EXERCISE DUE FRIDAY NOON, 10/4

(1) Data on a sample of California schools is available on the course web site, either as an R data file (`caschool.RData`) or an Excel file (`caschool.xlsx`). In R, after `load("caschool.RData"`, a data frame named `caschool` will be on your workspace. A pdf file describing the variables is available as `californiatestscores.docx`.

The principal components of a vector of random variables with covariance matrix Σ are the uncorrelated vector of variables $z = W'y$, where $\Sigma = W\Lambda W'$ is the eigenvalue decomposition of Σ. (W's columns are Σ's eigenvectors.) The rows of W determine how much each variable in $y = Wz$ "loads" on each of the principal components z.

(a) Find the eigenvector decomposition of the correlation matrix of the numeric variables in the data. (The first few columns are non-numeric.) [In R, these commands may be helpful: `cor()`, or `cov()` followed by `cov2cor()`, and `eigen()`.]

(b) Which principal components are most important in determining `testscr`, according to this decomposition? (I think two stand out.) Looking at the columns of W corresponding to these two components, which variables are strongly related to `testscr`?

Here's a sequence of R commands that check which principal components are most important in determining `testscr`.

```
> ev <- eigen(cor(caschool[ , -(1:5)]))
> dimnames(ev$vectors)[[1]] <- names(caschool)[-(1:5)]
> ev$vectors["testscr", ]
 [1]  0.401332321 -0.120991993  0.032842203 -0.031406307 -0.10573
 [6] -0.110538803 -0.217087897 -0.283478808 -0.050556731  0.02193
[11] -0.103401123  0.001055974 -0.810932894
```

The coefficients on `testscr` for the first and 13th are considerably bigger than any of the others. These two columns of the eigenvector matrix are

```
enrl_tot -0.1509110 -7.740888e-07
teachers -0.1465162  1.055164e-06
calw_pct -0.2875774 -4.499499e-09
meal_pct -0.3790833 -1.524986e-07
computer -0.1114922 -3.692731e-07
testscr   0.4013323 -8.109329e-01
comp_stu  0.1563686  9.388346e-08
expn_stu  0.1133054 -7.164313e-10
```

```
str       -0.1474874  4.695266e-08
avginc     0.3076185  1.650080e-08
el_pct    -0.3046274  1.808937e-07
read_scr   0.4029943  4.279097e-01
math_scr   0.3833833  3.991004e-01
```

The second has positive, nearly equal weights on `read_scr` and `math_scr`, and a negative weight, about twice the size, on `testscr`. Its associated eigenvalue is tiny, zero roughly to machine precision. This is just picking up the fact that `testscr` is the average of `math_scr` and `read_scr` The only reason the weights are not proportional to (1, -.5, -.5) is that the scores are scaled slightly differently in the conversion to a correlation matrix. The first column picks up substantial weights on many variables, including positive, similar weights on all three score variables. The other large positive weight is on `avginc`. Large negative weight appear on `meal_pct` and `el_pct`. `str` gets a negative weight, but it's small, fourth to last in absolute value.

(c) Using the same `caschool` data, estimate a linear regression of the `testscr` (school average test scores) on all the other non-numeric variables except the last two (`math_scr` and `read_scr`). [In R, the command

```
lm(testscr ~ x + y + z, data=caschool)
```

will work, with "`x + y + z`" replaced by a list of all the other variable names, separated by plus signs.]

Here's the part of `summary(lmout)` from R that gives coefficients and posterior standard errors.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.588e+02  9.748e+00  67.581  < 2e-16 ***
enrl_tot    -5.647e-04  1.648e-03  -0.343   0.7321
teachers     8.203e-03  3.636e-02   0.226   0.8216
calw_pct    -8.290e-02  5.834e-02  -1.421   0.1561
meal_pct    -3.739e-01  3.634e-02 -10.288  < 2e-16 ***
computer     1.570e-03  3.112e-03   0.505   0.6141
comp_stu     1.018e+01  7.767e+00   1.310   0.1908
expn_stu     1.597e-03  9.035e-04   1.767   0.0779 .
str         -1.525e-01  3.262e-01  -0.467   0.6404
avginc       6.147e-01  8.972e-02   6.851 2.71e-11 ***
el_pct      -1.995e-01  3.508e-02  -5.686 2.47e-08 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
```

(d) Explain why the last two are left out — the principal components decomposition may help.

As already noted, the last two are related to `testscr` trivially via an identity, and our regression would just recover the identity if we included those variables.

(e) If you were interested in what determines `testscr`, would you reach different conclusions from the principal components decomposition and the regression? Note that the principal components decomposition, being based on the correlation matrix, is scale-free, while the sizes of the regression coefficients are not. The t-statistics, as found easily with the `summary()` function applied to the regression output in R, are more comparable to the sizes of the coefficients in W.
The t statistics that exceed 2 are for the same variables that entered the first principal component: `avginc`, `meal_pct`, and the signs of the coefficients on variables other than `testscr` in the principal component vector match those in the regression. Note that a tightly fitting regression generally does not match this pattern. The 13th principal component, for example, has opposite-signed coefficients on `testscr` and the two component scores. The same-signed coefficient on the dependent variable that is found here arises when all the variables that get substantial weight are acting like noisy measures of a single underlying source of variation, which here might be "good school", or, maybe more plausibly, "student body from supportive backgrounds".

(2) (Stopping rule paradox) Suppose one has data on a single family from an isolated and unusual subculture where a) each family continues to have children until they have a boy, and then stop having children and b) the ratio of male to female births is not the usual .5 and is unknown. The data are just the family size n, which of course consists of $n-1$ female births and one male birth.

(a) What is the likelihood function for this single sample, as a function of the unknown probability p of a male birth?
$(1-p)^{n-1}p$

(b) Show that a flat-prior Bayesian posterior mean for p is biased for most values of p. You can do this analytically, but quicker, and OK for the exercise, is probably just to check it numerically. There will be one value of p for which the posterior mean is frequentist-unbiased. What is it? (3 significant figures accuracy OK).
The flat-prior Bayesian posterior mean, since the likelihood is proportional to a $\text{Beta}(n,2)$ pdf, is $2/(n+2)$. The mean of this estimator is

$$\sum_{n=1}^{\infty} \frac{2}{n+2}(1-p)^{n-1}p$$

The point at which it is unbiased is $p = .568$, approximately. for p's below that it is biased up, for p's above that it is biased down. The largest absolute bias is at $p = 1$, where it is a downward bias of one third. The largest upward bias is at around $p = .16$, where the bias is about .18.

(c) If instead of one family, data on a fairly large number N of families is available, taking an arithmetic average of the Bayesian posterior means for each family as calculated above is not a good way to estimate p. Why?
In a large sample, the average of the biased estimates will be biased by the same amount as the individual estimates, so the estimates will not converge to the truth as the sample size increases.

(d) The likelihood for the full set of observations on N families implies a posterior mean that is different from the average of the family-by-family posterior means. What is the posterior mean for the full likelihood?
If we multiply the likelihoods for all the individual families together, as is appropriate if the family data are i.i.d. across families sampled, the result is just the same likelihood we would get if we ignored the family groupings and just counted boys and girls:

$$p^N(1-p)^{\sum_j n_j - N},$$

where N is the number of families (also the number of boys) and $m = \sum_j n_j - N$ is the number of girls, where j indexes families. With a flat prior, since this is a $\text{Beta}(N+1, m+1)$ distribution, the posterior mean is $(N+1)/(N+m+2)$. This does converge in probability to the true probability, because $N/(N+m)$ is the average over the sample of the i.i.d. variable that is 1 for boys and 0 for girls. Since the ratio of $(N+1)/(N+m+2)$ to $N/(N+m)$ goes to 1 as N and m go to infinity, the posterior mean converges to the true value with probability one.

(e) There is a frequentist-unbiased estimate of $1/p$ for the case of data on a single family. What is it? Is there an unbiased estimate of p itself for this case? What about a procedure that takes an arithmetic average of the per-family unbiased estimates of $1/p$ when there are N families, then estimates p as the inverse of this average. Would it converge in probability to p as N increased? Can it be improved upon?
This may have involved a bit more calculus cleverness than I intended. It seems intuitively plausible that n, the number of children in the family, is an unbiased estimate of $1/p$. (This does involve treating n as $+\infty$ when $p = 0$.) Proving this is not hard, if you've seen this kind of argument before, but possibly hard otherwise. The expected value of n is

$$\sum_{n=1}^{\infty} np(1-p)^{n-1} = p\sum_{n=1}^{\infty} -\frac{d}{dp}(1-p)^n = -p\frac{d}{dp}\sum_{n=1}^{\infty}\frac{1-p}{p} = \frac{1}{p}.$$

So n is an unbiased estimator of $1/p$. Averaging it across a large number of randomly selected families will therefore give a result that converges in probability to $1/p$. Notice that the average of n across the N families is just $(N+m)/N$, which is the inverse of the maximum likelihood estimator of p.

The MLE, being average of $N+m$ i.i.d. variables, satisfies the CLT and therefore converges to the true p at the rate $(N+m)^{-1/2}$. One over this estimator of $1/p$ differs from the Bayesian posterior mean by a factor that converges to one at the rate $1/(N+m)$, so the two estimators are nearly the same in large samples. The frequentist unbiased estimator of $1/p$, because it is the MLE for that function of the parameter, is a function of the sufficient statistics. Thus there is no quick way to improve on it by taking its expectation conditional on the sufficient statistics (i.e. applying the Rao-Blackwell theorem, which we stated in class). Since it is the maximum likelihood estimator of $1/p$, it is also fully efficient as an estimator of that. So it can't be improved upon.

Note that, though n is unbiased in a frequentist sense, it gives $1/p = 1$ when $n = 1$. We know that $1/p \geq 1$, so any reasonable distribution for $1/p$ given the data in that case must have expectation *above* one. For characterizing post-sample beliefs, calling the $1/p = 1$ estimate "unbiased" is misleading, though nonetheless technically correct, since "unbiasedness" refers only to behavior of the estimator across hypothetical repeated samples, not to reasonable beliefs given the actual sample at hand.