HW5

Jinhong Du - 12243476

2020/01/03

Contents

9.9. Consider the model discussed in Section 9.2.4 having a random intercept and a random slope. Is the fit for subject i any different than using least squares to fit a line using only the data for subject i? Explain.

Let y_{ij} be the response of subject i at observation time j, x_i be a treatment indicator of whether the veteran receives the drug (1 = yes, 0 = no) and $t_j = \log(\text{week number} + 1)$. The model is $y_{ij} = (\beta_0 + u_{i1}) + (\beta_1 + u_{i2})t_j + \beta_2 x_i + \beta_3 t_j x_i + \epsilon_{ij} \sim \mathcal{N}(\beta_0 + \beta_1 t_j + \beta_2 x_i + \beta_3 t_j x_i, \sigma_\epsilon^2 + \sigma_{u_{i1}}^2 + t_j^2 \sigma_{u_{i2}}^2)$. If only using data for subject i, the model is $y_{ij} = \beta_0 + \beta_1 t_j + \beta_2 x_i + \beta_3 t_j x_i + \epsilon_{ij} \sim \mathcal{N}(\beta_0 + \beta_1 t_j + \beta_2 x_i + \beta_3 t_j x_i, \sigma_\epsilon^2)$. Although the estimated coefficients are similar for these two models, the estimated standard errors differ a lot.

9.11. When $\boldsymbol{X}_i$ and $\boldsymbol{V}_i$ in the linear mixed model are the same for each subject, show that the generalized least squares solution (9.10) can be expressed in terms of $\overline{\boldsymbol{y}} = \frac{1}{n}\sum_i \boldsymbol{y}_i$.

Proof.

$$
\begin{aligned}
\tilde{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{y} \\
&= \left(\sum_{i=1}^n \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right)^{-1} \sum_{i=1}^n \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{y}_i \\
&= \left[\left(\sum_{i=1}^n \boldsymbol{X}_1^\top \boldsymbol{V}_1^{-1}\boldsymbol{X}_1\right)^{-1} \boldsymbol{X}_1^\top \boldsymbol{V}_1^{-1}\right] \sum_{i=1}^n \boldsymbol{y}_i \\
&= \left[\left(\boldsymbol{X}_1^\top \boldsymbol{V}_1^{-1}\boldsymbol{X}_1\right)^{-1} \boldsymbol{X}_1^\top \boldsymbol{V}_1^{-1}\right] \overline{\boldsymbol{y}}
\end{aligned}
$$

□

9.15. For the random-effects one-way layout model (Section 9.3.2), show that $\hat{\beta}_0$ and $\tilde{u}_i$ are as stated there (i.e., with the known variances). Show that $\tilde{u}_i$ is a weighted average of 0 and the least squares estimate based on treating $\boldsymbol{u}$ as fixed effects. Give an application for which you would prefer $\hat{\beta}_0 + \tilde{u}_i$ to the fixed-effect estimate $\overline{y}_i$.

Proof. The model for the balanced one-way layout is $y_{ij} = \beta_0 + u_i + \epsilon_{ij}$ $i = 1, \ldots, c$, $j = 1, \ldots, n$. $\hat{\beta}_0 = \overline{y}$

$$
\begin{aligned}
\boldsymbol{V}_i^{-1} &= (\boldsymbol{Z}_i \sigma_u^2 \boldsymbol{I}_c \boldsymbol{Z}_i^\top + \sigma_\epsilon^2 \boldsymbol{I}_n)^{-1} \\
&= (\boldsymbol{1}_n \sigma_u^2 \boldsymbol{I}_c \boldsymbol{1}_n^\top + \sigma_\epsilon^2 \boldsymbol{I}_{nc})^{-1} = \frac{1}{\sigma_\epsilon^2}\boldsymbol{I}_n - \frac{\sigma_u^2}{\sigma_\epsilon^2(n\sigma_u^2 + \sigma_\epsilon^2)}\boldsymbol{1}_n\boldsymbol{1}_n^\top \\
\hat{\beta}_0 &= (\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{y} \\
&= \left[\sum_{i=1}^c \left(\frac{n}{\sigma_\epsilon^2} - \frac{n^2\sigma_u^2}{\sigma_\epsilon^2(n\sigma_u^2 + \sigma_\epsilon^2)}\right)\right]^{-1} \left[\sum_{i=1}^c \left(\frac{1}{\sigma_\epsilon^2} - \frac{n\sigma_u^2}{\sigma_\epsilon(n\sigma_u^2 + \sigma_\epsilon^2)}\right)\boldsymbol{1}_n^\top\right]\boldsymbol{y} = \overline{y} \\
\tilde{\boldsymbol{u}}_i &= \sigma_u^2 \boldsymbol{1}_n^\top (\sigma_u^2 \boldsymbol{1}_n \boldsymbol{1}_n^\top + \sigma_\epsilon^2 \boldsymbol{I}_n)^{-1}(\boldsymbol{y}_i - \boldsymbol{1}_n \overline{y}) \\
&= \sigma_u^2 \boldsymbol{1}_n^\top \left(\frac{1}{\sigma_\epsilon^2}\boldsymbol{I}_n - \frac{\sigma_u^2}{\sigma_\epsilon^2(n\sigma_u^2 + \sigma_\epsilon^2)}\boldsymbol{1}_n\boldsymbol{1}_n^\top\right)(\boldsymbol{y}_i - \boldsymbol{1}_n \overline{y}) \\
&= \sigma_u^2 \left(\frac{1}{\sigma_\epsilon^2}\boldsymbol{1}_n^\top - \frac{n\sigma_u^2}{\sigma_\epsilon^2(n\sigma_u^2 + \sigma_\epsilon^2)}\boldsymbol{1}_n^\top\right)(\boldsymbol{y}_i - \boldsymbol{1}_n \overline{y}) \\
&= \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2/n}(\overline{y}_i - \overline{y})
\end{aligned}
$$

and $\overline{y}_i - \overline{y}$ is the LS estimator for u of $\boldsymbol{y} - \overline{y}\mathbb{1} = \boldsymbol{u} + \boldsymbol{\epsilon}$. When the data are collected at different time and group means are relatively stable, $\boldsymbol{u}_i$ can capture the randomness caused by different time. □

9.18. For the REML approach for the normal null model described in Section 9.3.3, find $\boldsymbol{L}$, derive the distribution of $\boldsymbol{Ly}$, and find the REML estimator of σ^2.

$Y_1, \dots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, equivalently, $Y_i = \mu + \epsilon_i$ and $\epsilon_1, \dots, \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. The model is given by $\boldsymbol{y} = \boldsymbol{1}\mu + \boldsymbol{\epsilon}$. So $\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{1}(\boldsymbol{1}^\top\boldsymbol{1})^{-1}\boldsymbol{1}^\top$ and $\boldsymbol{Ly} = \boldsymbol{y} - \overline{y}\boldsymbol{1} \sim \mathcal{N}(\boldsymbol{L1}\mu, \sigma^2\boldsymbol{LL}^\top)$, i.e., $\boldsymbol{Ly} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{L})$. Notice that $\boldsymbol{L}$ has rank $n-1$, there exsit $\boldsymbol{Q} \in \mathbb{R}^{n\times(n-1)}$ and $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n-1})$ such that $\boldsymbol{QQ}^\top = \boldsymbol{L}$ and $\boldsymbol{z} = \boldsymbol{Q}^\top\boldsymbol{Ly}$. The log-likelihood function of $\boldsymbol{z}$ is

$$l(\sigma^2; \boldsymbol{z}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|\sigma^2\boldsymbol{I}_{n-1}|) - \frac{1}{2\sigma^2}\boldsymbol{z}^\top\boldsymbol{z}.$$

Taking derivative with respect to σ^2 yields

$$\frac{\partial l(\sigma^2; \boldsymbol{z})}{\partial \sigma^2} = -\frac{n-1}{2\sigma^2} + \frac{1}{2\sigma^4}\boldsymbol{z}^\top\boldsymbol{z} = 0.$$

So

$$\hat{\sigma}^2 = \frac{1}{n-1}\boldsymbol{z}^\top\boldsymbol{z} = \frac{1}{n-1}\boldsymbol{y}^\top\boldsymbol{L}^\top\boldsymbol{QQ}^\top\boldsymbol{Ly} = \frac{1}{n-1}\boldsymbol{y}^\top\boldsymbol{Ly} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

9.32. For the smoking prevention and cessation study (Section 9.2.3), fit multilevel models to analyze whether it helps to add any interaction terms. Interpret fixed and random effects for the model that has a SC×TV interaction.

As we can see, the estimated coefficients of both the random and the fixed effects do not change a lot. Also, the t value of the interaction term is small. Therefore, the interaction term is not necessary. In the model that has a SC×TV interaction, variances within class and school are about 0.06466 and 0.03864 respectively. While the estimated coefficients of the fixed effects are 1.70199, 0.30536, 0.64133, 0.18208 and -0.33094. From the estimated variances, there is more variability among classrooms within schools than among schools. Also, from the estimated coefficients, more knowledge on the smoking will help to discourage young people to smoke.

```r
library(lme4)
Smoking <- read.table("Smoking.dat", header=T)
fit1 <- lmer(y ~ PTHK + SC + TV + (1|school) + (1|class), data=Smoking)
summary(fit1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ PTHK + SC + TV + (1 | school) + (1 | class)
##    Data: Smoking
##
## REML criterion at convergence: 5374.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5202 -0.6975 -0.0177  0.6875  3.1630
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  class    (Intercept) 0.06853  0.2618
##  school   (Intercept) 0.03925  0.1981
##  Residual             1.60108  1.2653
## Number of obs: 1600, groups:  class, 135; school, 28
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.78493    0.11295  15.803
## PTHK         0.30524    0.02590  11.786
## SC           0.47147    0.11330   4.161
## TV           0.01956    0.11330   0.173
##
## Correlation of Fixed Effects:
##      (Intr) PTHK   SC
## PTHK -0.493
## SC   -0.503  0.025
## TV   -0.521  0.015 -0.002
```

```
fit2 <- lmer(y ~ PTHK + SC + TV + SC*TV + (1|school) + (1|class), data=Smoking)
summary(fit2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ PTHK + SC + TV + SC * TV + (1 | school) + (1 | class)
##    Data: Smoking
##
## REML criterion at convergence: 5373.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.49875 -0.69757 -0.01721  0.68241  3.14602
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  class    (Intercept) 0.06466  0.2543
##  school   (Intercept) 0.03864  0.1966
##  Residual             1.60229  1.2658
## Number of obs: 1600, groups:  class, 135; school, 28
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.70199    0.12543  13.569
## PTHK         0.30536    0.02589  11.794
## SC           0.64133    0.16095   3.985
## TV           0.18208    0.15724   1.158
## SC:TV       -0.33094    0.22459  -1.474
##
## Correlation of Fixed Effects:
##       (Intr) PTHK   SC     TV
## PTHK  -0.442
## SC    -0.634  0.015
## TV    -0.645  0.008  0.501
## SC:TV  0.448  0.005 -0.716 -0.700
```

9.37. The data file `Maculatum.dat` at the text website is from a study of salamander embryo development. These data refer to the spotted salamander (Ambystoma maculatum). One purpose of the study was to compare four rearing environments (very humid air, and water with low, medium, and saturated dissolved oxygen) on the age at hatching, for each of the embryos that survived to hatching. In the experiment, embryos from the same family were divided into four groups for the four treatments, and each group was reared together in the same jar. For the embryos that survived

to hatching, use multilevel modeling to compare the mean ages for the four treatments. Compare estimates and *SE* values to ones you would obtain with an ordinary linear model with treatment as the explanatory variable, ignoring the dependence due to embryos in the same jar being from the same family and due to the same family of embryos being in four jars.

The estimated coefficients are similar for the two models. While the SE of the estimated coefficients of the fixed effects in the first model are much larger than the SE of the second model.

```
Maculatum <- read.table("Maculatum.dat", header=T)
fit3 <- lmer(age ~ factor(trt) + (1|clutch) + (1|jar), data=Maculatum[Maculatum$hatch >0,])
summary(fit3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: age ~ factor(trt) + (1 | clutch) + (1 | jar)
##    Data: Maculatum[Maculatum$hatch > 0, ]
##
## REML criterion at convergence: 1271.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.9728 -0.4418  0.0159  0.3824  2.8984
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  jar      (Intercept) 2.561    1.600
##  clutch   (Intercept) 1.598    1.264
##  Residual             3.487    1.867
## Number of obs: 298, groups:  jar, 28; clutch, 12
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   17.0299     0.8428  20.206
## factor(trt)L   5.1186     1.1229   4.558
## factor(trt)M   2.3267     1.1103   2.095
## factor(trt)S   0.6932     0.9290   0.746
##
## Correlation of Fixed Effects:
##             (Intr) fct()L fct()M
## factr(trt)L -0.589
## factr(trt)M -0.600  0.524
## factr(trt)S -0.735  0.526  0.536
```

```
fit4 <- glm(age ~ factor(trt), data=Maculatum[Maculatum$hatch >0,])
summary(fit4)
```

```
##
## Call:
## glm(formula = age ~ factor(trt), data = Maculatum[Maculatum$hatch >
##     0, ])
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -10.3818   -1.3913   -0.3818    1.6087    8.9677
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     16.3818     0.3856  42.488  < 2e-16 ***
## factor(trt)L     4.6504     0.5297   8.780  < 2e-16 ***
## factor(trt)M     3.0095     0.5169   5.823 1.52e-08 ***
## factor(trt)S     0.9396     0.4708   1.996   0.0469 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 8.176125)
##
##     Null deviance: 3255.5  on 297  degrees of freedom
## Residual deviance: 2403.8  on 294  degrees of freedom
## AIC: 1477.8
##
## Number of Fisher Scoring iterations: 2
```

9.38. Refer to the previous exercise. For all the embryos, use a logistic GLMM to model the probability of hatching in terms of the treatment. Interpret results, and compare to those obtained with an ordinary logistic GLM that ignores the clustering.

The estimates of fixed effects have log-odds-ratio interpretations, within-subject for treatment effects. The estimated odds-ratio of treatment of water with low oxygen is $e^{2.631}$ times $e^{16.3818}$ the estimated odds-ratio of treatment of very humid air. The random effects have $\hat{\sigma}_u = 1.944$, a little strong associations exist among responses for the different jars.

For GLM, consider the latent variable threshold model

$$y_{is} = \begin{cases} 1 & , \boldsymbol{x}_{is}^\top \boldsymbol{\beta} + u_i - \epsilon_{is} \geq 0 \\ 0 & , \boldsymbol{x}_{is}^\top \boldsymbol{\beta} + u_i - \epsilon_{is} < 0 \end{cases}$$

where $\epsilon_{is} \overset{iid}{\sim} F$ and $F(x) = \frac{e^x}{1+e^x}$ is the logistic link function.

Notice that $F(x) \approx 1.7\Phi(x) = \Phi\left(\frac{x}{\sqrt{1.7}}\right)$, we have for the GLMM,

$$\begin{aligned} \mathbb{P}(y_{is} = 1) &= \mathbb{E}[\mathbb{E}(\mathbb{1}_{\{y_{is}=1\}} | u_i)] \\ &= \mathbb{E}[\mathbb{E}(\mathbb{1}_{\{\epsilon_{is} - u_i \leq \boldsymbol{x}_{is}^\top \boldsymbol{\beta}\}} | u_i)] \\ &= \mathbb{P}(\epsilon_{is} - u_i \leq \boldsymbol{x}_{is}^\top \boldsymbol{\beta}) \\ &\approx \Phi\left(\frac{\boldsymbol{x}_{is}^\top \boldsymbol{\beta}}{\sqrt{1.7 + \sigma_u^2}}\right) \\ \mathbb{P}(y_{is} = 1 | u_i) &= F(\boldsymbol{x}_{is}^\top \boldsymbol{\beta} + u_i) \approx \Phi\left(\frac{\boldsymbol{x}_{is}^\top \boldsymbol{\beta}}{\sqrt{1.7}}\right). \end{aligned}$$

While the first one represents the ordinary logistic GLM, the absolute values of the coefficients in GLM should be $\frac{\sqrt{1.7}\boldsymbol{\beta}}{\sqrt{1.7+\sigma_u^2}}$, which is smaller than the absolute values $\boldsymbol{\beta}$ of the GLMM. That is, the GLM shrinks the coefficients compared to the GLMM with extra random effects.

```r
library(glmmML)
fit.glmm <- glmmML(hatch ~ factor(trt), cluster=jar,
                   family=binomial,
                   data=Maculatum, method = "ghq",
                   n.points=70, start.sigma=9)
summary(fit.glmm)
```

```
## 
## Call:  glmmML(formula = hatch ~ factor(trt), family = binomial, data = Maculatum, 
## cluster = jar, start.sigma = 9, method = "ghq", n.points = 70) 
## 
## 
##                coef se(coef)      z Pr(>|z|)
## (Intercept)  -2.634   0.6671 -3.949 7.85e-05
## factor(trt)L  2.631   1.1240  2.341 1.92e-02
## factor(trt)M  2.997   1.1190  2.678 7.41e-03
## factor(trt)S  2.048   0.8901  2.301 2.14e-02
## 
## Scale parameter in mixing distribution:  1.944 gaussian 
## Std. Error:                              0.3259 
## 
##         LR p-value for H_0: sigma = 0:  6.133e-53 
## 
## Residual deviance: 763.2 on 812 degrees of freedom   AIC: 773.2
```

```r
fit.glm <- glm(hatch ~ factor(trt),
               family=binomial,
               data=Maculatum)
summary(fit.glm)
```

```
## 
## Call:
## glm(formula = hatch ~ factor(trt), family = binomial, data = Maculatum)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.2913  -1.0131  -0.6448   1.1705   1.8291  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)   -1.4649     0.1496  -9.792  < 2e-16 ***
## factor(trt)L   1.4812     0.2343   6.321 2.59e-10 ***
## factor(trt)M   1.7288     0.2361   7.322 2.44e-13 ***
## factor(trt)S   1.0654     0.1931   5.517 3.46e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1072.07  on 816  degrees of freedom
## Residual deviance:  996.37  on 813  degrees of freedom
## AIC: 1004.4
## 
## Number of Fisher Scoring iterations: 4
```

8. Explore the data in Chapter 9.2.3 (reference R code file: `Example6.Rmd`), explain why the standard error of SC and TV decrease a lot from using LMM to a regular linear model, while the standard error of PTHK does not change much.

Since $\mathrm{Var}(\boldsymbol{\beta}) = \left(\sum_{i=1}^n \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right)^{-1}$ in the LMM and $\mathrm{Var}\ (\boldsymbol{\beta}) = \left(\sum_{i=1}^n \boldsymbol{X}_i^\top \boldsymbol{X}_i\right)^{-1} \sigma^2$ in OLS, in general $\boldsymbol{V}_i = \boldsymbol{Z}_i \boldsymbol{\Sigma}_{\boldsymbol{u}} \boldsymbol{Z}_i + \sigma^2 \boldsymbol{I} \succeq \sigma^2 \boldsymbol{I}$, so the estimated standard errors tend to decrease from using LMM to OLS. However, since the magnitude of PTHK is much larger than SC and TV (SC and TV are binary), so the standard error of PTHK does not change much.

```r
library(lme4)
fit.lmm <- lmer(y ~ PTHK + SC + TV + (1|school) + (1|class), data = Smoking)
summary(fit.lmm)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ PTHK + SC + TV + (1 | school) + (1 | class)
##    Data: Smoking
##
## REML criterion at convergence: 5374.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5202 -0.6975 -0.0177  0.6875  3.1630
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  class    (Intercept) 0.06853  0.2618
##  school   (Intercept) 0.03925  0.1981
##  Residual             1.60108  1.2653
## Number of obs: 1600, groups:  class, 135; school, 28
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.78493    0.11295  15.803
## PTHK         0.30524    0.02590  11.786
## SC           0.47147    0.11330   4.161
## TV           0.01956    0.11330   0.173
##
## Correlation of Fixed Effects:
##      (Intr) PTHK   SC
## PTHK -0.493
## SC   -0.503  0.025
## TV   -0.521  0.015 -0.002
```

```r
fit.lm <- lm(y ~ PTHK + SC + TV, data = Smoking)
summary(fit.lm)
```

```
##
## Call:
## lm(formula = y ~ PTHK + SC + TV, data = Smoking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5635 -0.9130 -0.0626  0.8981  4.2416
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.73734    0.07866  22.088  < 2e-16 ***
## PTHK         0.32525    0.02589  12.561  < 2e-16 ***
## SC           0.47987    0.06529   7.350 3.15e-13 ***
## TV           0.04534    0.06518   0.696    0.487
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.303 on 1596 degrees of freedom
## Multiple R-squared:  0.1136, Adjusted R-squared:  0.112
## F-statistic: 68.21 on 3 and 1596 DF,  p-value: < 2.2e-16
```

3.15. Suppose patients can sense when the end is near, and drop out of the study just before they die. How would this affect model $y_i|n_i \sim$ Binomial(n_i, h_i)?

In such case, the observed y_i will tend to be smaller than the true one. So the $\hat{h}_i$ will be smaller and $\hat{S}_i$ be bigger, i.e., the estimated rate of survival will be larger than the true rate.

3.18. Show that $S_i(t) = e^{-H_i(t)}$ where $H_i(t) = \int_0^t h_i(s)\mathrm{d}s$. How does this relate to formula $S_i = \prod\limits_{j=1}^{i-1}(1-h_j)$.

Since $S_i(t) = \int_t^\infty f_i(s)\mathrm{d}s$, we have $1 - S_i(t) = \int_{-\infty}^t f_i(s)\mathrm{d}s$. Also, since $T_i \geq 0$, $S_i(0) = 1$. So

$$H_i(t) = \int_0^t h_i(s)\mathrm{d}s = \int_0^t \frac{f_i(s)}{S_i(s)}\mathrm{d}s = \int_0^t \frac{1}{S_i(s)}\mathrm{d}[1 - S_i(s)] = -\log[S_i(s)]\big|_0^t = -\log[S_i(t)],$$

i.e. $S_i(t) = e^{-H_i(t)}$. Notice that $S_i = e^{\sum_{j=1}^{i-1}\ln(1-h_j)}$ and $\ln(1-h_j) \approx -h_j$ when h_j is small enough, so $S_i \approx e^{-\sum_{j=1}^{i-1} h_j} = e^{-H_{i-1}}$ when $h_1, \dots, h_i \to 0^+$, which is similar to the continuous case.

3.19. Denoting $e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} = \alpha_i$, show that $S_i(t) = S_0(t)^{\alpha_i}$ (a relationship known as "Lehmann alternatives"), where $S_0(t)$ is the baseline survival function.

In the proportional hazards model, $h_i(t) = h_0(t)e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} = h_0(t)\alpha_i$. So

$$\begin{aligned} S_i(t) &= e^{-\int_0^t h_i(s)\mathrm{d}s} \\ &= e^{-\int_0^t h_0(s)\mathrm{d}s \cdot \alpha_i} \\ &= S_0(t)^{\alpha_i}. \end{aligned}$$

3.23. Use `coxph` to test the null hypothesis that Arm B is no better than Arm A for the NCOG data listed at the beginning of Section 3.6; data is in the file "ncogdata". Hint: The only explanatory variable is the Arm indicator.

As the likelihood ratio test is significant, we should reject the null hypothesis that Arm B is no better than Arm A.

```
ncog <- data.frame(t=c(7,34,42,63,64,74,83,84,91,108,112,129,133,133,139,140,140,146,149,
                       154,157,160,160,165,173,176,185,218,225,241,248,273,277,279, 297,319,
                      405,417,420,440,523,523,583,594,1101,1116,1146,1226,1349,1412,1417,
                       37,84,92,94,110,112,119,127,130,133,140,146,155,159,169,173,
                       179,194,195,209,249,281,319,339,432,469,519,528,547,613,633,
                       725,759,817,1092,1245,1331,1557,1642,1771,1776,1897,2023,2146,2297),
                   d=c(1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
                       1,0,1,1,1,1,1,1,0,1,0,1,1,1,1,1,0,1,1,1,0,1,0,0,0,1,
                       1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,
                       1,1,1,1,0,0,0,0,1,1,0,1,0,0,0,1,0,0,1,0,0,0,0),
                   arm=c(rep('A', 51), rep('B', 45)))
library(survival)
coxph(Surv(t,d) ~ arm, ncog)
```

```
## Call:
## coxph(formula = Surv(t, d) ~ arm, data = ncog)
##
##         coef exp(coef) se(coef)      z      p
## armB -0.5526    0.5754   0.2444 -2.261 0.0237
##
## Likelihood ratio test=5.23  on 1 df, p=0.02222
## n= 96, number of events= 73
```