Homework #8

Due: Tuesday, November 23 @ 6pm

Please remember to give R code, as well as answers, for any problems where you used R

Problem 1:

Consider the fungus growth data from HW7:

Laetisaric acid concentration (uG/mL)	0	0	3	3	6	6	10	10	20	20	30	30
Fungus growth (mm)	33.3	31	29.8	27.8	28	29	25.5	23.8	18.3	15.5	11.7	10

a. What is the mean and standard deviation of **both** fungus growth and acid concentration?

```
fungus <- data.frame(concentration = c(0,0,3,3,6,6,10,10,20,20,30,30),
                     growth = c(33.3,31,29.8,27.8,28,29,25.5,23.8,18.3,15.5,11.7,10))

# mean and sd of fungus growth (y)
mean_y <- mean(fungus$growth)
mean_y
```

```
## [1] 23.64167
```

```
sd_y <- sd(fungus$growth)
sd_y
```

```
## [1] 7.847114
```

```
# mean and sd of acid concentration (X)
mean_x <- mean(fungus$concentration)
mean_x
```

```
## [1] 11.5
```

```
sd_x <- sd(fungus$concentration)
sd_x
```

```
## [1] 10.88368
```

b. Using the values from (a) and the correlation coefficient calculated in HW7, calculate the estimates of β_1 and β_0 for the linear model: Growth $= \beta_0 + \beta_1$concentration. *Show your work*

```
# first correlation value r:
r <- cor(fungus$concentration, fungus$growth)
r
```

```
## [1] -0.9875349
```

```r
# b_1 = r(s_y / s_x)
b_1 <- r*(sd_y/sd_x)
b_1
```

```
## [1] -0.7120107
```

```r
# b_0 = mean_y - b_1*mean_x
b_0 <- mean_y - b_1*mean_x
b_0
```

```
## [1] 31.82979
```

c. Explain in words what the estimates β_1 and β_0 mean in this dataset.

\textcolor{red}{B_0 is the intercept. In this dataset, it means that at a concentration of 0 uG/mL we expect fungus growth of 31.82 mm. B_1 is the slope. In this dataset, it means that for every increase of 1 uG/mL in acid, we see a decrease of 0.712 mm of fungus growth.}

d. Now, using R's `lm()` function, calculate the estimates of β_1 and β_0 for the linear model: Growth = β_0 + $\beta_1$concentration. **How do your answers compare to (b)?**

```r
model <- lm(growth ~ concentration, data = fungus)
model
```

```
##
## Call:
## lm(formula = growth ~ concentration, data = fungus)
##
## Coefficients:
##   (Intercept)  concentration
##        31.830         -0.712
```

The values for the intercept (31.83) and slope (-0.712) were the same that I calculated by hand!

e. Using the equation for your regression line calculated in (b), how much growth (in mm) would you predict the fungus would have when exposed to 15 uG/mL? *Show your work*

```r
# growth <- b_0 + b_1*concentration
b_0 + b_1*15
```

```
## [1] 21.14963
```

f. Now, using R's `predict()` function and the linear model from (d), repeat (e). Is the answer the same?

```r
predict(model, data.frame(concentration = 15))
```

```
##        1
## 21.14963
```

Yes! the answers are the same.

g. Given the linear model generated for fungus growth and acid concentration, I estimate that at 35 uG/mL of laetisaric acid there will be approximately 6.9 mm of fungal growth. Do you agree with this statement? Why or why not?

```
predict(model, data.frame(concentration = 35))
```

```
##        1
## 6.909414
```

The prediction is correct given the linear model I generated, however a value of 35 uG/mL is outside of the range studied in this experiment, therefore we cannot assume it still has the same linear trend. Therefore I do not agree with this statement, we cannot predict a value given the data we have.

Problem 2:

Let's continue with the fungus data from **Problem 1**

a. Calculate the `SS(resid)` or residual sum of squares for the linear model you derived in **Problem 1**. There are many ways to do this, but the simplest is probably to use `deviance(model)` where `model` is the linear model output from `lm()`.

```
deviance(model)
```

```
## [1] 16.7812
```

b. Describe in words what the `SS(resid)` is measuring. (*Note: "residual sum of squares" is not an appropriate answer*)

The SS(resid) is the sum of all the squared residuals. What this means, is it is measuring the distance of each point to the regression line (residual), squaring it, and then adding them all up. This is an estimate of the variance of the data around the regression line. It is used to determine fit.

c. Calculate the coefficient of determination, r^2. Explain in words what this value means.

```
r^2
```

```
## [1] 0.9752252
```

The coefficient of determintation can be estimated by simply squaring r, the correlation coefficient. An r^2 of 0.975 means that 97.5% of the total variance in fungal growth can be explained by this linear model. In otherwords, can be explained by knowing the acid concentration. This is extremely high and indicates that our model fits the data very well!

d. Consider the null hypothesis that laetisaric acid has no effect on growth of the fungus. Assuming that the linear model is applicable, state in **symbols** the null hypothesis about the true regression line and an alternative hypothesis that laetisaric acid inhibits growth of the fungus.

$H_0 : \beta_1 = 0$

$H_A : \beta_1 < 0$

NOTE: H_A is directional because we are hypothesizing that the acid inhibits growth of fungus, not that it changes the growth of the fungus (good or bad). And because we are interested in inhibiting growth, our slope must be negative, not positive.

e. How many degrees of freedom are there for the test in (d)?

For regression, degrees of freedom are n-2. Since n = 12, df = 10

f. Calculate the test statistic for the test in (d) using `summary(model)` where `model` is the linear model output from `lm()`. **Be sure to write out the test statistic, just showing the model output is not good enough**

```
summary(model)
```

```
##
## Call:
## lm(formula = growth ~ concentration, data = fungus)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0896 -0.8498  0.2743  0.9004  1.4702
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   31.82979    0.55693   57.15 6.53e-14 ***
## concentration -0.71201    0.03589  -19.84 2.32e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.295 on 10 degrees of freedom
## Multiple R-squared:  0.9752, Adjusted R-squared:  0.9727
## F-statistic: 393.6 on 1 and 10 DF,  p-value: 2.321e-09
```

The test statistic for this test is -19.84.

g. Using the `*t()` functions or the output from (f), state your conclusion regarding the null hypothesis in this context (*Note: remember to consider your alternative hypothesis). Be sure to provide the p-value, alpha, whether you reject or not, and what your final conclusion is.

The p-value for our test is 1.16e-09. NOTE: the hypothesis tested in the model summary output is non-directional but our hypothesis is directional. So we need to divide the given p-value by 2. We can test this with the pt() function:

```
# because t is already negative, we want to look at lower tail - this is one directional
pt(-19.84,10)
```

```
## [1] 1.160672e-09
```

```
# two directional should give same value as model output
pt(-19.84,10)*2
```

```
## [1] 2.321343e-09
```

Problem 3:

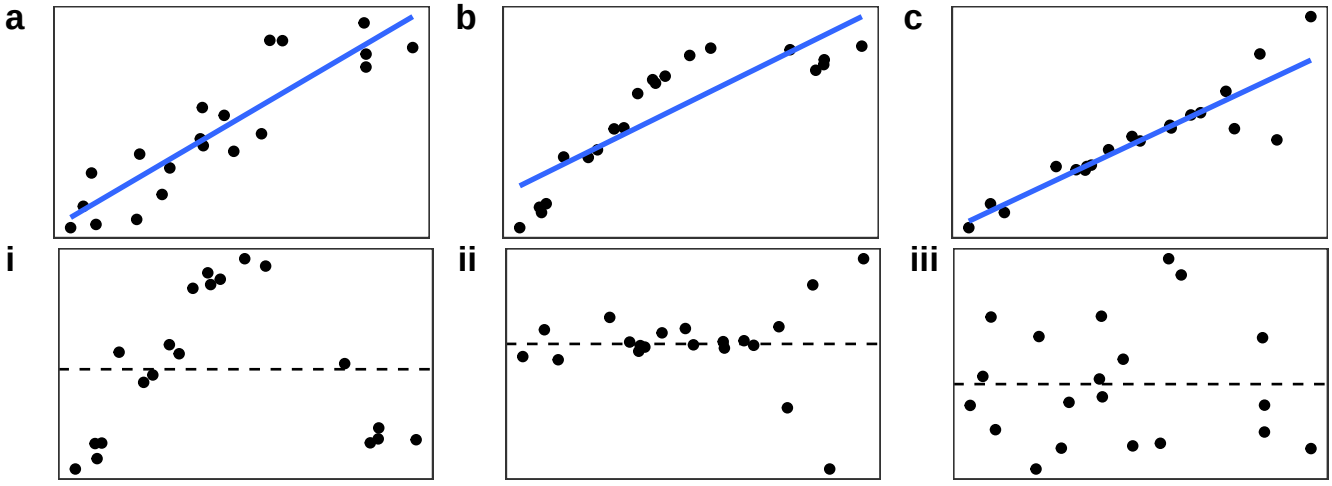
a. Describe the conditions/assumptions for using a linear model

1. Linearity - there is a linear relationship between the X and Y variables.

2. Independence - each observed pair must be independent of the others. (and randomly subsampled from a large population)

3. Normality - residuals are normally distributed

b. The following three residual plots (i), (ii), and (iii) were generated after fitting regression lines to the following three scatterplots (a), (b), and (c). Which residual plots goes with which scatterplot? **How do you know?**



a -> iii; b -> i; c -> ii

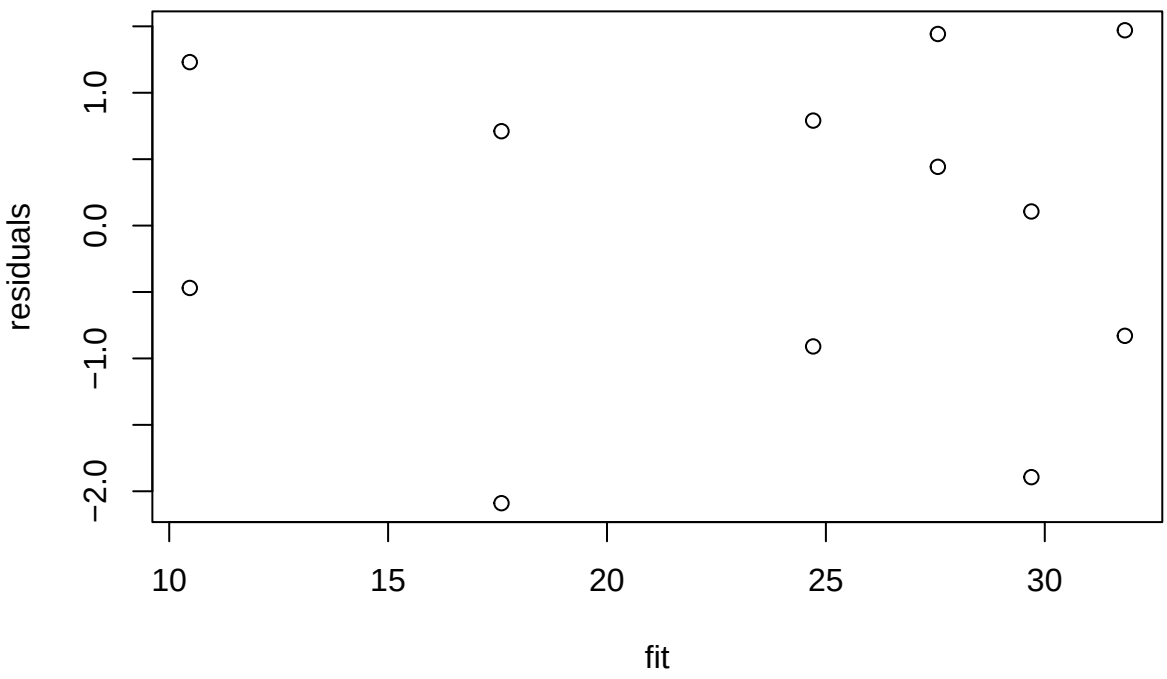
Residual plots show the distance between each point and the regression line. In plot a, all the points are somewhat equidistant from the regression line, therefore the residual plot has no clear shape with residuals centered around 0. The plot b shows a slight curve which the residuals exaggerate. Plot c shows increase in variance as x increases so the residuals will also get further from 0 as x increases.

c. Plot the residuals from the fungus data in **Problem 1**. Which condition/assumption are we checking with a residual plot? Is it met for the fungus data?

```
# get residuals with resid(model)
residuals <- resid(model)

# plot residuals against fitted values
fit <- fitted(model)

plot(fit, residuals)
```



The main assumption we are testing with the residual plot is looking for evidence of non-linearity. The residual plot for the fungus data shows no clear patterm and residuals center around 0, so this assumption is met. Which makes sense because the data looks quite linear.

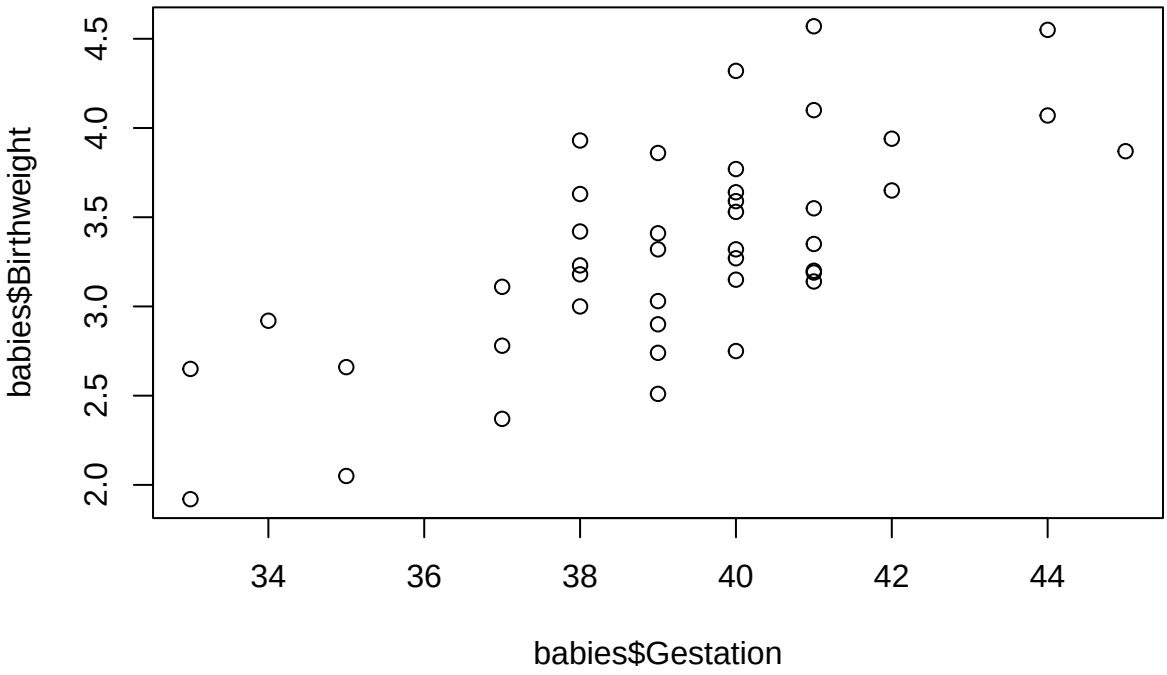
Problem 4:

The following dataset contains information about newborn babies and their parents. There is a lot of information here, but suppose we are interested in determining if the number of weeks at which a baby is born (i.e. `Gestation` (weeks)) can predict baby weight (i.e. `Birthweight` (kg)).

```
url <- "https://raw.githubusercontent.com/katiesevans/IGP_biostatistics/main/data/birthweight.csv"
babies <- read.csv(url)
```

a. Read in the data with the code above and plot gestation versus birth weight. Describe what you see. Does there appear to be a linear relationship between the variables? What is the correlation coefficient?

```
plot(babies$Gestation, babies$Birthweight)
```



There appears to be a positive linear correlation between gestation and birthweight. As gestation increases, birthweight also increases.

```
cor(babies$Gestation, babies$Birthweight)
```

```
## [1] 0.7083029
```

A strong correlation value of 0.708 provides further evidence of this strong positive linear relationship.

b. Assuming all assumptions are met, use R to generate a linear model of birth weight given gestational age. Write a few sentences detailing your results **as you would in a manuscript**. Make sure to include:

- What the question was
- What test was performed
- The equation of the linear model
- The variance explained
- Written explanation of the regression coefficient
- Confidence interval for the regression coefficient
- p-value of the test

```
summary(lm(Birthweight ~ Gestation, data = babies))
```

```
##
## Call:
## lm(formula = Birthweight ~ Gestation, data = babies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77203 -0.35999  0.00206  0.27387  0.96433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.0289     1.0015  -3.024  0.00434 **
```

```
## Gestation     0.1618     0.0255   6.346 1.54e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4316 on 40 degrees of freedom
## Multiple R-squared:  0.5017, Adjusted R-squared:  0.4892
## F-statistic: 40.27 on 1 and 40 DF,  p-value: 1.542e-07
```

A simple linear regression was performed to test if gestational age predicts a baby's weight at birth. The analysis indicated that gestational age does predict birth weight ($\beta_1 = 0.1618 \pm 0.051$; p-value = 1.54e-07) and that as a baby is born one week later (between 33 and 45 weeks), the weight at birth increases by 0.1618 kg. The model (weight = 0.1618(gestation) - 3.0289) explained 48.92% of the total variation in birth weight of babies in this study.

Note on confidence interval: you can use confint(model) or you can just estimate with +/- 2*SE (which is what I did here)