STSB6816 Test 2 of 2023

Mathematical Statistics and Actuarial Science; University of the Free State

2023/05/18

Time: 180 minutes; Marks: 50

MEMORANDUM

Instructions

- Answer all questions in a single R Markdown document. Please knit to PDF or Word at the end and submit both the PDF/Word document and the ".Rmd" file for assessment, in that order.
- Label questions clearly, as it is done on this question paper.
- All results accurate to about 3 decimal places.
- Show all derivations, formulas, code, sources, and reasoning.
- Intervals should cover 95% probability unless stated otherwise.
- No communication software, devices, or communication capable websites may be accessed prior to submission. You may not (nor even appear to) attempt to communicate or pass information to another student.

Introduction

The data is provided at https://ufs.blackboard.com. **It consists of the following columns: Response_ID, Respondent_Name, Respondent_ID, Year, Response_Text, Response_Numeric.**

A pastor is tracking their congregation's views on a particular matter. To do this they set up a survey where they ask whether people agree with a statement on that viewpoint using a 7 point Likert scale (Strongly Disagree, …, Strong Agree). The survey is sent out year after year for a few years and then the time comes to analyse it. The pastor has some concerns regarding the data and needs your help.

- It seems that mostly the responses come from the same people year after year (mostly the choir).
- They can't decide whether to model the responses as ordinal data using a categorical distribution, interval data using binomial distribution, or numeric using a normal distribution.

He asks you to build and compare these three as mixed effects models, each with Year as a linear fixed effect on the underlying scale and Respondent as a random intercept effect. Then he wants you to use the best model to determine whether the people are agreeing more with the statement.

Question 1

1.1) Explain why Respondent should be included in the model as an effect **at all**. **[3]**

The critical issue is that we have different numbers of observations per respondent (unbalanced) [1]. Respondent observations are likely correlated [1]. So respondents with lots of extreme responses can overly bias the results with respect to a typical future respondent [1].

1.2) Explain why Respondent should be included in the model as **random** effect. **[2]**

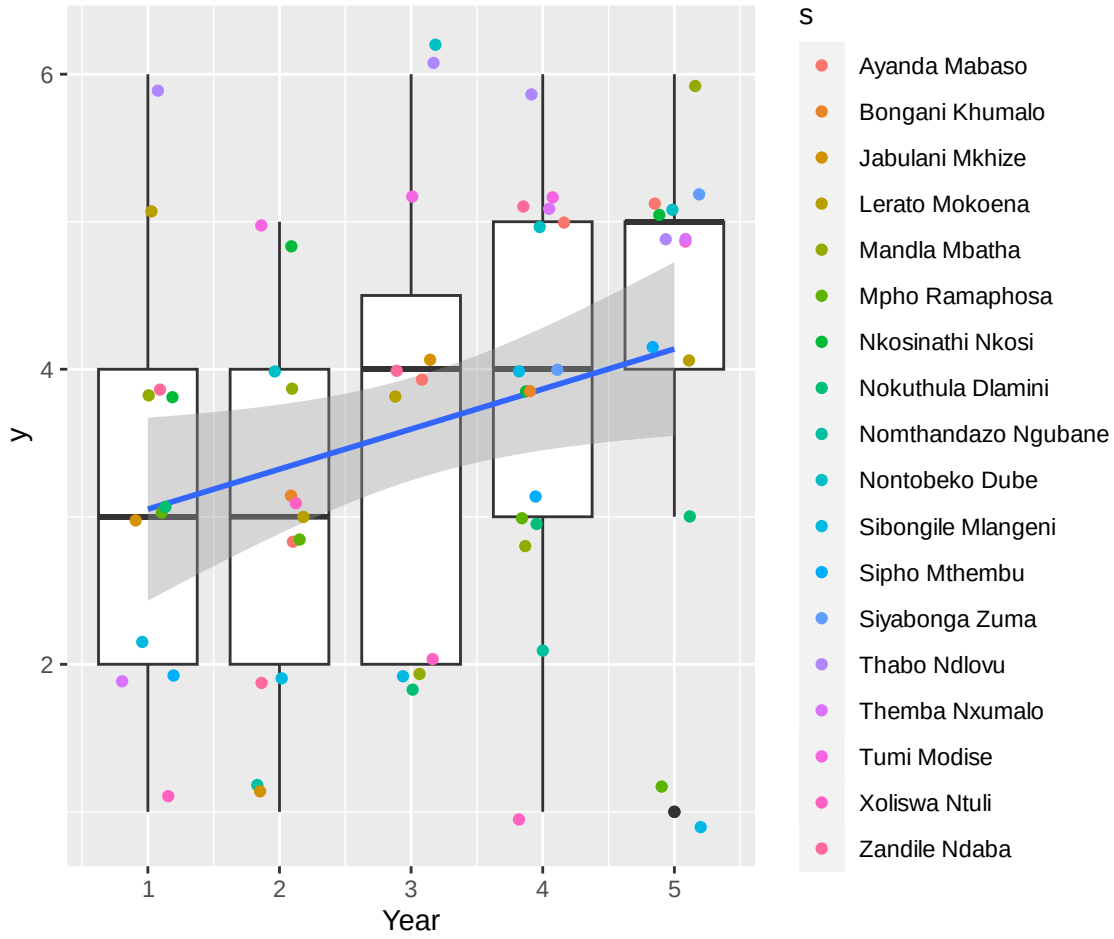
We are not interested in the views of specific respondents, but rather the views of a general congregant [2].

1.3) Explain what including Respondent only as an intercept term implies with regard to the assumed slopes of each congregant over time. **[2]**

Since the model does not specify a different slope for each congregant, we are assuming that they all have the same slope over time [2].

1.4) Import the data set into R and explore it visually. You could use a box plot with Year on the x axis perhaps. Discuss what you see. **[4]**

```
"STSB6816Test2Data2023.xlsx" |> openxlsx::read.xlsx("TestData") -> d

library(tidyverse)

data.frame(Year = d$Year, y = d$Response_Numeric, s = d$Respondent_Name) |>
  ggplot(aes(x = Year, y = y)) +
  geom_boxplot(aes(group = Year)) +
  geom_smooth(method = 'lm', formula = 'y~x') +
  geom_jitter(aes(colour = s), width = 0.2, height = 0.2)
```

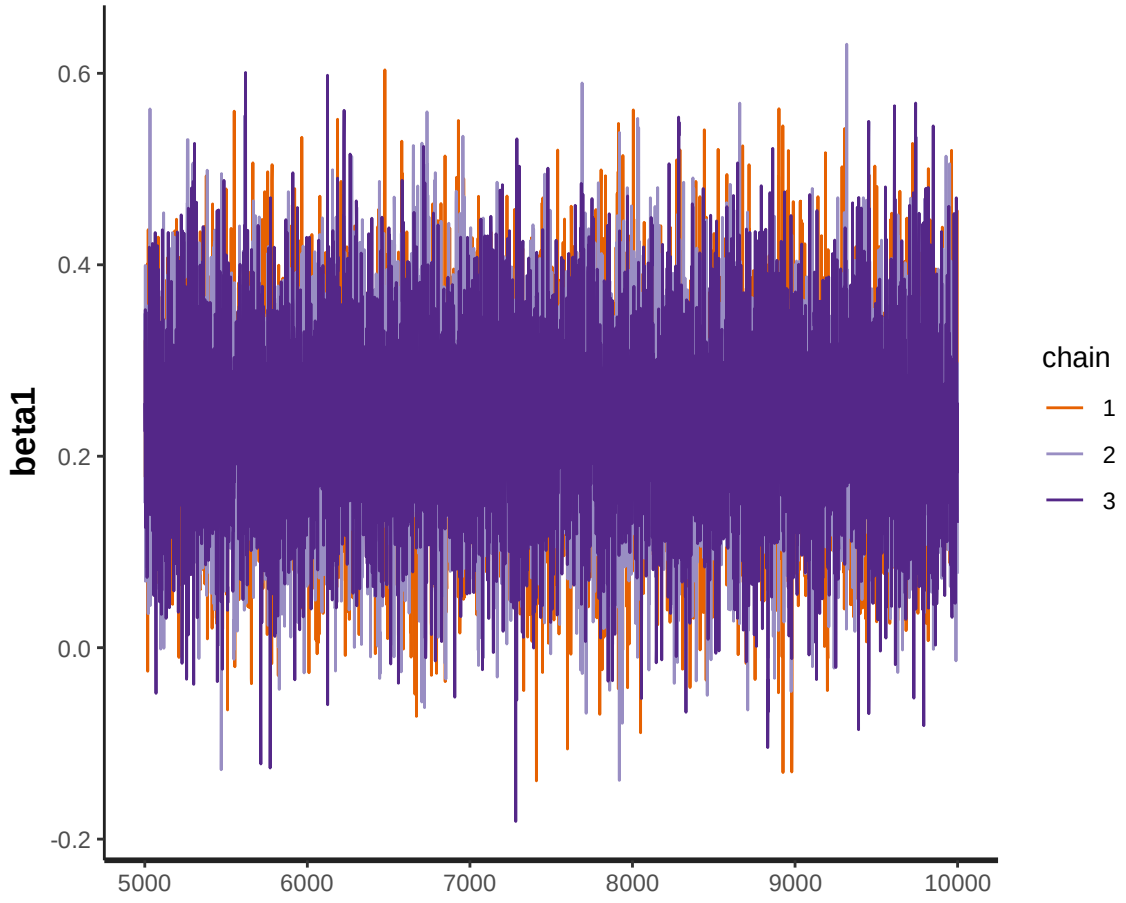


Loading data [1], box plot [1], and discussion saying something about a slight upward trend that may or may not be signficant - significance cannot be determined yet [2].

1.5) Fit a standard mixed effects model assuming that the numerically encoded responses follow a conditional normal distribution given the year number as a continuous linear predictor and respondent as a random intercept. Summarise the distribution of the coefficient of the year number. **[8]**

```
library(rstan)
mycores <- 3
options(mc.cores = mycores)

data {
  int n;
  vector[n] y;
  vector[n] x;
  int n_s;
  int subj_ind[n];
}
parameters {
  real beta0;
  real beta1;
  real<lower=0> sigma;
  real z[n_s];
  real<lower=0> tau;
}
transformed parameters {
  vector[n] mu;
  for (i in 1:n) {
```

```
    mu[i] = beta0 + beta1*x[i] + z[subj_ind[i]];
  }
}
model {
  y ~ normal(mu, sigma);
  z ~ normal(0, tau);
  target += -2*log(sigma) - 2*log(tau);
}
generated quantities {
  vector[n] log_lik;
  for (i in 1:n) {
    log_lik[i] = normal_lpdf(y[i] | mu[i], sigma);
  }
}

saveRDS(SimpleRE, file = 'SimpleRE.Rds')

Model1Fit <- sampling(SimpleRE,
                      list(n = nrow(d),
                           x = d$Year,
                           y = d$Response_Numeric,
                           n_s = max(d$Respondent_ID),
                           subj_ind = d$Respondent_ID),
                      iter = 10000,
                      chains = mycores)

pars_of_interest <- c('beta1')
Model1Fit |> traceplot(pars = pars_of_interest)
```



```
summary(Model1Fit, pars = pars_of_interest)$summary |> kable(digits = 3)
```

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
beta1	0.237	0.001	0.1	0.043	0.17	0.237	0.304	0.432	5037.645	1.001

Specifying the fixed effect model components correctly, including the data, parameters, and model components relating to ordinary regression [2]. Specifying the random effects model components correctly, including the data, parameters, and model components relating to random effects [3]. Implementing the model correctly using the provided data and giving a sensible summary of the key parameter [3]. Note that the generated log_lik is not required here and the associated marks full under a later question.

1.6) According to the above model, what is the probability that the agreement with the viewpoint is increasing by over 0.05 steps per year (i.e., $P[\beta_1 > 0.05]$)? **[3]**

```
Model1Sims <- rstan::extract(Model1Fit)
cat('\n\n$P[\\beta_1>0.05]=$', round(mean(Model1Sims$beta1 > 0.05),3), '\n\n')
```

$P[\beta_1 > 0.05] = 0.971$

Extracting the simulations [1]. Sensible calculation of the probability [2].

1.7) Adapt the mixed effects model to assume that the numerically encoded responses follow a conditional **binomial** distribution given the year number as a continuous linear predictor and respondent as a random intercept on the logistic scale. Fit the model and summarise the distribution of the coefficient of the year number. **[5]**

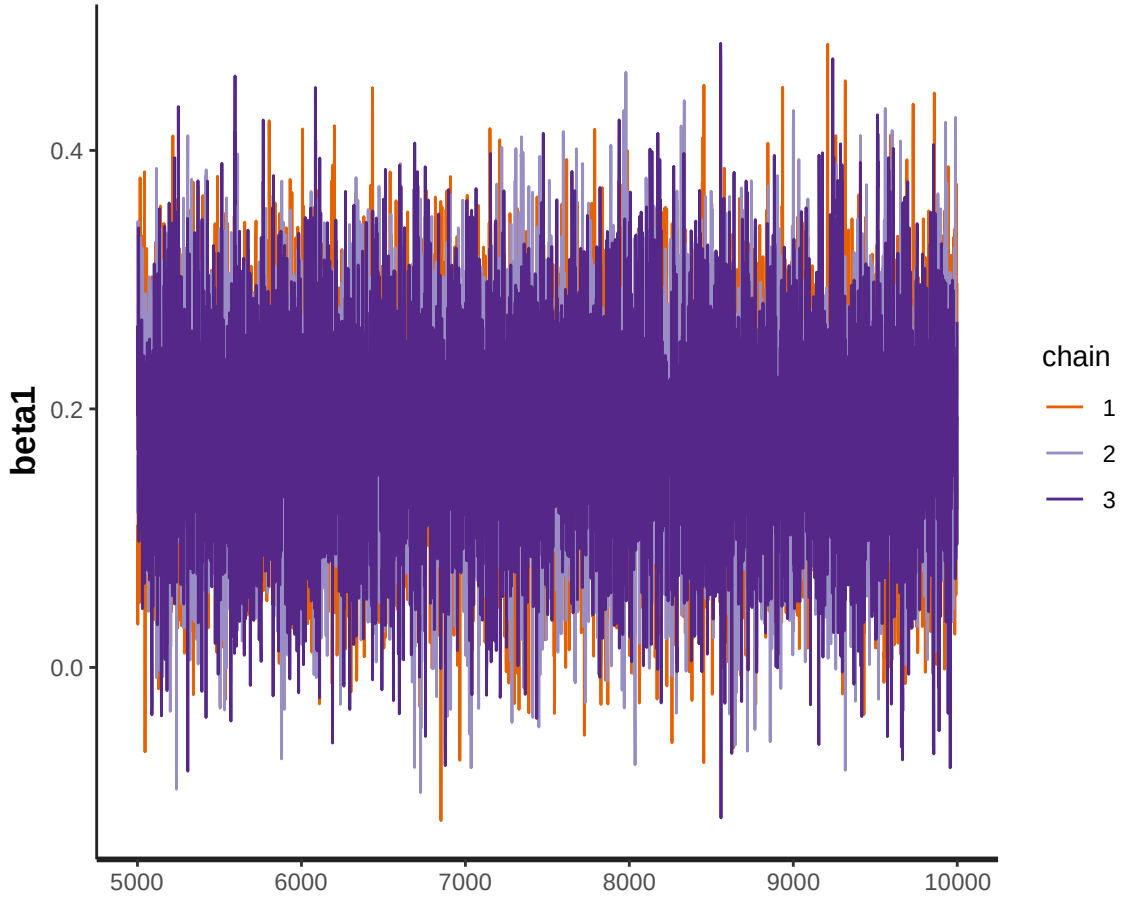
Hint: The likelihood component of the model can be expressed mathematically as

$$\begin{aligned} y_i &\sim binomial(6, \pi_i)\ i = 1 \ldots n \\ \pi_i &= logit^{-1}(\beta_0 + \beta_1 * x_i + z_{r_i}) \end{aligned}$$

where x is the year number and r is the respondent number.

```
data {
  int n;
  int y[n];
  vector[n] x;
  int n_s;
  int subj_ind[n];
}
parameters {
  real beta0;
  real beta1;
  real z[n_s];
  real<lower=0> tau;
}
transformed parameters {
  vector<lower=0,upper=1>[n] mu;
  for (i in 1:n) {
    mu[i] = inv_logit(beta0 + beta1*x[i] + z[subj_ind[i]]);
  }
}
model {
  y ~ binomial(6, mu);
  z ~ normal(0, tau);
  target += -2*log(tau);
```

```
}
generated quantities {
  vector[n] log_lik;
  for (i in 1:n) {
    log_lik[i] = binomial_lpmf(y[i] | 6, mu[i]);
  }
}

saveRDS(BinomialRE, file = 'BinomialRE.Rds')

Model2Fit <- sampling(BinomialRE,
                      list(n = nrow(d),
                           x = d$Year,
                           y = d$Response_Numeric,
                           n_s = max(d$Respondent_ID),
                           subj_ind = d$Respondent_ID),
                      iter = 10000,
                      chains = mycores)

Model2Fit |> traceplot(pars = pars_of_interest)
```




```
summary(Model2Fit, pars = pars_of_interest)$summary |> kable(digits = 3)
```

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
beta1	0.183	0.001	0.081	0.024	0.128	0.182	0.236	0.344	10350.76	1

Changing y to integer, dropping sigma, introducing inverse logit, and changing normal to binomial [4*1=4]. Implementing the model correctly using the provided data and giving a sensible summary of the key parameter [1].

1.8) Consider again the rate at which the agreement with the viewpoint is increasing per year (i.e., β_1), estimate the probability that this parameter differs between the models by more than 0.01 $\left(P\left[\left|\beta_1^{Model2} - \beta_1^{Model1}\right| > 0.01\right]\right)$. Also give a short statement (1 sentence) about what the calculated probability implies (if anything). **[4]**

```
Model2Sims <- rstan::extract(Model2Fit)
cat('\n\n$P\\left[\\left|\\hat{\\beta_1}^{Model2}-\\hat{\\beta_1}^{Model1}\\right| > 0.01\\right]=$', round(mean(abs(Model2Sims$beta1 - Model1Sims$beta1) > 0.01), 3), '\n\n')
```

$$P\left[\left|\widehat{\beta_1}^{Model2} - \widehat{\beta_1}^{Model1}\right| > 0.01\right] = 0.944$$

Extracting new simulations [1]. Sensible calculation of the probability by comparing simulations [2]. Statement saying that a high probability suggests disagreement between the models. [1]

1.9) Compare the two models using a criterion that considers model complexity and give a conclusion as to which model appears to offer a superior fit. Examples of acceptable criteria are LOOIC, DIC, and Bayes Factors, as well as variants of these. **[6]**

```
fits <- list(Normal = Model1Fit, Binomial = Model2Fit)

library(loo)
fits |> lapply(\(fit) {extract_log_lik(fit, merge_chains = FALSE)}) -> log_lik
log_lik |> lapply(\(ll) {relative_eff(exp(ll), cores = 1)}) -> r_eff
fits |> length() |> seq_len() |>
  lapply(\(i) {loo(log_lik[[i]], r_eff = r_eff[[i]], cores = 1)}) |>
  loo_compare() -> comparison
rownames(comparison) <- names(fits)[order(rownames(comparison))]
comparison |> knitr::kable(digits = 1)
```

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
Binomial	0.0	0.0	-102.2	4.2	11.5	1.6	204.3	8.3
Normal	-3.7	1.7	-105.9	5.5	14.2	2.0	211.7	11.0

Calculating a suitable statistic [4]. Statement saying that the (correctly identified) model with the lowest criterion value is preferred [2]. *Note that most of these marks are for the code that calculates the log likelihoods, either in Stan or in R, not for the last bit of code above to get the statistics.*

1.10) Using any of the models, what is the predicted standard deviation of a random future response in Year 6 of a random person who has not previously responded? **[4]**

```
# From Model 1:
nsims <- length(Model1Sims$beta1)
newRE <- rnorm(nsims, 0, Model1Sims$tau)
preds <- rnorm(nsims,
               Model1Sims$beta0 + Model1Sims$beta1*6 + newRE,
               Model1Sims$sigma) |>
  round(digits = 0)
preds <- pmin(pmax(preds, 0), 6)
sd(preds)
```

```
[1] 1.343505
```

```r
# From Model 2:
nsims <- length(Model2Sims$beta1)
newRE <- rnorm(nsims, 0, Model2Sims$tau)
preds <- rbinom(nsims, 6,
                plogis(Model2Sims$beta0 + Model2Sims$beta1*6 + newRE))
sd(preds)
```

```
[1] 1.37176
```

Predicting new random effects [1]. Incorporating Year 6 into the linear equation [1]. Predicting new ratings [1]. Calculating standard deviation [1].

1.11) Adapt the first mixed effects model to assume that the responses follow a conditional **ordered logistic** distribution given the year number as a continuous linear predictor and respondent as a random intercept on the logistic scale. This is also known as *ordinal regression*. Fit the model and summarise the distribution of the coefficient of the year number. **[5]**

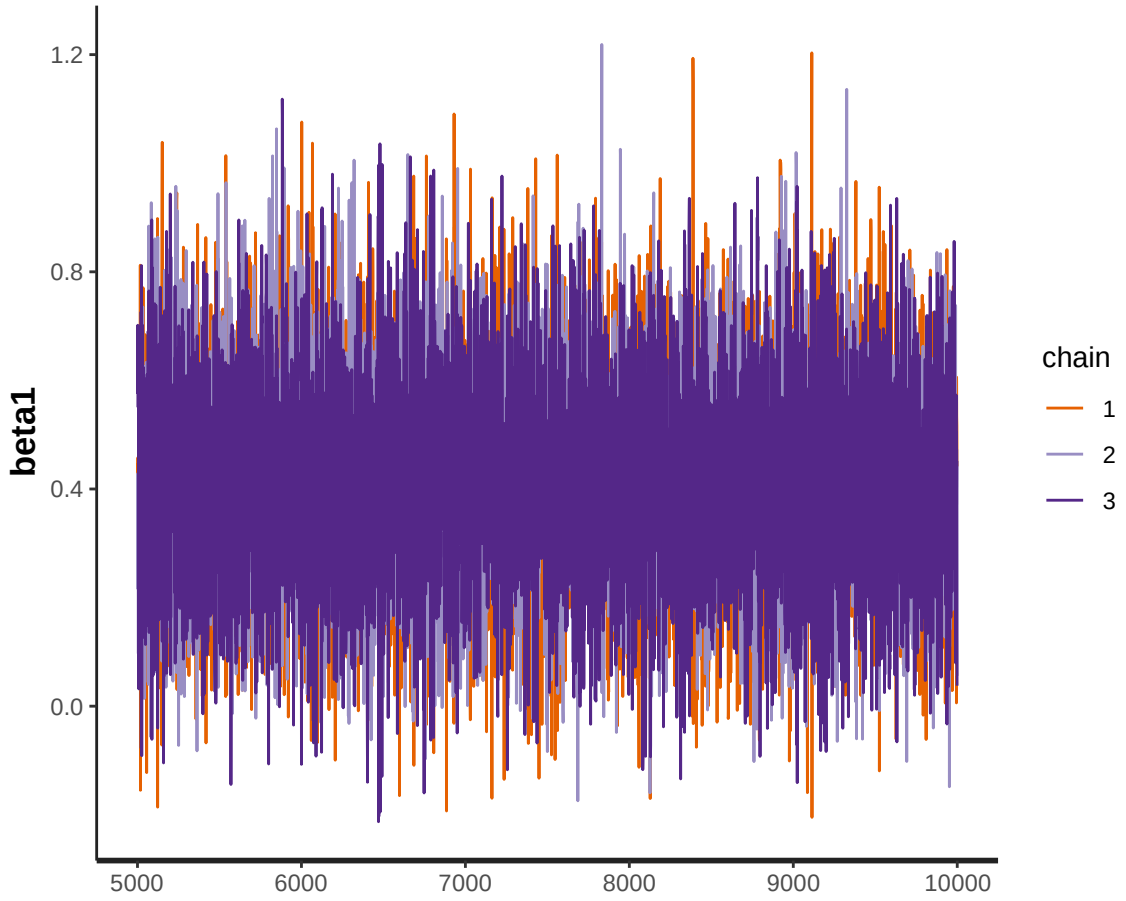
Hint: It is critical that a strict prior be placed on the thresholds (e.g. N(0,10)). The key components of the model can be expressed mathematically as

$$\begin{aligned} y_i &\sim ordered\ logistic(\mu_i, \boldsymbol{\theta})\ i = 1 \ldots n \\ \mu_i &= \beta_0 + \beta_1 * x_i + z_{r_i} \\ \theta_1 &< \theta_2 < \theta_3 < \theta_4 < \theta_5 \sim truncN(0,10) \end{aligned}$$

where x is the year number, r is the respondent number, and $\boldsymbol{\theta}$ is the set of thresholds.

```stan
data {
  int n;
  int y[n];
  vector[n] x;
  int n_s;
  int subj_ind[n];
}
parameters {
  real beta0;
  real beta1;
  real z[n_s];
  real<lower=0> tau;
  ordered[5] thresholds;
}
transformed parameters {
  vector[n] mu;
  for (i in 1:n) {
    mu[i] = beta0 + beta1*x[i] + z[subj_ind[i]];
  }
}
model {
  y ~ ordered_logistic(mu, thresholds);
  z ~ normal(0, tau);
  target += -2*log(tau);
  thresholds ~ normal(0, 10);
}
generated quantities {
  vector[n] log_lik;
  for (i in 1:n) {
    log_lik[i] = ordered_logistic_lpmf(y[i] | mu[i], thresholds);
  }
}
```

```
saveRDS(CatRE, file = 'CatRE.Rds')

Model3Fit <- sampling(CatRE,
                      list(n = nrow(d),
                           x = d$Year,
                           y = d$Response_Numeric,
                           n_s = max(d$Respondent_ID),
                           subj_ind = d$Respondent_ID),
                      iter = 10000,
                      chains = mycores,
                      control = list(max_treedepth = 12))

Model3Fit |> traceplot(pars = pars_of_interest)
```



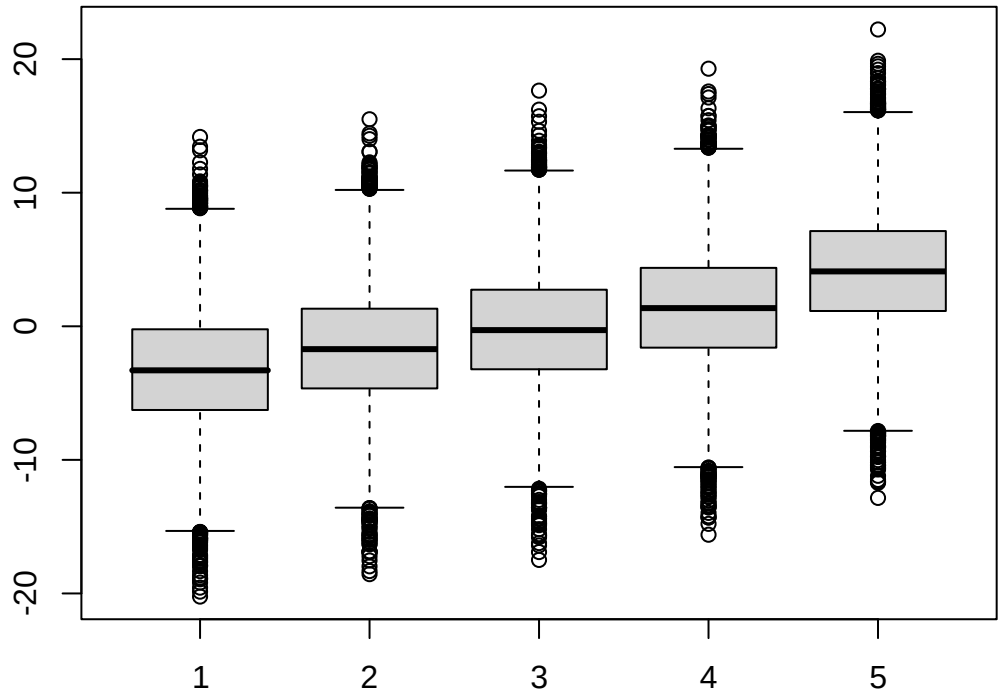

```
summary(Model3Fit, pars = pars_of_interest)$summary |> kable(digits = 3)
```

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
beta1	0.412	0.001	0.186	0.051	0.287	0.41	0.534	0.785	15903.73	1

Changing y to integer, replacing sigma with threshold vector, and changing normal to ordered logistic [3]. Implementing the model correctly using the provided data and giving a sensible summary of the key parameter [2].

1.12) Illustrate or estimate, and then analyse, the thresholds between response options as suggested by the ordinal regression. Are they evenly spaced (as the normal model assumes)? How do they relate to the thresholds implied by the binomial model?**[4]**

```
Model3Fit |> rstan::extract() -> postsims
boxplot(postsims$thresholds)
```




```
colMeans(postsims$thresholds)
```

```
| [1] -3.2531926 -1.6734577 -0.2445716  1.3904777  4.1336325
```

```
qlogis((1:5)/6)
```

```
| [1] -1.6094379 -0.6931472  0.0000000  0.6931472  1.6094379
```

Giving estimates or illustration of thresholds [2]. Saying they are not quite evenly spaced [1]. Saying that they are similar to the binomial thresholds but more flexible [1].

Points total

The points on the test add up to **50**
