Logistic Regression

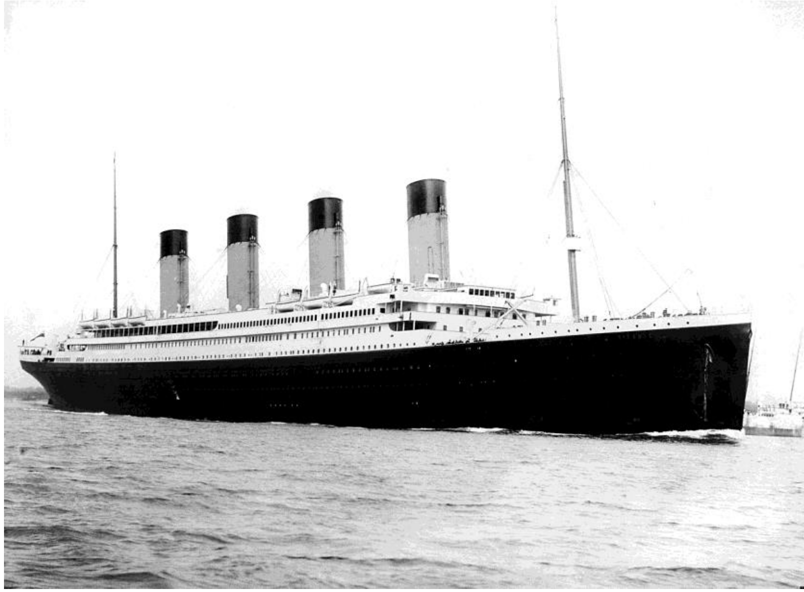
Some slides from Craig Burkett

STA303/STA1002: Methods of Data Analysis II, Summer 2016

Michael Guerzhoy

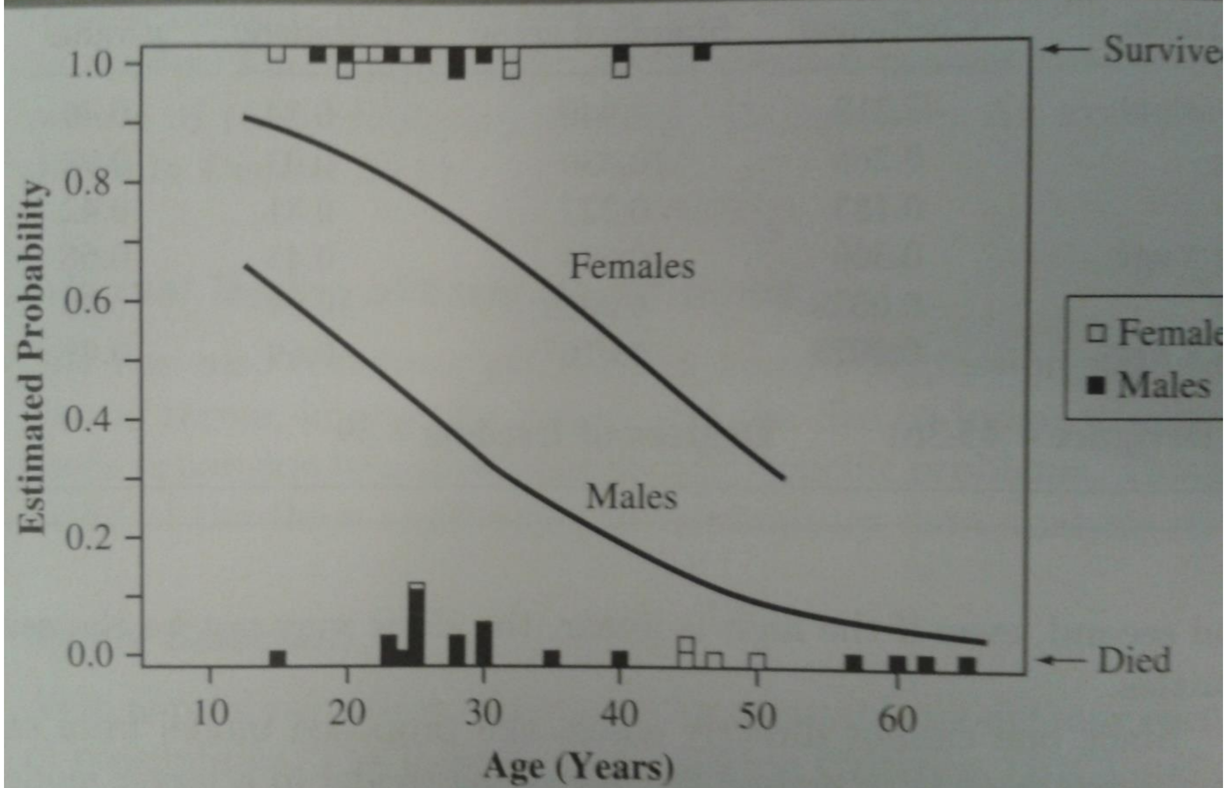
Titanic Survival Case Study

- The RMS *Titanic*
  - A British passenger liner
  - Collided with an iceberg during her maiden voyage
  - 2224 people aboard, 710 survived
- People on board:
  - 1st class, 2nd class, 3rd class passengers (the price of the ticket and also social class played a role)
  - Different ages
  - Different genders



Exploratory data analysis (in R)

explanatory variables



The Statistical Sleuth, 3rd ed

What's Wrong with Linear Regression?

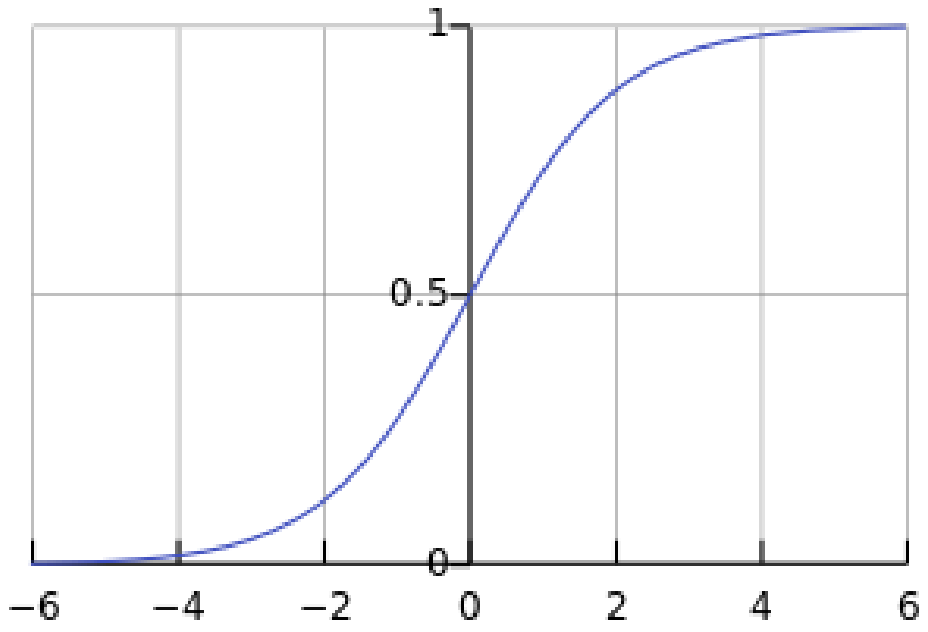
- $E[Y_i] = \beta_0 + \beta_1 X_1^{(i)} + \cdots + \beta_{k-1} X_{k-1}^{(i)}$
  $Y_i \sim Bernoulli(\pi_i)$ $(\pi_i = \pi(X_1, X_2 \dots))$
- We *can* match the expectation. But we'll have

$$Var[Y_i] = \pi_i(1 - \pi_i)$$

- Y_i is very far from normal (just two values)
- $Var[Y_i]$ is not constant
- Predictions for Y_i can have the right expectations, but will sometimes be outside of (0, 1)

Logistic Curve

$$s(y) = \frac{1}{1 + \exp(-y)}$$



Inputs can be in $(-\infty, \infty)$, outputs will always be in $(0, 1)$

Logistic Regression

- We will be computing

$$\frac{1}{1 + \exp(-\beta_0 - \beta_1 X_1 - \cdots - \beta_k X_k)}$$

to always get a value between 0 and 1

- Model:

$$Y_i \sim Bernoulli(\pi_i), \pi_i = \frac{1}{1+\exp(-\beta_0 - \beta_1 X_1^{(i)} - \cdots - \beta_k X_k^{(i)})}$$

- $Var(Y_i) = ?$

Log-Odds

$$\pi = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + ..)}}$$

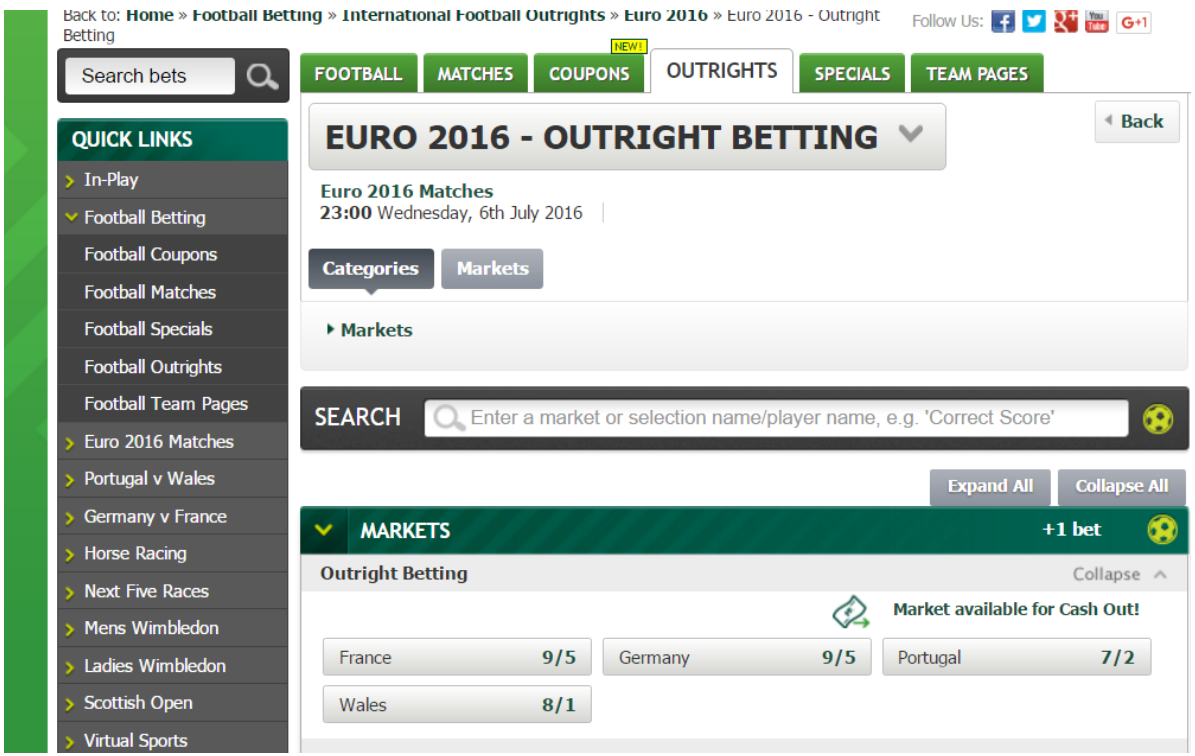
$$\Rightarrow \frac{1}{\pi} = 1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots)}$$

$$\Rightarrow \log\left(\frac{1}{\pi} - 1\right) = -(\beta_0 + \beta_1 x_1 + \cdots)$$

$$\Rightarrow \log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 x_1 + 000$$

Odds

- If the probability of an event is π, the odds of the event are $\frac{\pi}{1-\pi}$



Odds

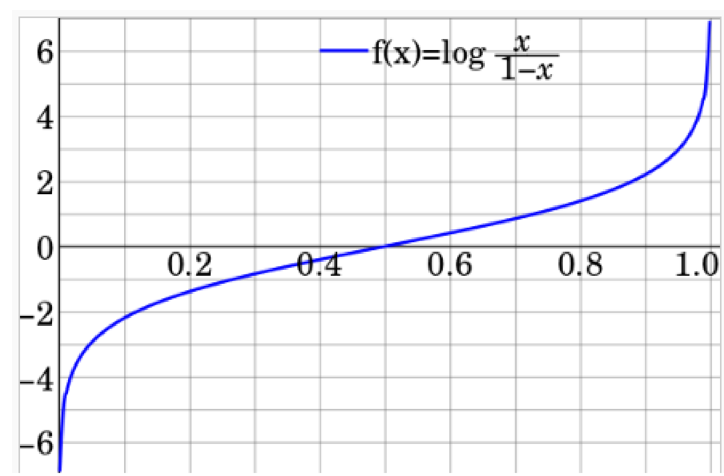
France	9/5	Germany	9/5	Portugal	7/2
Wales	8/1				

- You pay \$5 if France don't win, and get \$9 if they do.
- What's the probability of France winning assuming "true odds" are offered?

- (What about if the odds r=p/(1-p) are given as 0.6?)

The Log-Odds are Linear in X

- $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots$
- Generalized linear model: $g(\mu) = X\beta$, with a distribution imposed on Y
    - $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ with $Y \sim Bernoulli(\mu)$ is Logistic regression
    - g is the *link function*
    - $logit(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$



Maximum Likelihood

$Y_i \sim Bernoulli(\pi_i)$

$P(Y_i = y_i | \beta_0, \beta_1, \dots, \beta_{k-1}) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

$P(Y_1 = y_1, \dots, Y_n = y_n | \dots) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

$\log P(Y_1 = y_1, \dots, Y_n = y_n | \dots) = \sum_{i=1}^{n} (y_i \log \pi_i + (1-y_i)\log(1-\pi_i))$

$$\pi_i = \frac{1}{1+\exp(\beta_0 + \beta_1 X_1^{(i)} + \cdots + \beta_{k-1} X_{k-1}^{(i)})}$$

Now, take the derivative wrt the betas, and find the betas that maximize the log-likelihood….

Titanic Analysis

- (in R)

Interpretation of $\boldsymbol{\beta_0}$ - Linear Reg.

- In Linear Regression, if there are no other predictors, the least-squares (and Maximum Likelihood) estimate of Y is the mean
  - $E[Y] = \beta_0 \Rightarrow \bar{Y} = b_0$

```
y <- rnorm(100, 10, 25)
> summary(lm(y ~ 1))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.727      2.524   3.458 0.000804

> mean(y)
[1] 8.726752
```

Interpretation of $\boldsymbol{\beta_0}$ - Logistic Reg.

- MLE for $\pi = P(Y = 1)$ is also the sample mean $\bar{Y}$ (proportion of the time Y is 1/the marginal probability that Y is 1)
- In Logistic Regression, without predictors we have

$$\pi = \frac{1}{1 + e^{-\beta_0}}$$

$$\beta_0 = logit(\bar{Y})$$

- (in R)

Interpretation of $\boldsymbol{\beta_0}$ - Logistic Reg.

- Now, back to predicting with age. What's the interpretation of β_0 now?
- (in R)

Interpretation of $\boldsymbol{\beta_1}$ (Categorical Predictor)

```
fit <- glm(survived ~ sex, family= binomial, data=
titan)
> summary(fit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.44363    0.08762  -16.48   <2e-16
sexfemale    2.42544    0.13602   17.83   <2e-16

> exp(coef(fit))
(Intercept)   sexfemale
  0.2360704  11.3071844
```

- Interpretation: the odds of survival for women are 11.3 times higher than for men

Interpretation of $\boldsymbol{\beta}$ (Mixed Predictors)

$$logit(\pi) = \beta_0 + \beta_1 \cdot age + \beta_2 \cdot I_{female}$$

- In Linear Regression, β_1 would the increase in survival for a unit change in *age,* keeping other variables constant
- β_2 is the difference between group means, keeping other variables constant
- In logistic regression, β_1 is the increase in the log-Odds of survival, for a unit change in *age,* keeping other variables constant
- β_2 is the increase in the log-Odds of survival for women compared to men, keeping age constant

Quantifying Uncertainty

- (in R)

Interpreting Coefficients

```
fit <- glm(survived ~ age + sex, family= binomial,
data= titan)
> exp(coef(fit))
(Intercept)           age    sexfemale
  0.2936769     0.9957548   11.7128810
> exp(confint(fit))
Waiting for profiling to be done...
                2.5 %      97.5 %
(Intercept) 0.2037794  0.4194266
age         0.9855965  1.0059398
sexfemale   8.7239838 15.8549675
```

Controlling for age, we are 95% confident that females have a 772% to 1485% higher odds of survival, compared to males

Likelihood Ratio Test

- Can be used to compare any two nested models models
- Test statistic:
  - $LRT = \ 2\log(LMAX_{full}) - 2\log(LMAX_{reduced})$
    - Has a χ^2 distribution when the extra parameters in $LMAX_{full}$ are 0
  - LMAX: the maximum likelihood value
- A lot of the time what's computed is
  - $deviance = const - 2\log LMAX$
- Then:
  - $LRT = deviance_{reduced} - deviance_{full}$

Wald Test

- The test based on approximating the sampling distribution of the coefficients as normal
  - The SE’s that are show shown when you call summary
- Unreliable for “small” sample sizes

Model Assumptions

- Independent observations
- Correct form of model
  - Linearity between logits & predictor variables
  - All relevant predictors included
- For CIs and hypothesis tests to be valid, need large sample sizes

Titanic Dataset: Model Checking

- Independent observations?
  - Not really, for one thing, they were all on one ship!
- Large sample?
  - Yes

Titanic Dataset: Other Worries

- Do we have all relevant predictors?
  - I.e., might there be confounding variables we haven’t considered/don’t have available?