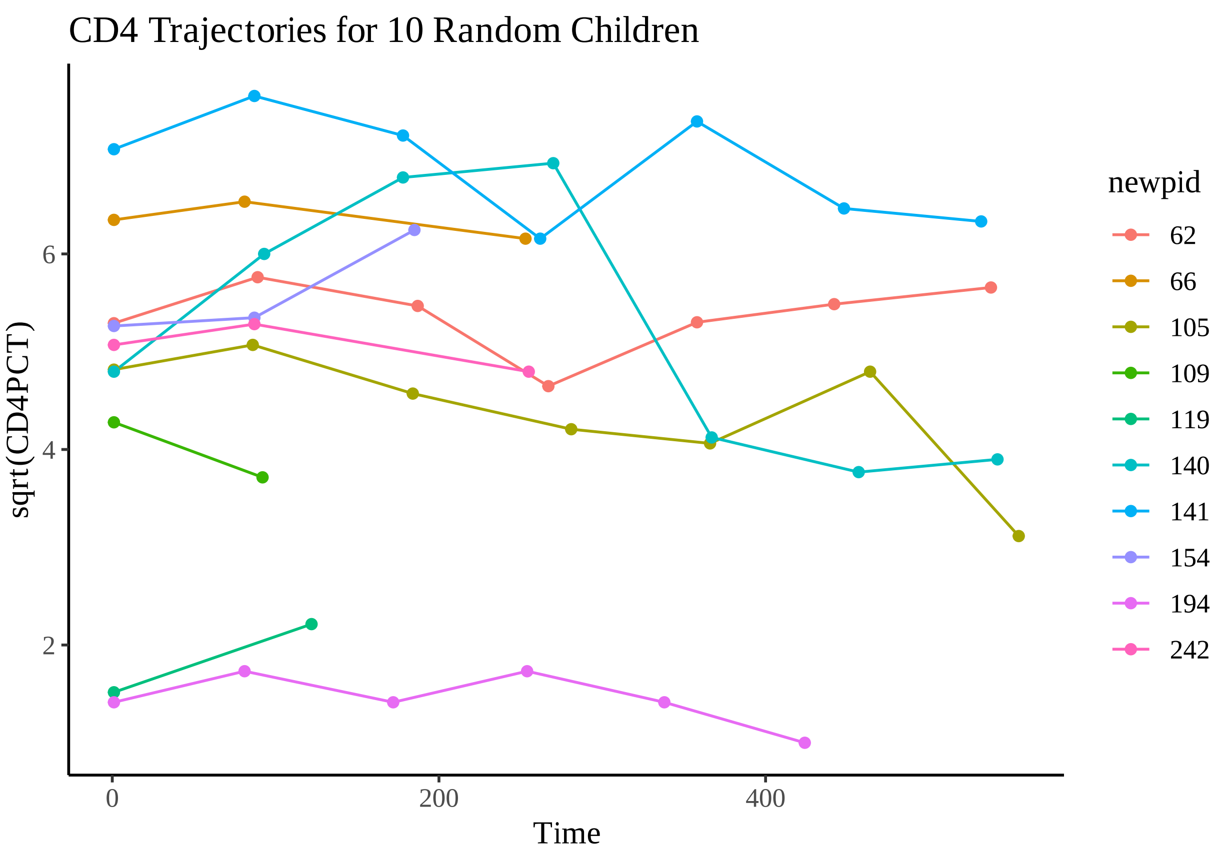
# PHP 2517 Homework #1

*Blain Morin*

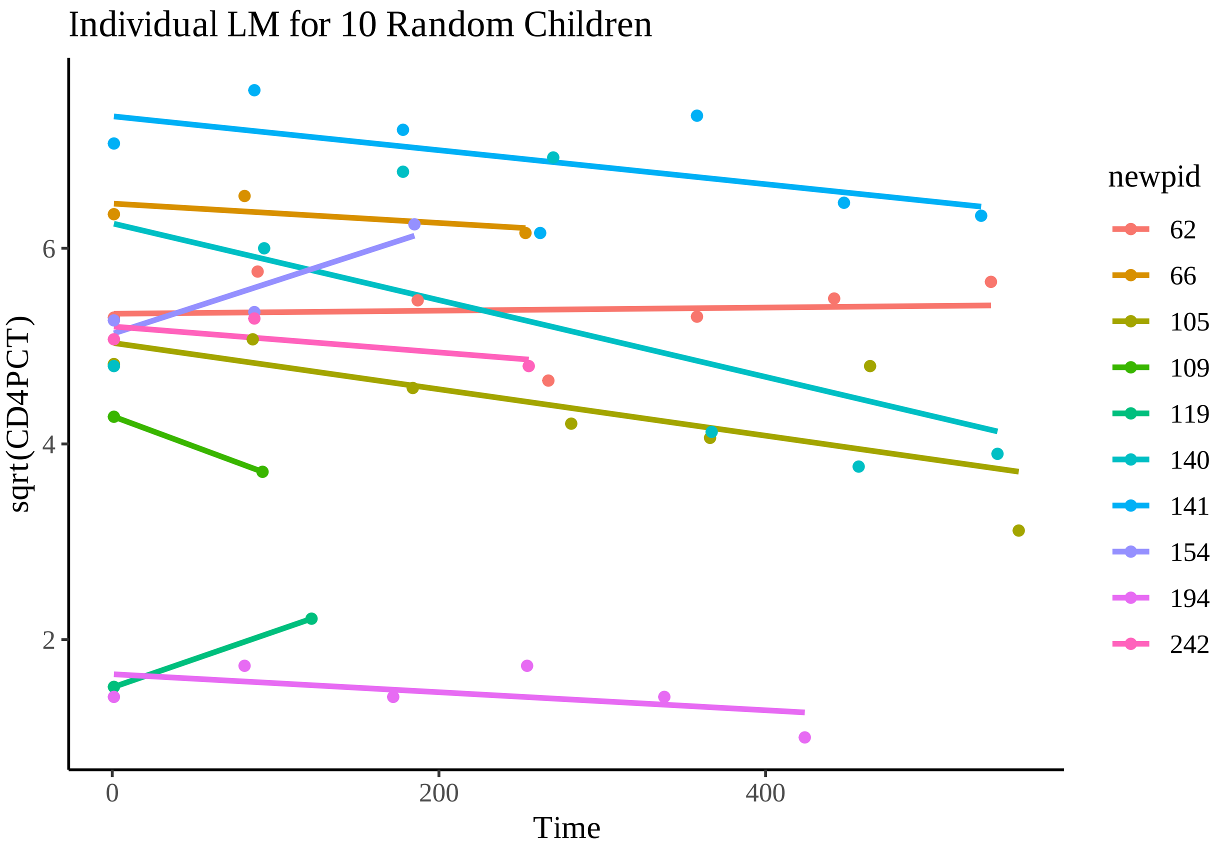
*February 11, 2019*

## Q1 GH Chapter 11: Exercise 4

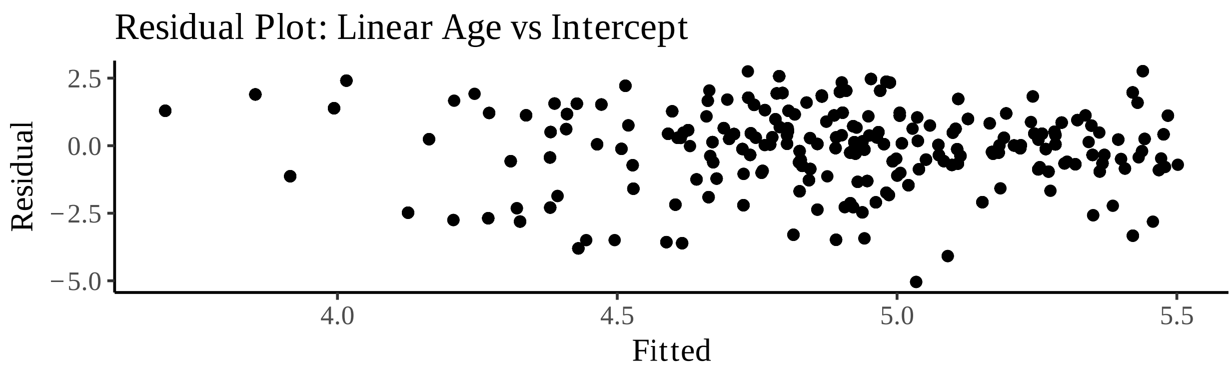
### a.) Graph the outcome (the CD4 percentage, on the square root scale) for 10 children as a function of time.

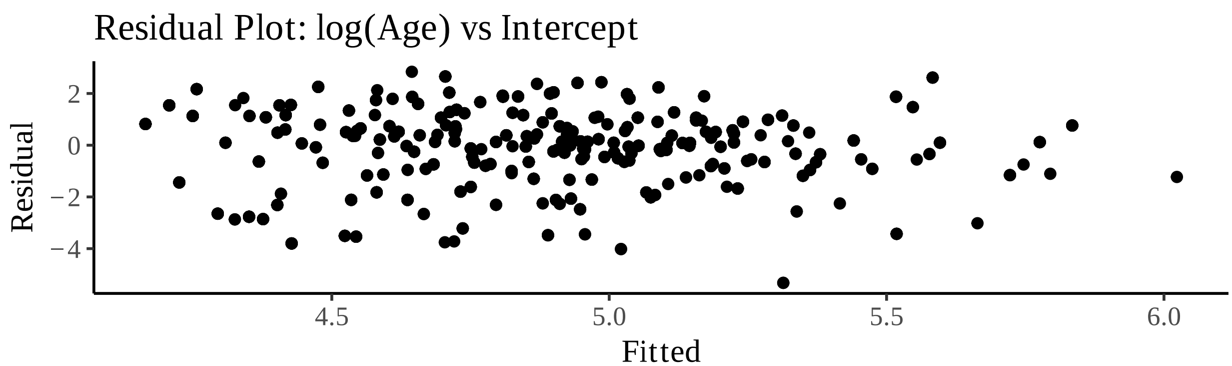


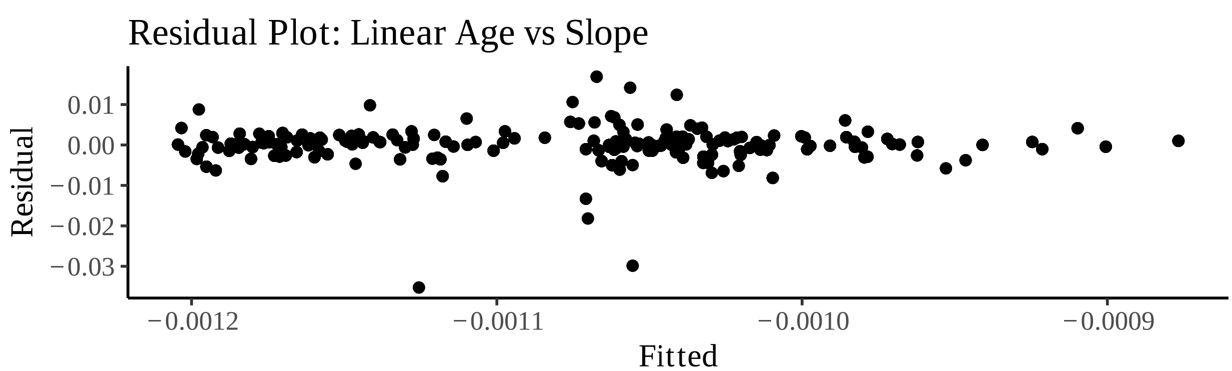
**b.) Each child's data has a time course that can be summarized by a linear fit. Estimate these lines and plot them for 10 children.**



**c.) Set up a model for the children's slopes and intercepts as a function of the treatment and age at baseline. Estimate this model using the two-step procedure–first estimate the intercept and slope separately for each child, then fit the between-child models using the point estimates from the first step.**



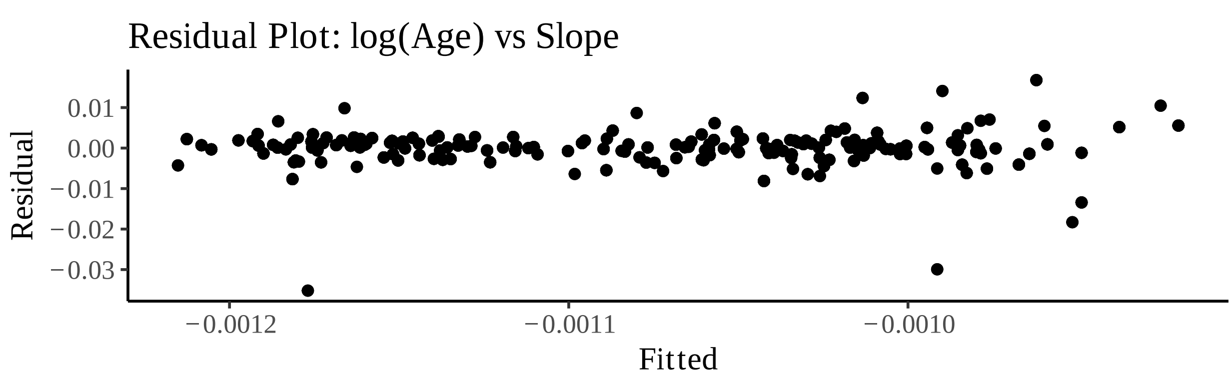



Table 1: Regression on Intercept and Slope

| | *Dependent variable:* | |
|---|---|---|
| | beta0 | beta1 |
| | (1) | (2) |
| Treatment = 2 | 0.451*** | −0.0001 |
| | (0.092) | (0.0003) |
| log(baseage) | −0.349*** | −0.0001 |
| | (0.065) | (0.0002) |
| Constant | 5.014*** | −0.001*** |
| | (0.089) | (0.0002) |
| Observations | 960 | 936 |
| $R^2$ | 0.052 | 0.0004 |
| Adjusted $R^2$ | 0.050 | −0.002 |
| Residual Std. Error | 1.428 (df = 957) | 0.004 (df = 933) |
| F Statistic | 26.155*** (df = 2; 957) | 0.170 (df = 2; 933) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |
| | | Individuals with only one observation do not have a slope estimate. |

# Q2 GH Chapter 12: Exercise 2

### a.) Write a model predicting CD4 percentage as a function of time with varying intercepts across children. Fit using lmer() and interpret the coefficient for time.

$$Level1 : CD4PCT_{ij} \sim \mathcal{N}(\alpha_j + \beta time_{ij}, \sigma^2)$$

$$Level2 : \alpha_j \sim \mathcal{N}(\mu, \tau^2)$$

```
rand.int.a = lmer(sqrt(CD4PCT) ~ time + (1 | newpid), data = cd4)
texreg(rand.int.a, caption.above = TRUE,
       caption = "Random Intercept Model",
       custom.model.names = "sqrt(CD4PCT)",
       float.pos = 'H')
```

Table 2: Random Intercept Model

| | sqrt(CD4PCT) |
|---|---|
| (Intercept) | $4.80^{***}$ |
| | (0.10) |
| time | $-0.00^{***}$ |
| | (0.00) |
| AIC | 2825.04 |
| BIC | 2844.50 |
| Log Likelihood | -1408.52 |
| Num. obs. | 960 |
| Num. groups: newpid | 221 |
| Var: newpid (Intercept) | 1.98 |
| Var: Residual | 0.59 |

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

### b.) Extend the model in (a) to include child-level predictors (that is, group-level predictors) for treatment and age at baseline. Fit using lmer() and interpret the coefficients on time, treatment, and age at baseline.

$$Level1 : CD4PCT_{ij} \sim \mathcal{N}(\alpha_j + \beta time_{ij}, \sigma^2)$$

$$Level2 : \alpha_j \sim \mathcal{N}(\mu + \gamma_0 Treat + \gamma_1 Age, \tau^2)$$

```r
rand.int.b = lmer(sqrt(CD4PCT) ~ time + treatmnt + log(baseage) + (1|newpid), data = cd4)

texreg(rand.int.b, caption.above = TRUE,
       caption = "Random Intercept Model with Child Level Predictors",
       custom.model.names = "sqrt(CD4PCT)",
       float.pos = 'H')
```

Table 3: Random Intercept Model with Child Level Predictors

| | sqrt(CD4PCT) |
|---|---|
| (Intercept) | $4.89^{***}$ |
| | (0.18) |
| time | $-0.00^{***}$ |
| | (0.00) |
| treatmnt2 | 0.32 |
| | (0.20) |
| log(baseage) | $-0.26^{*}$ |
| | (0.13) |
| AIC | 2825.90 |
| BIC | 2855.10 |
| Log Likelihood | -1406.95 |
| Num. obs. | 960 |
| Num. groups: newpid | 221 |
| Var: newpid (Intercept) | 1.93 |
| Var: Residual | 0.59 |

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

c.) Investigate the change in partial pooling from (a) to (b) both graphically and numerically.

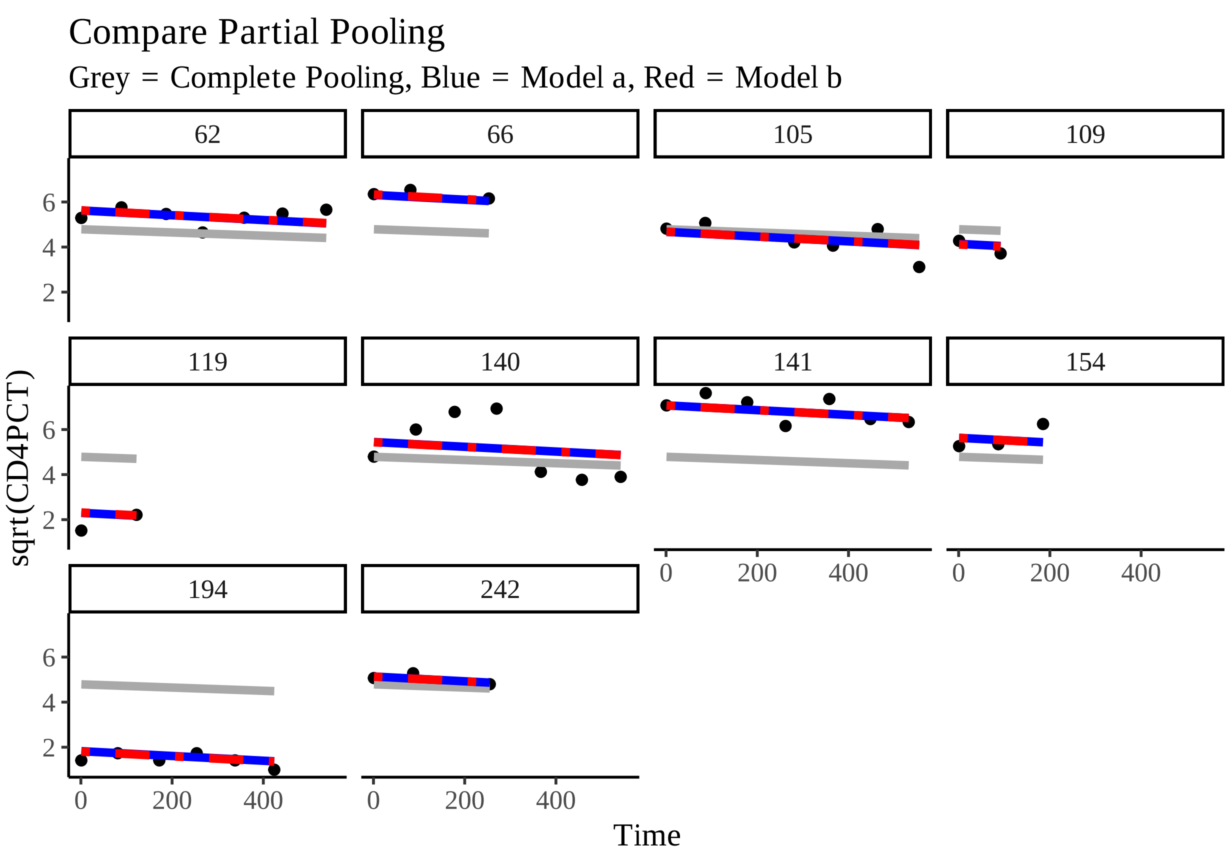
Since treatment and age at baseline are group level predictors, we expect that they may help explain some of the between group variation and leave the within group variation unchanged. First, we numerically compare the differences in the estimated within group variance ($\sigma^2$) and between group variance ($\tau^2$) between model (a) and model (b):

Table 4: Sources of Variation

| | sigma | tau |
|---|---|---|
| Model A | 0.77 | 1.981 |
| Model B | 0.77 | 1.926 |

We see that the within child standard error ($\sigma$) is basically unchanged between model (a) and model (b). We see that there is a slight reduction in the between children standard errors ($\tau$), from 1.981 to 1.926. Thus, the group level predictors, treatment and age at baseline, explain a small amount of the variation between children.

We also look at the difference in partial pooling graphically.




Graphically, we see that there is not much difference between the red and blue lines. Thus, there is not much difference in the partial pooling.

## Q3 GH Chapter 12: Exercise 4

Table 5: Random Intercept Model with County Level Predictor

| | log(radon) |
|---|---|
| (Intercept) | $1.329^{***}$ |
| | (0.060) |
| county.samples | $-0.005^{**}$ |
| | (0.002) |
| AIC | 2951.235 |
| BIC | 2971.555 |
| Log Likelihood | -1471.618 |
| Num. obs. | 1188 |
| Num. groups: county | 96 |
| Var: county (Intercept) | 0.127 |
| Var: Residual | 0.639 |

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

## Bonus

### a.)

$$var(\bar{Y}) = var(1/N * \sum_{j=1}^{J} \sum_{i=1}^{n} Y_{ij})$$

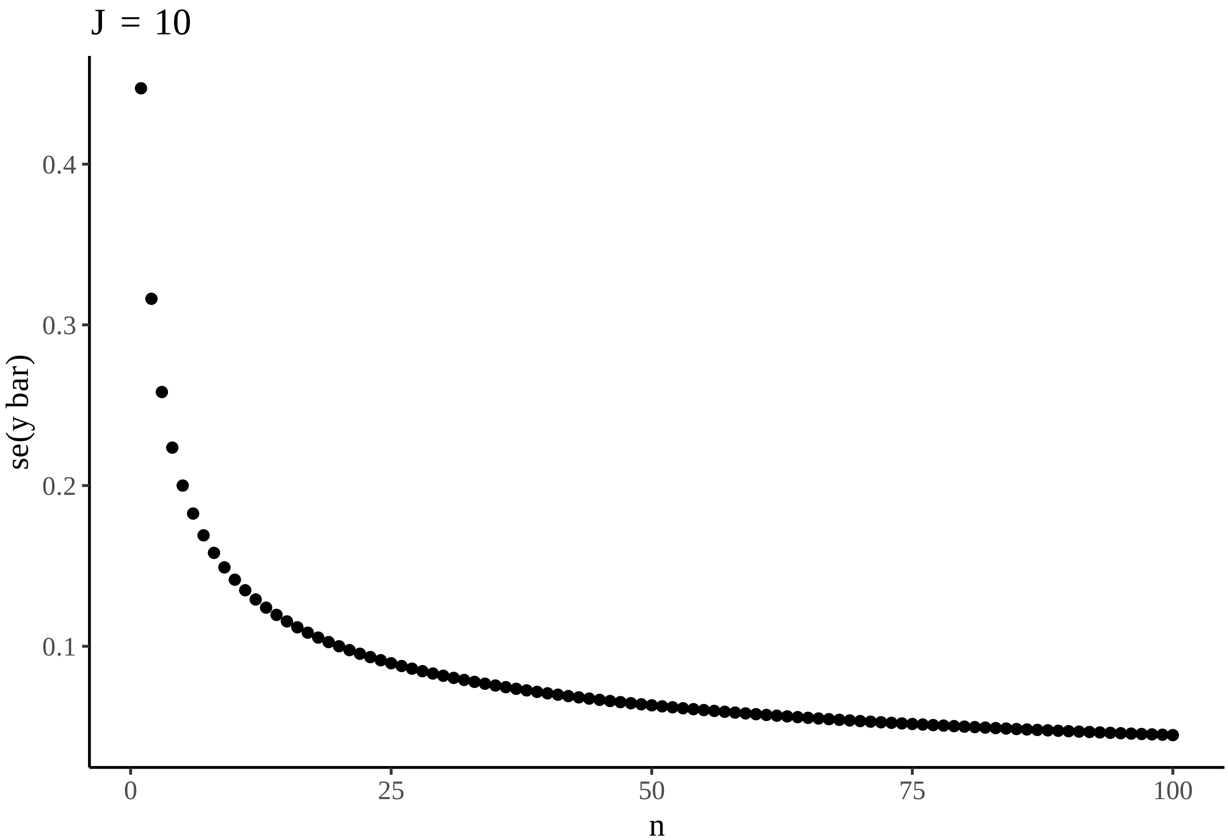
$$= 1/N^2 * \sum_{j=1}^{J} \sum_{i=1}^{n} var(Y_{ij})$$

$$= 1/N^2 * N(\sigma^2 + \tau^2)$$

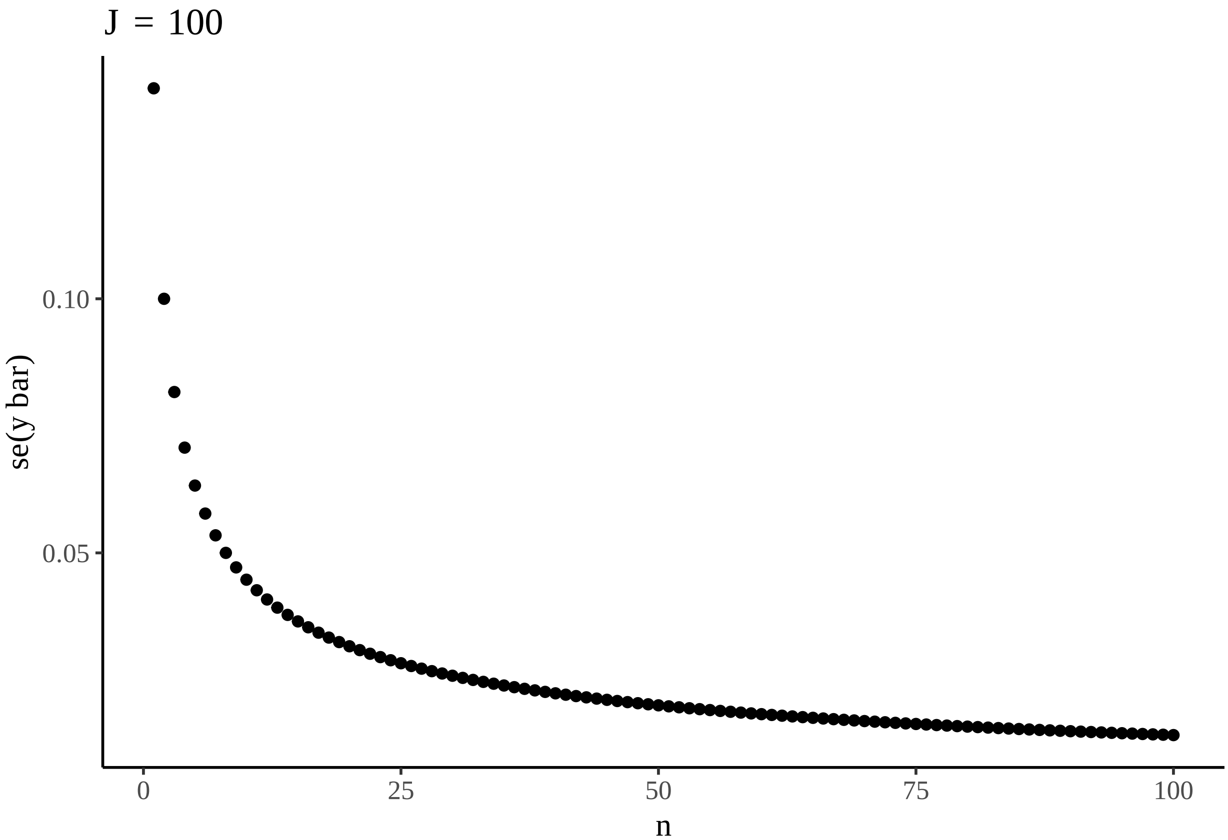
$$= \frac{(\sigma^2 + \tau^2)}{nJ}$$

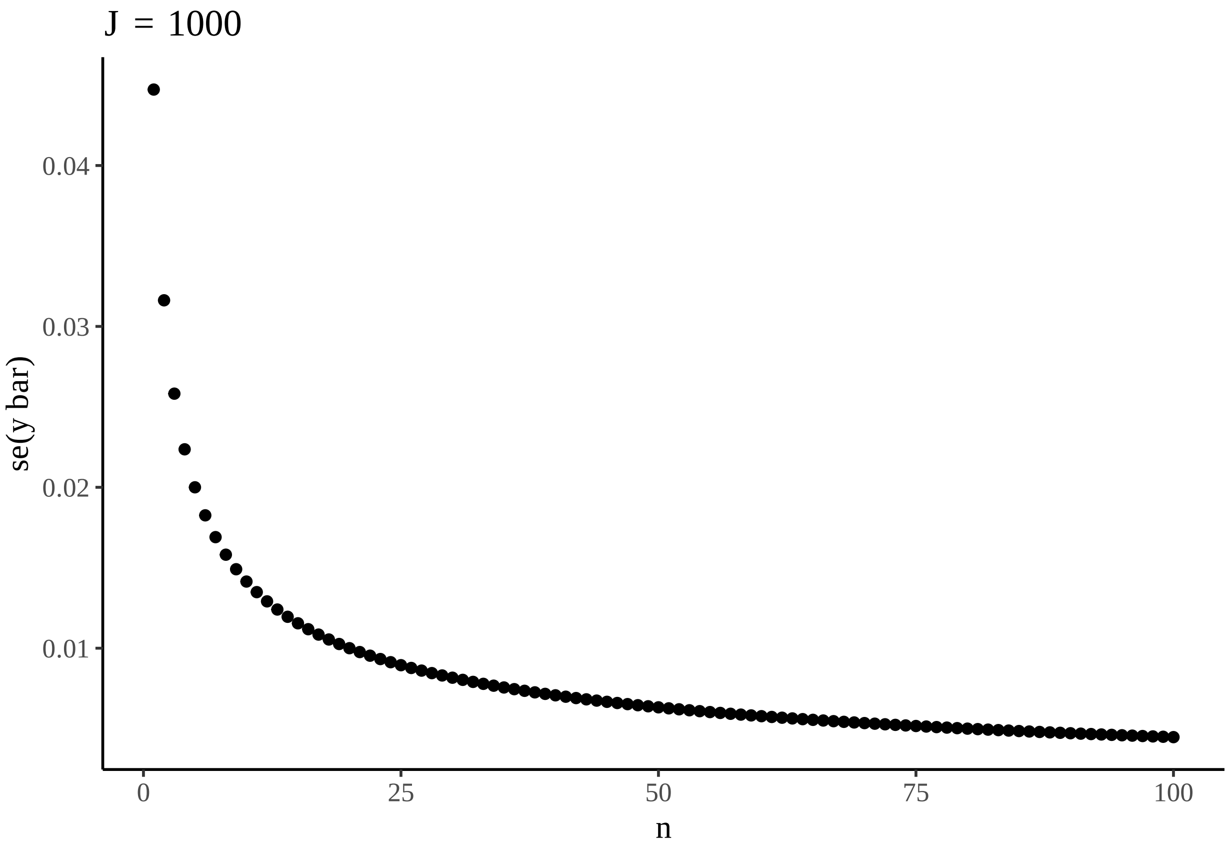
$$s.e.(\bar{Y}) = \sqrt{\frac{(\sigma^2 + \tau^2)}{nJ}}$$

b.)



**c.)**






### d.)

We see that the the standard error of the sample mean decreases at a decreasing rate as the number of observations within each cluster (n) increases. The same is true of the number of clusters (J).