**Greg Dimock   
RTI Exercise**

**Methodology:**

I began with examining pure summary statistics of each variable. I then performed contingency table analysis for all categorical variables just to explore if associations existed, which all variables appeared to have with the target (over\_50k) to varying levels of strength. For each of these I also looked at the contingency tables to reveal potential issues with quasi-complete separation. I then began to look into possible confounding with the Cochran Mantel Haenszel test and interactions with the Breslow Day test, but had difficulty with coding in these aspects. This would be a future step I would take to go back and continue with exploratory analysis.

For the next steps, I then split the data into 70% training, 20% validation, and 10% test by creating a uniform random variable between 0-1. I did not oversample because the event (over\_50k = 1) occurred in about 24% of the cases. My next step was to check that the continuous variables were linearly related to the logit by looking into the Box-Tidwell transformation. Age, education (continuous), and hours worked each week all did not satisfy the assumption of linearity, but the required exponents to ensure linearity were close enough to 1 that I left all at 1 for interpretability.

With a manageable number of variables I started going through the model building process by manually adding and removing variables based on their significance level, while also being aware of likely causes of multicollinearity (such as education\_num and education\_level). Due to the large sample size, I set a more conservative p-value of 0.005. After performing model comparison with several different candidate models, I took the best models forward to the validation dataset. From here I checked how they compared to each other, and using ROC-Test comparisons, chose the model with the fewest variables without giving up any predictive strength. A future step I would like to take is actually calculate the misclassification rate for the validation (and testing) datasets for the candidate models rather than just comparing the areas under the ROC curves.

The final model had the following predictors: age, sex, occupation, education (continuous), marital status, hours of work a week, capital gains, and capital losses. An argument could be made to keep the capital gains and capital losses out of the model if demographic variables are more of interest for the basis of the model. Finally, odds ratios were calculated for the final model.

Plotted below is a grouped bar graph showing how much more likely those with civilian spouses are to make over $50,000 a year than any other marriage group. I am new to R graphics so given more time, I am very interested in learning how to make more defining graphics.

