Lesson 26. Inference for Logistic Regression

1 Overview

- The simple logistic regression model (in logit form):

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X \qquad \pi = P(Y = 1)$$

- Recall the simple linear regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \varepsilon \sim \text{iid } N(0, \sigma_\varepsilon^2)$$

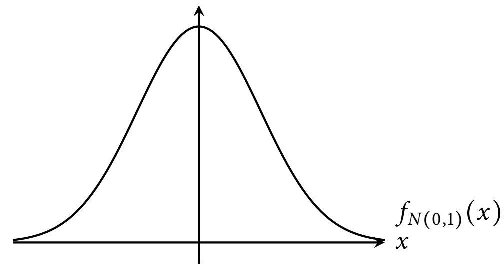
- How does inference for simple logistic regression compare to inference for simple linear regression?

	Linear regression	Logistic regression
Test for β_1	t-test	
CI for β_1	$\hat{\beta}_1 \pm t_{\alpha/2,n-2} SE_{\hat{\beta}_1}$	
Test for overall model	ANOVA F-test (change in SSE)	

2 z-test (Wald test) for the slope of a simple logistic regression model

- Question: Is the slope of the explanatory variable different from zero?
- Formal steps:
  1. State the hypotheses:
  2. Calculate the test statistic:

3. Calculate the p-value:
   - If the conditions for logistic regression hold, then the test statistic follows
   - p-value =



4. State your conclusion, based on the given significance level α

   If we reject H_0 (p-value $\le \alpha$):

   We reject H_0 because the p-value is less than the significance level α. We see evidence that $\underline{X}$ is significantly associated with $\underline{Y}$.

   If we fail to reject H_0 (p-value $> \alpha$):

   We fail to reject H_0 because the p-value is greater than the significance level α. We do not see evidence that $\underline{X}$ is significantly associated with $\underline{Y}$.

3 Confidence intervals for the slope of a simple logistic regression model

- The $100(1-\alpha)$% **confidence interval for the slope** β_1 is

- The $100(1-\alpha)$% **confidence interval for the odds ratio** e^{β_1} is

> **Example 1.** Continuing with the `MedGPA` data from previous lessons...
>
> We looked at a binary response variable (*Acceptance* = 1 if accepted, 0 if not) and a quantitative predictor (*GPA*) for 55 medical school applicants from a college in the Midwest.
>
> We fit the following logistic regression model:
>
> $$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 GPA \qquad \text{where} \quad P(Acceptance = 1)$$
>
> We get the following summary output from R:

```
Call:
glm(formula = Acceptance ~ GPA, family = binomial, data = MedGPA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7805  -0.8522   0.4407   0.7819   2.0967

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -19.207      5.629  -3.412 0.000644 ***
GPA            5.454      1.579   3.454 0.000553 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 75.791  on 54  degrees of freedom
Residual deviance: 56.839  on 53  degrees of freedom
AIC: 60.839

Number of Fisher Scoring iterations: 4
```

a. Is the association between *Acceptance* and *GPA* statistically significant? Use a significance level of 0.05. Report the relevant values from the summary output.
b. Give a 95% confidence interval on the odds ratio corresponding to a unit increase in *GPA*. What does this confidence interval mean?

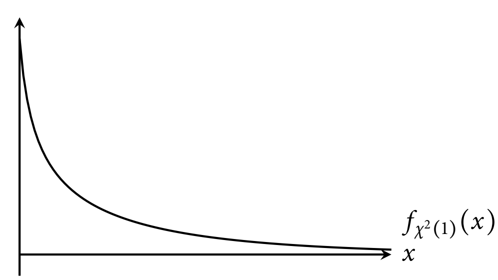
4 Likelihood ratio test (LRT) for utility of a simple logistic regression model

- A quick review of **maximum likelihood estimation** from SM239...
- The **likelihood** of the data, denoted by L, is the joint PDF of the data, regarded as a function of the unknown parameters with the data values fixed
- The **method of maximum likelihood** chooses parameter values to maximize L
    - The method of maximum likelihood is used to fit the logistic regression model
- Equivalently, we can minimize $-2 \log L$, which is called the **deviance**
- We compare nested logistic regression models by observing the change in deviance ("drop in deviance"), much like we observed changes in SSE for nested linear regression models
- Question: is the model useful?
- Formal steps:

1. State the hypotheses:

2. Calculate the test statistic:

3. Calculate the p-value:
    - If the conditions for logistic regression hold, then the the test statistic follows
    - p-value =



4. State your conclusion, based on the given significance level α

If we reject H_0 (p-value $\leq \alpha$):

We see significant evidence that the model is useful.

If we fail to reject H_0 (p-value $> \alpha$):

We do not see significant evidence that the model is useful.

5 Note: different tests for the same hypotheses?

- The z-test and the LRT we discussed above are asymptotically equivalent
- With smaller sample sizes, they may give different conclusions
  - If this happens, "trust" LRT
- The LRT is considered better than the z-test, but is more difficult computationally
- When we study multiple logistic regression models, we will see that these tests have different hypotheses (and uses)

Example 2. Continuing with the `MedGPA` data from Example 1...

Is the model useful? Use a significance level of 0.05.