Models for binary data: Logit

Linear Regression Analysis
Kenneth Benoit

August 21, 2012

Collinearity continued

- Stone (1945):

$$\mathrm{Var}(\hat{\beta}_k^{OLS}) = \frac{1}{N-K}\frac{\sigma_y^2}{\sigma_k^2}\frac{1-R^2}{1-R_k^2}$$

  - σ_y^2 is the estimated variance of Y
  - σ_k^2 is the estimated variance of the kth regressor
  - R_k^2 is the R^2 from a regression of the kth regressor on all the other independent variables
- So collinearity's main consequence is:
  - the variance of $\hat{\beta}_k^{OLS}$ decreases as the range of X_k increases (σ_k^2 higher)
  - the variance of $\hat{\beta}_k^{OLS}$ increases as the variables in $\mathbf{X}$ become more collinear (R_k^2 higher) and becomes infinite in the case of exact multicollinearity
  - the variance of $\hat{\beta}_k^{OLS}$ decreases as R^2 rises. sp that the effect of a high R_k^2 can be offset by a high R^2

Limited dependent variables

- Some dependent variables are limited in the possible values they may take on
  - might be binary (aka dichotomous)
  - might be counts
  - might be unordered categories
  - might be ordered categories
- For these methods, OLS and the CLRM will fail to provide desirable estimates – in fact OLS easily produces non-sensical results
- Focus here will be on binary and count limited dependent variables

Binary dependent variables

- Remember OLS assumptions
  - ϵ_i has a constant variance σ^2 (homoskedasticity)
  - ϵ_i are uncorrelated with one another
  - ϵ_i is normally distributed (necessary for inference)
  - Y is unconstrained on $\mathbb{R}$ – implied by the lack of restrictions on the values of the independent variables (except that they cannot be exact linear combinations of each other)
- This cannot work if $Y = \{0, 1\}$ only

$$
\begin{aligned}
E(Y_i) &= 1 \cdot P(Y_i = 1) + 0 \cdot P(Y_i = 0) = P(Y_i = 1) \\
&= \sum b_k X_{ik} = \mathbf{X_i b}
\end{aligned}
$$

- But if Y_i only takes two possible values, then $e_i = \hat{Y}_i - Y_i$ can only take on two possible values (here, 0 or 1)

Why OLS is unsuitable for binary dependent variables

- From above, $P(Y_i = 1) = X_i b$ – hence this is called a *linear probability model*
  - if $Y_i = 0$, then $(0 = X_i b + e_i)$ or $(e_i = -X_i b)$
  - if $Y_i = 1$, then $(1 = X_i b + e_i)$ or $(e_i = 1 - X_i b)$
- We can maintain the assumption that $\mathrm{E}(e_i) = 0$:

$$
\begin{aligned}
E(e_i) &= P(Y_i = 0)(-X_i b) + P(Y_i = 1)(1 - X_i b) \\
&= -(1 - P(Y_i = 1))P(Y_i = 1) + P(Y_i = 1)(1 - P(Y_i = 1)) \\
&= 0
\end{aligned}
$$

- As a result, OLS estimates are unbiased, but: they will not have a constant variance
- Also: OLS will easily predict values outside of $(0, 1)$ even without the sampling variance problems – and thus give non-sensical results

Non-constant variance

$$\begin{aligned}
Var(e_i) &= E(e_i^2) - (E(e_i))^2 \\
&= E(e_i^2) - 0 \\
&= P(Y_i = 0)(-X_i b)^2 + P(Y_i = 1)(1 - X_i b)^2 \\
&= (1 - P(Y_i = 1))(P(Y_i = 1))^2 + P(Y_i = 1)(1 - P(Y_i = 1))^2 \\
&= P(Y_i = 1)(1 - P(Y_i = 1)) \\
&= X_i b(1 - X_i b)
\end{aligned}$$

- Hence the variance of e_i varies systematically with the values of X_i
- Inference from OLS for binary dep. variables is therefore invalid

Back to basics: the Bernoulli distribution

- The Bernoulli distribution is generated from a random variable with possible events:
  1. Random variable Y_i has two mutually exclusive outcomes:

     $$Y_i = \{0, 1\}$$
     $$Pr(Y_i = 1 | Y_i = 0) = 0$$

  2. 0 and 1 are exhaustive outcomes:

     $$Pr(Y_i = 1) = 1 - Pr(Y_i = 0)$$

- Denote the population parameter of interest as π: the probability that $Y_i = 1$

$$\begin{aligned} Pr(Y_1 = 1) &= \pi \\ Pr(Y_i = 0) &= 1 - \pi \end{aligned}$$

Bernoulli distribution cont.

- Formula:

$$Y_i = f_{bern}(y_i|\pi) = \begin{cases} \pi^{y_i}(1-\pi)^{1-y_i} & \text{for } y_i = 0,1 \\ 0 & \text{otherwise} \end{cases}$$

- Expectation of Y is π

$$\begin{aligned} E(Y_i) &= \sum_i y_i f(Y_i) \\ &= 0 \cdot f(0) + 1 \cdot f(1) \\ &= 0 + \pi \\ &= \pi \end{aligned}$$

Introduction to maximum likelihood

- Goal: Try to find the parameter value $\tilde{\beta}$ that makes $E(Y|X,\beta)$ as close as possible to the observed Y
- For Bernoulli: Let $p_i = P(Y_i = 1|X_i)$ which implies $P(Y_i = 0|X_i) = 1 - P_i$. The probability of observing Y_i is then

$$P(Y_i|X_i) = P_i^{Y_i}(1 - P_i)^{1-Y_i}$$

- Since the observations can be assumed independent events, then

$$P(Y_i|X_i) = \prod_{i=1}^{N} P_i^{Y_i}(1 - P_i)^{1-Y_i}$$

- When evaluated, this expression yields a result on the interval $(0, 1)$ that represents the likelihood of observing this sample Y given X *if* $\hat{\beta}$ *were the "true" value*
- The MLE is denoted as $\tilde{\beta}$ for β that maximizes $L(Y|X, b) = \max L(Y|X, b)$

MLE example: what π for a tossed coin?

		0.5		
Y_i	**P^yi**	**(1-P)^(1-yi)**	**L**	**ln L**
0	1	0.5	0.5	-0.693147
1	0.5	1	0.5	-0.693147
1	0.5	1	0.5	-0.693147
0	1	0.5	0.5	-0.693147
1	0.5	1	0.5	-0.693147
1	0.5	1	0.5	-0.693147
0	1	0.5	0.5	-0.693147
1	0.5	1	0.5	-0.693147
1	0.5	1	0.5	-0.693147
1	0.5	1	0.5	-0.693147
Likelihood			0.0009766	
Log-Likelihood				-6.931472

		0.6		
Y_i	**P^yi**	**(1-P)^(1-yi)**	**L**	**ln L**
0	1	0.4	0.4	-0.916291
1	0.6	1	0.6	-0.510826
1	0.6	1	0.6	-0.510826
0	1	0.4	0.4	-0.916291
1	0.6	1	0.6	-0.510826
1	0.6	1	0.6	-0.510826
0	1	0.4	0.4	-0.916291
1	0.6	1	0.6	-0.510826
1	0.6	1	0.6	-0.510826
1	0.6	1	0.6	-0.510826
Likelihood			0.0017916	
Log-Likelihood				-6.324652

MLE example continued

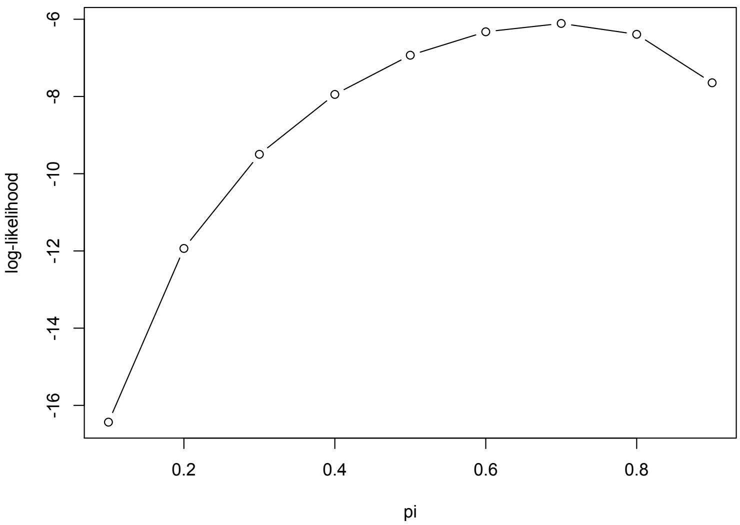
		0.7			
Y_i		**P^yi**	**(1-P)^(1-yi)**	**L**	**ln L**
0		1	0.3	0.3	-1.203973
1		0.7	1	0.7	-0.356675
1		0.7	1	0.7	-0.356675
0		1	0.3	0.3	-1.203973
1		0.7	1	0.7	-0.356675
1		0.7	1	0.7	-0.356675
0		1	0.3	0.3	-1.203973
1		0.7	1	0.7	-0.356675
1		0.7	1	0.7	-0.356675
1		0.7	1	0.7	-0.356675
Likelihood				0.0022236	
Log-Likelihood					-6.108643

		0.8			
Y_i		**P^yi**	**(1-P)^(1-yi)**	**L**	**ln L**
0		1	0.2	0.2	-1.609438
1		0.8	1	0.8	-0.223144
1		0.8	1	0.8	-0.223144
0		1	0.2	0.2	-1.609438
1		0.8	1	0.8	-0.223144
1		0.8	1	0.8	-0.223144
0		1	0.2	0.2	-1.609438
1		0.8	1	0.8	-0.223144
1		0.8	1	0.8	-0.223144
1		0.8	1	0.8	-0.223144
Likelihood				0.0016777	
Log-Likelihood					-6.390319

MLE example in R

```
> ## MLE example
> y <- c(0,1,1,0,1,1,0,1,1,1)
> coin.mle <- function(y, pi) {
+   lik <- pi^y * (1-pi)^(1-y)
+   loglik <- log(lik)
+   cat("prod L = ", prod(lik), ",  sum ln(L) = ", sum(loglik), "\n")
+   (mle <- list(L=prod(lik), lnL=sum(loglik)))
+ }
> ll <- numeric(9)
> pi <- seq(.1,.9,.1)
> for (i in 1:9) (ll[i] <- coin.mle(y, pi[i])$lnL)
prod L =  7.29e-08 ,  sum ln(L) =  -16.43418
prod L =  6.5536e-06 ,  sum ln(L) =  -11.93550
prod L =  7.50141e-05 ,  sum ln(L) =  -9.497834
prod L =  0.0003538944 ,  sum ln(L) =  -7.946512
prod L =  0.0009765625 ,  sum ln(L) =  -6.931472
prod L =  0.001791590 ,  sum ln(L) =  -6.324652
prod L =  0.002223566 ,  sum ln(L) =  -6.108643
prod L =  0.001677722 ,  sum ln(L) =  -6.390319
prod L =  0.0004782969 ,  sum ln(L) =  -7.645279
> plot(pi, ll, type="b")
```

MLE example in R: plot



From likelihoods to log-likelihoods

$$
\begin{aligned}
P(Y_i|X_i) &= \prod_{i=1}^{N} P_i^{Y_i}(1-P_i)^{1-Y_i} \\
\ln L(Y|X,b) &= \sum_{i=1}^{N}(Y_i \ln p_i + (1-Y_i)\ln(1-p_i))
\end{aligned}
$$

If $\tilde{b}$ maximizes $L(Y|X,b)$ then it also maximizes $\ln L(Y|X,b)$

Properties:

- asymptotically unbiased, efficient, and normally distributed
- invariant to reparameterization
- maximization is generally solved numerically using computers (usually no algebraic solutions)

Transforming the functional form

- Problem: the linear functional form is inappropriate for modelling probabilities
  - the linear probability model imposes inherent constraints about the marginal effects of changes in X, while the OLS assumes a constant effect
  - this problem is not solvable by "usual" remedies, such as increasing our variation in X or trying to correct for heteroskedasticity
- When dealing with limited dependent variables in general this is a problem, and requires a solution by choosing an alternative functional form
- The alternative functional form is based on a transformation of the core linear model

The logit transformation

Question: How to transform the functional form $X\beta$ to eliminate the boundary problems of $0 < p_i < 1$?

1. Eliminate the upper bound of $p_i = 1$ by using odds ratio:

$$0 < \frac{p_i}{(1 - p_i)} < +\infty$$

   this function is positive only, and as $p_i \to 1$, $\frac{p_i}{(1-p_i)} \to \infty$

2. Eliminate the lower bound of $p_i = 0$ by taking the logarithm of the odds ratio:

$$-\infty < \ln\left(\frac{p_i}{1 - p_i}\right) < +\infty$$

This transformation is known as logit and stands for the log of the odds ratio.

Expressing p_i in terms of the logit function

$$
\begin{aligned}
E(Y_i) = X_i\beta &= \ln\left(\frac{p_i}{1-p_i}\right) \\
X_i\beta &= \ln\left(\frac{p_i}{1-p_i}\right) \\
e^{X_i\beta} &= e^{p_i} - e^{1-p_i} \\
p_i &= \frac{e^{X_i\beta}}{1+e^{X_i\beta}} \\
&= \left(\frac{e^{-X_i\beta}}{e^{-X_i\beta}}\right)\left(\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right) \\
&= \frac{1}{1+e^{-X_i\beta}}
\end{aligned}
$$

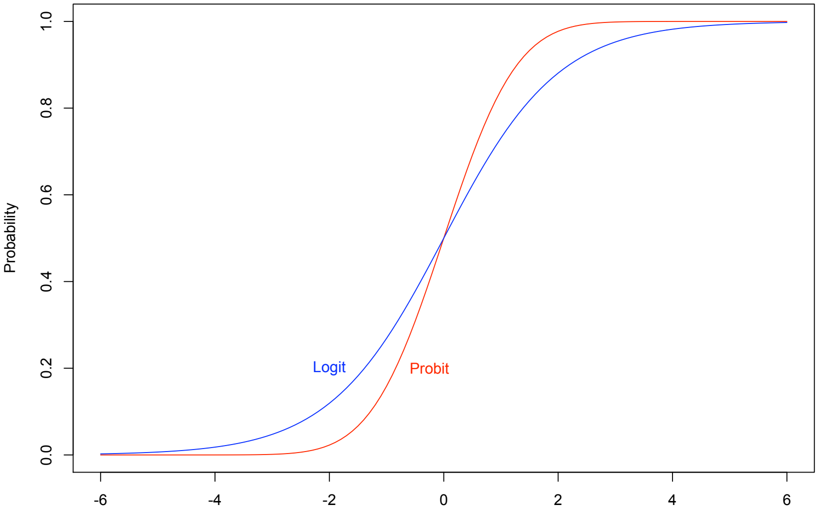
Alternative alternative functional forms

- The logit form is the most commonly used transformation of the linear $X\beta$, but other choices are possible
- Example: we could have used the cumulative distribution function of the normal distribution, defined as

$$\begin{aligned} F(z) &= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du \\ &= \Phi(z) \end{aligned}$$

  This functional form is known as the probit model, standing for "probability unit"
- Other possibilities include Urban, Gompertz, etc. found in Aldrich and Nelson p33

Logit versus probit



Back to the example

- Y = Won a seat (1=yes, 0=no)
- X_1 = `incumbency` (0=challenger, 1=incumbent)
- X_2 = `spending` (continuous variable, measures in euros)
- X_3 = `spendingXinc` (interaction of X_1 and X_2)
- Multiple binary logistic regression model:

$$\begin{aligned} \text{logit}(\pi) &= \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) \\ &= \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 \end{aligned}$$

- log (estimated odds of winning seat) = $\hat{\alpha} + \hat{\beta}_1 X_{\text{incumb}} + \hat{\beta}_2 X_{\text{spending}} + \hat{\beta}_3 X_{\text{inc}*\text{spending}}$

Estimated logit model: Campaign spending example

```
. logit wonseat incumb spend_total spend_totalXinc

Iteration 0:   log likelihood = -301.55276
Iteration 1:   log likelihood = -188.70741
Iteration 2:   log likelihood = -182.41553
Iteration 3:   log likelihood = -182.11942
Iteration 4:   log likelihood =   -182.119
Iteration 5:   log likelihood =   -182.119

Logistic regression                               Number of obs   =        463
                                                  LR chi2(3)      =     238.87
                                                  Prob > chi2     =     0.0000
Log likelihood =   -182.119                       Pseudo R2       =     0.3961

------------------------------------------------------------------------------
     wonseat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      incumb |   3.200883   .8391721     3.81   0.000     1.556136     4.84563
 spend_total |   .0001604   .0000232     6.92   0.000     .0001149    .0002058
spend_tota~c |  -.0000649   .0000428    -1.52   0.130    -.0001488     .000019
       _cons |  -3.901699    .429417    -9.09   0.000    -4.743341   -3.060057
------------------------------------------------------------------------------
```

(Note that this also “works” — but is wrong)

```
. regress wonseat incumb spend_total spend_totalXinc

      Source |       SS       df       MS              Number of obs =     463
-------------+------------------------------           F(  3,   459) =  123.16
       Model |  47.361442     3  15.7871473            Prob > F      =  0.0000
    Residual | 58.8372621   459  .128185756            R-squared     =  0.4460
-------------+------------------------------           Adj R-squared =  0.4423
       Total | 106.198704   462  .229867325            Root MSE      =  .35803

------------------------------------------------------------------------------
     wonseat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      incumb |    .560078   .0944139     5.93   0.000     .3745409     .745615
 spend_total |     .00002   2.31e-06     8.65   0.000     .0000155    .0000246
spend_tota~c |  -7.82e-06   4.47e-06    -1.75   0.081    -.0000166    9.67e-07
       _cons |  -.0498086   .0322345    -1.55   0.123    -.1131542    .0135369
------------------------------------------------------------------------------
```

Example in odds-ratios rather than logits

```
. logit wonseat incumb spend_total spend_totalXinc, or

Iteration 0:   log likelihood = -301.55276
Iteration 1:   log likelihood = -188.70741
Iteration 2:   log likelihood = -182.41553
Iteration 3:   log likelihood = -182.11942
Iteration 4:   log likelihood =   -182.119
Iteration 5:   log likelihood =   -182.119

Logistic regression                               Number of obs   =        463
                                                  LR chi2(3)      =     238.87
                                                  Prob > chi2     =     0.0000
Log likelihood =   -182.119                       Pseudo R2       =     0.3961

------------------------------------------------------------------------------
     wonseat | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      incumb |    24.5542    20.6052     3.81   0.000     4.740469    127.1834
 spend_total |    1.00016   .0000232     6.92   0.000     1.000115    1.000206
spend_tota~c |   .9999351   .0000428    -1.52   0.130     .9998512    1.000019
------------------------------------------------------------------------------
```

Interpreting exponentiated coefficients

- So odds of winning if you are incumbent are $e^{3.200883} = 24.554$ greater for incumbents than for challengers
- For *challengers*, the odds of winning increase by $e^{.0001604} = 1.00016$ for each €1 more spent
- So if a challenger spent €10,000 more, then his or her odds of winning would increase by $e^{10000 * .0001604} = 4.972884$
- If an incumbent spent €1 more, odds of winning would increase by $e^{.0001604 - .0000649} = 1.000096$
- If an incumbent spent €10,000 more, then his or her odds of winning would change by $e^{10000 * (.0001604 - .0000649)} = 2.598671$

Interpreting fitted probabilities

- As with linear regression models, often useful to present a selection of fitted probabilities to illustrate the model
- Formula for translating the estimated logit into estimated probability:

  $$\hat{\pi}_i = \frac{1}{1 + e^{-X_i\beta}}$$

  where $X_i\beta = \alpha + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
- This is the same as saying that

  $$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) == \quad \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$$

- Usually better to interpret interms of $\hat{\pi}$ rather than log odds
- By exponentiating a coefficient β_k, we get relative change in (un-logged) odds of $Y = 1$ for a one-unit increase in X_k

Unexponentiated coefficients

```
. logit wonseat incumb spend_total spend_totalXinc

Iteration 0:   log likelihood = -301.55276
Iteration 1:   log likelihood = -188.70741
Iteration 2:   log likelihood = -182.41553
Iteration 3:   log likelihood = -182.11942
Iteration 4:   log likelihood =   -182.119
Iteration 5:   log likelihood =   -182.119

Logistic regression                               Number of obs   =        463
                                                  LR chi2(3)      =     238.87
                                                  Prob > chi2     =     0.0000
Log likelihood =   -182.119                       Pseudo R2       =     0.3961

------------------------------------------------------------------------------
     wonseat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      incumb |   3.200883   .8391721     3.81   0.000     1.556136     4.84563
 spend_total |   .0001604   .0000232     6.92   0.000     .0001149    .0002058
spend_tota~c |  -.0000649   .0000428    -1.52   0.130    -.0001488     .000019
       _cons |  -3.901699    .429417    -9.09   0.000    -4.743341   -3.060057
------------------------------------------------------------------------------
```

Interpreting coefficients: example

In our example, the estimated probability of winning a seat for a challenger is therefore:

log(estimated odds of winning seat) =

$$-3.902 + 3.2009X_{\text{incumb}} + .00016X_{\text{spending}} - .00006X_{\text{inc}*\text{spending}}$$

- (for a challenger) Each additional €1 increases the log odds of winning by .00016
- (for a challenger) Each additional €1 multiplies the odds of being a volunteer by $e^{.00016} = 1.00016$
- (regardless of spending) Being an incumbent multiplies the odds of being a volunteer by $e^{3.200883} = 24.5542$

Interpreting coefficients on dummy variables

- In a multiple logistic regression these are adjusted odds ratios, adjusting or controlling for the other explanatory variables in the model
- *Holding constant the values of other X variables,* the log odds is β units higher for $X_{dummy} = 1$ than when $X_{dummy} = 0$
- The odds of $Y = 1$ for $X_{dummy} = 1$ are $e^{\beta_{dummy}}$ times the odds of $Y = 1$ for $X_{dummy} = 0$
- For polytomous categorical X variables, with c categories and $(c - 1)$ dummy variables, each estimated coefficient compares the odds for the category of interest to the reference category

More on interpreting logit coefficients

The problem: How do we interpret coefficents in terms of $\Pr(Y = 1)$ for a one-unit change in X?

1. We can compute fitted values on probabilities using the formula for $\hat{\pi}_i = \frac{1}{1+e^{-X_i\beta}}$:

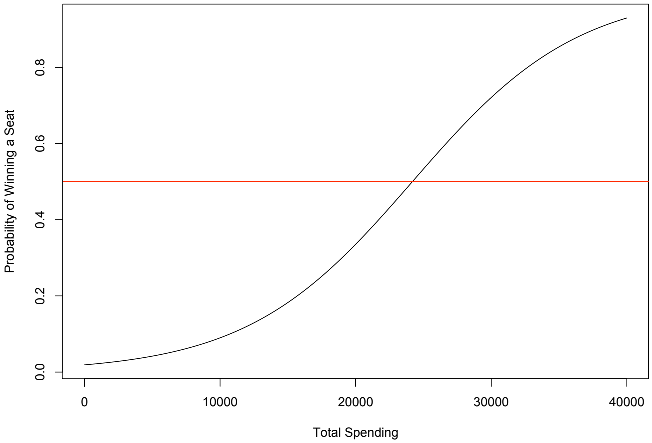
```
. clear
. set obs 9
obs was 0, now 9
. egen spendx = fill(0 5000 10000 15000 20000 25000 30000 40000)
. gen prchall = 1 / (1 + exp(-1*(-3.902 + .00016*spendx)))
. gen princ = 1 / (1 + exp(-1*(-3.902 + 3.2009*1 + .00016*spendx -.00006*1*spendx)))
. list, noobs clean

    spendx    prchall      princ
         0   .0198014   .3315684
      5000   .0430248   .4498937
     10000   .0909575   .5741736
     15000   .1821274   .6897391
     20000    .331369   .7856498
     25000   .5244804    .858015
     30000   .7105383   .9087859
     35000   .8452733   .9426163
     40000   .9240015   .9643911
```

Note: We have to make decisions about what to hold constant

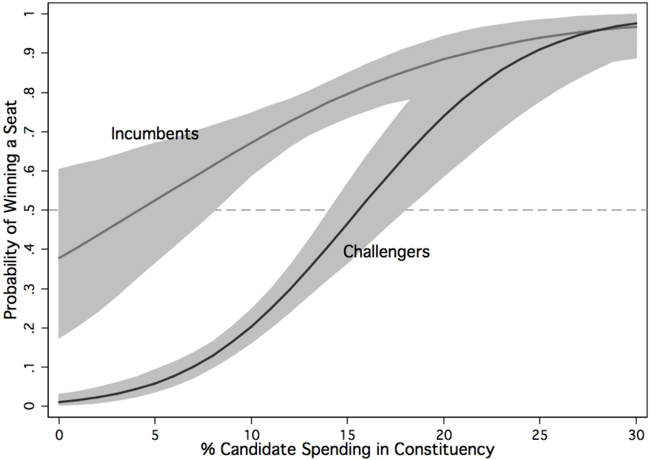
Interpreting logit coefficients cont.

2. We can use graphical methods plotting changes p_i by X:



Interpreting logit coefficients cont.

a slightly prettier version, with separate curves for both challengers and incumbents:



Interpreting logit coefficients cont.

3. We can compute first differences to show the effect of changes in X on p_i:

Change in % Spending (€)		Increase in Probability of Winning a Seat			
		Challengers		Incumbents	
From:	To:	Mean	S.E.	Mean	S.E.
0	5	**0.05**	(0.008)	**0.05**	(0.049)
5	10	**0.14**	(0.018)	**0.12**	(0.065)
10	15	**0.26**	(0.042)	**0.22**	(0.055)
5	15	**0.41**	(0.058)	**0.34**	(0.112)

- we have to choose plausible differences
- fitted values must be computed for each From, To point
- the calculation of standard errors is a different matter we have not yet covered (but will in Week 9)

Interpreting logit coefficients cont.

4. We can use derivative methods to show the instantaneous effect of changes in X_j on $P(Y_i = 1)$:

$$\begin{aligned} \frac{d\tilde{\pi}}{dX_j} &= \frac{d}{dX_j}\left[1 + e^{X_j\tilde{\beta}_j - X_*\tilde{\beta}_*}\right]^{-1} \\ &= \tilde{\beta}_j\tilde{\pi}(1-\tilde{\pi}) \end{aligned}$$

- so this depends on the size of $\tilde{\beta}$, which is the estimate of the coefficient
- it also depends on the size of $\tilde{\pi}$
  - this is maximized at $\tilde{\pi} = 0.5$
  - at $\tilde{\pi} = 0.5$, this quantity is $.5(1 - .5) = 0.25$
  - so $\tilde{\beta}/4$ is always the maximum instantaneous effect
  - this provides a very crude "rule of thumb" for interpreting logit coefficients: divide coefficient by four

Another example

- Socio-demographic determinants of infant mortality
- Response variable:
  Baby dies before first birthday ($1 = $ yes, $0 = $ no)
- Explanatory variables:

MATAGE	Maternal age in years
BI	Length of preceding birth interval in months (time between birth of child and last child)
URBAN	Type of region of residence (1=urban, 0=rural)
MATED	Maternal education ($1 = $ primary$+$, $0 = $ none)

Interaction between two categorical explanatory variables

Variable	$\hat{\beta}$
Constant	-1.70
MATAGE	0.04
BI	-0.03
URBAN	-0.70
MATED	-0.78
URBAN $\times$ MATED	0.50

- Interaction between mothers region of residence and level of education
- Logit = -1.70 + 0.04*MATAGE – 0.03*BI – 0.70*URBAN – 0.78*MATED + 0.50*URBAN*MATED

Interaction between two categorical explanatory variables

- Logit = -1.70 + 0.04*MATAGE – 0.03*BI – 0.70*URBAN – 0.78*MATED + 0.50*URBAN*MATED
- Let $A = -1.70 + 0.04 * \text{MATAGE}0.03 * \text{BI}$

	Education (MATED)	
Region (URBAN)	None (0)	Primary+ (1)
Rural (0)	A	$A - 0.78$
Urban (1)	$A - 0.70$	$A - 0.70 - 0.78 + 0.5 = A - 0.98$

Interaction between two categorical explanatory variables

- Convert the table of logits into a table of odds
- In this table, $B = \exp(A)$, which cancels out when we take ratios of odds
- Use the table to calculate a selection of odds ratios to examine joint effects of education and region on mortality risks

	Education (MATED)	
Region (URBAN)	None (0)	Primary+ (1)
Rural (0)	B	$B \exp(-0.78) = B \times 0.46$
Urban (1)	$B \exp(-0.70) = B \times 0.50$	$B \exp(-0.98) = B \times 0.38$

Some odds ratios to illustrate the interaction effects

	Education (MATED)	
Region (URBAN)	None (0)	Primary+ (1)
Rural (0)	B	$B \exp(-0.78) = B \times 0.46$
Urban (1)	$B \exp(-0.70) = B \times 0.50$	$B \exp(-0.98) = B \times 0.38$

- Conditional on MATED=0 (mother has no education)

$$\frac{\text{Odds(Urban)}}{\text{Odds(Urban)}} = \frac{0.50}{1} = 0.50 = \exp(-0.70)$$

- Conditional on MATED=0 (mother has no education)

$$\frac{\text{Odds(Urban)}}{\text{Odds(Urban)}} = \frac{0.38}{0.46} = 0.83 = \exp(-0.98 - (-0.78))$$

Some odds ratios to illustrate the interaction effects

	Education (MATED)	
Region (URBAN)	None (0)	Primary+ (1)
Rural (0)	B	$B\ \exp(-0.78) = B \times 0.46$
Urban (1)	$B\ \exp(-0.70) = B \times 0.50$	$B\ \exp(-0.98) = B \times 0.38$

- Conditional on URBAN=0 (rural)

$$\frac{\text{Odds(Primary+)}}{\text{Odds(None)}} = \frac{0.46}{1} = 0.46 = \exp(-0.78)$$

- Conditional on URBAN=1 (urban)

$$\frac{\text{Odds(Primary+)}}{\text{Odds(None)}} = \frac{0.38}{0.50} = 0.76 = \exp(-0.98 - (-0.70))$$

Some fitted probabilities to further illustrate the interaction

- For a 30-year old woman with 2 years since her last child

	Education (MATED)	
Region (URBAN)	None (0)	Primary+ (1)
Rural (0)	0.228	0.119
Urban (1)	0.128	0.100

- Combination of no education and rural residence increases chances of infant mortality

Interaction between two continuous explanatory variables

Variable	$\hat{\beta}$
Constant	-1.68
MATAGE	0.05
BI	-0.04
URBAN	-0.68
MATED	-0.80
URBAN $\times$ MATAGE	-0.0007

- Interaction between age of mother and time between birth of child and last child
- Logit $=$ –1.68 $+$ 0.05*MATAGE – 0.04*BI – 0.68*URBAN – 0.80*MATED – 0.0007*MATAGE*BI

Interaction between two continuous explanatory variables

- Logit = –1.68 + 0.05*MATAGE – 0.04*BI – 0.68*URBAN – 0.80*MATED – 0.0007*MATAGE*BI
- Let A = –1.68 – 0.68*URBAN – 0.80*MATED
- Make a table showing estimated logits for a selection of values of MATAGE and BI

Maternal age in years (MATAGE)	Length of preceding birth interval (BI)	
	12 months (low)	36 months (high)
20 (low)	$A + (20 \times 0.05) + (12 \times -0.04)$ $+(20 \times 12 \times -0.0007)$ $= A + 0.352$	$A + (20 \times 0.05) + (36 \times -0.04)$ $+(20 \times 36 \times -0.0007)$ $= A + -0.944$
40 (high)	$A + (40 \times 0.05) + (12 \times -0.04)$ $+(40 \times 12 \times -0.007)$ $= A + 1.184$	$A + (40 \times 0.05) + (36 \times -0.04)$ $+(40 \times 36 \times -0.0007)$ $= A - 0.448$

Interaction between two continuous explanatory variables

- Convert the table of logits into a table of odds
- In this table, $B = exp(A)$, which cancels out when we take ratios of odds

Maternal age in years (MATAGE)	Length of preceding birth interval (BI)	
	12 months (low)	36 months (high)
20 (low)	$B \times 1.42$	$B \times 0.39$
40 (high)	$B \times 3.27$	$B \times 0.64$

- Use the table to calculate a selection of odds ratios to examine joint effects of maternal age and birth interval on mortality risks

Some odds ratios to illustrate the interaction

Maternal age in years (MATAGE)	Length of preceding birth interval (BI)	
	12 months (low)	36 months (high)
20 (low)	$B \times 1.42$	$B \times 0.39$
40 (high)	$B \times 3.27$	$B \times 0.64$

- Conditional on MATAGE=20 (mother is 20 years old)

$$\frac{\text{Odds(BI} = 12)}{\text{Odds(BI} = 36)} = \frac{1.42}{0.39} = 3.64$$

- Conditional on MATAGE=40 (mother is 20 years old)

$$\frac{\text{Odds(BI} = 12)}{\text{Odds(BI} = 36)} = \frac{3.27}{0.64} = 5.11$$

- The effect of birth interval on infant mortality risks is greater for older than for younger mothers

Statistical significance in MLE

- Null hypothesis: $\beta_k = 0$
- Alternative hypothesis: $\beta_k \neq 0$
- Test statistic is the ratio of the estimated coefficient to its standard error:
$$z_k = \frac{\hat{\beta}_k}{\hat{\text{se}}(\hat{\beta}_k)}$$
- This z_k can be compared to the standard normal distribution ¡
- If $|z_k| > 1.96$, then reject H_0 at the $\alpha = .05$ significance level

Wald tests for single regression coefficients

- The Wald test statistic is the square of the z statistic:

$$\chi^2 = \left(\frac{\hat{\beta}_k}{\hat{\text{se}}(\hat{\beta}_k)} \right)^2$$

- Compare this to χ^2 distribution with df=1
- SPSS automatically calculates multivariate Wald test for polytomous categorical explanatory variables
- In Stata, `nltest`
- More on significance tests and model selection next week

Confidence intervals for coefficients

- Approximate 95% confidence intervals for β_k is:

$$\hat{\beta}_k + / - 1.96\hat{\sigma}_{\beta_k}$$

- Approximate 95% confidence interval for population odds ratio e^{β_k} is

$$e^{\hat{\beta}_k - 1.96\hat{\sigma}_{\beta_k}} \text{to} e^{\hat{\beta}_k + 1.96\hat{\sigma}_{\beta_k}}$$

- Note: This interval is asymmetric: its lower limit will be closer to the estimated odds ratio than upper limit will be
- To use the confidence interval to test H_0, reject H_0 if the interval contains 1.0

Likelihood ratio comparison test

- An alternative way of testing coefficients for significance
  - Individual coefficients
  - Several coefficients at once – including a categorical variable partitioned into multiple dummies, or combinations of separate variables
- Compare the likelihoods of two models: one including the variable(s) in question, one excluding them
- Likelihood $\propto$ probability of obtaining the observed pattern of results in the sample if that model were true (the larger the value, the better)
- Likelihood ratio test preferable to Wald test in small samples

Likelihood ratio comparison test

Consider two models:

- Model 1 is the simpler model, with likelihood L_1
- Model 2 is the more complex model, with likelihood L_2
  (nested do that M2 is M1 with some extra parameters)

H_0: more complex model is no better than simpler one

- If H_0 is true, then L_1 and L_2 will be similar – in other words, the ratio will be close to 1.0
- Instead of comparing "raw" likelihoods, we compare $-2 \log - \text{likelihood}$
- Likelihood ratio test statistic:

$$D = 2(\log L_2 - \log L_1 - \log L_1) = (-2\log L_1) - (-2\log L_2)$$

Likelihood ratio comparison test

Consider two models:

- If H_0 is true, then $D \sim \chi^2$ with degrees of freedom equal to the difference in the degrees of freedom in the two models (i.e. the number of extra parameters in the larger model)
- Small p-value for test statistic = evidence against H_0 — evidence that the bigger model is better, and that we should keep the extra variables
- Large p-value for test statistic = evidence for H_0 — evidence that the bigger model is no better, and that we should drop the extra variables

Goodness of fit

- Wald and likelihood ratio tests are tests of relative fit; compare nested models with more/fewer parameters
- Testing absolute fit is more difficult
- Need to, in some way, compare observed and expected values. For each unit (e.g.item respondent, in a survey data set), compare:
  - Observed value = value of Y (0 or 1)
  - Expected value = predicted probability that $Y = 1$, i.e. $\hat{\pi}_i$
- Various statistics exist, some much better than others
  - Pearson χ^2 goodness of fit test
  - Hosmer and Lemeshow goodness of fit test
  - Classification table and pseudo-R^2 measures

Pearson χ^2 goodness of fit test

- General form of Pearson χ^2

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

- For the logistic regression model, calculation is

$$\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\pi}_i)^2}{\hat{\pi}_i}$$

- When H_0 is true, test statistic follows a distribution with $\text{df} = n - k$ (where k is number of model parameters)
- Caution: this only works when expected values are each > 5 and probabilities are < 1
- So we cannot really use the statistic in this form, since we need to generate larger expected values

Hosmer & Lemeshow goodness of fit test

1. Arrange the observations in order or their predicted probabilities
2. Put them into g groups (denoted $j = 1, 2, \ldots, J$ of approximately equal sizes
   The idea is the units in the same group should have similar predicted probabilities, and therefore similar values on the explanatory variables
3. For each group, obtain
   - Number of cases with observed $Y = 1$, Y_{1j}
   - Sum of predicted probabilities that $Y = 1$, $\hat{\pi}_{1j}$
   - Number of cases with observed $Y = 0$, Y_{0j}
   - Sum of predicted probabilities that $Y = 0$, $\hat{\pi}_{0j}$

Hosmer & Lemeshow goodness of fit test

4. Calculate Hosmer and Lemeshow test statistic:

$$\chi^2 = \sum_{j=1}^{J} \left[\frac{(Y_{1j} - \hat{\pi}_{1j})^2}{\hat{\pi}_{1j}} + \frac{(Y_{0j} - \hat{\pi}_{0j})^2}{\hat{\pi}_{0j}} \right]$$

5. Obtain the p-value: test statistic $\sim \chi^2$ with df $= (G - 2)$
6. H_0: data were generated by the fitted model
   - If p is small, reject H_0, infer model is not a good fit
   - If p is large, fail to reject H_0, infer model is a good fit

Hosmer & Lemeshow example

From class/homework:

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	41.205	8	.000

Contingency Table for Hosmer and Lemeshow Test

		currently using a modern method of contraception = no		currently using a modern method of contraception = yes		
		Observed	Expected	Observed	Expected	Total
Step 1	1	405	362.272	96	138.728	501
	2	337	351.473	162	147.527	499
	3	350	366.788	182	165.212	532
	4	342	345.888	170	166.112	512
	5	318	330.320	181	168.680	499
	6	334	355.056	214	192.944	548
	7	329	346.588	221	203.412	550
	8	364	326.812	172	209.188	536
	9	307	306.454	220	220.546	527
	10	313	307.349	279	284.651	592

Classification table

- Classify:
  - $\hat{\pi}_i > 0.5$ as a predicted $\hat{Y}_i = 1$
  - $\hat{\pi}_i < 0.5$ as a predicted $\hat{Y}_i = 0$
- Then compare observed and predicted frequencies for $Y = 1$ and $Y = 0$

Classification Table[a]

Observed			Predicted: currently using a modern method of contraception — no	Predicted: currently using a modern method of contraception — yes	Percentage Correct
Step 1	currently using a modern method of contraception	no	3332	67	98.0
		yes	1819	78	4.1
	Overall Percentage				64.4

a. The cut value is .500

- A rather crude measure of how well the model fits the data, since it does not tell you how close your incorrect predictions were to correct predictions
- If proportion of $Y = 1$ is rare, then so all $\hat{\pi}_i > 0.5$, so fit may look very poor according to this diagnostic

Pseudo R^2 measures

- There are many of these, and little agreement on which one is best
- Broadly speaking, they involve comparing the likelihood of the null model (model containing only an intercept), L_N, with the likelihood of the model of interest, L_1, e.g.

$$\text{Pseudo}-R^2 = \frac{-2\log L_N - (-2\log L_1)}{-2\log L_N}$$

- SPSS reports two: Cox & Snell and Nagerlkerke, which are variations on the general idea
- Can be interpreted as proportional improvement in fit, but *not* as explained variance
- Not really common to rely on these – and are better avoided