Longitudinal Modeling with Logistic Regression

Longitudinal designs involve repeated measurements of the same individuals over time. There are two general classes of analyses that correspond to two definitions of change, the absolute level definition and the relative level definition.

Absolute Level Definition of Stability and Change with a Continuous Dependent Variable

Let's assume a continuous dependent variable for a moment. One type of question involves whether the scores on a variable increase or decrease over time. To answer this question, we can subtract one score from another, usually $Y_{2\text{-}1} = Y_2 - Y_1$, where the score at the first time point is subtracted from the score at the second time point. Referred to as *difference score*s, *gain scores*, or sometimes *changes scores*. The mean of the difference score indicates an average increase if it is positive and an average decrease if it is negative. A variable is considered stable on average to the extent the average difference is 0, and the degree of change is measured by the extent to which the average difference is greater or less than 0.

The statistical test to determine whether the average difference is different from zero is a paired (dependent-samples) *t*-test or within-subjects ANOVA. The average of the differences equals the differences of the averages, $E(Y_2 - Y_1) = \overline{Y}_2 - \overline{Y}_1$, so this test is also a test of whether the mean at Time 1 differs from the mean at Time 2. The downside to this definition of change is that regression toward the mean might be mistaken for changes over time and the change scores are sensitive to any random or temporary fluctuations in the values.

Relative Level Definition of Stability and Change with a Continuous Dependent Variable

An alternative way to think about stability and change is in terms of the correlation of a variable with itself over time. Stability is defined in this way by the size of the correlation between Y_2 and Y_1, with $r_{12} = 1.0$ indicating a perfectly stable variable. In this sense of the meaning of stability, there is some change if the correlation is less than perfect. This definition is distinct from the absolute definition of change, because the correlation between the two variables can be perfect even if there is a large increase or decrease from Time 1 to Time 2. Adding 5, 10, or even 100 points to all of the Y_2 scores would not change the correlation. This is because the correlation coefficient is more a function of the correspondence in the relative position of the scores in the sample than a metric of how well the scores match in absolute terms. The focus under this definition of stability and change has to do with the strength of the relationship over time, and it does not provide information about whether the scores increase or decrease over time. Longitudinal analyses involve a correlation between Y_1 and Y_2 or *autoregression* with Y_1 predicting Y_2. The residual in an autoregression represents the degree to which the variable is not stable over time or changing in the relative level sense.

Implications of the Definitions of Stability and Change for Longitudinal Analyses

The distinction between these two ways of thinking about stability and change is important for thinking about the types of questions answered by certain analyses (see Newsom, 2015, Chapter 4 for a more complete discussion). Many familiar repeated measures or longitudinal analyses, such as the paired t test, repeated measures ANOVA, regression with difference scores, growth curve models, latent difference score/change models, are a function of the degree of level stability or change. Others, such as ANCOVA, lagged regression, and cross-lagged panel models, are a function of the degree of relative stability or change (correlation). We can also distinguish between these conceptualizations of change in terms of analysis for binary variables. Below I consider use of logistic regression models to explore change in a binary variable over time and prediction of that change, distinguishing how they differ in terms of the level-change and relative-change concept.

Absolute Level Definition of Stability and Change with a Binary Dependent Variable

For binary variables, difference scores for Y also can be computed. The difference score for a binary variable, $Y_{2\text{-}1} = Y_2 - Y_1$ can only take on three possible values, -1, 0, and +1. A value of -1 represents a

decrease over time (moving from 1 to 0), a 0 indicates no change, a perfect match whether 0-0 or 1-1, and +1 represents an increase over time (moving from 0 to 1). Our usual test is not a test of whether the binary variable changes from Time 1 to Time 2 is not based on this difference score, however. The interest is usually in whether the proportion of one value of Y increases or decreases over time, such as comparing the proportion of "yes" votes at Time 1 and Time 2. This hypothesis, of marginal homogeneity that $\pi_{1+} = \pi_{+1}$., is usually tested using the McNemar chi-square.[1] Although not exactly equivalent. loglinear analysis also could be used to test the hypothesis that the proportions change over time. The likelihood ratio test in the loglinear model involves a test of the natural log of the ratio of observed and expected frequencies, or a difference between the log of the expected and log of the observed frequencies, whereas the McNemar chi-square does not involve a log transformation. [2]

When a log transformation is involved, the concept of difference of proportions gets trickier, and, depending how the statistical test is computed, value of the difference between proportions may not be the same. Mathematically, this is because the difference between two logged averages does not equal the average of the log of the differences between two scores. [3]

$$\ln\left(p_2 - p_1\right) \neq \left[\ln\left(p_2\right) - \ln\left(p_1\right)\right]$$

The binary nature of the dependent variable also raises the possibility of a test of a different hypothesis—that the number of voters changing their vote is different from the number of voters that stay consistent. Test of such a hypothesis would involve computing a new variable with 0 representing no change and one representing change (where the 0 to 1 and 1 to 0 change patterns are combined). The distinction from the McNemar test is subtle in that this hypothesis test focuses on a non-directional change.

Relative Level Definition of Stability and Change with a Binary Dependent Variable

Although there is still a distinction between stability as defined by correlation and as defined by absolute change when we are dealing with binary variables, the contrast is more complicated to conceptualize in terms of individual values. The proportion can remain unchanged over time without voters responding identically at both time points, for instance. But we can distinguish between the extent to which responses are identical (level stability) and the correlated (relative stability) over time. Suffice it to say that, with the exception of the trivial case in which there is a perfect match between Time 1 and Time 2 binary values, the correlation (i.e., are stable in the relative value sense) will not necessarily reflect the proportion of cases that match (i.e., are stable in level value sense).

Modeling Absolute Level Change with Logistic Regression

Conditional logistic regression. In addition to the McNemar test change in a binary variable over two or more time points can be assessed with a *conditional logistic* regression model. Both of these analyses are conceptually akin to the paired t test of whether the average difference between two time points is different from zero. They differ slightly from one another because the McNemar tests changes in the aggregate and conditional logistic regression focuses on changes at the individual level. The conditional logistic approach also models change using natural logarithms for predicted values. The conditional logistic model (Cox, 1958) examines change over time using a score for time x_t, as a predictor of Y_t, measured repeatedly at two (or more) time points.

$$\text{logit}\left[P\left(Y_{it} = 1\right)\right] = \alpha_i + \beta x_t$$

---

[1] Recall that the equivalence of marginal proportions is algebraically equivalent to the equivalence of the discordant cell proportions, that $p_{12} = p_{21}$. Refer to the "Matched Pairs Analysis" handout for more on this equivalence. The sign test, which compares the number of increasing cases ('+'s) to the number of decreasing cases ('-'s) is equivalent to McNemar's test.

[2] Recall that the quotient rule states that $\ln\left(n_{ij} / \mu_{ij}\right) = \ln\left(n_{ij}\right) - \ln\left(\mu_{ij}\right)$

[3] Although this discussion relies on the connection between the average score and the proportion, the natural log depends on the values used for a binary variable and the value of $\ln\left(\bar{Y}_2\right) - \ln\left(\bar{Y}_1\right)$ will not be the same if 1 and 2 are used rather than 0 and 1 for the coding of Y.

This is a typical logistic regression, except for two things. The dependent variable, Y_{it}, has values for each person and each time point, indicated by the subscript ij, and the intercept α_i has a different value for each person.[4] x_t is some arbitrary time code, usually beginning 0,1, for however many time points are available. The term "conditional logistic" is used because the regression is conditioned on subject. To conduct the analysis, data must be transformed into a long (or person × time) format. If there are 100 cases measured at two time points, the data set will have 200 records. In general, the long format data set has $n \times T$ records.

i	x_t	Y_{ij}
1	0	0
1	1	1
2	0	1
2	1	1
.		
.		
.		
99	0	1
99	1	0
100	0	0
100	1	1

Notice that, if we were using OLS (linear) regression with x_t predicting Y_{ij}, the slope for the predictor x_t would be equal to the average difference in Y between the two time points (because the definition of the slope is the change in Y for each unit change in X), and, thus, also equal to the difference between the means at Time 1 and Time 2, $\overline{Y}_2 - \overline{Y}_1$.

A standard logistic regression procedure can be used to estimate the conditional logistic model if the analysis can be stratified by subject (e.g., PROC LOGISTIC in SAS)[5] or survival analysis procedures can be used (special procedures are needed because the sample size will be double and each pair of scores will be correlated). The regression coefficient in the conditional logistic model represents the change in the logit for each unit change in x_t, with positive values indicating an increase over time and negative values indicating a decrease over time. The odds ratio is the odds increase or decrease with each unit increase in x_t. For two time points, it's the odds that $Y = 1$ increase from the first time point to the second, or,

$$\beta = \ln\left(\frac{n_{21}}{n_{12}}\right)$$

And notice that these are the same two cells, the discordant cells, that are compared in McNemar's test. The distinction is that the conditional logistic represents a test of the *subject-specific* (or conditional) *effect* and the McNemar's chi-squared represents a test of the *population-averaged* (or marginal) effect.[6] The involvement of the natural log makes them different. The two approaches to testing whether a binary variable has changed over time will generally lead to the same conclusion, but there are instances in which they may differ.

The intercept α_i gives the logit value of Y when X equals 0, or the first time point. The logistic transformation can be used to obtain the estimate of the probability that $Y = 1$ at the first time point, where

$$P\left(Y_{it} = 1\right) = \frac{e^{\alpha_i}}{1 + e^{\alpha_i}}$$

[4] If you are familiar with multilevel modeling, you will know this variation in the intercept to represent a random intercept model. Note that the slope is not varying, so it is not a random slope model.
[5] Agresti (2013) shows that the Cochran-Mantel-Haenszel (CMH) analysis, which examines the X-Y association stratifying by Z, is the same as the conditional logistic analysis.
[6] Technically, the subject-specific effect is limited to the intercept. With multilevel models for non-continuous outcomes, it is possible to have subject-specific estimates for slopes as well.

Generalized estimating equations. An alternative modeling approach to change over time, generalized estimating equations (GEE; Liang & Zeger, 1986) works much the same way. The GEE model accounts for correlated observations, and, for longitudinal data, it takes into account the dependence that occurs with multiple observations per person. The analysis uses quasi-likelihood estimation rather than the usual Newton-Raphson maximum likelihood estimation of logistic regression. And it couples the estimation with robust standard errors (Huber-White or "Sandwich" estimator; Huber, 1967; White, 1980) to account for the correlated residuals. The GEE approach is sometimes referred to as a population-averaged model, however, where the model is not conditional on subject.

$$\text{logit}\left[P\left(Y_{it}=1\right)\right]=\alpha+\beta x_t$$

But a version of the approach can assume "exchangeable units" within clusters, and estimate intercepts that vary across clusters (individuals in the longitudinal case). When intercepts but not slopes vary across individuals, the results are partially subject-specific. Predictors can be included to predict change from Time 1 to Time 2. Predictors that are measured at the person-level (e.g., ethnicity) are known as time-invariant covariates, and predictors measured at a specific time point (e.g., self-confidence when predicting passing grade) are known as time-varying covariates.

The GEE modeling approach is very popular and performs well statistically with longitudinal or clustered observations and can be used with missing data. The GEE model is not as flexible as a *multilevel regression* approach (e.g., Raudenbush & Bryk, 2002), also known as hierarchical linear modeling, random coefficient models, or *growth curve analysis* when applied to longitudinal data. A variant of growth curve models when estimated with structural equation modeling software are known as *latent growth curve models*. Growth curve models have several advantages over GEE models, including estimation of intercept and slope random effects that provide information about individual variation in change over time. For binary outcomes, multilevel models can provide population-averaged as well as subject-specific estimates.

Modeling absolute change with difference scores and ordinal logistic regression. Difference scores can be predicted directly using an ordinal logistic (or probit) model. In such a model, a difference score for Y would be computed by subtracting, with $Y_{2\text{-}1} = Y_2 - Y_1$ resulting in three values for $Y_{2\text{-}1}$ of -1, 0, and +1. The dependent variable is thus a ordinal value such that negative values indicate a decrease, zero indicates no change, and positive values indicate an increase over time. Any number of predictors can be included to predict the propensity to increase over time.

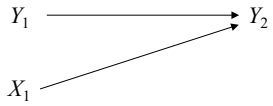
$$\text{logit}\left[P\left(Y\le j\mid X\right)\right]=\alpha_j+\beta X$$

The left side of the equation is a logit function of the probability of incrementing from either -1 to 0, or 0 to 1. We discuss ordinal logistic (and probit) models in the next section of the course.

Modeling Relative Level Change: The Lagged Regression Approach

The methods discussed above—conditional logistic, GEE, and multilevel models—estimate level changes over time. That is, they answer the questions about whether the observed variable Y_{it} increases from one time-point to the next. And when covariates are used to account for this change, the model describes who increase or decreases over time. Such models do not address questions about whether X may be a cause of Y or Y a cause of X. When experimental data are not available to investigate this question, longitudinal studies have advantages over cross-sectional studies for addressing this question.

The lagged regression model uses an independent variable measured at Time 1 to predict values at Time 2 controlling for the dependent variable measured at Time 1.



The analysis involves the relative level definition of change, because the path from Y_1 to Y_2 represents stability in the correlational sense.[7] The remaining variance in Y_2 not accounted for by Y_1 is what changes (again in the relative correlation sense). The model can be interpreted in two ways, as prediction of Y_2 while controlling for any preexisting differences in the dependent variable at Time 1 or and a prediction of change in Y_2. A special case of the model in ordinary least squares regression in which X_1 is binary, is an analysis of covariance (ANCOVA), and so the lagged regression model is sometimes referred to as the ANCOVA approach. This modeling approach has the advantage of accounting for regression toward the mean. The results from the conditional logistic and the lagged regression model will not always lead to the same conclusion (known as Lord's paradox), but that is because they address different questions about change.

Examples

Below I use SAS to illustrate several simple longitudinal models using the depression data from the LLSSE (a new data, dep2.sav, set with Time 1 predictors was created). The outcome is the binary depression variable from the 9-item CES-D using the recommended clinical cutoff. The lagged regression example is easily replicated in SPSS (`logistic regression`) or R (`glm`) using syntax I have already illustrated elsewhere.

Lagged logistic regression example with Time 1 negative social exchanges predicting depression depression at Time 2, controlling for Time 1 depression:

```
proc logistic data=one ;
  model w2dep = w1dep w1neg;
  run;

Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio        94.9888        2         <.0001
Score                   91.0140        2         <.0001
Wald                    81.4675        2         <.0001

The LOGISTIC Procedure

             Analysis of Maximum Likelihood Estimates

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

Intercept     1     -0.0665      0.1303        0.2607        0.6096
w1dep         1      1.6401      0.2005       66.8942        <.0001
w1neg         1      0.4519      0.2014        5.0333        0.0249


           Odds Ratio Estimates

              Point          95% Wald
Effect    Estimate      Confidence Limits

w1dep        5.156       3.480       7.638
w1neg        1.571       1.059       2.332
```

[7] Menard (2010) makes the important point that in logistic regression the association between the dependent variable measured at Time 1, Y_1, and the outcome is a nonlinear relationship because the predicted values are in logit form. He recommends a fully standardized logistic solution to interpret the Y_1 to Y_2 relationship as representing stability. Using a probit analysis with standardized solution or the linear binomial regression would be alternative approaches.

Sample Write-Up

Clinical depression measured at the six-month follow-up was regressed on negative social exchanges and clinical depression measured at baseline. Both baseline predictors significantly predicted clinical depression at follow-up. As expected, those who were depressed at baseline were more likely to be depressed at follow-up, $b = 1.64$, $SE = .13$, $OR = 5.16$, $p < .001$. Of greater interest was that those who reported more frequent negative social exchanges were more likely to be depressed at follow-up after controlling for initial depression, $b = .45$, $SE = .20$, $OR = 1.57$, $p = .02$, suggesting that negative social exchanges at baseline were predictive of an increase in clinical depression over the six-month interval.

Questions about absolute level change (difference of binary variable with no predictor) can be tested with the McNemar's test and conditional logistic. Conditional logistic requires configuration of the data to a long (person × period) data set. In SAS, ID is used as a strata to allow the intercept to vary by case, α_i. A survival analysis approach is an alternative method for obtaining a conditional regression, which would be needed in SPSS (`coxreg`) and R (`clogit` in the survival package). I only illustrate the conditional logistic here for didactic reasons, so I don't give code for either of these models. I recommend growth curve models (or possibly GEE if there are only two time points) instead of conditional logistic for this general approach.

```
/* mcnemar's test */
proc freq;
tables w1dep*w2dep /agree;
run;
```

w1dep(Time 1 depression) × w2dep(Time 2 depression)

Frequency / Percent / Row Pct / Col Pct	not depressed	depressed	Total
not depressed	146 22.43 48.50 75.65	155 23.81 51.50 33.84	301 46.24
depressed	47 7.22 13.43 24.35	303 46.54 86.57 66.16	350 53.76
Total	193 29.65	458 70.35	651 100.00

McNemar's Test

Statistic (S)	57.7426
DF	1
Pr > S	<.0001

```
/*reconfigure data set to be long format */
data one; set one;
id = _N_;
run;

DATA long ;
  SET one ;

  t = 0 ;
  dep = w1dep ;
  OUTPUT ;

  t = 1 ;
  dep = w2dep;
  OUTPUT ;

  keep id t dep ;
RUN;

/*conditional logistic regression */
proc logistic  data=long  order=data descending;
     strata id;
     MODEL dep=t ;
run;
```

```
The LOGISTIC Procedure
```

```
Conditional Analysis

        Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio        60.8670        1         <.0001
Score                   57.7426        1         <.0001
Wald                    51.3524        1         <.0001


       Analysis of Conditional Maximum Likelihood Estimates

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

t             1     -1.1933      0.1665       51.3524        <.0001


           Odds Ratio Estimates

              Point          95% Wald
Effect    Estimate      Confidence Limits

t            0.303       0.219       0.420
```

Notice that the score test from the conditional logistic matches the McNemar's test. Both are a test of the population-averaged (marginal) difference. The Wald test from the conditional logistic does not match either, because it is a test of the subject-specific (conditional) effect.

References and Further Readings

Agresti, A. (2013) *Categorical data analysis, third edition*. New York: Wiley.

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, *45*(3/4), 562-565.Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., & Pentz, M. A. (1998). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology*, *147*, 694-703.

Hanley, J. A., Negassa, A., & Forrester, J. E. (2003). Statistical analysis of correlated data using generalized estimating equations: an orientation. *American journal of epidemiology, 157*(4), 364-375.

Huber, Peter J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 221–233.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13-22.

Newsom (2012). Basic Longitudinal Analysis Approaches for Continuous and Categorical Variables. In Newsom, J.T., Jones, R.N., & Hofer, S.M. (Eds.) (2012). *Longitudinal Data Analysis: A Practical Guide for Researchers in Aging, Health, and Social Science.* New York: Routledge (pp. 168-170).

Newsom, J.T. (2015). *Longitudinal Structural Equation Modeling: A Comprehensive Introduction*. New York: Routledge.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica, 48*, 817–838.