Homework 5

Samuel Jones

15 April 2018

Homework 5

```
Read the introduction, data, and total examination score sections of
Goldstein et al. (1993). Don't bother replicating their results, as we do
not have the same data shown in the paper. The data can be obtained here:

install.packages("mlmRev", repos=(http://lib.stat.cmu.edu/R/CRAN/))
library(mlmRev)

data("Exam", package = "mlmRev")
head(Exam)

The variables in the data are:
-   school [: the student's school]{.p}
-   normexam[: the score on the General Certificate of Secondary Education (GCSE)
    examination at age 16]{.p}
-   schgend [: school gender type (mixed, boys only, girls only)]{.p}
-   schavg [: average London Reading Test intake score for the school]{.p}
-   vr [: verbal reasoning subscore of the London Reading Test at intake]{.p}
-   intake[: overall London Reading Test score at intake]{.p}
-   standLRT [: standardized intake score on the London Reading Test at age 11]{.p}
-   sex [: the student's gender]{.p}
-   type[: whether the school is mixed or single gender]{.p}
-   student [: student ID within school (not unique)]{.p}

    Write a short report telling the five stories of these data.
    Here's what I expect to be included in the report:
```

1. Create histograms of the number of students in each school, the scores on the GCSE exams, the average London Reading Test intake scores for the schools, and the standardized London Reading Test intake scores, summarizing what you see. Next, make tables of the school gender types, verbal reasoning subscores at intake, overall London Reading Test scores at intake, the student's gender, and the type of schools, summarizing what you see. (1 point)

```
#### Histograms of variables #### As usual, I have liberally re-used code from
#### (Davis, 2018)
```

```
png(filename = "HW5P1 Histograms.png", width = 1000, height = 1000, res = 100)
par(cex = 1.3, mfrow = c(2, 2), mar = c(5, 5, 2, 1), cex.lab = 1.5, cex.axis = 1.5)
hist(table(Exam$school), breaks = "FD", xlab = "Number of Students in School", main = "
hist(Exam$normexam, xlab = "Score on the GCSE", breaks = "FD", main = "")
hist(unique(Exam$schavg), xlab = "School Average LRT Intake Score", breaks = "FD",
    main = "", xlim = c(-1, 1), ylim = c(0, 20))
hist(Exam$standLRT, xlab = "Standardized LRT Intake Score", breaks = "FD", main = "")
dev.off()
```

- As seen in Fig. 1, the schools tend to have between 20 and 100 students, leaning towards the 60-80 range. There are 2 schools with less than 20 students and one school with over 180 students. The other 3 variables all appear to be nearly normal distributions.

```
table(Exam$type)
```

```
##
##  Mxd Sngl
## 2169 1890
```

- This table tells us that more than half of the of the students attend a mixed-gender school, but does not break down the single-gender schools.

```
table(Exam$schgend)
```

```
##
## mixed  boys girls
##  2169   513  1377
```

- As previously shown, over half of the students attend schools with mixed genders. This table, however, gives a little more granularity about the single-gender schools. There are almost 3 girls who attend an all-girl school for every boy who attends an all-boy school. If these two variables are used, they should show a high degree of correlation.

```
table(Exam$vr)
```

```
##
## bottom 25%    mid 50%    top 25%
##        640       2263       1156
```

- Most students are in the middle 50% for verbal reasoning scores. The top 25% outnumber the bottom 25% by almost 2 to 1.

```
table(Exam$intake)
```

```
##
## bottom 25%    mid 50%    top 25%
##       1176       2344        539
```

- Most of the students are also in the middle 50% for the London Reading Test score. The bottom 25% outnumber the top 25% by almost 2 to 1.

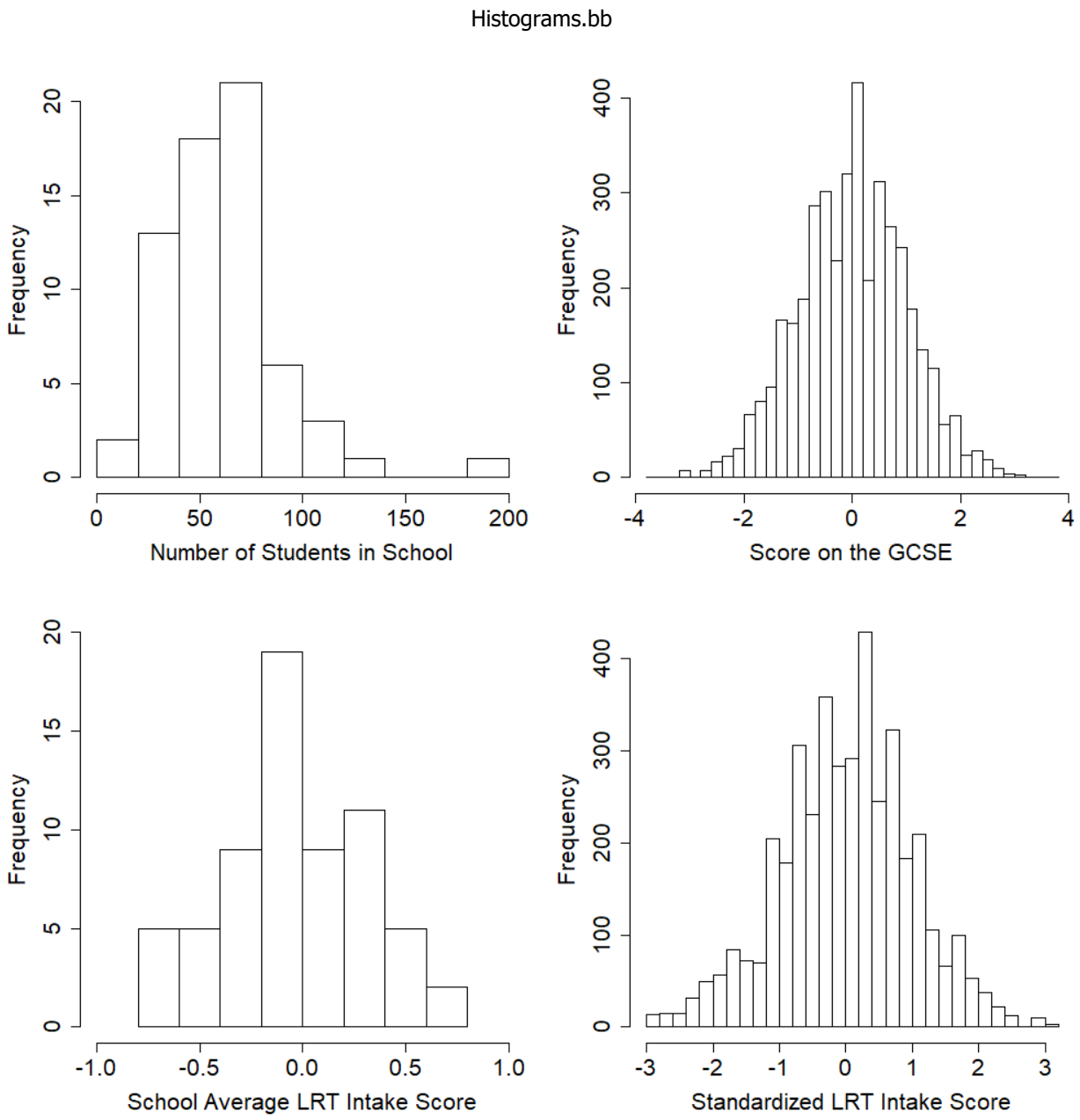


Figure 1: Histograms of the number of students in each school, score on the GCSE, school average LRT intake score and standardized LRT intake school.

	Complete Pooling	Partial Pooling
(Intercept)	−0.00	−0.01
	(0.02)	(0.05)
R^2	0.00	
Adj. R^2	0.00	
Num. obs.	4059	4059
RMSE	1.00	
AIC		11020.65
BIC		11039.58
Log Likelihood		-5507.33
Num. groups: school		65
Var: school (Intercept)		0.17
Var: Residual		0.85

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 1: Problem 2 Complete and Partial Pooling Results

```
table(Exam$sex)
```

```
##
##    F    M
## 2436 1623
```

- About two thirds of the students are female.

2. Conduct complete, no, and partial pooling regressions of the GCSE exam scores using only an intercept, summarizing the results. Examine the distribution of the no pooling and partial pooling intercepts. How are they similar or different? Provide the fitted intercept for an example school using no pooling and partial pooling. (1 point)

```
#### P-2 Complete, no, partial pooling #### Complete pooling
cp2 <- lm(normexam ~ 1, data = Exam)
# No pooling
np2 <- lm(normexam ~ factor(school) - 1, data = Exam)
# No pooling intercepts
np2.ints <- coef(np2)
# Average of the no pooling intercepts
np2.avg <- mean(np2.ints)
# Partial pooling
pp2 <- lmer(normexam ~ 1 + (1 | school), data = Exam)
# Partial pooling intercepts
pp2.ints <- unlist(coef(pp2)$school)
# Average of the partial pooling intercepts
pp2.avg <- mean(pp2.ints)
```

- As seen in Table 1, the GCSE score was approximately zero for complete pooling and -0.01 for partial pooling. It's difficult to display the results for the no pooling case due to the number of schools, but the GCSE score was -0.23.

```
#### P-2 Histograms of the No pooling and partial pooling intercepts ####
png(filename = "HW5P2 NP PP Ints.png", width = 1000, height = 1000, res = 100)
par(cex = 1.3, mfrow = c(2, 1), mar = c(5, 5, 2, 1), cex.lab = 1.5, cex.axis = 1.5)
hist(np2.ints, xlab = "No Pooling Intercepts", breaks = "FD", main = "", xlim = c(-1.5,
    1.5), ylim = c(0, 20))
hist(pp2.ints, xlab = "Partial Pooling Intercepts", breaks = "FD", main = "", xlim = c(-
    1.5), ylim = c(0, 20))
dev.off()
```

- As seen in Figure 2, the distribution appear to be very similar. They are both bimodal distributions with modes around -0.5 and zero. The partial pooling is slightly more compressed.
- School 5's intercepts are 0.40 for no pooling and 0.35 for partial pooling.

3. For the previous regression, explain when the partially pooled school intercepts will be closer to complete pooling or no pooling. (1 point)

- Partially pooled intercepts are closer to complete pooling intercepts when the sample size n_j in a school j is less than the dependent variable residual divided by the pooling group intercept as in the following equation:

$$\frac{\sigma_y^2}{\sigma_\alpha^2} = \frac{0.85}{0.17} = 5$$

4. Add the student's standardized intake score on the London Reading Test at age 11 as a predictor to the complete, no, and partial pooling regressions. How do the intercepts change as the predictor is added? How do the complete, no, and partial pooling regressions compare? Should we be concerned about any correlation between the students' standardized intake score and the school level intercepts (if there is any correlation)? (1 point)

```
#### P4 Add intake score on the London Reading Test to regressions #### Complete
#### pooling with predictor
cp4 <- lm(normexam ~ 1 + standLRT, data = Exam)
# No pooling with predictor
np4 <- lm(normexam ~ factor(school) - 1 + standLRT, data = Exam)
# No pooling intercepts
np4.ints <- coef(np4)
# [1:65] Average no pooling intercepts
np4.avg <- mean(np4.ints)
# Partial pooling with predictor
```

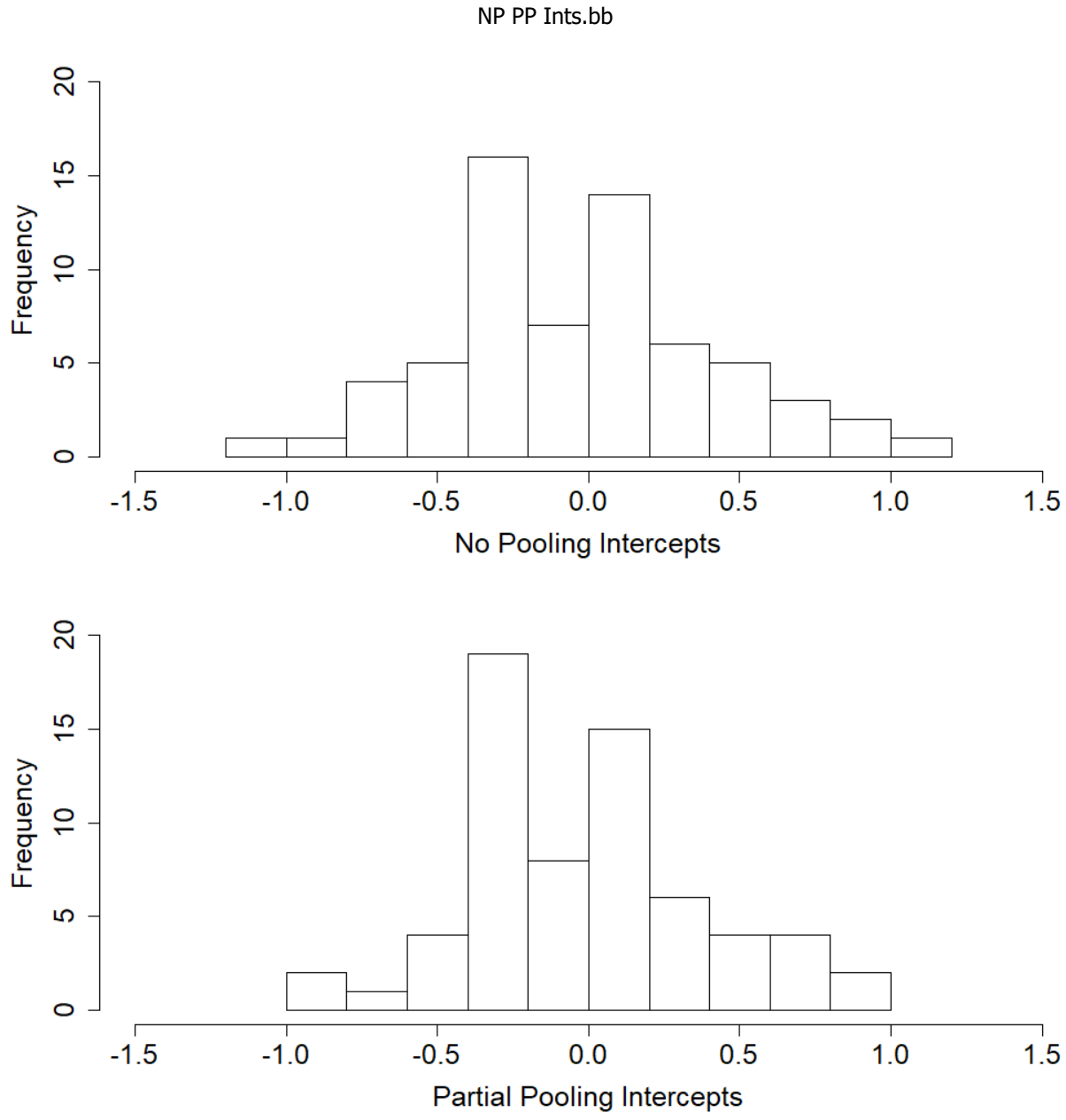


Figure 2: Histograms of the no pooling and parital pooling intercepts

	Complete Pooling	Partial Pooling
(Intercept)	-0.00	0.00
	(0.01)	(0.04)
standLRT	0.60^{***}	0.56^{***}
	(0.01)	(0.01)
R^2	0.35	
Adj. R^2	0.35	
Num. obs.	4059	4059
RMSE	0.81	
AIC		9376.77
BIC		9402.00
Log Likelihood		-4684.38
Num. groups: school		65
Var: school (Intercept)		0.09
Var: Residual		0.57

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 2: Problem 4 Complete and Partial Pooling Results

```
pp4 <- lmer(normexam ~ 1 + (1 | school) + standLRT, data = Exam)
# Get the partial pooling intercepts
pp4.ints <- unlist(coef(pp4)$school[1])
# Get the average of the partial pooling intercepts
pp4.avg <- mean(pp4.ints)
```

- As seen in Table 2, the complete pooling is still zero. The no pooling and partial pooling average intercepts are both 0.00.

```
#### P-2 Histograms of the No pooling and partial pooling intercepts ####
png(filename = "HW5P4 NP PP Ints.png", width = 1000, height = 1000, res = 100)
par(cex = 1.3, mfrow = c(2, 2), mar = c(5, 5, 2, 1), cex.lab = 1.5, cex.axis = 1.5)
hist(np2.ints, xlab = "No Pooling Intercepts", breaks = "FD", main = "", xlim = c(-1.5,
    1.5), ylim = c(0, 25))
hist(np4.ints, xlab = "No Pooling Intercepts w/ Standard LRT", breaks = "FD", main = "",
    xlim = c(-1.5, 1.5), ylim = c(0, 25))
hist(pp2.ints, xlab = "Partial Pooling Intercepts", breaks = "FD", main = "", xlim = c(-
    1.5), ylim = c(0, 25))
hist(pp4.ints, xlab = "Partial Pooling Intercepts w/ Standard LRT", breaks = "FD",
    main = "", xlim = c(-1.5, 1.5), ylim = c(0, 25))
dev.off()
```

- As seen in Figure 3, the additional regressor appears to have had a nearly identical effect. The partial pooling is still more compressed than the no pooling distribution.
- As seen in Table 2, the coefficient of the standardized LRT is 0.60 for complete pooling and 0.56 for partial pooling. I also looked up the value in the no pooling case. It was 0.56. So all three are very similar.

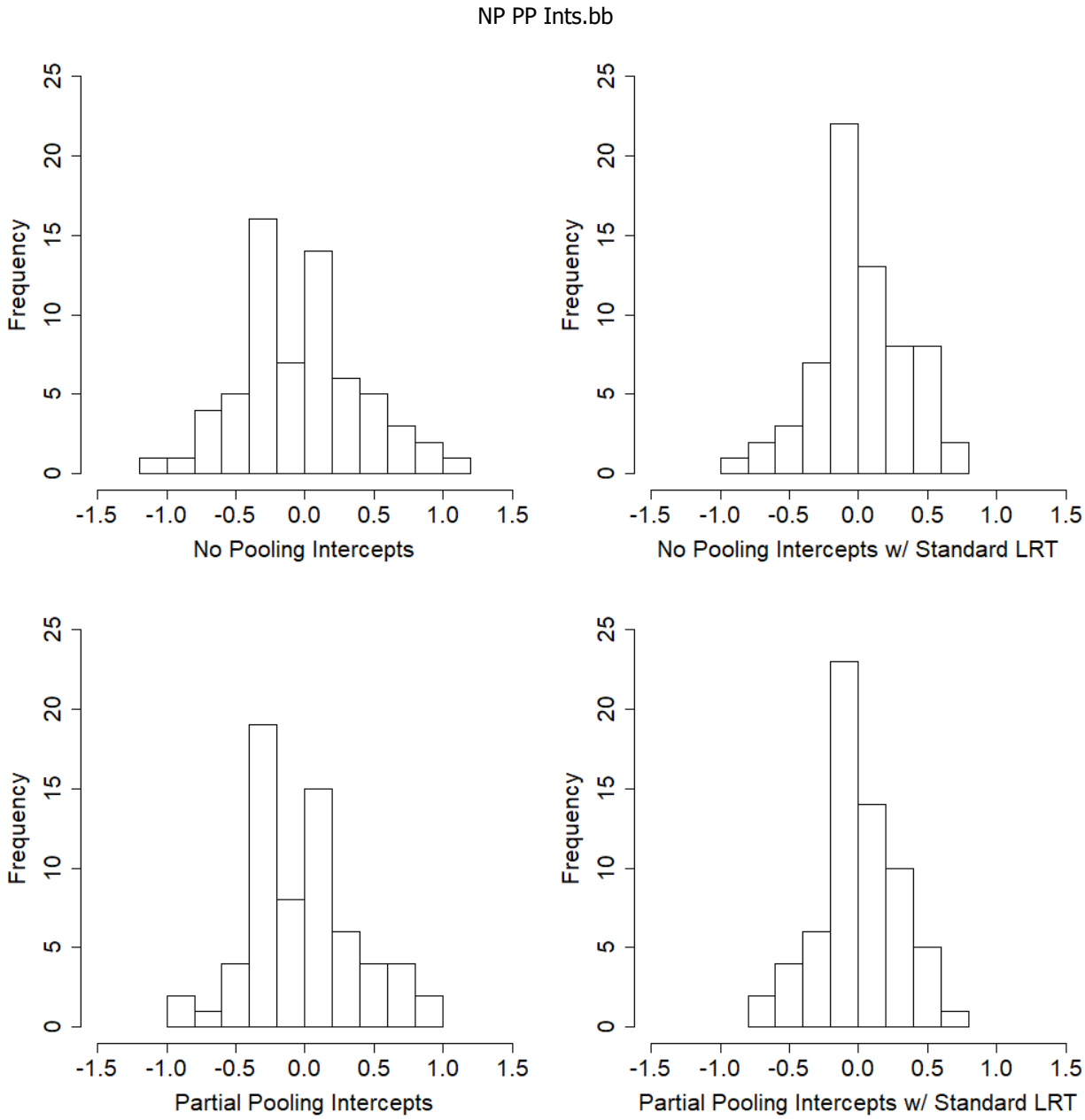


Figure 3: Histograms of the no pooling and parital pooling intercepts

	Partial Pooling w/o avg	Partial Pooling w/avg
(Intercept)	0.00	0.01
	(0.04)	(0.04)
standLRT	0.56***	0.56***
	(0.01)	(0.01)
schavg		0.36**
		(0.11)
AIC	9376.77	9371.86
BIC	9402.00	9403.41
Log Likelihood	-4684.38	-4680.93
Num. obs.	4059	4059
Num. groups: school	65	65
Var: school (Intercept)	0.09	0.08
Var: Residual	0.57	0.57

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 3: Problem 5 Partial Pooling Results with and without School Avg LRT

- I don't think we have to worry about correlation between them because the no pooling and partial pooling coefficients are very similar, 0.56 and 0.56.

5. Add the average London Reading Test intake score for the school as a group-level predictor in a multi-level model. How does the variance of the intercepts change? Plot the regression line of the average London Reading Test scores against the no pooled and partially pooled intercepts, summarizing what you see. Write the fitted regression line for an example school. (1 point)

```
#### P5 Add the avg London Reading Test intake score for the school #### Partial
#### pooling with predictors
pp5 <- lmer(normexam ~ 1 + schavg + (1 | school) + standLRT, data = Exam)
# Partial pooling intercepts
pp5.ints <- coef(pp5)$school[1] + coef(pp5)$school[3] * unique(Exam$schavg)
pp5.ints <- unlist(pp5.ints)
# Average the partial pooling intercepts
pp5.avg <- mean(pp5.ints)
```

- As seen in Table 3, very little changed. By grabbing the summaries for each, we can see that the variance of the school intercept variance went from 0.093 to 0.078. So it is slightly better with the school average.

```
#### P5 Plots #####
png(filename = "HW5P5 PP with without Avg LRT.png", width = 1000, height = 1000,
    res = 100)
par(cex = 1.3, mfrow = c(2, 1), mar = c(5, 5, 2, 1), cex.lab = 1.2, cex.axis = 1.2)
plot(unique(Exam$schavg), np4.ints, xlab = "School Average LRT Intake Score", ylab = "No
    pch = 19, ylim = c(-1, 1), xlim = c(-1, 1))
```

	Partial Pooling w/ avg	Partial Pooling varied
(Intercept)	0.01	-0.00
	(0.04)	(0.04)
schavg	0.36^{**}	0.29^{**}
	(0.11)	(0.11)
standLRT	0.56^{***}	0.55^{***}
	(0.01)	(0.02)
AIC	9371.86	9337.89
BIC	9403.41	9382.05
Log Likelihood	-4680.93	-4661.94
Num. obs.	4059	4059
Num. groups: school	65	65
Var: school (Intercept)	0.08	0.08
Var: Residual	0.57	0.55
Var: school standLRT		0.02
Cov: school (Intercept) standLRT		0.01

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 4: Problem 6 Partial Pooling Results when LRT varies at school level

```
abline(lm(np4.ints ~ unique(Exam$schavg)))
plot(unique(Exam$schavg), pp5.ints, xlab = "School Average LRT Intake Score", ylab = "Pa
    pch = 19, ylim = c(-1, 1), xlim = c(-1, 1))
abline(lm(pp5.ints ~ unique(Exam$schavg)))
dev.off()
```

- As seen in Figure 4, the additional regressor, as previously noted, eliminated some variance. There are observations that have clearly been pulled closer to the regressor lines.
- The regression line for school five is:

$$GCSE_{i5} = 0.01 + u_5 + 0.36(schavg_5) + 0.56(standLRT_i)$$

$$GCSE_{i5} = 0.01 + 0.44 + 0.36(0.21) + 0.56(standLRT_i)$$

$$GCSE_{i5} = 0.53 + 0.56(standLRT_i)$$

6. Extend the previous model by allowing the standardized intake score to vary at the school level, summarizing the results. Write the fitted regression line for an example school. Interpret the variance components of the new model. (1 point)

```
#### P6 Allow standardized intake score to vary at the school level ####
pp6 <- lmer(normexam ~ 1 + standLRT + schavg + (1 + standLRT | school), data = Exam)
```

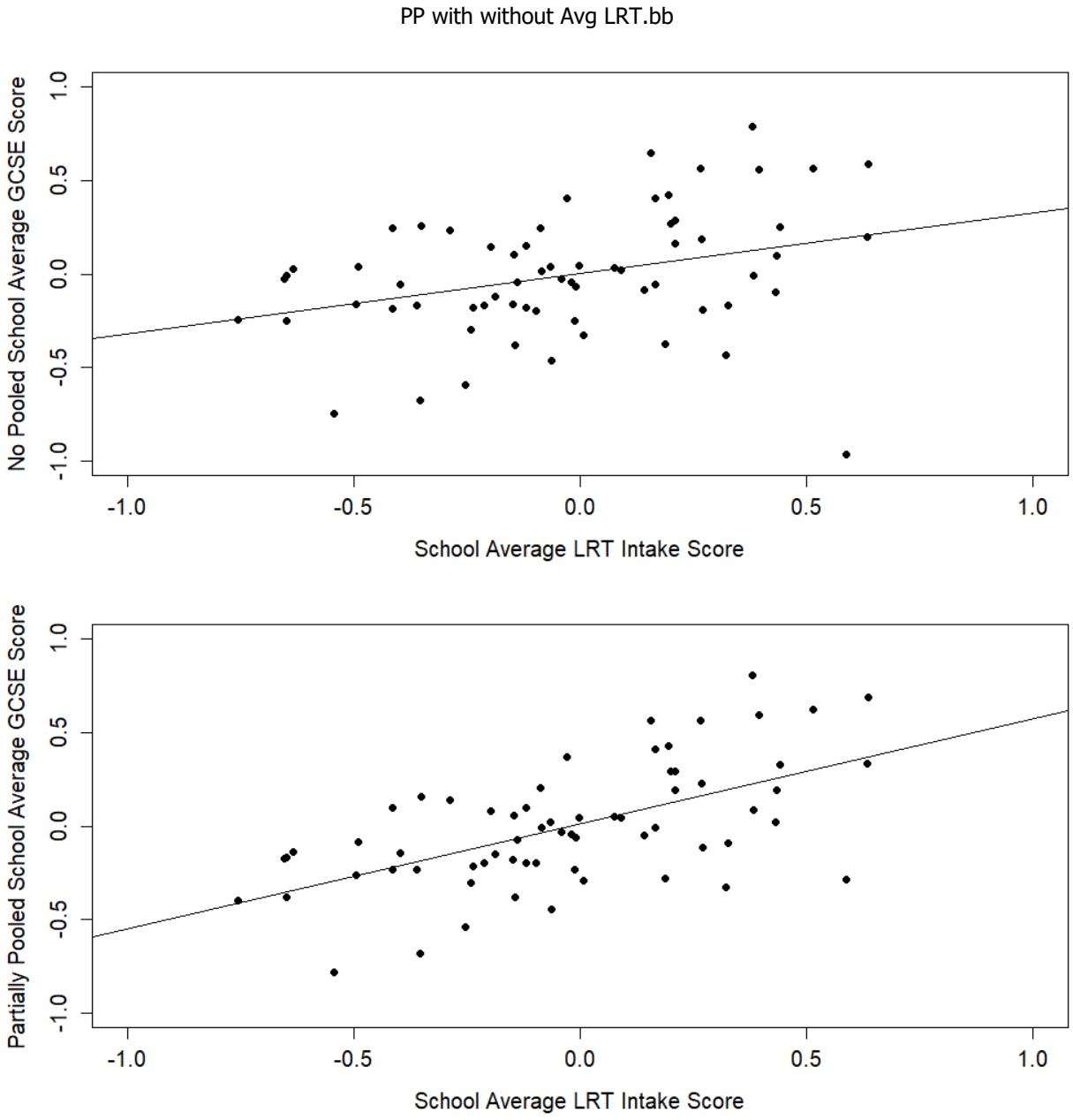


Figure 4: Regression lines of the average LRT scores against the no pooled and partially pooled intercepts

- As seen in Table 4, the intercept went from 0.01 to -0.00. The schavg coefficient went from 0.36 to 0.29. The standLRT coefficient went from 0.56 to 0.55. The new model also has 3 components with variance and one with covariance.
- The regression line for school five is now:

$$GCSE_{i5} = -0.00 + u_5 + 0.29(schavg_5) + 0.55(standLRT_i)$$

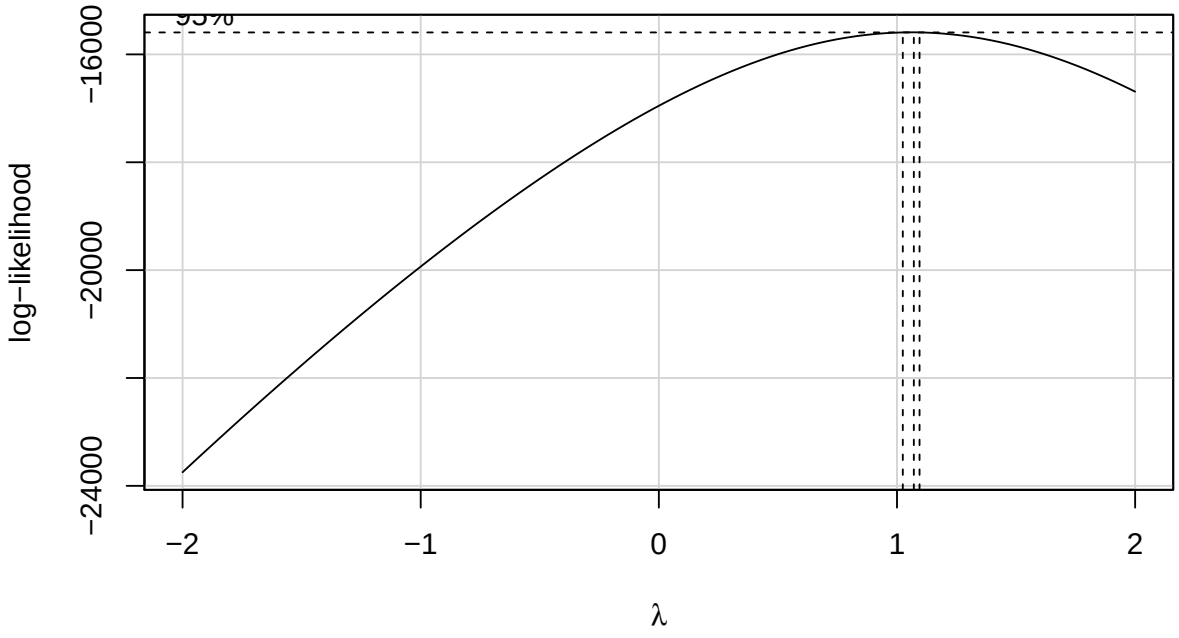
$$GCSE_{i5} = -0.00 + 0.92 + 0.29(0.21) + 0.55(standLRT_i)$$

$$GCSE_{i5} = 0.98 + 0.56(standLRT_i)$$

- As previously mentioned there are now three variance components. They are: varying intercepts $(\sigma^2_\alpha) = 0.08$, varying residuals $(\sigma^2_y) = 0.55$ and varying slopes $(\sigma^2_\beta) = 0.02$. The interclass correlation is $\rho = \frac{0.08}{0.08+0.55} = 0.13$. The covariance mentioned earlier is the interaction between the intercepts and slopes.

7. Use level 1 diagnostics to discover any problems with the first level of the previous model. Allow for potential transformations of the dependent and independent variables. What revised model do you suggest? (1 point)

```
#### P7 Level 1 diagnostics ####
bc7 <- with(Exam, boxCox(normexam ~ factor(school) * standLRT, data = Exam, family = "y
```



```
bc7$x[bc7$y == max(bc7$y)]
```

```
## [1] 1.070707
```

```
np7 <- lm(normexam ~ factor(school) * standLRT, data = Exam)
max(na.omit(cooks.distance(np7)))
```

```
## [1] 11.86334
```

```
median(na.omit(cooks.distance(np7)))
```

```
## [1] 8.224501e-05
```

- The maximum likelihood estimate for λ is 1.07. So, no transformation is necessary. There may be influential points according to Cook's distance, but it's not clear yet.

```
png(filename = "HW5P7 JK Plots.png", width = 1000, height = 1000, res = 100)
par(cex = 1.3, mar = c(5, 5, 2, 1), mfrow = c(2, 2), cex.lab = 1.5, cex.axis = 1.5)
plot(fitted(np7), rstudent(np7), xlab = "Fitted Values", ylab = "Jackknife Residuals",
    pch = 19, col = rgb(0, 0, 0, 0.1))
abline(h = 0, lty = 2, col = "red", lwd = 2)
lines(lowess(fitted(np7) ~ rstudent(np7)), col = "blue", lwd = 2)
plot(Exam$standLRT, rstudent(np7), xlab = "Standardized LRT score", ylab = "Jackknife LS
    pch = 19, col = rgb(0, 0, 0, 0.1))
abline(h = 0, lty = 2, col = "red", lwd = 2)
lines(lowess(Exam$standLRT ~ rstudent(np7)), col = "blue", lwd = 2)
qqPlot(np7, id.n = 3, distribution = "t", df = df.residual(np7), ylab = "Jackknife LS Re
dev.off()
```

- As seen in Figure 5, the plots all look as if no transformations are needed. There are no clear patterns and the LS residual vs. t quantiles looks normal.

8. Use level 2 diagnostics to discover problems with the second level of the previous model. Do the Empirical Bayes residuals look appropriate? Are there any potentially influential schools? (1 point)

```
#### P8 Level 2 diagnostics-Empirical Bayes ####
eb8 <- HLMresid(pp6, level = "school", type = "EB")
```

```
png(filename = "HW5P8 EB Hist.png", width = 1000, height = 1000, res = 100)
par(cex = 1.3, mar = c(5, 4, 2, 1), mfrow = c(2, 1), cex.lab = 1.3, cex.axis = 1.5)
hist(eb8[, 1], breaks = "FD", xlab = "Empirical Bayes Intercepts", main = "")
hist(eb8[, 2], breaks = "FD", xlab = "Emperical Bayes Slopes", main = "")
dev.off()
```

- As seen in Figure 6, the Emprical Bayes residual histograms look mostly normal. The intercepts are more dense at the low end and the slopes appear to have an outlier at the high end.

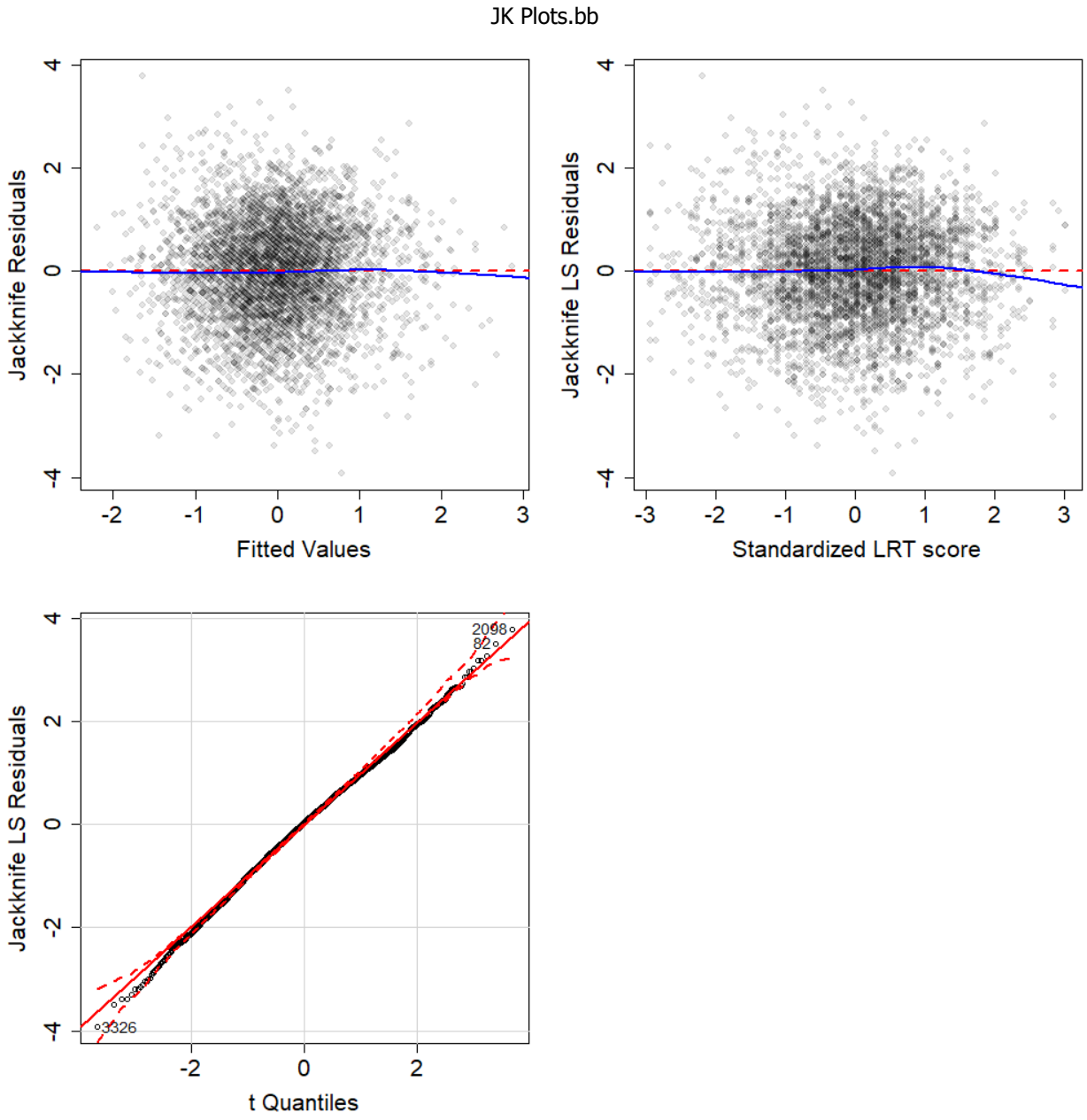


Figure 5: No Pooling Jacknife Residuals

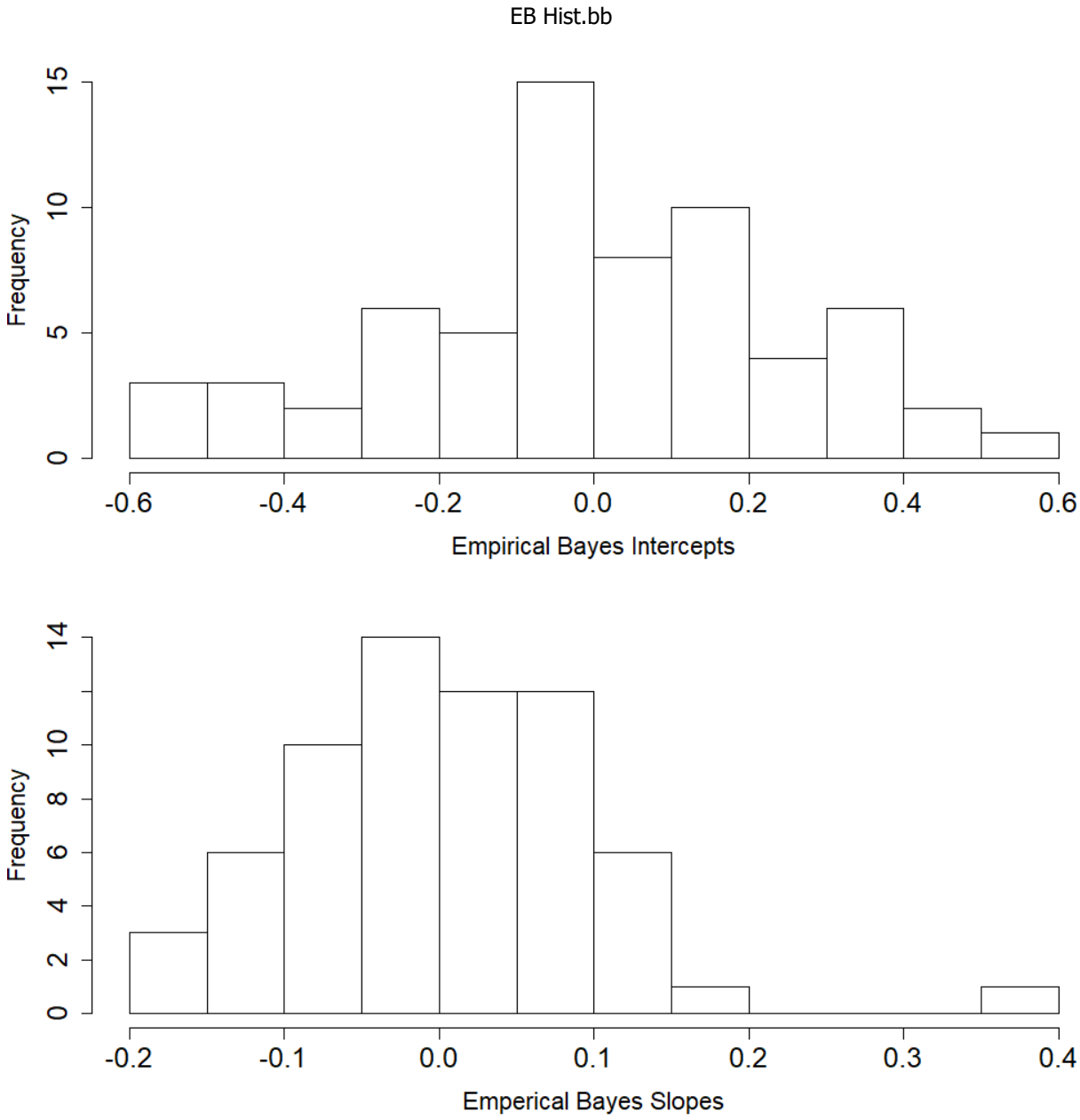


Figure 6: Empirical Bayes Histograms

```
png(filename = "HW5P8 EB Resid Quantiles.png", width = 1000, height = 1000, res = 100)
par(cex = 1.3, mar = c(5, 5, 2, 1), mfrow = c(2, 2), cex.lab = 1.5, cex.axis = 1.5)
plot(unique(Exam$schavg), eb8[, 1], pch = 19, col = rgb(0, 0, 0, 0.5), xlab = "School-le
    ylab = "Empirical Bayes Intercept Residuals", xlim = c(-1, 1), ylim = c(-1, 1))
abline(h = 0, col = "red", lty = 2)
lines(lowess(unique(Exam$schavg), eb8[, 1]), col = "blue", lwd = 2)
plot(unique(Exam$schavg), eb8[, 2], pch = 19, col = rgb(0, 0, 0, 0.5), xlab = "School-le
    ylab = "Empirical Bayes Slope Residuals", xlim = c(-1, 1), ylim = c(-0.5, 0.5))
abline(h = 0, col = "red", lty = 2)
lines(lowess(unique(Exam$schavg), eb8[, 2]), col = "blue", lwd = 2)
qqPlot(eb8[, 1], ylab = "Empirical Bayes Intercept Residuals", xlab = "Normal Quantiles"
qqPlot(eb8[, 2], ylab = "Empirical Bayes Slope Residuals", xlab = "Normal Quantiles")
dev.off()
```

- As seen in Figure 6, the plots shows some curvature. Transformation still seems unnecessary. Looking at the QQ plots, there appear to be three or four observations that are too low or too high and should probably be excluded from our calculations. These three or four schools could be overly influential.

9. Compare a no pooling dummy variable version of your revised model with a partial pooling version. How do the standard errors of your no pooling model change when using heteroskedasticity and cluster-robust standard errors? What does this suggest about potential model misspecification? Should we use our partial pooling or no pooling model? (1 point)

```
#### P9 Compare a np dummy variable w/ a pp version ####
np9 <- lm(normexam ~ standLRT + factor(school) - 1, data = Exam)
pp9 <- lmer(normexam ~ standLRT + 1 + (1 + standLRT | school), data = Exam)
het9 <- coeftest(np9, vcov = vcovHC(np9, type = "HC0"))
clust9 <- cluster.vcov(np9, Exam$school, df_correction = TRUE)
coeftest(np9, vcovCL)
summary(np9)
summary(pp9)
summary(het9)
summary(clust9)
```

- The standard errors for the standardized LRT for each model is as follows: 0.0125 for no pooling, 0.020 for partial pooling,for heteroskedasticity robust standard errors on the no pooling regression had a median of 0.095, and the cluster robust standard error had a median of 0.000017. Since the heteroskedastic robust was actually worse, it's probably best to stick with the partial pooling model.

EB Resid Quantiles.bb

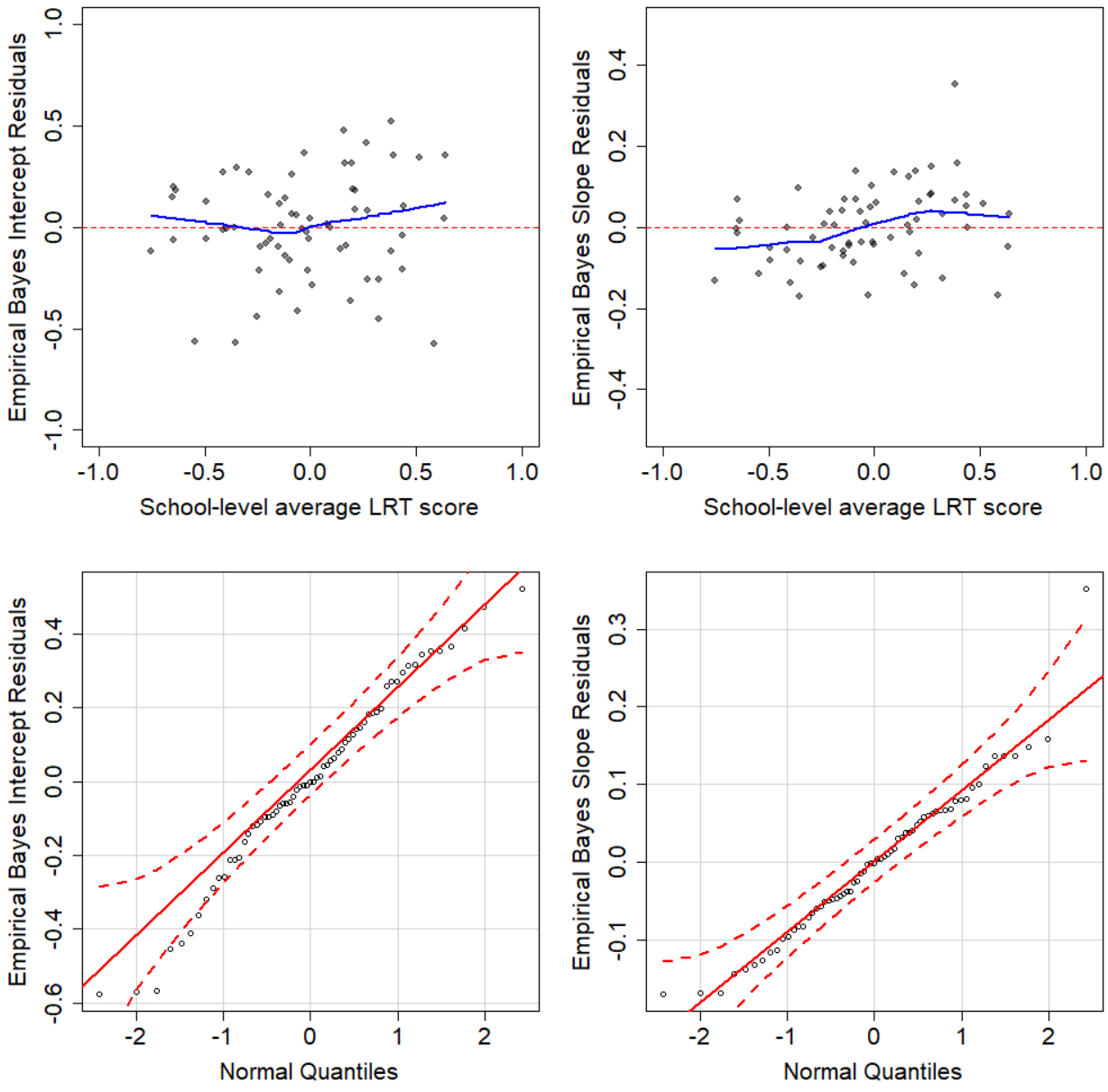


Figure 7: Empirical Bayes Histograms

10. Compare your model to a simpler version using level 1 and level 2 leave-one-out-cross-validation. Summarize their performance. Which model do you prefer? (1 point)

```
#### P10 Compare models #### Based on Week12.r (Davis, 2018)
simple <- c()
complex <- c()
for (i in 1:nrow(Exam)) {
    # Train complex model dropping the ith observation
    exam.complex <- lmer(normexam ~ standLRT + 1 + schavg + (1 + standLRT | school),
        data = Exam[-i, ])
    # Find the school from the dropped observation
    school <- Exam$school[i]
    # Let's break down the prediction into each of its parts There's the school
    # intercept
    sch.int <- coef(exam.complex)$school[school, 1]
    # Plus the school-average prediction of the intercept
    avg.int <- coef(exam.complex)$school[1, 3] * Exam$schavg[i]
    # Plus the school slope
    sch.slope <- coef(exam.complex)$school[school, 2] * Exam$standLRT[i]
    # The prediction is the sum of these three components
    pred.comp <- sch.int + avg.int + sch.slope
    # Calculate the squared residual
    comp.sq.resid <- (Exam$normexam[i] - pred.comp)^2
    # Add the squared residual to the cv vector
    complex <- append(complex, comp.sq.resid)
    # Train simple model dropping the ith observation
    exam.simple <- lmer(normexam ~ standLRT + 1 + schavg + (1 | school), data = Exam[-i,
        ])
    # Let's break down the prediction into each of its parts There's the school
    # intercept
    sch.int.1 <- coef(exam.simple)$school[school, 1]
    # Plus the school-average prediction of the intercept
    avg.int.1 <- coef(exam.simple)$school[1, 3] * Exam$schavg[i]
    # Plus the overall slope
    simp.slope1 <- coef(exam.simple)$school[school, 2] * Exam$standLRT[i]
    # The prediction is the sum of these three components
    pred.simp <- sch.int.1 + avg.int.1 + simp.slope1
    # Calculate the squared residual
    simp.sq.resid <- (Exam$normexam[i] - pred.simp)^2
    # Add the squared residual to the cv vector
    simple <- append(simple, simp.sq.resid)
}
comp.rMSE <- sqrt(sum(complex)/length(complex))
simp.rMSE <- sqrt(sum(simple)/length(simple))
```

- The simple model ended up with an rMSE of 0.758 vs an rMSE of 0.753 for the more complex model. So the two models are fairly close, but the complex model performed slightly better. I prefer the simple model because it has nearly the same level of performance, but takes less computation power.

11. Explain why econometricians, statisticians, and machine learning researchers are so hungry. (1 bonus point)

- They spend so little time sleeping, they have to increase their caloric intake to compensate for lack of rest.

Make sure your report includes reproducible R code, either as an appendix, as a footnote through Harvard's Dataverse (or some other site of your choice), or through CMU's dropbox.