THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Logistic Regression 2

Lecture 20

STA 371G

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:

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

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. Evaluating the model
2. Checking assumptions
3. Logistic regression with 2+ predictors
4. Interactions in logistic regression
5. Other applications of logistic regression

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:

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

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:

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

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. Evaluating the model
2. Checking assumptions
3. Logistic regression with 2+ predictors
4. Interactions in logistic regression
5. Other applications of logistic regression

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

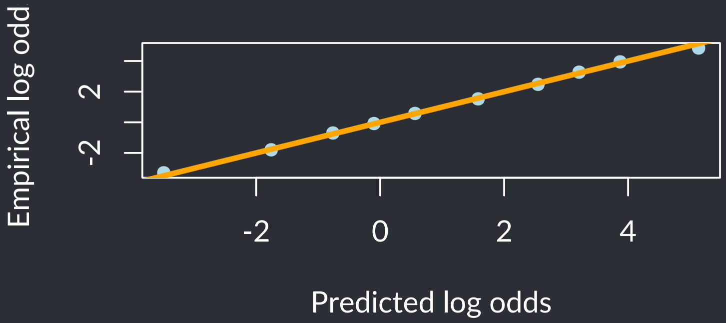
Then we'll calculate the empirical log odds within each group:

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.

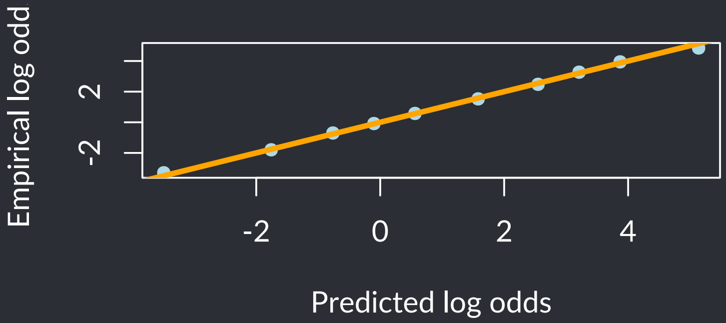
Checking assumptions: Linearity

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



Checking assumptions: Linearity

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



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

1. Evaluating the model
2. Checking assumptions
3. Logistic regression with 2+ predictors
4. Interactions in logistic regression
5. Other applications of logistic regression

Adding another predictor

- Just like with a linear regression model, we can add additional predictors to the model.
- Our interpretation of the coefficients in multiple logistic regression is similar to multiple linear regression, in the sense that each coefficient represents the predicted effect of one X on Y, holding the other X variables constant.

Adding another predictor

Let's add sexual orientation as a second predictor of gender, in addition to height:

```
model2 <- glm(male ~ height + orientation,
  data=my.profiles, family=binomial)
```

The `orientation` variable has three categories:

```
table(my.profiles$orientation)


bisexual      gay  straight
    2763     5568     51495
```

```
Call:
glm(formula = male ~ height + orientation, family = binomial,
    data = my.profiles)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-3.620   -0.481   0.198   0.530   4.022

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -46.08076    0.37167  -124.0   <2e-16 ***
height               0.66535    0.00537   124.0   <2e-16 ***
orientationgay       2.09556    0.07209    29.1   <2e-16 ***
orientationstraight  1.39972    0.06068    23.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 80654  on 59825  degrees of freedom
Residual deviance: 43722  on 59822  degrees of freedom
AIC: 43730

Number of Fisher Scoring iterations: 6
```

Interpreting coefficients

Our prediction equation is:

$$\log\left(\frac{p}{1-p}\right) = -46.08 + 0.67 \cdot \text{height} + 2.1 \cdot \text{gay} + 1.4 \cdot \text{straight}.$$

This means that:

- Our predicted log odds of being male for someone who is bisexual and has a height of 0" is -46.08 (the intercept).

Interpreting coefficients

Our prediction equation is:

$$\log\left(\frac{p}{1-p}\right) = -46.08 + 0.67 \cdot \text{height} + 2.1 \cdot \text{gay} + 1.4 \cdot \text{straight}.$$

This means that:

- Our predicted log odds of being male for someone who is bisexual and has a height of 0" is -46.08 (the intercept).
- Among people with the same sexual orientation, each additional inch of height corresponds to an increase in 95% in predicted odds of being male (i.e., multiplied by $e^{0.67} = 1.95$).

Interpreting coefficients

$$\log\left(\frac{p}{1-p}\right) = -46.08 + 0.67 \cdot \text{height} + 2.1 \cdot \text{gay} + 1.4 \cdot \text{straight}.$$

- Among people of the same height, being gay increases the predicted odds of being male by 713% (i.e., multiplied by $e^{2.1} = 8.13$) compared to being bisexual.

Interpreting coefficients

$$\log\left(\frac{p}{1-p}\right) = -46.08 + 0.67 \cdot \text{height} + 2.1 \cdot \text{gay} + 1.4 \cdot \text{straight}.$$

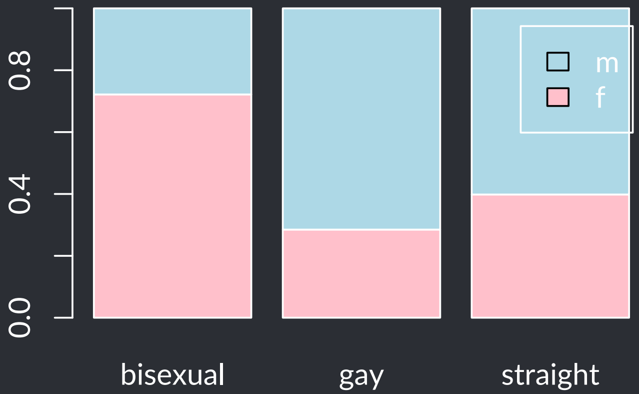
- Among people of the same height, being gay increases the predicted odds of being male by 713% (i.e., multiplied by $e^{2.1} = 8.13$) compared to being bisexual.
- Among people of the same height, being straight increases the predicted odds of being male by 305% (i.e., multiplied by $e^{1.4} = 4.05$) compared to being bisexual.

Understanding what's going on

```
crosstabs <- table(my.profiles$sex, my.profiles$orientation)
crosstabs
```

```
    bisexual   gay straight
  f     1994  1586    20509
  m      769  3982    30986
```

```
barplot(prop.table(crosstabs, 2), col=c("pink", "lightblue"),
  legend=T)
```

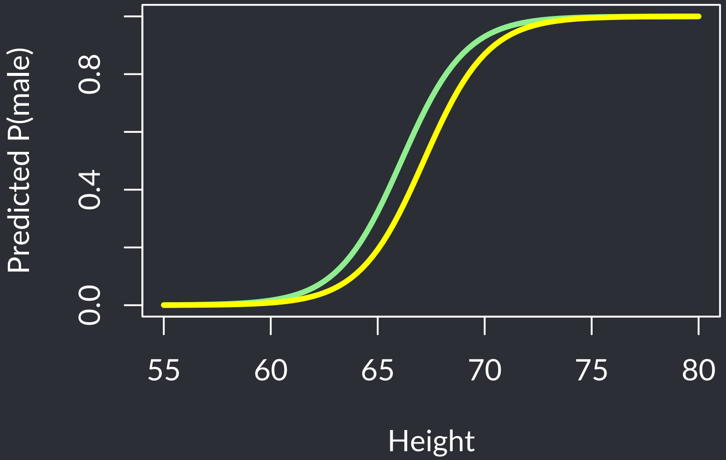


Converting back to probabilities

Because there is a nonlinear relationship between probability and odds, a particular percentage increase in odds does not correspond to a fixed change in probability. But it can be useful sometimes to compute some exemplar predicted probabilities to get a sense of the relationships:

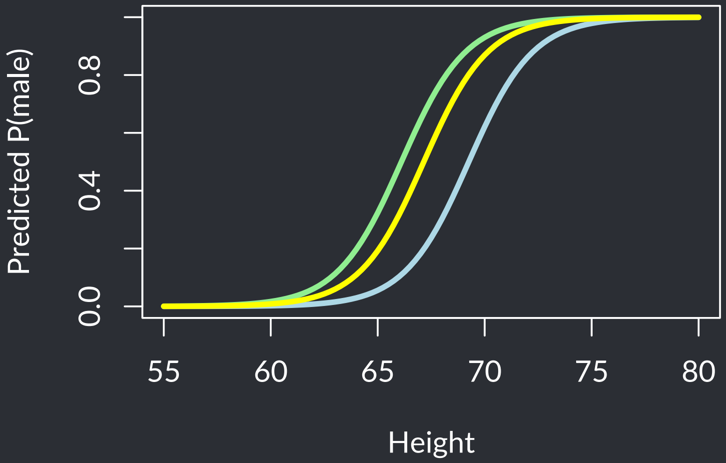
	Height			
	60"	64"	68"	72"
bisexual	0.002	0.029	0.302	0.861
gay	0.017	0.197	0.779	0.981
straight	0.008	0.109	0.637	0.962

We can also visualize this by plotting the three curves for straight (yellow), gay (green), and bisexual (blue) OkCupid users:



Where will the curve for bisexual OkCupid users be?

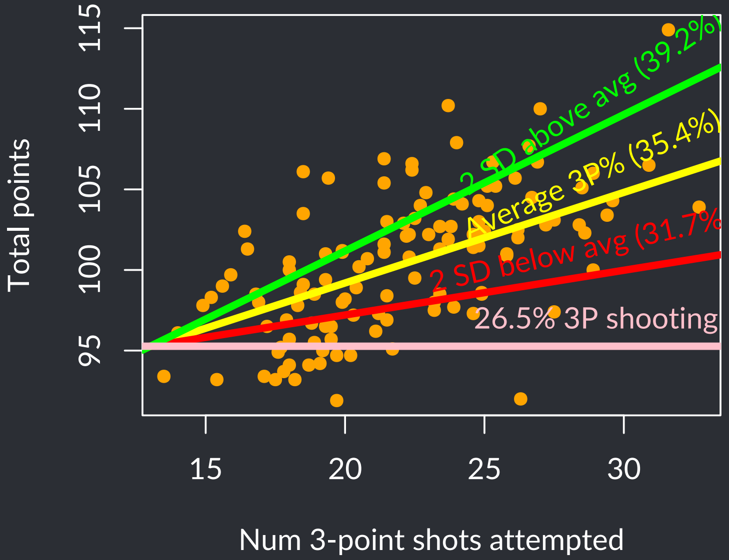
We can also visualize this by plotting the three curves for straight (yellow), gay (green), and bisexual (blue) OkCupid users:



1. Evaluating the model
2. Checking assumptions
3. Logistic regression with 2+ predictors
4. Interactions in logistic regression
5. Other applications of logistic regression

What would interactions do?

- In linear regression, an interaction between two predictors X_1 and X_2 means that the slope of X_1 will depend on the value of X_2.
- In other words, there will be differently-sloped regression lines predicting Y from X_1 depending on what the value of X_2 is.



Total points
95
100
105
110
115
15
20
25
30
Num 3-point shots attempted
2 SD above avg (39.2%)
Average 3P% (35.4%)
2 SD below avg (31.7%
26.5% 3P shooting

What would interactions do?

- We can add interactions to logistic regression and the interpretation is the same: the effect of X_1 on the probability of being male depends on the value of X_2.
- Let's try this out with $X_1 = \text{height}$ and $X_2 = \text{orientation}$.

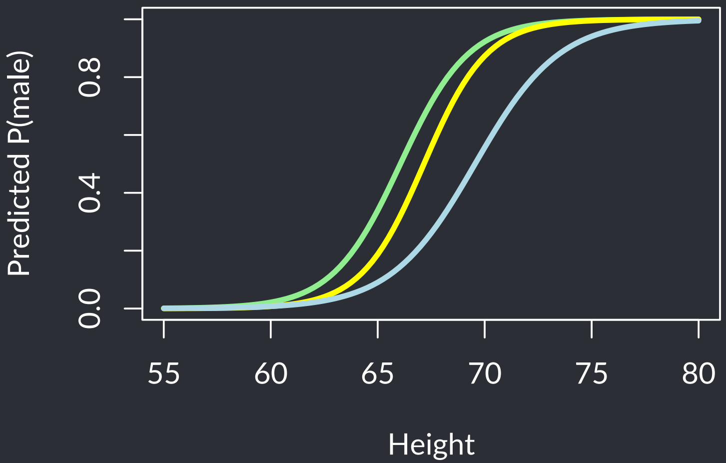
```r
int.model <- glm(male ~ height * orientation, data=my.profiles, family=binomial)
summary(int.model)
```

```

Call:
glm(formula = male ~ height * orientation, family = binomial,
    data = my.profiles)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-3.655  -0.470   0.194   0.521   4.064

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -35.3027     1.4050  -25.13  < 2e-16 ***
height                       0.5076     0.0206   24.67  < 2e-16 ***
orientationgay              -6.2727     1.8365   -3.42  0.00064 ***
orientationstraight        -10.2887     1.4596   -7.05  1.8e-12 ***
height:orientationgay        0.1218     0.0271    4.49  7.1e-06 ***
height:orientationstraight   0.1712     0.0214    8.01  1.2e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

The interaction model is:

$$\log\left(\frac{p}{1-p}\right) = -35.3 + 0.51 \cdot \text{height} - 6.27 \cdot \text{gay} - 10.29 \cdot \text{straight}$$
$$+ 0.12 \cdot \text{height} \cdot \text{gay} + 0.17 \cdot \text{height} \cdot \text{straight}.$$

Let's graph the equation for gay (green), straight (yellow), and bisexual (blue) users:



1. Evaluating the model
2. Checking assumptions
3. Logistic regression with 2+ predictors
4. Interactions in logistic regression
5. Other applications of logistic regression

What else can we use logistic regression for?

- **Finance:** Predicting which customers are most likely to default on a loan
- **Advertising:** Predicting when a customer will respond positively to an advertising campaign
- **Marketing:** Predicting when a customer will purchase a product or sign up for a service