Homework-6-Q1

BIOS507 Spring 2025 | Dr Lukemire | Elizabeth Nemeti Due: April 7 2025

Problem 1.

A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each 10-year age group, beginning with age 40 and ending with age 79. The dataset is "musclemass.txt". Note: These questions are adapted from questions 8.4 and 8.5 in the textbook.

- **Y (response)**: Mass
- **X1 (predictor)**: Age

```
data_path = "/Users/elizabethnemeti/Documents/GitHub/BIOS507-Coursework/Homeworks/Homework-6/"
data_file <- file.path(data_path, "musclemass.txt")
MUSCLEMASS_data <- read.table(data_file, header = TRUE)

str(MUSCLEMASS_data)
```

```
## 'data.frame':    60 obs. of  2 variables:
##  $ Mass: int  106 106 97 113 96 119 92 112 92 102 ...
##  $ Age : int  43 41 47 46 45 41 47 41 48 48 ...
```

```
head(MUSCLEMASS_data)
```

```
##   Mass Age
## 1  106  43
## 2  106  41
## 3   97  47
## 4  113  46
## 5   96  45
## 6  119  41
```

Part A. Fit a quadratic regression model with centered age. Plot the fitted regression function and the data. Does the quadratic regression function visually appear to be a good fit here? Report the R2 value.

```
MUSCLEMASS_data <- MUSCLEMASS_data %>%
  mutate(age_centered = Age - mean(Age), # center age to reduce multicollinearity between Age and Age^2
         age_centered_sq = age_centered^2) # ^2 for the quadratic part

head(MUSCLEMASS_data)
```

```
##   Mass Age age_centered age_centered_sq
## 1  106  43    -16.98333        288.4336
## 2  106  41    -18.98333        360.3669
## 3   97  47    -12.98333        168.5669
## 4  113  46    -13.98333        195.5336
## 5   96  45    -14.98333        224.5003
## 6  119  41    -18.98333        360.3669
```

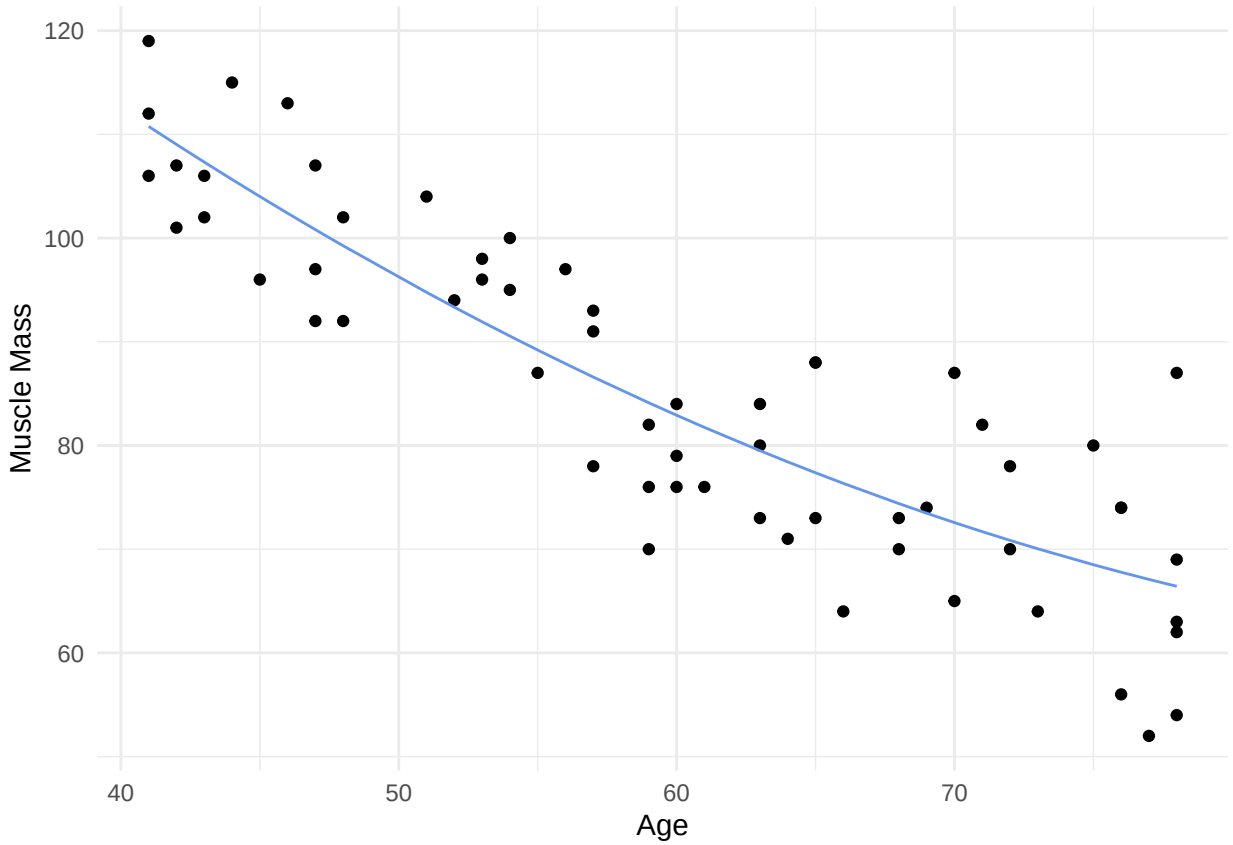
```r
# quadratic regression model: E[Mass] = ß0 + ß1(age_centered) + ß2 (age_centered_sq)

MUSCLEMASS_model <- lm(Mass ~ age_centered + age_centered_sq, data = MUSCLEMASS_data)
summary(MUSCLEMASS_model)
```

```
##
## Call:
## lm(formula = Mass ~ age_centered + age_centered_sq, data = MUSCLEMASS_data)
##
## 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 ***
## age_centered    -1.183958   0.088633 -13.358   <2e-16 ***
## age_centered_sq  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
```

```r
MUSCLEMASS_data <- MUSCLEMASS_data %>%
  mutate(mass_prediction = predict(MUSCLEMASS_model)) # get predicted vals

ggplot(MUSCLEMASS_data, aes(x = Age, y = Mass)) +
  geom_point(color = "black") +
  geom_line(aes(y = mass_prediction), color = "cornflowerblue") + # using Age for x axis not centered a
  labs(x = "Age",
       y = "Muscle Mass") +
  theme_minimal()
```



Yes, the quadratic regression does appear to be a good fit, there is a curved trend downwards in muscle mass as Age increases. R^2 is 0.7632, therefore the current model explains 76.32% of the variability in muscle mass.

Part B. Conduct an overall test for model fit using = 0.05. What are your findings?

```
summary(MUSCLEMASS_model)
```

```
##
## Call:
## lm(formula = Mass ~ age_centered + age_centered_sq, data = MUSCLEMASS_data)
##
## 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 ***
## age_centered    -1.183958   0.088633 -13.358   <2e-16 ***
## age_centered_sq  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
```

Here we're interested in whether age is important at all to the model.

H0: ß1 = ß2 = 0 (aka model explains doesn't explain muscle mass) HA: at least ß1 or ß2 != 0 (aka model can explain some variation in muscle mass)

We do an F-test, and our F-statistic for the model is 91.84 with 2 df, and a pvalue of 2.2e-16, where $p < 0.05$. Since the p value is less than 0.05, we reject the null hypothesis. Age is therefore significantly associated with muscle, and the quadratic model provides a better fit than an intercept only model.

Part C. Test whether the quadratic term can be dropped from the model using = 0.05. What are your findings?

```
MUSCLEMASS_linear <- lm(Mass ~ age_centered, data = MUSCLEMASS_data) # here's the linear model
# MUSCLEMASS_model <- lm(Mass ~ age_centered + age_centered_sq, data = MUSCLEMASS_data) # this was the

anova(MUSCLEMASS_linear, MUSCLEMASS_model)
```

```
## Analysis of Variance Table
##
## Model 1: Mass ~ age_centered
## Model 2: Mass ~ age_centered + age_centered_sq
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     58 3874.4
## 2     57 3671.3  1    203.13 3.1538 0.08109 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, we're using a nested model comparison to test whether adding the quadratic term (age_centered_sq) improves the model fit beyond a simple linear model (just age_centered term). Basically, do we need a quadratic or does a straight line suffice for good model fit. We use anova() to run an F-test and compare the reduced model and full model. The F-statistic is 3.15, with a p-value of 0.081, where $p > 0.05$. Since the p value is greater than 0.05, we fail to reject the null hypothesis. Therefore, a simple linear model suffices to explain the relationship between muscle mass and age, and there is not enough improvement in adding this term to warrant the added complexity.

Part D. Find and interpret a 95% confidence interval for the mean muscle mass for women age 50.

Since we found the quadratic model was not a significant improvement from the simple linear model, we'll move forward with the simple linear model.

```
age_mean <- mean(MUSCLEMASS_data$Age)
age_centered_50 <- 50 - age_mean # making sure 50 is centered
pt_50 <- data.frame(age_centered = age_centered_50)

predict(MUSCLEMASS_linear, newdata = pt_50, interval = "confidence", level = 0.95)
```

```
##        fit     lwr      upr
## 1 96.84679 94.0701 99.62348
```

Here we are interested in the mean muscle mass for women at the age of 50, not just one individual. The predicted mean muscle mass for a woman who is age 50 is 96.85 units, with a 95% confidence interval of (94.07, 99.62).

Part E. Find and interpret a 95% prediction interval for the muscle mass for a woman who is age 50.

Here we are interested in the muscle mass for an individual woman at age 50, not the mean. Interval therefore, will now be set to "prediction" instead of "confidence".

```
predict(MUSCLEMASS_linear, newdata = pt_50, interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 96.84679 80.25244 113.4411
```

The predicted muscle mass for a woman who is age 50 is 96.845 units, with a 95% prediction interval of (80.25, 113.44). The prediction interval is much wider than the confidence interval, this is due to greater uncertainty when predicting for an individual rather than the mean.

Part F. Fit the third-order model and test for the significance of the cubic term (using = 0.05)

```
MUSCLEMASS_data <- MUSCLEMASS_data %>%
  mutate(age_centered_cubic = age_centered^3) #adding the cubic term, just like earlier we added the sq

MUSCLEMASS_model_cubic <- lm(Mass ~ age_centered + age_centered_sq + age_centered_cubic, data = MUSCLEMA

summary(MUSCLEMASS_model_cubic)
```

```
##
## Call:
## lm(formula = Mass ~ age_centered + age_centered_sq + age_centered_cubic,
##     data = MUSCLEMASS_data)
##
## 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 ***
## age_centered       -1.2678894  0.2489231  -5.093 4.28e-06 ***
## age_centered_sq     0.0150390  0.0084390   1.782   0.0802 .
## age_centered_cubic  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
```

H0: ß3 = 0 (the cubic term does not improve model fit) HA: ß3 != 0 (the cubic term adds value)

Here, we're interested in testing whether whether the cubic term significantly improves model fit. R^2 is 0.7637, therefore this cubic model explains about 76.37% of the variance in muscle mass. This is a slight increase from the quadratic model R^2 at 76.32. The F-statistic is reported as 60.34 at 56 df, and the p value for the cubic term is 0.7193. As where $0.72 > 0.05$, we fail to reject the null hypothesis, and conclude that the cubic term does not significantly improve model fit.