Data Dictionary

1. Time: as the time (in years) since seroconversion, where a negative time denotes actual time before seroconversion.
2. Age: age at seroconversion (a baseline measurement), centred at 30 years of age, so that negative ages denote years younger than 30.
3. Packs: the number of packets of cigarettes smoked per day at time of measurement.
4. Drugs: a binary variable taking the values 1 or 0 to denote if the respondent takes recreational drugs or not respectively, measured at each time point.
5. Cesd: an index of depression measured at each time point, with time trends removed. Higher scores indicate greater depressive symptoms.
6. Sex: number of sexual partners reported at each time point. Looks to have been centred somehow and truncated at ± 5.

Setup

```
library(knitr)
library(tinytex)
library(kableExtra)
library(latex2exp)
library(tidyverse)
library(gridExtra)
library(nlme)
library(lmtest)
library(splines)
```

Load dataset and create engineered features for later use.

```
cd4_df <- read.table("cd4data.txt", header = TRUE)
cd4_df <- cd4_df %>%
  mutate(
    CD4sqrt = CD4^0.5,
    yr = round(Time),
    yr.f = factor(yr, levels=c(-3,-2,-1,0,1,2,3,4,5)),
    quarter = round(4*Time)/4,
    smoker = ifelse(Packs > 0, 1, 0)
  ) %>%
  arrange(ID, Time) %>%
  group_by(ID) %>%
  mutate(
    obsnum = 1:n()
  )
```

Exploratory Data Analysis

Look at how number of subjects and observations varied over the course of the study.

```
obs_per_year_df <- cd4_df %>% group_by(yr) %>% summarise(obs_cnt=n())
subs_per_year_df <- cd4_df %>% select(yr, ID) %>% distinct %>% group_by(yr) %>% summarise(sub_cnt=n())
sub_obs_df <- tibble(
  Year = obs_per_year_df$yr,
```

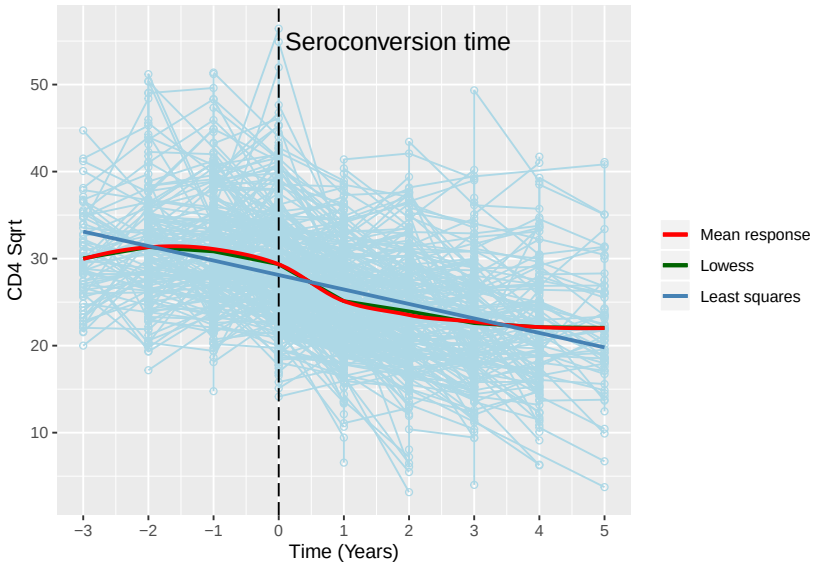
```
  Num.Observations = obs_per_year_df$obs_cnt,
  Num.Subjects = subs_per_year_df$sub_cnt
)
kable(t(sub_obs_df), caption="Study Observations Profile", escape = F, digits = 6) %>%
  kable_styling(latex_options = c("hold_position"))
```

Table 1: Study Observations Profile

Year	-3	-2	-1	0	1	2	3	4	5
Num.Observations	71	198	315	529	431	346	254	163	69
Num.Subjects	70	133	211	307	279	226	167	109	51

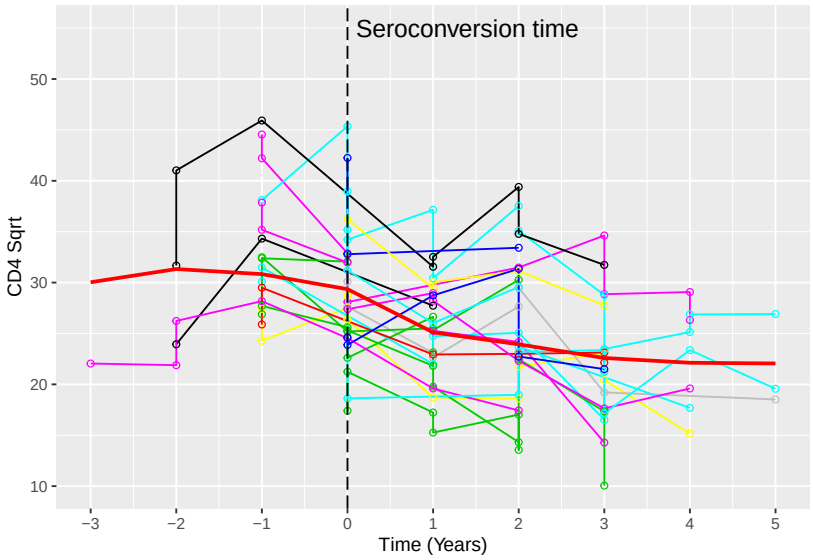
Plot response curves over time : build a dataframe for the means at each time point

```
x <- sort(unique(cd4_df$yr))
means_df <- cd4_df %>%
  select(yr, CD4sqrt) %>%
  group_by(yr) %>%
  summarise(mean_response = mean(CD4sqrt)) %>%
  arrange(yr)

y_lim = c(floor(min(cd4_df$CD4sqrt)), ceiling(max(cd4_df$CD4sqrt)))
y_scale <- seq(floor(min(cd4_df$CD4sqrt)), ceiling(max(cd4_df$CD4sqrt)))

ggplot(cd4_df)+
  geom_line(aes(x=yr, y=CD4sqrt, group=ID), color='lightblue')+
  geom_point(aes(x=yr, y=CD4sqrt, group=ID), shape=1, color='lightblue')+
  xlab('Time (Years)')+ ylab('CD4 Sqrt')+scale_x_continuous(breaks=x)+
  geom_vline(xintercept = 0, colour='black', lty=5)+
  annotate("text", label = "Seroconversion time", x = 0.1, y = 55, size = 5, colour = "black", hjust=0)+
  geom_line(data = means_df, aes(x=yr, y=mean_response, colour='red'), lwd=1)+
  geom_smooth(aes(x=yr, y=CD4sqrt, colour='darkgreen'), se = F, method='loess', span=0.75)+
  geom_smooth(aes(x=yr, y=CD4sqrt, colour='steelblue'), se = F, method='glm')+
  scale_colour_manual(values=c('red', 'darkgreen', 'steelblue'),
                      labels=c('Mean response', 'Lowess', 'Least squares'))+
  theme(legend.title = element_blank())
```



Plot a sample subset of responses:

```
# these were originally randomly generated
sample_sids <- c(20439,41829,41844,30693,40942,30820,20777,40286, 30489,30075,30827,
                 20205,41566,21083,41253,10302,40807,30835,20768,41194)
sample_df <- cd4_df %>% filter(ID %in% sample_sids)
ggplot(sample_df)+
  geom_line(aes(x=yr, y=CD4sqrt, group=ID), color=sample_df$ID)+
  geom_point(aes(x=yr, y=CD4sqrt, group=ID), shape=1, color=sample_df$ID)+
  scale_x_continuous(breaks=x)+
  geom_vline(xintercept = 0, colour='black', lty=5)+
  annotate("text", label = "Seroconversion time", x = 0.1, y = 55, size = 5, colour = "black", hjust=0)+
  xlab('Time (Years)')+ ylab('CD4 Sqrt')+
  geom_line(data = means_df, aes(x=yr, y=mean_response), colour='red', lwd=1)+
  theme(legend.title = element_blank())
```



Sample Variagram based on code from Week 2 lecture material. Use residuals from a simple spline fit.

```
cd4_df$resid <- resid(smooth.spline(cd4_df$yr, cd4_df$CD4sqrt))
# sample variogram from week 2 lecture material
vijk <- by(cd4_df, cd4_df$ID, function(df) {
  v <- outer(df$resid, df$resid,
             function(x, y) 0.5*(x-y)^2)
  v[lower.tri(v)]
})
uijk <- by(cd4_df, cd4_df$ID, function(df) {
  u <- outer(df$Time, df$Time,
             function(x, y) abs(x - y))
  u[lower.tri(u)]
})
uijk <- unlist(uijk)
vijk <- unlist(vijk)

vu.lowess <- lowess(uijk, vijk)
sigma2 <- var(cd4_df$resid)

plot(uijk, vijk, col = "gray50", pch = 18, cex = 0.4,
```

```
     xlim = c(0, 6), ylim = c(0, 50),
     xlab = "Lag", ylab = "Half squared differences")
lines(vu.lowess, col = "red", lwd = 2)
abline(h = sigma2, lty = 2)
```

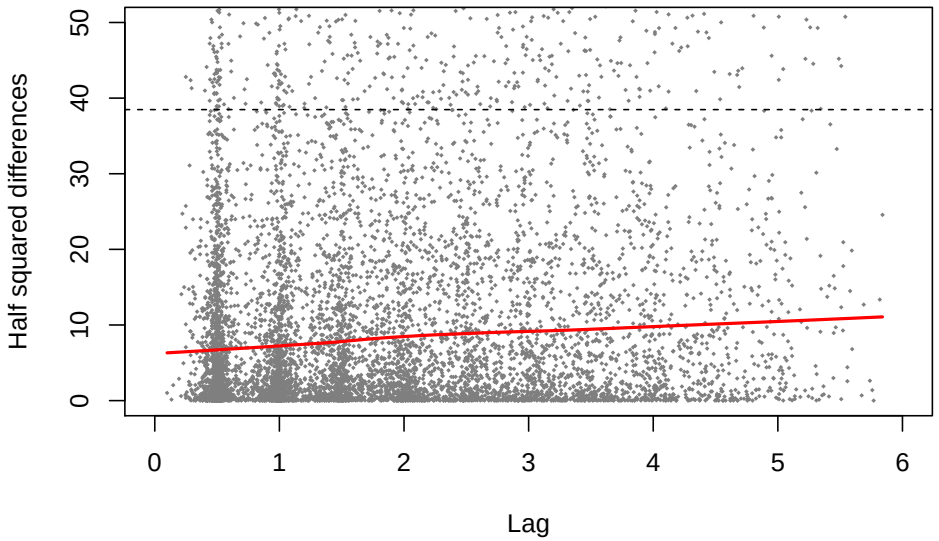


Table 2: **Variances of the residuals** from the simple spline fit above.

```
cd4.wide <- suppressWarnings(reshape(as.data.frame(cd4_df)[,c("ID", "resid", "yr")],
                    direction = "wide", v.names = "resid",
                    timevar = "yr", idvar = "ID"))
cv.cd4.wide <- cov(cd4.wide[,-1], use="pairwise.complete.obs")
cd4.resids.vars <- diag(cv.cd4.wide)
# re-order
cd4.resids.vars <- c(cd4.resids.vars[3:4], cd4.resids.vars[1:2], cd4.resids.vars[5:9])
cd4.resids.vars.df <- as.data.frame(cd4.resids.vars)
names(cd4.resids.vars.df) <- c('Variances')
kable(t(cd4.resids.vars.df), caption="Variances of Residuals", escape = F, digits = 6) %>%
  kable_styling(latex_options = c("hold_position"))
```

Table 2: Variances of Residuals

	resid.-3	resid.-2	resid.-1	resid.0	resid.1	resid.2	resid.3	resid.4	resid.5
Variances	27.98521	39.50176	40.56666	41.65118	28.78789	40.44546	37.82702	44.86887	53.37218

And the full variance-covariance matrix:

```
cv.cd4.wide
```

```
##          resid.-1  resid.0  resid.-3 resid.-2   resid.1   resid.2
## resid.-1 40.56666 20.68508 18.326055 25.28285 15.847902 16.085799
## resid.0  20.68508 41.65118 11.061741 20.43602 17.046089 13.198405
## resid.-3 18.32606 11.06174 27.985213 17.64018  8.873951 -3.473402
## resid.-2 25.28285 20.43602 17.640181 39.50176 11.738423 11.760714
## resid.1  15.84790 17.04609  8.873951 11.73842 28.787890 19.345113
## resid.2  16.08580 13.19840 -3.473402 11.76071 19.345113 40.445455
## resid.3  22.38065 13.50397  1.540104 14.42258 17.819749 26.889443
## resid.4  16.67129 18.52562        NA 39.67300 19.522630 26.481910
## resid.5  29.98802 21.97977        NA       NA 27.560134 33.393318
```

```
##            resid.3  resid.4  resid.5
## resid.-1 22.380652 16.67129 29.98802
## resid.0  13.503973 18.52562 21.97977
## resid.-3  1.540104       NA       NA
## resid.-2 14.422575 39.67300       NA
## resid.1  17.819749 19.52263 27.56013
## resid.2  26.889443 26.48191 33.39332
## resid.3  37.827020 32.41964 32.00380
## resid.4  32.419638 44.86887 33.81394
## resid.5  32.003805 33.81394 53.37218
```

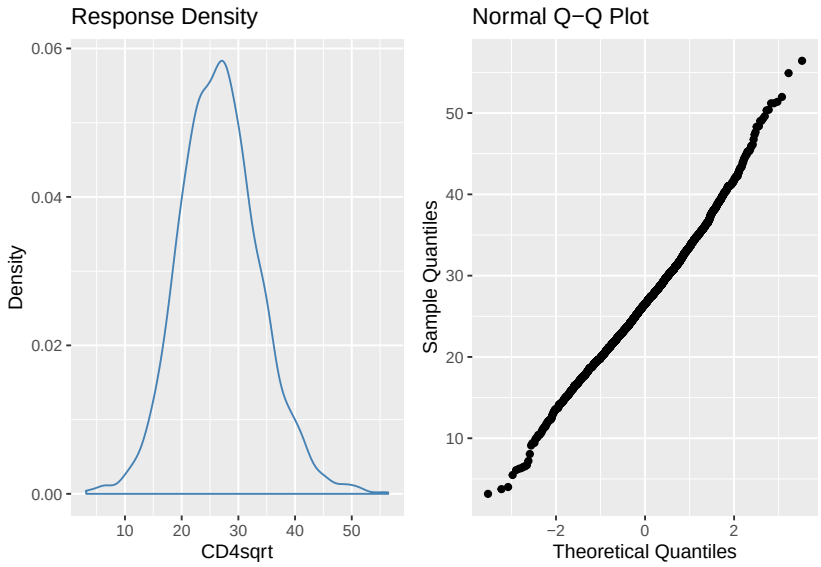
And the full correlation matrix:

```
cov2cor(cv.cd4.wide)
```

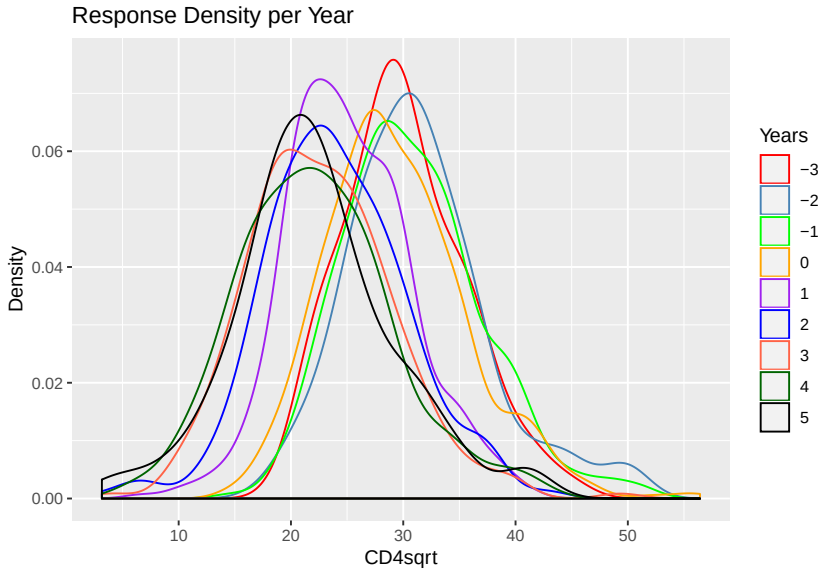
```
##           resid.-1   resid.0    resid.-3  resid.-2   resid.1    resid.2
## resid.-1 1.0000000 0.5032212  0.54390126 0.6315869 0.4637482  0.3971213
## resid.0  0.5032212 1.0000000  0.32400041 0.5038189 0.4922732  0.3215681
## resid.-3 0.5439013 0.3240004  1.00000000 0.5305551 0.3126424 -0.1032419
## resid.-2 0.6315869 0.5038189  0.53055509 1.0000000 0.3480941  0.2942325
## resid.1  0.4637482 0.4922732  0.31264235 0.3480941 1.0000000  0.5669327
## resid.2  0.3971213 0.3215681 -0.10324186 0.2942325 0.5669327  1.0000000
## resid.3  0.5713300 0.3402098  0.04733524 0.3731071 0.5400023  0.6874575
## resid.4  0.3907618 0.4285353          NA 0.9423544 0.5432014  0.6216442
## resid.5  0.6444744 0.4661782          NA        NA 0.7031030  0.7187325
##             resid.3   resid.4   resid.5
## resid.-1 0.57133002 0.3907618 0.6444744
## resid.0  0.34020984 0.4285353 0.4661782
## resid.-3 0.04733524        NA        NA
## resid.-2 0.37310710 0.9423544        NA
## resid.1  0.54000230 0.5432014 0.7031030
## resid.2  0.68745755 0.6216442 0.7187325
## resid.3  1.00000000 0.7869272 0.7122674
## resid.4  0.78692715 1.0000000 0.6909805
## resid.5  0.71226745 0.6909805 1.0000000
```

Response Variable Distribution:

```
pal <- c('red', 'steelblue', 'green', 'orange', 'purple', 'blue', 'tomato1', 'darkgreen', 'black')

plt.res.dens <- ggplot(cd4_df)+
  geom_density(aes(CD4sqrt), colour='steelblue')+
  ggtitle('Response Density') + ylab('Density')

plt.res.qq <- ggplot(cd4_df)+
  geom_qq(aes(sample=CD4sqrt))+
  xlab('Theoretical Quantiles')+
  ylab('Sample Quantiles')+
  ggtitle('Normal Q-Q Plot')

plt.res.dens.yr <- ggplot(cd4_df)+
  geom_density(aes(CD4sqrt, group=yr.f, colour=yr.f))+
  ggtitle('Response Density per Year') + ylab('Density') +
  scale_color_manual(name='Years', values = pal)

grid.arrange(plt.res.dens, plt.res.qq, ncol=2)
```

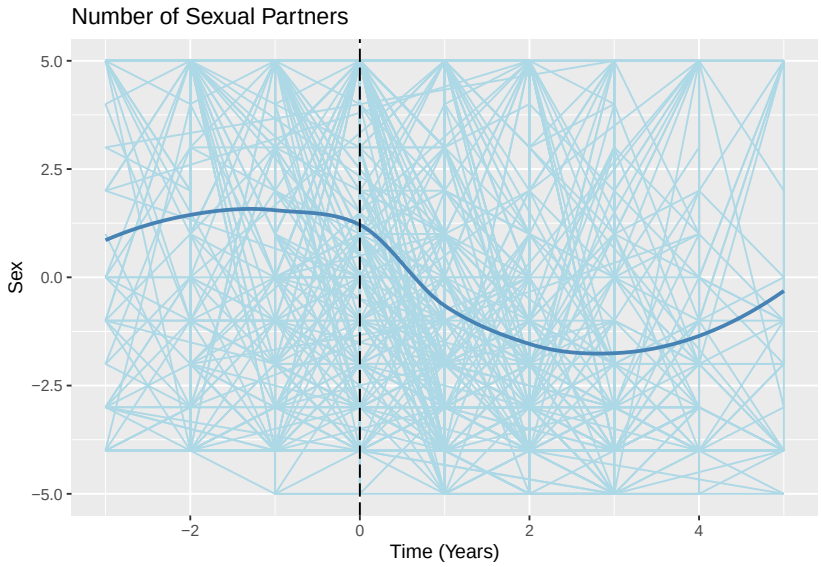



```
plt.res.dens.yr
```

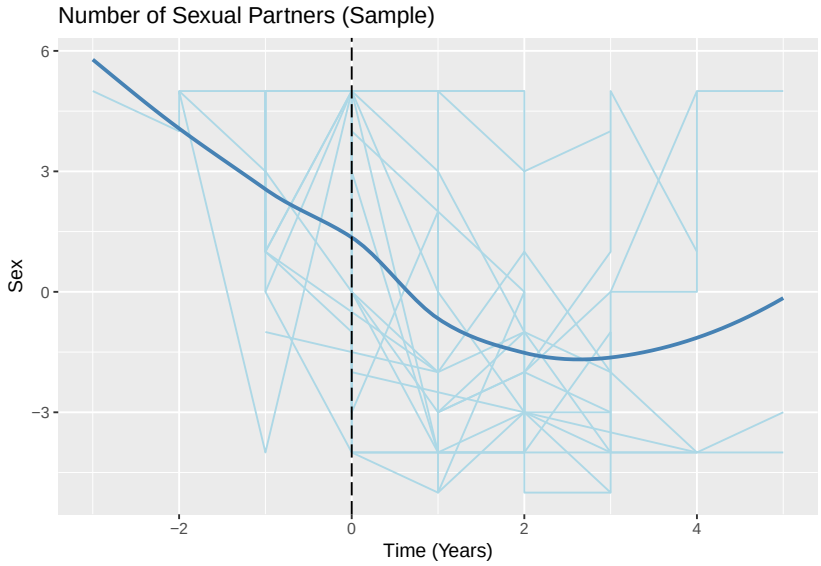


Time Trends for covariates : population and sample

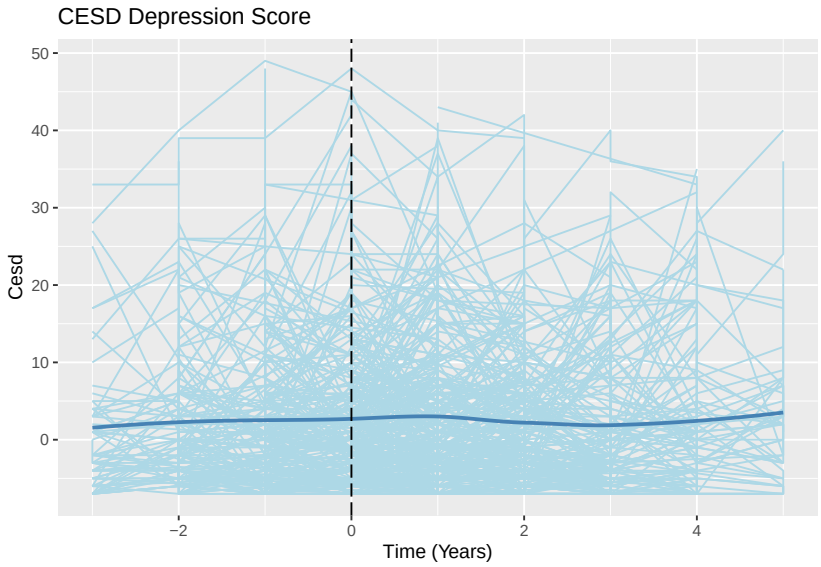
```
covariate_trend <- function(df, title, variable, colours=c('lightblue', 'steelblue', 'black')){
  g <- ggplot(df)+
    geom_line(aes_string(x = 'yr', y=variable, group='ID'), colour=colours[1])+
    geom_smooth(aes_string(x = 'yr', y=variable), colour=colours[2], se = F, method='loess', span=0.75)+
    ggtitle(title) + xlab('Time (Years)') +
    geom_vline(xintercept = 0, colour=colours[3], lty=5)
  return(g)
}
# partners
covariate_trend(cd4_df, 'Number of Sexual Partners', 'Sex')
```



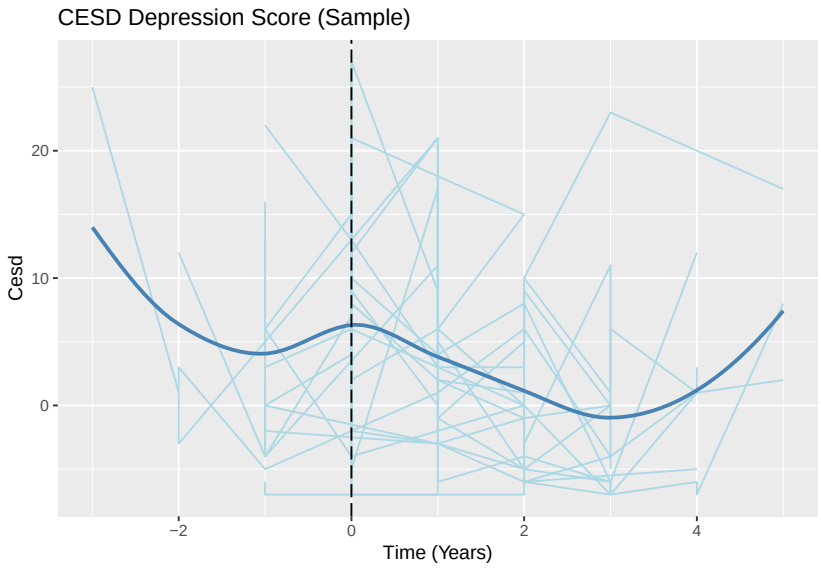

```
covariate_trend(sample_df, 'Number of Sexual Partners (Sample)', 'Sex')
```



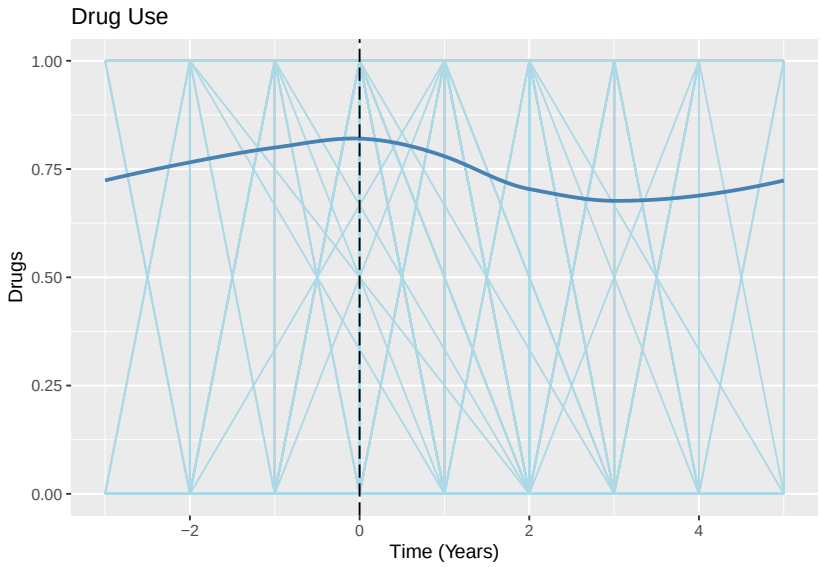

```
# depression score
covariate_trend(cd4_df, 'CESD Depression Score', 'Cesd')
```



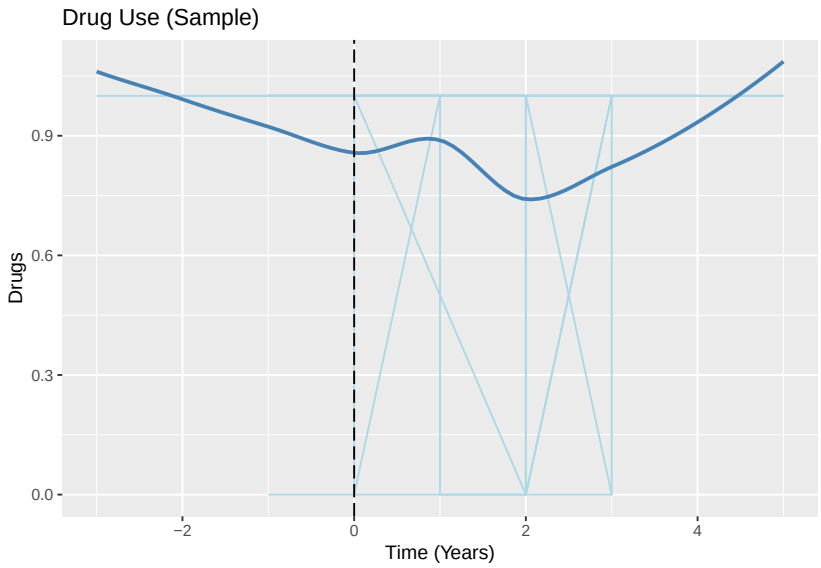

```
covariate_trend(sample_df, 'CESD Depression Score (Sample)', 'Cesd')
```



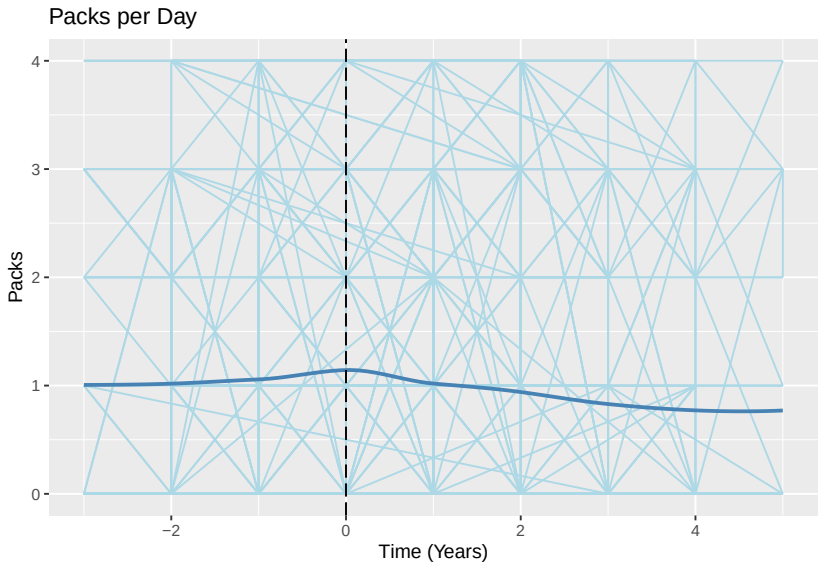

```
# drug use
covariate_trend(cd4_df, 'Drug Use', 'Drugs')
```



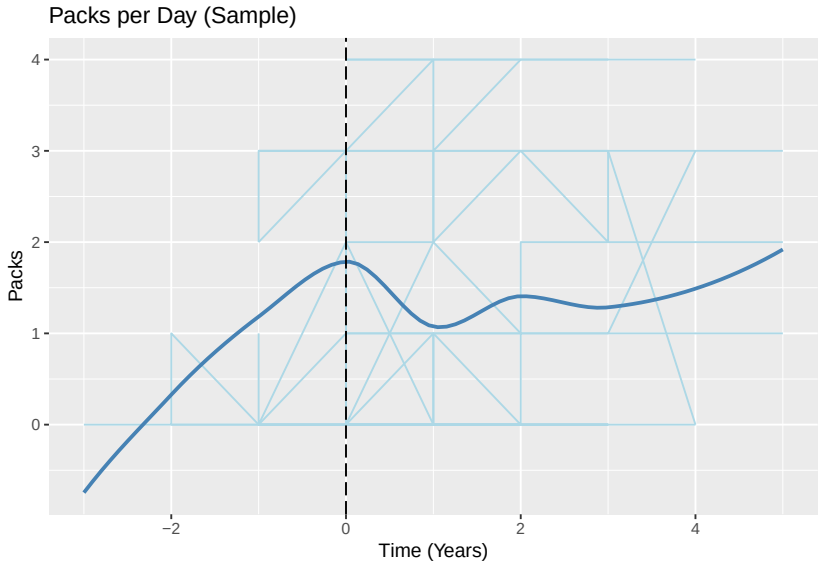

```
covariate_trend(sample_df, 'Drug Use (Sample)', 'Drugs')
```




```
# smoking
covariate_trend(cd4_df, 'Packs per Day', 'Packs')
```




```
covariate_trend(sample_df, 'Packs per Day (Sample)', 'Packs')
```

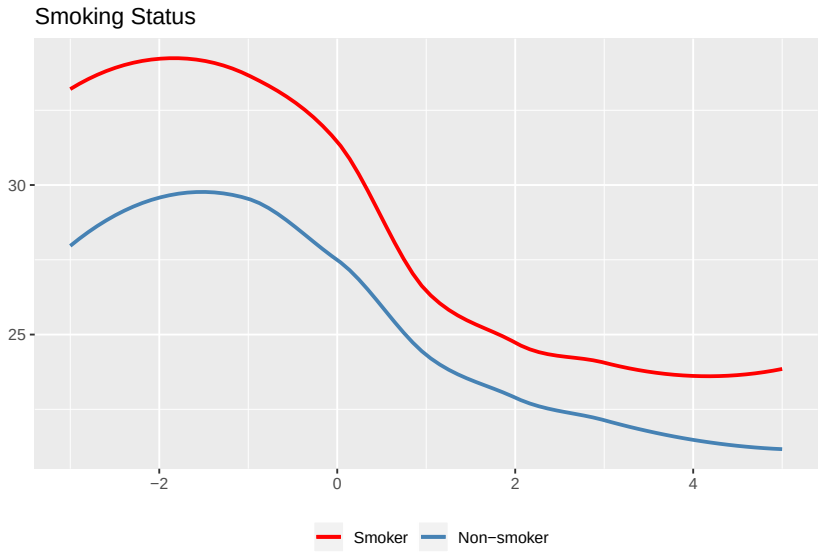


Covariates and the evolution of the response over time

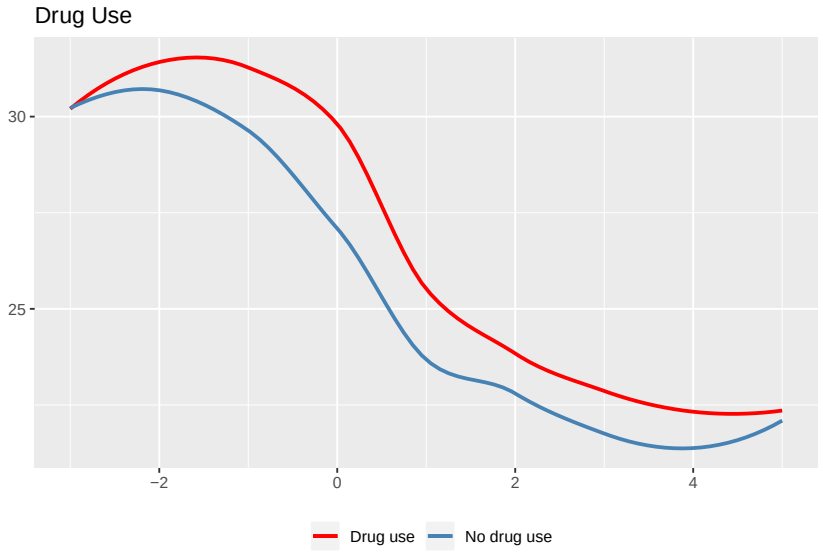
Not all of combinations of these plots were included in the main report due to space limitations.

```
covariate_response_profile <- function(df1, df2, title, labels, colours=c('red', 'steelblue')){
  g <- ggplot()+
    geom_smooth(data=df1, aes(x=yr, y=CD4sqrt, colour=colours[1]), se = F, method='loess', span=0.75)+
    geom_smooth(data=df2, aes(x=yr, y=CD4sqrt, colour=colours[2]), se = F, method='loess', span=0.75)+
    scale_colour_manual(values=colours, labels=labels)+
    scale_y_continuous(breaks = seq(20, 40, 5))+
    ggtitle(title) + xlab(NULL) + ylab(NULL) +
    theme(legend.title = element_blank(), legend.position = 'bottom')
  return(g)
}
# Smoker
covariate_response_profile(cd4_df %>% filter(smoker == T),
                           cd4_df %>% filter(smoker == F),
```

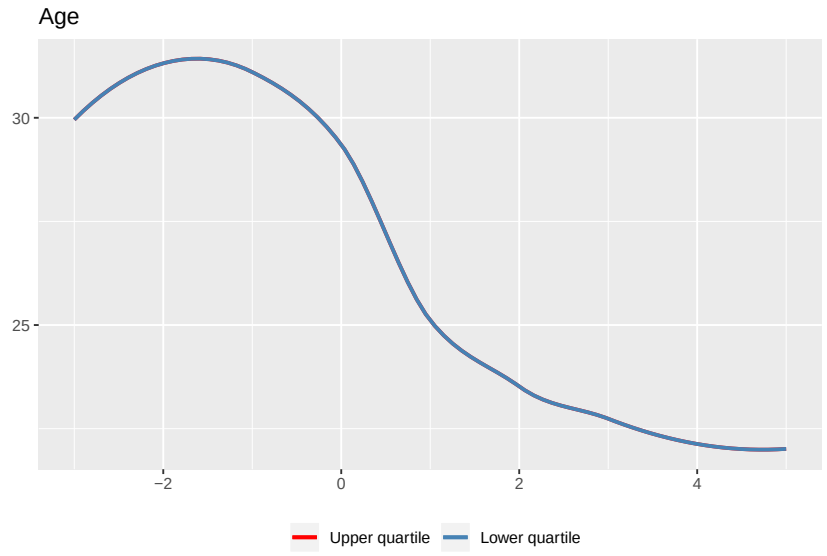
```
                        'Smoking Status', c('Smoker', 'Non-smoker'))
```




```
# drug use
covariate_response_profile(cd4_df %>% filter(Drugs == 1),
                           cd4_df %>% filter(Drugs == 0),
                           'Drug Use', c('Drug use', 'No drug use'))
```

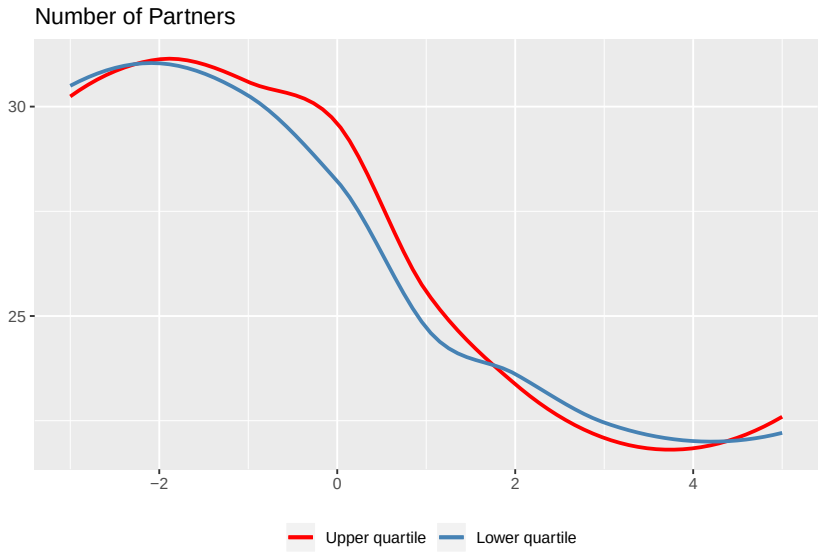



```
# age upper/lower quartiles
covariate_response_profile(cd4_df %>% filter(Age >= quantile(Age, 0.75)),
                           cd4_df %>% filter(Age <= quantile(Age, 0.25)),
                           'Age', c('Upper quartile', 'Lower quartile'))
```




```
# depression scores upper/lower quartiles
covariate_response_profile(cd4_df %>% filter(Cesd >= quantile(Cesd, 0.75)),
                           cd4_df %>% filter(Cesd <= quantile(Cesd, 0.25)),
                           'CESD Depression Score', c('Upper quartile', 'Lower quartile'))
```


```
# sex partners upper/lower quartiles
covariate_response_profile(cd4_df %>% filter(Sex >= quantile(Sex, 0.75)),
                           cd4_df %>% filter(Sex <= quantile(Sex, 0.25)),
                           'Number of Partners', c('Upper quartile', 'Lower quartile'))
```

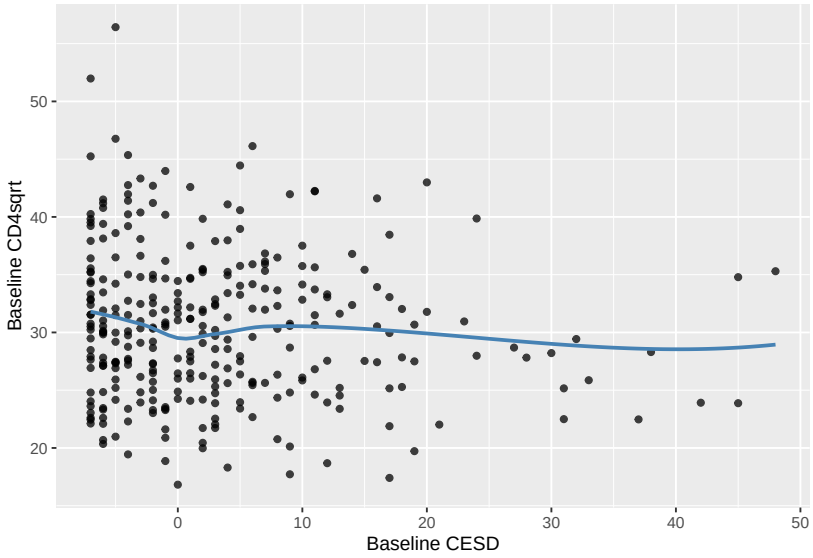


Separating Cross-sectional vs Longitudinal Effects

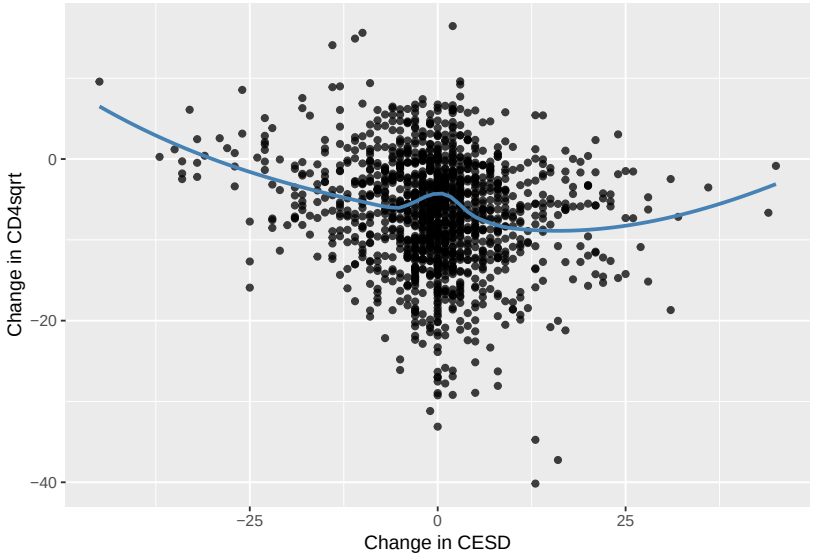
Plot covariate changes w.r.t baseline

```
baseline_ids <- cd4_df %>% filter(yr == 0) %>% select(ID) %>% distinct()
# there are 307 subjects with measurements at baseline

# data frame of baseline values for each covariate
baseline_df <- cd4_df %>% filter(ID %in% baseline_ids$ID & yr == 0) %>%
  group_by(ID) %>%
  top_n(n = 1, wt = -quarter) %>%
  arrange(ID) %>%
  mutate(
    CD4sqrti = CD4sqrt,
    Agei = Age,
    Cesdi = Cesd,
    Packsi = Packs,
    Drugsi = Drugs,
    Sexi = Sex
  ) %>%
  select(ID, CD4sqrti, Agei, Cesdi, Packsi, Drugsi, Sexi)

cd4_base_lines_df <- cd4_df %>% filter(ID %in% baseline_ids$ID& yr >= 0) %>%
  inner_join(baseline_df)

cross_trend_baseline <- function(df, variable, xlabel, span=0.75){
  g <- ggplot(baseline_df)+
    geom_point(aes_string(x = variable, y = 'CD4sqrti'), alpha=0.75)+
    geom_smooth(aes_string(x = variable, y = 'CD4sqrti'), color='steelblue', se = F, method='loess', spa
    xlab(xlabel) + ylab('Baseline CD4sqrt')
  return(g)
}

long_trend_baseline <- function(df, variable, xlabel, span=0.75){
  g <- ggplot(df)+
    geom_point(aes_string(x = variable, y = 'CD4sqrt - CD4sqrti'), alpha=0.75)+
    geom_smooth(aes_string(x = variable, y = 'CD4sqrt - CD4sqrti'), color='steelblue', se = F, method='
```

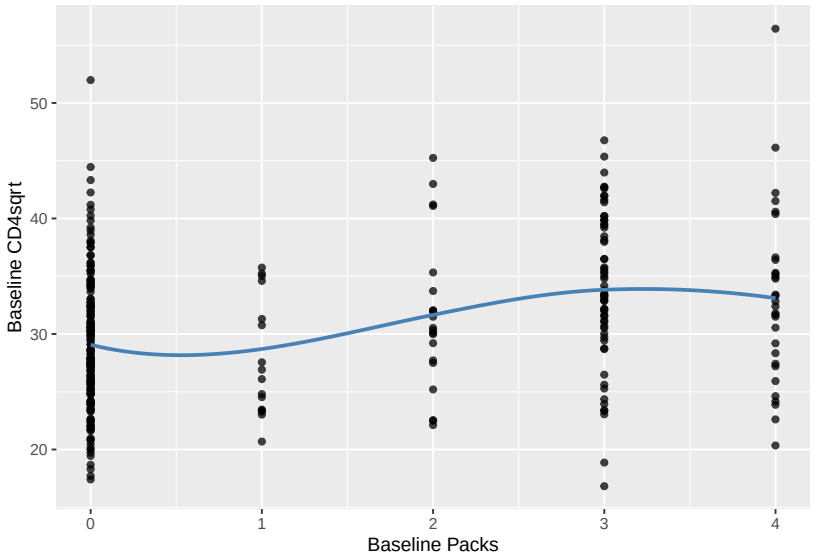
```
    xlab(xlabel) + ylab('Change in CD4sqrt')
  return(g)
}

# CESD
cross_trend_baseline(baseline_df, 'Cesdi', 'Baseline CESD')
```



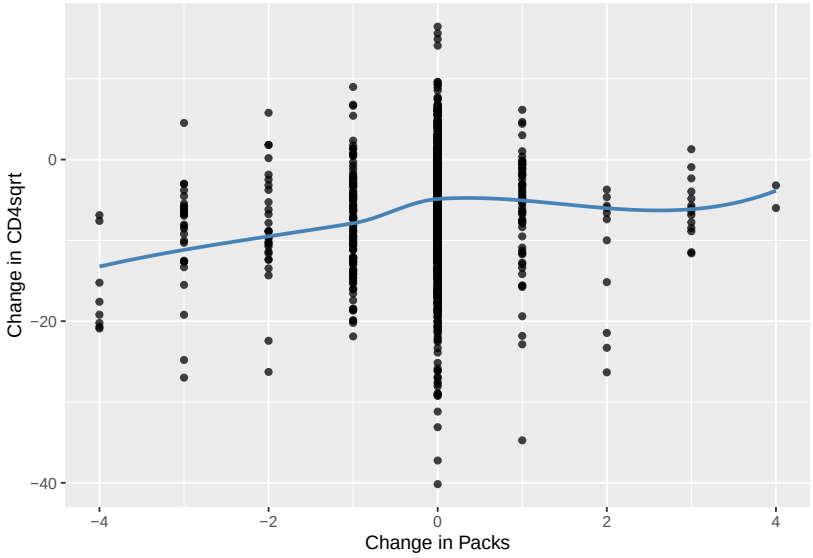

```
long_trend_baseline(cd4_base_lines_df, 'Cesd - Cesdi', 'Change in CESD')
```



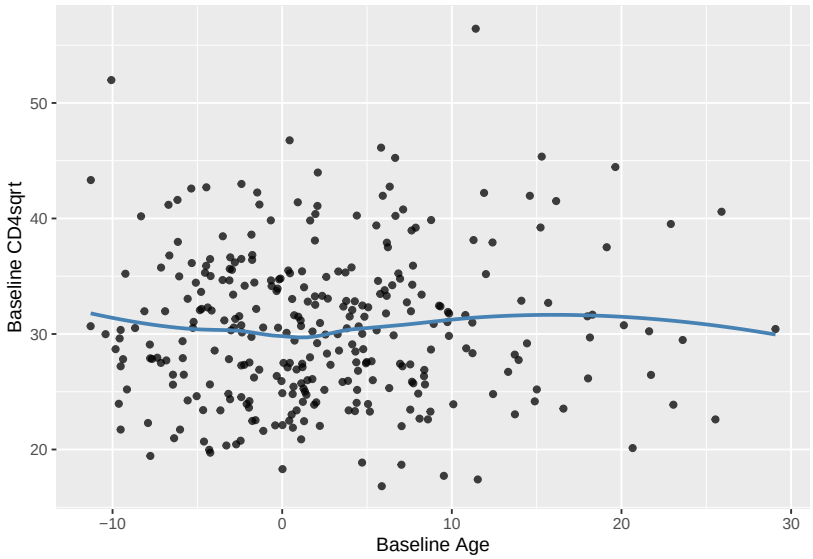

```
# Packs
cross_trend_baseline(baseline_df, 'Packsi', 'Baseline Packs')
```



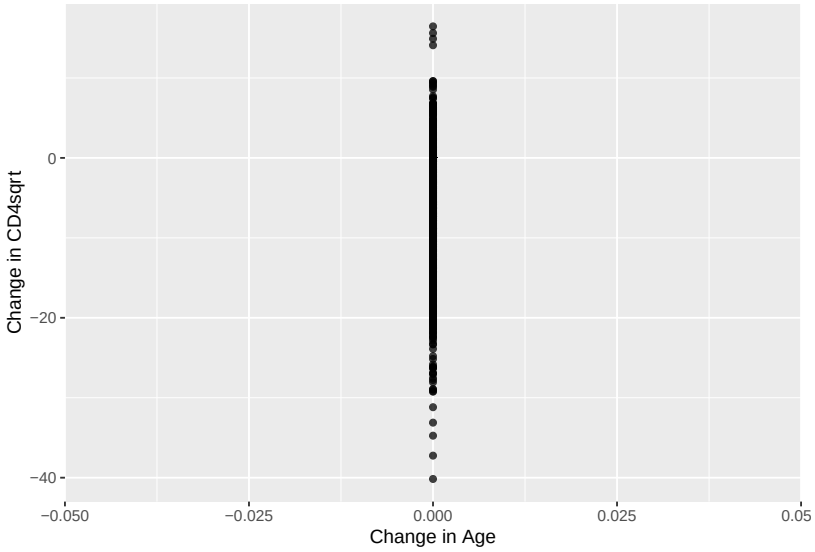

```
long_trend_baseline(cd4_base_lines_df, 'Packs - Packsi', 'Change in Packs', span=0.95)
```



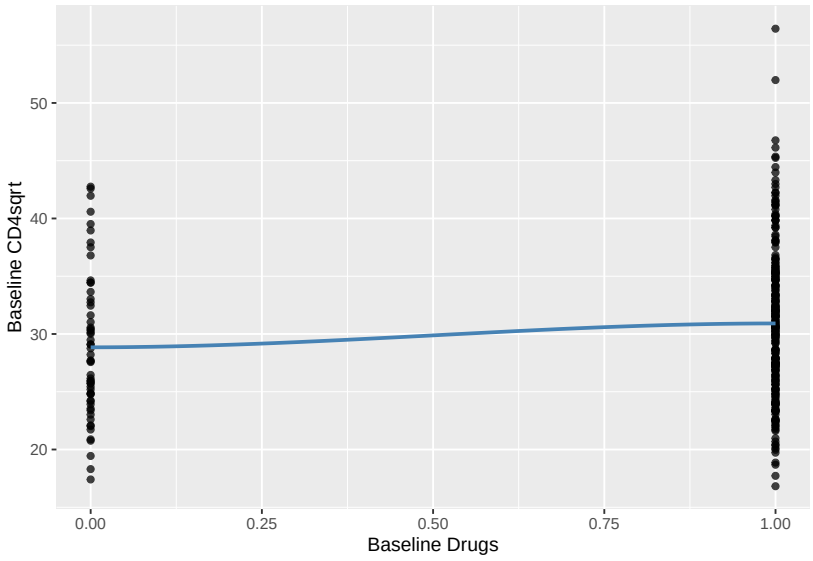

```
# Age
cross_trend_baseline(baseline_df, 'Agei', 'Baseline Age')
```



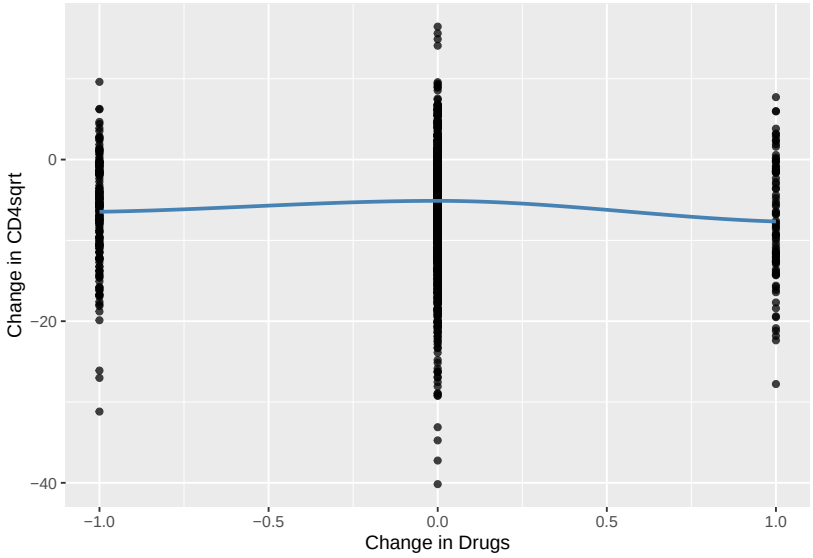

```
long_trend_baseline(cd4_base_lines_df, 'Age - Agei', 'Change in Age')
```



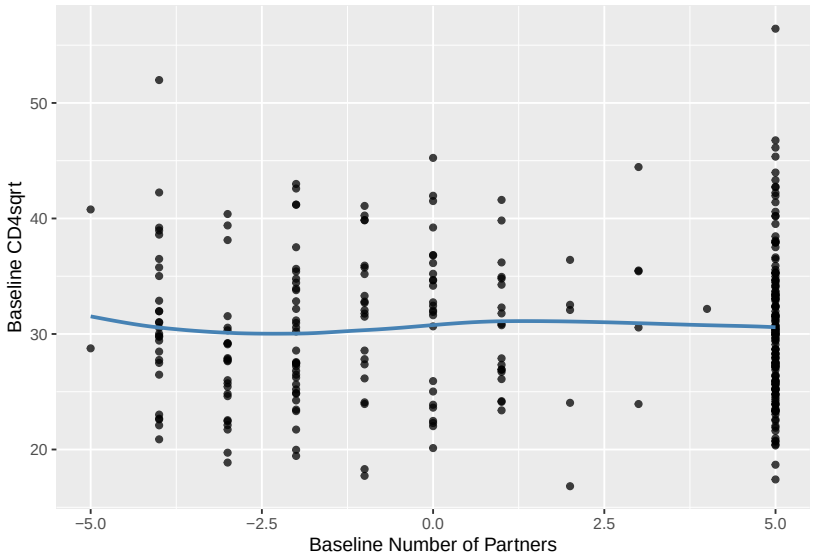

```
# Drugs
cross_trend_baseline(baseline_df, 'Drugsi', 'Baseline Drugs', span=0.95)
```



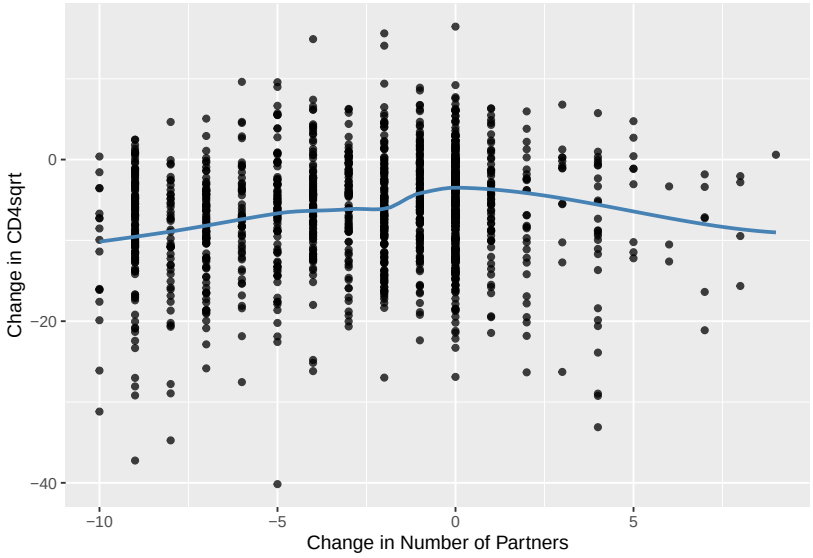

```
long_trend_baseline(cd4_base_lines_df, 'Drugs - Drugsi', 'Change in Drugs', span=0.95)
```




```
# Sex
cross_trend_baseline(baseline_df, 'Sexi', 'Baseline Number of Partners')
```

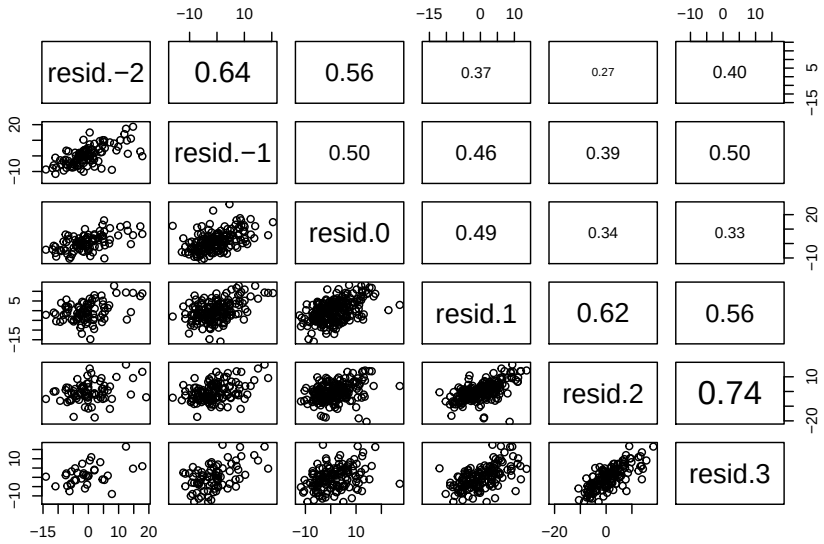



```
long_trend_baseline(cd4_base_lines_df, 'Sex - Sexi', 'Change in Number of Partners')
```



Use rounding to nearest year to avoid unbalanced issues and look at correlation plot:

```
panel.cor <- function(x, y, digits=2, prefix="", cex.cor){
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y,use="pairwise.complete.obs"))
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, sep="")
  if(missing(cex.cor)) cex <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex * r)
}
pairs(cd4.wide[,c(5, 2, 3, 6:8)], upper.panel = panel.cor)
```



Model Formulation

Initial 2 simple linear fixed effect models:

```
lm.basic1.fit <- lm(CD4sqrt ~ Time + smoker + Age + Drugs + Cesd + Sex, data = cd4_df)
sum.lm.basic1.fit <- summary(lm.basic1.fit)
sum.lm.basic1.fit
```

```
##
## Call:
## lm(formula = CD4sqrt ~ Time + smoker + Age + Drugs + Cesd + Sex,
##     data = cd4_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.7751  -4.2911  -0.3021   3.6952  27.5620
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.30720    0.28239  93.158  < 2e-16 ***
## Time        -1.62659    0.07026 -23.150  < 2e-16 ***
## smoker       2.87334    0.26732  10.749  < 2e-16 ***
## Age          0.02261    0.01702   1.329  0.18412
## Drugs        1.00294    0.30618   3.276  0.00107 **
## Cesd        -0.02923    0.01332  -2.194  0.02830 *
## Sex          0.01192    0.03715   0.321  0.74839
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.13 on 2369 degrees of freedom
## Multiple R-squared:  0.2479, Adjusted R-squared:  0.246
## F-statistic: 130.2 on 6 and 2369 DF,  p-value: < 2.2e-16
```

```
lm.basic2.smkr.int.fit <- lm(CD4sqrt ~ Time*smoker + Age + Drugs + Cesd + Sex, data = cd4_df)
sum.lm.basic2.smkr.int.fit <- summary(lm.basic2.smkr.int.fit)
sum.lm.basic2.smkr.int.fit
```

```
##
## Call:
## lm(formula = CD4sqrt ~ Time * smoker + Age + Drugs + Cesd + Sex,
##     data = cd4_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.4488  -4.2807  -0.3642   3.7588  27.1795
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.16067    0.28476  91.870  < 2e-16 ***
## Time        -1.45497    0.08531 -17.054  < 2e-16 ***
## smoker       3.26100    0.28842  11.306  < 2e-16 ***
## Age          0.02647    0.01701   1.556 0.119886
## Drugs        0.98109    0.30551   3.211 0.001339 **
## Cesd        -0.03024    0.01329  -2.275 0.022967 *
## Sex          0.01309    0.03706   0.353 0.723959
## Time:smoker -0.49813    0.14117  -3.529 0.000426 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.115 on 2368 degrees of freedom
## Multiple R-squared:  0.2519, Adjusted R-squared:  0.2497
## F-statistic: 113.9 on 7 and 2368 DF,  p-value: < 2.2e-16
```

```
anova(lm.basic1.fit, lm.basic2.smkr.int.fit)
```

```
## Analysis of Variance Table
##
## Model 1: CD4sqrt ~ Time + smoker + Age + Drugs + Cesd + Sex
## Model 2: CD4sqrt ~ Time * smoker + Age + Drugs + Cesd + Sex
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1   2369 89026
## 2   2368 88560  1    465.67 12.451 0.0004256 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3-knot piecewise linear model:

```
cd4_df$Time0 <- (cd4_df$Time)*(cd4_df$Time >= 0)
cd4_df$Time1 <- (cd4_df$Time)*(cd4_df$Time >= 1)
cd4_df$Time3 <- (cd4_df$Time)*(cd4_df$Time >= 3)
cd4_df$smoker.Time0 <- cd4_df$Time0 * (cd4_df$smoker == 1)
cd4_df$smoker.Time1 <- cd4_df$Time1 * (cd4_df$smoker == 1)
cd4_df$smoker.Time3 <- cd4_df$Time3 * (cd4_df$smoker == 1)
lm.basic3.3knots.fit <- lm(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                             smoker.Time0 + smoker.Time1 + smoker.Time3 +
                             Age + Drugs + Cesd + Sex, data = cd4_df)
sum.lm.basic3.3knots.fit <- summary(lm.basic3.3knots.fit)
sum.lm.basic3.3knots.fit$coefficients
```

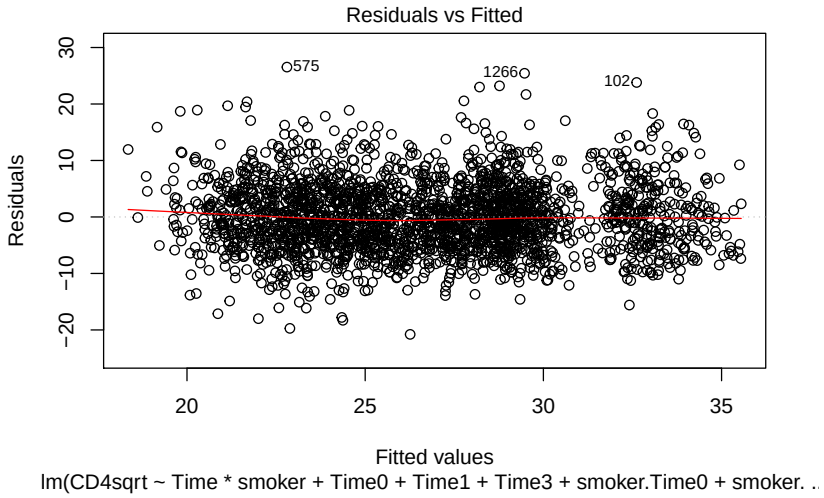
```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  27.62100224 0.42794520 64.5433153 0.000000e+00
## Time         -0.49509735 0.28322409 -1.7480764 8.058073e-02
## smoker        3.41618031 0.57565138  5.9344604 3.382017e-09
```

```
## Time0         -4.61848193 1.00864168 -4.5789124 4.916571e-06
## Time1          2.60944372 0.74984995  3.4799545 5.105892e-04
## Time3          0.82646767 0.19072683  4.3332533 1.530719e-05
## smoker.Time0 -1.11693076 1.61832253 -0.6901781 4.901500e-01
## smoker.Time1  0.82155255 1.20090047  0.6841138 4.939704e-01
## smoker.Time3  0.55113524 0.33201655  1.6599632 9.705460e-02
## Age           0.02634387 0.01681906  1.5663103 1.174099e-01
## Drugs         1.00121795 0.30277721  3.3067811 9.578813e-04
## Cesd         -0.03023167 0.01314667 -2.2995680 2.155955e-02
## Sex          -0.02650970 0.03706504 -0.7152212 4.745430e-01
## Time:smoker  -0.56652489 0.45668532 -1.2405148 2.149083e-01
```

```
AIC(lm.basic1.fit, lm.basic2.smkr.int.fit, lm.basic3.3knots.fit)
```

```
##                          df      AIC
## lm.basic1.fit             8 15368.25
## lm.basic2.smkr.int.fit    9 15357.79
## lm.basic3.3knots.fit     15 15303.75
```

Plot residuals from 3-knot model



Covariance Structures

Table 3 below shows the results of using the piecewise linear model as our 'maximal' model while exploring various candidate covariance structures.

```
# homogeneous variance
gls.homo.fit <- gls(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                      smoker.Time0 + smoker.Time1 + smoker.Time3 +
                      Age + Drugs + Cesd + Sex, data = cd4_df,
                    correlation = corCompSymm(form = ~ Time | ID))
# exponential model
gls.exp.fit <- gls(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                       smoker.Time0 + smoker.Time1 + smoker.Time3 +
                       Age + Drugs + Cesd + Sex, data = cd4_df,
                     correlation = corExp(form = ~ Time | ID))
# exponential model with nugget
gls.exp.nug.fit <- gls(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                         smoker.Time0 + smoker.Time1 + smoker.Time3 +
                         Age + Drugs + Cesd + Sex, data = cd4_df,
                       correlation = corExp(form = ~ Time | ID, nugget = T, value=c(2, 0.1)))
```

```
# Gaussian model with nugget
gls.gau.nug.fit <- gls(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                         smoker.Time0 + smoker.Time1 + smoker.Time3 +
                         Age + Drugs + Cesd + Sex, data = cd4_df,
                       correlation = corGaus(form = ~ Time | ID, nugget=T, value=c(1.5)))
# Gaussian model with nugget
gls.gau.nug.fit <- gls(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                         smoker.Time0 + smoker.Time1 + smoker.Time3 +
                         Age + Drugs + Cesd + Sex, data = cd4_df,
                       correlation = corGaus(form = ~ Time | ID, nugget=T, value=c(1.5)))

# accrue
variance.models <- c('Compound symmetry', 'Exponential Decay', 'Exponential Decay', 'Gaussian Decay', '
nugget.eff <- c(NA, 'No', 'Yes', 'No', 'Yes')
aics <- c(14444.61, 14454.11, 14264.99, 14863.40, 14277.90)

cov.struct.df <- tibble(
  Variance.Model = variance.models,
  Nugget = nugget.eff,
  AIC = aics
)
kable(cov.struct.df, caption="Covariance Structures Summary", escape = F, digits = 6) %>%
  kable_styling(latex_options = c("hold_position"))
```

Table 3: Covariance Structures Summary

Variance.Model	Nugget	AIC
Compound symmetry	NA	14444.61
Exponential Decay	No	14454.11
Exponential Decay	Yes	14264.99
Gaussian Decay	No	14863.40
Gaussian Decay	Yes	14277.90

Likelihood ratio tests performed against the compound symmetry model for both exponential and Gaussian models.

```
anova(gls.homo.fit, gls.exp.nug.fit)
```

```
##                 Model df      AIC      BIC    logLik   Test  L.Ratio
## gls.homo.fit        1 16 14444.61 14536.89 -7206.306
## gls.exp.nug.fit     2 17 14265.00 14363.04 -7115.497 1 vs 2 181.6172
##                 p-value
## gls.homo.fit
## gls.exp.nug.fit  <.0001
```

```
anova(gls.homo.fit, gls.gau.nug.fit)
```

```
##                 Model df      AIC      BIC    logLik   Test  L.Ratio
## gls.homo.fit        1 16 14444.61 14536.89 -7206.306
## gls.gau.nug.fit     2 17 14277.90 14375.94 -7121.950 1 vs 2 168.7124
##                 p-value
## gls.homo.fit
## gls.gau.nug.fit  <.0001
```

The next step is to refit this model using maximum likelihood and re-assess the fixed effects currently in this model through a sequence of drop-refit cycles.

```
gls.exp.nug.fit.ml <- update(gls.exp.nug.fit, method='ML')

# drop Age and refit
gls.exp1.nug.fit.ml <- gls(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                         smoker.Time0 + smoker.Time1 + smoker.Time3 + Drugs + Cesd + Sex,
                       correlation = corExp(form = ~ Time | ID, nugget = T, value=c(2, 0.1)),
                       data = cd4_df, method='ML')
sum.gls.exp1.nug.fit.ml <- summary(gls.exp1.nug.fit.ml)
anova(gls.exp.nug.fit.ml, gls.exp1.nug.fit.ml)
```

```
##                       Model df      AIC      BIC    logLik   Test   L.Ratio
## gls.exp.nug.fit.ml        1 17 14239.49 14337.63 -7102.745
## gls.exp1.nug.fit.ml       2 16 14237.68 14330.06 -7102.843 1 vs 2 0.1958589
##                       p-value
## gls.exp.nug.fit.ml
## gls.exp1.nug.fit.ml    0.6581
```

```
sum.gls.exp1.nug.fit.ml$tTable
```

```
##                     Value  Std.Error    t-value      p-value
## (Intercept)  28.67690237 0.46959962 61.0667068 0.000000e+00
## Time         -0.15264484 0.25780139 -0.5921025 5.538386e-01
## smoker        1.46566684 0.56018599  2.6163932 8.943018e-03
## Time0        -4.99998094 0.71140296 -7.0283387 2.723909e-12
## Time1         2.54524236 0.50207323  5.0694644 4.298791e-07
## Time3         0.51836481 0.14848976  3.4909128 4.901987e-04
## smoker.Time0 -0.29416816 1.14909809 -0.2559992 7.979738e-01
## smoker.Time1  0.68370781 0.80810374  0.8460644 3.976025e-01
## smoker.Time3  0.38167399 0.26605794  1.4345521 1.515471e-01
## Drugs         0.62123710 0.31388426  1.9791917 4.791033e-02
## Cesd         -0.04433498 0.01364592 -3.2489545 1.174541e-03
## Sex           0.09895057 0.03721560  2.6588470 7.893759e-03
## Time:smoker  -0.93307566 0.40577307 -2.2995012 2.156331e-02
```

```
gls.exp4.nug.fit.ml <- gls(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                           Drugs + Cesd + Sex,
                         correlation = corExp(form = ~ Time | ID, nugget = T, value=c(2, 0.1)),
                         data = cd4_df, method='ML')
# take out smoker.Time0
gls.exp2.nug.fit.ml <- gls(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                             smoker.Time1 + smoker.Time3 +
                             Drugs + Cesd + Sex,
                           correlation = corExp(form = ~ Time | ID, nugget = T, value=c(2, 0.1)),
                           data = cd4_df, method='ML')
anova(gls.exp1.nug.fit.ml, gls.exp2.nug.fit.ml)
```

```
##                       Model df      AIC      BIC    logLik   Test    L.Ratio
## gls.exp1.nug.fit.ml       1 16 14237.68 14330.06 -7102.843
## gls.exp2.nug.fit.ml       2 15 14235.75 14322.35 -7102.875 1 vs 2 0.06587116
##                       p-value
## gls.exp1.nug.fit.ml
## gls.exp2.nug.fit.ml    0.7974
```

```
summary(gls.exp2.nug.fit.ml)
```

```
## Generalized least squares fit by maximum likelihood
```

```
##   Model: CD4sqrt ~ Time * smoker + Time0 + Time1 + Time3 + smoker.Time1 +      smoker.Time3 + Drugs +
##   Data: cd4_df
##        AIC      BIC    logLik
##   14235.75 14322.35 -7102.875
##
## Correlation Structure: Exponential spatial correlation
##  Formula: ~Time | ID
##  Parameter estimate(s):
##     range    nugget
## 5.5045999 0.2943834
##
## Coefficients:
##                   Value Std.Error  t-value p-value
## (Intercept)  28.713698 0.4470470 64.22971  0.0000
## Time         -0.126077 0.2359790 -0.53427  0.5932
## smoker        1.374836 0.4339717  3.16803  0.0016
## Time0        -5.111527 0.5618721 -9.09731  0.0000
## Time1         2.614014 0.4237690  6.16849  0.0000
## Time3         0.525070 0.1461000  3.59391  0.0003
## smoker.Time1  0.505385 0.4093128  1.23472  0.2171
## smoker.Time3  0.361291 0.2538494  1.42325  0.1548
## Drugs         0.618203 0.3136013  1.97130  0.0488
## Cesd         -0.044284 0.0136417 -3.24619  0.0012
## Sex           0.099144 0.0372010  2.66510  0.0077
## Time:smoker  -1.003789 0.2971212 -3.37838  0.0007
##
##  Correlation:
##              (Intr) Time   smoker Time0  Time1  Time3  smk.T1 smk.T3
## Time          0.395
## smoker       -0.362 -0.138
## Time0        -0.398 -0.575 -0.004
## Time1         0.110  0.082  0.181 -0.802
## Time3         0.152  0.120 -0.061 -0.241 -0.034
## smoker.Time1  0.180  0.370 -0.487  0.027 -0.380  0.231
## smoker.Time3 -0.024  0.013  0.112 -0.037  0.162 -0.516 -0.429
## Drugs        -0.541 -0.031 -0.039 -0.006  0.043 -0.002 -0.006 -0.002
## Cesd         -0.068 -0.032  0.007 -0.017  0.036 -0.004 -0.035  0.001
## Sex           0.002  0.076 -0.043  0.111 -0.103 -0.107  0.063 -0.029
## Time:smoker  -0.116 -0.485  0.290 -0.006  0.296  0.002 -0.781  0.007
##              Drugs  Cesd   Sex
## Time
## smoker
## Time0
## Time1
## Time3
## smoker.Time1
## smoker.Time3
## Drugs
## Cesd         -0.014
## Sex          -0.138 -0.045
## Time:smoker   0.009  0.045 -0.044
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
```

```
## -3.30647137 -0.62678600 -0.02401535  0.64542868  4.33027322
##
## Residual standard error: 6.153606
## Degrees of freedom: 2376 total; 2364 residual
```

```
# take out smoker.Time1
gls.exp3.nug.fit.ml <- gls(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                             smoker.Time3 +
                             Drugs + Cesd + Sex,
                           correlation = corExp(form = ~ Time | ID, nugget = T, value=c(2, 0.1)),
                           data = cd4_df, method='ML')
anova(gls.exp2.nug.fit.ml, gls.exp3.nug.fit.ml)
```

```
##                      Model df      AIC      BIC    logLik   Test  L.Ratio
## gls.exp2.nug.fit.ml      1 15 14235.75 14322.35 -7102.875
## gls.exp3.nug.fit.ml      2 14 14235.28 14316.11 -7103.641 1 vs 2 1.530821
##                      p-value
## gls.exp2.nug.fit.ml
## gls.exp3.nug.fit.ml   0.216
```

```
summary(gls.exp3.nug.fit.ml)
```

```
## Generalized least squares fit by maximum likelihood
##   Model: CD4sqrt ~ Time * smoker + Time0 + Time1 + Time3 + smoker.Time3 +      Drugs + Cesd + Sex
##   Data: cd4_df
##        AIC      BIC    logLik
##   14235.28 14316.11 -7103.641
##
## Correlation Structure: Exponential spatial correlation
##  Formula: ~Time | ID
##  Parameter estimate(s):
##     range    nugget
## 5.4939855 0.2949111
##
## Coefficients:
##                   Value Std.Error  t-value p-value
## (Intercept)  28.613441 0.4396531 65.08186  0.0000
## Time         -0.233966 0.2192496 -1.06712  0.2860
## smoker        1.637719 0.3790024  4.32113  0.0000
## Time0        -5.130389 0.5618161 -9.13179  0.0000
## Time1         2.812942 0.3921217  7.17364  0.0000
## Time3         0.483629 0.1421790  3.40155  0.0007
## smoker.Time3  0.495774 0.2293822  2.16135  0.0308
## Drugs         0.620875 0.3136252  1.97967  0.0479
## Cesd         -0.043679 0.0136347 -3.20352  0.0014
## Sex           0.096165 0.0371309  2.58988  0.0097
## Time:smoker  -0.717558 0.1857015 -3.86404  0.0001
##
##  Correlation:
##              (Intr) Time   smoker Time0  Time1  Time3  smk.T3 Drugs
## Time          0.359
## smoker       -0.319  0.052
## Time0        -0.410 -0.630  0.010
## Time1         0.196  0.259 -0.004 -0.856
## Time3         0.116  0.038  0.061 -0.255  0.060
```

```
## smoker.Time3  0.060  0.204 -0.122 -0.028 -0.001 -0.475
## Drugs        -0.549 -0.031 -0.048 -0.006  0.045 -0.001 -0.005
## Cesd         -0.063 -0.021 -0.011 -0.016  0.024  0.004 -0.015 -0.014
## Sex          -0.010  0.057 -0.014  0.109 -0.085 -0.126 -0.003 -0.138
## Time:smoker   0.039 -0.339 -0.166  0.024 -0.002  0.299 -0.580  0.006
##              Cesd   Sex
## Time
## smoker
## Time0
## Time1
## Time3
## smoker.Time3
## Drugs
## Cesd
## Sex          -0.043
## Time:smoker   0.028  0.009
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -3.31648886 -0.63024169 -0.02186527  0.64301790  4.33187149
##
## Residual standard error: 6.152046
## Degrees of freedom: 2376 total; 2365 residual
```

```
# take out smoker.Time3
gls.exp4.nug.fit.ml <- gls(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                           Drugs + Cesd + Sex,
                            correlation = corExp(form = ~ Time | ID, nugget = T, value=c(2, 0.1)),
                            data = cd4_df, method='ML')
anova(gls.exp3.nug.fit.ml, gls.exp4.nug.fit.ml)
```

```
##                     Model df      AIC      BIC    logLik   Test  L.Ratio
## gls.exp3.nug.fit.ml     1 14 14235.28 14316.11 -7103.641
## gls.exp4.nug.fit.ml     2 13 14237.96 14313.01 -7105.981 1 vs 2 4.680938
##                     p-value
## gls.exp3.nug.fit.ml
## gls.exp4.nug.fit.ml  0.0305
```

```
summary(gls.exp4.nug.fit.ml)
```

```
## Generalized least squares fit by maximum likelihood
##   Model: CD4sqrt ~ Time * smoker + Time0 + Time1 + Time3 + Drugs + Cesd +      Sex
##   Data: cd4_df
##        AIC      BIC    logLik
##   14237.96 14313.01 -7105.981
##
## Correlation Structure: Exponential spatial correlation
##  Formula: ~Time | ID
##  Parameter estimate(s):
##     range    nugget
## 5.4497074 0.2932164
##
## Coefficients:
##                 Value Std.Error  t-value p-value
## (Intercept) 28.556413 0.4394516 64.98192  0.0000
```

```
## Time        -0.330707 0.2150764 -1.53763  0.1243
## smoker       1.736490 0.3765655  4.61139  0.0000
## Time0       -5.094444 0.5619903 -9.06500  0.0000
## Time1        2.811457 0.3922413  7.16767  0.0000
## Time3        0.628540 0.1253094  5.01591  0.0000
## Drugs        0.624050 0.3138564  1.98833  0.0469
## Cesd        -0.043203 0.0136425 -3.16678  0.0016
## Sex          0.096322 0.0371662  2.59167  0.0096
## Time:smoker -0.484140 0.1515726 -3.19411  0.0014
##
##  Correlation:
##             (Intr) Time   smoker Time0  Time1  Time3  Drugs  Cesd   Sex
## Time         0.356
## smoker      -0.315  0.079
## Time0       -0.410 -0.638  0.007
## Time1        0.196  0.264 -0.004 -0.856
## Time3        0.164  0.156  0.003 -0.304  0.068
## Drugs       -0.549 -0.030 -0.049 -0.006  0.045 -0.003
## Cesd        -0.062 -0.018 -0.013 -0.016  0.024 -0.004 -0.014
## Sex         -0.009  0.058 -0.015  0.109 -0.085 -0.144 -0.138 -0.043
## Time:smoker  0.090 -0.276 -0.293  0.010 -0.002  0.033  0.005  0.023  0.009
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -3.28265506 -0.62421825 -0.01781463  0.64676820  4.28889470
##
## Residual standard error: 6.161593
## Degrees of freedom: 2376 total; 2366 residual
```

Table 4 shows the model fit improvements made at each step when this is done.

```
removed.var <- c('Refit by ML', 'Remove Age', 'Remove smoker.Time0', 'Remove smoker.Time1', 'Remove smo
removed.aics <- c(14239.49, 14237.69, 14235.75, 14235.28, 14237.96)
fixed.ml.refit.df <- tibble(
  Action = removed.var,
  AIC = removed.aics
)
kable(fixed.ml.refit.df, caption="Fixed Effects Simplification Steps", escape = F, digits = 6) %>%
  kable_styling(latex_options = c("hold_position"))
```

Table 4: Fixed Effects Simplification Steps

Action	AIC
Refit by ML	14239.49
Remove Age	14237.69
Remove smoker.Time0	14235.75
Remove smoker.Time1	14235.28
Remove smoker.Time3	14237.96

Final GLS model:

```
gls.exp.fit <- gls(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 + Drugs + Cesd + Sex,
                   correlation = corExp(form = ~ Time | ID, nugget = T, value=c(2, 0.1)), data = cd4_df
AIC(gls.exp.fit)
```

```
## [1] 14259.54
```

Random Effects

Intercept Only:

```
me.exp.fit <- lme(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                    Drugs + Cesd + Sex, data = cd4_df,
                  random = ~ 1 | ID,
                  correlation = corExp(form = ~ Time | ID, nugget = T, value=c(2, 0.1)),
                  control = lmeControl(opt = 'optim', maxIter = 200))
summary(me.exp.fit)
```

```
## Linear mixed-effects model fit by REML
##  Data: cd4_df
##        AIC      BIC    logLik
##   14261.61 14342.38 -7116.806
##
## Random effects:
##  Formula: ~1 | ID
##         (Intercept) Residual
## StdDev:   0.8709348  6.11799
##
## Correlation Structure: Exponential spatial correlation
##  Formula: ~Time | ID
##  Parameter estimate(s):
##     range    nugget
## 5.2275789 0.2969823
## Fixed effects: CD4sqrt ~ Time * smoker + Time0 + Time1 + Time3 + Drugs + Cesd +      Sex
##                 Value Std.Error   DF  t-value p-value
## (Intercept) 28.555814 0.4397272 1998 64.93984  0.0000
## Time        -0.331342 0.2152309 1998 -1.53947  0.1238
## smoker       1.733834 0.3767272 1998  4.60236  0.0000
## Time0       -5.089535 0.5619440 1998 -9.05701  0.0000
## Time1        2.808323 0.3921278 1998  7.16175  0.0000
## Time3        0.627509 0.1254243 1998  5.00309  0.0000
## Drugs        0.623072 0.3138874 1998  1.98502  0.0473
## Cesd        -0.043207 0.0136426 1998 -3.16705  0.0016
## Sex          0.096282 0.0371749 1998  2.58996  0.0097
## Time:smoker -0.483620 0.1516019 1998 -3.19007  0.0014
##  Correlation:
##             (Intr) Time   smoker Time0  Time1  Time3  Drugs  Cesd   Sex
## Time         0.356
## smoker      -0.315  0.079
## Time0       -0.410 -0.638  0.007
## Time1        0.196  0.263 -0.004 -0.856
## Time3        0.164  0.156  0.003 -0.305  0.068
## Drugs       -0.549 -0.030 -0.049 -0.006  0.044 -0.003
## Cesd        -0.062 -0.018 -0.013 -0.017  0.024 -0.004 -0.014
## Sex         -0.009  0.058 -0.015  0.109 -0.085 -0.143 -0.138 -0.043
## Time:smoker  0.090 -0.276 -0.293  0.010 -0.002  0.034  0.005  0.023  0.009
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.29249705 -0.61275838 -0.01706265  0.63169992  4.21185980
##
## Number of Observations: 2376
```

```
## Number of Groups: 369
```

Intercept and Slope

```
me.exp2.fit <- lme(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                     Drugs + Cesd + Sex, data = cd4_df,
                   random = ~ Time | ID,
                   correlation = corExp(form = ~ Time | ID, nugget = T, value=c(2, 0.1)),
                   control = lmeControl(opt = 'optim', maxIter = 200))
summary(me.exp2.fit)
```

```
## Linear mixed-effects model fit by REML
##  Data: cd4_df
##        AIC      BIC    logLik
##   14230.02 14322.33 -7099.011
##
## Random effects:
##  Formula: ~Time | ID
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr
## (Intercept) 2.0964407 (Intr)
## Time        0.9845339 0.503
## Residual    5.3098138
##
## Correlation Structure: Exponential spatial correlation
##  Formula: ~Time | ID
##  Parameter estimate(s):
##     range    nugget
## 4.8780932 0.4217302
## Fixed effects: CD4sqrt ~ Time * smoker + Time0 + Time1 + Time3 + Drugs + Cesd +      Sex
##                 Value Std.Error   DF  t-value p-value
## (Intercept) 28.599133 0.4143627 1998 69.01957  0.0000
## Time        -0.264699 0.2147327 1998 -1.23269  0.2178
## smoker       1.878705 0.3560576 1998  5.27641  0.0000
## Time0       -5.332975 0.5557041 1998 -9.59679  0.0000
## Time1        2.899091 0.3877502 1998  7.47670  0.0000
## Time3        0.556828 0.1232737 1998  4.51700  0.0000
## Drugs        0.602292 0.3107112 1998  1.93843  0.0527
## Cesd        -0.041618 0.0134624 1998 -3.09141  0.0020
## Sex          0.092962 0.0365338 1998  2.54456  0.0110
## Time:smoker -0.603932 0.1609632 1998 -3.75198  0.0002
##  Correlation:
##             (Intr) Time   smoker Time0  Time1  Time3  Drugs  Cesd   Sex
## Time         0.384
## smoker      -0.309  0.052
## Time0       -0.411 -0.625  0.010
## Time1        0.210  0.270 -0.006 -0.860
## Time3        0.163  0.134  0.004 -0.281  0.065
## Drugs       -0.575 -0.038 -0.059 -0.003  0.045  0.002
## Cesd        -0.064 -0.018 -0.020 -0.017  0.023 -0.005 -0.013
## Sex         -0.010  0.065 -0.011  0.107 -0.084 -0.150 -0.137 -0.043
## Time:smoker  0.049 -0.297 -0.198  0.011 -0.002  0.039  0.019  0.026 -0.004
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
```

```
## -3.19926913 -0.57397965 -0.01855209  0.56955317  4.35946801
##
## Number of Observations: 2376
## Number of Groups: 369
```

Intercept, Slope and additional slope effects for each piecewise segment:

```
me.exp3.fit <- lme(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                     Drugs + Cesd + Sex, data = cd4_df,
                   random = ~ Time + Time0 + Time1 + Time3 | ID,
                   correlation = corExp(form = ~ Time | ID, nugget = T, value=c(2, 0.1)),
                   control = lmeControl(opt = 'optim', maxIter = 200))
```

```
AIC(me.exp.fit, me.exp2.fit, me.exp3.fit)
```

```
##             df      AIC
## me.exp.fit  14 14261.61
## me.exp2.fit 16 14230.02
## me.exp3.fit 28 14247.21
```

Drop Drugs covariate for final model

```
me.exp4.fit <- lme(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                     Cesd + Sex, data = cd4_df,
                   random = ~ Time | ID,
                   correlation = corExp(form = ~ Time | ID, nugget = T, value=c(2, 0.1)),
                   control = lmeControl(opt = 'optim', maxIter = 200))
summary(me.exp4.fit)
```

```
## Linear mixed-effects model fit by REML
##  Data: cd4_df
##        AIC      BIC    logLik
##   14226.69 14313.23 -7098.345
##
## Random effects:
##  Formula: ~Time | ID
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev   Corr
## (Intercept) 3.870145 (Intr)
## Time        1.194359 0.114
## Residual    4.101988
##
## Correlation Structure: Exponential spatial correlation
##  Formula: ~Time | ID
##  Parameter estimate(s):
##     range    nugget
## 0.7225396 0.5739228
## Fixed effects: CD4sqrt ~ Time * smoker + Time0 + Time1 + Time3 + Cesd + Sex
##                 Value Std.Error   DF  t-value p-value
## (Intercept) 29.052133 0.3327126 1999 87.31901  0.0000
## Time        -0.291148 0.2182508 1999 -1.33401  0.1824
## smoker       1.849093 0.3566748 1999  5.18425  0.0000
## Time0       -5.272980 0.5604966 1999 -9.40769  0.0000
## Time1        2.827089 0.3901273 1999  7.24658  0.0000
## Time3        0.554318 0.1232589 1999  4.49718  0.0000
## Cesd        -0.041219 0.0134607 1999 -3.06220  0.0022
## Sex          0.099269 0.0362553 1999  2.73805  0.0062
```

```
## Time:smoker -0.576220 0.1621713 1999 -3.55316  0.0004
##  Correlation:
##             (Intr) Time   smoker Time0  Time1  Time3  Cesd   Sex
## Time         0.399
## smoker      -0.427  0.057
## Time0       -0.492 -0.630  0.007
## Time1        0.296  0.276 -0.002 -0.861
## Time3        0.230  0.152  0.001 -0.282  0.057
## Cesd        -0.089 -0.018 -0.020 -0.017  0.024 -0.008
## Sex         -0.107  0.063 -0.023  0.105 -0.077 -0.155 -0.044
## Time:smoker  0.080 -0.295 -0.214  0.013 -0.004  0.038  0.025  0.004
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -3.717867678 -0.525988959 -0.001285655  0.527522235  5.105774184
##
## Number of Observations: 2376
## Number of Groups: 369
```

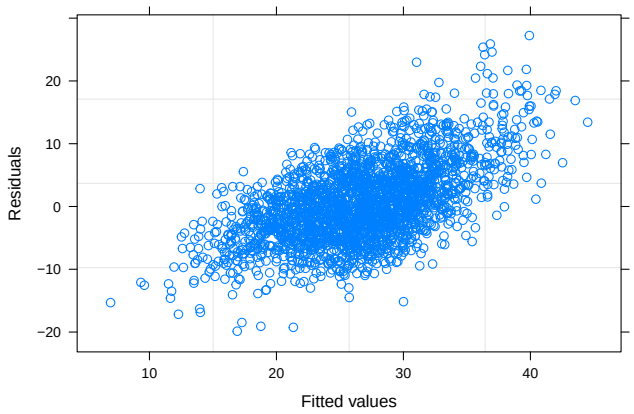
```
AIC(me.exp4.fit)
```

```
## [1] 14226.69
```

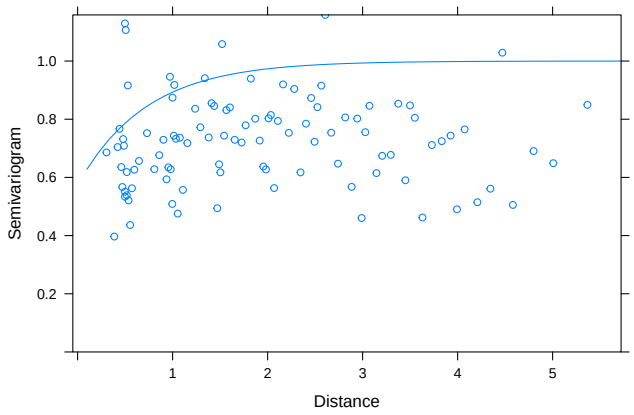
Diagnostic Plots

A residual plot, a variogram and an ACF for the residuals from this final model are shown below

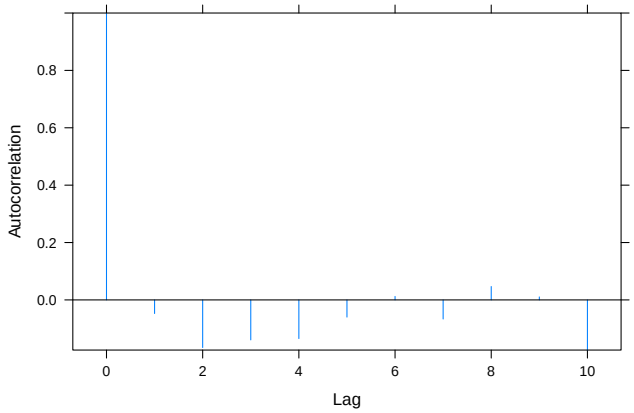
```
par(mfrow=c(1,2))
plot(me.exp4.fit, form = resid(., level = 0) ~ fitted(.))
```



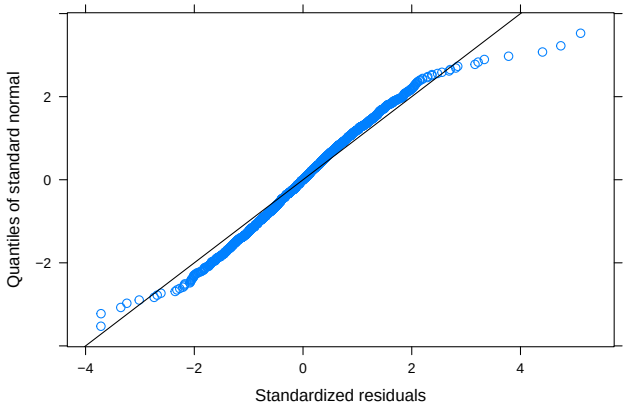

```
me.exp4.fit.variogram <- Variogram(me.exp4.fit, form = ~ Time | ID, nint = 100, robust = T)
plot(me.exp4.fit.variogram)
```



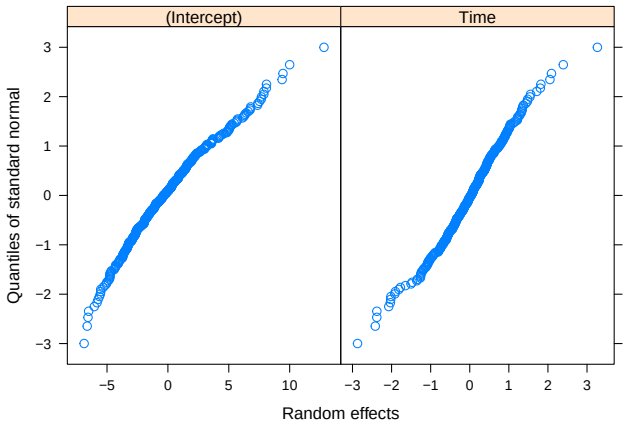

```
plot(ACF(me.exp4.fit))
```




```
par(mfrow=c(1,2))
qqnorm(me.exp4.fit, ~ resid(., type = "p"), abline = c(0, 1))
```




```
qqnorm(me.exp4.fit, ~ ranef(.))
```



Final Model Fixed Effects

```
sum.me.exp4.fit <- summary(me.exp4.fit)
round(sum.me.exp4.fit$tTable, digits = 4)
```

```
##                Value Std.Error   DF t-value p-value
## (Intercept) 29.0521    0.3327 1999 87.3190  0.0000
## Time        -0.2911    0.2183 1999 -1.3340  0.1824
## smoker       1.8491    0.3567 1999  5.1843  0.0000
## Time0       -5.2730    0.5605 1999 -9.4077  0.0000
## Time1        2.8271    0.3901 1999  7.2466  0.0000
## Time3        0.5543    0.1233 1999  4.4972  0.0000
## Cesd        -0.0412    0.0135 1999 -3.0622  0.0022
## Sex          0.0993    0.0363 1999  2.7381  0.0062
## Time:smoker -0.5762    0.1622 1999 -3.5532  0.0004
```

Components of Variability

```
sigma.b <- 3.870145 # (Intercept StdDev)
sigma.squ.b <- sigma.b^2 # 14.97802
sigma.squ.U.and.e <- 4.101988^2 # = 16.82631 (squared res std dev)
nugget <- 0.5739228
sigma.squ.e <- nugget * sigma.squ.U.and.e # 9.657003
sigma.squ.U <- sigma.squ.U.and.e - sigma.squ.e # = 7.169307
```

Application to Individual Trajectories

Subject 30119

```
idx.30119 <- which(unique(cd4_df$ID) == 30119)
var.cov.30119 <- getVarCov(me.exp4.fit, individual=c(idx.30119), type="marginal")
cor.30119 <- cov2cor(var.cov.30119[[1]])
kable(t(diag(var.cov.30119[[1]])), caption="Estimated Variances for Subject 30119", escape = F, digits =
  kable_styling(latex_options = c("hold_position"))
```

Table 5: Estimated Variances for Subject 30119

1	2	3	4	5	6	7	8	9	10	11	12
35.66	33.55	32.24	31.67	31.65	32.09	33.21	36.17	38.57	42.71	47.2	52.68

```
kable(t(cor.30119[1,]), caption="Estimated Correlations for Subject 30119", escape = F, digits = 2) %>%
  kable_styling(latex_options = c("hold_position"))
```

Table 6: Estimated Correlations for Subject 30119

1	2	3	4	5	6	7	8	9	10	11	12
1	0.61	0.53	0.48	0.45	0.41	0.36	0.29	0.26	0.21	0.17	0.13

Full covariance matrix:

```
var.cov.30119
```

```
## ID 30119
## Marginal variance covariance matrix
##          1       2      3      4      5      6      7      8       9
## 1  35.6590 21.0730 18.136 16.206 14.965 13.704 12.401 10.549  9.5093
## 2  21.0730 33.5450 19.488 17.030 15.672 14.463 13.356 11.936 11.1830
## 3  18.1360 19.4880 32.244 18.773 16.926 15.512 14.437 13.326 12.8190
## 4  16.2060 17.0300 18.773 31.671 19.069 17.030 15.718 14.713 14.3880
## 5  14.9650 15.6720 16.926 19.069 31.646 18.915 17.083 15.939 15.7010
## 6  13.7040 14.4630 15.512 17.030 18.915 32.089 19.353 17.609 17.3620
## 7  12.4010 13.3560 14.437 15.718 17.083 19.353 33.209 20.099 19.5790
## 8  10.5490 11.9360 13.326 14.713 15.939 17.609 20.099 36.165 24.4130
## 9   9.5093 11.1830 12.819 14.388 15.701 17.362 19.579 24.413 38.5680
## 10  8.0963 10.1810 12.194 14.077 15.600 17.426 19.649 23.684 26.7810
## 11  6.8452  9.3063 11.671 13.867 15.620 17.682 20.096 24.087 26.7690
## 12  5.5518  8.4071 11.146 13.680 15.693 18.039 20.739 24.996 27.6370
##         10      11      12
## 1   8.0963  6.8452  5.5518
## 2  10.1810  9.3063  8.4071
## 3  12.1940 11.6710 11.1460
## 4  14.0770 13.8670 13.6800
## 5  15.6000 15.6200 15.6930
## 6  17.4260 17.6820 18.0390
## 7  19.6490 20.0960 20.7390
## 8  23.6840 24.0870 24.9960
## 9  26.7810 26.7690 27.6370
## 10 42.7060 31.4410 31.7400
## 11 31.4410 47.2050 36.3210
## 12 31.7400 36.3210 52.6780
##   Standard Deviations: 5.9715 5.7918 5.6784 5.6277 5.6255 5.6648 5.7627 6.0137 6.2103 6.535 6.8706 7
```

Full correlation matrix:

```
cor.30119
```

```
##           1         2         3         4         5         6         7
## 1  1.0000000 0.6092848 0.5348636 0.4822465 0.4454697 0.4051114 0.3603646
## 2  0.6092848 1.0000000 0.5925652 0.5224840 0.4810007 0.4408240 0.4001655
## 3  0.5348636 0.5925652 1.0000000 0.5874764 0.5298669 0.4822404 0.4411858
## 4  0.4822465 0.5224840 0.5874764 1.0000000 0.6023299 0.5342000 0.4846511
## 5  0.4454697 0.4810007 0.5298669 0.6023299 1.0000000 0.5935467 0.5269723
## 6  0.4051114 0.4408240 0.4822404 0.5342000 0.5935467 1.0000000 0.5928468
## 7  0.3603646 0.4001655 0.4411858 0.4846511 0.5269723 0.5928468 1.0000000
## 8  0.2937402 0.3427017 0.3902508 0.4347269 0.4711373 0.5169137 0.5799661
## 9  0.2564189 0.3108982 0.3635204 0.4116680 0.4494231 0.4935164 0.5470764
```

```
## 10 0.2074715 0.2689963 0.3285993 0.3827660 0.4243480 0.4707403 0.5217681
## 11 0.1668432 0.2338668 0.2991542 0.3586281 0.4041466 0.4543088 0.5075714
## 12 0.1280965 0.1999953 0.2704401 0.3349100 0.3843576 0.4387527 0.4958531
##            8         9        10        11        12
## 1  0.2937402 0.2564189 0.2074715 0.1668432 0.1280965
## 2  0.3427017 0.3108982 0.2689963 0.2338668 0.1999953
## 3  0.3902508 0.3635204 0.3285993 0.2991542 0.2704401
## 4  0.4347269 0.4116680 0.3827660 0.3586281 0.3349100
## 5  0.4711373 0.4494231 0.4243480 0.4041466 0.3843576
## 6  0.5169137 0.4935164 0.4707403 0.4543088 0.4387527
## 7  0.5799661 0.5470764 0.5217681 0.5075714 0.4958531
## 8  1.0000000 0.6536855 0.6026589 0.5829710 0.5726844
## 9  0.6536855 1.0000000 0.6598877 0.6273784 0.6131577
## 10 0.6026589 0.6598877 1.0000000 0.7002634 0.6691875
## 11 0.5829710 0.6273784 0.7002634 1.0000000 0.7283677
## 12 0.5726844 0.6131577 0.6691875 0.7283677 1.0000000
```

BLUPs for Five

```
# get subject ids with 7 or more observations
sub.ids.7.plus <- cd4_df %>% group_by(ID) %>% summarise(num_obs = max(obsnum)) %>%
  filter(num_obs >= 7) %>% arrange(ID, num_obs) %>% select(ID)
cd4_7_df <- cd4_df %>% filter(ID %in% sub.ids.7.plus$ID)
target_sids <- c(30119, 40286, 20777, 10213, 10453)
target_df <- cd4_7_df %>% filter(ID %in% target_sids)
# build covariate matrix for selected subjects
X_df <- target_df %>%
  mutate(Intercept = rep(1, n())) %>%
  select(Intercept, Time, smoker, Time0, Time1, Time3, Cesd, Sex) %>%
  mutate(
    Time.Smoker = Time * smoker
  )

# define a method that returns a data frame containing the empirical BLUP for a subject
calculate_blup <- function(subjectId, df, mdl.fit){
  # extract subject design matrix
  X.sub <- as.matrix(df %>% ungroup() %>% filter(ID == subjectId) %>% select(-ID))
  # calculate fixed effects
  fe.sub <- X.sub %*% mdl.fit$coefficients$fixed
  # calculate random effects
  preds.sub <- mdl.fit$coefficients$random$ID[as.character(subjectId),]
  X.re <- X.sub[,c(1,2)]
  # do it componentwise and return data frame
  re.sub <- X.re %*% preds.sub
  blup.sub <- fe.sub + re.sub
  blup.df.sub <- tibble(
    ID = rep(subjectId, length(blup.sub)),
    Time = X.re[,2],
    Blup = blup.sub[,1]
  )
  return(blup.df.sub)
}

# apply to selected subjects
blup_df.30119 <- calculate_blup(30119, X_df, me.exp4.fit)
```

```
blup_df.40286 <- calculate_blup(40286, X_df, me.exp4.fit)
blup_df.20777 <- calculate_blup(20777, X_df, me.exp4.fit)
blup_df.10213 <- calculate_blup(10213, X_df, me.exp4.fit)
blup_df.10453 <- calculate_blup(10453, X_df, me.exp4.fit)
```

Plot BLUPs

```
ggplot(cd4_df)+
  ggtitle('Population Average and BLUPs for Selected Subjects')+ xlab('Time (Years)')+ ylab('CD4 Sqrt')+
  scale_x_continuous(breaks=x)+
  geom_vline(xintercept = 0, colour='black', lty=5)+
  annotate("text", label = "Seroconversion time", x = 0.1, y = 50, size = 4, colour = "black", hjust=0)+
  geom_smooth(aes(x=yr, y=CD4sqrt), colour='red', se = F, method='loess', span=0.75)+
  # 30119
  geom_point(data=target_df %>% filter(ID == 30119), aes(x=yr, y=CD4sqrt), color='blue', shape=3)+
  geom_line(data=blup_df.30119, aes(x=Time, y=Blup), color='blue', lwd=1)+
  annotate("text", label = "30119", x = blup_df.30119[1,]$Time, y = blup_df.30119[1,]$Blup, size = 5, colour = "blue", hjust=0,

  # 20777
  geom_point(data=target_df %>% filter(ID == 20777), aes(x=yr, y=CD4sqrt), color='darkorange', shape=3)+
  geom_line(data=blup_df.20777, aes(x=Time, y=Blup), color='darkorange', lwd=1)+
  annotate("text", label = "20777", x = blup_df.20777[1,]$Time, y = blup_df.20777[1,]$Blup, size = 5, colour = "darkorange", hju

  # 40286
  geom_point(data=target_df %>% filter(ID == 40286), aes(x=yr, y=CD4sqrt), color='darkgreen', shape=3)+
  geom_line(data=blup_df.40286, aes(x=Time, y=Blup), color='darkgreen', lwd=1)+
  annotate("text", label = "40286", x = blup_df.40286[1,]$Time, y = blup_df.40286[1,]$Blup, size = 5, colour = "darkgreen", hjus

  # 10213
  geom_point(data=target_df %>% filter(ID == 10213), aes(x=yr, y=CD4sqrt), color='steelblue', shape=3)+
  geom_line(data=blup_df.10213, aes(x=Time, y=Blup), color='steelblue', lwd=1)+
  annotate("text", label = "10213", x = blup_df.10213[1,]$Time, y = blup_df.10213[1,]$Blup, size = 5, colour = "steelblue", hjus

  # 10453
  geom_point(data=target_df %>% filter(ID == 10453), aes(x=yr, y=CD4sqrt), color='black', shape=3)+
  geom_line(data=blup_df.10453, aes(x=Time, y=Blup), color='black', lwd=1)+
  annotate("text", label = "10453", x = blup_df.10453[1,]$Time, y = blup_df.10453[1,]$Blup, size = 5, colour = "black", hjust=0,
```