MA 331 Final

Amane Chibana and Harry Wang

2024-05-06

Problem 1

```
pabmi = read_xls("./pabmi.xls")

with(pabmi,cor(PA,BMI))
```

```
## [1] -0.3854091
```

```
with(pabmi,cor.test(PA,BMI))
```

```
##
##  Pearson's product-moment correlation
##
## data:  PA and BMI
## t = -4.1348, df = 98, p-value = 7.503e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5408817 -0.2044696
## sample estimates:
##        cor
## -0.3854091
```

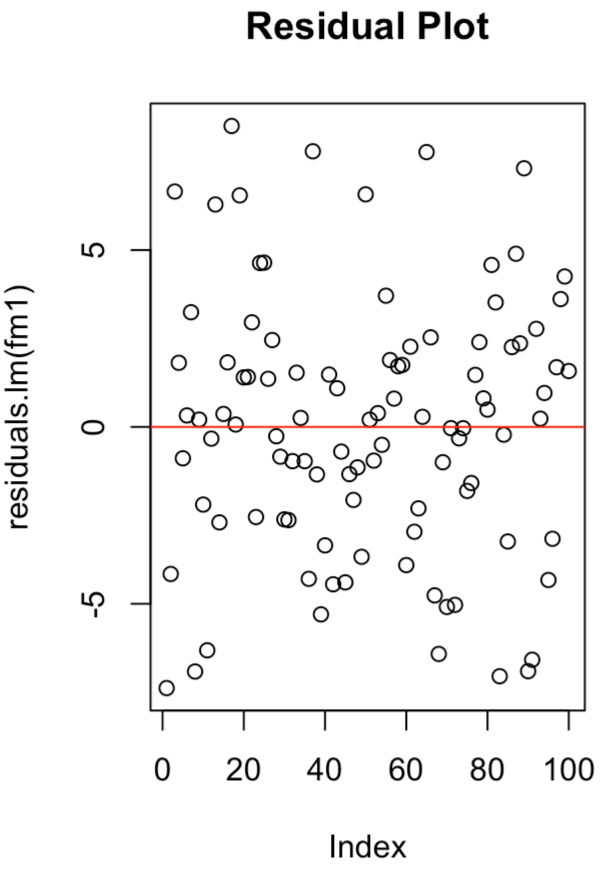
The p-value for testing the null hypothesis $H_0 : p(X, Y) = 0$ is 7.503e-05 , which is less than the significance level $\alpha = 0.05$. This result allows us to reject the null hypothesis, indicating that the correlation between PA and BMI is statistically significant and not zero.

Problem 2

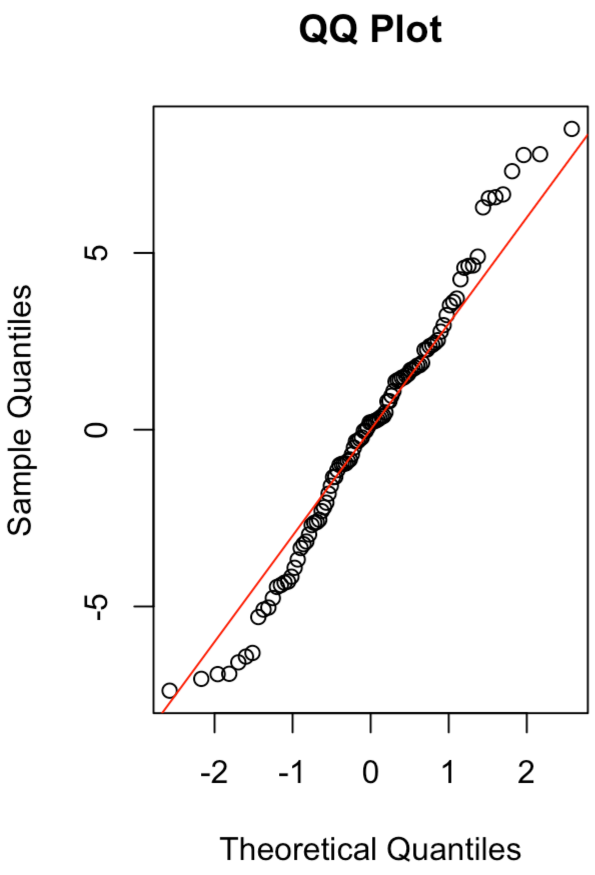
```
fm1 = lm(BMI~PA, data = pabmi)

par(mfrow=c(1,2))
plot(residuals.lm(fm1),main="Residual Plot")
abline(0,0,col="red")

par(mfrow=c(1,2))
```



```
qqnorm(residuals(fm1),main="QQ Plot")
abline(0,3, col="red")
```



From the residual plot, it appears that the residuals are randomly dispersed around the horizontal line at zero, which suggests that the linear model may be appropriate. The normal Q-Q plot shows that the residuals roughly follow a straight line, indicating that they are approximately normally distributed. This supports the assumption of normality in the linear regression model.Overall, the current regression model seems to be a good fit for the data based on these diagnostics.

Problem 3

```
kable(coefficients(summary(fm1)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	29.5782471	1.4119783	20.948089	0.0e+00
PA	-0.6546858	0.1583361	-4.134784	7.5e-05

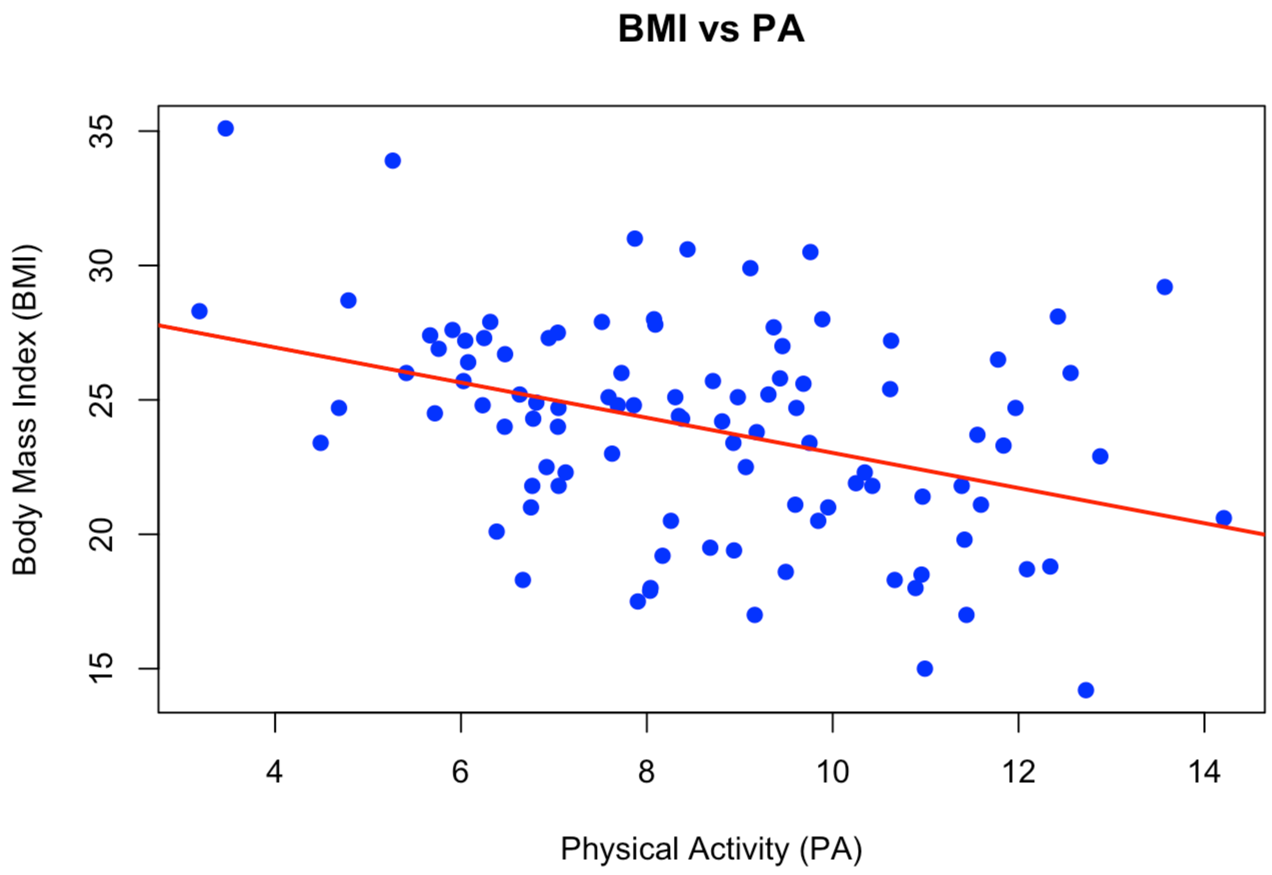
```
(sigma(fm1))^2
```

```
## [1] 13.35817
```

```
cat("Regression Equation: BMI = 29.578 - 0.655 x PA")
```

```
## Regression Equation: BMI = 29.578 - 0.655 x PA
```

```
plot(pabmi$PA, pabmi$BMI, main = "BMI vs PA", xlab = "Physical Activity (PA)", ylab = "Body Mass Index (BMI)", pc
h = 19, col = "blue")
abline(fm1, col = "red", lwd = 2)
```



Problem 4

```
kable(anova(fm1))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
PA	1	228.3772	228.37719	17.09644	7.5e-05
Residuals	98	1309.1007	13.35817	NA	NA

The hypothesis test for the slope parameter in the regression model of BMI as a function of PA yields a p-value of 7.5e-05, which is significantly less than the alpha level of 0.05. This indicates strong statistical evidence to reject the null hypothesis that the slope is zero. Therefore, we conclude that there is a significant negative relationship between PA (Physical Activity) and BMI (Body Mass Index), suggesting that increases in PA are associated with decreases in BMI.

Problem 5

```
kable(confint(fm1))
```

	2.5 %	97.5 %
(Intercept)	26.7762222	32.3802721
PA	-0.9688987	-0.3404729

The 95% confidence interval is from (-0.97,-0.34)

Problem 6

```
kable(anova(fm1))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
PA	1	228.3772	228.37719	17.09644	7.5e-05
Residuals	98	1309.1007	13.35817	NA	NA

T he Pr(>F) (p-value) is less than 0.05, you can reject the null hypothesis that the model with no predictors fits the data as well as your model. This means that the overall regression model is statistically significant, and the predictor (PA) provides a better fit to the data than the intercept-only model.

Problem 7

```
summary(fm1)
```

```
##
## Call:
## lm(formula = BMI ~ PA, data = pabmi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3819 -2.5636  0.2062  1.9820  8.5078
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.5782     1.4120  20.948  < 2e-16 ***
## PA           -0.6547     0.1583  -4.135  7.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.655 on 98 degrees of freedom
## Multiple R-squared:  0.1485, Adjusted R-squared:  0.1399
## F-statistic:  17.1 on 1 and 98 DF,  p-value: 7.503e-05
```

```
SSM = sum((predict(fm1) - mean(pabmi$BMI))^2) ; SSM
```

```
## [1] 228.3772
```

```
SST = sum((pabmi$BMI - mean(pabmi$BMI))^2); SST
```

```
## [1] 1537.478
```

The coefficient of determination is 0.15. SSM is 228.38 and SST is 1537.48

Problem 8

```
observation = data.frame(PA = 27.85)

predict(fm1,newdata=observation,interval="confidence",level=0.95)
```

```
##        fit      lwr      upr
## 1 11.34525 5.257584 17.43291
```

The predicted BMI for a PA of 27.85 is approximately 11.34525, with a 95% confidence interval ranging from about 5.257584 to 17.43291. This interval indicates where the true mean response is expected to fall with 95% confidence, assuming the model is correct and the assumptions hold.

Problem 9

```
observation = data.frame(PA = 31.25)

predict(fm1,newdata=observation,interval="prediction",level=0.9)
```

```
##        fit      lwr      upr
## 1 9.119317 0.597295 17.64134
```

The predicted BMI for a PA of 31.25 is approximately 9.119317, with a 90% prediction interval ranging from about 0.597295 to 17.64134. This interval indicates where the actual observed BMI is expected to fall with 90% confidence, assuming the model is correct and the assumptions hold.]

Problem 10

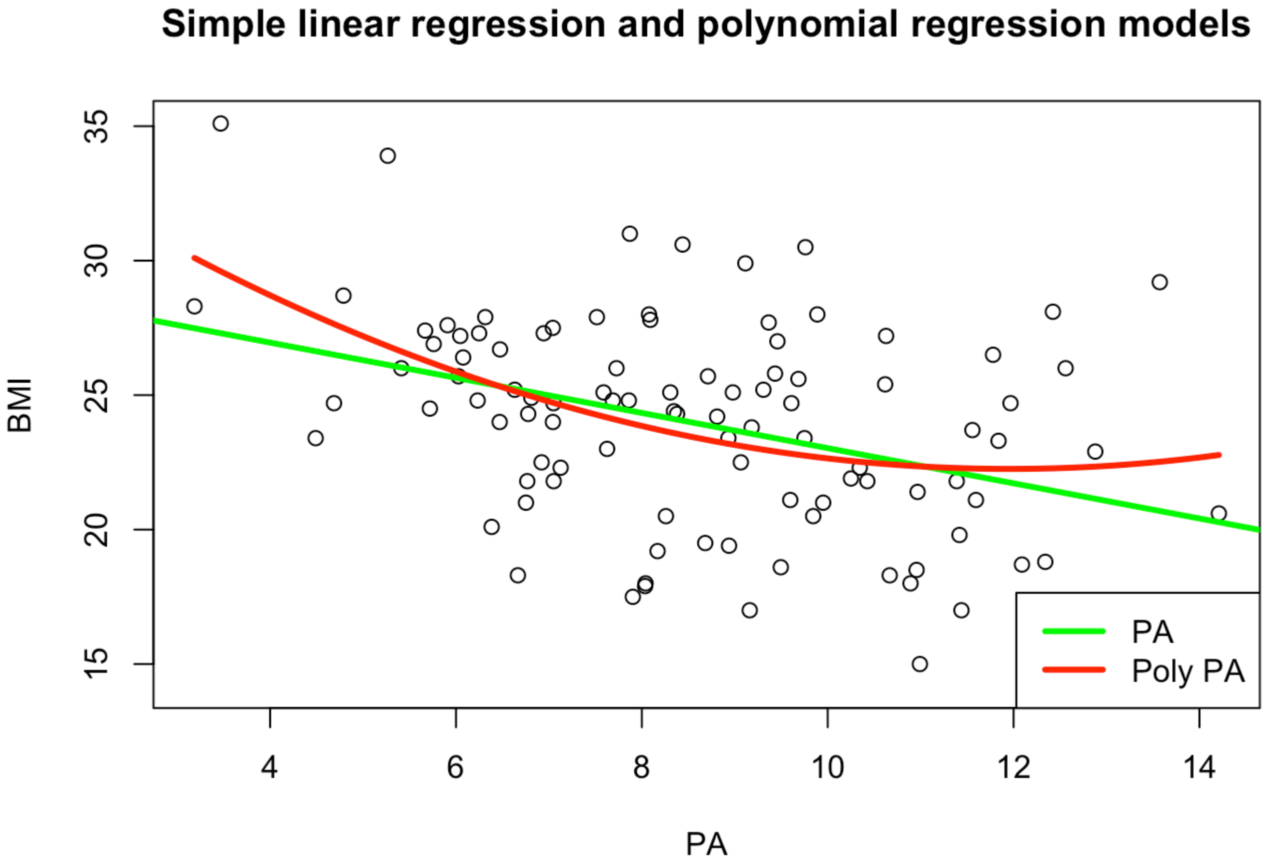
```
fm2= lm(BMI ~ poly(PA,2), data=pabmi)
summary(fm2)
```

```
##
## Call:
## lm(formula = BMI ~ poly(PA, 2), data = pabmi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1159 -2.3779  0.1315  2.2638  7.7518
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.9390     0.3612  66.285  < 2e-16 ***
## poly(PA, 2)1 -15.1122     3.6115  -4.184 6.28e-05 ***
## poly(PA, 2)2   6.6269     3.6115   1.835   0.0696 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.612 on 97 degrees of freedom
## Multiple R-squared:  0.1771, Adjusted R-squared:  0.1601
## F-statistic: 10.44 on 2 and 97 DF,  p-value: 7.838e-05
```

The regression polynomial equation is Y = 23.94 – 15.11x + 6.63x. If the adjusted for the polynomial model is higher than that for the linear model, the polynomial model is better. This indicates that including the squared term of PA provides a better fit to the data, capturing more of the variability in BMI. The reason a model with a higher adjusted R^2 is considered better is that it explains a greater proportion of the variance in the dependent variable, after adjusting for the number of predictors in the model, thus potentially capturing more complex relationships between the variables.

Problem 11

```
plot(BMI ~ PA, data=pabmi, main="Simple linear regression and polynomial regression models")
abline(fm1, col="green",lwd=3)
pavals <- seq(min(pabmi$PA), max(pabmi$PA), length.out = 100)
lines(pavals, predict(fm2, newdata = data.frame(PA = pavals)),col="red",lwd=3)
legend("bottomright",c("PA", "Poly PA"), col=c("green","red"),lwd=3)
```



Problem 12

$$Y = X\beta + \epsilon$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & (x_1 - \bar{x}) \\ 1 & (x_2 - \bar{x}) \\ \vdots & \vdots \\ 1 & (x_n - \bar{x}) \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Problem 13

$$X = \begin{bmatrix} 1 & (x_1 - \bar{x}) \\ 1 & (x_2 - \bar{x}) \\ \vdots & \vdots \\ 1 & (x_n - \bar{x}) \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$X' = \begin{bmatrix} 1 & 1 & \dots & 1 \\ (x_1 - \bar{x}) & (x_2 - \bar{x}) & \dots & (x_n - \bar{x}) \end{bmatrix}$$

$$X'X = \begin{bmatrix} n & \sum(x_i - \bar{x}) \\ \sum(x_i - \bar{x}) & \sum(x_i - \bar{x})^2 \end{bmatrix} = \begin{bmatrix} n & 0 \\ 0 & \sum(x_i - \bar{x})^2 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum(x_i - \bar{x})^2} \end{bmatrix}$$

$$X'Y = \begin{bmatrix} \sum y_i \\ \sum((x_i - \bar{x})y_i) \end{bmatrix}$$

$$\hat{\beta} = (X'X)^{-1} X'Y = \begin{bmatrix} \frac{\sum y_i}{n} \\ \frac{\sum((x_i - \bar{x})y_i)}{\sum(x_i - \bar{x})^2} \end{bmatrix}$$