The data in the file uissurv.dat on the assignment section of the course web page were collected from a randomized clinical trial of residential treatment for drug abuse. One of the main objectives was to compare the effectiveness of two programs, called the long and short residential treatment programs, in reducing relapse among former drug abusers. After giving up drug use and being admitted to the study, the subjects were randomly assigned to either the long or short residential treatment program. The time to event was time from admission into the program until the subject relapsed, (returned to using drugs). A censored time is indicated by a zero value of the status variable.

There is one line of data for each subject. Each line of data has information on the following variables in the following order:

- **id**: Subject identification number
- **age**: Age at enrollment (in years)
- **beck**: Beck depression score (0-54)
- **hercoc**: Drug use during the 3 months prior to enrollment in the study:
  - 1 = heroin and cocaine
  - 2 = heroin only
  - 3 = cocaine only
  - 4 = neither
- **ivhx**: History of intravenous drug use at time of admission to the study
  - 1 = never
  - 2 = previous
  - 3 = recent
- **ndrugtx**: number of prior treatments for drug abuse (0-40)
- **race**: Subjects race (0 = white 1 = non-white)
- **treat**: Assigned treatment (0 = short, 1 = long)
site Treatment site (0 = A, 1 = B)
time time from admission into the program until drug relapse
status 1=return to drug use, 0=censoring time

Code for reading these data into SAS is posted as uissurv.sas. Code for reading these data into a data frame in S-plus or R is posted as uissurv.ssc.

(a) Fit a proportional hazards model using the variables treat, beck, age, hercoc, ivhx, ndrugtx and race as covariates. You should recode the categorical variables hercoc and ivx as sets of binary variables. For previous drug use define

\[
\begin{align*}
X_1 &= \begin{cases} 1 & \text{if used both heroin and cocaine} \\ 0 & \text{otherwise} \end{cases} \\
X_2 &= \begin{cases} 1 & \text{if used only heroin} \\ 0 & \text{otherwise} \end{cases} \\
X_3 &= \begin{cases} 1 & \text{if used only cocaine} \\ 0 & \text{otherwise} \end{cases}
\end{align*}
\]

If you do this by defining a factor in S-Plus you should recode the hercoc variable before making the factor so that “neither” has the smallest value. For intravenous drug use history at admission use

\[
\begin{align*}
V_1 &= \begin{cases} 1 & \text{previous} \\ 0 & \text{otherwise} \end{cases} \\
V_2 &= \begin{cases} 1 & \text{recent} \\ 0 & \text{otherwise} \end{cases}
\end{align*}
\]

(a) Is there a significant difference in the long and short stay treatments with respect to discouraging return to drug use, after adjusting for the effects of the other covariates?
(b) Examine appropriate residuals for the model in part (a) to determine if there are any outliers. Show your plots and state your conclusion.
(c) Examine appropriate residuals for the model in part (a) to determine if there are any highly influential subjects?
(d) Refit the model in part (a) after removing any outliers from the data. Examine the scaled Schoenfeld residuals. State your conclusions.
(e) Explore how to use the continuous variables age, beck, and ndrugtx in the model by examining Martingale residuals. State your conclusions.
(f) Fit a model stratifying on site. Is it necessary to stratify on site for these data? Explain.
(g) Present the model that you feel best describes the effects of long or short treatment and the other covariates on time to return to drug use.
2. If you use the LIFEREG procedure in SAS to fit an accelerated failure time (AFT) model, it will fit a model of the form

\[ \log(T) = \mu - \beta X + \sigma \epsilon \]

where \( T \) is the failure time, \( X \) is a single covariate and \( \epsilon \sim G(0,1) \). \( G(0,1) \) denotes the standard extreme minimum value (or Gumbel) distribution. It has density function \( f(y) = \exp(y - \exp(y)) \).

(a) Derive the survivor function for \( T \). Do you recognize the form?

(b) Show that the survivor function \( S(t|X) \), for an individual with covariate value \( X \), is equal to a baseline survivor function multiplied by an acceleration factor, i.e.,

\[ \exp(X \beta) S_0(t \exp(X \beta)) \]

(c) Derive the hazard function. Does it correspond to a proportional hazards model?

3. Angioplasty is a procedure used to increase blood flow through a vessel. It is performed by inserting a tube into the vessel with a balloon on the leading end. At the point of obstruction, the balloon is inflated and used to remove the obstruction and stretch the vessel wall. In a study performed by the Iowa Heart Center, data were collected on 159 angioplasty procedures performed on blood vessels in the legs of 115 patients. A single angioplasty procedure was performed on 79 of the 115 patients, 30 patients had two procedures, 4 patients had three procedures and 2 patients had 4 procedures. The patients were examined at roughly three-month intervals. The response of interest was the time from when the angioplasty procedure was done until failure. A procedure was determined to fail if blood flow through the vessel was reduced to less than 20 percent of the flow measured immediately after the angioplasty procedure was performed. Other variables of interest to the researchers were age of the patient when the angioplasty procedure was performed (AGE), whether or not the patient had diabetes mellitus (DM) or hypertension (HTN). The data are posted in the file angioplasty.dat on the assignment section of the course web page. There is one line of data for each procedure. Values for the variables appear in the following order:

- **ID**: Patient identification number
- **DM**: Diabetes Mellitus (0=absence, 1=presence)
- **HTN**: Hypertension (0=absence, 1=presence)
- **Age**: Age of the patient when the angioplasty procedure was done
- **Time**: Time until failure or censoring
- **Status**: (0=censored, 1=failure)

Ages range from 27 to 91 years. Use a new variable AGE65=AGE-65 instead of AGE in the model, so the baseline hazard corresponds to a 65 year-old patient with no history of
diabetes, myocardial infarction, or hypertension. Code for entering these data into SAS is posted as angioplasty.sas, and code for entering test data into an S-PLUS data frame is posted as angioplasty.ssc.

(a) Fit a marginal Cox proportional hazards model to the data using AGE65, HTN, DM and DM*AGE65 as the covariates. In the estimation of the parameters, this procedure will use a partial likelihood function that incorrectly assumes independence among all failure times, and then a robust estimate of the covariance matrix of the parameter estimates is computed to account for correlation among results for procedures performed on the same patient. How do the robust estimates of the standard errors differ from those based on the assumption of completely independent failure times?

(b) Using the same set of covariates as in part (a), fit a frailty model to the data with a separate random effect for each patient? How do the standard errors from the frailty model differ from the robust standard errors produced in part (a)? Do the parameter estimates differ from the estimates in part (a)? Is there any indication that the frailty model does not adequately account for correlation among failure times for procedures performed on the same patient?

(c) Interpret the effects of age, diabetes and hypertension on the risk of failure. It is not enough to simply say that an effect is statistically significant. Describe the direction and magnitude of the effect.

4. In a study of the effects of energy-restricted diets on preventing skin cancer, mice were randomly assigned to either an ad lib (all you can eat) or an energy restricted diet. Restricting calorie intake cause the adrenal gland to produce more a substance that is thought to help prevent skin tumors caused by exposure to ultra violate radiation. To examine this, two thirds of the mice in each diet group were randomly selected to have their adrenal gland surgically removed. Within each diet group, half of the mice that had their adrenal gland removed received injections to restore levels of the particular substance thought to help prevent skin cancer. Consequently, there were six treatment groups corresponding to a 2 x 3 factorial arrangement; two diets (ad lib and energy restricted) and three levels of the surgery factor (no surgery, adrenal gland removed, adrenal gland removed and injections). Here we will only examine the effects of these six treatments on weight gain or loss over the first 27 weeks of the experiment. Each mouse was weighed at the beginning of the study and at 4, 8, 16, 20, 24, and 27 weeks. Weights for some mice are missing because some mice died before the end of the study. There are not equal numbers of mice in the six treatment groups because some mice were eliminated for various reasons at the beginning of the study. We will assume that the missing data are missing at random, which means that the missing results if we could see them would look like a random sample of results for mice in the respective treatment group. The data are posted in the file mweight.dat on the assignment section of the course web page. The variables are mouse id, treatment, and a set of seven weights corresponding to 0, 4, 8, 16, 20, 24 27 weeks. Code for entering these data into SAS is posted as mweight.sas, and code for entering these data into an S-PLUS data frame is posted as mweight.ssc. Both sets of code produce plots of average weights across time for the six treatment groups. Note that mice that are lost to follow-up can contribute to
average weights early in the experiment but not contribute to corresponding average weights at later inspection times.

(a) First run a two-way ANOVA at each of the seven time points. Since mice were randomly assigned to treatment groups, you would not expect any significant treatment effects at time zero. At what time point do the differences among treatments appear to stabilize? Describe the effects of the treatments on mean weight.

(b) Fit a marginal model to the data using a separate cubic trend in weight across time for each treatment. Allow for an arbitrary covariance structure for observations taken on the same mouse. Are there any indications of diet or surgery effects?

(c) Repeat part (a) using compound symmetry (CS), heterogeneous compound symmetry (CSH), first order autocorrelation model with homogeneous variances (AR(1)), first order autocorrelation model with heterogeneous variances (ARH(1)), a Toeplitz (TOEP) covariance structure, a heterogeneous Toeplitz (TOEPH) covariance structure, and a spatial power covariance structure (SP(POW)) for the repeated measurements of weights on a single mouse. Which covariance structure appears to be most appropriate? Which standard errors were most affected by changing the covariance structure?