Categorical: Logistic regression 2

Table of contents

1 Goals

1.1 Goals

1.1.1 Goals of this lecture

- More details about logistic regression
  - Dealing with coefficients and confidence intervals

  - Comparing models
  - R^2 measures

- Probit regression: Related model for binary outcomes

2 Review: Logistic regression

2.1 Logistic regression

2.1.1 Logistic regression

- Generalized linear model (GLiM) for **binary** (0,1) outcomes
  - Outcome has a *binomial* distribution
  - Link function is $logit = ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$
  - *Three metrics* to interpret: Probability, odds, logit
  - *Nonlinear* effects for probability and odds

3 Effects and confidence intervals

3.1 Confidence intervals

3.1.1 Review: Three forms of logistic regression

Probability:

$$\hat{p} = \frac{e^{(b_0+b_1X_1+b_2X_2+\cdots+b_pX_p)}}{1+e^{(b_0+b_1X_1+b_2X_2+\cdots+b_pX_p)}}$$

Odds:

$$\hat{odds} = \frac{\hat{p}}{1-\hat{p}} = e^{b_0+b_1X_1+b_2X_2+\cdots+b_pX_p}$$

Logit:

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p$$

3.1.2 So...

- Which metric are confidence intervals in?
  - And can we have confidence intervals in multiple metrics?
- This matters because "no effect" differs across the metrics
  - Odds: No effect = 1
  - Logit: No effect = 0

3.1.3 Output from logistic regression

```
Call:
glm(formula = Acceptance ~ GPAc, family = binomial(link = "logit"),
    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)   0.1736     0.3253   0.534 0.593488
GPAc          5.4542     1.5792   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
```

3.1.4 Confidence intervals for coefficients

	b	LowerCL	UpperCL
(Intercept)	0.174	-0.469	0.823
GPAc	5.454	2.696	8.966

- $\widehat{logit} = ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 0.174 + 5.454(GPAc)$

- Significant: Confidence interval doesn't include 0
  - Logit: No effect = 0

3.1.5 Confidence intervals for exp(coefficients)

	exp.b.	exp.LowerCL.	exp.UpperCL.
(Intercept)	1.19	0.625	2.278
GPAc	233.73	14.825	7829.246

- Each value is e raised to its power: e.g., $e^{b_0} = e^{0.174} = 1.19$
- $\widehat{odds} = \frac{\hat{p}}{1-\hat{p}} = e^{0.174+5.454(GPAc)}$
- Significant: Confidence interval doesn't include 1
  - Odds: No effect = 1

4 Model comparison

4.1 Model comparison

4.1.1 Why do we need model comparisons?

- **General**: Difference involves *more than one predictor*
  - X_1 versus X_1, X_2, interaction of X_1 and X_2
  - Adding several related predictors together
    * Several subscales of a larger measure
    * A set of several covariates
- **Logistic regression-specific**: Test of regression coefficients may not be reliable (Vaeth, 1985)
  - Especially when the coefficients are *small*

4.1.2 Testing effects in linear regression

- If you added a predictor, there were *two ways* to tell if that predictor was adding to the model:
  - Test of the regression coefficient (i.e., t-test)
  - R^2 for added prediction (i.e., R^2_{change})
    - F-test of R^2_{change}
    - R^2_{change} can also be used to test multiple predictors

4.1.3 Testing effects in logistic regression

- **Wald tests** of the regression coefficient may not be reliable
  - Wald tests: Estimate / standard error
  - z-tests, t-tests
- Analogue of the significance test for R^2_{change}
  - *Likelihood ratio test* (LR test)

4.1.4 A quick note on models

- Continuum of models, from the *worst model* to the *best model*
- Worst model
  - Null model
  - Intercept only model
  - 0 predictors
- Your model
  - The model you're considering
- Best model
  - Saturated model
  - Perfect model
  - n predictors

4.1.5 Likelihood ratio (LR) test

- **Likelihood** (i.e., from *maximum likelihood* estimation) of the model
  - Specifically, the **deviance**, which is $-2 \times log(likelihood)$
- What is **deviance**?
  - How far the model is from the *perfect model*
  - Kind of like $SS_{residual}$
- **Difference** in deviance between two models has a χ^2 distribution
  - How did we get from **ratio** to **difference**?
    * $log(x/y) = log(x) - log(y)$

4.1.6 Likelihood ratio (LR) test

$$\chi^2 = deviance_{model1} - deviance_{model2}$$

- Model 1: simpler model (fewer predictors, worse fit)
- Model 2: more complex model (more predictors, better fit)
- **Degrees of freedom** = difference in number of parameters
  - **Significant test**: Model 1 is significantly worse than Model 2
  - **NS test**: Model 1 and 2 are not significantly different, so go with simpler one (Model 1)

4.1.7 LR test example

- Model 1: `GPAc` predicts `Acceptance`
  - Residual deviance = 56.839
- Model 2: `GPAc` and `MCAT` predict `Acceptance`
  - Residual deviance = 54.014
- Degrees of freedom = 1
  - 1 additional thing estimated in model 2
- Smaller deviance means closer to perfect model
  - But is it *significantly* closer?

4.1.8 LR test example

- $\chi^2(1) = 56.839 - 54.014 = 2.825$
  - Critical value for $\chi^2(1) = 3.841$
  - $2.825 < 3.86$
    - NS test
    - Model 1 and model 2 are not significantly different, so go with the simpler one
    - Use the model with just `GPAc` as a predictor

4.1.9 Nested models

- LR tests are only appropriate for **nested models**
  - Simpler model is contained in the more complex model
- Comparing two models with different sets of predictors
  - AIC or BIC
  - Lower is better

5 (Pseudo) R^2 measures

5.1 (Pseudo) R^2 measures

5.1.1 What is R-squared again?

- In *linear regression*, R^2 is **ALL** of these things
  - *Explanatory power* of the model
  - *Proportion* of variance in outcome explained by predictor(s)
  - *Squared correlation* between observed outcome (Y) and predicted outcome ($\hat{Y}$)
  - *Always* between 0 and 1 (proportion)
  - *Always* increases or stays the same with more predictors
  - Based on the sums of squares

5.1.2 What about for logistic regression?

- In *logistic regression*, R^2 could be
  - *Explanatory power* of the model: **Still true because it's vague**
  - *Proportion* of variance in outcome explained by predictor(s): **Kind of, but not always mathematically true**

- *Squared correlation* between observed outcome (Y) and predicted outcome ($\hat{Y}$): **This is one approach**
- ~~*Always* between 0 and 1 (proportion)~~
- ~~*Always* increases or stays the same with more predictors~~
- ~~Based on the sums of squares~~

5.1.3 Many measures that try to approximate R^2

- **Why** so many measures?
  - Logistic regression is estimated via *maximum likelihood*
    - No sums of squares
  - *Heteroskedasticity* (in probability and odds metrics)
    - But not in logit metric…
  - **Base rate** influences how all these things work
    - Overall proportion of "events" in the sample

5.1.4 Many measures that try to approximate R^2

- Here are some of the best-behaved and most commonly-used
  - Squared correlation between predicted and observed values
  - Tjur
  - McFadden
    - Adjusted McFadden
  - Cox-Snell
    - Adjusted version: Nagelkerke
  - McKelvey Zavoina

5.1.5 Correlation between predicted and observed values

- Literally, the correlation between the *observed* (Y) values and the *predicted* ($\hat{Y}$) values
  - In the example, this value is 0.297
- Mathematically the same as Efron's R^2

5.1.6 Tjur's R^2

- $R^2_{Tjur} = |mean(\hat{p}|Y=0) - mean(\hat{p}|Y=1)|$
  - Average predicted probability of "not events" minus average predicted probability of "events"
  - Then take the absolute value
- In the example, this value is 0.302

5.1.7 Estimation

- GLiMs are estimated using **maximum likelihood**
  - Likelihood (L)
  - Log-likelihood (LL) = $ln(likelihood)$
  - Deviance = $-2 \times LL$
- For these R^2 measures, we'll often compare the deviance for *our model* to the deviance for a **null model**
  - This is the *worst* possible model
  - *No predictors*

5.1.8 McFadden R^2 (a.k.a. Likelihood ratio R^2, pseudo R^2)

- $R^2_{McFadden} = 1 - \frac{LL_{model}}{LL_{null}}$
- In the example, this value is 0.25
- Proportion of variance accounted for
  - Proportion of the way from null model to perfect model
- Adjusted version divides by the maximum possible value

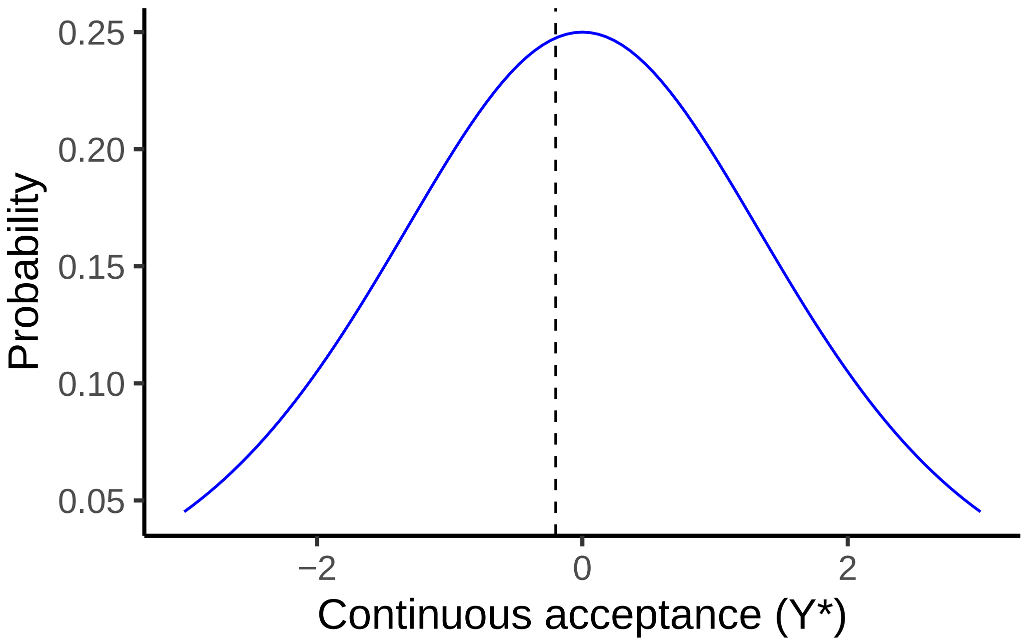
5.1.9 Cox-Snell R^2 (a.k.a. Generalized R^2)

- $R^2_{Cox-Snell} = 1 - \left(\frac{L_{null}}{L_{model}}\right)^{2/N}$
- In the example, this value is 0.291
- Very influenced by **base rate**
  - Adjusted version divides by the maximum possible value
  - Nagelkerke

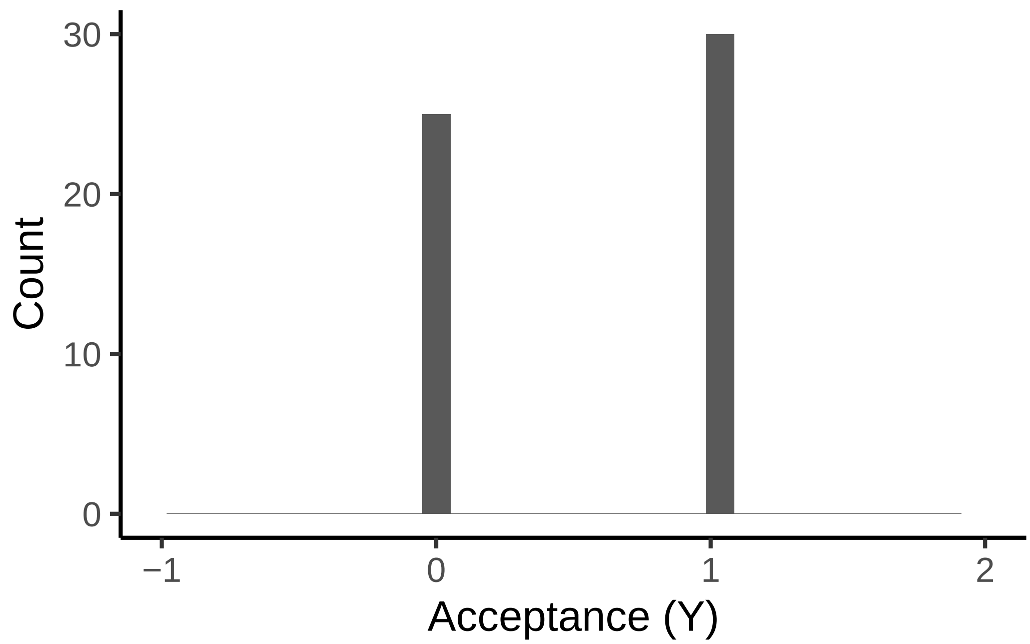
5.1.10 A quick note: Latent variable interpretation

- We've talked about logistic regression in terms of **binary outcome**
- *Binary manifestation* of a continuous variable
  - Agree / Disagree
    - Continuum of agreement: Switch from disagree to agree
  - Medical diagnosis (e.g., hypertension)
    - Continuous measure of blood pressure: ≥ 120 cutoff
- *Continuous latent variable* underlying the *binary observed one*

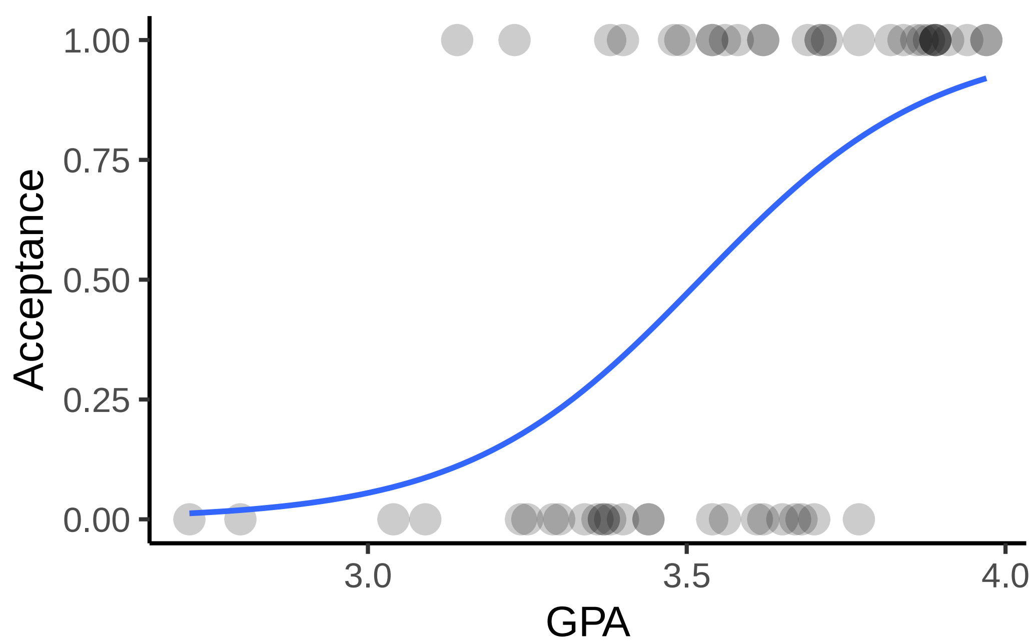
5.1.11 Figure: Latent variable underlying the binary



5.1.12 Figure: Latent variable underlying the binary



5.1.13 Figure: Acceptance vs GPA



5.1.14 Latent variable

$$Y^* = b_0 + b_1 X_1 + e$$

- b_0 here has a *different interpretation*
  - *Threshold* or cut point, not intercept: Location of vertical line
  - Common in SEM for binary or ordered category outcom
- How does this help us?
  - *Linear model* with *constant variance*
  - $e \sim \text{logistic}(0, \pi^2/3)$

5.1.15 McKelvey-Zavoina R^2

- $R^2_{MZ} = \frac{\sigma^2_{\hat{Y}}}{\sigma^2_{\hat{Y}} + \frac{\pi^2}{3}}$
  - Where $\sigma^2_{\hat{Y}}$ is the **variance of the predicted logit scores**
  - and $\pi^2/3$ is the **residual variance**
  - Interpreted as *proportion* of variance accounted for
- In the example, this value is 0.426

5.1.16 Which one to use?

- Some are recommended over others
  - McFadden (a.k.a. Likelihood ratio R^2, Pseudo R^2) = 0.25
    - Relatively invariant to *base rate*
    - Also usable for other GLiMs
      - Anything with a likelihood
  - McKelvey-Zavoina = 0.426
    - Only applicable to logistic regression because it involves the residual variance for the logistic distribution of the errors

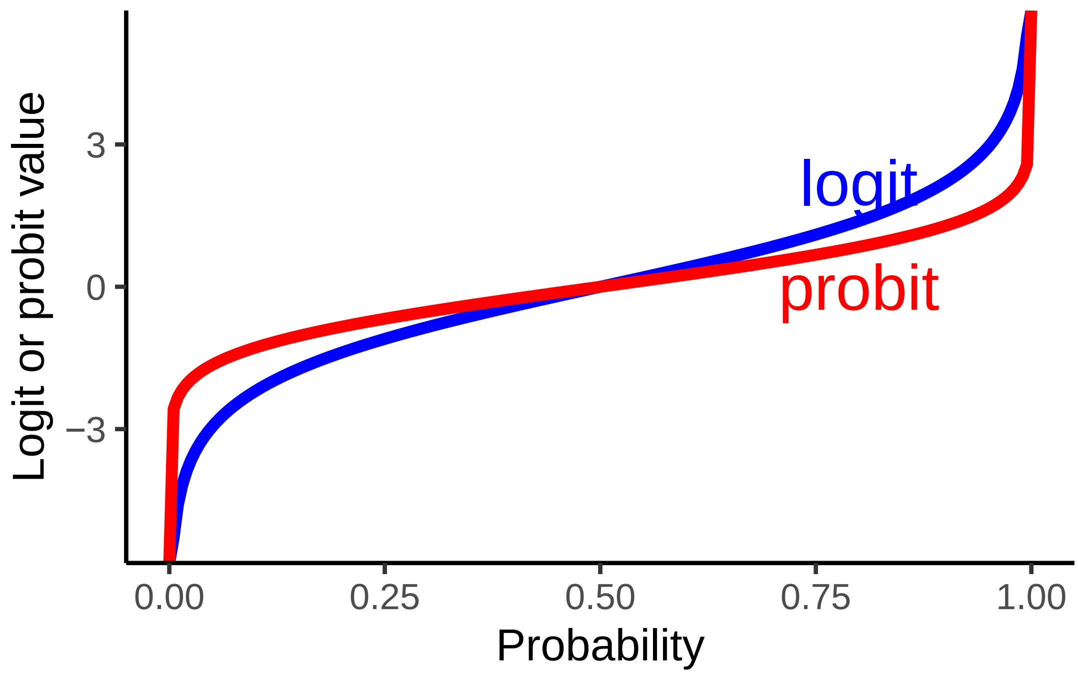
6 A slight variation: Probit

6.1 Probit regression

6.1.1 Probit regression

- Probit regression is an alternative to logistic regression
  - **Probit link function**
  - Binomial distribution
- "Probit" is the inverse normal distribution
  - Give it a probability, it returns the z-score for that probability
    * $probit(0.025) = -1.96$
    * $probit(0.5) = 0$
    * $probit(0.975) = 1.96$

6.1.2 Figure: Probit vs logit



6.1.3 Probit regression

- Often used in biological sciences

    - Sometimes used in SEM

- Gives results that are *very similar* to logistic regression
- Based on the *normal distribution*, which is nice
    - But you lose the odds ratio interpretation from logistic
    - Must use probability metric

7 Summary

7.1 Summary

7.1.1 Summary of this week

- Logistic regression
    - Confidence intervals
        * In different metrics
    - Comparing models
        * e.g., model with `GPA` vs model with `GPA` and `MCAT`
    - Pseudo-R^2 measures
        * Many options

7.1.2 Next week

- Extending this model to 3 or more outcome categories
    - Ordinal logistic regression for **ordered** categories
    - Multinomial logistic regression for **unordered** categories

7.1.3 Next few weeks

- Models for count outcomes
    - Poisson regression
    - Overdispersed Poisson regression
    - Negative binomial regression
    - Excess zeroes versions of these models