STA 712 Homework 4

Due: Friday, September 30, 12:00pm (noon) on Canvas.

Instructions: Submit your work as a single PDF. For this assignment, you may include written work by scanning it and incorporating it into the PDF. Include all R code needed to reproduce your results in your submission.

Optional: Delta method

Note: This question is optional! I have included it here in case you are interested in convergence of random variables, but it is slightly tangential to the course material and can be safely ignored if you choose.

Let $\theta \in \mathbb{R}^d$ be a parameter of interest, and $\widehat{\theta}$ be an estimate (e.g., the MLE). Our Wald tests and intervals depend on convergence in distribution to a normal:

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma),$$

where $\Sigma = Var(\widehat{\theta})$. We also know that if $\boldsymbol{a} \in \mathbb{R}^d$, then $\boldsymbol{a}^T \widehat{\theta} \approx N(\boldsymbol{a}^T \theta, \boldsymbol{a}^T \Sigma \boldsymbol{a})$.

But what if we are interested in a *nonlinear* function $g(\theta)$, for some $g : \mathbb{R}^d \to \mathbb{R}$? It turns out that, under certain conditions, $g(\widehat{\theta})$ is actually (approximately) normal too! Formally, if g is a continuously differentiable function, then

$$\sqrt{n}(g(\widehat{\theta}) - g(\theta)) \xrightarrow{d} N\left(0, \left(\frac{\partial g}{\partial \theta}\right)^T \Sigma \left(\frac{\partial g}{\partial \theta}\right)\right),$$

where $\frac{\partial g}{\partial \theta}$ is the gradient of g evaluated at θ. This is called the *(multivariate) delta method.*

The purpose of this problem is to derive the delta method in the univariate case (the same intuition applies to the multivariate case). In the univariate case, $d = 1$ and $\theta \in \mathbb{R}$. To prove the univariate delta method, we will use the following results:

- Taylor's theorem: If $g : \mathbb{R} \to \mathbb{R}$ is continuously differentiable, then there exists a continuous function h such that
$$g(x) = g(a) + g'(a)(x - a) + h(x)(x - a),$$
where $h(x) \to 0$ as $x \to a$.

- Slutsky's theorem: Suppose that X_n and Y_n are sequences of random variables with $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c \in \mathbb{R}$. Then
$$X_n Y_n \xrightarrow{d} cX$$
$$X_n + Y_n \xrightarrow{d} X + c$$

- If $Y_n \xrightarrow{d} c \in \mathbb{R}$, then $Y_n \xrightarrow{p} c$.

- Continuous mapping theorem: If g is a continuous function, and $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.

1. Now let us prove the univariate delta method: if $\sqrt{n}(\widehat{\theta}-\theta) \xrightarrow{d} N(0, \sigma^2)$, and g is a continuously differentiable function with $g'(\theta) \neq 0$, then

$$\sqrt{n}(g(\widehat{\theta}) - g(\theta)) \xrightarrow{d} N(0, \sigma^2[g'(\theta)]^2).$$

   (a) Using Taylor's theorem, show that

   $$\sqrt{n}(g(\widehat{\theta}) - g(\theta)) = \sqrt{n}g'(\theta)(\widehat{\theta} - \theta) + \sqrt{n}(\widehat{\theta} - \theta)h(\widehat{\theta}),$$

   for some continuous h such that $\lim_{\widehat{\theta} \to \theta} h(\widehat{\theta}) = 0$.

   (b) Using Slutsky's theorem, argue that $\sqrt{n}g'(\theta)(\widehat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2[g'(\theta)]^2)$.

   (c) Using Slutsky's theorem and the continuous mapping theorem, argue that

   $$\sqrt{n}(\widehat{\theta} - \theta)h(\widehat{\theta}) \xrightarrow{p} 0.$$

   (d) Using Slutsky's theorem, conclude that $\sqrt{n}(g(\widehat{\theta}) - g(\theta)) \xrightarrow{d} N(0, \sigma^2[g'(\theta)]^2)$.

   (e) Finally, let's apply the univariate delta method to logistic regression! Suppose that

   $$Y_i \sim Bernoulli(p_i)$$

   $$\log\left(\frac{p_i}{1-p_i}\right) = \beta^T X_i$$

   and we want to construct a confidence interval for p_i, the probability for the ith observation. In class, we constructed a confidence interval for $\beta^T X_i$, and transformed the endpoints. Another method would be to recognize that $\widehat{\beta}^T X_i$ is approximately normal, so by the delta method $\widehat{p}_i$ is approximately normal too.

   Using the univariate delta method, show that

   $$\widehat{p}_i \approx N(p_i, (X_i^T \mathcal{I}^{-1}(\beta) X_i)(1-p_i)^2).$$

   Note: Simply transforming the endpoints is a much more common technique to create a confidence interval for $\widehat{p}_i$. In practice it is very rare to actually use the delta method for this problem.

Deviance and likelihood ratio tests for linear regression

In class, we defined the *deviance* for a fitted model with estimated coefficients $\widehat{\beta}$ as

$$2\ell(\text{saturated model}) - 2\ell(\widehat{\beta}).$$

In a generalized linear model, deviance plays the same role as the residual sum of squares (SSE) in a linear regression model. The purpose of this question is to make that connection explicit, and to connect the likelihood ratio test with the nested F-test from linear regression.

2. Consider the linear regression model

$$Y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \beta^T X_i.$$

We observe data $(X_1, Y_1), ..., (X_n, Y_n)$, and calculate estimated coefficients $\widehat{\beta} \in \mathbb{R}^{k+1}$. In the fitted model, $\widehat{\mu}_i = \widehat{\beta}^T X_i$. In the saturated model, $\widehat{\mu}_i = Y_i$.

(a) Show that for this linear regression model, if σ^2 is known then the deviance is $\dfrac{SSE}{\sigma^2}$, where $SSE = \sum\limits_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$.

(Technically, this is the *scaled* deviance; there is also an *unscaled* deviance, which we will talk about later. For linear regression, the unscaled deviance is just SSE. For binary logistic regression, the scaled and unscaled deviances are the same).

(b) Suppose we want to test the hypotheses $H_0 : \beta = \beta^0$ vs. $H_A : \beta \neq \beta^0$, using a likelihood ratio test. Let $\widehat{\beta}$ be the estimated coefficients from the *full* model, and $\widehat{\beta}^0$ the estimated coefficients from the *reduced* model. Show that if σ^2 is known, then the likelihood ratio test statistic is

$$G = \frac{SSE_{reduced} - SSE_{full}}{\sigma^2} \sim \chi^2_q,$$

where q is the number of parameters tested.

(c) Using the previous questions, explain why

$$\frac{SSE_{full}}{\sigma^2} \sim \chi^2_{n-(k+1)}.$$

(d) Ok, but what happens if we *don't* know σ^2? The natural step is to estimate σ^2. Recall that in linear regression,

$$\widehat{\sigma}^2 = \frac{SSE_{full}}{n-(k+1)}.$$

Unfortunately, the statistic

$$\frac{SSE_{reduced} - SSE_{full}}{\widehat{\sigma}^2}$$

does not have a nice distribution. *Fortunately*, we can modify this statistic slightly so that everything works out! Recall the nested F statistic from linear regression:

$$F = \frac{(SSE_{reduced} - SSE_{full})/q}{\widehat{\sigma}^2} \sim F_{q,n-(k+1)},$$

where q is the number of parameters tested (the difference in the number of parameters between the full and reduced models). Also recall that an F_{ν_1,ν_2} distribution can be written as $\dfrac{S_1/\nu_1}{S_2/\nu_2}$, where $S_1 \sim \chi^2_{\nu_1}$ and $S_2 \sim \chi^2_{\nu_2}$ are independent.

Using the results from the previous questions, show that $F = \dfrac{(SSE_{reduced} - SSE_{full})/q}{\widehat{\sigma}^2}$ can indeed be written as $\dfrac{S_1/\nu_1}{S_2/\nu_2}$, where $S_1 \sim \chi^2_{\nu_1}$ and $S_2 \sim \chi^2_{\nu_2}$. (Proving independence is annoying, so we won't do that part here).

Data analysis

You are contacted by the US Small Business Administration (SBA), a government agency dedicated to helping support small businesses. The SBA provides loans to small businesses, but some businesses *default* on their loan (i.e., fail to pay it back). Researchers at the SBA are interested in predicting whether a business will default on the loan, and they have collected a random sample of 5000 different loans.

You can load the SBA data into R by

```
sba <- read.csv("https://sta712-f22.github.io/homework/sba_small.csv")
```

For each loan, we have the following variables:

- LoanNr_ChkDgt: Loan ID number that uniquely identifies each loan
- Name: Name of business receiving the loan
- City: City the business is based in
- State: State the business is based in (two-letter abbreviation)
- Zip: ZIP code the business is based in
- Bank: Name of bank making the loan
- BankState: State of the bank making the loan (two-letter abbreviation)
- NAICS: North American Industry Classification System code identifying the industry of the business receiving the loan
- ApprovalDate: Date of approval (YYYY-MM-DD) of the loan
- ApprovalFY: Fiscal year of approval of the loan
- Term: Length of the loan term (months)
- NoEmp: Number of employees of the business before receiving the loan
- NewExist: 1 if business already existed, 2 if business is new
- CreateJob: Number of jobs the business expects to create using the loan money
- RetainedJob: Number of jobs the business expects to retain because they received the loan
- FranchiseCode: For businesses that are franchises, a unique five-digit code identifying which brand they are a franchise of. 0 or 1 if the business is not a franchise.
- UrbanRural: 1 if business is in urban area, 2 if business is in rural area, 0 if unknown
- RevLineCr: Y if this is a revolving line of credit, N if not
- LowDoc: Y if loan was issued under the 'LowDoc Loan' program, which allows loans under $150,000 to be processed with a short one-page application. N if loan is issued with a standard application, which is much longer
- ChgOffDate: The date (YYYY-MM-DD) the loan was declared to be in default, if the borrower stopped paying it back

- DisbursementDate: Date (YYYY-MM-DD) the loan money was disbursed to the business
- DisbursementGross: The amount of money disbursed (loaned), in dollars
- BalanceGross: The amount of money remaining to be paid back, in dollars
- MIS_Status: Current loan status. CHGOFF = charged off, P I F = paid in full.
- ChgOffPrinGr: Amount of money charged off, if the borrower defaulted, in dollars
- GrAppv: Gross amount of loan approved by the bank, in dollars
- SBA_Appv: Amount of the loan guaranteed by the SBA, in dollars

Research questions: Researchers at the SBA are interested in the relationship between loan amount and whether the business defaults on the loan. They believe that whether the business is new vs. an existing business, and whether it is in an urban vs. rural environment, may also be related to the chance of defaulting. The SBA gives you the data, and asks the following questions:

- Is there a relationship between loan amount and the probability the business defaults on the loan, after accounting for whether or not the business is new, and whether it is in an urban or rural environment?
- Is there a difference in default rates between urban and rural businesses, after accounting for loan amount and urban vs. rural environment?
- The SBA is concerned when a loan has more than a 30% chance of default. What range of loan amounts should the SBA be concerned about for a new, urban business?

3. Here you will use logistic regression to answer the SBA's questions.

   (a) Which variables should we focus on to answer the SBA's questions? Which of these is our response variable, and which will be our explanatory variables, for logistic regression?

   (b) Perform univariate exploratory data analysis (EDA) for your selected variables in (a):

   - For categorical variables, present a table showing the number of observations in each category. *Note: check carefully how your explanatory variables are coded in the data! If a categorical variable is coded numerically, you may need to convert it to a* `factor` *in R before proceeding.*
   - For quantitative variables, present a histogram and summarize the distribution of the variable (give summary statistics and describe center, shape, spread, and any potential outliers)
   - Discuss whether there are any missing or erroneous values in the data, and if so how you will handle them. *Note: if we remove missing data, we should only remove rows with missing data in the columns we care about. Be careful with functions like* `drop_na`*, which will remove rows with missing values in ANY column.*

   (c) Perform multivariate EDA for your selected variables in (a):

   - Create empirical logit plots to summarize the relationship between quantitative predictors and your binary response. Details on creating empirical logit plots, with examples, can be found at `https://sta712-f22.github.io/homework/empirical_logits.html`
   - Using the empirical logit plots, discuss whether any transformations are needed on the explanatory variables.

- Use empirical logit plots to investigate potential interactions between variables
- Use a correlation matrix to summarize pairwise relationships between the quantitative explanatory variables. Should we be concerned with potential multicollinearity? (Ignore this if you have < 2 quantitative explanatory variables).

(d) Based on your exploratory data analysis, write down a logistic regression model that will allow you to answer the SBA's questions. Describe how you will use the model to answer their questions.

(e) Fit your model from (d), and report the equation of the fitted model. Interpret any estimated coefficients which address the research questions.

(f) Assess your model assumptions:

- Create quantile residual plots to check the shape assumption for any quantitative variables (you may use the `qresid` function in the `statmod` package)
- Calculate Cook's distance to check for any influential points (use a threshold of 0.5 or 1 to identify influential points)
- Calculate variance inflation factors to check for multicollinearity (see the `vif` function in the `car` package, and use a threshold of 5 or 10 to identify high multicollinearity). Note that if you have interaction terms in your model, you will see high VIFs for any variable involved in an interaction – this isn't a problem.

(g) Address any violations to the model assumptions (transformations for shape violations; report results with and without influential points; and combine or remove columns for high multicollinearity). If you made any changes to your model from (e), report and interpret your new fitted model here.

(h) Now let's address the first research question. Test whether there is a relationship between loan amount and the probability of default. (It is typical to use a Wald test when testing a single parameter, and a likelihood ratio test when testing multiple parameters). You should:

- State the null and alternative hypotheses in terms of one or more βs
- Calculate a test statistic and p-value
- Make a conclusion in the context of the original question

(i) Next, we will address the second research question. Test whether there is a difference in default rates between urban and rural businesses. If possible, also report and interpret a confidence interval for the difference between urban and rural businesses.

(j) Finally, we will address the third research question.

i. For what range of loan amounts is the estimated probability of default at least 0.3 for a new, urban business?

ii. **Optional:** The range in (i) involves an upper bound on the loan amount. Using the multivariate delta method (see Q1), create a 95% confidence interval for this upper bound.