Due: 5pm Friday, 22 Feb

1. The file temperature.txt contains data on worldwide average temperature from 1880 to 1987. We wish to estimate the linear slope, i.e. the average increase in temperature per year.

   (a) Fit the model $Temp_i = \beta_0 + \beta_1 Year_i + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. What are the estimated slope, $\hat{\beta}_1$ and its s.e.?

   (b) For these data, the assumption of independent errors is probably not reasonable. One reasonable alternative is that errors follow an ar(1) model. For that model, the first few rows and columns of the variance-covariance matrix of ϵ can be written as

   $$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

   I haven't written out the full 108 x 108 matrix, for obvious reasons, but I hope you see the pattern. Various lines of evidence suggest $\rho = 0.5$. What are the estimated slope and its s.e. if you assume $\rho = 0.5$?

   (c) Based on what you know about consequences of misspecifying the variance-covariance matrix, are these results surprising? Explain why or why not.

   (d) Reconsider the independent errors model in part 1a. Assume that the errors follow a AR(1) model with $\rho = 0.5$. Calculate the sandwich estimator of variance of the slope from part 1a.

   (e) Use the gls() function in the R nlme package (or comparable SAS proc mixed code, see me) to estimate ρ for these data. What is $\hat{\rho}$?

   R note: gls() has a syntax like lm() with a couple of extra arguments. For example, gls(formula, correlation=corAR1(), method='ML') will fit the specified formula using an AR1 correlation structure with a ML estimate of ρ. method ='ML' is important because the default is REML. logLik(object) extracts the log likelihood from a model object. You can use this on an lm or a gls output object or on many more model objects.

   (f) Use a likelihood ratio test to test the null hypothesis that $\rho = 0$. Report your test statistic and p-value.

   (g) Use a likelihood ratio test to test the null hypothesis that $\rho = 0.5$. Report your test statistic and p-value.

2. The data in birth2.txt are from a sociological study in Baltimore. The investigators believed that there is an association between loss of a child during pregnancy and the behaviour in school of a subsequent live-born child. The prospective research design (looking forward) would be to enroll some mothers who have lost a child and some who haven't, then look at school behaviour (problem or not). This is very inefficient because the proportion of school problems is low (thankfully!).

Instead, the investigators used a retrospective design, which is common in epidemiology and efficient at examining associations with rare events. The investigators identified 255 problem children and 110 control (non-problem) children. The mothers were then asked about the birth order (2nd child, 3rd child, ...) and whether they had lost the previous child. Birth order is 2 (2nd child), 3.5 (3rd or 4th child) and 5 (5th or higher child). The response is whether the previous child was lost or not. (The loss rates seem high to me, but that's what they were recorded as). The investigators want to know whether the odds of losing a child are associated with birth order and/or whether or not the subsequent child was a problem in school.

(a) Fit an appropriate generalized linear model treating problem as a factor and birth order as a continuous linear covariate. Do not worry about overdispersion for now. Report the odds ratio for being a problem child and the odds ratio for a 1 child increase in birth order. Is being a problem child associated with an increase or a decrease in the probability that your mom lost the previous child?

(b) Test whether the odds ratio for being a problem child = 1 and whether the odds ratio for birth order = 1. In both cases, report your test statistic and a p-value.

(c) Test whether the effect of birth order is the same for problem and control children. Report your test statistic and the p-value.

(d) Is a linear model for birth order adequate, or does the effect of birth order need a more complicated model? Report your test statistic and the p-value.

3. This question explores properties of retrospective designs. Consider a simplified version of the Baltimore study in question 2. The only two factors under study are Problem child: Y/N and Lost previous child: Y/N. The population counts are:

Lost	Problem child		Total
Child	Yes	No	
Yes	400	100	500
No	600	98,900	99,500
Total	1000	99,000	100,000

The investigators are interested in:
1) P[Problem = Y | Lost=Y]
2) P[Problem = Y | Lost=N]
3) P[Problem = Y]
4) The odds ratio describing the association of Problem child = Y and Lost child = Y

(a) Calculate these four quantities for the population

(b) Consider a random sample of 250 mothers from the population. Calculate the expected counts in that sample, then use those expected counts to calculate the four quantities.

(c) Consider a random sample of 255 mothers of a problem child and 100 mothers of a non-problem child. Again, calculate expected counts and use them to calculate the four quantities. Do not round fractional counts. This is the retrospective study design.

(d) Which of the four quantities can be correctly estimated in the random sample (part b)? If any are not correctly estimated, explain why they are not.

(e) Which of the four quantities can be correctly estimated in the retrospective design (part c)? If any are not correctly estimated, explain why they are not.