Intro to Logistic Regression

DAT 280

Dr. Alex Marsella

2024-03-13

Introduction

- Logistic Regression is a statistical technique used for binary (1 or 0) outcome variables.
- It predicts the probability of occurrence of an event by fitting data to what's called a "logit" function.
- Ideal for situations where the outcome is binary (e.g., yes/no, true/false).

Why Logistic Regression?

- **Binary Outcomes**: Suited for models with an outcome (Y) that takes on only two values.
- **Probabilities**: Provides us probabilities that Y takes on one value as opposed to the other, given some change in X.
- **Interpretability**: Easy to understand and explain the results once we do a little mathematical conversion to the coefficients.

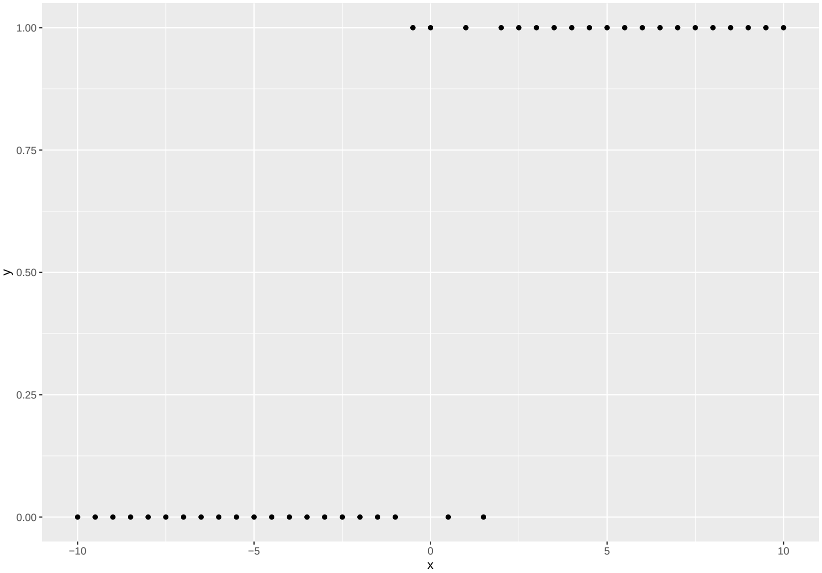
The Logistic Function

- Logistic regression uses the logistic function to model probabilities.
- The function maps any real-valued number into a value between 0 and 1.
- Formula: $P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\ldots\beta_k X_k)}}$
  - Note that the left-hand side of the formula is the **probability that Y equals 1**.
  - That is what we will be measuring and modelling as a function of some X variables.
- Logistic regression doesn't give us direct estimates of Y, it gives us probabilities of Y = 1, given some values of X.

Data where Y is binary

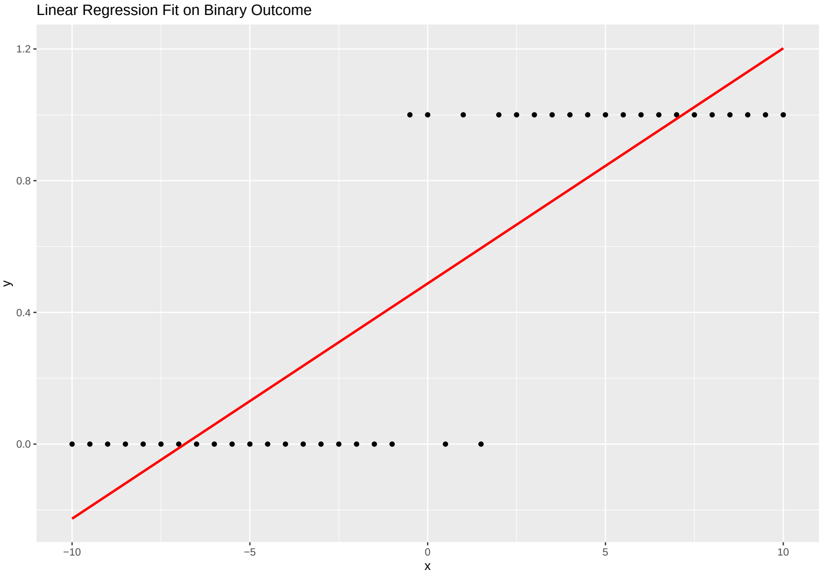
- Imagine a scenario where the outcome you are studying takes on one of two values.
  - "Will you vote in the next election?"
  - "Are you above the age of 65?"
  - "Do you smoke cigarettes?"
  - All questions with a Yes or No answer.
- We, (or the computer automatically), treat this variable as taking on a value of 1 or 0.
  - Usually, "Yes" is 1 and "No" is 0.

Data where Y is binary

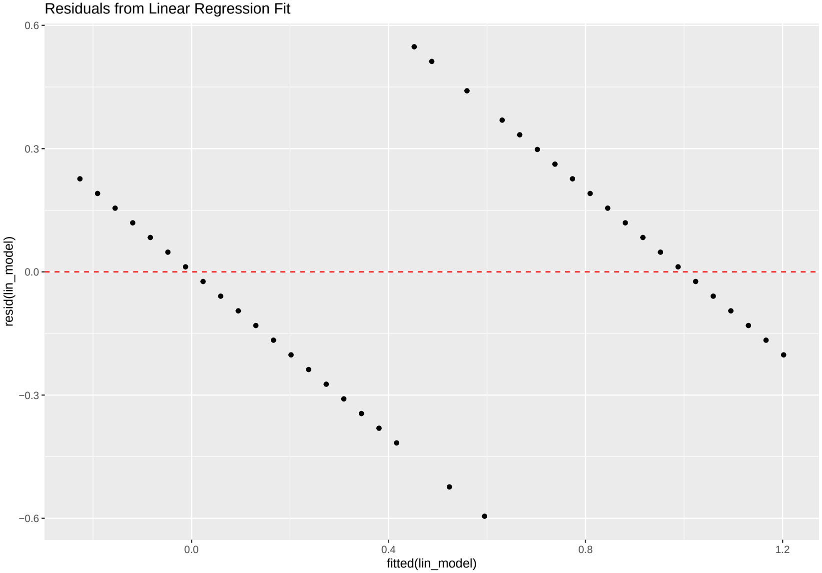


Linear Regression just won't do.

- Look what happens when we try to fit a line to this kind of data.

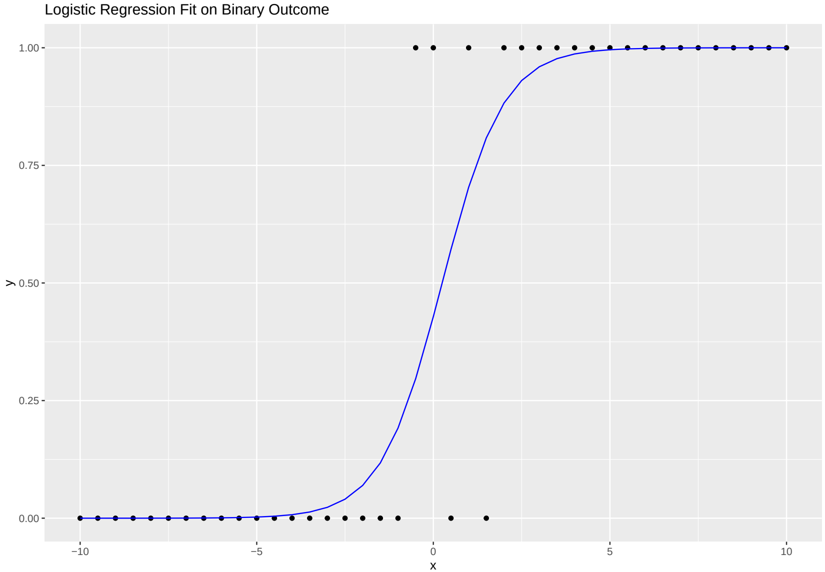


Look how bad these residuals are.



Logistic Regression

- Instead, we can use logistic regression.



Residuals from Linear Regression Fit

Example: Titanic Survival Prediction

- **Dataset**: Passenger data from the Titanic.
- **Objective**: Predict a passenger's likelihood of survival.
  - Survival either occurred or didn't! Our outcome variable is binary, 1 for "Survived" and 0 for "did not survive".
- **Features**: Class, sex, age, etc.

Data Exploration

```r
library(titanic)
data(titanic_train)
head(titanic_train)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                    Name
## 1                               Braund, Mr. Owen Harris
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) fe
## 3                                Heikkinen, Miss. Laina fe
## 4       Futrelle, Mrs. Jacques Heath (Lily May Peel) fe
## 5                              Allen, Mr. William Henry
## 6                                      Moran, Mr. James
##             Ticket    Fare Cabin Embarked
```

Fitting a Logistic Regression Model

- We use command `glm()` instead of `lm()` and set an option of `family = "binomial"`.
  - In statistics, a "binomial" is a variable that takes on one of two values. A *BI**NARY variable!

```
model <- glm(Survived ~ Pclass + Sex + Age,
             data = titanic_train, family = "binomial")
```

The model output.

- We don't interpret these the same way, though! This is not linear regression!

```r
summary(model)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age, family = "b
##     data = titanic_train)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.056006   0.502128  10.069  < 2e-16 ***
## Pclass      -1.288545   0.139259  -9.253  < 2e-16 ***
## Sexmale     -2.522131   0.207283 -12.168  < 2e-16 ***
## Age         -0.036929   0.007628  -4.841 1.29e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
```

STOP! Don’t try to interpret these numbers like you do with linear regression!

Positive vs. Negative Coefficients

- **Positive Coefficient ($\beta > 0$)**: Indicates that as the predictor increases, the probability of the outcome occurring increases.
- **Negative Coefficient ($\beta < 0$)**: Indicates that as the predictor increases, the probability of the outcome occurring decreases.

Understanding the Coefficients more deeply.

- **Log Odds**: The raw coefficients displayed represent the change in the log odds of the outcome for a one-unit increase in the predictor, holding all other predictors constant.

We convert the Log Odds to

- **Odds Ratio**: Exponentiating a coefficient (e^{β}) gives us the odds ratio (OR), a more intuitive measure of the predictor's effect.
  - An odds ratio measures "how many times as likely" something is.
  - For example, OR $= 1.7$ means increasing X by 1 is associated with $Y = 1$ being 1.7 times as likely than if X were not increased by 1.
  - **\textcolor{red}{Or, even easier, increasing X by 1 makes it Y=1 70% more likely to occur.}**

Odds Ratio (OR)

- **OR > 1**: A one-unit increase in the predictor is associated with an increase in the odds of the outcome.
- **OR < 1**: A one-unit increase in the predictor is associated with a decrease in the odds of the outcome.
- **OR $= 1$**: The predictor has no effect on the odds of the outcome.

Example Interpretation

Imagine a logistic regression model where the coefficient for age (β_{age}) is 0.05. The odds ratio is $e^{0.05} \approx 1.05$.

- This means for each additional year of age, the odds of the outcome (e.g., having a certain disease) increase by 5%, assuming all other factors are held constant.

Going back to Titanic

- "Sexmale" had a coefficient of -2.52.
  - "Sexmale" = 1 for males and 0 for females.
  - First takeaway: What does the sign on the coefficient mean?
- We need to "exponentiate" to turn it into an Odds Ratio.
  - $e^{\beta} = e^{-2.52} = 0.08$
  - Notice that a negative coefficient will always return a value less than 1 since e^0 = 1.
- Odds Ratio = 0.08
  - What does this mean?
  - Recall: An odds ratio measures "how many times as likely" something is or, the odds ratio minus 1 and converted to percent tells you "how much more likely something is".

Exercise: More interpreting

- Please convert the other two variables in our model to Odds Ratios and **write down an interpretation.**
- Work with the people around you, you will need a calculator (you can use R as a calculator!)
  - In R code, e^x is `exp(x)`.
- I will be around to help if you get stuck.