Lecture 7
Logistic Regression with Random Intercept

Logistic Regression

Odds: expected number of successes for each failure

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

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

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

Odds ratio

Log-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 Labor's force participation

ble 4.1: Maximum likelihood estimates for women's labor force participation

	Est	(SE)	OR=exp(β)	(95% CI)
β_1 [_cons]	1.34	(0.38)		
β_2 [husbinc]	−0.04	(0.02)	0.96	(0.92, 1.0)
β_3 [chilpres]	−1.58	(0.29)	0.21	(0.12, 0.37)

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]

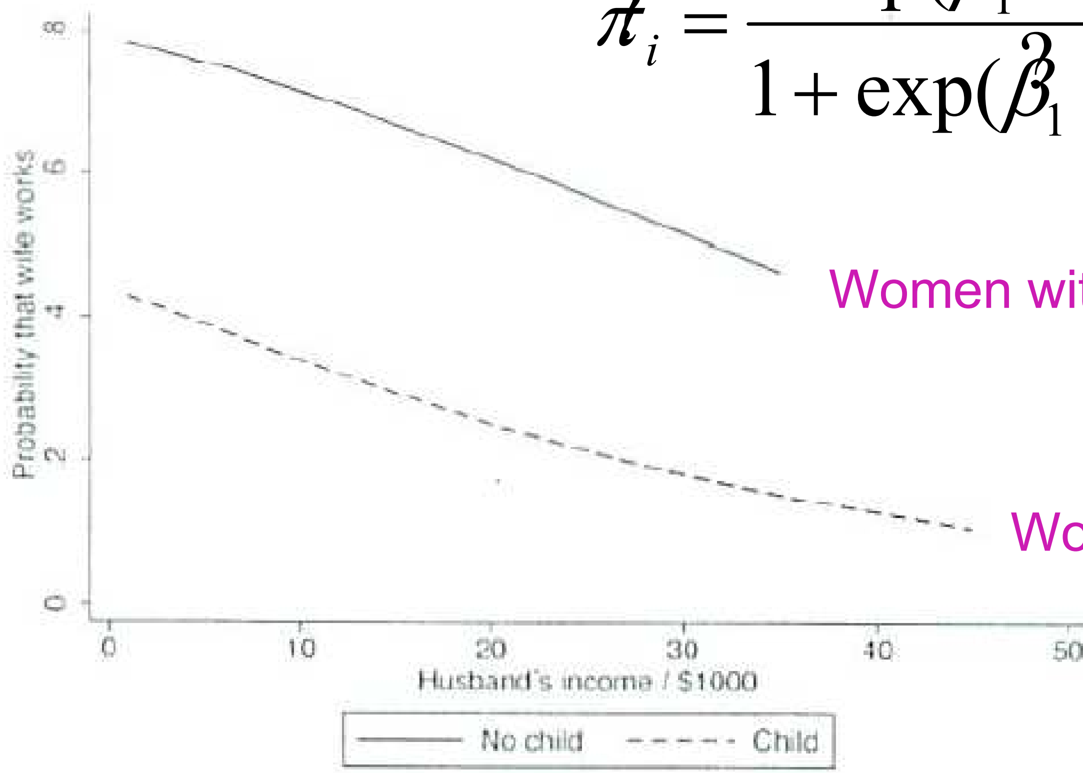
Standard errors

- 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

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i})}{1 + \exp(\hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i})}$$



Women without a child

Women with a child

Figure 1.1: Predicted probabilities from logistic regression model

Figure 4.2: predicted probabilities from logistic regression model, extrapolating outside the range of the data

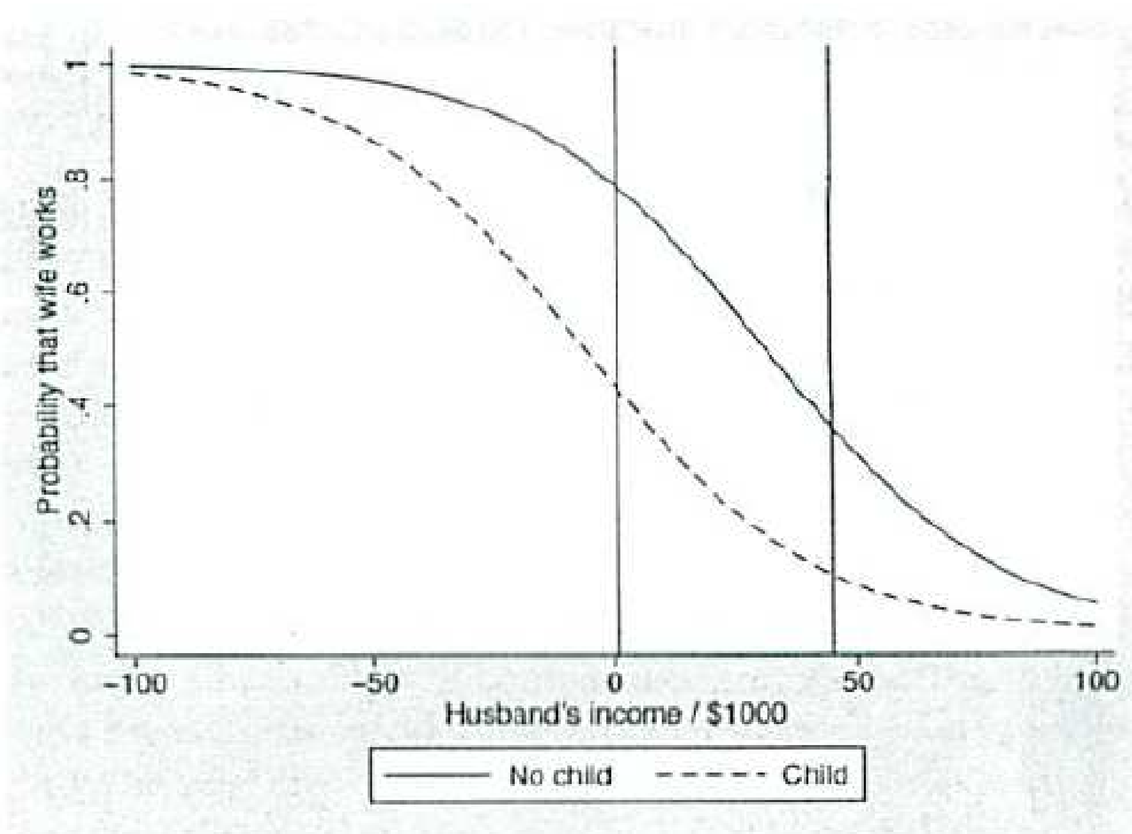


Figure 4.2: Predicted probabilities from logistic regression model, extrapolating outside the range of the data

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:

Latent continuous response

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

$$y_i^* > 0 \Rightarrow y_i = 1$$

$$y_i^* \leq 0 \Rightarrow y_i = 0$$

$$E(\varepsilon_i \mid x_i) = 0$$

Latent Response formulation of a logistic regression model

- In logistic regression the error ε_i is assumed to have a logistic cumulative density function given x,

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

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

$$Var[\varepsilon_i \mid x_i] = \frac{\pi^2}{3} \approx 3.29$$

Probit Regression

- When a latent-response formulation is used, it seems natural to assume that ε_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

$$\Pr(y_i = 1 \mid x_i) = \Pr(y_i^* > 1 \mid x_i) =$$

$$\Pr(\beta_1 + \beta_2 x_i + \varepsilon_i > 0) = \Pr(\varepsilon_i > -(\beta_1 + \beta_2 x_i)) =$$

$$\Pr(-\varepsilon_i \leq \beta_1 + \beta_2 x_i) = \Pr(\varepsilon_i \leq \beta_1 + \beta_2 x_i) = \Phi(\beta_1 + \beta_2 x_i)$$

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

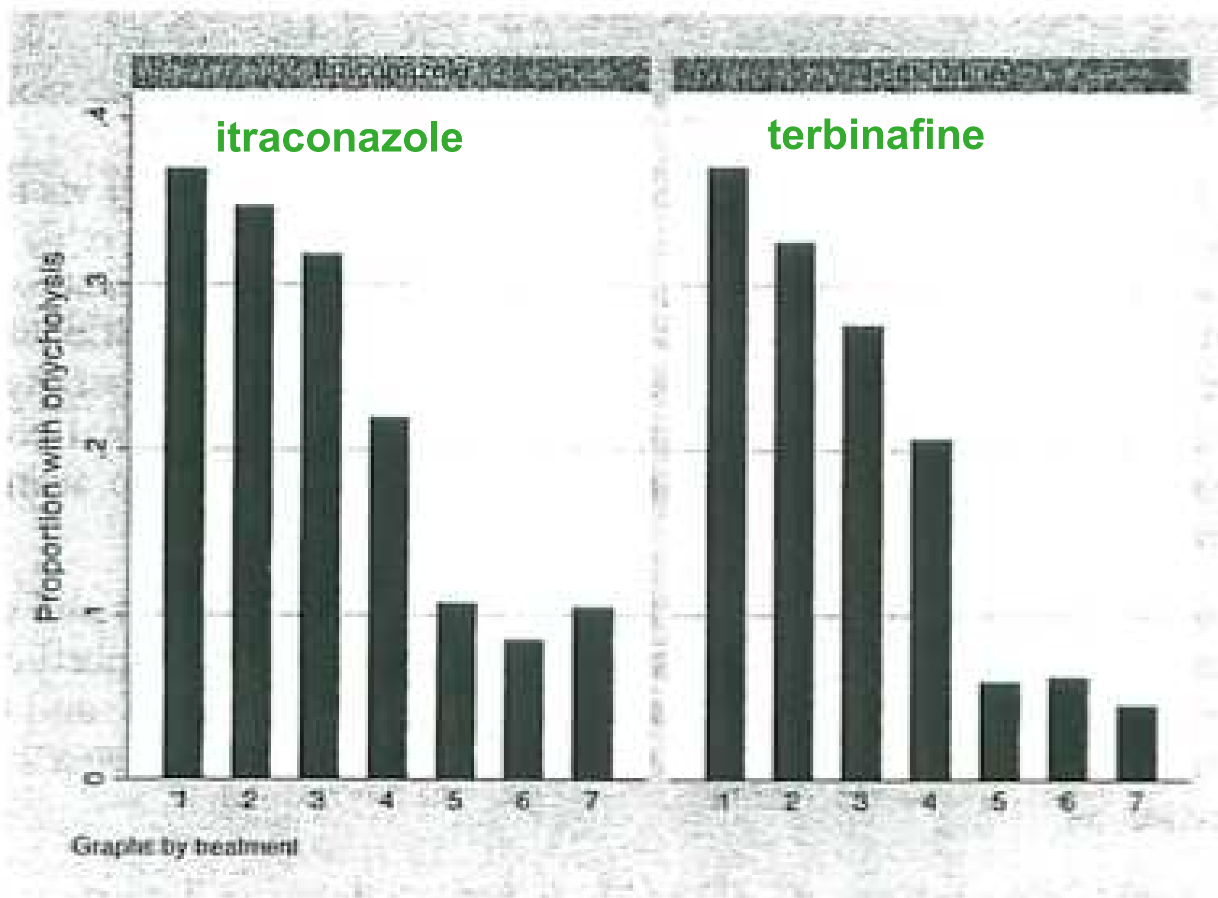


Figure 4.6: Bar plot of proportion of patients with toenail infection by visit and treatment group

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

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

treatment month

treat effect

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)

Fig 4.8: Proportions and fitted probabilities using ordinary logistic regression

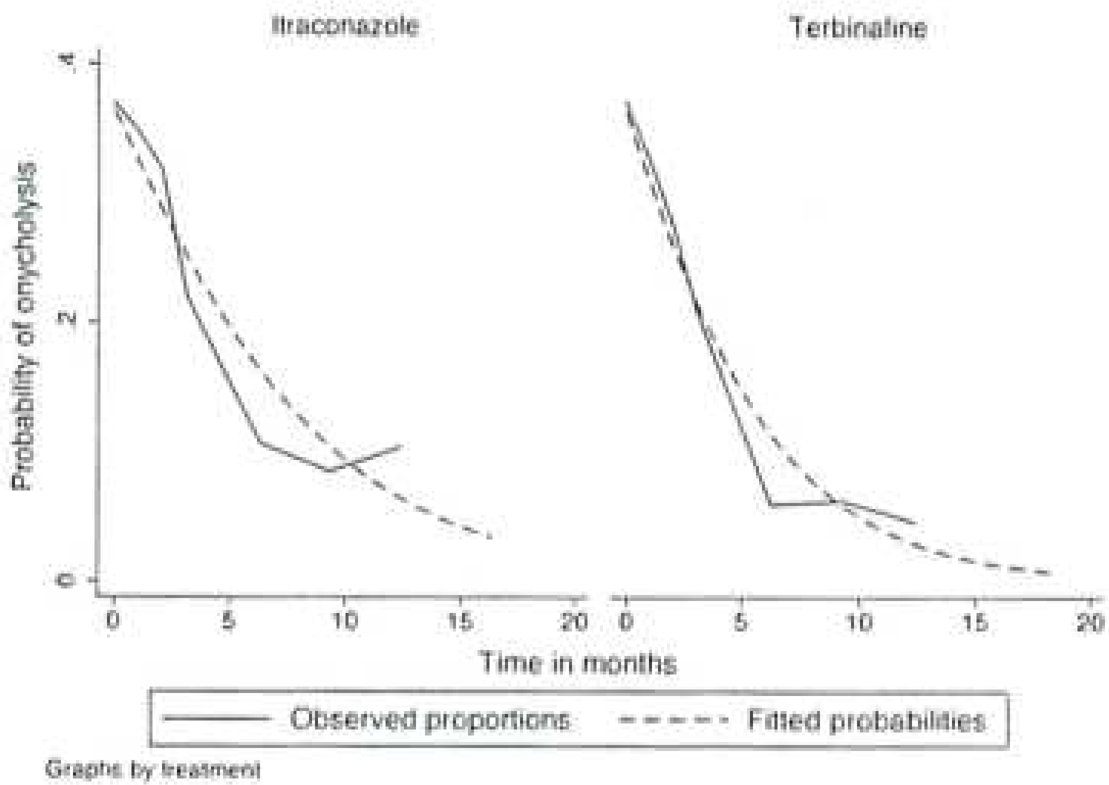


Figure 4.8: Proportions and fitted probabilities using ordinary logistic regression

This model makes the unrealistic assumption that the responses for a given patient are conditionally independent given the covariates

Logistic regression with random intercept (xtlogit,xtmelogit,gllamm)

$$y_{ij} \mid \pi_{ij} \sim Binomial(1, \pi_{ij})$$

$$\pi_{ij} = P(y_{ij} = 1 \mid x_{2j}, x_{3ij}, \varsigma_j)$$

$$\log it\{\pi_{ij}\} = \beta_1 + \beta_2 x_{2j} + \beta_3 x_{3ij} + \beta_4 x_{2j} x_{3ij} + \varsigma_j$$

$$\varsigma_j \sim N(0, \psi)$$

The random intercept represents the combined effect of all omitted **subject-specific** covariates that causes some subjects to be more prone to the disease than others

Table 4.2: Estimates for toenail data

Table 4.2: Estimates for toenail data

	Marginal effects				Conditional effects			
	Ordinary logistic		GEE logistic		Random int. logistic		Conditional logistic	
ieter	OR	(95% CI)	OR	(95% CI)†	OR	(95% CI)	OR	(95% CI)
part								
) [treatment]	1.00	(0.74, 1.36)	1.01	(0.61, 1.68)	0.85	(0.27, 2.65)		
) [month]	0.84	(0.81, 0.88)	0.84	(0.79, 0.89)	0.68	(0.62, 0.74)	0.68	(0.62, 0.75)
) [trt_month]	0.93	(0.87, 1.01)	0.93	(0.83, 1.03)	0.87	(0.76, 1.00)	0.91	(0.78, 1.05)
m part								
					16.08	(3.06)		
					0.83			
elihood	−908.01				−625.39		−188.94	

Variance of the random effects

ICC

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 itraconozole group and 13% lower (equal to 0.68x0.87=0.59) in the terbinafine group (for a patient with random intercept equal to zero)

Parameters Interpretation

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

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

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

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

Results: The odds decrease by 32% (100*(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

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

marginal prob

B:Random Intercept Logistic regression

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

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

$$P^*\left(y_{ij}=1 \mid x_{2j},x_{3ij}\right)=$$

Normal density

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

$$\neq P\left(y_{ij}=1 \mid x_{2j},x_{3ij},\varsigma_j\right)$$

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

Logistic regression as a Latent variable model

$$y_{ij}^{*} = \beta_1 + \beta_2 x_{2j} + \beta_3 x_{3ij} + \beta_4 x_{2j} x_{3ij} + (\varsigma_j + \varepsilon_{ij})$$

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

$$\xi_{ij} = (\varsigma_j + \varepsilon_{ij})$$

Residual variance of a marginal logistic regression

$$\operatorname{var}(\xi_{ij}) = \tau^2 + \boxed{\frac{\pi^2}{3}}$$

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

Intraclass correlation coefficient

Subject-specific versus population averaged logistic regression

Pop average slope
is attenuated
with respect to the
subject-specific slopes

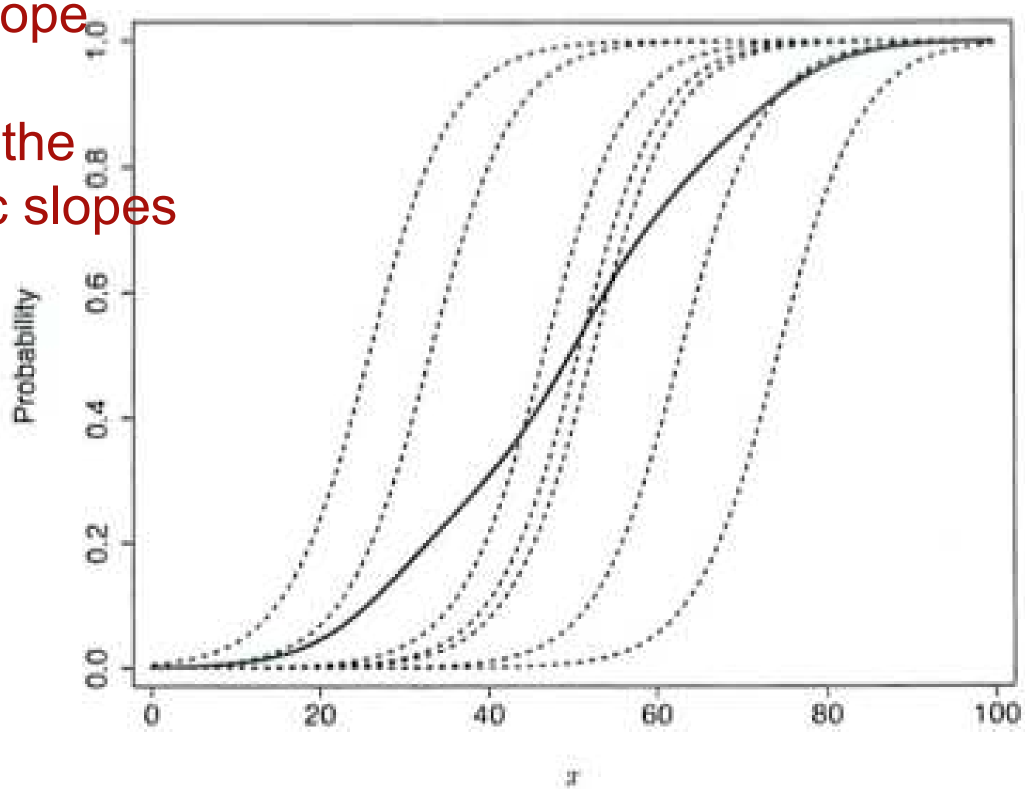


Figure 4.11: Subject-specific versus population-averaged logistic regression

Conditional and marginal probabilities for the random intercept logistic regression model

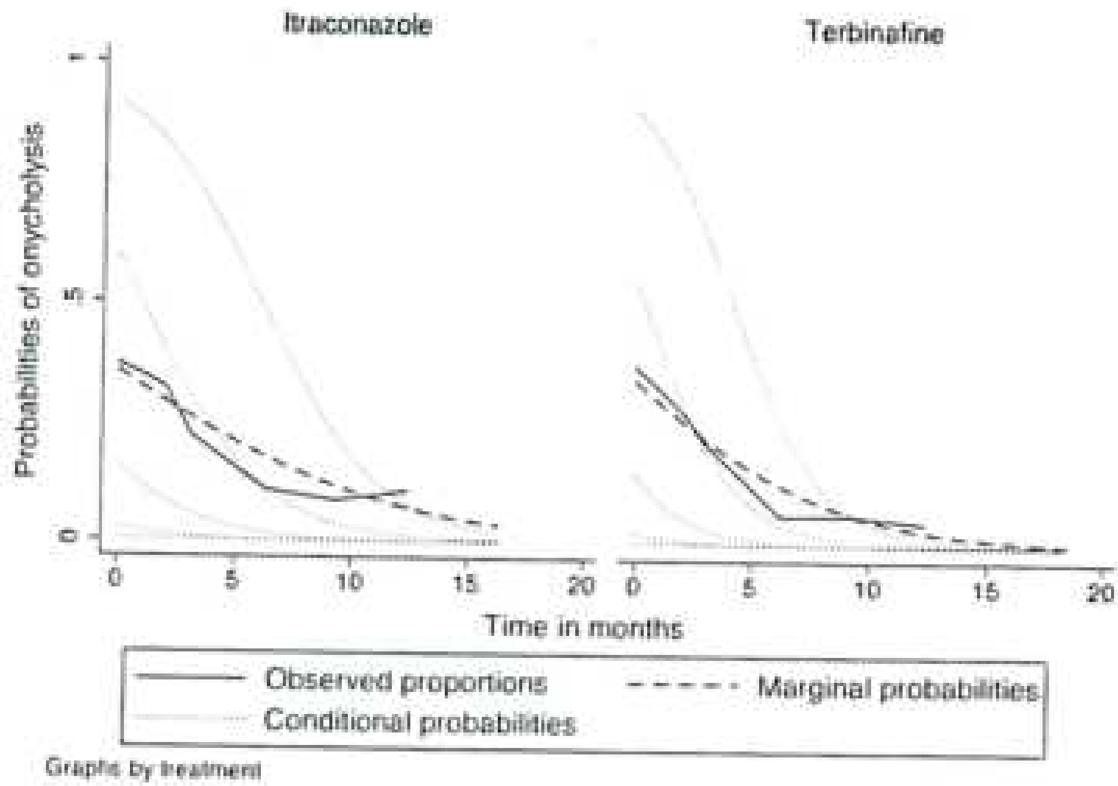


Figure 4.10: Conditional and marginal probabilities for the random-intercept logistic regression model

Logistic regression as a Latent variable model

$$y_{ij}^{*} = \beta_1 + \beta_2 x_{2j} + \beta_3 x_{3ij} + \beta_4 x_{2j} x_{3ij} + (\varsigma_j + \varepsilon_{ij})$$

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

$$\xi_{ij} = (\varsigma_j + \varepsilon_{ij})$$

$$\mathrm{var}(\xi_{ij}) = \tau^2 + \boxed{\frac{\pi^2}{3}}$$

Residual variance of a marginal logistic regression

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

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 women 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*

$$\operatorname{logit} P(Y_{ij}=1 \mid b_i) = \beta_1 + \beta_2 time_{ij} + \beta_3 time_{ij}^2 +$$

$$\boxed{+\beta_4(dose_i \times time_{ij}) + \beta_5(dose_i \times time_{ij}^2)} + b_{1i}$$

$$b_{1i} \sim N(0, \tau^2)$$

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 12.2 Parameter estimates and standard errors from a mixed effects logistic regression model, with randomly varying intercepts, for the amenorrhea data.

Variable	Estimate	SE	Z
Intercept	−3.8057	0.3050	−12.48
$time_{ij}$	1.1332	0.2682	4.22
$time_{ij}^2$	−0.0419	0.0548	−0.76
$dose_i \times time_{ij}$	0.5644	0.1922	2.94
$dose_i \times time_{ij}^2$	−0.1095	0.0496	−2.21
g_{11}	5.0646	0.5840	8.67

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 women assigned to the high dose, versus another women - 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

$$y_{ij}^{*} = \beta_1 + \beta_2 time_{ij} + \beta_3 time_{ij}^2 +$$

$$\boxed{+\beta_4(dose_i \times time_{ij}) + \beta_5(dose_i \times time_{ij}^2)} + b_{1i} + \varepsilon_{ij}$$

$$\boxed{\begin{array}{l} E[\varepsilon_{ij}] = 0 \\ Var[\varepsilon_{ij}] = \pi^2/3 \end{array}}$$

$$\rho = corr(y_{ij}^{*}, y_{ik}^{*}) = \frac{\tau^2}{\tau^2 + \pi^2/3}$$

$$\hat{\rho} = \frac{5.06}{5.06 + 3.29} = 0.61$$

- 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

$$\log itP(y_{ij}=1)=\beta_1+\beta_2 time_{ij}+\beta_3 time_{ij}^2+$$

$$\boxed{+\beta_4(dose_i \times time_{ij})+\beta_5(dose_i \times time_{ij}^2)}$$

$$\log OR(y_{ij},y_{ik})=\alpha_{jk}$$

$$OR(y_{ij},y_{ik})=\frac{P(y_j=1,y_k=1)P(y_j=0,y_k=0)}{P(y_j=1,y_k=0)P(y_j=0,y_k=1)}$$

Table 12.4 Parameter estimates and standard errors, obtained using GEE approach, from marginal logistic regression model for the amenorrhea data.

Variable	Estimate	SE	Z
Intercept	−2.2461	0.1765	−12.72
$\texttt{time}_{ij}$	0.7030	0.1581	4.45
$\texttt{time}^2_{ij}$	−0.0323	0.0318	−1.02
$\texttt{dose}_i \times \texttt{time}_{ij}$	0.3380	0.1097	3.08
$\texttt{dose}_i \times \texttt{time}^2_{ij}$	−0.0683	0.0284	−2.40
α_{12}	1.8475	0.1810	10.21
α_{13}	1.4851	0.1985	7.48
α_{14}	1.7605	0.2482	7.09
α_{23}	2.1610	0.1761	12.27
α_{24}	2.0665	0.2034	10.16
α_{34}	2.2783	0.1827	12.47

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