Why are tall parents more likely to have shorter children? The first regression.

Unit 7 Lecture 1

Jonathan Auerbach
STAT 489 Pre-Cap Prof Development
jauerba@gmu.edu

November 16, 2021

Learning Objectives

After this lecture, you will be able to:

1. Describe the regression phenomenon as stated by Galton.
2. Use the `ggplot2` and `geomtextpath` packages to visually compare the regression line and the identity line.
3. Derive the regression line from the bivariate normal distribution.
4. Derive the regression line using least squares.

These slides use the following R packages

Setup:

```
library("knitr")
library("HistData")
library("tidyverse")
library("geomtextpath")
theme_set(theme_bw(base_size = 20))
```

Why are tall parents likely to have shorter children?

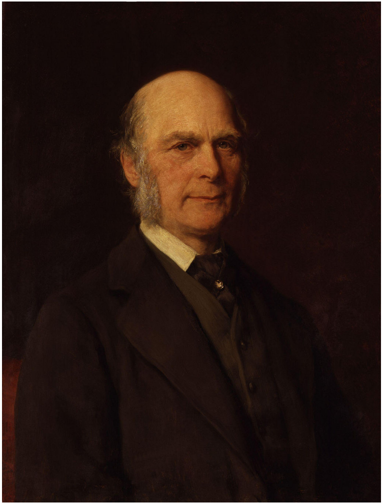
- Francis Galton's investigation of this question revolutionized statistical methodology.
  - Galton was Charles Darwin's half cousin and greatly influenced by Darwin's *On the Origin of Species* (1859).
  - He studied how physical characteristics were inherited, like height.

- In one study, Galton recorded the heights of 205 parents and their 928 adult children
  - The heights of women were multiplied by 1.08 to account for the fact that men are 8 percent taller than women, on average.
  - Galton then compared the average parent height to the height of each child.
  - He noticed that tall parents tend to have children who are shorter than they are.

Why are tall parents likely to have shorter children?

- Initially, Galton believed the children had "regressed towards mediocrity."
  - Galton concluded in his 1877 article *Typical Laws of Heredity* that regression was a force governing natural selection, opposing the force that creates new species.
- Galton later realized that the force of regression was an illusion (statistical artifact).
  - It was not the children who were abnormal in their regression towards mediocrity, but their parents who were abnormal in having an above average height to begin with.
  - He published his findings in his 1886 article *Regression Towards Mediocrity in Hereditary Stature*.
- Karl Pearson, Udny Yule, and other statisticians studied the regression phenomenon mathematically, resulting in the regression analysis that we teach in statistics courses today.

Galton and *Regression Towards Mediocrity* (1886)



ANTHROPOLOGICAL MISCELLANEA.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.

[WITH PLATES IX AND X.]

THIS memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.

It is some years since I made an extensive series of experiments on the produce of seeds of different size but of the same species. They yielded results that seemed very noteworthy, and I used them as the basis of a lecture before the Royal Institution on February 9th, 1877. It appeared from these experiments that the offspring did *not* tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small. The point of convergence was considerably below the average size of the seeds contained in the large bagful I bought at a nursery garden, out of which I selected those that were sown, and I had some reason to believe that the size of the seed towards which the produce converged was similar to that of an average seed taken out of beds of self-planted specimens.

The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it. This curious result was based on so many plantings, conducted for me by friends living in various parts of the country, from Nairn in the north to Cornwall in the south, during one, two, or even three generations of the plants, that I could entertain no doubt of the truth of my conclusions. The exact ratio of regression remained a little doubtful, owing to variable influences; therefore I did not attempt to define it. But as it seems a pity that no

Source: https://en.wikipedia.org/wiki/Francis_Galton#/media/File:Sir_Francis_Galton_by_Gustav_Graef.jpg

Galton cross-classified parent and child heights…

TABLE I.

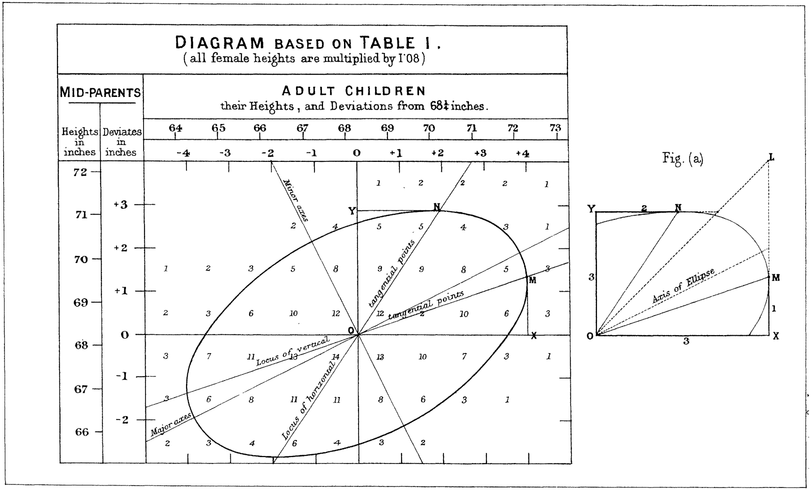
NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

… and visualized the relationship geometrically

Plate X.



Galton's Data from the HistData package

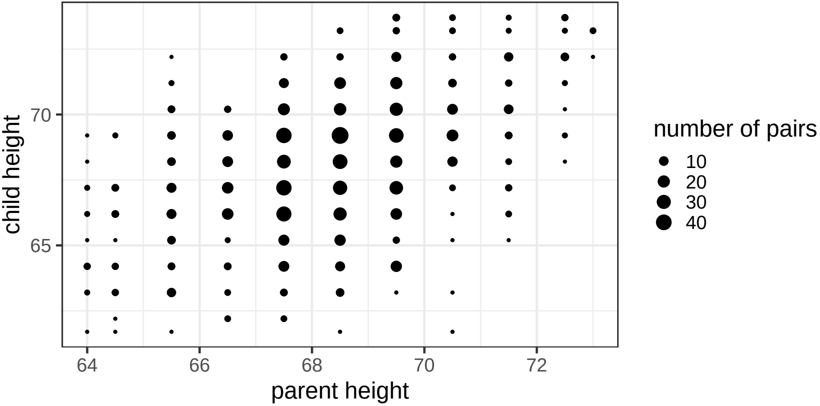
```
Galton %<>%
  group_by(parent, child) %>%
  summarize(num_pairs = n()) %>%
  ungroup()

Galton %>%
  top_n(5) %>%
  kable()
```

parent	child	num_pairs
67.5	66.2	36
67.5	67.2	38
67.5	69.2	38
68.5	68.2	34
68.5	69.2	48

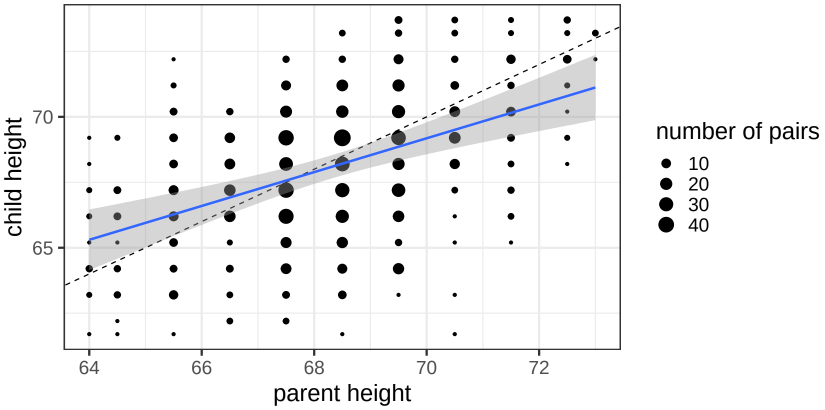
Taller parents tend to have taller children...

```
(galton_plot <- Galton %>% ggplot() +
   aes(x = parent, y = child, weight = num_pairs) +
    geom_point(aes(size = num_pairs)) +
    labs(x = "parent height", y = "child height",
         size = "number of pairs"))
```



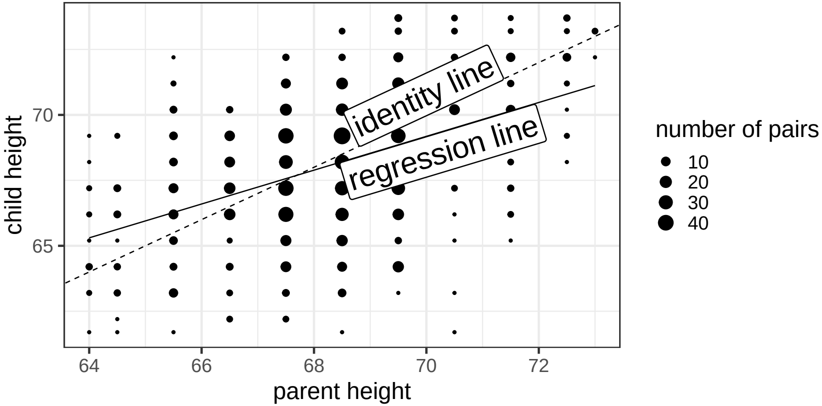
... but children appear to regress since slope of the best fit line is smaller than slope of the identity line

```
galton_plot +
    geom_abline(intercept = 0, slope = 1,
                linetype = 2) +
    geom_smooth(method = "lm")
```



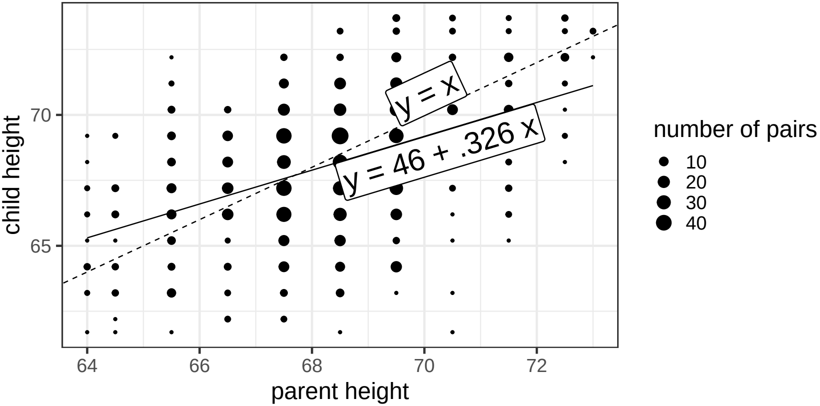
Add line labels with the geomtextpath package

```
galton_plot +
 geom_labelabline(label = "identity line", intercept = 0,
                  slope = 1, linetype = 2, size = 9,
                  hjust = .675, vjust = -.3) +
 geom_labelsmooth(label = "regression line", method = "lm",
                  size = 9, hjust = .8, vjust = 1.35)
```



Add line labels with the geomtextpath package

```
galton_plot +
 geom_labelabline(label = "y = x", intercept = 0,
                  slope = 1, linetype = 2, size = 9,
                  hjust = .675, vjust = -.3) +
 geom_labelsmooth(label = "y = 46 + .326 x", method = "lm",
                  size = 9, hjust = .8, vjust = 1.35)
```



Why are tall parents likely to have shorter children?

- Galton initially thought regression was a force governing natural selection, opposing the force that creates new species. He later realized that the force of regression was a statistical artifact:
  - Children only share some of the factors that made their parents tall.
  - By selecting tall parents, Galton unknowingly selected parents with unusual, height-promoting factors.
  - These factors were less likely to reoccur in these parents' children, resulting in shorter heights.
- Scientists have since found hundreds of genetic variants that influence height.
  - A recent study (2010) reports that more than 80 percent of height is due to genetic factors and 20 percent is due to environmental factors.
- Karl Pearson, Udny Yule, and other statisticians studied the regression phenomenon mathematically, resulting in the regression analysis that we teach in statistics courses today.

Karl Pearson (left) and Udny Yule (right)

Source: https://en.wikipedia.org/wiki/Karl_Pearson#/media/File:Karl_Pearson,_1912.jpg
https://en.wikipedia.org/wiki/Udny_Yule#/media/File:George_Udny_Yule.jpg

Pearson assumed a bivariate normal distribution

Recall $\rho = \frac{\mathsf{Cov}(X,Y)}{\sigma_X \sigma_Y}$, $\beta_1 = \frac{\mathsf{Cov}(X,Y)}{\sigma_X^2} = \rho\frac{\sigma_Y}{\sigma_X}$, and $\beta_0 = \mu_Y - \beta_1\mu_X$.

If

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \text{Normal}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}\right),$$

then

$$\begin{aligned} \mathbb{E}\left[Y|X=x\right] &= \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X) \\ &= \mu_Y + \beta_1(x-\mu_X) \\ &= \mu_Y - \beta_1\mu_X + \beta_1 x \\ &= \beta_0 + \beta_1 x \end{aligned}$$

The regression phenomenon happens when $\sigma_X \approx \sigma_Y$ and $\rho < 1$ because

$$\beta_1 = \rho\frac{\sigma_Y}{\sigma_X} \approx \rho < 1$$

Yule assumed a linear relationship

Yule used least squares to find the linear function of X that best fits Y.

$$\underset{\beta_0,\beta_1}{\text{argmin}}\ \mathbb{E}[(Y-(\beta_0-\beta_1 X))^2]$$

He solved the normal equations

$$\begin{cases} 0 \overset{\text{set}}{=} \frac{\partial}{\partial\beta_0}\mathbb{E}[(Y-(\beta_0-\beta_1 X))^2] = \mathbb{E}[(-2(Y-(\beta_0-\beta_1 X))] \\ 0 \overset{\text{set}}{=} \frac{\partial}{\partial\beta_1}\mathbb{E}[(Y-(\beta_0-\beta_1 X))^2] = \mathbb{E}[(-2(Y-(\beta_0-\beta_1 X)X)] \end{cases}$$

Rearranging the first equation yields $\beta_0 = \mathbb{E}[Y] - \beta_1\mathbb{E}[X]$

Multiplying both sides of the second equation by -2 and substituting the solution for β_0 results in the equation

$$0 = \mathbb{E}[XY] - (\beta_0)\mathbb{E}[X] - \beta_1\mathbb{E}[X^2] = \mathbb{E}[XY] - (\mathbb{E}[Y] - \beta_1\mathbb{E}[X])\mathbb{E}[X] - \beta_1\mathbb{E}[X^2]$$

Rearranging yields $\beta_1 = \frac{\mathbb{E}[XY]-\mathbb{E}[Y]\mathbb{E}[X]}{\mathbb{E}[X^2]-\mathbb{E}[X]^2} = \frac{\text{Cov}(X,Y)}{\sigma_X^2} = \rho\frac{\sigma_Y}{\sigma_X}$

Misinterpreting regression may be the most common statistical error

- Unusual observations are often the result of chance (at least in part) and become usual when measured again in later periods.
  - successful businesses, low performing students, high crime areas, etc.
- The relationship between observations can often be summarized by a regression line. Two common justifications are that
  1. the measurements follow a bivariate normal distribution.
  2. a best fit line well approximates their relationship.
- Misinterpreting the reversion from unusual to usual is called the "regression fallacy."
  - Despite being documented over 100 years ago, the regression fallacy is common today.
  - Milton Friedman (1992) suspected "... the regression fallacy is the most common fallacy in the statistical analysis of economic data..."

References

1. Allen, Hana Lango, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467.7317 (2010): 832.

2. Friedman, Milton. Do old fallacies ever die?. (1992): 2129-2132.

3. Galton, Francis. Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland. 15 (1886): 246-263.

4. Galton, Francis. Typical laws of heredity. Nature. 15 (1877): 492-495.

5. Stigler, Stephen M. Darwin, Galton and the statistical enlightenment. Journal of the Royal Statistical Society: Series A. 173.3 (2010): 469-482.