Recap

- Data set: 59946 profiles from users of OkCupid
- We predicted the sex of the user based on their height:

```
my.profiles <- profiles %>%
                mutate(male=ifelse(profiles$sex == "m", 1, 0)) %>%
                filter(height >= 55 & height <= 80)
model <- glm(male ~ height, data=my.profiles,
            family=binomial)
```

The regression output tells us that our prediction is

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

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

Logistic regression with 2+ predictors

Interactions in logistic regression

Hypothesis testing

Business applications

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 **the odds that Y=1**, 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:

```
xtabs(~ orientation, data=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.

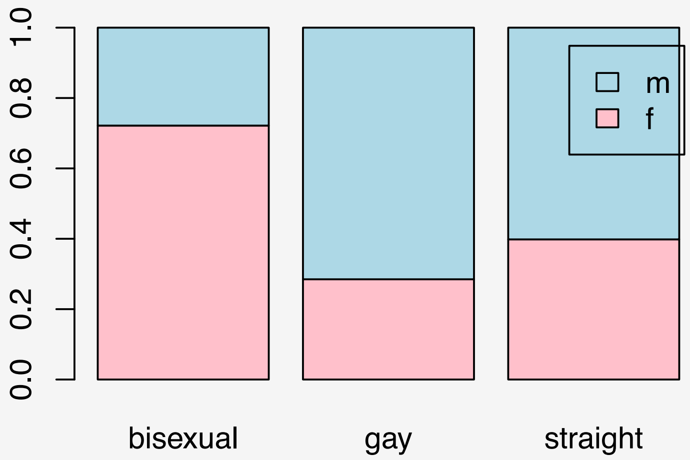
Interpreting coefficients

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

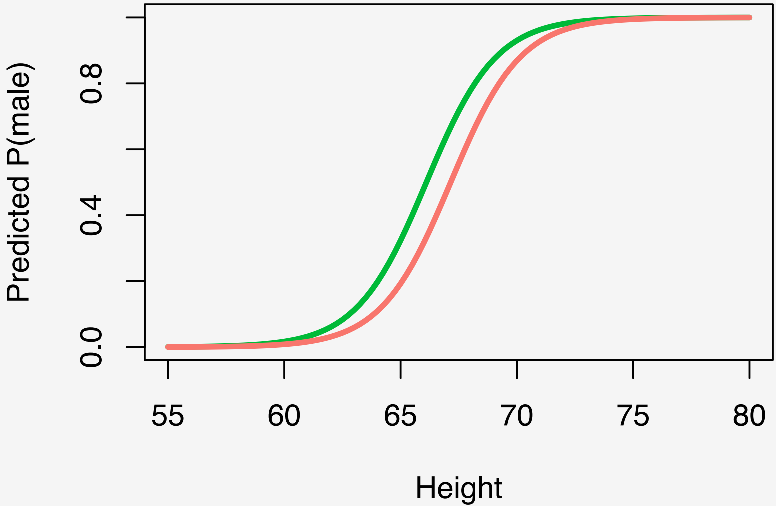
- Among people of the same height, being gay increases the predicted odds of being male by 713% (i.e., multiplied by $e^{2.1} = 8.13$) compared to being bisexual.
- Among people of the same height, being straight increases the predicted odds of being male by 305% (i.e., multiplied by $e^{1.4} = 4.05$)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

Understanding what's going on

```
xtabs(~ sex + orientation, data=my.profiles)

   orientation
sex bisexual   gay straight
  f     1994  1586    20509
  m      769  3982    30986

xtabs(~ sex + orientation, data=my.profiles) %>% prop.table(2)

   orientation
sex bisexual  gay straight
  f     0.72 0.28     0.40
  m     0.28 0.72     0.60
```

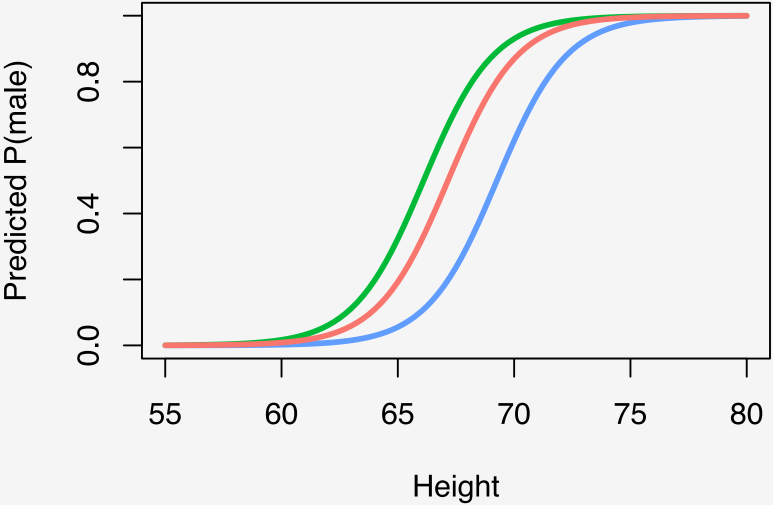

1.0
0.8
0.6
0.4
0.2
0.0
m
f
bisexual
gay
straight

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



Logistic regression with 2+ predictors

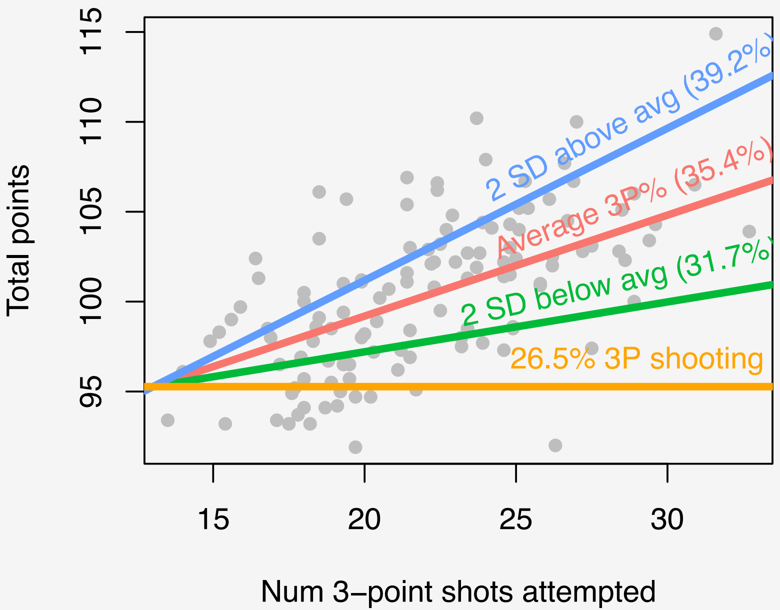
Interactions in logistic regression

Hypothesis testing

Business applications

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.


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

What would interactions do?

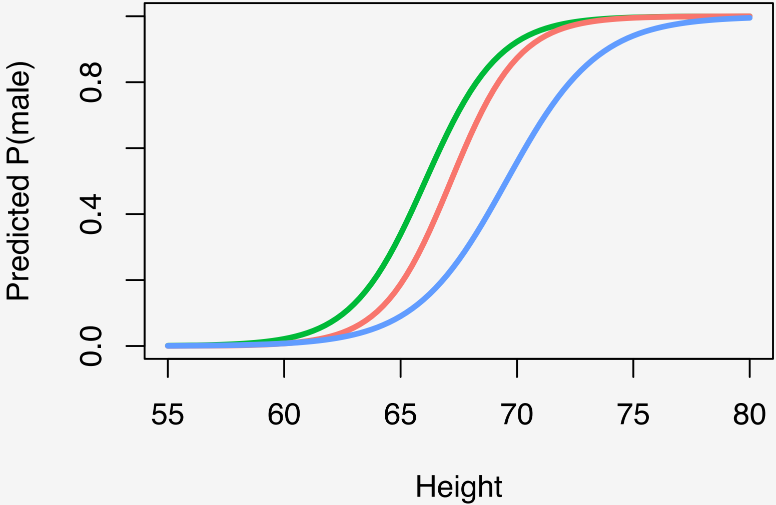
- We can add interactions to logistic regression and the interpretation is the same: the effect of X_1 on the odds of being male depends on the value of X_2.
- Let's try this out with X_1 = height and X_2 = orientation.

```
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)

    Null deviance: 80654  on 59825  degrees of freedom
```

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), red (straight), and blue (bisexual) users:



Logistic regression with 2+ predictors

Interactions in logistic regression

Hypothesis testing

Business applications

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. 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)

Example 1: Overall null hypothesis

The likelihood ratio test provides us with an overall p-value for the model testing the null hypothesis that $\beta_1 = \beta_2 = \cdots = 0$ (all of the slope coefficients are 0):

```
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 whether a a quantitative variable (e.g., height) is statistically significantly different from zero 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: Individual dummy variable

We can test whether the difference between two levels of a categorical variable (e.g. the difference between bisexual and straight) is statistically significantly different from zero by reading the *p*-value for `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
```

Logistic regression with 2+ predictors

Interactions in logistic regression

Hypothesis testing

Business applications

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