Logistic regression

ST552 Lecture 25

Charlotte Wickham

2019-03-08

Next week:

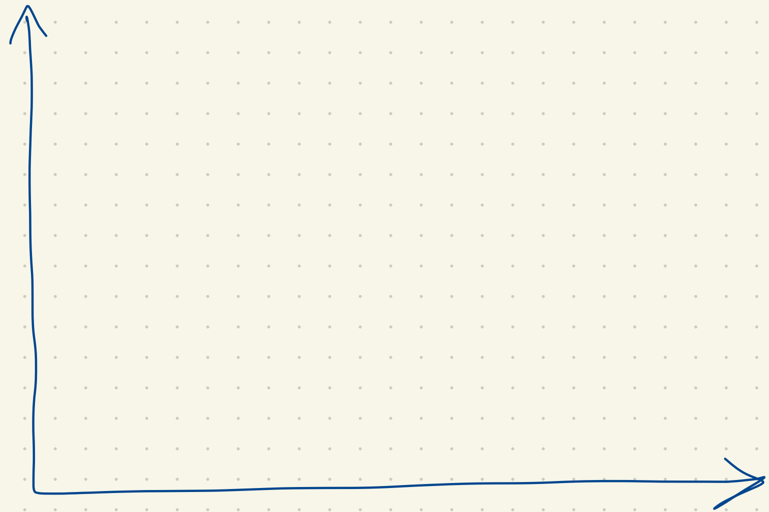
Lab 10 → Q2 from a previous year

- Mon – Non-linear regression
- Wed – Case study LAPTOPS
- Friday – ???

Extra Office Hours:

Week 10: Friday 2pm – class

Final's week Mon: 11am – 12:30pm



$$\sum_{j=1}^{k} \beta_j f_k(x)$$

send me
something ...

Logistic regression

Reading: 5.1 & 5.2 in Data Analysis Using Regression and Multilevel/Hierarchical Models, Gelman & Hill (http://search.library.oregonstate.edu/OSU:everything:CP71242639930001451)

Logistic regression is the standard way to model binary outcomes.

I.e. a response variable that only takes the values 0 or 1.

$$y_i = \begin{cases} 1, & \text{with probability } p_i \\ 0, & \text{with probability } 1 - p_i \end{cases}$$

Example: political preference from Gelman & Hill

Conservative parties generally receive more support among voters with higher incomes. We illustrate classical logistic regresssion with a simple analysis of this pattern from the National Election Study in 1992.

For each repondent, i*, in this poll, we label* $y_i = 1$ *if he or she preferred George Bush (the Republican candiadate for president) or 0 is he or she preferred Bill Clinton (the Democratic candidate), for now excluding repondents who preferred Ross Perot or other candidates.*

We predict preferences given the respondent's income level which is characterized on a five-point scale.

$$y_i = \begin{cases} 1, & \text{respondent } i \text{ preferred George Bush} \\ 0, & \text{respondent } i \text{ preferred Bill Clinton} \end{cases}$$

$x_i =$ Income class of respondent i: 0 (poor), 1, 2, 3, 4 or 5 (rich)

Our goal is to relate y_i to x_i.

Exploratory analysis in R

Can we fit a regression model?

Should we fit a regression model?

Logistic regression model

In logistic regression, the response is related to the explantories through the probability of the response being 1:

$y_i \sim \text{Bernoulli}(p_i)$

$$\text{logit}\,(\underbrace{P(y_i = 1)}_{p_i}) = \boxed{X_i\beta}$$

or equivalently

$E(y_i) = X\beta$

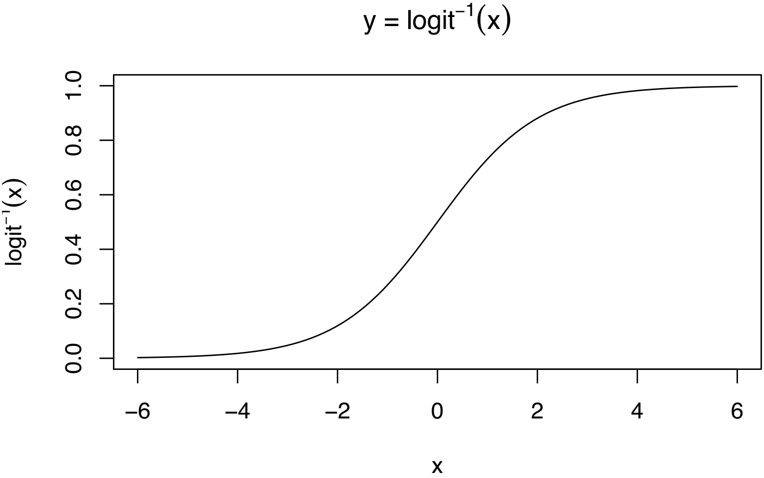
$$P(y_i = 1) = \text{logit}^{-1}\,(X_i\beta)$$

where $\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$

$X_i\beta$ is known as the linear predictor.

y_i are assumed to be i.i.d Bernoulli with probability p_i of success.

The inverse logit transforms continuous values to (0, 1)



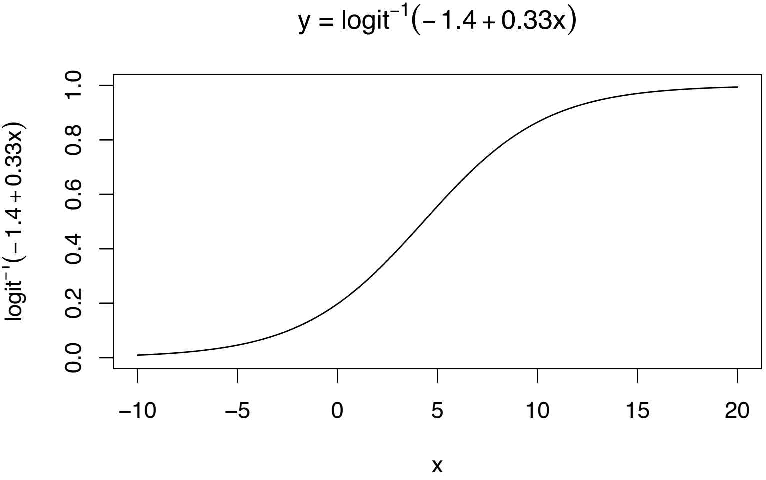
$\downarrow \quad X_i\beta$

Interpreting the logistic regression coefficients

```r
fit_1 <- glm(vote ~ income, family = binomial(link = "logit"),
  data = pres_1992)
summary(fit_1)
```

```
##
## Call:
## glm(formula = vote ~ income, family = binomial(link = "logit"),
##     data = pres_1992)
##
## 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)
```

The fitted model



Interpreting the logistic regression coefficients

Very generally, a coefficient greater than zero indicates increasing probability with increasing explanatory. A coefficient less than zero indicates decreasing probability with increasing explanatory.

But, the non-linear relationship with p_i makes it hard to interpret that exact value.

Three approaches:

At or near center of data

Divide by 4 rule

Odds ratios ← Classic

At or near center of data

```
invlogit <- function(x) 1/(1 + exp(-x))
# = Interpret at some x =
mean_inc <- with(pres_1992, mean(income, na.rm=T))
invlogit(-1.40 + 0.33*mean_inc)
```

```
## [1] 0.4049001
```

Estimated probability of supporting Bush for a respondent of average income is 0.4

At or near center of data

```
# = Interpret change in P for 1 unit change in x, at some :
invlogit(-1.40 + 0.33*3) - invlogit(-1.40 + 0.33*2)
```

```
## [1] 0.07590798
```

An increase in income from category 2 to category 3 is associated with an increase in the estimated probability of supporting Bush of 0.08

At or near center of data

```
logit_p <- (-1.40 + 0.33*3.1)
0.33*exp(logit_p)/(1 + exp(logit_p))^2
```

derivative of $P(Y_i = 1)$ w.r.t to X, at average income

```
## [1] 0.07963666
```

> *Each "small" unit of increase in income, at the average income, is associated with an increase in the estimated probability of supporting Bush of 0.08*

Divide by 4 rule

The logistic function reaches its maximum slope at its center, where the derivative is $\beta/4$.

```r
# = Interpret bound on change in P =
coef(fit_1)[2]/4
```

```
##     income
## 0.08149868
```

> *At most a one unit change in income is associated with an increase of P(Bush) of 0.08*

Odds ratios

odds ratio

$$\log\left(\boxed{\frac{P(y=1|x)}{P(y=0|x)}}\right) = \alpha + \beta x \quad = x\beta$$

A unit increase in x results in a β increase in the log odds ratio of supporting Bush.

> *A one unit increase in income is associated with a change in the log odds ratio of 0.33*

of supporting Bush

Inference & prediction

Coefficients are estimated with maximum likelihood.

Standard errors represent uncertainty in estimates.

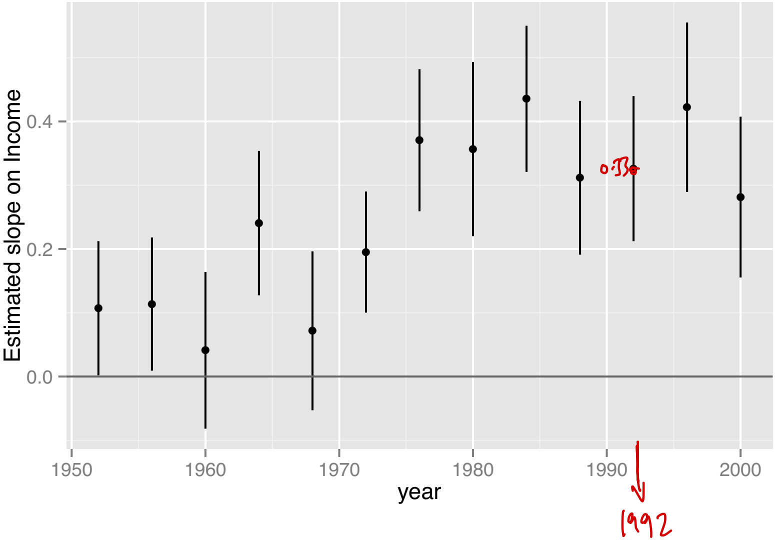
Assymptotically, estimates are Normally distributed under repeated sampling.

An approximate 95% confidence interval for estimates is:
estimate $\pm 2 \times$standard error

Predictions take the form of a predictive probability

$$\hat{p}_0 = \hat{P}(y_0 = 1) = \text{logit}^{-1}(x_0 \hat{\beta})$$

For a voter not in the survey with an income level of 5, the predicted probability of supporting Bush is 0.55


Estimated slope on Income
0.4
0.2
0.0
1950
1960
1970
1980
1990
2000
year
0.33
1992