MS COMPREHENSIVE EXAM
Linear Models and Advanced Regression
August 2015, 80 Points

Write your answers on separate sheets of paper with your name on each sheet.

1. In Stat 505 Homework 9, we analyzed a binary response: "Voted against higher fuel efficiency standards (1) or voted in favor (0)" collected on 100 US senators and obtained this output.

	Estimate	Std. Error	z value
(Intercept)	-0.17	0.33	-0.50
partyR	1.61	0.61	2.63
contrib	0.46	0.17	2.73
partyR:contrib	0.33	0.41	0.81

The explanatory variables were `party` (D or R) and amount of contribution the senator accepted from auto companies. For this fit, I transform contributions by taking $\log_2$ (which then ranged from 7 to 17), then centered by subtracting 12.66, the mean $\log_2$ contribution.

(a) Describe the default transformation used by R's `glm` function on binary response data. (5 pts)

(b) Explain what the intercept estimate above is estimating in terms of how some hypothetical senator would have voted. Include some information about the precision of that transformed estimate. (8 pts)

(c) Interpret the effect of party in the above model. (10 pts)

(d) Would you prefer a simpler model? Explain. (7 pts)

2. Consider the following mixed effects model:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \ \boldsymbol{\varepsilon}_i \sim iid(\boldsymbol{0}, \sigma^2\boldsymbol{V}), \ \boldsymbol{b}_i \sim iid(\boldsymbol{0}, \boldsymbol{\Psi}), \quad i = 1, 2, \ldots, 36$$

(a) Write out the variance of $\boldsymbol{y}_i$ using the above notation. State any additional assumptions needed. (8 pts)

Now consider a particular setting in which we might use the above model.
$\boldsymbol{y}_i$ contains log(median house prices) in the ith city from years 2001 through 2010.
$\boldsymbol{X}_i$ includes a column of ones and four columns of predictors:
log_income,   population,   incomeGrowth,   year
measured on each city for each year.

(b) Each $\boldsymbol{y}_i$ is a time series on one city of length 6.

    i. Under AR1 correlation, what does $\boldsymbol{\Omega}$ look like? (6 pts)

    ii. How would you examine the need for this structure? (7 pts)

(c) Suppose the random effects are set, and we wish to test to see if year and incomeGrowth could be removed from the full fixed effects model (simultaneously).

    i. How would you fit and compare the two models (1) using the four fixed effects listed versus (2) using only log_income and population? Explain the test you would use and possible conclusions from the test. (7 pts)

    ii. What distributional assumptions are needed? (5 pts)

(d) Suppose we have decided to use all four predictors shown plus a quadratic year effect. Consider three different specifications for random city effects:

- random intercept for each city (no random year effect)
- random intercept and slope for year for each city
- random intercept, slope, and quadratic terms for year for each city.

    i. How would you fit and compare the models to decide which is the best random effects model? (7 pts)

    ii. What distributional assumptions are needed? (5 pts)

(e) To add the quadratic term mentioned just above, how would you construct a column of quadratic year effects which do not complicate (or change) the previous interpretation of the linear year effect? (5 pts)