Module 4

# Reading

*PROJECT: \*\*\*Finalize rough draft of project analysis paper\*\*\* DUE TONIGHT*

*Read*

* (Cody) Chapter 11: Correlation
* (Cody) Chapter 12: Simple and Multiple Regression
* (Cody) Chapter 13: Binary Logistic Regression
* (OpenIntro) Chapter 8: Introduction to linear regression

*Watch*

* Marin Stats:
  + <https://youtu.be/vblX9JVpHE8> -- Simple Linear Regression
  + <https://youtu.be/tOzwEv0PoZk> -- Nonlinearity in Simple Linear Regression
  + <https://youtu.be/GI8ohuIGjJA> -- R-squared
* Calling Bullshit
  + <https://youtu.be/YAAHJm1pi1E> -- Calling Bullshit 3.1: Correlation and Causation
  + <https://youtu.be/BKQqKKjAwqM> -- Calling Bullshit 3.2: What is Correlation?
  + <https://youtu.be/WNsLcg2GQMY> -- Calling Bullshit 3.3: Spurious Correlation

Exercises:

* (Cody) Chapter 11 -- Exercise 3 (pg 139)
* (Cody) Chapter 12 -- Exercise 2,3, and 4 (pg 161-162)
* (Cody) Chapter 13 -- Exercise 1,2, and 3 (pg 175)
* [DUE Saturday November 14, 2020 @ 9:00AM (LINK)](https://elearn.stonehill.edu/webapps/assignment/uploadAssignment?content_id=_638657_1&course_id=_16488_1&group_id=&mode=cpview)

# SAS Workshop: Correlations

*Goal:* Explore hypothesis

*Pearson’s Correlation --*

*Ho*: The two variables are not linearly related.  
*Ha*: The two variables are linearly related.

*Spearman’s Correlation* –

*Ho*: The ranks of the two variables are not linearly related (ρ = 0).  
*HA*: The ranks of the two variables are linearly related (ρ ≠ 0).

*Additional Stats Guidance:*

Pearson’s Correlation -- <http://sites.utexas.edu/sos/guided/inferential/numeric/bivariate/cor/>

Spearman’s Rank Correlation -- <http://sites.utexas.edu/sos/guided/inferential/numeric/bivariate/rankcor/>

*Data: Preliminary 2020 Presidential Election Data --* [*https://github.com/alex/nyt-2020-election-scraper*](https://github.com/alex/nyt-2020-election-scraper)

*Methods:*

* 1. Load the preliminary election data using this code: <https://github.com/developing-bioinformatics/DAN602-Stats4Analytics/blob/main/snippets/election_regression.sas>. These commands will read the file directly from the GitHub repository csv file and do some initial data manipulation. Specifically, this code:
     + Changes the ‘vote\_differential’ variable to be negative when Biden is in the lead. This way the response is linear when the leading candidate switches during the vote count.
     + Filters for data from Pennsylvania. We will take a close look at the vote count in this state and then you can modify the code to explore other states later. Note that this code uses SQL syntax to find partial matches in the state column.   
         
       ‘create table WORK.filter as select \* from WORK.IMPORT where(state LIKE "Pennsylvania%");’  
         
       The ‘state LIKE “Pennsylvania%”’ syntax will find a partial match with “LIKE” and the “%” wildcard character. This is necessary here because the state column includes notation on how many Electoral College votes each state has. So we cannot match using equality against just the state name.
  2. Examine the scatterplot that this code generates. **Describe the relationship between vote\_differential and time.**
  3. Set up a correlation analysis. Tasks & Utilities 🡪 Statistics 🡪 Correlation. Use default statistics. **What is the correlation and p-value? Which method is being used?**
  4. Look at the code. How would you change this to use the Spearman’s correlation method? Try it by copying this code block into your program file and **report your results.**
  5. **Which method, Pearson’s or Spearman’s correlation, is a better representation for these data? Why?**

**Put all of the code into a single program file (if not already written that way) and save this file and turn it in with your answers to the questions above.**

# SAS Workshop: Regular Linear Regression

*Goal: Implement linear regression for cases where correlation analysis suggests there is a significant linear relationship between two quantitative parameters. Evaluate model fit and discuss meaning of results.*

*Additional Stats Guidance:*

*Linear Regression --* [*http://sites.utexas.edu/sos/guided/inferential/numeric/bivariate/cor/*](http://sites.utexas.edu/sos/guided/inferential/numeric/bivariate/cor/) *(below correlation)*

*Data: Preliminary 2020 Presidential Election Data --* [*https://github.com/alex/nyt-2020-election-scraper*](https://github.com/alex/nyt-2020-election-scraper)

*\*\*In the last section we looked at how the vote differential changed over the days after the election. Eventually the vote differential came to zero and the leading candidate switched. The trend towards this is clear and well supported from the correlation analysis.   
  
Could we use another representation of these data to tell whether we could have predicted the outcome? Can we tell that the vote differential would be 0 or lower before all votes were counted or all precincts reported?*

*Methods:*

* + 1. Set up a Linear Regression analysis. Tasks & Utilities 🡪 Statistics 🡪 Linear Regression. Set:
       - Dependent variable as “vote\_differential” (Can the votes remaining predict the vote differential?)
       - Continuous variable as “votes\_remaining”
       - Add a filter statement in the linear regression menu that filters the data for those observations with > 1 million votes remaining to be counted.  
           
         “votes\_remaining >= 1000000”
       - In the model tab edit the model effects to add “votes\_remaining” to the model effects. Keep “Intercept” checked.
    2. Run the regression model. **What is the R-squared value for this model?**  
         
       **What is the predicted vote differential when there are no votes remaining to count? (HINT: What is the y-intercept of this model?) What is the uncertainty (standard error) of this estimate?   
         
       With 1 million votes remaining to count could we conclude the correct outcome for the state of Pennsylvania?**
    3. Re-run the regression model for those observations with > 500,000 votes remaining to be counted.  
         
       **NOW:** **What is the predicted vote differential when there are no votes remaining to count? (HINT: What is the y-intercept of this model?) What is the uncertainty (standard error) of this estimate? How did these value change since the million votes left model and why?  
         
       With 500,000 votes remaining to count could we conclude the correct outcome for the state of Pennsylvania?**
    4. Re-run the regression models above but use the ‘percent\_reporting’ variable (which is the percent of precincts reporting data) as the model effect instead of ‘votes\_remaining’. **What differences do you observe in the results? (NOTE: You cannot compare the y-intercept values directly. Instead you have to solve the regression equation for the value for vote\_differential when percent\_reporting = 1. OR solve for the value of percent\_reporting when vote\_differential = 0.)**

**Put all of the code into a single program file (if not already written that way) and save this file and turn it in with your answers to the questions above.**

# SAS Workshop: Multiple Linear Regression

*Goal:* Implement multiple regression models and interpret results for model fit and variable contribution.

*Data: The IBM HR employee attrition database* [*https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset*](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset)

*Methods:*

* 1. Load the IBM HR employee attrition database with [*https://github.com/developing-bioinformatics/DAN602-Stats4Analytics/blob/main/snippets/attrition\_load.sas*](https://github.com/developing-bioinformatics/DAN602-Stats4Analytics/blob/main/snippets/attrition_load.sas)*a*
  2. Setup multiple linear regression. Tasks & Utilities 🡪 Statistics 🡪 Linear Regression. Set:
     + Dependent variable as “MonthlyIncome”
     + Continuous variable as “YearsAtCompany”
     + Model tab – Model effects, add “YearsAtCompany” to the model.
  3. Run the single variable setup. **What is the R-squared value? Look at the plot output; is there a linear relationship between years at company and monthly income? (Check the residuals\*)**
  4. Edit the model; add:
     + Continuous variable “Education” and “JobInvolvement” to the model effects
     + Model tab – Model effects, add “YearsAtCompany”, “Education”, and “JobInvolvement” to the model.
  5. Run the multiple regression. **What is the R-squared value for the new model? Did it change from the single variable model?   
       
     Look at the residual for each explanatory variable. Examine the statistics tables.   
       
     Which variables seem to be most related to MonthlyIncome? Are there any variables you would remove?   
       
     Are there any variables from the database you would add? Try one more combination and run the model again. Interpret your results here.**

**Put all of the code into a single program file (if not already written that way) and save this file and turn it in with your answers to the questions above.**

# SAS Workshop: Logistic Regression

*Goal:* Implement logistic regression models and interpret results for model fit.

***Hypotheses:*** *Each predictor will have its own set of hypotheses:*

*Ho: Controlling for all other predictors in the model, this predictor variable does not explain variation in the outcome.  
HA: Controlling for all other predictors in the model, this predictor variable does explain variation in the outcome.*

*\*\*For single variable logistic regression the hypothesis does not require the control statement.*

*\*\*The ‘outcome’ is some binary value that we would like to predict. Our hypothesis testing reveals whether there is a statistical association between the independent variables and the outcome*

*Additional Stats Guidance:*

*Logistic Regression --* [*http://sites.utexas.edu/sos/guided/inferential/categorical/logistic/*](http://sites.utexas.edu/sos/guided/inferential/categorical/logistic/)

*Data: The IBM HR employee attrition database* [*https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset*](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset)

*Methods:*

* 1. Load the IBM HR employee attrition database with [*https://github.com/developing-bioinformatics/DAN602-Stats4Analytics/blob/main/snippets/attrition\_load.sas*](https://github.com/developing-bioinformatics/DAN602-Stats4Analytics/blob/main/snippets/attrition_load.sas)
  2. Set up a Logistic Regression analysis between Attrition (Yes/No) and MonthlyIncome. Tasks & Utilities 🡪 Statistics 🡪 Binary Logistic Regression. Set:
     + Response variable as ‘Attrition’
     + Link function ‘Logit’ (\*this is what most people mean when they refer to Logistic regression’)
     + Explanatory Variables – Continuous variables as “MonthlyIncome”
     + Options Tab –
       - Select Default and additional plots (pick Effect plot and ROC plot)
  3. Run
  4. Review results.  **Examine the graph “Predicted probabilities for Attrition=Yes”. What does this graph show? Does income influence the probability of employee attrition?   
       
     Look at the ROC plot and the AUC (Area Under the Curve) value reported there. What does this show?**
  5. Look through the data for another quantitative variable that might be associated with employee attrition. Test the effect using logistic regression and **report your results here.  
       
     (HINT: Try some scatterplots first to see if it looks like there is an effect.)**

**Put all of the code into a single program file (if not already written that way) and save this file and turn it in with your answers to the questions above.**

# Post-Class:

\*\*Complete any remaining analyses from today’s SAS Lab\*\*

**Final Project**

***Peer Review – \*\*Groups to be assigned\*\**** *In the last portion of class and later on Slack present your project findings to your group (2-3 other students) and help each other with feedback or further development of your analysis. You may share any part of your project (writing, code, visualizations, slides) with your group. As a peer reviewer you can give any suggestions or provide any assistance relevant to improving the project.*

***Paper Revisions –*** *I will give you feedback on your paper drafts this week (before November 22, 2020). Take that feedback and make appropriate adjustments to your paper. If you’d like to meet this week schedule a time using* [*https://calendly.com/rharbert/scheduling*](https://calendly.com/rharbert/scheduling)*.*

***Project Presentations –*** *On our final meeting (December 5, 2020) you will each give a short presentation (~10 minutes with extra time for questions) about your project.*

*Format is flexible but it could be a slide show, a SAS demonstration, or a combination of the two. One effective strategy could be to prepare slides of background information and design a SAS program to run the analysis and produce appropriate visuals that you could walk through during your presentation. Another approach would be to prepare a slide show with background info, a description of your methods and hypotheses, and highlight the most important statistical results and graphics.*

*You may choose to assign the class some background reading about your project. A news article or blog post might be a good primer on your topic so everyone is ready to dive into your analysis.*

*Be prepared for follow up questions about your analysis and the choices you made regarding your statistics and representations of your data and findings.*

**Presentations and Final Paper DUE: December 5, 2020**