STATS 111/202

Lecture 10: GLM inference and confidence intervals

2/19/2025

Ana Maria Kenney

UC Irvine Department of Statistics

GLM Inference

4.2.3

GLM Inference

- Now we can select the random component and link function to fit a GLM
- Can interpret the regression coefficients
- But still want to know whether these coefficients are statistically significant? E.g.:
  - Does MCAT actually have an association with Acceptance?
  - Do Sex and MCAT interact?

```
glm(formula = Acceptance ~ MCAT + Sex + MCAT * Sex, family
= binomial(link = "logit"), data = mcat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7857  -0.9770   0.3549   0.9417   2.0304

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.1804     4.3247  -1.429    0.153
MCAT          0.1887     0.1212   1.557    0.119
SexM         -7.2122     7.1083  -1.015    0.310
MCAT:SexM     0.1697     0.1946   0.872    0.383

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 75.791  on 54  degrees of freedom
Residual deviance: 60.924  on 51  degrees of freedom
AIC: 68.924
```

GLM Inference

- Consider a GLM:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Let's say we want to test whether a single coefficient is equal to 0

- Our null is then:

$$H_0\colon \beta_j = 0, \qquad j = 1,2, \ldots, p$$

- Here we are testing if $\beta_j = 0$ **given all other coefficients are already in the model**

Recall in standard linear regression

- This is a standard setting in simple linear regression with a continuous response

- Let's say Y is continuous (assumed to be Normal), then under the null we have the true population mean μ but not variance σ^2

- Our test statistic is then

$$t = \frac{\hat{\beta}_j - \beta_j}{\widehat{SE}(\beta_j)}$$

- This follows an approximate t-distribution with n-p-1 degrees of freedom

Recall in standard linear regression

- Alternatively, we can compare a full model to a reduced model and create a F-statistic (following an F distribution)

- Here the reduced model follows:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \cdots + \beta_p X_p$$

- And the full model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Recall in standard linear regression

- Alternatively, we can compare a full model to a reduced model and create a F-statistic (following an F distribution)

- Here the reduced model follows:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \boxed{\beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1}} + \cdots + \beta_p X_p$$

- And the full model:

Notice that $\beta_j = 0$ in the reduced model (hence that term is missing)

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

What if we want to test several coefficients?

Recall in standard linear regression

- Let's say we want to test whether β_3 and β_5 are equal to 0

- Here our null follows:

$$H_0: \beta_3 = \beta_5 = 0 \ vs \ H_a: \boldsymbol{at\ least\ one}\ is\ not\ 0$$

- This is really testing if β_3 or β_5 are equal to 0, given all other coefficients are in the model ($\beta_0, \beta_1, \beta_2, \beta_4, \beta_6, \dots, \beta_p$ are not equal to 0)

- Then our reduced model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_6 X_6 \cdots + \beta_p X_p$$

- Our full model is:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Recall in standard linear regression

- Let's say we want to test whether β_3 and β_5 are equal to 0

- Here our null follows:

$$H_0: \beta_3 = \beta_5 = 0 \; vs \; H_a: \boldsymbol{at\ least\ one}\ is\ not\ 0$$

- This is really testing if β_3 or β_5 are equal to 0, given all other coefficients are in the model ($\beta_0, \beta_1, \beta_2, \beta_4, \beta_6, \ldots, \beta_p$ are not equal to 0)

- Then our reduced model is:

$$Y = \beta_0 + \beta_1 X_1 + \boxed{\beta_2 X_2 + \beta_4 X_4 + \beta_6 X_6} \cdots + \beta_p X_p$$

Missing β_3 and β_5 since this is under the null and $\beta_3 = \beta_5 = 0$

- Our full model is:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

GLM Inference

- Back to GLM setting:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Similarly, if we want to test single coefficients:

$$H_0: \beta_j = 0 \; for \; j = 1,2, \dots, p \; vs \; H_a: \beta_j \neq 0 \; or \; \beta_j > 0 \; or \; \beta_j < 0$$

- Under the null, the test statistics follows:

$$z = \frac{\hat{\beta}_j - \beta_j}{\widehat{SE}(\beta_j)}$$

- This follows an approximate standard normal distribution, can use Z test to find p-values

GLM Inference

- Back to GLM setting:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Similarly, if we want to test single coefficients:

$$H_0: \beta_j = 0 \; for \; j = 1,2, \ldots, p \; vs \; \boxed{H_a: \beta_j \neq 0} \; or \; \beta_j > 0 \; or \; \beta_j < 0$$

Default in R, going to focus on this for now

- Under the null, the test statistics follows:

$$z = \frac{\hat{\beta}_j - \beta_j}{\widehat{SE}(\beta_j)}$$

- This follows an approximate standard normal distribution, can use Z test to find p-values

Example

- Let's revisit the Framingham heart data with chd (coronary heart disease) as a response and sbp (systolic blood pressure), age, and their interaction as explanatory variables. The population model follows:

$$logit(\mu_i) = \beta_0 + \beta_1 SBP_i + \beta_2 Age_i + \beta_3 SBP_i \times Age_i$$

```
glm(formula = chdfate ~ sbp + age + sbp * age, family = binomial(link = "logit"),
    data = framingham)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -8.0951608  1.1959097  -6.769 1.30e-11 ***
sbp          0.0470097  0.0090109   5.217 1.82e-07 ***
age          0.1128184  0.0239063   4.719 2.37e-06 ***
sbp:age     -0.0006723  0.0001766  -3.808  0.00014 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5844.1  on 4698  degrees of freedom
Residual deviance: 5648.6  on 4695  degrees of freedom
AIC: 5656.6

Number of Fisher Scoring iterations: 4
```

Example

- Let's revisit the Framingham heart data with chd (coronary heart disease) as a response and sbp (systolic blood pressure), age, and their interaction as explanatory variables. The population model follows:

$$logit\ (\mu_i) = \beta_0 + \beta_1 SBP_i + \beta_2 Age_i + \boxed{\beta_3 SBP_i \times Age_i}$$

```
glm(formula = chdfate ~ sbp + age + sbp * age, family = binomial(link = "logit"),
    data = framingham)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.0951608  1.1959097  -6.769 1.30e-11 ***
sbp          0.0470097  0.0090109   5.217 1.82e-07 ***
age          0.1128184  0.0239063   4.719 2.37e-06 ***
sbp:age     -0.0006723  0.0001766  -3.808  0.00014 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5844.1  on 4698  degrees of freedom
Residual deviance: 5648.6  on 4695  degrees of freedom
AIC: 5656.6

Number of Fisher Scoring iterations: 4
```

- Let's say we want to test the interaction term
- That is, testing if the effect of age on chd depends on sbp (or the effect of sbp on chd depends on age)
- Is the effect of age modified by sbp (or the effect of sbp modified by age)?
- Our hypothesis:

$$H_0: \beta_3 = 0 \ vs \ H_a: \beta_3 \neq 0$$

What can we conclude?

Example

- Let's revisit the Framingham heart data with chd (coronary heart disease) as a response and sbp (systolic blood pressure), age, and their interaction as explanatory variables. The population model follows:

$$logit\ (\mu_i) = \beta_0 + \beta_1 SBP_i + \beta_2 Age_i + \boxed{\beta_3 SBP_i \times Age_i}$$

```
glm(formula = chdfate ~ sbp + age + sbp * age, family = binomial(link = "logit"),
    data = framingham)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.0951608  1.1959097  -6.769 1.30e-11 ***
sbp          0.0470097  0.0090109   5.217 1.82e-07 ***
age          0.1128184  0.0239063   4.719 2.37e-06 ***
sbp:age     -0.0006723  0.0001766  -3.808  0.00014 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5844.1  on 4698  degrees of freedom
Residual deviance: 5648.6  on 4695  degrees of freedom
AIC: 5656.6

Number of Fisher Scoring iterations: 4
```

- At a 0.05 significance level, we reject the null and conclude evidence for the alternative that $\beta_3 \neq 0$

GLM Inference

- Now let's extend to confidence intervals

- A $(1-\alpha) \times 100\%$ CI for population parameter θ is:

$$\hat{\theta} \pm z_{\alpha/2}\, \widehat{SD}(\hat{\theta})$$

- Here $\hat{\theta}$ is the point estimate of θ, $z_{\alpha/2}$ is the standard normal multiplier, and $\widehat{SD}(\hat{\theta})$ is the estimated standard deviation of the estimate $\hat{\theta}$

- In logistic regression, we are **not necessarily interested in CI for $\hat{\beta}$, but more specifically $e^{\hat{\beta}}$**

GLM Inference

- Now let's extend to confidence intervals

- A $(1-\alpha)\times 100\%$ CI for population parameter θ is:

$$\hat{\theta} \pm z_{\alpha/2}\,\widehat{SD}(\hat{\theta})$$

- Here $\hat{\theta}$ is the point estimate of θ, $z_{\alpha/2}$ is the standard normal multiplier, and $\widehat{SD}(\hat{\theta})$ is the estimated standard deviation of the estimate $\hat{\theta}$

- In logistic regression, we are **not necessarily interested in CI for $\widehat{\boldsymbol{\beta}}$, but more specifically $\boldsymbol{e^{\widehat{\beta}}}$**

How do we go from a CI for $\hat{\beta}$ to $e^{\widehat{\beta}}$?

GLM Inference

- Now let's extend to confidence intervals

- A $(1-\alpha)\times 100\%$ CI for population parameter θ is:

$$\hat{\theta} \pm z_{\alpha/2}\,\widehat{SD}(\hat{\theta})$$

- Here $\hat{\theta}$ is the point estimate of θ, $z_{\alpha/2}$ is the standard normal multiplier, and $\widehat{SD}(\hat{\theta})$ is the estimated standard deviation of the estimate $\hat{\theta}$

- In logistic regression, we are **not necessarily interested in CI for $\widehat{\boldsymbol{\beta}}$, but more specifically $e^{\widehat{\boldsymbol{\beta}}}$**

How do we go from a CI for $\hat{\beta}$ to $e^{\hat{\beta}}$?

Exponentiate!

Confidence interval review

- The interval (l,u) are all values of θ_0 which we will fail to reject under the null $\theta = \theta_0$

- For example, if the interval is (5,10), then we would fail to reject $\theta = \theta_0$ for all values of θ_0 in (5,10)

- We will reject the null for all values of θ_0 outside of (5,10)

- Going to use glmCI() function (on Canvas in lecture R code) to compute 95% CI for odds ratios

Example

- Let's revisit the Framingham heart data with chd (coronary heart disease) as a response and sbp (systolic blood pressure), age, and their interaction as explanatory variables. The population model follows:

$$logit(\mu_i) = \beta_0 + \beta_1 SBP_i + \beta_2 Age_i + \beta_3 SBP_i \times Age_i$$

	exp(Est)	ci95.lo	ci95.hi	z value	Pr(>|z|)
(Intercept)	0.0003	0.0000	0.0032	-6.7690	0e+00
sbp	1.0481	1.0298	1.0668	5.2170	0e+00
age	1.1194	1.0682	1.1731	4.7192	0e+00
sbp:age	0.9993	0.9990	0.9997	-3.8080	1e-04

Note what we are looking for is if 1 is in the confidence interval for the odds ratio. If 1 is in the interval, this is to say we are 95% confident that 1 can be a value for the odds ratio. If the odds ratio is equal to 1, this implies that the covariate in question has no effect on the outcome.

If 1 is not in the interval, this implies that we are 95% confident the odds ratio is not equal to 1. Which is to say the covariate in question does effect the outcome.

Confidence interval review

- We are further interested in linear combination of the coefficients.

- Example: Confidence interval for $\beta 1 + 15\beta 3$ or $\beta 0 + 20\beta 1$.

- The linContr.glm() function will create 95% confidence intervals for all the odds ratios in the model that use a combination of the coefficients.

- Function has 3 inputs.
  - contr.names = The names of the coefficients to be used in the contrast.
  - contr.coef = The coefficients to use in forming the linear combination (the covariate values).
  - model = The name of the model to pull the estimates from.