Module 04 Week 7 Problem Set #6

I confess that I started writing this, worked on the assignment PS#6, started rewriting it, gather publications about P-values, started rewriting it again… This seems to be the way many people handle the challenges surrounding defining and using P-values. I’ve tried to edit this to be readable and useful for your assignment. I think you’ll agree when you’ve thought about this, there are a lot of challenges here!

You are more likely to conduct Monte Carlo simulations in industry than stepwise (forward selection or backward elimination) regression, in general. I had hoped to give you an assignment based on running Monte Carlo simulations involving stepwise-regression, but could not quite make that work out. However, because this is an important topic I’m leaving the rest of this original section for you to consider. The actual assignment and online questions will not involve Monte Carlo simulation, nor will the Final Examination.

To support running Monte Carlo simulations you are provided the original paper on Monte Carlo simulation by Nicholas Metropolis and Stanislaw Ulam, “The Monte Carlo Method,” published in the Journal of the American Statistical Association in 1949. (One of my fondest memories of college was when a professor would pull out a original paper on a topic and teach the topic from that original paper.) Ulam is widely recognized as having invented the Monte Carlo method among many other things. He was a nuclear physicist and mathematician. You are not expected to understand and will not be tested on the content of this paper. The many topics included in Ulam’s paper are extremely complex. It is worthwhile to read this paper just to look at the differences in how topics were presented in that age of discovery. That is, much of this resulted from the study of nuclear fission, particle physics, statistical mechanics and other very advanced topics in the 1920’s to 1940’s.

In addition to Metropolis and Ulam’s paper, you are provided with a paper by McIntyre, Montgomery, Srinivasan, and Weitz from 1983, i.e. “Evaluating the Statistical Significance of Models Developed by Stepwise Regression”. This paper does exactly what its title implies and discusses the implications of that work. Because it was written in 1983 there were still considerable difficulties in computational power to overcome. Remember that the Apple IIe was released for sales in January 1983. It had a 6502 chip which gave it about 1.023 MHz in processing speed and about RAM of 48 kilobytes depending on its configuration. If you haven’t seen one before, here is a picture.



Needless to say, running a Monte Carlo simulation on a regression model with 1,000 observations and up to 40 variables was a challenge.

You are also provided the paper “The Use of an F-Statistic in Stepwise Regression Procedures” by Pope and Webster from 1972. Although we have not covered the derivation of the normal equation(s) in this course, they form an integral part of this area of data analytics and are included in part in this paper in discussing the sum of squares. Notice in this paper that SSR stands for Sum of Squares Regression, NOT Sum of Squares Error and SSE stands for Sum of Squares Residuals, NOT Sum of Squares Explained. If you happen to take the Coursera course in Machine Learning by Dr. Andrew Ng, you will go through this derivation and compare its results to the output from using the algorithm for gradient descent. That is, you can solve for a deterministic solution to the regression problem using the normal equation(s) (where a set of normal equations represents a system of linear equations rather than a single equation), or for a computational solution using a variety of methods or algorithms of which gradient descent is one of the most popular.

Both of these papers are intended to just raise your awareness of some of the issues in using stepwise regression, either forward selection or backward elimination. I will say again, this is a grey area that is easy to get trapped in by using more subjective and less objective criteria in an analysis. However, this is true in a lot of areas not just regression. In fact, regression is not the “stand alone” topic we have treated it as. For example, regression underlies or provides the foundation for many machine learning algorithms. This is just to say that it is important to fundamentally understand regression to become an expert at many other things. Also with regard to grey areas, I had originally wanted to walk you through reducing dimensionality of datasets as part of the Residency. I was vetoed on that topic. The topic includes feature selection, principal component analysis or PCA, linear discriminant analysis or LDA, and latent factor analysis. Latent factor analysis is notorious for being fraught with problems analogous to those commented on in stepwise regression. Latent factor analysis is a technique used extensively in marketing and sociology and relies heavily on subject matter experts input, which is undoubtedly objective.

# The Layoffs dataset

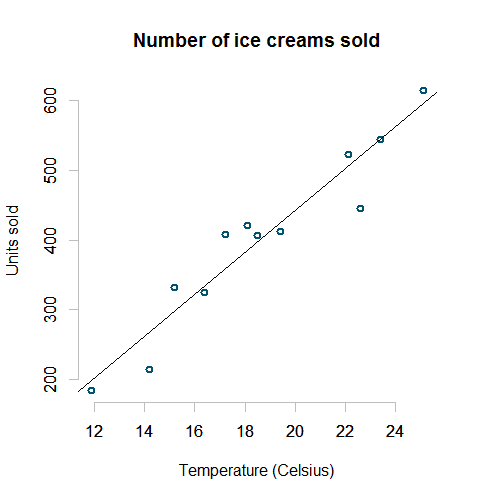
The layoffs dataset has 50 observations and seven independent variables (age, educ, married, head, tenure, manager, and sales). The dependent variable is “weeks,” i.e. the number of weeks a person has been jobless. Table 1 of McIntyre et al, lists the combinations of observations and predictors (independent variables) used in the simulations they ran in 1983. ~~Originally, I was concerned that we might have to develop some additional (artificial) data to do a good Monte Carlo simulation.~~ However, the layoffs dataset seems to fit in the middle of the simulations run by McIntyre et al. It is worthwhile to ask, “If you needed additional data to supplement the existing observations in the layoffs dataset, how would you obtain that (artificial) data”? I’ll let you ponder that for a bit while we finish laying out what is required for Problem Set #6.

# General Questions

Brereton in his article “P values and Ronald Fisher” tells the story of the tea tasting experiment that many of us learned when we first studied P-values (Brereton, 2020). I’m attaching that article for your files. On the other hand, other researchers report that the interpretation of Fisher’s “exact test” which included computation of P-values was less exact. In fact, the original interpretation seems to be that the result was merely sufficient to say an additional experiment was warranted in order to be conclusive.

The P-value has a long and sordid history with many misconceptions and misinterpretations to the point that many contemporary researchers have suggested not computing P-values at all and one peer-reviewed journal stating it would no longer accept articles where conclusions were based on P-values. This is some pretty serious stuff. Note that we will not cover the parts of these articles on Bayesian methods since we do not cover Bayes in this course. For now, let’s see if we can get to some of the misconceptions and misperceptions to give you better intuition about P-values.

1. Some researchers only recognize two possible outcomes from null hypothesis significance tests, i.e. the null hypothesis is rejected or we fail to reject the full hypothesis. However, other researchers recognize four possible outcomes from null hypothesis significance tests, i.e. select the choices below that best represent the four possible outcomes these other researchers recognize.
   1. Reject the null hypothesis
   2. Fail to reject the null hypothesis
   3. Accept the alternative hypothesis
   4. Commit a Type I error
   5. Commit a Type II error
   6. Rank the P-value to determine a result
   7. Use the P-value to determine a result based on a specified level of significance
   8. There are only two, not four possible outcomes to null hypothesis significance tests
2. There is a small dataset about ice cream sales on Mages’ blog at <https://www.magesblog.com/post/2015-08-18-visualising-theoretical-distributions/>. If you plot the basic data you’ll find something that looks like this:



I introduce this plot in ANA 500 and ask students if it looks like these data are linearly related. It really does look like it, but… This involves extrapolating from the lowest temperature data to where the trend line would cross the y-axis at x = 0 Celsius. At that point you can see that a linear relationship no longer makes any sense. That would require the ice cream vendor accepting an unrealistic number of units of ice cream returned by customers. That is, the number of units sold goes way negative. It isn’t until you get to a log-transformation or Poisson transformation (as shown at the Mages’ webpage) that you have a realistic curve for the data. However, if you do all the “correct” statistics for the linear relationship you’ll get an OLS regression that is statistically significant based on its P-value.

Misconception #3 from Goodman’s paper states, “A statistically significant finding is clinically important” (Goodman, 2020). This means that (select the best choice below):

* 1. You should not use only a statistical analysis to interpret results of a study. You should also include additional relevant information.
  2. It is sufficient to complete a statistical analysis and interpret the results of a study on that analysis.
  3. There is no way to integrate additional information into the results of a study that uses a statistical analysis.
  4. If a statistical analysis of data obtained from a study is completed then additional data or information should NOT be added to this analysis.

1. In Verdam et al’s paper, the authors state that, “The p value is not the probability of the null hypothesis” (Verdam et al, 2013). The authors go on to explain this statement by the probability is not the same as and further elaborate this with a great example of this situation. However, from our study of probability in 500 and review in 510 we know that there is a situation in which . Situations like this can lead to some of the misconceptions discussed. Select the choice below that best describes when this is true. (Hint, do not think about Verdam et al’s example. Only think about our study of probability to answer this question.)
   1. and are independent
   2. and are linearly related
   3. and are dependently related
   4. and are conditionally dependent

# The Problems

There are three problems in this problem set. For simplicity I’ll just call them Problem #1, Problem #2, and Problem #3. Use the results of the analyses you complete to answer the following questions.

# Problem #1

1. Determine the best single (independent) variable regression model for the Layoffs dataset. The computed slope of the “best” independent variable is \_\_\_\_\_\_\_\_. Don’t forget to round to two decimal places!
2. The way to determine which independent variable is “best” is by the (select the best choice below):
   1. P-value
   2. R-squared
   3. coefficient
   4. standard error

# Problem #2

Using the forward-selection procedure, determine the best regression model for the Layoffs dataset. Use your results to answer the following questions.

1. Which of the independent variables listed below is the first choice of independent variables?
   1. Age
   2. Manager
   3. Head
   4. Sales
   5. Educ
   6. Married
2. How did you determine which of the independent variables is the first choice?
   1. P-value
   2. Coefficient
   3. R-squared
   4. Standard error
3. What is the computed coefficient of the first choice rounded to two decimal places? 1.51
4. Which of the independent variables listed below is the second choice of independent variables?
   1. Age
   2. Manager
   3. Head
   4. Sales
   5. Educ
   6. Married
5. What is the computed coefficient of the second choice rounded to two decimal places?
6. Select the best choice below that describes how to stop the forward-selection process.
   1. None of the remaining independent variables are statistically significant at the 0.05 level
   2. None of the remaining independent variables have low standard errors
   3. None of the coefficients of the remaining independent variables are positive
   4. You just really pick an independent variable and continue until you believe you have captured the regression equation
7. The computed coefficient for the first choice of independent variables in the final forward-selection model is \_\_\_\_\_\_\_.
8. The computed coefficient for the second choice of independent variables in the final forward-selection model is \_\_\_\_\_\_\_.
9. How many of the original independent variables were selected for inclusion in the final forward-selection model for the Layoffs dataset?

# Problem #3

Using the backward-elimination procedure, determine the best regression model for the Layoffs dataset. Use your results to answer the following questions.

1. Which of the independent variables listed below is the first of the independent variables to be dropped in the backward-elimination procedure?
   1. Age
   2. Manager
   3. Head
   4. Sales
   5. Educ
   6. Married
2. How did you determine which of the independent variables to eliminate first?
   1. P-value
   2. Coefficient
   3. R-squared
   4. Standard error
3. What is the computed coefficient of the first eliminated independent variable rounded to two decimal places?
4. The computed coefficient for the second independent variable eliminated in the final backward-elimination model is \_\_\_\_\_\_\_.
5. How many of the original independent variables were selected for inclusion in the final forward-selection model for the Layoffs dataset?
6. Select the best choice below that describes how to proceed with the backward-elimination process.
   1. Run an OLS regression using all available independent variables, drop the independent variable with the highest P-value, repeat.
   2. Run an OLS regression with one independent variable at a time, drop the independent variable with the highest P-value, repeat.
   3. Run an OLS regression with all independent variables available, drop the independent variables with the 3 highest standard errors, repeat.
   4. You just really pick an independent variable, run an OLS to see if the coefficient of the variable is statistically significant and continue until you believe you have captured the regression equation.