# Lab #8 – Regression Modeling

A major strength of regression is that it allows you to assess the impact of a variable while adjusting for confounders. The dataset you will analyze in this lab comes from a major US university and its ongoing evaluation of salary differences between male and female faculty members. The data are titled “Professor Salaries” on p-web.

The dataset contains the following variables:

* **Rank** – a categorical variable with levels “AsstProf”, “AssocProf”, and “Prof”. Generally speaking, new professors are usually hired at the rank of Assistant Professor and after several years of productivity they are promoted to Associate Professor or released. The promotion to Full Professor occurs after several additional years of productivity.
* **Discipline** – A binary variable with levels “A” (theoretical departments) and “B” (applied departments)
* **Yrs.since.phd** – The number of years since the professor received their PhD
* **Yrs.service** – The number of years the professor has been working for the university
* **Sex** – Whether the professor is male or female
* **Salary** – The 9 month salary of the professor (in dollars)

The goal of this analysis will be to determine if a statistically significant male-female salary gap exists at this university after adjusting for potential confounding variables.

## Part #1 – Preliminary Analysis

In this section you will explore the variables in the data and how they relate to each other. This is an important step in any thorough statistical analysis.

### Question #1

Construct a histogram of the variable “Salary”. Do you see any outliers or skew? What could you do to address potential skew/outliers? (you don’t need to actually do this, regression doesn’t require a normally distributed response variable – only normally distributed errors).

### Question #2

Compare the mean salary of male and female professors using a two-sample t-test. Do males or female have higher salaries? Could the salary differences in these data be due to random chance?

### Question #3

The simple approach used in Question 2 rules out random chance, but it doesn’t address confounding. Use the definition of confounding to determine whether the variable “yrs.since.phd” is a confounder in the relationship between “Sex” and “Salary”. Describe your findings in 1-2 sentences and include any relevant summary statistics or hypothesis testing results.

### Question #4

Similar to Question 3, determine whether “Rank” is a confounding variable.

### Question #5 A

Using a method we’ve previously discussed, how could you address “Rank” being a cofounding variable in the relationship between “Salary” and “Sex”? Explain in 1-2 sentences (you don’t need to formally carry-out your described approach)

### Question #5 B

Could you reasonably use the same approach you described in Question 5-A to address “yrs.since.phd” as a confounding variable? Briefly explain why or why not.

## Part #2 – Model Building

So far we’ve established that there are multiple confounding variables influencing the relationship between “Sex” and “Salary”. In this section we will build a regression model that sufficiently adjusts for these variables.

### Question #6-A

Fit a regression model that uses “years.since.phd” to predict salary. Is there a linear relationship between the two variables sufficient or should a quadratic term be included? Briefly explain and use an ANOVA test to support your decision.

### Question #6-B

Is a cubic term necessary to adequately capture the relationship between “years.since.phd” and “salary”? Briefly explain and use an ANOVA test to support your decision.

### Question #7

Fit a multiple linear regression model that uses both “Sex” and “yrs.since.phd” to predict “Salary”. In this model, include “yrs.since.phd” as a linear effect. Using the “Graphs” menu under “Fit Regression Model” request a plot of residuals vs the variable: “yrs.since.phd”. Include this plot in your write-up and briefly explain what you see in the plot and how it might relate to your findings in Question #6.

### Question #8-A

Now fit a multiple regression model that predicts “Salary” and includes “Sex” and a quadratic effect for “yrs.since.phd”. See “How to include a quadratic term” at the end of this lab for how to do this. Report your model’s coefficients in your lab write-up (ie: replace the betas in the formula below with the actual regression effects)

### Question #8-B

Using the model you fit in Question 8-A, is there a significant difference in salary for male/female employees? Briefly explain, supporting your conclusion with a p-value.

### Question #9-A

Now add the variable “Rank” to your model that includes “Sex” and a quadratic effect of “yrs.since.phd”. Using the ANOVA table, does “Rank” significantly improve the model? (brief justification which includes a p-value).

### Question #9-B

Using the model you fit in Question 9-A, is there a significant difference in salary for male/female employees? Briefly explain, supporting your conclusion with a p-value.

### Question #9-C

Why does the p-value for the variable “Sex” change when “Rank” is added to the model? Briefly explain.

### Question #10-A

Now add “Discipline” as a predictor in the model. Use the ANOVA table to determine whether this variable improves the model.

### Question #10-B

Interpret the regression coefficient of the variable “Discipline”. Do professors in theoretical or applied disciplines have higher salaries (after adjusting for rank, sex, and yrs.since.phd)?

### Question #11

Now add “yrs.service” to the model and pay attention to a column in the coefficient table we’ve yet to look at titled “VIF”

VIF stands for Variance Inflation Factor, it is a measure of how much an explanatory variable overlaps with the other variables in the model. A large VIF indicates that a variable is essentially measuring the same effect as something else in the model. Including variable’s with high VIFs is generally not recommended, a rule of thumb of 5-10 is typically used for grounds to exclude a variable. Variable’s with VIF’s of 10 are often dropped as being redundant (note: this doesn’t apply to variables involved in quadratic or cubic terms, which should be related to things already in the model).

Using the description of the variables “yrs.service” and “yrs.since.phd”, explain the why VIF of “yrs.service” might be high.

### Question #12

Based upon modeling process you explored in Questions 6-11, decide upon a final model. Write the formula for your model in your lab write-up. Use this model to answer the question: “is there a difference in salary for male/female professors after adjusting for potential confounding variables?”

## PART #3 – Practicing on your own

### Question #13

For this part of the lab I’d like you to build a model for the outcome variable “Time” using the “Breast Cancer Data”. A description of the dataset can be found in the last set of course notes. Once you’ve decided upon a model, write its form (including the regression coefficients you’ve found) in your lab write-up. The groups with models that perform best on the data that I’ve withheld will receive a small amount of extra credit (added to their Exam 2 score).

How to include a quadratic term:

Navigate from STAT -> REGRESSION -> REGRESSION -> FIT REGRESSION MODEL

Click on the “Model…” button, see below for how to add the quadratic term:

