# Stat 21 Homework 3

Solutions

Due: Sunday, Feb 13th by midnight

## Contents



Use this file as the template for your submission. Do not delete anything from this template unless you are prompted to do so (e.g. where to write your name above, where to write your solutions or code below). Make sure you have installed the following packages in your version of RStudio: `tidyverse`, `knitr` **before** you attempt to knit this document.

Your completed assignment should be submitted as a single **PDF** using the link under Week 3 titled "Submit HW 3 to Gradescope". You must use R markdown to write up your solutions. For any homework problems that involve coding in R, you must provide **both** the code and a written answer interpreting the output within the context of the problem. You are allowed to work with your classmates on this homework assignment but you are expected to write up your own solutions. Every answer must be supported by a written statement unless otherwise specified. *A good rule of thumb is to make sure your answer is understandable to someone who hasn't read the problem question (or code output associated with it).*

# Part I: Non-R Problems

## Problem 1

For the data in the table below, find the estimated regression equation by hand. Make sure you show your work.

| X | Y |
|---|---|
| 1 | 4 |
| 3 | 9 |
| 5 | 8 |

**Solution** estimated equation should be $\hat{y} = x + 4$. student must show work through code or writing out

formulas

### Problem 2

Suppose we have two random variables $X$ and $Y$. What are the differences among the following assumptions regarding $X$ and $Y$:

- $X$ and $Y$ are uncorrelated,
- $X$ and $Y$ are independent,
- $X$ and $Y$ have the same variance, and
- $X$ and $Y$ have the same distribution?

**Solution:**

- uncorrelated means that there is no linear relationship between $X$ and $Y$
- independent means that the probability $X$ takes on some value in its sample space is unrelated to the probability that $Y$ takes on some value in its sample space. independent RVs are automatically uncorrelated.
- same variance means that the dispersion/spread of $X$ and $Y$ are the same, however their location or the probabilities with which the different values occur may differ
- same distribution means the possible values for $X$ and $Y$ are the same and occur with the exact same probabilities

### Problem 3

Sketch (by hand) residual plots (with predicted response values on the horizontal axis) that show each of the following:

- constant variance and linearity;
- non-constant variance and linearity;
- constant variance and non-linearity;
- non-constant variance and non-linearity.

You can draw these plots on paper and use CamScanner to take a photograph of your drawings. Once you knit this document to a PDF file, you can then convert your image files to PDF files and merge everything together using a website such as smallpdf.com.

**Solution:** Solutions may vary but should include clearly labeled residual plots with/without funneling behavior (spread) and/or with/without some patterned behavior (linearity)

### Problem 4

In a few sentences and/or equations, briefly explain the relationship of the sum of squared errors term to the correlation between predictor $X$ and response $Y$.

**Solution:** SSE is critical for calculating the coefficient of determination ($r^2$ which, for SLR), is just the square of the sample correlation between $X$ and $Y$

## Part II: R coding problems

**Make sure you have installed the package `Stat2Data` before you knit this document. It contains the data sets for problems 5 and 6.**

## Problem 5

Biologists know that the leaves on plants tend to get smaller as temperatures rise. The data set `LeafWidth` has data on samples of leaves from the species *Dodonaea viscosa* subsp. *angustissima*, which have been collected in a certain region of South Australia for many years. The variable `Width` is the average width, in mm, of leaves, taken at their widest points, that were collected in a given year.

```
data("LeafWidth")
LeafWidth %>% names
```

```
## [1] "Width"   "Length"  "LWRatio" "Area"    "Year"
```

```
LeafWidth %>% head
```

```
##       Width   Length   LWRatio     Area Year
## 1 0.6578947 85.13158 129.40000 70.32548 1974
## 2 0.9210526 53.15789  57.71429 48.96122 1918
## 3 0.9210526 59.86842  65.00000 57.72161 1993
## 4 0.9210526 63.68421  69.14286 37.36150 1946
## 5 0.9210526 64.21053  69.71429 62.55194 1956
## 6 0.9210526 86.97368  94.42857 67.52078 1964
```

(a) Fit the regression of `Width` on `Year`. What is the fitted regression model? (Define any symbols that you use.)

(b) Interpret the coefficient of `Year` in the context of this setting.

(c) What is the predicted width of these leaves in the year 1966?

**Solution:**

```
leafreg <- lm(Width ~ Year, data=LeafWidth)
leafreg %>% summary
```

```
##
## Call:
## lm(formula = Width ~ Year, data = LeafWidth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1214 -1.1253 -0.3136  0.9320  5.4144
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.723091   8.574977   4.399 1.61e-05 ***
## Year        -0.017560   0.004358  -4.029 7.43e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.424 on 250 degrees of freedom
## Multiple R-squared:  0.06098,    Adjusted R-squared:  0.05723
## F-statistic: 16.24 on 1 and 250 DF,  p-value: 7.425e-05
```

(a) $\hat{y} = 37.72 - 0.028x$ where $\hat{y}$ is the average leave width and $x$ is the year. for full credit, any symbols or abbreviations must be clearly defined.

(b) As time progresses, each year leaf widths decrease, on average, by 0.018 mm

(c) $37.72 - 0.028(1966) = -17.33mm$, [ COME BACK TO AND NOTE WEIRDNESS]

## Problem 6

The data set `RailsTrails` contains data from a sample of 104 homes that were sold in 2007 in the city of Northampton, Massachusetts. The goal was to see if a home's proximity to a biking trail would relate to its selling price. Perhaps, for example, proximity to the trail would add value to the home. And to see if any such effect has changed over time, the researchers took estimated prices for these homes at four different time points using the website Zillow.com. Here we focus on the estimated price of homes in 2007 using the price in thousands of 2014 dollars (variable name `Adj2007`). The variable `Distance` measures the distance in miles to the nearest entry point to the rail trail network.

```
data("RailsTrails")
RailsTrails %>% names
```

```
##  [1] "HouseNum"     "Acre"         "AcreGroup"    "Adj1998"      "Adj2007"
##  [6] "Adj2011"      "BedGroup"     "Bedrooms"     "BikeScore"    "Diff2014"
## [11] "Distance"     "DistGroup"    "GarageSpaces" "GarageGroup"  "Latitude"
## [16] "Longitude"    "NumFullBaths" "NumHalfBaths" "NumRooms"     "PctChange"
## [21] "Price1998"    "Price2007"    "Price2011"    "Price2014"    "SFGroup"
## [26] "SquareFeet"   "StreetName"   "StreetNum"    "WalkScore"    "Zip"
```

```
RailsTrails %>% head
```

```
##   HouseNum Acre  AcreGroup  Adj1998  Adj2007  Adj2011 BedGroup Bedrooms
## 1        1 0.28 > 1/4 acre 148.3625 233.8418 191.8211   3 beds        3
## 2        2 0.29 > 1/4 acre 135.2072 261.4203 206.9677   3 beds        3
## 3        3 0.36 > 1/4 acre 256.5283 401.0359 347.9472   3 beds        3
## 4        4 0.26 > 1/4 acre 231.6795 305.0861 257.4915   3 beds        3
## 5        5 0.31 > 1/4 acre 271.8762 298.7660 236.6253  4+ beds        4
## 6        6 0.31 > 1/4 acre 192.9444 275.7840 243.2982   3 beds        3
##   BikeScore   Diff2014   Distance    DistGroup GarageSpaces GarageGroup
## 1        35   62.36645 2.40000000 Farther Away            2         yes
## 2        44   68.96375 1.97000000 Farther Away            1         yes
## 3        66   82.13365 0.04337121       Closer            2         yes
## 4        61   44.57055 0.55473485 Farther Away            1         yes
## 5        53 -102.70320 0.59659091 Farther Away            0          no
## 6        36   18.54260 1.88000000 Farther Away            1         yes
##   Latitude Longitude NumFullBaths NumHalfBaths NumRooms  PctChange Price1998
## 1 42.31533 -72.69397            1            0        5  42.036518     101.5
## 2 42.29856 -72.67474            1            0        5  51.005956      92.5
## 3 42.34379 -72.68023            2            1        7  32.017377     175.5
## 4 42.34446 -72.67221            1            1        6  19.238025     158.5
## 5 42.34253 -72.66437            1            0        6 -37.775723     186.0
## 6 42.31874 -72.69097            1            1        6   9.610333     132.0
##   Price2007 Price2011 Price2014    SFGroup SquareFeet      StreetName StreetNum
## 1     203.5     181.1   210.729 <= 1500 sf      0.966 Acrebrook Drive       406
## 2     227.5     195.4   204.171 <= 1500 sf      0.960       Autumn Dr        57
## 3     349.0     328.5   338.662  > 1500 sf      1.725     Bridge Road        31
## 4     265.5     243.1   276.250  > 1500 sf      1.727     Bridge Road       200
## 5     260.0     223.4   169.173  > 1500 sf      1.576     Bridge Road       395
## 6     240.0     229.7   211.487 <= 1500 sf      1.320 Brierwood Drive        23
##   WalkScore  Zip
## 1         9 1062
## 2         5 1062
## 3        46 1062
## 4        40 1060
## 5        32 1062
```

```
## 6         12 1062
```

(a) Create a scatter plot of `Adj2007` vs `Distance` and describe the relationship.

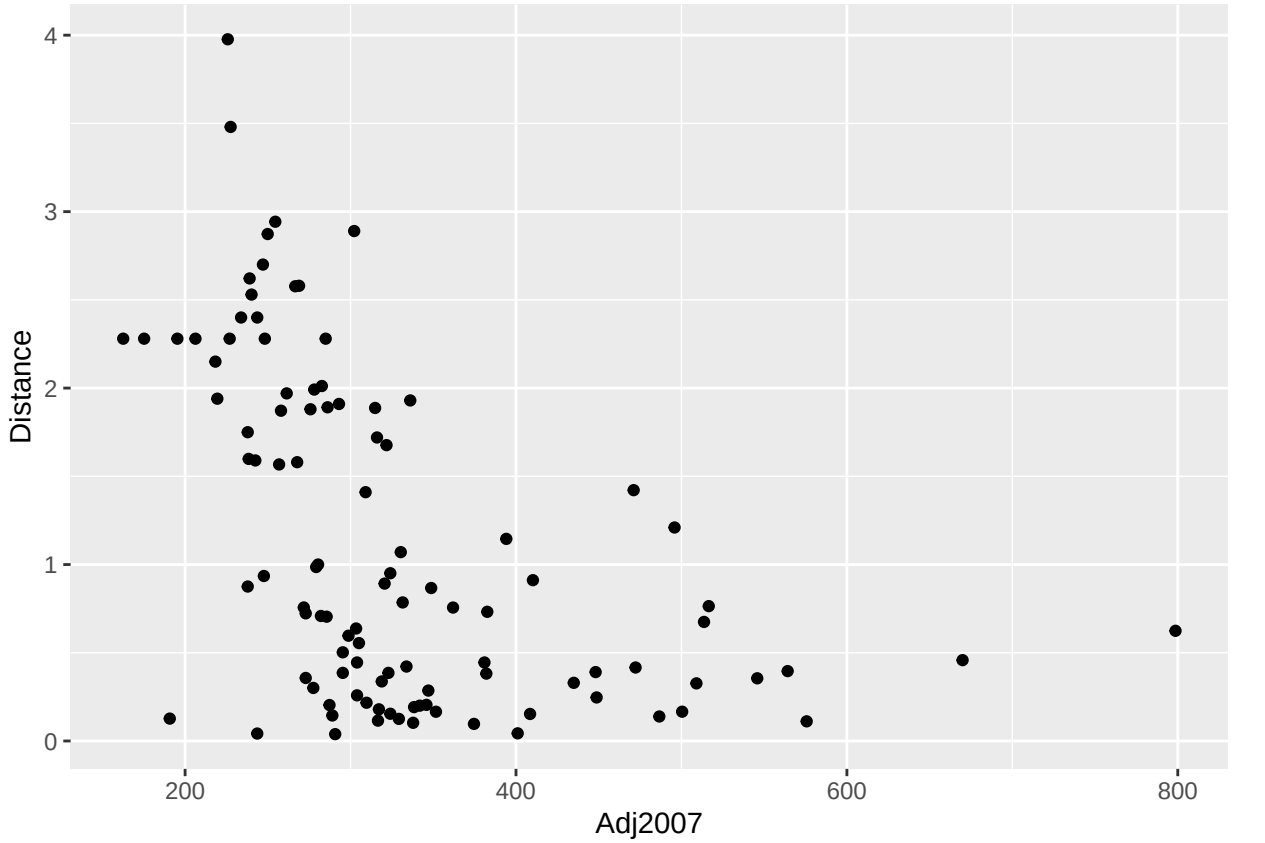
(b) Fit a SLR model and interpret the results, making a clear interpretation of the slope parameter. (Define any symbols that you use.)

(c) Report the regression standard error and discuss its meaning.

(d) Comment on the conditions of the model.

**Solutions:**

```
ggplot(data = RailsTrails, aes(x=Adj2007, y = Distance)) +
  geom_point()
```




```
railsreg <- lm(Adj2007 ~ Distance, data=RailsTrails)
railsreg %>% summary
```

```
##
## Call:
## lm(formula = Adj2007 ~ Distance, data = RailsTrails)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -190.55  -58.19  -17.48   25.22  444.41
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  388.204     14.052  27.626  < 2e-16 ***
```

```
## Distance     -54.427      9.659  -5.635 1.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 92.13 on 102 degrees of freedom
## Multiple R-squared:  0.2374, Adjusted R-squared:  0.2299
## F-statistic: 31.75 on 1 and 102 DF,  p-value: 1.562e-07
```

(a) negative trend, not linear, more curved relationship, concave up, but definitely some relationship between the two varbs

(b) $\hat{y} = 388.2 - 54.43x$ where $\hat{y}$ is the average adjusted price in 2007 dollars and $x$ is the distance to/from nearest rail trail network. for full credit, any symbols or abbreviations must be clearly defined. slope: on average, the adjusted price of a home decreases by \$54.43 for every additional mile between the home and the nearest rail trail entrance

(c) $\hat{\sigma} = 92.13$ is an estimate for the spread of the random noise/measurement error in predicting housing prices based on distance from rail trail network.

(d) For full credit must create a residuals plot to assess constant variance (and linearity if not already mentioned above) and a normal quantile plot to assess normality. also must mention whether or not independence and randomness are reasonable assumptions

## Problem 7

The number of pounds of steam used per month at a plant is thought to be related to the average monthly ambient temperature. The past year's usages (per 1000 lbs) and temperatures follow.

```
steam_data <- tibble(
  month = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"),
  temperature = c(21,24,32,47,50,59,68,74,62,50,41,30),
  usage = c(185.79,214.47,288.03,424.84,454.68,539.03,621.55,675.06,562.03,452.93,369.95,273.98))
```

(a) Fit a simple linear regression model to the data. (Define any symbols that you use.)

(b) Display the analysis-of-variance table for this model and test for the significance of the regression model. State your null and alternative hypotheses and interpret the conclusion in the context of the problem.

(c) Plant management believes that an increase in average ambient temperature of 1 degree will increase average monthly steam consumption by 10,000 lb. Do the data support this statement?

**Solution:**

```
steamreg <- lm(usage ~ temperature, data = steam_data)
steamreg %>% summary
```

```
##
## Call:
## lm(formula = usage ~ temperature, data = steam_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5629 -1.2581 -0.2550  0.8681  4.0581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.33209    1.67005  -3.792  0.00353 **
## temperature  9.20847    0.03382 272.255  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.946 on 10 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 7.412e+04 on 1 and 10 DF,  p-value: < 2.2e-16
```

```
steamreg %>% anova
```

```
## Analysis of Variance Table
##
## Response: usage
##             Df Sum Sq Mean Sq F value    Pr(>F)
## temperature  1 280590  280590   74123 < 2.2e-16 ***
## Residuals   10     38       4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) $\hat{y} = -6.33 + 9.21x$ where $\hat{y}$ is the average steam usage per 1000s lbs per month and $x$ is the average monthly temperature. for full credit, any symbols or abbreviations must be clearly defined.

(b) $H_0 : \beta_1 = 0, \quad H_A : \beta_1 \neq 0$ has a p-value of less than $2.2 \times 10^{16} \approx 0$ so at any significance level, we reject the null in favor of the alternative. That is, this data provides evidence that the linear relationship between average temperature and steam usage is non-trivial, or is "significant". this is one indicator of the model fitting the data well.

(c) based on the regression model, we expect an increase in 1 degree to increase steam consumption by an average of $9.12 \times 1000$ lbs which is close to the estimate from plant management (since $\hat{\sigma} = 4$, $(10 - 9.12)/2 \approx 0.44$ is actually within 1 standard deviation)

## Problem 8

Install the R package `MVP` before running the code chunk below.

```
library('MPV')
NFL_data <- table.b1
NFL_data %>% names
```

```
##  [1] "y"  "x1" "x2" "x3" "x4" "x5" "x6" "x7" "x8" "x9"
```

Use the code above to read this data into your R library. This data set concerns the performance of the 26 National Football League teams in 1976. It is suspected that the number of yards gained rushing by opponents (variable x8) has an effect on the number of games won by a team (variable y). Fit a simple linear regression model relating games won, y, to yards gained rushing by opponent, x8, and answer the following questions with this model.

(a) Display the analysis-of-variance table for this model and test for the significance of the regression model. State your null and alternative hypotheses and interpret the conclusion in the context of the problem.

(b) Find and interpret a 95% CI on the slope of the yards gained rushing.

(c) What percent of the total variability in the number of games won is explained by this model?

**Solution:**

```
nflreg <- lm(y ~ x8, data=NFL_data)
nflreg %>% anova
```

```
## Analysis of Variance Table
##
## Response: y
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x8         1 178.09 178.092  31.103 7.381e-06 ***
## Residuals 26 148.87   5.726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) $H_0 : \beta_1 = 0, \quad H_A : \beta_1 \neq 0$ has a p-value of $7.4 \times 10^{-6} \approx 0$ so at any significance level, we reject the null in favor of the alternative. That is, this data provides evidence that the linear relationship between the number of yards gained rushing by opponents and total number of games won by a team is non-trivial, or is "significant". this is one indicator of the model fitting the data well.

(b) $[-0.0096, -0.0044]$ and must include a correct problem-specific interpretation

(c) this is the definition of the coefficient of determination

---

The data shown below present the average number of surviving bacteria in a canned food product and the minutes of exposure to 300 degree Fahrenheit heat. Use this data to answer Problems 9-10.

```
bacteria_data <- tibble(bacteria_count = c(175, 108, 95, 82, 71, 50, 49, 31, 28, 17, 16, 11),
                        minutes_exposure = c(1,2,3,4,5,6,7,8,9,10,11,12))
```

## Problem 9

Fit a SLR model with the number of bacteria as the response.

(a) What is the estimated regression equation? (Define any symbols that you use.)

(b) Display the residual plot and calculate the coefficient of determination to comment on the fit of this model.

(c) Based on this model, what is the average effect on the bacterial growth per each additional minute of exposure?

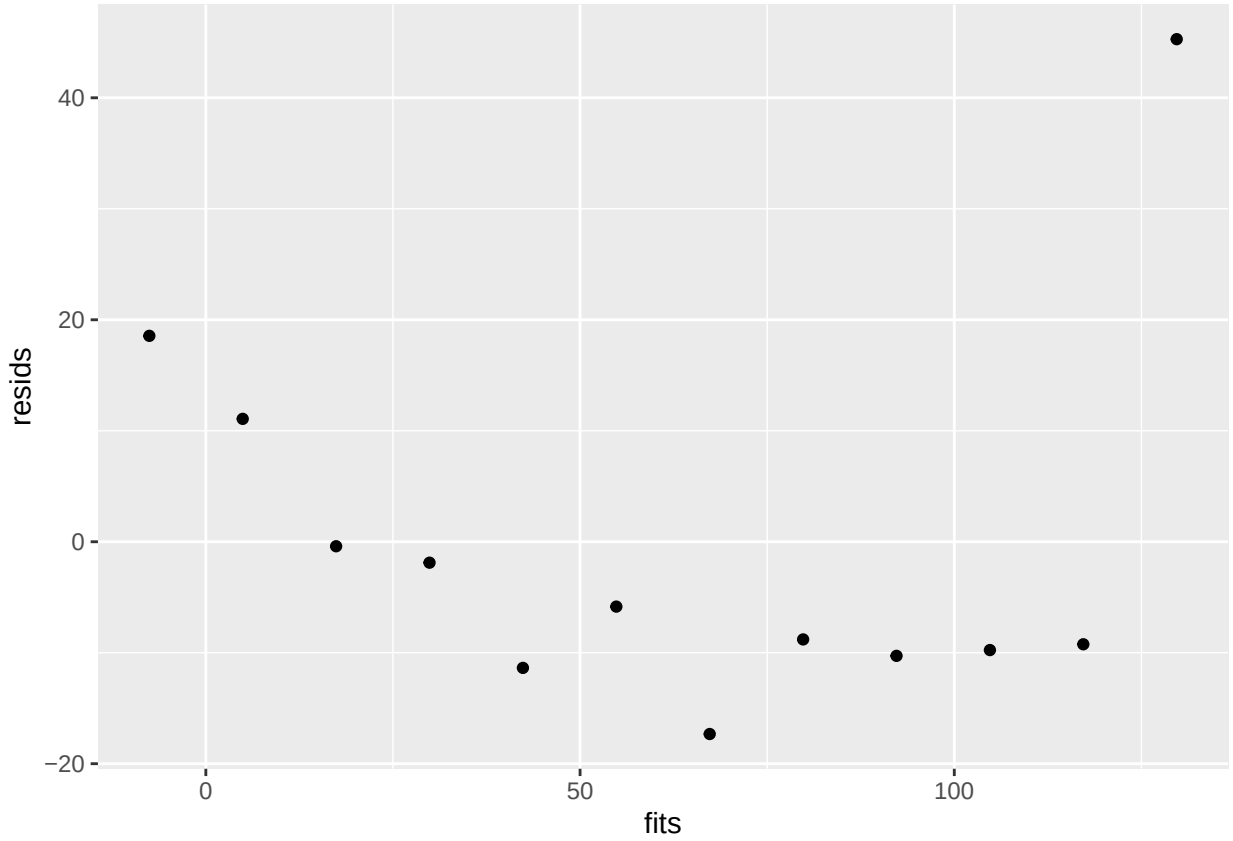
**Solution:**

```
bacteriareg <- lm(bacteria_count ~ minutes_exposure, data=bacteria_data)
bacteriareg %>% summary
```

```
##
## Call:
## lm(formula = bacteria_count ~ minutes_exposure, data = bacteria_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.323  -9.890  -7.323   2.463  45.282
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        142.20      11.26  12.627 1.81e-07 ***
## minutes_exposure   -12.48       1.53  -8.155 9.94e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.3 on 10 degrees of freedom
## Multiple R-squared:  0.8693, Adjusted R-squared:  0.8562
## F-statistic: 66.51 on 1 and 10 DF,  p-value: 9.944e-06
```

```
bacterial_data_all <- bacteria_data %>% mutate(resids = bacteriareg$residuals, fits = bacteriareg$fitte

ggplot(bacterial_data_all, aes(x=fits, y=resids)) +
  geom_point()
```




(a) $\hat{y} = 142.2 - 12.48x$ where $\hat{y}$ is the average bacterial count and $x$ is the minutes of exposure to heat. for full credit, any symbols or abbreviations must be clearly defined.

(b) $R^2 = 0.85$ which is pretty high indicating good model fit; residual plot however shows that there is an obvious issue with the linearity assumption

(c) decreases bacterial count by 12.48 on average

## Problem 10

Identify an appropriate transformation of this data that can be more appropriately modeled by a linear relationship. (Hint: Transform the response variable by taking the logarithm or taking the inverse and using this transformed data as the new response variable.)

(a) Write the estimated regression equation for the transformed data. (Define any symbols that you use.)

(b) Fit a SLR model to the transformed data and display the residual plot and the coefficient of determination. Interpret these in relation to the adequacy of this model.

(c) What is the average effect on the bacterial growth per each additional minute of exposure?

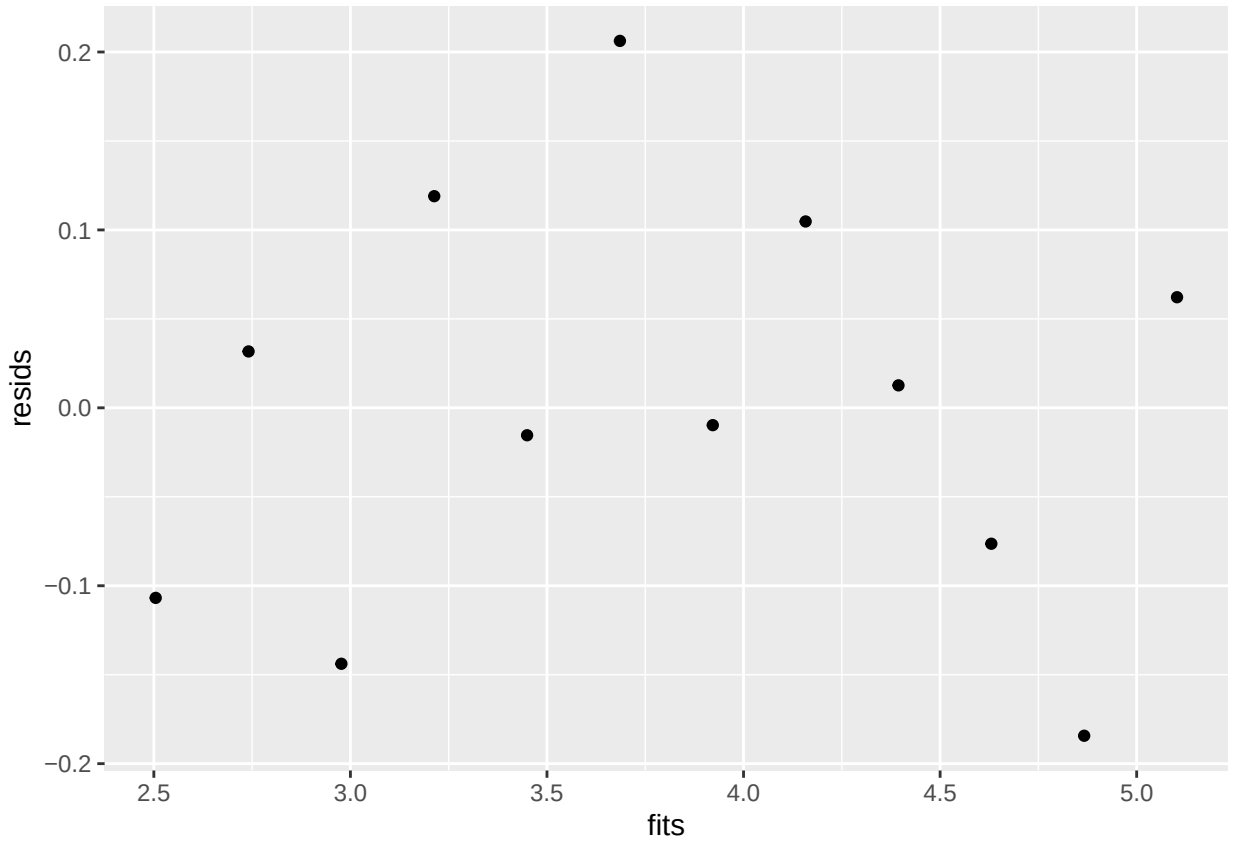
```
bacteriareg_transformed <- lm(log(bacteria_count) ~ minutes_exposure, data=bacteria_data)
bacteriareg_transformed %>% summary
```

```
##
```

```
## Call:
## lm(formula = log(bacteria_count) ~ minutes_exposure, data = bacteria_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.184303 -0.083994  0.001453  0.072825  0.206246
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.33878    0.07409   72.05 6.47e-15 ***
## minutes_exposure -0.23617    0.01007  -23.46 4.49e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1204 on 10 degrees of freedom
## Multiple R-squared:  0.9822, Adjusted R-squared:  0.9804
## F-statistic: 550.3 on 1 and 10 DF,  p-value: 4.489e-10
```

```
bacterial_transformed_all <- bacteria_data %>% mutate(transformed_y = log(bacteria_count),
                                                      resids = bacteriareg_transformed$residuals,
                                                      fits = bacteriareg_transformed$fitted.values)

ggplot(bacterial_transformed_all, aes(x=fits, y=resids)) +
  geom_point()
```




**Solution:**

(a) $\hat{y} = 5.338 - 0.236x$ where $\hat{y}$ is the average (natural) logarithm of the bacterial count and $x$ is the minutes

of exposure to heat. for full credit, any symbols or abbreviations must be clearly defined.

(b) key point is that residual plot looks much better for the transformed data

(c) $e^{\hat{\beta}_1} = 0.79$ represents the multiplicative change in $\hat{y}$ for a single unit increase in $x$