

HW17 – Q2

Nick Huo

2022-10-26

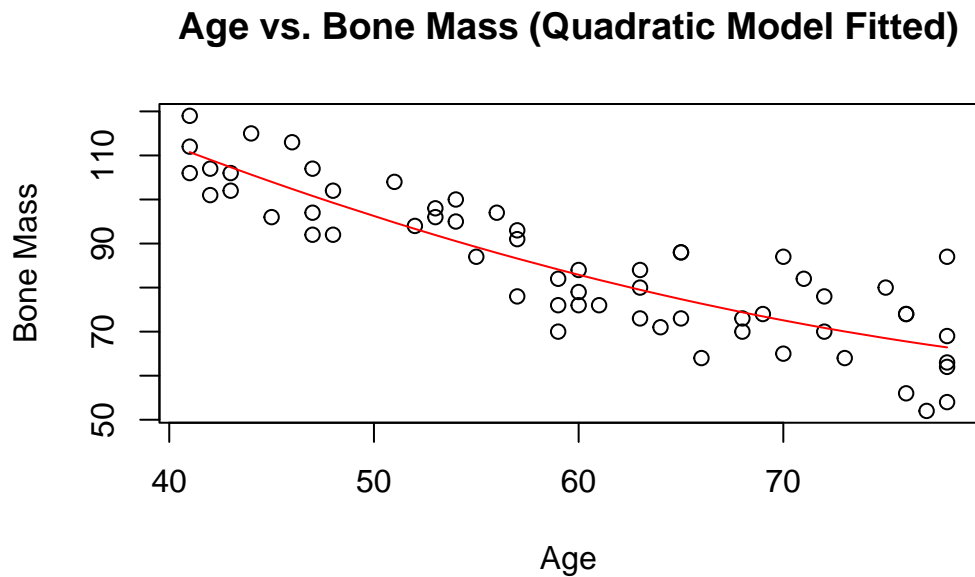
Q2

Set-up

```
ms <- read.csv("MuscleMass.csv")
ms$X_c <- ms$X - mean(ms$X)
attach(ms)
```

a.

```
fit <- lm(Y ~ X_c + I(X_c^2))
plot(X, Y, main="Age vs. Bone Mass (Quadratic Model Fitted)",
     xlab="Age", ylab="Bone Mass")
lines(sort(X), fitted(fit)[order(X)], col='red')
```



```
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X_c + I(X_c^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.086  -6.154  -1.088   6.220  20.578
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 82.935749   1.543146  53.745  <2e-16 ***
## X_c         -1.183958   0.088633 -13.358  <2e-16 ***
## I(X_c^2)     0.014840   0.008357   1.776   0.0811 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.026 on 57 degrees of freedom
## Multiple R-squared:  0.7632, Adjusted R-squared:  0.7549
## F-statistic: 91.84 on 2 and 57 DF,  p-value: < 2.2e-16
```

From the scatter plot and the regression fit line, it appears that the model is fitting the data quite well. The model has $R^2 = 0.763$ and $R^2_{adj} = 0.755$.

b.

H_0 : None of the predictors are significant.

H_a : At least one of the predictors are significant.

According to the regression model output, we see that the global F-test is significant as the p-value < 0.0001 is smaller than $\alpha = 0.05$ – there is strong evidence suggesting at least one of the predictors is significant. We are able to reject H_0 . Therefore, we conclude that at least one of the predictors is significant in predicting bone mass.

c.

```
newdata <- data.frame(X_c=48-mean(X))
predict(fit, newdata, level=0.95, interval="confidence")
```

```
##           fit        lwr        upr
## 1 99.25461 96.28436 102.2249
```

The estimated mean muscle mass for women aged 48 years old is 99.25. We are 95% confident that the mean muscle mass for women aged 48 years old is between 96.28 and 102.22.

d.

```
predict(fit, newdata, level=0.95, interval="prediction")
```

```
##           fit      lwr      upr
## 1 99.25461 82.9116 115.5976
```

The estimated muscle mass for a woman aged 48 years old is 99.25. We are 95% confident that muscle mass for a woman aged 48 years old is between 82.91 and 115.60.

e.

H_0 : The quadratic term is not significant.

H_a : The quadratic term is significant.

The p-value of 0.081 suggests that there is some evidence suggesting the quadratic term is significant as it is larger than but close to $\alpha = 0.05$. We failed to reject H_0 . Thus, we conclude that the quadratic term is not significant and we can drop the quadratic term from the model. So the squared difference of age vs. the mean age is not significant in explaining bone mass.

f.

```
mean(X)
```

```
## [1] 59.98333
```

We know that $x = X - \bar{X} = X - 59.98$.

Original Model: $\hat{Y} = 82.96 - 1.18(x) + 0.015(x)^2$.

$$\hat{Y} = 82.96 - 1.18(X - 59.98) + 0.015(X - 59.98)^2$$

Re-expressed Model: $\hat{Y} = 82.96 - 1.18X + 1.18 * 59.98 + 0.015(X^2 - 119.96X + 3598)$

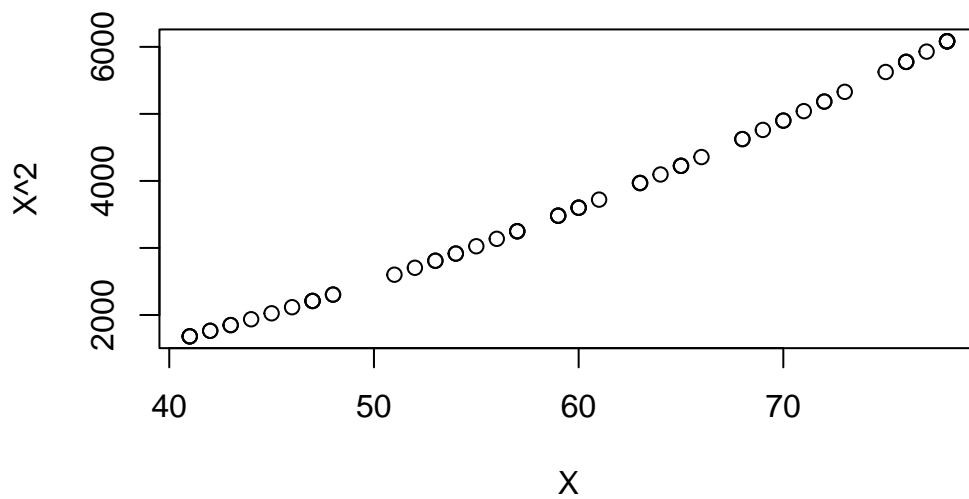
$$\hat{Y} = 207.71 - 0.619X + 0.015X^2$$

g.

```
cor(X, X^2)
```

```
## [1] 0.9960939
```

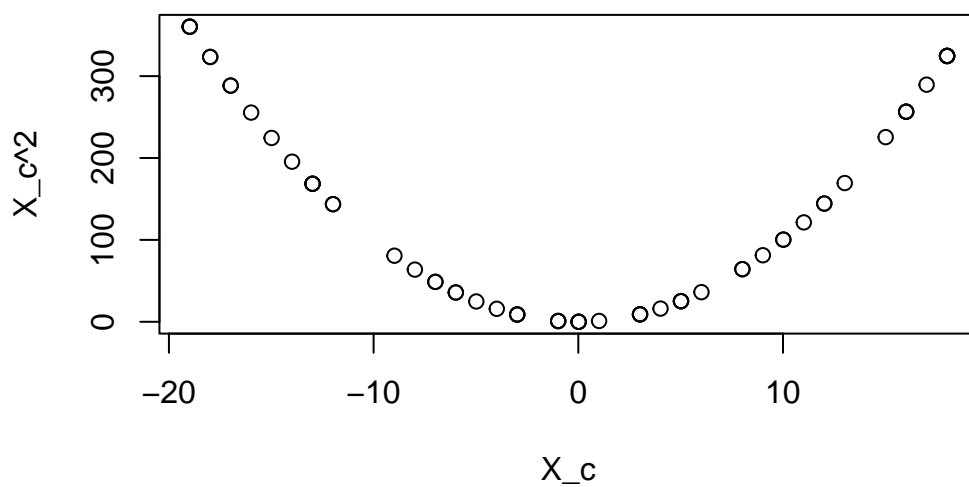
```
plot(X, X^2)
```



```
cor(X_c, X_c^2)
```

```
## [1] -0.03835694
```

```
plot(X_c, X_c^2)
```

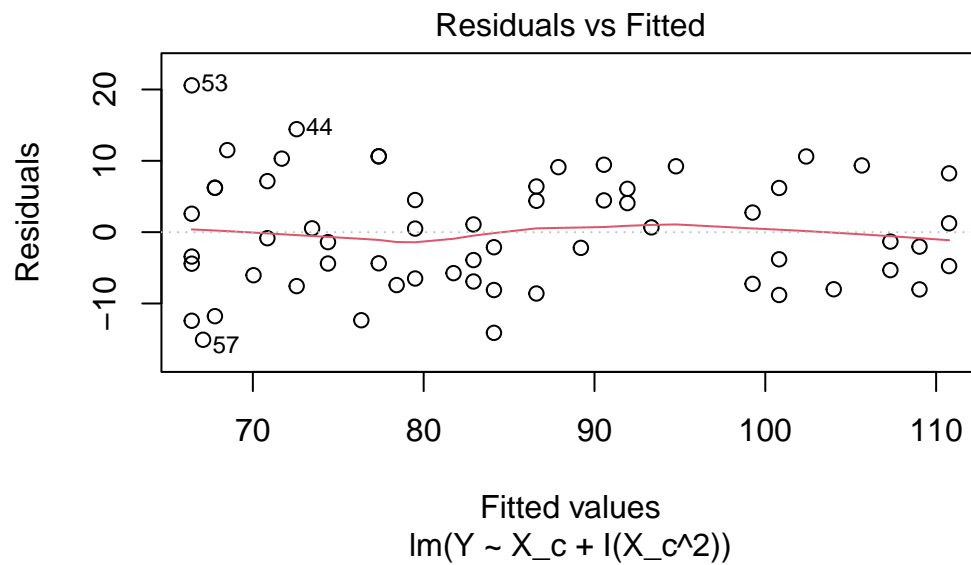


The correlation coefficient between X and X^2 is 0.996, which is very high. The correlation coefficient between x and x^2 is -0.038, very close to 0. This suggests that it is useful to use a centered variable since it reduces the correlation between X and X^2 . It wouldn't be helpful to include both X and X^2 as they are highly

correlated. But if we use x and x^2 , because they are not correlated, they could explain additional variability in the response.

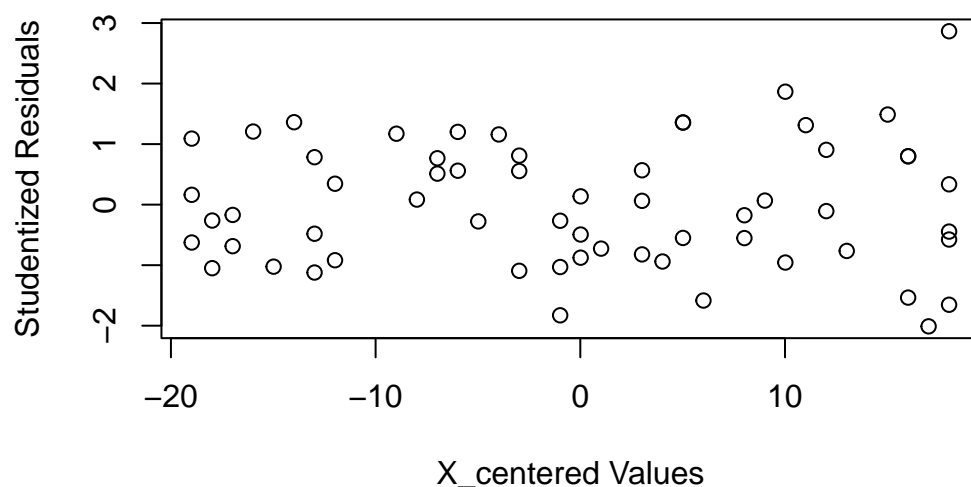
h.

```
library(MASS)
plot(fit, which=1)
```

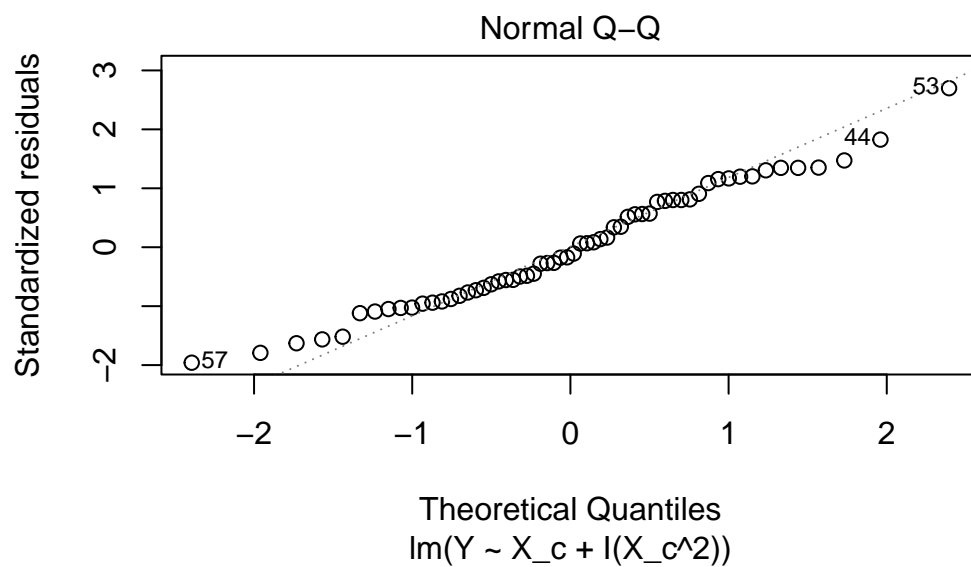


```
plot(X_c, studres(fit), main="Studentized Residuals vs. X_centered",
     xlab="X_centered Values", ylab="Studentized Residuals")
```

Studentized Residuals vs. X_centered



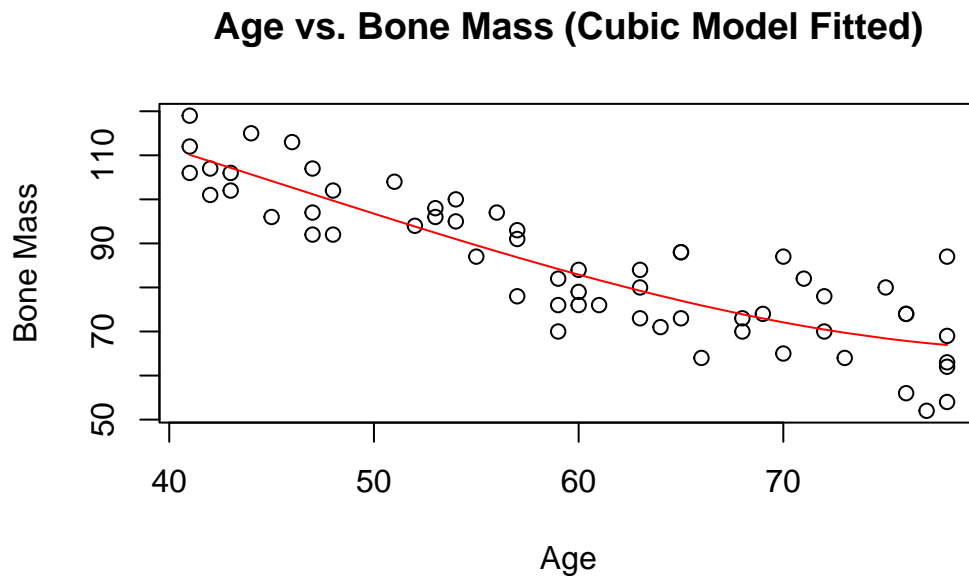
```
plot(fit, which=2)
```



From the studentized residuals plots, we notice that the variance is not changing and there is no clear pattern. These suggest that constant variance and linearity conditions are met. And looking at the QQ plot, we see that it mostly falls on the line, except at both tails, so it is approximately normal with a little bit of left-skewness.

i.

```
fit <- lm(Y ~ X_c + I(X_c^2) + I(X_c^3))
plot(X, Y, main="Age vs. Bone Mass (Cubic Model Fitted)",
     xlab="Age", ylab="Bone Mass")
lines(sort(X), fitted(fit)[order(X)], col='red')
```



```
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X_c + I(X_c^2) + I(X_c^3))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3671  -5.8483  -0.6755   6.1376  20.0637
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  82.9273444   1.5552264   53.322  < 2e-16 ***
## X_c          -1.2678894   0.2489231   -5.093  4.28e-06 ***
## I(X_c^2)       0.0150390   0.0084390    1.782   0.0802 .
## I(X_c^3)       0.0003369   0.0009327    0.361   0.7193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.087 on 56 degrees of freedom
## Multiple R-squared:  0.7637, Adjusted R-squared:  0.7511
## F-statistic: 60.34 on 3 and 56 DF, p-value: < 2.2e-16
```

H_0 : None of the predictors are significant.

H_a : At least one of the predictors are significant.

According to the regression model output, we see that the global F-test is significant as the p-value < 0.0001 is smaller than $\alpha = 0.05$ – there is strong evidence suggesting at least one of the predictors is significant. We are able to reject H_0 . Therefore, we conclude that at least one of the predictors is significant in predicting bone mass.

Now if we look at the cubic term alone. We can test for its significance with

H_0 : The cubic term is not significant.

H_a : The cubic term is significant.

The p-value of 0.72 suggests that there is little evidence suggesting the cubic term is significant as it is larger than $\alpha = 0.05$. We failed to reject H_0 . Thus, we conclude that the cubic term is not significant and we can drop the cubic term from the model. So the cubed difference of age vs. the mean age is not significant in explaining bone mass.