Linear regression 1

STA 032: Gateway to data science Lecture 23

Jingwei Xiong

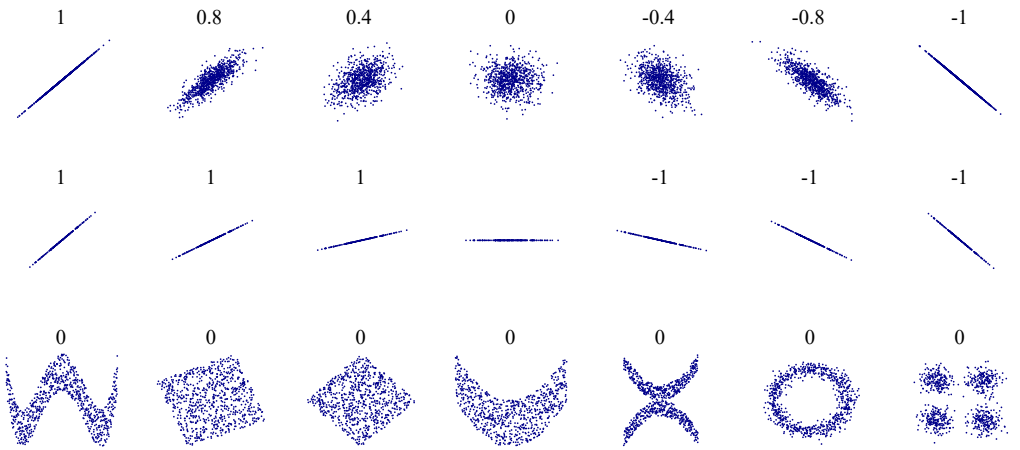
May 26, 2023

Summary:

- Sample correlation sampling distribution
- Conditional expectation
- Regression to the mean
- Regression line

Regression Intro

- Confidence interval, hypothesis tests are for single variables.
- It's common in data science application to study relationship between two or more variables.
- In lecture 10 we introduced correlation:
- Correlation is the association between two variables
- (Pearson) Correlation coefficient is a measure of **linear** correlation between two sets of data
- Ranges from -1 to 1



Recap: Correlation

- Sample correlation: $r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$
- Population correlation: ρ
- To calculate: `cor(vector1, vector2)` in R

> To propose statistical relationship about correlation, we use regressions.

Regression Analysis

- Regression analysis is a statistical technique used to model the relationship between a dependent variable and one or more independent variables.
- It is a powerful tool for understanding and predicting the behavior of variables in a dataset.
- Regression analysis allows us to analyze the impact of independent variables on the dependent variable and quantify their relationships.
- Understanding regression analysis is essential for conducting rigorous data analysis and making informed decisions based on statistical evidence.
- Regression occurs frequently in engineering and science

Case study: is height hereditary?

- Francis Galton studied the variation and heredity of human traits.
- He collected and studied height data from families to try to understand heredity.
- This is the dataset from which regression was born.
- A very specific question Galton tried to answer was: how well can we predict a child's height based on the parents' height?

> Historical note: Galton made important contributions to statistics and genetics, but he was also one of the first proponents of eugenics, a scientifically flawed philosophical movement favored by many biologists of Galton's time but with horrific historical consequences. You can read more about it here: https://pged.org/history-eugenics-and-genetics/.

Case study: is height hereditary?

We have access to Galton's family height data through the HistData package. This data contains heights on several dozen families: mothers, fathers, daughters, and sons. To imitate Galton's analysis, we will create a dataset with the heights of fathers and a randomly selected son of each family:

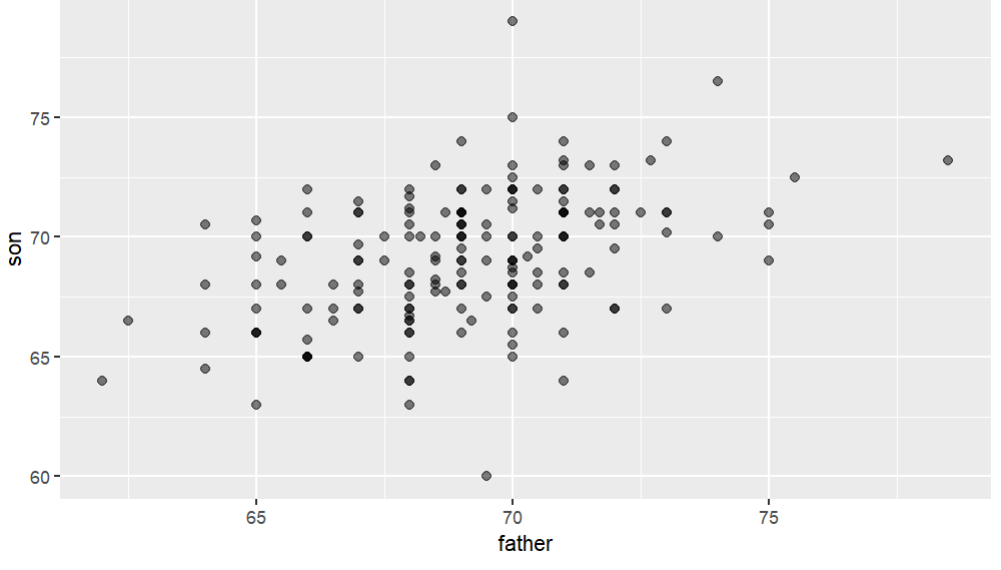
```
library(tidyverse)
library(HistData)
data("GaltonFamilies")

set.seed(1983)
galton_heights <- GaltonFamilies |>
  filter(gender == "male") |>
  group_by(family) |>
  sample_n(1) |>
  ungroup() |>
  select(father, childHeight) |>
  rename(son = childHeight)
```

Suppose we were asked to summarize the father and son data. Since both distributions are well approximated by the normal distribution, we could use the two averages and two standard deviations as summaries:

```
# A tibble: 1 x 4
  mean(father) sd(father) mean(son) sd(son)
         <dbl>      <dbl>     <dbl>   <dbl>
1         69.1       2.55      69.2    2.71
```

However, this summary fails to describe an important characteristic of the data: the trend that the taller the father, the taller the son.

```
galton_heights |> ggplot(aes(father, son)) +
  geom_point(alpha = 0.5, size = 2)
```



Sample correlation is a random variable

Before we continue connecting correlation to regression, let's remind ourselves about random variability.

- The sample correlation is the most commonly used estimate of the population correlation.
    - the correlation we compute and use as a summary is a random variable
- Sample correlation of the 179 pairs

```
galton_heights |> summarize(r = cor(father, son)) |> pull(r)
```

```
[1] 0.4334102
```

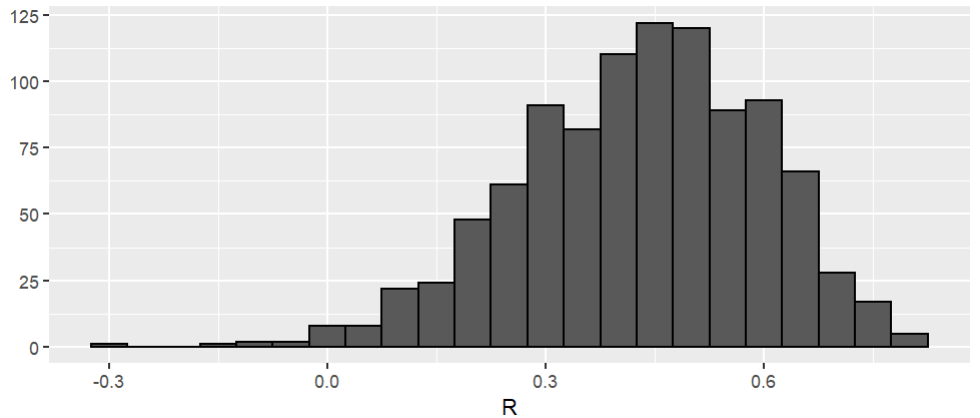
By way of illustration, let's assume that the 179 pairs of fathers and sons is our entire population. A less fortunate geneticist can only afford measurements from a random sample of 25 pairs. The sample correlation can be computed with:

```
R <- sample_n(galton_heights, 25, replace = TRUE) |>
  summarize(r = cor(father, son)) |> pull(r)
```

```
R <- sample_n(galton_heights, 25, replace = TRUE) |>
  summarize(r = cor(father, son)) |> pull(r)
```

R is a random variable. We can run a Monte Carlo simulation to see its distribution:

```
B <- 1000
N <- 25
R <- replicate(B, {
  sample_n(galton_heights, N, replace = TRUE) |>
    summarize(r=cor(father, son)) |>
    pull(r)
})
qplot(R, geom = "histogram", binwidth = 0.05, color = I("black"))
```



We see that the expected value of R is the population correlation: (Here we assume 179 is our entire population)

```
mean(R)
```

```
[1] 0.4295175
```

and that it has a relatively high standard error relative to the range of values R can take:

```
sd(R)
```

```
[1] 0.166288
```

So, when interpreting correlations, remember that correlations derived from samples are estimates containing **uncertainty**.

Also, note that because the sample correlation is an average of independent draws, the central limit actually applies. Therefore, for large enough N, the distribution of `R` is approximately normal with expected value ρ. The standard deviation, which is somewhat complex to derive, is $\sqrt{\frac{1-\rho^2}{N-2}}$.

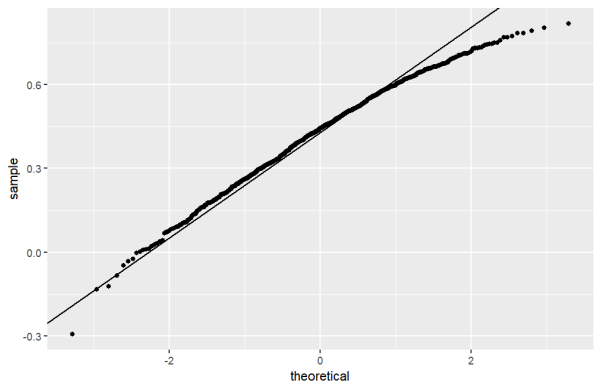
$$R \approx N(\rho, \sqrt{\frac{1-\rho^2}{N-2}})$$

Central limit applies. For large enough N, the distribution of R is approximately normal with expected value ρ, standard deviation $\sqrt{\frac{1-\rho^2}{N-2}}$.

$$R \approx N(\rho, \sqrt{\frac{1-\rho^2}{N-2}})$$

In our example, $N = 25$ does not seem to be large enough to make the approximation a good one. If you increase N, you will see the distribution converging to normal.

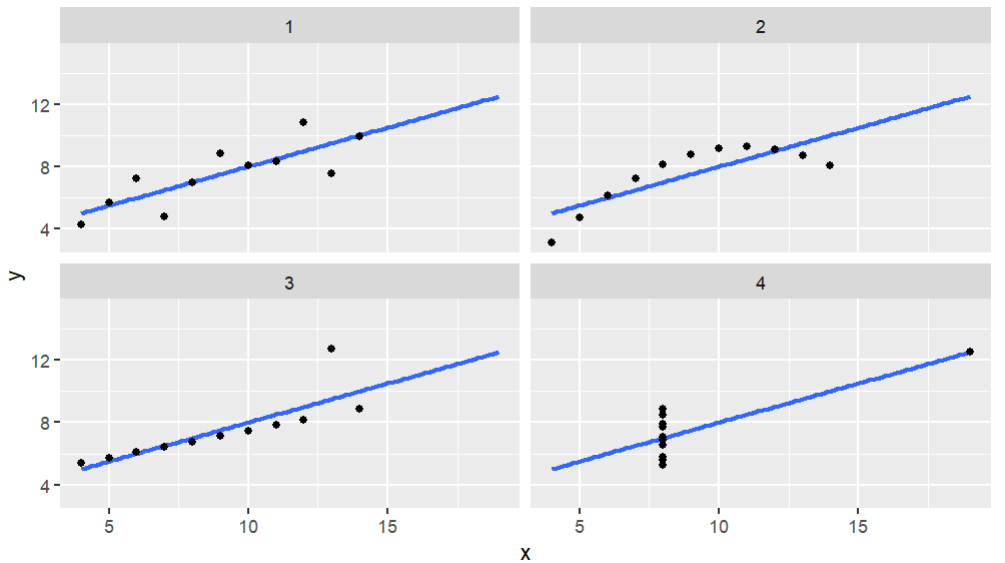
```
ggplot(aes(sample=R), data = data.frame(R)) +
  stat_qq() +
  geom_abline(intercept = mean(R), slope = sqrt((1-mean(R)^2)/(N-2)))
```



> QQplot: If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the identity line y = x. More: https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

Correlation is not always a useful summary

Correlation is not always a good summary of the relationship between two variables. The following four artificial datasets, referred to as Anscombe's quartet, famously illustrate this point. All these pairs have a correlation of 0.82:



Conditional expectations

- Suppose we want to guess the height of a randomly selected son and we don't know his father's height.

Because the distribution of sons' heights is approximately normal, we know the average height, 69.2, is the value with the highest proportion and would be the prediction with the highest chance of minimizing the error.

- But if we are told that the father is taller than average, say 72 inches tall, do we still guess 69.2 for the son?

If we were able to collect data from a very large number of fathers that are 72 inches, the distribution of their sons' heights would be normally distributed.

> This implies that the average of the distribution computed on this subset would be our best prediction.

- In general, we call this approach **conditioning**.
    - we stratify a population into groups and compute summaries in each group

- To provide a mathematical description of conditioning, consider we have a population of pairs of values $(x_1, y_1), \ldots, (x_n, y_n)$, for example all father and son heights in England.
- The expectation of son heights: $EY = 1/n \sum_{i=1}^{n} y_i$
- We are interested in only the subset of a population with a specific x_i value, 72 inches in our example.
- The y_i in the subpopulation have a distribution, referred to as the **conditional distribution**,

$$Y|X = x$$

- It has an expected value referred to as the **conditional expectation**:

$$\mathrm{E}(Y \mid X = x)$$

with x representing the fixed value that defines that subset, for example 72 inches here.

- Similarly, we denote the standard deviation of the strata with

$$\mathrm{SD}(Y \mid X = x) = \sqrt{\mathrm{Var}(Y \mid X = x)}$$

- Here we are interested in computing the average son height *conditioned* on the father being 72 inches tall: $E(Y|X = 72)$ using the sample collected by Galton.
- We previously learned that the sample average is the preferred approach to estimating the population average.
- However, a challenge when using this approach to estimating conditional expectations is that for continuous data we don't have many data points matching **exactly one value** in our sample. For example, we have only:

```
sum(galton_heights$father == 72)
```

```
[1] 8
```

fathers that are exactly 72-inches. If we change the number to 72.5, we get even fewer data points:

```
sum(galton_heights$father == 72.5)
```

```
[1] 1
```

Naive way: round continuous by integers

A practical way to improve these estimates of the conditional expectations, is to define strata of with similar values of x.

In our example, we can round father heights to the nearest inch and assume that they are all 72 inches. If we do this, we end up with the following prediction for the son of a father that is 72 inches tall:

```
conditional_avg <- galton_heights |>
  filter(round(father) == 72) |>
  summarize(avg = mean(son)) |>
  pull(avg)
conditional_avg
```

```
[1] 70.5
```

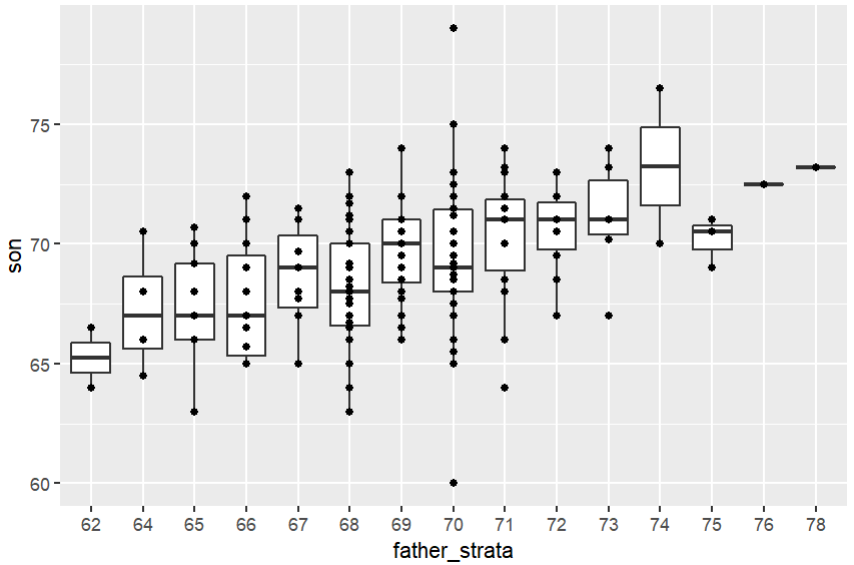
Note that a 72-inch father is taller than average -- specifically, (72.0 - 69.1)/2.5 = 1.1 standard deviations taller than the average father.

Our prediction 70.5 is also taller than average, but only 0.49 standard deviations larger than the average son.

The sons of 72-inch fathers have **regressed** some **to the average height**. We notice that the reduction in how many SDs taller is about 0.5, which **happens to be the correlation**. As we will see in a later section, this is not a coincidence.

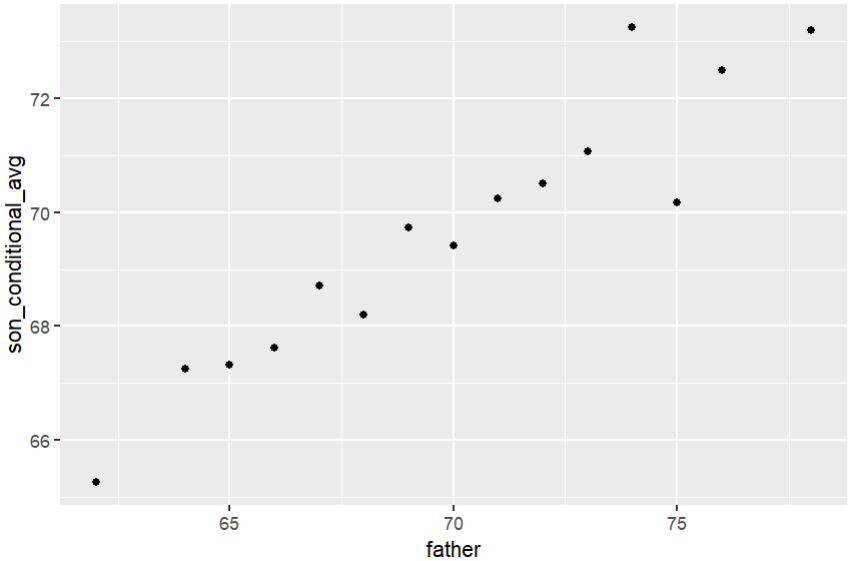
If we want to make a prediction of any height, not just 72, we could apply the same approach to each strata. Stratification followed by boxplots lets us see the distribution of each group:

```
galton_heights |> mutate(father_strata = factor(round(father))) |>
  ggplot(aes(father_strata, son)) +
  geom_boxplot() +
  geom_point()
```



Not surprisingly, the centers of the groups are increasing with height.

Not surprisingly, the centers of the groups are increasing with height. Furthermore, these centers appear to follow a linear relationship. Below we plot the averages of each group.



If we take into account that these averages are random variables with standard errors, the data is consistent with these points following a straight line.

The fact that these conditional averages follow a line is not a coincidence.

The regression line

If we are predicting a random variable Y knowing the value of another $X = x$ using a **regression line**:

For every standard deviation, σ_X, that x increases above the average μ_X, Y increase ρ standard deviations σ_Y above the average μ_Y with ρ the correlation between X and Y.

The formula for the regression is therefore:

$$\left(\frac{Y - \mu_Y}{\sigma_Y}\right) = \rho\left(\frac{x - \mu_X}{\sigma_X}\right)$$

We can rewrite it like this:

$$Y = \mu_Y + \rho\left(\frac{x - \mu_X}{\sigma_X}\right)\sigma_Y$$

If there is perfect correlation, the regression line predicts an increase that is the same number of SDs. If there is 0 correlation, then we don't use x at all for the prediction and simply predict the average μ_Y. For values between 0 and 1, the prediction is somewhere in between. If the correlation is negative, we predict a reduction instead of an increase.

Regression toward mediocrity

Note that if the correlation is positive and lower than 1, our prediction is closer, in standard units, to the average height than the value used to predict, x, is to the average of the xs. This is why we call it **regression**: the son regresses to the average height.

In fact, the title of Galton's paper was: **Regression toward mediocrity in hereditary stature**. To add regression lines to plots, we will need the above formula in the form:

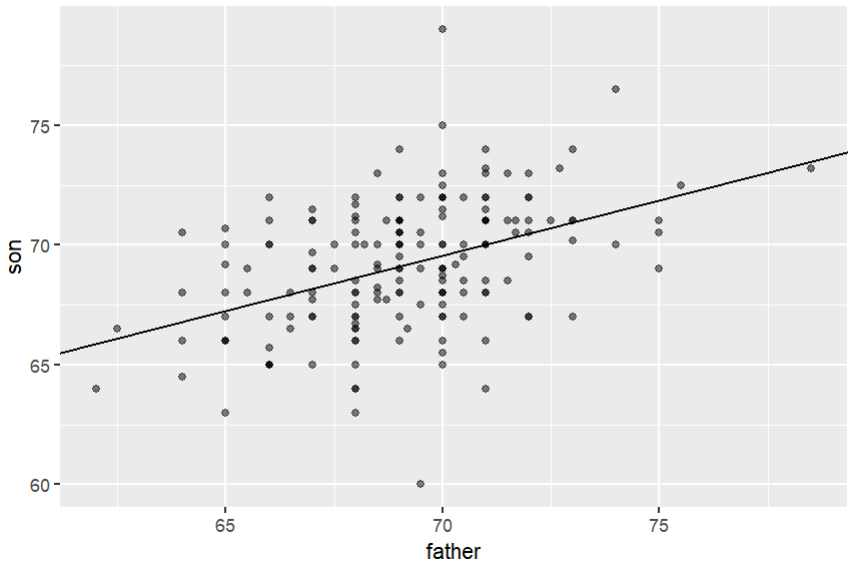
$$y = b + mx \text{ with slope } m = \rho\frac{\sigma_y}{\sigma_x} \text{ and intercept } b = \mu_y - m\mu_x$$

Here we add the regression line to the original data:

```
mu_x <- mean(galton_heights$father)
mu_y <- mean(galton_heights$son)
s_x <- sd(galton_heights$father)
s_y <- sd(galton_heights$son)
r <- cor(galton_heights$father, galton_heights$son)

slope = r * s_y/s_x
intercept = mu_y - r * s_y/s_x * mu_x
```

Here we add the regression line to the original data:

```
galton_heights |>
  ggplot(aes(father, son)) +
  geom_point(alpha = 0.5) +
  geom_abline(slope = slope, intercept = intercept)
```



Regression improves precision

The regression line is much more stable than the conditional mean.

> The conditional average is computed on a relatively small subset, but regression use all the data.

Let's compare the two approaches to prediction that we have presented:

1. Round fathers' heights to closest inch, stratify, and then take the average.
2. Compute the regression line and use it to predict.

We use a Monte Carlo simulation sampling $N = 50$ families:

```
B <- 1000;N <- 50
set.seed(1983)
conditional_avg <- replicate(B, {
  dat <- sample_n(galton_heights, N)
  dat |> filter(round(father) == 72) |>
    summarize(avg = mean(son)) |>
    pull(avg)
  })

regression_prediction <- replicate(B, {
  dat <- sample_n(galton_heights, N)
  mu_x <- mean(dat$father)
  mu_y <- mean(dat$son)
  s_x <- sd(dat$father)
  s_y <- sd(dat$son)
  r <- cor(dat$father, dat$son)
  mu_y + r*(72 - mu_x)/s_x*s_y
})
```

Although the expected value of these two random variables is about the same:

```
mean(conditional_avg, na.rm = TRUE)
```

```
[1] 70.49368
```

```
mean(regression_prediction)
```

```
[1] 70.50941
```

The standard error for the regression prediction is substantially smaller:

```
sd(conditional_avg, na.rm = TRUE)
```

```
[1] 0.9635814
```

```
sd(regression_prediction)
```

```
[1] 0.4520833
```

Variance explained

By bivariate normal theory (You will learn that in STA 131), the standard deviation of the **conditional** distribution is:

$$\text{SD}(Y \mid X = x) = \sigma_Y \sqrt{1 - \rho^2}$$

- To see why this is intuitive, notice that without conditioning, $\text{SD}(Y) = \sigma_Y$, we are looking at the variability of all the sons.
- But once we condition, we are only looking at the variability of the sons with a tall, 72-inch, father. This group will all tend to be somewhat tall so the standard deviation is reduced.
- Specifically, the variance (SD square) reduced to $1 - \rho^2 = 1 - 0.25$ = 0.75 of what it was originally. We could say that father heights **"explain"** 25% of the **variability observed** in son heights.

> But it is important to remember that the "variance explained" statement only makes sense when the data is approximated by a bivariate normal distribution.

Warning: there are two regression lines

We computed a regression line to predict the son's height from father's height. We used these calculations:

```
mu_x <- mean(galton_heights$father)
mu_y <- mean(galton_heights$son)
s_x <- sd(galton_heights$father)
s_y <- sd(galton_heights$son)
r <- cor(galton_heights$father, galton_heights$son)
m_1 <-  r * s_y / s_x
b_1 <- mu_y - m_1*mu_x
```

which gives us the function $\mathrm{E}(Y \mid X = x) = 37.3 + 0.46\ x$.

- What if we want to predict the father's height based on the son's?
- It is important to know that this is not determined by computing the inverse function: $x = \{\mathrm{E}(Y \mid X = x) - 37.3\ \} / 0.5$.
- We need to compute $\mathrm{E}(X \mid Y = y)$. Since the data is approximately bivariate normal, the theory described above tells us that this conditional expectation will follow a line with slope and intercept:

```
m_2 <-  r * s_x / s_y
b_2 <- mu_x - m_2 * mu_y
```

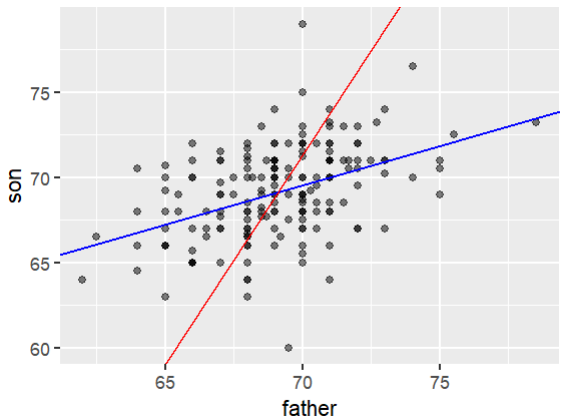
- We need to compute $\mathrm{E}(X \mid Y = y)$

```
m_2 <-  r  s_x / s_y
b_2 <- mu_x - m_2  mu_y
```

So we get $\mathrm{E}(X \mid Y = y) = 40.9 + 0.41y$. Again we see regression to the average: the prediction for the father is closer to the father average than the son heights y is to the son average.

Here is a plot showing the two regression lines, with blue for the predicting son heights with father heights and red for predicting father heights with son heights:

```
galton_heights |>
  ggplot(aes(father, son)) +
  geom_point(alpha = 0.5) +
  geom_abline(intercept = b_1, slope = m_1, col = "blue") +
  geom_abline(intercept = -b_2/m_2, slope = 1/m_2, col = "red")
```



Interpreting Linear models

- One reason linear models are popular is that they are interpretable.

In the case of Galton's data, we can interpret the data like this: due to inherited genes, the son's height prediction grows by β_1 for each inch we increase the father's height x. Because not all sons with fathers of height x are of equal height, we need the term ε, which explains the remaining variability.

This remaining variability includes the mother's genetic effect, environmental factors, and other biological randomness.

- Given how we wrote the model above, the intercept β_0 is not very interpretable as it is the predicted height of a son with a father with no height.

- Due to regression to the mean, the prediction will usually be a bit larger than 0. To make the slope parameter more interpretable, we can rewrite the model slightly as:

$$Y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i,\ i = 1, \ldots, N$$

with $\bar{x} = 1/N \sum_{i=1}^{N} x_i$ the average of the x. In this case β_0 represents the height when $x_i = \bar{x}$, which is the height of the son of an average father.