Project1(B) Report

2025-02-27

Project 1: Multivariate Non-Normal Distributions and Correlated Data

Distributions and correlation

```r
set.seed(123)  # for reproducibility

# Define sizes and parameters
n     <- 200    # number of subjects
t     <- 4      # number of repeated measurements per subject
beta0 <- -1.0   # intercept on logit scale
beta1 <-  0.3   # effect of time on logit scale
sigma <-  1.0   # std dev of random intercept (b_i)

# Create a data frame with one row per subject-time combination
# We'll store each subject's random intercept in 'b_i'.
dat <- data.frame(
  id   = rep(1:n, each = t),
  time = rep(1:t,     times = n)
)

# Simulate one random intercept per subject
# Then replicate that intercept across all time points for that subject.
b_i <- rnorm(n, mean = 0, sd = sigma)
dat$b_i <- b_i[dat$id]   # match the random intercept to each row

# Compute probability p_ij = logistic(beta0 + beta1*time_j + b_i)
# Then draw Y_ij ~ Bernoulli(p_ij).
dat$p_ij <- plogis(beta0 + beta1 * dat$time + dat$b_i)
dat$Y    <- rbinom(n * t, size = 1, prob = dat$p_ij)

head(dat)
```

Data generation method

```
##   id time        b_i      p_ij Y
## 1  1    1 -0.5604756 0.2208920 1
## 2  1    2 -0.5604756 0.2767830 0
## 3  1    3 -0.5604756 0.3406328 1
## 4  1    4 -0.5604756 0.4108444 0
## 5  2    1 -0.2301775 0.2828887 0
## 6  2    2 -0.2301775 0.3474703 0
```

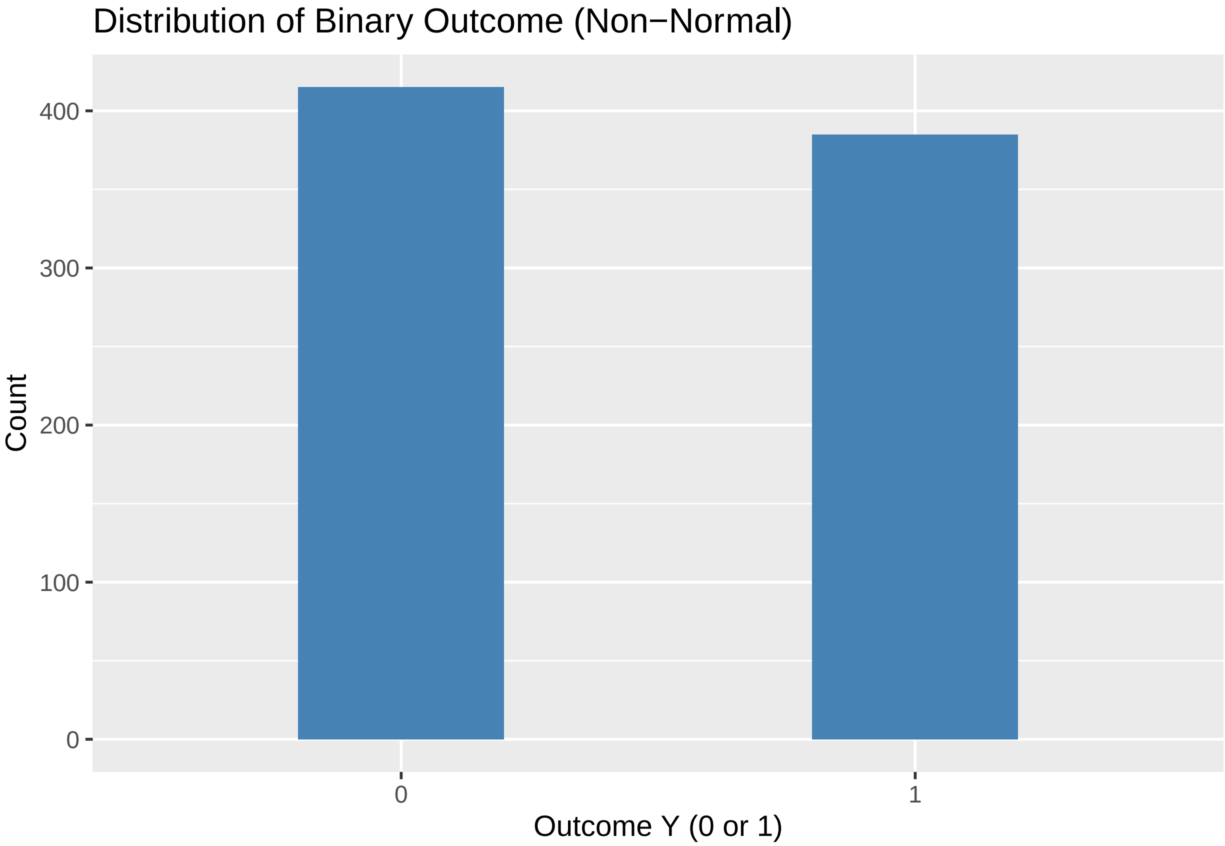
```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)

#    Simply show the binary response Y.
ggplot(dat, aes(x = factor(Y))) +
  geom_bar(width = 0.4, fill = "steelblue") +
  xlab("Outcome Y (0 or 1)") +
  ylab("Count") +
  ggtitle("Distribution of Binary Outcome (Non-Normal)")
```



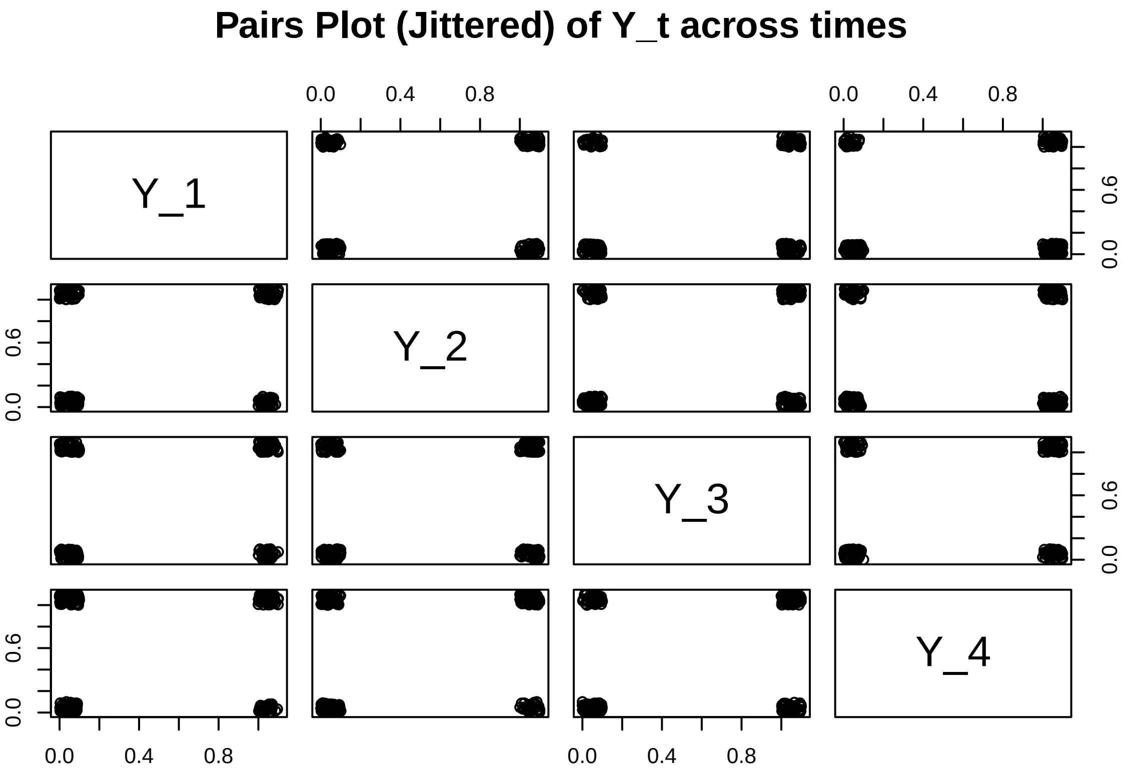
```
# each row becomes a subject
dat_wide <- dat %>%
  select(id, time, Y) %>%
  pivot_wider(names_from = time,
              values_from = Y,
              names_prefix = "Y_")

# Now dat_wide has columns: id, Y_1, Y_2, ... up to Y_t
pairwise_corr <- cor(dat_wide[, -1])
pairwise_corr
```

```
##           Y_1       Y_2       Y_3       Y_4
## Y_1 1.0000000 0.1232673 0.1908684 0.1661559
## Y_2 0.1232673 1.0000000 0.1498906 0.1391663
## Y_3 0.1908684 0.1498906 1.0000000 0.1933134
## Y_4 0.1661559 0.1391663 0.1933134 1.0000000
```

```
#> This matrix shows the sample correlation among times 1..t,

# which clearly prove that outcomes are not independent
pairs(
  dat_wide[, -1] + matrix(runif(nrow(dat_wide)*(ncol(dat_wide)-1), 0, 0.1),
                          nrow(dat_wide), ncol(dat_wide)-1),
  main = "Pairs Plot (Jittered) of Y_t across times"
)
```



```
# ----------------------------
# 2) Fit a naive logistic regression
# ----------------------------

naive_glm <- glm(Y ~ time, data = dat, family = binomial)
summary(naive_glm)
```

Simulation study 1

```
##
## Call:
## glm(formula = Y ~ time, family = binomial, data = dat)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.58530    0.17561  -3.333 0.000859 ***
## time         0.20372    0.06396   3.185 0.001448 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1107.9  on 799  degrees of freedom
## Residual deviance: 1097.7  on 798  degrees of freedom
## AIC: 1101.7
##
## Number of Fisher Scoring iterations: 4
```

```
# install.packages("geepack")
library(geepack)

gee_fit <- geeglm(Y ~ time,
                  family = binomial,
                  id = id,
                  corstr = "exchangeable",
                  data = dat)
summary(gee_fit)
```

Simulation study 2

```
##
## Call:
## geeglm(formula = Y ~ time, family = binomial, data = dat, id = id,
##     corstr = "exchangeable")
##
##  Coefficients:
##             Estimate  Std.err  Wald Pr(>|W|)
## (Intercept) -0.58552  0.16939 11.95 0.000547 ***
## time         0.20372  0.05835 12.19 0.000481 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate  Std.err
## (Intercept)        1 0.008393
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.1601 0.03668
## Number of clusters:   200  Maximum cluster size: 4
```