STRINGFEST ANALYTICS

INTERMDIATE EXCEL STATISTICS FOR BUSINESS ANALYTICS

George Mount

Data Analyst & Educator at Stringfest Analytics

George works as an independent analyst and data analytics educator with the goal to help clients manage their data so they think more creatively. He serves as a technical expert and lead curriculum developer for Thinkful's data analytics program and is the instructor of the DataCamp course "Survey and Measure Development in R."

George blogs about data, innovation, and career development at georgejmount.com. He holds a master's degree in information systems with a certificate of achievement in quantitative methods from Case Western Reserve University

COURSE OBJECTIVES

- Test for differences across multiple groups and at multiple points in time
- Model a causal relationship between two variables
- Make graphical representations of one or more variables
- Make compelling business recommendations using inferential statistics

Intermediate Excel
Statistics for Business
Analytics

WHY WOULD WE DO THIS IN EXCEL?

Intermediate Excel
Statistics for Business
Analytics

“You get to look at the data every step of the way, building confidence while learning the tricks of the trade.”

-- John Foreman

FOLLOWING ALONG

- Each section is a sub-folder
- Demos = follow along with me
- Drills = try it yourself
  - Refresh your memory with the demo notes

Intermediate Excel
Statistics for Business
Analytics

HAVE YOU INSTALLED THE DATA ANALYSIS TOOLPAK?

ON WINDOWS:

- File
- Options
- Add-ins
- Go
- Check on Analysis ToolPak
- OK

ON MAC:

- Tools
- Excel Add-ins
- Check on Analysis ToolPak
- Click OK

1. EXPECTED VALUES AND REPEATED MEASURES

(Hypothetical) warm-up

- File: `housing.xlsx`
- How would you check for a significant difference in prices of homes with and without air conditioning?
  - *At the 95% confidence level* (a constant for the course)

Warm-up

- File: `housing.xlsx`
  - *What about a relationship in homes with air conditioning versus homes with a full basement?*

CHI SQUARE TEST OF INDEPENDENCE

(χ^2)

ASSUMPTIONS

1. Two variables are categorical
2. Each subject contributes data to one and only one cell

	no	yes	Total
no	248	107	355
yes	125	66	191
Total	**373**	**173**	**546**

HYPOTHESES

Ho: No relationship exists between variables exists
Ha: A relationship between the variables exists

DEMO

- File: `housing.xlsx`
  - Is there a relationship in homes with air conditioning versus homes with a full basement?
    - *Don’t forget about the demo notes!*

DRILL

- File: `computers.xlsx`
  - Is there a relationship between having a CD-ROM and being a “premium” computer brand?
    - *Don’t forget the demo notes!*

QUESTIONS?

The acorn becomes the oak

- How do we measure differences in time across *same* individuals?
  - *Repeated measures*

House at time 1

Intervention
(install AC)

House at time 2

PAIRED SAMPLE T-TEST

ASSUMPTIONS

1. The data is paired
2. Independence of observations
3. The dependent variable is continuous
4. The data is continuous at times 1 and 2

HYPOTHESES

Ho: No difference on average between time 1 and time 2
Ha: A difference on average between time 1 and time 2

DEMO

- Demo: `bp.xlsx`
  - Is there a difference after the intervention?

DRILL

- Demo: `tomography.xlsx`
  - For which groups is there a significant difference from volume 1 to volume 2?

DRILL

- Congratulations on replicating a research study!

	Volume 1 (mL)	Volume 2 (mL)	*p*-value	Scan interval (days)
Group 1	4525.8 ± 1056.4	4539.9 ± 1009.6	0.751	361 (293, 365)
Group 2	4657.6 ± 1138.4	4639.6 ± 1102.8	0.744	279 (30, 365)
Group 3	3234.7 ± 947.1	3198.0 ± 978.6	0.371	182 (24, 365)

Data are presented as the mean±the standard deviation, unless otherwise stated.

The median interval between the two CT scans is presented with the minimum and maximum values.

https://doi.org/10.1371/journal.pone.0182849.t002

QUESTIONS?

PARAMETRIC AND NON-PARAMETRIC TESTS

WILCOXON SIGNED-RANK TEST

ASSUMPTIONS

1. The data is paired
2. Independence of observations
3. The dependent variable is continuous

HYPOTHESES

Ho: The median difference between time 1 and time 2 is zero

Ha: The median difference between time 1 and time 2 is not zero

DEMO

- Demo: `cortisol.xlsx`
  - Is there a difference in morning versus evening doses?
    - Multiply the *rank* of each observation by its *sign*
    - Compare the sum of all positive versus negative ranks
    - If test statistic is *less than* critical value, we reject the null

WHAT JUST HAPPENED?

Parametric	Non-parametric
Assumptions are made about the population *parameters*	No assumptions made about the populations
More rigid, more powerful, less flexible	Less rigid, less powerful, more flexible
Test statistic is based on probability distribution	Test statistic is arbitrary

QUESTIONS?

2. WORKING WITH MULTIPLE GROUPS

EDA, PART DEUX

There’s ALWAYS room for descriptives!

- Central tendency
  - *Expected value* = mean
- Variability
  - Variance, standard deviation, range
- Distribution
  - Skewness, kurtosis

Every picture tells a story

- Visualizing distributions with histograms and box plots
- Demo: `iris-viz.xlsx`

DEMO

- File: `outliers.xlsx`
  - What makes an outlier, an outlier?

DRILL

- File: `abalone-viz.xlsx`
  - Visualize the distribution of `shucked_wgt` by `sex`

COMPARING THE MEANS OF MORE THAN TWO GROUPS

ANALYSIS OF VARIANCE

ASSUMPTIONS

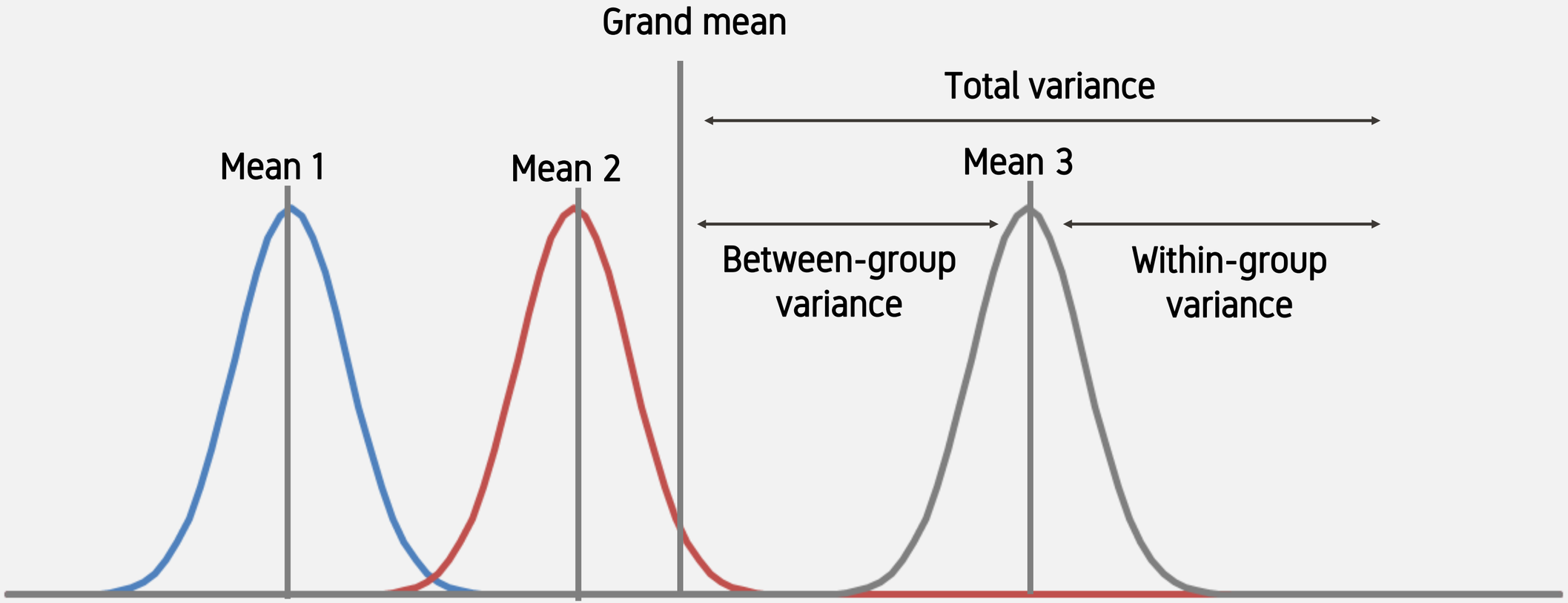
1. Subjects are randomly sampled
2. Observations are independent
3. Normality of each group
4. Population variance is equal for all groups

HYPOTHESES

Ho: No difference in population means of all groups
Ha: A difference in population means of all groups

WHY ANOVA? WHY NOT ANOME?

BETWEEN-GROUP vs WITHIN-GROUP



DEMO

- Demo: `abalone-anova.xlsx`
  - Is there a difference in shucked weight across all groups?

DEMO

- Demo: `abalone-anova.xlsx`
  - Is there a difference in shucked weight across all groups?
  - What pairs are actually different?
    - Pairwise t-tests
    - “Post-hoc”
  - *Watch out for that p!*
    - Experimentwise error

SOLD: $200,000

SOLD: $200,000

$\alpha = .05$

If the null were true (i.e. no real difference in means), we would find a significant difference in 5% of our samples *due to random error.*

IT HAPPENS

Yes, it's true that a team at Google couldn't decide between two blues, so they're testing 41 shades between each blue to see which one performs better. I had a recent debate over whether a border should be 3, 4 or 5 pixels wide, and was asked to prove my case. I can't operate in an environment like that. I've grown tired of debating such minuscule design decisions. There are more exciting design problems in this world to tackle.

https://stopdesign.com/archive/2009/03/20/goodbye-google.html

BONFERRONI CORRECTION

$$\text{Corrected p-value} = \frac{\propto}{n}$$

1 – confidence level (usually 5%)

Number of groups compared

DEMO

- Demo: `abalone-posthoc.xlsx`
  - What groups are different? (Pairwise t-tests)
  - How do we adjust for experimentwise error? (Bonferroni correction)

DRILL

- `iris-anova.xlsx`
  - Is there a significant difference in petal lengths across groups?
    - Which groups?

QUESTIONS?

PEARSON CORRELATION

ASSUMPTIONS

1. Two variables are normally distributed
2. Relationship between two variables is linear
3. No influential cases

HYPOTHESES

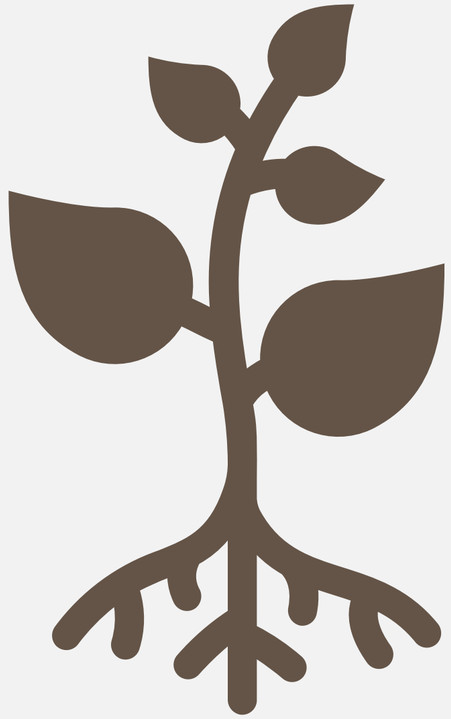
Use this rule of thumb for now:

Correlation coefficient	Interpretation
-1.0	Perfect negative (linear) relationship
-.7	Strong negative relationship
-.5	Moderate negative relationship
-.3	Weak negative relationship
0	No (linear) relationship
+.3	Weak positive relationship
.5	Moderate positive relationship
.7	Strong positive relationship
+1.0	Perfect positive (linear) relationship

Correlations

- Demo: `iris-corr.xlsx`
  - Printing a correlation matrix
  - Visualizing a bivariate relationship: scatter plots
    - X-axis: *independent variable*
    - Y-axis: *dependent variable*

WHICH CAME FIRST: THE INDEPENDENT OR DEPENDENT VARIABLE?



Independent variable:
Not affected by experiment

Dependent variable:
Affected by change in independent variable

Every picture tells a story

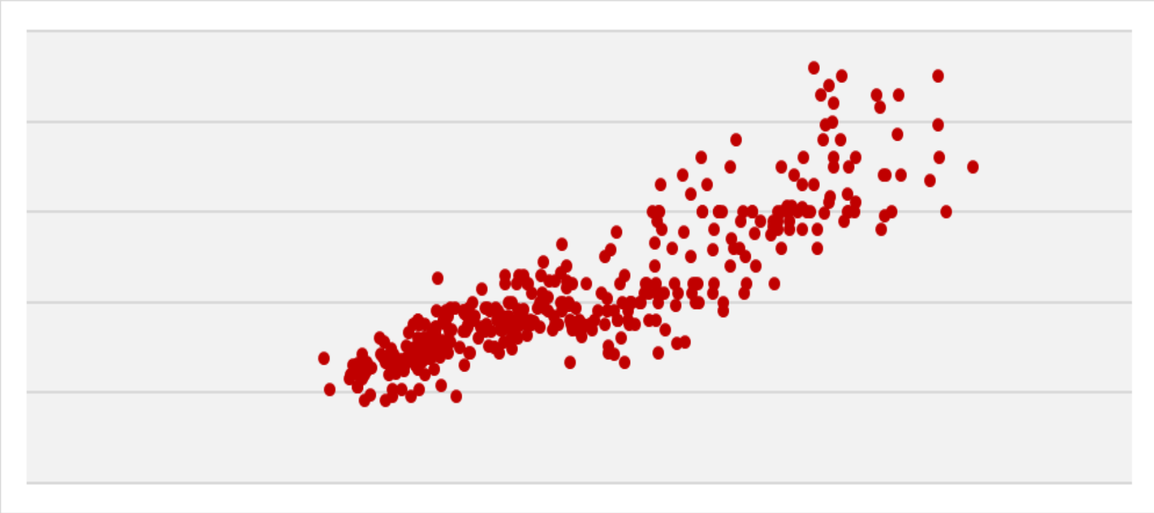
- Be careful about linearity!
- Demo: `anscombe.xlsx`

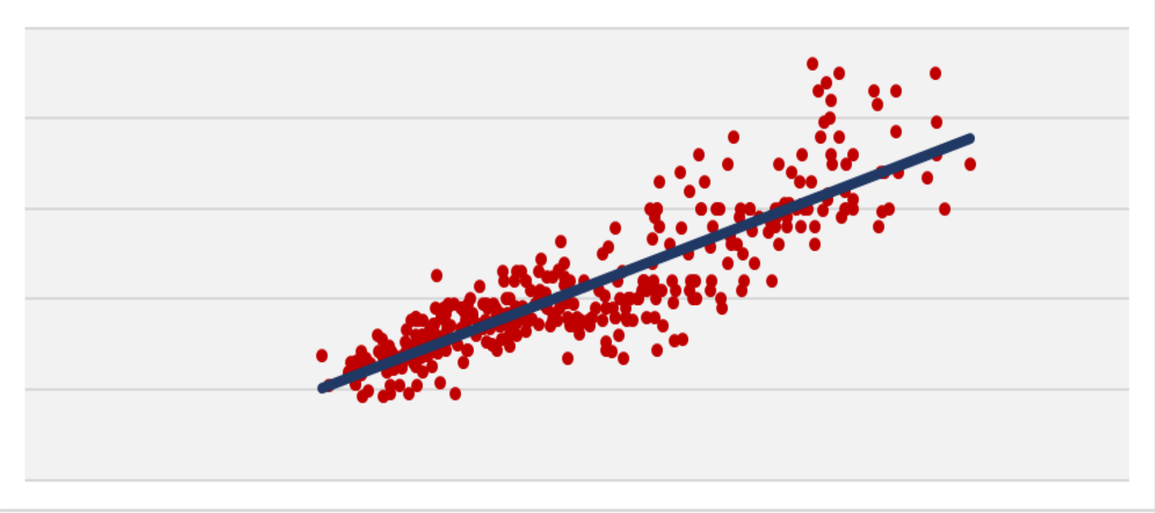
DRILL

- `mpg.xlsx`
  - Produce a correlation matrix
    - What is the strength of the relationship between weight and acceleration?
    - Plot the relationship.

3. UP AND RUNNING WITH LINEAR REGRESSION

Correlation	Regression
Indicates the extent to which two variables move together linearly	Indicates the estimated impact of a unit change of the independent variable *X* on the dependent variable *Y*.





ASSUMPTIONS

1. Linear relationship between independent and dependent variable
2. No influential cases
3. Values of residuals are independent
4. Variance of residuals is constant
5. Values of residuals are normally distributed

EXPLICIT WARNING: MATH AHEAD

LINEAR REGRESSION EQUATION

Dependent / predictor variable

Y intercept

Slope coefficient

Independent / response variable

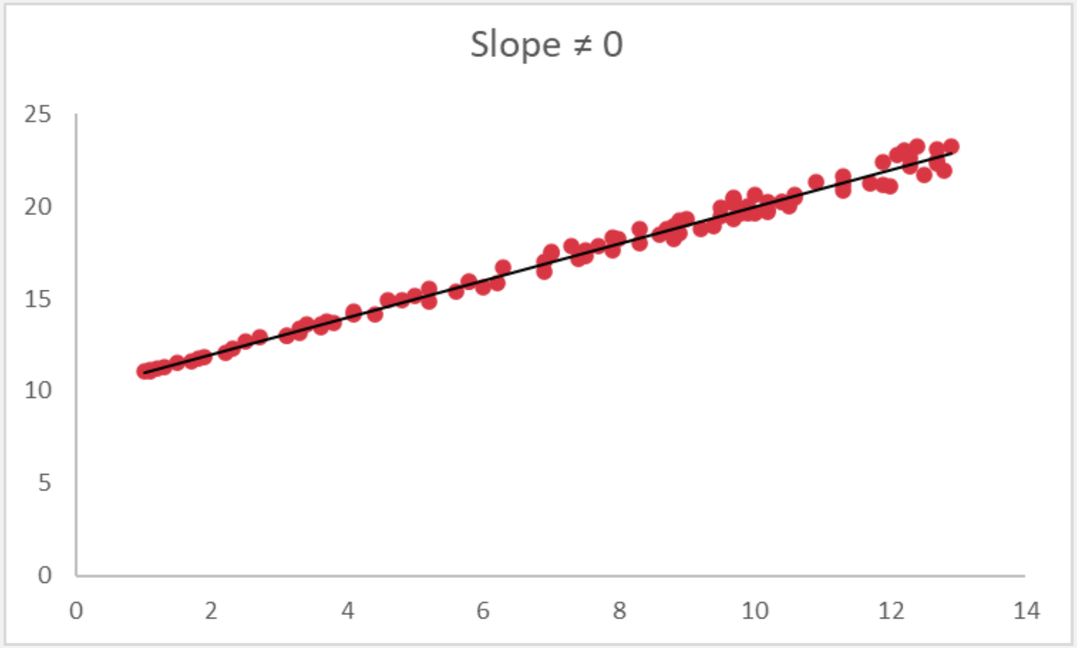
Error term

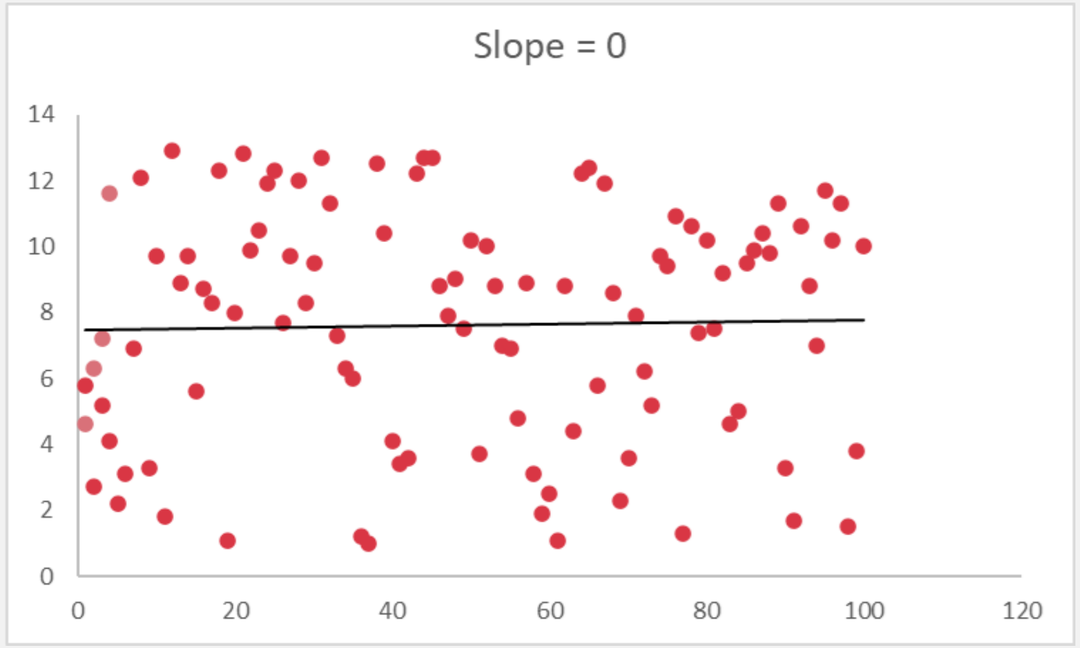
$$Y_i = \beta_0 + \beta_1 * X_i + \varepsilon_i$$

HYPOTHESES

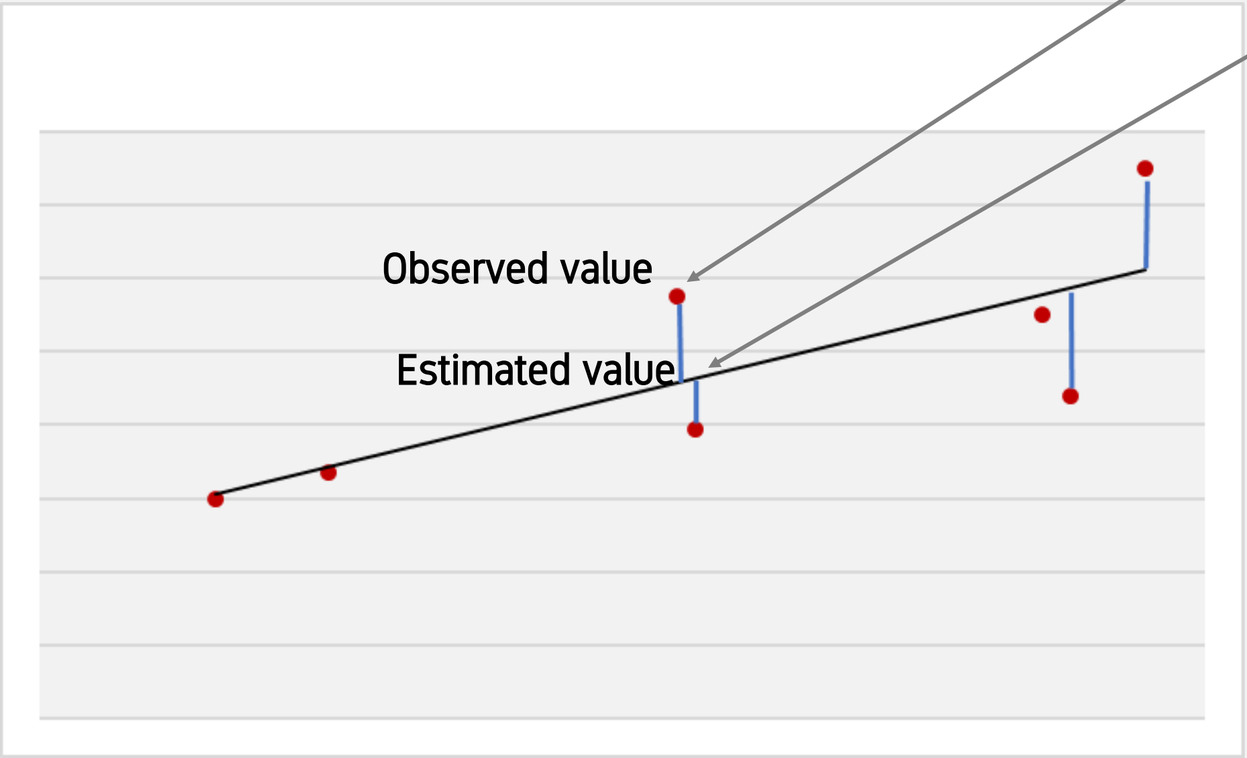
Ho: No relationship between X and Y. The slope equals zero.

Ha: A relationship between X and Y. The slope does not equal zero.





$$Residual = Y - \hat{Y}$$



DEMO

- `mpg-regression.xlsx`
  - Is there a significant relationship between weight (X) and acceleration (Y)?

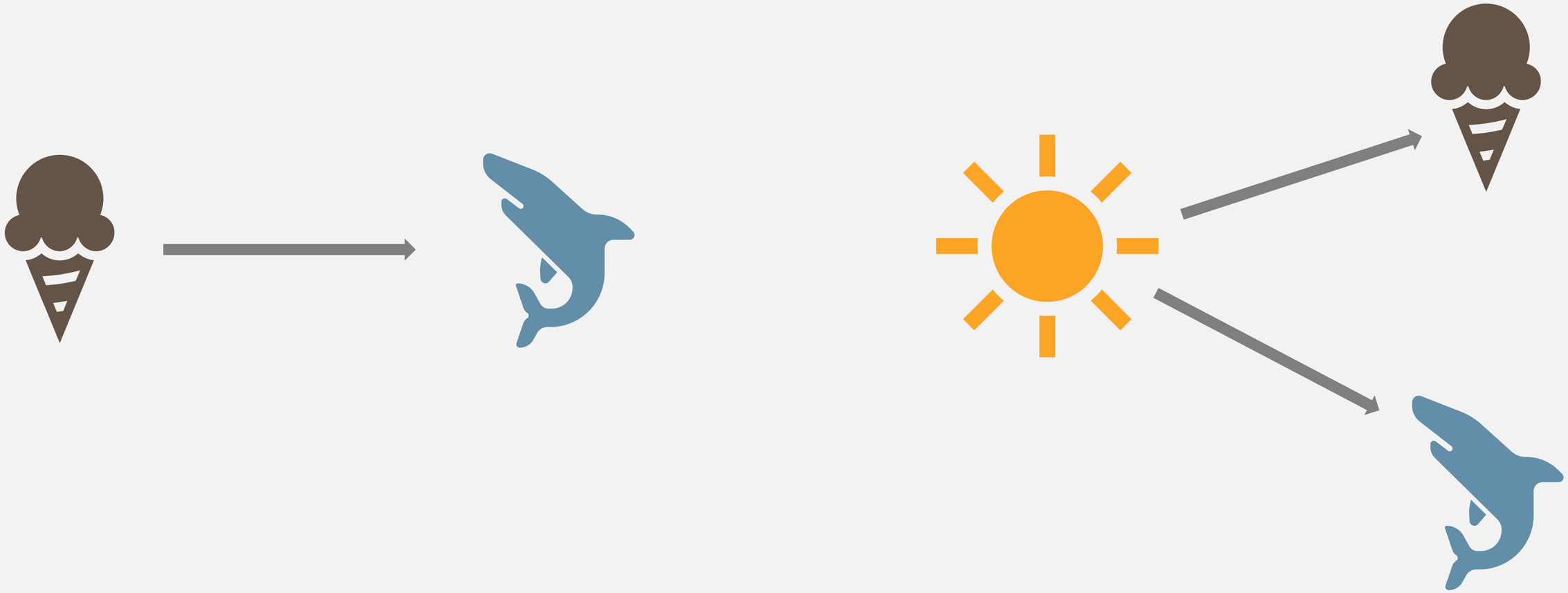
DRILL

- `mpg-regression-drill.xlsx`
  - Is there a significant linear trend between weight (X) and displacement (Y)?

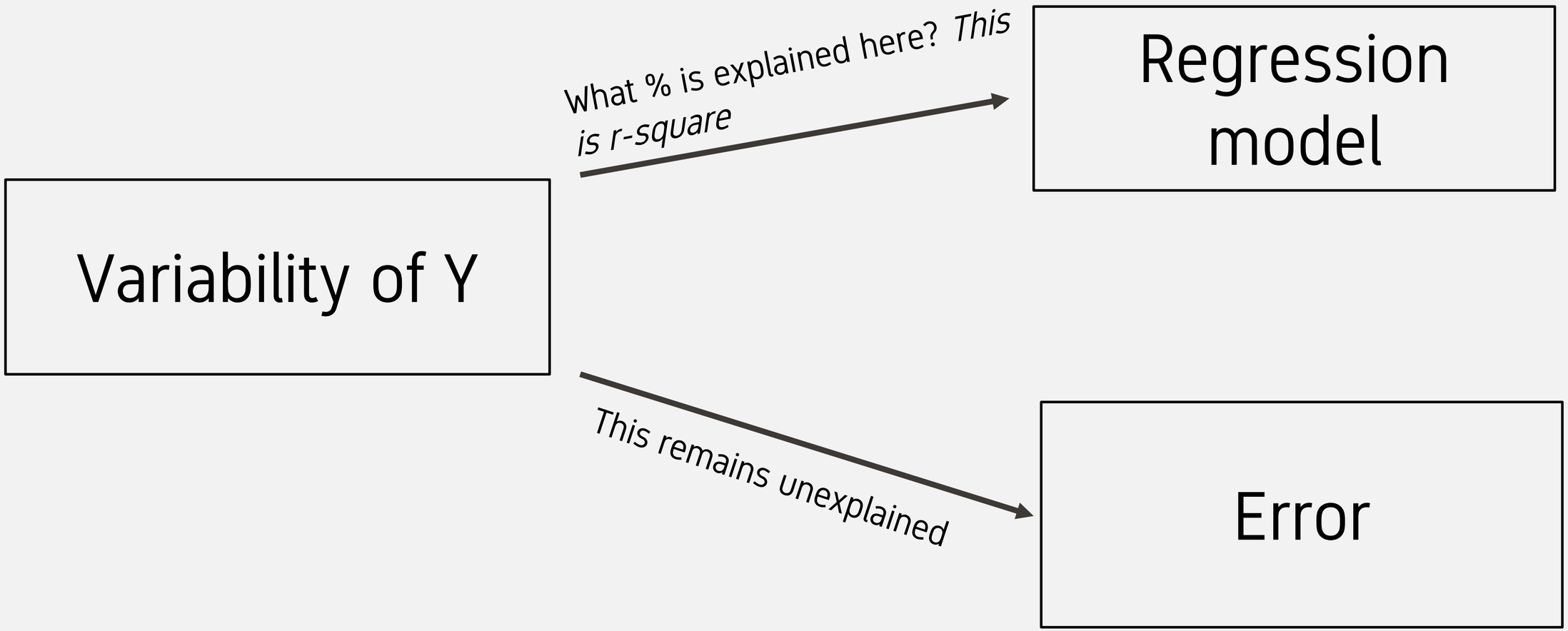
DRILL

- Researchers have found that the results of a linear regression of shark attacks (Y) on ice cream consumption (X) returns significant results.
  - *Do ice cream cones influence shark attacks?*

THE DEPENDENT, THE INDEPENDENT, AND THE CONFOUNDING



MODEL DIAGNOSTICS: R-SQUARE



INTERPRETING R-SQUARE

R-square value	Interpretation
.05	
.66	
.92	

INTERPRETING R-SQUARE

R-square value	Interpretation
.05	5% of the variability in Y is explained by X
.66	66% of the variability in Y is explained by X
.92	92% of the variability in Y is explained by X

MAKING POINT PREDICTIONS

$$\hat{Y} = \beta_0 + \beta_1 * X_i$$

$$\hat{Y} = 10 + .5 * \mathbf{4}$$

$$\mathbf{12} = 10 + 2$$

DEMO

- `mpg-regression-diagnostics.xlsx`
  - Locate and evaluate r-square
  - What is the predicted mpg for a car weighing 3,000 pounds?

DRILL

- `housing-regression-diagnostics-drill.xlsx`
  - Locate and evaluate r-square
  - What is the predicted displacement of a car weighing 3,000 pounds?

QUESTIONS?

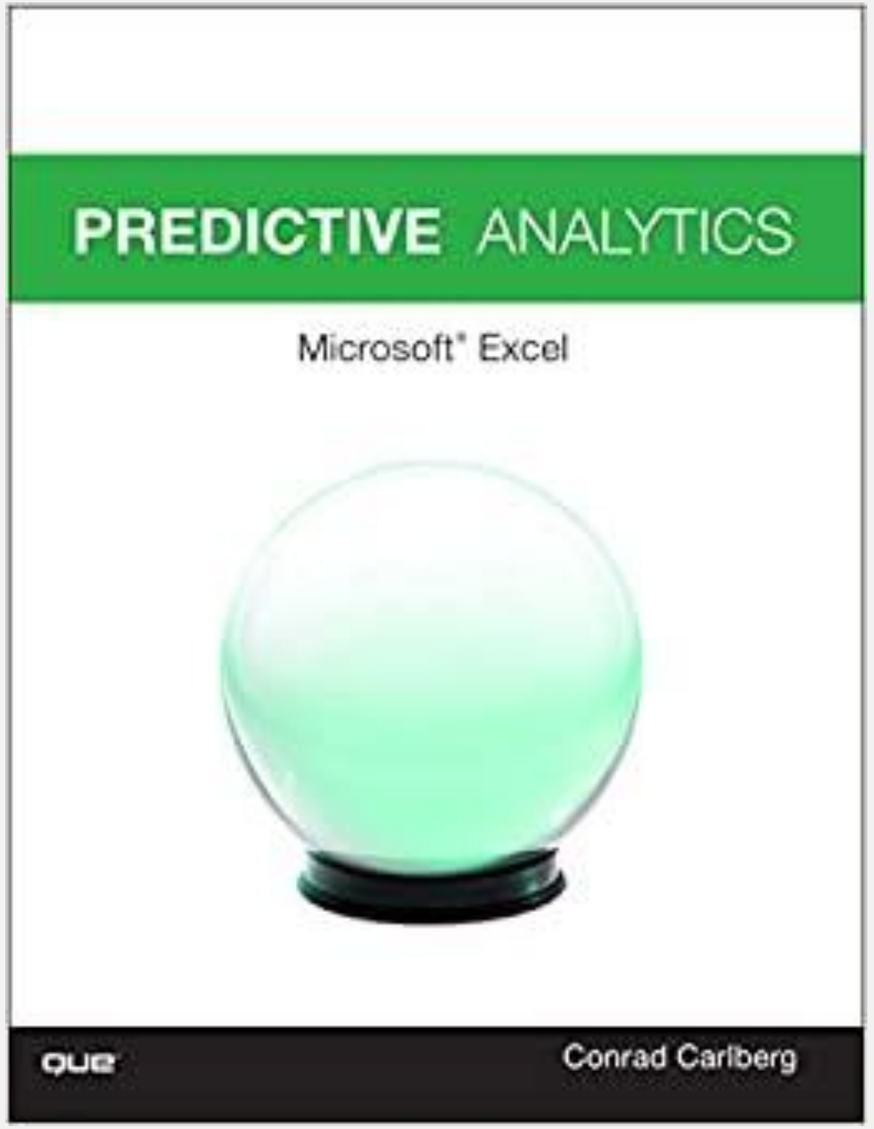
4. CONCLUSION

Future learning

- Continue exploring linear regression
  - Assumptions
  - Multiple regression
  - Regression with categorical variables
- Logistic regression
- Simulation and optimization

Predictive Analytics: Microsoft Excel, by Conrad Carlberg

- On O'Reilly Learning at https://learning.oreilly.com/library/view/predictive-analytics-microsoft/9780134682921/



***Data Smart: Using Data Science to Transform Information into Insight,* by John Foreman**

- On O'Reilly Learning at https://learning.oreilly.com/library/view/data-smart-using/9781118661468/

STRINGFEST ANALYTICS

LET'S TALK

LINKEDIN

linkedin.com/in/gjmount

EMAIL ADDRESS

george@stringfestanalytics.com

WEBSITE

stringfestanalytics.com

GITHUB

github.com/summerofgeorge

QUESTIONS?

5. BONUS

CAPSTONE

File: `capstone.xlsx`

Using the `hdma` dataset:

1. Is there a relationship between a public bad credit rating (`pbcr`) and being denied a mortgage application (`deny`) ?
2. Is there a difference in average housing expense to income ratio (`hir`) across mortgage credit score levels (`mcs`)?
3. Is there a significant relationship of loan-value ratio (`lvr`) on housing expense to income ratio (`hir`)?

Using the `ais` dataset:

4. Is there a significant relationship of hemaglobin concenration (hg) on hematocrit (hc)?

STATISTICAL SIGNIFICANCE OF CORRELATIONS

Ho: The correlation coefficient between these two variables equals zero.
Ha: The correlation coefficient between these two variables does not equal zero.

STATISTICAL SIGNIFICANCE OF CORRELATIONS

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

r = correlation coefficient

n = sample size

If abs(test statistic) > critical value,
reject the null

DEMO

- `Correlation-significance.xlsx`
  - Is the correlation between sepal length and sepal width statistically different from zero?