UWEO StatR201 – Winter 2013: Homework 2

Due: Thursday January 31, before Class (grace period 1 additional week)

Assaf Oron, *assaf@uw.edu*

Reading related to this assignment: Lectures 3-4; Dobson and Barnett Sections 1.6, 3.1, 3.2, 3.4 and (only browse) Ch. 7; Hastie, Tibshirani and Friedman Sections 4.4.1-4.4.2.

Instructions:

- Please submit online in the class dropbox. Please submit either ***.pdf is accepted as the main submission (with code pasted verbatim), or *.rmd.** (tip: to save time when using knitr, use `'cache=TRUE'` in the chunk header for any code chunks that run big simulations – this will prevent them from re-running each time you recompile the code).
- **Starred (*) questions and question-parts are not required.** You may submit them if you choose, or do any part of them without submitting.
- **Grading is determined chiefly by effort, not by correctness.** If your submission shows evidence of independent, honest effort commensurate with the amount of homework assigned – you will receive full credit.

0. R Cool Trick of the Week

Please insert a mathematical expression or symbol to least one caption in one of your plots, and similarly at least one instance of a variable's value embedded in a caption. If you manage to do *both*, plus strings, in a single caption (following the trick provided by Eli and now available on Lecture 3 notes) – all the better, but not required.

Note: no need to add a plot. Just "embed" these features in some of your plots.

1. R Annoyance of the Week: Custom panel titles in lattice

Using the workaround shown on Lecture 4 (via `dimnames`), generate at least one lattice multi-panel plot, with the panel headings changed to be more meaningful and descriptive. Please use some `lattice` function other than `'xyplot'`.

2. Linear Regression simulation.

We started playing with this simulation in class. Continue at home with **ensembles** of 1000 instances each:

a. For the Null case (x,y independent), check **how often the p-value falls below 0.05** (if this happens ~5% of the time or less, the regression's t-test is said to be ***unbiased***), and **what is the most extreme p-value you get in each ensemble.**

Do it for sample sizes of n=10 and 100, and for 4 distributions of *y*:

- Normal
- t with 4 d.f., normalized by dividing by the square root of 2 (use `rt`)

- Exponential (scale=1)

- Cauchy (`rcauchy`), divided by 2. **Note: this one *might* behave strangely.**

(so overall, 8 different ensembles of 1000 runs each).

To get started, use this code:

```
x=matrix(rnorm(10000),nrow=10)
y=matrix(rnorm(10000),nrow=10)
betas=mapply(function(z,w) {summary(lm(w~z))$coef[2,]},
              split(x,col(x)),split(y,col(y)))
```

This returns a 4-column matrix with the β point estimates, SEs, t-statistics and p-values.

Note: for each sample size, randomize *x* only once, then keep it the same as you change *y* (*a simulation tip: don't change too many things at once*).

b. Repeat the exercise for a true-effect univariate case (one x variable).

For each of the 8 cases, generate an ensemble of

`y=2*x + (random noise),` and calculate:

- The empirical **power:** in what proportion of runs is the Null rejected and for the right reason (i.e., showing a positive effect)? You can do it by comparing the t-statistic to the 97.5% percentile of the t-distribution.

- The empirical **bias**: (ensemble average of effect estimate) – (true value, which is 2)

- The **interval coverage**: in what proportion of the 1000 runs, does **the true value (=2)** fall inside **the estimated 95% confidence interval around the effect estimate?** In other words: for each run take the point estimate, add and subtract the SE * (the 97.5% t-distribution multiplier), and check if the resulting interval contains the true value.

You can present the results via simple tables, or plots, whichever you like more.

***Starred option: qqnorm the distribution of effect estimates, and evaluate under what distributions and sample sizes it seems "close enough" to normal.**

c.* Repeat all this for a bivariate case, with two mildly correlated covariates (say, r=0.4 between them. focus on only one of the covariates, to save work). One quick way to generate two Normal variables with a pre-set correlation, is via `mvrnorm` in the MASS package.

3. Estimator-Performance Simulation.

a. Simulate the Uniform(0,β) i.e., the "snow-day bus stop" scenario described in class. Refresher: the goal is to estimate the distribution's upper limit β (set the true value to 1, to simplify matters). Compare the MLE and "intuitive" (MME) estimators' performance, on bias, variance and RMSE (root-mean-square-error), using sample sizes of n=5, 25 and 125.

To make future work easier, prepare generic functions to gauge estimator performance. For example,

```
rmse=function(x,ref) sqrt(mean((x-ref)^2))
```

Comment on the results. At what sample size does the MLE start beating the "intuitive" estimator? Note: the "intuitive" moment-based estimator is *not* described in the lecture notes. You will have to actually *play* Lecture 3 to see it.

Starred option: add visual comparisons – boxplots and/or qqnorms. Don't forget to mark the true value with a horizontal/vertical line.

b*. As mentioned in class, this problem's MLE is hopelessly biased. A combination of statistical theorems (usually studied at the 1st-year-graduate level) shows that under most cases, an *unbiased* estimator which is an algebraic function of the MLE, will be the best among all unbiased estimators (i.e., it will have the smallest variance since the bias is already 0). Find this **"modified MLE"** – either yourselves using common-sense considerations and/or probability calculations, or by searching online for the expectation of the maximum of a uniform i.i.d. sample – and compare its performance to the other two, on the sample sizes used above. (hint: the modification should be very simple).

4. The Likelihood always goes up! We learned about Likelihood-ratio tests on Lecture 4, and started to see how the Likelihood/"Deviance" might play a role in **model selection.** Here's a curious property that is often missed: **show that when you add a covariate to an existing model, you are essentially guaranteed that the likelihood under the new MLE will be greater (="better") than under the old MLE.**

The proof does **not** need to be formal (hint: recall that the simpler model is equivalent to assuming that the additional covariate has a coefficient of **zero**).

5. Logistic Regression: the Basics.

This is a simple logistic regression, just to get acquainted with interpretation, likelihood-ratio tests, and useful descriptives/diagnostics for this important model type.

Download the file "HosmerLemeshowHeart.csv", taken from the website containing datasets from *Hosmer and Lemeshow (2000) Applied Logistic Regression: Second Edition.* (http://www.umass.edu/statdata/statdata/stat-logistic.html). The question of interest is how coronary heart disease (coded as 0 or 1) depends upon age.

a. Regress disease status upon age, show the summary. Write a brief text interpretation of the results as done in Lecture 4, including the baseline probability at some meaningful reference age (hint: center the age covariate when running the regression), the odds-ratio per year, and the 95% CI for that odds-ratio.

b*. Examine whether there might be higher-order (nonlinear) effects of age, by adding quadratic, cubic, etc. terms. Use the likelihood-ratio test to determine significance. Again, it's highly advisable to center any variable for which higher-order polynomial terms are added.