THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Logistic regression 3

Lecture 18

STA 371G

Reminder

- Remember that Part 3 of the project is due on Friday at 11:59 PM!

1. Logistic regression with 2+ predictors
2. Interactions in logistic regression
3. Hypothesis testing when there are 2+ predictors
4. Other applications of logistic regression

1. Logistic regression with 2+ predictors
2. Interactions in logistic regression
3. Hypothesis testing when there are 2+ predictors
4. 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

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

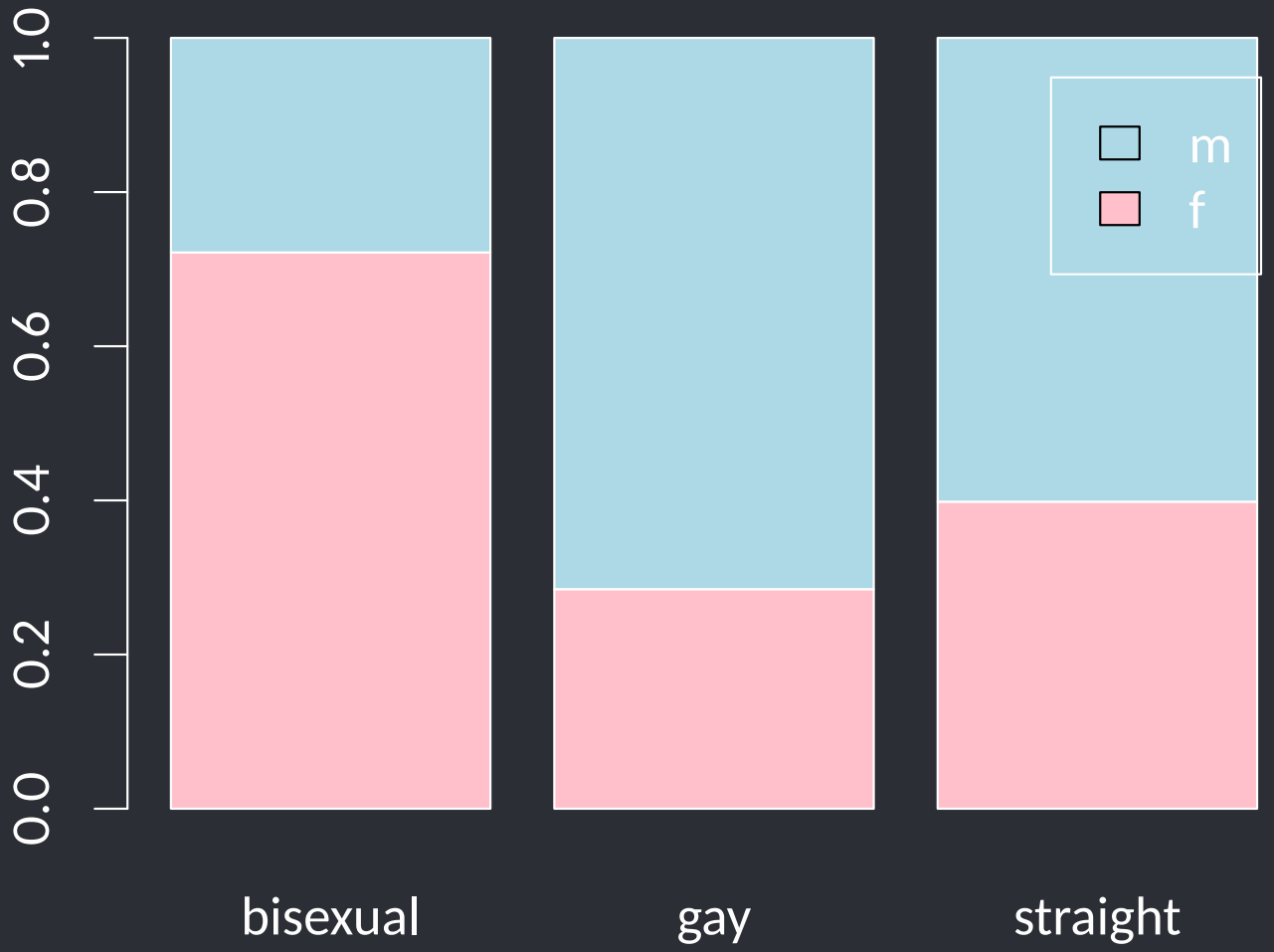
- Among people of the same height, being gay increases the predicted odds of being male by 713% (i.e., multiplied by $e^{2.1} = 8.13$) compared to being bisexual.
- Among people of the same height, being straight increases the predicted odds of being male by 305% (i.e., multiplied by $e^{1.4} = 4.05$)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

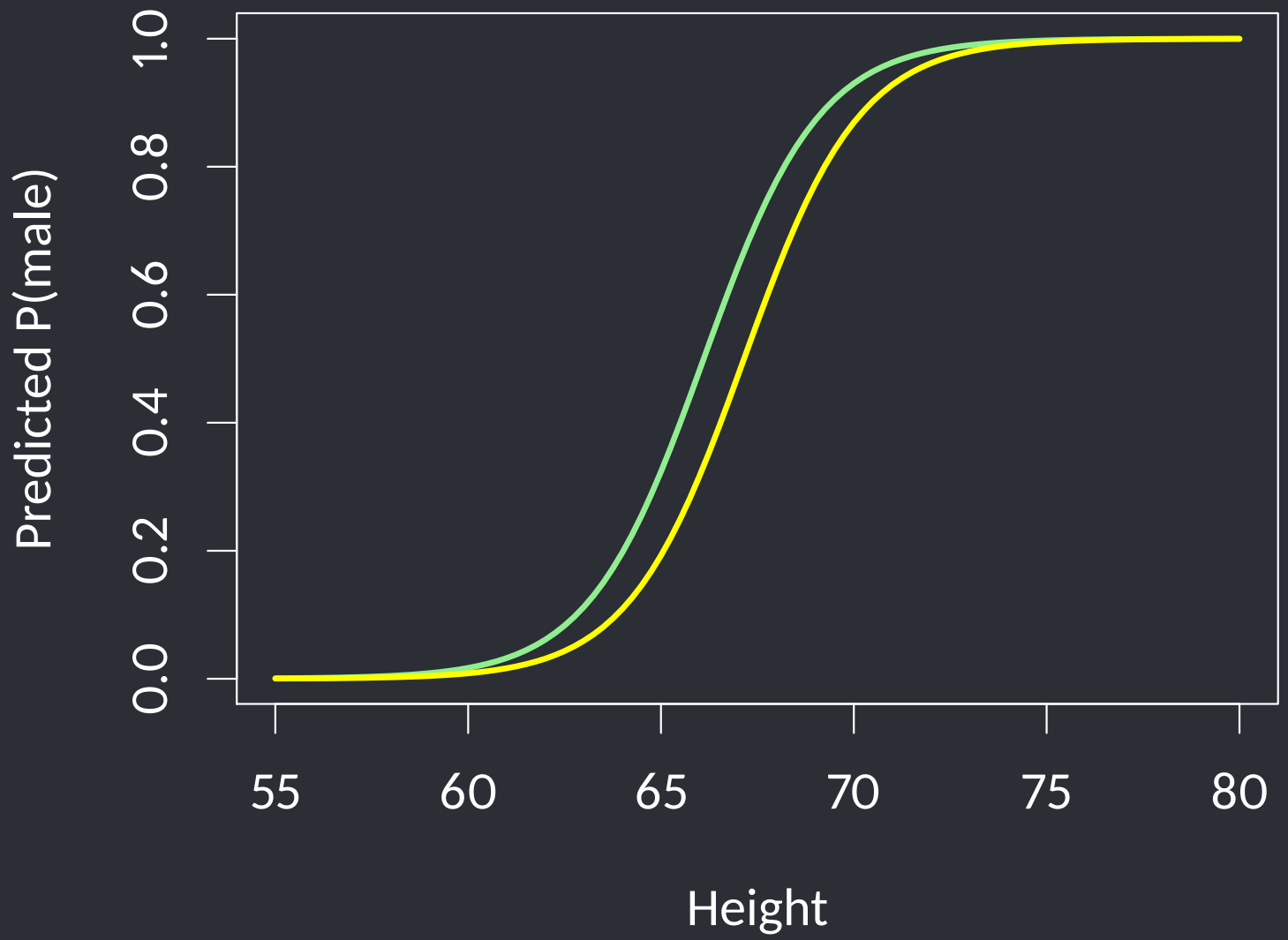


Converting back to probabilities

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:

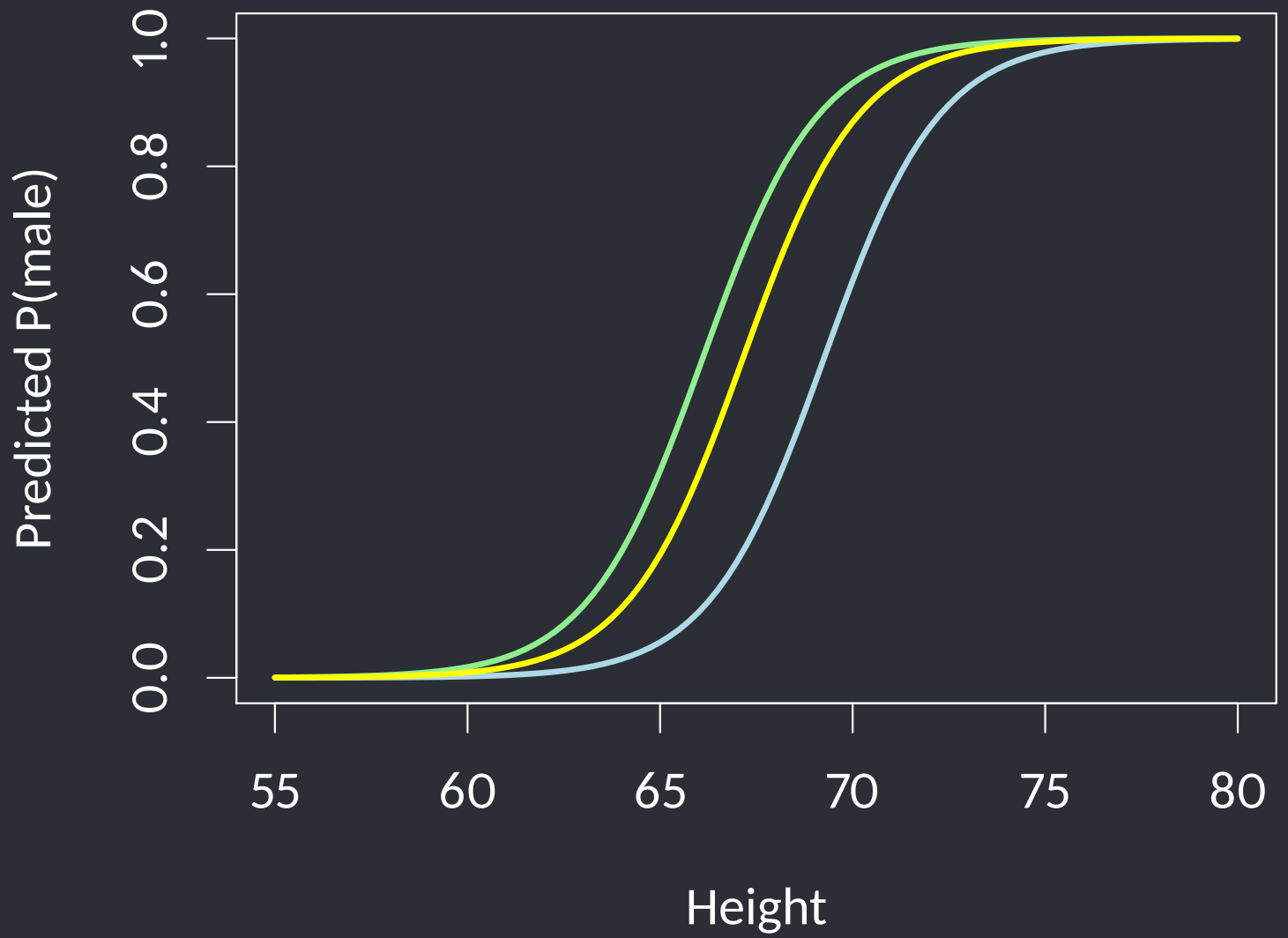
	Height			
	60"	64"	68"	72"
bisexual	0.002	0.029	0.302	0.861
gay	0.017	0.197	0.779	0.981
straight	0.008	0.109	0.637	0.962

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?

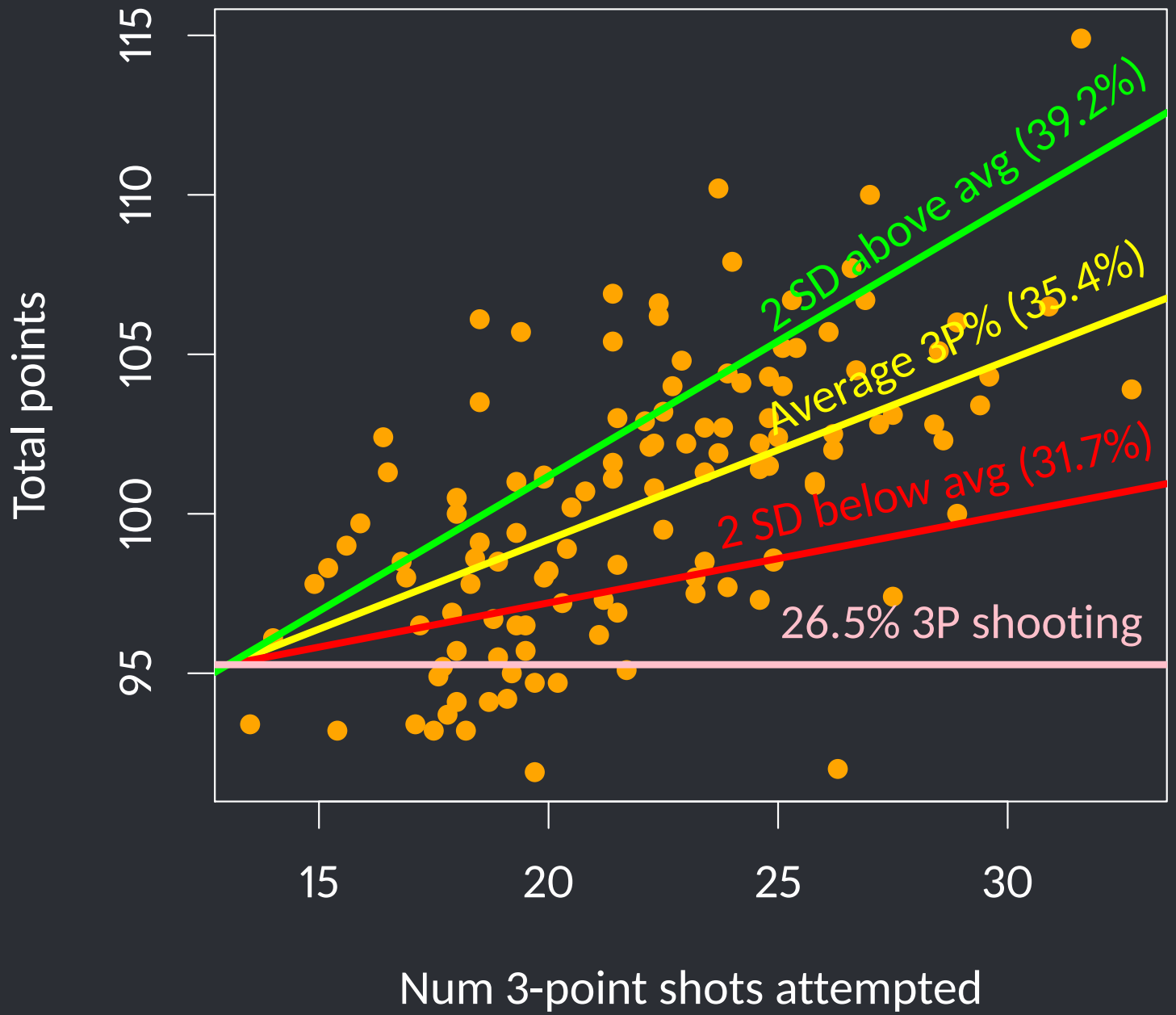
We can also visualize this by plotting the three curves for straight (yellow), gay (green), and bisexual (blue) OkCupid users:



1. Logistic regression with 2+ predictors
2. Interactions in logistic regression
3. Hypothesis testing when there are 2+ predictors
4. 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 $X_1 =$ height and $X_2 =$ orientation.

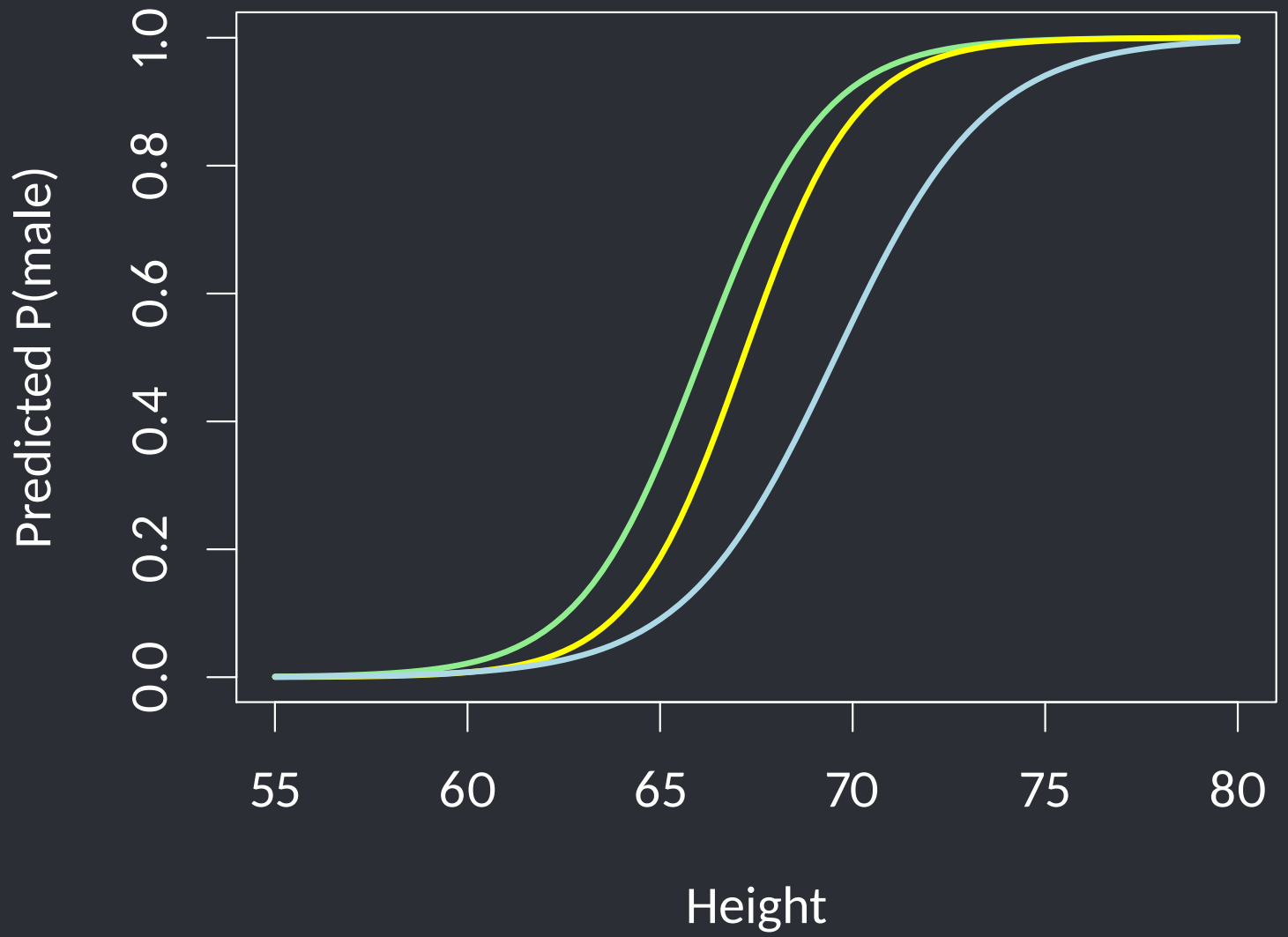
```
int.model <- glm(male ~ height * orientation, data=my.profiles, family=binomial)
summary(int.model)
```

```
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), yellow (straight), and blue (bisexual) users:



1. Logistic regression with 2+ predictors
2. Interactions in logistic regression
3. Hypothesis testing when there are 2+ predictors
4. Other applications of logistic regression

Four kinds of hypotheses to test

1. Overall null hypothesis: $\beta_1 = \beta_2 = \cdots = 0$ (all of the slope coefficients are 0, the model has no predictive power at all)

Four kinds of hypotheses to test

1. Overall null hypothesis: $\beta_1 = \beta_2 = \cdots = 0$ (all of the slope coefficients are 0, the model has no predictive power at all)
2. Quantitative variable null hypothesis: $\beta_i = 0$ (there is no relationship between gender and a particular predictor variable, holding constant the other predictors)

Four kinds of hypotheses to test

1. Overall null hypothesis: $\beta_1 = \beta_2 = \cdots = 0$ (all of the slope coefficients are 0, the model has no predictive power at all)
2. Quantitative variable null hypothesis: $\beta_i = 0$ (there is no relationship between gender and a particular predictor variable, holding constant the other predictors)
3. Categorical variable null hypothesis: $\beta = 0$ for all dummy variables corresponding to this categorical variable (there is no relationship between gender and a particular predictor variable, holding constant the other predictors)

Four kinds of hypotheses to test

1. Overall null hypothesis: $\beta_1 = \beta_2 = \cdots = 0$ (all of the slope coefficients are 0, the model has no predictive power at all)
2. Quantitative variable null hypothesis: $\beta_i = 0$ (there is no relationship between gender and a particular predictor variable, holding constant the other predictors)
3. Categorical variable null hypothesis: $\beta = 0$ for all dummy variables corresponding to this categorical variable (there is no relationship between gender and a particular predictor variable, holding constant the other predictors)
4. Individual dummy variable coefficient null hypothesis: $\beta_i = 0$ (there is no difference in predicted probability of being male between this level and the reference level, holding constant other predictors)

Likelihood ratio test

The likelihood ratio test lets us test a null hypothesis of the form: Model A has no more predictive power than Model B.

We can use this to test null hypothesis that don't correspond to *p*-values that we can read off the regression output. (And remember that there's no R^2 or Adjusted R^2 in logistic regression to compare models!)

Example 1: Overall null hypothesis

We'll test the overall null hypothesis by comparing the model to a "null model" with no variables:

```
library(lmtest)
lrtest(model2)

Likelihood ratio test

Model 1: male ~ height + orientation
Model 2: male ~ 1
  #Df LogLik Df Chisq Pr(>Chisq)
1   4 -21861
2   1 -40327 -3 36932     <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example 2: Quantitative variable

We can test the significance of a quantitative variable (e.g., height) by reading the *p*-value for `height` off of the regression output:

```
summary(model2)

...
                    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

...
```

Example 3: Categorical variable

We'll test the significance of a categorical variable by comparing the model with `orientation` to the model without it:

```
model1 <- glm(male ~ height, data=my.profiles, family=binomial)
model2 <- glm(male ~ height + orientation, data=my.profiles, family=binomial)
lrtest(model1, model2)

Likelihood ratio test

Model 1: male ~ height
Model 2: male ~ height + orientation
  #Df LogLik Df Chisq Pr(>Chisq)
1   2 -22319
2   4 -21861  2   915     <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example 4: Individual dummy variable

We can test the significance of the difference between two levels of a categorical variable (e.g. the difference between bisexual and straight) reading the *p*-value for `height:orientationstraight` off of the regression output:

```
summary(model2)

...
                    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

...
```

1. Logistic regression with 2+ predictors
2. Interactions in logistic regression
3. Hypothesis testing when there are 2+ predictors
4. 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