Interpreting Regression Coefficients

Auguste Rodin, *The Thinker* (1903)

SML201: Introduction to Data Science, Spring 2020

Michael Guerzhoy

Linear Regression

```
##
## Call:
## lm(formula = lifeExp ~ log(gdpPercap) + continent, data = gap.1982)
##
## Coefficients:
##       (Intercept)     log(gdpPercap)  continentAmericas
##            12.264              5.342              7.299
##     continentAsia    continentEurope   continentOceania
##             6.473              9.563              9.534
```

- An increase of 1 in log(gdpPercap) is associated with an increase of 5.3 years in predicted life expectancy
  - An increase of 1 in log(gdpPercap) is an increase by a factor of 2.7 in gdpPercap
- All other things (i.e., gdpPercapita) being equal, we'll predict a country in Europe will have life expectancy that's (9.6-7.3 = 2.3) years higher than a country in the Americas
  - i.e., we'll predict LE that's 2.3 higher for a country that has the same gdpPercap, but is in Europe and not in the Americas

Association vs. Causation

- Does this mean that increases in GDP per capita *cause* increases in life expectancy?

Association vs. Causation

- Does this mean that increases in GDP per capita *cause* increases in life expectancy? Not necessarily.
  - Reverse causation: increased life expectancy causes increase in GDP per capita
  - Common cause: no natural/human-cause disaster cause both increased LE and increased GDP per capita, and are not distributed equally across continents
  - The causation goes both ways
  - It’s a coincidence

Log-Odds

$$p = \frac{1}{1 + e^{-(a_0 + a_1 x_1 + ..)}}$$

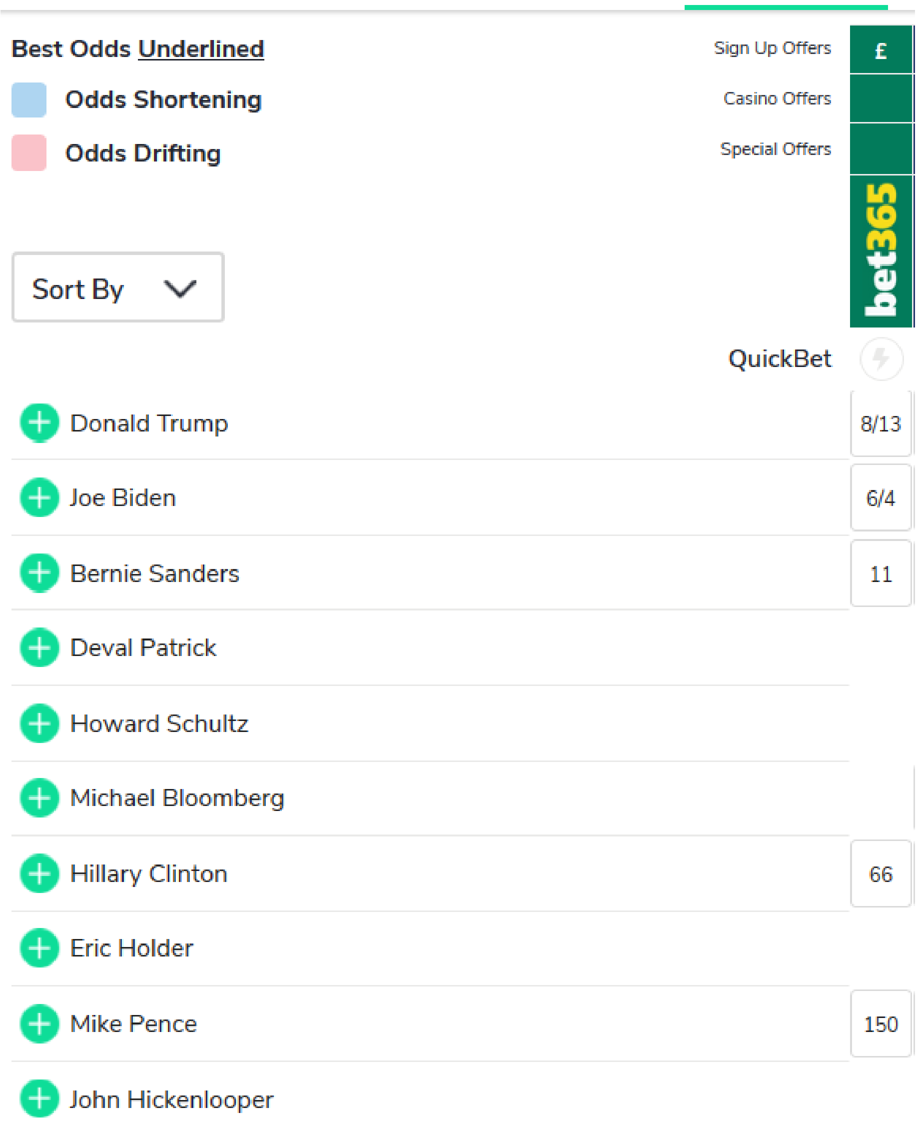
$$\Rightarrow \frac{1}{p} = 1 + e^{-(a_0 + a_1 x_1 + \cdots)}$$

$$\Rightarrow \log\left(\frac{1}{p} - 1\right) = -(a_0 + a_1 x_1 + \cdots)$$

$$\Rightarrow \log\left(\frac{p}{1-p}\right) = a_0 + a_1 x_1 + \cdots$$

"Log-odds" of the answer being 1

Odds of the answer being 1



You pay $4 if Biden doesn't win, and get $6 if he does

$$\frac{p}{1-p} = \frac{4}{6}$$
$$6p = 4(1-p)$$
$$(6+4)p = 4$$
$$p = \frac{4}{10}$$

Probability of Biden win assuming the odds are "fair" (i.e., player and house don't stand to win on average because the odds represent the probabilities)

Out of 10 times, Biden would win four times and not win six times

Logistic Regression

```
##
## Call:  glm(formula = Survived ~ Sex + Age, family = binomial, data = titanic)
##
## Coefficients:
## (Intercept)      Sexmale          Age
##     1.11388     -2.50000     -0.00206
##
## Degrees of Freedom: 886 Total (i.e. Null);  884 Residual
## Null Deviance:       1183
## Residual Deviance: 916   AIC: 922
```

- An increase of 1 year in Age is associated with a decrease of 0.002 in the log-odds of survival, all other things (i.e, sex) being the same
  - Corresponds to different things in terms of the change in probability
- All other things (i.e., age) being equal, men will have predicted log-odds of survival that are 2.5 lower
  - I.e., all other things being equal, the odds of survival for a woman are exp(2.5)=12 times higher
    - Note: not the same as the probability