Longitudinal Data Analysis Exercises

Joschka Hüllmann

2019-01-15

Exercise One

The goal of this manuscript is to reproduce Marinov's analysis and compare its underlying model to other model choices. In particular, I reproduce the exact model and compare it to other conditional unit effects models, i.e. a random effects model and the least squares dummy variables (LSDV), to a standard OLS, and to general equation estimation (GEE) models with varying working correlation structure.

Marinov's article explores the effect of various covariates, e.g. the presence of international sanctions, and economic wealth and growth, on the a leader's ability to retain power. He bases his analysis on a data set aggregated from various sources. According to the manuscript, the data set includes 160 countries from the year 1947 to 1999 ($N = 160; T = 52$; actual data frame $N = 160; T = 49; N(observations) = 5766$). The dependent variable Y is binary with the leader either surviving in office ($Y = 0$) or giving up his position ($Y = 1$). The author is interested in what covariates reduce the time to event of a leader giving up his position. Although this suggests a a survival model, Marinov, following Beck, Katz and Tucker (1998), argues that a conditional logit model is equivalent to a cox proportional hazards model (c.f. appendix 1). Besides three cubic splines to capture the three phases of a leader's stability early, middle and late phase of his leadership, as well as the time t in office, interacting with regime type, no explicit time variance on the other covariates is assumed. Not looking into any dynamic effects, the author models unit effects conditional on the countries, i.e. assuming institutional or other effects inherent to each country.

In summary, the assumptions of the author for the data generating process are as follows:

- panel type data with $N >> T$
- binary outcome $Y = 1$ or $Y = 0$
- no time variance on covariates other than years as a covariate interacting with regime dummies
- three cubic splines to deal with differences in the three phases of leadership.
- conditional unit effects on the countries

Strictly following the assumptions, it is clear, that a conditional unit effects model is to be specified with binary outcome, thus suggesting a logit model conditional on the country. The authors choose a fixed effect logit model over random effects, which I guess is due to the panel data $N >> T$ (c.f. slides day 1, AM, p. 51). The specified model shoud look like the following, f being the *logit* function and α_i being the country-level intercept:

$$Y_{it} = f(X_{it}\beta + \alpha_i + u_{it})$$

Alternatively to the proposed model by the authors, we can specify a naive estimator without unit-level effects, which would remove the α_i from the equation. The variance would end up in the coefficients and the error. As a result, not considering existing unit-level effects may bias the estimates and lead to an inconsistent estimator (c.f. appendix 2). Conversely, we can model random effects, if we assume that the unit level effects are uncorrelated to the covariates. In this case, the α_i would be a second error term (random effect) instead of a fixed effect, so the error looks like $u_{it} = \alpha_i + \eta_{it}$. If the independence assumptions holds ($cov(x_{it}, \alpha_i) = 0$), it is more efficient than the fixed effects model. However, if the assumption does not hold, it leads to a biased and inconsistent model (omitting fixed effects similar to appendix 2). Analytically, this can be tested with the Hausman test. For completeness, the model could also be specified using an LSDV model ("brute force"). However, with $N = 160; T = 52$, the LSDV model suffers from the incidental parameter problem with few observations (small T) within each unit, so that inconsistent estimates are reached. Hence, the differences between estimates in the conditonal logit model and the LSDV model (for small T the estimates are biased $\hat{\sigma}^2 = \sigma^2 \frac{T-1}{T}$, c.f. appendix 3).

Contrary to a conditional unit effect model, we can change our assumptions about the influence and correlation between the outcomes and covariates, and in particular the dependence between Y and time t, and specify a marginal population average model, i.e. generalised equation estimation model (GEE). At the core of a GEE is the working correlation matrix, which describes the structure of covariance (correlation) and is specified by the analyst. In general the formula as follows:

$$U(\beta) = \sum_{i=1}^{N} D_i' V_i^{-1} [Y_i - \mu_i] = 0$$

with the variance (after the standard GLM assumption) decomposed into:

$$V_i = \frac{(A_i^{1/2}) R_i(\alpha) (A_i^{1/2})}{\phi}$$

Typical specifications for the working correlation matrix ϕ include: independent, exchangeable, autoregressive, stationary, unstructured (empirical) (c.f. appendix 4 for how they look like).

R Examples

Listing 1 shows the loading and filtering of Marinov's original data set.

```
library(RCurl)
library(dplyr)
library(geepack)
library(survival)
library(glmmML)
df<-read.csv(textConnection(getURL("https://raw.githubusercontent.com/PrisonRodeo/GSERM-Oslo-2019-git/ma
df2 <- filter(df, is.na(fail)==FALSE, is.na(sanctionsl1)==FALSE,
              is.na(growthpc)==FALSE, is.na(lngdppc)==FALSE,
              is.na(democl1)==FALSE, is.na(democlnt)==FALSE,
              is.na(mixedl1)==FALSE, is.na(mixedlnt)==FALSE,
              is.na(age)==FALSE, is.na(ot3)==FALSE,
              is.na(X_spline1)==FALSE,is.na(ccode)==FALSE)
```

Listing 2 shows the results of reproducing Marinov's conditional logit model as LSDV and clogit model. Note, that the interpretation of Rsquared is not straight-forward.

```
fit.clogit <- clogit(fail~sanctionsl1+forcel1+growthpc+lngdppc+democl1+
           democlnt+mixedl1+mixedlnt+age+ot3+X_spline1+X_spline2+
           X_spline3+strata(ccode), data=df)
summary(fit.clogit)
fit.lsdv <- glm(fail~sanctionsl1+forcel1+growthpc+lngdppc+democl1+
           democlnt+mixedl1+mixedlnt+age+ot3+X_spline1+X_spline2+
           X_spline3+as.factor(ccode), data=df,family=binomial)
cbind(fit.lsdv$coefficients[1:11],fit.lsdv$residuals[1:11])
```

Listing 3 shows the results of the random effects model. Note, that the fixed effects model is unconditional and thus equivalent to the LSDV model, suffering from the incidental parameters problem, too. The random effects model has similar estimates of the coefficients and standard errors as the conditional fixed effects model. It is an interesting question what this means. Does it mean that the within-unit estimates for the model explains a lot of the variance (this would be the case for comparing with an unconditional unit effect model, how is it with the conditional unit effects model)? A comparison of the likelihoods should inform this...

```
fit.glmm.fe <- glmmboot(fail~sanctionsl1+forcel1+growthpc+lngdppc+democl1+
           democlnt+mixedl1+mixedlnt+age+ot3+X_spline1+X_spline2+
           X_spline3, data=df, family="binomial", cluster=ccode)
summary(fit.glmm.fe)
fit.glmm.re <- glmmML(fail~sanctionsl1+forcel1+growthpc+lngdppc+democl1+
           democlnt+mixedl1+mixedlnt+age+ot3+X_spline1+X_spline2+
           X_spline3, data=df,family="binomial",cluster=ccode)
summary(fit.glmm.re)
```

Listing 4 shows the results of various GEE models (unstructured is not included, because the computation does not conclude on my computer and stationary is not included, because it requires a user-defined function, i.e. not implemented by default). As the GEE model is a population model, it is expected that the results change. In the marginal population average model, the primary covariate of interest, sanctions, still is positively associated with leaders failing. Across the GEE models with varying working correlation matrices, minor differences in estimation exist. The R package geepack defaults to the sandwiched (robust) standard errors (c.f. appendix 5), based on the observed empirical variance. For further reference, it would be interesting to look into what we need to change about the data generating process and our assumptions to establish equivalence between GEE and the conditional fixed effects. For starters, GEE with independence is equal to GLM, adding the logit and conditioning on country level effects, may lead us towards such an equivalence.

```
fit.gee.in <- geeglm(fail~sanctionsl1+forcel1+growthpc+lngdppc+democl1+
             democlnt+mixedl1+mixedlnt+age+ot3+X_spline1+X_spline2+
             X_spline3,data=df2,id=ccode,family=gaussian,
             corstr="independence")
fit.gee.ex <- geeglm(fail~sanctionsl1+forcel1+growthpc+lngdppc+democl1+
             democlnt+mixedl1+mixedlnt+age+ot3+X_spline1+X_spline2+
             X_spline3,data=df2,id=ccode,family=gaussian,
             corstr="exchangeable")
fit.gee.ar <- geeglm(fail~sanctionsl1+forcel1+growthpc+lngdppc+democl1+
             democlnt+mixedl1+mixedlnt+age+ot3+X_spline1+X_spline2+
             X_spline3,data=df2,id=ccode,family=gaussian,
             corstr="ar1")
summary(fit.gee.in)
summary(fit.gee.ex)
summary(fit.gee.ar)
```

In general, Marinov, acknowledging unobserved unit-level effects, choose a proper model. In case the author assumes a dependence of Y_i on time t and can assume (or compute) the covariance-variance matrix and is interested in marginal population average effects, GEE models are a viable alternative.

Exercise Two

The exercise deals with the subject of leadership tenure and how age, female and region may influence the duration of tenure. In the following, I will consider parametric survival models (exponential, weibull), the semiparametric cox proportional hazards model and a discrete time survival model (time dummies). The data set is provided by the *Archigos* project with $N = 2990$ world leaders, a time frame between 1875-2003 and total observations of $NT = 15,244$. It contains time-varying information on how long the leader was in office. Listing 1 contains the preprocessing.

```
library(survival)
setwd("C:\\dev\\workspace\\GSERM-Oslo-2019-git\\Exercises-Solution")
df<-read.csv("..\\Exercises\\GSERM-Oslo-2019-ExTwo.csv")
```

The basic parametric model is as follows, with $f(t)$ being the survival density, $S(t)$ the survival function (or risk to die at particular time unit), $h(t)$ the hazard function, and L the likelihood:

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t)}{\Delta t}$$

$$S(t) = \Pr(T \leq t) = 1 - F(t)$$

$$h(t) = \frac{f(t)}{S(t)} = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

$$L = \prod_{i=1}^{N} [f(T_i)]^{C_i} [S(T_i)]^{1-C_i}$$

For descriptive analysis a kaplan-meier estimator can be calculated with $S(t_k) = \prod_{t<t_k} \frac{n_t - d_t}{n_t}$, with n_t being number of observations at risk (size of risk set) at time t, and d_t denoting the number of observations that experience the event at time t.

The exponential model assumes a constant hazard function $h(t) = \lambda$ with integrated hazard $H(t) = \lambda t$, survival function $S(t) = \exp(-\lambda t)$, density $f(t) = \lambda \exp(-\lambda t)$. The weibull model extends the exponential model by a parameter p with $h(t) = \lambda p (\lambda t)^{p-1}$, and $S(t) = \exp(-\lambda t)^p$, and $f(t) = \lambda p (\lambda t)^{p-1} \exp(-\lambda t)^p$, whereas $p = 1$ is equal to the exponential model, and $p > 1$ means a rising hazard, and $0 < p < 1$ means a declining hazard. The cox proportional hazards model is semiparametric and does not assume an underlying distribution. Instead an empirical baseline hazard is assumed (that everyone shares), and based on the order of events (not the duration), a partial likelihood is estimated with

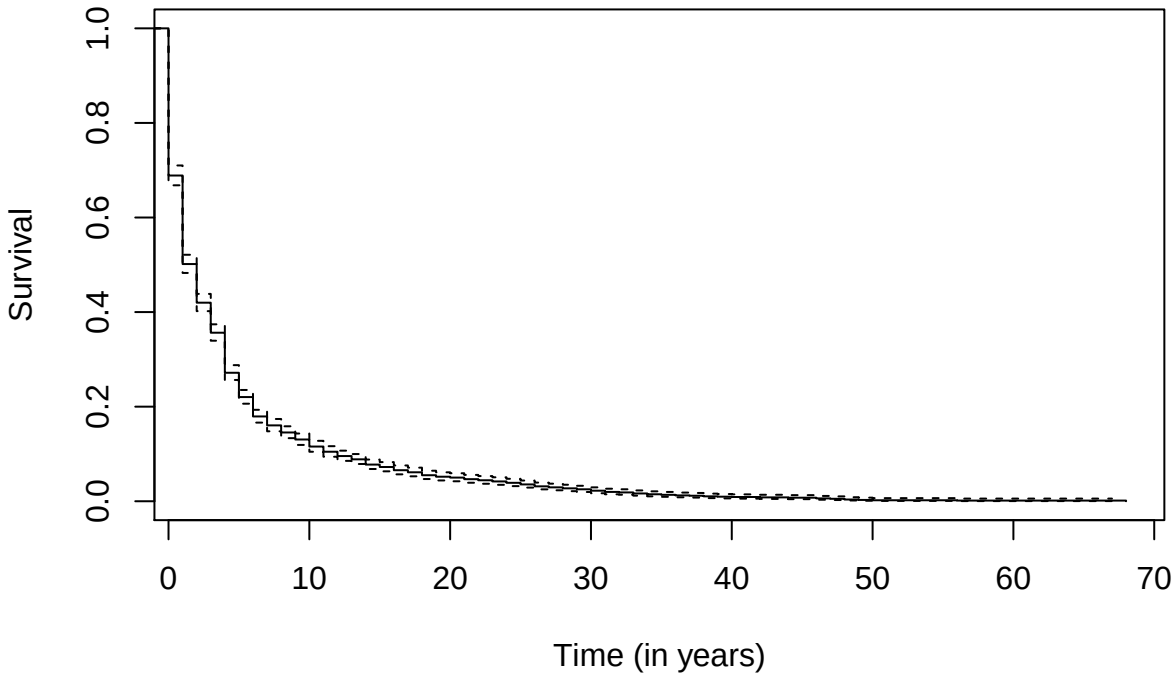
$$L_i = \frac{\exp(X_i \beta)}{\sum_{l \in R_j} \exp(X_l \beta)}$$

(see also appendix 1). A discrete time model, as the name suggests, assumes events at discrete time, so that between two timepoints there can be no events. The conditional probability for survival is given as $\Pr(T > t | T \geq t) = 1 - h(t)$, which implies $S(t) = \prod_{j=0}^{t} [1 - h(t-j)]$ (skipping the enumeration of probabilities from $0 \leq T \leq t$) in the product). The density is then $f(t) = h(t)S(t) = h(t) \prod_{j=1}^{t-1} [1 - h(t-j)]$ (again skipping the enumeration), and the likelihood $L = \prod_{i=1}^{N} (h(t) \prod_{j=1}^{t-1} [1 - h(t-j)])^{Y_{it}} (\prod_{j=0}^{t} [1 - h(t-j)])^{1-Y_{it}}$

R Examples

The kaplan-meier estimate of the survival function is given in figure 1.

```
lead.s <- with(df, Surv(tenurestart,tenureend,leftoffice))
plot(lead.s,xlab="Time (in years)", ylab="Survival")
```



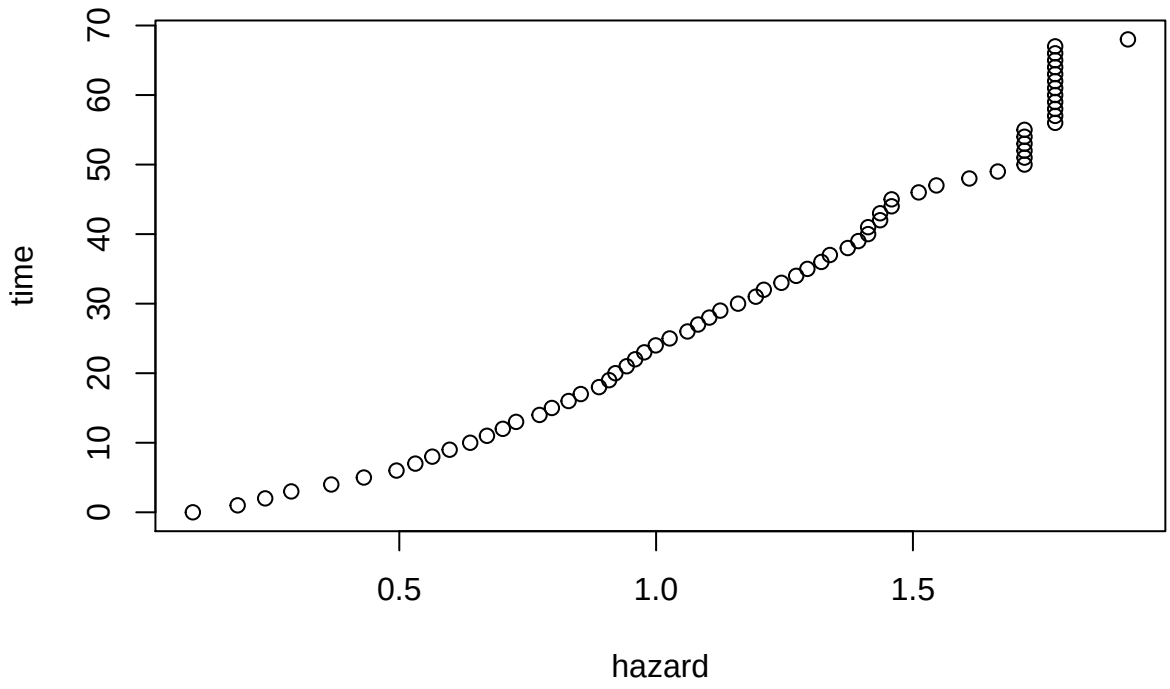
Listing 2 shows the estimate of the exponential model, with some extra processing to convert it to a duration model and fix the error that duration cannot be zero. Plus one is a naive fix and might require better adjustment.

```
# maximum tenureend, last entry is the duration to event
df2 <- df %>% group_by(leadid) %>% slice(max(which(tenureend == max(tenureend)))) %>% filter(is.na(age)=
df2$tenureend = df2$tenureend + 1
lead.s <- with(df2, Surv(tenureend,leftoffice))
lead.exp.AFT<-survreg(lead.s ~ age + fename + LatinAm + Europe + Africa + Asia + MidEast,data=df2,dist=
lead.exp.PH<-(-lead.exp.AFT$coefficients)
lead.exp.HRs<-exp(-lead.exp.AFT$coefficients)
lead.wei.AFT<-survreg(lead.s ~ age + fename + LatinAm + Europe + Africa + Asia + MidEast,data=df2,dist=
lead.wei.PH<-(-lead.exp.AFT$coefficients)
lead.wei.HRs<-exp(-lead.exp.AFT$coefficients)
lead.exp.AFT
lead.wei.AFT
```

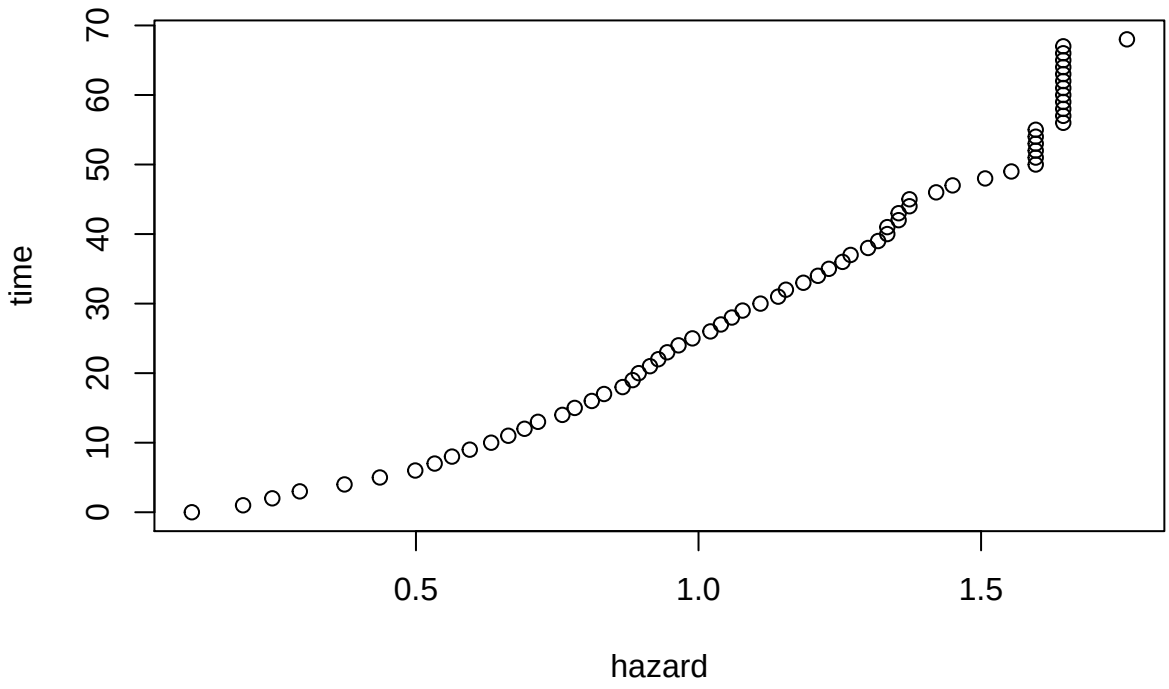
The coefficients in the AFT model show the change of duration, with a negative coefficient meaning a decrease of duration. Hence, females are marginally less likely to lose leadership, same with higher age. While in Middle East people tend to be holding longer on to their position, Europe and Americas see more change in leaders (Democracy, yay). To check the effect of gender in the different regions, an interaction effect can be modeled. For the weibull model a scale parameter of $scale = 0.866$ is estimated, meaning that the parameter p as in the formula is $p > 1$, indicating a rising hazard over time.

The cox proportional hazards implementation can deal with the original interval based data set and is illustrated in Listing 3. The estimated coefficiens are similar to the weibull AFT, with the exception of age, which has a different sign.

```
lead.s <- with(df, Surv(tenurestart,tenureend,leftoffice))
lead.cox.br<-coxph(lead.s~ age + fename + LatinAm + Europe + Africa + Asia + MidEast,
                data=df,na.action=na.exclude, method="breslow")
lead.cox.ef<-coxph(lead.s~ age + fename + LatinAm + Europe + Africa + Asia + MidEast,
                data=df,na.action=na.exclude, method="efron")
#lead.cox.ex<-coxph(lead.s~ age + fename + LatinAm + Europe + Africa + Asia + MidEast,
#                 data=df,na.action=na.exclude, method="exact")
summary(lead.cox.br)
summary(lead.cox.ef)
lead.cox.br.basehaz<-basehaz(lead.cox.br,centered=FALSE)
lead.cox.ef.basehaz<-basehaz(lead.cox.ef,centered=FALSE)
plot(lead.cox.br.basehaz)
```




```
plot(lead.cox.ef.basehaz)
```



Listing 4 shows the time dummy model. The intercept has a high coeffient and the sign for Europe changed.

```
lead.dummy<-glm(leftoffice~ age + fename + LatinAm + Europe + Africa + Asia + MidEast
+as.factor(tenureend),data=df2,na.action=na.exclude, family="binomial")
hat.dummy<-predict(lead.dummy,df2,type="response")
summary(lead.dummy)
```

Listing 5 shows a basic stratified proportional hazards model by country. The data frame can be transformed and the cluster estimate used with region. The results can be compared with the other cox model, which has the regions as dummies.

```
lead.s <- with(df, Surv(tenurestart,tenureend,leftoffice))
lead.coxs.ef<-coxph(lead.s~ age + fename,
                    data=df,na.action=na.exclude, method="efron", cluster(ccode))
summary(lead.coxs.ef)
```

Summarising, the cox proportional hazards model is well suited to estimate the effect of the covariates on the duration. The coefficients are similar to the weibull model, both of which show that Europe and North America have a quicker change of leaders compared to the other countries. Females manage to stay longer in power.

TODO:

- better diagnostics for the surival models, including joint plots
- incl. better scaling for the plots
- better data processing for parametric models (also include more detailed comparison of underlying statistics and reported estimates)
- include interaction models
- include stratified and frailty models, e.g. by country or region

- compare results of stratified model with the naive coxph model
- show equivalence of cox and logit glm again

Appendix

Appendix 1 - Equivalence Cox and CLogit

Equivalency of Cox and Conditional Logit (c.f. slides day 4, PM, p. 42).

logit conditional likelihood:

$$Pr(Y_i = j) = \frac{\exp(X_{ij}\beta)}{\sum_{l=1}^{J} \exp(X_{il}\beta)}$$

cox partial likelihood:

$$L_k = \frac{\exp(X_k\beta)}{\sum_{l \in R_j} \exp(X_l\beta)}$$

Appendix 2 - Bias of omitting unit level effects

I think that omitting the unit level effect (similar to omitted variable bias) leads to biased estimates and an inconsistent estimator.

Normal OLS (regression model):

$y_i = X_i\beta + u_i$

With fixed unit effect (population model):

$y_i = X_i\beta + \alpha_i + u_i$

Now estimating only normal OLS with substituting the assumed unit effects:

$\hat{\beta} = (X'X)^{-1}X'Y$

$\hat{\beta} = (X'X)^{-1}X'(X\beta + A + U)$

$\hat{\beta} = (X'X)^{-1}X'\beta + (X'X)^{-1}X'A + (X'X)^{-1}X'U$

$\hat{\beta} = \beta + (X'X)^{-1}X'A + (X'X)^{-1}X'U$

$E[\hat{\beta}] = \beta + E[(X'X)^{-1}X'A] + E[(X'X)^{-1}X'U]$

$E[\hat{\beta}] = \beta + (X'X)^{-1}E[X'A] + (X'X)^{-1}E[X'U]$

with $E[X'U] = 0$ by assumption leads to:

$E[\hat{\beta}] = \beta + (X'X)^{-1}E[X'A]$

$E[\hat{\beta}] = \beta + \text{bias}$

Appendix 3

The incidental parameters problem and how the conditional logit model deals with the it by conditioning out the unit-level effect is described here (conditional maximum likelihood estimator):

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. Econometrica: Journal of the Econometric Society, 1-32.

Abrevaya, J. (1997). The equivalence of two estimators of the fixed-effects logit model. Economics Letters, 55(1), 41-43.

Greene, W. (2004). The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. The Econometrics Journal, 7(1), 98-119.

Lancaster, T. (2000). The incidental parameter problem since 1948. Journal of econometrics, 95(2), 391-413.

and a helpful link: https://www.stata.com/statalist/archive/2007-10/msg00926.html

Appendix 4

$$R_i(\alpha) = \begin{pmatrix} 1.0 & \alpha & \cdots & \alpha \\ \alpha & 1.0 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \cdots \\ \alpha & \cdots & \alpha & 1 \end{pmatrix}$$

(for the following c.f. slides day 3, AM, pp. 8ff)

Independent: no within-unit temporal correlation, equivalent to GLM, with $\alpha = 0$.

Exchangeable: constant temporal correlation within units across time points with $\alpha_{ts} = \alpha \forall t \neq s$.

Autoregressive: conditional within-unit correlation as an exponential function of lag (closer together means higher correlation) with $\alpha_{ts} = \alpha^{|t-s|} \forall t \neq s$.

Stationary: conditional within-unit correlation as a function of lag, up to lag p, and zero afterwards with

$$\alpha = \begin{cases} \alpha_{ts} & \text{for } t \leq p, \\ 0 & \text{otherwise.} \end{cases}$$

Unstructured: conditional within-unit correlation is completely data dependent with $T(T-1)/2$ free parameters (computationally demanding).

Appendix 5

c.f. slides Day 3, AM, p. 12.

Standard GEE model:

$$\sum_{\text{Normal}} = N(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{D}_i)$$

Robust GEE model:

$$\sum_{\text{Robust}} = N(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{D}_i)^{-1} (\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{S}_i \hat{V}_i^{-1} \hat{D}_i) (\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{D}_i)^{-1}$$