INTERPRETATION OF REGRESSION OUTPUT: DIAGNOSTICS, GRAPHS AND THE BOTTOM LINE

Wesley Johnson and Mitchell Watnik
University of California
USA

A standard approach in presenting the results of a statistical analysis of regression data in scientific journals is to focus on the question of statistical significance of regression coefficients. The reporting of p-values in conjunction with a description of the various positive and negative associations between the response and the factors in question ensues. The real question of interest beyond these initial assessments ought to be, "how well does the treatment work?" The point of view taken here will be that this standard presentation, while important, constitutes only a first order approximation to a complete analysis, and that the bottom line ought to involve the quantification of regression effects on the scale of observable quantities. This will mainly be accomplished graphically. It is also emphasized that diagnostic assessment of the compatibility of the data to the model should be based on similar considerations.

INTRODUCTION

Emphasis on the estimation of unobservable quantities such as parameters in a regression model is ubiquitous. Scientific literature is replete with presentations and conclusions that focus almost entirely upon statistical results that indicate that a treatment is either effective or not at the .05 or .01 levels, but with no attempt whatsoever to try to quantify the magnitude of the effect or moreover to indicate whether the effect is of any practical import. The issue of sample size and its relevance to either statistically significant or insignificant results is generally missing beyond the requisite sample size calculations that are performed in the process of writing NIH (or similar) grant proposals. We wonder where the fault for this might lie. Perhaps it is a failure at the level of instruction in basic statistics courses. Are we, as statistics instructors, giving sufficient emphasis to the distinction between statistical significance and practical import in our presentations of data analysis? We think not.

Standard statistical textbooks give careful and methodical presentations of didactic material in step by step approaches to building various models and to the implementation of these models using modern computational methods. There is generally heavy emphasis on the estimation of parameters and on the formal method of constructing confidence intervals and making statistical tests of hypotheses. While most of this is laudable, what is missing from most presentations is a discussion of the real scientific problem of interest and its importance, and discussion of how the statistical analysis that has been performed might shed light on this scientific problem. Instead of statistical methods being developed and presented in a way to serve science, it seems to be too often the case that it is the development of the statistical methods by statisticians that tends to determine how scientific results are presented.

This fundamental disconnect is perhaps understandable, since most statisticians were probably not trained in substantive areas of science beyond Statistics or perhaps Mathematics, including the authors of this article. However, it is a fact that scientists depend on their statistics instructors and statistics books to guide them and thus are often led to believe that the statistician's presentation of data analysis is all there is and could or should be. A fundamental question that we would ask is, why have scientists seem to be so willing to accept this state of affairs? Why have they not demanded more of their statistical partners in the scientific endeavour?

In the remaining sections of this article, we will attempt to illustrate part what we believe is missing from the presentation of statistical methods in standard regression problems. The foundational basis for what we are about to present can be found in the collective writings of Seymour Geisser, much of which can be found in Geisser (1993). Articles by Bedrick, Christensen and Johnson (1997, 2000) illustrate much of what we discuss below, and excerpts will be taken from them. We would like to be clear before proceeding that we make no claims to

originality. Our main purpose is to emphasize the need for greater emphasis on the quantification of ultimate conclusions based on regression output.

GUIDING PRINCIPLES

Let's consider a generic regression problem. It could be a simple linear or non-linear regression problem, regression in the context of a generalized linear model, or regression in survival analysis, just to name a few. The observed responses are labelled as Y, and the corresponding vector of covariates is labelled as x where the first component of x is a one and corresponds to the intercept term. The model involves a regression coefficients vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$, and the mean or median of the response Y is a function of the form $g(x\beta)$. For the standard non-linear model, $Y = g(x\beta) + e$ and the mean and median are usually presumed to be $g(x\beta)$, since the error distribution is generally taken to be symmetric about zero. In the case of a linear model, $g(x\beta)= x\beta$. When considering a log-normal regression model in survival analysis with T denoting the time to failure, we have $Y= \ln(T)= x\beta +e$ and the natural quantity of interest is the median of T, which is $g(x\beta) = \exp(x\beta)$. In a logistic regression model, Y is a Bernoulli random variable with mean $g(x\beta)$ satisfying $\text{logit}(g(x\beta))= x\beta$.

Standard statistical inferences in this context revolve around testing that various β_i `s are zero. The full extent of interpretive efforts for the linear model generally involves the simple assertion that a unit increase in the corresponding x_i will result in an increase in mean response of β_i , other variables being fixed. In the logistic regression context, β_i is interpreted as the effect of a unit increase in x_i on the log odds of success, or $\exp(\beta_i)$ is interpreted as the effect of increasing x_i by one unit on the odds of success. While this can be effective for some problems and variables, its value is limited in our estimation.

At the very least, it is sensible to select a range of values for x_i that has some practical meaning, and to interpret the difference in mean or median response over that range of values rather than for a single unit. For example, if we were regressing college GPA on SAT score, and the resulting estimate of slope were .005, then an increase in SAT score of one unit will result in an increased expected college GPA of .005, while an increase in SAT score of 100 results in an increase of half a grade point. Obviously one can do this arithmetic in their head, but why not simply make it explicit for the reader, particularly since SAT scores are reported in increments of 10 points? This would further convey the impression that the writer was interested to determine whether *meaningful*, rather than *significant*, differences had been obtained. For example, if the above estimated regression coefficient were .0005 and highly statistically significant, the practical import of the result becomes evident when it is realized that an increase of 100 points on the SAT only results in an increase in GPA of .05.

It is our fervent belief that the scale of interest for making statistical inferences is $g(x\beta)$, or perhaps some relevant function of it, when this is indeed possible. We believe the ultimate inference should revolve around a discussion of how $g(x\beta)$ varies as x varies over the intersection of the range of plausible values of x and the space spanned by the covariate vectors seen in the data. This can be accomplished by either presenting a table of median or mean values over a range of possible values for the collection of covariates. For example, if there are k covariates, then for each variate, two "representative" values can be selected, and a 2^k table of estimated mean or median values (and their corresponding confidence intervals), can be presented. In some instances, when interest focuses on a treatment effect for example, a corresponding table of relative medians comparing the medians under treatment and control can be obtained. Even better, a corresponding template of plots of mean or median response against an important continuous variate for say treatment and control on the same plot can be presented. The same plot is calculated under the 2^{k-2} possible varying circumstances of the other variates. If some of the variates are of lesser interest, then a smaller dimensional array can be obtained with average or typical values being used for the less interesting variates.

The issue of assessing the influence of deleting observations that have been identified as potentially influential by some criterion also deserves mention in this context. Suppose a Cook's distance measure has determined that a particular case may be influential. Suppose it has been determined that the identified case was not an error and thus at the outset there is no obvious

justification for simply deleting it. How then do we assess the actual impact of deleting this case? Obviously, the case should be deleted and the ultimate inferences of interest should be obtained and compared with those based on the full data set. So in the context of this section, we recommend obtaining the same tables of estimated median values and the same templates of curves, and that these be compared with those obtained from the full data set. If the curves and estimated medians are very similar from both analyses, there is nothing obvious to worry about. If on the other hand estimated medians change on the same order of magnitude as they change by altering the important covariates, the identified case is obviously very important and requires further study and discussion.

ILLUSTRATIONS

To illustrate the above ideas, suppose that a cross-sectional study has been performed to assess the effect of smoking, adjusted for age (women between ages 20 and 40 were selected), on the length of menstrual cycle. It is natural to fit a standard linear regression model on the logs of cycle lengths. Suppose the log transformation was indeed appropriate and standard diagnostic plots indicated that variances were reasonably homogeneous and that normality was not unreasonable.

As a first step in the analysis, the standard test for smoking effect and age effect are performed (the latter may not be performed by some since it not really necessary). Suppose that the standard tests indicate that both smoking and age are statistically significant at the .01 level of significance based on what a reasonably large sample size. It is then reported that indeed smokers have significantly longer cycles than non-smokers. A scientist with the appropriate biological background can then speculate about the implications of this in terms of women's health. However, we would argue that without taking the next step to quantify the additional length of cycle for smokers that this would be difficult to do with any accuracy. Assuming a model with no interaction between smoking and age, the effect on the log scale is β_1 and this has been estimated. However, it is unreasonable to expect the scientist to make the conversion in their heads so we would argue that the effect should be presented on the original scale. Since the median cycle length for a woman with covariate x is $\exp(x\beta)$, the relative median for a smoker compared with a non-smoker is $\exp(\beta_1)$, and inferences for this quantity are straight forward. If there were an interaction with age, then the relative median would be $\exp(\beta_1 + \beta_3 \text{ AGE})$, where β_3 is the coefficient corresponding to the cross-product term in the model including the interaction. A plot of the estimate of this function over the range of ages from 20 to 40 years would be appropriate and would convey the best picture possible of how the effect of smoking is moderated by age over the range of ages in the data.

If one were even more interested in a comparison of the actual lengths of cycles for smokers and non-smokers as they varied by age, one could plot the estimated median cycle lengths as a function of age for smokers and for non-smokers on the same plot. In this way, the effect of smoking on estimated cycle length can be pictured and quantified for each age. For example, suppose that the estimated median cycle lengths for a smoker and a non-smoker of age 20, respectively, were 35 and 28 days, and that the same quantities for a 40 year old were 40 and 28 days. We believe that this information, in conjunction with the knowledge that the smoking and age effects were statistically significant, would be of much greater use to the medical researcher than simply knowing that there were positive effects in conjunction with the actual estimated effect on the log scale, as might be traditionally reported.

In non-linear regression models, it seems even more evident that the above approach is necessary. Consider the O-ring data presented in Lavine (1991), where the effect of temperature on O-ring failure in space shuttles is discussed. The goal is to model the probability of at least one failure (out of a possible of six) as a function of temperature at launch. It was determined after the failure of the Challenger launch that cold temperatures greatly increased the probability of O-ring failure and this was ultimately determined to be the cause of the Challenger's failure. This is an example where it is clear, in retrospect, that the ultimate goal is to calculate the probability of at least one O-ring failure for the temperature at the time of launch. The temperature at the time of the Challenger launch was 31 degrees and the modelled probability of failure was nearly one

based on the data available at that time. If these data had been analysed prior to the launch, the conclusion that temperature was negatively associated with probability of failure by itself would have been terribly incomplete.

Now consider the problem discussed by Bedrick et al (1997) where they discuss the relationship between death at the time of trauma surgery and various risk factors that are assessed at the time on entry into the emergency room, including a continuous injury severity score (ISS), a measure how ``sick'' they were, called the revised trauma score (RTS), an indicator of whether the injury was blunt or penetrating (TYPE), and their age (AGE). All of these factors turn out to be statistically important. Bedrick et al present a series of four plots of the estimated probability of death versus ISS, with the plots for blunt and penetrating injuries on the same plot. The plots are arranged in 2 by 2 table format with left to right corresponding to a change in age from 10 to 60 years of age, and with a change in top to bottom corresponding to a change in RTS from a relatively ``sick'' individual with an RTS of 3.34 to an RTS of 5.74 for a relatively ``healthy'' individual. It is possible to visualize the effect of blunt versus penetrating by simply looking at each plot. The interaction between this TYPE and AGE is viewed by comparing how the difference in blunt versus penetrating curves changes as we move from the frame for ten year olds to the frame for 60 year olds, holding the RTS value fixed, etc. For this particular problem, the curves comparing blunt versus penetrating injury types are virtually identical for the 60 year old, regardless of the value of RTS, while they are noticeably different from one another when considering 10 year olds. For ten year olds, penetrating wounds generally result in between a 10 and 20\% higher estimated probability of dying than those with blunt wounds. The sicker individuals have roughly double the probability of dying than do the ones that are more healthy. For example, a 50 year old with a blunt wound and RTS of 5.74 is estimated to have a .2 chance of dying, while the comparable individual with an RTS of 3.74 has probability .45 of dying. The issue of practical import of these statistically significant results is thus immediate by viewing these plots. Moreover, the need for a sample size calculation seems irrelevant here since the non-significant variates showed absolutely no effect that would be of practical import according to these graphs, and the variates that were statistically significant indeed have a very large practical import.

The bottom line for this analysis is that the chances of surviving trauma surgery are quite variable. The quantification of this variability that is reflected by varying the various risk factor combinations results in a useful tool for assessing the true severity of any given situation. While the surgeons who are actually performing the surgery can certainly intuit and project based on their experience, this kind of quantification serves the purpose of actually measuring the relative seriousness of the situation. With this knowledge, surgeons and emergency room personnel in general will hopefully be better able to administer care to patients if only by perhaps by making it easier to know how to prioritise the treatment of patients during those times when there are more patients than surgeons available.

For these data, a case deletion analysis was also performed. A particular individual was identified by Cook's distance as being influential compared with the rest of the data. That particular individual was relatively old, but relatively healthy and not too severely injured, yet they died. However, removal of this case from the data resulted in virtually the same inferences both quantitative and qualitative, so there was no need to pursue it further.

Another example is provided by Bedrick et al (2000). In this article, the authors present a survival analysis of data corresponding to the time of natural abortion in dairy cattle. The main risk factor of interest is the infection status of the cows; it is believed that *neospora caninum* infection (INF) is a causative agent for abortion and it is of interest to determine, for those cows actually aborting, whether they abort sooner or later as a result of the infection. Other confounding variables of interest are the number of days open (DO) between the last calving and the current pregnancy, the ease/difficulty of the previous calving (CE), and whether or not prostaglandin (PR) had been administered. The ultimate goal of the experiment is to quantify the effects of these factors on the timing of abortions and to use this information to determine whether or not to modify dairy management practices to minimize fetal wastage.

All the described factors, except PR, are statistically significant, including a DO by CE interaction. The estimated median time to abortion among infected cows is roughly a month later

than among non-infected cows. While DO has no discernible effect when the previous calving was difficult, the estimated median time to abortion for cows with an easy previous calving is about a month later for those with 135 versus only 45 days open. These effects have practical import given that abortions can only occur over a 250 day period. Bedrick et al considered a two by two template of plots with two curves on each corresponding to infected and non-infected cows, and where calving ease is varied in plots going left to right and days open is varied from top to bottom. The identical template was plotted for PR, yes and no, but the plots were virtually identical. So we are able to discern that the lack of statistical significance for PR corresponds to a complete lack of practical import as well. The bottom line for this analysis is that the administration of PR seems to have no statistical or practical import, that with a difficult previous calving, DO is a practically unimportant variable, while it is practically important when the previous calving was easy, and that infection delays abortions noticeably, regardless of the other circumstances.

The bottom-bottom line is now for the scientist and dairy manager to determine how to use this information in order to reduce the impact of abortion on the cost of running this kind of business. Clearly, there is no reason to modify the administration of PR. Since we are not experts, we can only speculate about further utility of this knowledge. Perhaps the cows can be categorized according to their specific covariate information and then watched more carefully or even treated during periods (early versus late) when they might be expected to abort. Nonetheless, we believe we have provided the necessary information for the scientist/dairy manager to mitigate dairy protocols. A case deletion analysis was similarly performed here. Deletion of the most influential cases had little impact on the above inferences.

CONCLUSIONS

We would like to emphasize our belief that the purpose of statistical endeavours is to serve scientific endeavours. That is not to imply that statisticians have a lesser role to play in science, but that the primary reason that Statistics exists as a discipline is because statistical problems are created naturally in the pursuit if scientific knowledge. We believe that the analysis of data should reflect this fact and have attempted to illustrate how the ultimate conclusions of a data analysis might better serve the scientific intent of the exercise.

REFERENCES

Bedrick, E.J., Christensen, R., &. Johnson, W.O (1997). Bayesian Binomial Regression: Predicting Survival at a Trauma Center. *The American Statistician, 51*, 211-218.

Bedrick, E.J., Christensen, R., &. Johnson, W.O (2000). Bayesian accelerated failure time analysis with application to veterinary epidemiology. *Statistics in Medicine*, *19*, 221-237.

Geisser, S. (1993). *Predictive Inference: An Introduction*. New York: Chapman & Hall.

Lavine, M. (1991). Problems in extrapolation illustrated with space shuttle o-ring data. *Journal of the American Statistical Association*, *86,* 919-921.