THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Logistic Regression 1

Lecture 22

STA 371G

Announcements

- You'll get Midterm 2 back later this week (a few students still have to take the test).

Announcements

- You'll get Midterm 2 back later this week (a few students still have to take the test).
- Get started on your project with your group as soon as you can, so you have plenty of time to iron out any issues.

Near, far, wherever you are....

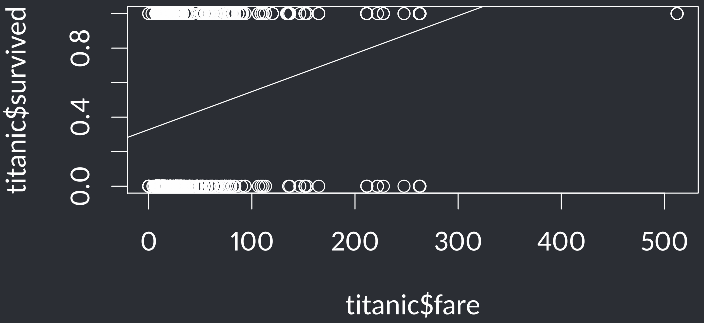
- How much did the ticket price affect whether someone survived the Titanic?
- We have a data set of 1045 passengers on the Titanic; 427 survived
  - **fare**: the amount of money paid for the ticket
  - **survived**: a dummy variable indicating whether the passenger surived (1) or not (0)

Near, far, wherever you are....

- How much did the ticket price affect whether someone survived the Titanic?
- We have a data set of 1045 passengers on the Titanic; 427 survived
  - **fare**: the amount of money paid for the ticket
  - **survived**: a dummy variable indicating whether the passenger surived (1) or not (0)
- What is unusual about this data set?

What goes wrong with linear regression?

```r
plot(titanic$fare, titanic$survived)
model <- lm(survived ~ fare, data=titanic)
abline(model)
```



The idea behind logistic regression

- Instead of predicting whether someone survives, let's predict the *probability* that they survive
- 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
- 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(survived ~ fare, data=titanic,
             family=binomial)
summary(model)
```

How to interpret the curve?

The regression output tells us that our prediction is

$$\log\left(\frac{P(\text{survival})}{1 - P(\text{survival})}\right) = -0.794 + 0.012 \cdot \text{fare}.$$

How to interpret the curve?

The regression output tells us that our prediction is

$$\log\left(\frac{P(\text{survival})}{1 - P(\text{survival})}\right) = -0.794 + 0.012 \cdot \text{fare}.$$

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

$$P(\widehat{\text{survival}}) = \frac{e^{-0.794 + 0.012 \cdot \text{fare}}}{1 + e^{-0.794 + 0.012 \cdot \text{fare}}}$$

Making predictions

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

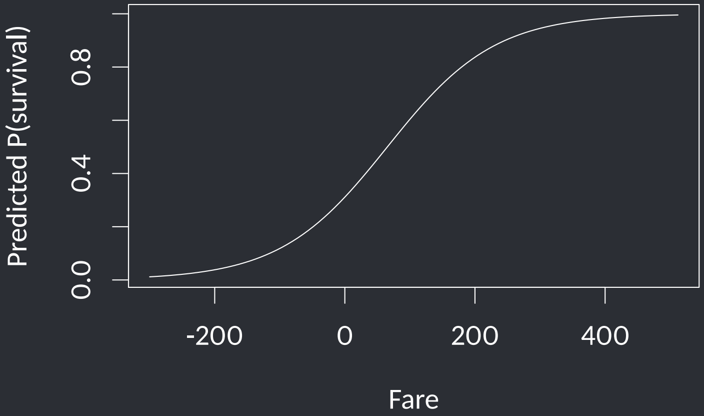
```
predict(model, list(fare=50), type="response")

    1
0.453
```

$$\frac{e^{-0.794+0.012\cdot 50}}{1+e^{-0.794+0.012\cdot 50}} = 0.453$$

How to interpret the curve?

$$\widehat{P(\text{survival})} = \frac{e^{-0.794+0.012\cdot\text{fare}}}{1+e^{-0.794+0.012\cdot\text{fare}}}$$



Interpreting the coefficients

Our prediction equation is:

$$\log\left(\frac{P(\text{survival})}{1 - P(\text{survival})}\right) = -0.794 + 0.012 \cdot \text{fare}.$$

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

- When fare $= 0$, we predict that the log odds will be -0.794

Interpreting the coefficients

Our prediction equation is:

$$\log\left(\frac{P(\text{survival})}{1 - P(\text{survival})}\right) = -0.794 + 0.012 \cdot \text{fare}.$$

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

- When fare $= 0$, we predict that the log odds will be -0.794

Interpreting the coefficients

Our prediction equation is:

$$\log\left(\frac{P(\text{survival})}{1 - P(\text{survival})}\right) = -0.794 + 0.012 \cdot \text{fare}.$$

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

- When fare $= 0$, we predict that the log odds will be -0.794 , so the probability of survival is predicted to be 31%.
- When fare increases by £1, we predict that the log odds will increase by 0.012.

Interpreting the coefficients

Let's rewrite the prediction equation as:

$$\text{Predicted odds of survival} = e^{-0.794+0.012\cdot\text{fare}}$$

Increasing the fare by £1 will *multiply* the odds by $e^{0.012} = 1.012$; i.e., increase the odds by 1.2%.

Interpreting the coefficients

Let's rewrite the prediction equation as:

$$\text{Predicted odds of survival} = e^{-0.794+0.012\cdot\text{fare}}$$

Increasing the fare by £1 will *multiply* the odds by $e^{0.012} = 1.012$; i.e., increase the odds by 1.2%.

Increasing the fare by £10 will *multiply* the odds by $e^{10\cdot 0.012} = 1.129$; i.e., increase the odds by 12.9%.

Testing the null hypothesis

As in regular linear regression, the overall null hypothesis is that $\beta_1 = 0$; we can test this by using the *p*-value for that variable on the output.

Testing the null hypothesis

As in regular linear regression, the overall null hypothesis is that $\beta_1 = 0$; we can test this by using the p-value for that variable on the output.

Since p is very small, we can reject the null hypothesis that $\beta_1 = 0$; i.e., there is a statistically significant relationship between fare and survival.

How good is our model?

- Unfortunately, the typical R^2 metric isn't available for logistic regression.

How good is our model?

- Unfortunately, the typical R^2 metric isn't available for logistic regression.
- However, there are many "pseudo-R^2" metrics that indicate model fit.

Take 1: How many cases did we accurately predict?

We could use our model to make a prediction of survival (or not), based on the probability. Suppose we say that our prediction is:

$$\text{Prediction} = \begin{cases} \text{survival}, & \text{if } P(\widehat{\text{survival}}) \geq 0.5, \\ \text{tragedy}, & \text{if } P(\widehat{\text{survival}}) < 0.5. \end{cases}$$

Take 1: How many cases did we accurately predict?

We could use our model to make a prediction of survival (or not), based on the probability. Suppose we say that our prediction is:

$$\text{Prediction} = \begin{cases} \text{survival}, & \text{if } P(\widehat{\text{survival}}) \geq 0.5, \\ \text{tragedy}, & \text{if } P(\widehat{\text{survival}}) < 0.5. \end{cases}$$

Now we can compute the fraction of people whose survival we correctly predicted:

```r
predicted.survival <- (predict(model, type='response') >= 0.5)
actual.survival <- (titanic$survived == 1)
sum(predicted.survival == actual.survival) / nrow(titanic)
```

```
[1] 0.644
```

Take 1: How many cases did we accurately predict?

64.4% sounds pretty good—what should we compare it against?

Take 1: How many cases did we accurately predict?

64.4% sounds pretty good—what should we compare it against?

We should compare 64.4% against what we would have gotten if we just predicted the most common outcome (not surviving) for everyone, without using any other information:

Take 1: How many cases did we accurately predict?

64.4% sounds pretty good—what should we compare it against?

We should compare 64.4% against what we would have gotten if we just predicted the most common outcome (not surviving) for everyone, without using any other information:

```
1 - sum(actual.survival) / nrow(titanic)
```

```
[1] 0.591
```

Take 2: McFadden's pseudo-R^2

To get a metric on the usual 0-1 scale (like regular R^2), McFadden's pseudo-R^2 can be used (this is what is described in the reading):

$$\text{pseudo-}R^2 = 1 - \frac{\text{residual deviance}}{\text{null deviance}} = 1 - \frac{1339.941}{1413.571} = 0.052$$

```
Call:
glm(formula = survived ~ fare, family = binomial, data = titanic)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-2.231  -0.928  -0.898   1.335   1.528

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.7941     0.0844   -9.41  < 2e-16 ***
fare          0.0121     0.0017    7.13  9.9e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1413.6  on 1044  degrees of freedom
Residual deviance: 1339.9  on 1043  degrees of freedom
AIC: 1344

Number of Fisher Scoring iterations: 4
```

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