HW2

Shay Lebovitz

1/22/2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(broom)
```

2.27

a

```
muscle <- read_delim('data/CH01PR27.txt', delim = " ", col_names = F)
```

```
## Parsed with column specification:
## cols(
##   X1 = col_character(),
##   X2 = col_character()
## )
```

```
muscle <- lapply(muscle, as.numeric)
muscle <- as_tibble(muscle)
muscle <- muscle %>%
  transmute(X = X2, Y = X1)

muscle_fit <- lm(Y ~ X, data = muscle)
summary(muscle_fit)
```

```
##
## Call:
## lm(formula = Y ~ X, data = muscle)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1368  -6.1968  -0.5969   6.7607  23.4731
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.3466     5.5123   28.36   <2e-16 ***
## X            -1.1900     0.0902  -13.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.173 on 58 degrees of freedom
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.7458
## F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

$H_0 : B_1 < 0$ vs. $H_1 : B_1 >= 0$
$B_1 = -1.1900z$, $s_{B_1} = 0.0902$
$(-1.1900 - 0)/0.0902 = -13.19$

$t(.95, 58) = 1.672 < |-13.19|$
Therefore, we reject the null hypothesis that there is not a negative relationship between muscle mass and age in favor of the alternative. The p-value of this test is $8.24 * 10^{-19}$, which can be written as 0^+.

b

Just because there is a statistically significant p-value for the test that $B_0 = 0$ does not mean that B_0 has any significant meaning. In this case, it is certainly not meaningful, because the data only has middle-aged - elderly woman, who show declines in muscle mass. However, we all know that babies have practically no muscle mass and develop muscle until their 20's - 30's. Thus, we know that linear relationship does not extend much past the data. In general, one cannot make meaninful predictions far outside the range of the data.

c

A 95% confidence interval of the difference in expected muscle mass for women differing by one year is essentially a 95% for Beta1.
$-1.1900 + -0.0902 * 2.0017 = (-1.3706, -1.0094)$
There is no need to know the specific ages because the regression is linear, meaning the slope at one X value is the same as the slope at any other X value.

2.28

a

```r
alpha <- 0.05
age60 <- data.frame(X = 60)
predict(muscle_fit, age60, interval = 'confidence', level = 1-alpha,
        se.fit = TRUE)
```

```
## $fit
##        fit      lwr      upr
## 1 84.94683 82.83471 87.05895
##
## $se.fit
## [1] 1.055154
##
```

```
## $df
## [1] 58
##
## $residual.scale
## [1] 8.173177
```

There is a 95% probability that the mean muscle mass of of women aged 60 is between 82.83 and 87.06.

b

```
predict(muscle_fit, age60, interval = 'prediction', level = 1-alpha)
```

```
##        fit      lwr     upr
## 1 84.94683 68.45067 101.443
```

There is a 95% probability that a 60 year old woman's muscle mass is between 68.45 and 101.44. This is a fairly wide range and doesn't give us much information, so it is not precise.

c

```
CI <- predict(muscle_fit, age60, interval = "confidence",
              level = 1 - alpha, se.fit = TRUE)
Yh.hat <- CI$fit[1]
SE.Yh.hat <- CI$se.fit
W <- sqrt(2 * qf(1 - alpha, 2, 58))
LowerBound <- Yh.hat - W * SE.Yh.hat
UpperBound <- Yh.hat + W * SE.Yh.hat
Band <-c(LowerBound, UpperBound)
Band
```

```
## [1] 82.29593 87.59774
```

The confidence interval is very slightly wider in this case than in part (a). This is as it should be, because the confidence band takes into consideration the entire regression line, whereas in (a) only a single point is considered. Thus, there is more room for variation than in part (a) and hence a wider confidence interval.

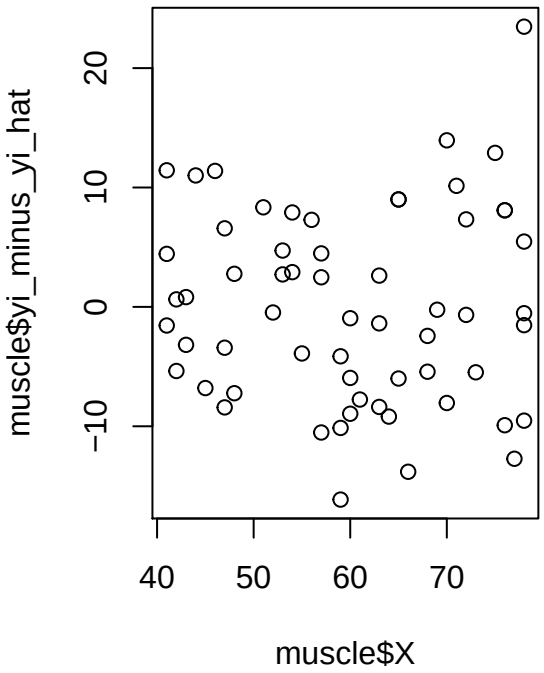
2.29

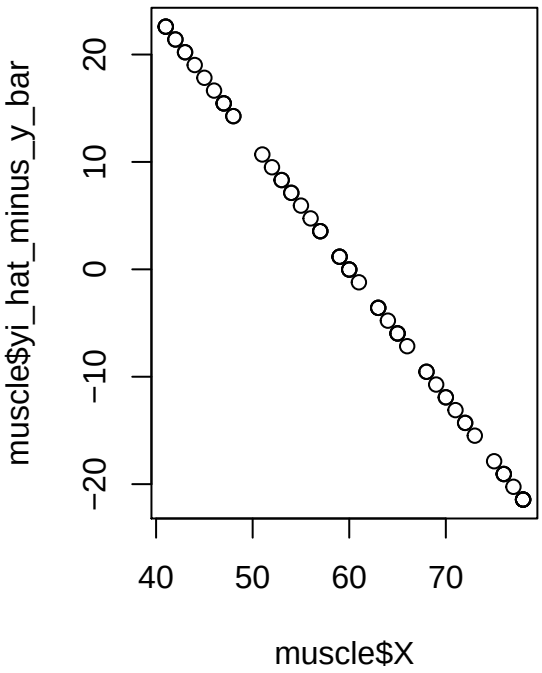
a

```
muscle <- muscle %>%
  mutate(yi_hat = 156.35 - 1.19*X) %>%
  mutate(yi_minus_yi_hat = Y - yi_hat) %>%
  mutate(y_bar = mean(Y)) %>%
  mutate(yi_hat_minus_y_bar = yi_hat - y_bar)

par(mfrow = c(1,2))

plot(muscle$X, muscle$yi_minus_yi_hat)
plot(muscle$X, muscle$yi_hat_minus_y_bar)
```





I think it is pretty hard to tell from these graphs whether SSE or SSR is the larger proportion of SST. They look to be fairly equal. This tells me that R^2 will be moderate, likely in the range of 0.6-0.8.

b

```
anova(muscle_fit)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X          1 11627.5 11627.5  174.06 < 2.2e-16 ***
## Residuals 58  3874.4    66.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c

$H_0 : B_1 = 0$ vs. $H_a : B_1 \neq 0$
$F^* = MSR/MSE = 11627.5/66.8 = 174.06$
$F^* > F(0.95, 1, 58)$, $174.06 = 4.01$
Therefore we can reject the null hypothesis that $B_1 = 0$ in favor of the alternative. There is statistical evidence at the 95% level that B_1 is not 0.

d

The variation that remains unexplained by age is the error sum of squares, or SSE. The proportion of total variation is $SSE/SST = 25\%$. This is relatively small, meaning that the regression accoutned for most of the variation. Given the variability of data in the real world, this number is as expected.

e

$R^2 = SSR/SST = 11627.5/(11627.5 + 3874.4) = 0.7501$
$r = -\sqrt{(R^2)} = -0.8661$