


3-6-2019

Logistic Regression: An Inferential Method for Identifying the Best Predictors

Rand Wilcox

University of Southern California, rwilcox@usc.edu

Follow this and additional works at: <https://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

Recommended Citation

Wilcox, R. (2018). Logistic regression: An inferential method for identifying the best predictors. *Journal of Modern Applied Statistical Methods*, 17(2), eP3061. doi: [10.22237/jmasm/1551906905](https://doi.org/10.22237/jmasm/1551906905)

This Invited Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

INVITED ARTICLE

Logistic Regression: An Inferential Method for Identifying the Best Predictors

Rand Wilcox

University of Southern California
Los Angeles, CA

When dealing with a logistic regression model, there is a simple method for estimating the strength of the association between the j^{th} covariate and the dependent variable when all covariates are entered into the model. There is the issue of determining whether the j^{th} independent variable has a stronger or weaker association than the k^{th} independent variable. This note describes a method for dealing with this issue that was found to perform reasonably well in simulations.

Keywords: Regression, binary data, strength of association

Introduction

Let Y denote some outcome variable of interest and let X_1, \dots, X_p denote p covariates. A basic issue is determining whether X_j is a better predictor of the typical value of Y than X_k , $1 \leq j < k \leq p$. In the regression literature, there are numerous methods for estimating the relative importance of the p covariates, which were reviewed by Wilcox (2018). A simple strategy is to compare the correlation of X_j with Y to the correlation between X_k and Y . However, a concern with this strategy is that the strength of the associations can depend on the covariates included in the model. Methods for dealing with this issue have been developed, but generally they do not indicate the strength of the empirical evidence that X_j , say, is a better predictor than X_k when all p covariates are included in a linear regression model. Tibshirani, Taylor, Lockhart, and Tibshirani (2016) as well as Lee, Sun, Sun, and Taylor (2016)

doi: 10.22237/jmasm/1551906905 | Accepted: October 21, 2018; Published: March 6, 2019.

Correspondence: Rand Wilcox, rwilcox@usc.edu

Rand Wilcox is a Professor in the Department of Psychology, University of Southern California. His primary interests are robust and nonparametric statistical methods.

RAND WILCOX

derived methods aimed at dealing with this latter issue assuming normality and homoscedasticity. Wilcox (2018) derived a robust method that allows heteroscedasticity.

The focus here is on the situation where Y is binary and the conditional probability of $Y = 1$ is given by the logistic regression model. That is, it is assumed that

$$P(Y = 1 | X_1, \dots, X_p) = \frac{A}{A+1}, \quad (1)$$

where $A = \exp(\beta_0 + \sum \beta_j X_j)$ for unknown constants β_0, \dots, β_p . An obvious speculation is that a simple modification of the method in Wilcox (2018) can be used for the situation at hand. However, simulations revealed that this is not the case; it performs poorly. This note describes two modifications of Wilcox's method aimed at dealing with this issue.

The Proposed Method

Momentarily consider the case of a single explanatory variable and let \hat{Y} be some estimate of the typical value of Y given X . Explanatory power (e.g., Wilcox, 2017) is

$$\eta^2 = \frac{\tau^2(\hat{Y})}{\tau^2(Y)}, \quad (2)$$

where τ^2 is some measure of variation; η is known as the explanatory measure of the strength of the association. When the ordinary least squares estimator is used and τ^2 is taken to be the variance, η^2 is the coefficient of determination. In particular, when $p = 1$, $\eta^2 = \rho^2$, where ρ is Pearson's correlation.

Note that for two independent variables, X_j and X_k , determining which is more important can be approached by testing

$$H_0 : \eta_j = \eta_k. \quad (3)$$

In terms of Tukey's three-decision rule (e.g., Wilcox, 2017), if this null hypothesis is rejected, make a decision about whether η_j is greater than or less than η_k . Otherwise, no decision is made.

LOGISTIC REGRESSION

Given X_{1j}, \dots, X_{nj} , let $A_{ij} = \exp(\beta_0 + \beta_j X_{ij})$. Then, based on the logistic regression model, $A_{ij}/(A_{ij} + 1)$ corresponds to $\pi_{ij} = P(Y = 1 | X_{ij})$. Let σ_j^2 denote the population variance associated with $\pi_{1j}, \dots, \pi_{nj}$. Note that for the situation at hand, to test (3) it suffices to test

$$H_0 : \sigma_j = \sigma_k. \quad (4)$$

Let b_0, \dots, b_p be estimates of β_0, \dots, β_p , respectively, based on the random sample $(Y_i, X_{i1}, \dots, X_{ip})$, $i = 1, \dots, n$. Let

$$\hat{\pi}_{ij} = \frac{\hat{A}_{ij}}{\hat{A}_{ij} + 1} \quad (5)$$

be the estimate of π_{ij} where $\hat{A}_{ij} = \exp(b_0 + b_j X_{ij})$. Then σ_j^2 is estimated with

$$\hat{\sigma}_j^2 = \frac{1}{n-1} \sum (\hat{\pi}_{ij} - \bar{\pi}_j)^2, \quad (6)$$

where

$$\bar{\pi}_j = \frac{1}{n} \sum \hat{\pi}_{ij}.$$

The value of β_j can depend on which other independent variables are included in the model. This issue is addressed here by including all of the independent variables when estimating β_j with b_j , which yields an estimate of σ_j^2 via (6).

A percentile bootstrap method is used to test (4). First, generate a bootstrap sample by sampling with replacement n vectors from n vectors $(Y_1, X_{11}, \dots, X_{1p}), \dots, (Y_n, X_{n1}, \dots, X_{np})$ yielding $(Y_1^*, X_{11}^*, \dots, X_{1p}^*), \dots, (Y_n^*, X_{n1}^*, \dots, X_{np}^*)$. Let $\hat{\sigma}_j^*$ be the bootstrap estimate of σ_j . Repeat this process B times yielding $\hat{\sigma}_{jb}^*$ ($b = 1, \dots, B$). Let C denote the proportion of times $\hat{\sigma}_{1b}^*$ is less than $\hat{\sigma}_{2b}^*$. That is,

$$C = \frac{1}{B} \sum \mathbf{I}(\hat{\sigma}_{1b}^* < \hat{\sigma}_{2b}^*),$$

RAND WILCOX

where the indicator function $I(\hat{\sigma}_{1b}^* < \hat{\sigma}_{2b}^*) = 1$ when $\hat{\sigma}_{1b}^* < \hat{\sigma}_{2b}^*$, otherwise $I(\hat{\sigma}_{1b}^* < \hat{\sigma}_{2b}^*) = 0$. From Liu and Singh (1997) a (generalized) p -value is $2\min(C, 1 - C)$. This will be called method P henceforth.

Wilcox (2018) found that method P performs poorly when Y is continuous and a robust regression estimator is used. An alternative method was found that performed reasonably well which differs from the bootstrap method used here in two crucial ways. First, Wilcox's method uses two independent bootstrap samples. The first yields $\hat{\sigma}_1^*$ and the second is used to compute $\hat{\sigma}_2^*$. The same bootstrap sample is used to get $\hat{\sigma}_1^*$ and $\hat{\sigma}_2^*$. The second difference is that C is replaced by

$$\frac{2}{B^2 - B} \sum_{j=1}^B \sum_{k=1}^B I(j < k) I(\hat{\sigma}_{1b}^* < \hat{\sigma}_{2b}^*). \quad (7)$$

An obvious speculation is Wilcox's method continues to perform well for the situation at hand, but simulations revealed that this is not the case.

The proposed method for testing (4) is readily generalized to comparing pairs of non-overlapping explanatory variables. Consider, for example, the case of $p = 4$ explanatory variables where the goal is to test

$$H_0 : \sigma_{12} = \sigma_{34}. \quad (8)$$

Then, proceed as before but with $\hat{A}_{12} = \exp(b_0 + b_1X_{i1} + b_2X_{i2})$ when estimating σ_{12} , and $\hat{A}_{34} = \exp(b_0 + b_3X_{i3} + b_4X_{i4})$ when estimating σ_{34} . Bootstrap estimates of σ_{12} and σ_{34} are computed as previously described.

Simulation Results

Simulations were used as a partial check on the ability of method P to control the probability of a Type I error when there are $p = 4$ independent variables. The independent variables were generated from a multivariate normal distribution having a common correlation ρ . Three values for ρ were used: 0.0, 0.5, and 0.8. The sample size was taken to be 50 and 100. Given the goal of testing (4), when $j = 1$ and $k = 2$, two choices for slopes were used: $(\beta_1, \beta_2, \beta_3, \beta_4) = (0, 0, 1, 1)$ and $(1, 1, 0, 0)$. Simulations were also run when testing (8); the slopes were then taken to be $(0, 0, 0, 0)$ and $(1, 1, 1, 1)$.

LOGISTIC REGRESSION

Table 1. Simulation estimates of the actual Type I error probability when testing (4), $\alpha = 0.05$

n	ρ	$(\beta_1, \beta_2, \beta_3, \beta_4)$	$B = 200$	$B = 500$
50	0.0	(0, 0, 1, 1)	0.004	0.002
50	0.5	(0, 0, 1, 1)	0.047	0.044
50	0.8	(0, 0, 1, 1)	0.061	0.045
100	0.0	(0, 0, 1, 1)	0.001	0.002
100	0.5	(0, 0, 1, 1)	0.061	0.053
100	0.8	(0, 0, 1, 1)	0.042	0.051
50	0.0	(1, 1, 0, 0)	0.070	0.065
50	0.5	(1, 1, 0, 0)	0.076	0.069
50	0.8	(1, 1, 0, 0)	0.064	0.057
100	0.0	(1, 1, 0, 0)	0.072	0.066
100	0.5	(1, 1, 0, 0)	0.069	0.058
100	0.8	(1, 1, 0, 0)	0.062	0.053

Table 2. Simulation results when testing (8), $\alpha = 0.05$

n	ρ	$(\beta_1, \beta_2, \beta_3, \beta_4)$	$B = 200$
50	0.0	(0, 0, 0, 0)	0.005
50	0.5	(0, 0, 0, 0)	0.002
50	0.8	(0, 0, 0, 0)	0.002
100	0.0	(0, 0, 0, 0)	0.004
100	0.5	(0, 0, 0, 0)	0.004
100	0.8	(0, 0, 0, 0)	0.005
50	0.0	(1, 1, 1, 1)	0.070
50	0.5	(1, 1, 1, 1)	0.068
50	0.8	(1, 1, 1, 1)	0.048
100	0.0	(1, 1, 1, 1)	0.072
100	0.5	(1, 1, 1, 1)	0.062
100	0.8	(1, 1, 1, 1)	0.056

Compiled in Table 1 are the estimates of the actual Type I error probability when testing (4) at the 0.05 level. The estimates are based on 2000 replications. Two choices for B were used: 200 and 500. Results in Wilcox (2018) suggest that $B = 200$ might suffice, which was the motivation for considering it here.

Although the seriousness of a Type I error can depend on the situation, Bradley (1978) suggests that when testing at the 0.05 level, as a general guide the actual level should be between 0.025 and 0.075. When $B = 200$, estimates are less than 0.075 in all situations except one, where it is 0.076. The difficulty is when there is no association with the independent variables and simultaneously the covariates have a common correlation of zero, the actual Type I error probability is

RAND WILCOX

estimated to be substantially less than 0.025. Increasing n , B , or both does not correct this problem. This was expected based on results in Wilcox (2018).

Contained in Table 2 are results when testing (8) with $B = 200$. In this case the estimates never exceed 0.075. Again, when there is no association and the covariates have a common correlation of zero, the estimates are substantially less than 0.025 as was expected.

Illustration

Method P is illustrated with data from the Well Elderly 2 study (Clark et al., 2011), which was generally focused on an intervention program aimed at improving the emotional and physical wellbeing of older adults. The focus was on data collected after intervention. One issue was the association between a measure of depressive symptoms (CESD) and two independent variables: a measure of life satisfaction (LSIZ) and the cortisol awakening response (CAR), which is the change in cortisol upon awakening and measured again 30-45 minutes later. Both enhanced and reduced CARs are associated with various psychosocial factors, including depression and anxiety disorders (e.g., Bhattacharyya, Molloy, & Steptoe, 2008; Pruessner, Hellhammer, Pruessner, & Lupien, 2003). A CESD score greater than 15 is regarded as an indication of mild depression. A score greater than 21 indicates the possibility of major depression. The goal was to understand the relative importance of the two independent variables in terms of the probability that a CESD score is greater than 15.

The explanatory strength of the associations was estimated to be 0.002 and 0.201 for CAR and LSIZ, respectively. The sample size is $n = 243$. The p -value when testing (4) is less than 0.001. ($B = 500$ was used.)

However, the logistic regression model assumes that the probability of the event under consideration has a monotonic association with the independent variables. As a partial check on this assumption, Figure 1 shows a plot of the regression surface based on a nonparametric smoother (e.g., Wilcox, 2017, section 15.5.4). Note that the plot suggests that the association is not monotonic. However, a monotonic association does appear to be reasonable when the CAR is negative, ignoring CAR values that are positive. And the same is true when the CAR is positive, ignoring CAR values that are negative. Focusing only on CAR values less than zero, the estimates of the explanatory strength of the associations were 0.070 and 0.189 for the CAR and LSIZ, respectively, and the p -value when testing (4) is 0.012. For positive CAR values, ignoring negative CAR values, the estimates were 0.121 and 0.199 and the p -value is 0.236. So, it appears that LSIZ is more important

LOGISTIC REGRESSION

than the CAR when the CAR is negative. When the CAR is positive, LSIZ is estimated to be more important, but the empirical evidence supporting this conclusion is weak. For this latter situation, the sample size is now $n = 94$.

It is known that leverage points, meaning outliers among the independent variables, can have an inordinate impact on the estimates of the slopes yielding a misleading indication of the nature of the association among the bulk of the data. The analysis just described was conducted again with leverage points removed resulting in the same conclusions.

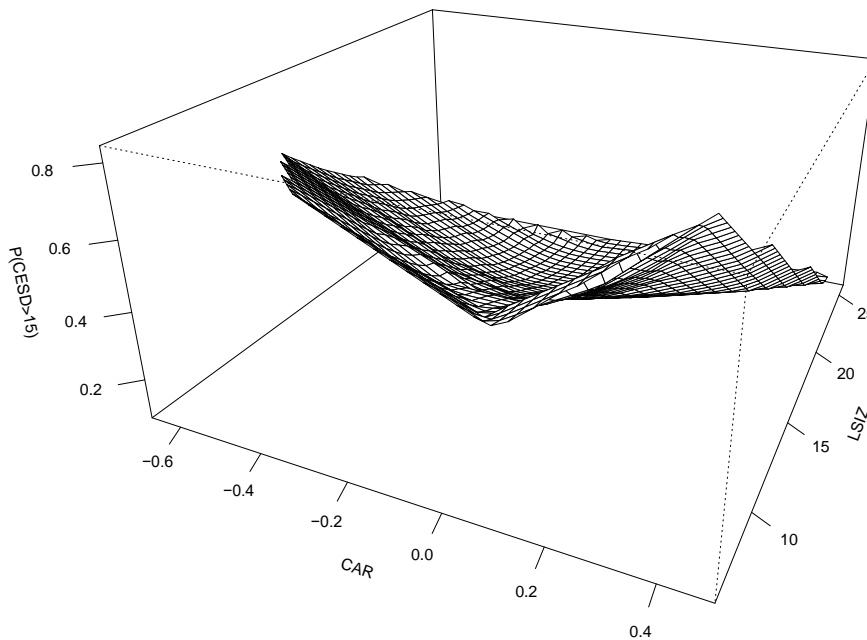


Figure 1. Shown is a smooth where the goal is to estimate the probability that CESD is greater than 15 given the cortisol awakening response (CAR) and a measure of life satisfaction (LSIZ)

Conclusion

An issue not addressed is testing (4) for every $j < k$ in a manner that controls familywise error rate (FWE), meaning the probability of one or more Type I errors. Some simulations were run based on the Bonferroni method. Reject (4) if the p -value is less than or equal to α/C , where C is the number of tests to be performed. With $n = 50$ and $B = 500$, situations were found where the actual FWE was estimated to be greater than 0.08 when testing at the 0.05 level. Increasing $B = 1000$ did not correct this problem. Increasing the sample size to $n = 100$, still using $B = 1000$, resulted in estimates less than 0.075 among the situations considered, but this issue is in need of further study.

As was illustrated, situations are encountered where the logistic regression model can be inappropriate. For $p = 2$ independent variables, a smooth of the regression surface might suggest how to deal with this issue. But of course, when $p > 2$, dealing with this concern is more difficult. One possibility is to use a partial residual plot as suggested by Fowlkes (1987). An analog of this approach can be applied via the R function `logrchk`, which is stored in the file `Rallfun-v35` described below. The resulting plot might suggest modifications of the logistic regression model that provides a more satisfactory approximation of the true association. But the extent to which this approach provides a satisfactory solution for the situation at hand is unclear.

The R function `logIVcom` applies the proposed method and has been stored in the file `Rallfun-v35`, which can be downloaded at <https://dornsife.usc.edu/labs/rwilcox/software>. The function will be added to the R package `WRS` as well, which can be installed at <https://github.com/nicebread/WRS>.

References

- Bhattacharyya, M. R., Molloy, G. J., & Steptoe, A. (2008) Depression is associated with flatter cortisol rhythms in patients with coronary artery disease. *Journal of Psychosomatic Research*, 65(2), 107-113. doi: 10.1016/j.jpsychores.2008.03.012
- Bradley, J. V. (1978) Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x
- Clark, F., Jackson, J., Carlson, M., Chou, C.-P., Cherry, B. J., Jordan-Marsh M., ... Azen, S. P. (2011). Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: results of the Well Elderly 2

LOGISTIC REGRESSION

Randomised Controlled Trial. *Journal of Epidemiology & Community Health*, 66(9), 782-790. doi: 10.1136/jech.2009.099754

Fowlkes, E. B. (1987). Some diagnostics for binary logistic regression via smoothing. *Biometrika*, 74(3), 503-515. doi: 10.1093/biomet/74.3.503

Lee, J., Sun, D., Sun, Y., & Taylor, J. (2016). Exact post-selection inference with the lasso. *The Annals of Statistics*, 44(3), 907-927. doi: 10.1214/15-aos1371

Liu, R. G., & Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association*, 92(437), 266-277.

Pruessner, M., Hellhammer, J. C., Pruessner, J. C., & Lupien, S. J. (2003). Self-reported depressive symptoms and stress levels in healthy young men: associations with the cortisol response to awakening. *Psychosomatic Medicine*, 65(1), 92-99. doi: 10.1097/01.psy.0000040950.22044.10

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514), 600-620. doi: 10.1080/01621459.2015.1108848

Wilcox, R. (2017). *Modern statistics for the social and behavioral sciences: A practical introduction* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC. doi: 10.1201/9781315154480

Wilcox, R. R. (2018). Robust regression: An inferential method for determining which independent variables are most important. *Journal of Applied Statistics*, 45(1), 100-111. doi: 10.1080/02664763.2016.1268105