Biostat/Stat 571 Exercise #6

Answer Key

February, 2010

Question 1

Part a

In the mean model

$$E[Y_{ij}] = \beta_0 + \beta_1 \cdot \texttt{TX}_i + \beta_2 \cdot \texttt{post}_{ij} + \gamma \cdot \texttt{TX}_i \cdot \texttt{post}_{ij},$$

the parameter γ is interpreted as the additional treatment effect, measured at follow-up, over the initial treatment effect at baseline.

Part b

The model suggests that we have the following set-up for the means of the 2 treatment groups at baseline and follow-up:

	baseline	follow-up
TX = 0	β_0	$\beta_0 + \beta_2$
TX = 1	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \gamma$.

For each individual, we observe 2 observations $\mathbf{Y}_i = (Y_{0i}, Y_{1i})$. Thus, we can use a repeated measures model for $\mathbf{Y}_i$ that assumes multivariate normality as follows:

$$\mathbf{Y}_i = \begin{pmatrix} Y_{0i} \\ Y_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{0i} \\ \mu_{1i} \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_0\sigma_1\rho \\ \sigma_0\sigma_1\rho & \sigma_1^2 \end{pmatrix} \right),$$

where $\mu_0 = \beta_0 + \beta_1 \text{TX}_i$ and $\mu_1 = (\beta_0 + \beta_2) + (\beta_1 + \gamma)\text{TX}_i$. Alternatively, since we are assuming joint normality, we can see that each of the cells in the above table is marginally normally distributed:

$$\begin{aligned}
Y_{i0,TX=0} &\sim N(\mu_{00}, \sigma_0^2) \equiv N(\beta_0, \sigma_0^2) \\
Y_{i0,TX=1} &\sim N(\mu_{01}, \sigma_0^2) \equiv N(\beta_0 + \beta_1, \sigma_0^2) \\
Y_{i1,TX=0} &\sim N(\mu_{10}, \sigma_1^2) \equiv N(\beta_0 + \beta_2, \sigma_1^2) \\
Y_{i1,TX=1} &\sim N(\mu_{11}, \sigma_1^2) \equiv N(\beta_0 + \beta_1 + \beta_2 + \gamma, \sigma_1^2).
\end{aligned}$$

The maximum likelihood estimators for $\mu_{00}, \mu_{01}, \mu_{10}$, and μ_{11} are the appropriate sample means. To get the MLE for γ, notice that:

$$\begin{aligned}\gamma &= [(\beta_0 + \beta_1 + \beta_2 + \gamma) - (\beta_0 + \beta_1)] - [(\beta_0 + \beta_2) - \beta_0] \\ &= (\mu_{11} - \mu_{01}) - (\mu_{10} - \mu_{00}).\end{aligned}$$

Consequently, the MLE for γ is given by:

$$\begin{aligned}\hat{\gamma}^{(1)} &= (\hat{\mu}_{11} - \hat{\mu}_{01}) - (\hat{\mu}_{10} - \hat{\mu}_{00}) \\ &= (\overline{Y_{i1,TX=1}} - \overline{Y_{i0,TX=1}}) - (\overline{Y_{i1,TX=0}} - \overline{Y_{i0,TX=0}}) \\ &= \overline{Y_{i1,TX=1} - Y_{i0,TX=1}} - \overline{Y_{i1,TX=0} - Y_{i0,TX=0}}.\end{aligned}$$

Part c

Under the assumption that $\sigma_0^2 = \sigma_1^2 = \sigma^2$, the variance of $\hat{\gamma}^{(1)}$ is given by:

$$\begin{aligned}\mathrm{V}(\hat{\gamma}^{(1)}) &= \mathrm{V}\left(\overline{Y_{i1,TX=1} - Y_{i0,TX=1}} - \overline{Y_{i1,TX=0} - Y_{i0,TX=0}}\right) \\ &= \mathrm{V}\left(\overline{Y_{i1,TX=1} - Y_{i0,TX=1}}\right) + \mathrm{V}\left(\overline{Y_{i1,TX=0} - Y_{i0,TX=0}}\right) \\ &= \frac{1}{m}\mathrm{V}(Y_{i1,TX=1} - Y_{i0,TX=1}) + \frac{1}{m}\mathrm{V}(Y_{i1,TX=0} - Y_{i0,TX=0}) \\ &= \frac{2}{m}(\sigma^2 + \sigma^2 - 2\rho\sigma^2) \\ &= \frac{4\sigma^2(1-\rho)}{m},\end{aligned}$$

since study participants are independent of each other.

Part d

Consider the bivariate likelihood for $\mathbf{Y}_i$ that is given in part (b). We can factor the likelihood into two components: the marginal distribution of Y_{0i} and the conditional distribution $Y_{1i|0i}$. This is given by:

$$f(\mathbf{Y}_i) = f(Y_{0i})f(Y_{1i|0i}),$$

where

$$\begin{aligned}Y_{0i} &\sim N(\beta_0, \sigma_0^2) \\ Y_{1i|0i} &\sim N\left(\beta_0 + \beta_2 + \gamma \mathrm{TX}_i + \rho\frac{\sigma_1}{\sigma_0}(Y_{0i} - \beta_0), \sigma_1^2(1-\rho^2)\right).\end{aligned}$$

To estimate γ, we see that the contribution to the full likelihood by the marginal distribution of Y_{0i} will factor out, and so can be ignored. Assuming ρ is known, we can reparametrise the above conditional distribution of $Y_{1i|0i}$ as follows:

$$Y_{1i|0i} \sim N(\xi + \gamma \mathrm{TX}_i + \phi Y_{0i}, \tau^2).$$

To get the maximum likelihood estimate for γ, notice that the above estimate is essentially a simple linear regression problem. So, the MLE will be the same as the OLS estimate. We can set up two equations, based on

the following, and solve them simulataneously for ξ and γ:

$$\text{TX} = 0 \quad : \quad \sum_{i=1}^{m} (Y_{i1,TX=0} - (\xi + \phi Y_{i0,TX=0}))^2 \tag{1}$$

$$\text{TX} = 1 \quad : \quad \sum_{i=1}^{m} (Y_{i1,TX=1} - (\xi + \gamma + \phi Y_{i0,TX=1}))^2. \tag{2}$$

For (1), we can see that the OLS estimate for ξ will be

$$\hat{\xi} = \overline{Y_{i1,TX=0}} - \phi \overline{Y_{i0,TX=0}}.$$

Given this, from (2), we can see that the estimator for γ will be:

$$\begin{aligned}
\hat{\gamma}^{(2)} &= \overline{Y_{i1,TX=1}} - \phi \overline{Y_{i0,TX=1}} - \hat{\xi} \\
&= \overline{Y_{i1,TX=1}} - \phi \overline{Y_{i0,TX=1}} - \overline{Y_{i1,TX=0}} - \phi \overline{Y_{i0,TX=0}} \\
&= \overline{Y_{i1,TX=1} - \phi Y_{i0,TX=1}} - \overline{Y_{i1,TX=0} - \phi Y_{i0,TX=0}}.
\end{aligned}$$

Part e

Under the assumption that $\sigma_0^2 = \sigma_1^2 = \sigma^2$, we have that $\phi = \rho$. Hence, the variance of $\hat{\gamma}^{(2)}$ is given by:

$$\begin{aligned}
\text{V}(\hat{\gamma}^{(2)}) &= \text{V}(\overline{Y_{i1,TX=1} - \phi Y_{i0,TX=1}} - \overline{Y_{i1,TX=0} - \phi Y_{i0,TX=0}}) \\
&= \text{V}(\overline{Y_{i1,TX=1} - \rho Y_{i0,TX=1}}) + \text{V}(\overline{Y_{i1,TX=0} - \rho Y_{i0,TX=0}}) \\
&= \frac{1}{m}\text{V}(Y_{i1,TX=1} - \rho Y_{i0,TX=1}) + \frac{1}{m}\text{V}(Y_{i1,TX=0} - \rho Y_{i0,TX=0}) \\
&= \frac{2}{m}(\sigma^2 + \rho^2\sigma^2 - 2\rho^2\sigma^2) \\
&= \frac{2\sigma^2(1-\rho^2)}{m},
\end{aligned}$$

since study participants are independent of each other.

Part f

If we assume that the model in part (d) holds, then we see that $\hat{\gamma}(\alpha)$ is unbiased for γ for any value of α:

$$\begin{aligned}
E(\hat{\gamma}(\alpha)) &= E(Y_{i1,TX=1}) - \alpha E(Y_{i0,TX=1}) - E(Y_{i1,TX=0}) + \alpha E(Y_{i0,TX=0}) \\
&= (\beta_0 + \beta_2 + \gamma) - \alpha(\beta_0 + \beta_2) - \beta_0 + \alpha\beta_0 \\
&= \gamma.
\end{aligned}$$

Using the same arguments as parts (c) and (e), the variance of $\hat{\gamma}(\alpha)$ (assuming $\sigma_0^2 = \sigma_1^2 = \sigma^2$) is given by:

$$\text{V}(\hat{\gamma}(\alpha)) = \frac{2\sigma^2(\alpha^2 - 2\alpha\rho + 1)}{m}.$$

The estimator $\hat{\gamma}(0) = \overline{Y_{i1,TX=1}} - \overline{Y_{i1,TX=0}}$ arises from a model that only uses that follow-up data in the estimation of γ. The variance of $\hat{\gamma}(0)$ is

$$\mathrm{V}(\hat{\gamma}(0)) = \frac{2\sigma^2}{m}.$$

We see that $\hat{\gamma}(1)$ will be more precise than $\hat{\gamma}(0)$ when

$$\frac{4\sigma^2(1-\rho)}{m} < \frac{2\sigma^2}{m}.$$

That is, when $\rho > \frac{1}{2}$. So, the estimate from the model that includes the baseline outcome information is more precise when the within person correlation is greater than $\frac{1}{2}$. To find the optimal α, we want to minimize $V(\hat{\gamma}(\alpha))$ with respect to α. We see that

$$\frac{\partial}{\partial\alpha}\alpha^2 - 2\alpha\rho + 1 = 2\alpha - 2\rho =^{set} 0$$

implies that we have a minimum at $\alpha = \rho$.

Part g

From parts (a)-(f), it seems evident that when we have repeated measurements on individuals, then an optimal analysis will incorporate information regarding the within-subject variability in the outcome measure. A flexible (semi-parametric) approach that we could adopt would be to use the General Linear Model for Correlated Data (GLMCD) using a weighting scheme that is the inverse of the variance-covariance matrix for the vector of subject-specific observations.

Question 2

In the previous exercise we considered data from the Multicenter Aids Cohort Study (MACS). Scientific interest is in whether the rate of decline in CD4 count is associated with baseline viral load measurement. The available data, after removing observations with missing baseline viral load or CD4 counts consists of 1457 observations on 226 subjects. Each subject has at least 3 measurements recorded.

Part a

The data consist of repeated measures on 226 subjects, and consequently, it is desirable to ensure that we account for within-subject variability in any regression analysis. In the last exercise, we looked at the correlation structure of the repeated measures and on the basis of an empirical correlation matrix and a variogram, it was concluded that either an exchangeable or autoregressive correlation structure may be sufficient. Here, we will adopt the autoregressive correlation structure.

Also in Exercise 5, we noted that the distribution of baseline viral load is heavily skewed. In particular, there are a few very high baseline viral loads which will likely be highly influential in the estimation of the regression coefficients. Consequently, baseline viral load was log transformed. In addition, in order to aid interpretation of certain coefficients, the log transformed baseline viral load was centered at its median value. The median log viral load is 10.10940, which corresponds to a viral load of approximately 30,000 copies per milliliter.

We fit the following model with an autoregressive correlation structure, random intercepts and measurement error:

$$E[Y_{ij}|\text{month}_{ij}, X_i] = \beta_0 + \beta_1 \cdot \text{month}_{ij} + \beta_2 \cdot X_i + \beta_3 \cdot \text{month}_{ij} \cdot X_i,$$

where X_i represents the i^{th} subjects' centered log baseline viral load. From this model, we have the following output:

```
Linear mixed-effects model fit by maximum likelihood
 Data: macs
       AIC      BIC    logLik
  22803.07 22846.51 -11393.53

Random effects:
 Formula: ~1 | id
        (Intercept) Residual
StdDev:    142.0596 251.1175

Correlation Structure: Exponential spatial correlation
 Formula: ~month | id
 Parameter estimate(s):
     range     nugget
41.0730002  0.3074062
Fixed effects: cd4 ~ month * log.vload0
                    Value Std.Error   DF   t-value p-value
(Intercept)      757.4284 18.163643 1457  41.70025  0.0000
month             -7.0923  0.453009 1457 -15.65588  0.0000
log.vload0       -33.9797  9.935336  224  -3.42008  0.0007
month:log.vload0  -0.4697  0.248757 1457  -1.88815  0.0592
```

From this model, there are four parameters which we can interpret as follows:

- β_0: Expected CD4 count at baseline (ie. seroconversion) for an individual from a population where baseline viral load equals 10.30822 (ie. the median log viral load among the 226 subjects).
- β_1: Change in the expected CD4 count associated with an increase in time of one month, for an individual from a population where baseline viral load is 10.10940. We could also interpret this as the rate of change in the expected CD4 count, over the period of a month, for an individual from a population where baseline log viral load is 10.10940. In this case, we find that subjects where baseline viral load is equal to the median for the sample have CD4 counts that deteriorate over time, at a rate of roughly 7 CD4 cells per mm^3 per month.
- β_2: Difference in the expected CD4 count at baseline comparing two populations whose baseline log viral load differs by one unit (on the natural log scale). An additive unit increase on the log scale is equivalent to an e-fold (2.7-fold) multiplicative increase on the original scale.
- β_3: The difference in the rate of change of CD4 count over the period of one month associated with an e-fold increase in baseline viral load. Equivalently, if we compare two populations whose baseline viral load differs by a multiplicative factor of e, then we expect the population with the higher baseline viral load to have a rate of change, over the period of one month, that differs by β_3. In this case, the rate of change

will decrease by approximately 0.5 units suggesting that subjects with higher baseline viral load have worse progression (in terms of CD4 counts) than subjects with lower baseline viral loads.

From the output, we see that the p-value associated with the interaction term is 0.0592. Comparing this to the usual critical value of 0.05, we find that there is insufficient evidence to indicate that there is an association between baseline (log) viral load and the rate of decline in CD4.

Part b

The above model makes the strong assumption of linearity about the impact of log baseline viral load. We can attempt to allow the dependency of CD4 count on log baseline viral load to be more flexible by including a series of factors which represent quartiles of the log baseline viral load distribution. Towards this end, the range of (centered) log baseline viral load have been split into 4 equal ranges. Table 1 provides the ranges as well as the number of subjects (out of 226) that fall into each range. Since the log-transformation is a monotone one, we can also translate (approximately) the ranges on the log scale onto the original scale. Again, using a linear

Table 1: Log viral load categories, based on splitting the range of the centered log(viral load) for 226 subjects.

Category	1	2	3	4
Range (log)	(-4.41, -2.37]	(-2.37, -0.34]	(-0.34, 1.70]	(1.70, 3.73]
Range (original)	(300, 2300]	(2300, 17550]	(17550, 134000]	(134000, 1027000]
Number of subjects	32	66	92	36

mixed model with an autoregressive correlation structure, random intercepts and measurement error, we fit the following mean model:

$$E(Y_{ij}|\text{month}_{ij}, X_i) = \beta_0 + \beta_1 \cdot \texttt{month}_i + \sum_{k=2}^{4} \beta_{2,k} \cdot X_i(k) + \sum_{k=2}^{4} \beta_{3,k} \cdot \texttt{month}_i \cdot X_i(k),$$

where $X_i(k)$ is a binary indicator that the centered log baseline viral load for the i^{th} subject is in category k, where $k = 2, 3, 4$. Consequently, category 1 serves as the reference (comparison) group. The resulting output is provided below:

```
Linear mixed-effects model fit by maximum likelihood
 Data: macs
       AIC      BIC    logLik
  22809.09 22874.24 -11392.54

Random effects:
 Formula: ~1 | id
        (Intercept) Residual
StdDev:    139.1644 252.1283

Correlation Structure: Exponential spatial correlation
 Formula: ~month | id
```

```
 Parameter estimate(s):
     range     nugget
42.0998788  0.3063569
Fixed effects: cd4 ~ month * log.vload0.cat
                          Value Std.Error   DF   t-value p-value
(Intercept)            916.4385  48.31625 1455 18.967501  0.0000
month                   -5.0257   1.19554 1455 -4.203694  0.0000
log.vload0.cat1       -166.7018  58.73222  222 -2.838336  0.0050
log.vload0.cat2       -189.0272  56.06619  222 -3.371501  0.0009
log.vload0.cat3       -185.4405  66.23679  222 -2.799661  0.0056
month:log.vload0.cat1   -1.8773   1.46049 1455 -1.285404  0.1989
month:log.vload0.cat2   -2.3785   1.38790 1455 -1.713771  0.0868
month:log.vload0.cat3   -3.1067   1.65603 1455 -1.875984  0.0609
```

In this model, there are 8 parameters, although the interpretations of $\beta_{2,k}$ and $\beta_{3,k}$ can be generalized for $k = 2, 3, 4$. As indicated above, the reference group for this model is now log viral load category 1.

- β_0: Expected CD4 count at baseline (ie. seroconversion) for an individual from a population where baseline log viral load is given by category 1 (ie. between 300 and 2300 virus copies per ml).
- β_1: Change in the expected CD4 count associated with an increase in time of one month for an individual from a population where baseline log viral load is given by category 1. We could also interpret this as the rate of change in the expected CD4 count over the period of a month for an individual from a population where baseline log viral load is given by category 1. In this case, we find that subjects where baseline viral load is in category 1 have CD4 counts that deteriorate over time at a rate of roughly 5.0 CD4 cells per mm^3 per month.
- $\beta_{2,k}$: Difference in the expected CD4 count at baseline comparing two populations whose baseline log viral load are given by category k and category 1, for $k = 2, 3, 4$.
- $\beta_{3,k}$: The difference in the rate of change of CD4 count over the period of one month, comparing two populations whose baseline log viral load are given by category k and category 1, for $k = 2, 3, 4$. For example, the rate of change in the expected CD4 count is estimated to be 2.4 units lower for individuals whose baseline log viral load is given by category 3 than for individuals whose baseline log viral load is given by category 1.

The p-value associated with testing the null hypothesis that $H_0 : \beta_{3,2} = \beta_{3,3} = \beta_{3,4}$ is 0.257. Consequently, there is insufficient evidence (at the 0.05 level) to reject the null hypothesis and, therefore, insufficient evidence to suggest that there is an association between baseline (log) viral load and the rate of decline of CD4.

Part c

Finally, we can allow the viral load components (both main effects and interaction terms) to take on richer functional forms, by allowing a more flexible class of models. In particular, we fit the following general mean model:

$$E(Y_{ij}|X_i) = \gamma_0(X_i) + \gamma_1(X_i) \cdot \texttt{month}_{ij}.$$

Here, instead of assuming linearity (part (a)) or categorising (part (b)), we incorporate natural splines into the model via the $\gamma_0(\cdot)$ and $\gamma_1(\cdot)$ functions. The basis for the natural splines are based on 2 knots at -0.731 and 0.757,

which represent the 33 and 67 percentiles of the centered log-transformed baseline viral load distributions. These correspond approximately to 11800 and 52400 copies of the virus per ml. The following is the output from the resulting fit:

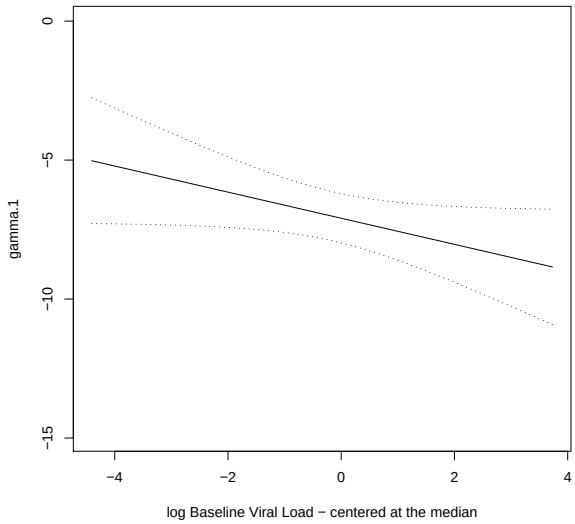
```
Linear mixed-effects model fit by maximum likelihood
 Data: macs
       AIC      BIC    logLik
  22805.08 22870.24 -11390.54

Random effects:
 Formula: ~1 | id
        (Intercept) Residual
StdDev:    136.0649 252.1396

Correlation Structure: Exponential spatial correlation
 Formula: ~month | id
 Parameter estimate(s):
     range     nugget
41.8132786  0.3050983
Fixed effects: cd4 ~ month * ns(log.vload0, knots = c(-0.731, 0.757))
                                                 Value Std.Error   DF   t-value p-value
(Intercept)                                  979.9785  76.90811 1455 12.742200  0.0000
month                                         -3.8415   1.93321 1455 -1.987129  0.0471
ns(log.vload0, knots = c(-0.731, 0.757))1   -235.3953  76.10289  222 -3.093119  0.0022
ns(log.vload0, knots = c(-0.731, 0.757))2   -427.1893 186.72804  222 -2.287762  0.0231
ns(log.vload0, knots = c(-0.731, 0.757))3   -122.0797  99.07148  222 -1.232239  0.2192
month:ns(log.vload0, knots = c(-0.731, 0.757))1   -1.2681   1.89869 1455 -0.667881  0.5043
month:ns(log.vload0, knots = c(-0.731, 0.757))2   -8.2314   4.69153 1455 -1.754521  0.0796
month:ns(log.vload0, knots = c(-0.731, 0.757))3   -4.0719   2.49339 1455 -1.633083  0.1027
```

Figure 1(b) provides a plot of the estimated function $\gamma_1(X_i)$ versus the centered log baseline viral load (X_i) along with approximate 95% confidence intervals. The model given in part (a) specified $\gamma_1(X_i)$ as a linear function of X_i: $\gamma_1(X_i) = \beta_1 + \beta_3 X_i$. Although figure 1(b) suggests that there may be a steeper downward trend for very high baseline viral loads, we see that the pointwise confidence intervals are also very wide in this range. This reflects the lack of subjects in the dataset with very high viral loads. Given figure 1(b), it does not seem unreasonable that the model given in part (a) is appropriate. Figure 1(a) provides the corresponding plot of $\gamma_1(X_i)$ for the linear model in part (a).

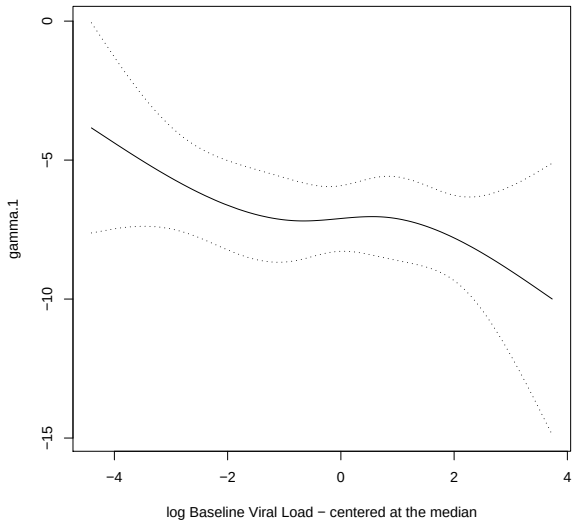
Part d

Now, we incorporate cubic splines into the model. The basis for the cubic splines are based on 2 knots at -0.731 and 0.757, which represent the 33 and 67 percentiles of the centered log-transformed baseline viral load distributions. These correspond approximately to 11800 and 52400 copies of the virus per ml. The following is the output from the resulting fit:

```
Linear mixed-effects model fit by maximum likelihood
 Data: macs
```



(a) Linear mean model (part (a))



(b) Natural Splines mean model (part (c))

Figure 1: Estimated $\gamma_1(X_i)$ function based on the specified mean model.

```
      AIC      BIC    logLik
  22808.16 22895.03 -11388.08

Random effects:
 Formula: ~1 | id
        (Intercept) Residual
StdDev:    144.3522 246.4502

Correlation Structure: Exponential spatial correlation
 Formula: ~month | id
 Parameter estimate(s):
    range    nugget
39.179053  0.320427
Fixed effects: cd4 ~ month * bs(log.vload0, knots = c(-0.731, 0.757))
                                                  Value Std.Error   DF   t-value p-value
(Intercept)                                    992.8634 102.82696 1453  9.655672  0.0000
month                                           -6.6019   2.54972 1453 -2.589283  0.0097
bs(log.vload0, knots = c(-0.731, 0.757))1      -95.6093 205.58448  220 -0.465061  0.6423
bs(log.vload0, knots = c(-0.731, 0.757))2     -243.3302 128.93242  220 -1.887269  0.0604
bs(log.vload0, knots = c(-0.731, 0.757))3     -256.6787 152.38239  220 -1.684438  0.0935
bs(log.vload0, knots = c(-0.731, 0.757))4     -406.3492 172.98171  220 -2.349087  0.0197
bs(log.vload0, knots = c(-0.731, 0.757))5      -76.0078 216.77563  220 -0.350629  0.7262
month:bs(log.vload0, knots = c(-0.731, 0.757))1    5.6649   5.09815 1453  1.111160  0.2667
```

```
month:bs(log.vload0, knots = c(-0.731, 0.757))2   -2.5081   3.21320 1453 -0.780558  0.4352
month:bs(log.vload0, knots = c(-0.731, 0.757))3   -0.2466   3.76543 1453 -0.065481  0.9478
month:bs(log.vload0, knots = c(-0.731, 0.757))4    1.3511   4.35045 1453  0.310561  0.7562
month:bs(log.vload0, knots = c(-0.731, 0.757))5   -8.3616   5.43518 1453 -1.538416  0.1242
```

Figure 2 provides a plot of the estimated function $\gamma_1(X_i)$ versus the centered log baseline viral load (X_i) along with approximate 95% confidence intervals. We notice that this curve is wigglier than the curve presented in part (c). As for part (c), this figure suggests that there may be a steeper downward trend for very high baseline viral loads, but we see that the pointwise confidence intervals are also very wide in this range. Again, this reflects the lack of subjects in the dataset with very high viral loads. From this figure, the model given in part (a) no longer seems appropriate.

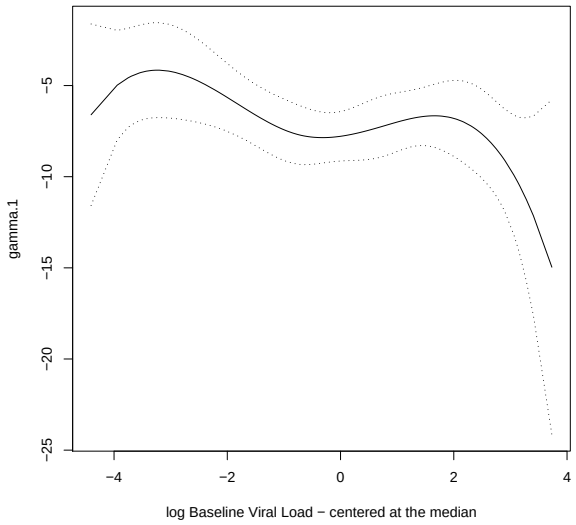


Figure 2: Estimated $\gamma_1(X_i)$ function based on the specified mean model.

Part e

There are a multitude of other models that can be fit. Here, we fit two additional models: one with only random intercepts, and one with random intercepts and slopes. Below is the output from fitting those two models:

```
Linear mixed-effects model fit by maximum likelihood
 Data: macs
       AIC      BIC    logLik
  22942.41 22974.98 -11465.20

Random effects:
```

```
 Formula: ~1 | id
        (Intercept) Residual
StdDev:    213.3492 186.4047

Fixed effects: cd4 ~ month * log.vload0
                    Value Std.Error   DF   t-value p-value
(Intercept)      746.3470 16.443838 1457  45.38764  0.0000
month             -6.9269  0.316321 1457 -21.89836  0.0000
log.vload0       -35.7910  8.988696  224  -3.98178  0.0001
month:log.vload0  -0.3798  0.173747 1457  -2.18616  0.0290


Linear mixed-effects model fit by maximum likelihood
 Data: macs
       AIC      BIC    logLik
  22816.65 22860.08 -11400.32

Random effects:
 Formula: ~1 + month | id
 Structure: General positive-definite
            StdDev     Corr
(Intercept) 239.303086 (Intr)
month         5.515425 -0.445
Residual    165.600953

Fixed effects: cd4 ~ month * log.vload0
                    Value Std.Error   DF   t-value p-value
(Intercept)      746.2187 17.604520 1457  42.38790  0.0000
month             -6.9497  0.474419 1457 -14.64877  0.0000
log.vload0       -34.9735  9.623243  224  -3.63428  0.0003
month:log.vload0  -0.4349  0.260099 1457  -1.67224  0.0947
```

In general, we notice that our parameter estimates are not all that sensitive to the specification of the variance. We notice that the point estimates for the interaction term (the term of scientific interest) are slightly smaller in the models fit above compared to the model fit in part (a). We also notice that our inference would change depending on which model we choose to fit (inference from the model with random intercepts alone indicates a significant result at the 0.05 level). Thus, we can clearly see that our assumptions regarding the correlation structure does have an influence on inference.

R Code

```
## Question 2

library(nlme)
library(splines)

macs<-read.table("MACS-cd4-vload0.data", header=FALSE)
```

```
macs<-macs[,c(1,2,4,6)]
names(macs)<-c("id", "month", "cd4", "vload0")

# Remove those subjects for whom the baseline viral load measurement and/or cd4 count is missing
# Remove subjects with less than 3 cd4 measurements; 13 subjects
n.obs <- unlist( lapply( split( macs$id, macs$id), length ) ) # number of observations per person
drop <- n.obs[ n.obs < 3 ]
drop.id <- as.numeric (names(drop) )
macs$id[is.element(macs$id, drop.id)] <- NA
macs <- na.omit( macs )
macs$log.vload0 <- log( macs$vload0 ) - median( log( macs$vload0 ) )

# Part a
mod1<-lme(cd4~month*log.vload0, method="ML", random=reStruct(~1|id, pdClass="pdSymm", REML=F),
correlation = corExp(form=~month|id, nugget=TRUE), data=macs)
summary(mod1)

# Part b
cutoffs<-min(macs$log.vload0) + c(1:3 / 4) * (max(macs$log.vload0)-min(macs$log.vload0))
macs$log.vload0.cat<-rep(0, 1685)
for(i in 1:3) {
macs$log.vload0.cat[macs$log.vload0 > as.numeric(cutoffs[i])] <- i
}
macs$log.vload0.cat <- as.factor( macs$log.vload0.cat )
apply( (table( macs$id, macs$log.vload0.cat ) !=0 ), 2, sum ) # number of subjects per log(VL) quartile

mod2<-lme(cd4~month*log.vload0.cat, method="ML", random=reStruct(~1|id, pdClass="pdSymm", REML=F),
correlation = corExp(form=~month|id, nugget=TRUE), data=macs)
summary(mod2)
mod2b<-lme(cd4~month+log.vload0.cat, method="ML", random=reStruct(~1|id, pdClass="pdSymm", REML=F),
correlation = corExp(form=~month|id, nugget=TRUE), data=macs)
summary(mod2b)
anova(mod2, mod2b)

# Part c
knots.log.vload0<-as.vector(quantile(macs$log.vload0, prob=c(1/3, 2/3)))
mod3<-lme(cd4~month*ns(log.vload0, knots=c(-0.731, 0.757)), method="ML", random=
reStruct(~1|id, pdClass="pdSymm", REML=F), correlation = corExp(form=~month|id, nugget=TRUE),
data=macs)
summary(mod3)

gamma.1.coef.a<-fixef(mod1)[c(2,4)] # linear gamma_1
varcov.a<-mod1$varFix[c(2,4),c(2,4)]
design.mat.a<-cbind(1, macs$log.vload0)
gamma.1.a<-design.mat.a%*%gamma.1.coef.a
gamma.1.var.a<-design.mat.a%*%varcov.a%*%t(design.mat.a)
gamma.1.se.a<-sqrt(diag(gamma.1.var.a))
gamma.1.upper.a<-gamma.1.a+(qnorm(0.975)*gamma.1.se.a)
```

```
gamma.1.lower.a<-gamma.1.a-(qnorm(0.975)*gamma.1.se.a)

gamma.1.coef.c<-fixef(mod3)[c(2,6,7,8)] # natural spline gamma_1
varcov.c<-mod3$varFix[c(2,6,7,8),c(2,6,7,8)]
design.mat.c<-cbind(1, ns(macs$log.vload0, knots=c(-0.731, 0.757)))
gamma.1.c<-design.mat.c%*%gamma.1.coef.c
gamma.1.var.c<-design.mat.c%*%varcov.c%*%t(design.mat.c)
gamma.1.se.c<-sqrt(diag(gamma.1.var.c))
gamma.1.upper.c<-gamma.1.c+(qnorm(0.975)*gamma.1.se.c)
gamma.1.lower.c<-gamma.1.c-(qnorm(0.975)*gamma.1.se.c)

ooo<-order(macs$log.vload0)
plot(macs$log.vload0[ooo], gamma.1.a[ooo], xlab="log Baseline Viral Load - centered at the
median", ylab="gamma.1", ylim=range(c(gamma.1.lower.c, gamma.1.upper.c, gamma.1.lower.a,
gamma.1.upper.a) ), type='l')
lines(macs$log.vload0[ooo], gamma.1.upper.a[ooo], lty=3)
lines(macs$log.vload0[ooo], gamma.1.lower.a[ooo], lty=3)

plot(macs$log.vload0[ooo], gamma.1.c[ooo], xlab="log Baseline Viral Load - centered at the
median", ylab="gamma.1", ylim=range(c(gamma.1.lower.c, gamma.1.upper.c) ), type='l')
lines(macs$log.vload0[ooo], gamma.1.upper.c[ooo], lty=3)
lines(macs$log.vload0[ooo], gamma.1.lower.c[ooo], lty=3)

# Part d
macs$log.vload0.sq<-macs$log.vload0^2
macs$log.vload0.cu<-macs$log.vload0^3
macs$log.vload0.k1.cu<-(macs$log.vload0-knots.log.vload0[1])^3
macs$log.vload0.k2.cu<-(macs$log.vload0-knots.log.vload0[2])^3

mod4<-lme(cd4~month*bs(log.vload0, knots=c(-0.731, 0.757)), method="ML", random=
reStruct(~1|id, pdClass="pdSymm", REML=F), correlation = corExp(form=~month|id, nugget=TRUE),
data=macs)
summary(mod4)

gamma.1.coef.d<-fixef(mod4)[c(2,8,9,10,11,12)] # cubic spline gamma_1
varcov.d<-mod4$varFix[c(2,8,9,10,11,12),c(2,8,9,10,11,12)]
design.mat.d<-cbind(1, bs(macs$log.vload0, knots=c(-0.731, 0.757)))
gamma.1.d<-design.mat.d%*%gamma.1.coef.d
gamma.1.var.d<-design.mat.d%*%varcov.d%*%t(design.mat.d)
gamma.1.se.d<-sqrt(diag(gamma.1.var.d))
gamma.1.upper.d<-gamma.1.d+(qnorm(0.975)*gamma.1.se.d)
gamma.1.lower.d<-gamma.1.d-(qnorm(0.975)*gamma.1.se.d)

plot(macs$log.vload0[ooo], gamma.1.d[ooo], xlab="log Baseline Viral Load - centered at the
median", ylab="gamma.1", ylim=range(c(gamma.1.lower.d, gamma.1.upper.d) ), type='l')
lines(macs$log.vload0[ooo], gamma.1.upper.d[ooo], lty=3)
lines(macs$log.vload0[ooo], gamma.1.lower.d[ooo], lty=3)
```

```
# Part e
# random intercepts only
mod5a<-lme(cd4~month*log.vload0, method="ML", random=reStruct(~1|id, pdClass="pdSymm", REML=F),
data=macs)
summary(mod5a)

# random intercepts and slopes
mod5b<-lme(cd4~month*log.vload0, method="ML", random=reStruct(~1+month|id, pdClass="pdSymm",
REML=F), data=macs)
summary(mod5b)
```