# Correlation

Lectures 30 & 31

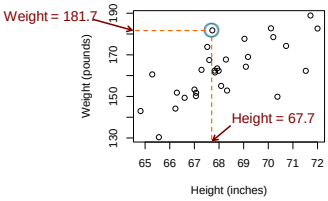
Reading: GW 16.1-16.3; SS 12.0-12.2

## Relationships Between Continuous Variables

- Some studies measure multiple variables
  - Any paired-sample experiment
  - Training & testing performance; personality variables; neurological measures
  - Continuous independent variables
- How are these variables related?
  - Positive relationship: tend to be both large or both small
  - Negative relationship: when one is large, other tends to be small
  - Independent: value of one tells nothing about other

## Scatterplots

- Graph of relationship between two variables, *X* and *Y*
- One point per subject
  - Horizontal coordinate *X*
  - Vertical coordinate *Y*




## Correlation

- Measure of how closely two variables are related
  - Population correlation: ρ (rho)
  - Sample correlation: *r*
  - Also called a "Pearson" correlation

## History of the Correlation

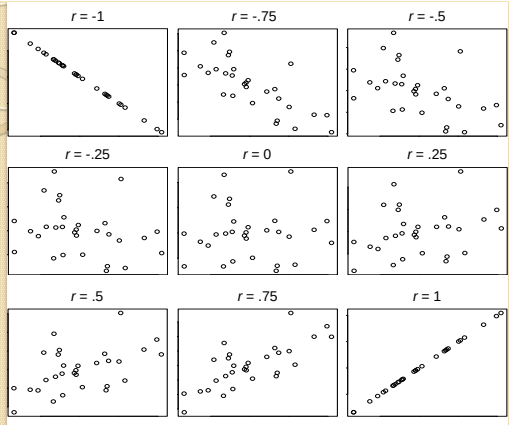
- Developed by Karl Pearson

And Francis Galton

- Both were pioneers of mathematical statistics…
- but also "eugenicists" who believed non-European races were inferior: *"My view – and I think it may be called the scientific view of a nation, is that of an organized whole, kept up to… efficiency by insuring that its numbers are substantially recruited from the better stocks, and… by way of war with inferior races."*

## Correlation

- Measure of how closely two variables are related
  - Population correlation: ρ (rho)
  - Sample correlation: *r*
  - Also called a "Pearson" correlation
- Direction
  - $r > 0$: positive relationship; big *X* goes with big *Y*
  - $r < 0$: negative relationship; big *X* goes with small *Y*
- Strength
  - ±1 means perfect relationship
    - Data lie exactly on a line
    - If you know *X*, you know *Y*
  - 0 means no relationship
    - Independent: Knowing *X* tells nothing about *Y*




## Computing Correlation

1. Get z-scores for both samples
2. Multiply all pairs
3. Get average by dividing by $n - 1$

$$r = \frac{\sum (z_X \times z_Y)}{n - 1}$$

- Positive relationship
  - Positive $z_X$ tend to go with positive $z_Y$
  - Negative $z_X$ tend to go with negative $z_Y$
  - $z_X \cdot z_Y$ tends to be positive
- Negative relationship
  - Positive $z_X$ tend to go with negative $z_Y$
  - Negative $z_X$ tend to go with positive $z_Y$
  - $z_X \cdot z_Y$ tends to be negative

## Computing Correlation

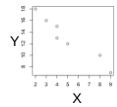
$$r = \frac{\sum (z_X \times z_Y)}{n - 1}$$

| X | Y |
|---|---|
| 5 | 8 |
| 8 | 10 |
| 3 | 4 |
| 9 | 13 |
| 2 | 2 |
| 4 | 7 |
| 4 | 5 |


## Computing Correlation

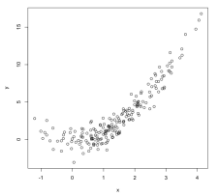
$$r = \frac{\sum (z_X \times z_Y)}{n - 1}$$

| X | Y |
|---|---|
| 5 | 12 |
| 8 | 10 |
| 3 | 16 |
| 9 | 7 |
| 2 | 18 |
| 4 | 13 |
| 4 | 15 |




## Correlation and Linear Relationships

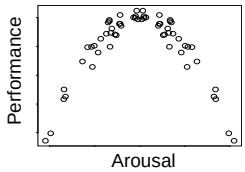
- Correlation measures how well data fit on straight line
  - Assumes linear relationship between $X$ and $Y$
- Will under-estimate the strength of non-linear relationships



$r = .75$

## Correlation and Linear Relationships

- Correlation measures how well data fit on straight line
  - Assumes linear relationship between $X$ and $Y$
- Will under-estimate non-linear relationships
- Not useful for non-monotonic relationships



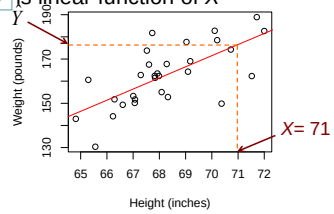

$r = 0$

## Predicting One Variable from Another

- Knowing one measure gives information about others from same subject
  - Knowing a person's weight tells about their height
- Goal: Come up with a rule or function that uses $X$ to compute best estimate of $Y$
- $Y$ (Y-hat) $\hat{Y}$
  - Predicted value of $Y$
  - Function of $X$
  - Best prediction of $Y$ based on $X$

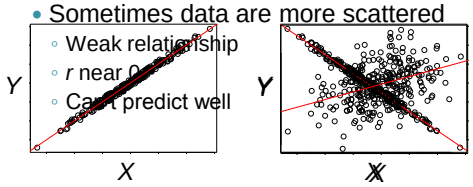
## Linear Prediction

- Simplest way to predict one variable from another
- Straight line through data
- $\hat{Y}$ is linear function of $X$

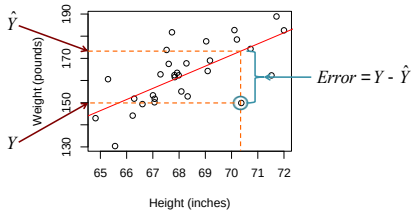



## How Good is the Prediction?

- Sometimes data fall nearly on a perfect line
  - Strong relationship between variables
  - $r$ near ±1
  - Good prediction
- Sometimes data are more scattered
  - Weak relationship
  - $r$ near 0
  - Can't predict well




## How Good is the Prediction?

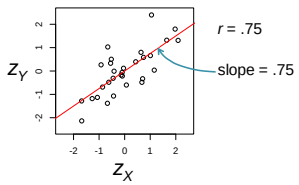


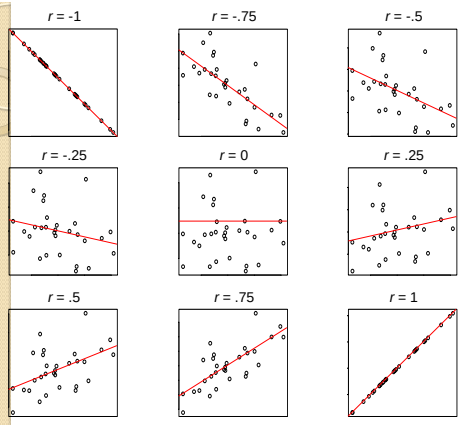

- Goal: Keep error close to zero
  - Minimize mean squared error:

$$MS_{\text{Error}} = \frac{\sum \left(Y - \hat{Y}\right)^2}{n - 1}$$

## Correlation and Prediction

- Best prediction line minimizes MSE
  - Closest to data; best "fit"
- Correlation determines best prediction line
  - Slope = $r$ when plotting z-scores: $z_{\hat{Y}} = r \times z_X$



## Explained Variance

Without knowing *X*

$\hat{Y} = M_Y$

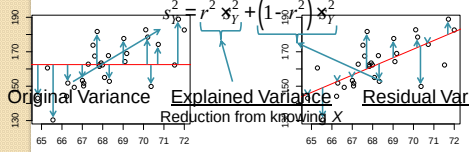
$MS_{\text{Error}} = s_Y^2$

Knowing *X*

$z_{\hat{Y}} = r \times z_X \longrightarrow \hat{Y} = M_Y + r \times z_x \times s_y$

$MS_{\text{Error}} = \left(1 - r^2\right) \times s_Y^2$

$s_Y^2 = r^2 \times s_Y^2 + \left(1 - r^2\right) \times s_Y^2$



Original Variance | Explained Variance | Residual Variance

Reduction from knowing *X*

## Spearman Correlation

- Useful in two situations

1) When data are ordinal (ranks) rather than interval or ratio. In this case, merely obtain the regular correlation on the data
2) When the relationship between two variables is monotonic (always increasing or decreasing) but not necessarily linear. In this case, first convert the data to ranks, then obtain the regular (pearson) correlation on the data

## Inference for a correlation

- Null hypothesis: $\rho = 0$
- Alternative hypothesis: $\rho \neq 0$
- The distribution of *r,* divided by its SE, follows a t-distribution:

$$t(n-2) = \frac{r}{\sqrt{1-r^2} \Big/ \sqrt{n-2}}$$

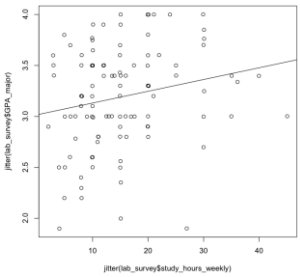
- We can also get a 95% confidence interval of *r,* which tells us the likely value of $\rho$ in the population

## Correlation in R

```
> cor(lab_survey$study_hours_weekly,lab_survey$GPA_major)
[1] NA
> cor(lab_survey$study_hours_weekly,lab_survey$GPA_major,use='pairwise')
[1] 0.2015451

> cor.test(lab_survey$study_hours_weekly,lab_survey$GPA_major)

        Pearson's product-moment correlation

data:  lab_survey$study_hours_weekly and lab_survey$GPA_major
t = 2.1085, df = 105, p-value = 0.03737
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.01215169 0.37697843
sample estimates:
      cor
0.2015451
```

## Correlation in R

```
> plot(lab_survey$study_hours_weekly,lab_survey$GPA_major)
> abline(lm(lab_survey$GPA_major~lab_survey$study_hours_weekly))
```




## Properties of Correlation

- Measures relationship between two continuous variables
  - How well data are fit by a straight line

$$r = \frac{\sum (z_X \times z_Y)}{n - 1}$$

- Sign of *r* shows direction of relationship
- Magnitude of *r* shows strength of relationship
  - Strongest relationships have $r = \pm 1$; weak relationships have $r \approx 0$
- Best prediction line minimizes error of prediction ($MS_{\text{Error}}$)
  - Correlation gives slope of line (when using z-scores): $z_{\hat{Y}} = r \times z_X$
- $r^2$ equals proportion of variance in one variable explained by other
  - Reduction from original variance ($s_Y^2$) to residual variance ($MS_{\text{Error}}$)