Applied Statistics

Chiara Seghieri,
Laboratorio MeS
Istituto di Management
Scuola Superiore Sant’Anna
c.seghieri@santannapisa.it

12-02-2021

Introduction

Sometimes we have a situation where the dependent variable Y is qualitative in nature and it takes two mutually exclusive values:

•*Y* = get into college (success), or not (failure); *X* = gender

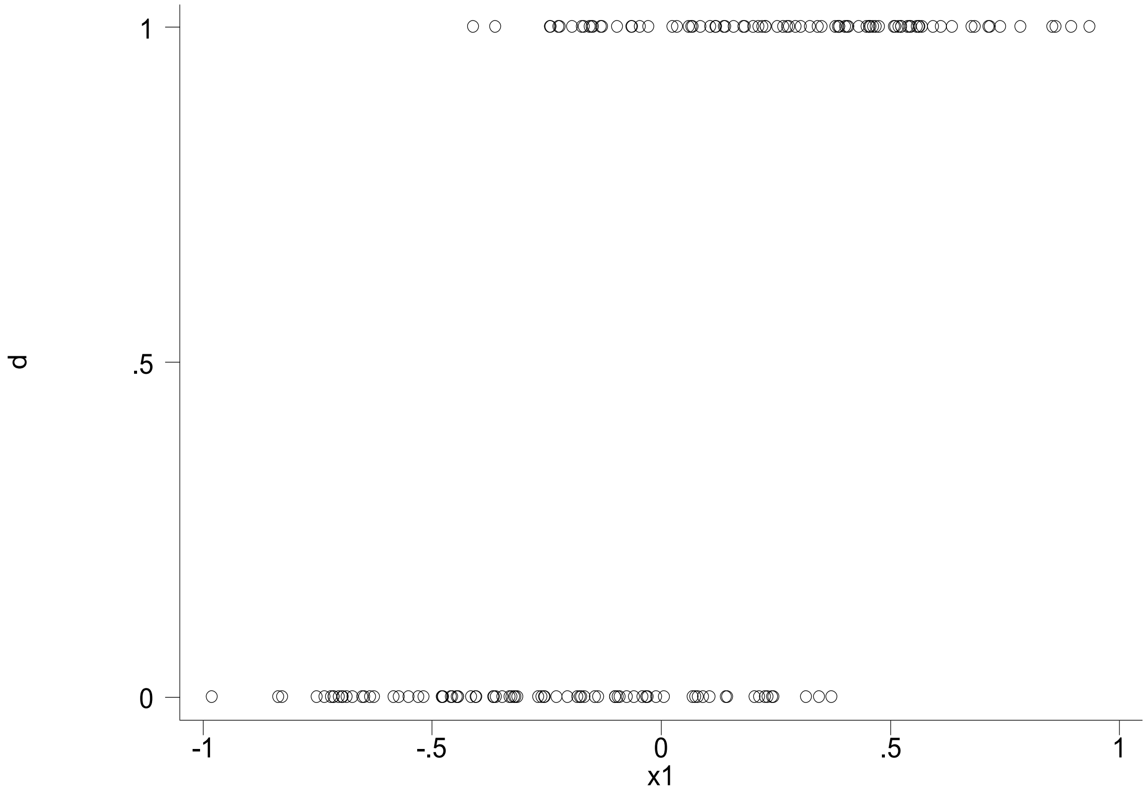
•*Y* = person smokes, or not; *X* = income

•*Y* = mortgage application is accepted, or not; *X* = income, house characteristics, marital status, race

•The basic aim will be to describe the way in which Y (i.e. get into college) varies by gender, race,…

Remember that: μ(Y) = E (Y) is equal to the proportion of observations for which we observe a success (**Σ**1/n). Therefore, the logistic regression model estimate the proportion of success in a population given the X

P(y_i = 1) prob of success for the observation i



The aim is modelling:

$$p_i = P(y_i = j) = F(x_i'\beta)$$

y_i is the binary dependent variable , x_i is the vector of the k regressors, b the parameters and p_i is the probability that the observation i choose j (0,1).

Binary Logistic Regression Model

Y = Binary response *X = predictor*

p = proportion of 1's (yes,success) at any X

Pr(Y = 1|X = x) as a function of x

We denote:

$$P(Y=1)=p(x)$$

$$P(Y=0)=1-p(x)$$

reflecting its dependence on values x_1, x_2,..,x_k of the predictors

NB Average of Y is P and the variance is not constant!

$$\begin{aligned} E(y_i \mid x_i) &= 1 \cdot p_i + 0 \cdot (1 - p_i) \\ &= p_i \\ &= F(x_i'\beta). \end{aligned}$$

var (y_i)=E(y_i)*(1-E(y_i))

Linear Probability Model

The most obvious idea is to let the probability of success be a linear function of x.

$$E(y_i|x) = P(y_i = 1|x = x_i) = \alpha + \beta x \qquad \text{i=1,2,…,n}$$

As example, examine choice of whether a family owns a house:

- Where
  - Y = 1 if family owns a house
  - Y = 0 if family does not own a house
  - X = family income,

  i=1,…,n families

Linear Probability Model

We can estimate such a model by OLS.

- However, we don't get good results:

Probabilities fall between 0 and 1, but linear functions take values over the entire real line.

During the fitting process, $p(x)$ falls outside the (0,1) range for some **x** values. The model can be valid over a restricted range of *x* values.

Problems with LPM

LPM assumes that probabilities increase linearly with the explanatory variables

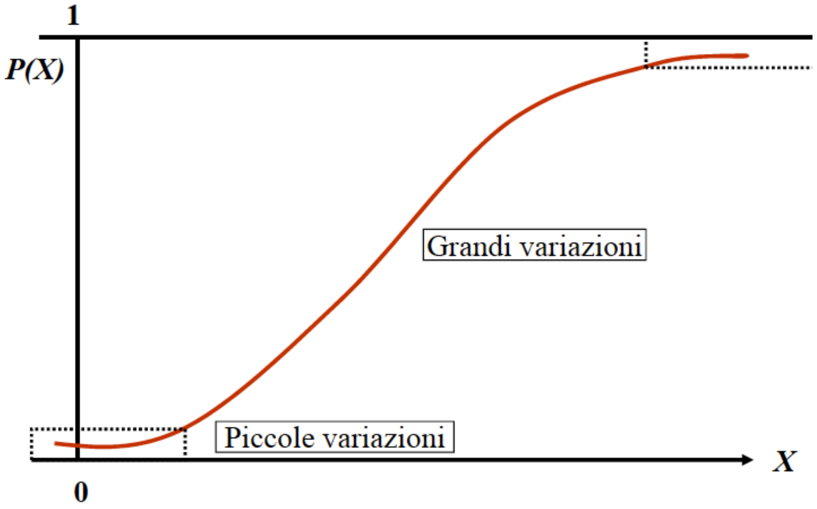
- Each unit increase in an X has the same effect on the probability of Y occurring regardless of the level of the X.
- Moreover, in many situations we empirically see "diminishing returns" — changing p by the same amount requires a bigger change in x when p is already large (or small) than when p is close to 1/2. Linear models can't do this.

EX: Let p(x) denote the probability of buying a new house when annual family income is *x*. An increase of $50,000 in annual income would have less effect when *x is* $1,000,000 (for which p(x) is near 1) than when *x is* $50,000.

What we would like is a model that produces valid probabilities and one where the probabilities are nonlinear in X. As Aldrich and Nelson (1984) note, we want a model that "approaches zero at slower and slower rates as X gets small and approaches one at slower and slower rates as X gets very large."

$$P(y = 1) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

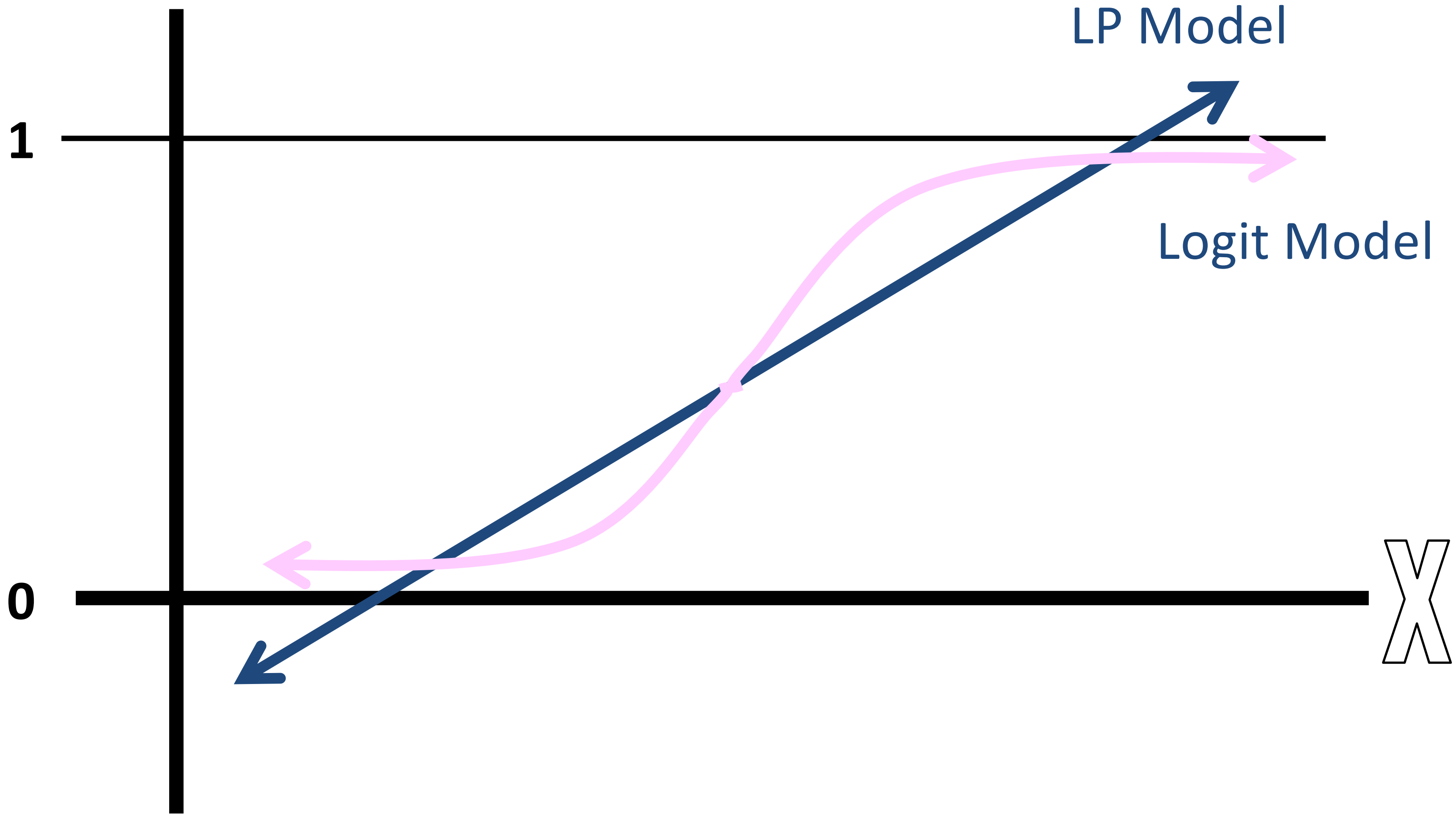
logistic function:
sigmoid function (or an "S-shaped" function)



[P(y = 1)] varies according to the logistic function in [0, 1]

β indicates if [P(y = 1)] increase (**β** > 0) or decrease (**β** < 0) for an increase x

Comparing LP and Logit Models



The logit model

Y~Bin(n,p)

Probability form

$$P(Y=1)=\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$

we linearize the expression by applying the natural log to the two members of the equation (Logit form):

$$\log\frac{p(Y=1)}{\underbrace{1-p(Y=1)}_{P(Y=0)}}=\alpha+\beta x$$

Logistic regression

N.B.: This is natural log (aka “ln”)

Probabilities, odds, and logits:

The logit model has three equivalent forms:

Probabilities: $P(Y = 1) = \dfrac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$

Logits (Log-Odds):

$$\ln \frac{P(y=1)}{P(y=0)} = \alpha + \beta x$$

$$logit[P(y = 1) = \alpha + \beta x]$$

⬇ =1-P(y=1)

Odds:

$$\frac{P(y=1)}{P(y=0)} = e^{\alpha + \beta x}$$

Odds, Odds Ratios

- **The odds** of "success" is the ratio: $\omega = \dfrac{p}{1-p}$

- consider two groups with success probabilities:

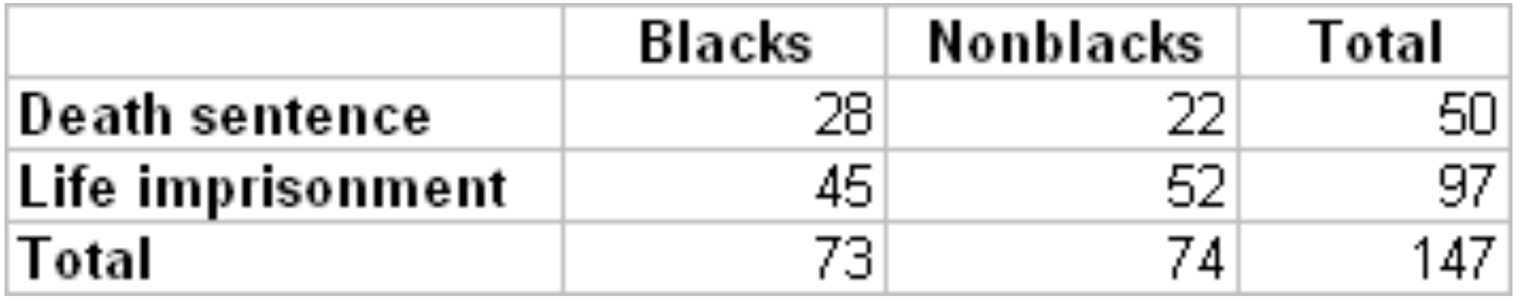
  p_1 and p_2

- **The odds ratio** (OR) is a measure of the odds of success in group 1 *relative* to group 2

$$\theta = \frac{\omega_1}{\omega_2} = \frac{p_1 / (1-p_1)}{p_2 / (1-p_2)}$$

Odds for Independent Variable Groups

- We can compute the **odds of receiving a death penalty** by race:



	Blacks	Nonblacks	Total
Death sentence	28	22	50
Life imprisonment	45	52	97
Total	73	74	147

- The odds of receiving a death sentence if the defendant was Black =p/1-p= (28/73)/(1-(28/73)) = 0.6222
- The odds of receiving a death sentence if the defendant was not Black = 0.4231

The Odds Ratio Measures the Effect

- The impact of being black on receiving a death penalty is measured by the odds ratio which equals:

  = the odds if black / the odds if not black

  = 0.6222 / 0.4231 = 1.47

- Which can be interpreted as:
  - Blacks are 1.47 times more likely to receive a death sentence as non blacks
  - The risk of receiving a death sentence are 1.47 times greater for blacks than non blacks
  - The odds of a death sentence for blacks are 47% higher than the odds of a death sentence for non blacks. (1.47 - 1.00)
  - A one unit change in the independent variable race (nonblack to black) increases the odds of receiving a death penalty by a factor of 1.47.

Maximum likelihood estimations

- Logit and Probit models are nonlinear in the coefficients β therefore these models can't be estimated by the standard OLS (you could use non linear OLS but not efficient!)

- MLE is an alternative to OLS. It consists of finding the parameters values which is the most consistent with the data we have.

- In Statistics, the likelihood is defined as the joint probability to observe a given sample, given the parameters involved in the generating function.

- One way to distinguish between OLS and MLE is as follows:

OLS adapts the model to the data you have: you only have one model derived from your data. MLE instead supposes there is an infinity of models and chooses the model most likely to explain your data.

The method of maximum likelihood yields values for the unknown parameters that maximize the probability of obtaining the observed set of data

Parameter interpretation

$$\log \frac{p(Y=1)}{1-p(Y=1)} = \alpha + \beta x$$

Logistic slope coefficients can be interpreted as the effect of a unit of change in the X variable on the logits with the other variables in the model held constant. That is how a one unit change in X effects the log of the odds when the other variables in the model held constant.

Parameter interpretation (ctd).

- Exponentiating both sides of the logit link function we get the following:

$$\left(\frac{p_i}{1-p_i}\right) = \text{odds} = \exp(\beta_0 + \beta_1 X_1) = e^{\beta 0}\, e^{\beta 1X1}$$

- The odds increase **multiplicatively** by $e^{\beta 1}$ for every 1-unit increase in x, <u>x continuous</u>.

- The odds at X = x+1 are e^{β} times the odds at X = x, furthermore, $(e^{\beta} -1)*100$ gives the percent increase in the odds of a success for each 1-unit increase in x if estimate of $\beta > 0$. $(1-e^{\beta})*100$ estimate of $\beta < 0$

Parameter interpretation (ctd).

for *x* (*i.e. income in thousands of dollars-continuos*) and Y=buying a house or not. The estimate of β is equal to 0.012. If income is increased by $10, this increases the odds of buying a house by about 13%

$$(e^{10\times 0.012}-1)\times 100\% = 12.75\%$$

- if estimate of β> 0 then percent increase in odds for a unit change in *x* is

$$(e^{\hat{\beta}}-1)\times 100\%$$

- if estimate of β < 0 then percent decrease in odds for a unit change in *x* is

$$(1-e^{\hat{\beta}})\times 100\%$$

Parameter interpretation (ctd).

Example: for *x* (***dummy variable coded 1 for female, 0 male***), y satisfaction. The odds ratio is

$$\theta = \frac{p_f / (1 - p_f)}{p_m / (1 - p_m)} = \frac{\omega_f}{\omega_m} = \exp(\hat{\beta}_2) = \exp(0.67) = 1.95$$

- holding the other variable constants, women's odds of buying a house is nearly twice those of men.

Goodness of Fit Measures

- In ML estimations, there is no such measure as the R^2

- But the log likelihood measure can be used to assess the goodness of fit. But note the following :
  - Given the number of observations, the better the fit, the higher the LL measures
- The philosophy is to compare two models looking at their LL values. One is meant to be the constrained model, the other one is the unconstrained model.

Goodness of Fit Measures

- A model is said to be constrained when the parameters associated with some variable are set to zero.
- A model is said to be unconstrained when the parameters associated with some variable are allowed to be different from zero.

- For example, we can compare two models, one with no explanatory variables, one with all our explanatory variables. The one with no explanatory variables implicitly assume that all parameters are equal to zero. Hence it is the constrained model because we (implicitly) constrain the parameters to be null.

The likelihood ratio test (LR test)

- The most used measure of goodness of fit in ML estimations is the likelihood ratio. The likelihood ratio is the difference between the unconstrained model and the constrained model. This difference is distributed χ^2.
- If the difference in the LL values is (no) important, it is because the set of explanatory variables brings in (un)significant information. The null hypothesis H_0 is that the model brings no significant information as follows:

$$LR = 2\left[\ln L_{unc} - \ln L_{c}\right]$$

- High LR values will lead the observer to reject hypothesis H_0 and accept the alternative hypothesis H_a that the set of explanatory variables does significantly explain the outcome.

Other usage of the LR test

- The LR test can also be generalized to compare any two models, the unconstrained one being *nested* in the constrained one.

- Any variable which is added to a model can be tested for its explanatory power as follows :
  - `logit [model contraint]`
  - `est store [name1]`
  - `logit [model non contraint]`
  - `est store [name2]`
  - `lrtest name2 name1`

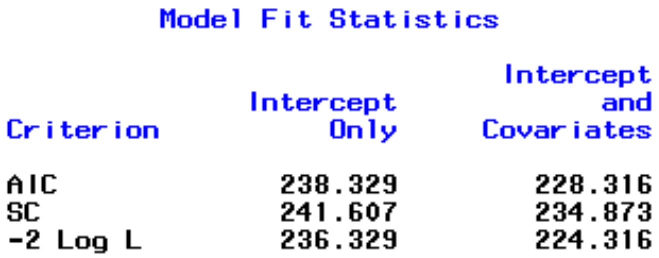
The McFadden Pseudo R^2

- We also use the McFadden Pseudo R^2 (1973). Its interpretation is analogous to the OLS R^2. However it remains generally low.

- pseudo-R^2 also compares The likelihood ratio is the difference between the unconstrained model and the constrained model and is comprised between 0 and 1.

$$\text{Pseudo } R^2_{MF} = \frac{\left[\ln L_c - \ln L_{unc}\right]}{\ln L_{unc}} = 1 - \frac{\ln L_{unc}}{\ln L_c}$$

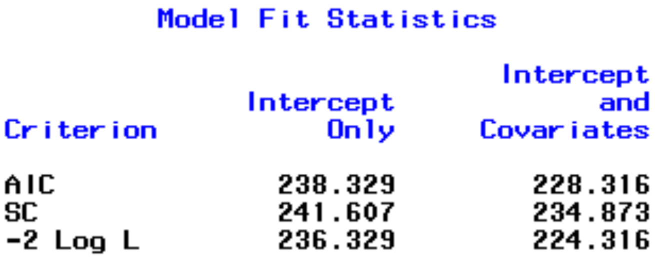
Model Fit Statistics



```
                     Model Fit Statistics

                                             Intercept
                         Intercept                 and
         Criterion            Only          Covariates

         AIC               238.329             228.316
         SC                241.607             234.873
         -2 Log L          236.329             224.316
```

Both AIC & SC are deviants of the -2 Log L that penalize for model complexity (the number of predictor variables).

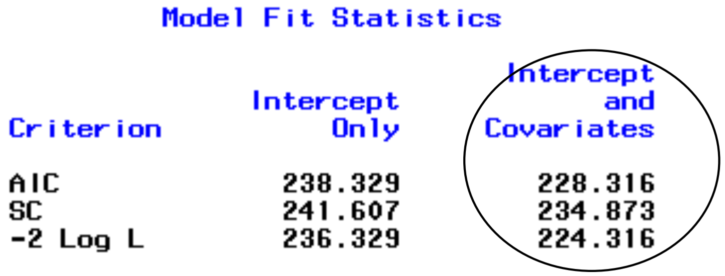
Model Fit Statistics



```
                Model Fit Statistics

                                         Intercept
                          Intercept            and
Criterion                      Only     Covariates

AIC                         238.329        228.316
SC                          241.607        234.873
-2 Log L                    236.329        224.316
```

AIC   Akaike Information Criterion. Used to compare non-nested models. Smaller is better. AIC is only meaningful in relation to another model’s AIC value.

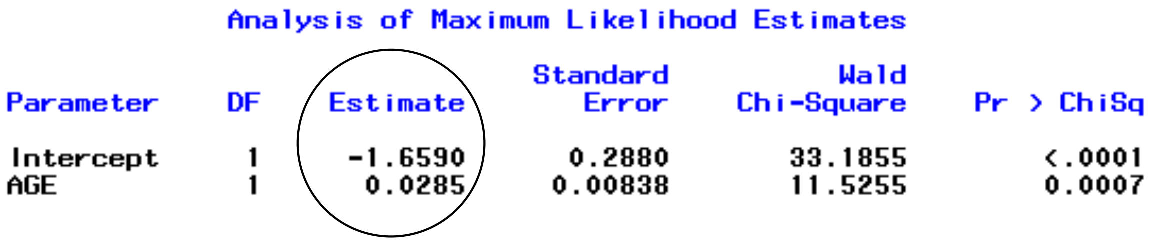
Model Fit Statistics



```
                  Model Fit Statistics

                                           Intercept
                          Intercept              and
Criterion                      Only       Covariates

AIC                         238.329          228.316
SC                          241.607          234.873
-2 Log L                    236.329          224.316
```

Choose either AIC or SC (not both) and use the values under the heading ‘Intercept and Covariates’ to compare to competing models.

The model equation.



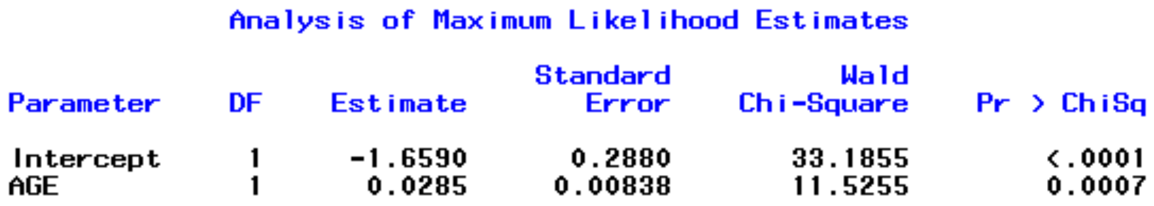
```
                    Analysis of Maximum Likelihood Estimates

                                      Standard          Wald
Parameter     DF     Estimate          Error      Chi-Square     Pr > ChiSq

Intercept      1      -1.6590         0.2880        33.1855         <.0001
AGE            1       0.0285        0.00838        11.5255         0.0007
```

$$P(y=1) = \frac{\exp(-1.659 + .0285x_1)}{1+\exp(-1.659 + .0285x_1)}$$

$$\ln\left(\frac{\pi}{1-\pi}\right) = -1.659 + .0285x_1$$

$$\text{logit}(y) = -1.659 + .0285x_1$$

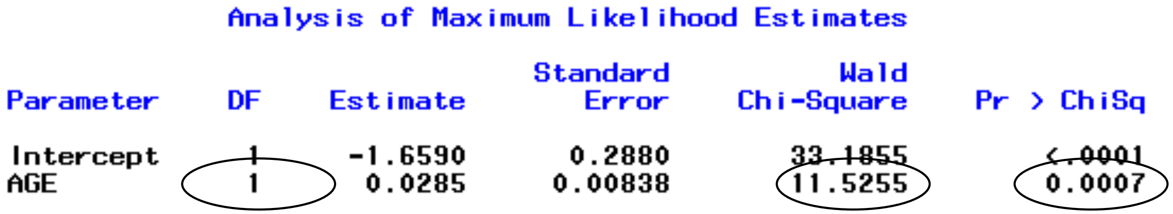
Inference: The Coefficients.



```
                  Analysis of Maximum Likelihood Estimates

                                         Standard          Wald
Parameter      DF      Estimate           Error      Chi-Square      Pr > ChiSq

Intercept       1       -1.6590          0.2880         33.1855          <.0001
AGE             1        0.0285         0.00838         11.5255          0.0007
```

Instead of a *t*-test for the significance of a coefficient (like in linear regression), we have a Wald Chi-Squared test.

Inference: The Coefficients.



```
                    Analysis of Maximum Likelihood Estimates

                                       Standard          Wald
Parameter    DF       Estimate            Error    Chi-Square    Pr > ChiSq

Intercept     1        -1.6590           0.2880       33.1855        <.0001
AGE           1         0.0285          0.00838       11.5255        0.0007
```

Furthermore: (exp(.0285)-1)*100% = 2.89%.

We can state that the odds of being in poor health increase by 2.89% with each additional year in age.

Ex: Y=1 not winner (of election)

logit nowin leader age scandal

```
Logit estimates                                   Number of obs   =       5036
                                                  LR chi2(3)      =     265.97
                                                  Prob > chi2     =     0.0000
Log likelihood = -1214.2961                       Pseudo R2       =     0.0987

------------------------------------------------------------------------------
     nowin    |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      leader |  -1.029759    .516245     -1.99   0.046     -2.04158    -.0179374
      age    |  -.0420528   .0035401    -11.88   0.000    -.0489913    -.0351143
     scandal |   2.839299   .3194128      8.89   0.000     2.213261     3.465337
       _cons |  -1.487179   .0859136    -17.31   0.000    -1.655566    -1.318791
------------------------------------------------------------------------------
```

A positive coefficient: the log-odds of not winning are decreasing as a function of being in a leadership position and with increasing of age and increase as a function of being involved in a scandal.

logistic nowin leader age scandal

```
Logit estimates                                   Number of obs   =       5036
                                                  LR chi2(3)      =     265.97
                                                  Prob > chi2     =     0.0000
Log likelihood = -1214.2961                       Pseudo R2       =     0.0987

------------------------------------------------------------------------------
     Nowin    | Odds Ratio   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      leader |    .357093   .1843475     -1.99   0.046     .1298234     .9822225
      age    |   .9588192   .0033943    -11.88   0.000     .9521895      .965495
     scandal |   17.10377   5.463164      8.89   0.000     9.145496     31.98723
```

Scandal: the odds of not winning of candidate who is involved in a scandal are about 17 times higher than a candidate who is not involved.

Age: (1-(.9588))*100=4% decrease in the odds of not winning for a unit increase in age

Checking assumptions

0. Independent data points

    (no tests for that, just think about your data)

    Problem: likelihood function is wrong otherwise + confidence intervals too small

1. Influential data points

2. No multi-collinearity (Stata: “collin”, VIF)

3. All relevant variables included