

PS6 Solutions

Jingle Fu

Problem 1

Solution (a).

Yes, there are missing values in the data set. The initial number of observations is 99457, The dataset's reported missing values indicate that only *age* has missing observations (119 missing values). After dropping these, we end up with 99338 observations.

```
1 rm(list = ls())
2 library(tidyverse)
3 library(ggplot2)
4 library(dplyr)
5 library(broom)
6 library(stats)
7 library(stargazer)
8 library(car)
9
10 set.seed(2024)
11
12 dat <- read.csv("dat_SalesCustomers.csv")
13 variables_to_check <- c("category", "price", "gender", "age", "payment_
    method")
14 missing_counts <- sapply(dat[variables_to_check], function(x) sum(is.na(
    x)))
15 print("Number of missing values in each variable:")
16 print(missing_counts)
17
18 dat_clean <- dat[complete.cases(dat[variables_to_check]), ]
19
20 num_observations <- nrow(dat_clean)
21 print(paste("Number of observations after removing missing values:", num
    _observations))
22
```

Solution (b).

Define

$$\text{paid_in_cash}_i = \mathbf{1}\{\text{payment_method}_i = \text{Cash}\}$$

$$\text{male}_i = \mathbf{1}\{\text{gender}_i = \text{Male}\}$$

The fraction of transactions carried out in cash is

$$\frac{1}{n} \sum_{i=1}^n \text{paid_in_cash}_i.$$

Empirically, this is about 44.69%.

The fraction of overall sales carried out in cash is

$$\frac{\sum_{i=1}^n \text{paid_in_cash}_i \cdot \text{price}_i}{\sum_{i=1}^n \text{price}_i}.$$

Empirically, this fraction is about 44.79%.

These results indicate that cash payments represent nearly half of all transactions and sales value.

```

1 dat_clean$paid_in_cash <- ifelse(dat_clean$payment_method == "Cash", 1,
  0)
2 dat_clean$male <- ifelse(dat_clean$gender == "Male", 1, 0)
3 fraction_cash_transactions <- mean(dat_clean$paid_in_cash)
4 print(paste("Fraction of transactions carried out in cash:",
5             round(fraction_cash_transactions * 100, 2), "%"))
6
7 total_sales <- sum(dat_clean$price)
8 cash_sales <- sum(dat_clean$price[dat_clean$paid_in_cash == 1])
9 fraction_cash_sales <- cash_sales / total_sales
10 print(paste("Fraction of overall sales carried out in cash:",
11             round(fraction_cash_sales * 100, 2), "%"))
12

```

Solution (c).

We now consider only the first $n = 1000$ observations. Let the categories be divided into five mutually exclusive groups: Clothes and Shoes (C), Cosmetics (Cos), Food (F), Technology (T), and Other (O). Define indicator variables:

$$\begin{aligned}
 d_{C,i} &= \mathbf{1}\{\text{category}_i = \text{Clothes and Shoes}\} \\
 d_{Cos,i} &= \mathbf{1}\{\text{category}_i = \text{Cosmetics}\} \\
 d_{F,i} &= \mathbf{1}\{\text{category}_i = \text{Food}\} \\
 d_{T,i} &= \mathbf{1}\{\text{category}_i = \text{Technology}\} \\
 d_{O,i} &= 1 - (d_{C,i} + d_{Cos,i} + d_{F,i} + d_{T,i}).
 \end{aligned}$$

The fraction of transactions in category j is

$$\frac{1}{1000} \sum_{i=1}^{1000} d_{j,i}.$$

The fraction of sales in category j is

$$\frac{\sum_{i=1}^{1000} d_{j,i} \cdot price_i}{\sum_{i=1}^{1000} price_i}.$$

Empirically:

- Transactions fraction: Clothes/Shoes: 43.8%, Cosmetics: 14.8%, Food: 14.0%, Technology: 5.0%, Other: 22.4%.
- Sales fraction: Clothes/Shoes: 70.58%, Cosmetics: 2.72%, Food: 0.32%, Technology: 23.9%, Other: 2.49%.

The result shows that most transactions and sales are in the Clothes/Shoes category. Technology, though having the lowest transaction fraction, has the second-highest sales fraction, meaning that it has the highest average price.

```

1 dat_1000 <- dat_clean[1:1000, ]
2
3 dat_1000$clothes_shoes <- ifelse(dat_1000$category %in% c("Clothing", "
  Shoes"), 1, 0)
4 dat_1000$cosmetics <- ifelse(dat_1000$category == "Cosmetics", 1, 0)
5 dat_1000$food <- ifelse(dat_1000$category %in% c("Food", "Food &
  Beverage"), 1, 0)
6 dat_1000$technology <- ifelse(dat_1000$category == "Technology", 1, 0)
7
8 dat_1000$other_category <- ifelse(dat_1000$clothes_shoes + dat_1000$
  cosmetics + dat_1000$food + dat_1000$technology == 0, 1, 0)
9
10 all(dat_1000$clothes_shoes + dat_1000$cosmetics + dat_1000$food + dat_
  1000$technology + dat_1000$other_category == 1)
11
12 fraction_transactions <- c(
13   "Clothes and Shoes" = mean(dat_1000$clothes_shoes),
14   "Cosmetics" = mean(dat_1000$cosmetics),
15   "Food" = mean(dat_1000$food),
16   "Technology" = mean(dat_1000$technology),
17   "Other" = mean(dat_1000$other_category)
18 )
19 print("Fraction of transactions in each category:")
20 print(round(fraction_transactions * 100, 2))
21
22 total_sales_1000 <- sum(dat_1000$price)
23 sales_clothes_shoes <- sum(dat_1000$price[dat_1000$clothes_shoes == 1])
24 sales_cosmetics <- sum(dat_1000$price[dat_1000$cosmetics == 1])
25 sales_food <- sum(dat_1000$price[dat_1000$food == 1])
26 sales_technology <- sum(dat_1000$price[dat_1000$technology == 1])
27 sales_other <- sum(dat_1000$price[dat_1000$other_category == 1])
28

```

```

29 fraction_sales <- c(
30   "Clothes and Shoes" = sales_clothes_shoes / total_sales_1000,
31   "Cosmetics" = sales_cosmetics / total_sales_1000,
32   "Food" = sales_food / total_sales_1000,
33   "Technology" = sales_technology / total_sales_1000,
34   "Other" = sales_other / total_sales_1000
35 )
36
37 print("Fraction of sales in each category:")
38 print(round(fraction_sales * 100, 2))
39

```

Solution (d).

For a sample $\{(y_i, x_i)\}_{i=1}^n$, the log-likelihood function is:

$$\ell(\beta; Z_n) = \sum_{i=1}^n [y_i \log(\Phi(x_i' \beta)) + (1 - y_i) \log(1 - \Phi(x_i' \beta))].$$

To derive the Hessian, focus on a single observation i and then we will take expectations:

$$\ell_i(\beta) = y_i \log(\Phi(t_i)) + (1 - y_i) \log(1 - \Phi(t_i)), \quad t_i = x_i' \beta.$$

First Derivative (Score):

First, take the derivative with respect to t_i :

$$\frac{\partial \ell_i(\beta)}{\partial t_i} = y_i \frac{\phi(t_i)}{\Phi(t_i)} - (1 - y_i) \frac{\phi(t_i)}{1 - \Phi(t_i)},$$

where $\phi(\cdot)$ is the standard normal PDF.

Combine the fractions:

$$\frac{\partial \ell_i(\beta)}{\partial t_i} = \phi(t_i) \left[\frac{y_i}{\Phi(t_i)} - \frac{1 - y_i}{1 - \Phi(t_i)} \right].$$

Find a common denominator $\Phi(t_i)(1 - \Phi(t_i))$:

$$\frac{\partial \ell_i(\beta)}{\partial t_i} = \frac{\phi(t_i)}{\Phi(t_i)(1 - \Phi(t_i))} (y_i - \Phi(t_i)).$$

Since $y_i - \Phi(t_i) = y_i - p_i$, we have:

$$\frac{\partial \ell_i(\beta)}{\partial t_i} = \frac{\phi(t_i)}{p_i(1 - p_i)} (y_i - p_i).$$

To get the gradient w.r.t. β , use chain rule:

$$\frac{\partial \ell_i(\beta)}{\partial \beta} = \frac{\partial \ell_i(\beta)}{\partial t_i} x_i = \frac{\phi(t_i)}{p_i(1 - p_i)} (y_i - p_i) x_i.$$

So,

$$s_n(\beta) = -\frac{1}{n} \nabla_{\beta} \ell(\beta; Z_n) = -\frac{1}{n} \sum_{i=1}^n \left[y_i \frac{\phi(x'_i \beta)}{\Phi(x'_i \beta)} - (1 - y_i) \frac{\phi(x'_i \beta)}{1 - \Phi(x'_i \beta)} \right] x'_i.$$

Second Derivative (Hessian):

Now differentiate again with respect to β :

$$\frac{\partial^2 \ell_i(\beta)}{\partial \beta \partial \beta'} = \frac{\partial}{\partial \beta} \left(\frac{\phi(t_i)}{p_i(1 - p_i)} (y_i - p_i) x_i \right).$$

Since $t_i = x'_i \beta$, $\frac{\partial t_i}{\partial \beta} = x_i$. Thus, second derivatives w.r.t. β come through differentiating w.r.t. t_i , then applying chain rule again:

$$\frac{\partial^2 \ell_i(\beta)}{\partial \beta \partial \beta'} = \left(\frac{\partial^2 \ell_i(\beta)}{\partial t_i^2} \right) x_i x'_i.$$

We have:

$$\frac{\partial \ell_i(\beta)}{\partial t_i} = \frac{\phi(t_i)}{p_i(1 - p_i)} (y_i - p_i).$$

Take the derivative w.r.t. t_i :

$$\frac{\partial^2 \ell_i(\beta)}{\partial t_i^2} = \frac{\partial}{\partial t_i} \left[\frac{\phi(t_i)}{p_i(1 - p_i)} (y_i - p_i) \right].$$

This involves the product rule and quotient rule. However, *the key simplification occurs when we take expectations at the true parameter β_0* . Under the true model, $E[y_i] = p_i$, so $E[y_i - p_i] = 0$. Terms involving $(y_i - p_i)$ vanish when taking expectation.

At the true parameter, the Fisher information (which is $-E[\partial^2 \ell_i(\beta_0)/\partial \beta \partial \beta']$) simplifies dramatically. Instead of going through the full complex algebra of the second derivative in the y_i form, we use the known result from standard Probit derivations:

Under correct specification, the expected Hessian w.r.t. t_i at β_0 is known to be:

$$E \left[\frac{\partial^2 \ell_i(\beta_0)}{\partial t_i^2} \right] = -\frac{\phi(t_i)^2}{p_i(1 - p_i)}.$$

Thus:

$$E \left[\frac{\partial^2 \ell_i(\beta_0)}{\partial \beta \partial \beta'} \right] = E \left[\frac{\partial^2 \ell_i(\beta_0)}{\partial t_i^2} x_i x'_i \right] = E \left[-\frac{\phi(t_i)^2}{p_i(1 - p_i)} x_i x'_i \right].$$

Multiplying by -1 , the Fisher Information matrix (which is H in the problem) is:

$$H = E \left[\frac{\phi(x'_i \beta_0)^2}{\Phi(x'_i \beta_0)[1 - \Phi(x'_i \beta_0)]} x_i x'_i \right].$$

Since $1 - \Phi(t_i) = \Phi(-t_i)$:

$$H = E \left[\frac{\phi(x'_i \beta_0)^2}{\Phi(x'_i \beta_0)\Phi(-x'_i \beta_0)} x_i x'_i \right].$$

Step 1: Characterizing the MLE $\hat{\beta}$

The MLE $\hat{\beta}$ sets the gradient to zero:

$$\nabla_{\beta} \ell(\hat{\beta}; Z_n) = 0.$$

Substituting back:

$$\sum_{i=1}^n \left[y_i \frac{\phi(x'_i \hat{\beta})}{\Phi(x'_i \hat{\beta})} - (1 - y_i) \frac{\phi(x'_i \hat{\beta})}{1 - \Phi(x'_i \hat{\beta})} \right] x_i = 0.$$

This is a system of k nonlinear equations in the k unknowns $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_k)'$.

Step 2: No Closed-Form Solution

Unlike in linear regression or the logit model (even the logit doesn't have a closed form), the Probit model does not admit a closed-form solution for $\hat{\beta}$. The equation above must be solved using numerical optimization techniques such as the Newton-Raphson algorithm or other iterative methods.

Step 3: Numerical Optimization

A common iterative procedure is:

Algorithm 1: Newton-Raphson Method

Input: Initialize β_0 , tolerance level $\varepsilon > 0$

```

1 for  $m = 1$  to  $M$  do
2   | Given  $\beta^m$ , compute  $\nabla_{\beta} \ell(\beta^m; Z_n)$  and  $[H(\beta^m; Z_n)]$ ;
3   | Set  $\beta^{(m+1)} = \beta^m - [H(\beta^m; Z_n)]^{-1} \nabla_{\beta} \ell(\beta^m; Z_n)$ ;
4   | if  $\|\beta^{m+1} - \beta^m\| < \varepsilon$  then
5   |   |  $\hat{\beta} = \beta^{m+1}$ ;
6   | else
7   |   | Proceed to the next iteration;
8   | end
9 end
```

where $H(\beta; Z_n)$ is the Hessian matrix of second derivatives evaluated at β . Convergence is achieved when changes in β or the norm of the gradient are below a given tolerance.

The regression result is as follows:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_{price} \\ \hat{\beta}_{male} \\ \hat{\beta}_{age} \\ \hat{\beta}_{cosmetics} \\ \hat{\beta}_{food} \\ \hat{\beta}_{technology} \end{bmatrix} = \begin{bmatrix} 0.0682 \\ 0.000112 \\ -0.0502 \\ -0.00183 \\ -0.2879 \\ -0.1195 \\ 0.0640 \\ -0.4195 \end{bmatrix}.$$

Interpretation:

- The coefficient on price is positive but very small, suggesting a tiny positive association of price with the probability of cash payment (not statistically significant).
- male is negative, but not significant, suggesting no strong gender effect on the probability of cash usage.
- age coefficient is negative and small, not statistically significant either.
- Some category dummies (like Clothes/Shoes) are significantly different from zero, indicating that the reference category (likely "Other") differs in payment method probability.

```

1 X <- as.matrix(cbind(1, dat_1000[, c("price", "male", "age", "clothes_
   shoes", "cosmetics", "food", "technology")]))
2 y <- dat_1000$paid_in_cash
3
4 neg_log_likelihood <- function(beta, X, y) {
5   X_beta <- X %*% beta
6   log_phi_Xb <- pnorm(X_beta, log.p = TRUE)
7   log_phi_minus_Xb <- pnorm(-X_beta, log.p = TRUE)
8   ll <- sum(y * log_phi_Xb + (1 - y) * log_phi_minus_Xb)
9   return(-ll)
10 }
11
12 neg_log_likelihood_grad <- function(beta, X, y) {
13   X_beta <- X %*% beta
14   phi_Xb <- dnorm(X_beta)
15   Phi_Xb <- pnorm(X_beta)
16   Phi_minus_Xb <- pnorm(-X_beta)
17   epsilon <- 1e-16
18   Phi_Xb <- pmax(Phi_Xb, epsilon)
19   Phi_minus_Xb <- pmax(Phi_minus_Xb, epsilon)
20   gradient <- -t(X) %*% ((y * phi_Xb / Phi_Xb) - ((1 - y) * phi_Xb / Phi_
   _minus_Xb))
21   return(as.vector(gradient))
22 }
23
24 initial_beta <- rep(0, ncol(X))
25
26 result <- optim(par = initial_beta, fn = neg_log_likelihood, gr = neg_
   log_likelihood_grad, X = X, y = y, method = "BFGS")
27
28 if (result$convergence == 0) {
29   cat("Optimization converged.\n")
30 } else {
31   cat("Optimization did not converge.\n")

```

Table 1: Optimization model

	<i>Dependent variable:</i>
	paid_in_cash
price	0.0001 (0.0001)
male	−0.050 (0.081)
age	−0.002 (0.003)
clothes_shoes	−0.288** (0.130)
cosmetics	−0.120 (0.133)
food	0.064 (0.135)
technology	−0.420 (0.314)
Constant	0.068 (0.148)
Observations	1,000
Log Likelihood	−685.217
Akaike Inf. Crit.	1,386.434
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	


```

32 }
33
34 beta_hat <- result$par
35 print("Estimated coefficients (beta_hat):")
36 print(beta_hat)
37
38 ### Programming Method
39 model <- glm(paid_in_cash ~ price + male + age + clothes_shoes +
40             cosmetics + food + technology,
41             data = dat_1000, family = binomial(link = "probit"))
42 stargazer(model, type = "latex", title = "Optimization model", out = "d.
43           tex")
44
45 beta_hat2 <- coef(model)
46 print("Estimated coefficients (beta_hat):")
47 print(beta_hat2)
48

```

Solution (e).

We define

$$\gamma_1(\beta) = \Phi(x'_2\beta) - \Phi(x'_1\beta),$$

where x'_1 is a vector for a 30-year-old male buying Clothes/Shoes for 500 TRY, and x'_2 is the same vector but with age increased to 60 years old. Only the age element of x_i changes from 30 to 60.

For our estimated $\hat{\beta}$,

$$\gamma_1(\hat{\beta}) \approx -0.02096.$$

This suggests that increasing age from 30 to 60 reduces the probability of cash payment by about 2.1 percentage points for this specific profile.

For $\gamma_2(\beta)$, we do not condition on category. We take a weighted average of the partial effects across the five categories, with weights given by their share in total sales:

$$\gamma_2(\hat{\beta}) = \sum_j w_j [\Phi(x'_{2,j}\hat{\beta}) - \Phi(x'_{1,j}\hat{\beta})],$$

where w_j is the sales fraction for category j .

Empirically,

$$\gamma_2(\hat{\beta}) \approx -0.02077,$$

very close to $\gamma_1(\hat{\beta})$, indicating a similar overall effect once categories are averaged by their sales importance.

```

1 x_age_30 <- c(1, 500, 1, 30, 1, 0, 0, 0)
2
3 x_age_60 <- x_age_30
4 x_age_60[4] <- 60 # Update age to 35

```

```

5
6 prob_age_30 <- pnorm(sum(x_age_30 * beta_hat))
7 prob_age_60 <- pnorm(sum(x_age_60 * beta_hat))
8
9 gamma_1 <- prob_age_60 - prob_age_30
10 print(paste("Gamma_1 (effect of age increasing by 5 years):", gamma_1))
11
12 gamma_c <- numeric(length(fraction_sales))
13 names(gamma_c) <- names(fraction_sales)
14
15 for (cat in names(fraction_sales)) {
16   clothes_shoes <- ifelse(cat == "Clothes and Shoes", 1, 0)
17   cosmetics <- ifelse(cat == "Cosmetics", 1, 0)
18   food <- ifelse(cat == "Food", 1, 0)
19   technology <- ifelse(cat == "Technology", 1, 0)
20
21   x_age_30_2 <- c(1, 500, 1, 30, clothes_shoes, cosmetics, food,
22                 technology)
23   x_age_60_2 <- x_age_30_2
24   x_age_60_2[4] <- 60 # Update age to 35
25
26   prob_age_30_2 <- pnorm(sum(x_age_30_2 * beta_hat))
27   prob_age_60_2 <- pnorm(sum(x_age_60_2 * beta_hat))
28
29   gamma_c[cat] <- prob_age_60_2 - prob_age_30_2
30 }
31 gamma_2 <- sum(fraction_sales * gamma_c)
32 print(paste("Gamma_2 (weighted effect over categories):", gamma_2))
33

```

Solution (f).

Consider the linear model

$$y_i = x_i' \beta + u_i$$

where $u_i \mid x_i \sim N(0, 1)$.

Step 1: Define the Objective Function

Define $\mathcal{B} = \{\beta \in \mathbb{R} : \|\beta\| \leq c\}$ for some very large c .

The objective function is given by:

$$\hat{\beta} = \arg \min_{\beta \in \mathcal{B}} Q_n(\beta) = \arg \min_{\beta \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^n (y_i - x_i' \beta)^2$$

Step 2: Express the Limiting Behavior of the Objective Function

The expected form of the objective function $Q_n(\beta)$ is:

$$\begin{aligned}
 Q(\beta) &= E[(y_i - x_i'\beta)^2] \\
 &= E[(x_i'\beta_0 + u_i - x_i'\beta)^2] \\
 &= E[(x_i'(\beta_0 - \beta) + u_i)^2] \\
 &= E[(x_i'(\beta_0 - \beta))^2] + 2E[x_i'(\beta_0 - \beta)u_i] + E[u_i^2] \\
 &= E[(x_i'(\beta_0 - \beta))^2] + \sigma^2 \\
 &= (\beta_0 - \beta)'E[x_i x_i'](\beta_0 - \beta) + 1
 \end{aligned}$$

since $E[u_i | x_i] = 0$ and $\sigma^2 = 1$.

Step 3: Show Minimization of $Q(\beta)$ at β_0

As $E[x_i x_i']$ is positive definite, the function $Q(\beta)$ is minimized at β_0 because:

$$(\beta_0 - \beta)'E[x_i x_i'](\beta_0 - \beta) \geq 0$$

which is zero if and only if $\beta = \beta_0$.

Step 4: Prove Uniform Convergence of $Q_n(\beta)$ to $Q(\beta)$

It's obvious that $m(x_i, y_i, \beta) = (y_i - x_i'\beta)^2$ satisfies the first three conditions of the Uniform Law of Large Numbers (ULLN).

We then prove the fourth one:

$$\begin{aligned}
 E \left[\sup_{\beta \in \mathcal{B}} \|m(x_i, y_i, \beta)\| \right] &\leq E[|y_i|^2] + \sup_{\beta \in \mathcal{B}} 2E[|y_i| |x_i| \|\beta\|] + \sup_{\beta \in \mathcal{B}} E[|x_i|^2 \|\beta\|^2] \\
 &< \infty
 \end{aligned}$$

By ULLN:

$$Q_n(\beta) \xrightarrow{p} Q(\beta), n \rightarrow \infty$$

Step 5: Demonstrate the Consistency of $\hat{\beta}_n$

According to extremum estimator theory, if $Q_n(\beta)$ converges uniformly to $Q(\beta)$ and $Q(\beta)$ has a unique global minimum at β_0 , then:

$$\hat{\beta}_n \xrightarrow{p} \beta_0$$

Consistency of Probit Estimator:

First, we define $f(w_i; \theta) = \Phi(x_i'\beta)^{y_i} \Phi(-x_i'\beta)^{1-y_i}$, then $\log f(w_i; \theta) = y_i \log \Phi(x_i'\beta) + (1 - y_i) \log \Phi(-x_i'\beta)$.

Step 1: Theorem of Consistency

Theorem 1. If $Q_n(w_i; \theta)$ is a function of w_i and θ such that:

(A) Parameter space $\Theta \in \mathbb{R}^k$ is compact, $\theta_0 \in \Theta$;

(B) $Q_n(w_i; \theta)$ is continuous in $\theta \in \Theta$ for all w_i .

(C) $Q_n(\theta)$ converges in probability to $Q(\theta)$ uniformly in $\theta \in \Theta$, and $Q(\theta)$ has a unique global minimum at θ_0 .

Define $Q_n(\hat{\theta}_n) = \max_{\theta \in \Theta} Q_n(\theta)$.

Then, $\hat{\theta}_n \xrightarrow{p} \theta_0$.

Proof for Theorem 1.

Let N be a neighbourhood in \mathbb{R}^k containing θ_0 . Then $\overline{N} \cap \Theta$ is compact $\Rightarrow \max_{\theta \in \overline{N} \cap \Theta} Q(\theta)$ exists.

Denote $\varepsilon = Q(\theta_0) - \max_{\theta \in \overline{N} \cap \Theta} Q(\theta)$.

Define event A_n as:

$$A_n : \left| \frac{1}{n} Q_n(\theta) - Q(\theta) \right| < \frac{\varepsilon}{2}.$$

This implies that:

$$\begin{cases} Q(\hat{\theta}_n) > \frac{1}{n} Q_n(\hat{\theta}_n) - \frac{\varepsilon}{2} \\ \frac{1}{n} Q_n(\hat{\theta}_n) > Q(\theta_0) - \frac{\varepsilon}{2} \end{cases}$$

But, as $Q_N(\hat{\theta}_n) \geq Q_n(\theta_0)$,

$$Q(\hat{\theta}_n) > Q(\theta_0) - \varepsilon \Rightarrow \hat{\theta}_n \in N.$$

Thus, $\mathbb{P}[A_n] \leq \mathbb{P}[\hat{\theta}_n \in N]$. Since we have $\lim_{n \rightarrow \infty} \mathbb{P}[A_n] = 1$ by (C), we have $\mathbb{P}[\hat{\theta}_n \in N] \rightarrow 1$.

Hence $\hat{\theta}_n \xrightarrow{p} \theta_0$. □

In our case, we take the parameter space $\mathcal{B} = \{\beta \in \mathbb{R}^k : \|\beta\| < c\}$ for some large c , then, we have our compact parameter space Θ , (A) is satisfied.

As we take $Q_n(\theta) = \frac{1}{n} \ell(\beta; Z_n) = \frac{1}{n} \sum_{i=1}^n \log f(w_i; \beta)$, it's continuous in θ for all w_i , (B) is satisfied.

So, we need to prove two conditions for the theorem to hold:

1. $Q_n(\theta)$ converges in probability to $Q(\theta)$ uniformly in $\theta \in \Theta$.
2. (Identification) $Q(\theta)$ has a unique global minimum at θ_0 .

Step 2: Identification of Probit Model

Definition 1 (Identification). The information matrix $I(\theta)$ is defined as:

$$I(\theta) = E \left[\frac{\partial^2 \ell^2(w_i; \theta)}{\partial \theta \partial \theta'} \right].$$

If $I(\theta)$ is positive definite, then θ is identified.

If θ is identified, it means that if $\theta \neq \theta_0$, then $f(w_i; \theta) \neq f(w_i; \theta_0)$.

Lemma 1 (Information Inequality).

If θ is identified, and $\mathbb{E} [\log f(w_i; \theta) < \infty]$ for all θ , then $Q(\theta) = \mathbb{E} [\log f(w_i; \theta)]$ has a unique maximum at θ_0 .

Proof for Lemma 1.

For a random variable Y , by Jensen's inequality, we have:

$$-\log \mathbb{E}[Y] < \mathbb{E} [-\log Y].$$

In our case, we define a new random variable for $\theta \neq \theta_0$:

$$Y = \frac{f(w_i; \theta)}{\mathbb{E}[f(w_i; \theta_0)]}.$$

Then, we have:

$$\begin{aligned} Q(\theta_0) - Q(\theta) &= \mathbb{E} [-\log Y] > -\log \mathbb{E}[Y] \\ &= -\log \int f(w_i; \theta) d\theta = 0 \end{aligned}$$

□

Step 3: Prove uniform convergence of Probit Model

To prove the uniform convergence of the Probit model, we give the second theorem:

Theorem 2 (Uniform Law of Large Numbers).

If x_i (or say, data) are i.i.d, and $\log f(w_i; \theta)$ is a function of x_i, y_i, θ such that:

- (a) Parameter space $\Theta \in \mathbb{R}^k$ is compact, $\theta_0 \in \Theta$;
- (b) $m(w_i; \theta)$ is continuous at each $\theta \in \Theta$ with probability to 1;
- (c) There exist a dominant function $d(w_i)$ such that $\|\log f(w_i; \theta)\| \leq \|d(w_i)\| \forall \theta \in \Theta$;
- (d) $E[d(w_i)] < \infty$.

then, $\mathbb{E} [\log f(w_i; \theta)]$ is continuous and

$$\sup_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n \log f(w_i; \theta) - \mathbb{E} [\log f(w_i; \theta)] \right\| \xrightarrow{p} 0.$$

Proof for Theorem 2.

For $\forall \theta_0 \in \Theta$, we define $\mathcal{B}(\theta_0, \delta) = \{\theta \in \Theta : \|\theta - \theta_0\| < \delta\}$ and

$$\Delta_\delta(w_i; \theta) = \sup_{\theta \in \mathcal{B}(\theta_0, \delta)} (\log f(w_i; \theta) - \mathbb{E} [\log f(w_i; \theta)]).$$

For $\theta_0 \in \Theta$, we have:

$$\mathbb{E} [\Delta_\delta(w_i; \theta) \rightarrow 0] \text{ as } \delta \rightarrow 0.$$

because

1. $\Delta_\delta(w_i; \theta_0) \rightarrow \log f(w_i; \theta_0) - \mathbb{E} [\log f(w_i; \theta_0)]$ almost surely as $\delta \rightarrow 0$, because:

$$\mathbb{P} \left[\lim_{\delta \rightarrow 0} \sup_{\theta \in \mathcal{B}(\theta_0, \delta)} \log f(w_i; \theta) = \log f(w_i; \theta_0) \right] = 1$$

by condition (b) and that $\mathbb{E} [\log f(w_i; \theta)]$ is continuous at θ_0 .

2. By condition (c) and (d), we have:

$$\Delta_\delta(w_i; \theta_0) \leq 2 \sup_{\theta \in \mathcal{B}(\theta_0, \delta)} |\log f(w_i; \theta)| \leq 2d(w_i)$$

So, for all $\theta \in \Theta$, $\varepsilon > 0$, $\exists \delta_\varepsilon(\theta)$, such that

$$\mathbb{E} [\Delta_{\delta_\varepsilon(\theta)}(w_i; \theta)] < \varepsilon.$$

Obviously, we can cover the entire parameter space with a finite number of $\mathcal{B}(\theta, \delta_\varepsilon(\theta)) : \theta \in \Theta$, which is:

$$\mathcal{B}(\theta_k, \delta_\varepsilon(\theta_k)) : k = 1, 2, \dots, K \text{ s.t. } \Theta = \bigcup_{k=1}^K \mathcal{B}(\theta_k, \delta_\varepsilon(\theta_k)).$$

Note that:

$$\begin{aligned} & \sup_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^n \log f(w_i; \theta) - \mathbb{E} [\log f(w_i; \theta)] \right] \\ &= \max_k \sup_{\theta \in \mathcal{B}(\theta_k, \delta_\varepsilon(\theta_k))} \left[\frac{1}{n} \sum_{i=1}^n \log f(w_i; \theta) - \mathbb{E} [\log f(w_i; \theta)] \right] \\ &\leq \max_k \frac{1}{n} \left[\sum_{i=1}^n \sup_{\theta \in \mathcal{B}(\theta_k, \delta_\varepsilon(\theta_k))} \log f(w_i; \theta) - \mathbb{E} [\log f(w_i; \theta)] \right] \\ &\leq \mathcal{O}_p(1) + \max_k \mathbb{E}^* \Delta_{\delta_\varepsilon(\theta_k)}(w_i; \theta_k) \\ &= \mathcal{O}_p(1) + \varepsilon. \end{aligned}$$

where the second inequality holds by the Weak Law of Large Numbers (WLLN) because:

$$\left| \sup_{\theta \in \mathcal{B}(\theta_k, \delta_\varepsilon(\theta_k))} \log f(w_i; \theta) \right| \leq d(w_i); \mathbb{E}[d(w_i)] < \infty.$$

and the third inequality holds by the definition of $\delta_\varepsilon(\theta_k)$. By analogous argument, we can

prove that:

$$\inf_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^n \log f(w_i; \theta) - \mathbb{E} [\log f(w_i; \theta)] \right] \geq O_p(1) - \varepsilon.$$

Combing the two results, we have:

$$\left| \frac{1}{n} \sum_{i=1}^n \log f(w_i; \theta) - \mathbb{E} [\log f(w_i; \theta)] \right| \rightarrow O_p(1) = 0.$$

□

Finishing the proof of Theorem 2, we can find that the Probit model still have to satisfy conditions (c) and (d) to hold the theorem.

Step 4: Proof of Conditions (c) and (d) for ULLN

In this part, we show that identification and the uniform convergence of the Probit model are combined by the existence of $\mathbb{E}[x_i x_i']$ and its nonsingularity.

Proof for ULLN conditions (c) and (d).

For this proof, we take two steps:

Step 1: $\mathbb{E} [|\log f(w_i; \theta)|]$ is finite.

Let $\theta \neq \theta_0$, then

$$\begin{aligned} \mathbb{E} \left[(x_i'(\theta - \theta_0))^2 \right] &= (\theta - \theta_0)' \mathbb{E}[x_i x_i'] (\theta - \theta_0) > 0 \\ &\Rightarrow x_i'(\theta - \theta_0) \neq 0 \\ &\Rightarrow x_i' \theta \neq x_i' \theta_0 \end{aligned}$$

Since Φ is strictly monotone, this gives us $\Phi(x_i' \theta) \neq \Phi(-x_i' \theta)$. So that $f(w_i; \theta) = \Phi(x_i' \beta)^{y_i} \Phi(-x_i' \beta)^{1-y_i} \neq f(w_i; \theta_0)$.

We know that $\frac{d \log \Phi(v)}{dx} = \frac{\phi(v)}{\Phi(v)}$ is convex and asymptotic to 0 as $v \rightarrow \infty$ and to $-v$ as $v \rightarrow -\infty$.

We take the mean-value expansion around $\theta = 0$:

$$\begin{aligned} |\log \Phi(x_i' \theta)| &= \left| \log \Phi(0) + \lambda(x_i' \tilde{\theta}) x_i' \theta \right| \\ &\leq |\log \Phi(0)| + \left| \lambda(x_i' \tilde{\theta}) x_i' \theta \right| \\ &\leq |\log \Phi(0)| + C \left(1 + |x_i' \tilde{\theta}| \right) |x_i' \theta| \\ &\leq |\log \Phi(0)| + C (1 + \|x_i\| \|\theta\|) \|x_i\| \|\theta\| \end{aligned}$$

where λ is the reverse Mills ratio.

Since $1 - \Phi(v) = \Phi(-v)$ and y are bounded, we have:

$$|\log f(w_i; \theta)| \leq |\log \Phi(0)| + C (1 + \|x_i\| \|\theta\|) \|x_i\| \|\theta\|$$

where C is a constant.

Thus, we could say that $\mathbb{E}[|\log f(w_i; \theta)|]$ is finite.

Step 2: $\mathbb{E}[d(w_i)]$ exist, and is finite.

Based on Step 1, we could directly take

$$d(w_i) = C(1 + \|x_i\|^2).$$

It's obvious that $\mathbb{E}[d(w_i)]$ is finite. □

Combining Lemma 1, Lemma 2, Theorem 1, and Theorem 2, we could say that the Probit model estimator is consistent.

Solution (g).

```

1 M <- 100
2 n <- nrow(dat_1000)
3 beta_age_bootstrap <- numeric(M)
4 gamma_1_bootstrap <- numeric(M)
5 gamma_2_bootstrap <- numeric(M)
6 set.seed(2024)
7
8 for (m in 1:M) {
9   indices <- sample(1:n, size = n, replace = TRUE)
10  dat_bootstrap <- dat_1000[indices, ]
11
12  model_boot <- glm(paid_in_cash ~ price + male + age + clothes_shoes +
13    cosmetics + food + technology, data = dat_bootstrap, family =
14    binomial(link = "probit"))
15  beta_hat_boot <- coef(model_boot)
16
17  beta_age_bootstrap[m] <- beta_hat_boot["age"]
18
19  x_age_30 <- c(1, 500, 1, 30, 1, 0, 0, 0)
20  x_age_60 <- x_age_30
21  x_age_60[4] <- 60 # Update age to 35
22  prob_age_30 <- pnorm(sum(x_age_30 * beta_hat_boot))
23  prob_age_60 <- pnorm(sum(x_age_60 * beta_hat_boot))
24  gamma_1_bootstrap[m] <- prob_age_60 - prob_age_30
25
26  gamma_c <- numeric(length(fraction_sales))
27  names(gamma_c) <- names(fraction_sales)
28
29  for (cat in names(fraction_sales)) {
30    clothes_shoes <- ifelse(cat == "Clothes and Shoes", 1, 0)
31    cosmetics <- ifelse(cat == "Cosmetics", 1, 0)
32    food <- ifelse(cat == "Food", 1, 0)
33    technology <- ifelse(cat == "Technology", 1, 0)

```



```
33   x_age_30_2 <- c(1, 500, 1, 30, clothes_shoes, cosmetics, food,
34   technology)
35   x_age_60_2 <- x_age_30
36   x_age_60_2[4] <- 60 # Update age to 35
37
38   prob_age_30_2 <- pnorm(sum(x_age_30_2 * beta_hat_boot))
39   prob_age_60_2 <- pnorm(sum(x_age_60_2 * beta_hat_boot))
40   gamma_c[cat] <- prob_age_60_2 - prob_age_30_2
41 }
42
43 gamma_2_bootstrap[m] <- sum(fraction_sales * gamma_c)
44 }
45 hist(beta_age_bootstrap, main = "Bootstrap Distribution of Coefficient
46   on Age", xlab = "Coefficient on Age", breaks = 20)
47 hist(gamma_1_bootstrap, main = "Bootstrap Distribution of Gamma_1", xlab
48   = "Gamma_1", breaks = 20)
49 hist(gamma_2_bootstrap, main = "Bootstrap Distribution of Gamma_2", xlab
50   = "Gamma_2", breaks = 20)
```

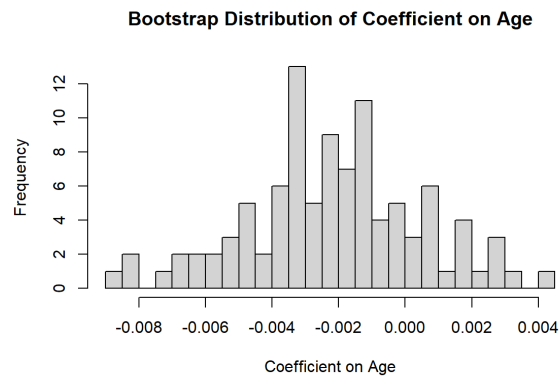


Figure 1: Bootstrap Distribution of Coefficient on Age

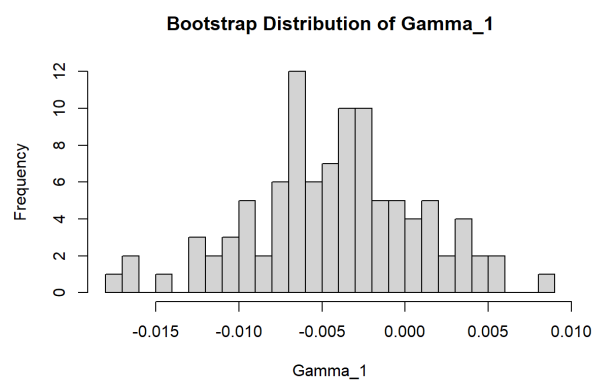


Figure 2: Bootstrap Distribution of Gamma_1

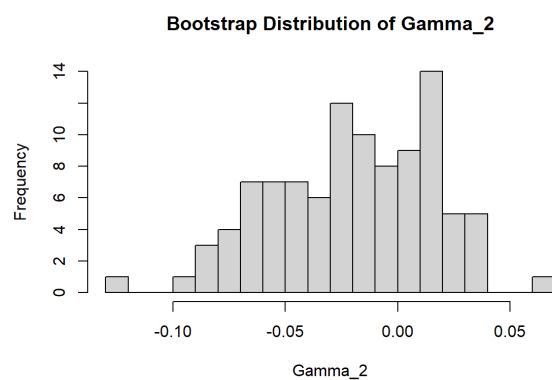


Figure 3: Bootstrap Distribution of Gamma_2

Solution (h).

In the dataset result, the given histograms for the bootstrap distributions and the asymptotic distributions show approximately symmetric, bell-shaped distributions. This suggests that the normal approximation may be reasonable.

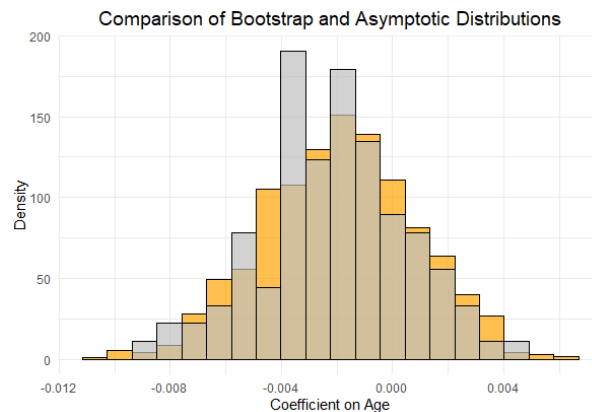


Figure 4: Comparison of Bootstrap and Asymptotic Distributions

```

1 X <- as.matrix(cbind(1, dat_1000[, c("price", "male", "age", "clothes_
   shoes", "cosmetics", "food", "technology")]))
2 X_beta_hat <- X %*% beta_hat
3
4 log_phi_Xb <- dnorm(X_beta_hat, log = TRUE)
5 log_Phi_Xb <- pnorm(X_beta_hat, log.p = TRUE)
6
7 log_Phi_minus_Xb <- pnorm(-X_beta_hat, log.p = TRUE)
8 log_factor <- 2 * log_phi_Xb - log_Phi_Xb - log_Phi_minus_Xb
9 factor <- exp(log_factor)
10 factor[!is.finite(factor)] <- 0
11 factor <- as.vector(factor)
12 X_weighted <- sweep(X, 1, factor, FUN = "*")
13
14 H_hat <- t(X) %*% X_weighted / nrow(X)
15 V_hat <- solve(H_hat)
16
17 variance_beta_age <- V_hat[4, 4]
18 beta_age_sd <- sqrt(variance_beta_age / nrow(X))
19
20 if (!is.finite(beta_age_sd)) {
21   stop("Standard error for the coefficient on age is not finite.")
22 }
23
24 mean_age <- beta_hat[4]
25 simulated_draws <- rnorm(1000, mean = mean_age, sd = beta_age_sd)
26
27 library(ggplot2)
28

```

```

29 bootstrap_data <- data.frame(Distribution = "Bootstrap", Values = beta_
    age_bootstrap)
30 asymptotic_data <- data.frame(Distribution = "Asymptotic", Values =
    simulated_draws)
31 combined_data <- rbind(bootstrap_data, asymptotic_data)
32
33 ggplot(combined_data, aes(x = Values, fill = Distribution)) +
34   geom_histogram(aes(y = ..density..),
35     bins = 20, alpha = 1, position = "identity", color = "
    black") +
36   scale_fill_manual(values = c("Bootstrap" = "grey", "Asymptotic" = "red
    ")) +
37   labs(title = "Comparison of Bootstrap and Asymptotic Distributions",
38     x = "Coefficient on Age", y = "Density") +
39   theme_minimal() +
40   theme(plot.title = element_text(hjust = 0.5, size = 14),
41     legend.title = element_blank(),
42     legend.position = "topright") +
43   guides(fill = guide_legend(reverse = TRUE))
44

```

Solution (i).

The Delta Method states that if

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, H^{-1}),$$

and $g(\cdot)$ is a continuously differentiable function at β_0 , then

$$\sqrt{n}(g(\hat{\beta}) - g(\beta_0)) \xrightarrow{d} N(0, \nabla_{\beta} g(\beta_0)' H^{-1} \nabla_{\beta} g(\beta_0)).$$

In our case, $g(\beta) = \gamma_1(\beta)$.

Computing the Gradient $\nabla_{\beta} \gamma_1(\beta)$:

We have:

$$\gamma_1(\beta) = \Phi(x_2' \beta) - \Phi(x_1' \beta).$$

The gradient with respect to β is:

$$\nabla_{\beta} \gamma_1(\beta) = \frac{\partial}{\partial \beta} [\Phi(x_2' \beta)] - \frac{\partial}{\partial \beta} [\Phi(x_1' \beta)].$$

Since $\frac{d}{dt} \Phi(t) = \phi(t)$, we get:

$$\nabla_{\beta} \gamma_1(\beta) = \phi(x_2' \beta) x_2 - \phi(x_1' \beta) x_1.$$

Asymptotic Distribution of $\gamma_1(\hat{\beta})$:

Applying the Delta Method at β_0 :

$$\sqrt{n}(\gamma_1(\hat{\beta}) - \gamma_1(\beta_0)) \xrightarrow{d} N(0, \nabla_{\beta}\gamma_1(\beta_0)' H^{-1} \nabla_{\beta}\gamma_1(\beta_0)).$$

In finite samples, we replace β_0 with $\hat{\beta}$, and H with its estimator \hat{H} , thus:

$$\gamma_1(\hat{\beta}) \overset{approx}{\sim} N\left(\gamma_1(\hat{\beta}), \frac{1}{n} \nabla_{\beta}\gamma_1(\hat{\beta})' \hat{H}^{-1} \nabla_{\beta}\gamma_1(\hat{\beta})\right),$$

where \hat{H} and $\nabla_{\beta}\gamma_1(\hat{\beta})$ are computed from the sample and the estimated parameters. This gives us an asymptotic approximation to the finite sample distribution of $\gamma_1(\hat{\beta})$.

To summarize, the asymptotic variance of $\gamma_1(\hat{\beta})$ is:

$$\widehat{\text{Var}}(\gamma_1(\hat{\beta})) = \frac{1}{n} \nabla_{\beta}\gamma_1(\hat{\beta})' \hat{H}^{-1} \nabla_{\beta}\gamma_1(\hat{\beta}).$$

Empirical Implementation:

1. Estimate $\hat{\beta}$ using the Probit model.
2. Compute $\nabla_{\beta}\gamma_1(\hat{\beta}) = \phi(x_2'\hat{\beta})x_2 - \phi(x_1'\hat{\beta})x_1$.
3. Compute \hat{H}^{-1} (the inverse of the estimated H matrix).
4. Approximate:

$$\gamma_1(\hat{\beta}) \sim N\left(\gamma_1(\hat{\beta}), \frac{1}{n} \nabla_{\beta}\gamma_1(\hat{\beta})' \hat{H}^{-1} \nabla_{\beta}\gamma_1(\hat{\beta})\right).$$

The comparison shows that both approximations yield similar conclusions: the effect is not statistically significant, and the distributions are roughly symmetric.

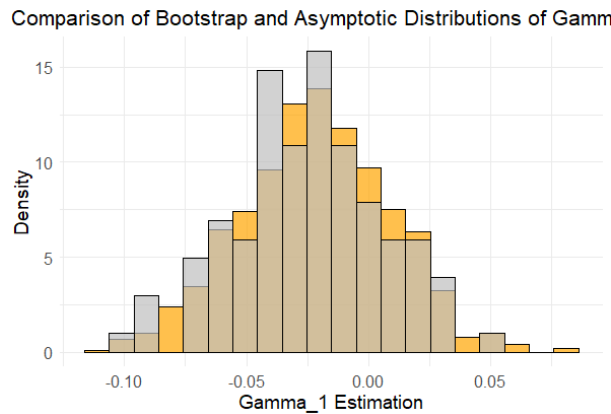


Figure 5: Comparison of Bootstrap and Asymptotic Distributions of Gamma_1

```

1 phi_age_60 <- dnorm(sum(x_age_60 * beta_hat))
2 phi_age_30 <- dnorm(sum(x_age_30 * beta_hat))
3 grad_g <- phi_age_60 * x_age_60 - phi_age_30 * x_age_30
4
5 print(grad_g)
6
7 # Compute asymptotic variance

```

```

8 var_gamma_1 <- t(grad_g) %*% V_hat %*% grad_g / nrow(dat_1000)
9 gamma_1_sd <- sqrt(var_gamma_1)
10
11 simulated_gamma <- rnorm(1000, mean = gamma_1, sd = gamma_1_sd)
12
13 bootstrap_data2 <- data.frame(Value = gamma_1_bootstrap, Distribution =
14   "Bootstrap")
15
16 simulated_data2 <- data.frame(Value = simulated_gamma, Distribution = "
17   Asymptotic")
18
19 # Combine data
20 combined_data2 <- rbind(bootstrap_data2, simulated_data2)
21
22 # Create the plot
23 ggplot(combined_data2, aes(x = Value, fill = Distribution)) +
24   geom_histogram(aes(y = ..density..), bins = 20, position = "identity",
25     alpha = 0.7, color = "black") +
26   scale_fill_manual(values = c("Bootstrap" = "grey", "Asymptotic" = "
27     orange")) +
28   labs(title = "Comparison of Bootstrap and Asymptotic Distributions of
29     Gamma_1",
30     x = "Gamma_1 Estimation", y = "Density") +
31   theme_minimal() +
32   theme(plot.title = element_text(hjust = 0.5, size = 16),
33     legend.title = element_blank(),
34     legend.position = "topright",
35     axis.title = element_text(size = 14),
36     axis.text = element_text(size = 12)) +
37   guides(fill = guide_legend(reverse = TRUE))

```

Solution (j).

We test

$$H_0 : \gamma_1(\beta) = 0 \quad \text{vs.} \quad H_1 : \gamma_1(\beta) \neq 0.$$

Under the asymptotic approximation,

$$\gamma_1(\hat{\beta}) \approx N\left(\gamma_1(\beta_0), \frac{1}{n}\hat{V}\right),$$

where

$$\hat{V} = \nabla_{\beta}\gamma_1(\hat{\beta})'\hat{H}^{-1}\nabla_{\beta}\gamma_1(\hat{\beta}).$$

The t-statistic is:

$$t = \frac{\gamma_1(\hat{\beta})}{\sqrt{\hat{V}/n}}.$$

Empirically, $t \approx -0.68$.

A 95% confidence interval for $\gamma_1(\beta)$ is:

$$\left[\gamma_1(\hat{\beta}) - 1.96\sqrt{\frac{\hat{V}}{n}}, \gamma_1(\hat{\beta}) + 1.96\sqrt{\frac{\hat{V}}{n}} \right].$$

Empirically, the 95% Confidence Interval for $\gamma_1(\beta)$ is: -0.0815 to 0.0396, covering 0.

So, we conclude that the expected probabilities of cash payment for a 30 year-old and a 60 year-old male buying clothes for 500 TRY are not significantly different.

```

1 t_statistic <- gamma_1 / gamma_1_sd
2 critical_value <- qnorm(0.975) # 1.96 for 95% confidence
3
4 if (abs(t_statistic) > critical_value) {
5   conclusion <- "Reject the null hypothesis."
6 } else {
7   conclusion <- "Fail to reject the null hypothesis."
8 }
9
10 print(paste("t-statistic:", round(t_statistic, 2)))
11 print(paste("Conclusion:", conclusion))
12
13 lower_bound <- gamma_1 - critical_value * gamma_1_sd
14 upper_bound <- gamma_1 + critical_value * gamma_1_sd
15
16 print(paste("95% Confidence Interval for gamma_1:", round(lower_bound,
17   4), "to", round(upper_bound, 4)))

```