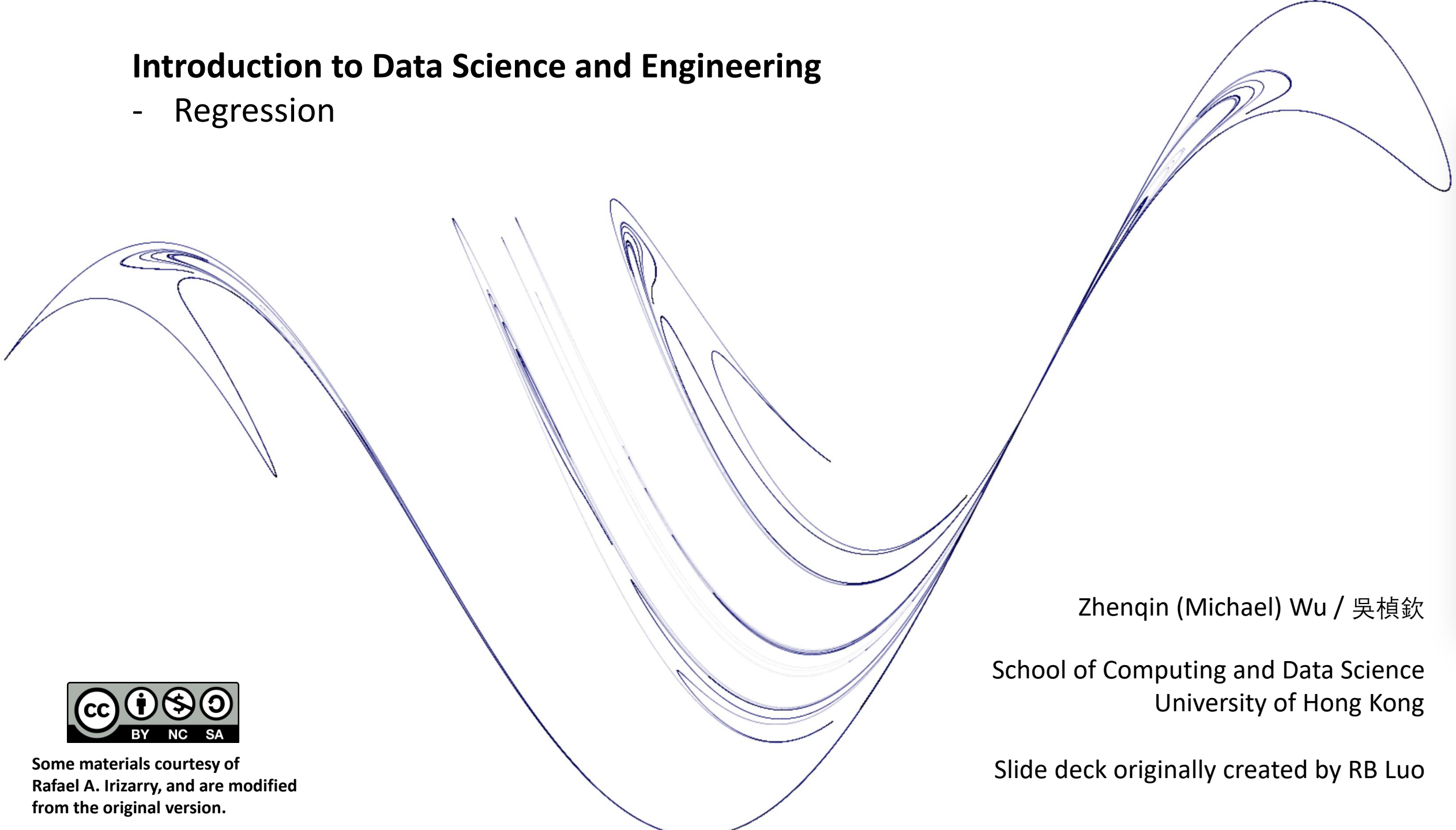
Introduction to Data Science and Engineering

- Regression



Zhenqin (Michael) Wu / 吳楨欽

School of Computing and Data Science
University of Hong Kong

Slide deck originally created by RB Luo

Some materials courtesy of Rafael A. Irizarry, and are modified from the original version.

In this lecture

- Correlation
- Regression
- Association versus causation

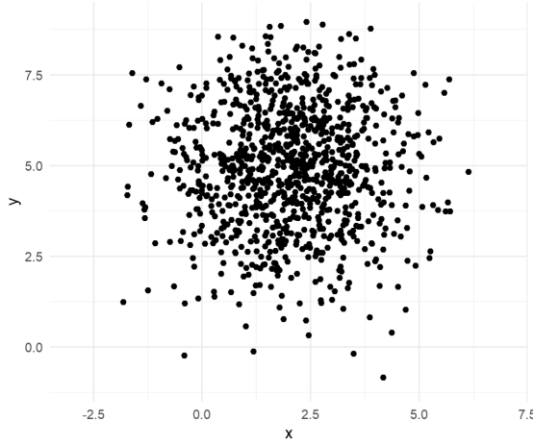
In this lecture

- **Correlation**
- Regression
- Association versus causation

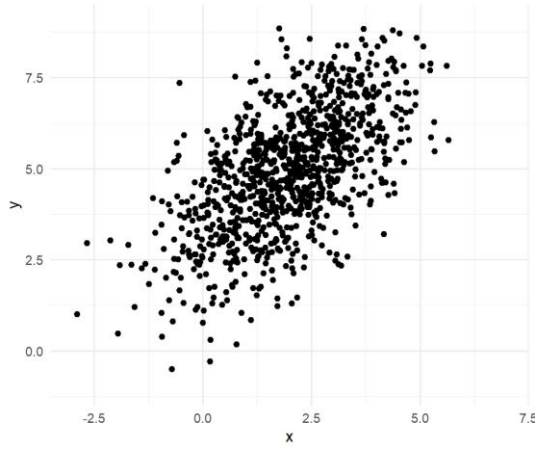
Association / Correlation

- When we have two or more variables, how can we study / characterize their relationship?

```r
mu <- c(2, 5)
sigma <- matrix(c(2, 0, 0, 3), nrow = 2)

set.seed(42)
samples <- mvrnorm(n = 1000, mu = mu, Sigma = sigma)

samples <- data.frame(x = samples[,1], y = samples[,2])
samples |> ggplot(aes(x, y)) +
  geom_point() +
  scale_x_continuous(limits = c(-3, 7)) +
  scale_y_continuous(limits = c(-1, 9)) +
  theme_minimal()
```




```r
sigma_correlated <- matrix(c(2, 1.5, 1.5, 3), nrow = 2)

set.seed(42)
samples <- mvrnorm(n = 1000, mu = mu, Sigma = sigma_correlated)

samples <- data.frame(x = samples[,1], y = samples[,2])
samples |> ggplot(aes(x, y)) +
  geom_point() +
  scale_x_continuous(limits = c(-3, 7)) +
  scale_y_continuous(limits = c(-1, 9)) +
  theme_minimal()
```



Is height hereditary?

```r
data("GaltonFamilies")

galton_heights <- GaltonFamilies |>
  filter(gender == "male") |>
  group_by(family) |>
  sample_n(1) |>  # Sample one child per family
  ungroup() |>
  dplyr::select(father, childHeight) |>
  rename(son = childHeight)
```

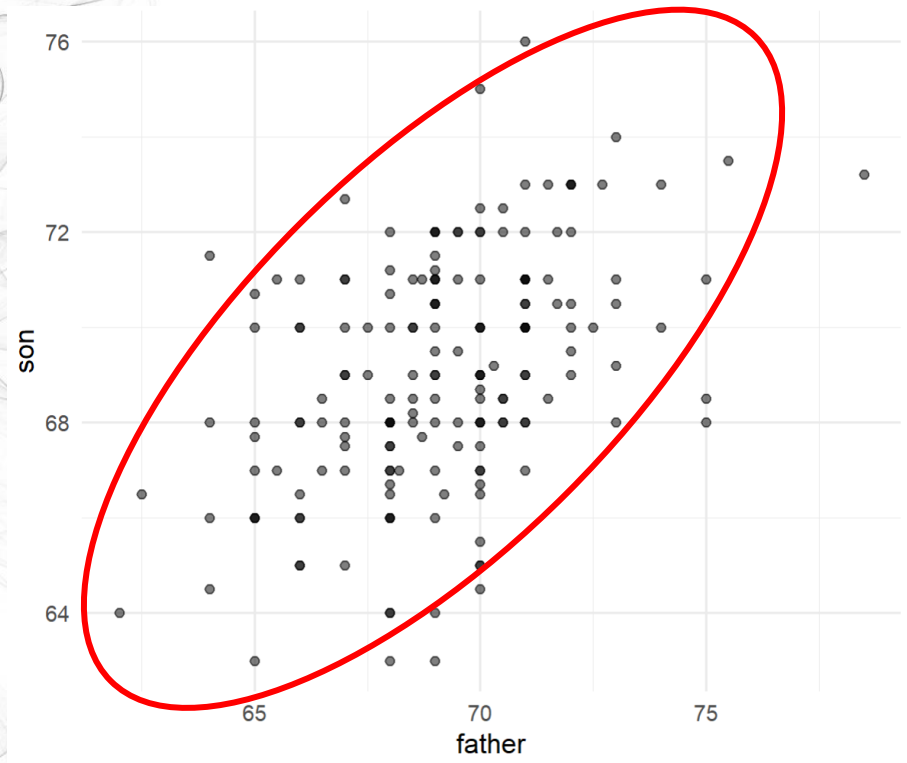
```
> galton_heights |> head(3)
# A tibble: 3 × 2
  father   son
   <dbl> <dbl>
1   78.5  73.2
2   75.5  73.5
3   75    71
```

```
> galton_heights |>
+     summarize(mean(father), sd(father), mean(son), sd(son))
# A tibble: 1 × 4
  `mean(father)` `sd(father)` `mean(son)` `sd(son)`
           <dbl>        <dbl>       <dbl>     <dbl>
1           69.1         2.55        69.3      2.50
```

- We will use Galton's family height data to demonstrate.
- This data contains heights of family members: mothers, fathers, daughters, and sons.
- Fathers' and sons' heights have similar means and standard deviations

Is height hereditary?

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



- We will use Galton's family height data to demonstrate.
- This data contains heights of family members: mothers, fathers, daughters, and sons.
- Fathers' and sons' heights have similar means and standard deviations
- We can roughly see that these two variables have similar trends:
  - The taller the father, the taller the son

Correlation coefficient

$$\rho = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_x}{\sigma_x}\right)\left(\frac{y_i - \mu_y}{\sigma_y}\right)$$

- For two variables, correlation coefficient is summary of how two variables move together
- For a pair of data point x_i and y_i: if they are both larger than mean or smaller than mean, we will have a positive correlation.
- The $(n-1)$ versus n difference, is again, due to degree of freedom.

```
> x <- galton_heights$father
> y <- galton_heights$son
> n <- length(x)
> mean(scale(x) * scale(y))
[1] 0.4485471
> mean((scale(x) * scale(y)) * n / (n-1))
[1] 0.451067
> cor(x, y)
[1] 0.451067
> cor(y, x)
[1] 0.451067
```

Correlation coefficient

$$\rho = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_x}{\sigma_x}\right)\left(\frac{y_i - \mu_y}{\sigma_y}\right)$$

- $\frac{x_i - \mu_x}{\sigma_x}$ is the z-score of x_i; if x follows a normal distribution, this value will be standard normal (0-mean, 1-SD).

- Note that this is also a sample mean; with large n, ρ will converge to its ground truth (LLN), and follows a normal distribution (CLT)

Correlation coefficient

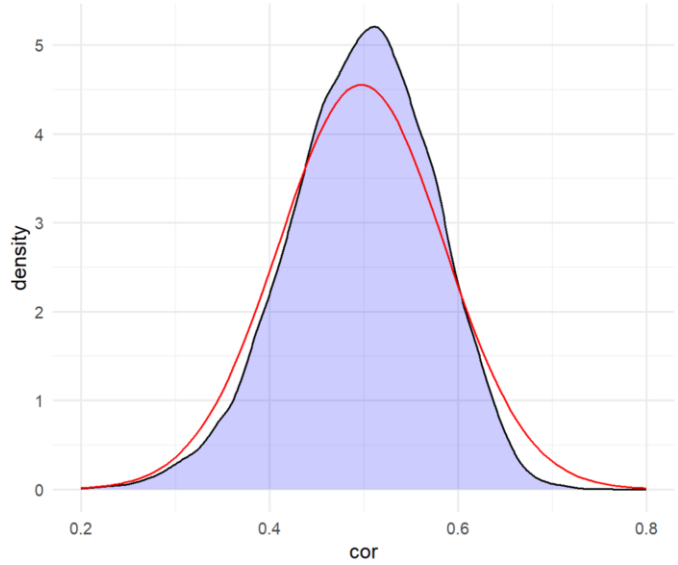
$$\rho = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_x}{\sigma_x}\right)\left(\frac{y_i - \mu_y}{\sigma_y}\right)$$

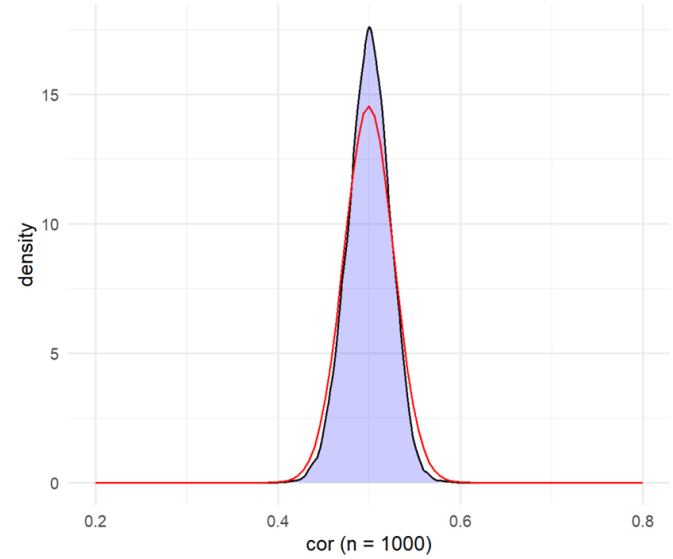
- $\frac{x_i - \mu_x}{\sigma_x}$ is the z-score of x_i; if x follows a normal distribution, this value will be standard normal (0-mean, 1-SD).
- Note that this is also a sample mean; with large n, ρ will converge to its ground truth (LLN), and follows a normal distribution (CLT).
- The normality of ρ does not require normality of x and/or y (CLT).

$$SD(\rho) = \sqrt{\frac{1 - \rho^2}{n - 2}}$$

Correlation coefficient

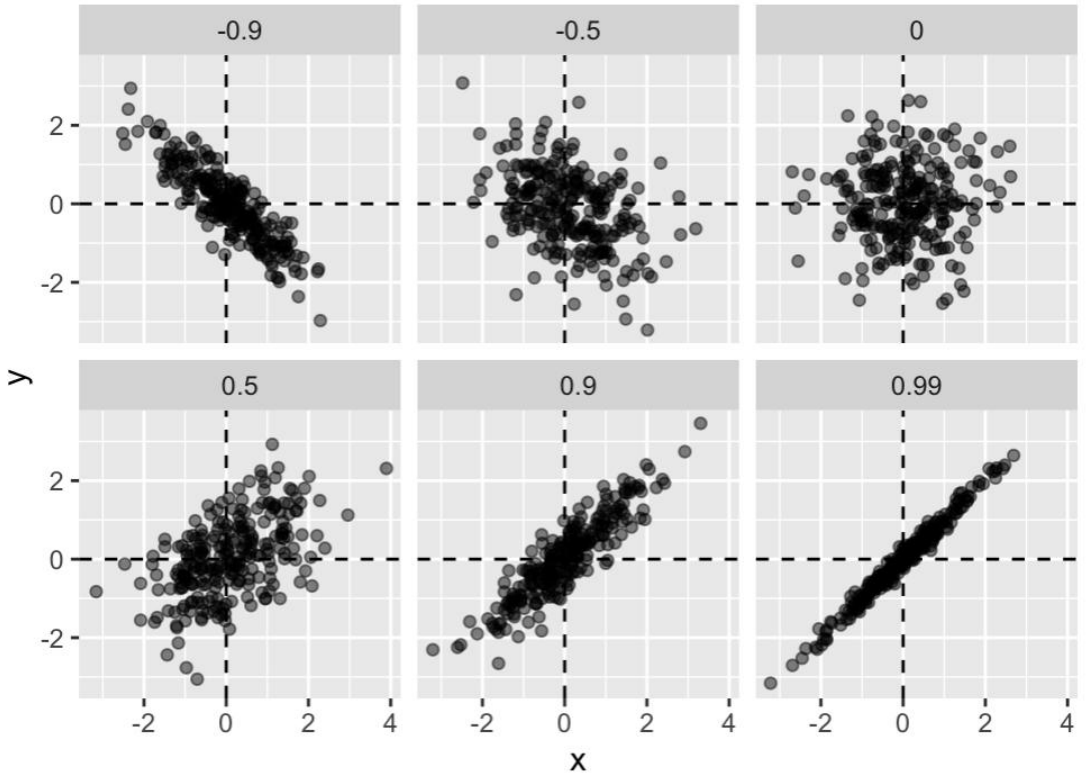
```r
n <- 100
mu <- c(0, 0)
sigma <- matrix(c(1, 0.5, 0.5, 1), nrow = 2)

sample_correlation <- function(){
  data <- mvrnorm(n = n, mu = mu, Sigma = sigma)
  cor(data[,1], data[,2])
}
sampled_cors <- data.frame(cor=replicate(1e4, sample_correlation()))
sampled_cors |>
  summarize(mean=mean(cor), sd=sd(cor))
cor_hat <- mean(sampled_cors$cor)
sampled_cors |>
  ggplot(aes(cor)) +
  geom_density(fill="blue", alpha=0.2) +
  stat_function(
    fun=dnorm,
    args=list(mean=cor_hat, sd=sqrt((1 - cor_hat**2)/(n - 2))),
    color="red") +
  ylab("density") +
  scale_x_continuous(limits = c(0.2, 0.8)) +
  theme_minimal()
```





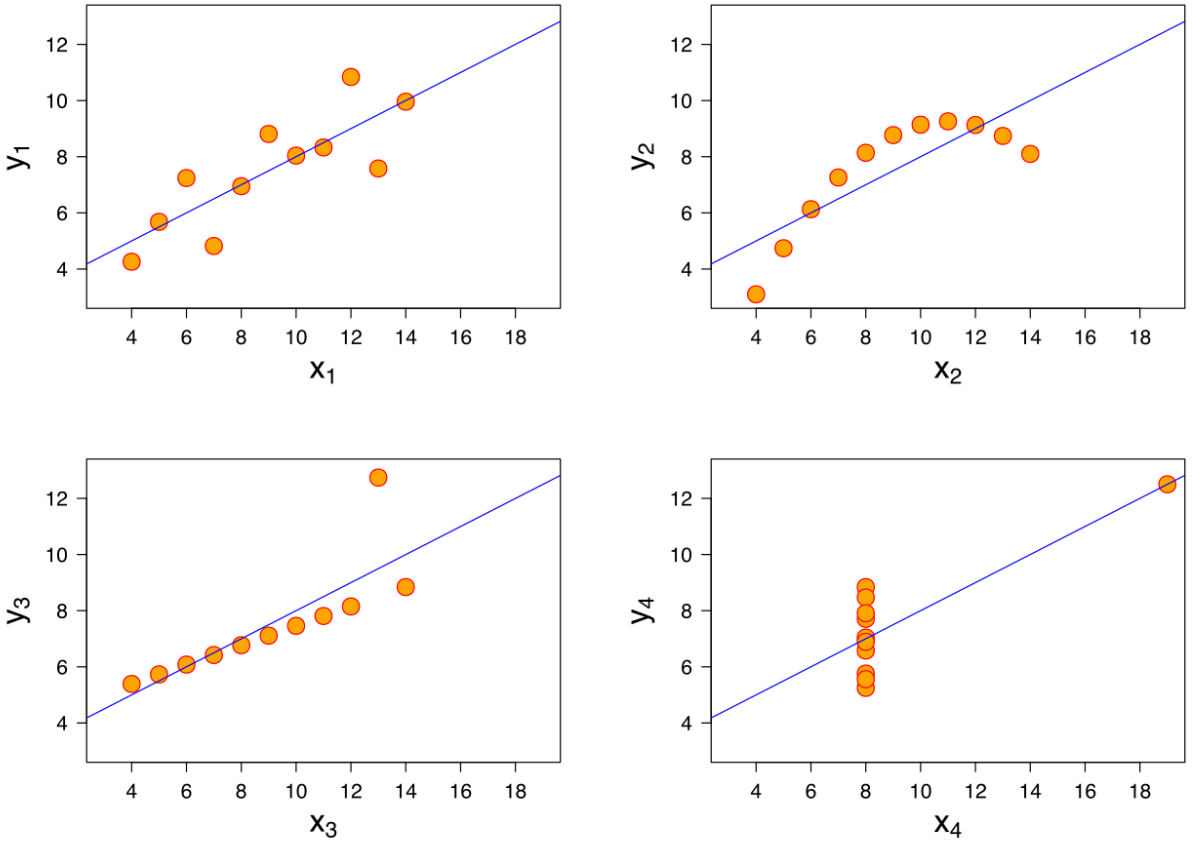
Correlation coefficient

- ρ is always between -1 and 1 (why?).
- `cor(x, x) = 1; cor(x, -x) = -1`
- What data looks like for different values of ρ



Correlation coefficient

- Anscombe's quartet:
- They have almost identical x-mean, x-sd, y-mean y-sd
- All cases have ρ of 0.816

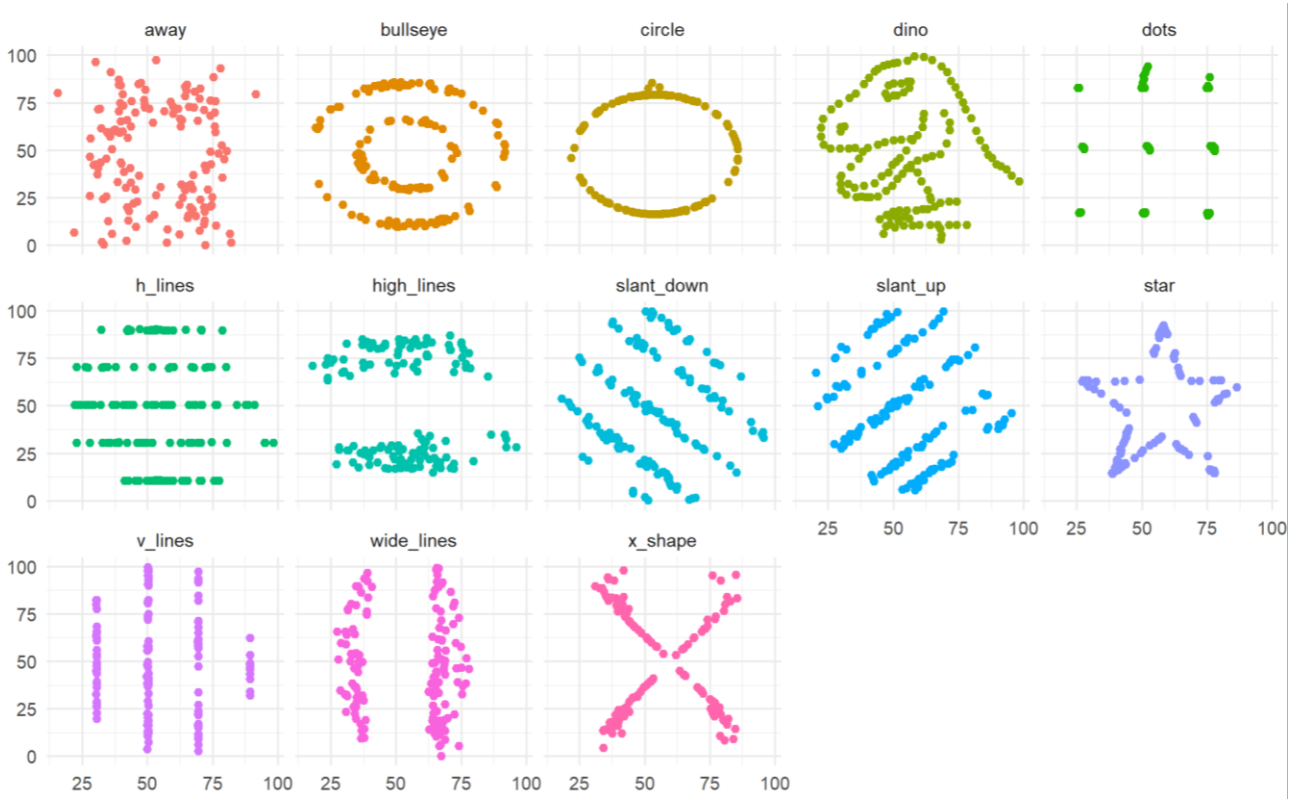


https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Correlation coefficient

```
library(datasauRus)
datasaurus_dozen |>
  ggplot(aes(x, y, color = dataset)) +
  geom_point(show.legend = FALSE) +
  facet_wrap(~dataset, ncol = 5) +
  theme_minimal()
```

- Almost identical x-mean, x-sd, y-mean y-sd
- Very similar ρ



https://juliasilge.com/blog/datasaurus-multiclass/

Correlation coefficient

How to not misinterpret:

- **Visualize**: scatter plot is always a good helper
- Other correlation coefficients:
  - ρ is also commonly known as Pearson correlation coefficient
  - Spearman's rank correlation coefficient: rank instead of z-score

```
cor(x, y, method="spearman")
```

https://juliasilge.com/blog/datasaurus-multiclass/

In this lecture

- Correlation
- **Regression**
- Association versus causation

The regression line

- For a father who is 180cm tall, how tall will his son be?
  - We can look at data from the male population that are around 180cm tall and summarize their sons' heights (modeling the conditional distribution).
  - Alternatively, we can build a regression model if we believe that there is a linear association between fathers' heights (x) and sons' heights (y) .

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

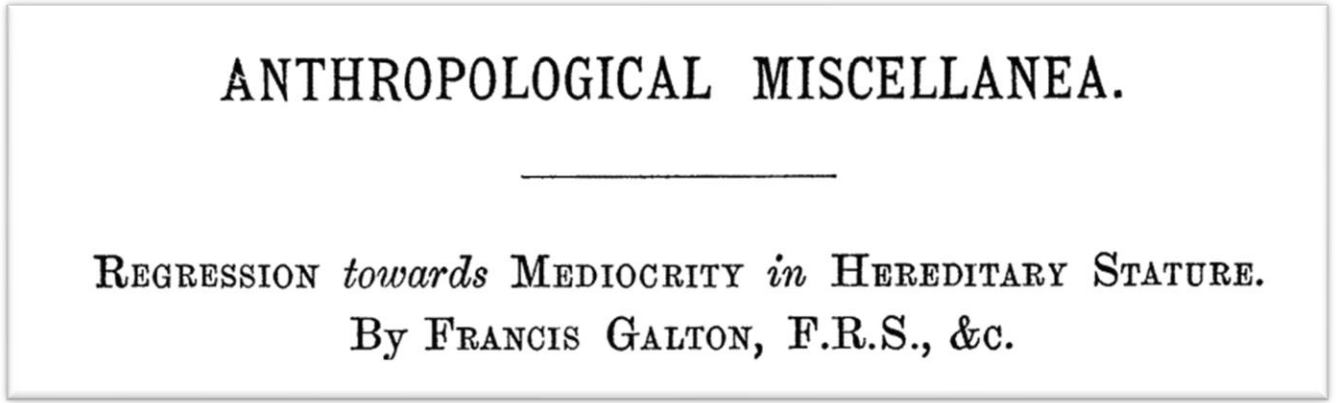
$$Y=\mu_Y+\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\sigma_Y$$

The regression line

- If we use $y = b + mx$ to model the relationship between these two variables:

$$m = \rho \frac{\sigma_Y}{\sigma_X} \qquad\qquad b = \mu_Y - m\mu_X$$

- In the father and son case, SD of heights are the same, because $\rho \in [-1, 1]$; y is always going to be closer to its mean than x



ANTHROPOLOGICAL MISCELLANEA.

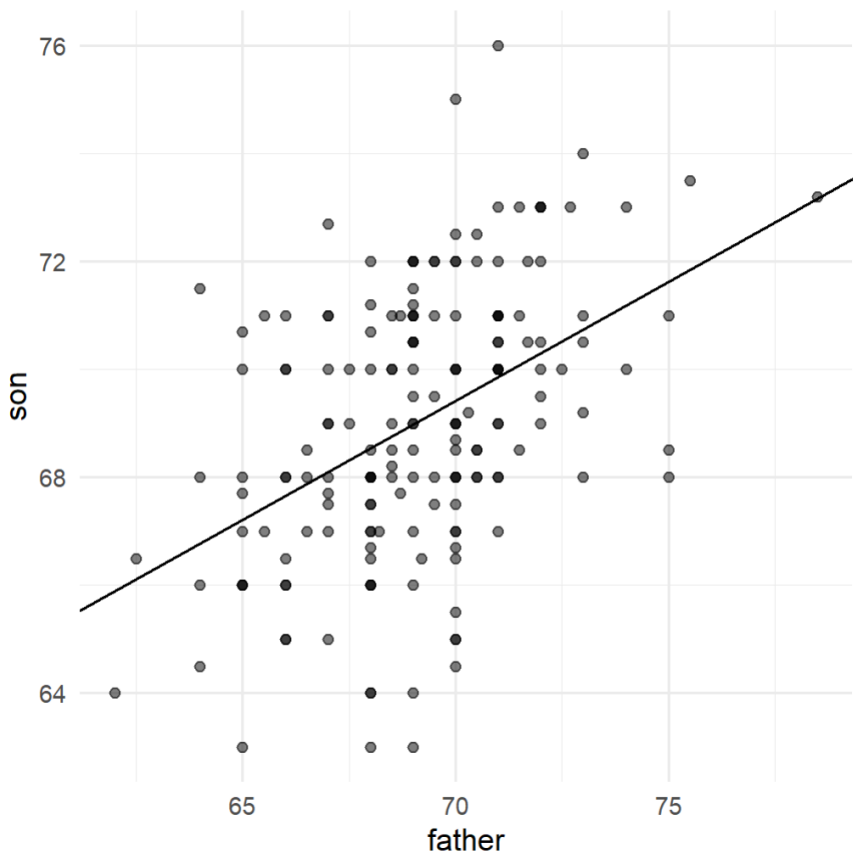
REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.

By FRANCIS GALTON, F.R.S., &c.

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

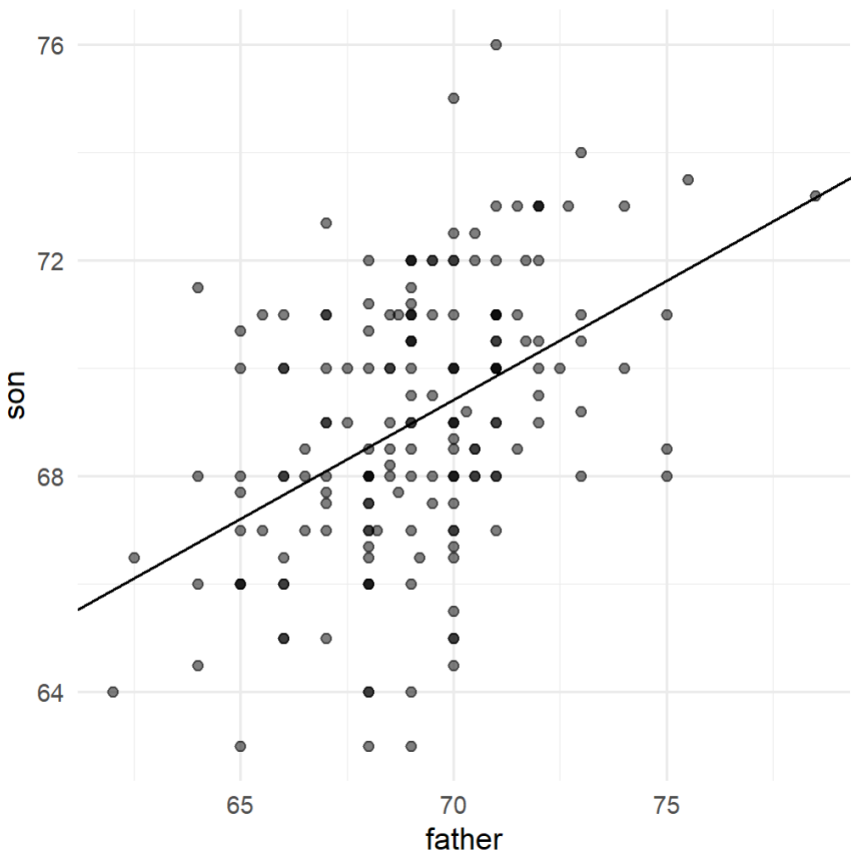
The regression line

```
mu_x <- mean(x)
s_x <- sd(x)
mu_y <- mean(y)
s_y <- sd(y)
rho <- cor(x, y)

galton_heights |>
  ggplot(aes(father, son)) +
  geom_point(alpha=0.5) +
  geom_abline(
    slope=rho * s_y/s_x,
    intercept=mu_y - mu_x * (rho * s_y/s_x)) +
  theme_minimal()
```



The regression line

```r
model <- lm(son ~ father, data=galton_heights)
galton_heights |>
  ggplot(aes(father, son)) +
  geom_point(alpha=0.5) +
  geom_abline(
    slope=model$coefficients[2],
    intercept=model$coefficients[1]) +
  theme_minimal()
```




```r
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = T
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

Arguments

`formula`	an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'.
`data`	an optional data frame, list or environment (or object coercible by `as.data.frame` to a data frame) containing the variables in the model. If not found in `data`, the variables are taken from `environment(formula)`, typically the environment from which `lm` is called.

The regression line

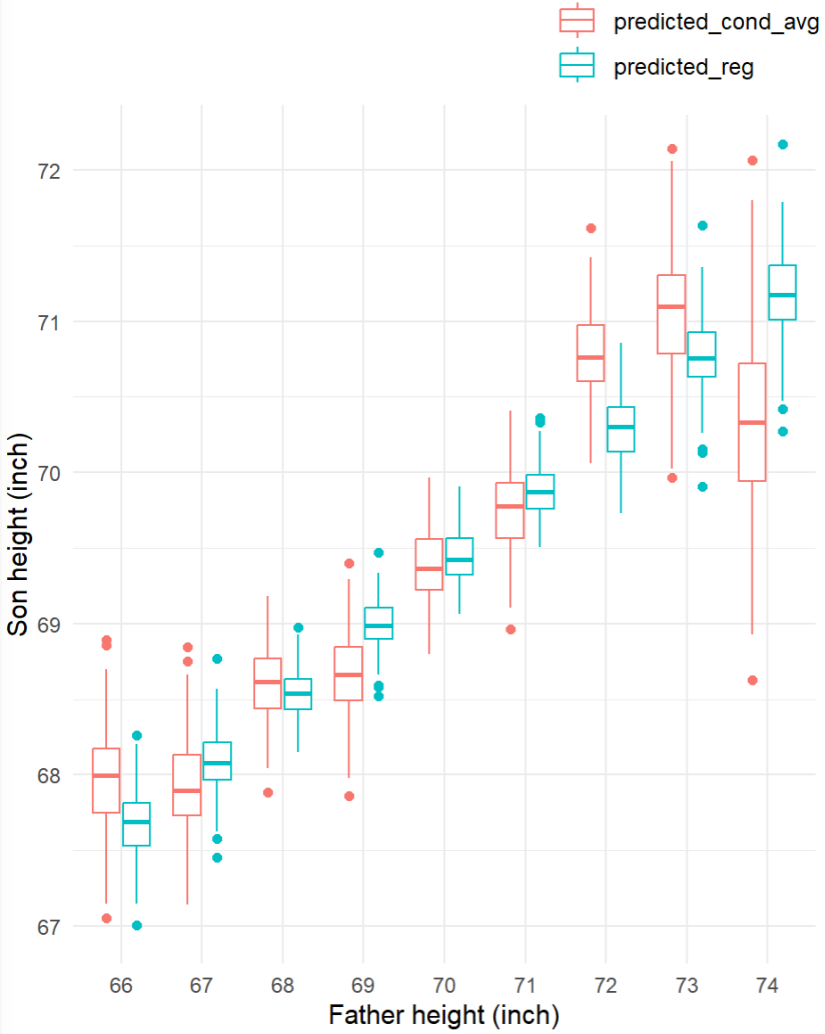
- Both the slope m and the intercept b are random variables, and they follow CLT.

$$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Derived from Ordinary Least Squares (OLS)
- The parameters will, therefore, have uncertainty => the predictions based on the regression line will have uncertainty.

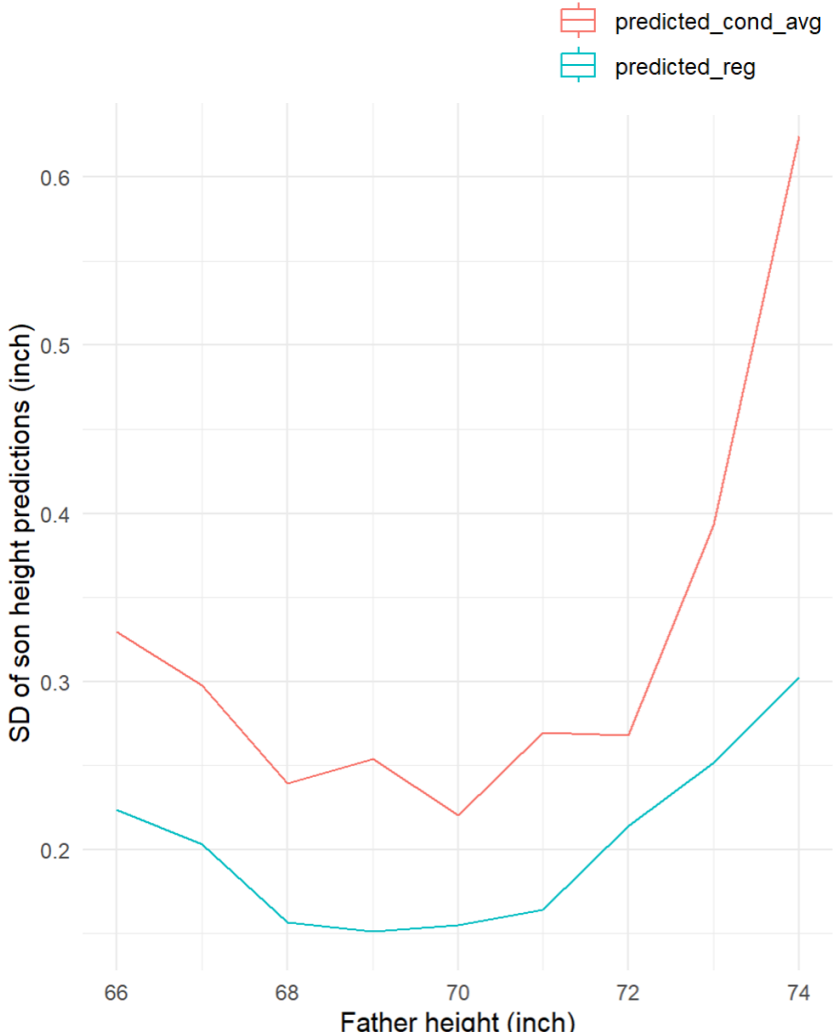
Regression improves prediction precision

```r
rep <- 200
N <- 100

predict_using_conditional_avg <- function(father_height){
  dat <- sample_n(galton_heights, N)
  dat |> filter(between(father, father_height-1, father_height+1)) |>
    summarize(avg = mean(son)) |>
    pull(avg)
}

predict_using_regression <- function(father_height){
  dat <- sample_n(galton_heights, N)
  model <- lm(son ~ father, data=dat)
  predict(model, newdata=data.frame(father=c(father_height)))
}

predictions <- expand.grid(father_height=seq(66, 74, 1), i_sample=seq(1:rep))
predictions <- predictions |>
  mutate(predicted_cond_avg = sapply(father_height, predict_using_conditional_avg)) |>
  mutate(predicted_reg = sapply(father_height, predict_using_regression))

predictions |>
  pivot_longer(c("predicted_cond_avg", "predicted_reg")) |>
  filter(!is.na(value)) |>
  mutate(father_height=factor(father_height)) |>
  ggplot(aes(x=father_height, y=value, color=name)) +
  geom_boxplot() +
  xlab("Father height (inch)") +
  ylab("Son height (inch)") +
  theme_minimal()
```



Regression improves prediction precision

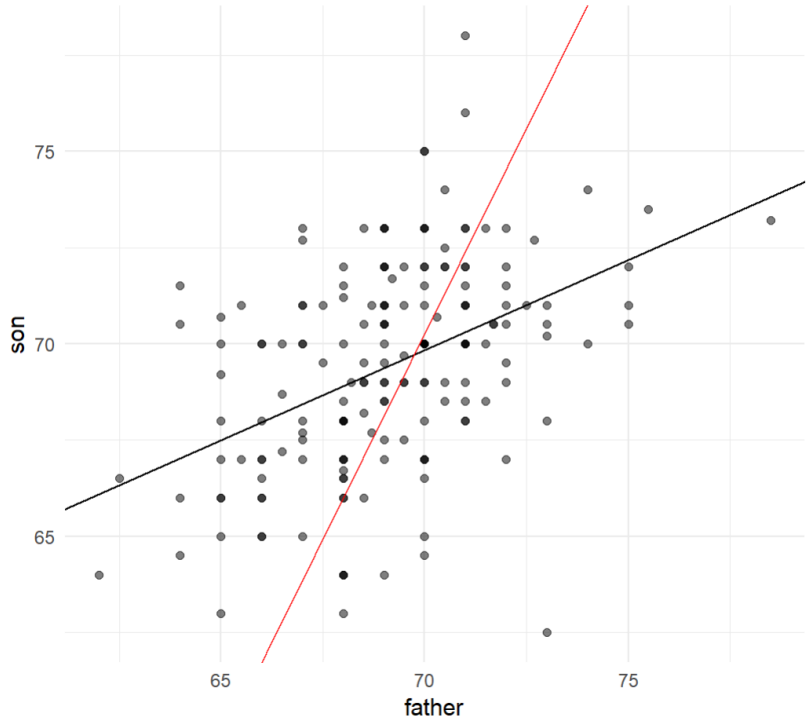
```
predictions |>
  pivot_longer(c("predicted_cond_avg", "predicted_reg")) |>
  group_by(father_height, name) |>
  summarize(prediction_sd=sd(value)) |>
  ggplot(aes(x=father_height, y=prediction_sd, color=name)) +
  geom_line() +
  xlab("Father height (inch)") +
  ylab("SD of son height predictions (inch)") +
  theme_minimal()
```



The regression line

- Regression coefficients are not commutative:
  - The regression line $y = b + mx$ does NOT translate to $x = \frac{y}{m} - \frac{b}{m}$

```
> model <- lm(son ~ father, data=galton_heights)
> coef(model)
(Intercept)      father
 37.0671243   0.4684554
> model <- lm(father ~ son, data=galton_heights)
> coef(model)
(Intercept)         son
 40.5427394   0.4112533
```



Wait, how come that I am taller than my dad?

- The underlying formula is actually:

$$y = b + mx + \epsilon$$

- Here ϵ (residue) is the part of y that is independent of (cannot be explained by x. (e.g., mother's height)

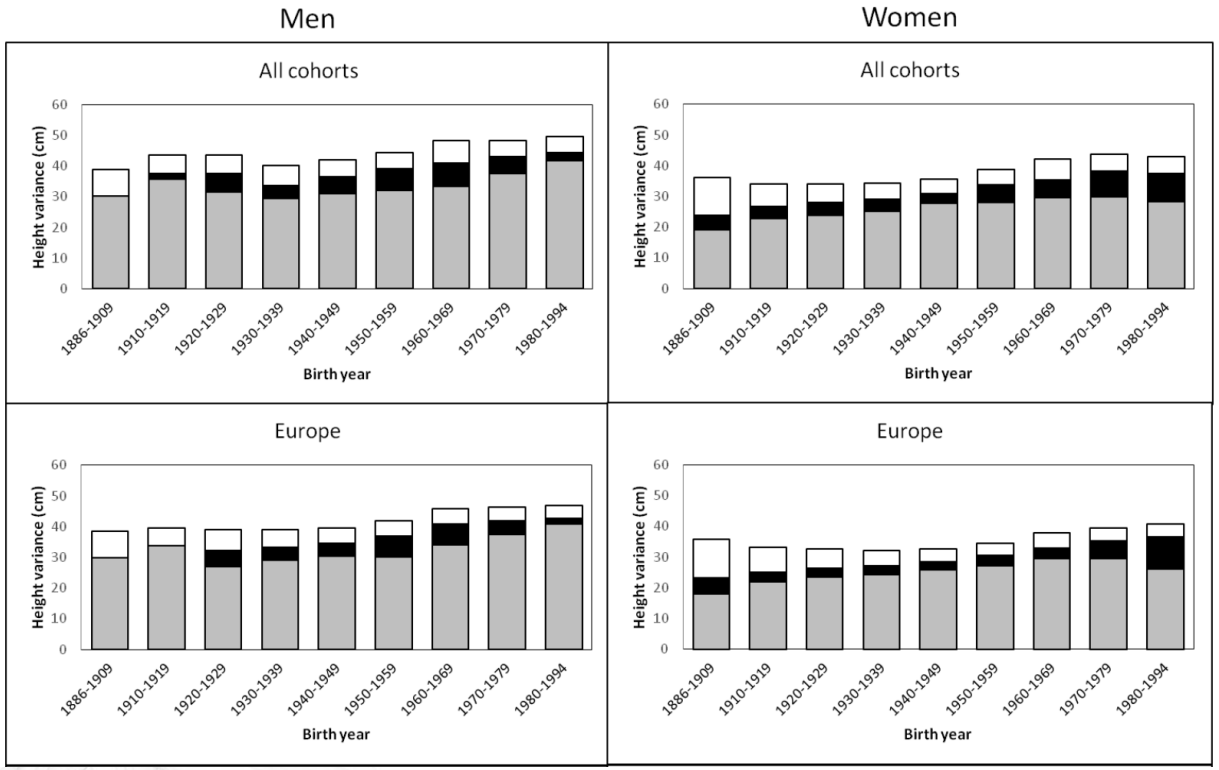
$$Var(y) = m^2 Var(x) + Var(\epsilon)$$

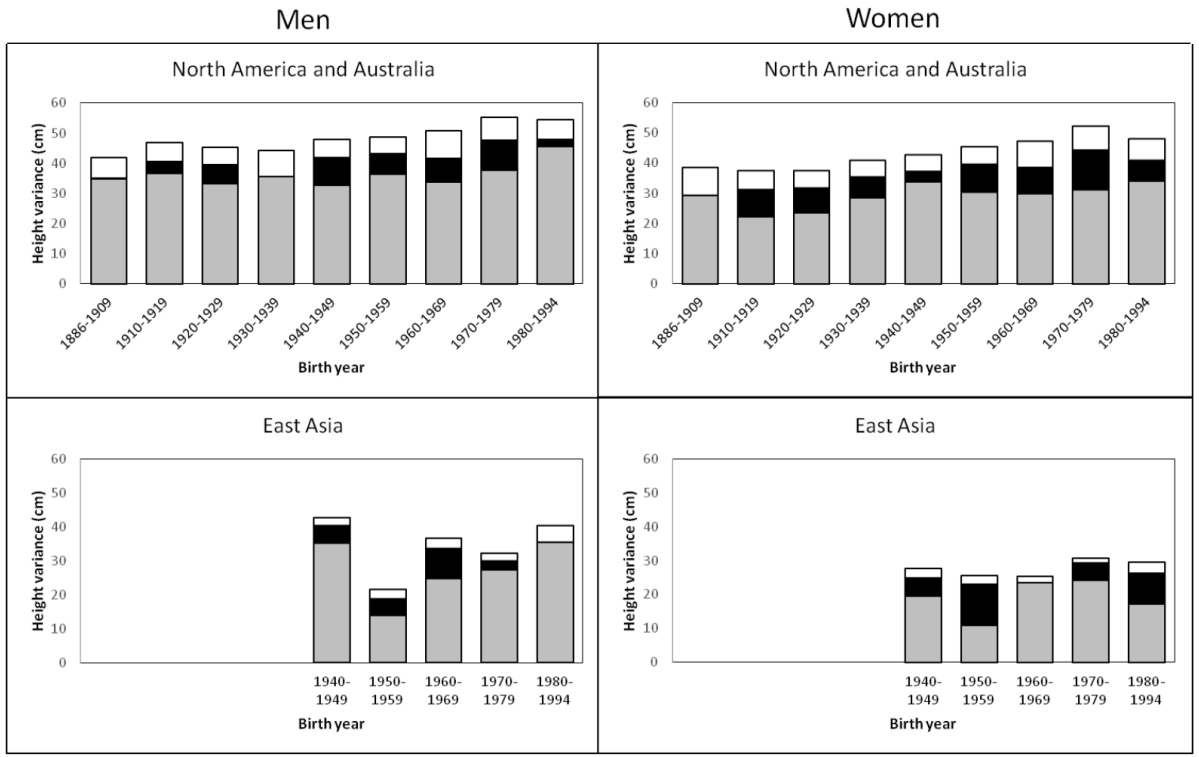
$$Var(y) = (\rho \frac{\sigma_Y}{\sigma_X})^2 Var(x) + Var(\epsilon)$$

$$Var(y) = \rho^2 Var(y) + Var(\epsilon)$$

- Fathers' heights explained ρ^2 (~25%) of the variance of sons' heights

Is height hereditary?





- Additive genetic (grey), shared environmental (black) and unique environmental (white) variances of height across birth-year cohorts for the pooled data and by geographic-cultural region.
- (TL;DR) Yes, 70~80% is explained by genetic factors.

Jelenkovic, Aline, et al. "Genetic and environmental influences on adult human height across birth cohorts from 1886 to 1994." Elife 5 (2016): e20320.

Is height hereditary?

- Baseline: average of your mother's and father's heights, and:
  - Male: +6.5 cm
  - Female: -6.5cm

- What's your delta from the baseline? Share it through the quiz ☺

In this lecture

- Correlation
- Regression
- **Association versus causation**

If there’s only one thing you remember after 10 years, I hope that will be:

Association is not causation

Association is not causation

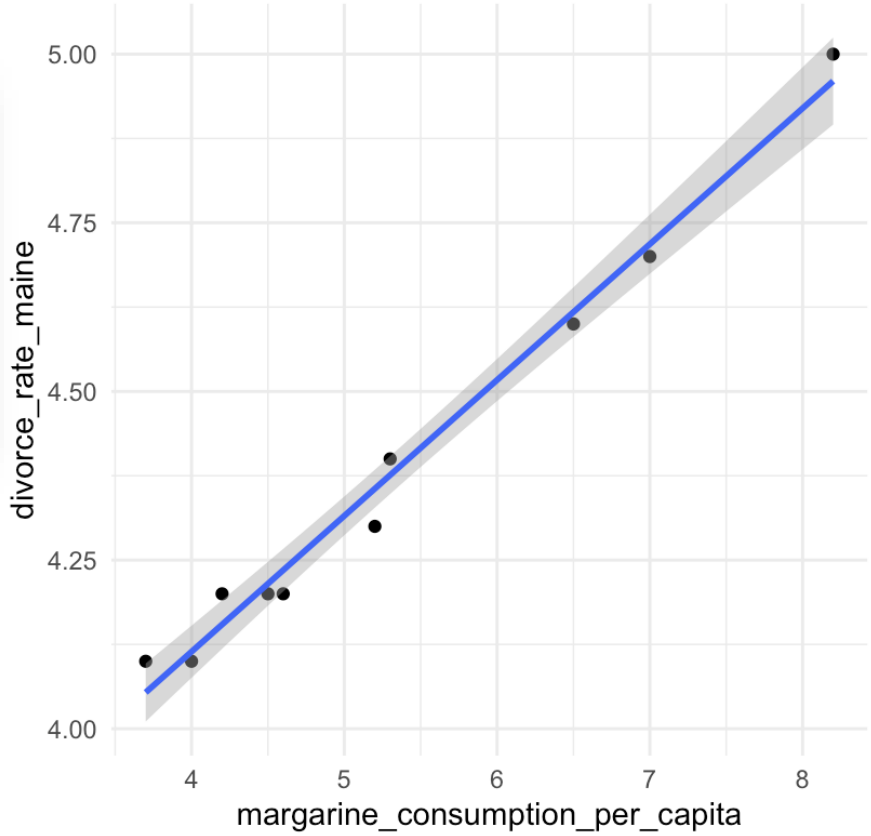
Association/Correlation ≠ Causation

Causality is an influence by which one event, process, state, or object (a cause) **contributes to** the production of another event, process, state, or object (an effect) where **the cause is at least partly responsible for the effect**, and the effect is at least partly dependent on the cause.

Spurious correlation

```
data("divorce_margarine")
divorce_margarine |> head(3)

divorce_margarine |>
  ggplot(aes(margarine_consumption_per_capita, divorce_rate_maine)) +
  geom_point() +
  geom_smooth(method="lm") +
  theme_minimal()
```

```
> divorce_margarine |> head(3)
  divorce_rate_maine margarine_consumption_per_capita year
1                5.0                               8.2 2000
2                4.7                               7.0 2001
3                4.6                               6.5 2002
```



More examples at https://tylervigen.com/spurious-correlations

Spurious correlation

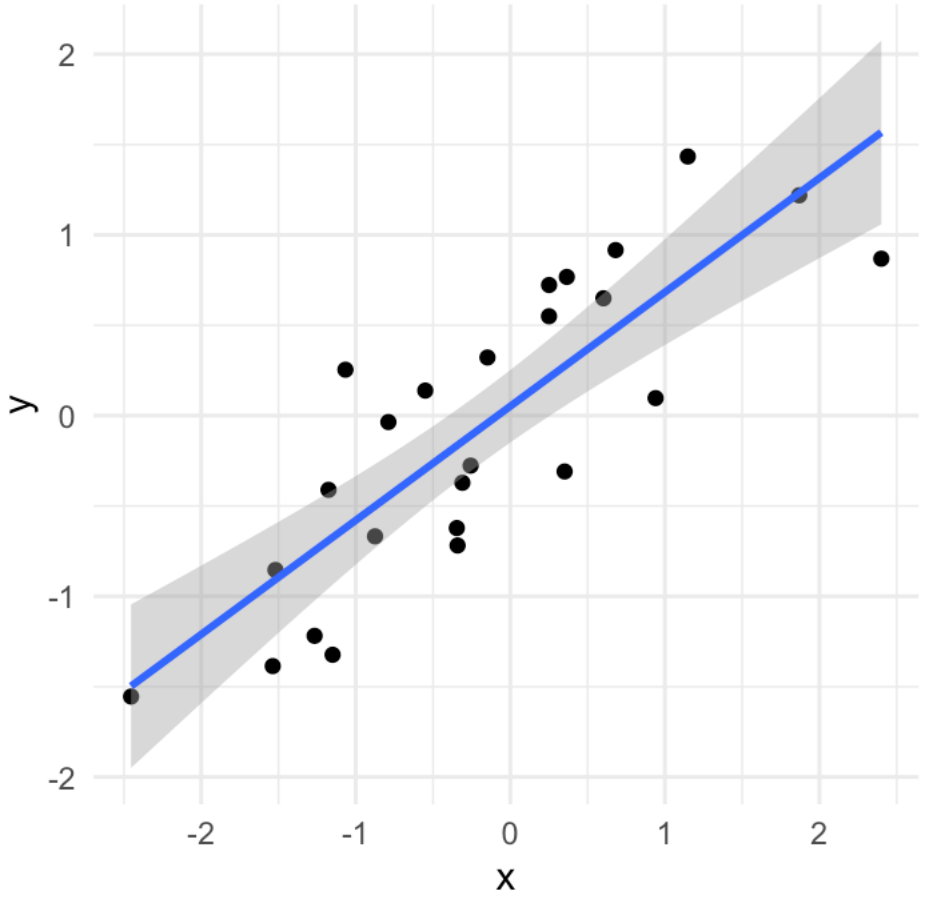
```r
set.seed(42)
N <- 25
g <- 1000000
sim_data <- data.frame(
  group = rep(1:g, each=N),
  x = rnorm(N * g),
  y = rnorm(N * g))
sim_data |>
  group_by(group) |>
  summarize(cor = cor(x, y)) |>
  arrange(desc(cor)) |>
  head(5)
```

```
# A tibble: 5 × 2
   group   cor
   <int> <dbl>
1 117480 0.834
2 529225 0.810
3 548963 0.782
4 909788 0.781
5 550169 0.779
```

- Here we created 1 million groups: each group contains 25 data points.
- They don't have causal relationship. Actually, they are independent.
- However, when computing the correlation between X and Y, many showed pretty strong correlations.

Spurious correlation

```
sim_data |>
  filter(group == 117480) |>
  ggplot(aes(x, y)) +
  geom_point() +
  geom_smooth(method="lm") +
  theme_minimal()
```



- We can do statistical test to verify if null hypothesis (no correlation) can be rejected:

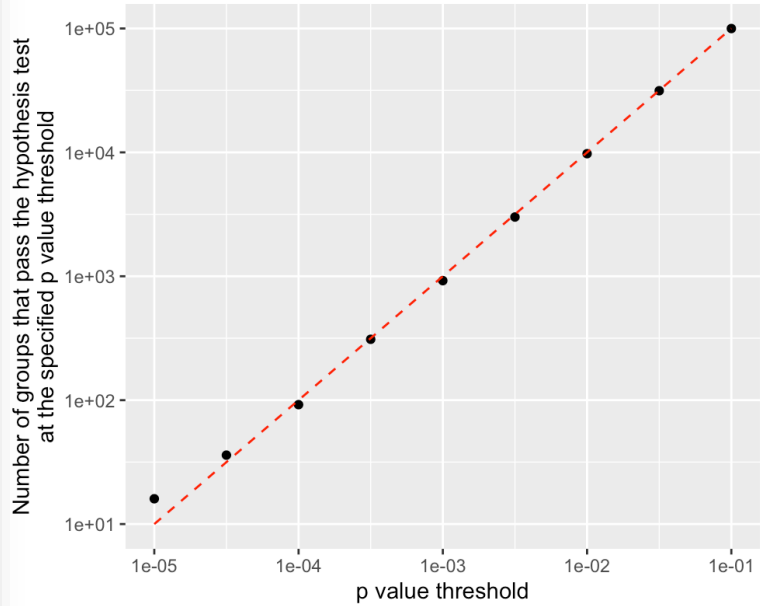
$$\mathrm{SD}(\rho) = \sqrt{\frac{1-\rho^2}{n-2}}$$

$t = \rho/\mathrm{SD}(\rho)$, test against the t-distribution

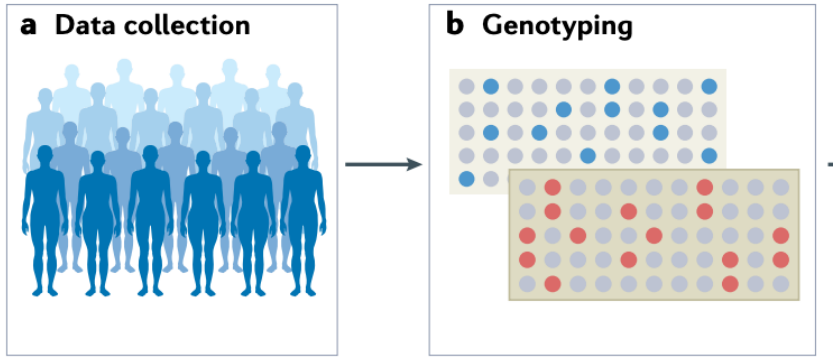
```
> rho <- cor(subset$x, subset$y)
> t_stat <- rho / sqrt((1 - rho**2)/(N - 2))
> print(t_stat)
[1] 7.241302
> p_value <- 2 * pt(t_stat, df=N-1, lower.tail=FALSE)
> print(p_value)
[1] 1.756772e-07
```

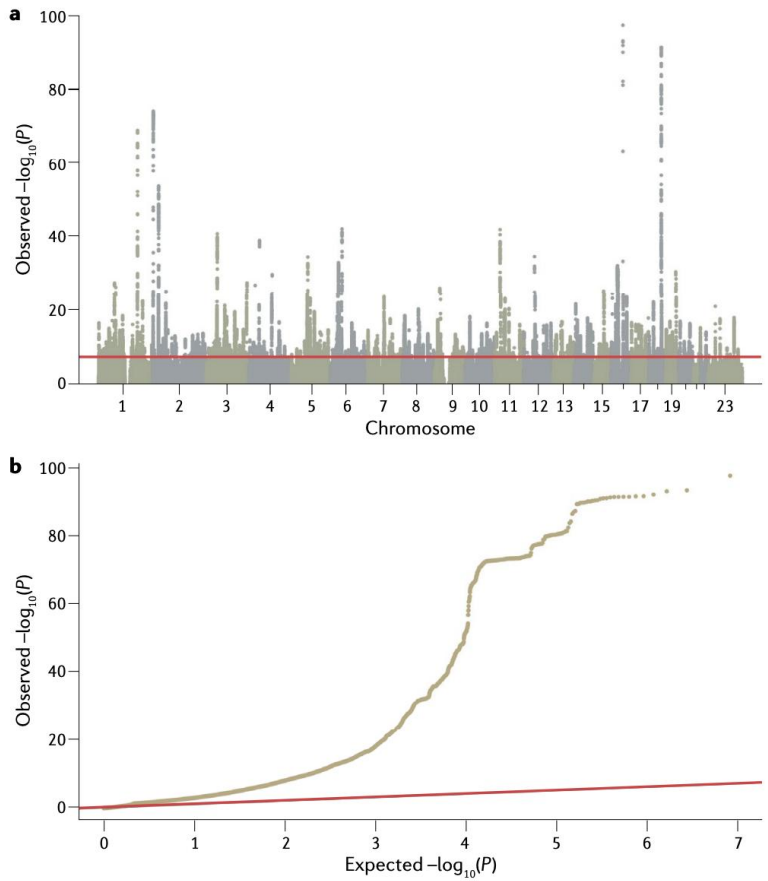
What does p-value mean?

```r
sim_data_with_p_value <- sim_data |>
  group_by(group) |>
  summarize(cor_pvalue = cor.test(x, y, method = "pearson")$p.value)

get_n_points_at_p_threshold <- function(p){
  sim_data_with_p_value |>
    filter(cor_pvalue <= p) |>
    pull(group) |>
    length()
}
df <- data.frame(p_thresholds = 10 ** seq(-5, -1, 0.5))
df |>
  mutate(n_points = sapply(p_thresholds, get_n_points_at_p_threshold)) |>
  ggplot(aes(p_thresholds, n_points)) +
  geom_point() +
  stat_function(fun = function(x) 1e6 * x, color = "red", linetype = "dashed") +
  scale_x_log10() +
  scale_y_log10() +
  xlab("p value threshold") +
  ylab("Number of groups passing the hypothesis test at the specified p-value thresholds")
```

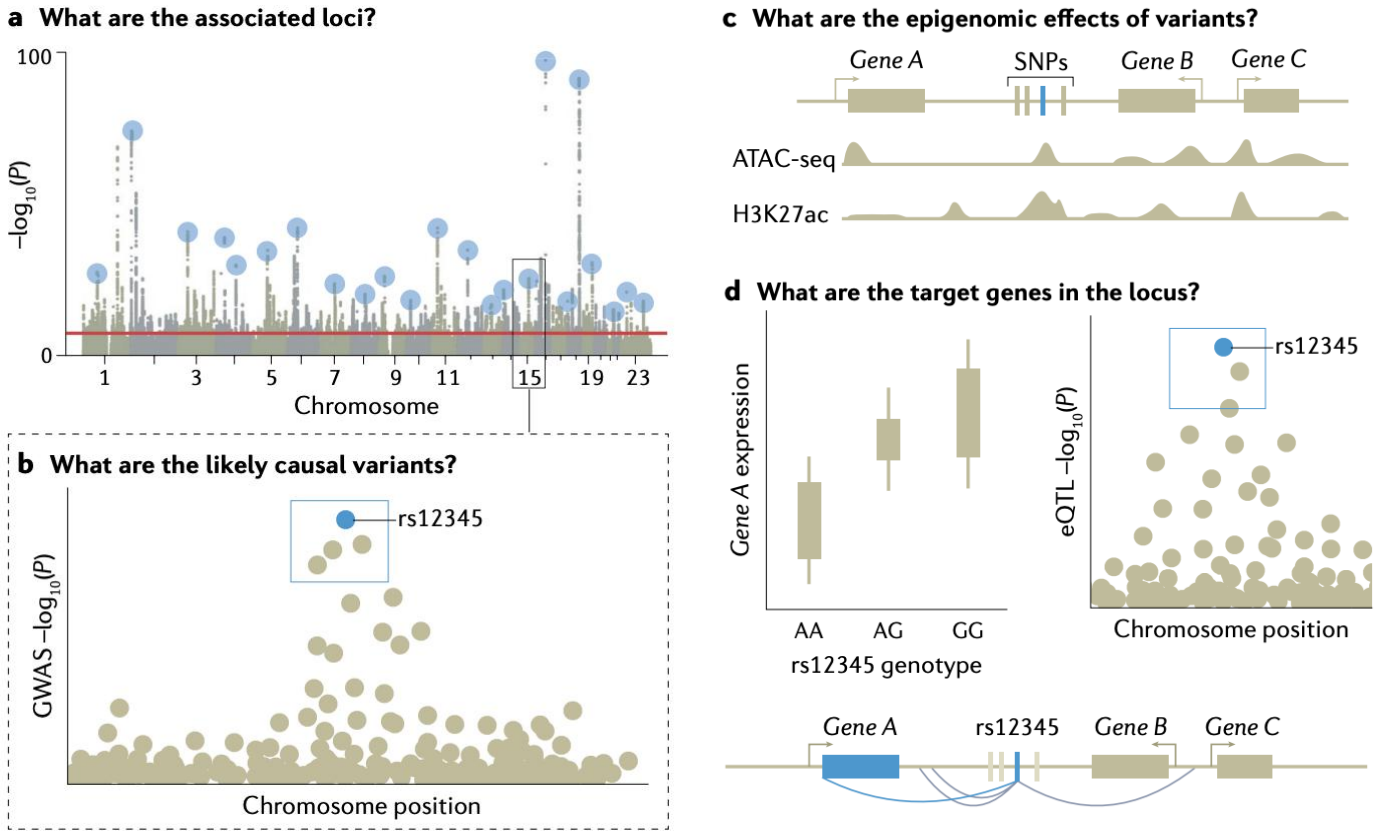


- This applies to all other statistical tests.
- When we say p-value = 0.01, then on average (expectation) 1 out of 100 randomly generated datasets will pass the test.





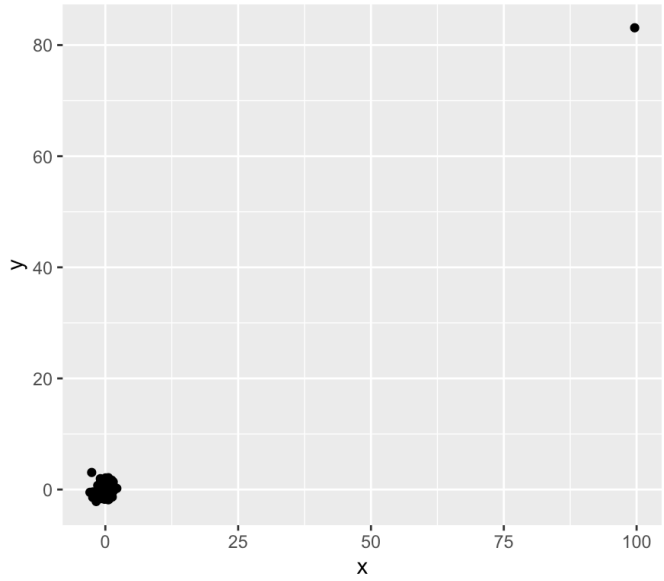
Genetic variants associated with BMI



GWAS: In genomics, a genome-wide association study is an observational study of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait (e.g., disease).

Uffelmann, Emil, et al. "Genome-wide association studies." *Nature Reviews Methods Primers* 1.1 (2021): 59.

Outliers




```
> set.seed(42)
> x <- rnorm(100,100,1)
> y <- rnorm(100,84,1)
> x[-42] <- scale(x[-42])
> y[-42] <- scale(y[-42])
> qplot(x, y)
```

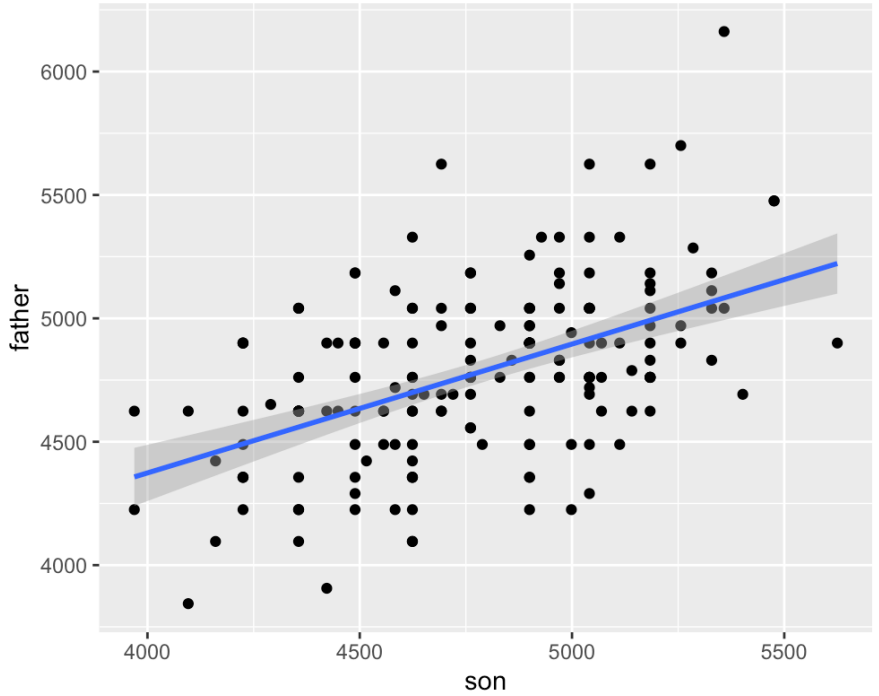
```
> cor(x,y)
[1] 0.9883282
```

```
> cor(x[-42], y[-42])
[1] 0.02799956
```

- Measurements from two independent outcomes, *X* and *Y*, and we standardize the measurements.
- However, imagine we make a mistake and forget to standardize entry 42.
- The correlation is very high, driven by a single outlier

Reversing cause and effect

- If we regress father's height on son's height, the correlation/slope will be significantly positive.
- Even though this correlation is true, we cannot say that it is because son is tall so that father is tall.



Confounders

- Confounders are perhaps the most common reason that leads to associations begin misinterpreted
- If X and Y are correlated, we call Z a confounder if changes in Z causes changes in both X and Y.

  X: Length of queue at MTR exit

  Y: Length of queue at lunch

  Z: number of students coming to school today

- The tricky part of confounders is that we never know if there is another confounder that we are not aware of.

Confounders

- 1973 admission data of six U.C. Berkeley majors show that more men were admitted than women: 44% men were admitted compared to 30% women.

```
> data(admissions)
> two_by_two <- admissions |> group_by(gender) |>
+     summarize(total_admitted = round(sum(admitted / 100 * applicants)),
+               not_admitted = sum(applicants) - sum(total_admitted)) |>
+     select(-gender)
>
> chisq.test(two_by_two)$p.value
[1] 1.055797e-21
```

- But closer inspection shows a paradoxical result. Here are the percent admissions by major

```
> admissions |> select(major, gender, admitted) |>
+     pivot_wider(names_from = "gender", values_from = "admitted") |>
+     mutate(women_minus_men = women - men)
# A tibble: 6 × 4
  major   men women women_minus_men
  <chr> <dbl> <dbl>           <dbl>
1 A        62    82              20
2 B        63    68               5
3 C        37    34              -3
4 D        33    35               2
5 E        28    24              -4
6 F         6     7               1
```

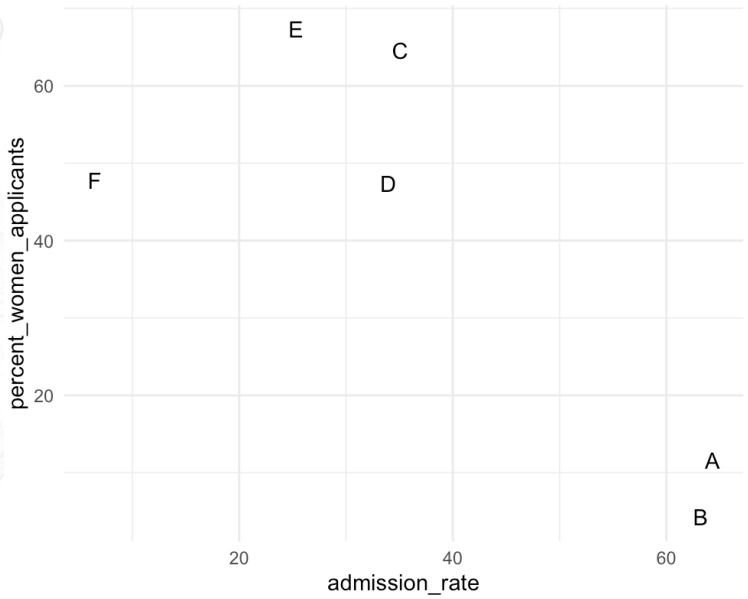
Four out of the six majors favor women

Bickel, Peter J., Eugene A. Hammel, and J. William O'Connell. "Sex Bias in Graduate Admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation." *Science*187.4175 (1975): 398-404.

Confounders

- Major is a confounder here that affects both male/female applicant ratio and admission rate.

```
> admissions |>
+   group_by(major) |>
+   summarize(admission_rate = sum(admitted * applicants)/sum(applicants),
+             percent_women_applicants = sum(applicants * (gender=="women")) /
+               sum(applicants) * 100) |>
+   ggplot(aes(admission_rate, percent_women_applicants, label = major)) +
+   geom_text() +
+   theme_minimal()
```

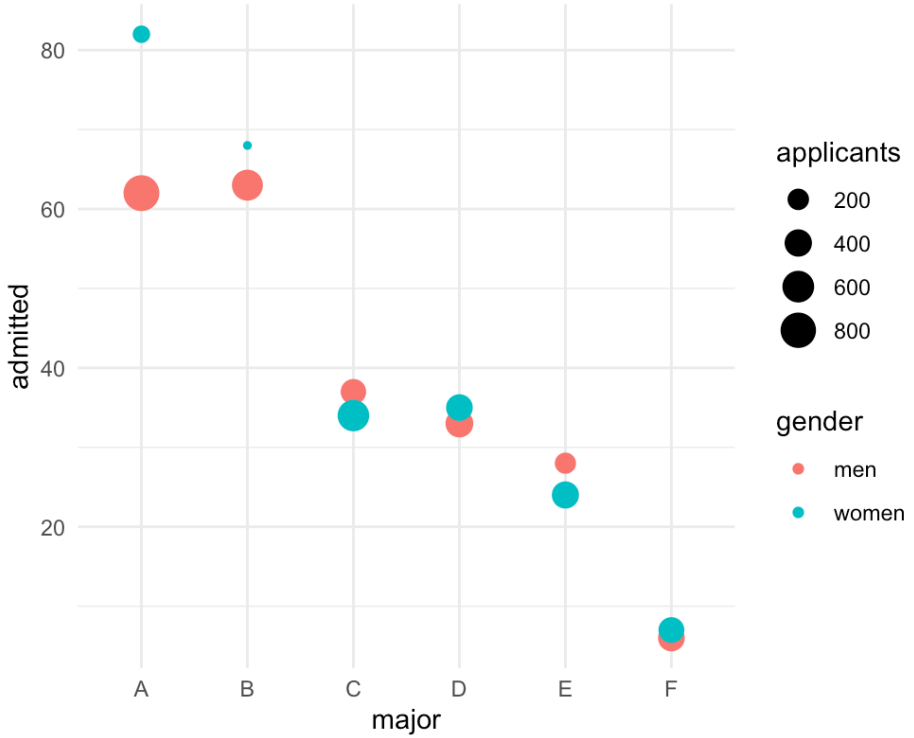


- "Easy" majors (A, B) have predominantly male applicants
- Female students prefer "hard" majors (C, E)

Confounders

- Conditioning on major, there doesn't seem to be a major gender bias.

```
> admissions |>
+   ggplot(aes(major, admitted,
+              col = gender, size = applicants)) +
+   geom_point() +
+   theme_minimal()
```

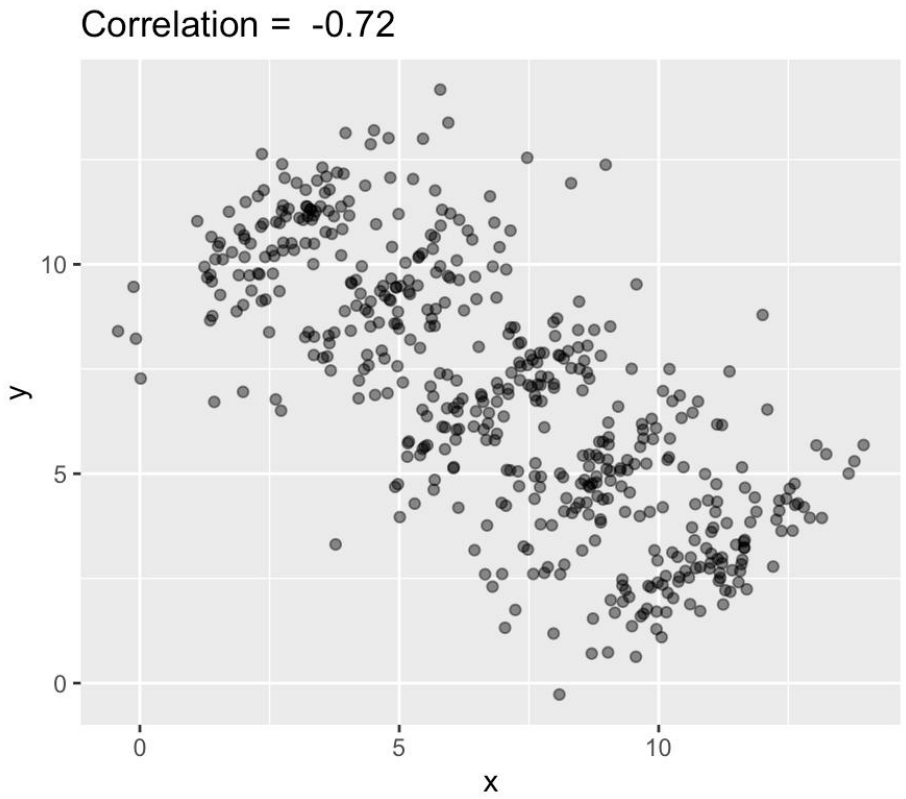


- If we average the difference by major, we find that the percent is actually 3.5% higher for women

```
> admissions |>  group_by(gender) |> summarize(average = mean(admitted))
# A tibble: 2 × 2
  gender average
  <chr>    <dbl>
1 men       38.2
2 women     41.7
```

Simpson's paradox

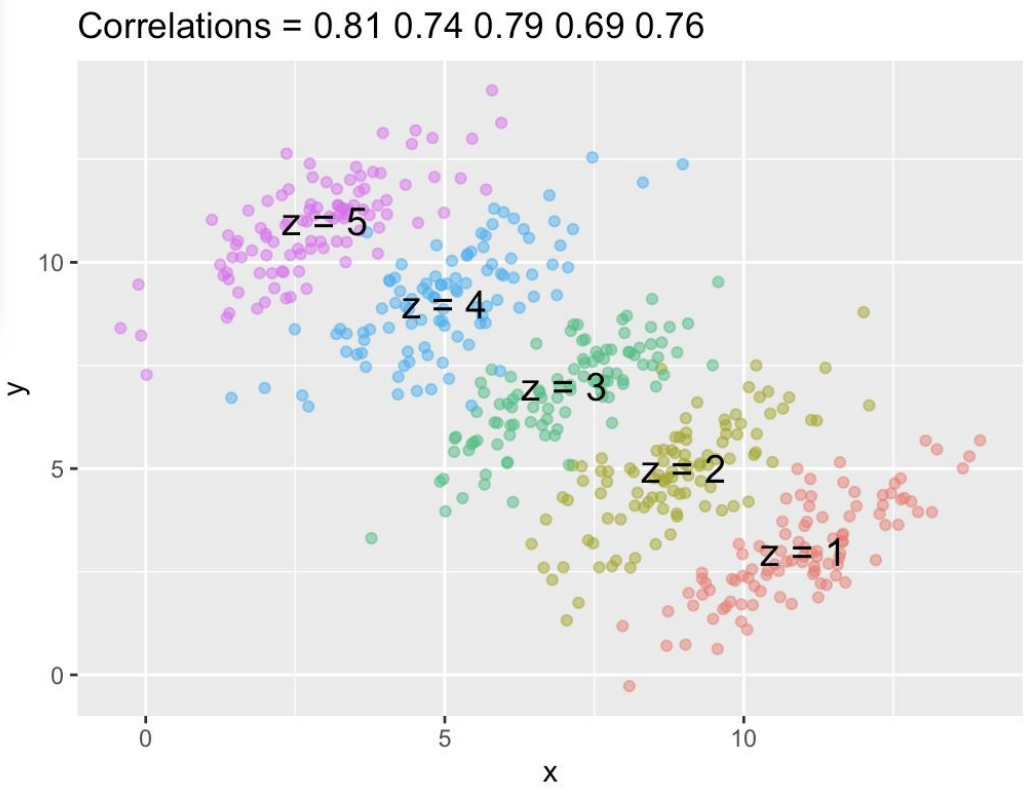
```
> N <- 100
> Sigma <- matrix(c(1,0.75,0.75, 1), 2, 2)*1.5
> means <- list(c(x = 11, y = 3),
+               c(x = 9, y = 5),
+               c(x = 7, y = 7),
+               c(x = 5, y = 9),
+               c(x = 3, y = 11))
> dat <- lapply(means, function(mu){
+     res <- MASS::mvrnorm(N, mu, Sigma)
+     colnames(res) <- c("x", "y")
+     res
+ })
> dat <- do.call(rbind, dat) |>
+     as_tibble() |>
+     mutate(z = as.character(rep(seq_along(means), each = N)))
> dat |> ggplot(aes(x, y)) + geom_point(alpha = 0.5) +
+     ggtitle(paste("Correlation = ", round(cor(dat$x, dat$y), 2)))
```



- The case we have just covered is an example of Simpson's paradox.
- Suppose you have three random variables *X*, *Y*, and *Z.*
- When plotting X and Y, you can see a negative correlation.

Simpson's paradox

```
> means <- do.call(rbind, means) |>
+     as_tibble() |>
+     mutate(z = as.character(seq_along(means)))
>
> corrs <- dat |> group_by(z) |> summarize(cor = cor(x, y)) |> pull(cor)
> dat |> ggplot(aes(x, y, color = z)) +
+     geom_point(show.legend = FALSE, alpha = 0.5) +
+     ggtitle(paste("Correlations =",  paste(signif(corrs,2), collapse="
")))) +
+     annotate("text", x = means$x, y = means$y, label = paste("z =", means
$z), cex = 5)
```



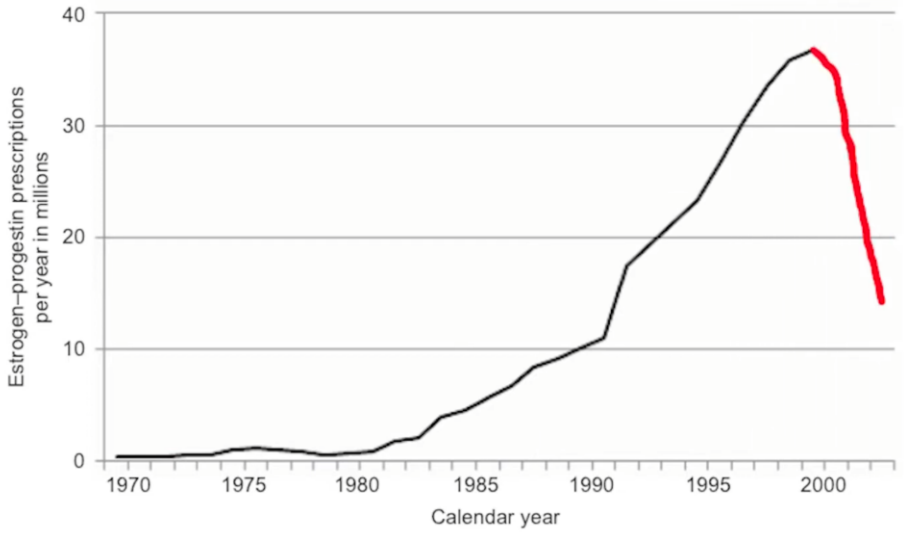
- The case we have just covered is an example of Simpson's paradox.
- Suppose you have three random variables *X*, *Y*, and *Z.*
- But when stratify by Z, X and Y are positively correlated.

How do we derive true causal relationship

- Randomized Controlled Trials (RCTs): the gold standard
  - We don't control for confounders; we just use random sampling to make sure that confounder effects are the same between groups
  - The treatment (X) is applied after random assignment of groups, so it is strictly independent of all possible confounding factors (Z).
- Twins, difference-in-differences
  - Minimize chance of potential confounders
- Causal inference methods: propensity score matching, graphical modeling, etc.

Observational study vs RCT

- In 1980s, large observational studies show that postmenopausal women on hormone replacement therapies had lower CVD risk.
- However, this was later found to be confounded by other factors; women on therapies are wealthier and more health-conscious.
- Randomized controlled trials later in 1990s and 2000s show that such therapies might increase CVD risks.

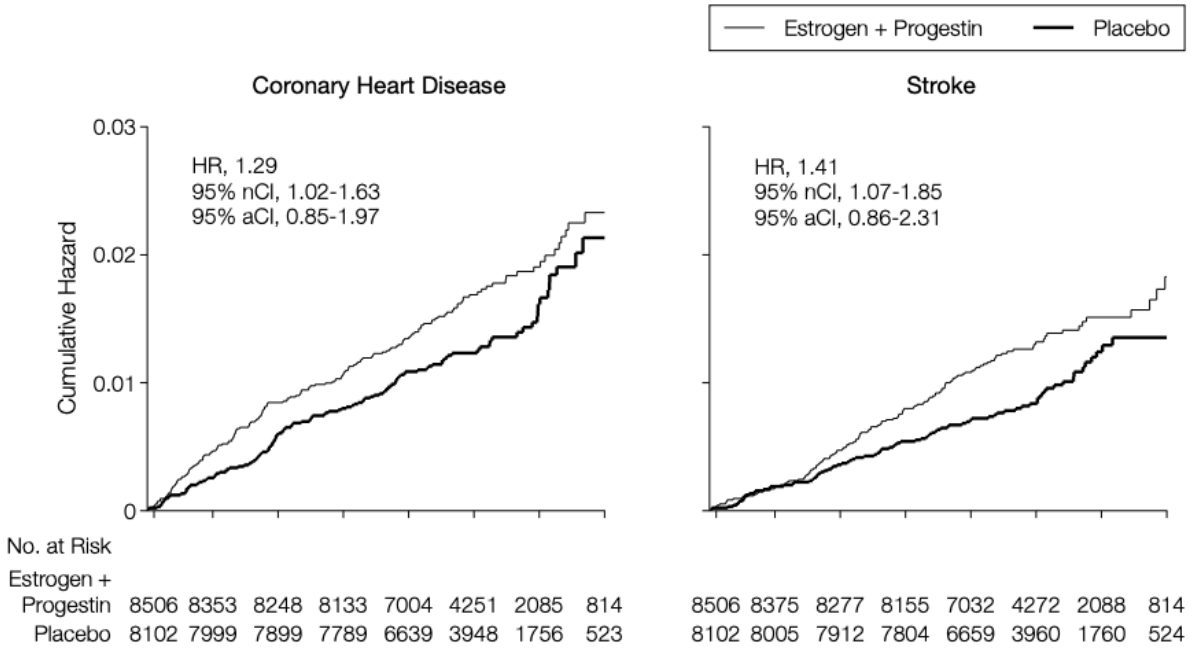


To learn more: https://www.youtube.com/watch?v=MVYWqWu2Za4

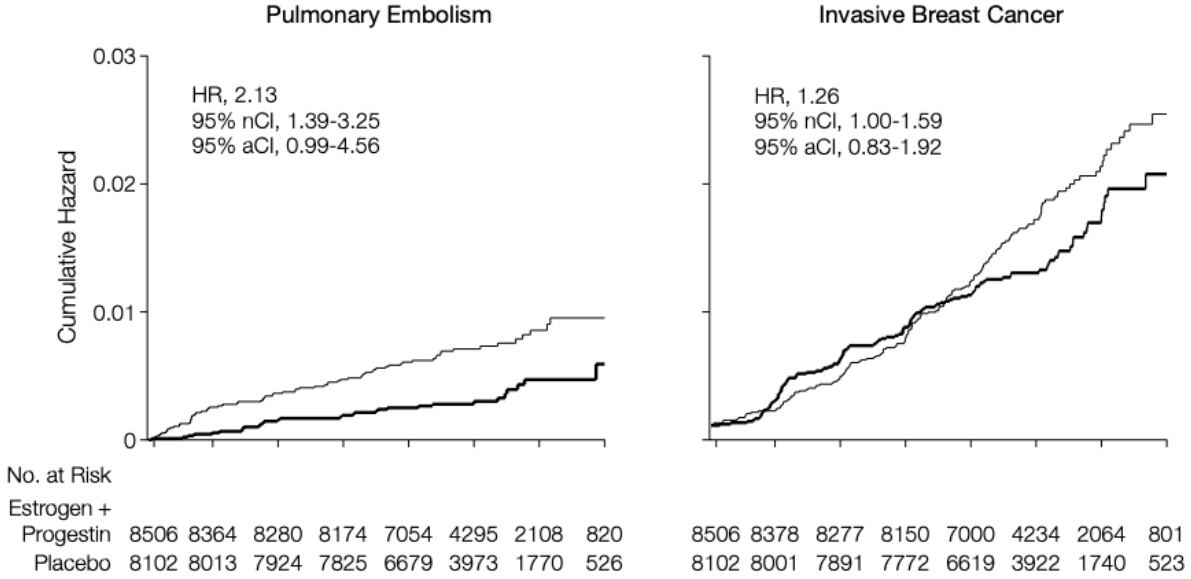
Table 1. Baseline Characteristics of the Women's Health Initiative Estrogen Plus Progestin Trial Participants (N = 16 608) by Randomization Assignment*

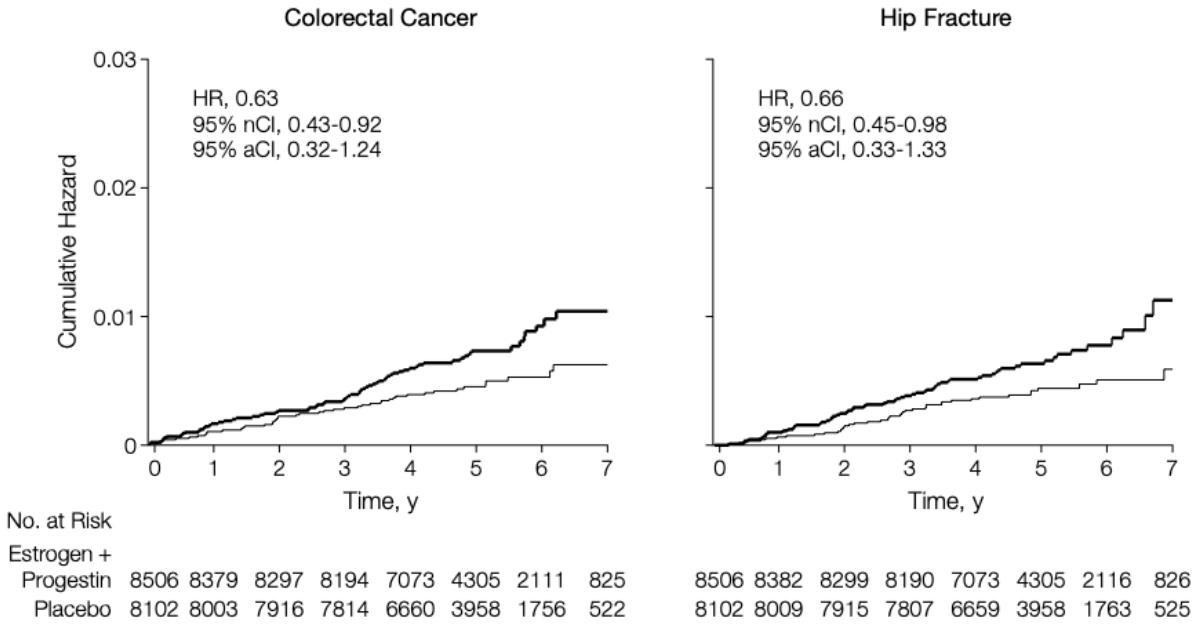
Characteristics	Estrogen + Progestin (n = 8506)	Placebo (n = 8102)	*P* Value†
Age at screening, mean (SD), y	63.2 (7.1)	63.3 (7.1)	.39
Age group at screening, y			
50-59	2839 (33.4)	2683 (33.1)	.80
60-69	3853 (45.3)	3657 (45.1)	
70-79	1814 (21.3)	1762 (21.7)	
Race/ethnicity			
White	7140 (83.9)	6805 (84.0)	.33
Black	549 (6.5)	575 (7.1)	
Hispanic	472 (5.5)	416 (5.1)	
American Indian	26 (0.3)	30 (0.4)	
Asian/Pacific Islander	194 (2.3)	169 (2.1)	
Unknown	125 (1.5)	107 (1.3)	
Hormone use			
Never	6280 (73.9)	6024 (74.4)	.49
Past	1674 (19.7)	1588 (19.6)	
Current‡	548 (6.4)	487 (6.0)	
Duration of prior hormone use, y			
<5	1538 (69.1)	1467 (70.6)	.25
5-10	426 (19.1)	357 (17.2)	
≥10	262 (11.8)	253 (12.2)	
Body mass index, mean (SD), kg/m^2§	28.5 (5.8)	28.5 (5.9)	.66
Body mass index, kg/m^2			
<25	2579 (30.4)	2479 (30.8)	.89
25-29	2992 (35.3)	2834 (35.2)	
≥30	2899 (34.2)	2737 (34.0)	
Systolic BP, mean (SD), mm Hg	127.6 (17.6)	127.8 (17.5)	.51
Diastolic BP, mean (SD), mm Hg	75.6 (9.1)	75.8 (9.1)	.31
Smoking			
Never	4178 (49.6)	3999 (50.0)	.85
Past	3362 (39.9)	3157 (39.5)	
Current	880 (10.5)	838 (10.5)	
Parity			
Never pregnant/no term pregnancy	856 (10.1)	832 (10.3)	.67
≥1 term pregnancy	7609 (89.9)	7233 (89.7)	
Age at first birth, y‖			
<20	1122 (16.4)	1114 (17.4)	.11
20-29	4985 (73.0)	4685 (73.0)	
≥30	723 (10.6)	621 (9.7)	
Treated for diabetes	374 (4.4)	360 (4.4)	.88
Treated for hypertension or BP ≥140/90 mm Hg	3039 (35.7)	2949 (36.4)	.37
Elevated cholesterol levels requiring medication	944 (12.5)	962 (12.9)	.50
Statin use at baseline¶	590 (6.9)	548 (6.8)	.66
Aspirin use (≥80 mg/d) at baseline	1623 (19.1)	1631 (20.1)	.09
History of myocardial infarction	139 (1.6)	157 (1.9)	.14
History of angina	238 (2.8)	234 (2.9)	.73
History of CABG/PTCA	95 (1.1)	120 (1.5)	.04
History of stroke	61 (0.7)	77 (1.0)	.10
History of DVT or PE	79 (0.9)	62 (0.8)	.25
Female relative had breast cancer	1286 (16.0)	1175 (15.3)	.28
Fracture at age ≥55 y	1031 (13.5)	1029 (13.6)	.87

Rossouw, Jacques E., et al. "Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial." *Jama* 288.3 (2002): 321-333.



- RCT shows higher risk of CHD, Stroke in patients taking estrogen + progestin





Rossouw, Jacques E., et al. "Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial." *Jama* 288.3 (2002): 321-333.

In this lecture

- Correlation
- Regression
- Association versus causation
    - Spurious correlation, outliers, confounders