GLM

Advanced regression

Chuvakin Sergey

«School of Advanced Studies»

January 13, 2021

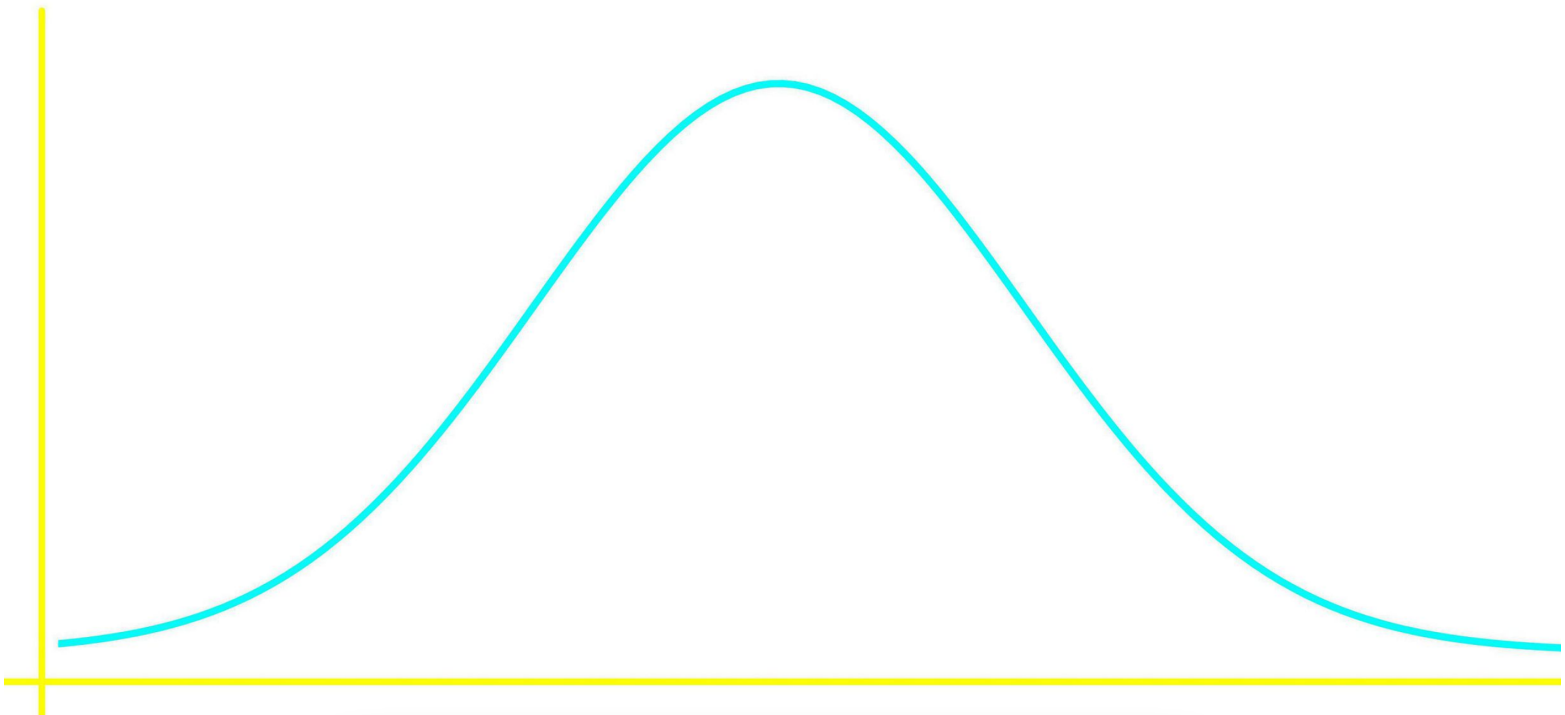
Outline

- Assumptions
- Normal distribution
- types
- link function
- logit
- estimators
- probit (normal distribution)
- interpretation

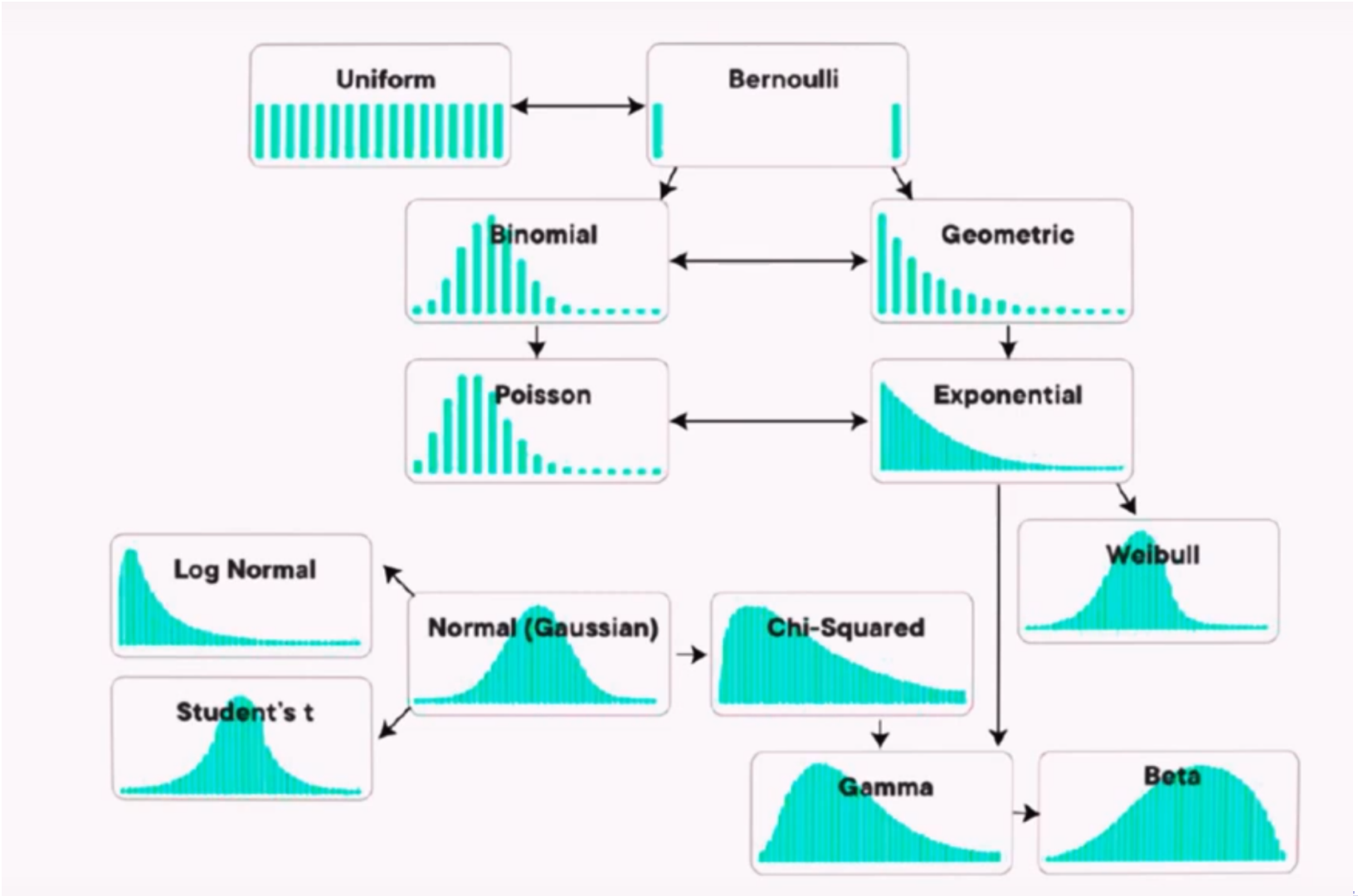
Assumptions

- Linearity of data
- Sample should be *randomly* selected for population
- X matrix should not be correlated within
- X marrix should not be correlated with error
- Variance of error should be constant
- *Normality of Y*

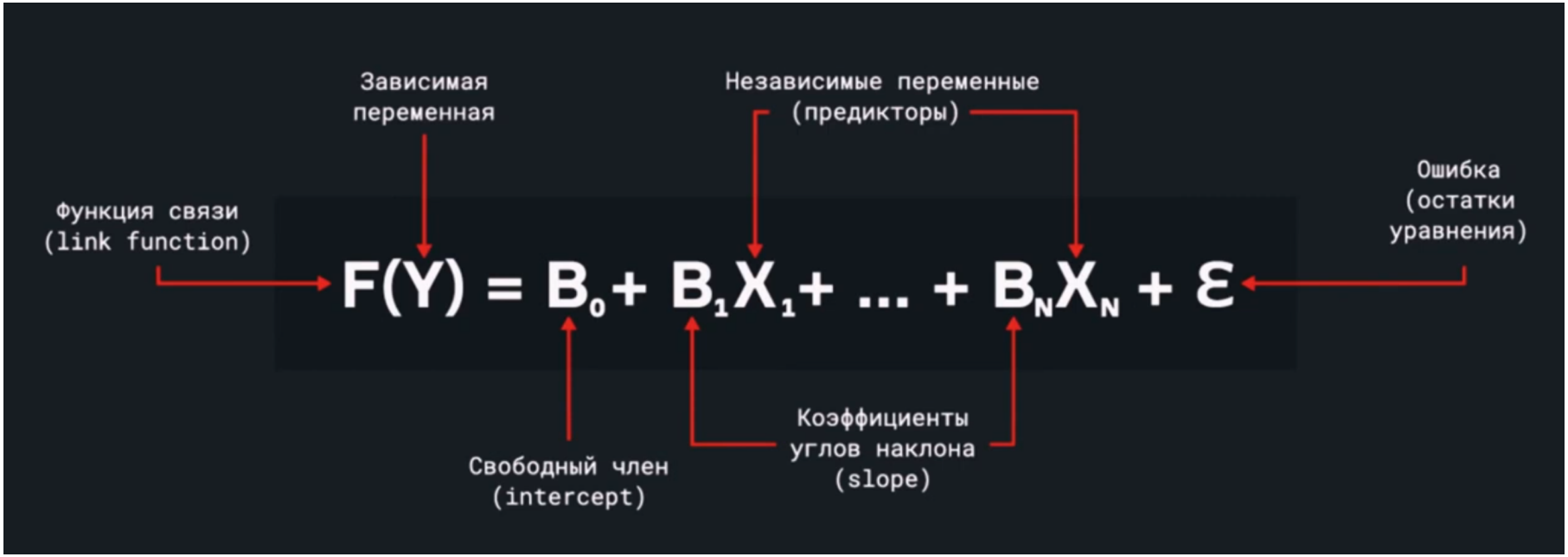
Normall Distribution



Types


Uniform
Bernoulli
Binomial
Geometric
Poisson
Exponential
Weibull
Log Normal
Normal (Gaussian)
Chi-Squared
Student's t
Gamma
Beta

Idea


Зависимая переменная
Независимые переменные (предикторы)
Функция связи (link function)
F(Y) = B₀ + B₁X₁ + ... + BₙXₙ + ε
Ошибка (остатки уравнения)
Свободный член (intercept)
Коэффициенты углов наклона (slope)

Link functons

- Identity
- log (logit)
- probit
- poisson
- negative binomial
- etc

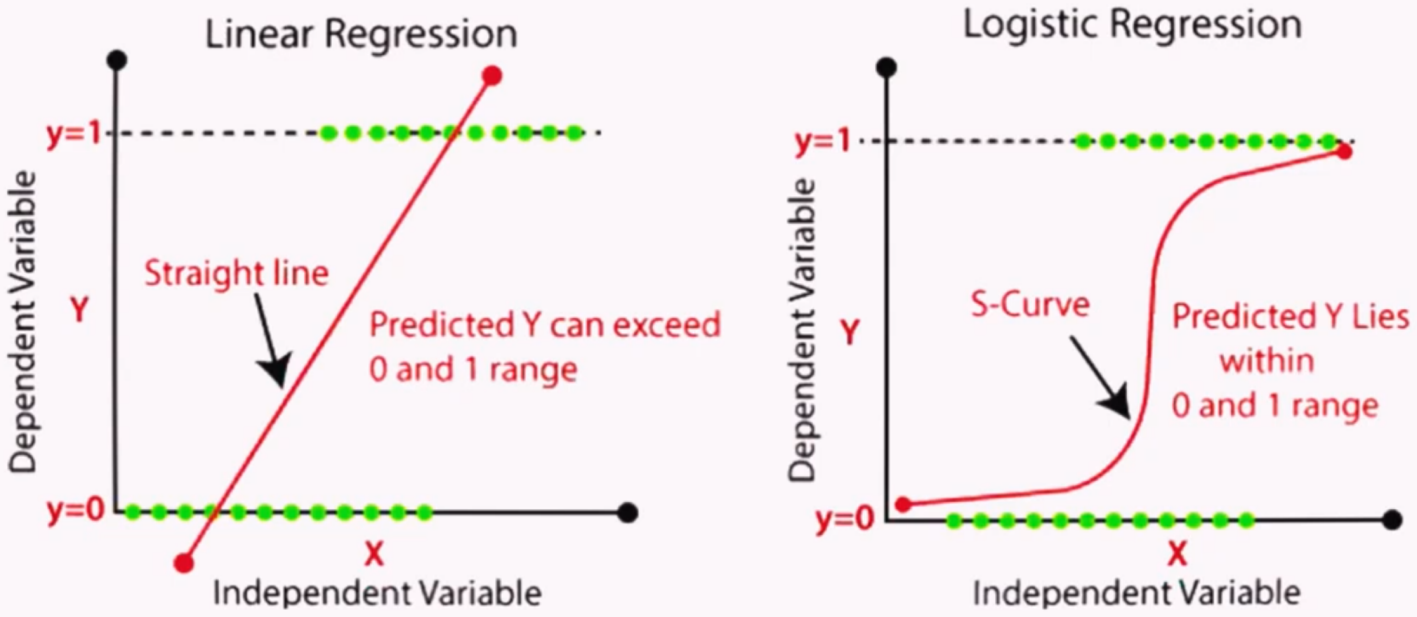
Logit

This type of regression suits for binary varaible.

$$Y = log(\frac{p}{1-p}) \tag{1}$$

Coef output not an absolute straightforward number to interpret - it's a chance.
Chance is a probability relation. 1 means that there is equal probability for success and for fail.

Logit


Linear Regression
y=1
y=0
Dependent Variable
Y
Straight line
Predicted Y can exceed
0 and 1 range
X
Independent Variable
Logistic Regression
y=1
y=0
Dependent Variable
Y
S-Curve
Predicted Y Lies
within
0 and 1 range
X
Independent Variable

Estimator

MLE - Maximul likelihood Estimator more suits to logit and other GLM

$$\text{likelihood} = \hat{y} * y + (1-\hat{y}) * (1-y) \tag{2}$$

$$\text{log-likelihood} = log(\hat{y}) * y + log(1-\hat{y}) * (1-y) \tag{3}$$

$$\text{maximize: } \Sigma_i^n log(\hat{y}_i) * y_i + log(1-\hat{y}_i) * (1-y_i) \tag{4}$$

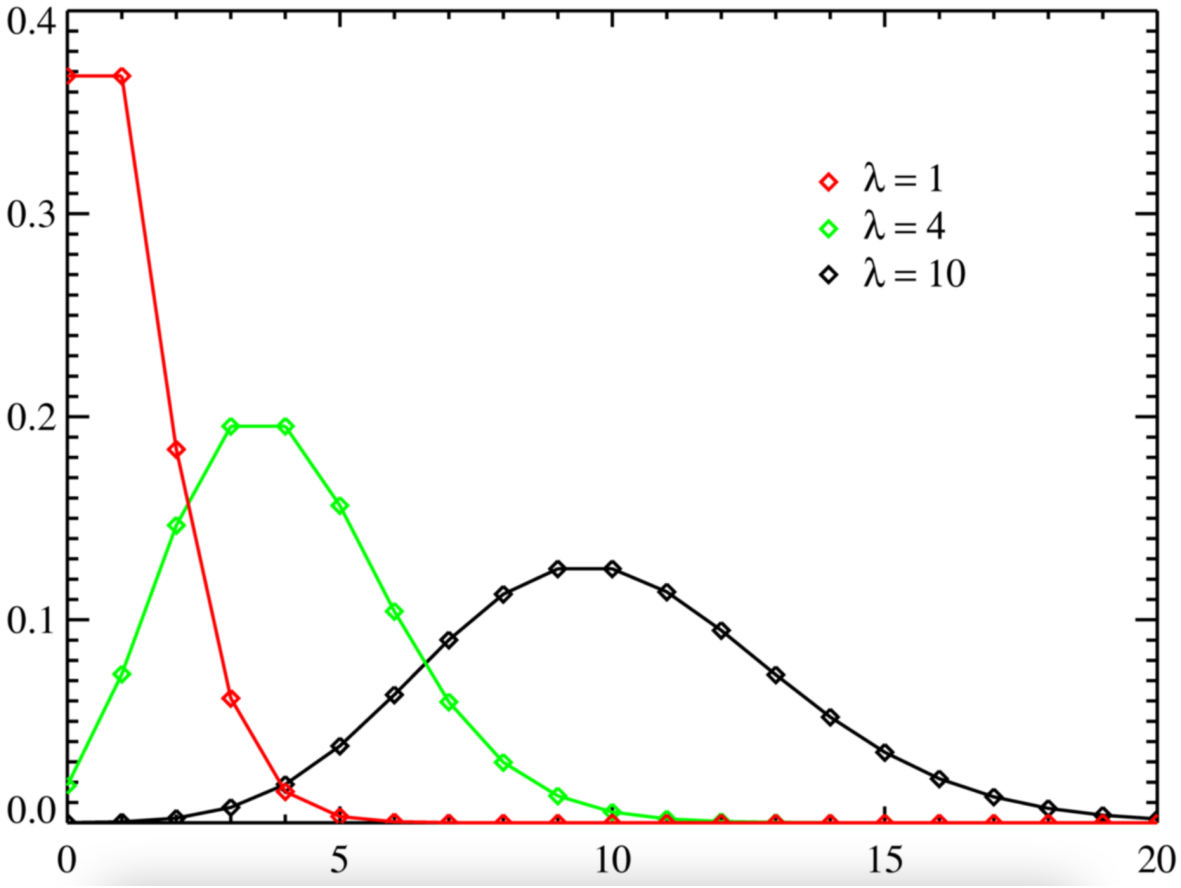
$$\text{minimize: } \Sigma_i^n - (log(\hat{y}_i) * y_i + log(1-\hat{y}_i) * (1-y_i)) \tag{5}$$

Probit

Probit - absolutely the same as logit, but instread of sigmoid generates normal distribution. It hardly could be interpreted as easily as logit, so it's the reason why it so unpopular.

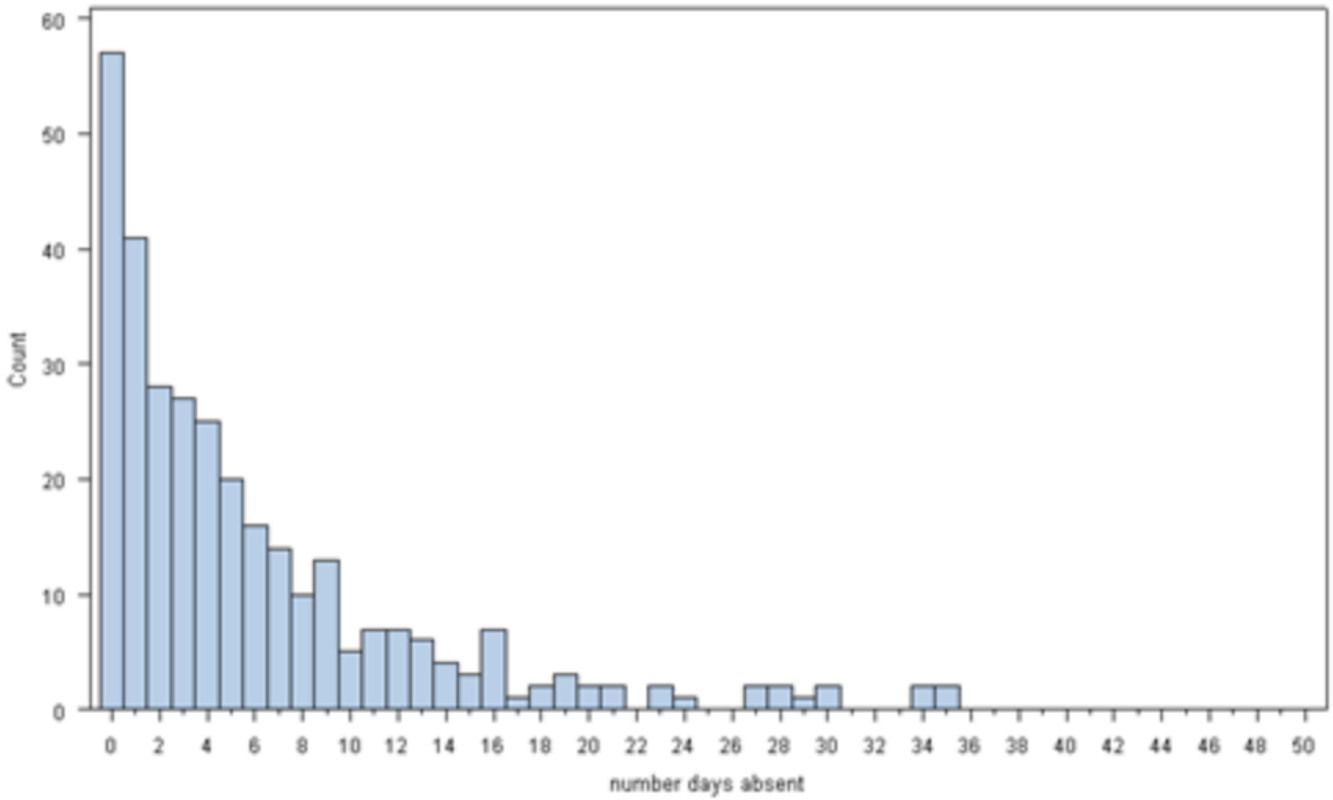
Poisson

Usually used for count data. But it's not dealing with zeros.



Negative Binomial

Negative binomial is a mix of poisson and Gamma distribution.



Output

```
Call:
glm(formula = Survived ~ Sex + Pclass, family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2030  -0.7036  -0.4519   0.6719   2.1599

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.2946     0.2974  11.077   <2e-16 ***
Sexmale      -2.6434     0.1838 -14.380   <2e-16 ***
Pclass       -0.9606     0.1061  -9.057   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.7  on 890  degrees of freedom
Residual deviance:  827.2  on 888  degrees of freedom
AIC: 833.2

Number of Fisher Scoring iterations: 4
```

Output

```r
logit %>% coef %>% exp
```

```
 (Intercept)      Sexmale       Pclass
  26.9677456   0.0711192   0.3826812
```