STAT 313/513 Assignment 2

Does weight mean fat? p. 45

1. Using R, create a scatterplot of Fat versus Weight. Note that fat is measured as a percentage and weight is in kg.

```r
> plot(Fat ~ Weight, data = fat, pch = 16, xlab = 'Weight (kg)',
+     ylab = 'Fat (%)', col = 'gray')
```

![Scatterplot of Fat versus Weight](scatterplot.png)

2. Use R to regress Fat on Weight and display the R code and output. Then, using the output write the equation of the line that predicts the response variable Fat from the predictor variable Weight.

```r
> m1 = lm(Fat ~ Weight, data = fat)
> summary(m1)

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 26.88558 | 4.67036 | 5.757 | 2.33e-05 *** |
| Weight | 0.02069 | 0.06414 | 0.323 | 0.751 |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.574 on 17 degrees of freedom
Multiple R-squared: 0.006081, Adjusted R-squared: -0.05238
F-statistic: 0.104 on 1 and 17 DF, p-value: 0.751

\[ \hat{y} = 26.886 + 0.021x \]

\[ \text{Fat} = 26.886 + 0.021 \times \text{Weight} \]

3. Write the “word equation” of the model in part 2.
Fat = Weight

4. If there was no linear relationship between Fat and Weight, how likely is it (i.e., what is the probability) that you would get an estimated regression slope as different from zero as the one that you obtained in part 2? Cite and present any relevant R output used.

The referenced probability is the p-value for the slope coefficient, $p = 0.751$.

5. How strong is the evidence for a significant linear relationship between Fat and Weight? Cite an appropriate test statistic and p-value.

There is no significant evidence that the slope is significantly different from zero ($t(17) = 0.323, p = 0.751$). Equivalently, we could cite the F-statistic, $F(1,17) = 0.104$ with the same p-value.

6. What proportion of the variability in the Fat data is explained by the linear relationship between Fat and Weight? Cite and present any relevant R output used.

$$R^2 = 0.006081 = 0.608\%$$

Only 0.608% of the variability in Fat % is explained by the regression relationship between at and Weight.
1. Use R to graphically illustrate how Flowers and DBH are related. Note that DBH is measured in mm.

![Graph showing the relationship between number of flowers and DBH](image)

2. Use R to regress Flowers on DBH and display the R code and output.

```r
> m1 = lm(Flowers ~ DBH, data = trees)
> summary(m1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -481.1604    86.2409  -5.579 1.09e-06 ***
DBH            4.5128     0.4224  10.683 2.78e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 210.6 on 48 degrees of freedom 
Multiple R-squared: 0.7039, Adjusted R-squared: 0.6978 
F-statistic: 114.1 on 1 and 48 DF,  p-value: 2.784e-14
```

3. Assuming you still have your plot from part 1 still open (if not, make the plot again), type

```r
abhline(a = __ , b = __ )
```

where you fill in the blanks with the estimated y-intercept and slope. What does it do? Display the result. Is a straight line a good fit to this data? Where does the line tend to underestimate (and/or overestimate) the number of flowers?
> plot(Flowers ~ DBH, data = trees, pch = 16, col = 'gray',
+     xlab = 'DBH (mm)', ylab = 'Number of Flowers')
>
> abline(m1)# or abline(-481.1604, 4.5128)

The straight line fit is not very good in this situation. The line underestimates the number of flowers when DBH is small (under 100 mm) or large (around 300 mm). For middle values of DBH (around 150 to 250 mm) the line overestimates the number of flowers.

4. Refer to the model output of part 2. Write the equation of the line that predicts Flowers from DBH.

\[ \hat{y} = -481.160 + 4.513x \]

5. Test the null hypothesis that \( \beta = 4 \). That is, is there significant evidence that \( \beta \neq 4 \)?

\[ t = \frac{4.5128 - 4}{0.4224} = 1.214 \]

with \( df = 48 \).

\[ P = Pr\{t > 1.214 \text{ or } t < -1.214\} = 0.2307 \] using the computer.

Using R:

> t.stat = (summary(m1)$coef[2,1] - 4)/summary(m1)$coef[2,2]
> t.stat
[1] 1.214048
> 2*(1 - pt(t.stat, df = 48))
[1] 0.2306688

There is no significant evidence that \( \beta \neq 4 \) (\( t(48) = 1.214, p = 0.231 \)).
6. Consider the R code used to make your plot in part 1. Add the following options to your code: `col = Sex, pch = 16` and then repeat the `abline` command from part 3. What does this new plot reveal?

We can see from the new plot that the line systematically under/overestimates the number of flowers for one of the sexes. (It underestimates the number of flowers for the males.)

7. Provide a point and 95% interval estimate for the number of flowers on a tree with DBH = 280 mm. Explain why you chose the type of interval that you did.

```r
> new.data = data.frame(DBH = 280)
> predict(ml, newdata=new.data, interval = 'prediction')
  fit  lwr  upr
1 782.4352 348.2946 1216.576
```

We estimate that a tree with DBH of 280 mm will have about 782 flowers. We further estimate that 95% of all trees with DBH =280 mm will have between around 348 to 1217 flowers. (NOT PART OF THE SOLUTION TO THE PROBLEM: We are 95% confident that the average number of flowers for all trees with DBH = 280 is between 686 and 878 flowers).

We know we are dealing with a prediction interval here because we are only trying to estimate the number of flowers for one particular tree with a DBH of 280 mm.

8. Is it appropriate to estimate the number of flowers on a tree with DBH = 390 mm? Explain.

No, this would be an extrapolation. The largest trees studied had a DBH of around 325 mm, so we shouldn’t make inference for trees larger than this.