Galtons Data

Gary R Seamans

2018

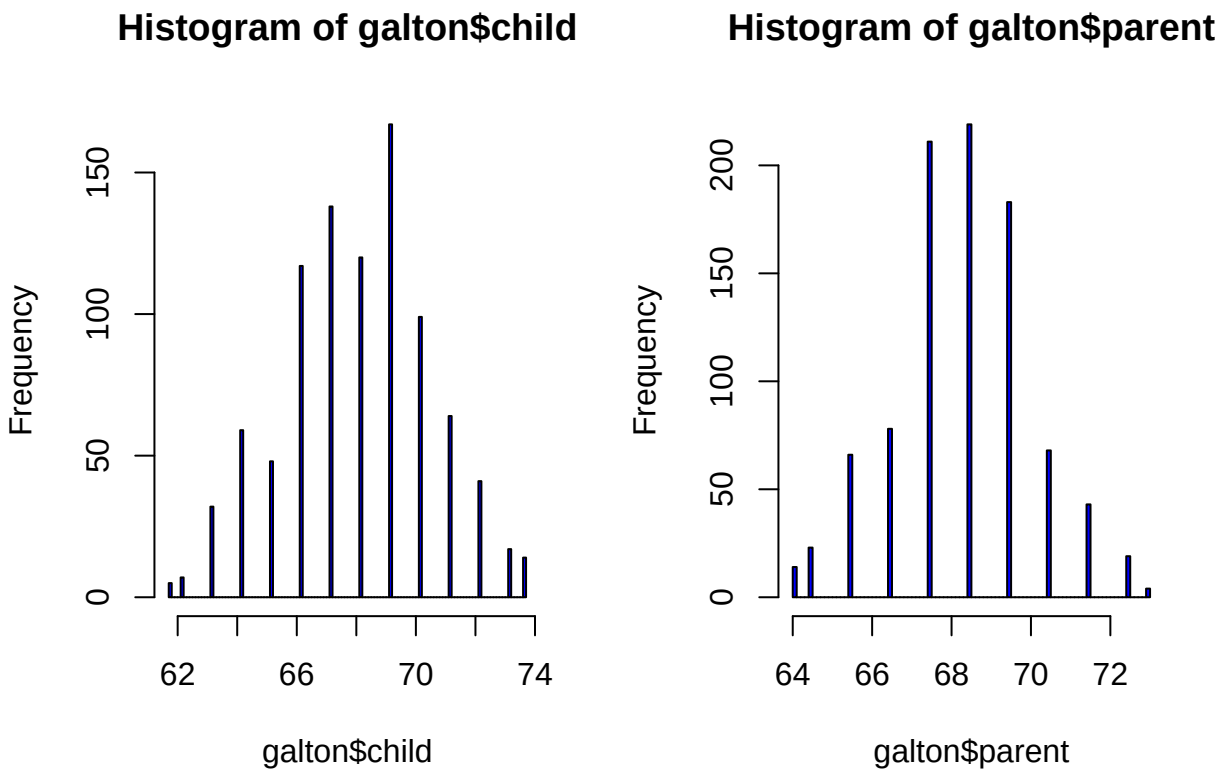
```
library(datasets)
library(prettydoc)
library(ggplot2)
library(reshape2)
library(UsingR)
library(dplyr)
library(manipulate)
set.seed(500)
```

Galton's data

- Parent distribution is all heterosexual couples.
- Correction for gender via multiplying female heights by 1.08.
- Overplotting is an issue from discretization.

Overplotting, the plotting of data over a previous plot, can occur for several reasons. Rounding is a common cause as is dividing up continuous variables and placing them into discrete bins. There a number of ways to deal with overplotting. One is to add a small amount of random noise to the plot, jitter.

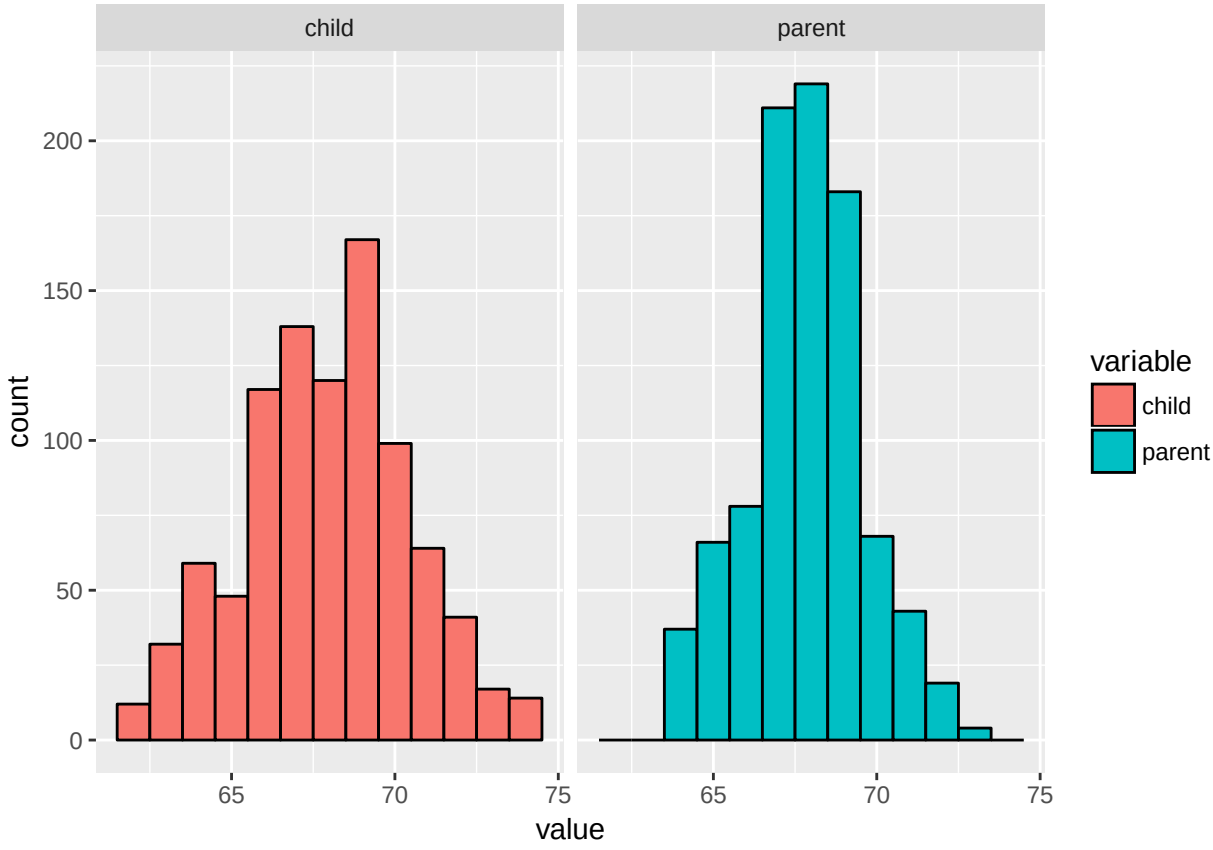
```
data(galton)
par(mfrow=c(1,2))
hist(galton$child,col="blue",breaks=100)
hist(galton$parent,col="blue",breaks=100)
```




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

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

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

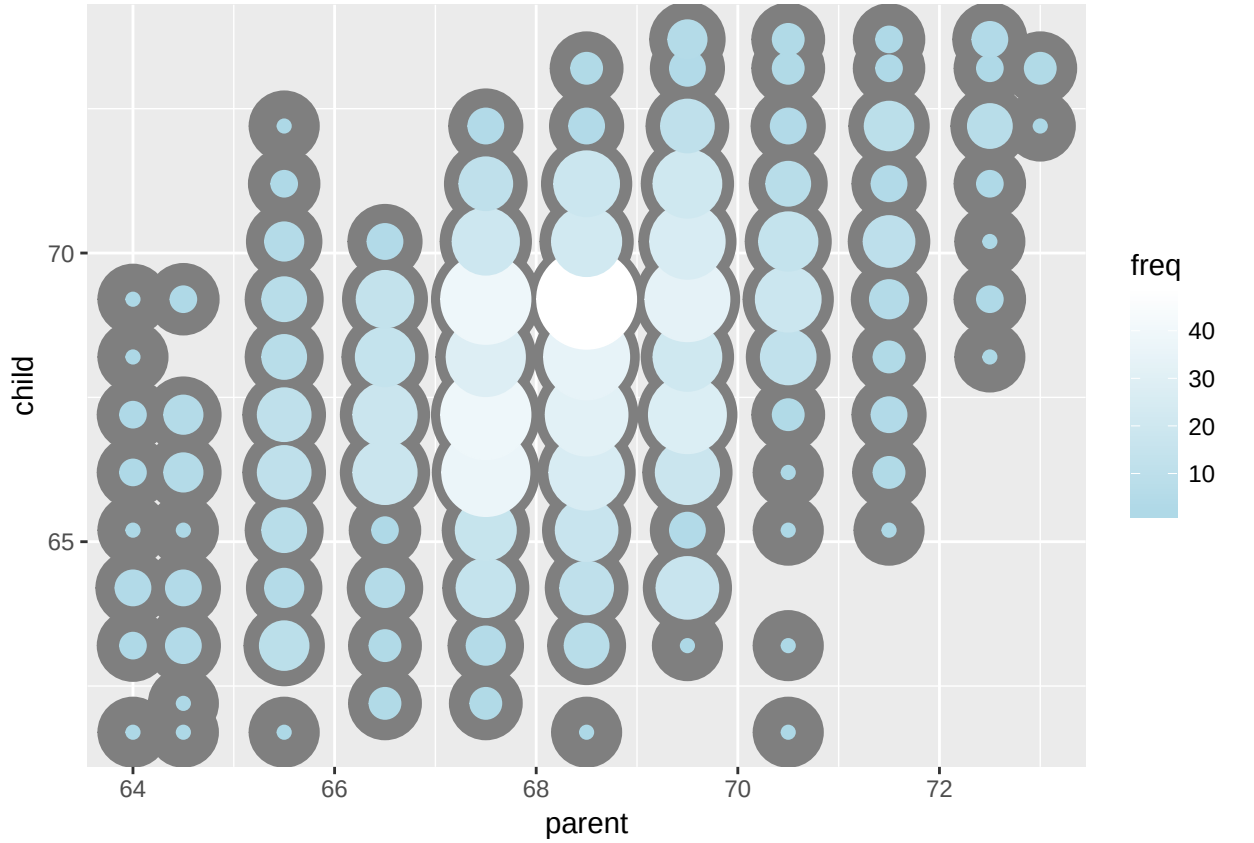


Finding the middle via least squares

```
myHist <- function(mu){
    mse <- mean((galton$child - mu)^2)
    g <- ggplot(galton, aes(x = child)) + geom_histogram(fill = "salmon", colour = "black", binwidth=1)
    g <- g + geom_vline(xintercept = mu, size = 3)
    g <- g + ggtitle(paste("mu = ", mu, ", MSE = ", round(mse, 2), sep = ""))
    g
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
```

Comparing child to parent heights

```
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))
g <- ggplot(filter(freqData, freq > 0), aes(x = parent, y = child))
g <- g  + scale_size(range = c(2, 20), guide = "none" )
g <- g + geom_point(colour="grey50", aes(size = freq+20))
g <- g + geom_point(aes(colour=freq, size = freq))
g <- g + scale_colour_gradient(low = "lightblue", high="white")
g
```



Regression through the origin

- X_i are the partents' heights.
- Consider picking the slope β that minimizes $\sum (Y_i - X_i\beta)$
- This is using the origin as a pivot point picking the line that mines the sum of the squared vertical distances of the points to the line.

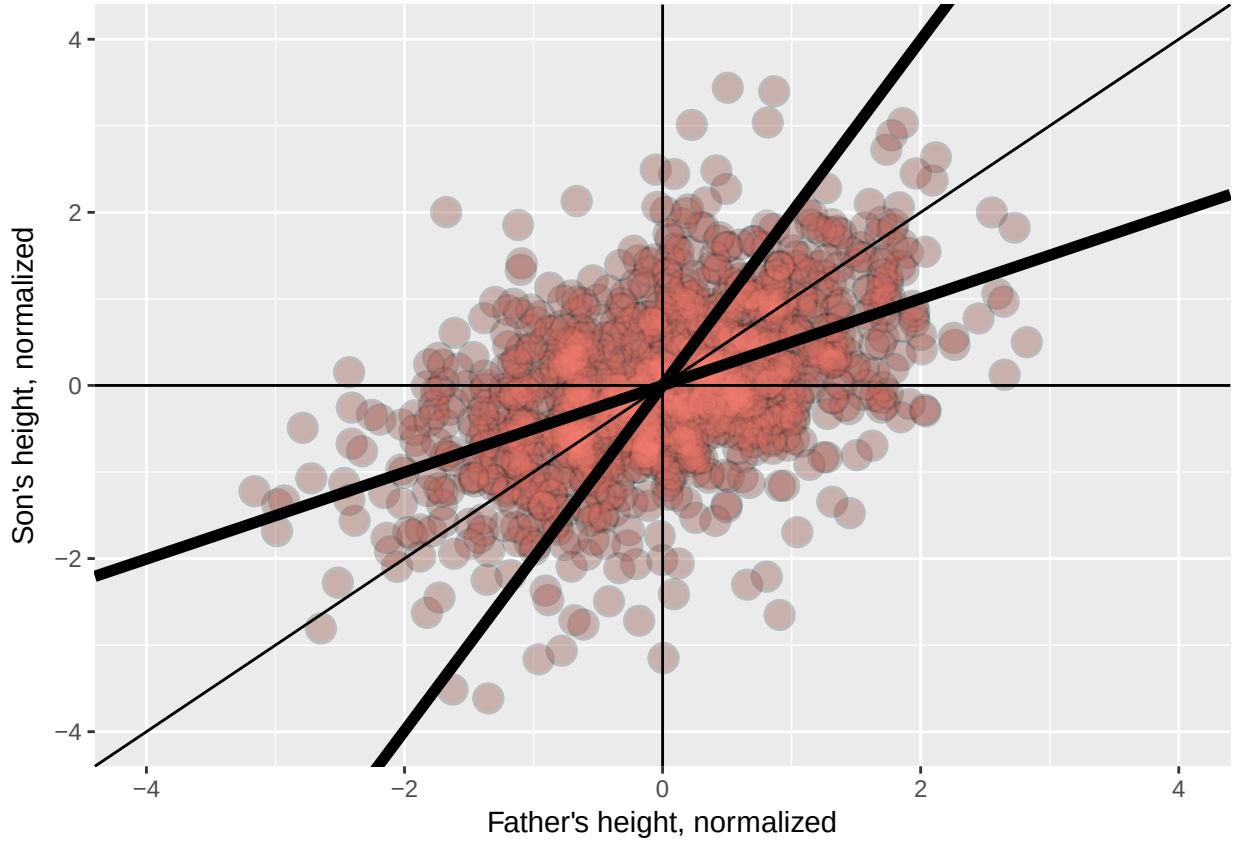
```
myPlot <- function(beta){ y <- galton$child - mean(galton$child) x <- galton$parent - mean(galton$paren
```

```
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
```

Regression to the mean example

```r
y <- (father.son$sheight - mean(father.son$sheight)) / sd(father.son$sheight)
x <- (father.son$fheight - mean(father.son$fheight)) / sd(father.son$fheight)
rho <- cor(x, y)
g <- ggplot(data.frame(x, y), aes(x = x, y = y))
g <- g + geom_point(size = 5, alpha = .2, colour = "black")
g <- g + geom_point(size = 4, alpha = .2, colour = "salmon")
g <- g + xlim(-4,4) + ylim(-4,4)
g <- g + geom_vline(xintercept = 0)
g <- g + geom_hline(yintercept = 0)
g <- g + geom_abline(intercept = 0, slope = 1)
g <- g + geom_abline(intercept = 0, slope = rho, size = 2)
g <- g + geom_abline(intercept = 0, slope = 1 / rho, size = 2)
g <- g + xlab("Father's height, normalized")
g <- g + ylab("Son's height, normalized")
g
```



Note: You normalize data by subtracting its mean and dividing by its standard deviation.

Quiz

Question 1

Consider the data set given below. Give the value of μ that minimizes the least squares equation $\sum_{i=1}^{n} w_i \left(x_i - \mu\right)^2$

```r
## Data set
x <- c(0.18, -1.54, 0.42, 0.95)
## Weights
w <- c(2, 1, 3, 1)

## Calculate the mean of the sum of $w\timesx$ the devide by the sum of $w$ to get the mean
sum(w * x)/sum(w)
```

```
## [1] 0.1471429
```

Question 2

Consider the following data set. Fit the regression through the origin and get the slope treating y as the outcome and x as the regressor. (Hint, do not center the data since we want regression through the origin, not through the means of the data.)

```r
x <- c(0.8, 0.47, 0.51, 0.73, 0.36, 0.58, 0.57, 0.85, 0.44, 0.42)
y <- c(1.39, 0.72, 1.55, 0.48, 1.19, -1.59, 1.23, -0.65, 1.49, 0.05)

## Subtract 1 to omit the calculated intercept and go through the origin
q2 <- lm(y ~ x - 1)

q2
```

```
##
## Call:
## lm(formula = y ~ x - 1)
##
## Coefficients:
##      x
## 0.8263
```

Question 3

Do data(mtcars) from the datasets package and fit the regression model with mpg as the outcome and weight as the predictor. Give the slope coefficient.

```r
data("mtcars")

lm(mtcars$mpg ~ mtcars$wt)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars$wt)
##
## Coefficients:
## (Intercept)    mtcars$wt
##      37.285       -5.344
```

Question 4

Consider data with an outcome (Y) and a predictor (X). The standard deviation of the predictor is one half that of the outcome. The correlation between the two variables is .5. What value would the slope coefficient

for the regression model with Y as the outcome and X as the predictor?

$$Cor\left(X,Y\right)=\frac{Cov\left(X,Y\right)}{S_xS_y}$$

where S_x and S_y are the estimates of standard deviation for the X and Y observations so:

```
.5*(1/.5)
```

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

Question 5

Students were given two hard tests and scores were normalized to have empirical mean 0 and variance 1. The correlation between the scores on the two tests was 0.4. What would be the expected score on Quiz 2 for a student who had a normalized score of 1.5 on Quiz 1?

```
cor <- 0.4
stq2 <- 1.5
## In the normalized data the slope = correlation so
cor * stq2
```

```
## [1] 0.6
```

Question 6

Consider the data given by the following. What is the value of the first measurement if x were normalized (to have mean 0 and variance 1)?

```
x <- c(8.58, 10.46, 9.01, 9.64, 8.86)
xm <- (x - mean(x))/sd(x)
xm
```

```
## [1] -0.9718658  1.5310215 -0.3993969  0.4393366 -0.5990954
```

Question 7

Consider the following data set (used above as well). What is the intercept for fitting the model with x as the predictor and y as the outcome?

```
x <- c(0.8, 0.47, 0.51, 0.73, 0.36, 0.58, 0.57, 0.85, 0.44, 0.42)
y <- c(1.39, 0.72, 1.55, 0.48, 1.19, -1.59, 1.23, -0.65, 1.49, 0.05)

model <- lm( y ~ x)
model$coefficients
```

```
## (Intercept)           x
##    1.567461   -1.712846
```

Question 8

You know that both the predictor and response have mean 0. What can be said about the intercept when you fit a linear regression?

- **It must be identically 0.**

- It must be exactly one.
- It is undefined as you have to divide by zero.
- Nothing about the intercept can be said from the information given.

Question 9

Consider the following data. What value minimizes the sum of the squared distances between these points and itself? Since the value is $\mu = \bar{X}$ take the mean of the values.

```
x <- c(0.8, 0.47, 0.51, 0.73, 0.36, 0.58, 0.57, 0.85, 0.44, 0.42)
mean(x)
```

```
## [1] 0.573
```

Question 10

Let the slope having fit Y as the outcome and X as the predictor be denoted as β_1. Let the slope from fitting X as the outcome and Y as the predictor be denoted as λ_1. Suppose that you divide β_1 by λ_1; in other words consider $\frac{\beta_1}{\lambda_1}$. What is this ratio always equal to?

$\beta = Cor\left(X, Y\right) \times \frac{\sigma_y}{\sigma_x}$ where σ is the respective squared distance for x and y. Since $Cor\left(X, Y\right) = Cor\left(Y, X\right)$ they cancel out so:

- **Var(Y)/Var(X)**
- 1
- Cor(Y,X)
- 2SD(Y)/SD(X)