Running head: WRITTEN ASSIGNMENT 2: CHAPTER 2

Written Assignment 2

Literature Review

The reliability of regression as a forecasting tool

College ID #:0249501

(804) 387-7315

Liberal Arts Capstone (LIB-495-OL013)

Thomas Edison State College

October, 2015

**Part 1: Literature Review**

**Introductory paragraph**

Regression analysis is a statistical process of estimating relationships among variables. It includes many techniques for the model and analyzing several variables when the focus is mainly to check the relationship between a dependent variable and independent variables. Regression assists in comprehending how the typical value of the dependent variable changes when any of the independent variables varies while the other independent variables remain fixed. Analysts widely use regression for prediction and forecasting where there its use is a substantial overlap with the field of machine learning. The model, therefore, is used to understand the-the relationship among the independent variables and the dependent variables. There are few and very restricted circumstances where regression is used to infer the causal relationship between independent and dependent variables. The familiar methods are linear regression and ordinary least squares.

**History of regression**

One of the early forms of regression was the method of least squares which was published by Legendre and Gauss at the beginning of the 1800s. Gauss used the method to the problem of determining from astronomical observations the orbits of bodies about the sun. Regression was used by Francis Galton in the 19th century to describe the biological phenomenon that is, that the heights of descendants of tall ancestors seemed to regress towards a healthy average. Regression analysis was publicized in the 1870s with the pioneering work of Francis Galton. However, least squares can be traced back to the early 1800s and Karl Gauss, who used the work to predict astronomical phenomena (Armstrong J). Simultaneous equations provide forecasts that were more accurate than those from those from simple regression models.

Regression was once viewed as a complex and cumbersome model, especially during the days of Friedman. Friedman. It has, however, evolved into a simple process that people can do in seconds. To understand the great strides that had been made one has to go back to the days when Friedman took 40 hours using a computer to calculate regression estimates and test statistics. Friedman also conducted tests using alloys that he believed would withstand extremely high temperatures. The first alloy, however, broke after 2 hours while the second did the same after 3 hours. Friedman concluded that it was prudent to focus on test outputs instead of statistically significant inputs. He noted that the more complex the regression, the more skeptical he was. (Friedman & Schwartz 1991). The simplicity is due to the technological revolution.

Regression analysis is arguably the most efficient model of analysis. It provides very useful forecasts and accurate predictions when dealing with valid data, small data, and large amounts of variables and even when using causal relationships (Allen & Fildes 2001). There are, however, illusions that many analysts continue to cling to which only serve to undermine its accuracy.

*Regression provides the best linear unbiased estimates*. Statisticians have wasted no time in pursuing their desire to prove that regression leads to the best estimates of relationships

Studies have, however, proved that regression estimates produce forecasts that are often less accurate than forecasts from ‘unit weights’ models. Schmidt (1971) became one of the first people to test the idea. He discovered that ‘unit weights’ were superior to regression weights. There are conditions under which regression is not effective, about equal weights (Einhorn & Hogarth, 1975).

A good characteristic of regression estimates is that they become more conservative as uncertain increases. Unfortunately, this uncertainty is often ignored. For instance, the coefficients can ‘get credit’ for important excluded variables that happen to be correlated with predictor variables. There is also evidence pointing out that regressions over-estimate change (Iannidis 2005, 2008). To solve the problem, one needs to combine forecasts and to include a naïve model among the alternative models. For time-series problems, it is prudent to dump the forecast more toward the naïve model. The action reflects the uncertainty that increases in the future.

*Regression models are sufficient*. Analysts derive forecasts from what is believed to be the best. The principle of forecasting is that the no-change model is always accurate. The principle often assists in shrinking each coefficient toward having no effect. The regression model is not sufficient, and this, therefore, implies the need for combined forecasting. One can find two or more valid forecasting models and then calculate the averages. This method is most effective where there is a substantial difference in the methods, models and data. It reduces the error in the analysis.

*Complexity illusion*. Most analysts believe that complexity is synonymous to accuracy. That the more complex the model is, the more accurate, it is. Evidence from studies indicates that simplicity is more favorable as compared to complexity. However, despite the overwhelming evidence towards simplicity, many analysts continue to cling on complexity. Such fact begs the question why is there an in their a persistent urge for complexity. Hogarth (2012) explains that the urge to remain complex is perhaps due to the academics’ preference for complex solutions. Christ (1960) was able to discover that simultaneous equations when used, resulted in forecasts that were more accurate than forecasts from simple regression models when tested with artificial data. The situation was different when tested using real data as it was less accurate. Furthermore, the aim of statistics is to communicate and they should therefore as simply as possible. However, the continued complexity acts as a barrier to the much-desired communication.

*That fit implies accuracy*. Analysts have often made an assumption that models with better fits provide more accurate forecasts than those with inferior fits. The assumption is in total disregard of research which indicates that fit bears little relationship to forecast accuracy, especially time-series. Zellner (2001) concluded that fit improves as complexity increases while accuracy decreases. The continued belief in such illusions by even some of the most learned statisticians is a basis for inaccuracies in regression.

*The control*. Users of regression often assume that by putting variables into the equation they are in control of the variable. The control, however, can only be exercised by using experimental and then adjusting the dependent variable. When using real data, the users have no control of the variables and should cease assuming that they have the control.

**Using regression for decision–making**

Regression analysis is an objective to analyze data. Decisions based on regression are less likely to be subject to bias. They are consistent and are easy to explain unlike judgemental decisions based on the same data. They are not arbitrary but well documented (Grove & Meehl 1996; Armstrong 2001). This makes them very useful when it comes to making decisions. The accuracy of regression in decision-making is, however, reduced by two illusions, that is statistical significance and correlations.

*Statistical significance illusion*. This is based on the belief that statistical significance improves decision-making. Though it is not based on any studies, many analysts continue to hold onto the belief. On the contrary, there are numerous cases where statistical significance harmed decision-making. Schmidt (1996) challenged the analysts to articulate any contribution that significance testing has brought about development of cumulative scientific knowledge.

*Correlation illusion*. Many researchers often make a grievous mistake of taking correlation to mean causation. The fact that one variable correlates to another does not automatically signify that it is the cause of the variable. A comparison of findings from experimental of findings experimental with those from analyzes of non-experimental data showed a huge difference (Armstrong & Patnaik 2009). This leads to the conclusion that analysis from non-experimental is at times misleading. The illusion has misled many people into making wrong decisions about medication, food among other areas of life. The misleading decisions often end up as costly

This is, however, not surprising since Sir Harold Jeffreys had warned of this illusion (Zellner 2001). He even went to refer to it as the ‘most fundamental fallacy of all’.

**Conclusion**

From the study of Soyer and Hogarth, echoing Freedman, emphasizes that scientific theories should be tested for their predictive ability. It would be helpful if software packages would focus on testing by making it easy to simulate the forecasting situation. Software packages should allow for meaningful comparison of methods (Armstrong & Collopy 1992).

Regression analysis is clearly one of the most important tools for researchers. However, it is not the only avenue. Researchers made scientific discoveries about casualty prior to the availability of regression analysis as shown by Friedman (1991) in his paper aptly titled ‘Statistical models and shoe leather'. He demonstrates how major gains were made in epidemiology in the 1800s. For instance, John Snow's discovery of cholera in London in the 1850s came about due to the clarity of prior reasoning, the bringing together of many different lines of evidence and his determination to get all the available data. The characteristics portrayed by John Snow are a reflection of real science as described by Freedman.

**Part 2: Annotated Bibliography**

**References**

Allen, P. G. & Fildes, R. (2001). *Econometric forecasting*. In J. S. Armstrong. Principles of Forecasting, Boston: Kluwer Academic Publishers

Armstrong, J. S. (2001). *Judgmental bootstrapping: Inferring experts’ rules for forecasting*.” In J. S. Armstrong.

Armstrong, J. S. & Collopy, F. (1992*). Error measures for generalizing about forecasting methods*: Empirical comparisons, International Journal of Forecasting, 8, 69-80.

Armstrong, J. S. & Patnaik, S. (2009). Using Quasi-experimental data to develop principles for persuasive 9 advertising. Journal of Advertising Research, 49, No. 2, 170-175.

Christ, C. F. (1960). Simultaneous equation estimation: Any verdict yet? Econometrica, 28, 835-845 Ecological rationality: Intelligence in the world (in press). Oxford: Oxford University Press

Einhorn, H. J. & R. Hogarth (1975). *Unit weighting schemes for decision making*. Organizational Behavior and Human Performance, 13, 171-192.

Friedman, M.A. & Schwartz, A. J. (1991). *Alternative approaches to analyzing economic data*. American EconomicReview, 81, Appendix 48-49.

Hogarth, R. M. (2012). When simple is hard to accept. In P. M. Todd & Gigerenzer, G. (Eds.),

Ioannidis, J. P. A. (2005). *Why most published research findings are false*. PLoS Medicine, 2,

696-701.

Ioannidis, J. P. A. (2008). *Why most discovered true associations are inflated.* Epidemiology, 19, 640-6

Grove, W.M. & Meehl, P.E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: the clinical – statistical controversy. Psychology, Public Policy, and Law, 2, 293-323.

Schmidt, F. L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. Educational and Psychological Measurement, 31, 699-714

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods, 1, 115-129.

Zellner, A. (2001). Keep it sophisticatedly simple. In Keuzenkamp, H. & McAleer, M. Eds. Simplicity, Inference, and Modelling: Keeping it Sophisticatedly Simple. Cambridge University Press, Cambridge