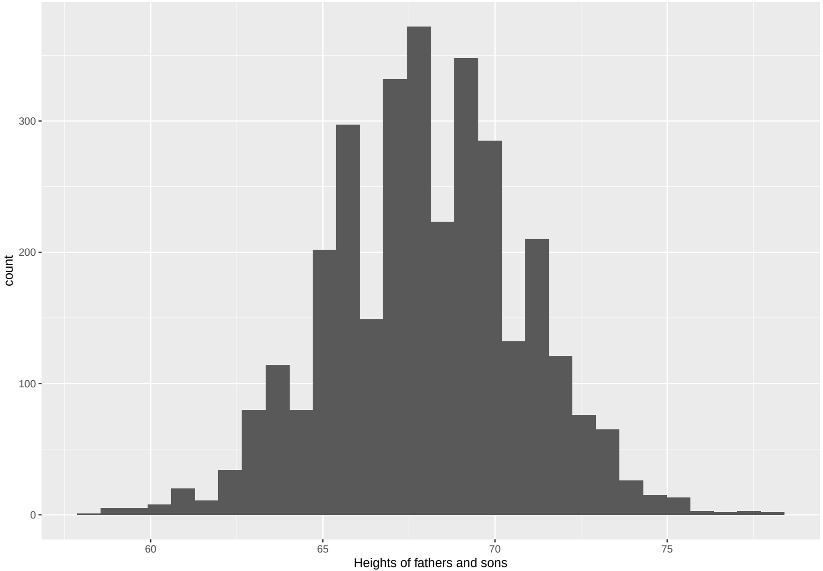
The complex inner life of simple regression

Matthew Rudd

Math for Data Science Conference, 12/1/20

Pearson-Lee height data



Pearson-Lee height data

- Best guess: the average height, 68.02 inches

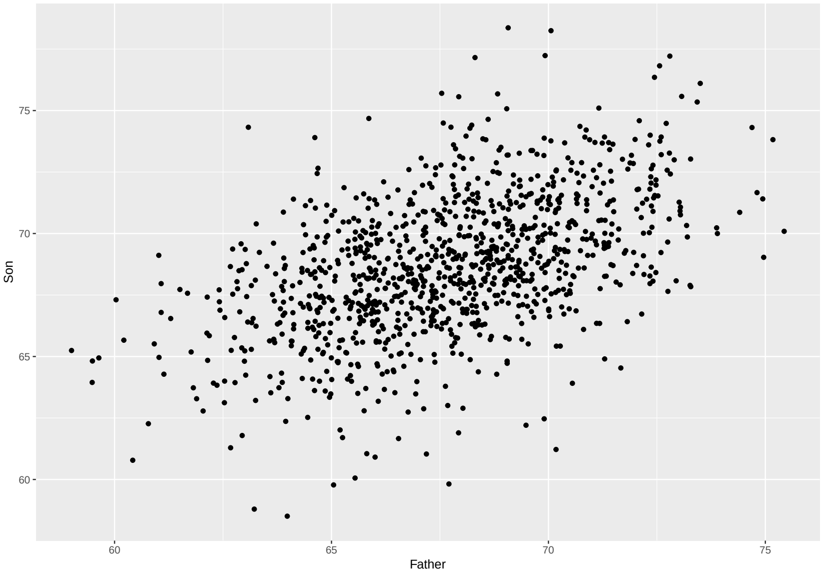
Pearson-Lee height data

- Best guess: the average height, 68.02 inches
- Probably off by 1 or 2 SDs, 2.8 to 5.6 inches

Pearson-Lee height data

- Best guess: the average height, 68.02 inches
- Probably off by 1 or 2 SDs, 2.8 to 5.6 inches
- For better predictions, use more information!

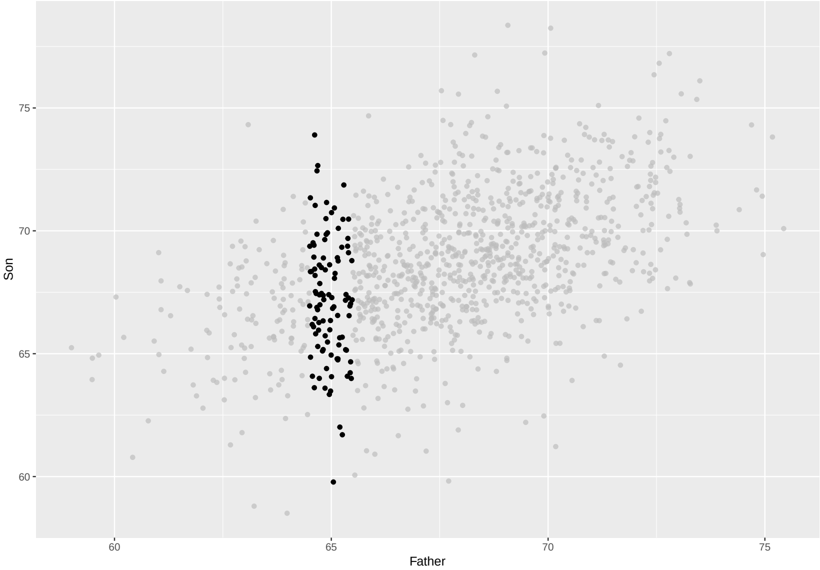
Pearson-Lee height data



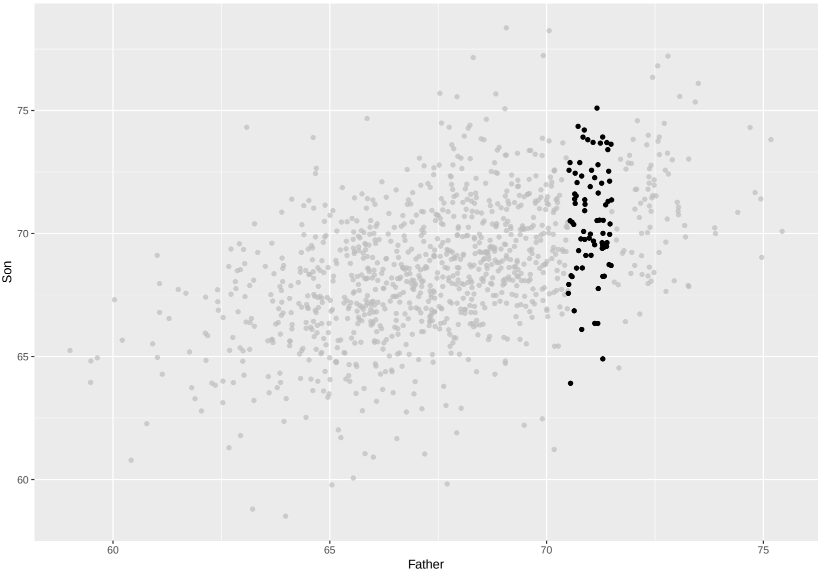
Pearson-Lee height data

Father's height	Number of sons	Average height	SD
62	15	65.59	1.78
63	36	66.51	2.91
64	60	66.70	2.31
65	101	67.22	2.53
66	139	67.66	2.35
67	134	68.14	2.24
68	157	69.09	2.76
69	142	69.44	2.30
70	115	69.77	2.49
71	77	70.54	2.31
72	50	70.68	2.33
73	28	72.05	2.76

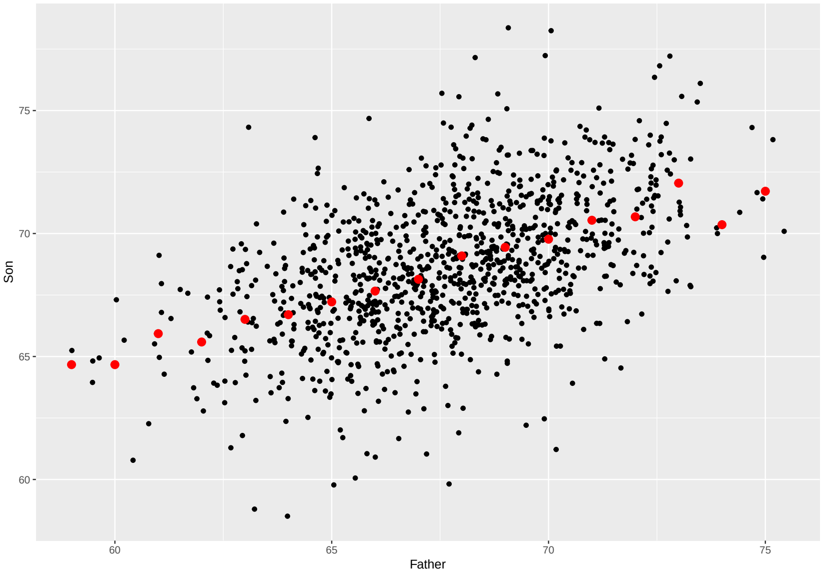
Pearson-Lee height data



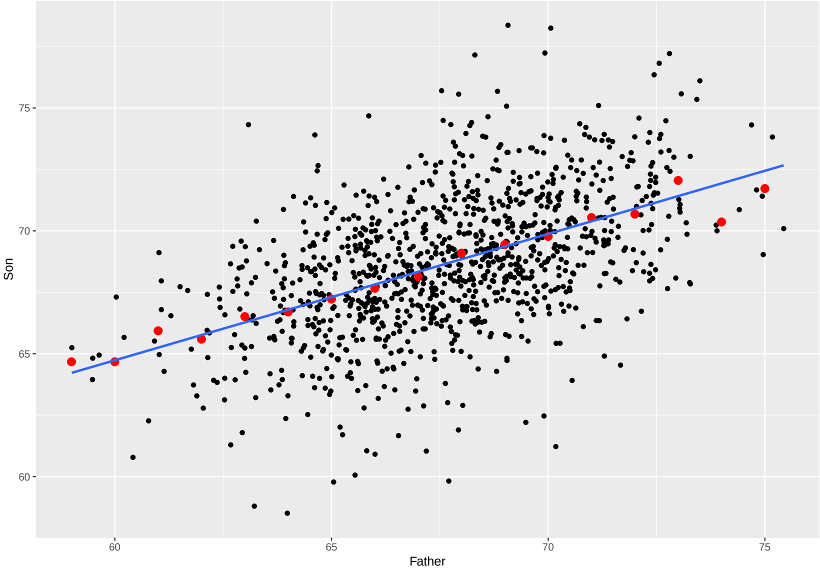
Pearson-Lee height data



Pearson-Lee height data



Pearson-Lee height data



Pearson-Lee height data

- The average height of a group of sons depends linearly on the father's given height

Pearson-Lee height data

- The average height of a group of sons depends linearly on the father's given height
- Using this data,

$$\text{Average height of sons} = 33.89 + .514 \times \text{Father's height}$$

Pearson-Lee height data

- The average height of a group of sons depends linearly on the father's given height
- Using this data,

$$\text{Average height of sons} = 33.89 + .514 \times \text{Father's height}$$

- This is *simple linear regression*.

Pearson-Lee height data

```
##
## Call:
## lm(formula = Son ~ Father, data = heights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8772 -1.5144 -0.0079  1.6285  8.9685
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.88660    1.83235   18.49   <2e-16 ***
## Father       0.51409    0.02705   19.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 2.437 on 1076 degrees of freedo
## Multiple R-squared:  0.2513, Adjusted R-squared:  0.2506
## F-statistic: 361.2 on 1 and 1076 DF,  p-value: < 2.2e-16
```

The simple linear regression model

blah β

The Gauss-Markov Theorem

you know, OLS is BLUE and whatnot

Chebyshev's Theorem

75% of observations are within 2 SDs – no matter what!