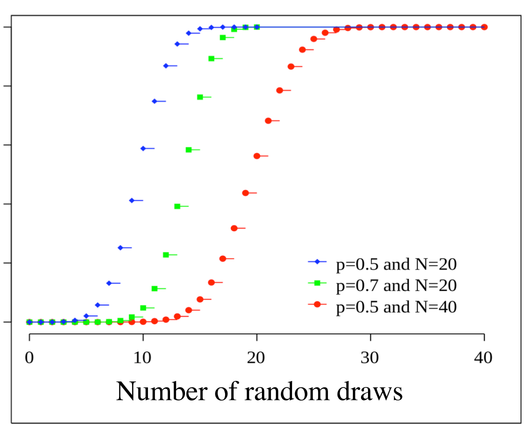
# Logistic Regression

**Lecture 6**
CLPS 2908 Multivariate Statistical Techniques
February 14, 2019

## Categorical DVs?

**Multiple Regression**

- Assumes normal error distributions — but binary outcomes have binomial error distributions
  - violates homoscedasticity



# Categorical DVs?

**Multiple Regression**

- Assumes normal error distributions — but binary outcomes have binomial error distributions
  - violates homoscedasticity
- MR → impossible predicted values: < 0 (no), > 1 (yes)

One option: Discriminant Function Analysis; later

**Logistic regression**

- No distributional assumptions about Y or ε
- Bounded between 0 and 1 (probability of event)
- No constraints on predictors [unlike logit; later]
- Fits <u>curved</u> regression line to *p(event)* ← *predictors*
- Maximum Likelihood estimation instead of least-squares (no linear variance decomposition)

# Logistic Regression Logic

$X_1, X_2, \ldots$

$u = B_0 + B_1X_1 + \ldots + B_pX_p$

$\rightarrow e^u$

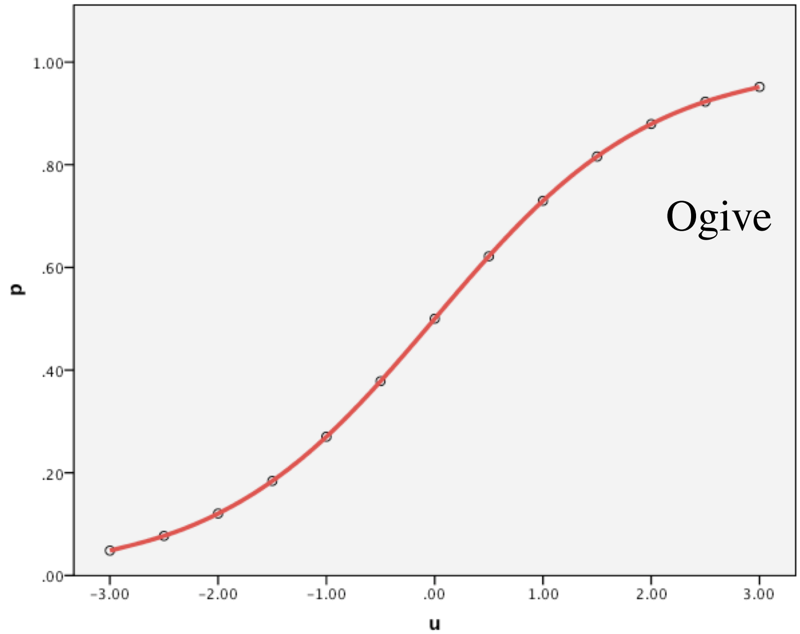
allows us to estimate:

$$\boxed{p(\text{event}) = \frac{e^u}{1+e^u}}$$

$$p = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

| u | $e^u$ | p |
|---|---|---|
| 3 | 20.10 | .95 |
| 2 | 7.40 | .88 |
| 1 | 2.72 | .70 |
| .5 | 1.65 | .62 |
| 0 | 1.0 | .50 |
| -.5 | .61 | .38 |
| -1 | .37 | .27 |
| -2 | .14 | .12 |
| -3 | .05 | .05 |

$p$ as a function of ln-transformed $u$ $\quad p = \dfrac{e^u}{1+e^u}$




# Model

- ML estimation: ***B* values are chosen** so as to minimize total prediction errors (over all $X_j$), subject to this constraint:

$$\sum_{i=1}^{n}(y_i - \hat{p}_i)x_{ij} = 0$$

= no error-predictor correlation

- Iterative process of re-estimating covariance matrix and B values *(→ best possible B parameters)*

- Log-likelihood (total remaining error ~ summed residuals)

$$\sum_{i=1}^{n}(y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i))$$

What contributes to LL being maximally negative (large error):
**predicted *p* is extreme (e.g., .05 or .95) but incorrect.**

*Example:* If $p$ = .05 for a given person whose real value $y_i$ is 1, then the LL term is:1*(-3.03) [right term drops out because 1-y = 0 in this case].

# Comparisons among Predictors in LR

- Weights $B_1$, $B_2$, etc. are again semi-partial weights

- The ***Wald*** statistic tests their significance (slightly conservative)

$$Wald_i = \frac{B_i^2}{SE_i^2}$$

- **Semi-partial *r***

$$r_{semi-partial} = \sqrt{\frac{Wald - 2df}{-2LL_0}}$$

  - Don't have 1 as their maximum, so they are typically **smaller** than what a normal correlation would give you.
  - Nevertheless, allow us to compare various predictors to each other
  - **Familiar:** Compare with zero-order *r*s to gain a sense of shared/suppressed predictive power.

- SPSS does not display these
- Compute them using this formula or the EXCEL file in the canvas folder.
- Null model $-2LL_0$ requires full output option.

Semi-partial calculator

Calculating semi-partial r's in logistic regression

| | | |
|---|---|---|
| WALD | 2 | [Wald must be greater than 2 for the computation to work] |
| df | 1 | For continuous predictors, df = 1; for categorical predictors with k categories, df = k-1 |
| -2LL(null) | 266.11 | |
| r | 0.0000 | |

# Comparisons (cont'd)

- **B**s are a little difficult to interpret in LR because they are "stuck" in the power of e

$$p(\text{event}) = \frac{e^u}{1+e^u}$$

- Pull them out: $u$ = ln(odds of event) = "log-odds" = logit

$$u = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1X_1 + \ldots + b_jX_j$$

- So every B-fold unit change of $X$ goes along with a change of the log-odds (still not particularly intuitive…).

Odds in statistics = "odds for" = $f$(event) vs. $f$(nonevents)
e.g., odds of rolling a "4" are 1/5

Odds in gambling = "odds against" (5/1)

# Exp(B)

- ***Exp(B)***, for a given $X$, is somewhat more interpretable: it is the factor by which the **actual odds** change with every B-fold unit change in X.
  - *Exp(B)* **around 1** is unimpressive;
    if **> 1**, predictor **increases** *p* (event);
    if **< 1**, predictor **decreases** *p* (event).

$$\frac{p}{1-p} = e^{b_0 + b_1 X_1 + \ldots + b_j X_j}$$

- *Exp(B)* is just the $ln$-transformation of raw B ($e^B$);
  *B* is semi-partial, so *exp(B)* must be semi-partial too.
- *That is, exp($B_1$)* represents the impact of its $X_1$ on $Y$ **while holding constant all other predictors**.

Exponent B effects.xls

# Overall Success

- $R^2$ analogs

$$R^2_{Cox-Snell} = 1 - \left( \frac{l(\text{null})}{l(\text{current})} \right)^{\frac{2}{W}}$$

- Menard suggests to compute **$R^2$ of predicted & observed $Y_i$ values:** [0;1] * 0,1 (SPSS can save predicted values, making it easy to compute the correlation)
- Most common:
  Case classification (“confusion matrix”)

Predicted Y

| Actual Y | | |
|---|---|---|
| | **32** | **5** |
| | **11** | **22** |

## Confusion Matrix

| Predicted / Real | "Present" | "Absent" | |
|---|---|---|---|
| **Present** | 30 | 20 | 50 |
| Absent | 10 | 40 | 50 |
| | 40 | 60 | |

**Two types of errors:**

False alarm = 10/50 = 20%

Misses = 20/50 = 40%

Diagnosticity:
30/40

## Two Ways to Take Control

- SPSS allows you to define contrasts for categorical predictors in this subcommand: /CONTRAST (*predictor*)= [HELMERT, POLYNOMIAL, etc.] (Read about it in the SPSS syntax reference file.)
- You can change classification cut-off: /CRITERIA = CUT(0.5)

Additional Literature:

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: Wiley.

Menard, S. W. (2002). *Applied Logistic Regression Analysis*. Sage.

## Model Building and Testing

- -2LL ratios for models with/without *Xj*

$$LR = -2\ln\left(\frac{l(reduced)}{l(full)}\right) \quad LR = -2(LL(reduced) - LL(full))$$

- Define predictor sets, guided by hypotheses
- Best subsets (time consuming, out of fashion)
- Goodness of fit ( $\chi^2$ ).
  - Actually, **badness of fit**

$$\sum_{i=1}^{n} \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)}$$

- Run the analysis with and without the predictor and compare the classification results.

## Sample Output

**LOGREG output.pdf**
(with annotations)

```
LOGISTIC REGRESSION VARIABLES natur6
/METHOD=ENTER degree6 contr6 source6
/CLASSPLOT
/CASEWISE
/PRINT = all.
```