Two Quantitative Variables: Association

Stat 120

January 16 2023

Describing associations between two quantitative variables

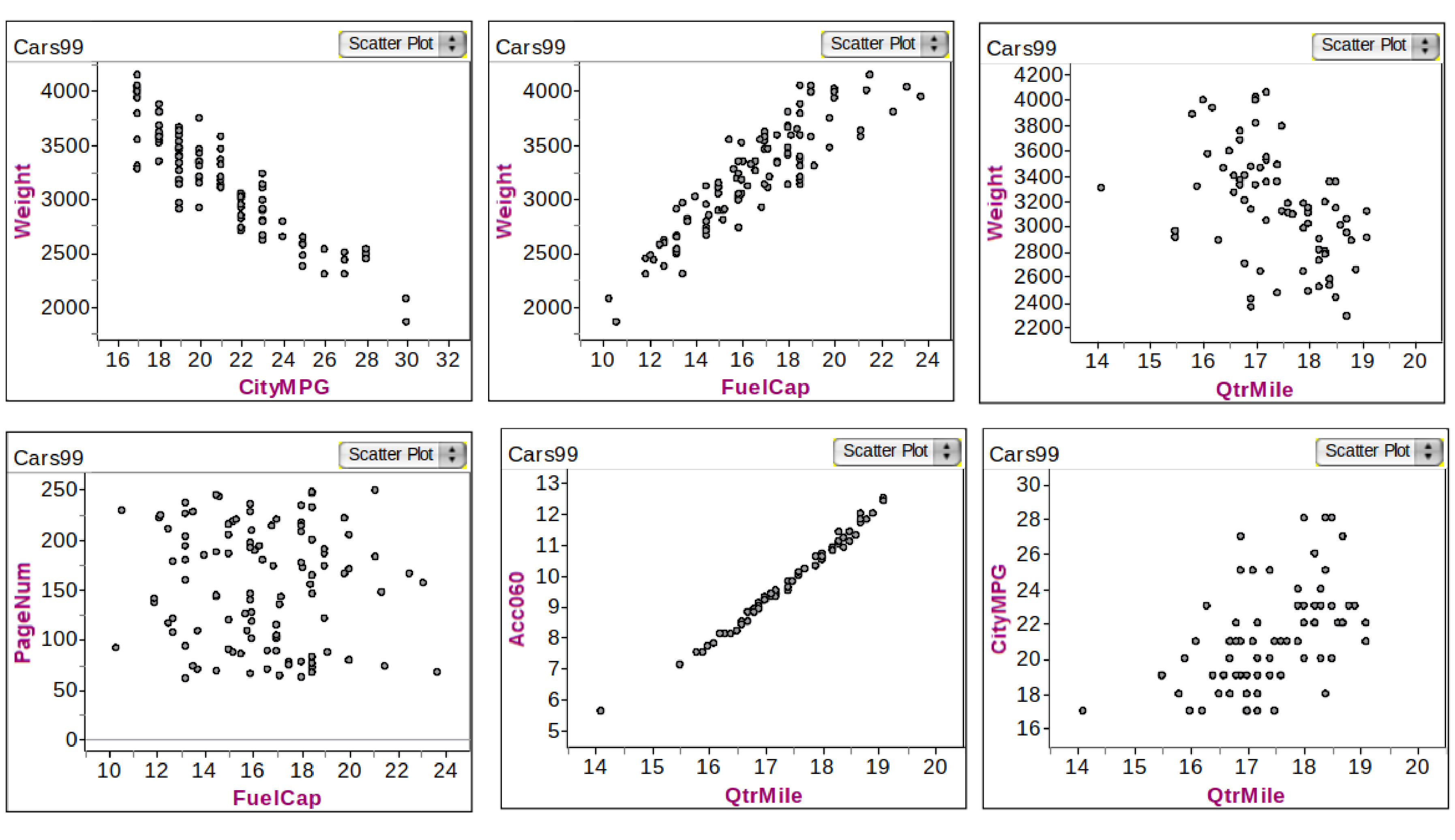
Data: each case i has two measurements

- x_i is explanatory variable
- y_i is response variable

A **scatterplot** is the plot of (x_i, y_i).

- form? linear or non-linear
- direction? positive, negative, no association
- strength? amount of variation in y around a "trend"

Example: Associations in Car dataset



Various Associations of quantitative variables in Cars data

Direction

positive association: as x increases, y increases

- age of the husband and age of the wife
- height and diameter of a tree

negative association: as x increases, y decreases

- number of cigarettes smoked per day and lung capacity
- depth of tire tread and number of miles driven on the tires

Correlation Coefficients

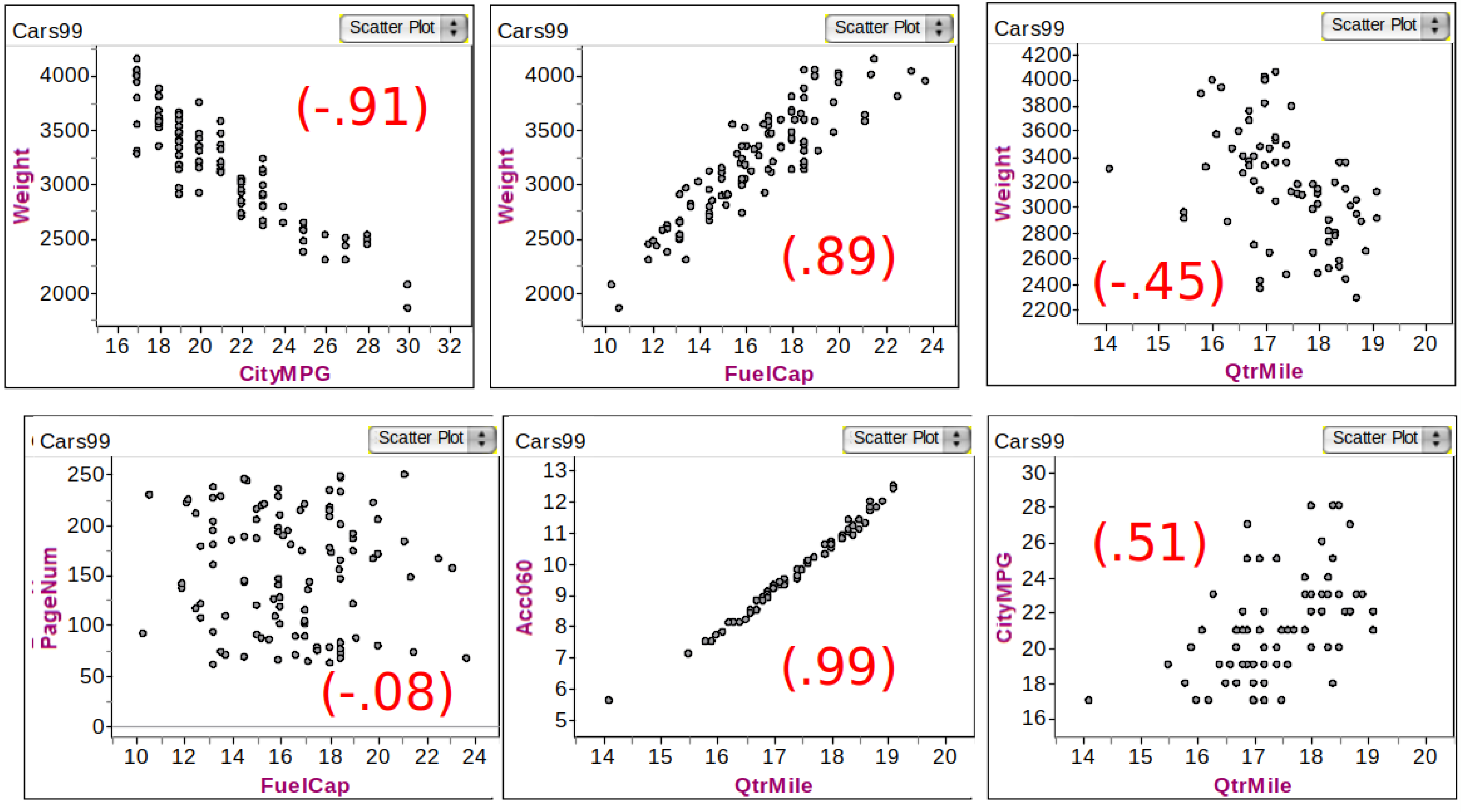
Correlation coefficient: *denoted r (sample) or ρ (population)*

- ***Strength of linear association***
  - $r \approx \pm 1:$ ***strong***
  - $r \approx 0:$ ***weak***
- ***Direction of linear association***
  - $r > 0:$ ***positive***
  - $r < 0:$ ***negative***

Correlation can be heavily affected by outliers. Plot your data!

```
# R-code
cor(data$x, data$y)  # order of x and y doesn't matter!
```

Car Correlations



Correlations of various variables in Cars data

Linear Regression

Goal: To find a straight line that best fits the data in a scatterplot

The estimated regression line is

$$\hat{y} = a + bx$$

- ***x is the explanatory variable***
- ***$\hat{y}$ is the predicted response variable.***

Slope: increase in predicted y for every unit increase in x

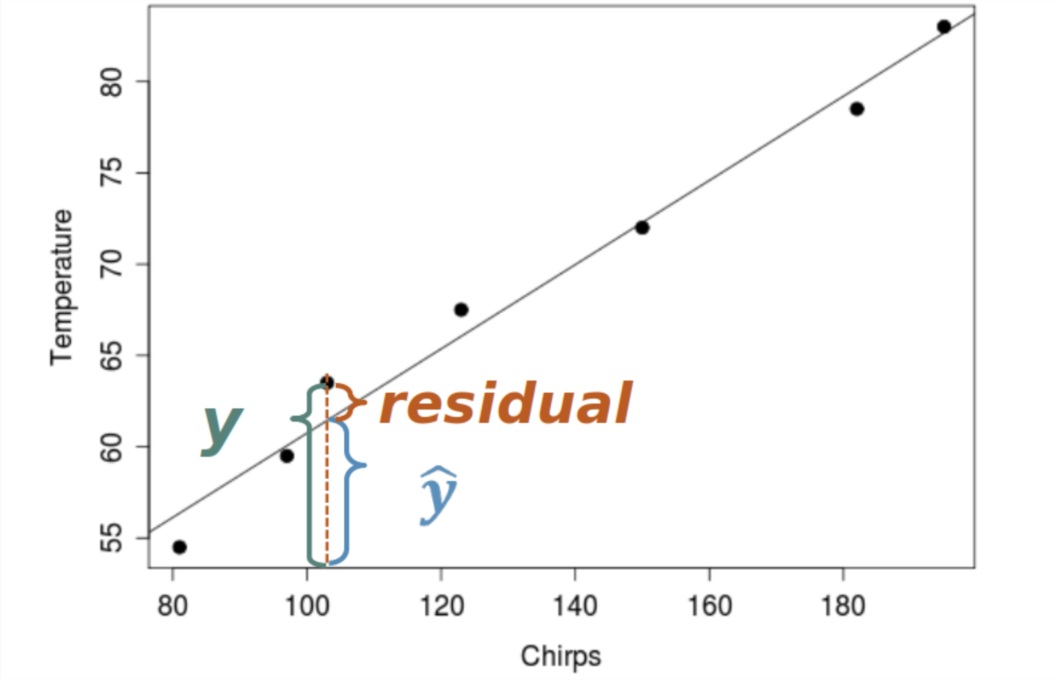
$$b = \frac{\text{change } \hat{y}}{\text{change } x}$$

Intercept: predicted y value when $x = 0$

$$\hat{y} = a + b(0) = a$$

Residuals

- ***Geometrically***, *residual is the vertical distance from each point to the line*
- ***Mathematically***, $y - \hat{y}$ *is the residual of* y *at* x
- *If the model is linear, measure how much variation in the response is explained by the model.*



Residuals

Least Squares Line

The Least squares line is the line which minimizes the sum of squared residuals. Want to minimize:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2$$

“least squares line” = “regression line”

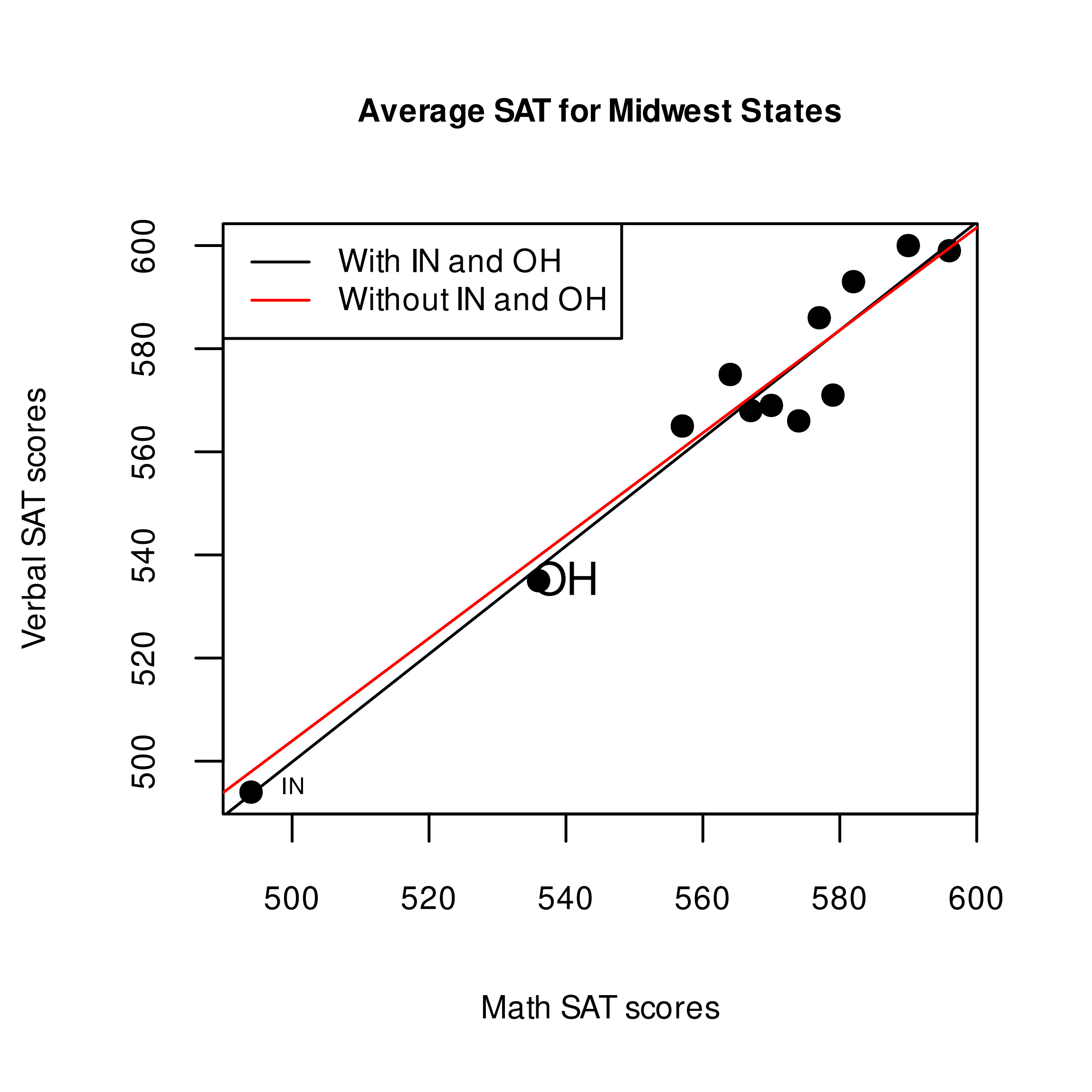
Regression Caution!

- *Do not use the regression equation or line to predict values far from those that were used to create it --> Extrapolation!*
- *The regression line/equation should only be used if the association is approximately linear*
- *Unlike correlation, for linear regression it does matter which is the explanatory variable and which is the response*

Outliers Detection

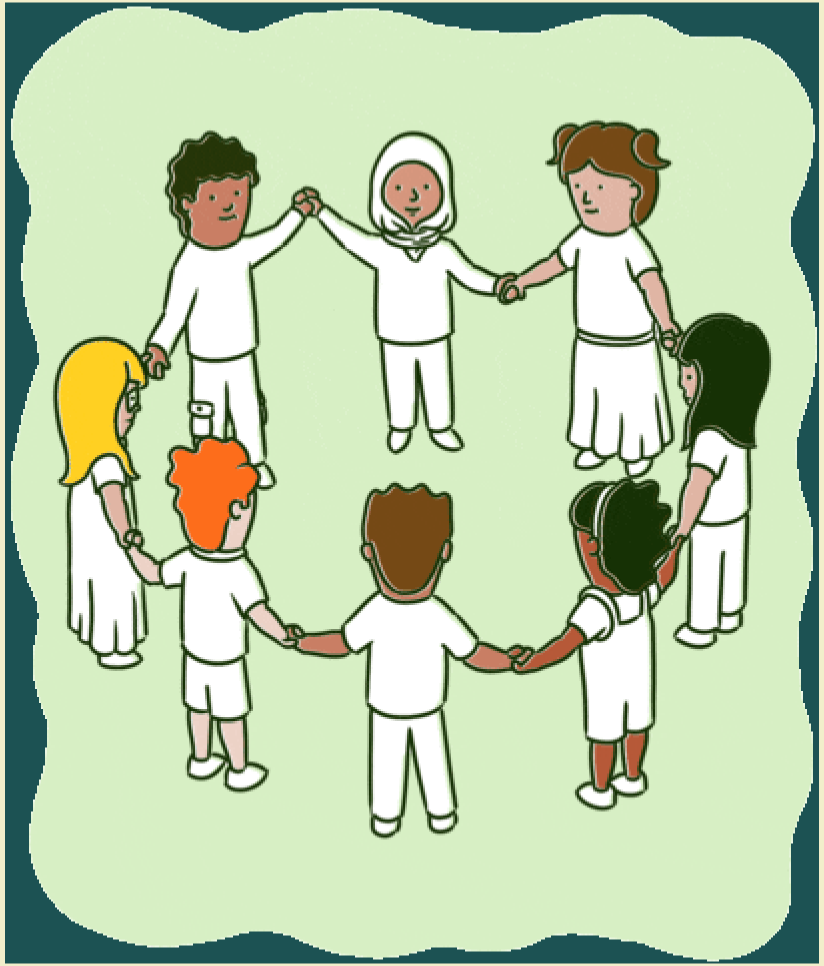
Outliers can be very influential on the regression line

- remove the points and see if the regression line changes significantly



Your Turn 1

05:00



Go to our class moodle and skim through class activity 1

Feel free to talk to your neighbor

Regression line of Blood Alcohol Content (BAC) data

Regression of BAC on number of beers

```
bac.lm <- lm(BAC ~ Beers, data=bac)
summary(bac.lm)

Call:
lm(formula = BAC ~ Beers, data = bac)

Residuals:
      Min        1Q    Median        3Q       Max
-0.027118 -0.017350  0.001773  0.008623  0.041027

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.012701   0.012638  -1.005    0.332
Beers        0.017964   0.002402   7.480 2.97e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 0.02044 on 14 degrees of freedo
Multiple R-squared:  0.7998,    Adjusted R-squared:  0.7
F-statistic: 55.94 on 1 and 14 DF,  p-value: 2.969e-06
```

Slope, $b = 0.0180$:

- **Estimate** column and **Beers** row

Intercept, $a = -0.0127$:

- **Estimate** column and **Intercept** row

Regressing BAC on number of beers

$$\widehat{BAC} = -0.0127 + 0.0180(Beers)$$

Slope Interpretation?

- ***Each additional beer consumed is associated with a*** 0.0180 ***unit increase in BAC***

y-intercept Interpretation?

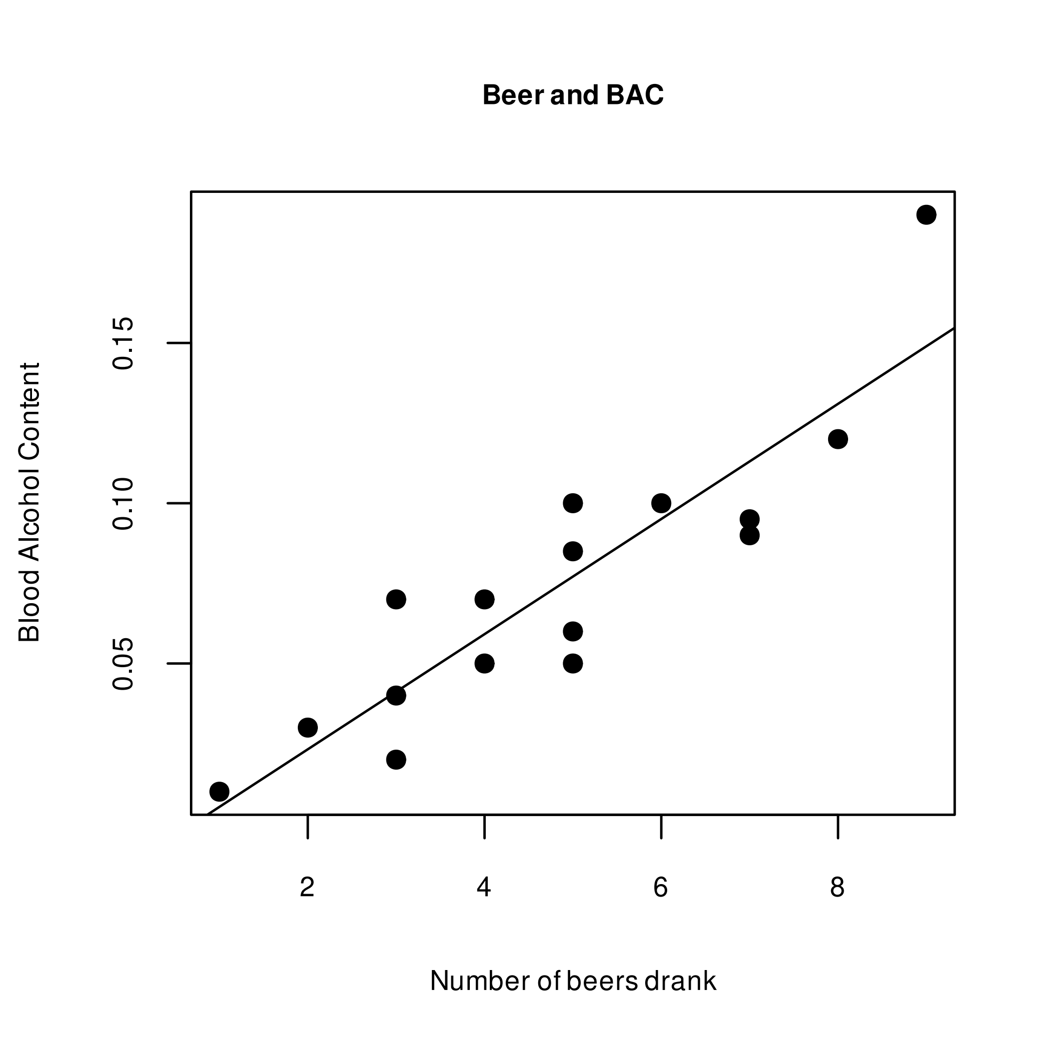
- ***Predicted BAC with 0 beers consumed***

Regressing BAC on number of beers

```r
plot(BAC ~ Beers, data=bac, pch=19,
     main="Beer and BAC",
     xlab="Number of beers drank",
     ylab = "Blood Alcohol Content")
abline(bac.lm) # adds regression line
```

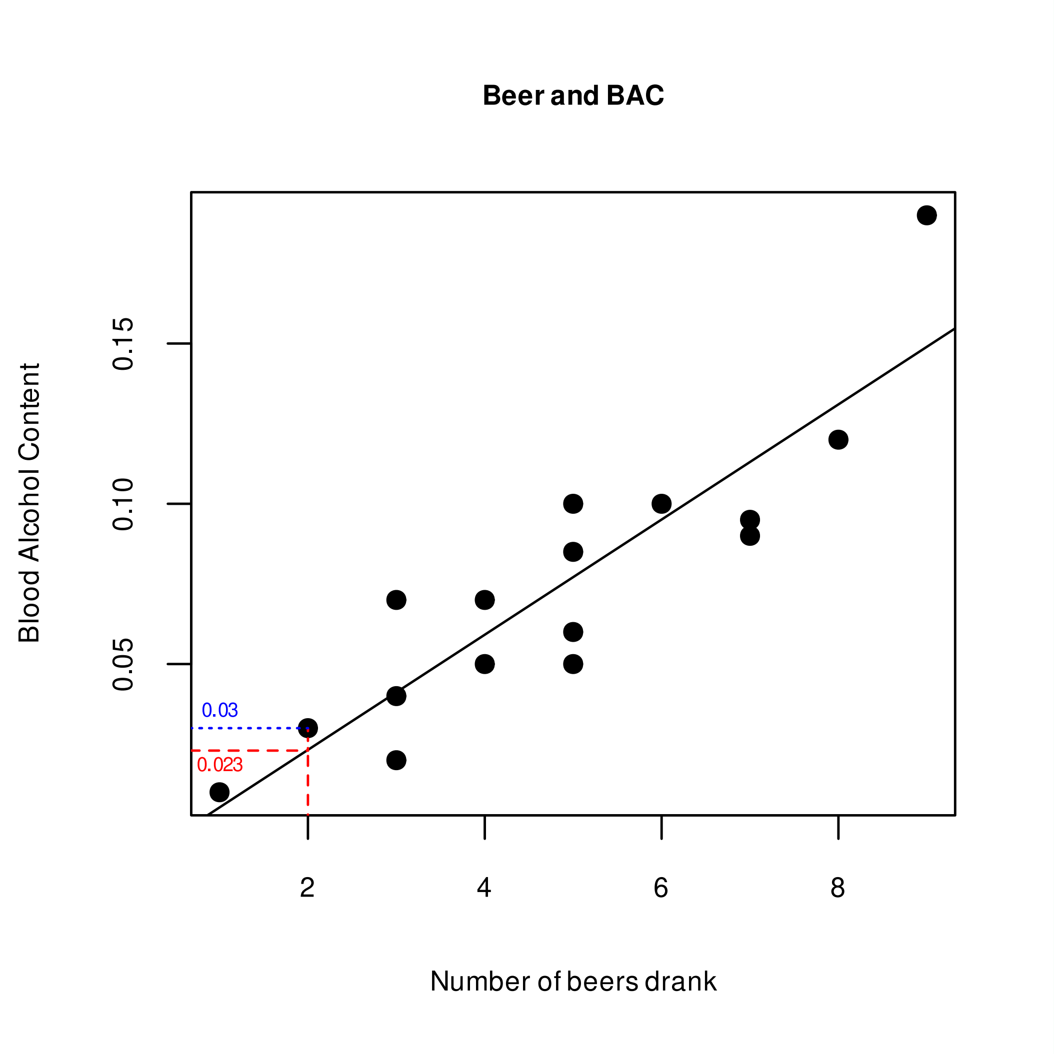
If your friend drank 2 beers, what is your best guess at their BAC after 30 minutes?

$$\widehat{BAC} = -0.0127 + 0.0180(2) = 0.023$$

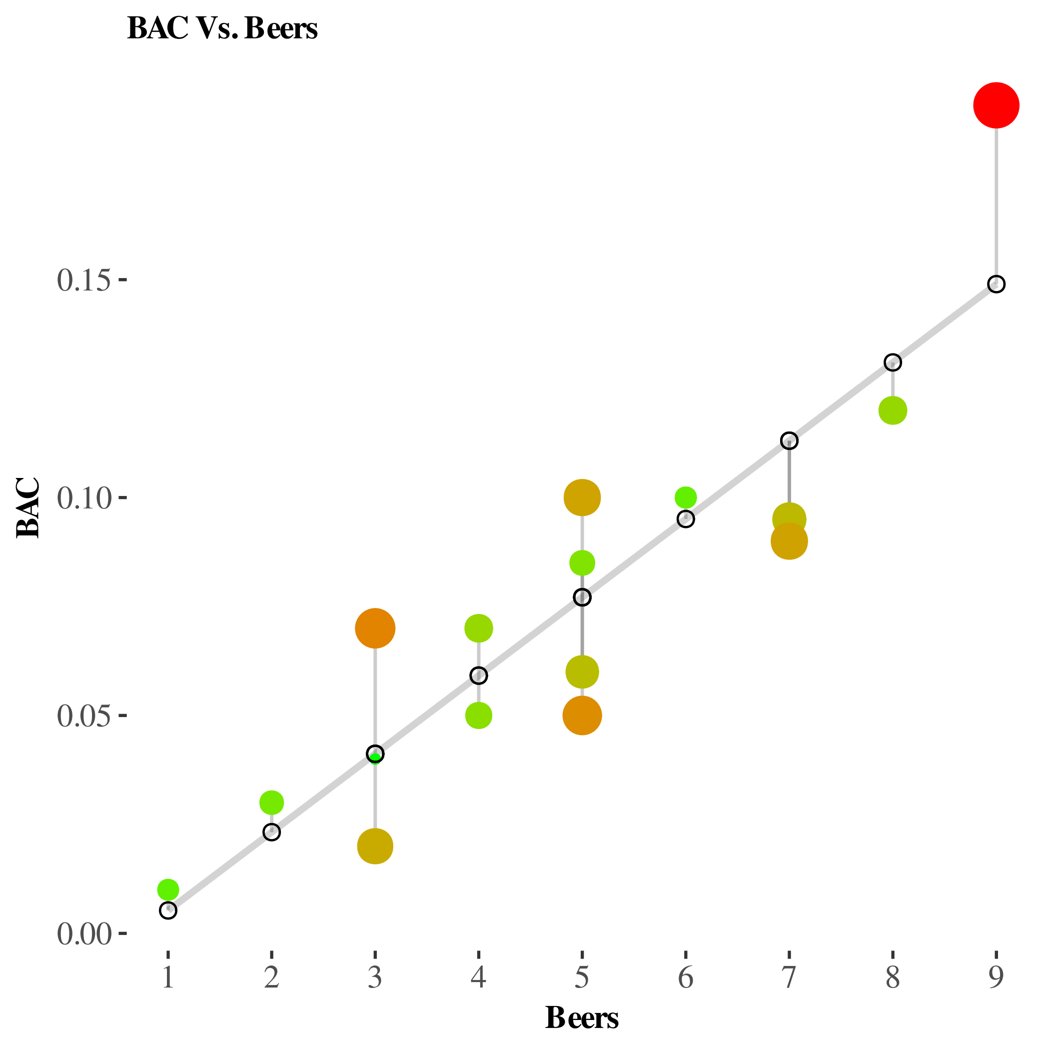


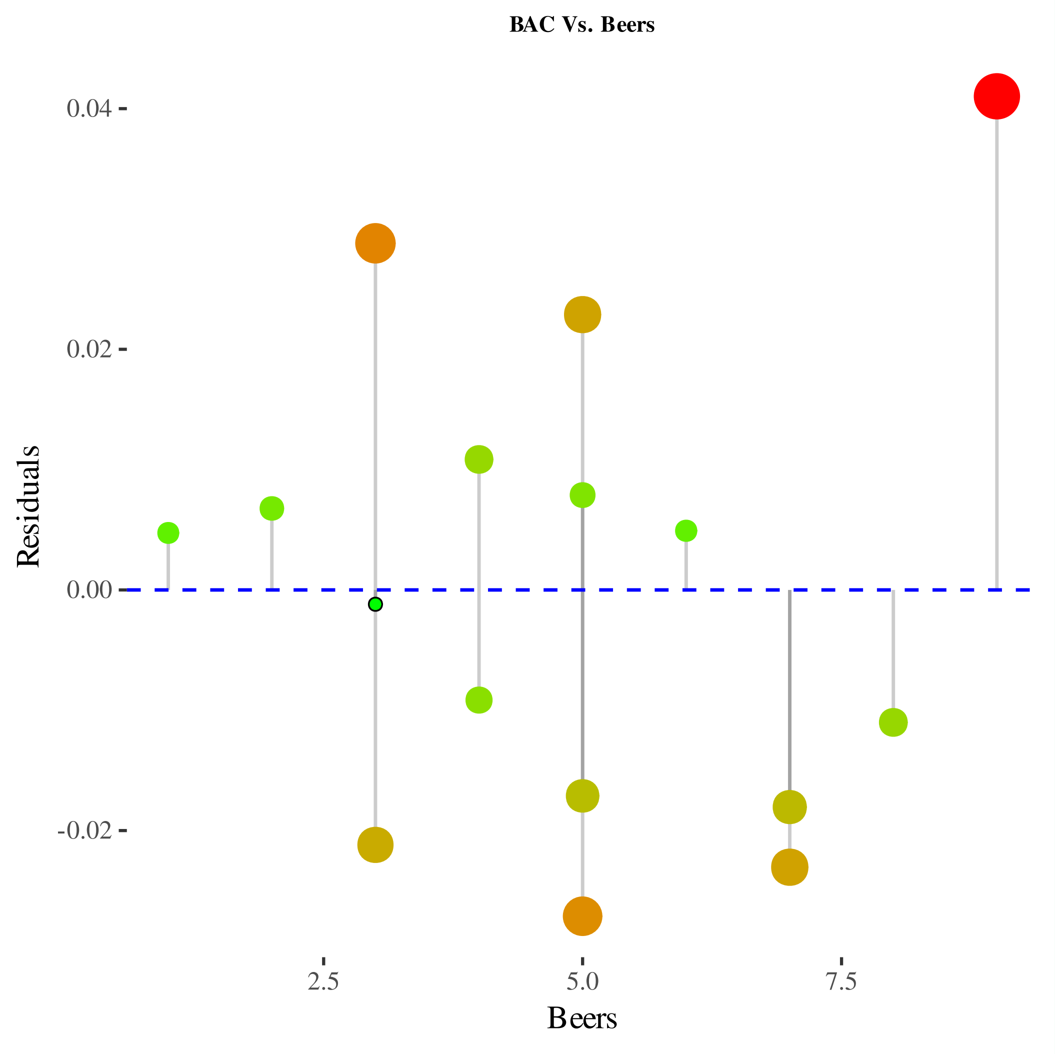
Regressing BAC on number of beers

Find the residual for the student in the dataset who drank 2 beers and had a BAC of 0.03. The residual is about $y - \hat{y} = 0.03 - 0.023 = 0.007$



Residuals Plot





R-squared

R-squared is proportion (or percentage) of variability observed in the response y which can be explained by the explanatory variable x.

$$R^2 = 1 - \text{unexplained variation} = 1 - \frac{s^2_{\text{residuals}}}{s^2_y}$$

- $R^2 = r^2$ in simple linear regression model (One explanatory variable)

BAC : $R^2 = 0.7998$

- ***The number of beers consumed explains about 80.0% of the observed variation in BAC***
- ***What factors (variables) besides number of beers drank might explain the other roughly 20% of variation in BAC?***

R-squared

Called **Multiple R-squared** in the summary output

```
summary(bac.lm)

Call:
lm(formula = BAC ~ Beers, data = bac)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.027118 -0.017350  0.001773  0.008623  0.041027 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.012701   0.012638  -1.005    0.332    
Beers        0.017964   0.002402   7.480 2.97e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02044 on 14 degrees of freedom
Multiple R-squared:  0.7998,	Adjusted R-squared:  0.7855 
F-statistic: 55.94 on 1 and 14 DF,  p-value: 2.969e-06
```

Additional Comments

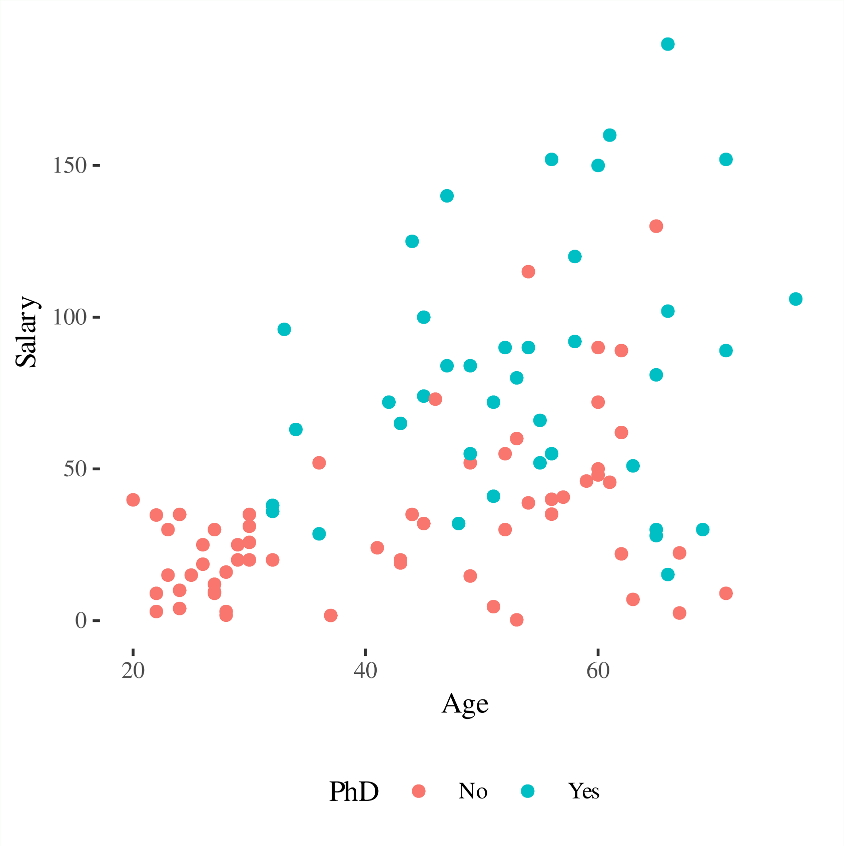
Include confounding variables when appropriate

- ***augment scatterplot with colors for each category***

```
ggplot(data, aes(x=x,y=y,color=z)) + geom_point()
```

- split (subset) data by categories, run regressions for each group.
- look for outliers that affect the fitted model and correlation
- fit the model/correlation with and without case(s) to see the affects

Adding a categorical variable




```
ggplot(salarydata, aes(x=Age,
                       y=Salary,
                       color=PhD)) +
  geom_point()
```

- Visually split the data by PhD status
- Potentially different trends

Adding a categorical variable

Adding a categorical variable: stats by group

*Can also use **filter** function available under **dplyr** package to divide responses into the groups of interest*

```
library(dplyr)
table(salarydata$PhD)

 No Yes
 61  39
```

```
salary.NoPhD <- filter(salarydata, PhD == "No")
salary.PhD <- filter(salarydata, PhD == "Yes")
```

```
cor(salary.NoPhD$Salary,salary.NoPhD$Age)
[1] 0.4759365
cor(salary.PhD$Salary,salary.PhD$Age)
[1] 0.2376678
```

Outliers: Average SAT by state

```
library(dplyr)
sat <- read.csv("https://math.carleton.edu/Stats215/RLabManual/sat.csv")
sat.MW <- filter(sat, region == "Midwest") # just MW states
cor(sat.MW$math, sat.MW$verbal)
[1] 0.9731605

sat.lm <- lm(math ~ verbal, data=sat.MW)
sat.lm

Call:
lm(formula = math ~ verbal, data = sat.MW)

Coefficients:
(Intercept)       verbal
    -23.584        1.047
```

```
summary(sat.lm)$r.squared
[1] 0.9470413
```

Correlation = 0.9732, Regression Slope = 1.0469, R-squared = 94.7%

Outliers: Average SAT by state, excluding Indiana and Ohio

```
which(sat.MW$verbal < 550)
[1]  2 10
cor(sat.MW$math[-c(2,10)], sat.MW$verbal[-c(2,10)])
[1] 0.8465318
sat.lm.noIO <- lm( math ~ verbal, data=sat.MW, subset = -c(2,10))
sat.lm.noIO

Call:
lm(formula = math ~ verbal, data = sat.MW, subset = -c(2, 10))

Coefficients:
(Intercept)       verbal
     6.1453       0.9956
```

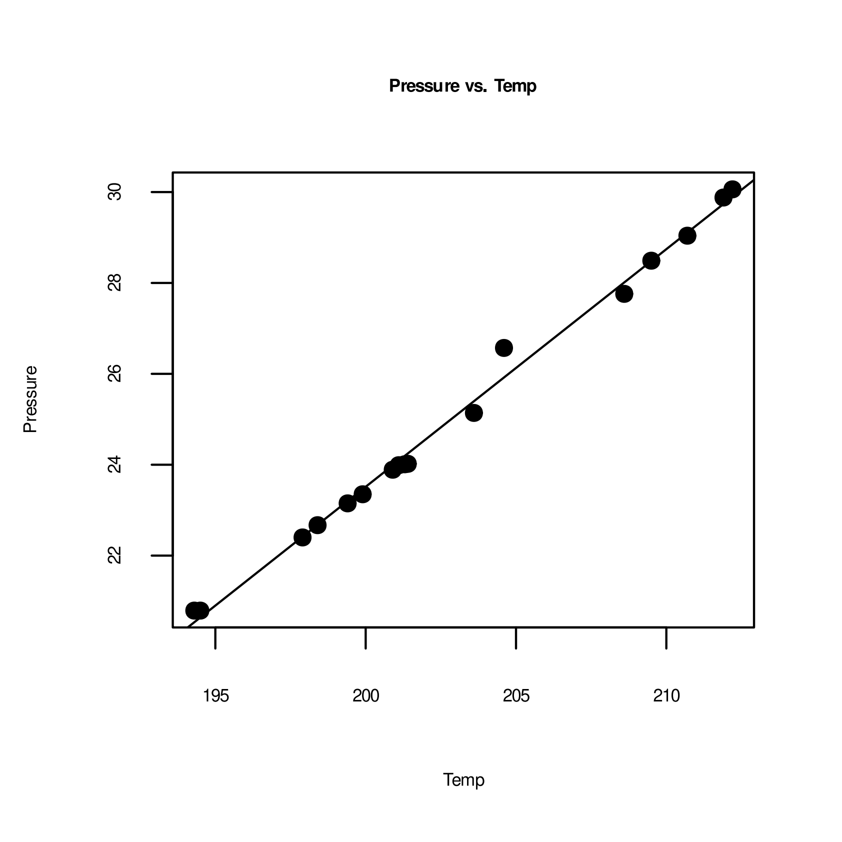
```
summary(sat.lm.noIO)$r.squared
[1] 0.7166161
```

Correlation = 0.8465, Regression slope = 0.9956 , R-squared = 71.66%

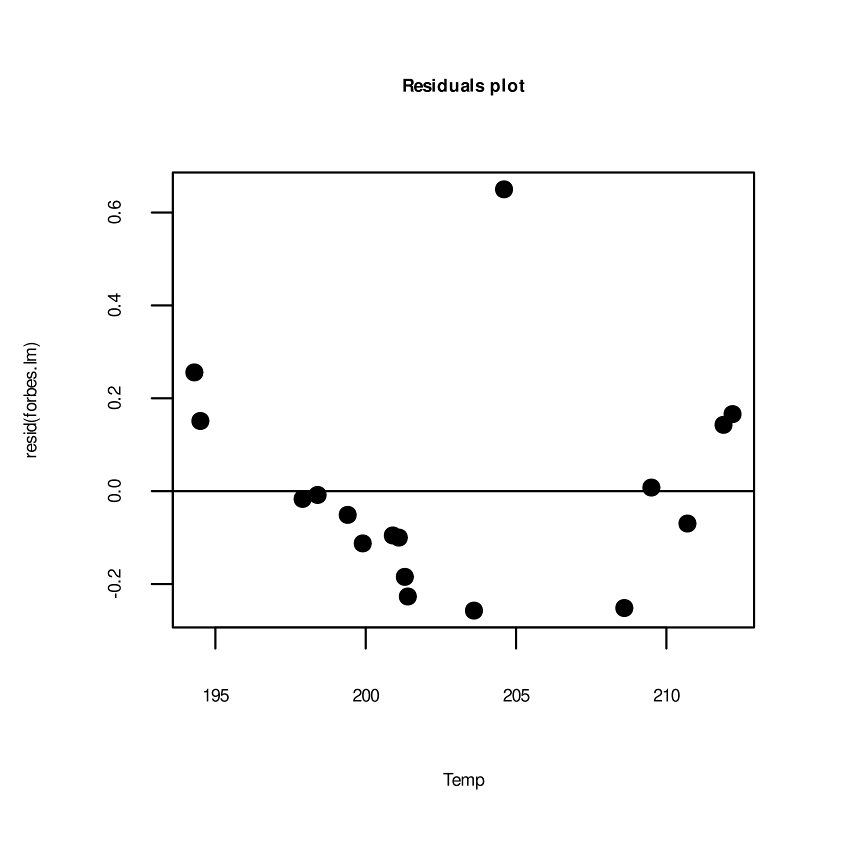
Non-linear Patterns

James D. Forbes 1857 experiment: Can atmospheric pressure be determined from the boiling point of water? Is the relationship linear?

Seems Linear!



Curvature!



Residuals Plot

While the scatterplot of pressure vs. temp may look linear relationship, the residuals plot reveals that there is curvature in the relationship.

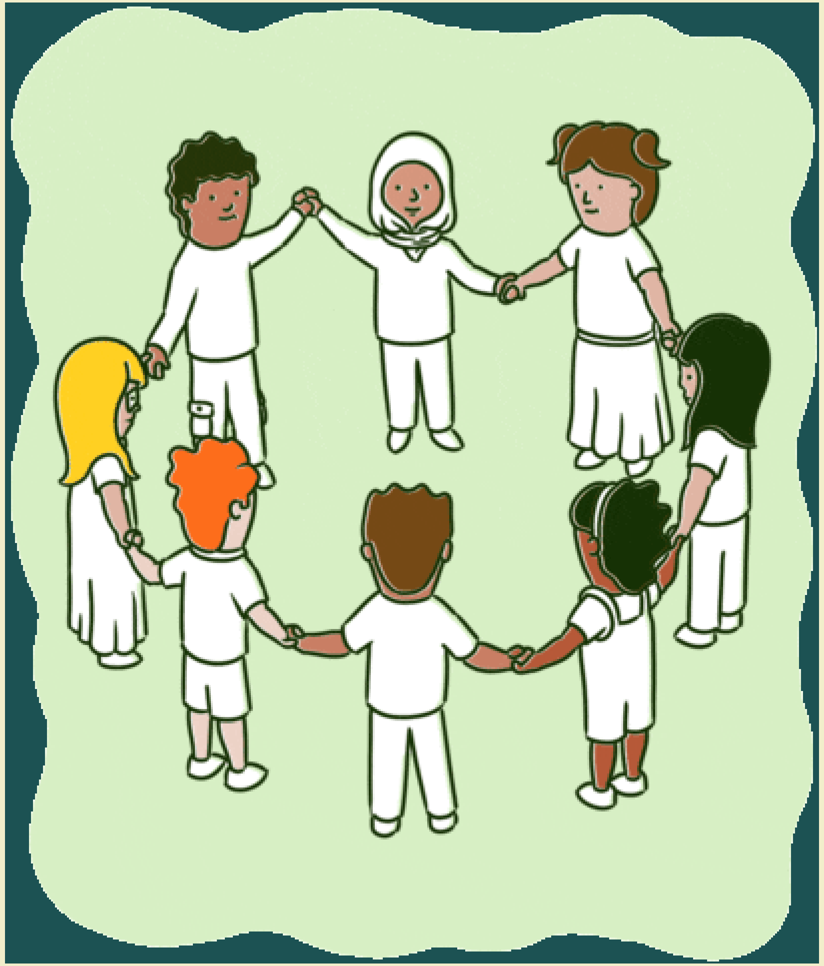
Correlation is almost 1, so why not use linear regression!!

- ***Because the true nature of the relationship is not linear***
- ***Would systematically overestimate pressure for midrange temps and underestimate pressure for high/low temps.***

However, temp and log(pressure) have a linear relationship and we can apply linear model after transformation of the variables!!

Your Turn 2

10:00



Go over the remaining portion of in class activity and let me know if you have any questions!