



# Master of Science in Business Analytics

**Machine Learning I**  
**OPAN 6602**

**Week 2 Live Session**

**Instructor:**  
**Tommy Jones, PhD**

*GEORGETOWN*  
*UNIVERSITY* / *McDONOUGH*  
*SCHOOL of BUSINESS*

---

**“All models are wrong. Some Models are useful.”**

– George Box

**“An approximate answer to the right question is worth far more than a precise answer to the wrong one.”**

– John Tukey

---

# Agenda

1. Finish Week 1 Example
2. Template for Supervised Learning Scripts/Notebooks
3. Identifying Regression Pathologies
4. Guided Exercise

# Agenda

1. **Finish Week 1 Example**
2. Template for Supervised Learning Scripts/Notebooks
3. Identifying Regression Pathologies
4. Guided Exercise

# From Week 1: Automatic Variable Selection

Reminder: Good for prediction only tasks, sketchy for explanation.

It's not unusual for forward selection, backwards elimination, and stepwise selection to give the same result.

Variable selection in Python works slightly differently than described in the asynchronous material.

# Agenda

1. Finish Week 1 Example
2. **Template for Supervised Learning Scripts/Notebooks**
3. Identifying Regression Pathologies
4. Guided Exercise

# Open OPAN\_6602\_Template.ipynb

# Agenda

1. Finish Week 1 Example
2. Template for Supervised Learning Scripts/Notebooks
3. **Identifying Regression Pathologies**
4. Guided Exercise

# Multicollinearity

- What is it?
  - Two or more predictors are highly linearly correlated
- What is its effect?
  - Coefficient estimates are inconsistent.
- How to detect it?
  - Variance inflation factor.
  - $VIF > 3$  is cause for concern
- How to correct for it?
  - Omit one variable.
  - Transform one variable.

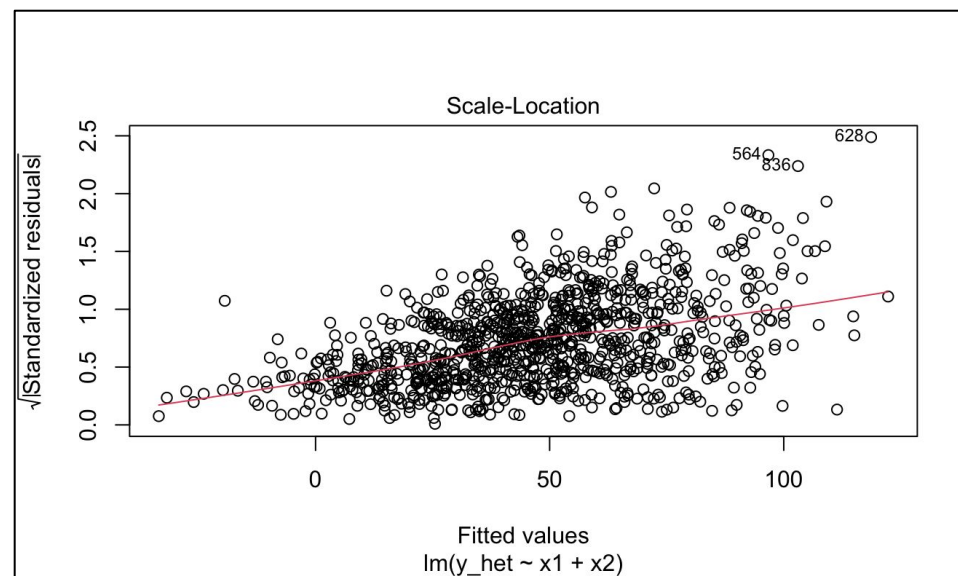
$$VIF_i = \frac{1}{1 - R_i^2}$$

**where:**

$R_i^2$  = Unadjusted coefficient of determination for regressing the  $i$ th independent variable on the remaining ones

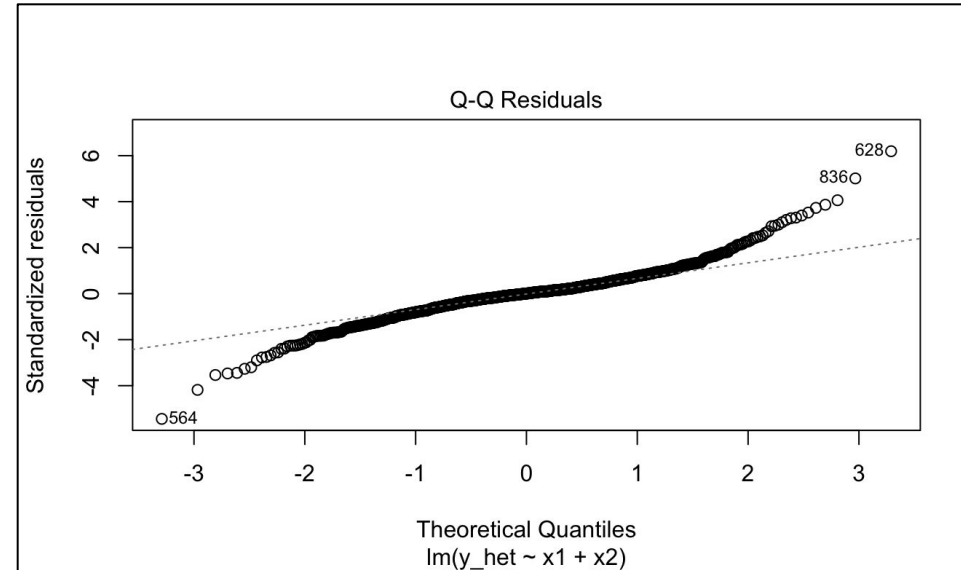
# Non-Constant Error Variance AKA Heteroskedasticity

- What is it?
  - The variance of the error term changes with the dependent variable or independent var(s)
- What is its effect?
  - SEs are overly optimistic.
- How to detect it?
  - Plotting.
  - Breusch-Pagan test.
- How to correct for it?
  - "Robust" standard errors
  - Variable transformations



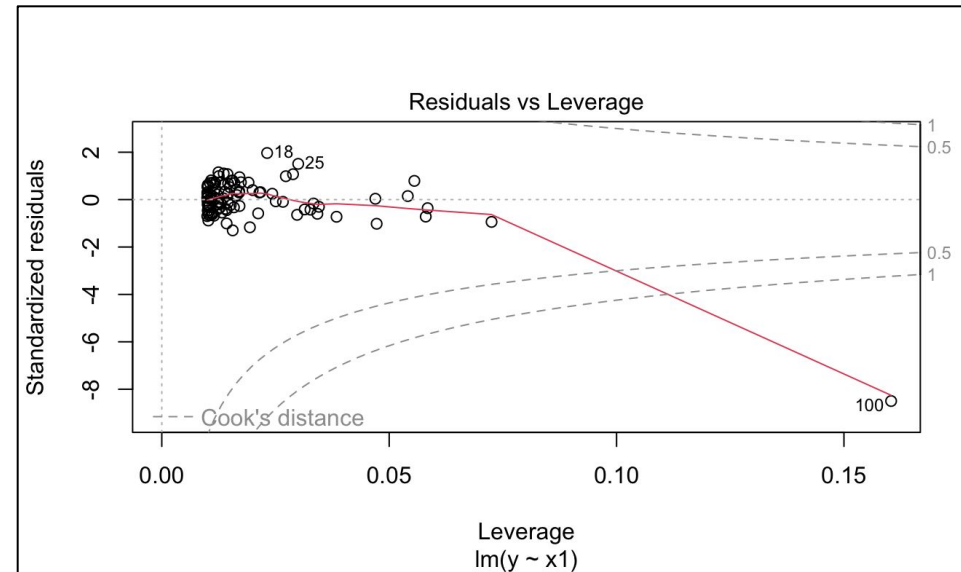
# Error Term Has Non-Normal Distribution

- What is it?
  - The residuals do not follow a normal distribution
- What does it mean?
  - It implies your model is misspecified.
  - *Possibly* not a linear relationship.
- How to detect it?
  - Plotting
  - Formal tests (e.g., Shapiro Wilk)
- How to correct for it?
  - Variable transformations



# Influential Observations

- What is it?
  - One of your observations is very outside of the norm of others
- What is its effect?
  - It biases your regression line.
- How to detect it?
  - Cook's distance
- How to correct for it?
  - Remove it as an outlier... **maybe**
  - Weighted least squares
  - Median regression



# Two Criteria for Removing Outliers

If you want to remove an outlier, one of the following needs to be true

1. **Measurement error**  
e.g., a 100 foot tall human
2. **Not part of the population of interest**  
e.g., a fire at a factory means that factory's output data is not representative of other factories in your company.

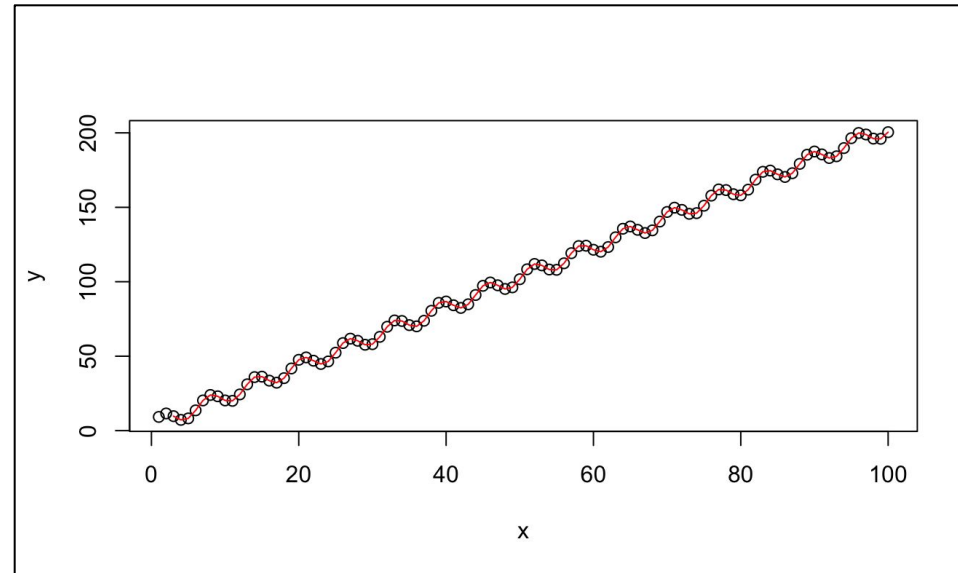
If your goal is purely prediction, you might be able to relax the above. But still... be careful.

# Omitted Variable Bias

- What is it?
  - You didn't include (or have) a variable that matters
- What is its effect?
  - Inconsistent (i.e., asymptotically biased) coefficient estimates
- How to detect it?
  - No statistical test. Understand your domain. Think hard about what you include.
  - Can sometimes show up in residual plots
- How to correct for it?
  - Collect data on all relevant variables

# Autocorrelation

- What is it?
  - Errors for observations are correlated
- What is its effect?
  - Coefficients inconsistent.  
You'll get big prediction misses down the line
- How to detect it?
  - Visual inspection
  - Durbin-Watson test
- How to correct for it?
  - In a time series context: add lags of Y or X as appropriate



# Agenda

1. Finish Week 1 Example
2. Template for Supervised Learning Scripts/Notebooks
3. Identifying Regression Pathologies
4. **Guided Exercise**

# Open OPAN\_6602\_Week\_02\_Exercise.ipynb