BST 232: Methods I
Homework #3
Due October 9, 2024, 11:59pm on Canvas.

Readings

- Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N. and Altman, D.G., 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology, 31(4),* pp.337-350.
- ASA Statement on Statistical Significance and P-Values: `https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108#d1e849`

Question 1: Maximum likelihood estimation for the linear model with logistic errors

Recall that, when closed form maximum likelihood estimates are unavailable, we can use the Newton-Raphson algorithm for estimation. For a generic parameter vector $\boldsymbol{\theta} \in \mathbb{R}^q$, this algorithm is given by

- Start with initial values $\boldsymbol{\theta}^{(0)}$
- For $r \geq 1$
  - Set $\boldsymbol{\theta}^{(r+1)} \leftarrow \boldsymbol{\theta}^{(r)} + \left[\boldsymbol{I}\left(\boldsymbol{\theta}^{(r)}\right)\right]^{-1} \boldsymbol{U}(\boldsymbol{\theta}^{(r)})$
  - If $d(\boldsymbol{\theta}^{(r)}, \boldsymbol{\theta}^{(r+1)}) <$ tol then stop and set $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(r+1)}$. Else repeat.

  where d is a user-specified distance metric and tol is a user-specified constant used to gauge algorithm convergence.

In the above, $\boldsymbol{U}(\boldsymbol{\theta})$ is the score function and $\boldsymbol{I}(\boldsymbol{\theta})$ is the observed information matrix.

Write an R function (that does not rely on any existing optimization functions/packages) to estimate the regression coefficients for the example given in the notes of the linear model with errors distributed $\varepsilon_i \overset{iid}{\sim}$ Logistic($\mu = 0, s = 1$). The function should allow for an arbitrary number of predictors and should use the euclidean distance as the distance measure in the convergence check. Include the code for the function in your response to this problem. Then, apply the function to the HERS data to estimate the regression coefficients for the simple linear model with log(LDL)

as the outcome and BMI as the covariate. You may omit any observations with missingness in LDL or BMI. Use tol=0.001 and a vector of starting values $\boldsymbol{\beta}^{(0)} = (4.94, 0)'$ (note that we set the starting value for the intercept at the mean of log(LDL)). Report your estimates and comment on any differences from the ones you got using OLS in Homework 2, question 3a.

Question 2: Origins of the regression F-test

Over time, statisticians have relied on several different "systems" to construct hypothesis tests. *Wald-style* tests, as discussed in lecture, use the distribution of the estimator itself. By contrast, *likelihood ratio* (LR) tests compare the distribution of the estimator to the distribution of the data under the null hypothesis. While more complicated to construct and prove, likelihood ratio tests often achieve better finite-sample performance. In lecture we discussed two strategies for testing a hypothesis that a group of regression coefficients are all equal to zero. One strategy was the Wald Test, but the other was using the F-test. The F-test statistic has the form

$$F = \frac{(SSE_R - SSE_F)/(df_R - df_F)}{SSE_F/df_F} \sim F(df_R - df_F, df_F) \tag{1}$$

where subscripts R and F indicate quantities from the reduced model and full model, respectively, and *df* refers to the degrees of freedom of the SSE. Note that while this representation of the F-statistic looks different than what we showed in class, the two work out to be equivalent. In this question, we'll show that the regression F-test is actually a likelihood ratio test.

a. Write out the likelihood $\mathcal{L}(\boldsymbol{\beta}, \sigma^2)$ of the linear model in terms of matrices and vectors, assuming that $\boldsymbol{\varepsilon} \sim \text{Normal}(\mathbf{0}, \sigma^2\mathbf{I}_n)$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

b. Consider plugging the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ into the likelihood from (a). Rewrite the likelihood $\mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ in terms of $SSE = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ and simplify algebraically (your answer should contain no $\hat{\boldsymbol{\beta}}$ or $\hat{\sigma}^2$ terms other than those within the SSE term). Hint: recall that the MLE $\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$.

c. Use the likelihood from (b) to compute the estimated likelihood ratio $LR = \mathcal{L}(\hat{\boldsymbol{\beta}}_R, \hat{\sigma}^2_R)/\mathcal{L}(\hat{\boldsymbol{\beta}}_F, \hat{\sigma}^2_F)$, which compares the likelihood of some reduced model R missing some coefficients to that of the full model F with all coefficients. Write this ratio in terms of the sums of squares SSE_R and SSE_F.

d. While the finite-sample distribution of LR is unknown, the F-statistic in Equation (1) has a known distribution. Fortunately, as you will learn in BIOSTAT 231: Inference I, the likelihood ratio is invariant under monotonic transformation. Find a monotonic transformation g such that $g(LR) = F$ to prove that the classical linear regression F-test is a likelihood ratio test.

Question 3: Testing for effect modification

The data contained in the file `epa.dat` was generated by the Environmental Protection Agency. The amount of magnesium uptake was measured on different subjects, with each subject receiving one of two treatments and being measured at seven times (TIME=0,1,2,3,4,5,6). It was anticipated that the two treatments used may result in different regression equations for time. For simplicity, we here assume repeated measures over time from the same subject are independent. You will learn how to relax this assumption in the future longitudinal data course.

a. A model is postulated in which magnesium uptake is regressed against time in a quadratic regression, controlling for treatment but assuming that the effect of time is the same for the two treatments. Write out the this regression model, and fit the model to the EPA data.

b. Suppose, based on your model in (a), you are interested in testing whether mean uptake varies with time. What is the null hypothesis of interest (please write out this hypothesis in terms of the regression coefficients)? Test this null hypothesis, and state your conclusions.

c. Interest also focuses on determining whether simply controlling for treatment is adequate, or whether it modifies the effect of time. Consider the more general model:

$$E(Y_i) = \begin{cases} \beta_0 + \beta_1 \text{TIME}_i + \beta_2 \text{TIME}_i^2 \text{ for treatment 1} \\ \gamma_0 + \gamma_1 \text{TIME}_i + \gamma_2 \text{TIME}_i^2 \text{ for treatment 2.} \end{cases}$$

Write this structure out in one unified model for uptake. Based on this representation, set up the general linear hypothesis for testing:

$$H_0 : \beta_1 = \gamma_1 \text{ and } \beta_2 = \gamma_2.$$

That is, set up $\mathbf{C}$ and $\boldsymbol{\beta}$ such that this hypothesis corresponds to $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, where $\boldsymbol{\beta}$ contains all of the regression coefficients from the model for both treatments.

d. Test the hypothesis considered in (c) using either an F-test or a Wald test. What do you conclude in terms of the scientific question of interest?

Question 4: Multiple comparisons

A researcher is interested in determining whether the expression levels of several genes are significantly different between two conditions (e.g., diseased vs. healthy). The researcher fits a regression

model and performs hypothesis tests on 50 genes. The p-values for the 50 tests are given in the file `gene_pvals.csv`.

a. Bonferroni Correction:

   i. Apply the Bonferroni correction to the p-values. Provide the adjusted p-values and indicate which hypotheses are significant at the 0.05 level.

   ii. Discuss the advantages and disadvantages of the Bonferroni correction in this context.

b. Holm's Method:

   i. Apply Holm's method to the p-values. Provide the adjusted p-values and indicate which hypotheses are significant at the 0.05 level.

   ii. Compare Holm's method with the Bonferroni correction. Which is more powerful in this case, and why?

c. False Discovery Rate (FDR):

   i. Apply the Benjamini-Hochberg procedure to control the FDR at the 5% level. Provide the adjusted p-values and indicate which hypotheses are significant at the 0.05 level.

   ii. Discuss the interpretation of controlling the FDR in this context and the trade-off between FDR control and family-wise error rate (FWER) control.

d. Critical Thinking:

   i. In your own words, explain how the choice of multiple comparisons correction method impacts the replicability of the study's findings.

   ii. Discuss the role of sample size in the context of multiple comparisons. How does increasing the sample size affect the power and the need for multiple comparisons correction?