Logistic Regression

Predicting whether a police stop results in a frisk

“The New York City stop-and-frisk policy allowed police to stop and search individuals when any officer had "reasonable suspicion" — a standard that is lower than the "probable cause" needed to justify an arrest. …

Civil rights activists, such as the Center for Constitutional Rights, argued that this stop-and-frisk policy led to the New York City Police Department (NYPD) unfairly targeting people of color. Recent New York leaders such as Mayor Bill de Blasio made changes to that policy, decreasing the frequency of stop-and-frisks happening on a daily basis.”

https://dataspace.sites.grinnell.edu/nypd1.html

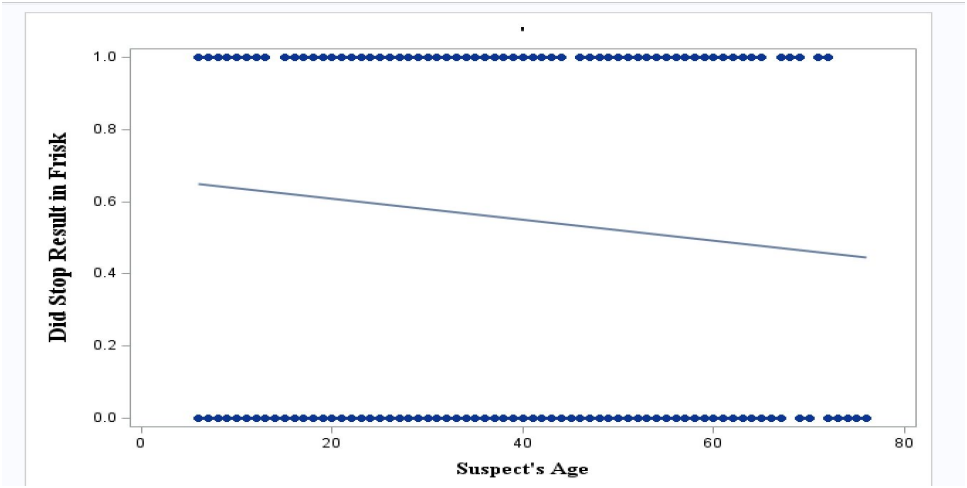
The response variable is binary: it takes two values

Y = 1 if the police stop results in a frisk
Y = 0 if the police stop does not result in a frisk

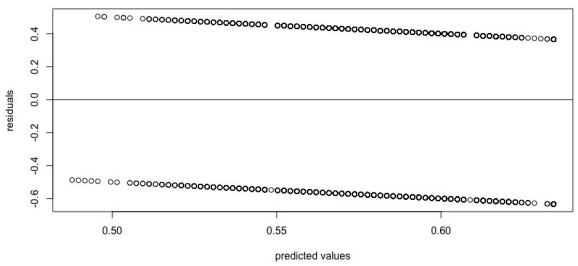
One explanatory variable, suspect’s age, is quantitative.
We will later look at the suspect’s sex and race as categorical explanatory variables.

problems?
- response b/w 0+1
- < 0
- residuals look large
- > 1
- predictions seem weird

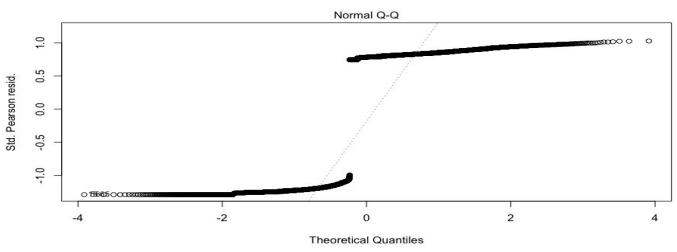
What if we treat the *binary* response as *quantitative*? i.e., what if we fit a simple linear regression model?

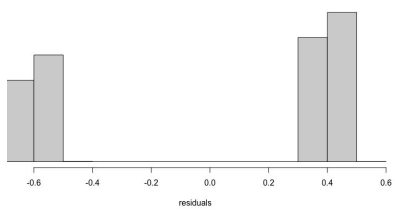


What if we treat the *binary* response as *quantitative*? i.e., what if we fit a simple linear regression model?



conditions don't seem to be met





Modeling the Mean Response

In linear regression, we model the mean response:

$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \varepsilon \sim N(0, \sigma_\varepsilon)$

μ_Y

For a binary variable, what does the sample "mean" represent?

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{6072}{6072+4122} = 59.6\%$

Status	
Frisk (Y = 1)	No Frisk (Y = 0)
6072	4122

- In logistic regression, we use the sample proportion to model the the probability of "success", which changes as a function of X.

Using a transformation to fix problem

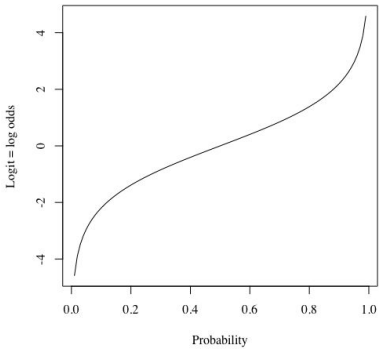
In logistic regression, instead of constructing a model for Y, we construct a model for a function of the prob. of success

- Let π be the probability of "success" frisk given stop
- Since π is a probability, it can take values from 0 to 1.
- Possible model: $\pi = \beta_0 + \beta_1 X$
  - Problem: $\hat{\pi} < 0 \quad \hat{\pi} > 1$
- Second possible model: Instead of modeling the probability π directly, model the "odds" of success
- $\frac{\pi}{1-\pi} = \beta_0 + \beta_1 X$
  - Range of values: 0 to ∞
  - Problem: $\hat{\pi}/1-\hat{\pi} < 0$

The Logit Transformation

The **logit function** is defined as

$$\text{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$$



i.e., the logit is the log of the odds of success.

Range of values of logit: $-\infty$ to ∞

The Logistic Regression model is given by $\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$

The Logistic Regression Model

$$\text{Logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

- The logit is a one-to-one transformation, i.e.,
  - For every value of π (with the exception of 0 and 1), there is one, and only one, value of the logit
  - For every value of the logit (with the exception of $-\infty$ and $+\infty$), there is one, and only one, value of π

Probability Form: We can rewrite our model so that the *probability of success* (π) is a function of the model parameters and the predictor X.

$$P(Y=1) = \pi = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Four Probabilities (definitions on pg. 419)

Let $P(success)$ = the probability of success. There are four version of this probability of interest:

	TRUE Value	FITTED Value
Actual Probability	p = true P(success) for given x	$\hat{p}$ = # success / sample size for given x
Model Probability	π = true P(success) from this model	$\hat{\pi}$ = estimated P(success) from model

If the model is exactly correct, then $p = \pi$

Interpreting The Logistic Regression Parameters

In the logit form:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

We are modeling the log-odds of "success":

An age increase of 1 year for a suspect is associated with a log-odds of the police stop resulting in a frisk increase by $\hat{\beta}_1$ on average

We can calculate the "odds" of success by:

$$\frac{\pi}{1-\pi} = e^{\beta_0 + \beta_1 X}$$

As X increases by 1 unit, the odds of success increases e^{β_1} times.

Interpretation

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.3055	0.070	-4.346	0.000	-0.443	-0.168
arsenic	0.3791	0.039	9.840	0.000	0.304	0.455

$e^{0.3791} = 1.46$

As arsenic levels increase by 1 unit, the odds of switching increases 1.46 times on avg.

Test and Confidence Interval for β_1

In the Simple Logistic Regression model,

$$\ln\left(\frac{\pi}{1-\pi}\right) = \text{logit}(\pi) = \beta_0 + \beta_1 X$$

The null and alternative hypotheses are given by:

$$H_0: \beta_1 = 0 \quad \text{v.} \quad H_a: \beta_1 \neq 0$$

The test statistic has the form:

$$Z = \frac{\hat{\beta}_1 - 0}{SE_{\hat{\beta}_1}}$$

use Normal distn

The confidence interval for the β_1 :

$$\hat{\beta}_1 \pm Z^* SE_{\hat{\beta}_1}$$

Checking the Conditions

Conditions for Logistic Regression:

1. Linearity: The logit is a linear function of the explanatory variable X

2. Independence: The observations are not related to each other – no paired data, no clustered (grouped) data

3. Random: The data are obtained via a random process, either through (a) a random sample was obtained, or (b) random assignment to groups.

Conditions NOT required:

1. Constant Variance: Variance will NOT be constant!!
2. Normality: Population and sample will NOT be Normal…

Interpretation

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.0027	0.079	0.035	0.972	-0.153	0.158
arsenic	0.4608	0.041	11.134	0.000	0.380	0.542
dist100	-0.8966	0.104	-8.593	0.000	-1.101	-0.692

$e^{0.4608} = 1.59$

$e^{-0.8966} = 0.41$

As the arsenic level increases the associated odds of switching increase by 1.59 times on avg.

As the distance to nearest well increases the associated odds of switching decreases by $100(1-0.41)\% = 59\%$ on avg.