# RegressionModelsNotes

Coursera Course by John Hopkins University

INSTRUCTORS: Dr. Brian Caffo, Dr. Roger D. Peng, Dr. Jeff Leek

## Contents

# Intro

This course covers regression analysis, least squares and inference using regression models. Special cases of the regression model, ANOVA and ANCOVA will be covered as well. Analysis of residuals and variability will be investigated. The course will cover modern thinking on model selection and novel uses of regression models including scatterplot smoothing.

## GitHub Link for Lectures

### Link to the GitHub for this course

## Course Book

### Regression Models for Data Science in R, through Leanpub

Further Reading: **Advanced Linear Models for Data Science**

## Instructor's Note

"*We believe that the key word in Data Science is 'science'. Our course track is focused on providing you with three things:*
1) *An introduction to the key ideas behind working with data in a scientific way that will produce new and reproducible insight*
2) *An introduction to the tools that will allow you to execute on a data analytic strategy, from raw data in a database to a completed report with interactive graphics*
3) *Giving you plenty of hands on practice so you can learn the techniques for yourself.*

*Regression Models represents a both fundamental and foundational component of the series, and it presents the single most practical data analysis toolset. Using only a bare minimum of mathematics,*

*we will attempt to provide you with the fundamentals for the application and practice of regression. We are excited about the opportunity to attempt to scale Data Science education. We intend for the courses to be self-contained, fast-paced, and interactive, and we intend to run them frequently to give people with busy schedules the opportunity to work on material at their own pace.*

*Brian Caffo and the Data Science Track Team*"

### Data Science Specialization Community Site

#### The site is created using GitHub Pages

In addition, Johns Hopkins has **a site on Statistical Methods and Applications for Research in Technology** that Dr. Caffo helps manage.

**Reminder to commit (01) delete this line *AFTER* committing**

## Least Squares and Linear Regression

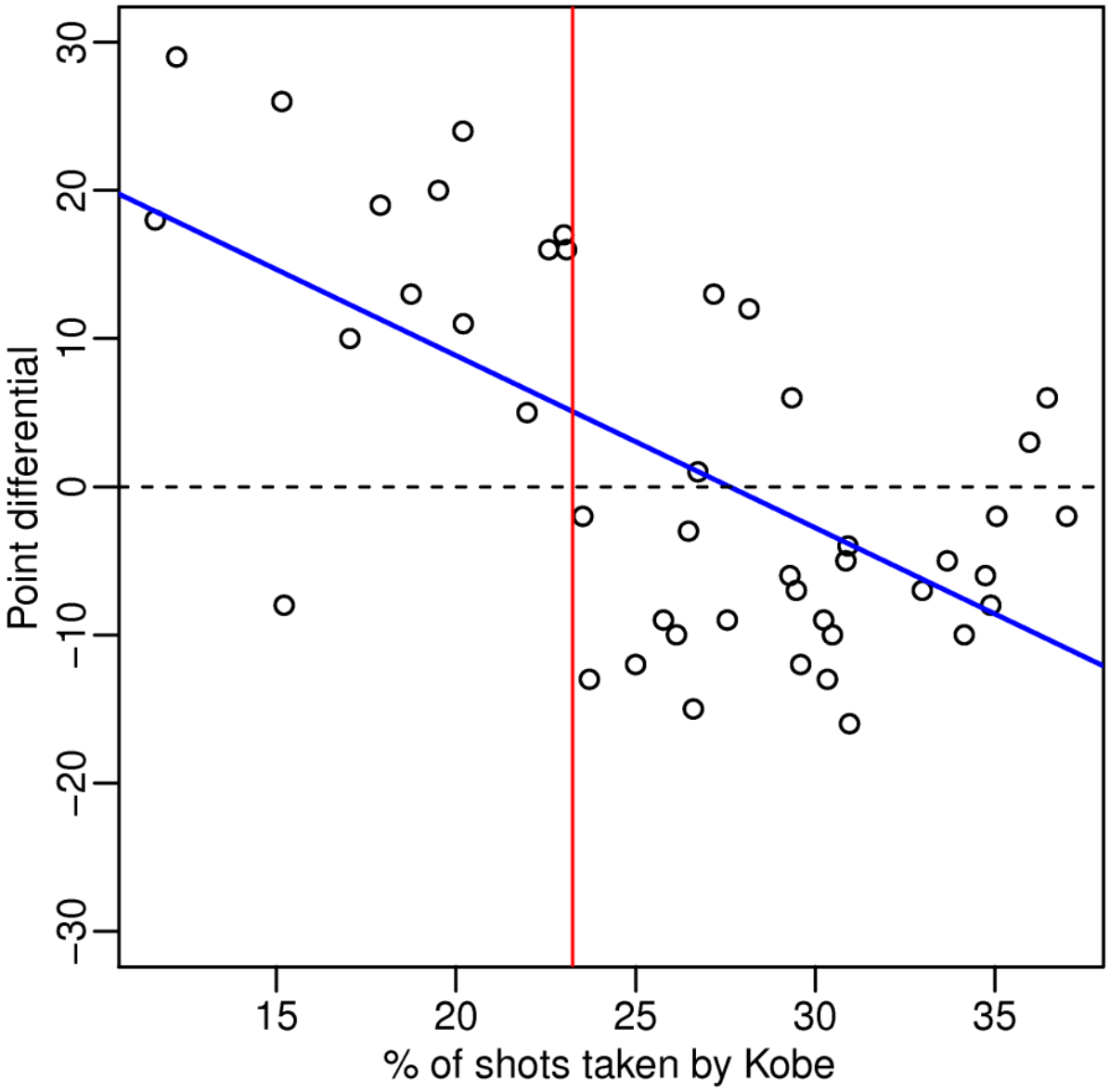
### Regression

#### Introduction to Regression

- The simplicity and intrepretability offered by regression models should make them a first tool of choice for any practical problem.
- First discovered by **Francis Galton** who coined most of the terminology we use today.

#### Relevant Simply Statistics Post

**Simply Statistics is a blog by Jeff Leek, Roger Peng and Rafael Irizarry, who wrote this post**




- "Data supports claim that if Kobe stops ball hogging the Lakers will win more"

- "Linear regression suggests that an increase of 1% in percent of shots taken by Kobe results in a drop of 1.16 (+/- 0.22) in score differential."
  + Standard error given as "+/- 0.22"

**Questions for this Class**

In reference to Galton's parent/children height data, which can be accessed from the `galton` dataset in the `UsingR` package.
Consider trying to answer the following kinds of questions:
* To use the parents' heights to predict childrens' heights.

* To try to find a parsimonious (explain the data), easily described mean relationship between parent and children's heights.
* To investigate the variation in childrens' heights that appears unrelated to parents' heights (residual variation).
* To quantify what impact genotype information has beyond parental height in explaining child height.
* To figure out how/whether and what assumptions are needed to generalize findings beyond the data in question.
* Why do children of very tall parents tend to be tall, but a little shorter than their parents and why children of very short parents tend to be short, but a little taller than their parents? (This is a famous question called "Regression to the mean".)

### Introduction to Basic Least Squares

- Let's look at the data first used by Francis Galton in 1885.

- Galton was a statistician who invented the term and concepts of regression and correlation, founded the journal Biometrika, and was the cousin of Charles Darwin.

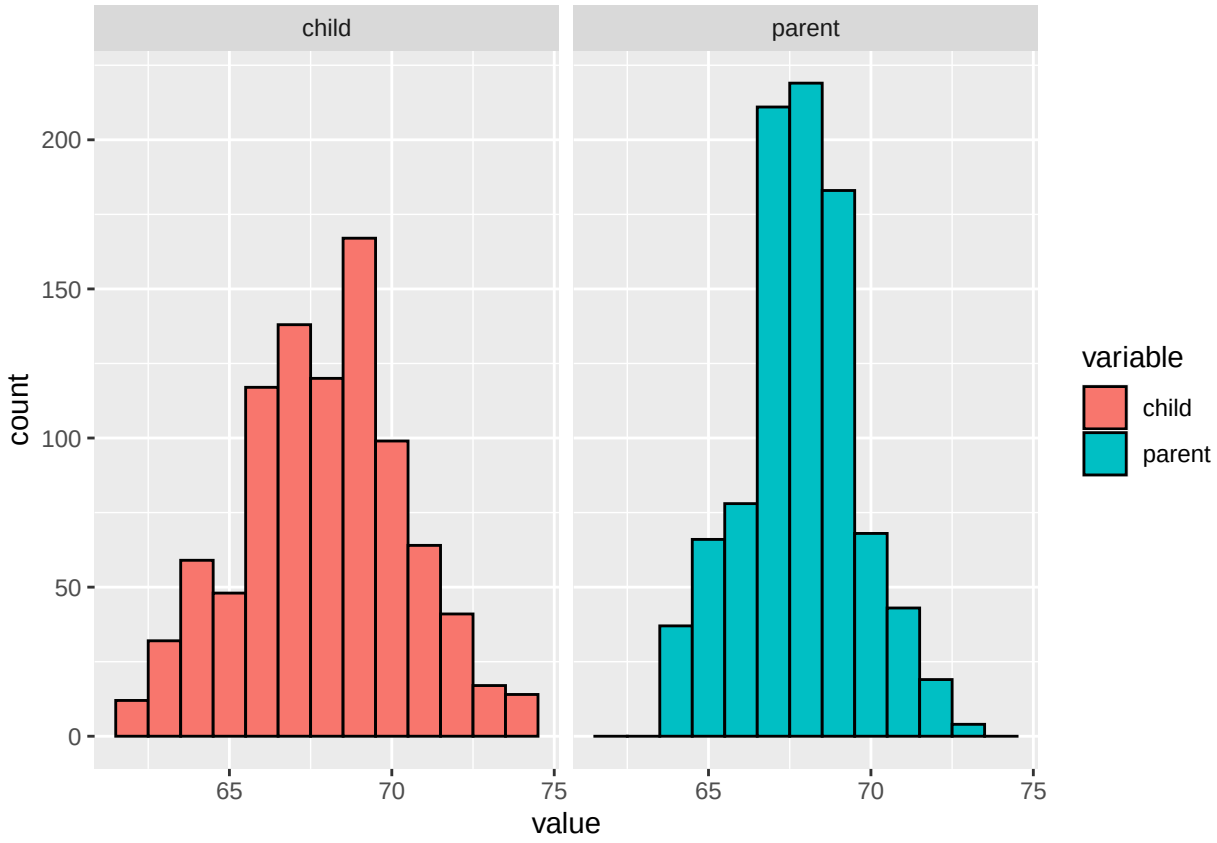
- Let's look at the marginal (parents disregarding children and children disregarding parents) distributions first.
  + Parent distribution is all heterosecual couples.
  + Correction for gender via multiplying female heights by 1.08.
  + Overplotting is an issue from discretization.

```
library(UsingR); data(galton); library(reshape2); library(tidyverse)
```

```
long <- melt(galton)
```

```
## No id variables; using all as measure variables
```

```
plot <- ggplot(long, aes(x = value, fill = variable)) +
        geom_histogram(colour = "#000000", binwidth = 1)
plot + facet_grid(.~variable)
```



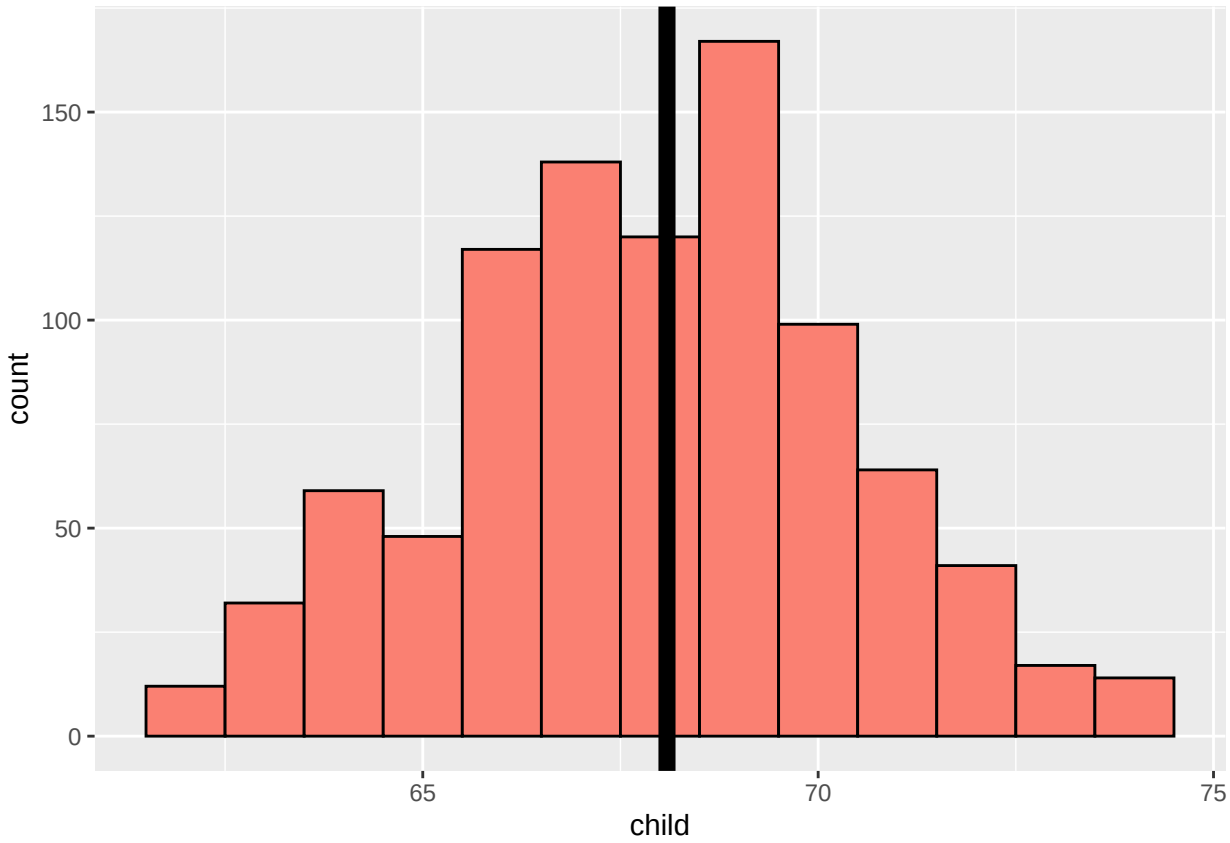

**Finding the Middle via Least Squares**

- Consider only the children's heights
  + How could one describe the "middle"?
  + One definition, let $Y_i$ be the height of child $i$ for $i = 1, ..., n = 928$, then define the middle as the value of $\mu$ that minimizes
  $\sum_{i=1}^{n}(Y_i - \mu)^2$

- This is the physical center of mass of the histogram.

- The result of this is that $\mu = \bar{Y}$

```
ggplot(galton, aes(x = child)) +
       geom_histogram(fill = "salmon", colour = "#000000", binwidth = 1) +
       geom_vline(xintercept = mean(galton$child), size = 3)
```




- The above plot of child heights has a mean of 68.0884698

### Technical Details

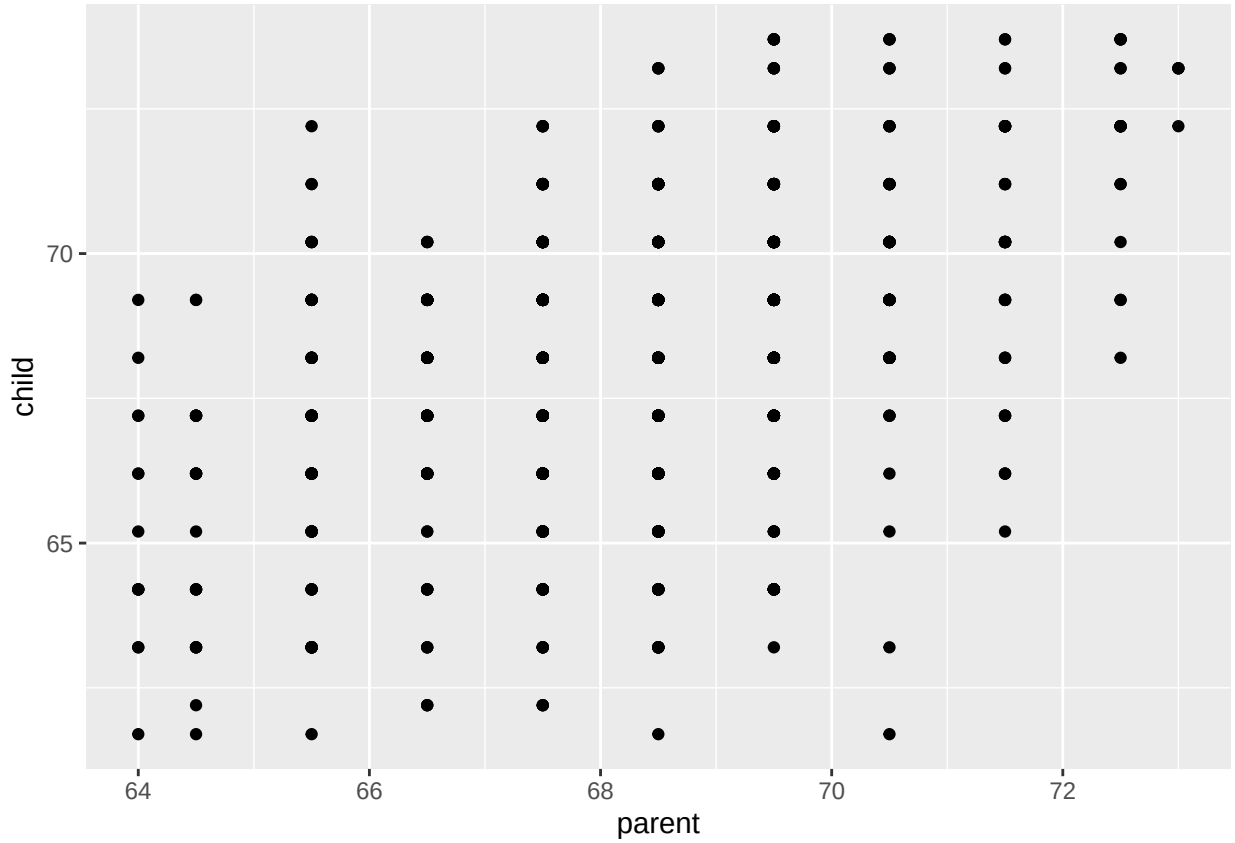
Proof that $\bar{Y}$ is the minimizer for $\sum_{i=1}^{n}(Y_i - \mu)^2$
$\sum_{i=1}^{n}(Y_i - \mu)^2 = \sum_{i=1}^{n}(Y_i - \bar{Y} + \bar{Y} - \mu)^2$
$= \sum_{i=1}^{n}(Y_i - \bar{Y}^2 + 2\sum_{i=1}^{n}(Y_i - \bar{Y})(\bar{Y} - \mu) + \sum_{i=1}^{n}(\bar{Y} - \mu)^2$
$= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu)\sum_{i=1}^{n}(Y_i - \bar{Y}) + \sum_{i=1}^{n}(\bar{Y} - \mu)^2$
$= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu)(\sum_{i=1}^{n} Y_i - n\bar{Y}) + \sum_{i=1}^{n}(\bar{Y} - \mu)^2$
$= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + 0 + \sum_{i=1}^{n}(\bar{Y} - \mu)^2$
$\geq \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Therefore, $\sum_{i=1}^{n}(Y_i - \mu)^2$ is minimized when $\bar{Y} = \mu$

### Introductory Data Example

### Comparing Childrens' Heights and Their Parents' Heights

```
ggplot(galton, aes(x = parent, y = child)) + geom_point()
```

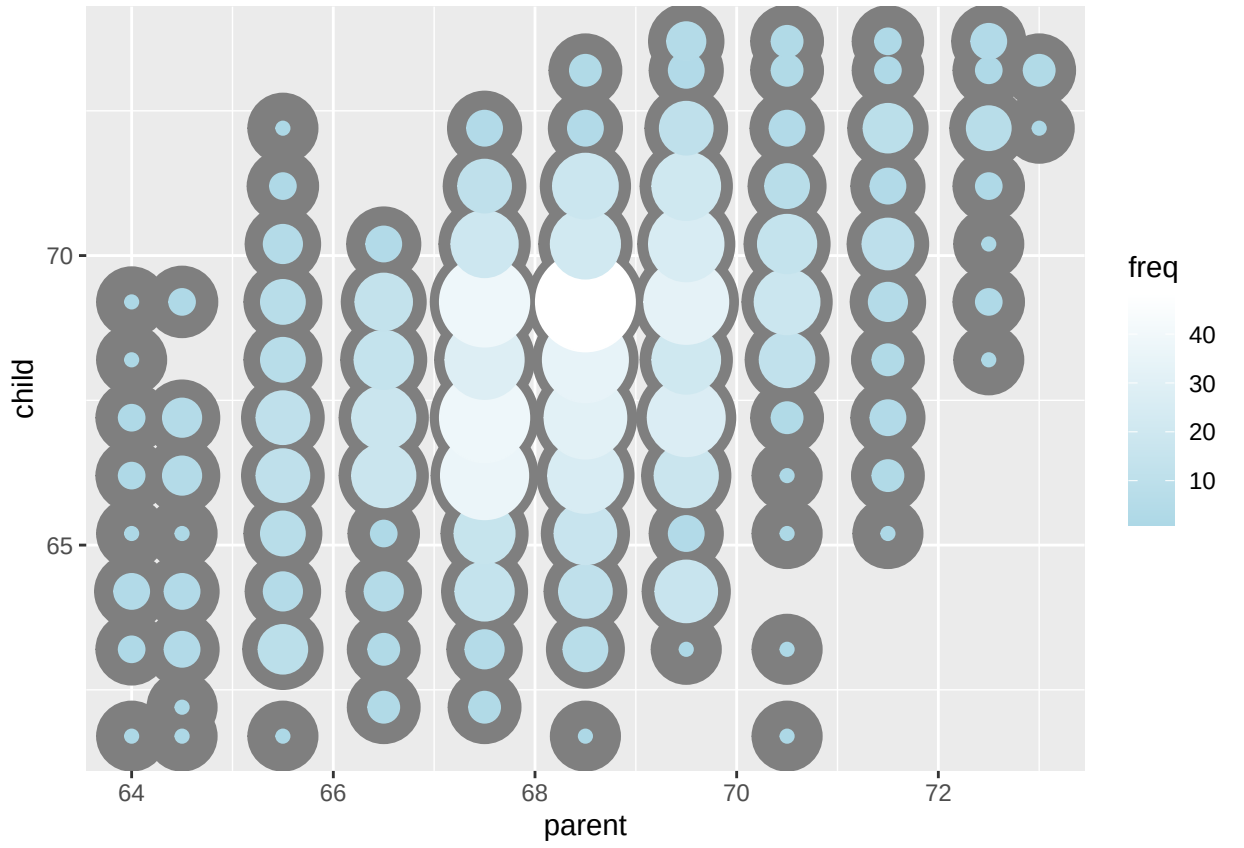



- These points are overplotted, there are multiple overlays at each point, so let's make a better plot

```
freqData <- as.data.frame(table(galton$child, galton$parent))
names(freqData) <- c("child", "parent", "freq")
freqData$child <- as.numeric(as.character(freqData$child))
freqData$parent <- as.numeric(as.character(freqData$parent))
plot <- ggplot(filter(freqData, freq > 0), aes(x = parent, y = child)) +
        scale_size(range = c(2, 20), guide = "none") +
        geom_point(colour = "grey50",
                   aes(size = freq + 20, show_guide = FALSE)) +
        geom_point(aes(colour = freq, size = freq)) +
        scale_colour_gradient(low = "lightblue", high = "#FFFFFF")
```

```
## Warning: Ignoring unknown aesthetics: show_guide
```

```
plot
```




**Regression Through the Origin**

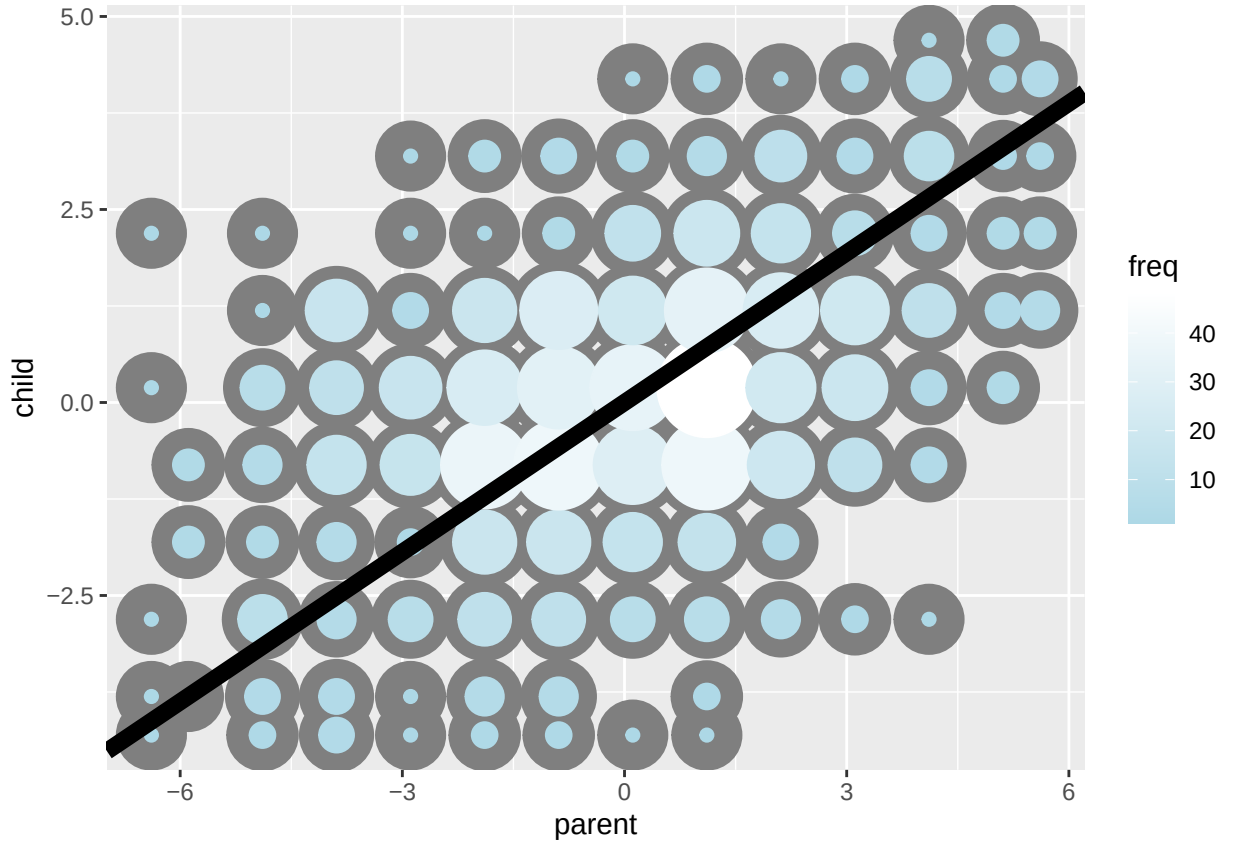
- Suppose that $X_i$ are the parents' heights
- Consider picking the slope $\beta$ that minimizes $\sum_{i=1}^{n}(Y_i - X_i\beta)^2$
- This is exactly using the orgin as a pivot point picking the line that minimizes the sum of squared vertical distances of the points to the line
- Subtract the means so that the orgin is the mean of the parent and children's heights
  + A plot with a regression line going through true (0,0) often doesn't make sense, so subtracting the means realigns the orgin to be in the middle of the data

```
freqData <- as.data.frame(table(galton$parent - mean(galton$parent),
                                galton$child - mean(galton$child)))
names(freqData) <- c("child", "parent", "freq")
freqData$child <- as.numeric(as.character(freqData$child))
freqData$parent <- as.numeric(as.character(freqData$parent))
plot <- ggplot(filter(freqData, freq > 0), aes(x = parent, y = child)) +
        scale_size(range = c(2, 20), guide = "none") +
        geom_point(colour = "grey50",
                   aes(size = freq + 20)) +
```

```
        geom_point(aes(colour = freq, size = freq)) +
        scale_colour_gradient(low = "lightblue", high = "#FFFFFF") +
        geom_abline(intercept = 0,

                slope = lm(
                        I(child - mean(child)) ~
                                I(parent - mean(parent)) - 1,
                        data = galton)$coeff,

                    size = 3)
plot
```



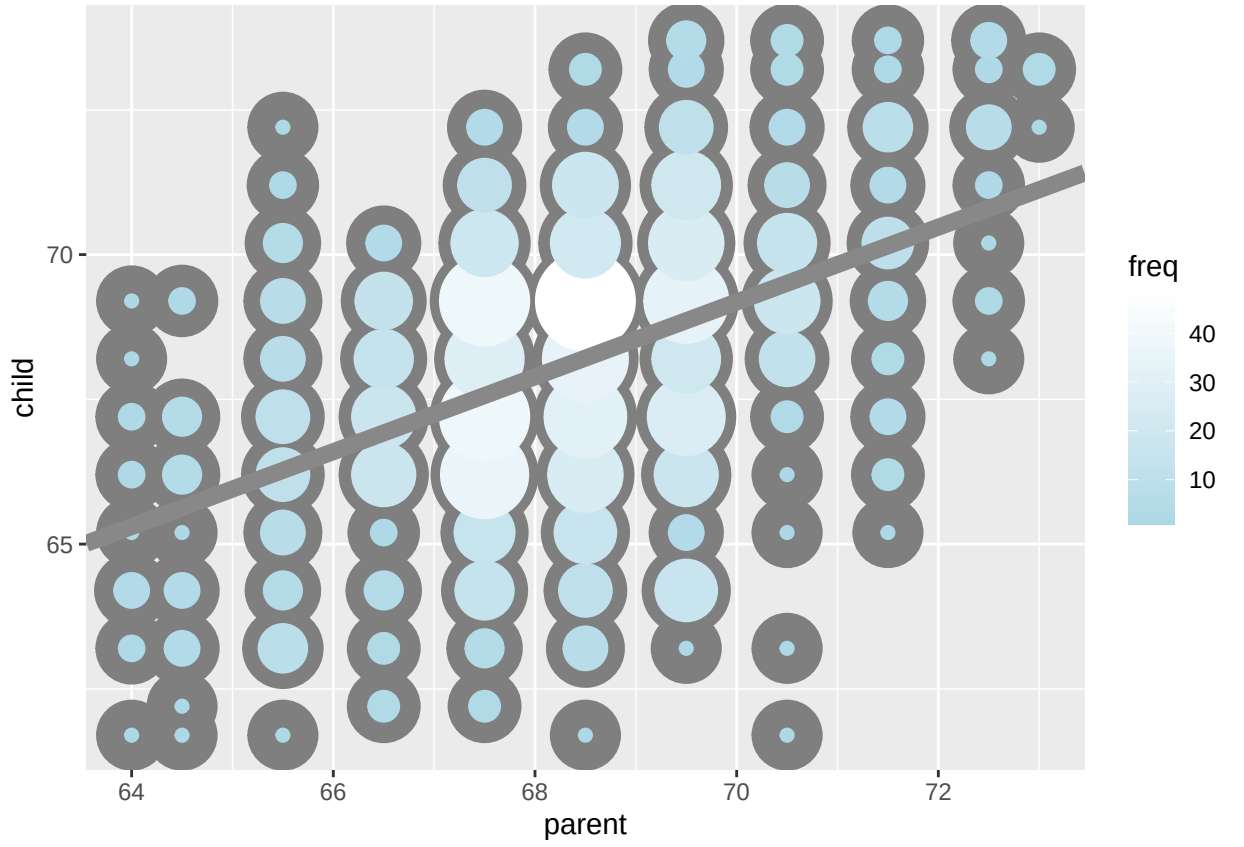

- In the next few lectures we'll talk about why this is the solution

```
lm(I(child - mean(child)) ~ I(parent - mean(parent)) - 1, data = galton)
```

```
##
## Call:
## lm(formula = I(child - mean(child)) ~ I(parent - mean(parent)) -
##     1, data = galton)
##
## Coefficients:
## I(parent - mean(parent))
##                   0.6463
```

- The `I` function just ignores the intercept, since we already adjusted for that
- We can also fit a line to an un-adjusted model

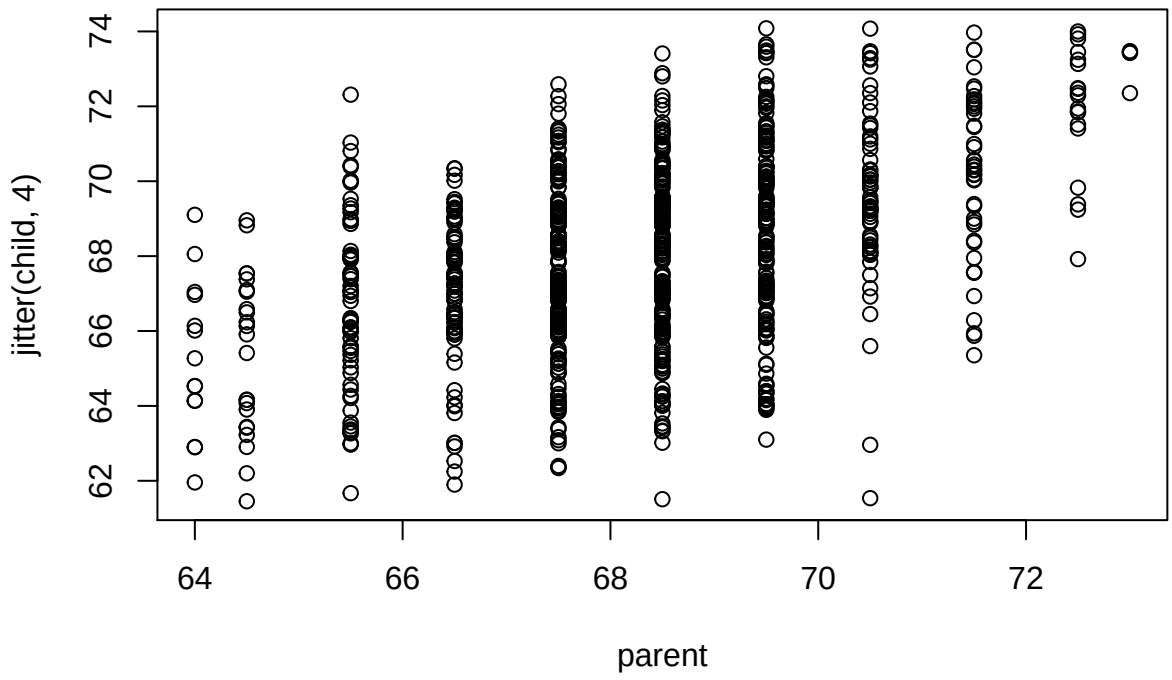
```
freqData <- as.data.frame(table(galton$child, galton$parent))
names(freqData) <- c("child", "parent", "freq")
freqData$child <- as.numeric(as.character(freqData$child))
freqData$parent <- as.numeric(as.character(freqData$parent))
plot <- ggplot(filter(freqData, freq > 0), aes(x = parent, y = child)) +
        scale_size(range = c(2, 20), guide = "none" ) +
        geom_point(colour = "grey50", aes(size = freq + 20)) +
        geom_point(aes(colour = freq, size = freq)) +
        scale_colour_gradient(low = "lightblue", high = "#FFFFFF")
lm1 <- lm(galton$child ~ galton$parent)
plot + geom_abline(intercept = coef(lm1)[1], slope = coef(lm1)[2],
                   size = 3, colour = "#888888")
```




### Lesson with `swirl()`: Introduction

- Another way we could have gotten past overlapping plot points is to use the `jitter` function

```
plot(jitter(child,4) ~ parent, galton)
```




## Linear Least Squares

- Also called **Ordinary Least Squares (OLS)**; it fits a line through some data.

### Notation and Background

### Notation

- The empirical mean is defined as
  $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- If we subtract the mean from data points, we get data that has a mean of 0. That is, if we define:
  $\tilde{X}_i = X_i - \bar{X}$.
  + The mean of $\tilde{X}_i$ is 0

- This process is called "**centering**" the random variables

- Recall from the previous lecture that the mean is the elast squares solution for minimizing $\sum_{i=1}^{n}(X_i - \mu)^2$

**The Emprical Standard Deviation adn Variance**

- Define the empirical variance as
  $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)$

- The empirical standard deviation is defined as $S = \sqrt{S^2}$.
  + Notice that the standard deviation has the same units as the data.

- The data defined by $\frac{X_i}{s}$ have an empirical standard deviation of 1. + This is called "**scaling**" the data.

**Normalization**

- The data defined by
  $Z_i = \frac{X_i - \bar{X}}{s}$
  have an empirical mean of 0 and an empirical standard deviation of 1.

- The process of centering then scaling the data is called "**normalizing**" the data.

- Normalized data are centered at 0 and have units equal to standard deviations of the original data.

- For example, a value of 2 from normalized data is saying that data point was two standard deviations larger than the mean.

**The Empirical Covariance**

- Consider now when we have pairs of data, $(X_i, Y_i)$

- Their empirical covariance is
  $Cov(X,Y) = \frac{1}{n-1}\sum_{i=1} n(X_i - \bar{X})(Y_i - \bar{Y})$
  $= \frac{1}{n-1}(\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y})$

- The correlation is defined as
  $Cor(X,Y) = \frac{Cov(X,Y)}{S_x S_y}$
  + Where $S_x$ and $S_y$ are the estimates of standard deviaitons for the X observations and Y observations, respectively.

**Some Facts About Correlation**

- $Cor(X,Y) = Cor(Y,X)$

- $-1 \leq Cor(X,Y) \leq 1$

- $Cor(X,Y) = 1$ and $Cor(X,Y) = -1$ only when the $X$ or $Y$ observations fall perfectly on a positive or negative sloped line, repectively.

- $Cor(X, Y)$ measures the strength of the linear relationship between the $X$ and $Y$ data, with stronger relationships as $Cor(X, Y)$ heads towards either -1 or 1 {
- $Cor(X, Y) = 0$ implies no linear relationship

### Linear Least Squares

### Fitting the Best Line

- Let $Y_i$ be the $i^{th}$ child's height and $X_i$ be the $i^{th}$ (average over the pair of) parents' heights.

- Consider finding the best line
  + Child's Height = $\beta_0$ + Parent's Height * $\beta_1$
  $\sum_{i=1}^{n} Y_i - (\beta_0 + \beta_1 X_i)^2$
- the least squares model fit to the line $Y = \beta_0 + \beta_1 X$ through the data pairs $(X_i, Y_i)$ with $Y_i$ as the outcome obtains the line $Y = \hat{\beta}_0 + \hat{\beta}_1 X$ where
  $\hat{\beta}_1 = Cor(Y, X)\frac{Sd(Y)}{Sd(X)}$
  $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

- $\hat{\beta}_1$ has the units of $Y/X$, $\hat{\beta}_0$ has the units of $Y$.

- The line passes through the point $(\bar{X}, \bar{Y})$

- The slope of the regression line with $X$ as the outcome and $Y$ as the predictor is $\frac{Cor(Y,X)Sd(X)}{Sd(Y)}$

- The slope si the same one you would get if you centered the data, $(X_i - \bar{X}, Y_i - \bar{Y})$, and made a regression through the orgin

- If you normalized the data, $(\frac{X_i - \bar{X}}{Sd(X)}, \frac{Y_i - \bar{Y}}{Sd(Y)})$, the slope is $Cor(Y, X)$

### Linear Least Squares Coding Example

```
y <- galton$child
x <- galton$parent
beta1 <- cor(y,x) * sd(y) / sd(x)
beta0 <- mean(y) - beta1 * mean(x)

#Showing the computations by hand are the same as coef from lm function
rbind(c(beta0, beta1), coef(lm(y~x)))
```

```
##      (Intercept)         x
## [1,]    23.94153 0.6462906
## [2,]    23.94153 0.6462906
```

- `lm` stands for *linear model*

```
#The slope is the same in centered data
yc <- y - mean(y)
xc <- x - mean(x)
beta1 <- sum(yc * xc) / sum(xc^2)
c(beta1, coef(lm(y ~ x))[2])
```

```
##                   x
## 0.6462906 0.6462906
```

```
lm(yc ~ xc - 1)$coef #minus 1 gets rid of intercept
```

```
##        xc
## 0.6462906
```

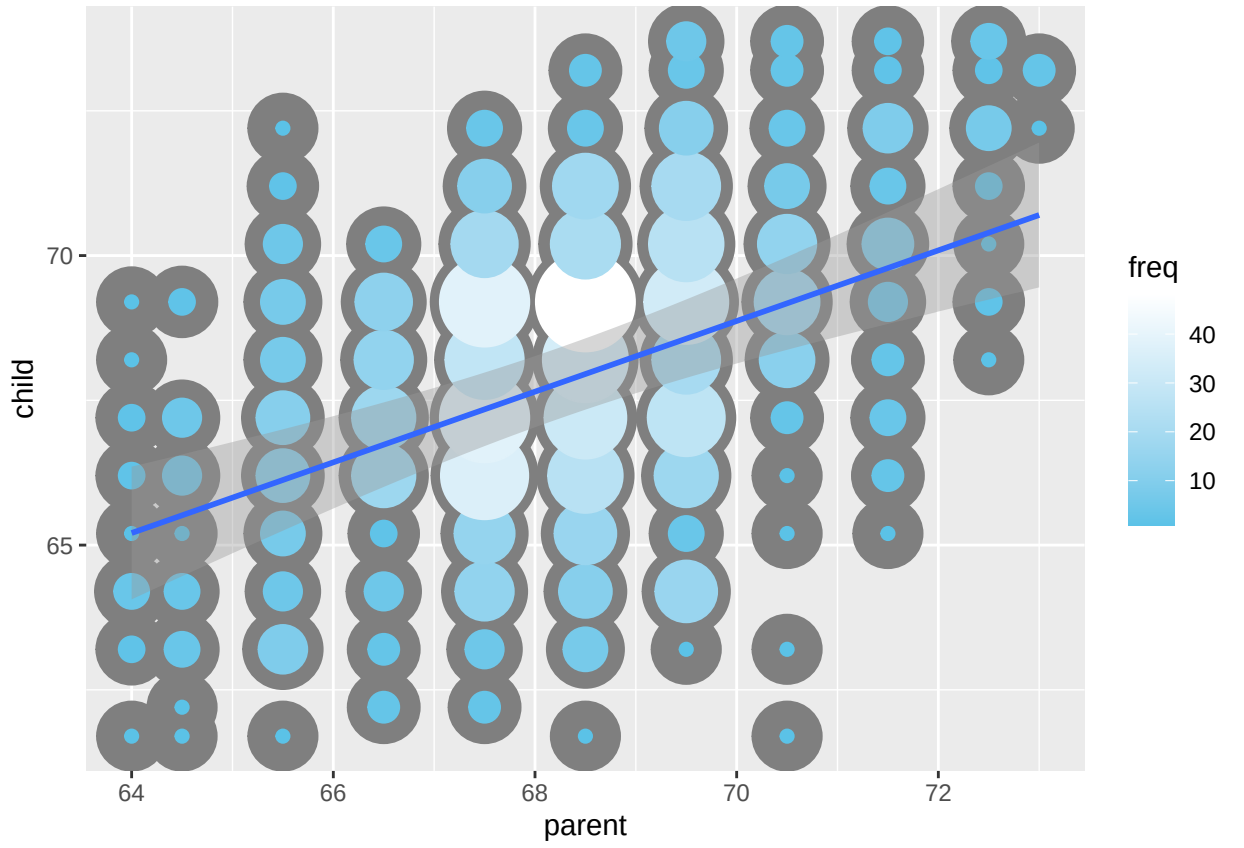
```
#Normalizing variables results in the slope being the correlation
yn <- (y - mean(y))/sd(y)
xn <- (x - mean(x))/sd(x)
results <- cbind(cor(y,x), lm(yn ~ xn)$coef[2], cor(yn, xn))
colnames(results) <- c("cor(y,x)", "Slope(yn ~ xn)", "cor(yn, xn)")
results
```

```
##     cor(y,x) Slope(yn ~ xn) cor(yn, xn)
## xn 0.4587624      0.4587624   0.4587624
```

### Adding a Linear Regression to ggplot

```
plot <- ggplot(filter(freqData, freq > 0), aes(parent, child)) +
        scale_size(range = c(2, 20), guide = "none") +
        geom_point(colour = "grey50", aes(size = freq + 20)) +
        geom_point(aes(colour = freq, size = freq)) +
        scale_colour_gradient(low = "#5BC2E7", high = "#FFFFFF")

#Adding smoother
#y ~ x is assumed if not given
plot + geom_smooth(method = "lm", formula = y ~ x)
```




- A confidence interval is also given around the line automaticly

**Technical Details**

**Brian Caffo discusses the proof for least squares regression beta_1 value in this video**

**Lesson with `swirl()`: Least Squares Estimation**

(No new content)

**Reminder to commit (03) delete this line *AFTER* committing**

## Regression to the Mean

**Regression to the Mean**

**Lesson with `swirl()`: Residuals**

**Reminder to commit (04) delete this line *AFTER* committing**

**Quiz 1**

**Reminder to commit (S1) delete this line *AFTER* committing**

## Linear Regression & Multivariable Regression

### Statistical Linear Regression Models

**Statistical Linear Regression Models**

**Interpreting Coefficients**

**Linear Regression for Prediction**

**Lesson with `swirl()`: Introduction to Multivariable Regression**

**Reminder to commit (05) delete this line *AFTER* committing**

### Residuals

**Residuals**

**Residuals, Coding Example**

**Residual Variance**

**Lesson with `swirl()`: Residual Variation**

**Reminder to commit (06) delete this line *AFTER* committing**

### Inference in Regression

**Inference in Regression**

**Coding Example**

**Prediction**

**Lesson with `swirl()`: MultiVar Examples**

**Reminder to commit (07) delete this line *AFTER* committing**

### Quiz 2

**Reminder to commit (S2) delete this line *AFTER* committing**

## Multivariable Regression, Residuals, & Diagnostics

### Multivariable Regression

**Multivariable Regression Part 1**

**Multivariable Regression Part 2**

**Multivariable Regression Continued**

**Reminder to commit (08) delete this line *AFTER* committing**

### Multivariable Regression Tips and Tricks

**Multivariable Regression Examples Part 1**

**Multivariable Regression Examples Part 2**

**Multivariable Regression Examples Part 3**

**Multivariable Regression Examples Part 4**

**Lesson with `swirl()`: MultiVar Examples2**

**Lesson with `swirl()`: MultiVar Examples3**

**Reminder to commit (09) delete this line *AFTER* committing**

### Adjustment

**Adjustment Examples**

**Reminder to commit (10) delete this line *AFTER* committing**

### Residuals Again

**Residuals and Diagnostics Part 1**

**Residuals and Diagnostics Part 2**

**Residuals and Diagnostics Part 3**

**Lesson with `swirl()`: Residuals Diagnostics and Variation**

**Reminder to commit (11) delete this line *AFTER* committing**

### Model Selection

**Model Selection Part 1**

**Model Selection Part 2**

**Model Selection Part 3**

**Reminder to commit (12) delete this line *AFTER* committing**

### Practice Exercise in Regression Modeling

### Quiz 3

**Reminder to commit (S3) delete this line *AFTER* committing**

## Logistic Regression and Poisson Regression

### GLMs

### Logistic Regression

**Logistic Regression Part 1**

**Logistic Regression Part 2**

**Logistic Regression Part 3**

**Lesson with `swirl()`: Variance Inflation Factors**

**Lesson with `swirl()`: Overfitting and Underfitting**

**Reminder to commit (13) delete this line *AFTER* committing**

### Poisson Regression

**Poisson Regression Part 1**

**Poisson Regression Part 2**

**Lesson with `swirl()`: Binary Outcomes**

**Lesson with `swirl()`: Count Outcomes**

**Reminder to commit (14) delete this line *AFTER* committing**

### Hodgepodge

**Mishmash**

**Hodgepodge**

**Reminder to commit (15) delete this line *AFTER* committing**

### Quiz 4

**Reminder to commit (S4) delete this line *AFTER* committing**

## Course Project

**Reminder to commit (P1) delete this line *BEFORE* committing**