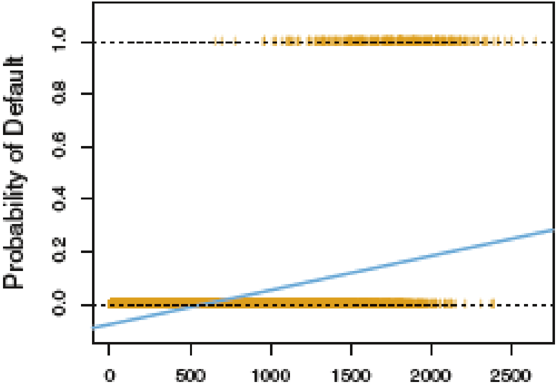
Logistic Regression Basics

Roberta De Vito

BROWN
Public Health

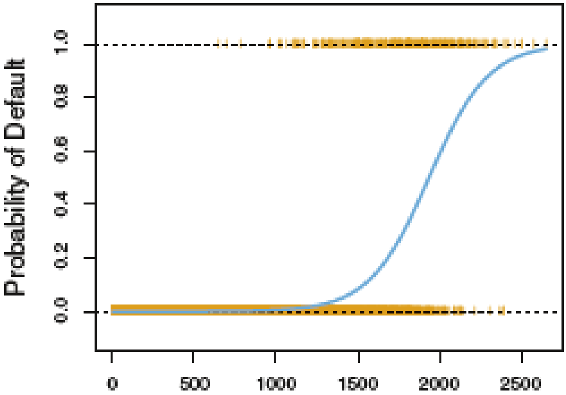
Function to be used



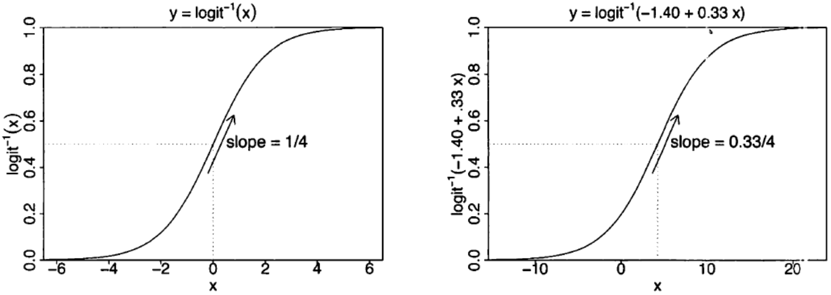
Function to be used

- $Y \in \{0, 1\}$
- $p(Y = 1|X) = \beta_0 + \beta_1 X$
- $p(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$

Function to be used



Function to be used



Question on prismia

Odds

The odds of $Y = 1$ are

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

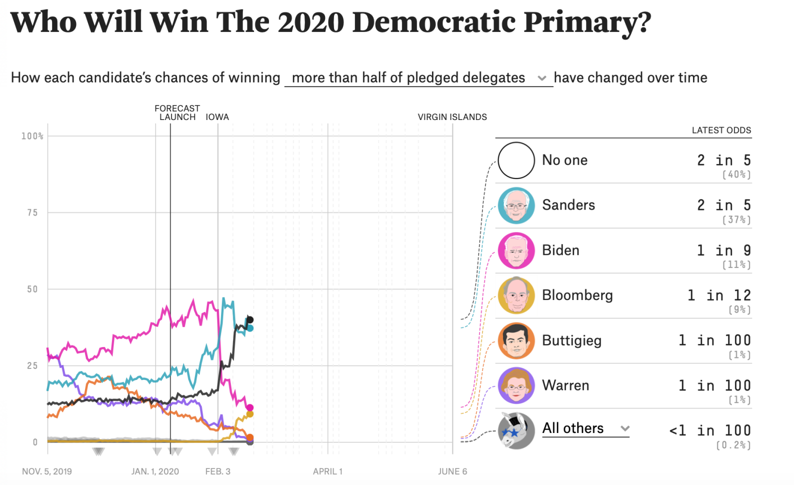
Odds can take values in all of $\mathbb{R}_+$.

Odds II

if one out of ten people will vote Republican, what is the odds of voting the Democratic?

1. $1/5$
2. $1/9$
3. 9

Odds II



Who Will Win The 2020 Democratic Primary?
How each candidate's chances of winning more than half of pledged delegates have changed over time
FORECAST LAUNCH
IOWA
VIRGIN ISLANDS
100%
75
50
25
0
NOV. 5, 2019
JAN. 1, 2020
FEB. 3
APRIL 1
JUNE 6
LATEST ODDS
No one 2 in 5 (40%)
Sanders 2 in 5 (37%)
Biden 1 in 9 (11%)
Bloomberg 1 in 12 (9%)
Buttigieg 1 in 100 (1%)
Warren 1 in 100 (1%)
All others <1 in 100 (0.2%)

Logistic Regression

The logistic regression model relates the log-odds to the covariates through the model

$$logit(P(Y = 1|X)) = \beta X,$$

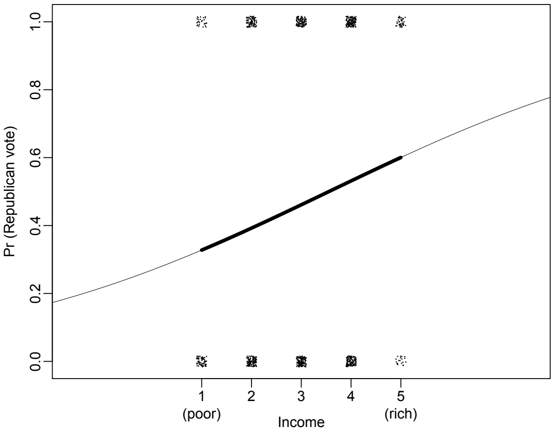
where

$$logit(P(Y = 1|X)) = \log \left(\frac{P(Y = 1|X)}{1 - P(Y = 1|X)} \right).$$

Example: National Election Study data set

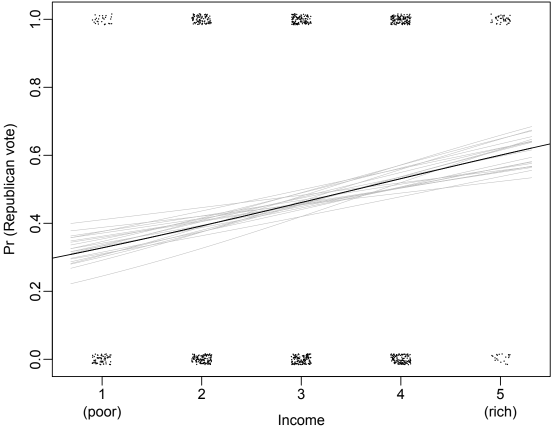
`ID`	code for the identification of the sample
`Vote`	1 Bush (Republican), 0 Clinton (Democratic)
`Race`	Ethnicity: 0=white, 1=black, 0.5=other
`Age`	18, 65+
`Educ1`	Education 1 = no high school, 2 = high school graduate, 3 = some college, 4 =college grad
`Sex`	0=male, 1=female
`income`	income (1=0-16th percentile, 2= 17-33rd percentile, 3=34-67th percentile, 4=68-95th percentile, 5=96-100th percentile)
`year`	year of voting

A first look to the data



Do you expect that the income will increase the probability to vote for the Republican?

Best Logistic Line



Do you expect that the income will increase the probability to vote for the Republican?

The logistic results

```
Coefficients:
            Estimate Std. Error
(Intercept) -1.40213    0.18946
income       0.32599    0.05688
```

The logistic results

```
Coefficients:
            Estimate Std. Error
(Intercept) -1.40213    0.18946
income       0.32599    0.05688
```

```
> invlogit(-1.40 + 0.33*3)
[1] 0.3989121
```

The logistic results

```
Coefficients:
            Estimate Std. Error
(Intercept) -1.40213    0.18946
income       0.32599    0.05688
```

Any other idea?

The logistic results

```
Coefficients:
            Estimate Std. Error
(Intercept) -1.40213    0.18946
income       0.32599    0.05688
```

```
> invlogit(-1.40 + 0.33*mean(income, na.rm=T))
[1] 0.4049001
>
> mean(income, na.rm=T)
[1] 3.075488
```

Interpreting the logistic regression coefficients

- the intercept can only be interpreted assuming zero values for the other predictors
- A difference of 1 in income corresponds to a positive difference of 0.33 in the logit $P(Y) = 1$
  1. evaluate how the probability differs with a unit difference in x near the central value
  2. compute the derivative of the logistic curve at the central value
- the "divide by 4 rule"
- odds ratio

Comparing two proportions

Let μ_j be the proportion of successes in in group $j = 0, 1$ Some commonly used quantities to compare the proportions are:

- Risk difference: $\mu_1 - \mu_0$.
- Relative risk: $\frac{\mu_1}{\mu_0}$
- Odds ratio: $OR(\mu_1, \mu_2) = \frac{\frac{\mu_1}{1-\mu_1}}{\frac{\mu_0}{1-\mu_0}}$

Inference

```
> summary(fit.1)

Call:
glm(formula = vote ~ income, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2756  -1.0034  -0.8796   1.2194   1.6550

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.40213    0.18946  -7.401 1.35e-13 ***
income       0.32599    0.05688   5.731 9.97e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1591.2  on 1178  degrees of freedom
Residual deviance: 1556.9  on 1177  degrees of freedom
  (368 observations deleted due to missingness)
AIC: 1560.9

Number of Fisher Scoring iterations: 4
```

- maximum likelihood estimation
- standard error
- is it significant the income coefficient?
- predictions

Inference

- Can construct confidence intervals for β, $[\hat{\beta} - 1.96\hat{S}E(\hat{\beta}), \hat{\beta} + 1.96\hat{S}E(\hat{\beta})]$.
- Can construct confidence intervals for e^{β}, $[e^{\hat{\beta}-1.96\hat{S}E(\hat{\beta})}, e^{\hat{\beta}+1.96\hat{S}E(\hat{\beta})}]$.
- Create a Wald test for $H_0 : \beta_k = \alpha$ using the test statistic

$$\frac{\hat{\beta}_k - \alpha}{\hat{S}E(\hat{\beta}_k)}$$

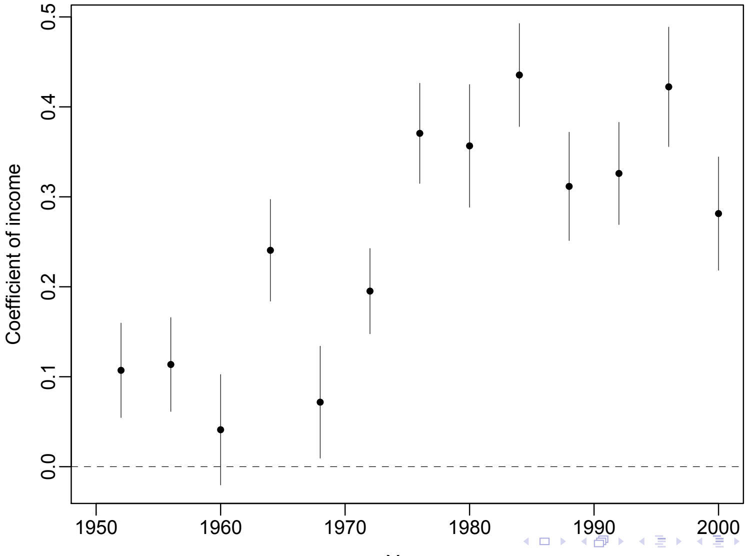
Inference II

```
> summary(fit.1)

Call:
glm(formula = vote ~ income, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2756  -1.0034  -0.8796   1.2194   1.6550

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.40213    0.18946  -7.401 1.35e-13 ***
income       0.32599    0.05688   5.731 9.97e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1591.2  on 1178  degrees of freedom
Residual deviance: 1556.9  on 1177  degrees of freedom
  (368 observations deleted due to missingness)
AIC: 1560.9

Number of Fisher Scoring iterations: 4
```

Inference II

```
> with(fit.1, null.deviance - deviance)
[1] 20.47077
> with(fit.1, df.null - df.residual)
[1] 1
> with(fit.1, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))
[1] 6.054897e-06
```

Coefficients with standard errors



Latent-data formulation

$$
\begin{aligned}
y_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases} \\
z_i &= X_i\beta + \epsilon_i,
\end{aligned}
$$

$$\Pr(\epsilon_i < x) = \text{logit}^{-1}(x) \text{ for all } x.$$

Latent-data formulation