22s:152 Applied Linear Regression

Ch. 14 (sec. 1) and Ch. 15 (sec. 1 & 4): Logistic Regression

Logistic Regression

- **Maximum likelihood estimates (MLEs)**
  - The regression coefficients in the logistic regression model are estimated through maximum likelihood estimation.
  - To use maximum likelihood estimation, you must completely specify the joint distribution of the observations.

– **Example**:

Consider the usual regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

$$\text{with } \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

The density function for a single observation is:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left(-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}'\boldsymbol{\beta})^2\right)$$

The joint density for all the independent observations is:

$$\prod_{i=1}^{n} f(y_i) = \frac{1}{(2\pi)^{n/2}\sigma^n}\, exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{Y} - X\boldsymbol{\beta})^T(\boldsymbol{Y} - X\boldsymbol{\beta})\right)$$

After observing the data (and considering it fixed), this joint density is seen as a function of the parameters $\boldsymbol{\beta}$ and σ^2 (we will choose the 'best' parameters for the given observed data).

We will find values of $\boldsymbol{\beta}$ and σ^2 that maximize this likelihood function.

These 'best' parameters will make our observed data <u>most likely</u> to have been observed.

* For example, if we're drawing from a normal distribution with unknown μ and we observe $Y = 9.5$, then set $\hat{\mu} = 9.5$

* For example, if we're drawing from a normal distribution with unknown μ and we observe $Y_1 = 10$ and $Y_2 = 11$, then set $\hat{\mu} = 10.5$

The joint density function shown on the earlier page is called the likelihood function and is denoted as

$$L(\boldsymbol{\beta}, \sigma^2; Y_1, Y_2, \ldots, Y_n).$$

We switch from maximizing the likelihood to maximizing the log-likelihood...

It turns out that finding the parameter values that maximize the *log*-likelihood (which is often easier to work with) also maximize the likelihood.

The log-likelihood for the normal case...

$$l(\boldsymbol{\beta}, \sigma^2; Y_1, Y_2, \ldots, Y_n)$$

$$= log(L(\boldsymbol{\beta}, \sigma^2; Y_1, Y_2, \ldots, Y_n))$$

$$= -\tfrac{n}{2} log(2\pi) - \tfrac{n}{2} log(\sigma^2)$$

$$-\tfrac{1}{2\sigma^2}(\boldsymbol{Y} - X\boldsymbol{\beta})^T(\boldsymbol{Y} - X\boldsymbol{\beta})$$

$$= -\tfrac{n}{2} log(2\pi) - \tfrac{n}{2} log(\sigma^2)$$

$$-\tfrac{1}{2\sigma^2} \textstyle\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}))^2$$

We wish to maximize this function with respect to $\boldsymbol{\beta}$ and σ^2 so we will take partial derivatives and solve the equations.

The resulting equations from the partial derivatives set equal to 0 are called the likelihood equations.

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2; \boldsymbol{Y})}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})) = 0$$

For all the other coefficients...

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2; \boldsymbol{Y})}{\partial \beta_j} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ji} (Y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})) = 0$$

And finally,

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2; \boldsymbol{Y})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}))^2 = 0$$

Solution for the case of normal errors:
$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}$ (same as OLS)

$\hat{\sigma^2} = \frac{RSS}{n}$ (but this estimator is biased)

So, we usually use $\hat{\sigma^2} = \frac{RSS}{n-k-1}$

– So, the procedure for getting MLEs is to ...

  * Completely specify your joint density (i.e. the likelihood function)

    Is the data distributed normal? gamma? Weibull? log-normal? etc.

  * Take partial derivatives with respect to each parameter

  * Solve the likelihood equations to get the parameter estimates

– What about MLEs for logistic regression parameters?

– MLEs for logistic regression parameters

As Y_i is discrete and in $\{0, 1\}$, we write the probability mass function for a single observation as a Bernoulli distribution:

$$p(Y_i = y_i|\boldsymbol{x}_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

$$\text{for} \;\; y_i \in \{0, 1\}$$

$$\text{where} \;\; \pi_i = P(Y_i = 1|\boldsymbol{x}_i) = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}.$$

(we're modeling the probability of a *success* as a function of the covariates $\boldsymbol{x}_i$)

The joint distribution for all the independent observations is:

$$\prod_{i=1}^{n} p(Y_i = y_i|\boldsymbol{x}_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} (1-\pi_i)$$

where $\frac{\pi_i}{1-\pi_i}$ is the odds for $P(Y_i = 1|\boldsymbol{x}_i)$

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0+\beta_1 x_{1i}+\cdots+\beta_k x_{ki}} = e^{\boldsymbol{x}_i'\boldsymbol{\beta}}$$

So, the likelihood (or joint distribution) is written as a function of the observed y-values and x-values, and the unknown parameters.

And we can write the log-likelihood as:

$$l(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} y_i \boldsymbol{x}_i' \boldsymbol{\beta} - \sum_{i=1}^{n} \ln(1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}})$$

where $\boldsymbol{x}_i$ holds the covariate values for observation i.

Taking partial derivatives with respect the regression coefficients, we find

$$\frac{\partial l(\boldsymbol{\beta}; \boldsymbol{y})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(y_i - \left(\frac{e^{\boldsymbol{x}_i' \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}}} \right) \right) \boldsymbol{x_i}$$

Or using matrix notation and setting equal to 0,

$$\frac{\partial l(\boldsymbol{\beta}; \boldsymbol{y})}{\partial \boldsymbol{\beta}} = X'(\boldsymbol{y} - \boldsymbol{p}) = \mathbf{o}$$

where $p_i = P(Y_i = 1 | \boldsymbol{x}_i) = \frac{e^{\boldsymbol{x}_i' \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}}}$.

$$\frac{\partial l(\boldsymbol{\beta};\boldsymbol{y})}{\partial \boldsymbol{\beta}} = X'(\boldsymbol{y}-\boldsymbol{p}) = \mathbf{o}$$

Unlike in OLS, the expected value of $\boldsymbol{Y}$ (i.e. $\boldsymbol{p}$), is not a linear function of the $\boldsymbol{\beta}$ parameters, so we have to use an iterative procedure to solve the likelihood equations.

For this, we need the asymptotic variance-covariance matrix for $\hat{\boldsymbol{\beta}}$. And part of that matrix includes the variance in $\boldsymbol{Y}$. Here, that is $V = V(\boldsymbol{Y}) = diag\{p_i(1-p_i)\}$.

Applying the Newton-Raphson method...

1. Select an initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$
2. At each $(l+1)$st iteration, compute new estimates as
   $\hat{\boldsymbol{\beta}}^{(l+1)} = \hat{\boldsymbol{\beta}}^{(l)} + (X'V^{(l)}X)^{-1}X'(\boldsymbol{Y}-\boldsymbol{p}^{(l)})$
3. Continue until convergence of $\hat{\boldsymbol{\beta}}$

This is also called the *Iterative Weighted Least Squares* (IWLS) procedure.

At convergence, we have found the MLEs (maximum likelihood estimates) for $\boldsymbol{\beta}$.

and

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, var(\hat{\boldsymbol{\beta}})) \text{ for large n.}$$

- **Example**: Logistic Regression ANCOVA

  Binary response Y:
  Either 0 or 1

  Two predictor variables:
  1 continuous variable x_1
  1 categorical variable x_2

Data from a breast cancer data set:

Binary response *Menopause*:
0 (not present) or 1 (present)

Predictor variables:
1 continuous variable *age*
1 categorical variable *highed*
0 - lower education
1 - higher education

```
> glm.out=glm(menopause ~ age + highed,
                           family=binomial(logit))
> summary(glm.out)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.51368    1.97708  -6.329 2.46e-10 ***
age           0.28383    0.04117   6.894 5.42e-12 ***
highed       -0.70705    0.36576  -1.933   0.0532 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 313.01  on 337  degrees of freedom
Residual deviance: 193.16  on 335  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 199.16

Number of Fisher Scoring iterations: 7
```

Only 7 iterations were needed for convergence of estimates.

age is a significant predictor of the probability of having entered menopause.

For women in a given *highed* level (i.e. holding *highed* constant), increasing *age* by 1 year is associated with an increase in the odds of having entered menopause by a factor of $e^{0.28282} = 1.3267$

---

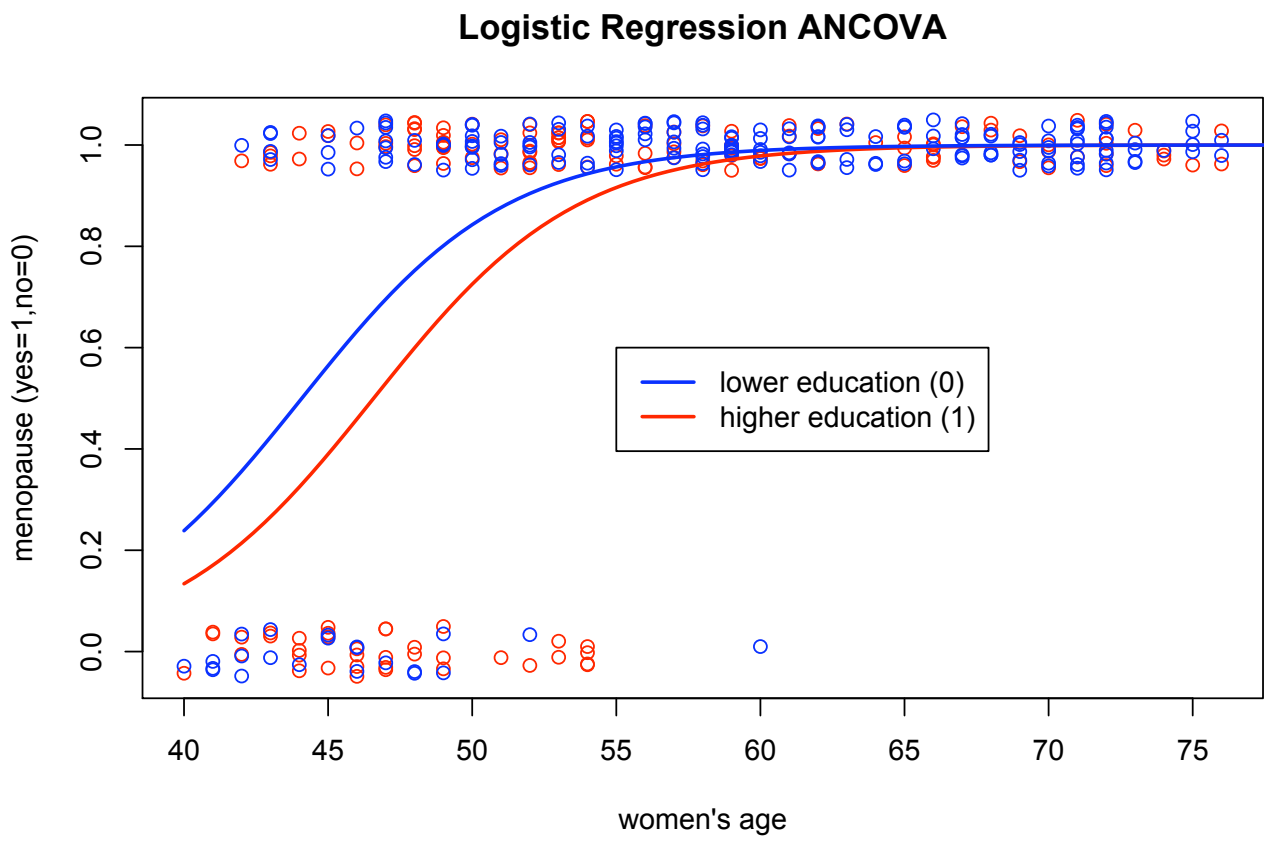
highed is a 'almost' a significant predictor (p=0.0532) and $\hat{\beta}_{highed} = -0.70705$.

For women in a given *age* group (i.e. holding *age* constant), an increase in *highed* (i.e. having more education, or changing *highed* from 0 to 1) is associated with a decrease in the probability of having entered menopause (because $\hat{\beta}_{highed}$ is negative).

Specifically,...

The odds of a woman with higher education having entered menopause is $e^{-0.70705} = 0.4931$ times the odds of a woman with lower education having entered menopause.

What do the fitted curves look like?




```
> glm.out$coefficients
(Intercept)         age      highed
-12.5136809   0.2838317  -0.7070464
```

Additive model:
In terms of interpretation of the parameters, a 1-unit change in *age* has the same affect on both groups because there is no interaction in this model.

The fitted curves are not on top of each other because one is 'shifted' to the right of the other. The shift is present because of the *highed* effect.

The predicted probability of a woman having entered menopause at a given age, is different between the two groups.

Interaction model:
If a model was fit that included interaction, these two curves could feasibly crossover each other.

But, as with classical regression models, there are a variety of forms of interaction, and the shape of the fitted curves (and how they relate to each other) will depend on your data.

Code for ANCOVA plot:

```
> plot(age,jitter(menopause,factor=0.25),main="Logistic
      Regression ANCOVA", xlab="women's age",
      ylab="menopause (yes=1,no=0)",type="n")
> points(age[highed==1],
    jitter(menopause[highed==1],factor=0.25),col="red")
> points(age[highed==0],
    jitter(menopause[highed==0],factor=0.25),col="blue")

> fitted.function.high=function(x)
  {1/(1+exp(-(glm.out$coefficients[1] +
    glm.out$coefficients[2]*x+glm.out$coefficients[3])))}

> fitted.function.low=function(x)
      {1/(1+exp(-(glm.out$coefficients[1] +
           glm.out$coefficients[2]*x)))}

> curve(fitted.function.high,40,80,add=T,lwd=2,col="red")
> curve(fitted.function.low,40,80,add=T,lwd=2,col="blue")

> legend(55,0.6,c("lower education (0)","higher
        education (1)"),lwd=c(2,2),col=c("blue","red"))
```