THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Logistic Regression 1

Lecture 19

STA 371G

okcupid

Have an account? Sign in

DATING DESERVES BETTER

On OkCupid, you're more than just a photo. You have stories to tell, and passions to share, and things to talk about that are more interesting than the weather. Get noticed for who you are, not what you look like. Because you deserve what dating deserves: better.

JOIN okc

By clicking Join, you agree to our Terms of Service. Learn about how we process and use your data in our Privacy Policy and how we use cookies and similar technology in our Cookie Policy.

GET THE APP

The OkCupid data set

- The OkCupid data set contains information about 59946 profiles from users of the OkCupid online dating service.
- We have data on user age, height, sex, income, sexual orientation, education level, body type, ethnicity, and more.
- OkCupid often publishes their own analyses of their data—see `https://theblog.okcupid.com/tagged/data`.
- Let's see if we can predict the sex/gender of the user based on their height.

What's wrong with this regression?

$$\widehat{\text{sex}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{height}$$

What's wrong with this regression?

$$\widehat{\text{sex}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{height}$$

The Y variable here is categorical (two levels—everyone in this data set is either labeled male or female), so regular linear regression won't work here.

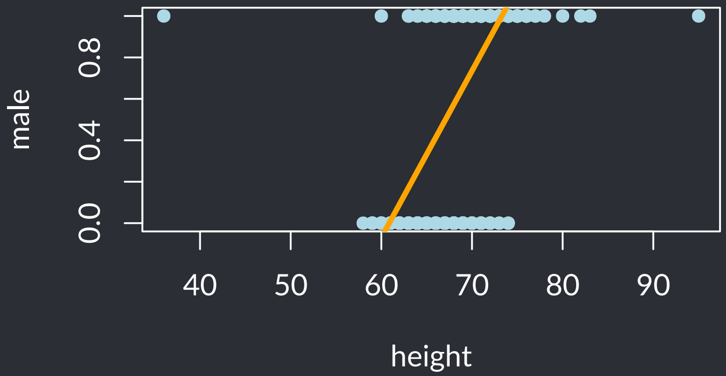
But what if we just do it anyway?

Let's first create a dummy variable to convert sex to a quantitative dummy variable:

```
profiles$male <- ifelse(profiles$sex == "m", 1, 0)
```

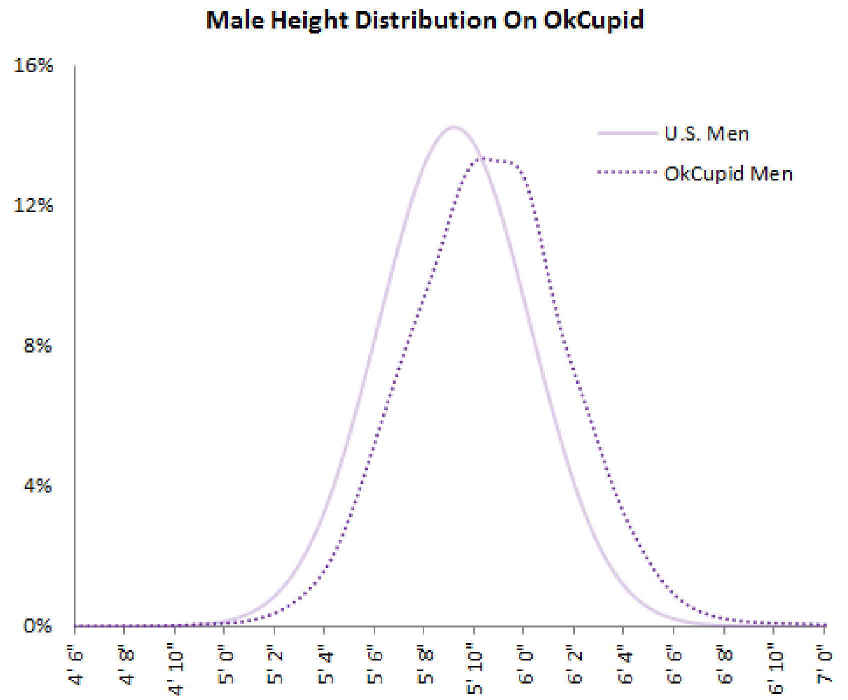
We could do this with 1 representing either male or female (it wouldn't matter).

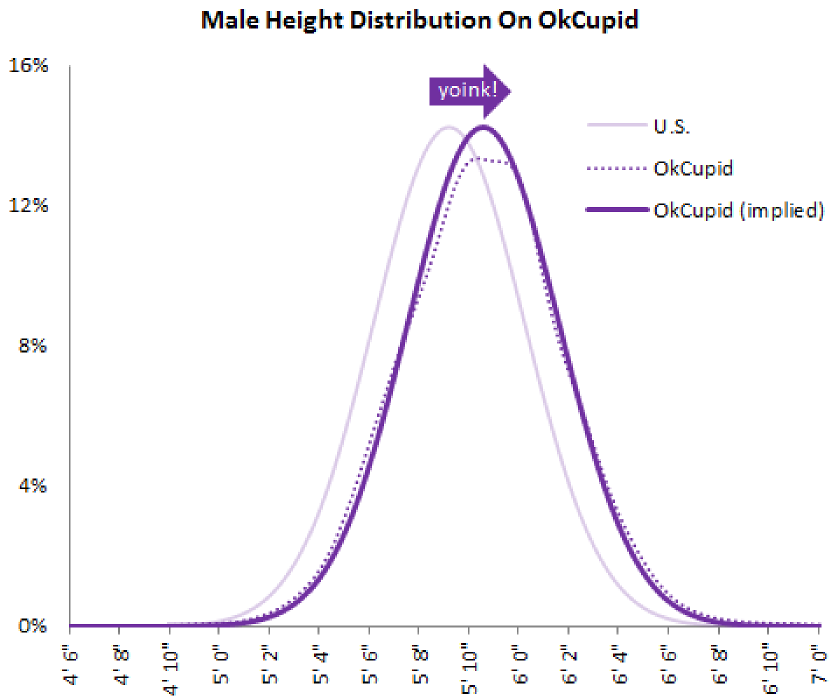
But what if we just do it anyway?



A line is a spectacularly bad fit to this data—it's not even close to linear. And what does it mean to predict that male = 0.7 (or 1.2)?

What challenges might we run into with this data?


Male Height Distribution On OkCupid
16%
12%
8%
4%
0%
U.S. Men
OkCupid Men
4' 6"
4' 8"
4' 10"
5' 0"
5' 2"
5' 4"
5' 6"
5' 8"
5' 10"
6' 0"
6' 2"
6' 4"
6' 6"
6' 8"
6' 10"
7' 0"

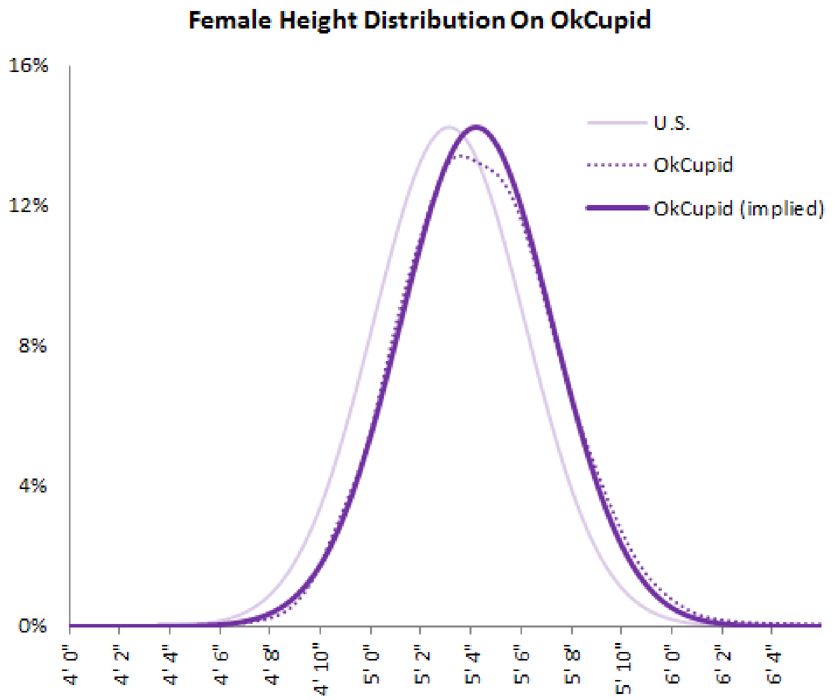


Male Height Distribution On OkCupid
16%
12%
8%
4%
0%
yoink!
U.S.
OkCupid
OkCupid (implied)
4' 6"
4' 8"
4' 10"
5' 0"
5' 2"
5' 4"
5' 6"
5' 8"
5' 10"
6' 0"
6' 2"
6' 4"
6' 6"
6' 8"
6' 10"
7' 0"

So men lie about their height—by an average of about 2 inches! And many men round up to 6 feet.

So men lie about their height—by an average of about 2 inches! And many men round up to 6 feet.

Women do too!



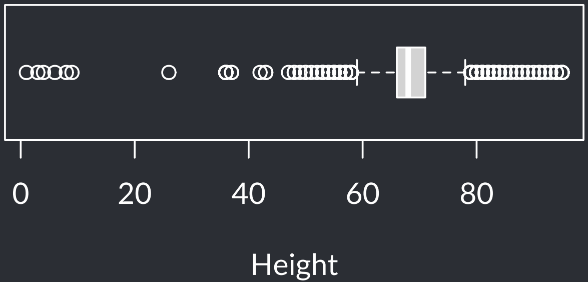
Female Height Distribution On OkCupid
16%
12%
8%
4%
0%
U.S.
OkCupid
OkCupid (implied)
4' 0"
4' 2"
4' 4"
4' 6"
4' 8"
4' 10"
5' 0"
5' 2"
5' 4"
5' 6"
5' 8"
5' 10"
6' 0"
6' 2"
6' 4"

We don’t really have any tools at our disposal to correct for this, but let’s still proceed with the analysis (with some caution) since the exaggeration seems about the same regardless of gender.

Cleaning the data

There are definitely some weird values for height:

```
boxplot(profiles$height, horizontal=T, xlab="Height")
```



Cleaning the data

Let’s consider only heights between 55 and 80 inches (4’7" and 6’8"), inclusive. This is arbitrary, but it excludes only 117 cases out of 59946.

```
my.profiles <- subset(profiles,
                      height >= 55 & height <= 80)
```

The idea behind logistic regression

- Instead of predicting whether someone is male, let's predict the *probability* that they are male
- In logistic regression, one level of Y is always called "success" and the other called "failure." Since $Y = 1$ for males, in our setup we have designated males as "success." (You could also set $Y = 1$ for females and call females "success.")
- Let's fit a curve that is always between 0 and 1.

Odds

- When something has “even (1/1) odds,” the probability of success is 1/2

Odds

- When something has “even (1/1) odds,” the probability of success is 1/2
- When something has “2/1 odds,” the probability of success is 2/3

Odds

- When something has “even (1/1) odds,” the probability of success is 1/2
- When something has “2/1 odds,” the probability of success is 2/3
- When something has “3/2 odds,” the probability of success is 3/5

Odds

- When something has “even (1/1) odds,” the probability of success is 1/2
- When something has “2/1 odds,” the probability of success is 2/3
- When something has “3/2 odds,” the probability of success is 3/5
- In general, the odds of something happening are $p/(1-p)$

The logistic regression model

Logistic regression models the log odds of success p as a linear function of X:

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

This fits an S-shaped curve to the data (we'll see what it looks like later).

Let’s try it

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

How to interpret the curve?

The regression output tells us that our prediction is

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

How to interpret the curve?

The regression output tells us that our prediction is

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

Let's solve for $P(\text{male})$:

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

Making predictions

We can use `predict` to automate the process of plugging into the equation:

```
predict(model, list(height=69), type="response")

   1
0.77
```

$$\frac{e^{-44.45+0.66\cdot 69}}{1+e^{-44.45+0.66\cdot 69}} = 0.77$$

Making predictions

We can use `predict` to automate the process of plugging into the equation:

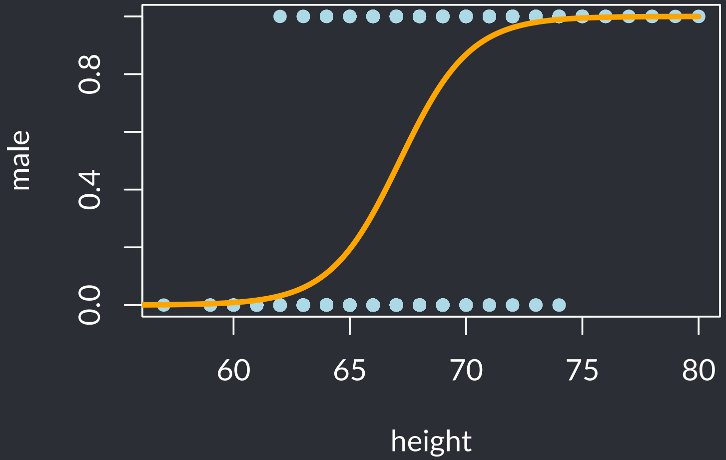
```
predict(model, list(height=69), type="response")

   1
0.77
```

$$\frac{e^{-44.45+0.66\cdot 69}}{1+e^{-44.45+0.66\cdot 69}} = 0.77$$

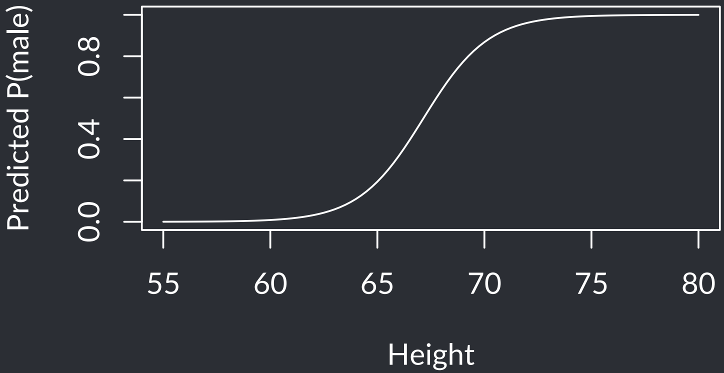
We predict that someone that is 5’9" has a 77% chance of being male.

Visualizing the model



How to interpret the curve?

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



Interpreting the coefficients

Our prediction equation is:

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

Let's start with some basic, but not particularly useful, interpretations:

- When height $= 0$, we predict that the log odds will be -44.45

Interpreting the coefficients

Our prediction equation is:

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

Let's start with some basic, but not particularly useful, interpretations:

- When height $= 0$, we predict that the log odds will be -44.45

Interpreting the coefficients

Our prediction equation is:

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

Let's start with some basic, but not particularly useful, interpretations:

- When height $= 0$, we predict that the log odds will be -44.45 , so the probability of male is predicted to be very close to 0%.
- When height increases by 1 inch, we predict that the log odds of being male will increase by 0.66.

Interpreting the coefficients

Let's rewrite the prediction equation as:

$$\text{Predicted odds of male} = e^{-44.45+0.66\cdot\text{height}}$$

Increasing height by 1 inch will *multiply* the odds by $e^{0.66} = 1.94$; i.e., increase the odds by 94%.

Interpreting the coefficients

Let's rewrite the prediction equation as:

$$\text{Predicted odds of male} = e^{-44.45+0.66\cdot\text{height}}$$

Increasing height by 1 inch will *multiply* the odds by $e^{0.66} = 1.94$; i.e., increase the odds by 94%.

Increasing height by 2 inches will *multiply* the odds by $e^{2\cdot 0.66} = 3.76$; i.e., increase the odds by 276%.