) = SSE \implies SST = \frac{SSE}{1-R^2} =$

```
SST <- SSE / (1 - RR); SST
```

```
## [1] 447.2352
```

thus

```
SSM <- SST - SSE; SSM
```

```
## [1] 81.97267
```

## Problem 7

We have $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$. Our testing statistic $f \sim F_{1,n-2} = \frac{SSM}{SSE/(n-2)} = \frac{SST-SSE}{SSE/(n-2)} =$

```r
f <- (SST-SSE) / (SSE/(n-2)); f
```

```
## [1] 17.28044
```

```r
# which gives us p-value:
1 - pf(f, 1, n-2)
```

```
## [1] 8.290511e-05
```

$< 0.01$. Thus we reject $H_0$ and assume $H_a$.

# Problem 8

We can estimate $Y^*$ as $y^* = \beta_0 + \beta_1 x^*$ (ignoring $\varepsilon^*$ to get expected value). When $x^* = 27.55, 31.5$, $y^* =$

```r
b0 + (27.55 * b1)
```

```
## [1] 7.618108
```

```r
b0 + (31.5 * b1)
```

```
## [1] 6.575331
```

```r
SS <- (1 / (n-2)) * sum((PA - mean(PA))^2)
SEuY <- sqrt((1 / n) + ((27.55 - mean(BMI))^2)/sum((BMI-mean(BMI))^2)*SS)
MOE <- pt(1 - (cl / 2), n-2) * SEuY; MOE
```

```
## [1] 0.1988727
```

```r
sprintf('Our confidence interval for x* = 27.55 is: [%f, %f]',
    b0 + b1 * 27.55 - MOE,
    b0 + b1 * 27.55 + MOE)
```

```
## [1] "Our confidence interval for x* = 27.55 is: [7.419235, 7.816980]"
```

```r
SEuY <- sqrt((1 / n) + ((31.5 - mean(BMI))^2)/sum((BMI-mean(BMI))^2)*SS)
sprintf('Our confidence interval for x* = 31.5 is: [%f, %f]',
    b0 + b1 * 31.5 - MOE,
    b0 + b1 * 31.5 + MOE)
```

```
## [1] "Our confidence interval for x* = 31.5 is: [6.376458, 6.774203]"
```

# Problem 9

```r
b0 + (27.55 * b1) + e
```

```
## [1] 9.796104
```

```r
b0 + (31.5 * b1) + e
```

```
## [1] 8.753327
```

```r
SEuY <- sqrt(1 + (1 / n) + ((27.55 - (mean(BMI)))^2/sum((BMI-mean(BMI))^2))*SS)
MOE <- pt(1 - (cl / 2), n-2) * SEuY; MOE
```

```
## [1] 0.7271771
```

```r
sprintf('Our prediction interval for x* = 27.55 is: [%f, %f]',
    b0 + b1 * 27.55 - MOE,
    b0 + b1 * 27.55 + MOE)
```

```
## [1] "Our prediction interval for x* = 27.55 is: [6.890931, 8.345285]"
```

```
SEuY <- sqrt(1 + (1 / n) + ((31.5 - (mean(BMI)))^2/sum((BMI-mean(BMI))^2))*SS)
MOE <- pt(1 - (cl / 2), n-2) * SEuY; MOE
```

```
## [1] 0.7983737
```

```
sprintf('Our prediction interval for x* = 31.5 is: [%f, %f]',
    b0 + b1 * 31.5 - MOE,
    b0 + b1 * 31.5 + MOE)
```

```
## [1] "Our prediction interval for x* = 31.5 is: [5.776957, 7.373704]"
```

# Problem 10

```
nodel <- lm(PA ~ poly(BMI, 2), pabmi)
sn <- summary(nodel); sn
```

```
##
## Call:
## lm(formula = PA ~ poly(BMI, 2), data = pabmi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2330 -1.7188 -0.1299  1.2164  6.4088
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8.5991     0.2466  34.871  < 2e-16 ***
## poly(BMI, 2)1  -9.0539     2.1918  -4.131  9.2e-05 ***
## poly(BMI, 2)2  -0.4103     2.1918  -0.187    0.852
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.192 on 76 degrees of freedom
## Multiple R-squared:  0.1837, Adjusted R-squared:  0.1622
## F-statistic: 8.549 on 2 and 76 DF,  p-value: 0.0004477
```

```
a0 <- sn$coefficients[1, 1]
a1 <- sn$coefficients[2, 1]
a2 <- sn$coefficients[3, 1]
sprintf('Our new regression equation is: Y =  %f + %fX1 + %fX2',
    a0, a1, a2)
```

```
## [1] "Our new regression equation is: Y =  8.599063 + -9.053876X1 + -0.410261X2"
```

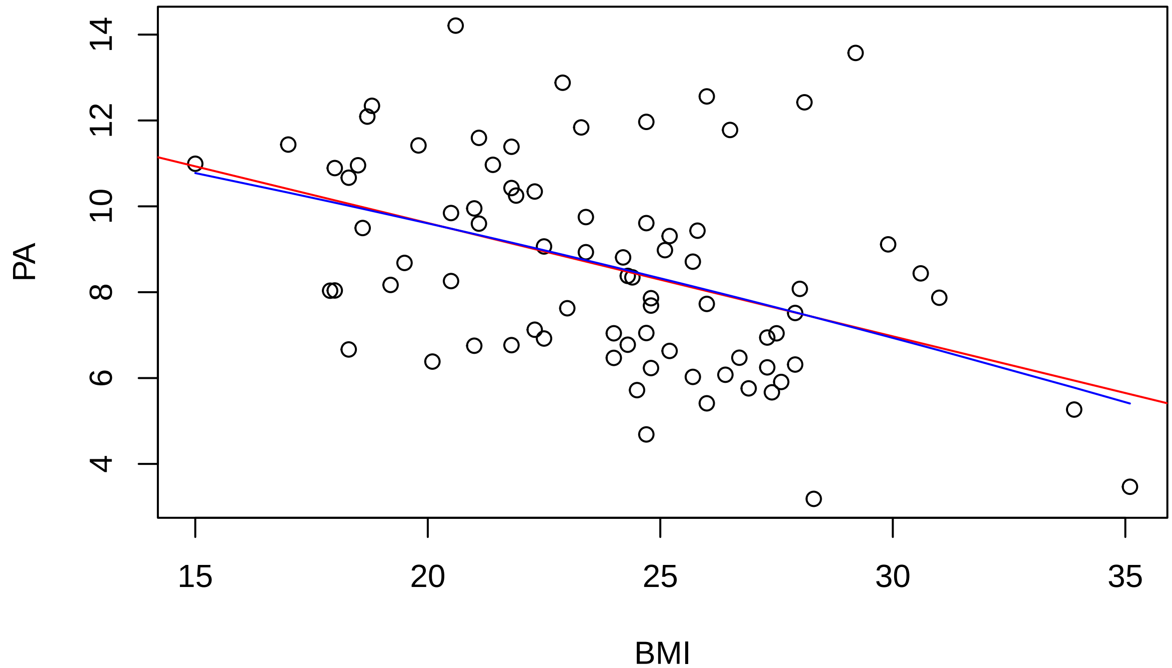
```
RRn <- sn$adj.r.squared
sprintf('We will check adjusted R^2s %f vs %f.',
    sm$adj.r.squared, RRn)
```

```
## [1] "We will check adjusted R^2s 0.172681 vs 0.162181."
```

The adjusted $R^2$ of the original model is higher, which makes the original model a better fit to the data.

## Problem 11

```r
plot(BMI, PA)
abline(b0, b1, col="red")
lines(sort(BMI), fitted(nodel)[order(BMI)], col='blue')
```




## Problem 12

$$Y = \tilde{X}\beta + \varepsilon \implies \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

## Problem 13

We know

$$(\tilde{X}'\tilde{X}) = \begin{bmatrix} 1 & 1 & \dots & 1 \\ (x_1 - \bar{x}) & (x_2 - \bar{x}) & \dots & (x_n - \bar{x}) \end{bmatrix} * \begin{bmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n}(x_i - \bar{x}) \\ \sum_{i=1}^{n}(x_i - \bar{x}) & \sum_{i=1}^{n}(x_i - \bar{x})^2 \end{bmatrix}$$

If we know that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, this becomes

$$\begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^{n}(x_i - \bar{x})^2 \end{bmatrix}$$

and if we know the property of the inverse of a diagonally dominant matrix, we can find $(\tilde{X}'\tilde{X})^{-1} =$

$$\begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^{n}(x_i-\bar{x})^2 \end{bmatrix}.$$

We can find

$$\tilde{X}'Y = \begin{bmatrix} 1 & 1 & \dots & 1 \\ (x_1-\bar{x}) & (x_2-\bar{x}) & \dots & (x_n-\bar{x}) \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i(x_i-\bar{x}) \end{bmatrix}.$$

Using these two equations, we can find $\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y =$

$$\begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^{n}(x_i-\bar{x})^2 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i(x_i-\bar{x}) \end{bmatrix} = \begin{bmatrix} \frac{\sum_{i=1}^{n} y_i}{n} \\ \frac{\sum_{i=1}^{n} y_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \end{bmatrix}$$

# Problem 14

## i

$\hat{\beta} \sim N(\beta, (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y\sigma^2)$.

## ii

$\mu_{\hat{\beta}} = \beta$; $\sigma^2_{\hat{\beta}} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y\sigma^2$.