SOCIAL DATA SCIENCE

Causation & Prediction

Sebastian Barfort

August 14, 2016

University of Copenhagen
Department of Economics

What is the objective of empirical policy research?

1. causation: what is the effect of a particular variable on an outcome?
2. prediction: find some function that provides a good prediction of y as a function of x

Today: Introduction.

$$y = \alpha + \beta x + \varepsilon$$

causation: $\hat{\beta}$ problem

prediction: $\hat{y}$ problem

Causal Inference

Most econometric theory is focused on estimating **causal effects**

Causal effect: what is the effect of some policy on an outcome we are interested in?

Examples of causal questions:

- what is the effect of immigration on native wages?
- what is the effect of democracy on growth?
- what is the effect of newspaper coverage on stock prices?

Variable of interest (often called *treatment*): D_i

Outcome of interest: Y_i

Potential outcome framework

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1, \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

The observed outcome Y_i can be written in terms of potential outcomes as

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

$Y_{1i} - Y_{0i}$ is the *causal* effect of D_i on Y_i.

But we never observe the same individual i in both states. This is the **fundamental problem of causal inference**.

SELECTION BIAS I

We need some way of estimating the state we do not observe (the counterfactual)

Usually, our sample contains individuals from both states

So why not do a naive comparison of averages by treatment status?

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] =& E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] + \\ & E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] \end{aligned}$$

SELECTION BIAS II

$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] = E[Y_{1i} - Y_{0i}|D_i = 1]$: the average *causal* effect of D_i on Y.

$E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$: difference in average Y_{0i} between the two groups. Likely to be different from 0 when individuals are allowed to self-select into treatment. Often referred to as selection bias.

RANDOM ASSIGNMENT SOLVES THE PROBLEM

Random assignment of D_i solves the problem because random assignment makes D_i independent of potential outcomes

That means that $E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$ and thus that the selection bias term is zero

Intuition: with random assignment, non-treated individuals can be used as counterfactuals for treated (*what would have happened to individual i had he not received the treatment*?)

This allows us to overcome the fundamental problem of causal inference

RANDOMIZATION

no causation without manipulation

WHO RANDOMIZES?

As mentioned, we need to worry when individuals are allowed to self-select

This means that a lot of thought has to go into the randomization phase

Randomization into treatment groups has to be manipulated by someone

But what about effect of immutable characteristics such as race, gender, etc.?

QUASI EXPERIMENTS

Quasi-experiments: randomization happens by “accident”

- Differences in Differences
- Regression Discontinuity Design
- Instrumental variables

Randomized controlled trials: randomization done by researcher

- Survey experiments
- Field experiments

Note: difficult to say one is strictly better than the other.
Randomization can be impractical and/or unethical.

Can you come up with an example where randomization would be unethical?

External & internal validity

Internal validity: Refers to the validity of causal conclusions

External validity: Refers to the extent to which the conclusions of a particular study can be generalized beyond a particular setting

Imai (2016): RCTs trade off external and internal validity

Samii (2016): No such tradeoff.

OBSERVATIONAL STUDY

In many cases, social scientists are unable to randomize treatment assignment for ethical or logistic reasons

Observational study: No random manipulation of treatment

Strategy: Statistical control (control variables, fixed effects, matching, etc)

Risk selection/confounding bias.

Pritchett, Lant and Justin Sandefur. 2015. "Learning from Experiments When Context Matters." *American Economic Review*, 105(5): 471-75.

> *We analyze the trade-off between internal and external validity faced by a hypothetical policymaker weighing experimental and non- experimental evidence. Empirically, we find that for several prominent questions in development economics, relying on observational data analysis from within context produces treat- ment effect estimates with lower mean-square error than relying on experimental estimates from another context.*

RACIAL DISCRIMINATION IN THE LABOR MARKET

Does racial discrimination exist in the labor market?

Experiment: In response to newspaper ads, researchers send out resumes of fictitious job candidates, varying only the names of the job applicants while leaving all other information in the resumes unchanges

Names were randomized between stereotypically black- and white-sounding names (Lakisha, Jamal, Emily, Greg, etc.)

```
library("readr")
gh.link = "https://raw.githubusercontent.com/"
user.repo = "kosukeimai/qss/"
branch = "master/"
link = "CAUSALITY/resume.csv"
data.link = paste0(gh.link, user.repo, branch, link)
df = read_csv(data.link)
```

firstname	sex	race	call
Allison	female	white	0
Kristen	female	white	0
Lakisha	female	black	0
Latonya	female	black	0
Carrie	female	white	0

CONTINGENCY TABLE

```
library("dplyr")
df.table = df %>%
  count(race, call)
```

race	call	n
black	0	2278
black	1	157
white	0	2200
white	1	235

```
library("dplyr")
df.table = df %>%
  group_by(race, call) %>%
  summarise(n = n()) %>%
  mutate(freq = n / sum(n))
```

race	call	n	freq
black	0	2278	0.9355236
black	1	157	0.0644764
white	0	2200	0.9034908
white	1	235	0.0965092

BY GENDER

```
library("dplyr")
df.table = df %>%
  group_by(race, sex, call) %>%
  summarise(n = n()) %>%
  mutate(freq = n / sum(n))
```

race	sex	call	n	freq
black	female	0	1761	0.9337222
black	female	1	125	0.0662778
black	male	0	517	0.9417122
black	male	1	32	0.0582878
white	female	0	1676	0.9010753
white	female	1	184	0.0989247
white	male	0	524	0.9113043
white	male	1	51	0.0886957

SYSTEMATIC DIFFERENCES?

Difference in means estimator

```
lin.model = lm(call ~ race == "black",
               data = df)
```

Table 5:

	Dependent variable:
	call
race == ”black”	−0.032***
	(0.008)
Constant	0.097***
	(0.006)
Observations	4,870
R^2	0.003
Adjusted R^2	0.003
Residual Std. Error	0.272 (df = 4868)
F Statistic	16.931*** (df = 1; 4868)
Note:	*p<0.1; **p<0.05; ***p<0.01

EXAMPLE: REGRESSION DISCONTINUITY DESIGN

Eggers, Andrew and Alexander Fouirnaies. 2014. "The Economic Impact of Economic News."

> *We start from the observation that news media pay considerable attention to a binary distinction between recession and non-recession: by a convention observed in essentially every industrialized country, a recession is announced when an economy contracts for two consecutive quarters. In cases where growth is essentially zero, the distinction between a recession and a non-recession becomes highly arbitrary. Nevertheless (as we confirm below), the media treat fundamentally comparable situations quite differently, producing anxious headlines announcing a recession if growth is barely negative for two consecutive quarters but not if growth is even slightly positive.*

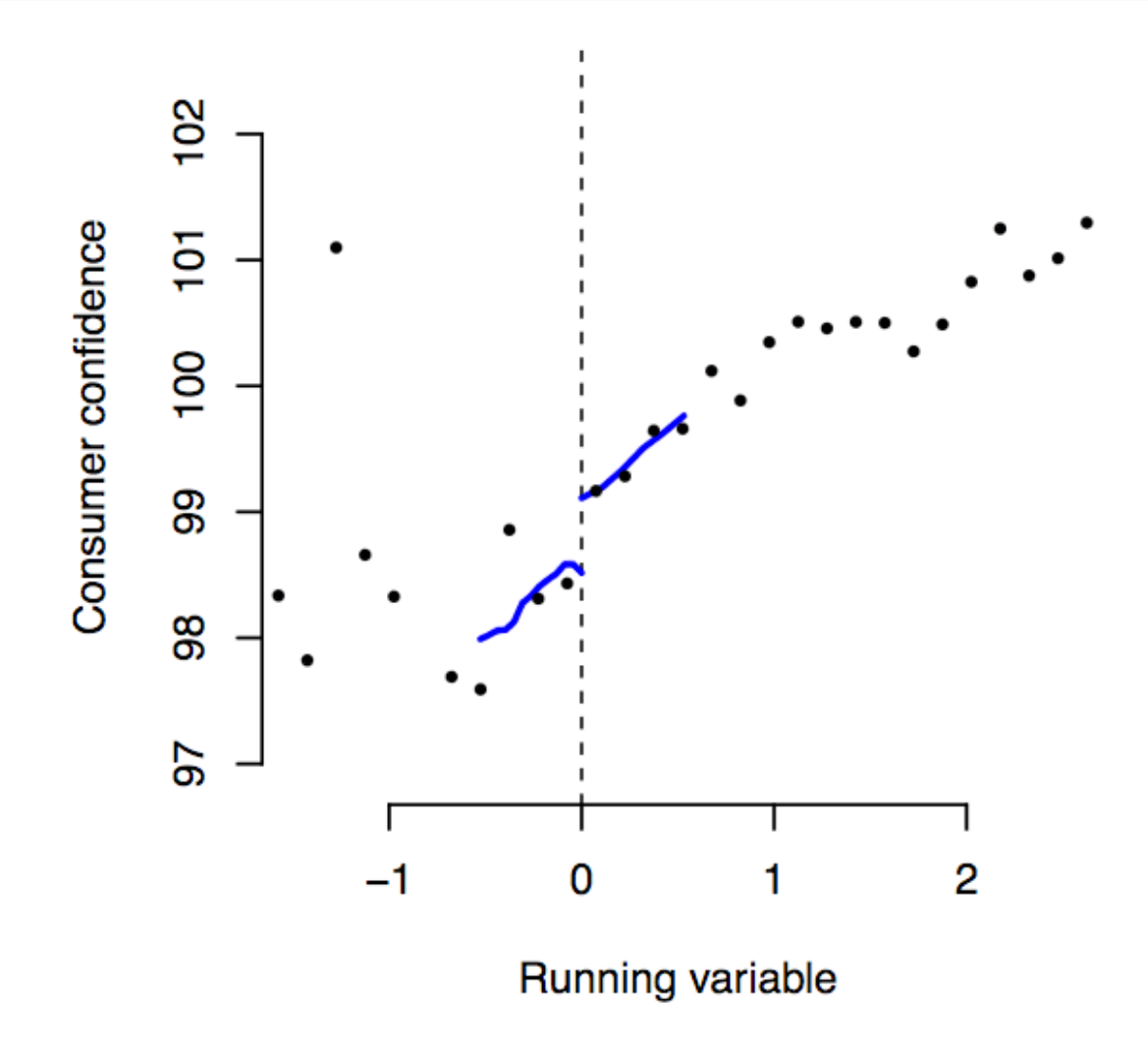
IDENTIFICATION STRATEGY

Discontinuous relationship between recession announcements in media and underlying economic fundamentals

Some countries are assigned to *recession* while others are not even when they have the same underlying economic fundamentals (exogenous assignment of treatment)

Result: Announcing a recession reduces both consumer confidence and growth in private consumption in the quarter during which the recession is announced

RESULTS


Consumer confidence
97
98
99
100
101
102
−1
0
1
2
Running variable

SUMMARY

Causal questions are of key interest to policy makers and academics

The key focus is on inference: we want to know about the causal effect of D on Y *in the population of interest*

When you are interested in a **causal question** you need to think carefully about randomization of treatment (this is often referred to as your identification strategy)

Due to the fundamental problem of causal inference, we can't estimate individual-level causal effects. Instead, we estimate averages.

Is causality the only thing policy makers, firms and social scientists should be interested in?

Prediction

PREDICTION

Many policy problems are not about causality but rather about prediction

Sometimes called *prediction policy problems*

- How many people will sign up for Obamacare?
- Who will win the U.S general election in November?
- Who should the Department of Economics hire in the future?

WHO PREDICTS?

- Local governments -> pension payments/crime/etc
- Google -> whether you will click on an ad
- Netflix -> what movies you will watch
- Insurance companies -> what your risk of death is
- You? -> will *Social Data Science* be a fun/rewarding/interesting course to follow?

WHY PREDICT? GLORY!

Netflix Awards $1 Million Prize and Starts a New Contest

THE NETFLIX CONTEST

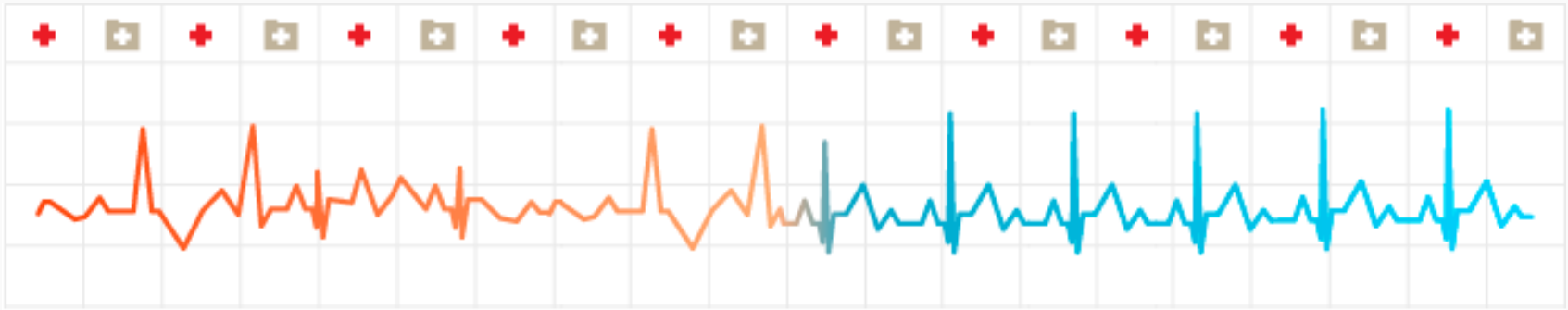
Competition started in October 2006. Training data is ratings for 18K movies by 400K Netflix customers, each rating between 1 and 5

Training data is very sparse - about 98% missing

Objective is to predict the rating for a set of 1 million customer-movie paris that are missing in the training data

Winner: Averaged 800 models (but the solution never actually implemented)

WHY PREDICT? RICHES!



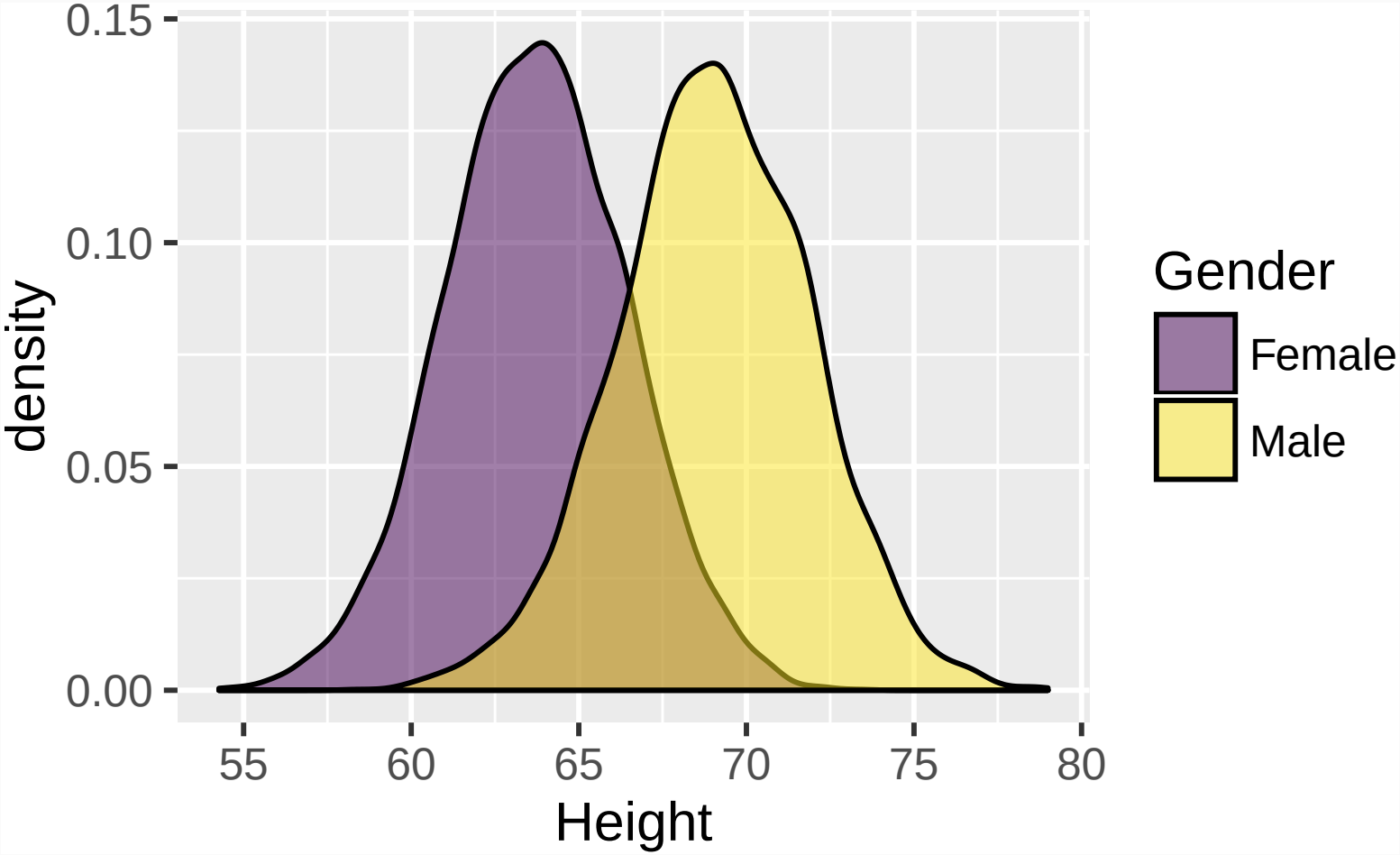
http://www.heritagehealthprize.com/c/hhp

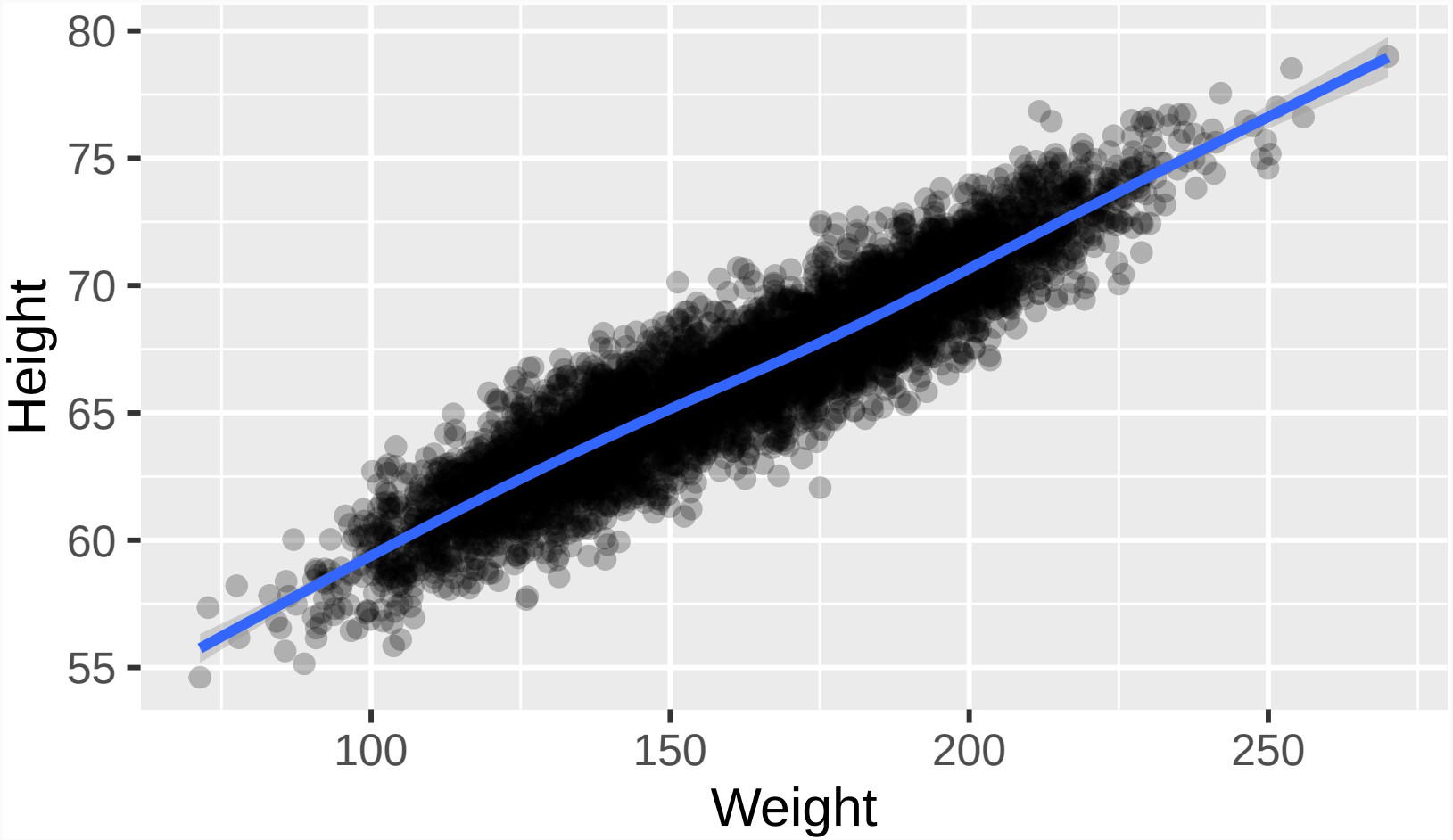
EXAMPLE: PREDICTING GENDER FROM WEIGHT/HEIGHT

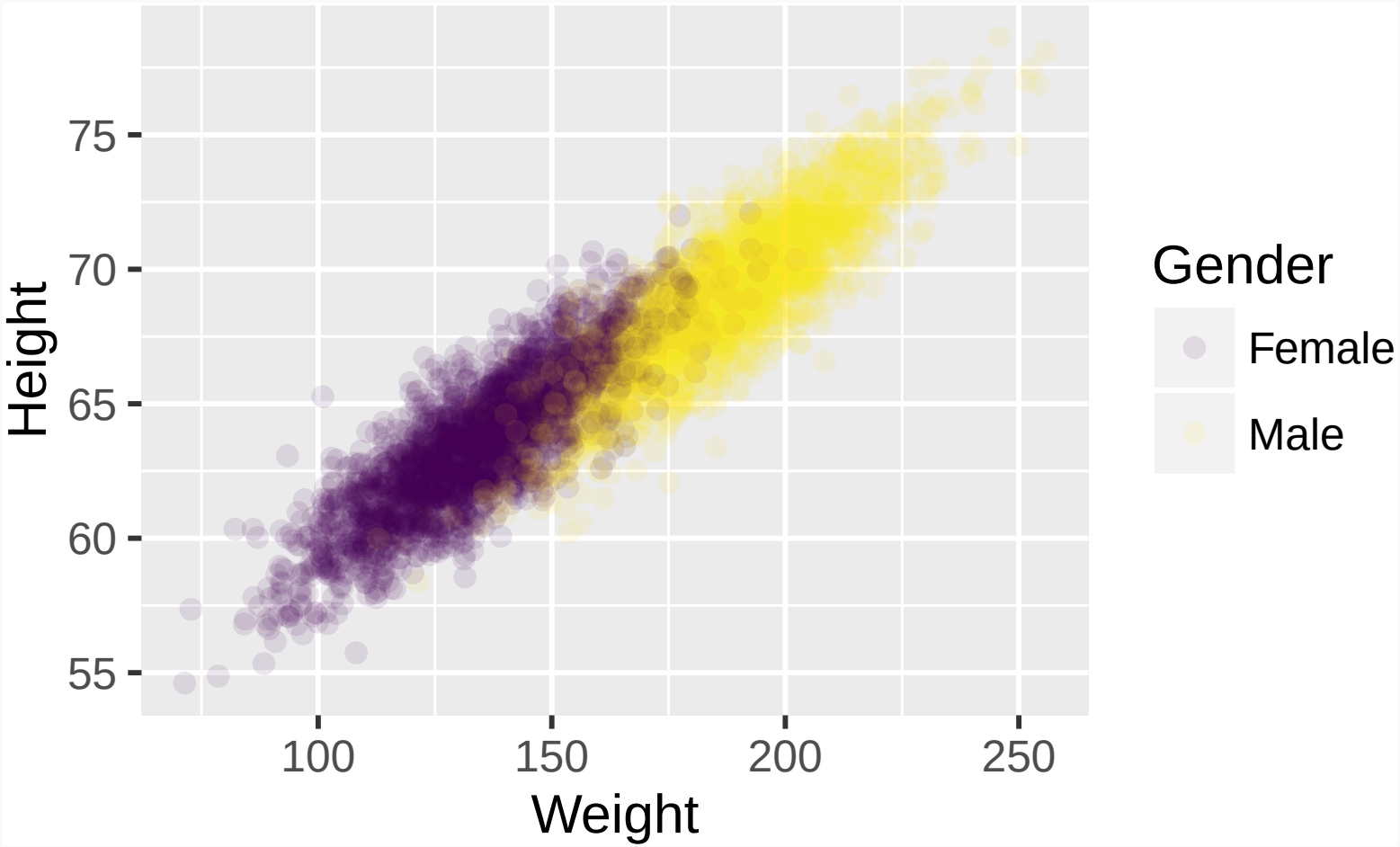
Can we predict gender based on information on an individual's weight/height?

```
gh.link = "https://raw.githubusercontent.com/"
user.repo = "johnmyleswhite/ML_for_Hackers/"
branch = "master/02-Exploration/"
link = "data/01_heights_weights_genders.csv"
data.link = paste0(gh.link, user.repo, branch, link)
df = read_csv(data.link)
```

Gender	Height	Weight
Male	73.84702	241.8936
Male	68.78190	162.3105
Male	74.11011	212.7409
Male	71.73098	220.0425
Male	69.88180	206.3498


0.15
0.10
0.05
0.00
density
55
60
65
70
75
80
Height
Gender
Female
Male


80
75
70
65
60
55
Height
100
150
200
250
Weight


Height
75
70
65
60
55
100
150
200
250
Weight
Gender
Female
Male

LOGIT MODEL

```
df = df %>% mutate(gender = Gender == "Male")
logit.model = glm(gender ~ Height + Weight,
                  data = df,
                  family = binomial(link = "logit"))
```

Logit estimates

$$P(Y_i = 1|X_i = x_i) = \frac{1}{1 + e^{-x_i\beta}}$$

This probability is .5 when $x_i\beta = 0$

So we can classify predicted gender based on height and weight

$$\hat{y} = \begin{cases} 1 & \text{if } x_i\beta \geq 0, \\ 0 & \text{otherwise} \end{cases}$$

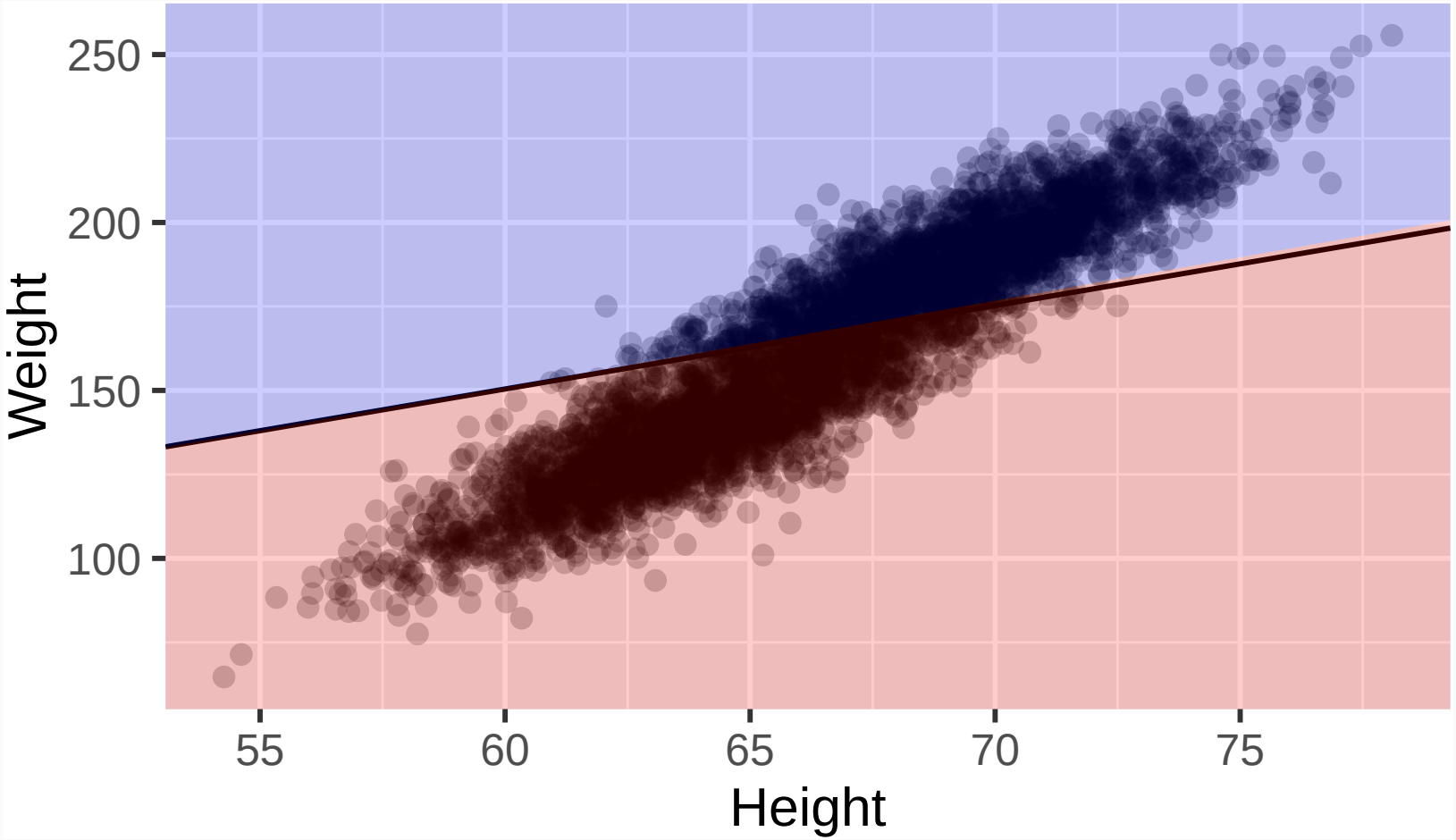
Intercept:

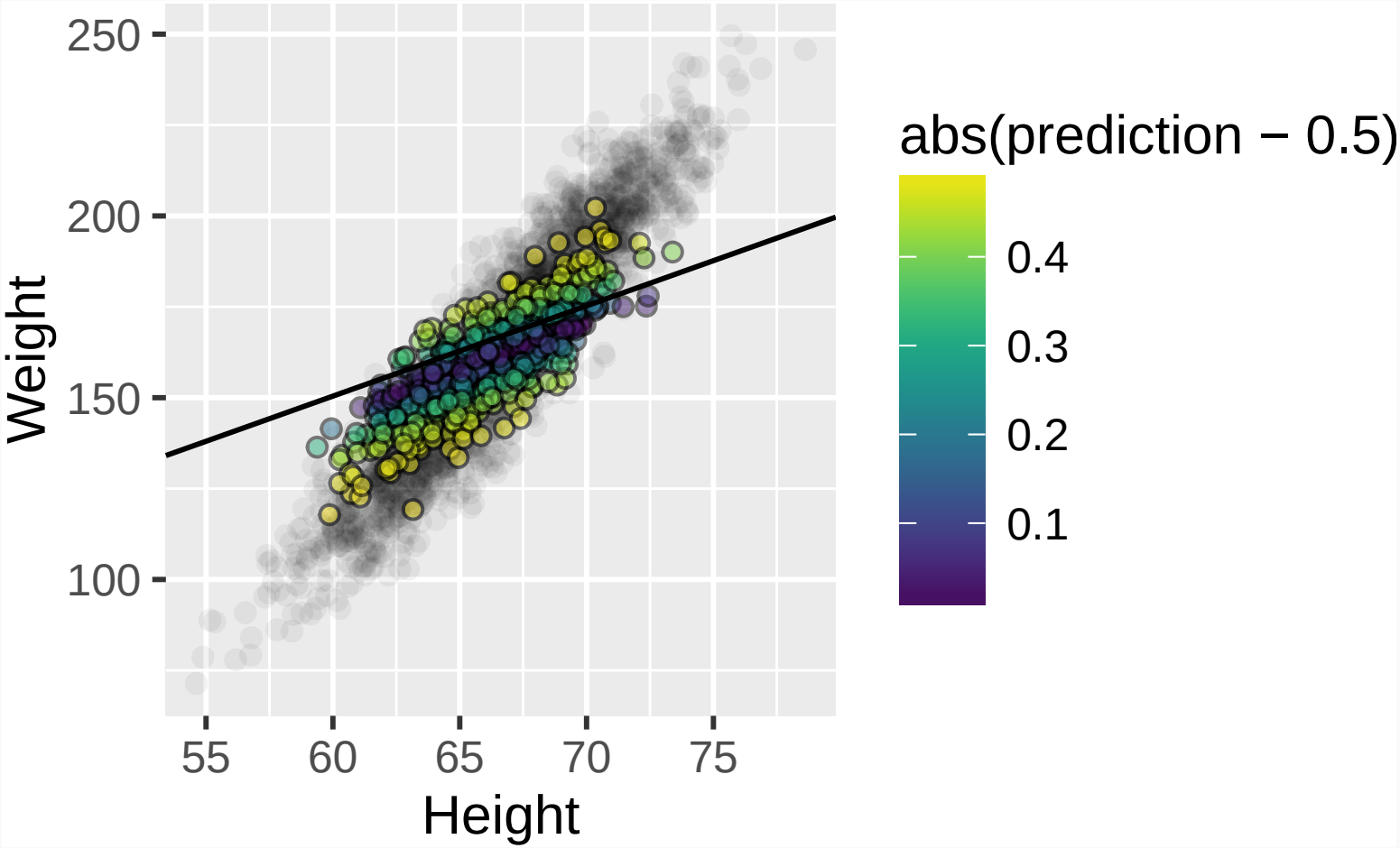
$$W = \frac{-\alpha - \beta_1 H}{\beta_2}$$

Slope:

$$-\frac{\beta_1}{\beta_2}$$

So now we have a classifier: For each combination of weight and height we're able to predict gender based on our logistic model


250
200
150
100
Weight
55
60
65
70
75
Height


250
200
150
100
Weight
55
60
65
70
75
Height
abs(prediction − 0.5)
0.4
0.3
0.2
0.1

How many cases were misclassified?

```
df.class = df %>%
  mutate(
    pred.prob = predict(logit.model),
    pred.cat = ifelse(pred.prob >= .5,
                      "Male", "Female"),
    classified = ifelse(prediction.cat == Gender,
                        "correct", "incorrect")
  ) %>%
  group_by(classified) %>%
  summarise(n = n()) %>%
  mutate(freq = n / sum(n))
```

classified	n	freq
correct	9194	0.9194
incorrect	806	0.0806

```
df.class = df %>%
  mutate(
    pred.prob = predict(logit.model),
    pred.cat = ifelse(pred.prob >= .5,
                      "Male", "Female")
  ) %>%
  group_by(Gender, pred.cat) %>%
  summarise(n = n()) %>%
  mutate(freq = n / sum(n))
```

Gender	pred.cat	n	freq
Female	Female	4708	0.9416
Female	Male	292	0.0584
Male	Female	583	0.1166
Male	Male	4417	0.8834

HOW WELL DID WE DO?

Overall, we correctly classified 92% of the cases in our dataset

Is this a good prediction model?

Is looking at the classification rate on all the data the correct strategy?

Standard empirical techniques are not optimized for prediction problems because they focus on inference

Standard result in econometrics: When there is no omitted variable bias ($E(\varepsilon) = 0$) and the model is homoskedastic ($V(\varepsilon_i) = \sigma^2$) then the OLS estimator is BLUE (Best Linear Unbiased Estimator).

Keywords: *unbiased* ($E(\hat{\beta}) = \beta$) and *best* (smallest variance among the class of all linear unbiased estimators)

But what about **biased** estimators?

OLS is designed to minimize *in sample error*: the error rate you get on the same data set you used to build your predictor.

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

But for prediction we are interested in minimizing out of sample error: the error rate you get on a new data set

Too see this, consider a prediction at a new point, x_0. Our prediction for y_0 is then $\hat{f}(x_0)$ and the mean squared error (MSE) can be decomposed as

$$E[(y_0 - \hat{f}(x_0))^2] = [\text{Bias}(\hat{f}(x_0))]^2 + V(\hat{f}(x_0)) + \sigma^2$$

By ensuring zero bias, OLS picks a corner solution. This is generally not optimal for prediction.

What do we mean by the *variance* and *bias* of an estimator?

Bias ($(\hat{f}(x_0)) = E[\hat{f}(x_0) - f(x_0)]$): Bias refers to the error that is introduced by approximating a real-life problem with a simple model. It won't fit the new data well.

Variance ($V(\hat{f}(x_0))$): Referes to model complexity. If the model is too complex then small changes to the data will cause the solution to change a lot.

Machine learning techniques were developed specifically to maximize prediction performance by providing an empirical way to make this bias-variance trade off

But generally, that means that all our models are somewhat biased (making inference irrelevant)

Cross validation: Split data in test and training data. Train model on training data, test it on test data

Regularization: A technique used in an attempt to solve overfitting problems

Supervised Learning: Models designed to infer a relationship from **labeled** training data.

- linear model selection (OLS, Ridge, Lasso)
- Classification (logistic, KNN, CART)

Unsupervised Learning: Models designed to infer a relationship from **unlabeled** training data.

- PCA
- KNN

SUMMARY

Statistical learning models are designed to optimally trade off bias and variance

This makes them more efficient for prediction than OLS

But also **generally biased** (so they are generally not meant for inference)

Statistical learning models can also be used for **exploratory data analysis**

Models in R

Linear regression is a simple approach to supervised learning. Assumes that the dependence of Y on $X_1, ..., X_n$ is linear

Assumes a model of

$$Y = \alpha + \beta X + \varepsilon$$

Where α and β are unknown constants to be estimated from the data.

When we've obtained these estimates, we can predict values of the dependent variable by plugging in new values of X

$$\hat{y} = \hat{\alpha} + \hat{\beta} X$$

```r
library("readr")
gh.link = "https://raw.githubusercontent.com/"
user.repo = "sebastianbarfort/sds_summer/"
branch = "gh-pages/"
link = "data/bball.csv"
data.link = paste0(gh.link, user.repo, branch, link)
df = read_csv(data.link)
```

PREDICT POINTS

```
model.1 = lm(
  pts ~ height + weight + fg.pct + ft.pct,
  data = df)
```

USEFUL FUNCTIONS AND PACKAGES

- `stargazer`: convert model to LaTeX
- `broom`: convert model to tidy format
- `summary`: summarise model output
- `predict`: predict new *y* based on *x*

```
library("stargazer")
stargazer(model.1)
```

Table 9:

	pts
height	−3.690
	(2.971)
weight	0.009
	(0.046)
fg.pct	47.940***
	(15.709)
ft.pct	11.371
	(7.869)
Constant	4.149
	(14.855)
N	54

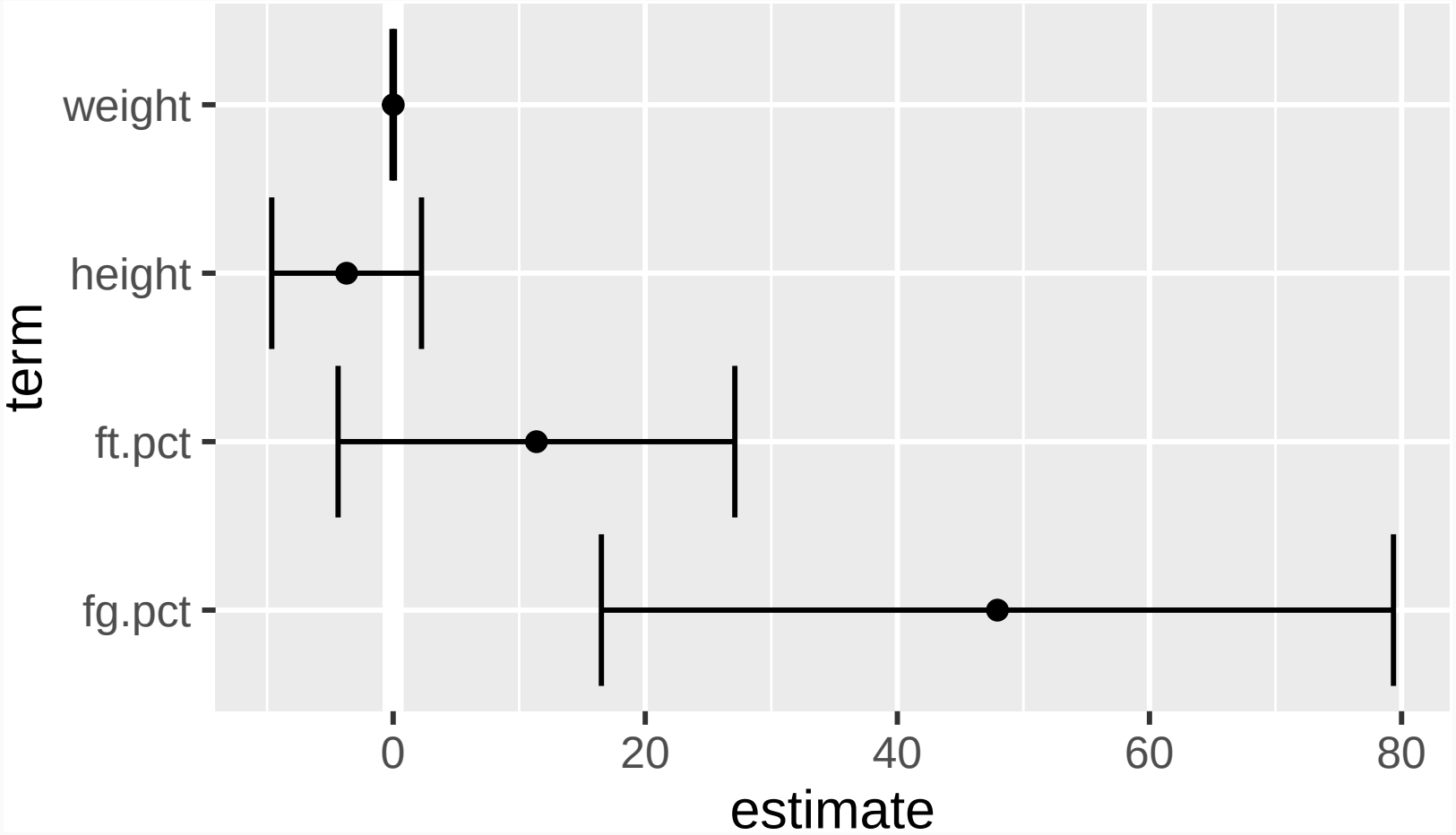
***p < .01; **p < .05; *p < .1

```r
library("broom")
output.1 = model.1 %>% tidy
```

term	estimate	std.error	statistic	p.value
(Intercept)	4.1487067	14.8550055	0.2792801	0.7812054
height	-3.6904991	2.9707798	-1.2422661	0.2200511
weight	0.0094585	0.0462972	0.2042986	0.8389664
fg.pct	47.9401992	15.7091307	3.0517411	0.0036685
ft.pct	11.3710193	7.8685361	1.4451251	0.1547880

```
p = ggplot(output.1, aes(x = term, y = estimate))
p = p + geom_hline(aes(yintercept = 0),  size = 2,
                   colour = "white") +
  geom_point() +
  geom_errorbar(aes(ymin=estimate-2*std.error,
                    ymax=estimate+2*std.error)) +
  coord_flip()
```

COEFFICIENT PLOT

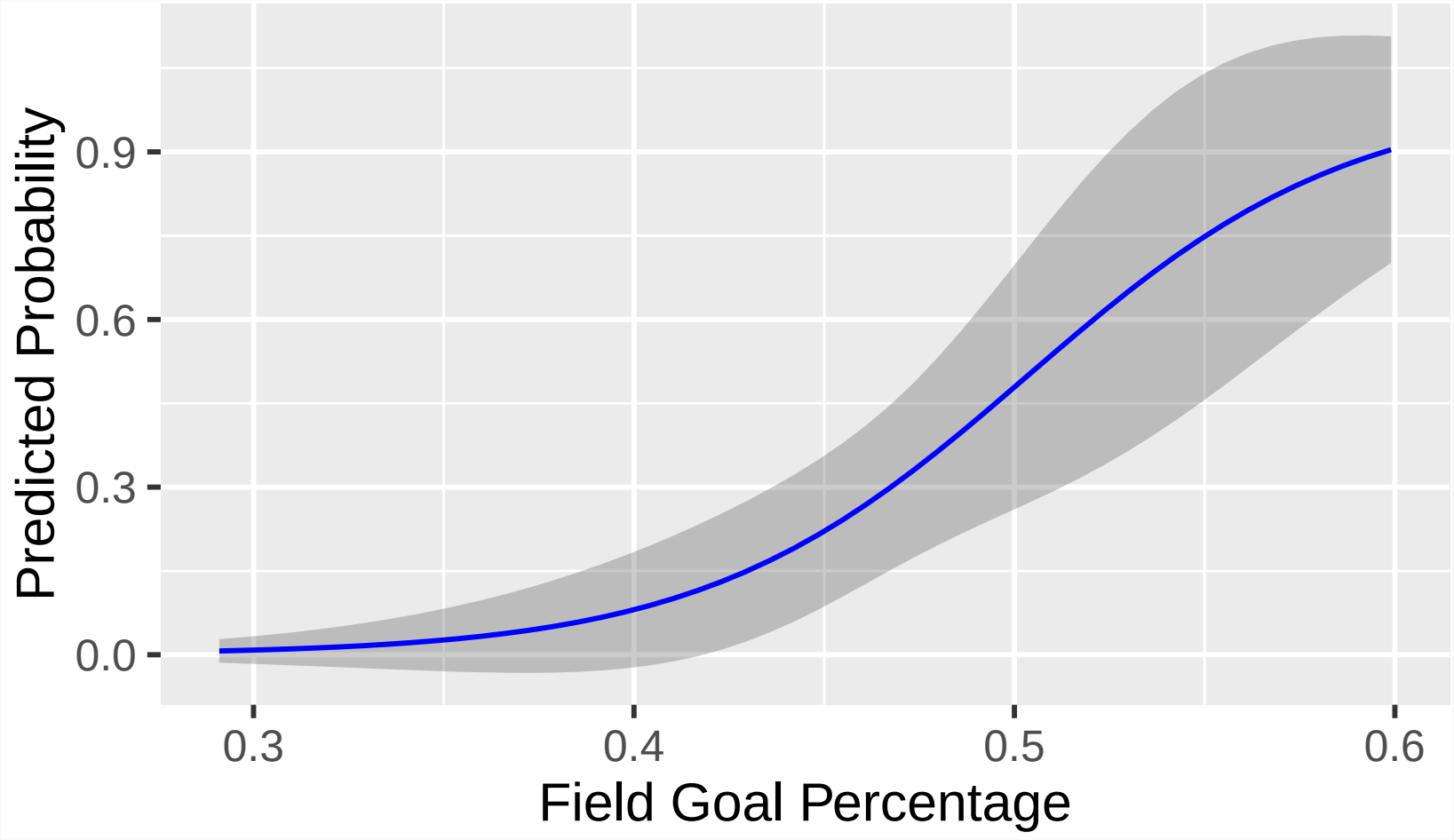


```
df = df %>% mutate(pts.high = pts > 13)
model.2 = glm(
  pts.high ~ height + weight + fg.pct + ft.pct,
  data = df,
  family = binomial(link = "logit"))
```

```
library("modelr")
df.plot = df %>%
  data_grid(fg.pct = seq_range(fg.pct, 50),
            .model = model.2)
preds = predict(model.2,
                newdata = df.plot,
                type = "response",
                se = TRUE)
df.plot$pred.full = preds$fit
df.plot$ymin = df.plot$pred.full - 2*preds$se.fit
df.plot$ymax = df.plot$pred.full + 2*preds$se.fit
```

fg.pct	height	weight	ft.pct	pred.full	ymin
0.2910000	6.65	212.5	0.7535	0.0067145	-0.0143930
0.2972857	6.65	212.5	0.7535	0.0077752	-0.0158256
0.3035714	6.65	212.5	0.7535	0.0090020	-0.0173481
0.3098571	6.65	212.5	0.7535	0.0104204	-0.0189527
0.3161429	6.65	212.5	0.7535	0.0120595	-0.0206270
0.3224286	6.65	212.5	0.7535	0.0139527	-0.0223526

```
p = ggplot(df.plot, aes(x = fg.pct, y = pred.full)) +
    geom_ribbon(aes(y = pred.full,
                    ymin = ymin,
                    ymax = ymax),alpha = 0.25) +
    geom_line(color = "blue")
```


Predicted Probability
0.9
0.6
0.3
0.0
0.3
0.4
0.5
0.6
Field Goal Percentage