Stat 430 Homework 5

due: December 11, 2009 at 5pm

This question will run you through a multiple (generalized) linear regression analysis of a model of virus growth in cell culture. The model is meant to emulate a biological experiment where virus (like HIV) are added to cells multiplying on growth medium in wells. The viruses infect the cells, and the infected cells begin producing more virus. Because virus production is sloppy, many of these viruses are defective, i.e. non-infectious. Periodically, the biologist samples a portion of the solution out of the well and counts the number of infectious *and* non-infectious, defective virus, the *response*. In this analysis, however, there is no experimental data, only data produced by a computer program which it is hoped emulates the true biological process. The computer program takes many input *parameters* and produces the predicted *response*.

A portion of this dataset was first introduced in the lecture notes (not in lecture) as an illustration of simple linear regression. The dataset was produced for a sensitivity analysis, whose goal, not unlike that of an experimental study, is to determine which parameters have the greatest impact on the response. The difference from an experimental study is that the model is known and the data is generated by the model, not experiment. A sensitivity analysis can be used to probe and "feel out" the model when it is essentially a black box.

1. The first part of any data analysis is to explore the data. `R` can help by giving you access to diverse plotting and summarization functions.

   (a) The data contains the output for 100 simulations of the model (a computer program generated these results). The first 9 columns are the values of the input parameters, i.e. in regression terminology the design matrix, for the 100 experiments. The input parameters and their meaning are listed in the following table.

Parameter	Meaning
β_W	production rate of virus from infected cell
η_W	proportion of produced virus that are functional
ξ_W	rate at which functional virus become defective
ν_W	clearance rate of defective virus
γ_W	infection rate
δ_W	death rate of infected cells
α	growth rate of uninfected cells
$C_{\max}$	maximum number of cells that can fit in a well
init_W	initial number of infected cells at experiment start

   The next 10 columns are the computer program output, predicted number of virus, at several times points after the experiment start (days 2 through 11).

   Plot the time course of virus count for trials $1, 2, 4, 18$, and 19 in a single plot to get a feeling for how the virus growth profile changes as the parameters vary. The response is clearly not linear in time. Why would we be justified in using a linear model?

(b) A sensitivity analysis is a very carefully designed study of parameter effects on the response. Because the sample is produced by a computer, there are few limitations on how the predictor variables are set, unlike in an experimental or observational study. This sensitivity analysis used a latin hypercube design, which we briefly discussed in a lecture on experimental designs for ANOVA studies. To understand a little about how the design chooses input values, produce the following plots

1. A histogram of the β_W predictor. All other predictors have similar histograms. What distribution does this look like?
2. A scattergram of predictor β_W vs. γ_W. Other pairs of predictors look similar. Do you see any sign of a reslationship between these predictors?

(c) In what follows, we will focus on predicting the number of viruses at 2 days post inoculation (DPI). It should be intuitive that the initial number of infectious virus should have a direct effect on the number of virus at later time points. We will use this predictable relationship to explore whether some kind of data transformation is needed for these data. Produce a scatterplot of init_W vs. `Y.2` and another of init_W vs $\ln(Y.2)$. Why is ln a good transformation to use on the viral count data?

(d) [optional] Plot all 9 scatterplots of each predictor against the log of viral count at 2 DPI. To get all scatterplots in one plot, use `par(mfrow=c(3,3))`, and then issue the 9 plot commands. Which predictors do you think have an effect on the chosen response?

2. Continuing our analysis, in this question you will estimate model parameters.

(a) Using `R`, create a design matrix, call it `X`, that includes an intercept term and the predictors β_W and α only. Pull out the logged virus count at 2 days post infection (DPI) as a response and store it in vector `Y`. Use the `solve()` function to solve the normal equations

$$X^T X \hat{\beta} = X^T Y$$

for $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_{\beta_W}, \hat{\beta}_\alpha)$. Please note that `R` uses the operator `%*%` for matrix multiplication. Interpret the results. How do the inputs β_W and α affect the amount of virus at 2 DPI? What happens if you try to fit all the predictors with a larger design matrix X?

(b) Use `R` to manually find the unscaled covariance matrix $(X^T X)^{-1}$, the estimated scaling factor S^2, and report 95% confidence intervals for $\hat{\beta}_0, \hat{\beta}_{\beta_W}$, and $\hat{\beta}_\alpha$.

(c) Now use `R`'s `lm()` function to fit all predictors simultaneously in a multiple linear regression model for the expected log virus count at 2 DPI (the same response used above). Look at a `summary()` of the fit to report which predictors have a significant impact on the response. Would you characterize the fit as good? If any conclusions differ from the preceding manual analysis, suggest an explanation.

(d) Take a look at the unscaled covariance matrix (store the result of `fit.s <- summary(fit)` and access the component `fit$cov.unscaled`; you can list all its components with the `names()` function). Both η_W and init_W are significant parameters in the model. Calculate the correlation of their estimates and interpret it.

3. We will now make some inferences about the model.
   (a) Can you reject null hypothesis
   $$H_0 : \beta_{\beta_W} = \beta_{\eta_W} = \cdots = \beta_{C_{\max}} = \beta_{\text{init}_W} = 0?$$
   (b) Verify the t value computed for β_{β_W} and show that the F test for $H_0 : \beta_{\beta_W} = 0$ (in the presence of all other predictors) produces the same result as reported in the summary.
   (c) [optional] Further investigation has suggested the values $\beta_W = 3500, \eta_W = 0.007, \xi_W = 2.77, \nu_W = 0, \gamma_W = 4.6, \delta_W = 0, \alpha = 2, C_{\max} = 1.2 \times 10^6$, and $\text{init}_W = 250$. Predict a response at 2 DPI using R's `predict()` function and find the 95% confidence interval. Why is this prediction useless? What did we do wrong?

4. We do some diagnostics to check for potential problems.
   (a) Use R's function `rstudent()` to plot the studentized residuals (were incorrectly called "standardized residuals" in lecture) as a function of the predicted response.
   (b) Identify the most obvious outlier (R's `identify()` with same `x` and `y` arguments as the plot function can help) with the largest studentized residual. Refit the full model without the outlier and then test whether the predicted value of the outlier according to the new model is significantly different from the observed outlier response. Use a Bonferroni correction to account for the fact that you have effectively done 100 such tests, one for each predicted value, when you chose the one with largest studentized residual.
   (c) Check a normal probability plot of the studentized residuals. Any problems?

5. Although sensitivity analysis is less interested in parameter estimation and prediction than most applications of multiple linear regression, you will proceed to model refinement for the exercise of it. The goal now is to identify the simplest model by excluding unimportant predictors. The simple model has the advantage of improved parameter estimation and more precise predictions. Generally, before model refinement, you should remove outliers where justified and deal with non-normality issues. If the outliers are bad data (perhaps data entry problems or program bugs), then you would remove them before continuing. In this case, there is no obvious problem with the data, so we leave all data intact.

   Also, we will ignore any evidence of non-normality.

   (a) In *backward elimination* you fit the full model with all predictors (you already have that). Then, you remove the predictor with the largest p-value at least as large as $p_{\text{threshold}}$, and refit the model. Repeat until p-values for all predictors are smaller than $p_{\text{threshold}}$. A reasonable threshold value is $p_{\text{threshold}} = 0.15$. Find the best model according to this proposed backward selection procedure.
   (b) [optional] Redo question 3c with the simplified model. Are the results now trustworthy?

6. For this question, use the mining data, a report of the number of fractures in upper seams of coal mines. It also records data about the seams, including `X1`, the shortest distance between the seam floor and the next lower seam, `X2`, the percent extraction of the lower seam, `X3`, the lower seam height, and `X4`, the age of the mine.

   (a) The *residual deviance*, as reported by `R`'s `summary()` applied to a `glm()` fitted object, is the log likelihood ratio statistic

   $$\lambda(\hat{\beta}) = -2\ln\left(\frac{L(\hat{\beta})}{L(\hat{\mu})}\right)$$

   where the "perfect" model with maximized likelihood $L(\hat{\mu})$ fits a separate mean μ_i to every observation Y_i. Notice that if the generalized linear model, with mean

   $$\mu_i = \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}\right) = e^{x_i^T \beta}$$

   for observation i, fits the data well, then $\lambda(\hat{\beta}) \approx 0$. Does a linear model with log link function and all predictors fit the data reasonably well, or is there evidence of a lack-of-fit of the data to the assumed generalized linear model?

   (b) We discussed use of the likelihood ratio statistic $-2\ln\left(\frac{L(\hat{\beta}_0)}{L(\hat{\beta})}\right)$ for testing whether a restricted model with parameter vector β_0 fits the data significantly worse than an unrestricted model with parameter vector β. Under the null hypothesis that β_0 is the true model, it has an asymptotic χ^2 distribution with degrees of freedom equal to the difference in number of parameters between the two models. Notice that this statistic can be written as a difference in deviances:

   $$\begin{aligned}\lambda(\hat{\beta}_0) - \lambda(\hat{\beta}) &= -2\ln\left(\frac{L(\hat{\beta}_0)}{L(\hat{\mu})}\right) + 2\ln\left(\frac{L(\hat{\beta})}{L(\hat{\mu})}\right) \\ &= -2\ln\left(\frac{L(\hat{\beta}_0)}{L(\hat{\mu})} \times \frac{L(\hat{\mu})}{L(\hat{\beta})}\right) = -2\ln\left(\frac{L(\hat{\beta}_0)}{L(\hat{\beta})}\right)\end{aligned}$$

   Thus, you may use differences in deviances to compare pairs of nested models.

   Use these differences in deviances, starting from the `null deviance` (reported by `summary()`) which is for the model $\mu_i = e^{\beta_0}$ to build a model by adding the predictors one-at-a-time in order `X1,X2,X3`, and `X4`. Only include predictors that significantly improve the fit, and test the subsequent predictors in the context of all previous predictors that warranted addition to the model. Don't worry about multiple testing and use a significance level of 0.05. Summarize how the important predictors impact the number of fractures in upper coal mine seams?

7. [optional] In this question you will fit a logistic growth curve using nonlinear regression. The data are the results of 6 independent experiments of the virus growth experiment without the virus. The response is the number of cells in the well during 19 days. Plotting the data (`t` against `cell`) suggests that a logistic growth curve would work well

for this data. The `R` library `nlme` provides funtion `nls()` for fitting the logistic curve, `SSlogis()`. The logistic function as parameterized in `SSlogis(input, Asym, xmid, scal)` is

$$\frac{\text{Asym}}{1+\exp\left(\frac{\text{xmid}-\text{input}}{\text{scal}}\right)},$$

where `Asym`, `xmid`, and `scal` are three parameters of the model to be fitted, and `input` are the predictor values chosen by the experimenter. In our case, the predictors are the times at which cell counts were taken. Initialize the three parameters at reasonable values by looking at a scatterplot of the data and fit the model using `nls(cell` $\sim$ `SSlogis(t, Asym, xmid, scal), data=d)`, where `d` is the data in the provided file. The parameterization used in the model of the preceding questions is

$$\frac{C_{\max}C(0)e^{\alpha t}}{C_{\max}+C(0)(e^{\alpha t}-1)}$$

where $C(0)$ is the initial number of cells, α is the growth rate, and $C_{\max}$ is the maximum number of cells. Are the values of α used in the sensitivity analysis in reasonable agreement with this data?