

Homework Six

Josh Virene

Problem 1

Problem 1(a)

There will be 84 rows, this is because there are 42 students and they are tested for either method A or method B and assigned a 1 if they are included in the method, and a 0 if they are not included in the method.

Problem 1(b)

Response variable: Spelling test score

Variables modeled as random effects: Student ID number

Variables modeled as fixed effects: Method (A or B), Month (1 through 8)

Problem 1(c)

The variables in this model that should interact in order to investigate the research question “scores improve more in one method more quickly than in the other method” are **month and method**.

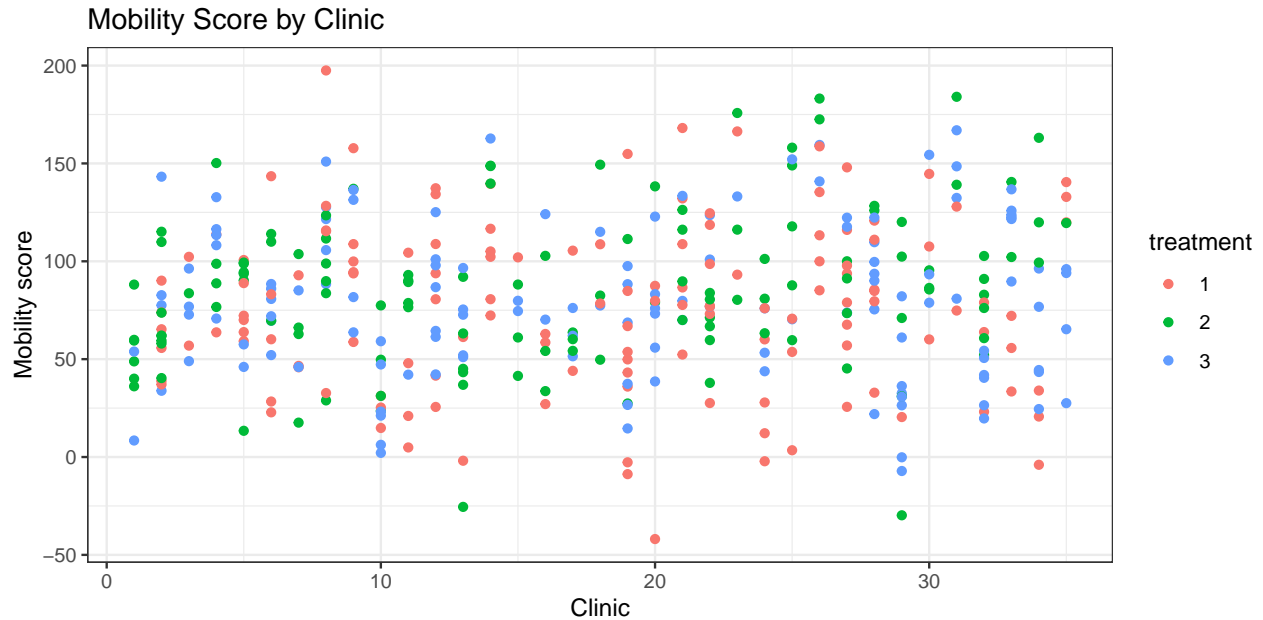
Problem 1(d)

Two sources of random variation that are appropriate for this model are:

1. Student ID number (accounts for variation in test score among the students).
2. Error (this accounts for variation in test scores among students when all other factors such as month word_type etc. are held constant- the variation not explained by the model)

Problem 2

Problem 2(a)



Using only this plot, it is hard to determine whether variation exists across clinics or across treatment methods. The data are all over the plot for both of these treatment and clinic so any patterns are not discernible.

Problem 2(b)

```
## Analysis of Variance Table
##
## Response: mobility_change
##           Df Sum Sq Mean Sq F value Pr(>F)
## treatment  2   5574   2787.0   1.6939  0.1851
## Residuals 394 648259  1645.3
##
## $lsmeans
##   treatment lsmean    SE df lower.CL upper.CL
##   1          76.6  3.47 394     69.8     83.4
##   2          85.7  3.64 394     78.5     92.8
##   3          79.3  3.48 394     72.5     86.1
##
## Confidence level used: 0.95
##
## $contrasts
##   contrast estimate    SE df t.ratio p.value
##   1 - 2      -9.05  5.03 394   -1.801  0.1707
##   1 - 3      -2.66  4.91 394   -0.542  0.8505
##   2 - 3       6.39  5.04 394    1.269  0.4137
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

This model violates the independence assumption because it fails to account for the clinic variable in the data by using some form of random effect model. The errors differ systematically across clinics, so there is correlation in errors, violating the independence assumption which can introduce bias. Accounting for this in the other models addresses this violation.

Problem 2(c)

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: mobility_change ~ treatment + (1 | clinic)
## Data: mobility
##
## REML criterion at convergence: 3957.1
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.2568 -0.5989 -0.0261  0.6783  3.0651
##
## Random effects:
## Groups   Name      Variance Std.Dev.
## clinic  (Intercept) 604.2    24.58
## Residual                1097.4   33.13
## Number of obs: 397, groups: clinic, 35
##
## Fixed effects:
##              Estimate Std. Error    df t value Pr(>|t|)
## (Intercept)   74.093      5.101  54.009   14.526 < 2e-16 ***
## treatment2    16.096      4.279  369.202    3.761 0.000197 ***
## treatment3     8.197      4.148  367.160    1.976 0.048876 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) trtmn2
## treatment2 -0.403
## treatment3 -0.408  0.483
##
## Type III Analysis of Variance Table with Satterthwaite's method
##              Sum Sq Mean Sq NumDF DenDF F value    Pr(>F)
## treatment  15560  7779.8      2  368.5   7.0894 0.0009525 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## $lsmeans
## treatment lsmean SE    df lower.CL upper.CL
## 1          74.1 5.10 55.0     63.9     84.3
## 2          90.2 5.17 58.1     79.8    100.5
## 3          82.3 5.10 54.9     72.1     92.5
##
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.95
##
## $contrasts
```

```
## contrast estimate SE df t.ratio p.value
## 1 - 2 -16.1 4.28 370 -3.758 0.0006
## 1 - 3 -8.2 4.15 368 -1.975 0.1199
## 2 - 3 7.9 4.29 370 1.842 0.1575
##
## Degrees-of-freedom method: kenward-roger
## P value adjustment: tukey method for comparing a family of 3 estimates
```

iii. This model includes treatment in addition to clinic as a random effect whereas in the previous model, the only variable in the model was treatment. Due to the difference in the structure of these models the pairwise comparison output is different. One important difference is that the standard errors in the model that includes clinic as a random effect are lower than the original model where clinic was not included.

iv. Adding clinic as a random effect makes a large difference in the standard error of these models because the model is now accounting for clinics whereas the model without attributes variation in mean mobility to error (there was more error variance causing the standard errors to be higher). The clinic variable is reducing the error attributed to noise in the model.

v. The two variances in this model, obtained from the summary statistics are as follow: $\sigma_\epsilon^2 = 1097.4$
 $\sigma_\alpha^2 = 604.2$

The error variance, σ_ϵ^2 is the variance in individual values of y for some combination of values for x and alpha, or some combination of fixed and random factors.

The random effect variance σ_α^2 is quantifying the variability in means in the response variable that is attributable to differences in levels across clinics.

Problem 2(d)

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: mobility_change ~ treatment + (treatment | clinic)
## Data: mobility
##
## REML criterion at convergence: 3951
##
## Scaled residuals:
## Min 1Q Median 3Q Max
## -3.1552 -0.5596 -0.0347 0.6382 3.1622
##
## Random effects:
## Groups Name Variance Std.Dev. Corr
## clinic (Intercept) 657.81 25.648
## treatment2 415.79 20.391 -0.39
## treatment3 82.35 9.075 0.03 0.91
## Residual 1025.93 32.030
## Number of obs: 397, groups: clinic, 35
##
## Fixed effects:
## Estimate Std. Error df t value Pr(>|t|)
## (Intercept) 73.901 5.240 29.946 14.103 9.22e-15 ***
## treatment2 16.639 5.474 30.173 3.040 0.00486 **
## treatment3 8.900 4.365 52.477 2.039 0.04653 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) trtmn2
## treatment2 -0.506
## treatment3 -0.365  0.551

## Type III Analysis of Variance Table with Satterthwaite's method
##          Sum Sq Mean Sq NumDF  DenDF F value    Pr(>F)
## treatment 9672.1    4836      2  48.471  4.7138 0.01347 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- i. This model having random slopes means that both the slope and the intercept vary by subject. This is different than the random effects model because in that model, it was only the intercept that could vary by subject.
- ii. $\sigma_{\alpha\beta 2}^2 = 20.391$ This estimate quantifies the standard deviation in mobility attributed to treatment 2 across clinics.

Problem 2(e)

The p value for treatment method is 0.01347. The interpretation of this p value within the context of this study in relation to imaginary repeated trials is that it assesses the statistical significance of treatment, when treatment is a variable that takes on a value for the B_1 coefficient that holds across these trials. This is different than the clinic variable because clinic is a random effect, so this means that the clinics being sampled do not need to be the same across these repeated samples.

Problem 3

Problem 3(a)

The plot that exhibits the largest random effect variance is plot c (bottom left). Looking at this plot, the mean values for the response variable vary largely when considering group and treatment; the reason this is not error variance is because the values are clustered more closely looking at the main effect for each factors.

Problem 3(b)

The output with the largest error variance is plot a (top left). This plot has the largest error variance because the large differences in the response variable are not well accounted for by assessing whether there are main effects for group or treatment. To clarify this is looking across groups while holding treatment constant, or looking across treatments holding groups constant. Because it does not appear variance is accounted for by fixed effects for either factor in this model, the variance in the response variable will be attributed to error and this model will return a large error variance.

Problem 3(c)

The output with the largest fixed effect for treatment will be plot b (top right). This plot has a large fixed effect for treatment because the values for the response variable differ by treatment level (1 and 2) across all groups in the model. As an example of this, for group 5.0 the mean value for y is roughly 100 in treatment 1 and around 130 in treatment 2.

Problem 4

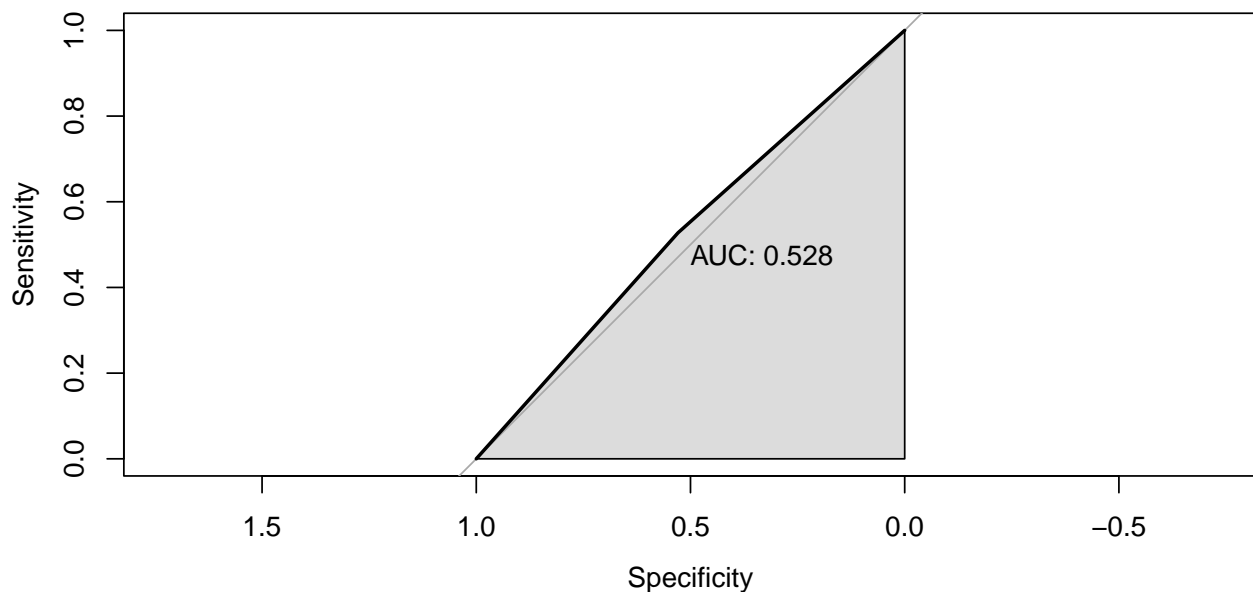
Problem 4(a)

```
## [1] 4.925991
## [1] 0.3922157
```

- i. The average age of children in the study is **4.925991** years
- ii. The standard deviation of children in the study is **0.3922157** years

Problem 4(b)

```
##
## Call:
## glm(formula = correct ~ word_type, data = spelling)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.5411  -0.4844   0.4589   0.4589   0.5156
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.48444    0.01664  29.107  <2e-16 ***
## word_type    0.05667    0.02354   2.407   0.0162 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.249311)
##
##      Null deviance: 449.71  on 1799  degrees of freedom
## Residual deviance: 448.26  on 1798  degrees of freedom
## AIC: 2611.9
##
## Number of Fisher Scoring iterations: 2
```



```

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##           0 464 436
##           1 413 487
##
##           Accuracy : 0.5283
##           95% CI : (0.505, 0.5516)
##       No Information Rate : 0.5128
##       P-Value [Acc > NIR] : 0.09732
##
##           Kappa : 0.0567
##
## Mcnemar's Test P-Value : 0.45023
##
##           Sensitivity : 0.5276
##           Specificity : 0.5291
##       Pos Pred Value : 0.5411
##       Neg Pred Value : 0.5156
##           Prevalence : 0.5128
##       Detection Rate : 0.2706
##       Detection Prevalence : 0.5000
##       Balanced Accuracy : 0.5284
##
##       'Positive' Class : 1
##

```

- ii. The slope for word type = 0.05667 ($\hat{\beta}_1 = 0.05667$). The interpretation of this slope is there is a 0.05667 change in log odds for a one unit change in word type. This is a binary variable where if word_type = 1, it sounds like its first letter and if it does not sound like its first letter, word_type = 0.

This means that log odds of a child choosing the correct first letter of the word being spelled increase by 0.05667 if a word sounds like its first letter.

- iii. The logodds for the two outcomes are calculated using the model as follows: $\log\hat{odds} = \hat{\beta}_0 + \hat{\beta}_1 x$
 $\log\hat{odds} = \hat{\beta}_0 + \hat{\beta}_1(\text{wordtype})$

where $\hat{\beta}_1 = 1$ when the start of the word sounds like the first letter $\hat{\beta}_1 = 0$ when the start of the word does not sound like the first letter

When the start of the word sounds like the first letter ($\hat{\beta}_1 = 1$) $\log\hat{odds} = 0.48444 + 0.05667(1)$ logodds = 0.54111

When the start of the word sounds like the first letter ($\hat{\beta}_1 = 0$) $\log\hat{odds} = 0.48444 + 0.05667(0)$ log odds = 0.48444

Using these log odds, the probabilities are computed by converting these into odds, then calculating probability using the formula $P = (\text{odds} / 1 + \text{odds})$

When the start of the word sounds like the first letter formula: $\exp(0.4844+0.05667)/(1+\exp(0.4844+0.05667))$

Probability = 0.6320613

When the start of the word does not sound like the first letter: formula: $\exp(0.4844)/(1+\exp(0.4844))$

Probability = 0.6187863

- iv. From this ROC curve, the **AUC=0.528**, this does not seem like a good AUC as it is very close to 0.500; this value means the prediction made by the model is as good as random guessing or even flipping a coin. From this, it is clear that only using the word_type variable to predict whether a child spelled a word correctly or not is not a very good model, and that additional information is needed to make this prediction.
- v. The specificity, sensitivity, false negative rate, and false positive rate are reported below:
 - Sensitivity: This is the true positive rate, the rate at which the predicted outcome is equal to 1 (correct) when the true outcome is equal to 1. $P(Y'_i = 1|Y_i = 1)$ **In this model the sensitivity = 0.5276**
 - Specificity: This is the true negative rate, the rate at which the predicted outcome is equal to 0 (incorrect) when the true outcome is equal to 0. $P(Y'_i = 0|Y_i = 0)$ **In this model the specificity = 0.5291**
 - False negative rate: This is the rate at which the predicted outcome is equal to 0 when the true outcome is equal to 1 $P(Y'_i = 1|Y_i = 0)$ **In this model the false negative rate is 0.4724**
 - False positive rate: This is the rate at which the predicted outcome is equal to 1 when the true outcome is equal to 0 $P(Y'_i = 1|Y_i = 0)$ **In this model the false positive rate is 0.4709**

Overall, this tells us the model is not very accurate; this is mainly shown by the false positive and negative rates being 0.47. This means that the model is returning a negative when the true outcome is positive (false negative) nearly half the time (47.24%) and returning a positive when the true outcome is negative nearly half the time (47.09%). Further supporting this, the true positive and negative rates are low, (52.76% and 52.91% respectively) these values indicate the success rate of the model in correctly predicting positive and negative outcomes. Overall, just under half of the predictions made by this model are incorrect, so the model is very inaccurate and a better model that incorporates more of the data provided by this data set will be needed.

Problem 4(c)

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: correct ~ word_type + (1 | subject)
## Data: spelling
##
##      AIC      BIC    logLik deviance df.resid
##  1437.6   1454.1   -715.8   1431.6     1797
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -4.8831 -0.3638  0.1132  0.3375  4.7903
##
## Random effects:
##  Groups Name      Variance Std.Dev.
##  subject (Intercept) 9.431    3.071
## Number of obs: 1800, groups:  subject, 75
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.02416    0.37566  -0.064 0.948711
## word_type    0.53179    0.14050   3.785 0.000154 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
##
## Correlation of Fixed Effects:
##      (Intr)
## word_type -0.184
```

The logodds are computed as follows:

$$\log\hat{odds} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \log\hat{odds} = \hat{\beta}_0 + \hat{\beta}_1(\text{wordtype})$$

where $\hat{\beta}_1 = 1$ when the start of the word sounds like the first letter $\hat{\beta}_1 = 0$ when the start of the word does not sound like the first letter

When the start of the word sounds like the first letter ($\hat{\beta}_1 = 1$) $\log\hat{odds} = -0.02416 + 0.53179(1) = 0.55595$

When the start of the word sounds like the first letter ($\hat{\beta}_1 = 0$) $\log\hat{odds} = -0.02416 + 0.53179(0) = -0.02416$

Using these log odds, the probabilities are computed by converting these into odds, then calculating probability using the formula $P = (\text{odds} / 1 + \text{odds})$

When the start of the word sounds like the first letter: formula: $\exp(0.55595)/(1+\exp(-0.02416 + 0.53179(1)))$

Probability = 0.6355149

When the start of the word sounds like the first letter: formula: $\exp(-0.02416)/(1+\exp(-0.02416))$

Probability = 0.6187863

Problem 4(d)

These results are very different because in the first model, there was not a random intercept for subject whereas in the second model, a random intercept for subject was included. The second model has word_type as a fixed effect, meaning word type is tested in the model at specific and defined levels that the researchers are interested in. This model includes subject number as well; this refers to which child's response is recorded. For random factors (in this model subject), the effects themselves are not of interest, but the variation or variability introduced by subject is of interest. Researchers want to introduce this eliminate the bias that arises from systematic differences in error across subject number.

Problem 4(e)

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: correct ~ word_type + age + order + letterwrite + (1 | subject)
## Data: spelling
##
##      AIC      BIC   logLik deviance df.resid
##  1383.4   1416.3  -685.7   1371.4     1794
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -5.8515 -0.3565  0.0960  0.3014  5.5942
##
## Random effects:
## Groups Name Variance Std.Dev.
## subject (Intercept) 4.839    2.2
## Number of obs: 1800, groups: subject, 75
##
```

```
## Fixed effects:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.55692    3.69540  -2.586 0.009705 **
## word_type    0.53906    0.14713   3.664 0.000248 ***
## age          1.09895    0.78905   1.393 0.163694
## order        -0.06767    0.02126  -3.183 0.001458 **
## letterwrite  6.62775    1.08467   6.110 9.94e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) wrd_ty age    order
## word_type  -0.036
## age        -0.979  0.009
## order      -0.022 -0.022 -0.008
## letterwrite 0.191  0.038 -0.370 -0.034
```

The slope for letter write, at 6.62775 is much larger than the slopes for the other independent variables in the model. In order to explain this, first the interpretation for this slope must be provided. This slope means that the log odds of a child getting the first letter of the word correctly increase by 6.62775 when the proportion of letters a child writes correctly upon request increases by one.

The statistical explanation for why this slope is larger than the others is that proportion of letters written correctly has the largest change in the log odds for the response variable out of all of the variables.

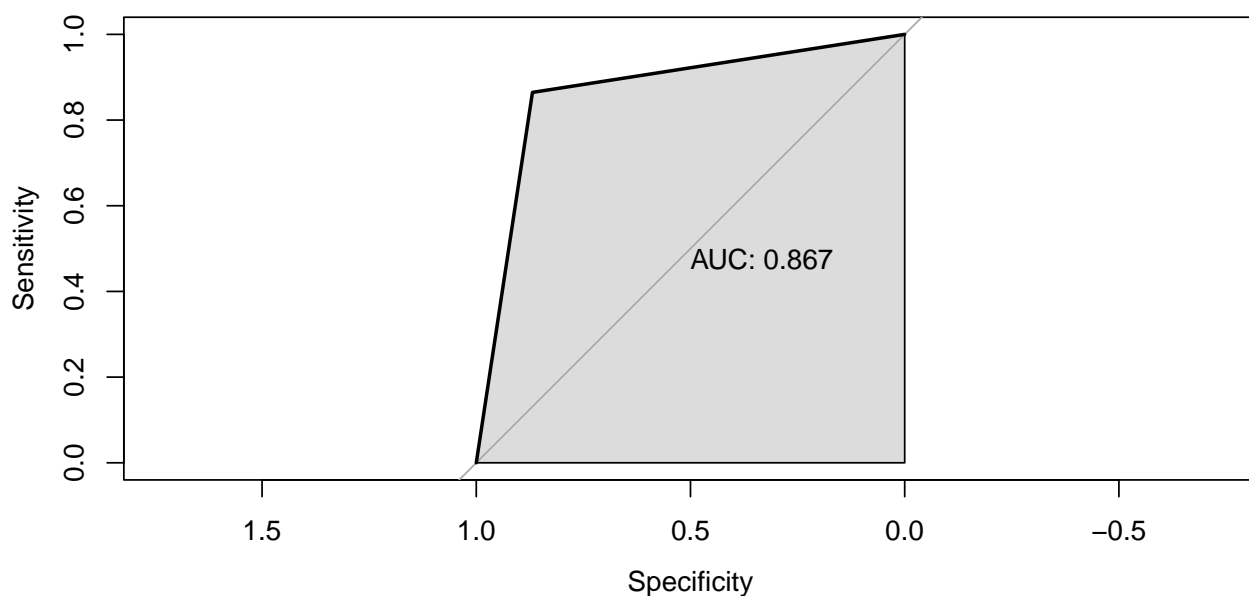
The practical explanation of this, considering children and their spelling; for the letters b,d,g,j,k,p higher proportions of these that children can write correctly upon request correspond to a higher success rate with which children write the first letter of all words correctly, especially for words that involve these letters.

Another consideration is that the variables in this model are not all of the same units / scale (i.e., age and letter write). This means that a one unit in change for these variables will not have the same impact, changing age by one year may only correspond to a small percentage change in the odds a child will get the first letter of a word correctly whereas a one unit change in a proportion of the above letters a child can write correctly will have a much larger impact.

Problem 4(f)

The slope for age turned out to be non-significant, this is likely because within the data relative to age, the other variables such as word_type, order, and letterwrite are picking up a lot of the signal and the model does not depend on age very much to predict the log odds of the response variable. In models with fewer explanatory variables, age will have more signal (the model will depend on age more to predict the response variable outcome) and become statistically significant.

Problem 4(g)



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##           0 762 125
##           1 115 798
##
##           Accuracy : 0.8667
##           95% CI : (0.8501, 0.882)
##           No Information Rate : 0.5128
##           P-Value [Acc > NIR] : <2e-16
##
##           Kappa : 0.7332
##
##           McNemar's Test P-Value : 0.5613
##
##           Sensitivity : 0.8646
##           Specificity : 0.8689
##           Pos Pred Value : 0.8740
##           Neg Pred Value : 0.8591
##           Prevalence : 0.5128
##           Detection Rate : 0.4433
##           Detection Prevalence : 0.5072
##           Balanced Accuracy : 0.8667
##
##           'Positive' Class : 1
##
```

The specificity, sensitivity, false negative rate, and false positive rate are reported below:

- Sensitivity: This is the true positive rate, the rate at which the predicted outcome is equal to 1 (correct) when the true outcome is equal to 1. $P(Y'_i = 1|Y_i = 1)$ **In this model the sensitivity = 0.8646**
- Specificity: This is the true negative rate, the rate at which the predicted outcome is equal to 0

(incorrect) when the true outcome is equal to 0. $P(Y'_i = 0|Y_i = 0)$ **In this model the specificity = 0.8689**

- False negative rate: This is the rate at which the predicted outcome is equal to 0 when the true outcome is equal to 1 $P(Y'_i = 1|Y_i = 0)$ **In this model the false negative rate is 0.1354**
- False positive rate: This is the rate at which the predicted outcome is equal to 1 when the true outcome is equal to 0 $P(Y'_i = 1|Y_i = 0)$ **In this model the false positive rate is 0.1311**

Comparing this model to the previous model, the additional independent variables of age, order, and letterwrite being added to the model make the predictions much more accurate. the false positive and negative rates decreased to around 0.13, which is much lower than the previous rates of 0.47. This model will return a negative when the true outcome is positive 13.54% of the time, and return a positive when the true outcome is negative 13.11% of the time. The true positive and negative rates have significantly increased compared to the first model at 86.46% and 86.89% respectively. Because of these indicators, this model predicts the outcomes correctly; **this model is much better at making predictions compared to the previous model**

Problem 4(h)

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: correct ~ word_type + age + order + letterwrite + (age | subject)
## Data: spelling
##
##      AIC      BIC    logLik deviance df.resid
##  1387.3   1431.2   -685.6   1371.3     1792
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -5.8480 -0.3566  0.0945  0.3009  5.5914
##
## Random effects:
##  Groups Name      Variance Std.Dev. Corr
##  subject (Intercept) 1.15219  1.0734
##      age           0.05217  0.2284   1.00
## Number of obs: 1800, groups: subject, 75
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.51959    3.61519  -2.633 0.008458 **
## word_type    0.53900    0.14712   3.664 0.000249 ***
## age          1.09655    0.77524   1.414 0.157223
## order        -0.06767    0.02126  -3.183 0.001458 **
## letterwrite  6.61373    1.08034   6.122 9.25e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) wrd_ty age      order
## word_type    -0.037
## age          -0.978  0.009
## order        -0.023 -0.022 -0.008
```

```
## letterwrite 0.185 0.038 -0.367 -0.034
```

The issue with including a random slope for age across subjects is that there is little variance in age in this data set, so the variance in age across subjects will not be very large. This model will have a random slope for age across subjects which would not make much sense if the age across subjects does not change very much. The implication of low variance in age across subjects is that the difference in age between subject one and subject two is not going to have a large difference when making comparison to the difference in age between subject one and subject three. To support this, the age range in the dataset is: minimum age = 4.02, maximum age = 5.67. In order to justify having a random slope for age, the ages would have to vary much more across subjects, (i.e., range from 5-50), then the study design and the questions would be very different than what is being considered in this study.

Appendix

```
library(knitr)
#install the tidyverse library (do this once)
#install.packages("tidyverse")
library(tidyverse)
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, fig.width = 8, fig.height = 4, tidy = FALSE)
#code for prob2a
library(readr)
mobility <- read_csv("mobility.csv")

mobility$treatment = as.factor(mobility$treatment)
##create the plot:
library(ggplot2)

ggplot() +
  theme_bw() +
  geom_point(aes(
    x = clinic,
    y = mobility_change,
    color = treatment),
    data = mobility) +
  xlab("Clinic") +
  ylab("Mobility score") +
  ggtitle("Mobility Score by Clinic")

#code for prob2b
##anova and pairwise comparisons
library(emmeans)
model1 <- lm(mobility_change ~ treatment, data = mobility)
anova(model1)
lsmeans(model1, pairwise ~ "treatment")
#code for prob2c

##Getting the packages
install.packages("lme4")
library(lme4)
install.packages("lmerTest")
library("lmerTest")
install.packages("pbkrtest")
```

```

library(pbkrtest)
library(emmeans)
##fitting the model
model2 <- lmer(mobility_change ~ treatment + (1|clinic), data = mobility)
summary(model2)
#ANOVA table
anova(model2)
##Pairwise comparisons
lsmeans(model2, pairwise ~ "treatment")

#code for prob2d
model3 <- lmer(mobility_change ~ treatment + (treatment|clinic), data = mobility)
summary(model3)
anova(model3)

#code for prob2e

#code for prob4a

##import the dataset
library(readr)
spelling <- read_csv("spelling.csv")
##
mean(spelling$age)
sd(spelling$age)
#code for prob4b
install.packages("pROC")
library(pROC)
install.packages("caret")
library(caret)
##create the model
model4 <- glm(correct ~ word_type, data = spelling)
summary(model4)

##create the plot
model4_predict <- predict(model4, type = "response")
model_classify <- ifelse(model4_predict>0.5,1,0)
Model4_ROC <- roc(spelling$correct,model_classify,plot=TRUE, auc.polygon=TRUE,print.auc=TRUE, show.thresh=TRUE)

confusionMatrix(factor(model_classify),
                   factor(spelling$correct), positive = "1")
#code for prob4c
library(lme4)
spelling$subject <- as.factor(spelling$subject)
model5 <- glmer(correct ~ word_type + (1|subject), data = spelling, family = "binomial")
summary(model5)
#code for prob4d

#code for prob4e
library(lme4)
model6 <- glmer(correct ~ word_type + age + order + letterwrite + (1|subject), data = spelling, family = "binomial")
summary(model6)
#code for prob4g

```

```

model6 <- glmer(correct ~ word_type + age + order + letterwrite + (1|subject), data = spelling, family = "binomial")
model6_predict <- predict(model6, type = "response")
model6_classify <- ifelse(model6_predict>0.5,1,0)
Model6_ROC <- roc(spelling$correct,model6_classify,plot=TRUE, auc.polygon=TRUE, print.auc=TRUE, show.thresh=TRUE)

install.packages("caret")
library(caret)

confusionMatrix(factor(model6_classify),
                    factor(spelling$correct), positive = "1")
#code for prob4h
model7 <- glmer(correct ~ word_type + age + order + letterwrite + (age|subject), data = spelling, family = "binomial")
summary(model7)

```