Final Exam

Please give complete solutions, showing your work and explaining clearly. Lack of details can result in point deduction.

All figures are given at the end of the document.

NAME: Your Name
DUE DATE: April 23, 9:00am

Question 1: Exploratory Data Analysis.

a) (5 points) Is the boxplot in Figure 1 left-skewed, right-skewed, or symmetric? Explain why.

The boxplot is left-skewed. Because there are lots of outliers in the negative side of the axis. Also, the median is upper near zero, so we see there is a big body near zero and a small tail skewed to the negative ray of axis.

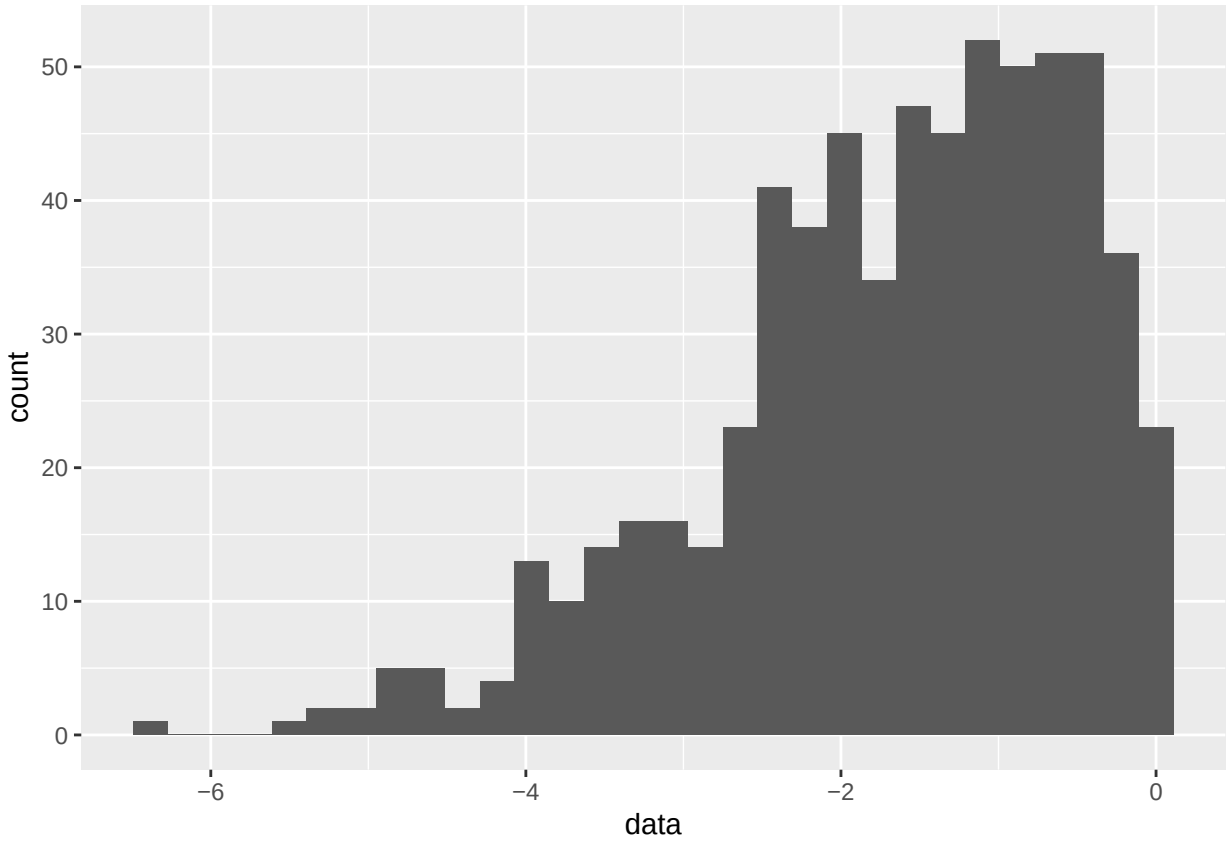
b) (10 points) Generate 1000 units from a random variable from a normal distribution with mean equals to 0 and variance equals to 3. Now adjust this random variable such that the new data are similar to the distribution presented in the boxplot. Plot (in ggplot) the histogram of the new data points.

```
sample <- rnorm(1000, 0, sqrt(3))
sample <- sample - 0.6
data <- sample[sample <= 0]
ggplot() + geom_histogram(aes(data))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Variable	$\hat{\beta}_j$	SE($\hat{\beta}_j$)
constant	4.1769	0.3790
X_1	-0.0127	0.0223
X_2	0.1100	0.0366

Table 1: Linear Regression model output



Question 2: Linear Regression and Generalized Linear Model.

This question explores the use of linear regression and the Generalized Linear Model in general.

a) (10 points) Suppose to perform a linear regression model and to look at the output in Table 1. Are your estimates significant or not? Explain and motivate your procedure. Write down the null and alternative hypothesis to generate the significance of these estimates. Suppose that X_1 is a random variable from a normal distribution with mean 0 and variance 4 and X_2 is a random variable from a normal distribution with mean 0 and variance 5. Generate 1000 samples from these two distributions.

t-statistic is $\hat{\beta}_j$ SE($\hat{\beta}_j$). We will calculate t-statistc for x1 and x2 in the following code and get the p-values then.

```
t1 <- -0.0127 / 0.0223
t2 <- 0.1100 / 0.0366
p1 <- 2*pnorm(t1)
p2 <- 2*pnorm(t2, lower.tail = FALSE)
```

The null hypothesis is B1 = 0, alternative hypothesis is B1 != 0; The null hypothesis is B2 = 0, alternative

hypothesis is B2 != 0; As above, we calcualte p-value for X1 is 0.569 and p-value for X2 is 0.00265. Therefore, X1 is not significant and X2 is significant.

```
x1 <- rnorm(1000, 0, 2)
x2 <- rnorm(1000, 0, sqrt(5))
```

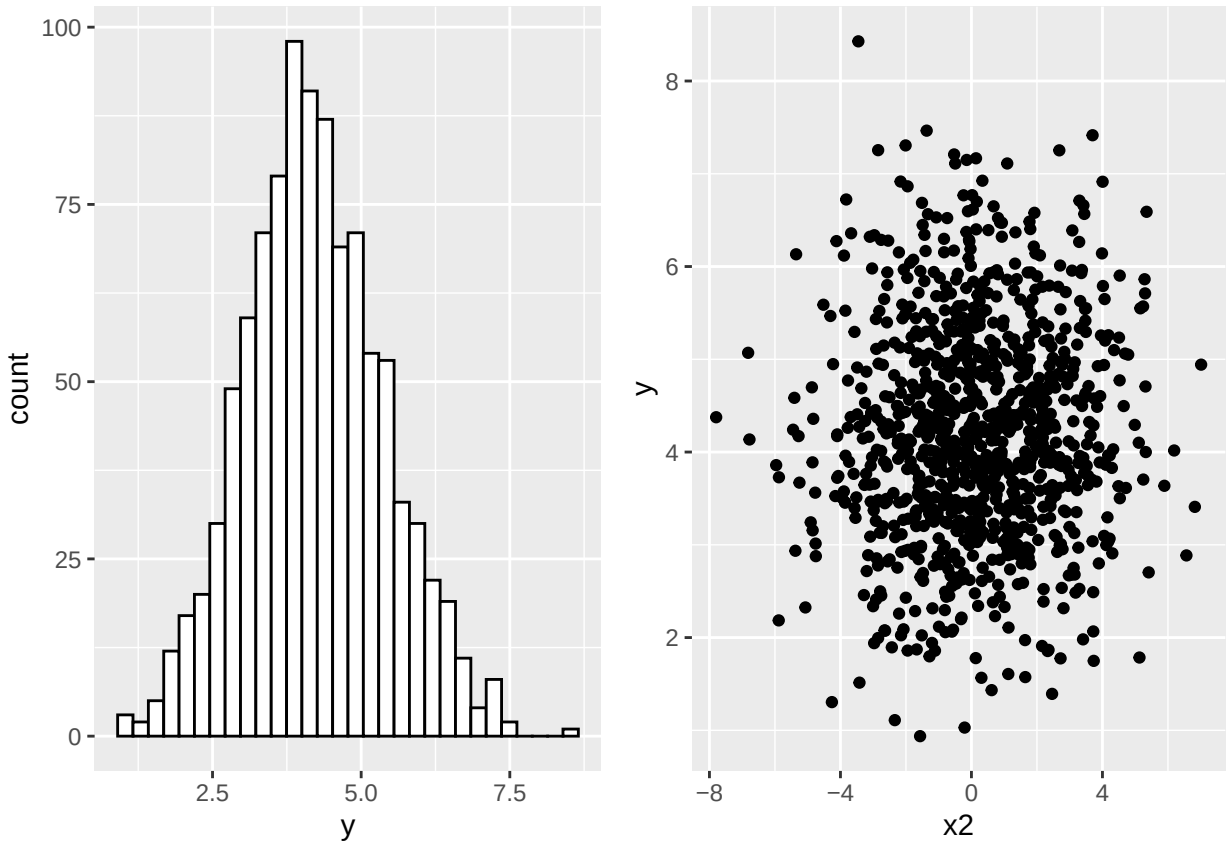
b) (10 points) Generate y by assuming the linear regression model obtained in Table 1 with the variance of the error term equals to $\sigma^2 = 1.2$ (use the distribution property of linear regression and not the linear equation), plot side-by-side (using ggplot) the histogram of y and the scatterplot of y and X_2. Is the distribution of y symmetric or not? Do you expect the scatterplot to have this shape? Explain.

$E(Y) = E(\beta_0 + \beta_1 x_1 + \beta_2 x2 + \epsilon) = E(\beta_0) + E(\beta_1)E(x_1) + E(\beta_2)E(x_2) + E(\epsilon) = 4.1769 - 0.0127 * E(x_1) + 0.11 * E(x_1) + 0$

$var(Y) = var(\beta_0 + \beta_1 x_1 + \beta_2 x2 + \epsilon) = \beta_1^2 var(x_1) + \beta_2^2 var(x_2) + var(\epsilon)$ With the above equations, we can sample from the distribution of Y in a normal distribution with E(Y) and var(Y). The distribution of y is symmetric. Yes. Since y is sampled from a normal distribution, y is symmetric as normal distribution is symmetric around its mean. The scatterplot is also in the right shape. Since both x2 and y are sampled from normal distribution, we should have symmetric view centered around (0, 4) with standard deviation of x2 and y horizontally and vertically. The plot is almost symmetric either looking from x-axis or y-axis. It is a multinomial distribution.

```
e_y = -0.0127 * mean(x1) + 0.11 * mean(x2) + 4.1769
var_y = -0.0127^2*var(x1) + 0.11^2*var(x2) + 1.2
y <- rnorm(1000, e_y, sqrt(var_y))
df <- data.frame(x2, y)
g1 <- ggplot(df, aes(x=y)) + geom_histogram(color="black", fill="white")
g2 <- ggplot(df, aes(x=x2, y=y)) + geom_point()
grid.arrange(g1, g2, ncol = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



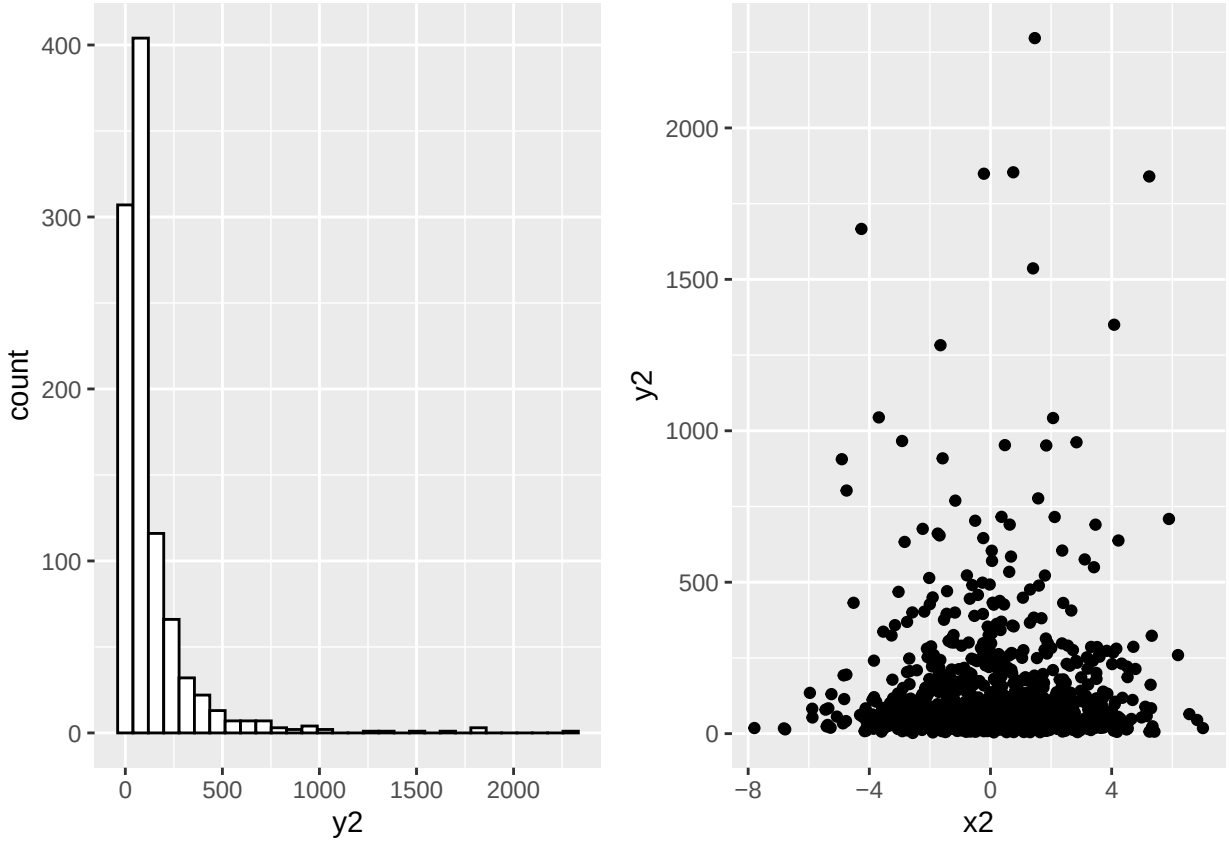
c) (20 points) Now suppose that the model in Table 1 is a Poisson regression model (so it is a generalized linear model, with a Poisson link) and we want to predict the number of defect products y. First, write down the Poisson model. Second, interpret the coefficient. Finally, by using the distributions in a for X_1 and X_2, generate y from the new Poisson regression model (use the distribution property of Poisson regression). Plot side-by-side (using ggplot) the histogram of the new y and the scatterplot of the new y and X_2. Explain what you have obtained.

The poisson model of Y: Y ~ Poisson(k, λ). *lambda* here is the expected number of defect products from a period of time. K is discrete integers that means the number of defect products. P(Y=k) means the probability that number of defect products is k. $logE(Y|X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ -> $E(Y|X) = e^{\beta_0+\beta_1 x_1+\beta_2 x_2}$. Therefore $Y_{poi} = e^{y_{norm}}$. Here β_0 is the intercept, β_1 and β_2 are coefficients of x_1, x_2. The exponentiated β_1 coefficient is the multiplicative term to use to calculate the estimated defect products (Y) when x_1 increases by 1 unit. The exponentiated β_2 coefficient is the multiplicative term to use to calculate the estimated defect products (Y) when x_2 increases by 1 unit. The exp(Intercept), exponentiated β_0 is the baseline rate, and all other estimates would be relative to it.

The new Y is derived from the original y from normal distribution. The histogram and scatter plot are not symmetric and the historgram is right skewed. Since we exponentialize the original y random variable samples, all values will be positive and original y corresponding to negavtive values will be very small and positive ones will be very large. Hence, the histogram is right skewed. x_2 is significant. In the scatter plot, we can see that most y values concetrate below 500 and even below 200. There are outliers above 1000 and 2000. We can also see that when x proceeds around zero, the y value is higher than those far left and right. Outliers also appear around 0. Hence, the linear relationship between x_2 and y is broken since we have an exponential term in the Poisson model. When the exponential term is added, as x_2 goes to negatvie, the rate of change is decreseaing, while as x_2 goes to positive, the rate of change is increasing.

```
y2 <- exp(rnorm(1000, e_y, sqrt(var_y)))
df2 <- data.frame(x2, y2)
g3 <- ggplot(df2, aes(x=y2)) + geom_histogram(color="black", fill="white")
g4 <- ggplot(df2, aes(x=x2, y=y2)) + geom_point()
grid.arrange(g3, g4, ncol = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Question 3: Multi-level Models.

We are performing two different multilevel models, plotted in Figure 2.

a) (10 points) Write down what type of different multi-level models are described in Figure 2 for A and B. Suppose you do not have any group-level covariates, write down the distributions of the two models described in Figure 2 and write down the linear models for model A and B. Give a written interpretation for these two models.

The lines in figure A are parallel with different intercepts, which means that the group in A are only affect the intercepts of the linear regression models. Therefore, A is a Random intercepts model. In Figure B, not only the intercepts are different across different groups, the regression lines have joined each other, which means that the slopes are different across different groups. Therefore, B is a Random intercepts and slopes model. Lets denote intercept variable as α, and slope variable as β The distribution of model A: $Y \sim N(\alpha_{j[i]} + \beta x_i, \sigma_y^2)$ The linear model of A: $Y = \alpha_{j[i]} + \beta x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma_y^2)$ The distribution of model B: $Y \sim N(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2)$ The linear model of B: $Y = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma_y^2)$

b) (5 points) How will you code these two models in R? Write the code here (without performing it)

assuming that you do not have any group-level covariates.

model_a <- lmer(formula = y ~ (1|group) + x, data=data_a)

model_b <- lmer(formula = y ~ ((1+x)|group) + x, data=data_b)

Question 4: Factor Analysis and Principal Component Analysis.

a) (10 points) Suppose that the factor loading matrix is equal to

$$\Lambda = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & -0.3 \\ 0.6 & -0.4 \\ 0.3 & 0.7 \\ 0.4 & 0.6 \end{bmatrix}$$

Compute the covariance matrix of your data (i.e., Σ_X), excluding the diagonal matrix. Explain all of your steps.

The covaraince matrix of my data excluding the diagonal matrix is $\Lambda\Lambda'$.

X = u + Lf + e

u is the mean of population, L is the loding matrix lambda, f is the explanatory variables and e is vector of specific factors.

Var(X) = E[(X-u)(X-u)'] = E[(Lf +e)(Lf+e)'] = E[LFF'L'] + E[LFe'] + E[eF'L'] + E[ee']

```
= LE[FF']L' + LE[Fe'] + E[eF']L' + E[ee']
```

Since E[Fe'] = 0 and E[eF'] = 0 and E[FF'] = I, which is identity matrix, we have Var(X) = LL' + E[ee']

If we exclude the diagonal matrix E[ee'] here, we have LL', which is $\Lambda\Lambda'$.

The λ' is the transpose of λ. We can acquire it by simply change first column to first row and second column to second row.

$\Lambda\Lambda'$ is symmetric. The matrix multiplication is operated by product and sum between rows and columns. $M_1 2$ of new matrix is acquired by mutiplying the element of first row of λ with corresponding element of the second column of λ' and sum each result of the multiplication.

$\Lambda\Lambda'$ = [0.50, 0.10, 0.10, 0.50, 0.50

```
0.10, 0.34, 0.42, -0.06, 0.02

0.10, 0.42, 0.52, -0.1, 0

0.50, -0.06, -0.10, 0.58, 0.54

0.50, 0.02, 0, 0.54, 0.52]
```

b) (10 points) Now, suppose that the diagonal elements of your covariance matrix are equal to $diag = (0.8730926, 1, 0.8489338, 1, 0.7080494)$. Estimate the diagonal elements of the error covariance matrix, Ψ. Explain all of your steps.

The covariance matrix is $\Lambda\Lambda'$ + E[ee'] as we explanied above.

The diagonal of Ψ := covariance matrix - $\Lambda\Lambda' = diag(0.8730926, 1, 0.8489338, 1, 0.7080494)$ -

```
diag([0.50, 0.10, 0.10, 0.50, 0.50
      0.10, 0.34, 0.42, -0.06, 0.02
      0.10, 0.42, 0.52, -0.1, 0
      0.50, -0.06, -0.10, 0.58, 0.54
      0.50, 0.02, 0, 0.54, 0.52])
  =  [0.3730926, 0, 0, 0, 0
      0, 0.66, 0, 0, 0
      0, 0, 0.3289338, 0, 0
      0, 0 ,0 , 0.42, 0
      0, 0, 0, 0, 0.1880494]
```

c) (10 points) Suppose now that the factor loading matrix is equal to

$$\Lambda = \begin{bmatrix} 0.144 & 0.692 \\ 0.583 & 0.023 \\ 0.721 & -0.006 \\ -0.133 & 0.750 \\ 0.006 & 0.721 \end{bmatrix}$$

Compute the covariance matrix of the data, Σ_X, excluding the diagonal. Explain why you get this result in terms of factor analysis theory.

[0.4996 0.099868 0.099672 0.499848 0.499796

0.099868 0.340418 0.420205 -0.060289 0.020081

0.099672 0.420205 0.519877 -0.100393 0

0.499848 -0.060289 -0.100393 0.580189 0.539952

0.499796 0.020081 0 0.539952 0.519877]

We can see each attribute x_i as a linear combination of two factors in λ. For example, $x_1 = \alpha_1 + 0.144F_1 + 0.692F2 + e_1$ and etc. And we only have 2 vectors here, namely F_1, F_2. The deduction is same as what we have in sub question (a).

I calculate this matrix in the same step as above using the λ. Since the covariance of the data does not change if the data we pick are fixed, the results are same if we approximate this one. The loading matrix here is different from above as we may have multiple ways to factorize the current data into explanatory factors, so the loading matrix changes while the covriance matrix stays the same.

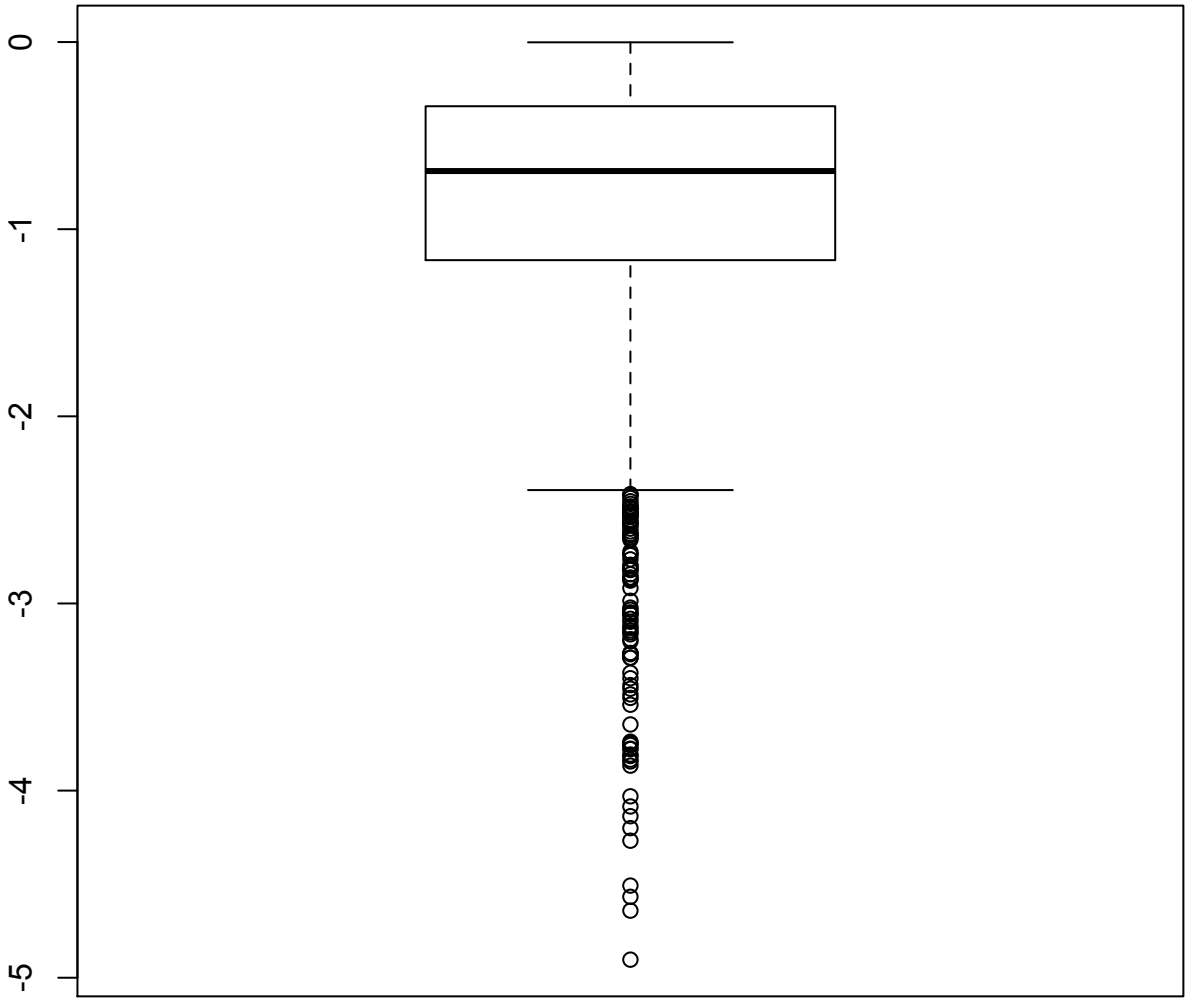


Figure 1: Boxplot

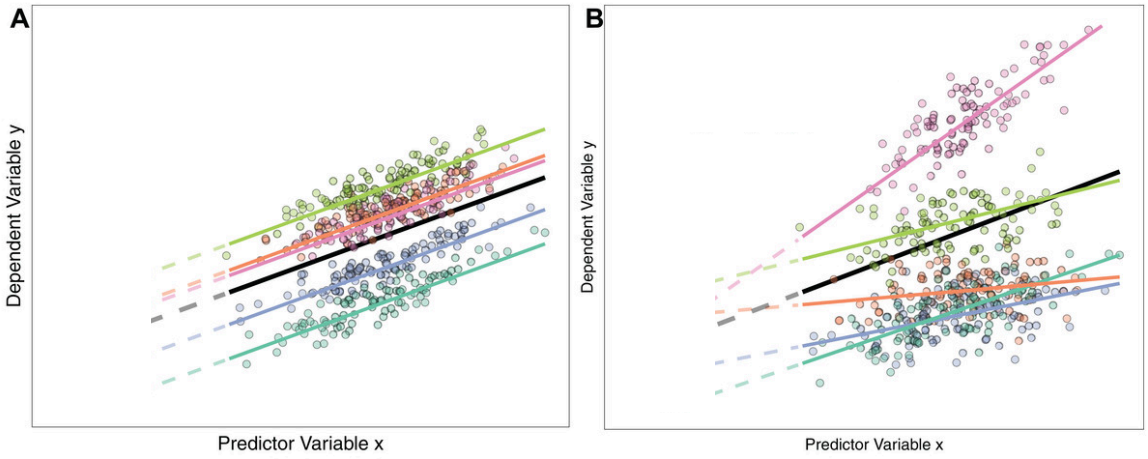


Figure 2: Multilevel models