36-663 Project Part 1

Mengran He

18 November, 2016

1. Exploratory Data Analysis

There are 414 players and 20 unique questions. According to figure 1a, there are more than 132 players who try to answer all 20 questions and there are more than 65 players trying to answer 2 or less questions. Furthermore, figure 1b represents the number of players categorized by partial IP (organizations ip2 and computers within organizations ip3). In particular, each player can be uniquely identified by ip because ip represents the web address which denotes a specific computer, while ip2 and ip3 are group-level variables which are composed by players. Figure 1b shows that among 229 distinct organizations ip2, there are 219 organizations containing 5 or less players and 1 organization containing 39 players. Moverover, there are 253 distinct groups of computers ip3. According to figure 1c, there are 248 groups of computers, each has 5 or less players and 1 group has about 40 players.

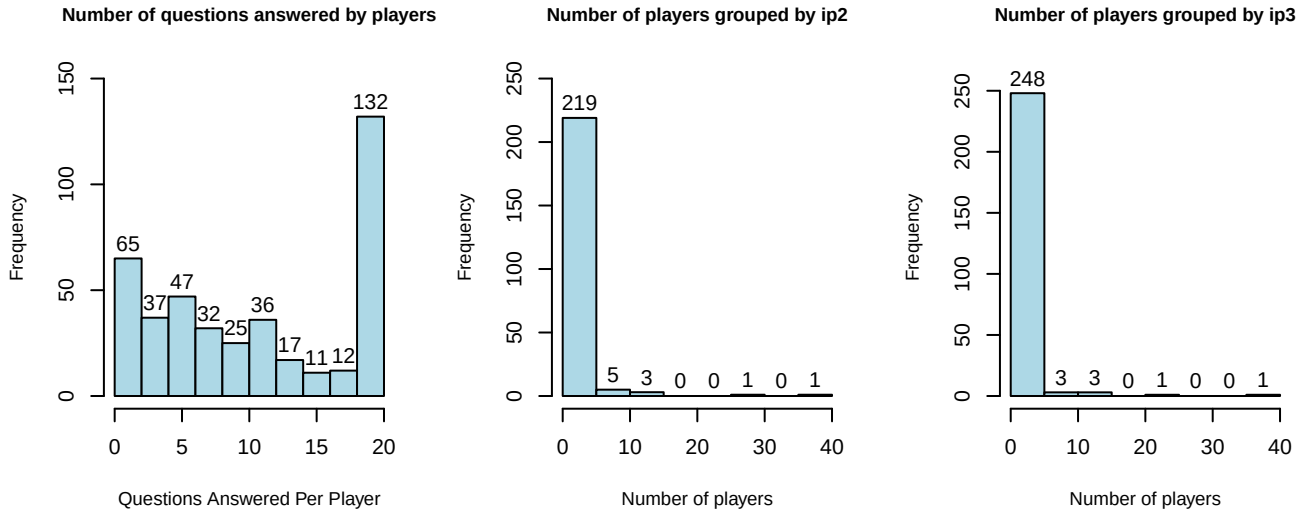


Figure 1: a) Histogram of questions answered by 414 players, b) Histogram of players grouped by organizations, c) Histogram of players specified by groups of computers

As shown in figure 2, there are 15 distinct experiments. According to the blue color barplot, there are 4 experiments answered by less than 30 players, 5 experiments answered by 30 players, and 6 experiments answered by more than 30 players. In addition, based on pink color barplot, each experiment contains all 20 types of questions.

Table 1 summarizes the distribution of proportion-correct scores for 414 players. Specifically, among 414 players, the lowest percentage of correct answers is 0% and the highest percentage value is 100%. The median is 15.38%, average value of percentage rate is 21.77%, and the variance for the percentage rage is 0.05. In addition, 9.66% of players, namely 4 out of 414 players (SID: 161828,162085,162485,163315) have 100% of correct response.

Table 1: Summary of proportion of correct scores

Min	1stQu.	Median	Mean	3rdQu.	Max	Var
0	0	0.1538	0.2177	0.3649	1	0.0515

After calculating the average value of reaction times for each players, we have figure 3a describing the distribution of average of reaction times across players. In particular, there are about 204 players who have average reaction times within 2 to 4 seconds, about 139 players who have average value of reaction times within 4 to 6 seconds, and 1 player who has average reaction time about 12 seconds. Then, we have the histogram of average reaction times across 20 questions in figure 3b. There are 5 questions which reflect the average value of reaction times within 3.7 to 3.75 seconds. Regardless of

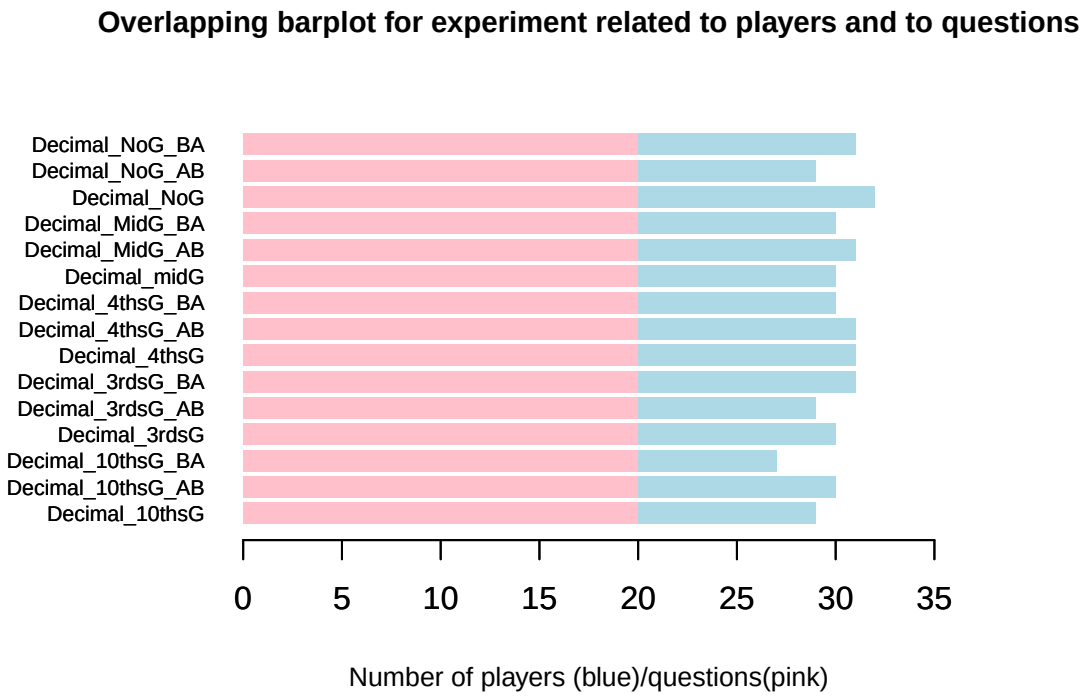


Figure 2: Overlaying Barplot represents the relationship between experiments and players or questions

the players and questions, the grand mean of reaction times is 3.733 seconds.

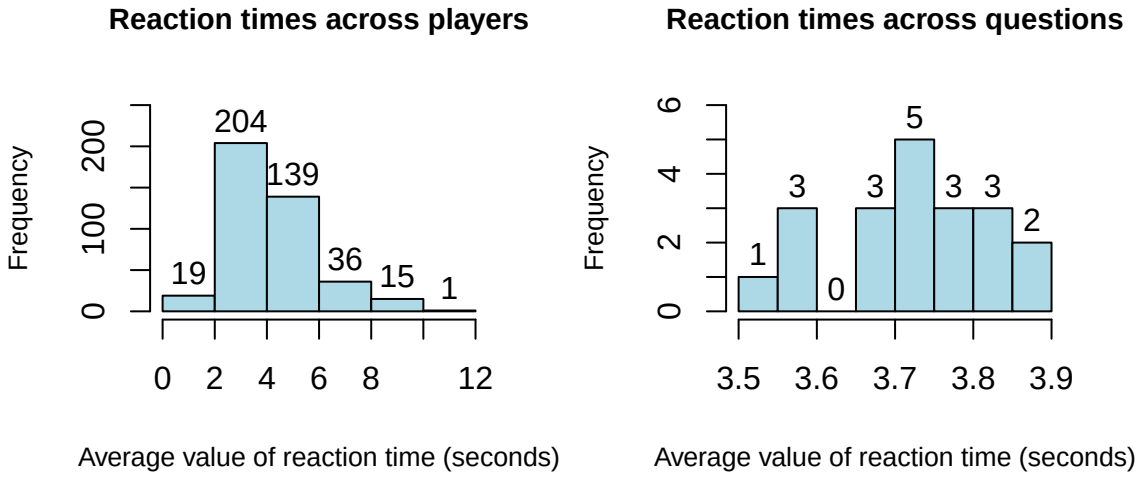


Figure 3: a) Histogram of reaction time across players, b) Histogram of reaction time across questions

2. Logistic regression for question difficulty

(a) Fit logistic regression model

The logistic regression model of predicting the probability that a question will be answered correctly given currentQuestion as predictor is summarized in model 1 without intercept: $resp_i = \beta_{1i} currentQuestion_i + \epsilon_i$. Model 1 suggests that all levels of questions are statistically significant, except question0.5 with P-value less than α. Figure 4 represents the marginal model plot for model 1, the blue line of data and red line of model are on top of each other, so model 1 predicts quite well. Furthermore, table 2 summarizes the distribution of predicted probability of answering question correctly based on model 1. It shows that the minimum of predicted probability of getting answer right is 19.53%, the maximum of predicted probability is 46.3%, the average of predicted probability is 27.46% and the variance of these predictions is 0.004. If we don't remove intercept, the point estimate of intercept β_0 means the expected value of $log(\frac{p}{1-p})$ when x = 0. However, the

predictor currentQuestion is a categorical variable which has 20 levels, so it doesn't make any sense if currentQuestion = 0. If we take out the intercept, there are 20 estimated coefficients in the model, representing 20 levels of currentQuestions individually.

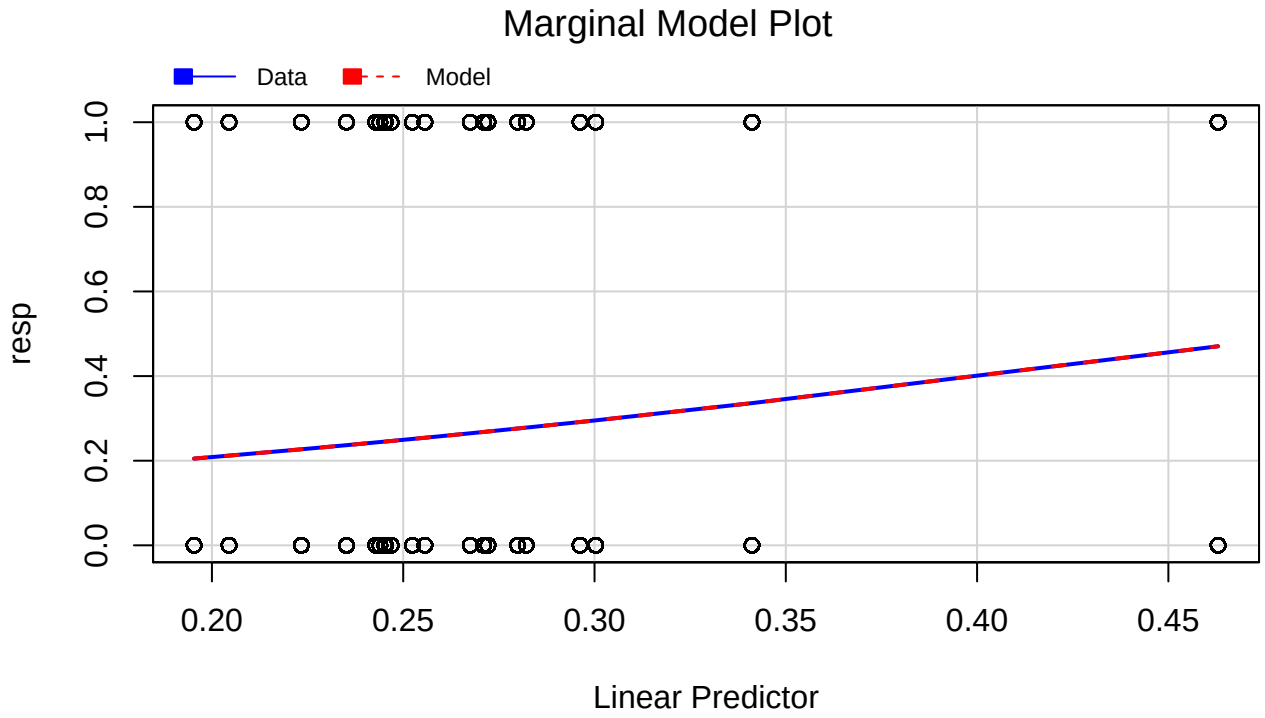


Figure 4: Marginal Model plot for model 1

Table 2: Summary of predicting probability of getting correct answer

	min	max	median	mean	var
Probability	0.1953	0.463	0.2557	0.2746	0.004061

(b) Plot the coefficients against the fraction of participants who got the corresponding question right

The plots of coefficients against the fraction of participants who got the corresponding question right and coefficients against the logit of the fraction are generated in figure 5. Figure 5b, coefficients vs logit of fraction is better than figure 5a. The estimated coefficients β_i are generated from the logistic regression model given types of questions as predictors, β_i represents how much the logit of fraction changes if player shifts from one question to next question. Therefore, the plot of coefficients against the logit of the fraction of correct answer is more suitable for presenting the linear relationship between response and currentQuestion as shown in figure 5b.

(c) Describe what methods are used to find the best model and interpret that model.

In order to predict the probability that a question will be answered correctly, we chose five predictors from the dataset which are currentQuestion, experimentName, levelName, ip2, and ip3. By comparing the values of deviance and AIC, we found that the logistic model modelexip3: $resp = \beta_0 + \beta_{1i} experimentName_i + \beta_{2i} ip3_i + \epsilon_i$ has smallest AIC (8489) and deviance score (7911), other candidate models are in appendix II. In addition, the marginal model plot of modelexip3 in figure 6 is relatively better than others (shown in appendix II), since the data line and model line are on top of each other except for large linear predictors (>0.6). Moreover, the binned residual plot of modelexip3 in figure 7a has less points outside of 95% confidence interval than other models(shown in appendix II), but pattern exists in this plot.

Modelexip3 evaluates the question difficulty by analyzing the effects of question types, player × question interaction and group of computers on question response. This model reveals that all the levels of currentQuestion and experimentName

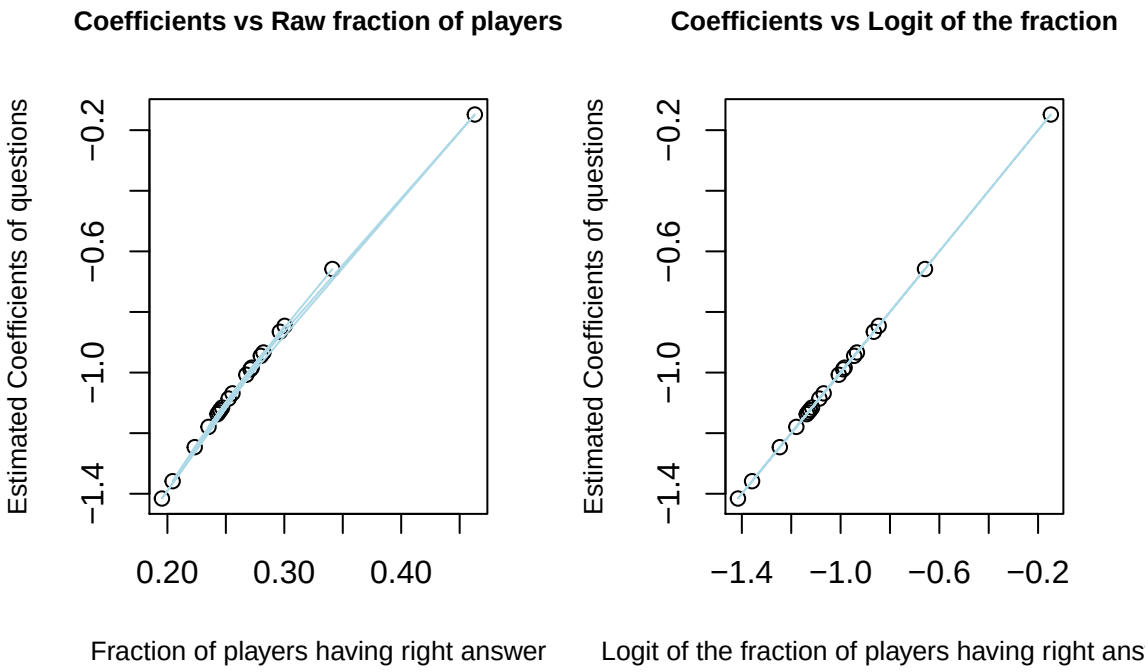


Figure 5: a) Plot of coefficients vs Raw fraction of players, b) Plot of coefficients vs Logit of the fraction

are statistically significant, but the intercept and several levels of ip3 are not significant. In addition, the cook's distance vs leverage plot in figure 7b reveals that there are several influential points in this model which will affect the overall fitting.

3. Logistic regression for player proficiency

(a) Fit logistic regression model giving the probability that a player will provide a correct answer

In order to measure player's proficiency, we fitted a logistic regression model by regressing question response on players' id in model 5: $resp_i = \beta_{1i} SID_i + \epsilon_i$ without intercept. It measures the effect of each player on question response. Because there are 414 distinct player id, this model contains 414 estimated coefficients, among which 129 of them are statistically significant, which means the logit of probability that a player will provide a correct answer won't change much from one player to next player for the rest of 285 players. Figure 8a shows the residual plot of model 5 which looks good for logitic model. Figure 8b indicates that there is a strong pattern and half of them are outside the boundary, even though the range of residual is small, this model is not adquate to fit the data.

(b) Plot the coefficients against the proportion correct for each player.

Figure 9a represents the plot of estimated coefficients against the proportion correct for each player and there are 414 points corresponding to player's id. Figure 9b shows the plot of estimated coefficients against the logit of the proportion correct. There are 288 observations in figure 9b, because the rest of 126 players have 0 correct response and the logit of proportion can't be applied to their fraction of correct response. By comparison, the plot from figure 9b is better, as it shows the linear relationship between logit of proportion correct and coefficients, the coefficients represents the change of logit of probability that a player will provide a correct answer when shifiting from one player to next player. In figure 9a, it shows that there are plenty of points on the upper right and lower left cornors, caused by overfitting and the line in the middle is not straight.

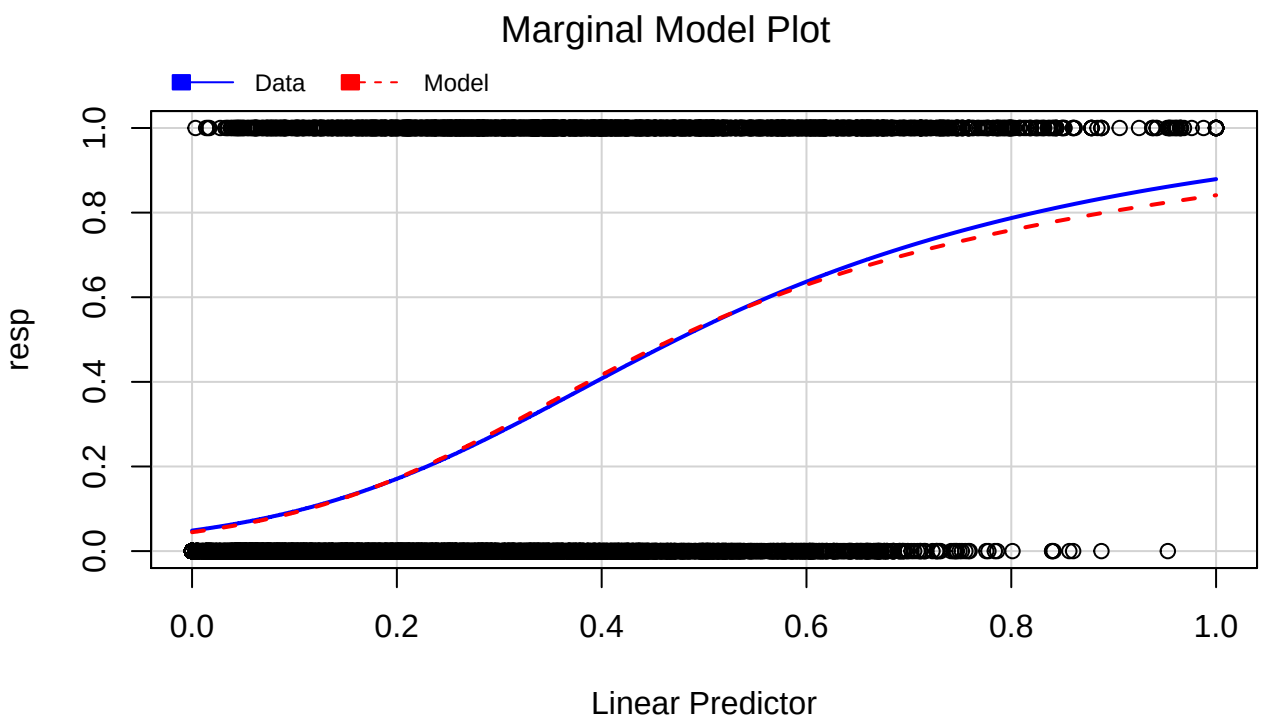


Figure 6: Marginal model plot of modelexip3

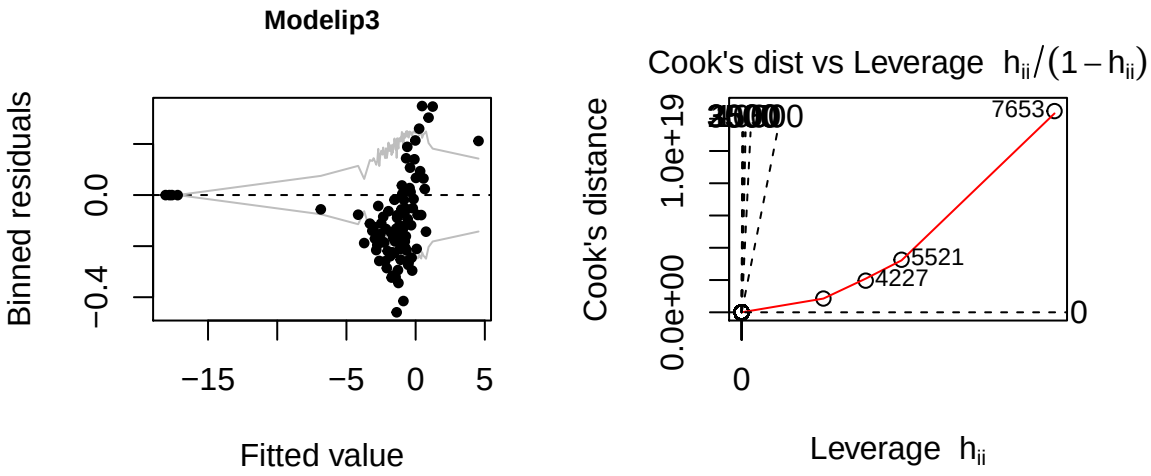


Figure 7: a) Binned residual plot of modelexip3, b) Cooks D vs Leverage plot

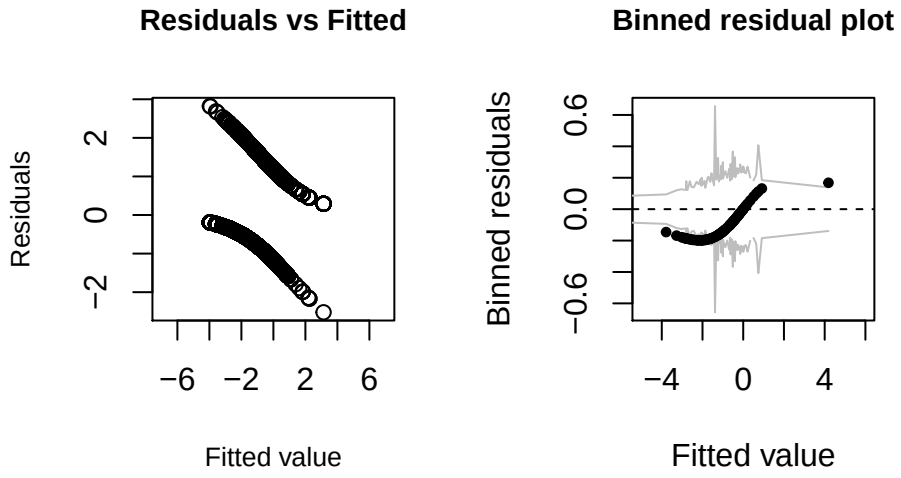


Figure 8: a)Residual plot of model 5, b) Binned residual plot of model 5

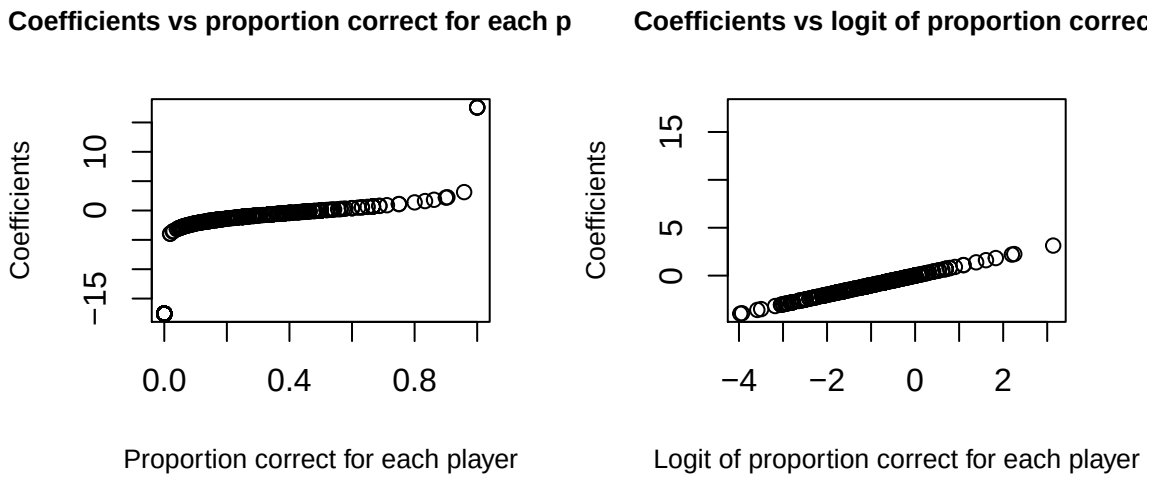


Figure 9: a) Plot of coefficients against proportion correct, b) Plot of coefficients against logit of proportion correct

4. Mixed Effects Models

(a) Fit a mixed effects logistic regression predicting the probability of a correct response.

We generated the mixed effect logistic regression model by assigning currentQuestion as individual-level predictor and SID as group-level variable. Model 6 has two levels, such that level 1: $resp_i = \alpha_{0j[i]} + \alpha_{1i} currentQuestion_i + \epsilon_i$; level 2: $\alpha_{0j} = \beta_0 + \eta_j$. The summary of model 6 indicates that all of fixed effects α_{1i} from currentQuestion are statistically significant. In addition, the randomness of the model prediction is generated by group-level variable SID.

```
glmer(formula = resp ~ questions - 1 + (1 | sid), data = dtf,
    family = binomial)
             coef.est coef.se
questions0.1  -1.02     0.13
questions0.13 -1.44     0.13
questions0.16 -1.63     0.16
questions0.2  -1.61     0.16
questions0.25 -1.37     0.13
questions0.3  -1.64     0.16
questions0.33 -1.69     0.14
questions0.38 -1.91     0.17
questions0.4  -1.48     0.16
questions0.43 -1.62     0.13
questions0.5  -0.32     0.11
questions0.6  -1.51     0.16
questions0.63 -1.51     0.13
questions0.66 -1.71     0.14
questions0.7  -1.72     0.16
questions0.75 -1.35     0.13
questions0.8  -1.66     0.16
questions0.83 -1.39     0.15
questions0.88 -1.93     0.14
questions0.9  -1.29     0.13

Error terms:
 Groups   Name        Std.Dev.
 sid      (Intercept) 1.17
 Residual             1.00
---
number of obs: 8257, groups: sid, 414
AIC = 8613.1, DIC = 7031.7
```

```
deviance = 7801.4
```

The QQplot of random effects in figure 10 suggests that the data on two tails slightly deviates from the normality assumption line. In addition, figure 11 indicates that there is pattern in the binned residual plot and some points are outside of 95% confidence interval. Therefore, model 6 doesn't fit very well.

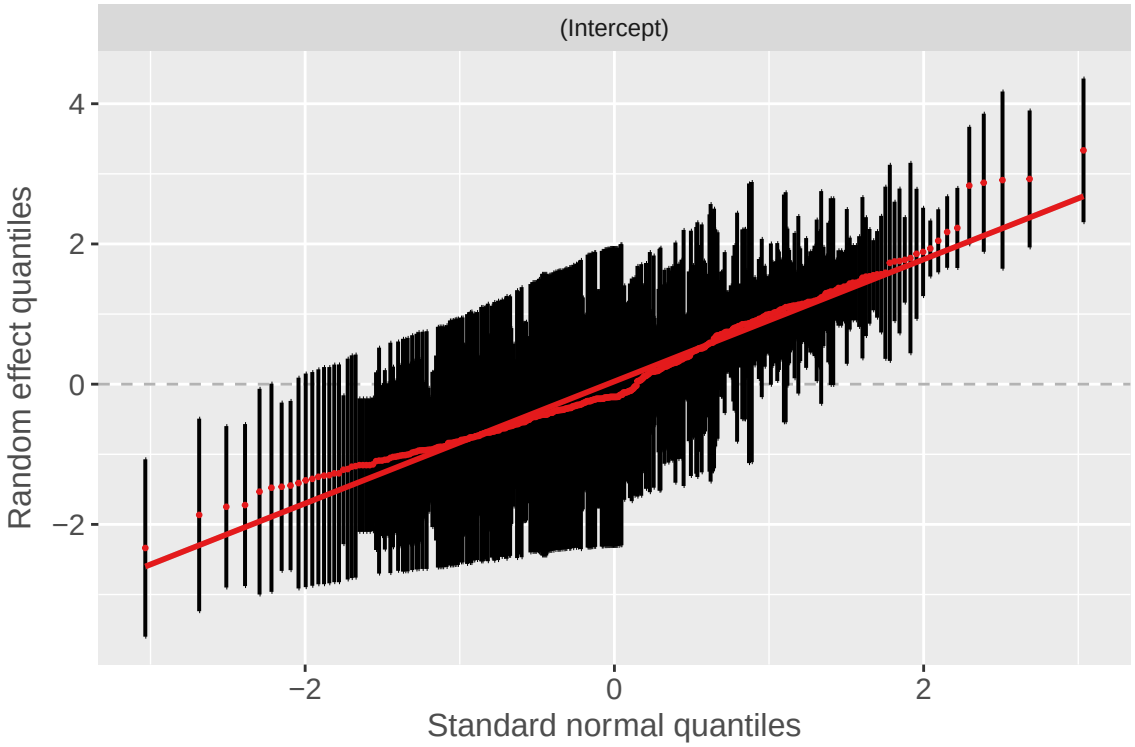


Figure 10: Diagnostic plot of random effects in model 6

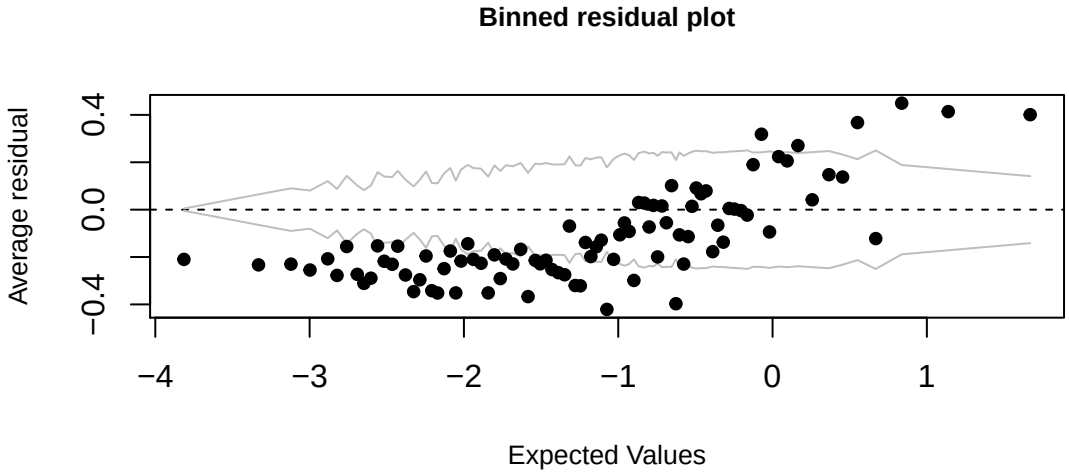


Figure 11: Binned residual plot of model 6

(b) Plot the fixed effects against the fraction of players who got the corresponding question correct or against the logit of that fraction.

Figure 12 expresses the relationship between fixed effects from model 6 and logit of the fraction of players who got the question correct. On the horizontal axis, when the fraction of players who got the question right gets large, its corresponding logit value will also increase. On the vertical axis, the fixed effects represent the effects of 20 types of question on logit of fraction of response answer. So, the coefficients α_{1i} represents the changes in logit of probability of a correct response when player shifts from one question to next question.

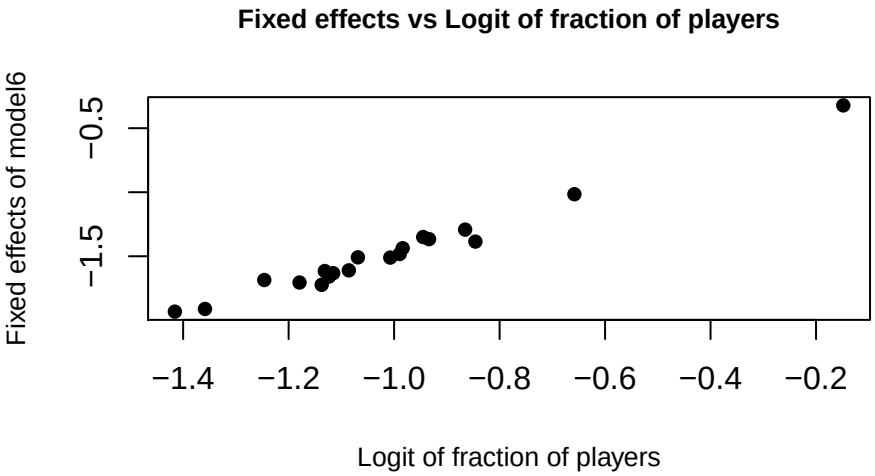


Figure 12: Plot of fixed effects from model 6 vs logit of fraction of players who got question right

(c) Plot the random effects against the proportion correct for each player or against the logit of the proportion correct.

The plot of random effects vs the logit of the proportion correct for each player is shown in figure 13. Random effects η_j in model 6 represents to the randomness in predicting the logit of probability of a correct response, they are drawn from the distribution of players. Figure 13 shows that there is positive relationship between random effects and logit of the proportion correct, but as we can see, some points are lined up because of variations in the data.

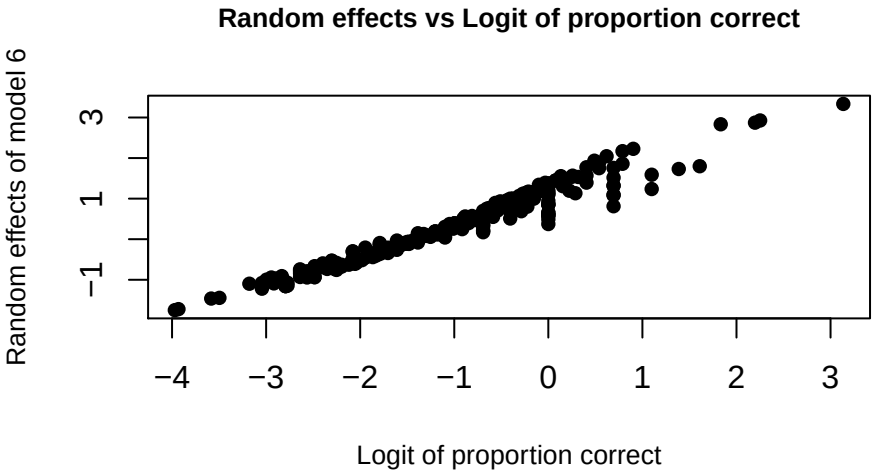


Figure 13: Plot of random effects from model 6 vs logit of proportion correct for each player

(d) Describe the methods you used to find the best model.

We tried additional three models for predicting the probability of a correct response. Model 7: $resp = \alpha_{0j[i]} + \alpha_{1j[1]} currentQuestion_i + \alpha_{2j[i]} experimentName_i + \epsilon_i$, where $\alpha_{0j} = \beta_{00} + \eta_j$; Model 8: $resp = \alpha_{0j[i]} + \alpha_{1j[i]} experimentName_i + \epsilon_i$, where $\alpha_{0j} = \beta_{00} + \eta_j$; Model 9: $resp = \alpha_{0j[i]} + \alpha_{1j[i]} currentQuestion_i + \epsilon_i$, where $\alpha_{0j} = \beta_{00} + \beta_{01} avgAccuracy_j + \eta_{0j}$. In general, currentQuestion and experimentName are considered to be individual-level predictors, and avgAccuracy is group-level predictor as it varies by participants. In order to find the best model, we first compared the QQplot of random effects of these three models as shown in appendix II. The QQplots show that their random effects all follows normal distribution as points genearlly follow the straight line. Then we compared the values of AIC, BIC and deviance for three models. From table 4, we can see that model 9 has smallest AIC, BIC and deviance. Therefore, model 9 is the best model, it indicates that the response is affected by player's average accuracy and type of question as fixed effects and randomness caused by players.

Table 3: Comparison on AIC, BIC and deviance

	Model7	Model8	Model9
AIC	8584	8755	7923
BIC	8830	8867	8077
deviance	8514	8723	7879

(e) Add a second random intercept.

If we add ip3, the groups of computers, as the second random effect, we will have new model 10, where avgAccuracy is group-level predictor, currentQuestion is individual-level predictor and random effects are composed by SID and ip3. Since model 10 has AIC = 7911.1, BIC = 8072.6 and deviance = 7865.1 which are all smaller than previous best model 9, so by adding a second random intercept does help. If we exclude random effect from SID, we have model 11, which has AIC = 7929.9, BIC = 8084.3 and deviance = 7885.9. By comparison, the values of AIC, BIC and deviance from model 11 are all greater than those in model 10, by the rule of 3 units difference in considering a better model, we conclude that model 10 is better than model 11, and figure 14 the QQplot of random effect of model 10 also support this argument. Even though, ip3 is composed by individual players SID, they play different random effects on the logistic regression model, so we keep the random effect of SID. Therefore, our best model with smallest values of AIC, BIC and deviance is model 10.

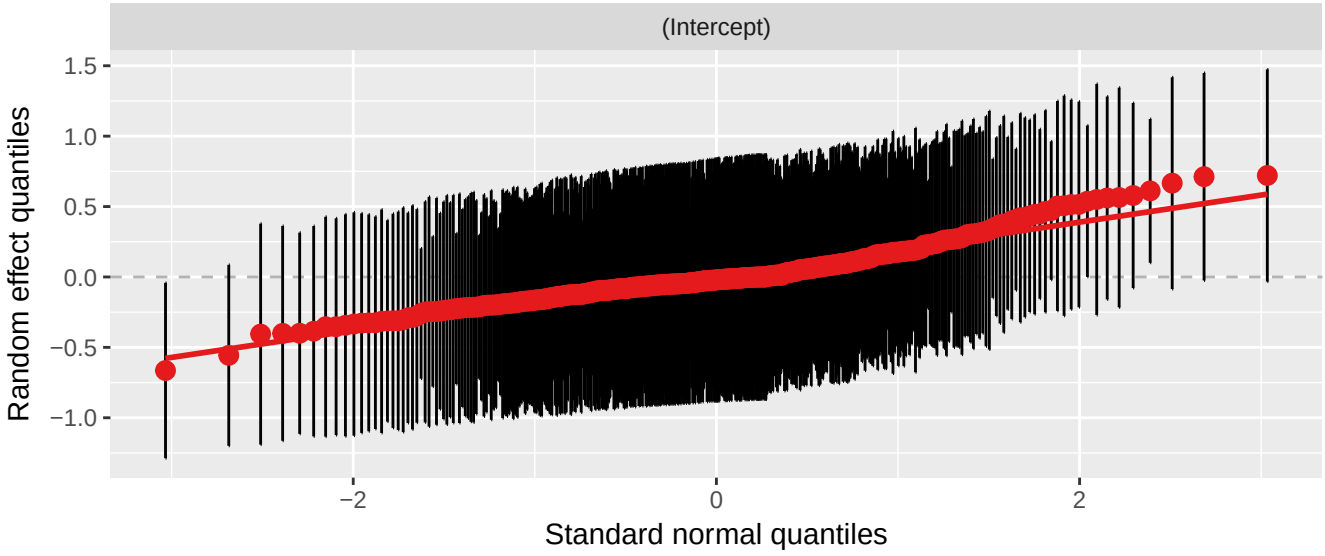


Figure 14: QQplot of random effect of model 10

5. Summary

This report analyzed a dataset of 8257 observations and 31 variables, which records the performance of 414 players playing 20 types of questions in a computer game. The goal is to predict the probability of question difficulty based on this dataset. In particular, there are 132 players answering all 20 questions and only 4 out of 414 players get 100% accuracy rate.

We first fitted logistic regression model for predicting the probability that a question will be answered correctly and tried to find the best model by conducting diagnostic analysis and comparing the values of AIC and deviance, the smaller the better. We found out that the logit of the probability of correct response has positive linear relationship with type of questions. As player moves from one question to next question, the probability that he will answer next question right depends on the coefficient of that question. Then, we fitted another logistic regression model for predicting the probability that a player will provide a correct answer. We found that the logit of the probability is linearly related to players specified by players' id. Accordingly, we can assume that the probability of question difficulty is related to players' id and types of question. Then, we fitted mixed effect logistic regression model by specifying the fixed effects and random effects. We found the best model by comparing diagnostic plots, AIC and BIC. Finally, we identified that the probablity of a correct response is related to player's average accuracy, type of question givien the random effects of ip3 and SID. Specifically,

type of question is individual level predictor which has fixed effect on the correct response, player's average accuracy is group-level predictor which also has fixed effect on the correct response, while group of computers and player's id act as group-level variables which produce random error for the prediction of probability.

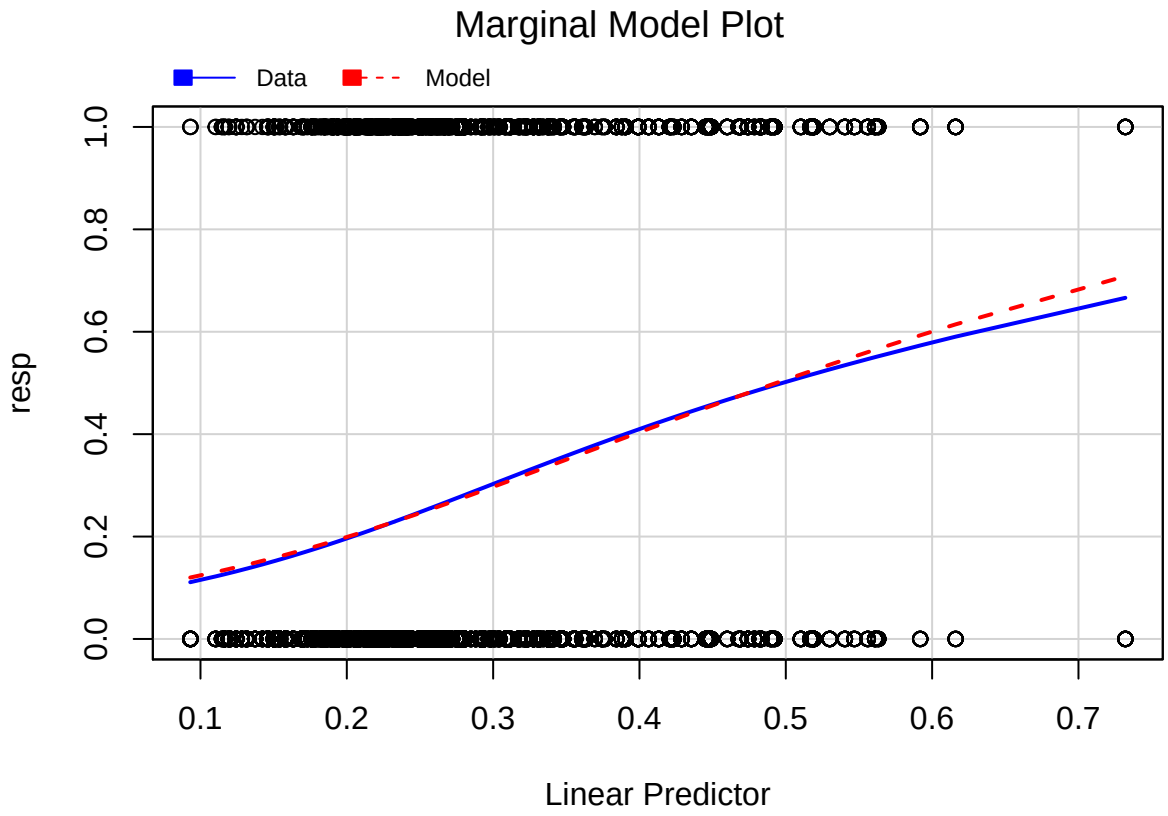
Even though we found that the probability of quesiton difficulty is associated with type of question and individual's average performance, there exist some drawbacks in our logistic model. The diagnostic plot indicates that the prediction is not sufficient enough and the model shows convergence issues. In addition, due to large size of parameters, R doesn't function very well on automated model selection, so there might exist better model generated by more comprehensive testing.

Appendix I: References

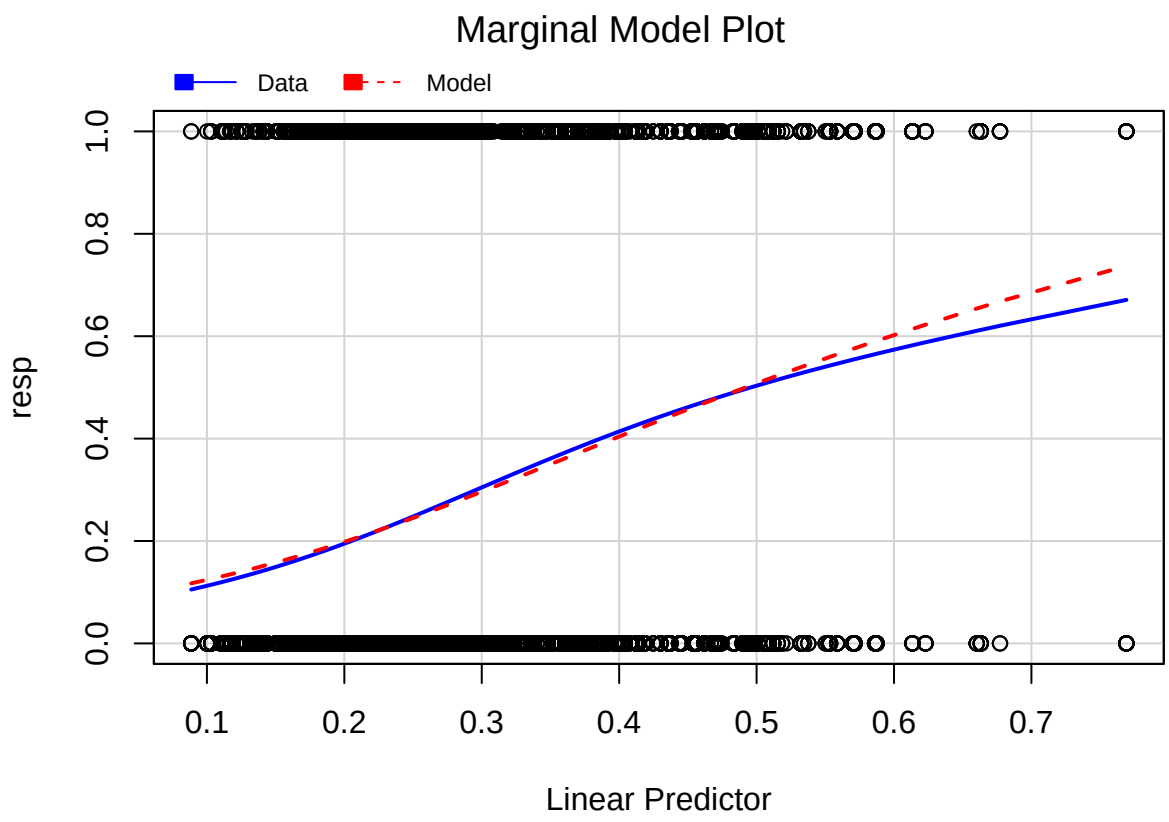
1. Gelman, Andrew, and Jennifer Hill. Data Analysis Using Regression and Multilevel / Hierarchical Models. 10th ed. Cambridge: Cambridge University Press, 2006. Print.
2. "How Do I Interpret Odds Ratios in Logistic Regression?" N.p., n.d. Web. 15 Nov. 2016.
3. "Deviance (statistics)." Wikipedia. Wikimedia Foundation, n.d. Web. 16 Nov. 2016.
4. "Visualizing (generalized) Linear Mixed Effects Models with Ggplot #rstats #lme4." R-bloggers. N.p., 2014. Web. 16 Nov. 2016.

Appendix II

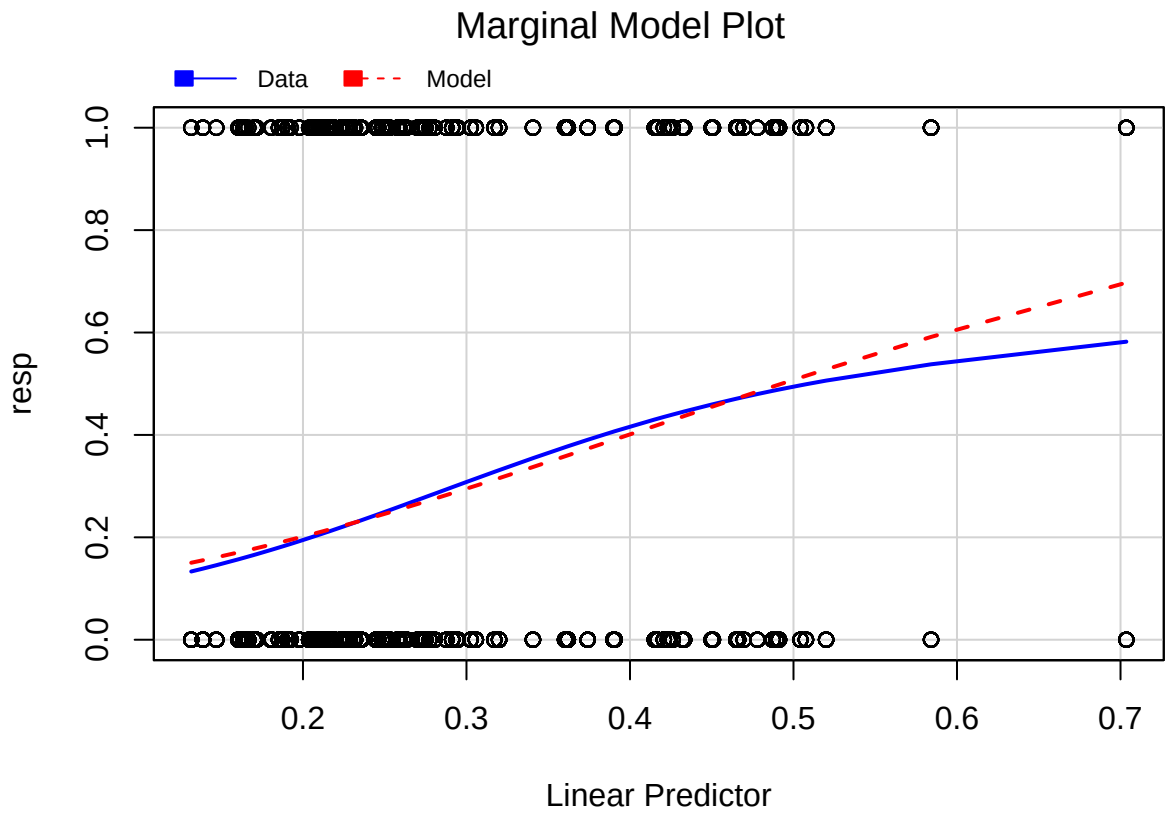
```
#### Q2(c) find best model to predict the probability that a question will be
#### answered. There are other candidate models:
model2 <- glm(resp ~ questions + experimentName, data = dtf, family = "binomial")
residual2 <- resid(model2)
predict2 <- predict(model2)
mmps(model2)  #marginal model plot
```



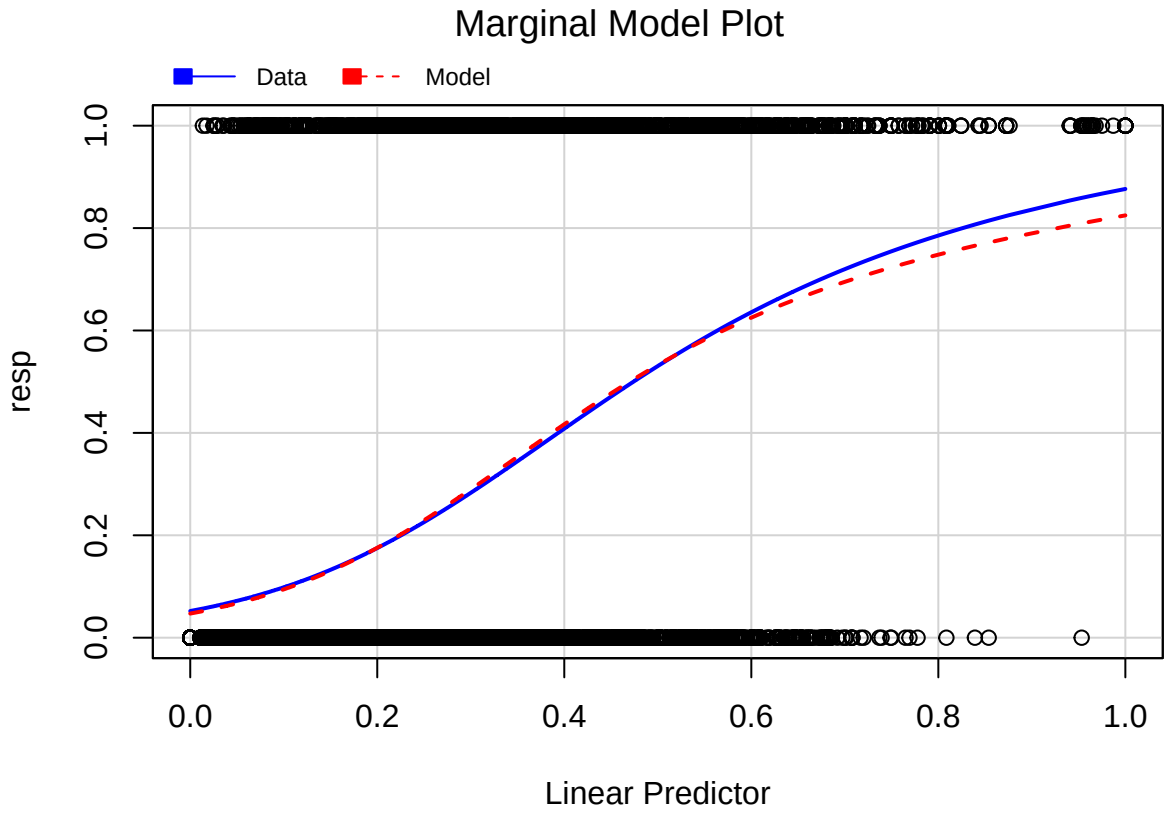

```
model3 <- glm(resp ~ questions + experimentName + levelName, data = dtf, family = "binomial")
residual3 <- resid(model3)
predict3 <- predict(model3)
mmps(model3)  #marginal model plot
```



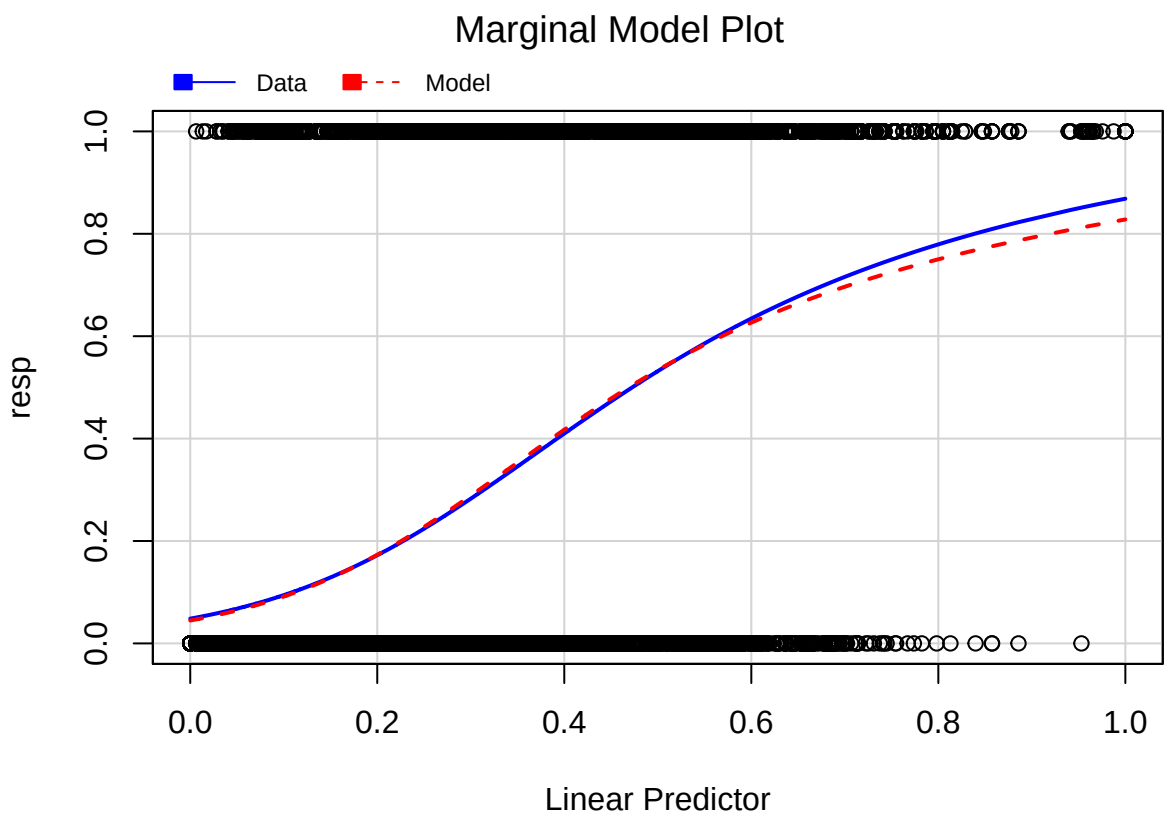

```
model4 <- glm(resp ~ questions + levelName, data = dtf, family = "binomial")
residual4 <- resid(model4)
predict4 <- predict(model4)
mmps(model4)  #marginal model plot
```



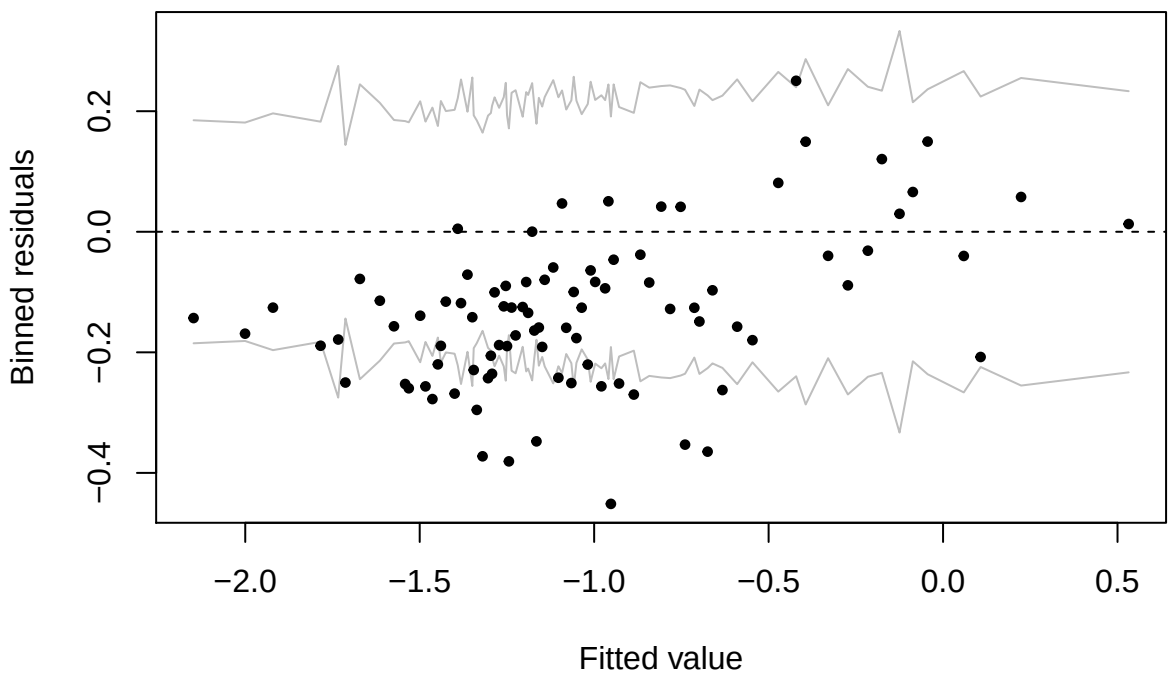
```r
Ip2 <- as.factor(dtf$ip2)
modelip <- glm(resp ~ questions + Ip2, data = dtf, family = "binomial")
residualip <- resid(modelip)
predictip <- predict(modelip)
mmps(modelip)  #marginal model plot
```



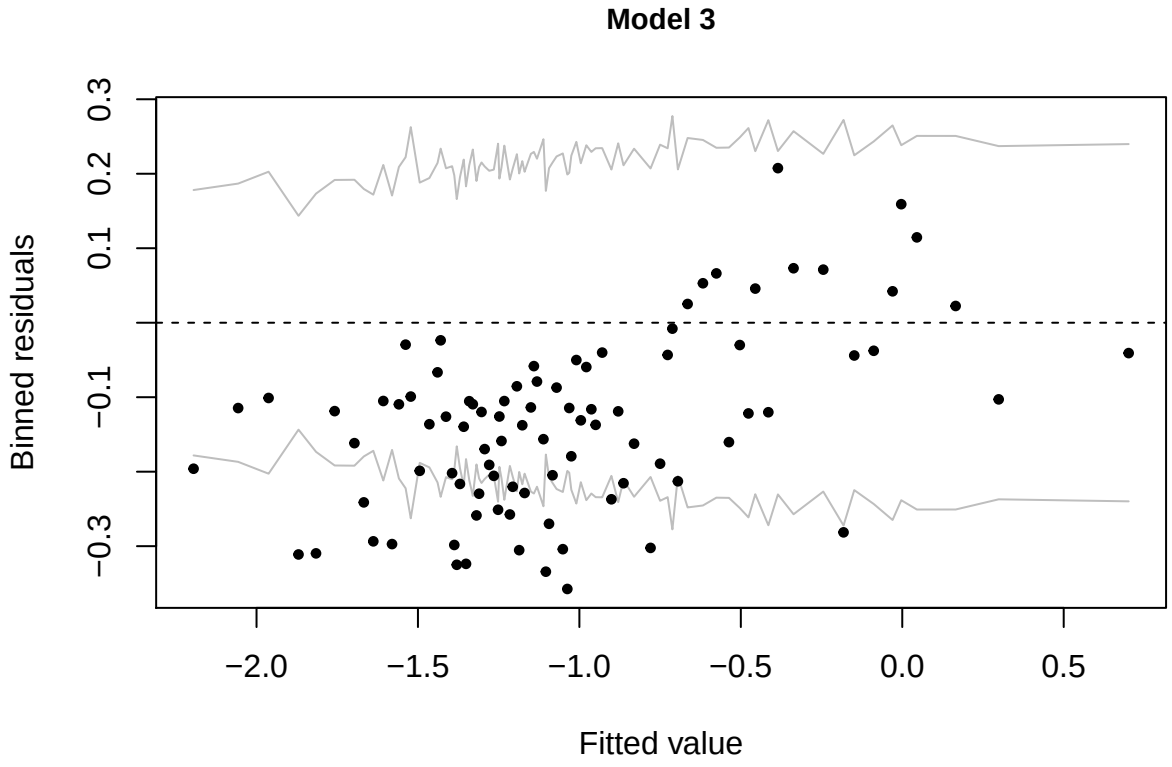

```r
Ip3 <- as.factor(dtf$ip3)
modelip3 <- glm(resp ~ questions + Ip3, data = dtf, family = "binomial")
residualip3 <- resid(modelip3)
predictip3 <- predict(modelip3)
mmps(modelip3)  #marginal model plot
```



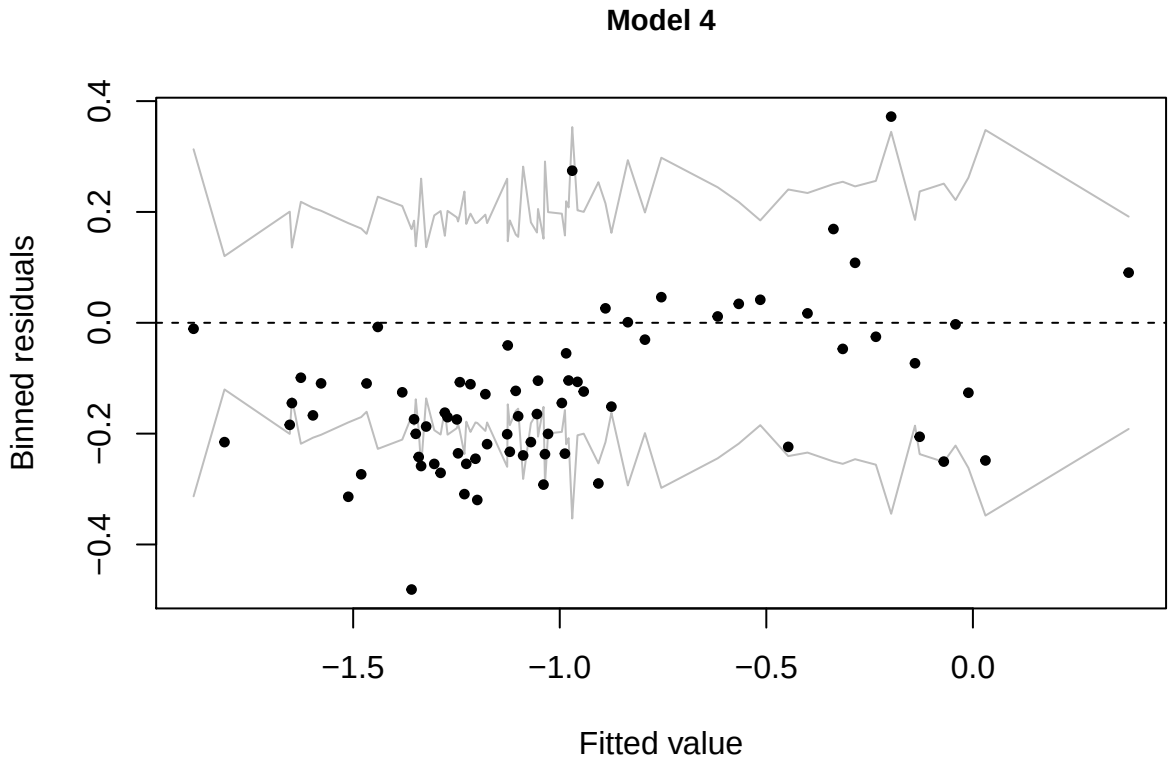

```
# binned residual plot for other candidate models:
binnedplot(predict2, residual2, xlab = "Fitted value", ylab = "Binned residuals",
    main = "Model 2", cex.pts = 0.6, cex.main = 0.9)
```



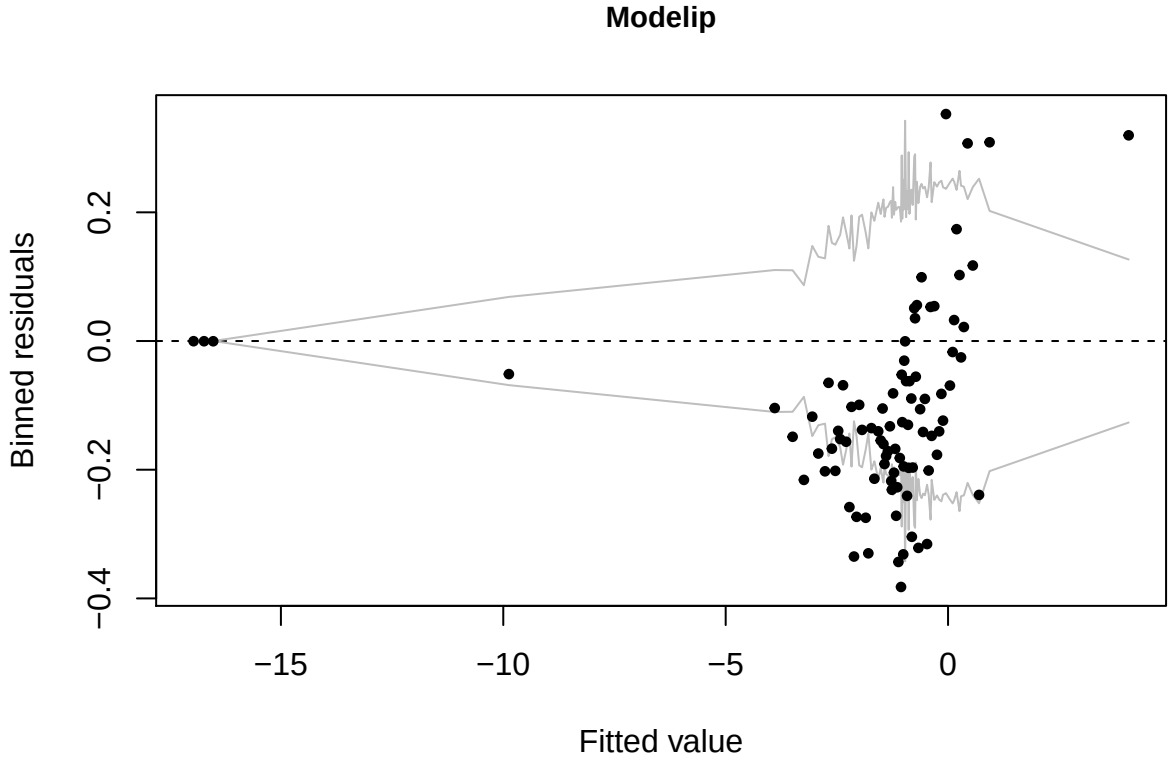

```
binnedplot(predict3, residual3, xlab = "Fitted value", ylab = "Binned residuals",
    main = "Model 3", cex.pts = 0.6, cex.main = 0.9)
```



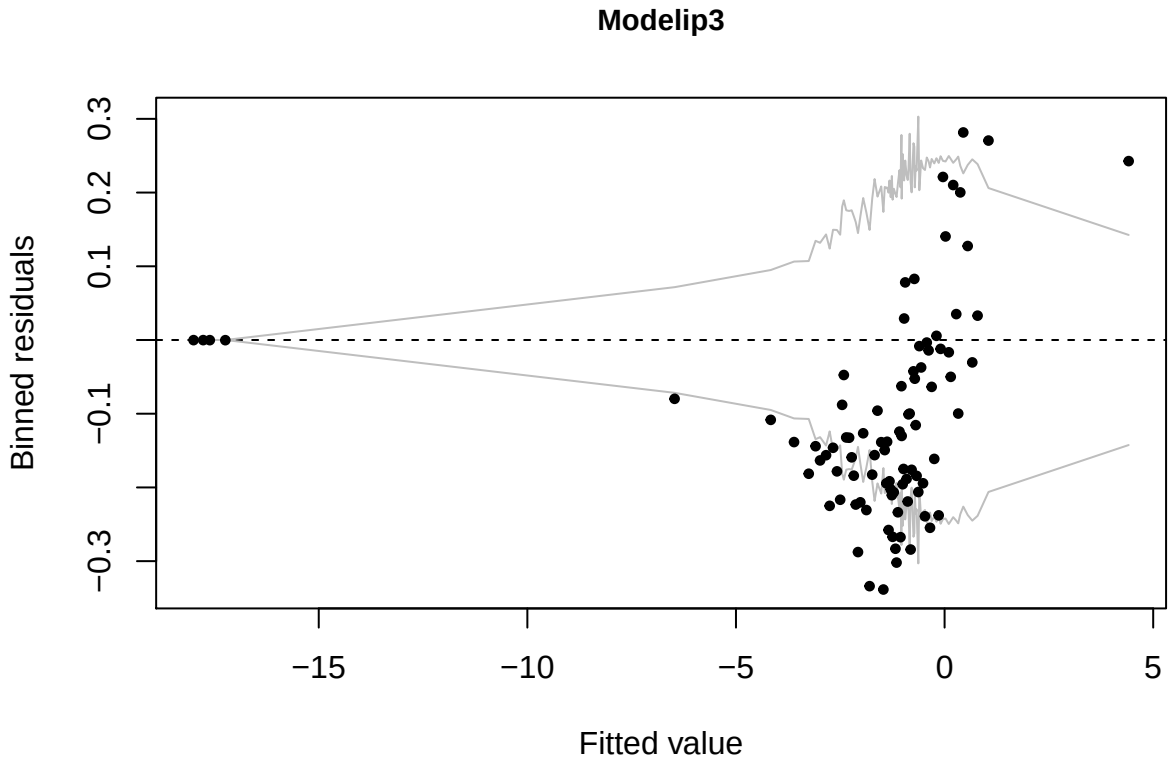

```
binnedplot(predict4, residual4, xlab = "Fitted value", ylab = "Binned residuals",
    main = "Model 4", cex.pts = 0.6, cex.main = 0.9)
```



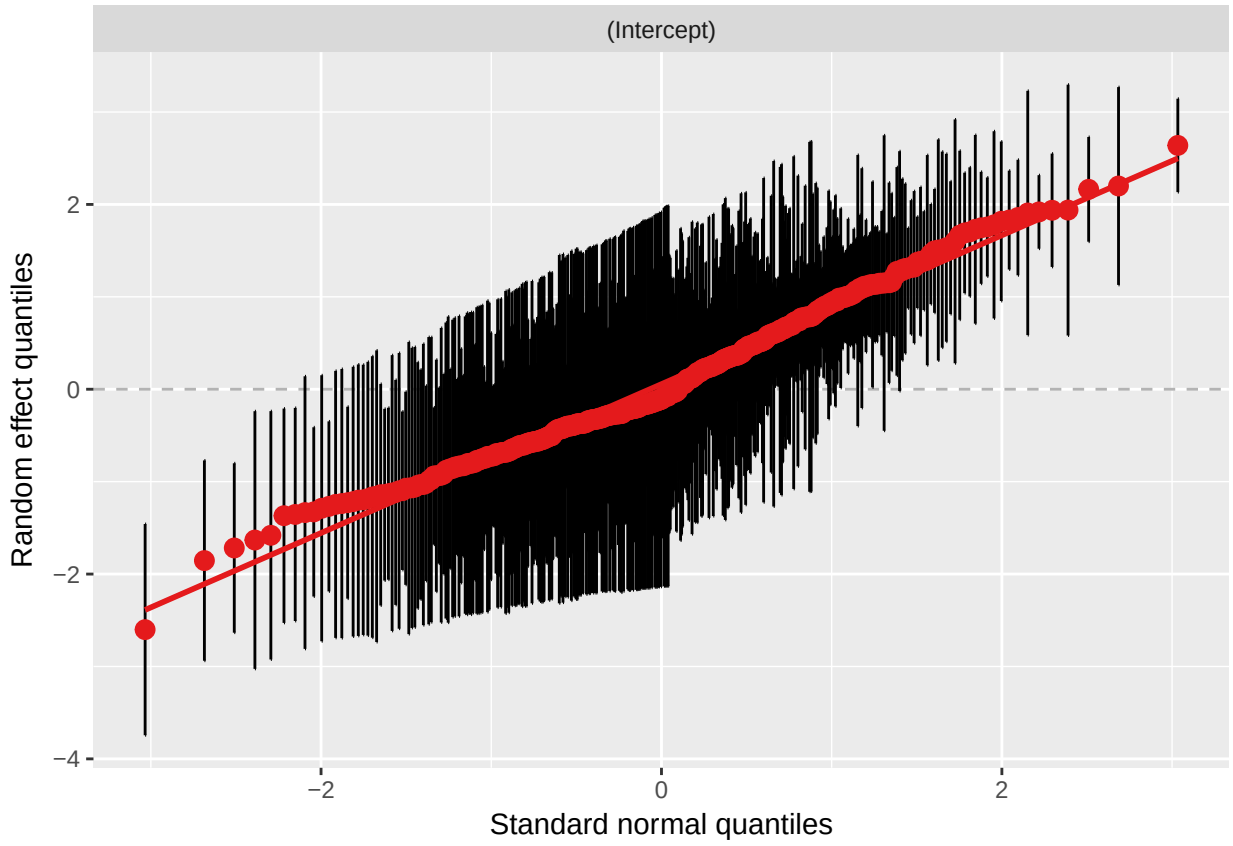

```
binnedplot(predictip, residualip, xlab = "Fitted value", ylab = "Binned residuals",
    main = "Modelip", cex.pts = 0.6, cex.main = 0.9)
```



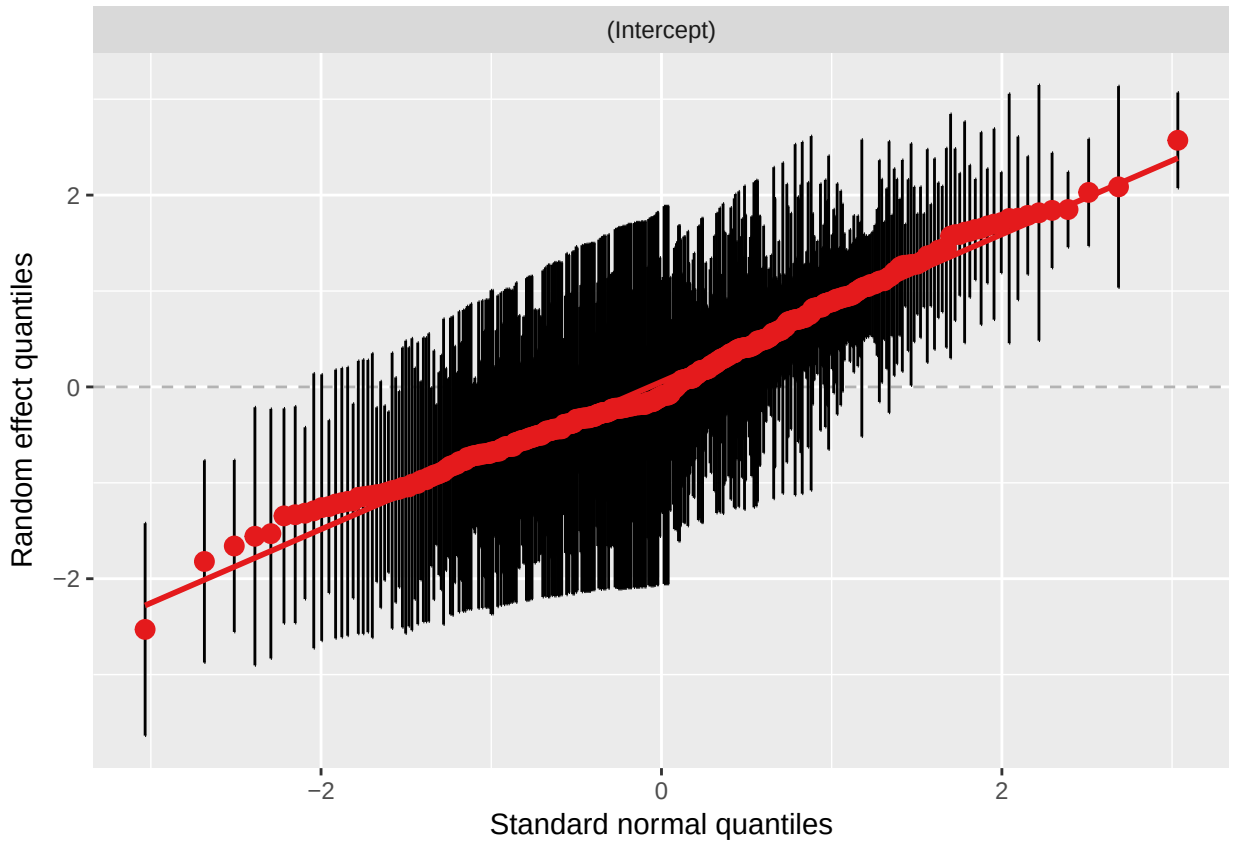

```
binnedplot(predictip3, residualip3, xlab = "Fitted value", ylab = "Binned residuals",
    main = "Modelip3", cex.pts = 0.6, cex.main = 0.9)
```




```
# Q4 find best mixed effect logistic model
sjp.glmer(model7, type = "re.qq")  # glmer(resp ~ questions + experimentName + (1|SID)
```




```
sjp.glmer(model8, type = "re.qq")  #glmer(resp ~ experimentName - 1 + (1|SID), data = dtf, family = 'binomial
```



```
sjp.glmer(model9, type = "re.qq")  #glmer(resp ~ avgAccuracy + questions + (1|SID) - 1
```