Lecture 7
Logistic Regression with Random Intercept
Logistic Regression

Odds: expected number of successes for each failure

\[
\log \left( \frac{P(y_i \mid x_i)}{1 - P(y_i \mid x_i)} \right) = \beta_1 + \beta_2 x_i
\]

\[
\log \{ Od(y_i = 1 \mid x_i = a + 1) \} - \log \{ Od(y_i = 1 \mid x_i = a) \} = \beta_2
\]

\[
\frac{Od(y_i = 1 \mid x_i = a + 1)}{Od(y_i = 1 \mid x_i = a)} = \exp(\beta_2)
\]

Log-odds ratio

Odds ratio
Women Employment status
(womenlf.dta)

- “Workstat”: employment status (0: not working, 1: working part-time, 2: working full time)
- “Husbinc”: husband income in $1000
- “Childpres”: child present in the household (dummy variable)
Logistic regression model

\[
\text{logit}P(y_i = 1 \mid x_{2i}, x_{3i}) = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i}
\]

Table 4.1: Maximum likelihood estimates for women’s labor force participation
Parameter’s interpretation in logistic regression

• Women who don’t have a child at home are 5 times more likely to be working (1/0.21) than women that have a child at home controlling for husbands income

• Within the two groups of women (the ones that have a don’t have a child), each extra $1,000 of husband’s income reduces the odds of working by about 4% [(1-0.96)X100]
Standard errors

- Standard errors of exponentiated regression coefficients should generally not be used for confidence intervals or hypothesis tests.
- Instead the 95% confidence intervals of the above output were computed by taking the exponentials of the confidence limits for the regression coefficient

\[ \exp\{\hat{\beta} \pm 1.96 \times SE(\hat{\beta})\} \]
Visualization of the predictive probabilities

\[ \pi_i = \frac{\exp(\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i})}{1 + \exp(\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i})} \]

Women without a child

Women with a child

Figure 4.1: Predicted probabilities from logistic regression model
Figure 4.2: predicted probabilities from logistic regression model, extrapolating outside the range of the data.
Latent Response formulation of a logistic regression model

• These models assume that underlying the observed dichotomous response (whether the women works or not), there is an **unobserved or latent continuous response**, representing the propensity to work. If this latent response is greater than zero, then the observed response is 1:

\[
y_i^* = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i
\]

\[
y_i^* > 0 \implies y_i = 1
\]

\[
y_i^* \leq 0 \implies y_i = 0
\]

\[
E(\varepsilon_i \mid x_i) = 0
\]
Latent Response formulation of a logistic regression model

• In logistic regression the error $\varepsilon_i$ is assumed to have a logistic cumulative density function given $x$,

$$\Pr(\varepsilon_i < \tau \mid x_i) = \frac{\exp(\tau)}{1 + \exp(\tau)}$$

$$E[\varepsilon_i \mid x_i] = 0$$

$$Var[\varepsilon_i \mid x_i] = \frac{\pi^2}{3} \approx 3.29$$
When a latent-response formulation is used, it seems natural to assume that $\mathcal{E}_i$ has a normal distribution given $x$, as is usually done in linear regression. If a standard (mean zero variance 1) normal distribution is assumed, the model becomes a probit model

$$
\Pr(y_i = 1 \mid x_i) = \Pr(y_i^* > 1 \mid x_i) = \\
\Pr(\beta_1 + \beta_2 x_i + \mathcal{E}_i > 0) = \Pr(\mathcal{E}_i > -(\beta_1 + \beta_2 x_i)) = \\
\Pr(-\mathcal{E}_i \leq \beta_1 + \beta_2 x_i) = \Pr(\mathcal{E}_i \leq \beta_1 + \beta_2 x_i) = \Phi(\beta_1 + \beta_2 x_i)
$$

Standard normal cumulative distribution function
Which treatment is best for toenail infection (toenail.dat)?

- Randomized, double-blind clinical trial of two competing antifungal treatments for toenail infection (250mg/day terbinafine and 200 mg/day itraconazole)
- 378 patients were randomly allocated to two treatment groups and evaluated at seven visits at weeks (0,4,8,12,24,36, and 48)
- Outcome: onycholosis (the degree of separation of the nail plate from the nail bed) which has been dichotomized ("moderate or severe" versus "none or mild")
The data set includes the following variables

- **Patient:** patient identifier
- **Outcome:** onycholosis (0, none to mild, 1 moderate or severe)
- **Treatment:** 0:itraconazole; 1:terbinafine
- **Visit:** visit number (1,2,…..,7)
- **Month:** exact timing of the visit in months
Research question

• Do patients receiving one treatment experience a greater decrease in their probability of having onycholosis than those receiving the other treatment?

• The data set is not balanced since all patients did not attend all planned visits.
• 224 have complete data
• 21 missed the 6-th visit
• 10 missed the 5-th visit
• Monotone pattern of missing data: most of the patients dropped at one of the visit and never returned
MLE and Missing at Random

• A nice feature of MLE for incomplete data is that all the information is used. Thus not only patients who attended all the visits, but also patients with missing visits contribute information.

• This is true as long as the data are Missing at Random (MAR).
Missing at Random and Missing Completely at Random

• MAR: used to describe situations where response and explanatory variables are recorded but the response may be missing with a probability independent of its unobserved value

• MCAR: if the probability is also independent of the explanatory variables
Barplot of the proportion of patients with toenail infection by visit and treatment group

![Graph showing proportion of patients with toenail infection by visit and treatment group using itraconazole and terbinafine.](image)

Figure 4.6: Bar plot of proportion of patients with toenail infection by visit and treatment group
Marginal or Population average probabilities

- The figure shows the estimated average (or marginal) probabilities of oncholysis given: 1) time since randomization; and 2) treatment group.
- We are not attempting to estimate individual subject’s personal probabilities, which might well vary substantially, but are considering the population averages, given the covariates.
Marginal logistic regression model

(i) is the occasion, (j) is the patient

\[
\log \text{it} \left\{ P(y_{ij} = 1 \mid x_{2j}, x_{3ij}) \right\} = \beta_1 + \beta_2 x_{2j} + \beta_3 x_{3ij} + \beta_4 x_{2j} x_{3ij}
\]

This model allows for:

• difference between groups at baseline (beta2)
• linear changes in the log-odds of infection over time with slopes (beta3) for the itraconozole group and slope (beta3+beta4) for the terbinafine group
• beta4 is the difference in the rate of improvement (on the log odds scale) between treatment groups (treatment effect)
This model makes the unrealistic assumption that the responses for a given patient are conditionally independent given the covariates.
Logistic regression with random intercept
(\texttt{xtlogit,xtmelogit,gllamm})

\[ y_{ij} \mid \pi_{ij} \sim \text{Binomial}(1, \pi_{ij}) \]

\[ \pi_{ij} = P(y_{ij} = 1 \mid x_{2j}, x_{3ij}, \zeta_j) \]

\[ \logit\{\pi_{ij}\} = \beta_1 + \beta_2 x_{2j} + \beta_3 x_{3ij} + \beta_4 x_{2j} x_{3ij} + \zeta_j \]

\[ \zeta_j \sim N(0, \psi) \]

The random intercept represents the combined effect of all omitted \textbf{subject-specific} covariates that causes some subjects to be more prone to the disease than others.
Table 4.2: Estimates for toenail data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Marginal effects</th>
<th>Conditional effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ordinary logistic</td>
<td>GEE logistic</td>
</tr>
<tr>
<td></td>
<td>OR (95% CI)</td>
<td>OR (95% CI)‡</td>
</tr>
<tr>
<td>part</td>
<td></td>
<td></td>
</tr>
<tr>
<td>treatment</td>
<td>1.00 (0.74, 1.36)</td>
<td>1.01 (0.61, 1.68)</td>
</tr>
<tr>
<td>month</td>
<td>0.84 (0.81, 0.88)</td>
<td>0.84 (0.79, 0.89)</td>
</tr>
<tr>
<td>trt_month</td>
<td>0.93 (0.87, 1.01)</td>
<td>0.93 (0.83, 1.03)</td>
</tr>
<tr>
<td>likelihood</td>
<td>-908.01</td>
<td></td>
</tr>
</tbody>
</table>

Variance of the random effects: 625.39

ICC: -188.94
Results

- Random Intercept model: significant treatment effect, with terbinafine having a greater downward slope for the log odds than itraconazole

- Odds ratio is 0.68 per month in the itraconazole group and 13% lower (equal to $0.68 \times 0.87 = 0.59$) in the terbinafine group (for a patient with random intercept equal to zero)
Parameters Interpretation

\[
\frac{Odds(y_{ij} = 1 \mid x_{2j} = 0, x_{3ij} = a + 1, \zeta_j)}{Odds(y_{ij} = 1 \mid x_{2j} = 0, x_{3ij} = a, \zeta_j)} = \exp(\beta_3)
\]

Odds of infection per month in the itraconazole group for each patient

\[
\frac{Odds(y_{ij} = 1 \mid x_{2j} = 1, x_{3ij} = a + 1, \zeta_j)}{Odds(y_{ij} = 1 \mid x_{2j} = 1, x_{3ij} = a, \zeta_j)} = \exp(\beta_3 + \beta_4)
\]

Odds of infection per month in the terbinafine group for each patient

Results: The odds decrease by 32% \((100 \times (1 - OR))\) in the itraconazole group and by 42% in the terbinafine group and this difference is statistically significant at the 5% level
Marginal and Individual Probabilities

- Marginal (ordinary) logistic regression models the overall (population-averaged) probabilities.
- Random effects logistic regression models the individual (subject-specific) probabilities.
Marginal and Individual probabilities

A: Marginal Logistic regression

$$\text{logit}\left\{ P(y_{ij} = 1 \mid x_{2j}, x_{3ij}) \right\} = \beta_1 + \beta_2 x_{2j} + \beta_3 x_{3ij} + \beta_4 x_{2j} x_{3ij}$$

marginal prob

B: Random Intercept Logistic regression

$$\text{logit}\left\{ P(y_{ij} = 1 \mid x_{2j}, x_{3ij}, \zeta_j) \right\} = \beta_1 + \beta_2 x_{2j} + \beta_3 x_{3ij} + \beta_4 x_{2j} x_{3ij} + \zeta_j$$

individual prob
The population average probabilities implied by the random-intercept model can be obtained by averaging the subject-specific probabilities over the random-intercept distribution. Since the random intercepts are continuous, this averaging is accomplished by integration

\[
P^*\left( y_{ij} = 1 \mid x_{2j}, x_{3ij} \right) = \\
= \int P(y_{ij} = 1 \mid x_{2j}, x_{3ij}, \varsigma_j) \phi(\varsigma_j; 0, \psi)d\varsigma_j \neq \\
\neq P\left( y_{ij} = 1 \mid x_{2j}, x_{3ij}, \varsigma_j \right)
\]

The difference between the population-averaged and subject specific effects is due to the fact that average of non linear function is not the same as the non linear function of the average.

Subject-specific curves for different values of the random effect can be obtained with the stata command `gllapred`.\[\text{Normal density}\]
Logistic regression as a Latent variable model

\[ y_{ij}^* = \beta_1 + \beta_2 x_{2j} + \beta_3 x_{3ij} + \beta_4 x_{2j} x_{3ij} + (\zeta_j + \epsilon_{ij}) \]

\[ y_{ij} = 1 \iff y_{ij}^* > 0 \]

\[ \xi_{ij} = (\zeta_j + \epsilon_{ij}) \]

Residual variance of a marginal logistic regression

\[ \text{var}(\xi_{ij}) = \tau^2 + \frac{\pi^2}{3} \]

\[ \rho = \frac{\tau^2}{\tau^2 + \pi^2 / 3} \]

Intraclass correlation coefficient
Subject-specific versus population averaged logistic regression

Pop average slope is attenuated with respect to the subject-specific slopes.
Conditional and marginal probabilities for the random intercept logistic regression model

Figure 4.10: Conditional and marginal probabilities for the random-intercept logistic regression model
Logistic regression as a Latent variable model

\[ y_{ij}^* = \beta_1 + \beta_2 x_{2j} + \beta_3 x_{3ij} + \beta_4 x_{2j} x_{3ij} + (\zeta_j + \epsilon_{ij}) \]

\[ y_{ij} = 1 \Leftrightarrow y_{ij}^* > 0 \]

\[ \xi_{ij} = (\zeta_j + \epsilon_{ij}) \]

\[ \text{Residual variance of a marginal logistic regression} \]

\[ \text{var}(\xi_{ij}) = \tau^2 + \frac{\pi^2}{3} \]

\[ \rho = \frac{\tau^2}{\tau^2 + \pi^2 / 3} \]

\[ \text{Intraclass correlation coefficient} \]
Clinical Trial of Contracepting Women

- In this trial, women received an injection of either 100mg or 150mg of depot-medroxyprogesterone acetate (DMPA) on the day of the randomization and three additional injection at 90-day intervals.
- There was a final follow-up visit 90 days after the four injections, this is, one year after the first injection.
- Throughout the study, each woman completed a menstrual diary which was used to determine whether a women experience amenorrhea, the absence of menstrual bleeding for a specified number of days.
Drop-out

- A total of 1151 women completed the menstrual diaries.
- More than 1/3 of the women dropped out before the completion of the trial; 17% dropped out after receiving only one injection of DMPA; 13% dropped out after receiving only 2 injections; and 7% dropped out after receiving 3 injections.
- For women who dropped out before the end of the 90-day injection interval, a determination of whether or not they experienced amenorrhea was made.
Goal of the analysis

• To determine subject-specific changes in the risk of amenorrhea over the course of the study (12 months), and the influence of the dosage of DMPA on changes in a woman’s risk of amenorrhea.
A Mixed effects logistic regression model

- (i) is the women, (j) is the injection interval
- Time = (1, 2, 3, 4) for the 4 consecutive time intervals
- Dose = 1, if randomized to 150mg DMPA and 0 otherwise
- Note that there is not baseline measure of amenorrhea prior receiving the treatment. However, due to randomization, we assume that the baseline risk (at time = 0) is the same in both groups and omit a main effect of dose from the model

\[
\log it P(Y_{ij} = 1 | b_i) = \beta_1 + \beta_2 time_{ij} + \beta_3 time_{ij}^2 + \\
+ \beta_4 (dose_i \times time_{ij}) + \beta_5 (dose_i \times time_{ij}^2) + b_{1i} \\
b_{1i} \sim N(0, \tau^2)
\]
A Mixed effects logistic regression model

• By including a random intercept we assume that there is a random heterogeneity in women’s propensity or underlying risk of amenorrhea that persists throughout the entire duration of the study
Table: parameter estimates and standard errors from a mixed effects logistic regression model, with random intercept for the amenorrhea data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>SE</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.8057</td>
<td>0.3050</td>
<td>-12.48</td>
</tr>
<tr>
<td>$t_{ij}$</td>
<td>1.1332</td>
<td>0.2682</td>
<td>4.22</td>
</tr>
<tr>
<td>$t_{ij}^2$</td>
<td>-0.0419</td>
<td>0.0548</td>
<td>-0.76</td>
</tr>
<tr>
<td>$dose_i \times t_{ij}$</td>
<td>0.5644</td>
<td>0.1922</td>
<td>2.94</td>
</tr>
<tr>
<td>$dose_i \times t_{ij}^2$</td>
<td>-0.1095</td>
<td>0.0496</td>
<td>-2.21</td>
</tr>
<tr>
<td>$g_{11}$</td>
<td>5.0646</td>
<td>0.5840</td>
<td>8.67</td>
</tr>
</tbody>
</table>

Variance of the random intercept
Parameters interpretation

• There is evidence that the subject-specific log-odds of amenorrhea increase over the 12 months of the trial, and that subject-specific changes in the risk of amenorrhea depend on the dose of DMPA.

• For example, for a women assigned to the low dose of DMPA, the log odds of amenorrhea increase approximately linearly, with an increase in the log odds of 1.09 (1.1332-0.0419) at 3 months, 2.10 (2x1.13320-4x0.0419) at 6 months, 3.02 (3x1.1332-9x0.0419) at 9 months, and 3.86 (4x1.1332 - 16x0.0410) at 12 months.
Parameters Interpretation

- These increases in risk corresponds to odds of 3 (or \( \exp(1.09) \)), 8.2 (or \( \exp(2.10) \)), 20.5 (or \( \exp(3.02) \)), and 47.5 (or \( \exp(3.86) \)) at 3, 6, 9, and 12 months.

- On the other end, for the women assigned to the high dose of DMPA, the log odds of amenorrhea increases quadratically, with an increase in 1.55 ((1.1332-0.0419) + (0.5644-0.1095) ) at 3 months, 2.79 at 6 months, 3.73 at 9 months, and 4.37 at 12 months.

- That is, the early trend shows a decline toward the end. These increases in risk correspond to odds equal to 4.7 (or \( \exp(1.55) \)), 16.3 (or \( \exp(2.79) \)), 41.7 (or \( \exp(3.73) \)), and 79 (or \( \exp(4.37) \)) at 3, 6, 9, and 12 months.
Interpretation of the interaction terms

• Because treatment (high versus low doses of DMPA) is a subject-specific variable, this makes the interpretation of the fixed effects for the (dose x time) interactions more difficult.

• The interaction effects must be given an interpretation in terms of a contrast of the increases in log odds of amenorrhea for two different women, who happen to have the same underlying risk of experiencing amenorrhea prior randomization, but who differ in terms of dose (i.e. one assigned to low dose and the other to high dose).
Interpretation of the interaction term

- From the estimates of the fixed effects in the Table, the ratio of increased odds of amenorrhea (odds ratio) at 12 months for a woman assigned to the high dose, versus another woman - who happen to have the same risk of amenorrhea prior the randomization (e.g. the same value of the random effect)- but who was assigned to the low dose, is 1.66 (or exp(4.37-3.86)), with 95% CI 1.03 to 2.66
Variance of the random intercept

- The estimated variance of the random intercept is 5.06. This implies that there is substantial variability in the propensity to experience amenorrhea, since approximately 95% of the women have a baseline risk of amenorrhea that varies within the range

\[
\frac{\exp(-3.8 - 1.96\sqrt{5.06})}{1 + \exp(-3.8 - 1.96\sqrt{5.06})} = 0.0003
\]

\[
\frac{\exp(-3.8 + 1.96\sqrt{5.06})}{1 + \exp(-3.8 + 1.96\sqrt{5.06})} = 0.65
\]
Variance of the random intercept: latent variable formulation

\[ y_{ij}^* = \beta_1 + \beta_2 \text{time}_{ij} + \beta_3 \text{time}_{ij}^2 + \]
\[ + \beta_4 (\text{dose}_i \times \text{time}_{ij}) + \beta_5 (\text{dose}_i \times \text{time}_{ij}^2) + b_i + \varepsilon_{ij} \]

\[ E[\varepsilon_{ij}] = 0 \]
\[ \text{Var}[\varepsilon_{ij}] = \pi^2 / 3 \]

\[ \rho = \text{corr}(y_{ij}^*, y_{ik}^*) = \frac{\tau^2}{\tau^2 + \pi^2 / 3} \]
\[ \rho = \frac{5.06}{5.06 + 3.29} = 0.61 \]

- Marginal intra-class correlation coefficient between the “latent” responses
A cautionary note

- There is usually not much information available on the random effects, beyond a random intercept, when the number of repeated measurements is relatively small.
- Thus convergence problems during estimation are often encountered when random effects beyond a random intercept are included in the logistic regression for longitudinal data.
Marginal logistic regression

\[
\log \text{it} P(y_{ij} = 1) = \beta_1 + \beta_2 \text{time}_{ij} + \beta_3 \text{time}_{ij}^2 + \\
\beta_4 (dose_i \times \text{time}_{ij}) + \beta_5 (dose_i \times \text{time}_{ij}^2)
\]

\[
\log \text{OR}(y_{ij}, y_{ik}) = \alpha_{jk}
\]

\[
\text{OR}(y_{ij}, y_{ik}) = \frac{P(y_j = 1, y_k = 1)P(y_j = 0, y_k = 0)}{P(y_j = 1, y_k = 0)P(y_j = 0, y_k = 1)}
\]
Table 12.4 Parameter estimates and standard errors, obtained using GEE approach, from marginal logistic regression model for the amenorrhea data.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>SE</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.2461</td>
<td>0.1765</td>
<td>-12.72</td>
</tr>
<tr>
<td>time$_{ij}$</td>
<td>0.7030</td>
<td>0.1581</td>
<td>4.45</td>
</tr>
<tr>
<td>time$_{ij}^2$</td>
<td>-0.0323</td>
<td>0.0318</td>
<td>-1.02</td>
</tr>
<tr>
<td>dose$<em>i \times$ time$</em>{ij}$</td>
<td>0.3380</td>
<td>0.1097</td>
<td>3.08</td>
</tr>
<tr>
<td>dose$<em>i \times$ time$</em>{ij}^2$</td>
<td>-0.0683</td>
<td>0.0284</td>
<td>-2.40</td>
</tr>
<tr>
<td>$\alpha_{12}$</td>
<td>1.8475</td>
<td>0.1810</td>
<td>10.21</td>
</tr>
<tr>
<td>$\alpha_{13}$</td>
<td>1.4851</td>
<td>0.1985</td>
<td>7.48</td>
</tr>
<tr>
<td>$\alpha_{14}$</td>
<td>1.7605</td>
<td>0.2482</td>
<td>7.09</td>
</tr>
<tr>
<td>$\alpha_{23}$</td>
<td>2.1610</td>
<td>0.1761</td>
<td>12.27</td>
</tr>
<tr>
<td>$\alpha_{24}$</td>
<td>2.0665</td>
<td>0.2034</td>
<td>10.16</td>
</tr>
<tr>
<td>$\alpha_{34}$</td>
<td>2.2783</td>
<td>0.1827</td>
<td>12.47</td>
</tr>
</tbody>
</table>
Marginal versus random effects logistic regression

• The estimated regression coefficients from a Marginal model are smaller (in absolute value) than the estimated regression coefficients from a random effects model.

• The ratio of population odds of amenorrhea at 12 months (odds ratio) for women on the high versus low dose is 1.30 (95% CI 0.98, 1.71).

• These differences in odds ratio are due to different interpretation of the parameters between these two classes of models.
Marginal versus random effects logistic regression

- The estimates of the fixed effect dose in the RE model describe the effect of a high versus low dose conditionally to a specific women’s risk of amenorrhea.
- The corresponding effect in the M model describe the effects of dose on the prevalence of amenorrhea in the population of women assigned to high versus low doses.