Categorical Data Analysis
Lecture 11 & 12

Graduate School of Advanced Science and Engineering
Rei Monden

2025.1.16

Multicategory logit models

Multicategory logit models

Today we generalize logistic models to response variables with 3 or more categories.

We will learn about two such models:

- Nominal logistic regression model.
  When the response variable is *nominal*.
- Ordinal logistic regression model.
  When the response variable is, well, *ordinal*.

Multicategory logit models

Notation:

- c: number of categories of the response variable Y.
- $\pi_1, \ldots, \pi_c$: response probabilities, satisfying $\sum_j \pi_j = 1$.

Nominal logistic regression model

Nominal logistic regression model

AKA baseline-category logit model.

This model is based on a set of $(c-1)$ logits of category j ($j = 1, \dots, c-1$) against a baseline category, say, the cth.

For example, suppose we have only one predictor x. Then the nominal regression model consists on the following set of $(c-1)$ equations:

$$\log\left(\frac{\pi_j}{\pi_c}\right) = \alpha_j + \beta_j x,$$

for $j = 1, \dots, c-1$.

Important: Each equation has its own set of parameters $\{\alpha_j, \beta_j\}$.

Nominal logistic regression model

$$\boxed{\log\left(\frac{\pi_j}{\pi_c}\right) = \alpha_j + \beta_j x}$$

Although not immediately clear, we can compare *any* two response categories.

For example, if $c = 3$ then:

- $\log\left(\frac{\pi_1}{\pi_3}\right) = \alpha_1 + \beta_1 x$
- $\log\left(\frac{\pi_2}{\pi_3}\right) = \alpha_2 + \beta_2 x$
-

$$\begin{aligned}\log\left(\frac{\pi_1}{\pi_2}\right) &= \log\left(\frac{\pi_1}{\pi_3}\right) - \log\left(\frac{\pi_2}{\pi_3}\right)\\ &= (\alpha_1 + \beta_1 x) - (\alpha_2 + \beta_2 x)\\ &= (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)x.\end{aligned}$$

Nominal logistic regression model – Example in R

x	y
1.24	I
1.30	I
1.30	I
⋮	⋮
3.68	O
3.71	F
3.89	F

59 rows in total.

Predictor: x, alligator's length in cm (continuous variable).
Outcome: y, primary food type (F = fish, I = invertebrate, O = other).

```
# Import data frame from file:
allig.df <- read.table("Alligators.dat", header = TRUE)
```

We will use "O" as the baseline category.

Nominal logistic regression model – Example in R

```
library(VGAM)

allig.fit <- vglm(y ~ x,
                  family = multinomial(refLevel = "O"),
                  data   = allig.df
                  )

coef(allig.fit, matrix = TRUE)
summary(allig.fit)
```

Output:

```
            log(mu[,1]/mu[,3]) log(mu[,2]/mu[,3])
(Intercept)           1.617731           5.697444
x                    -0.110109          -2.465446
------------------------------------------------
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1   1.6177     1.3073   1.237  0.21591
(Intercept):2   5.6974     1.7937   3.176  0.00149 **
x:1            -0.1101     0.5171  -0.213  0.83137
x:2            -2.4654     0.8996      NA       NA
```

Nominal logistic regression model – Example in R

```
            log(mu[,1]/mu[,3]) log(mu[,2]/mu[,3])
(Intercept)           1.617731           5.697444
x                    -0.110109          -2.465446
```

The estimated model is (1 = "F", 2 = "I", 3 = "O")

$$\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_3}\right) = 1.618 - 0.110x$$
$$\log\left(\frac{\hat{\pi}_2}{\hat{\pi}_3}\right) = 5.697 - 2.465x,$$

and also

$$\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) = (1.618 - 5.697) + [-0.110 - (-2.465)]x$$
$$= -4.080 + 2.355x.$$

Nominal logistic regression model – Example in R

Interpretation is all in all similar to the dichotomous case.

For example, from

$$\log\left(\frac{\widehat{\pi}_1}{\widehat{\pi}_2}\right) = -4.080 + 2.355x$$

we conclude that the odds of an alligator eating fish ($y = 1$) over invertebrates ($y = 2$) are multiplied by $e^{2.355} = 10.5$ for each 1m extra of length.

Nominal logistic regression model – Example in R

Is the alligator length independent from its diet?

As usual, we can compare our model with the null model that excludes x:

```
allig.null <- vglm(y ~ 1,
                   family = multinomial(refLevel = "O"),
                   data   = allig.df
                  )

c(deviance(allig.fit), deviance(allig.null))
lrtest(allig.fit, allig.null) # lrtest from the VGAM package
```

Output:

```
[1]  98.34124 115.14186
----------------------------------
  #Df  LogLik Df  Chisq Pr(>Chisq)
2 116 -57.571  2 16.801  0.0002248 ***
----------------------------------
```

Conclusion:
$\chi^2(2) = 115.14 - 98.34 = 16.80$, $p < .001$: At $\alpha = 5\%$, we conclude that there is evidence supporting a relation between the alligator's length and diet.

Estimating response probabilities

In general, we can solve each equation

$$\log\left(\frac{\pi_j}{\pi_c}\right) = \alpha_j + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p$$

with respect to π_j for $j = 1, \ldots, c-1$:

$$\pi_j = \frac{\exp(\alpha_j + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p)}{1 + \sum_{h=1}^{c-1}\exp(\alpha_h + \beta_{h1}x_1 + \cdots + \beta_{hp}x_p)}.$$

Finally, from $\sum_j \pi_j = 1$ we get

$$\begin{aligned}\pi_c &= 1 - (\pi_1 + \cdots + \pi_{c-1}) \\ &= \frac{1}{1 + \sum_{h=1}^{c-1}\exp(\alpha_h + \beta_{h1}x_1 + \cdots + \beta_{hp}x_p)}.\end{aligned}$$

Estimating response probabilities

For the alligator example,

$$\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_3}\right) = 1.618 - 0.110x \qquad \log\left(\frac{\hat{\pi}_2}{\hat{\pi}_3}\right) = 5.697 - 2.465x,$$

we have that

$$\begin{aligned}
\hat{\pi}_1 &= \frac{\exp(1.618 - 0.110x)}{1 + \exp(1.618 - 0.110x) + \exp(5.697 - 2.465x)} \\
\hat{\pi}_2 &= \frac{\exp(5.697 - 2.465x)}{1 + \exp(1.618 - 0.110x) + \exp(5.697 - 2.465x)} \\
\hat{\pi}_3 &= \frac{1}{1 + \exp(1.618 - 0.110x) + \exp(5.697 - 2.465x)}.
\end{aligned}$$

Example, if $x = 1.5$ then $(\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3) = (.34, .58, .08)$.

Estimating response probabilities

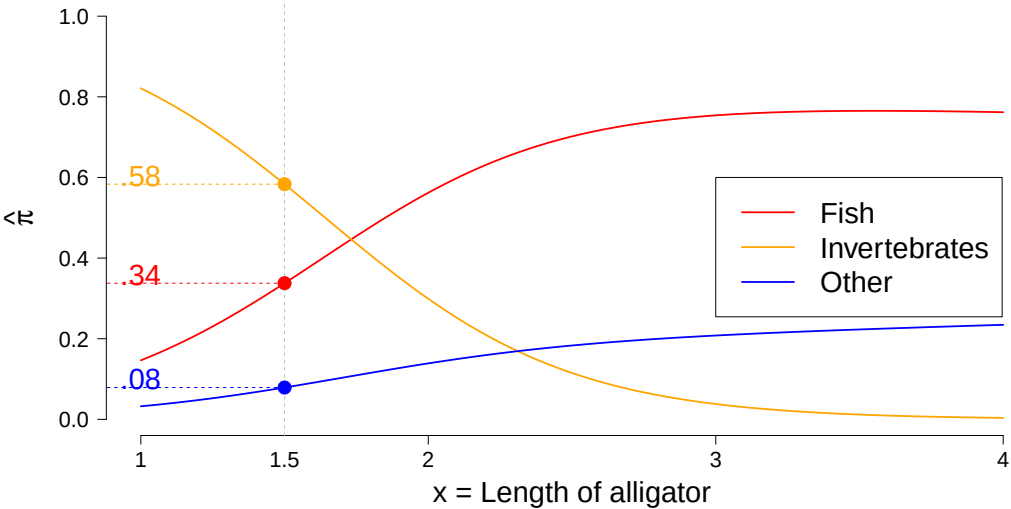
In R:

```
predict(allig.fit,
        newdata = data.frame(x = 1.5),
        type    = "response")
```

Output:

```
         F         I          O
1 0.337664 0.5833332 0.07900282
```

Estimating response probabilities



Refitting the model I

Recall that

$$\begin{aligned}\log\left(\frac{\widehat{\pi}_1}{\widehat{\pi}_2}\right) &= (1.618 - 5.697) + [-0.110 - (-2.465)]x \\ &= -4.080 + 2.355x.\end{aligned}$$

Instead of comparing the two non-baseline categories indirectly, we could have also refitted the model with group 2 ("I") as the baseline category.

The model fit remains the same, but the coefficients change.

Refitting the model I

$$\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) = (1.618 - 5.697) + [-0.110 - (-2.465)]x$$
$$= -4.080 + 2.355x.$$

```
allig2.fit <- vglm(y ~ x,
                   family = multinomial(refLevel = "I"), # change the baseline category
                   data   = allig.df
                  )

coef(allig2.fit, matrix = TRUE)                          # the first column compares "F" to "I"
```

Output:

```
            log(mu[,1]/mu[,2]) log(mu[,3]/mu[,2])
(Intercept)          -4.079713          -5.697444
x                     2.355337           2.465446
```

Refitting the model II

Another thing we can control is the *order* of the levels of variable y.

By default in R, the levels are ordered alphabetically (F→I→O), and the last level is chosen as the baseline.

But we can use whatever level we prefer, by creating a factor for variable y.

Suppose we would like that R orders the levels in this order: I→O→F.

Refitting the model II

```
y.factor <- factor(allig.df$y, levels = c("I", "O", "F"))

allig3.fit <- vglm(y.factor ~ x,
                   family = multinomial(refLevel = "O"),
                   data   = allig.df
                   )

coef(allig3.fit, matrix = TRUE)                          # column 1 = "I" to "O"
                                                         # column 2 = "F" to "O"
```

Output:

```
            log(mu[,1]/mu[,2]) log(mu[,3]/mu[,2])
(Intercept)           5.697444           1.617731
x                    -2.465446          -0.110109
```

Ordinal logistic regression model

Ordinal logistic regression model

AKA cumulative logistic regression model.

This model relies on Y displaying ordered categories.

This model has fewer parameters and is easier to interpret than the nominal logit model.

Ordinal logistic regression model

Define the cumulative probability for category j by

$$P(Y \leq j) = \pi_1 + \cdots + \pi_j,$$

for $j = 1, \ldots, c$.

Clearly,

$$P(Y \leq 1) \leq P(Y \leq 2) \leq \cdots \leq P(Y \leq c) = 1.$$

We can now define $(c-1)$ logits of the cumulative probabilities (*cumulative logits*):

$$\underbrace{\log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right)}_{\text{logit}(P(Y \leq j))} = \log\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_c}\right)$$

for $j = 1, \ldots, c-1$.

Ordinal logistic regression model

$$\boxed{\text{logit}(P(Y \leq j)) = \log\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_c}\right)}$$

Each cumulative logit contrasts categories $\{Y_1, \ldots, Y_j\}$ against $\{Y_{j+1}, \ldots, Y_c\}$.

This is conceptually different from comparing all categories to a baseline.

Ordinal logistic regression model

$$\boxed{P(Y \leq j) = \text{logit}^{-1}\left[\log\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_c}\right)\right]}$$

Once all the cumulative probabilities $P(Y \leq j)$ are available $(j = 1, \ldots, c-1)$, we can use differences to compute direct probabilities for each response category:

$$P(Y = 1) = P(Y \leq 1)$$
$$P(Y = j) = P(Y \leq j) - P(Y \leq j-1), \text{for } j = 2, ..., c-1$$
$$P(Y = c) = 1 - P(Y \leq c-1).$$

Ordinal logistic regression model

Adding predictors is simple.

In fact, it is simpler than the nominal logit model because we assume a constant x effect across all cumulative logits.

For example, for one predictor only x:

$$\text{logit}(P(Y \leq j)) = \alpha_j + \beta x,$$

for $j = 1, \ldots, c-1$.

Ordinal logistic regression model

$$\boxed{\text{logit}(P(Y \leq j)) = \alpha_j + \beta x}$$

Similarly to dichotomous logistic regression, and for any $j = 1, \ldots, c-1$,

$$\exp(\beta) = \frac{\text{cumulative odds}_{x+1}}{\text{cumulative odds}_x}.$$

More generally,

$$\exp[\beta(a-b)] = \frac{\text{cumulative odds}_a}{\text{cumulative odds}_b}.$$

This means that, for any $j = 1, \ldots, c-1$, the cumulative odds are proportional to the difference between the two x values (a and b).

This is the proportional odds property.

Ordinal logistic regression model – Example in R

Gender	Political party	Political ideology				
		Very liberal	Slightly liberal	Moderate	Slightly conservative	Very conservative
Female	Democrat	25	105	86	28	4
	Republican	0	5	15	83	32
Male	Democrat	20	73	43	20	3
	Republican	0	1	14	72	32

Predictors:

- x_1, political party (0 = Democrats, 1 = Republicans)
- x_2, gender (0 = females, 1 = males)

Outcome: y, political ideology (categorical, 5 levels).

Ordinal logistic regression model – Example in R

```
# Import data frame from file:
politic.df <- read.table("Polviews.dat", header = TRUE)
politic.df
```

Output:

```
  gender party y1  y2 y3 y4 y5
1 female   dem 25 105 86 28  4
2 female repub  0   5 15 83 32
3   male   dem 20  73 43 20  3
4   male repub  0   1 14 72 32
```

Ordinal logistic regression model – Example in R

```
library(VGAM)

politic.fit <- vglm(cbind(y1, y2, y3, y4, y5) ~ party + gender,
                    family = cumulative(parallel = TRUE),
                    data   = politic.df
                    )

coef(politic.fit, matrix = TRUE)
```

Output:

```
            logitlink(P[Y<=1]) logitlink(P[Y<=2]) logitlink(P[Y<=3]) logitlink(P[Y<=4])
(Intercept)        -2.12232620         0.16891554          1.85715563          4.65005237
partyrepub         -3.63365808        -3.63365808         -3.63365808         -3.63365808
gendermale          0.04731236         0.04731236          0.04731236          0.04731236
```

Ordinal logistic regression model – Example in R

```
library(VGAM)

politic.fit <- vglm(cbind(y1, y2, y3, y4, y5) ~ party + gender,
                    family = cumulative(parallel = TRUE),
                    data   = politic.df
                    )

summary(politic.fit)
```

Output:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1 -2.12233    0.16875 -12.577   <2e-16 ***
(Intercept):2  0.16892    0.11481   1.471    0.141
(Intercept):3  1.85716    0.15103  12.297   <2e-16 ***
(Intercept):4  4.65005    0.23496  19.791   <2e-16 ***
partyrepub    -3.63366    0.21785 -16.680   <2e-16 ***
gendermale     0.04731    0.14955   0.316    0.752
```

Ordinal logistic regression model – Example in R

The estimated model is

$$\begin{aligned}
\text{logit}(P(Y \leq 1)) &= -2.122 - 3.634\text{party} + 0.047\text{gender} \\
\text{logit}(P(Y \leq 2)) &= 0.169 - 3.634\text{party} + 0.047\text{gender} \\
\text{logit}(P(Y \leq 3)) &= 1.857 - 3.634\text{party} + 0.047\text{gender} \\
\text{logit}(P(Y \leq 4)) &= 4.650 - 3.634\text{party} + 0.047\text{gender}.
\end{aligned}$$

For instance, interpreting the *party* effect:

> *For any* $j = 1, \ldots, 4$*, the estimated odds that a Republican's political ideology is in the liberal direction (i.e.,* $Y \leq j$*) rather than the conservative direction (i.e.,* $Y > j$*) are* $\exp(-3.634) = 0.0026$ *times the estimated odds for Democrats.*

That is, Democrats answer much more likely in the liberal direction.

Ordinal logistic regression model – Example in R

Estimated category probabilities:

```
data.frame(politic.df[, c("gender", "party")],
           fitted(politic.fit)
           )
```

Output:

```
  gender party    y1    y2    y3    y4    y5
1 female   dem 0.107 0.435 0.323 0.126 0.009
2 female repub 0.003 0.027 0.114 0.590 0.266
3   male   dem 0.112 0.442 0.317 0.121 0.009
4   male repub 0.003 0.028 0.119 0.593 0.257
```

For example, $P(Y = 2)$ for the (female, Democrat) group:

- $P(Y \leq 2) = \frac{\exp(0.169)}{1+\exp(0.169)} = .542$
- $P(Y \leq 1) = \frac{\exp(-2.122)}{1+\exp(-2.122)} = .107$
- $P(Y = 2) = P(Y \leq 2) - P(Y \leq 1) = .542 - .107 = .435.$

Ordinal logistic regression model – Example in R

To learn about the *party* effect, we can compare our model to the *gender*-only model via the likelihood-ratio test:

```
library(VGAM)

politic.fit2 <- vglm(cbind(y1, y2, y3, y4, y5) ~ gender,
                     family = cumulative(parallel = TRUE),
                     data   = politic.df
                     )

lrtest(politic.fit, politic.fit2)
```

Output:

```
----------------------------------
  #Df   LogLik Df  Chisq Pr(>Chisq)
2  11 -236.827  1 403.25  < 2.2e-16 ***
----------------------------------
```

Conclusion:

$\chi^2(1) = 403.25$, $p < .001$: We reject the *gender*-only model in favor of the model including both *gender* and *party* effects.

Thus, *party* has a statistically significant effect.

Exercise 11-1

The nominal logistic regression model is used to predict the preference for President (Democrat, Republican, Independent) using the annual income (in $\$10,000$ dollars) as predictor x.

The estimated model equations are

$$log(\hat{\pi}_D/\hat{\pi}_I) = 3.3 - 0.2x$$
$$log(\hat{\pi}_R/\hat{\pi}_I) = 1.0 + 0.3x.$$

a. Which group was used as the baseline?
b. State the prediction equation for $log(\hat{\pi}_R/\hat{\pi}_D)$. Interpret its slope.
c. State the prediction equation for $\hat{\pi}_I$.
d. Find the range of x for which $\hat{\pi}_R > \hat{\pi}_D$.

Exercise 11-2

Research was conducted to relate job satisfaction, Y, with three explanatory variables x_1, x_2, and x_3 as shown below:

	Variables	Response categories			
		1	2	3	4
Y	Job satisfaction	Less satisfied	···	···	Very satisfied
x_1	Earnings compared to others	Much less	···	···	Much more
x_2	Having freedom about how you work	Very true	···	···	Not at all true
x_3	Work environment allows productivity	Strongly agree	···	···	Strongly disagree

For simplicity, let's treat each predictor as being a continuous variable.
The estimated prediction equation is $\text{logit}[\hat{P}(Y \leq j)] = \hat{\alpha}_j - 0.54x_1 + 0.60x_2 + 1.19x_3$.

a. Summarize each partial effect by indicating whether subjects tend to be more satisfied, or less satisfied, as (i) x_1, (ii) x_2, (iii) x_3, increases.
b. Report the settings for x_1, x_2, x_3 at which a subject is most likely to have highest job satisfaction.

Exercise 11-3

Does marital happiness depend on family income? Data from the 2002 General Social Survey are available in file 'happy.csv'. The outcome is variable 'happiness' (categorical with three groups: 'not' happy, 'pretty' happy, 'very' happy), and the predictor is 'income' (1 = below average income; 2 = average income; 3 = above average income).

a. Fit a nominal logistic regression model, with baseline-category equal to 'very' happy (provide all the R code you used).
b. Report the prediction equation for each happiness group.
c. Interpret the income effect in the first equation ('not' happy vs 'very' happy).
d. Report the likelihood ratio test statistic and P-value for testing that marital happiness is independent of family income. Interpret.
e. Estimate the probability that a person with average family income reports a very happy marriage.

Exercise 11-4

Refer to the previous exercise.

a. Fit the ordinal logistic regression model to these data (provide all the R code you used). Report the estimated model equations.
b. Explain why there are two different intercepts but only one income effect.
c. Interpret the income effect.
d. Report a test statistic and P-value for testing that marital happiness is independent of family income. Interpret.
e. Estimate the probability that a person with average family income reports a very happy marriage.

Next lecture

In Lecture 13, we will cover Chapter 7.1 $\sim$ 7.4, but we will skip the following sections: 7.1.5-7.1.7.

In Lecture 14, we will cover Chapter 8.1 and 8.5, but we will skip the following sections: 8.5.4

Exercise 11-1 (Japanese/日本語)

年収 ($\$10,000$ ドル単位) を説明変数 x とし, 大統領の投票政党 (Democrat, Republican, Independent) を予測する為に名義ロジスティックモデルを適用させた.

その結果, 以下のようなモデル方程式が推定された.

$$log(\hat{\pi}_D/\hat{\pi}_I) = 3.3 - 0.2x$$
$$log(\hat{\pi}_R/\hat{\pi}_I) = 1.0 + 0.3x.$$

a. どのグループがベースラインとして用いられているか?
b. $log(\hat{\pi}_R/\hat{\pi}_D)$ を予測する方程式を示し, その傾きを解釈せよ.
c. $\hat{\pi}_I$ を予測する方程式を示せ.
d. $\hat{\pi}_R > \hat{\pi}_D$ における x の範囲を求めよ.

Exercise 11-2

仕事の満足度 Y と以下の説明変数 x_1, x_2, x_3 に関する研究が行われた:

	変数	回答カテゴリー			
		1	2	3	4
Y	仕事の満足度	不満足	…	…	大変満足
x_1	他者に比べた収入	かなり少ない	…	…	かなり多い
x_2	仕事のやり方の自由度が高い	とても当てはまる	…	…	全く当てはまらない
x_3	生産性を高める職業環境	とても当てはまる	…	…	全く当てはまらない

簡便の為, 説明変数は連続変数として扱う.
推定された予測モデルは $\text{logit}[\widehat{P}(Y \leq j)] = \widehat{\alpha}_j - 0.54x_1 + 0.60x_2 + 1.19x_3$ と得られた.

a. (i) x_1, (ii) x_2, (iii) x_3 が増えるに従って仕事の満足度は上がるか下がるかに着目して各変数の効果を要約せよ.
b. x_1, x_2, x_3 がどのような回答の時に仕事の満足度が最も高くなる傾向があるか述べよ.

Exercise 11-3

夫婦の幸福度は家族の収入に依存するかについて考える. 2002 年の General Social Survey の結果が Moodle の Example data のフォルダーの'happy.csv' に保存されている. 目的変数は'happiness(幸福度)' (3 グループのカテゴリカル変数: 'not' happy, 'pretty' happy, 'very' happy) であり, 説明変数は'income(収入)' (1 = below average income; 2 = average income; 3 = above average income).

a. ベースラインカテゴリーを'very' happy として名義ロジスティック回帰モデルを適用せよ (分析に用いた R コードもレポートに示すこと).
b. 各'happiness' グループの予測モデルを示せ.
c. 最初の方程式 ('not' happy vs 'very' happy) における収入の効果を解釈せよ.
d. 夫婦の幸福度が年収と独立であるという仮説の尤度比検定統計量とその P 値を報告せよ.
e. 平均的な年収の家族で夫婦の幸福度が'very happy' と回答する確率を推定せよ.

Exercise 11-4

引き続き Exercise 11-3 について考える.

a. Exercise 11-3 で用いたデータに対して順序ロジスティック回帰を適用し, その R コードもレポートに表示せよ. 推定されたモデル式を報告せよ.
b. なぜ収入の効果が 1 つなのに対して 2 つの異なる切片が生じるのか説明せよ.
c. 収入の効果を解釈せよ.
d. 夫婦の幸福度が年収と独立であるという仮説の検定統計量とその P 値を報告し, 解釈せよ.
e. 平均的な年収の家族で夫婦の幸福度が'very happy' と回答する確率を推定せよ.

Final Exam

- The final exam will be given on Thursday, Jauary 30th 8:45 10:30 at **EDU** K201.
- The exam will consist entirely of multiple-choice questions, and you will need a black pen to fill out the answer sheet. Please ensure you bring your own black pen to the exam.
- The questions will cover the material discussed in the lectures up to and including Lecture 14.

最終筆記試験

- 最終筆記試験は第 1 回の講義でアナウンスした通り、2025 年 1 月 30 日 (木)8:45 10:30に教 K201 で行います。
- 試験はすべて選択式の問題で構成され、回答用紙を記入するために黒のペンが必要です。必ずご自身で黒のペンを持参してください。
- テスト範囲は、第 14 回講義までに扱った内容です。