STA 2201S 2014 Assignment 2. *due Friday, February 28 at the beginning of class*

When answering questions requiring numerical work, the results are to be reported in a narrative summary, in your own words. Tables and Figures may be included, but must be formatted along with the text. DO NOT include in this summary printouts of computer code with the relevant selections highlighted. All computer code used to obtain the results summarized in the response should be provided as an appendix. In this appendix you may highlight the relevant results.

1. *Deviance and scaled deviance for the Gamma distribution:* Suppose y_j follows a Gamma distribution with shape parameter ν and mean parameter μ_j, and that $g(\mu_j) = x_j^{\mathrm{T}}\beta = \eta_j$, where $j = 1, \dots, n$, $\beta = (\beta_1, \dots, \beta_p)$, $p < n$.

   (a) Derive the scaled deviance, defined as

   $$D = 2\sum_{j=1}^{n}\{\log f(y_j; \tilde{\eta}_j) - \log f(y_j; \hat{\eta}_j)\}, \quad \hat{\eta}_j = \eta_j(\hat{\beta}),$$

   where $\tilde{\eta}_j$ is the maximum likelihood estimator of $g^{-1}(\mu_j)$ when no linear constraint is invoked, and $\hat{\eta}_j = x_j^{\mathrm{T}}\hat{\beta}$, the maximum likelihood estimator under the generalized linear model.

   (b) Show that the scaled deviance is asymptotically equivalent to

   $$\nu\sum_{j=1}^{n}\left(\frac{y_j - \hat{\mu}_j)}{\hat{\mu}_j}\right)^2.$$

   (c) Compare the estimator of ν^{-1} proposed in the text: $\hat{\phi} = \frac{1}{n-p}\sum(y_j - \hat{\mu}_j)^2/V(\hat{\mu}_j)$ with the maximum likelihood estimator of ν.

2. *Agresti, 2002, Problem 6.20.* Logistic regression is applied increasingly to large financial databases, such as for credit scoring to model the influence of predicts on whether a consumer is creditworthy. The UCI machine-learning data archive `ftp://ftp.ics.uci.edu/pub/machine-learning-databases/statlog/german/` has a data set that includes 20 covariates and 1000 observations. The response of interest is the credit score (1= Good, 2 = Bad). Build a model for credit-worthiness using the predictors running account, duration of credit, payment of previous credits, intended use, gender, and marital status. Several of the variables are categorical (see "german.doc" in the archive); for example the first attribute takes values "A11", "A12", "A13", and "A14", to indicate status of existing checking account (< 0, $0 - 200$, > 200, no account). Some people have converted these to numeric categories, such as "1", "2", "3", "4"; see `http://www4.stat.ncsu.edu/~boos/var.select/german.credit.html` for the SAS code that does the conversion. For `R` it should be fine to work with the original categorical variables.

   Write up the conclusions of your model building exercise in two parts: (i) a one-page summary for a bank manager, (ii) a statistical report for the quants.

3. *Ref: Rosenbaum (2006)*: The Fatal Accident Reporting System records information on every road accident in which there was at least one fatality. Evans (1986) ‘double pairs design’ restricts attention to the subset of crashes with two people in the front seat of a car, in which exactly one of these two people died and exactly one was wearing a seat belt. The summary data is given in Table 1.

   (a) Assuming that we are interested in the question of whether or not seat belts save lives, what potential confounding variables are controlled for in Evans’ design? What potential confounding variables are not controlled for in this design?

   (b) For the data in Table 1, compute the relative risk of the Driver dying and the Passenger surviving, when the Driver is not belted (and the Passenger is) relative to when the Driver is belted (and the Passenger is not), and give an estimated standard error for your estimate.

   (c) Table 2 gives more detailed information on the breakdown by age of the data in Table 1. Is there evidence that the relative risk changes with age?

References

Rosenbaum, P.R. (2006). Differential effects and generic biases in observational studies. *Biometrika* **93**, 573–586.

Evans, L. (1986). The effectiveness of safety belts in preventing fatalities.

Table 1: From Rosenbaum (2006), quoting Evans (1986).

	Driver Not Belted, Passenger Belted	Driver Belted, Passenger Not Belted
Driver Died, Passenger Survived	189	153
Driver Survived, Passenger Died	111	363

Table 2: From Rosenbaum (2006), quoting Evans (1986).

	Age stratum s (Driver, Passenger)	Driver Not Belted, Passenger Belted	Driver Belted, Passenger Not Belted
Driver D, Passenger S	$s = 1$	75	36
Driver S, Passenger D	$(16 - 24, 16 - 24)$	22	92
Driver D, Passenger S	$s = 2$	6	6
Driver S, Passenger D	$(16 - 24, 25 - 34)$	4	20
Driver D, Passenger S	$s = 3$	2	4
Driver S, Passenger D	$(16 - 24, \geq 35)$	2	17
Driver D, Passenger S	$s = 4$	12	8
Driver S, Passenger D	$(25 - 34, 16 - 24)$	6	15
Driver D, Passenger S	$s = 5$	22	24
Driver S, Passenger D	$(25 - 34, 25 - 34)$	17	30
Driver D, Passenger S	$s = 6$	3	6
Driver S, Passenger D	$(25 - 34, \geq 35)$	6	21
Driver D, Passenger S	$s = 7$	4	8
Driver S, Passenger D	$(\geq 35, 16 - 24)$	0	8
Driver D, Passenger S	$s = 8$	5	9
Driver S, Passenger D	$(\geq 35, 25 - 34)$	2	16
Driver D, Passenger S	$s = 9$	60	52
Driver S, Passenger D	$(\geq 35, \geq 35)$	52	144