THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Logistic regression 2

Lecture 17

STA 371G

Team evaluations for Project Part 2

- If you forgot to submit your team evaluation for Project Part 2, you can still submit it anytime today, up until 11:59 PM!

Midterm 2!

- Midterm 2 will be held in class on Wednesday, April 10.

Midterm 2!

- Midterm 2 will be held in class on Wednesday, April 10.
- It covers material through Lecture 18 (Wednesday, April 3).

Midterm 2!

- Midterm 2 will be held in class on Wednesday, April 10.
- It covers material through Lecture 18 (Wednesday, April 3).
- You can bring two pages of notes with you (front and back).

Midterm 2!

- Midterm 2 will be held in class on Wednesday, April 10.
- It covers material through Lecture 18 (Wednesday, April 3).
- You can bring two pages of notes with you (front and back).
- The TAs will hold an optional review session in GSB 3.130 on Sunday, April 7 at 6 PM.

Midterm 2!

- Midterm 2 will be held in class on Wednesday, April 10.
- It covers material through Lecture 18 (Wednesday, April 3).
- You can bring two pages of notes with you (front and back).
- The TAs will hold an optional review session in GSB 3.130 on Sunday, April 7 at 6 PM.
- Don't forget about the weekly TA help sessions on Tuesday nights at 6 PM (also in GSB 3.130)!

Last time

- The OkCupid data set contains information about 59946 profiles from users of the OkCupid online dating service.
- We predicted sex (as a binary categorical variable) from height using logistic regression, and came up with the prediction equation:

$$\log \text{odds} = \log\left(\frac{P(\text{male})}{1 - P(\text{male})}\right) = -44.45 + 0.66 \cdot \text{height}.$$

or, solving for $P(\text{male})$,

$$\widehat{P(\text{male})} = \frac{e^{-44.45+0.66\cdot\text{height}}}{1 + e^{-44.45+0.66\cdot\text{height}}}$$

1. Hypothesis testing
2. Evaluating the model
3. Checking assumptions

Testing the null hypothesis

As in regular linear regression, the overall null hypothesis is that $\beta_1 = 0$; we can test this by using the p-value for that variable on the output.

Testing the null hypothesis

As in regular linear regression, the overall null hypothesis is that $\beta_1 = 0$; we can test this by using the *p*-value for that variable on the output.

Since p is very small, we can reject the null hypothesis that $\beta_1 = 0$; i.e., there is a statistically significant relationship between height and sex.

1. Hypothesis testing

2. Evaluating the model

3. Checking assumptions

How good is our model?

- Unfortunately, the typical R^2 metric isn't available for logistic regression.

How good is our model?

- Unfortunately, the typical R^2 metric isn't available for logistic regression.
- However, there are many "pseudo-R^2" metrics that indicate model fit.

How good is our model?

- Unfortunately, the typical R^2 metric isn't available for logistic regression.
- However, there are many "pseudo-R^2" metrics that indicate model fit.
- But: most of these pseudo-R^2 metrics are difficult to interpret, so we'll focus on something simpler to interpret and communicate.

How many cases did we accurately predict?

We could use our model to make a prediction of sex based on the probability. Suppose we say that our prediction is:

$$\text{Prediction} = \begin{cases} \text{male}, & \text{if } \widehat{P(\text{male})} \geq 0.5, \\ \text{female}, & \text{if } \widehat{P(\text{male})} < 0.5. \end{cases}$$

How many cases did we accurately predict?

We could use our model to make a prediction of sex based on the probability. Suppose we say that our prediction is:

$$\text{Prediction} = \begin{cases} \text{male}, & \text{if } \widehat{P(\text{male})} \geq 0.5, \\ \text{female}, & \text{if } \widehat{P(\text{male})} < 0.5. \end{cases}$$

Now we can compute the fraction of people whose sex we correctly predicted:

```
predicted.male <- (predict(model, type="response") >= 0.5)
actual.male <- (my.profiles$male == 1)
sum(predicted.male == actual.male) / nrow(my.profiles)

[1] 0.83
```

How many cases did we accurately predict?

83% sounds pretty good—what should we compare it against?

How many cases did we accurately predict?

83% sounds pretty good—what should we compare it against?

We should compare 83% against what we would have gotten if we just predicted the most common outcome (male) for everyone, without using any other information:

How many cases did we accurately predict?

83% sounds pretty good—what should we compare it against?

We should compare 83% against what we would have gotten if we just predicted the most common outcome (male) for everyone, without using any other information:

```
sum(actual.male) / nrow(my.profiles)

[1] 0.6
```

How many cases did we accurately predict?

83% sounds pretty good—what should we compare it against?

We should compare 83% against what we would have gotten if we just predicted the most common outcome (male) for everyone, without using any other information:

```
sum(actual.male) / nrow(my.profiles)

[1] 0.6
```

In other words, our model provided a “lift” in accuracy from 60% to 83%.

The confusion matrix

Sometimes it is useful to understand what kinds of errors our model is making:

- True positives: predicting male for someone that is male
- True negatives: predicting female for someone that is female
- False positives: predicting male for someone that is female
- False negatives: predicting female for someone that is male

(If we had designated female as 1 and male as 0, these would have switched!)

The confusion matrix

```
table(predicted.male, actual.male)

              actual.male
predicted.male FALSE  TRUE
         FALSE 19466  5494
         TRUE   4623 30243

prop.table(table(predicted.male, actual.male), 2)

              actual.male
predicted.male FALSE TRUE
         FALSE  0.81 0.15
         TRUE   0.19 0.85
```

1. Hypothesis testing
2. Evaluating the model
3. Checking assumptions

Checking assumptions

- Independence
- Linearity
- Normality of residuals ✗
- Homoscedasticity / Equal variance ✗

With logistic regression, we don't need to check the last two assumptions (since Y is binary).

Checking assumptions: Independence

Like with linear regression, we check independence by thinking about the data conceptually: are the predictions the model makes likely to be independent from each other?

Checking assumptions: Independence

Like with linear regression, we check independence by thinking about the data conceptually: are the predictions the model makes likely to be independent from each other?

✓ Yes! Each case is a completely different person whose heights and genders are unrelated.

Checking assumptions: Linearity

Look at the logistic regression model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X + \epsilon$$

We need an approximately linear relationship between the log odds of success and X, or, equivalently, a linear relationship between the log odds of success and what is predicted from our linear model on the right side of the equation.

Checking assumptions: Linearity

To do this, we segment the predicted log odds into groups by deciles (bottom 10%, next 10%, up until the highest 10%):

```
quantile(predict(model), probs=seq(0, 1, 0.1))

   0%    10%    20%    30%    40%    50%    60%    70%
-8.04 -2.75 -1.42 -0.76 -0.10  0.56  1.88  2.55
  80%    90%   100%
 3.21  3.87  8.50
```

Checking assumptions: linearity

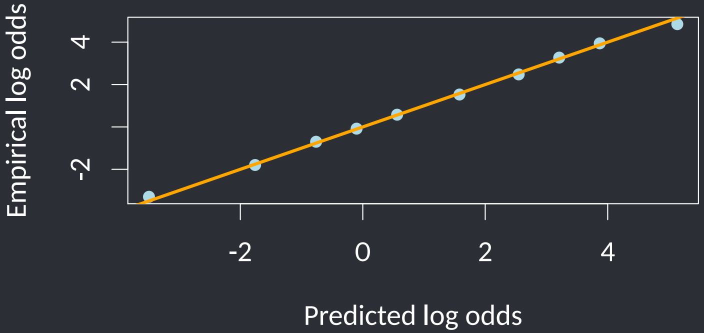
Then we'll calculate the empirical log odds within each group:

Predicted log odds	# males	Total	$p = P(\text{male})$	Log odds
$[-8.04, -2.75]$	256	7182	0.04	-3.3
$[-2.75, -1.42]$	1090	7659	0.14	-1.8
$[-1.42, -0.76]$	1579	4759	0.33	-0.7
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[3.87, 8.5]$	5168	5208	0.99	4.85

Then we'll plot the empirical log odds against the mean of each decile; we'd like to see approximately the line $y = x$; this is called an empirical logit plot.

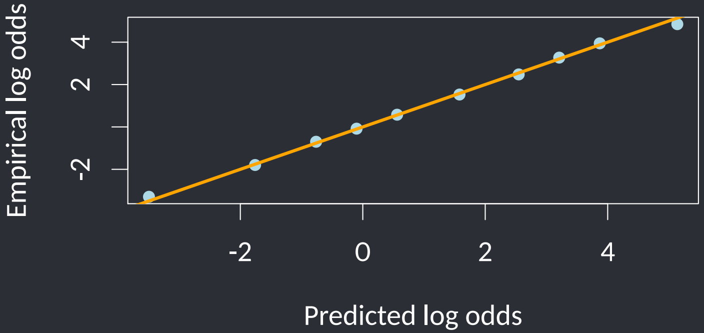
Checking assumptions: Linearity

```
empirical.logit.plot(model)
```



Checking assumptions: Linearity

```
empirical.logit.plot(model)
```



✓ Yes! This is approximately along the line $y = x$.