Lecture 7: Correlation

Madeleine Udell
Stanford University

Upcoming deadlines

This Friday (4/22): Assignment 3

[Get started ASAP!]

Next Tuesday (4/26): Project proposal

[Required project meetings happening this week]

Next Friday (4/29): Assignment 4

[Shorter than Assignments 2 and 3]

Following Thursday (5/5): Quiz 1

[Everything up to and including linear regression]

Demo

https://colab.research.google.com/github/stanford-mse-125/demos/blob/main/correlation.ipynb

The standard deviation line

1. Goes through the point of averages.

2. Climbs [or falls] at the rate of one vertical SD for each horizontal SD.

Correlation

Measure of association between two variables

Quantifies the dispersion of the points around the SD line. Ranges from -1 to 1.

Definition: the correlation is the average of the products of the variables, when both are measured in standard units.

Correlation properties

Correlation is

- Scale invariant
- A measure of linear dependence
- Sensitive to outliers

Association is not causation

Examples from *Statistics* by Freedman et al.

For school children, shoe size is strongly correlated with reading skills. Does learning new words make your feet grow?

Association is not causation

Examples from *Statistics* by Freedman et al.

For school children, shoe size is strongly correlated with reading skills. Does learning new words make your feet grow?

Age is a confounding factor!

Association is not causation

Examples from *Statistics* by Freedman et al.

During the Great Depression of 1929-1933, better-educated people tended to have shorter spells of unemployment.

Does education protect you against unemployment?

Association is not causation

Examples from *Statistics* by Freedman et al.

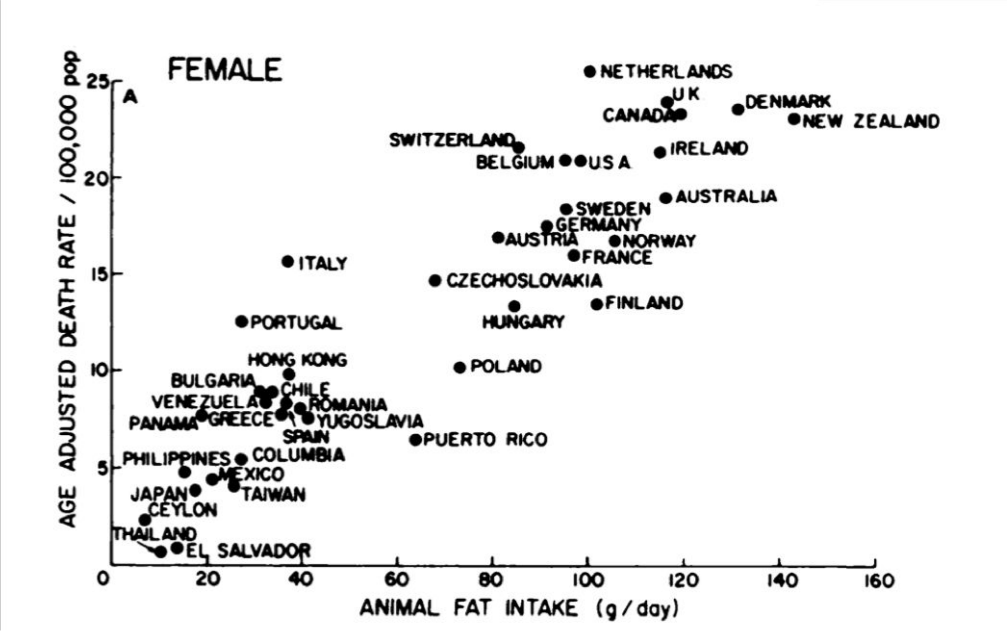
During the Great Depression of 1929-1933, better-educated people tended to have shorter spells of unemployment.

Does education protect you against unemployment?

Age is again a confounding factor. Employers tended to prefer younger job-seekers, and younger people were better educated.

Association is not causation

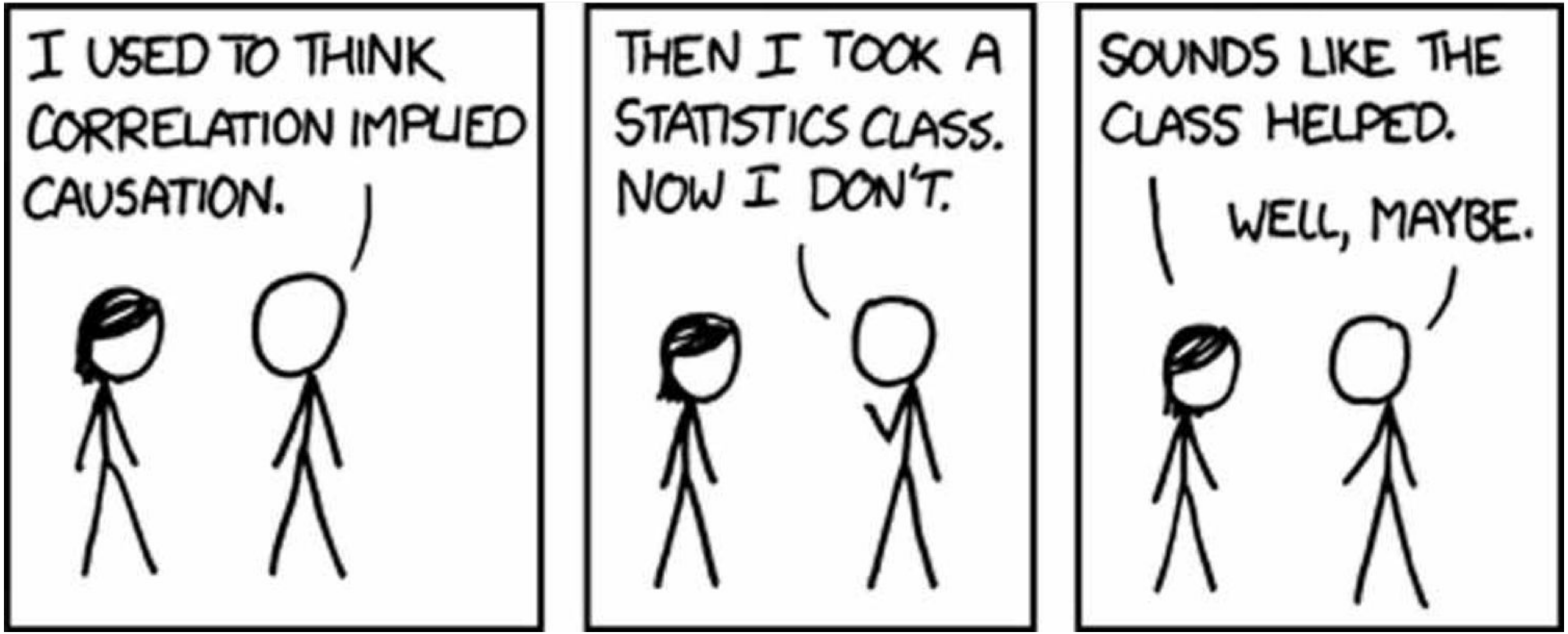
Fat in the diet and breast cancer



Association is not causation

Fat in the diet and breast cancer

Fat is relatively expensive so high fat intake occurs primarily in rich countries. Rich countries differ in a lot of ways from poorer ones.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Regression

Describes the relationship between two [or more] variables.

Regression

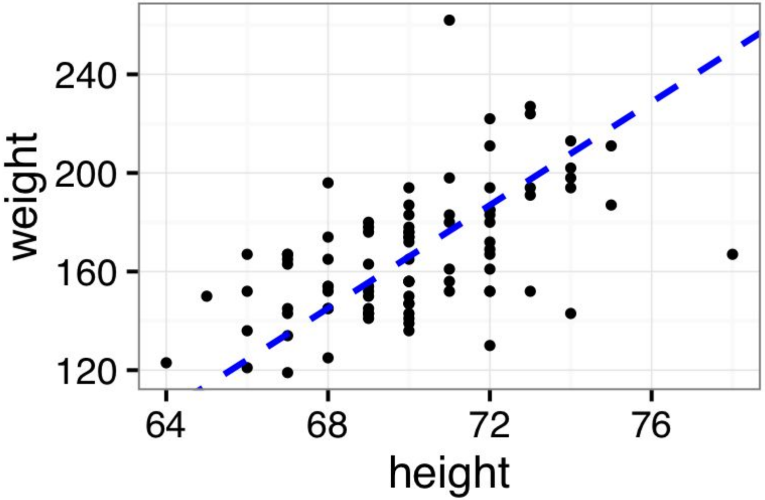
```r
head(measures)
```

```
## # A tibble: 6 × 2
##   weight height
##    <int>  <int>
## 1    169     72
## 2    150     70
## 3    167     67
## 4    167     66
## 5    152     73
## 6    156     70
```

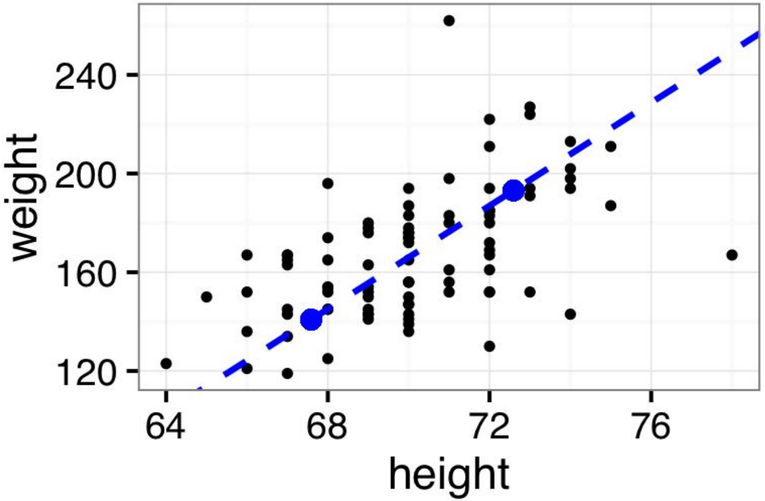
Regression

Estimates the average value of **y** for each value of **x**.

```
p <- qplot(data = measures, x = height, y = weight,
           geom = "point", size=I(0.75)) +
  sd_line(measures$height, measures$weight)
p
```



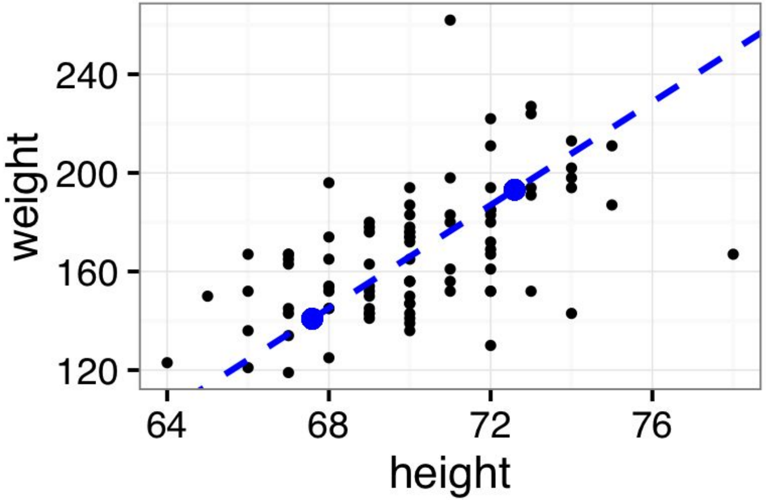
```
q <- p +
  geom_point(aes(x = mean(height) + sd(height),
                 y = mean(weight) + sd(weight)),
             color = "blue", size = 2) +
  geom_point(aes(x = mean(height) - sd(height),
                 y = mean(weight) - sd(weight)),
             color = "blue", size = 2)
q
```



Regression line

```r
tall <- measures %>%
  filter(height >= 72 & height <= 73)
avg_tall <- mean(tall$weight)

short <- measures %>%
  filter(height >= 67 & height <= 68)
avg_short <- mean(short$weight)
```

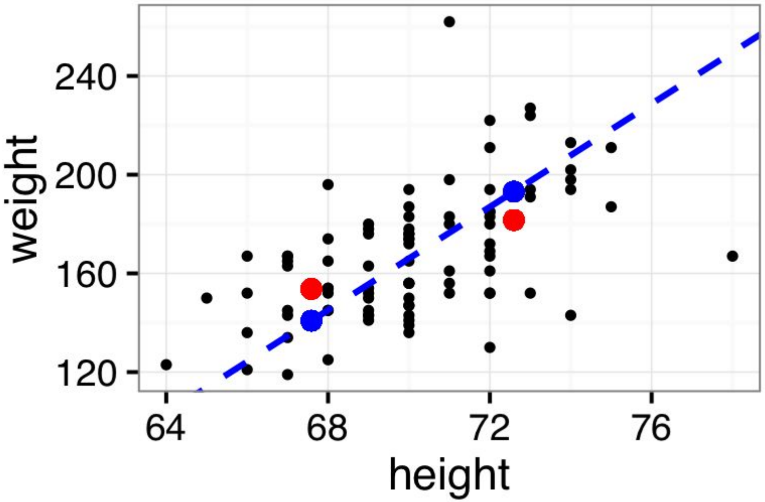


```r
tall <- measures %>%
  filter(height >= 72 & height <= 73)
avg_tall <- mean(tall$weight)

short <- measures %>%
  filter(height >= 67 & height <= 68)
avg_short <- mean(short$weight)
```

Draw points for `avg_tall` and `avg_short` on the plot.

```r
q <- q +
  geom_point(aes(x = mean(height) + sd(height)),
             y = avg_tall, color = "red", size = 2) +
  geom_point(aes(x = mean(height) - sd(height)),
             y = avg_short, color = "red", size = 2)
q
```



Regression to the mean

```
avg_tall
```

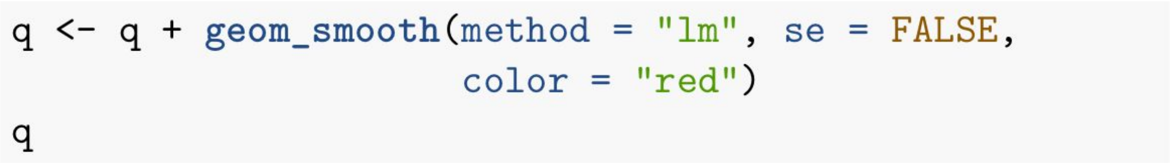
```
## [1] 181.6316
```

```
avg_tall - mean(measures$weight)
```

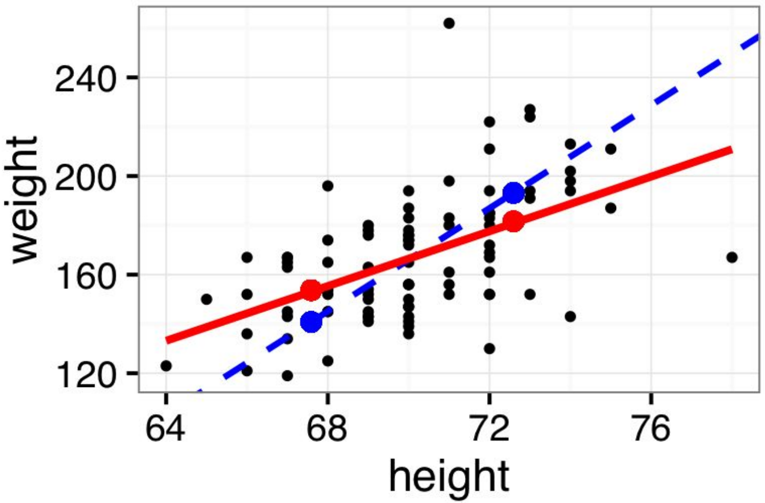
```
## [1] 14.64294
```

```
sd(measures$weight)
```

```
## [1] 26.16855
```



```
q <- q + geom_smooth(method = "lm", se = FALSE,
                     color = "red")
q
```



Both the SD line and the regression line cross the point of averages.

But, the regression line is shallower.

The slope of the regression line is the slope of the SD line, multiplied by the correlation.

Regression to the mean

A one standard deviation increase on the x-axis results in less than a one standard deviation increase on the y-axis.

```
(avg_tall - mean(measures$weight)) / sd(measures$weight)
```

```
## [1] 0.5595626
```

```
(avg_short - mean(measures$weight)) / sd(measures$weight)
```

```
## [1] -0.5075847
```

```
cor(measures$height, measures$weight)
```

```
## [1] 0.5309282
```

Why does it scale by r?

Some intuition

Why does it scale by r?

Some intuition

r = 0

No [linear] association between x and y

Why does it scale by r?

Some intuition

r = 0

No [linear] association between x and y

r = 1

Points lie on the SD line

r = -1

Points lie on the SD line

Regression

Estimates the average value of **y** for each value of **x**.

Regression line

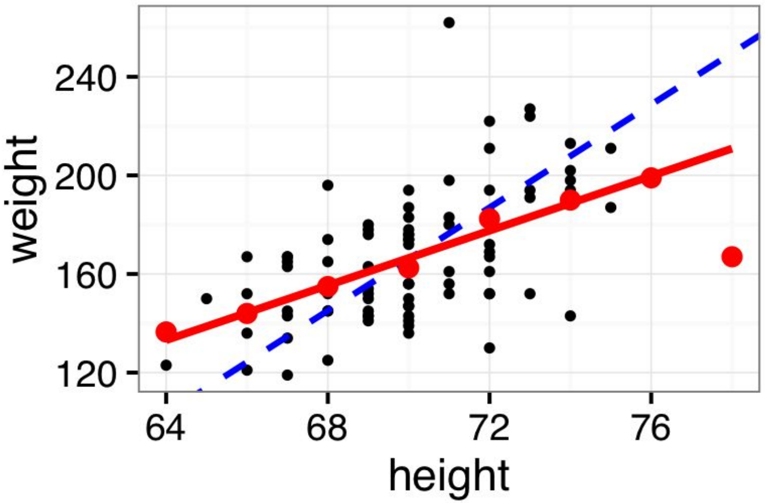
```
measures <- measures %>%
  mutate(rounded_height = round_any(height, 2))
head(measures)
```

```
## # A tibble: 6 × 3
##   weight height rounded_height
##    <int>  <int>          <dbl>
## 1    169     72             72
## 2    150     70             70
## 3    167     67             68
## 4    167     66             66
## 5    152     73             72
## 6    156     70             70
```

```r
avgs <- measures %>%
  group_by(rounded_height) %>%
  summarize(avg_weight = mean(weight))
head(avgs)
```

```
## # A tibble: 6 × 2
##   rounded_height avg_weight
##            <dbl>      <dbl>
## 1             64   136.5000
## 2             66   144.0000
## 3             68   154.9286
## 4             70   162.5000
## 5             72   182.4231
## 6             74   190.0000
```

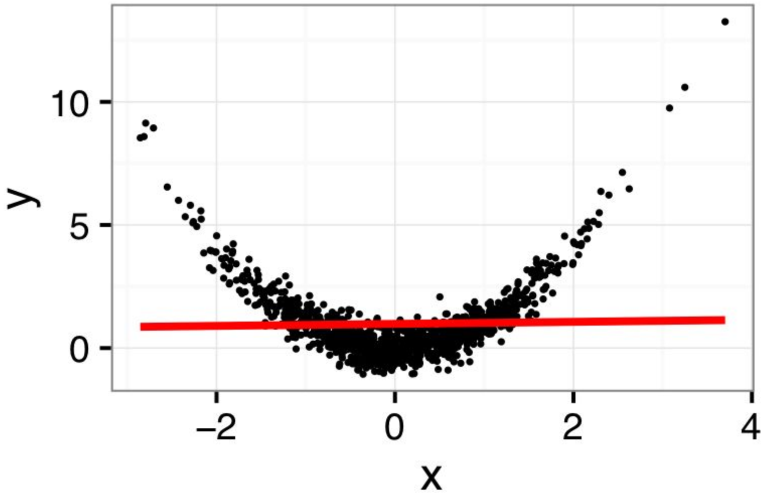
```r
q <- p +
  geom_smooth(method = "lm", se = FALSE,
                              color = "red") +
  geom_point(
    data = avgs,
    aes(rounded_height, avg_weight),
    color = "red", size = 2)
q
```



Regression line

Captures *linear* trends

```r
p +
  geom_smooth(method = "lm", se = FALSE,
              color = "red")
```



```r
cor(x, y)
```

```
## [1] 0.02802503
```

Regression as prediction

`weight` ≈ -220 + 5.5 * `height`

Regression as prediction

`weight` ≈ -220 + 5.5 * `height`

`height` = 70 inches

`weight` ≈ [-220 + 5.5 * 70] = 165 pounds

The regression fallacy

Ascribing spurious causal explanations for regression-to-the-mean effects.

The regression fallacy

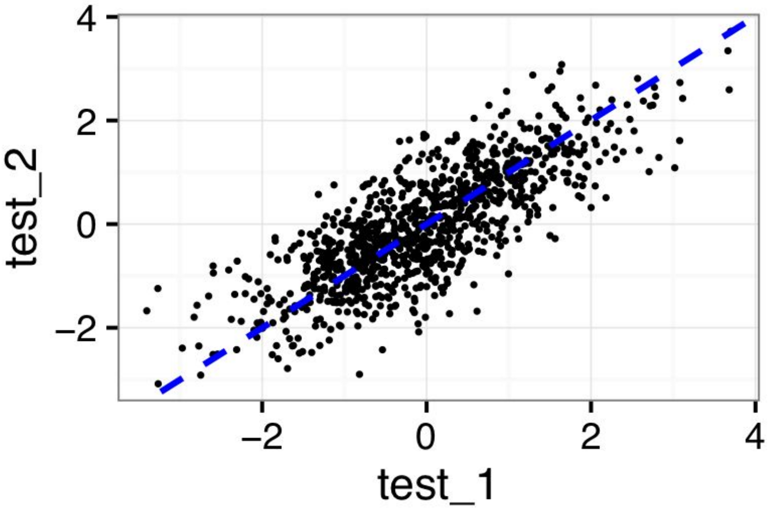
The *Sports Illustrated* cover Jinx

Individuals who appear on the cover of *Sports Illustrated* will subsequently perform badly.

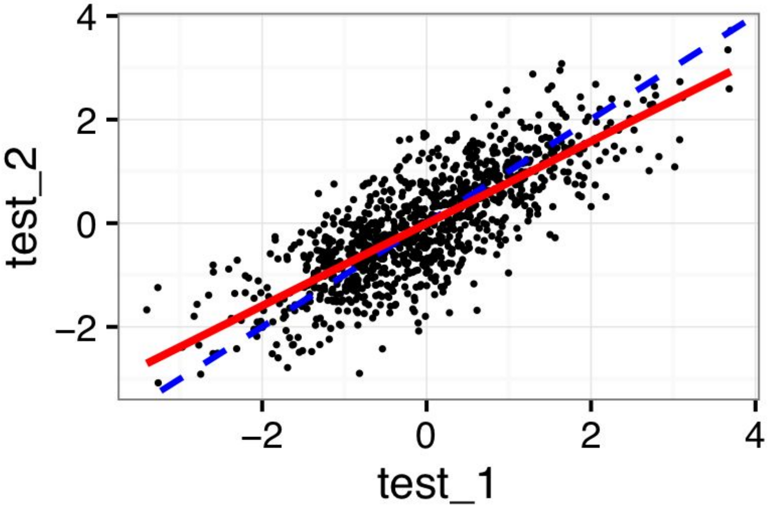
[Discuss with neighbors]

```r
ability <- rnorm(1000, 0, 1)
test_1 <- ability + rnorm(1000, 0, .5)
test_2 <- ability + rnorm(1000, 0, .5)

p <- qplot(test_1, test_2, size=I(0.25)) +
  sd_line(test_1, test_2)
p
```



```
p <- p +
  geom_smooth(method = "lm", se = FALSE,
              color = "red")
p
```


test_2
test_1