STRINGFEST ANALYTICS

ADVANCED 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

- Model a causal relationship between multiple independent variables and a categorical or continuous dependent variable
- Use simulation and optimization techniques to model business scenarios
- Build and evaluate forecasts
- Make compelling business recommendations using inferential statistics

WHY WOULD WE DO THIS IN EXCEL?

Advanced 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

Advanced 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. REGRESSION ANALYSIS AND PREDICTIVE MODELS

Warm-up

- File: `mpg-warmup.xlsx`
  - Calculate descriptive statistics and correlations for these variables
  - Draw scatterplots between mpg/weight, mpg/horsepower and mpg/displacement
  - *There's always room for descriptive statistics*

MULTIPLE LINEAR REGRESSION

EXPLICIT WARNING: MATH AHEAD

MULTIPLE REGRESSION EQUATION

Dependent / predictor variable

Y intercept

Slope coefficient(s)

Independent / response variable(s)

Error term(s)

$$Y_i = \beta_0 + \beta_1 * X_{1i} + \beta_1 * X_{1i} + \cdots + \beta_k * X_{ki} + \varepsilon_i$$

HYPOTHESES

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

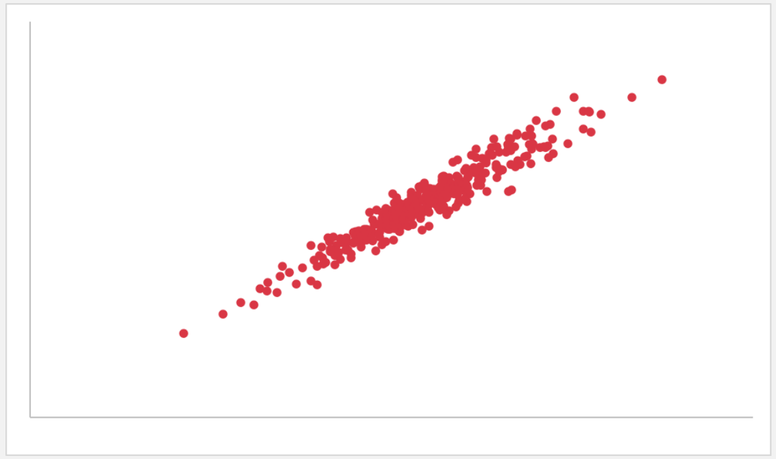
Ha: A relationship between X's and Y. The slope does not equal zero.

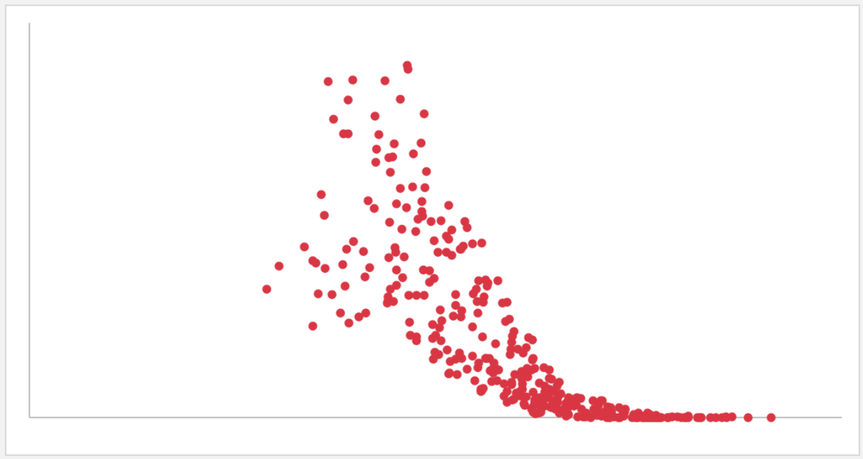
ASSUMPTIONS

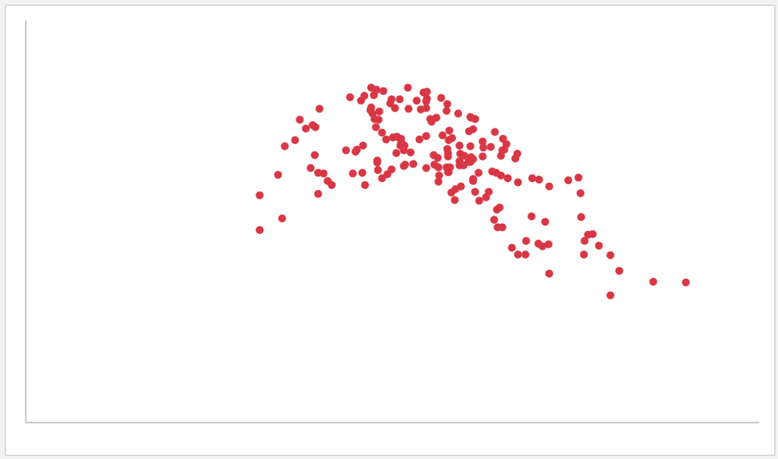
1. Linear relationship between independent and dependent variables
2. No influential cases
3. Variance of residuals is constant
4. Values of residuals are normally distributed
5. No multicollinearity

ASSUMPTIONS

Linear relationship between independent and dependent variables



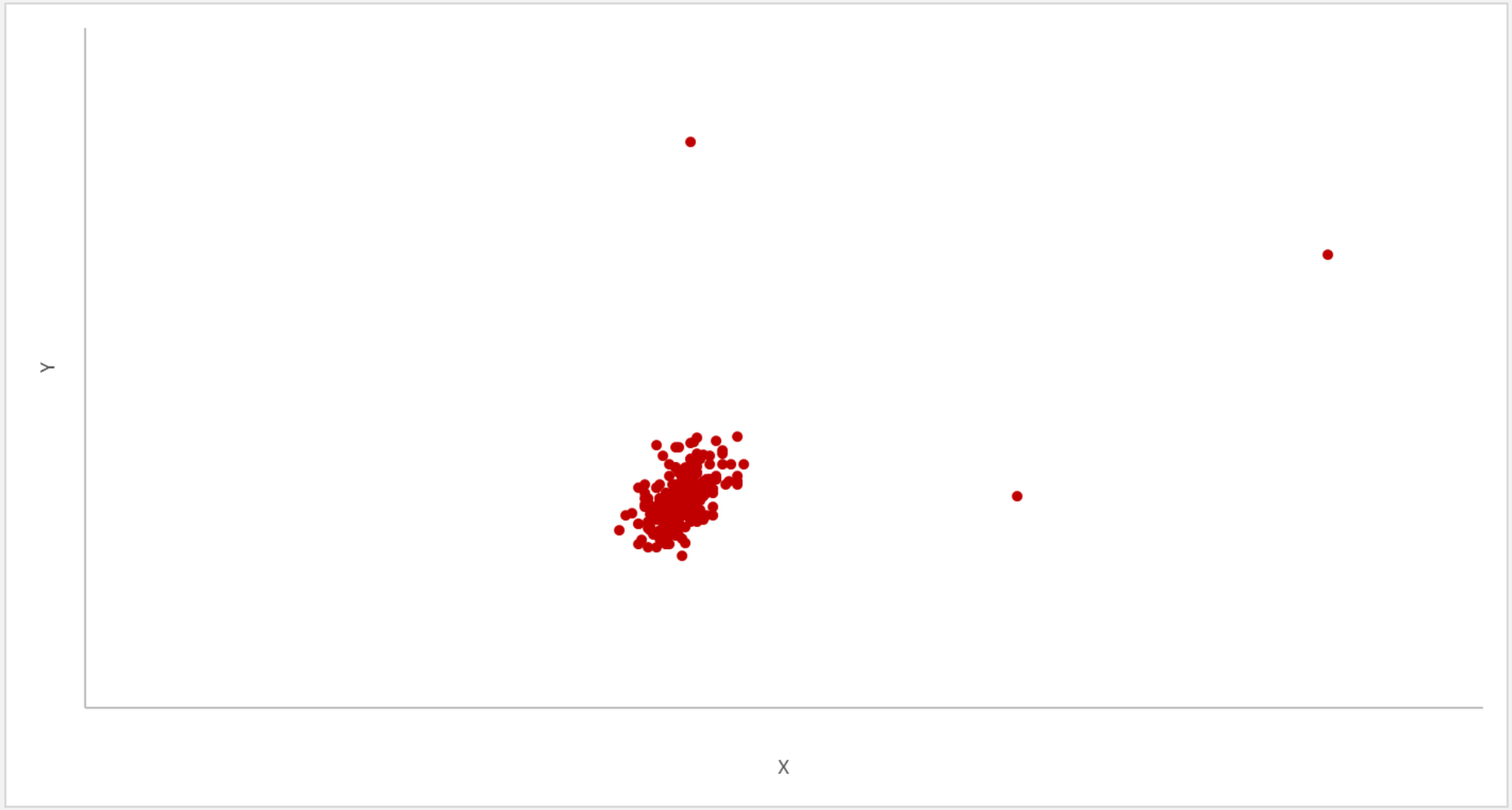




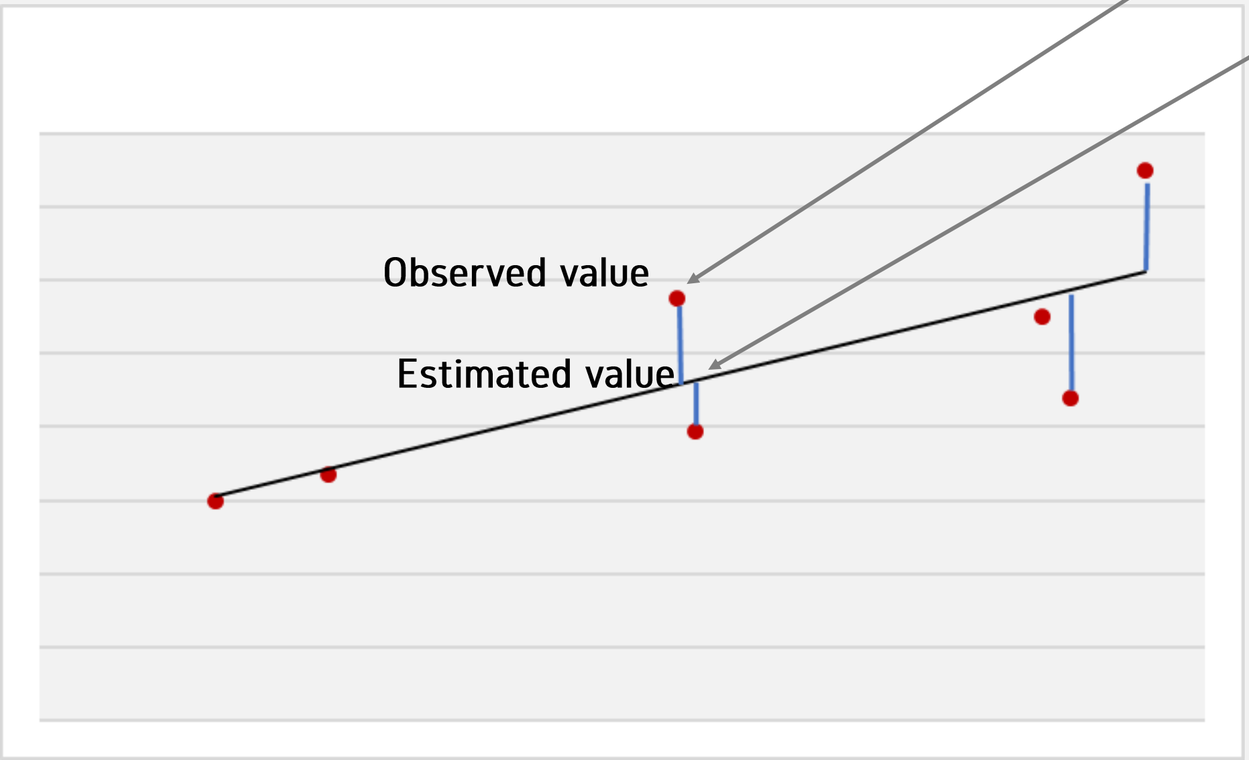
ASSUMPTIONS

No influential cases

Which of the below are actually influencing the line?

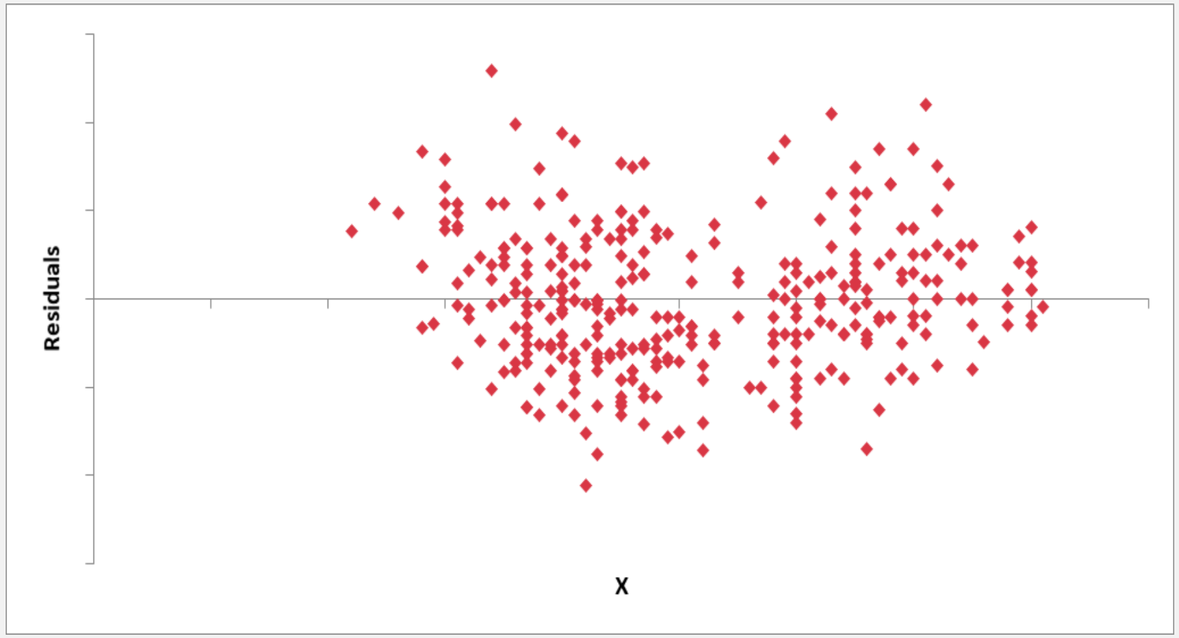


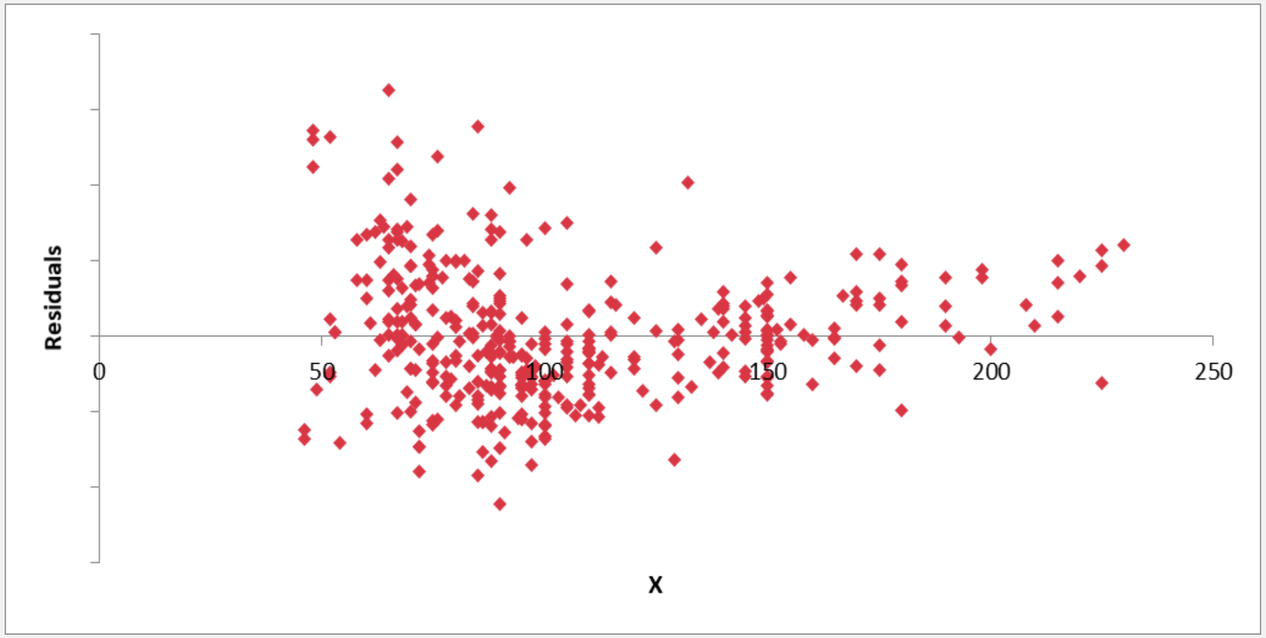
$$Residual = Y - \hat{Y}$$



ASSUMPTIONS

Variance of residuals is constant





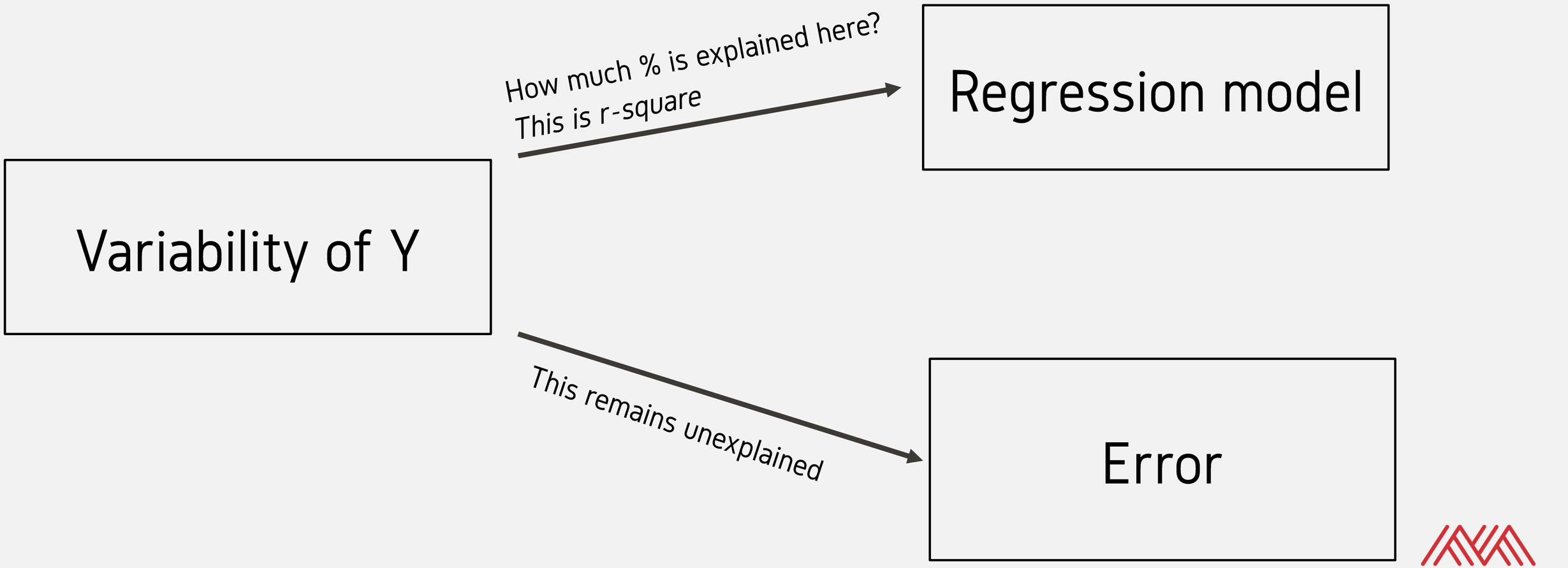
DEMO

- `mpg-regression.xlsx`
- Is there a significant relationship of displacement, horsepower and weight to mpg?
- *Check assumptions: linearity, no influential cases, independence of residuals*

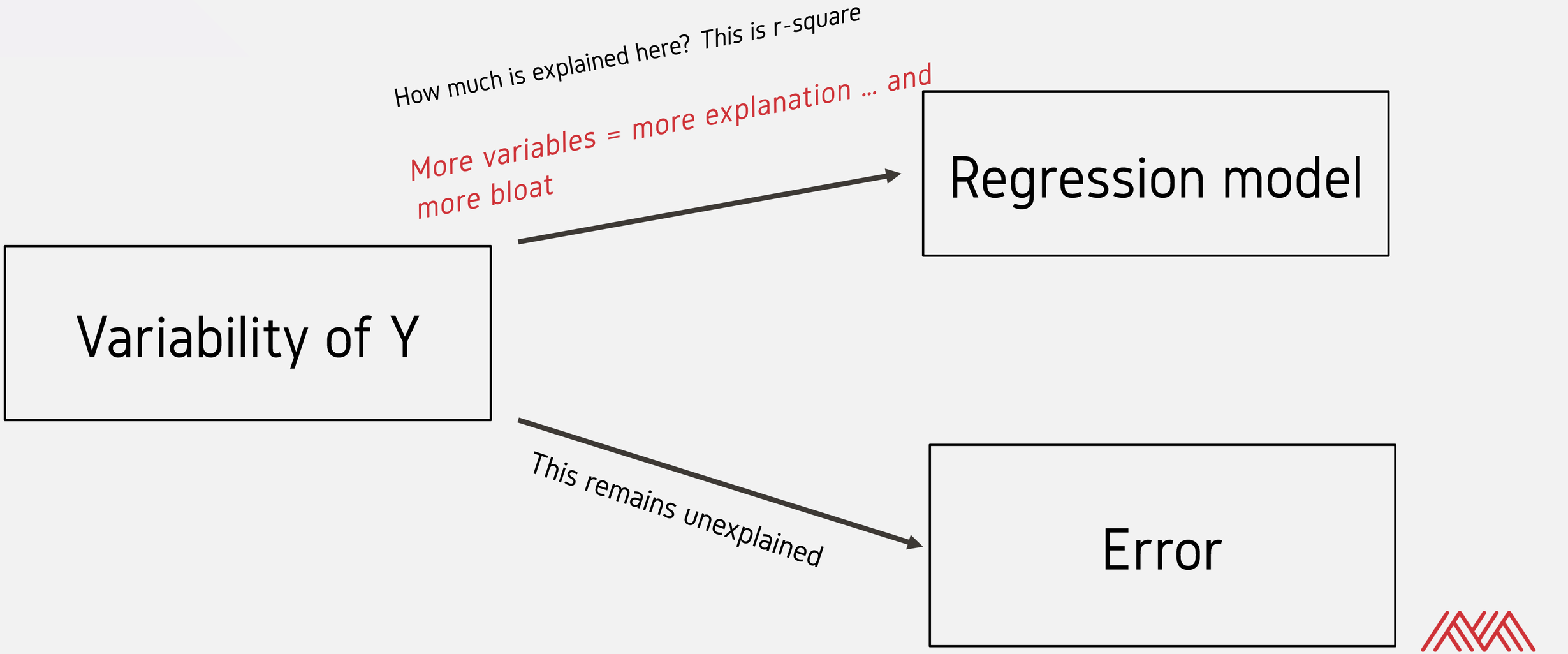
DRILL

- `penguins-linear.xlsx`
- Is there a significant relationship of culmen length, culmen depth and flipper length on body mass?
- *Check assumptions: linearity, no influential cases, independence of residuals*
  - Unhide influential-cases worksheet when read. No peeking!

MODEL DIAGNOSTICS: R-SQUARE



ADJUSTED R-SQUARE AND MODEL PARSIMONY



MAKING POINT PREDICTIONS

$$\hat{Y} = \beta_0 + \beta_1 * X_{i1} + \beta_1 * X_{i2}$$

$$\hat{Y} = 10 + .5 * 4 + 1.5 * 2$$

$$\mathbf{15} = 10 + 2 + 3$$

DEMO

- `mpg-regression-diagnostics.xlsx`
- What is the *adjusted* R-square of this model?
- What is the expected MPG of a car weighing 3,000 pounds that has 200 horsepower?

DRILL

- `penguins-linear-diagnostics.xlsx`
- What is the R-square of this model?
- What is the expected body mass of a penguin with a flipper length of 200 mm?

QUESTIONS?

INTERACTION TERMS

When IV's conspire: *interaction terms*

- High heat is uncomfortable
- High humidity is uncomfortable
- High heat * high humidity is *even more* uncomfortable
  - The effect of heat (IV) on discomfort (DV) is *different at different values* of humidity (another IV)

REGRESSION WITH INTERACTION

Dependent / predictor variable

Y intercept

Main effects

Interaction effect

Error term

$$Y_i = \beta_0 + \beta_1 * X_{1i} + \beta_2 * X_{2i} + \boldsymbol{\beta_3 * X_{1i} * X_{2i}} + \varepsilon_i$$

DEMO

- `airquality-interaction.xlsx`
- What is the effect of solar radiation and temperature on ozone? What about solar radiation * temperature?
  - Run the regression without the interaction term first (*parsimony*)

DRILL

- `wine-interaction.xlsx`
- What is the influence of fixed and volatile acidity on pH? What about fixed * volatile?

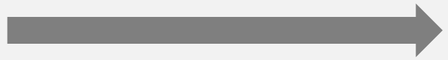
LINEAR REGRESSION WITH CATEGORICAL IV'S

What is 2*USA? 4.5*Europe?

- Linear regression assumes continuous independent variables
- How can we use categorical variables?
  - We can *encode* them as a series of 0-1 values

DUMMY-CODING

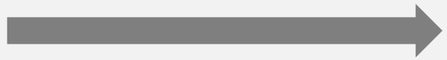
Sex	Height
Male	72
Female	67
Female	62
Male	74
Female	71
Male	68



Male	Female	Height
0	0	72
0	1	67
0	1	62
0	0	74
0	1	71
0	0	68

REGRESSION WITH DUMMIES

Sex	Height
Male	72
Female	67
Female	62
Male	74
Female	71
Male	68



Male	Female	Height
0	0	72
0	1	67
0	1	62
0	0	74
0	1	71
0	0	68

$$Height = \beta_0 + \beta_1 X_1$$

X_1 = 1 when female; otherwise 0.

DEMO

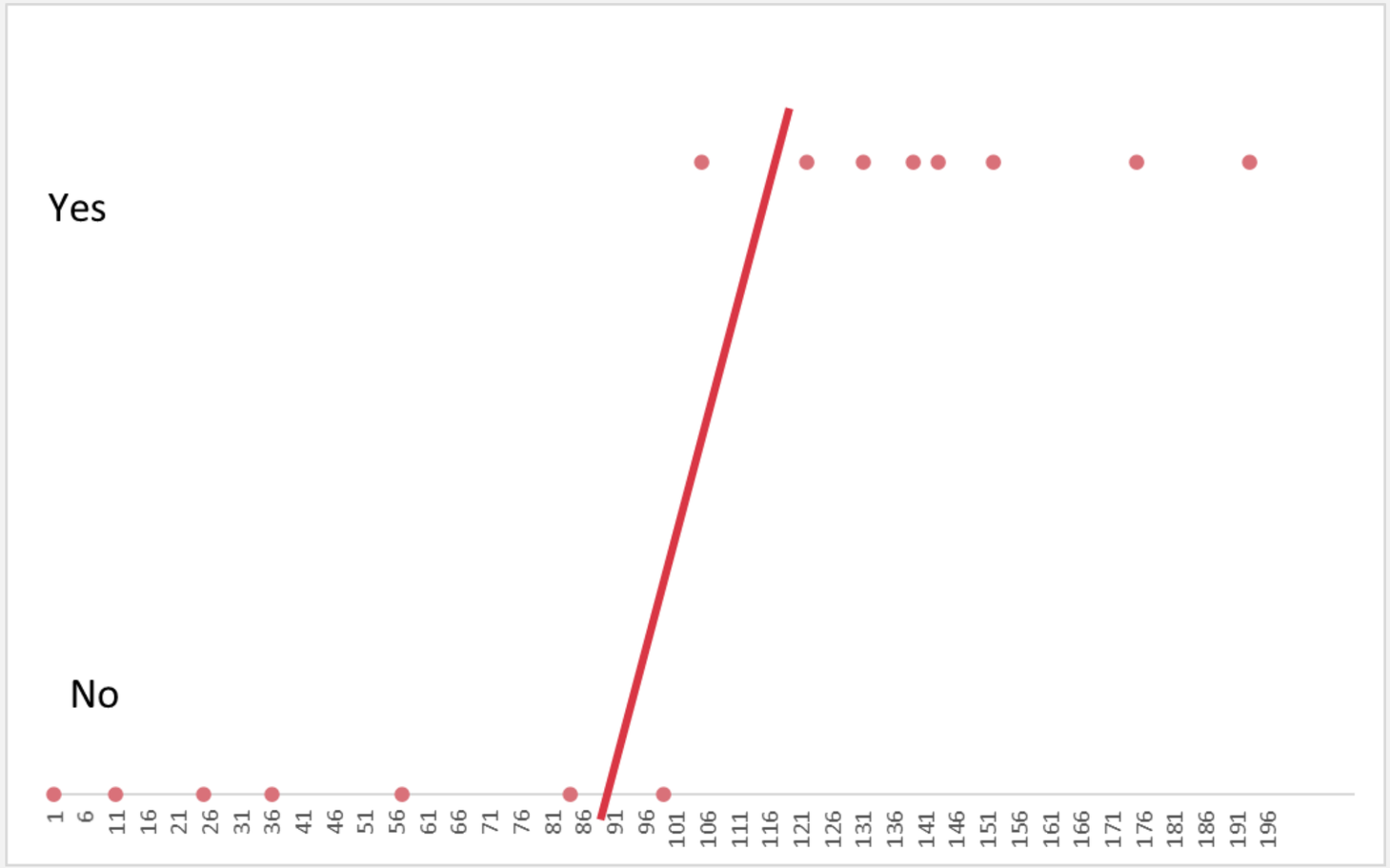
- `mpg-dummy.xlsx`
- Does the car's origin have significant influence on its mileage?
  - What do we expect the mileage to be for each origin for a car weighing 3,000 pounds?

DRILL

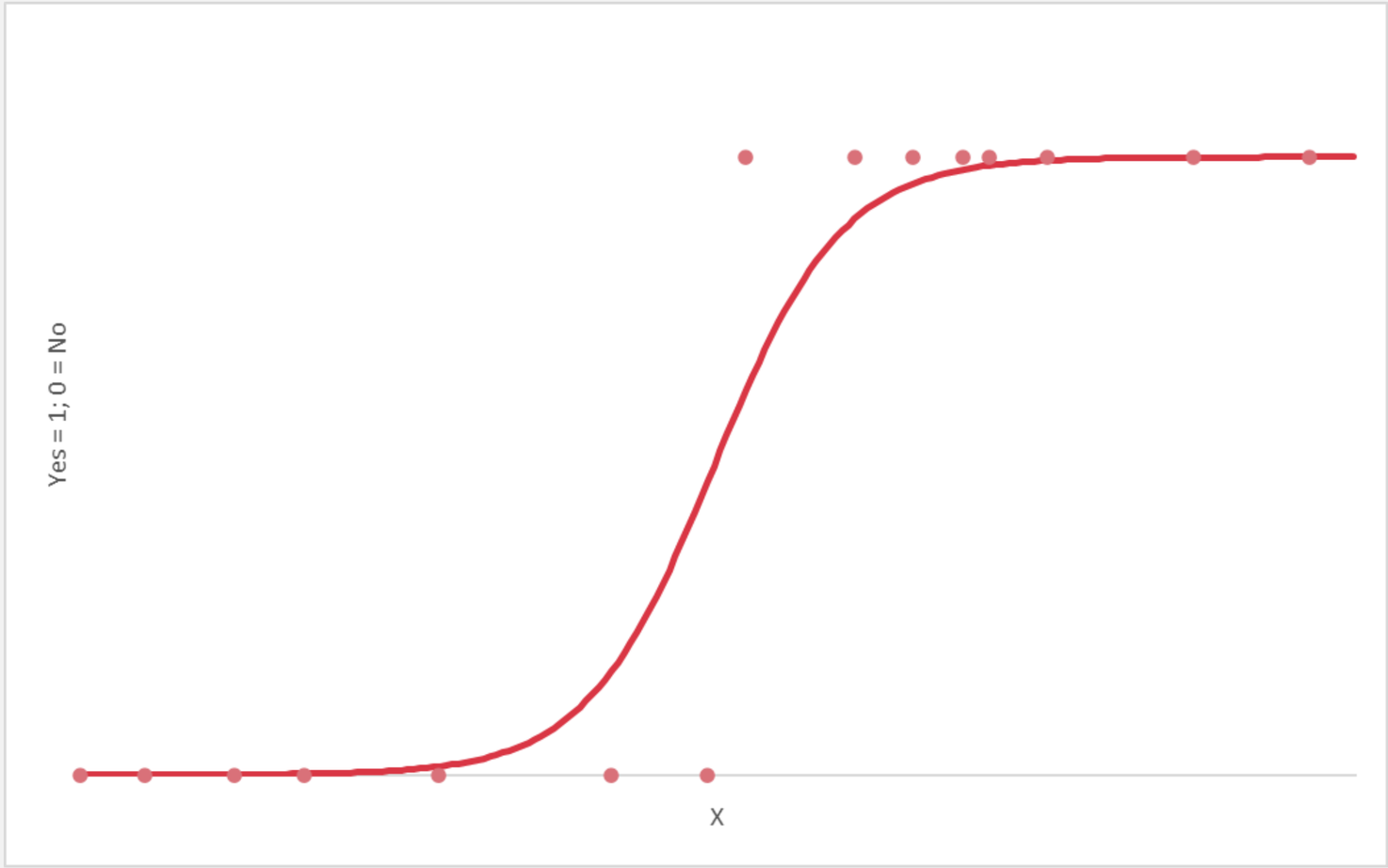
- `penguins-dummy.xlsx`
- Regress sex on height

LOGISTIC REGRESSION

HOW DO YOU MODEL A BINARY OUTCOME?



FITTING A BINARY OUTCOME: LIKE ANTS ON A LOGARITHM



ASSUMPTIONS

1. Binary dependent variable
2. Observations are independent
3. Large sample size
4. Linearity of independent variables and log odds
5. No influential cases
6. No multicollinearity

LOGISTIC REGRESSION EQUATION

Dependent / predictor variable (*expressed in probability)*

Y intercept

Independent / response variable(s)

Slope coefficient(s)

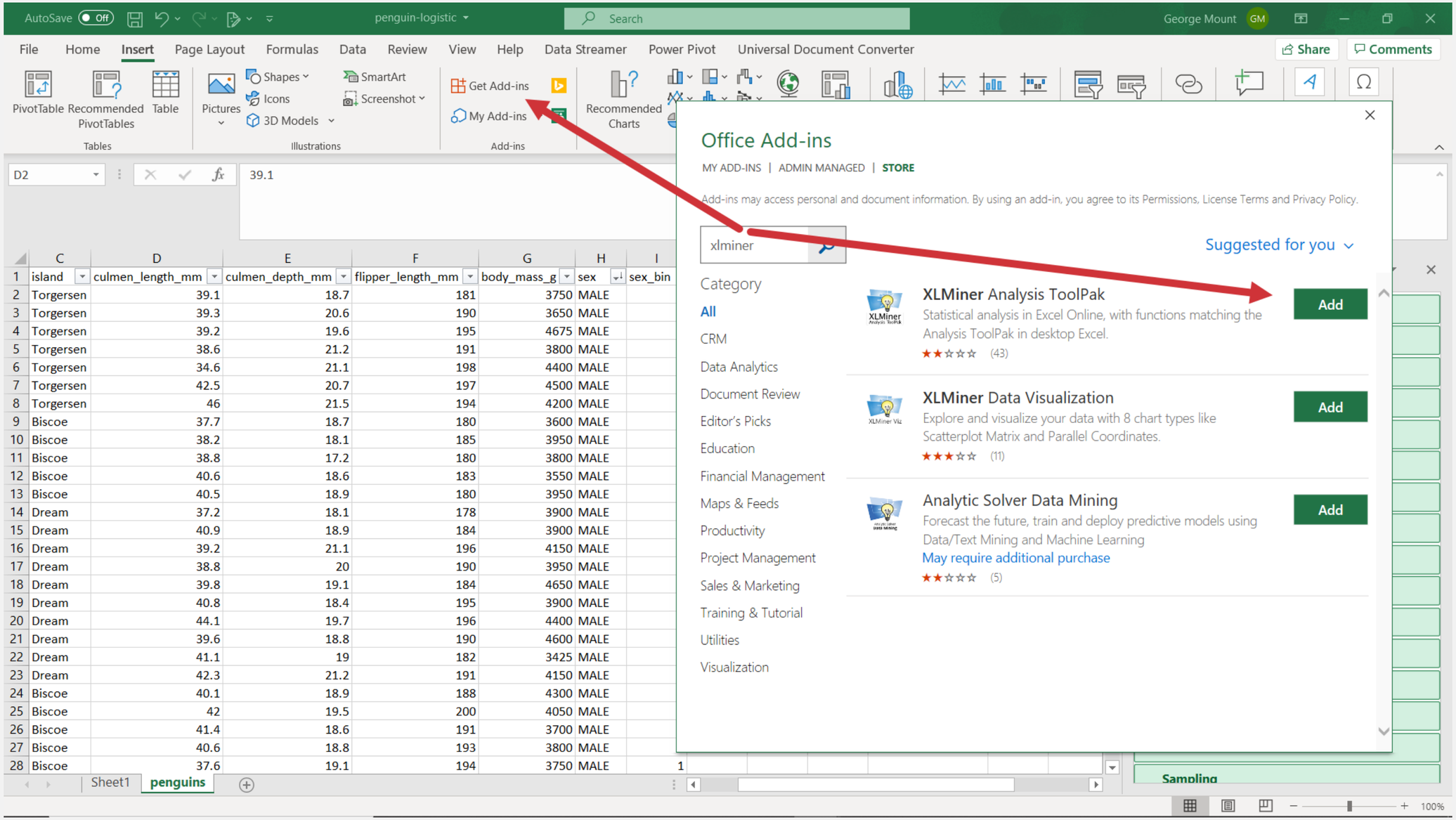
$$P = \frac{e^{\beta 0 + \beta 1 * x1i + \cdots + \beta k * xki}}{1 + e^{\beta 0 + \beta 1 * x1i + \cdots + \beta k * xki}}$$

HYPOTHESES

Ho: No relationship between X's and Y. The Y values you predict are no closer to the actual Y values than you would expect by chance.

Ha: A relationship between X's and Y. The slope does not equal zero. The Y values you predict are closer to the actual Y values than you would expect by chance.

INSTALLING XLMINER



	C	D	E	F	G	H	I
1	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex	sex_bin
2	Torgersen	39.1	18.7	181	3750	MALE	
3	Torgersen	39.3	20.6	190	3650	MALE	
4	Torgersen	39.2	19.6	195	4675	MALE	
5	Torgersen	38.6	21.2	191	3800	MALE	
6	Torgersen	34.6	21.1	198	4400	MALE	
7	Torgersen	42.5	20.7	197	4500	MALE	
8	Torgersen	46	21.5	194	4200	MALE	
9	Biscoe	37.7	18.7	180	3600	MALE	
10	Biscoe	38.2	18.1	185	3950	MALE	
11	Biscoe	38.8	17.2	180	3800	MALE	
12	Biscoe	40.6	18.6	183	3550	MALE	
13	Biscoe	40.5	18.9	180	3950	MALE	
14	Dream	37.2	18.1	178	3900	MALE	
15	Dream	40.9	18.9	184	3900	MALE	
16	Dream	39.2	21.1	196	4150	MALE	
17	Dream	38.8	20	190	3950	MALE	
18	Dream	39.8	19.1	184	4650	MALE	
19	Dream	40.8	18.4	195	3900	MALE	
20	Dream	44.1	19.7	196	4400	MALE	
21	Dream	39.6	18.8	190	4600	MALE	
22	Dream	41.1	19	182	3425	MALE	
23	Dream	42.3	21.2	191	4150	MALE	
24	Biscoe	40.1	18.9	188	4300	MALE	
25	Biscoe	42	19.5	200	4050	MALE	
26	Biscoe	41.4	18.6	191	3700	MALE	
27	Biscoe	40.6	18.8	193	3800	MALE	
28	Biscoe	37.6	19.1	194	3750	MALE	1

DEMO

- `occupancy.xlsx`
- Is there a significant relationship of temperature, humidity and light to occupancy?

DRILL

- `penguin-logistic.xlsx`
- Is there a significant relationship of culmen length, culmen depth, flipper length and body mass and sex?
  - *Use the 0-1 sex_bin variable*

MAKING POINT PREDICTIONS

$$\hat{P} = \frac{e^{-10+.01*750}}{1+e^{-10+.01*750}}$$

$$.08 = \frac{e^{-10+.01*750}}{1+e^{-10+.01*750}}$$

% PREDICTIVE ACCURACY

Actual	Predicted % chance	Predicted	Predicted right?
0	8.00%	0	TRUE
0	68.56%	1	FALSE
1	47.13%	0	FALSE
0	31.08%	0	TRUE
0	73.63%	1	FALSE
0	27.96%	0	TRUE
0	7.59%	0	TRUE
0	89.61%	1	FALSE
0	19.36%	0	TRUE
1	61.49%	1	TRUE

DEMO

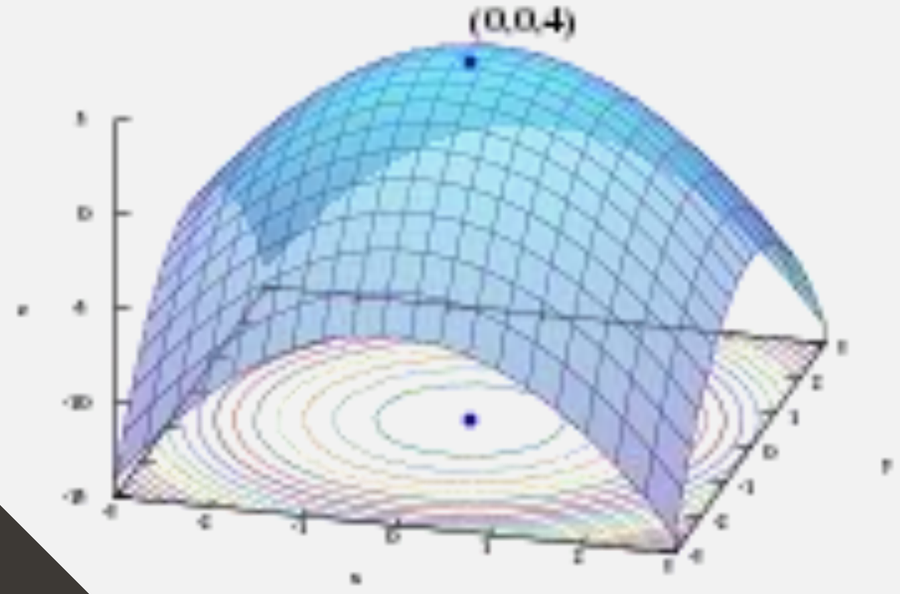
- `occupancy-diagnostics.xlsx`
- Do we predict that a room with temperature 25, humidity 30 and light 400?
- What is the predictive accuracy of our dataset?

DRILL

- `penguin-logistic-diagnostics.xlsx`
- Do we predict that a penguin with a culmen length of 40 mm and a flipper length of 175 mm is male or female?
- What is the predictive accuracy of our dataset?

QUESTIONS?

2. OPTIMIZATION AND SIMULATION



OPTIMIZATION: GETTING THE MOST BANG FOR YOUR BUCK

https://en.wikipedia.org/wiki/Mathematical_optimization

DEMO

- File: `grades.xlsx`
- “What grade do I need on the final to get an A for the class?”
- Use Goal Seek

DRILL

- File: `sales-price.xlsx`
- How many doo-hickeys do you need to sell to raise $1,000?

Limitations of Goal Seek

- No *constraints*
  - Inputs must be integers
  - Inputs must not exceed a given amount
- No *max/min* objectives
  - Maximize profits given a mix of resources and costs
  - Minimize distance traveled given a route of stops

HAVE YOU INSTALLED THE SOLVER ADD-IN?

ON WINDOWS:

- File
- Options
- Add-ins
- Go
- Check on Solver Add-in
- OK

ON MAC:

- Tools
- Excel Add-ins
- Check on Solver Add-in
- Click OK

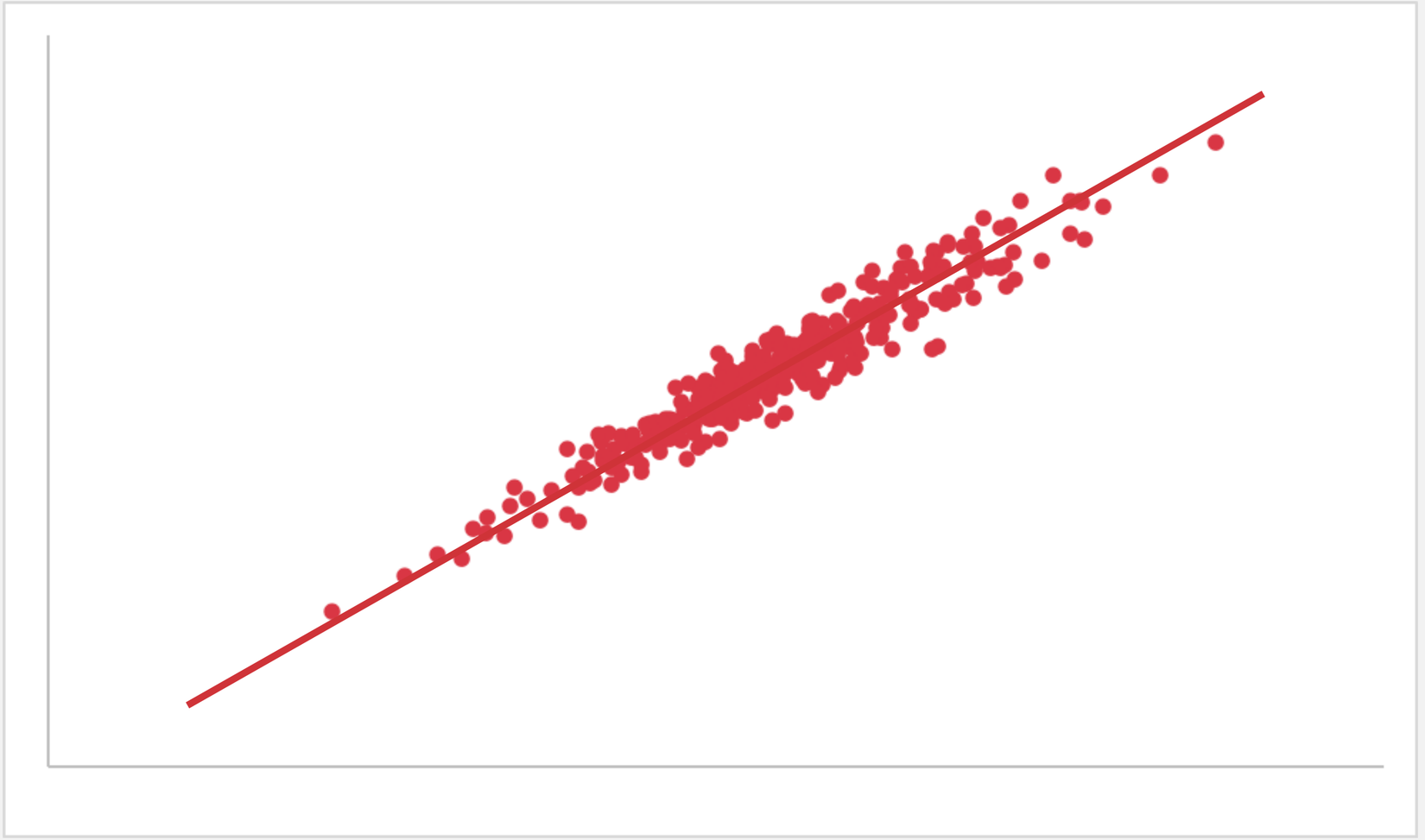
DEMO

- File: `product-mix.xlsx`
  - How many units of Product A and Product B should we produce to maximize revenue, *given the amount of materials we have*?
  - Use Solver

DRILL

- File: `unit-production.xlsx`
  - How many units of Product A and Product B should we produce to maximize revenue, given the available labor, parts and shipping units?
  - Integer constraint: You can't produce a quarter of a unit!

QUESTIONS?



DEMO

- File: `solver-regression.xlsx`
  - Fit a regression line using Solver
  - Minimize the sum of squared errors

SIMULATION: WHEN IN DOUBT, PLAY IT OVER AND OVER ON A COMPUTER

Pick a number, any number (following the distribution)…

- Demo: `inverse-distribution.xlsx`
- What is the “inverse” of a probability distribution?

DEMO

- `widget-sales.xlsx`
- Our goal is to sell $10,000.
  - Sales follow a normal distribution of $1,000 a day on average with a standard deviation of $300.
- How many days can we expect it to take to sell to $10,000?

DEMO

- `monte-carlo.xlsx`
- Sales follow a normal distribution of $7,500 with a standard deviation of $1,500.
- Fixed costs always equal $2,000.
- Variable costs follow a normal distribution of $3,000 with a standard deviation of $1,000.
- Run 1,000 simulations based on these statistics. What percent of time is a loss expected?

QUESTIONS?

3. TIME SERIES AND FORECASTING

“I have seen the future and it is very much like the present, only longer.” –Kehlog Albran, *The Profit*

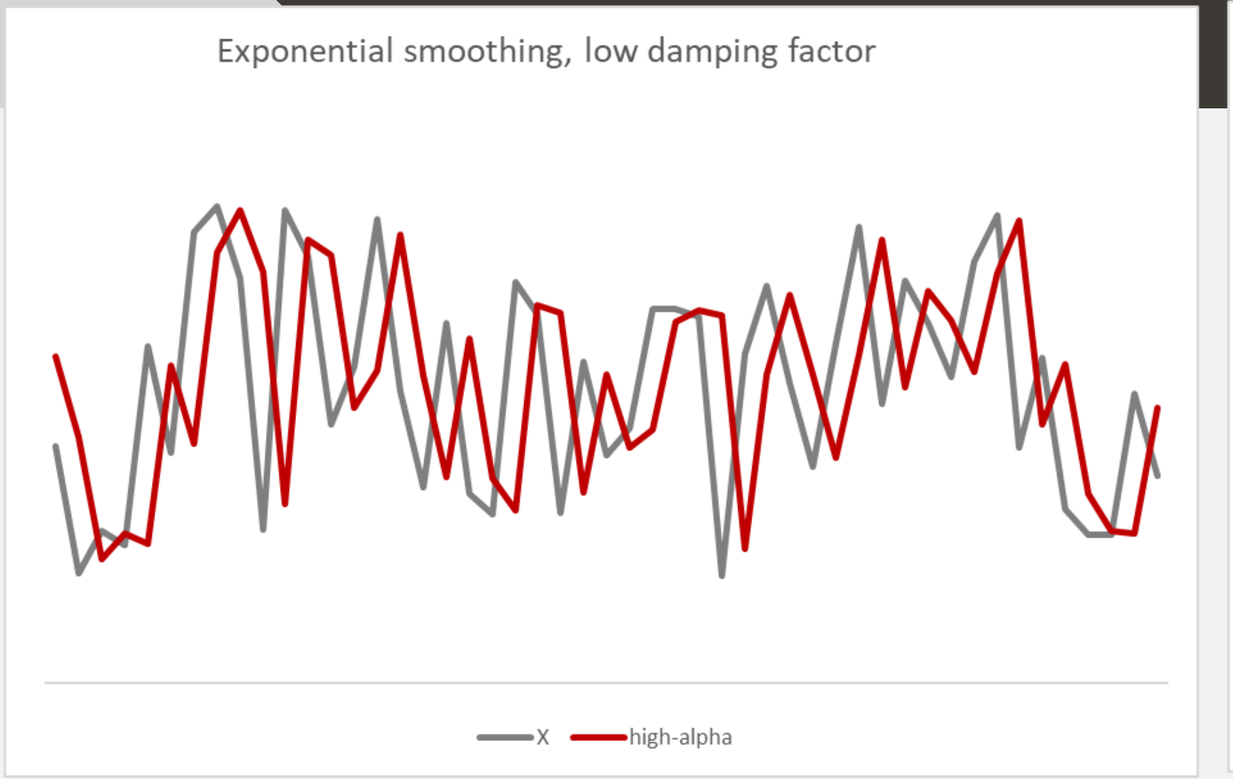
“Forecasting is the art of saying what will happen, and then explaining why it didn’t.” -- Anonymous

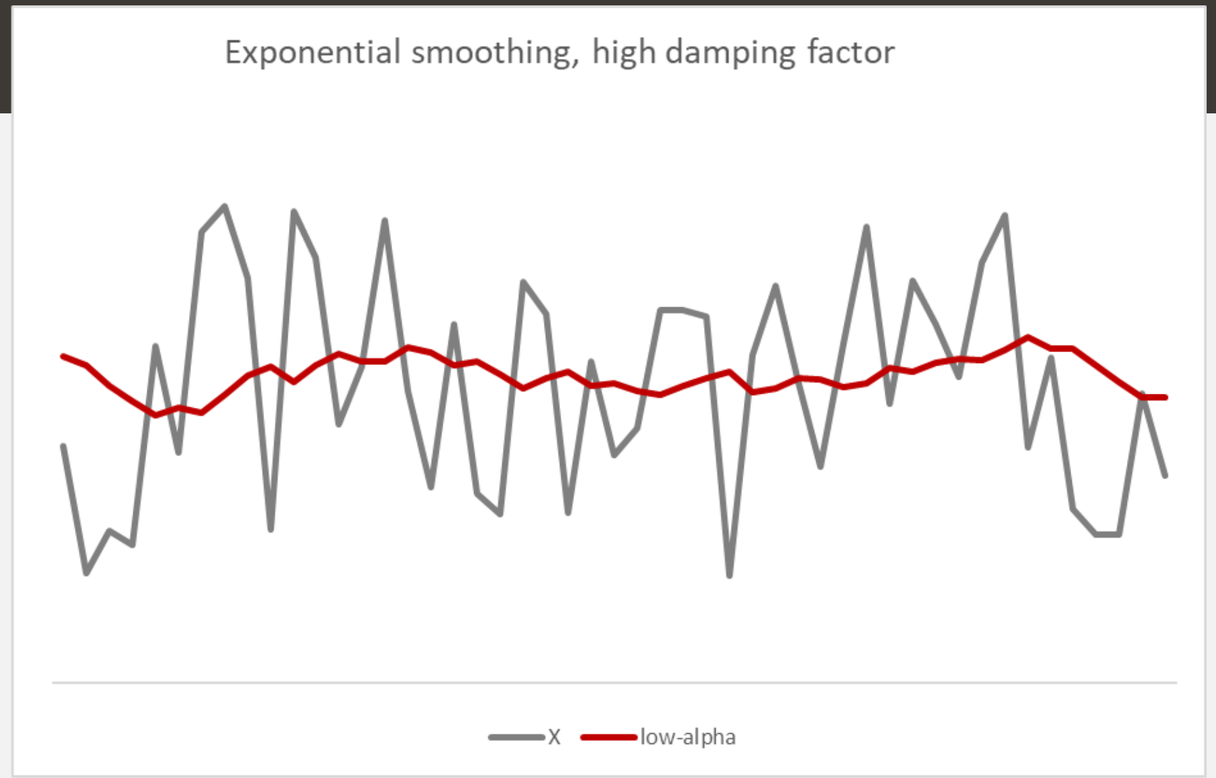
How do we try to predict the future?

- Qualitative: surveys, market research

- Quantitative:
  - Just use last period's value ("naïve")
  - Take a rolling average of the last *X* months
  - Assign heavier weight to more recent data points and smaller weight to less recent data points (exponential smoothing)

Exponential smoothing: damping factors





More weight to more recent data points, less to past

Less weight to more recent data points, more to past

How do we know if we've predicted the future well?

- Mean absolute error (MAE)
- Root mean squared error (RMSE)
- **Mean absolute percentage error (MAPE): what is the absolute percentage error (actual <> forecast) of an observation on average?**

DEMO

- `sp500.xlsx`
- What is the MAPE of
  - a naïve forecast?
  - a four-day rolling-average forecast?
  - an exponential smoothing forecast with a 20% damping factor?
    - Can we do any better on damping factor?

DEMO

- `female-births.xlsx`
- What is the MAPE of
  - a naïve forecast?
  - a four-day rolling-average forecast?
  - an exponential smoothing forecast with a 20% damping factor?
    - Can we do any better on damping factor?

QUESTIONS?

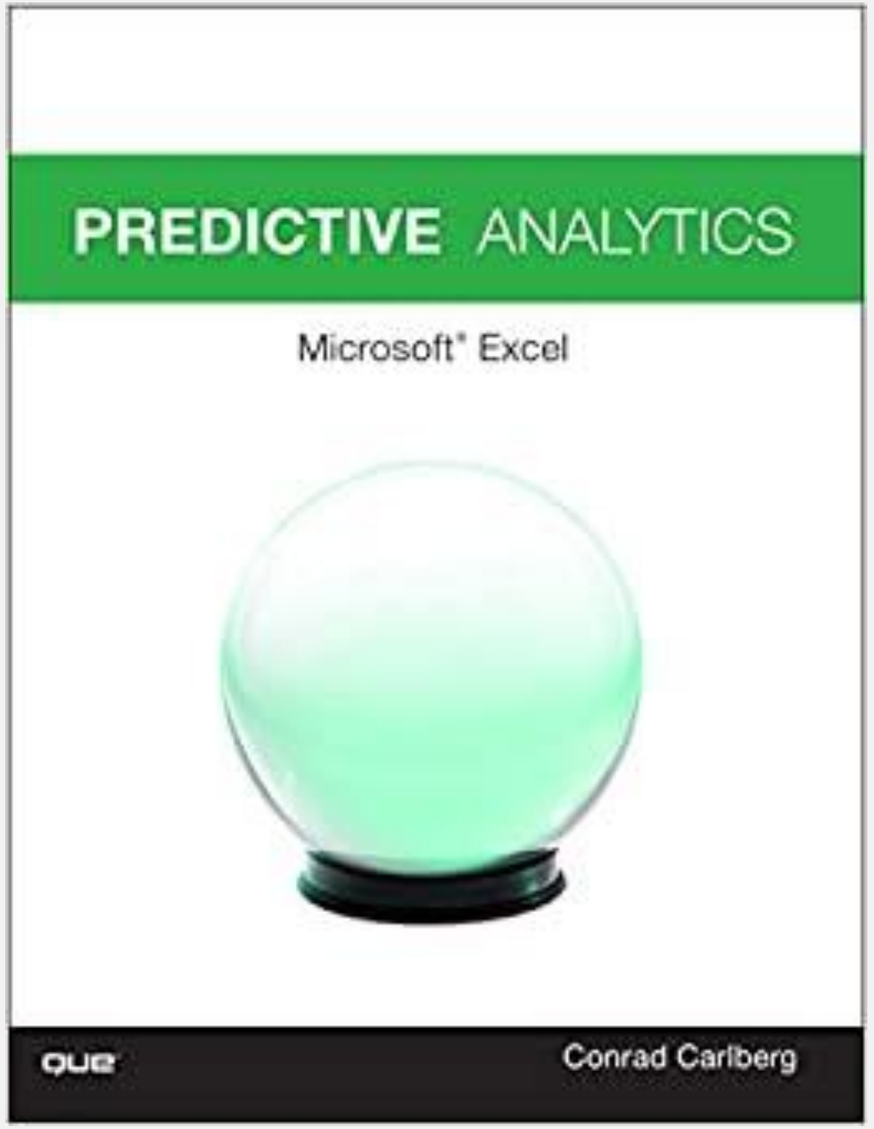
4. CONCLUSION

Future learning

- Statistical programming with R
- Further exploratory data analysis & data preparation
  - Outlier detection, treatment and removal
  - Handling missing values
  - Dimensionality reduction
  - Forecasting for trends and seasonality

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?