Семинар 5.

Модели бинарного и упорядоченного выбора.

1. Рассмотрим модель бинарного выбора $\mathbb{P}(y_i = 1) = F(\alpha+\beta d_i)$, где d — фиктивная переменная (принимающая значения 0 и 1). Ниже представлены результаты 100 наблюдений:

	$y = 0$	$y = 1$
$d = 0$	20	32
$d = 1$	36	12

   (a) Оцените параметры α, β с помощью ММП, используя logit–модель.

   (b) Проверьте гипотезу $H_0 : \beta = 0$ с помощью LR–теста.

   Решение: 1. Оценивание параметров.

   В logit-модели $F(z) = \Lambda(z) = e^z/(1 + e^z)$. Введём вспомогательный параметр $\gamma = \alpha + \beta$. Тогда вероятности для двух групп:

$$P(y = 1 \mid d = 0) = \Lambda(\alpha), \qquad P(y = 1 \mid d = 1) = \Lambda(\gamma).$$

   Обозначим частоты из таблицы:

$$n_{00} = 20,\ n_{01} = 32,\ n_{10} = 36,\ n_{11} = 12.$$

   Функция правдоподобия (предполагаем независимость наблюдений):

$$L(\alpha, \gamma) = \big[1 - \Lambda(\alpha)\big]^{n_{00}} \big[\Lambda(\alpha)\big]^{n_{01}} \big[1 - \Lambda(\gamma)\big]^{n_{10}} \big[\Lambda(\gamma)\big]^{n_{11}}.$$

   Логарифмическая функция правдоподобия:

$$\ln L(\alpha, \gamma) = n_{00} \ln\big(1 - \Lambda(\alpha)\big) + n_{01} \ln \Lambda(\alpha) + n_{10} \ln\big(1 - \Lambda(\gamma)\big) + n_{11} \ln \Lambda(\gamma).$$

   Поскольку параметры α и γ входят в разные слагаемые независимо, максимизацию можно проводить отдельно. Для логит-модели удобно перейти к вероятностям $p_0 = \Lambda(\alpha)$ и $p_1 = \Lambda(\gamma)$. Тогда

$$\ln L = n_{00} \ln(1 - p_0) + n_{01} \ln p_0 + n_{10} \ln(1 - p_1) + n_{11} \ln p_1.$$

   Максимизация по p_0 и p_1 даёт хорошо известные оценки для биномиальных выборок:

$$\hat{p}_0 = \frac{n_{01}}{n_{00} + n_{01}} = \frac{32}{52} = \frac{8}{13}, \qquad \hat{p}_1 = \frac{n_{11}}{n_{10} + n_{11}} = \frac{12}{48} = \frac{1}{4}.$$

В самом деле, приравнивая производную $\frac{\partial \ln L}{\partial p_0} = 0$:

$$-\frac{n_{00}}{1-p_0} + \frac{n_{01}}{p_0} = 0 \implies \frac{n_{01}}{p_0} = \frac{n_{00}}{1-p_0} \implies p_0 = \frac{n_{01}}{n_{00}+n_{01}}.$$

Аналогично для p_1.

Теперь перейдём обратно к α и γ, используя обратную логит-функцию:

$$\alpha = \text{logit}(p_0) = \ln \frac{p_0}{1-p_0}, \qquad \gamma = \text{logit}(p_1) = \ln \frac{p_1}{1-p_1}.$$

Вычисляем:

$$\hat{\alpha} = \ln \frac{8/13}{5/13} = \ln \frac{8}{5} \approx 0.4700, \qquad \hat{\gamma} = \ln \frac{1/4}{3/4} = \ln \frac{1}{3} \approx -1.0986.$$

Поскольку $\gamma = \alpha + \beta$, получаем оценку β:

$$\hat{\beta} = \hat{\gamma} - \hat{\alpha} = \ln \frac{1}{3} - \ln \frac{8}{5} = \ln \frac{5}{24} \approx -1.5686.$$

Таким образом, искомые оценки:

$$\hat{\alpha} = \ln \frac{8}{5}, \quad \hat{\beta} = \ln \frac{5}{24}.$$

2. Проверка гипотезы $\beta = 0$ с помощью LR-теста.

Гипотеза $H_0 : \beta = 0$ означает равенство вероятностей в обеих группах: $p_0 = p_1 = p$. Оценим ограниченную модель. Общее число наблюдений с $y = 1$ равно $n_{01} + n_{11} = 32 + 12 = 44$, общий объём выборки 100, поэтому

$$\tilde{p} = \frac{44}{100} = 0.44, \qquad \tilde{\alpha} = \text{logit}(0.44) = \ln \frac{0.44}{0.56} = \ln \frac{11}{14} \approx -0.2412.$$

При этом $\tilde{\beta} = 0$ по условию.

Вычислим значения логарифмических функций правдоподобия для ограниченной и неограниченной моделей. Удобно использовать вероятности.

Для неограниченной модели:

$$\ln \hat{p}_0 = \ln \frac{8}{13} \approx -0.4855, \quad \ln(1-\hat{p}_0) = \ln \frac{5}{13} \approx -0.9555,$$
$$\ln \hat{p}_1 = \ln \frac{1}{4} \approx -1.3863, \quad \ln(1-\hat{p}_1) = \ln \frac{3}{4} \approx -0.2877.$$

Тогда

$$\begin{aligned}\ln L_u &= 20\ln(1-\hat{p}_0) + 32\ln\hat{p}_0 + 36\ln(1-\hat{p}_1) + 12\ln\hat{p}_1 \\ &= 20(-0.9555) + 32(-0.4855) + 36(-0.2877) + 12(-1.3863) \\ &= -19.11 - 15.536 - 10.357 - 16.636 = -61.639.\end{aligned}$$

Для ограниченной модели ($\tilde{p} = 0.44$):

$$\ln\tilde{p} = \ln 0.44 \approx -0.82098, \qquad \ln(1-\tilde{p}) = \ln 0.56 \approx -0.57982.$$

Общее число наблюдений с $y = 0$ равно $20 + 36 = 56$, с $y = 1$ равно 44. Поэтому

$$\ln L_r = 56\ln(1-\tilde{p}) + 44\ln\tilde{p} = 56(-0.57982) + 44(-0.82098) = -32.470 - 36.123 = -68.593.$$

Статистика LR-теста:

$$LR = -2(\ln L_r - \ln L_u) = -2\big(-68.593 + 61.639\big) = -2(-6.954) = 13.908.$$

При нулевой гипотезе статистика асимптотически распределена как χ^2_1. Критическое значение при уровне значимости 5% равно 3.84. Поскольку $13.91 > 3.84$, гипотеза $H_0 : \beta = 0$ отвергается. Следовательно, влияние фиктивной переменной d на вероятность $y = 1$ статистически значимо.

2. Ниже представлены результаты 250 наблюдений:

y	0	1	2	3	4
n	50	40	45	80	35

Используя данные, найдите оценки максимального правдоподобия неизвестных параметров упорядоченной *probit*–модели. [Подсказка: Рассматривайте вероятности как неизвестные параметры.]

Решение:

Рассмотрим упорядоченную probit-модель для случая, когда отсутствуют объясняющие переменные. Пусть латентная переменная y^* определяется как

$$y^* = \mu + \varepsilon,$$

где $\varepsilon \sim N(0,1)$, а μ — неизвестный параметр сдвига (константа). Наблюдаемая переменная y принимает значения $0, 1, 2, 3, 4$ в соответствии с порогами $c_1 <$

$c_2 < c_3 < c_4$:

$$y = \begin{cases} 0, & y^* \le c_1, \\ 1, & c_1 < y^* \le c_2, \\ 2, & c_2 < y^* \le c_3, \\ 3, & c_3 < y^* \le c_4, \\ 4, & y^* > c_4. \end{cases}$$

Для идентификации модели необходимо наложить нормировочное условие. Обычно фиксируют один из порогов или константу. В данном решении положим $c_1 = 0$. Тогда остальные параметры μ, c_2, c_3, c_4 подлежат оцениванию.

Вероятности каждого исхода выражаются через функцию стандартного нормального распределения Φ:

$$\begin{aligned} P(y=0) &= \Phi(0-\mu) = \Phi(-\mu), \\ P(y=1) &= \Phi(c_2-\mu) - \Phi(-\mu), \\ P(y=2) &= \Phi(c_3-\mu) - \Phi(c_2-\mu), \\ P(y=3) &= \Phi(c_4-\mu) - \Phi(c_3-\mu), \\ P(y=4) &= 1 - \Phi(c_4-\mu). \end{aligned}$$

По данным известны частоты исходов, которые приравниваем к соответствующим вероятностям (согласно методу максимального правдоподобия оценки вероятностей совпадают с выборочными долями). Выборочные доли:

$$\hat{p}_0 = \frac{50}{250} = 0.2, \quad \hat{p}_1 = \frac{40}{250} = 0.16, \quad \hat{p}_2 = \frac{45}{250} = 0.18, \quad \hat{p}_3 = \frac{80}{250} = 0.32, \quad \hat{p}_4 = \frac{35}{250} = 0.14.$$

Накопленные вероятности:

$$\hat{F}_0 = 0.2, \quad \hat{F}_1 = 0.36, \quad \hat{F}_2 = 0.54, \quad \hat{F}_3 = 0.86, \quad \hat{F}_4 = 1.$$

Из соотношения $P(y \le 0) = \Phi(-\mu) = \hat{F}_0 = 0.2$ находим μ:

$$-\mu = \Phi^{-1}(0.2) \approx -0.8416 \quad \Longrightarrow \quad \hat{\mu} = 0.8416.$$

Далее последовательно определяем пороги:

$$P(y \le 1) = \Phi(c_2 - \mu) = \hat{F}_1 = 0.36 \Rightarrow c_2 - \mu = \Phi^{-1}(0.36) \approx -0.3585$$
$$\Rightarrow \hat{c}_2 = \mu - 0.3585 = 0.8416 - 0.3585 = 0.4831,$$
$$P(y \le 2) = \Phi(c_3 - \mu) = \hat{F}_2 = 0.54 \Rightarrow c_3 - \mu = \Phi^{-1}(0.54) \approx 0.1004$$
$$\Rightarrow \hat{c}_3 = \mu + 0.1004 = 0.8416 + 0.1004 = 0.9420,$$
$$P(y \le 3) = \Phi(c_4 - \mu) = \hat{F}_3 = 0.86 \Rightarrow c_4 - \mu = \Phi^{-1}(0.86) \approx 1.0803$$
$$\Rightarrow \hat{c}_4 = \mu + 1.0803 = 0.8416 + 1.0803 = 1.9219.$$

Таким образом, оценки максимального правдоподобия параметров упорядоченной probit-модели при нормировке $c_1 = 0$ равны:

$$\hat{\mu} = 0.8416, \quad \hat{c}_2 = 0.4831, \quad \hat{c}_3 = 0.9420, \quad \hat{c}_4 = 1.9219.$$

3. Пусть $y_i^* = x_i^T\beta + \varepsilon_i$, где $\varepsilon_i \sim i.i.d.(0,1)$. Известно, что

$$y_t = \begin{cases} 0, & y^* \le c_1, \\ 1, & c_1 < y^* \le c_2, \\ 2, & y^* > c_2. \end{cases}$$

Для модели упорядоченного выбора рассчитайте предельные эффекты:

(a) $\frac{\partial P(y_i=0|x_i)}{\partial x_{ij}}$,

(b) $\frac{\partial P(y_i=1|x_i)}{\partial x_{ij}}$,

(c) $\frac{\partial P(y_i=2|x_i)}{\partial x_{ij}}$.

Решение:

Обозначим через F и f функцию распределения и плотность ошибки ε_i. Вероятности исходов:

$$P(y_i = 0 \mid x_i) = F(c_1 - x_i'\beta),$$
$$P(y_i = 1 \mid x_i) = F(c_2 - x_i'\beta) - F(c_1 - x_i'\beta),$$
$$P(y_i = 2 \mid x_i) = 1 - F(c_2 - x_i'\beta).$$

Дифференцируя по x_{ij} (учитывая, что $\frac{\partial(x_i'\beta)}{\partial x_{ij}} = \beta_j$), получаем искомые предель-

ные эффекты:

$$\frac{\partial P(y_i = 0 \mid x_i)}{\partial x_{ij}} = -\beta_j \, f(c_1 - x_i'\beta),$$
$$\frac{\partial P(y_i = 1 \mid x_i)}{\partial x_{ij}} = \beta_j \big[f(c_1 - x_i'\beta) - f(c_2 - x_i'\beta) \big],$$
$$\frac{\partial P(y_i = 2 \mid x_i)}{\partial x_{ij}} = \beta_j \, f(c_2 - x_i'\beta).$$

Если ошибка имеет стандартное нормальное распределение (что соответствует условию $\varepsilon_i \sim i.i.d.(0,1)$), то $f = \varphi$ — плотность стандартного нормального распределения, $F = \Phi$. В этом случае полученные выражения представляют предельные эффекты для упорядоченной probit-модели.