Midterm 1

Ishita Dutta

5/3/2021

1.(25 points)

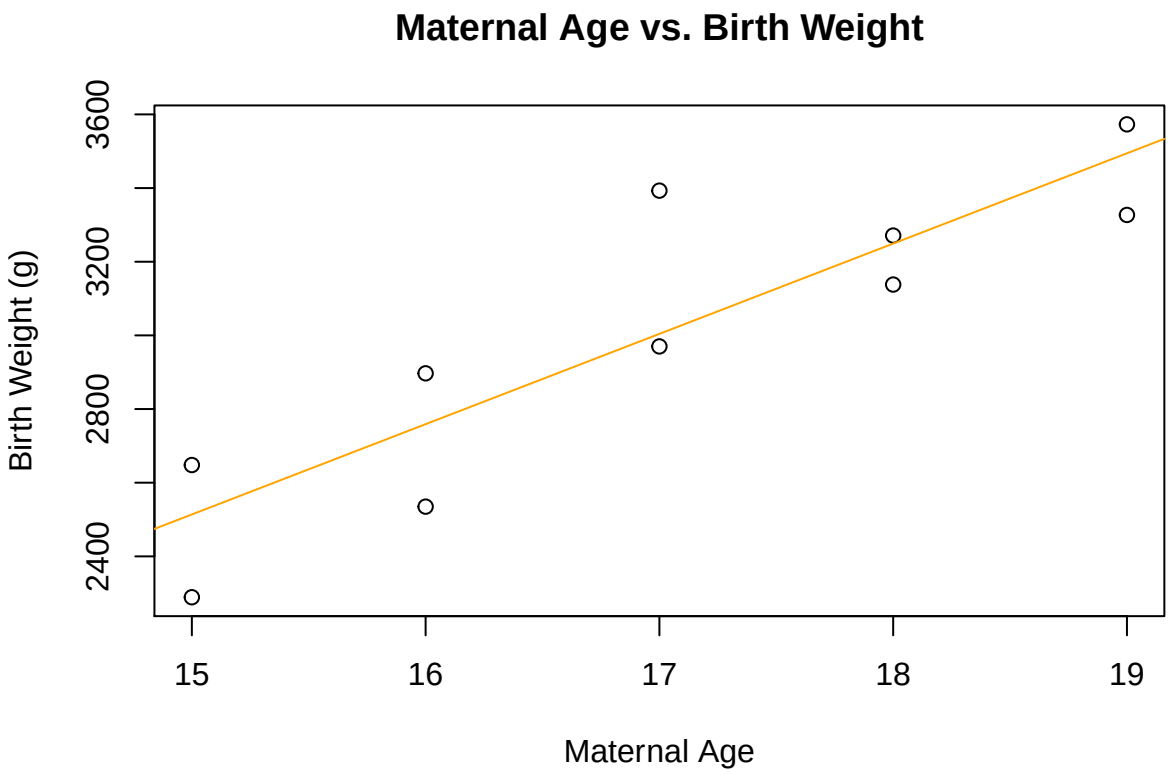
In a study on the risk of teen mothers having low birth weight babies published in 1998 in the Journal of School Health, the authors reported the following data set, where x represents maternal age (in years) and Y represents birth weight of baby (in grams).

```
X1 = c(15,17,18,15,16,19,17,16,18,19)
Y1 = c(2289, 3393, 3271, 2648, 2897, 3327, 2970, 2535, 3138, 3573)
n1 = length(X1)
```

1. a)

A researcher considers a simple linear regression with x as the predictor and Y as the response. Note that, in the data, each x value corresponds to two different Y values. Why are there such differences? Explain briefly using the simple linear regression model.

```
fit1 = lm(Y1~X1)
plot(X1, Y1, xlab="Maternal Age", ylab="Birth Weight (g)" , main = "Maternal Age vs. Birth Weight")
abline(fit1, col = "orange")
```



Solution: The reason there is a difference in the weight from the regression line is due to the error not being factored by our regression line. The orange line in the figure represents the regression line only taking into account B0 and B1, not E1, or the residual error from the observed values. The points on the graph are our obeserved values, which take error from the regression line into account, thus causing the difference.

1. b)

Under the simple linear regression model, the researcher considers the mean birth weights of two babies. The first baby's mother's maternal age is 17; the second baby's mother's maternal age is 16. Express the difference of E(x = 17) - E(x = 16) in terms of regression coefficients.

Solution: The difference is the change in expected values from when the mother is 17 as opposed to 16. We used our regression line here to figure out what these values are since we are looking at the expected values from our data as opposed to the actual data. The coefficient B1 will represent how much we can expect an increase in the weight of the baby every year the mother is older.

1. c)

Find an estimate of the difference in (1) of part b. Show your steps on how you find theanswer to receive full credit.

```
b1hat1 = t(X1-mean(X1))%*%(Y1-mean(Y1))/sum((X1-mean(X1))^2)
b0hat1 = mean(Y1) - b1hat1*mean(X1)
```

```r
E17 = b0hat1 + (b1hat1 * 17)
cat("E17: ", E17)
```

```
## E17:  3004.1
```

```r
E16 = b0hat1 + (b1hat1 * 16)
cat("E16: ", E16)
```

```
## E16:  2758.95
```

```r
cat(E17, " - ", E16, " :", E17 - E16)
```

```
## 3004.1  -  2758.95  : 245.15
```

1. d)

Find the standard error (s.e.) of the estimate in part c. Again, you need to show your calculation in order to receive full credit.

```r
y_hat1 = fit1$fitted.values
MSE1 = (sum((Y1-y_hat1)^2))/(n1-2)
MSE1
```

```
## [1] 42151.56
```

```r
SSTO1 = sum((Y1-mean(Y1))^2)
SSTO1
```

```
## [1] 1539183
```

1. e)

Find a 95% confidence interval for the difference in (1). Again, you need to show your calculation in order to receive full credit.

```r
confint(lm(Y1~X1), parm = "Interval: ",  level = 0.95)
```

```
##            2.5 % 97.5 %
## Interval:     NA     NA
```

2. (25 points)

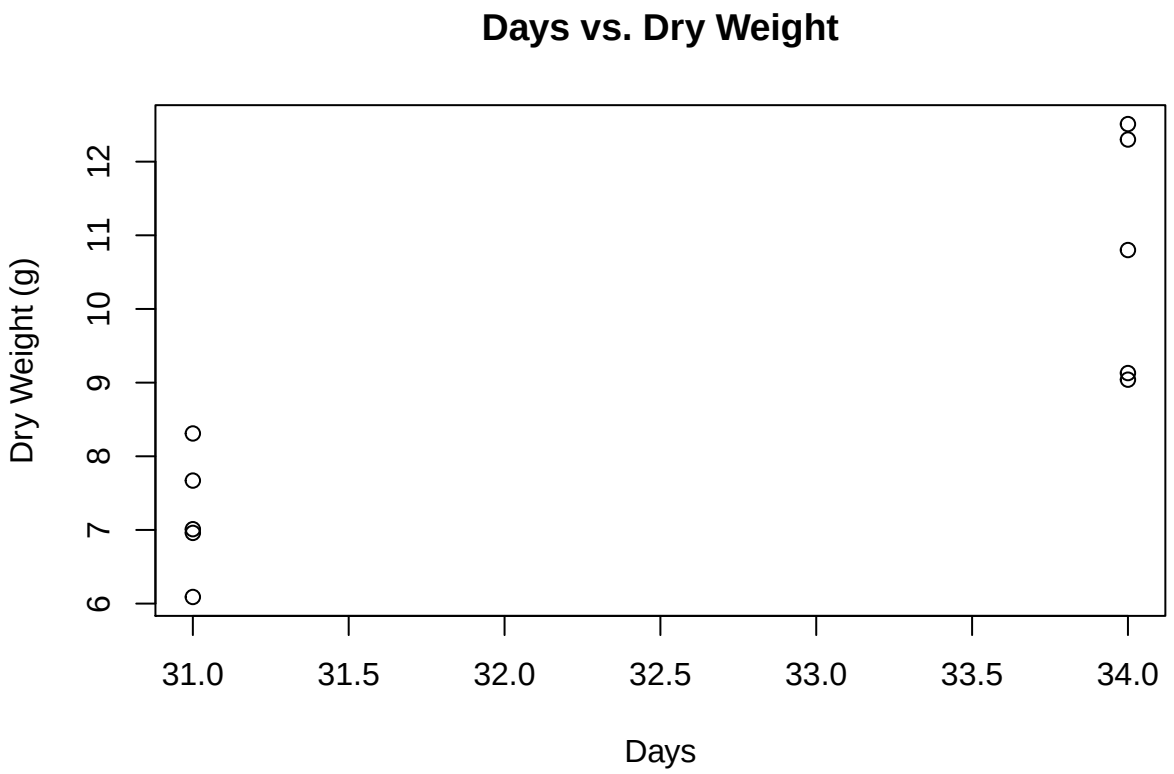
A botanist placed 10 one-week-old soybean seedlings in individual pots. After31 days of growth, she harvested, dried, and weighed 5 of the soybean plants (randomly selected); 3 days later, she harvested, dried, and weighed the rest of the soybean plants. So,in the end, she had data presented in the following table.

```r
X2 = c(31, 31, 31, 31, 31, 34, 34, 34, 34, 34)
Y2 = c(8.31, 7.67, 6.09, 7.01, 6.96, 9.13, 12.51, 12.30, 10.80, 9.04)
```

2. a)

Make a scatter plot of Y=Dry weight against x=Days of growth. From the plot, does it look like that the constant variance assumption in the simple linear regression model is satisfied?

```r
plot(X2, Y2, xlab="Days", ylab="Dry Weight (g)" , main = "Days vs. Dry Weight")
```



Solution: It seems that the constant variance assumption is satisfied, as we can see a relatively equal distribution of points at the x values we tested(31, and 34), which are in line with the simple regression model, showing the uncertainty that is E.

2. b)

In order to test the hypothesis (sd1)^2 = (sd2)^2 v (sd1)^2 != (sd2)^2, where (sd1)^2 is the variance of the regression error corresponding to xi= 31, and (sd2)^2is the variance of the regression error corresponding to xi= 34, the Brown - Forsythe (B-F) test is considered. The residuals are divided to two parts with the first part corresponding to xi = 31 and the second part to xi = 34. Compute the tBF statistic. Show your steps in order to receive full credit.

```r
#1.
fit2 = lm(Y2~X2)
e2 = fit2$residuals
#2.
group2_1 <- X2 == 31
group2_2 <- !group2_1
#3.
d2_1 <- abs(e2[group2_1] - median(e2[group2_1]))
d2_2 <- abs(e2[group2_2] - median(e2[group2_2]))
#4.
n2 <- length(e2)
n2_1 <- length(d2_1)
n2_2 <- length(d2_2)
#5.
s2 <- sqrt((sum((d2_1-mean(d2_1))^2)+sum((d2_2-mean(d2_2))^2))/(n2 - 2))
tstar <- (mean(d2_1)-mean(d2_2))/s2/sqrt(1/n2_1+1/n2_2)
cat("Statistic: ", tstar)
```

```
## Statistic:  -1.773669
```

2. c)

Carry out the B-F test using alpha = 0.05 as the level of significance. Is your conclusion consistent with your observation in part a?

```r
tcrit <- qt((1 - (0.05/2)), df = n2-2)
cat("Critical Value: ", tcrit)
```

```
## Critical Value:  2.306004
```

```r
cat("Result: ", abs(tstar) <= tcrit)
```

```
## Result:  TRUE
```

Solution: Because |tstar| < tcrit, we conclude the null(Ho), where the error variance is constant as opposed to the alternate(Ha) of the error variance not being constant. This is consistent with my answer in part a.

2. d)

Regardless of your test result in part c, 'propose a transformation that would stabilize the variance, that is, after the transformation, the two parts (see part b) of the transformed Y (corresponding to x= 31 and x= 34, respectively) spread about equally. Explain briefly why you propose this transformation.

Solution: No transformation, as the variance is a constant, not a function. Performing a transformation means that y* = y, and therefore there is no difference.

2. e)

Make a plot of the transformed data to show the effect of the transformation proposed in part d(again, the plot may be made by a computer, or draw by hand).

```
plot(X2, Y2, xlab="Days", ylab="Dry Weight (g)" , main = "Days vs. Dry Weight")
abline(fit2, col = "blue")
```