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

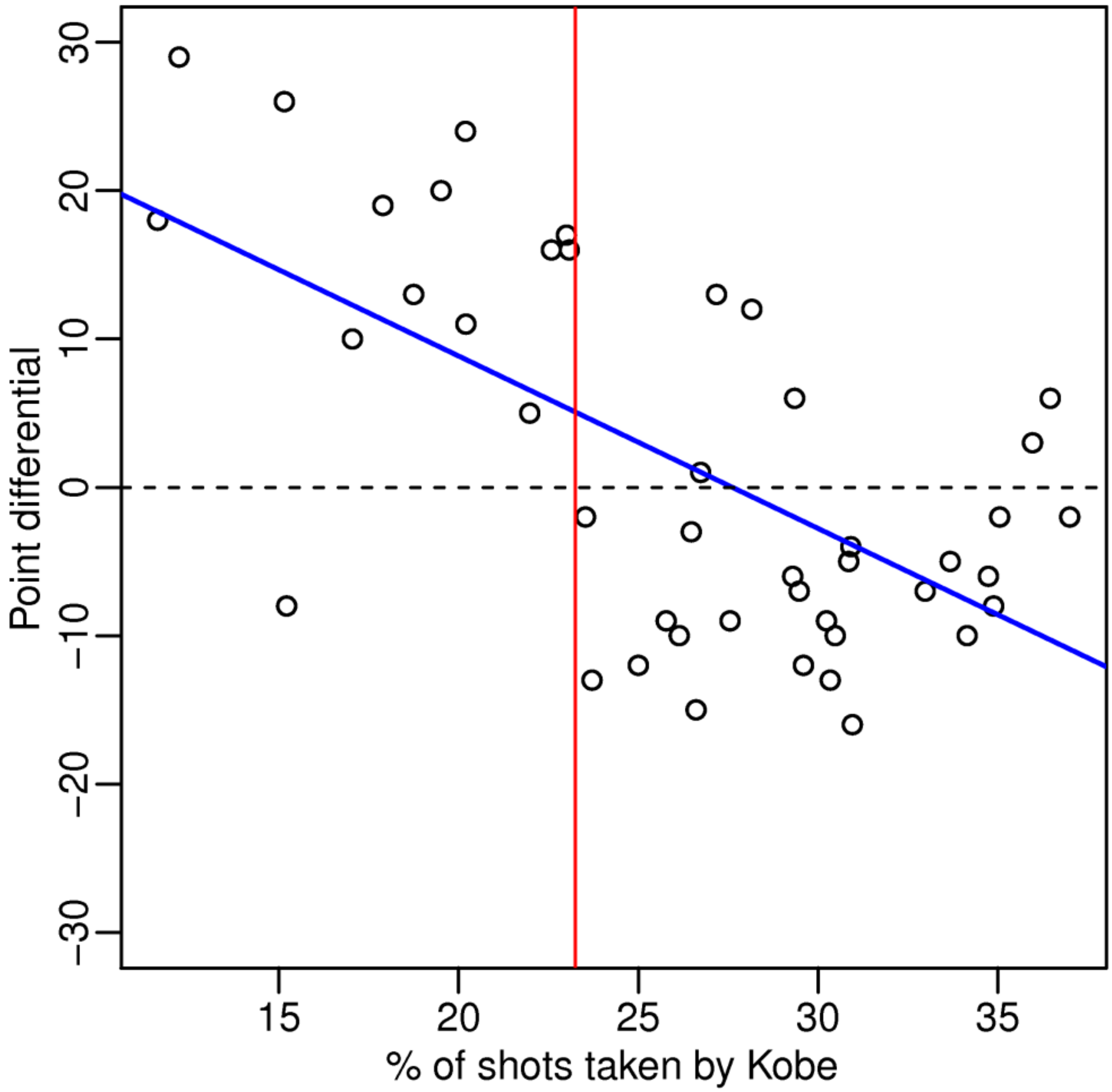
Introduction to Regression

- 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.

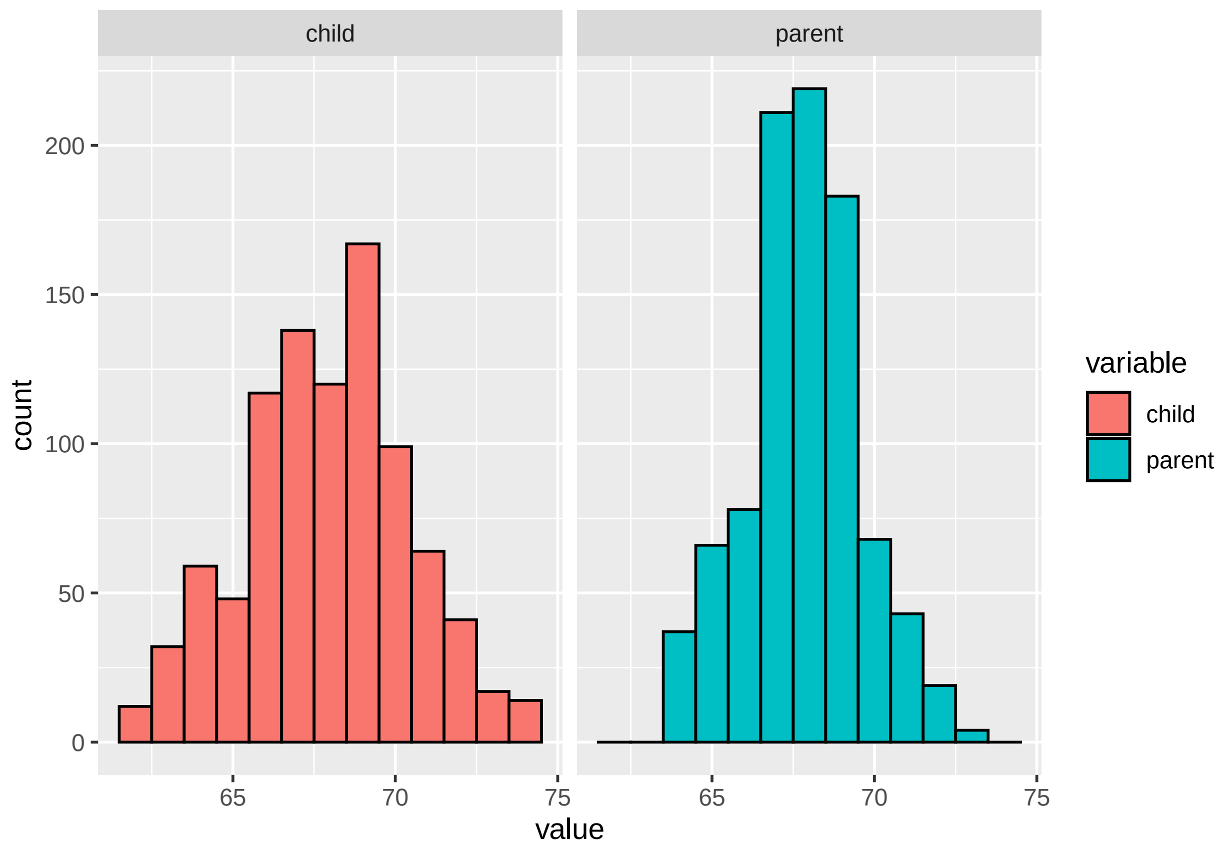
- Let's look at the marginal (parents disregarding children and children disregarding parents) distributions first.
  + Parent distribution is all heterosecual couples.
  + Correction for gender via multiplying female heights by 1.08.
  + Overplotting is an issue from discretization.

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

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

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

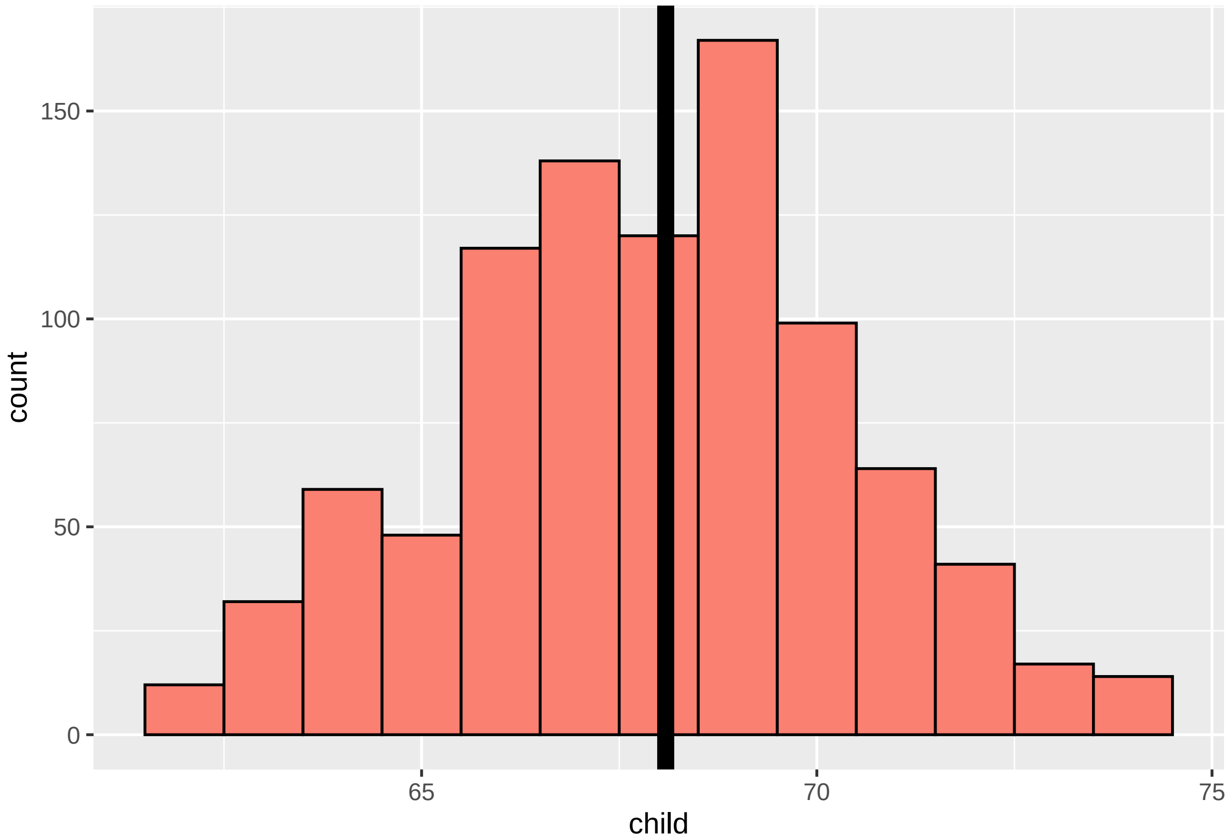
```
plot <- ggplot(long, aes(x = value, fill = variable)) +
        geom_histogram(colour = "#000000", binwidth = 1)
plot + facet_grid(.~variable)
```



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 μ 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

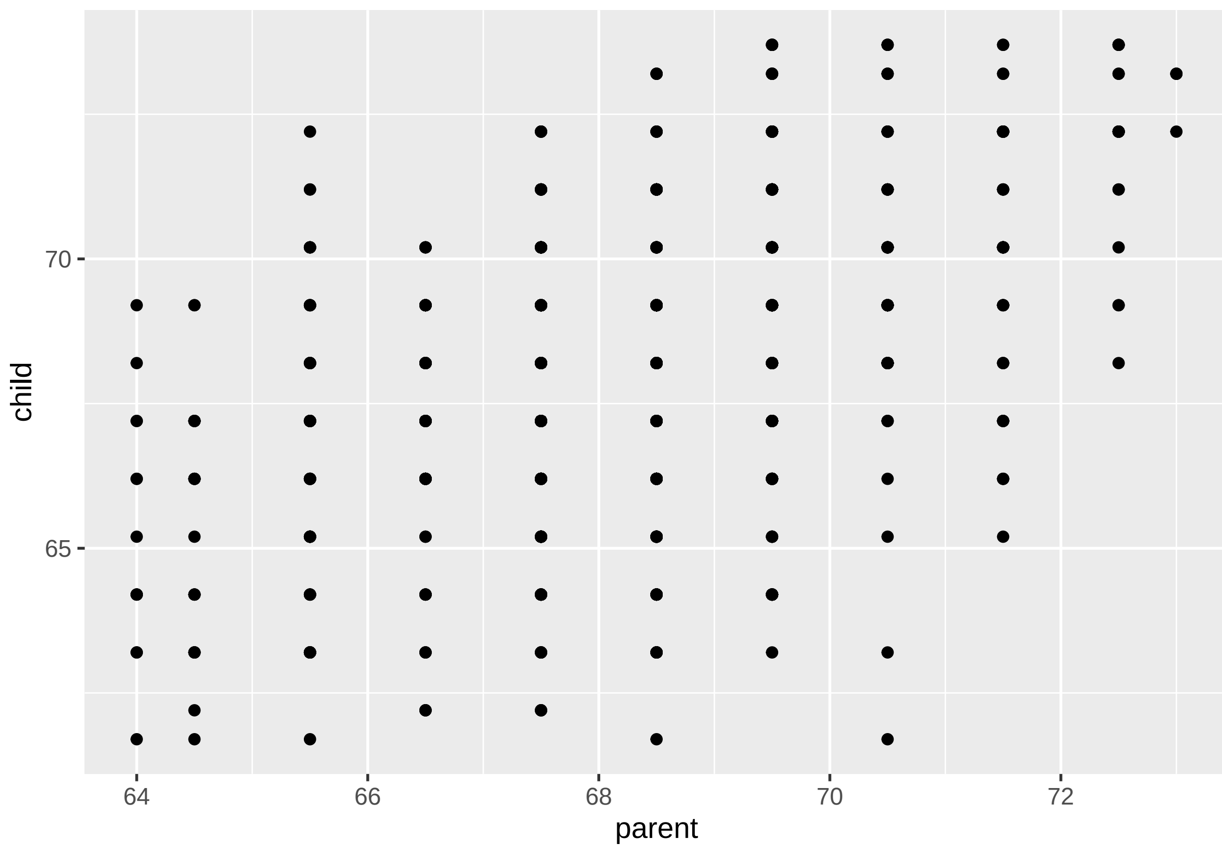
Proof that $\bar{Y}$ is the minimizer for $\sum_{i=1}^{n}(Y_i - \mu)^2$
$\sum_{i=1}^{n}(Y_i - \mu)^2 = \sum_{i=1}^{n}(Y_i - \bar{Y} + \bar{Y} - \mu)^2$
$= \sum_{i=1}^{n}(Y_i - \bar{Y}^2 + 2\sum_{i=1}^{n}(Y_i - \bar{Y})(\bar{Y} - \mu) + \sum_{i=1}^{n}(\bar{Y} - \mu)^2$
$= \sum_{i=1}^{n}(Y_i - \bar{Y})^2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

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()
```

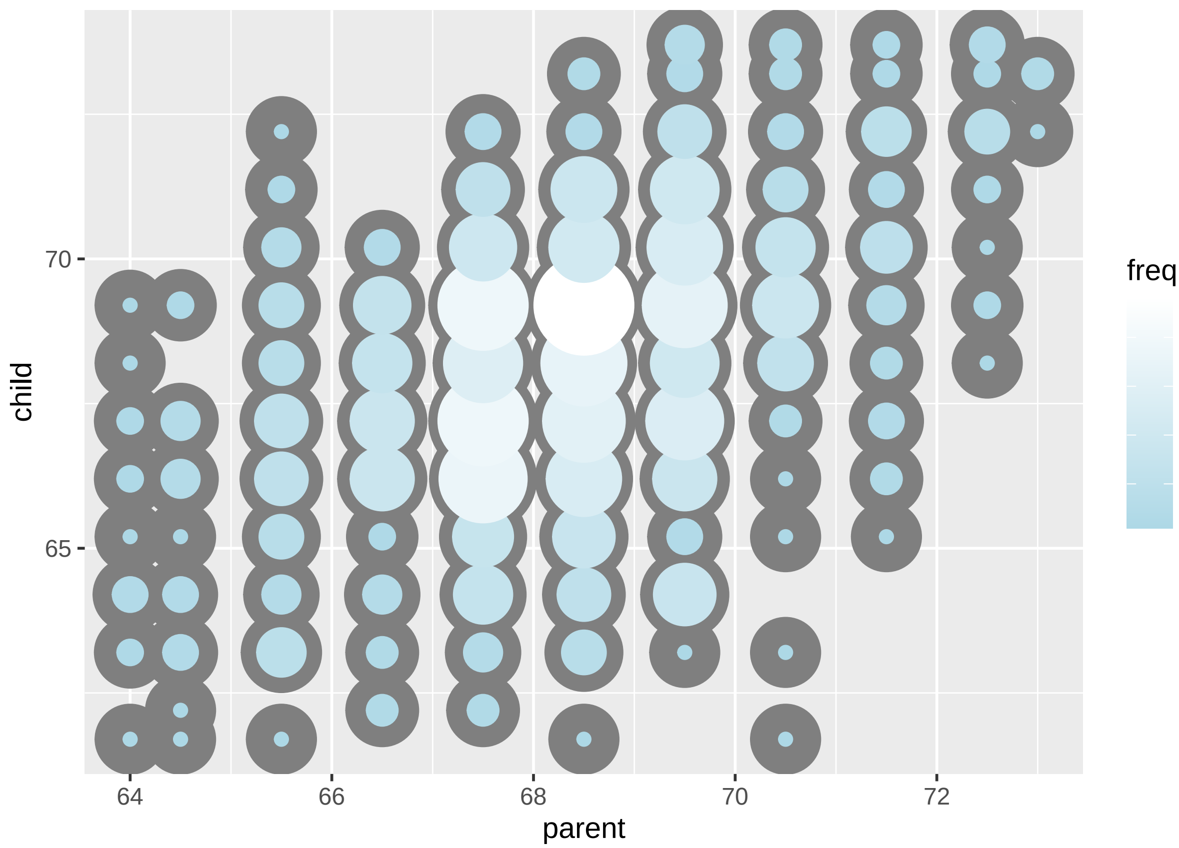


- These points are overplotted, there are multiple overlays at each point, so let's make a better plot

```
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

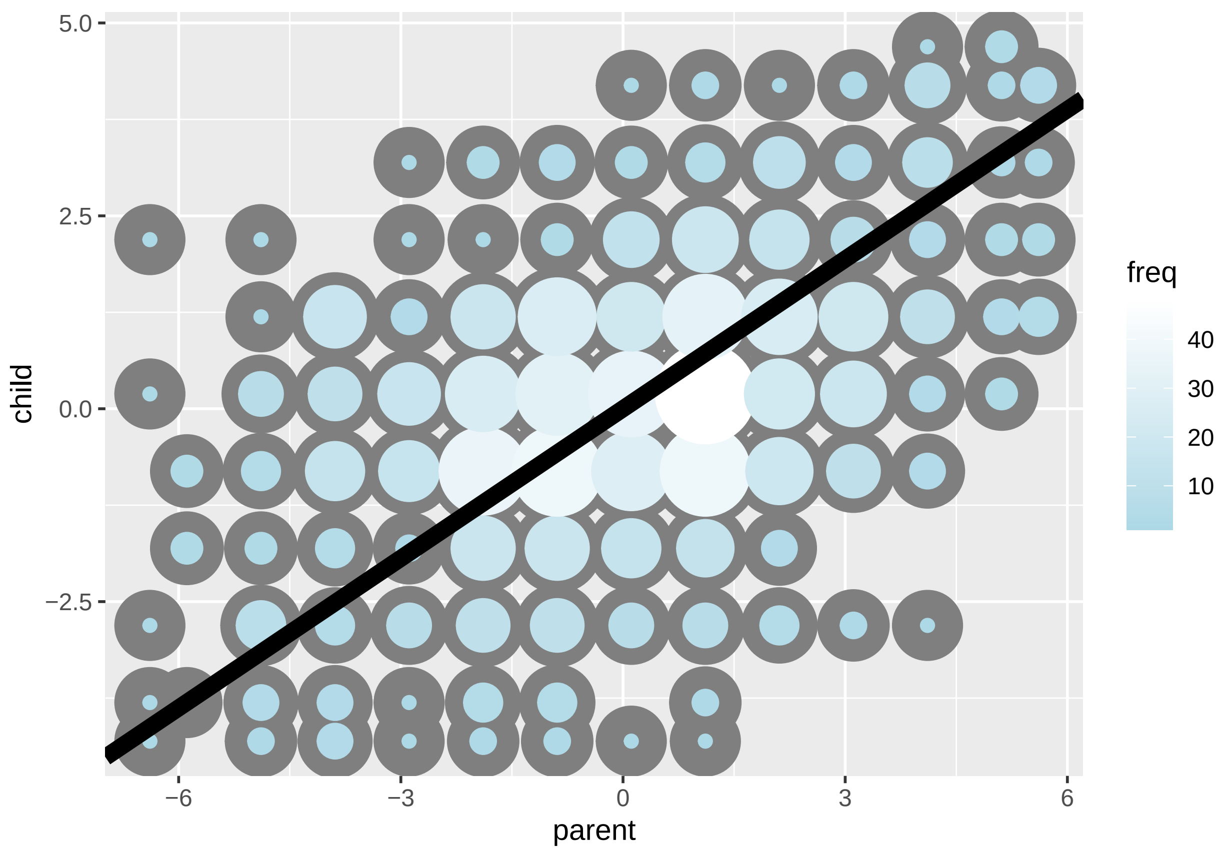
- Suppose that X_i are the parents' heights
- Consider picking the slope β that minimizes $\sum_{i=1}^{n}(Y_i - X_i\beta)^2$
- This is exactly using the orgin as a pivot point picking the line that minimizes the sum of squared vertical distances of the points to the line
- Subtract the means so that 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)) +
```

```r
        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

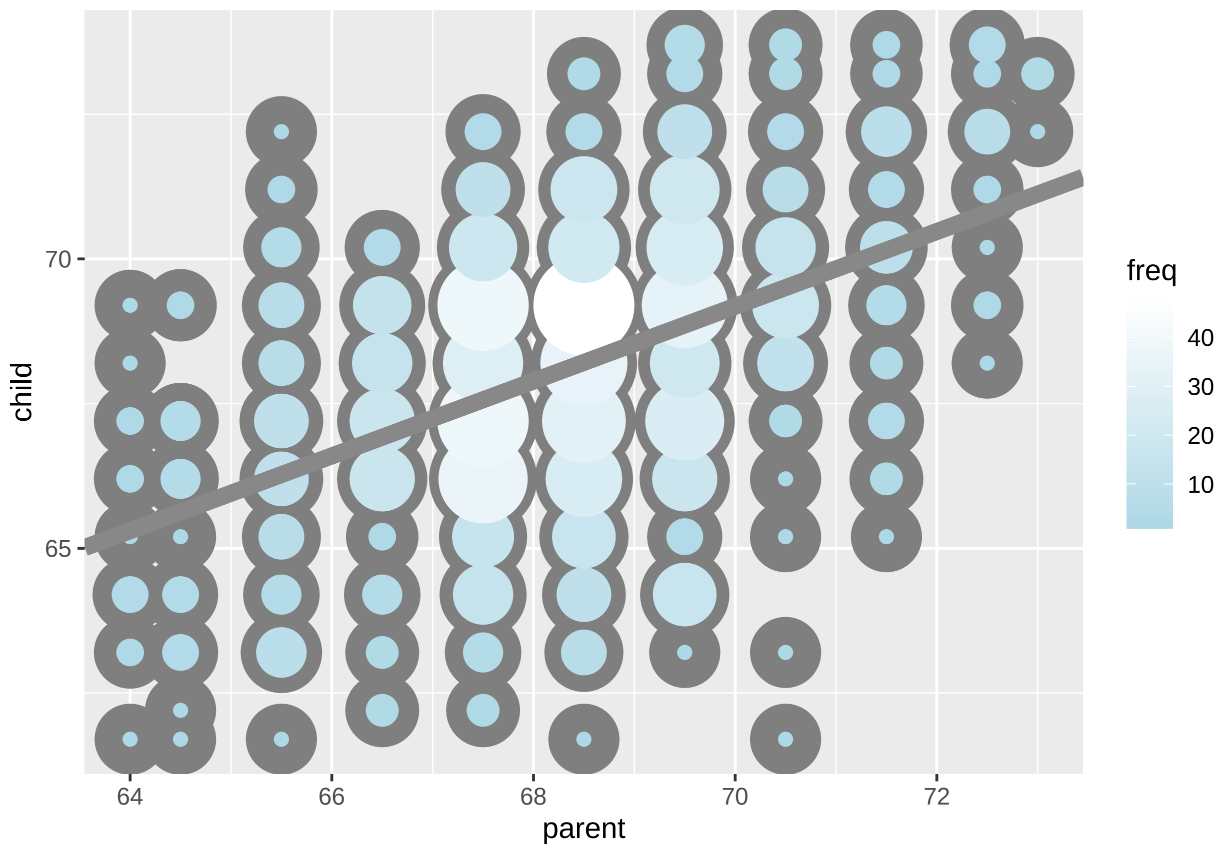
```r
lm(I(child - mean(child)) ~ I(parent - mean(parent)) - 1, data = galton)
```

```
##
## 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

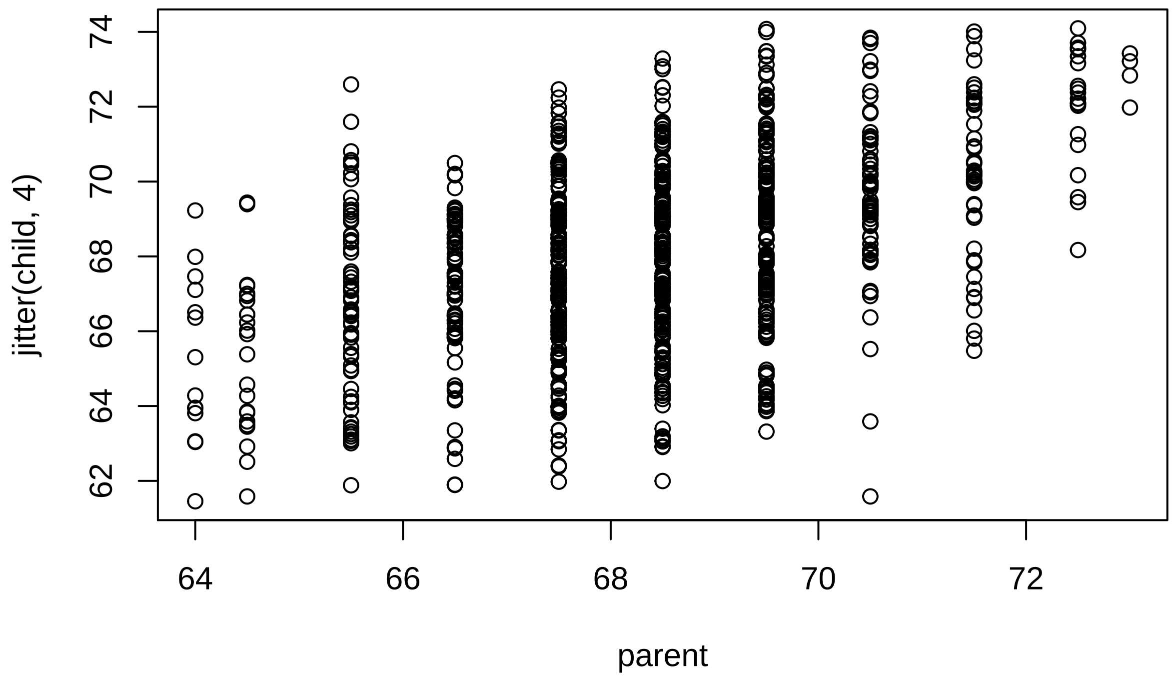
```
freqData <- as.data.frame(table(galton$child, galton$parent))
names(freqData) <- c("child", "parent", "freq")
freqData$child <- as.numeric(as.character(freqData$child))
freqData$parent <- as.numeric(as.character(freqData$parent))
plot <- ggplot(filter(freqData, freq > 0), aes(x = parent, y = child)) +
        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)
```



Reminder to commit (02) delete this line *AFTER* committing

Linear Least Squares

Notation and Background

Linear Least Squares

Linear Least Squares Coding Example

Technical Details

Lesson with `swirl()`: Least Squares Estimation

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