Checking Logistic Regression Models

Grinnell College

October 28, 2026

Model Selection and Model Fit

On Monday, we compared possible logistic regression models.

- What if the models we considered all miss something?
- How can we find where a model fits poorly?
- Does an unusual observation change our conclusions?

A useful model should describe the data adequately as well as compare favorably with alternatives.

Deviance

The **saturated model** fits the recorded responses exactly.

$$D_{\text{residual}} = 2\big(\ell_{\text{saturated}} - \ell_{\text{fitted}}\big),$$

$$D_{\text{null}} = 2\big(\ell_{\text{saturated}} - \ell_{\text{intercept-only}}\big)$$

- Smaller residual deviance means a closer fit to the saturated model.
- Deviance decreases as we add terms to a nested model.

How much improvement would justify adding more terms?

Marijuana Use

Students at one high school near Dayton reported whether they had *ever* used marijuana.

```
> Marijuana
   race gender yes  no
1 white female 420 620
2 white   male 483 579
3 other female  25  55
4 other   male  32  62
```

Each row is a binomial count at one combination of race and gender. "Other" is the dataset's race category.

An Additive Logistic Model

$$\text{logit}(\pi) = \alpha + \beta_1 \,\text{gender} + \beta_2 \,\text{race}.$$

```
> fit <- glm(yes/(yes + no) ~ gender + race,
  weights = yes + no, family = binomial, data = Marijuana)
>
> round(coef(summary(fit)), 4)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.8303     0.1685 -4.9266   0.0000
gendermale     0.2026     0.0852  2.3785   0.0174
racewhite      0.4437     0.1677  2.6466   0.0081
```

What restriction does this model impose on the gender odds ratio?

Goodness of Fit in R

```
> summary(fit)

Call:
glm(formula = yes/(yes + no) ~ gender + race, family =
    binomial,
    data = Marijuana, weights = yes + no)

Coefficients:
            Estimate Std. Error z value   Pr(>|z|)
(Intercept)  -0.8303     0.1685   -4.93 0.00000084 ***
gendermale    0.2026     0.0852    2.38     0.0174 *
racewhite     0.4437     0.1677    2.65     0.0081 **
---

    Null deviance: 12.752784  on 3  degrees of freedom
Residual deviance:  0.057982  on 1  degrees of freedom
AIC: 30.41
```

- No evidence that the additive model fits poorly.
- This does not establish that the model is exactly correct.

Observed and Fitted Counts

With grouped data in particular, we are able to compare observed and fitted values in a meaningful way

$$\widehat{\text{successes}}_i = n_i\hat{\pi}_i, \qquad \widehat{\text{failures}}_i = n_i(1-\hat{\pi}_i).$$

```
> fitted_counts
   race gender yes fitted_yes  no fitted_no
1 white female 420    420.714 620   619.286
2 white   male 483    482.286 579   579.714
3 other female  25     24.286  55    55.714
4 other   male  32     32.714  62    61.286
```

All eight fitted counts exceed about 5, supporting the chi-square approximation here.

The Omitted Interaction

With two binary predictors, adding the interaction saturates the four-group model.

```
> anova(fit, fit_interaction, test = "LRT")
Analysis of Deviance Table
Model 1: yes/(yes + no) ~ gender + race
Model 2: yes/(yes + no) ~ gender + race + gender:race
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         1      0.058
2         0      0.000  1    0.058     0.81
```

This tests whether the gender odds ratio varies by race. It does **not** test whether the main effects are zero.

Grouped and Ungrouped Data

The same subjects can be recorded in two ways.

	Grouped	Individual
Rows	4 groups	2,276 people
Response	Yes/no counts	0 or 1
Trials per row	n_i	1
Fitted probabilities	$\hat{\pi}_i$	Same within each group

Do you expect the coefficients to change? What about the residual deviance?

The Same Estimated Effects

After expanding the counts into individual responses:

```
> coef_compare
             grouped individual
(Intercept) -0.8303    -0.8303
gendermale   0.2026     0.2026
racewhite    0.4437     0.4437
```

The response format changes; the additive model still assigns one probability to each race–gender combination.

Different Residual Deviances

```
> deviance_compare
                deviance   df
grouped         0.057982    1
individual 3086.598201 2273
```

- Saturated model for group: one probability for each of four groups.
- Saturated model for individuals: one probability for each person's outcome.

The fitted model is the same, but the saturated benchmark changes. Specifically, the individual saturated model is going to have a much stronger fit, hence greater distance from the considered model

What Can the Saturated Model Fit?

Consider two people with identical predictors: one says yes, the other no.

- Grouped saturated model: $\hat{\pi} = 1/2$ for the pair.
- Individual saturated model: $\hat{\pi}_1 = 1$ and $\hat{\pi}_2 = 0$.

On the individual-data likelihood scale,

$$L_{\text{common } p} = p(1-p) \leq \tfrac{1}{4}, \qquad L_{\text{individual saturation}} = 1.$$

The individual saturated model can fit which person said yes, rather than just how many said yes. Binomial combinatorial factors cancel in each deviance ratio.

Nested Model Comparisons Still Agree

Subtracting deviances cancels the saturated-model benchmark:

$$D_{\text{smaller}} - D_{\text{larger}} = 2(\ell_{\text{larger}} - \ell_{\text{smaller}}).$$

```
> c(grouped = deviance(fit) - deviance(fit_interaction),
    individual = deviance(fit_people) -
+      deviance(fit_people_interaction))
   grouped individual
  0.057982   0.057982
```

The interaction test agrees in either representation. The usual likelihood-ratio theory applies under its regularity conditions.

Continuous Predictors and Individual Responses

For individual Bernoulli data, do **not** automatically compare residual deviance with residual degrees of freedom using a chi-square distribution.

- Each recorded response is only 0 or 1.
- More people do not produce large binomial counts at each distinct predictor setting.
- Compare specific larger models to investigate possible lack of fit: a quadratic term, an interaction, or another relevant predictor.

Judge the practical change in predictions as well as the test result.

Graduate Admissions at University of Florida

Model admission probability by department, with no gender effect:

$$\text{logit}(\pi_{ik}) = \alpha + \beta_k.$$

```
> summary(adm_fit)
Call:
glm(formula = cbind(yes, no) ~ department, family = binomial,
    data = Admissions)
Coefficients:
                     Estimate Std. Error z value     Pr(>|z|)
(Intercept)           -0.8337     0.1645   -5.07 0.0000004023
departmentastronomy    0.9515     0.5130    1.85      0.06363
.
.
.
departmentzoology     -1.2808     0.3273   -3.91 0.0000909872
---

    Null deviance: 449.830  on 45  degrees of freedom
Residual deviance:  44.735  on 23  degrees of freedom
AIC: 241.4
```

Residuals Locate Lack of Fit

For group i, the observed and fitted success counts are y_i and $n_i\hat{\pi}_i$.

$$\text{Standardized residual}_i = \frac{y_i - n_i\hat{\pi}_i}{\widehat{\text{SE}}(Y_i - n_i\hat{\pi}_i)}.$$

- Positive: more successes than predicted.
- Negative: fewer successes than predicted.
- Magnitudes around 2 or 3 suggest groups worth examining.

These are investigation prompts: sparse counts and examining many groups complicate their interpretation.

Admission Data

See text

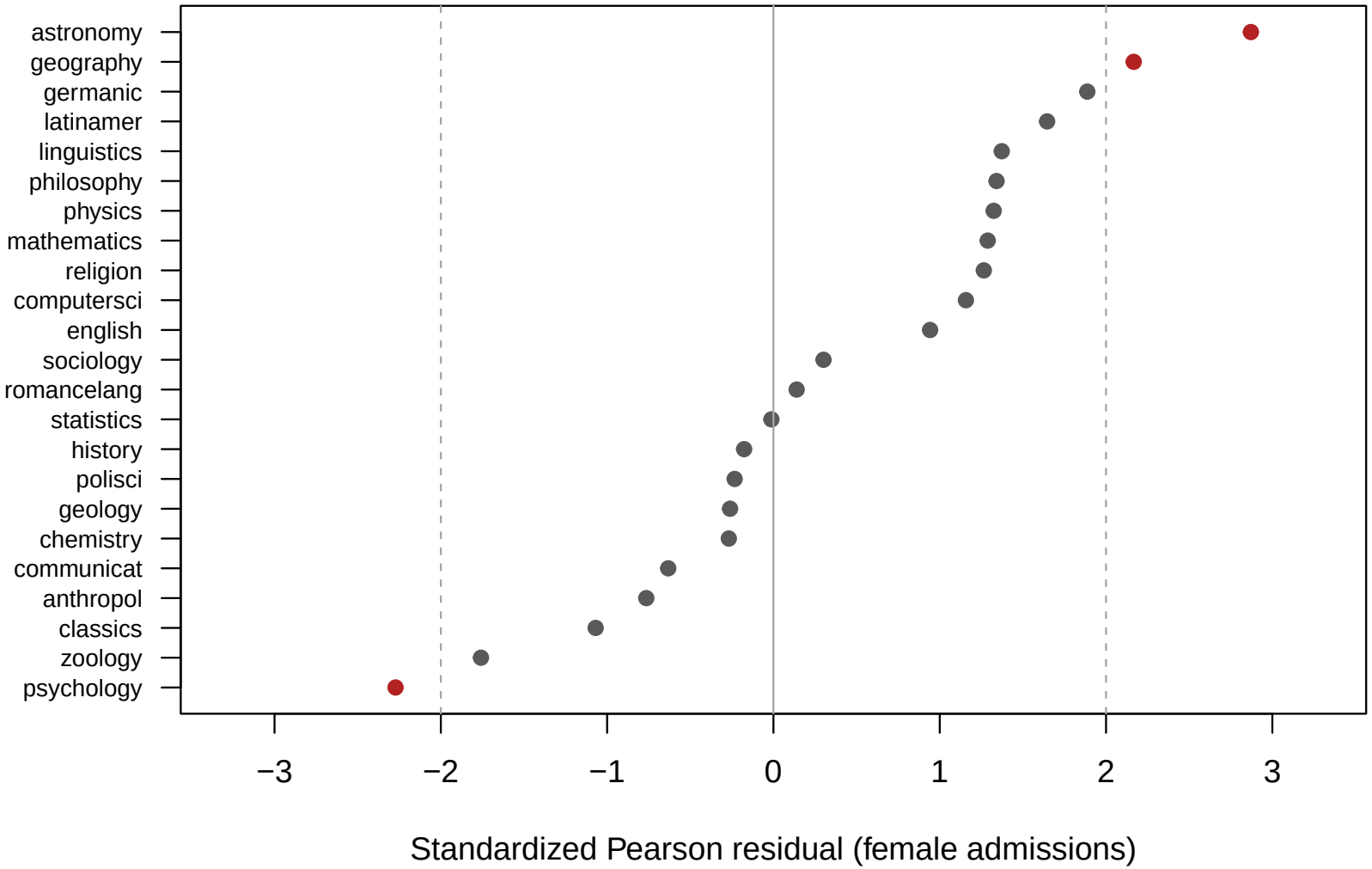
Admissions Residuals in R

```
> Admissions$std_res <- rstandard(adm_dept, type = "pearson")
> subset(Admissions, gender == "female" & abs(std_res) >
+     2, select = c(department, yes, no, std_res))
   department yes  no std_res
4   astronomy   6   0  2.8710
16  geography   9   1  2.1664
36 psychology   2 123 -2.2722
```

These are residuals for female admissions under the department-only model.

Which departments differ, and in which direction?

Admissions Residuals by Department



The model misses different departments in different directions.

Adding a Common Gender Effect

```
> adm_gender <- update(adm_dept, . ~ . + gender)
> anova(adm_dept, adm_gender, test = "LRT")
Analysis of Deviance Table
Model 1: cbind(yes, no) ~ department
Model 2: cbind(yes, no) ~ department + gender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        23       44.7
2        22       42.4  1     2.38     0.12
```

A common gender effect provides little improvement. It cannot describe effects that vary across departments. Some admission counts are sparse, so treat the global chi-square goodness-of-fit approximation cautiously. These comparisons alone do not establish a cause for the differences.

Residuals and Influence

- A **residual** describes a discrepancy between an observed and fitted response.
- **Influence** describes how much an observation changes fitted coefficients or predictions.

Refitting without a flagged group can help assess sensitivity.

An unusual group may have little influence. A flag is a reason to investigate, rather than an instruction to remove data.

Heart Disease and Blood Pressure

Eight blood-pressure groups; response is heart disease during six years of follow-up among men aged 40–59.

Blood Pressure	No	Yes
111.50	156	3
121.50	252	17
131.50	284	12
141.50	271	16
151.50	139	12
161.50	85	8
176.50	99	16
191.50	43	8

These are grouped binomial counts, despite using numerical blood-pressure scores

Heart Disease and Blood Pressure

```
> summary(heart_fit)
Call:
glm(formula = y / (n+y) ~ bp, family = binomial, weights =
    n+y, data = HeartBP)

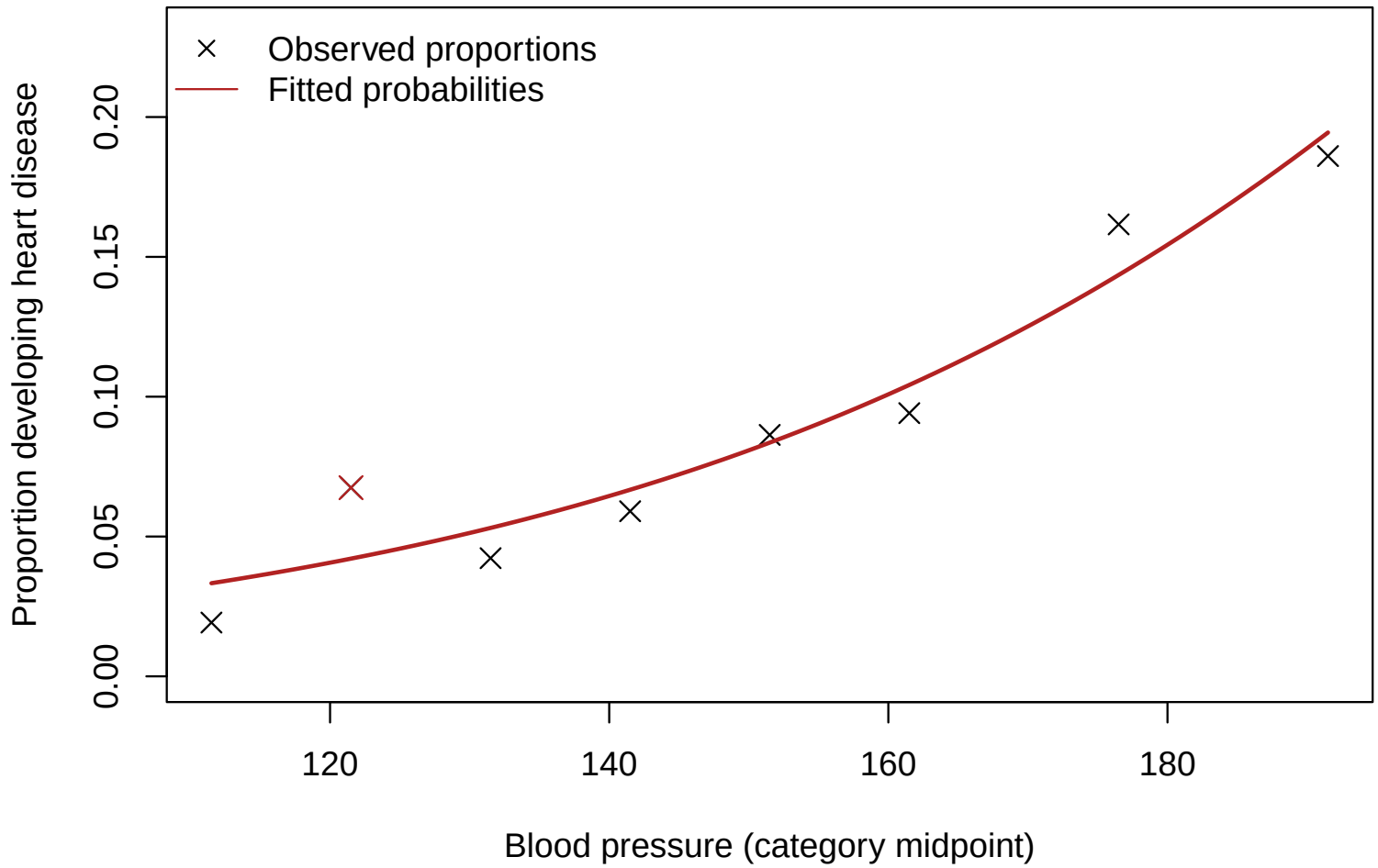
Coefficients:
            Estimate Std. Error z value             Pr(>|z|)
(Intercept) -6.08203    0.72432   -8.40 < 0.000000000000002
bp           0.02434    0.00484    5.03            0.0000005

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 30.0226  on 7  degrees of freedom
Residual deviance:  5.9092  on 6  degrees of freedom
AIC: 42.61
```

Diagnostic Measures

See text

Observed Proportions and Fitted Probabilities



The second group stands out, while the overall pattern is reasonably described by the fitted curve.

Sensitivity to One Group

```
> heart_without2 <- update(heart_fit, data = HeartBP[-2,
+     ])
> rbind(all_groups = coef(heart_fit), without_group2 =
    coef(heart_without2))
               (Intercept)        bp
all_groups         -6.0820 0.024338
without_group2     -7.0841 0.030309
```

- The estimated blood-pressure slope increases when group 2 is omitted.
- The fitted association remains positive in both fits.
- This removes an entire binomial group, not one person.

Review

1. What does the marijuana residual-deviance test tell us about the interaction?
2. Why does expanding counts into individual responses change residual deviance?
3. Why might adding one gender coefficient fail to address the admissions lack of fit?
4. How does a large residual differ from an influential observation?