UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
Basic Exam - Applied Statistics
Thursday, January 17, 2019

Work all problems. 60 points are needed to pass at the Masters Level and 75 to pass at the Ph.D. level.

1. Below are some concerns you might have while doing multiple linear regression analyses. For each:

   - How would you identify this problem? Name or briefly describe a diagnostic measure you would use.
   - How would know know if you had the problem? Describe or sketch a result of this diagnostic that would confirm your concern.
   - How would you address the problem in your analysis? Describe one approach.

   (a) (9 points) You suspect two predictors are highly correlated

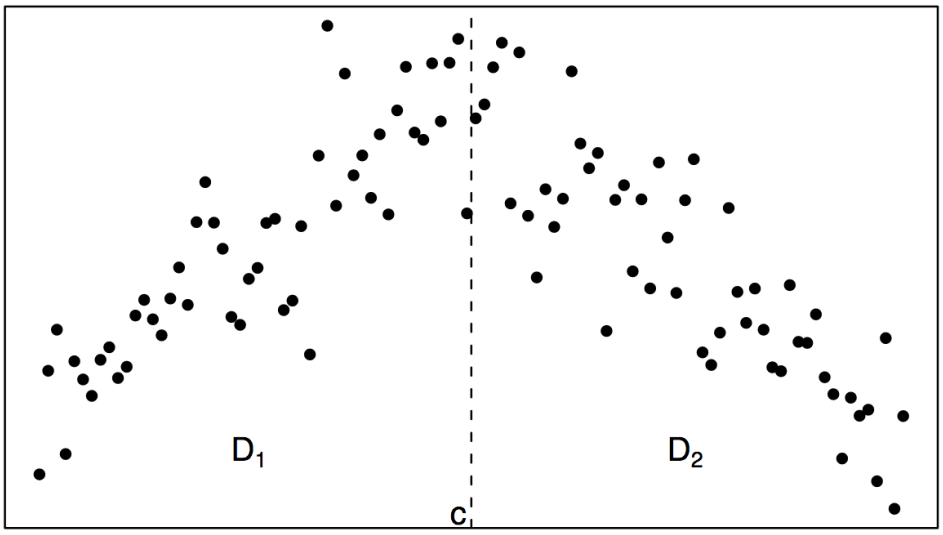
   (b) (9 points) You suspect the errors are non-normal

   (c) (9 points) You suspect 1 point may be an outlier and heavily influencing your results

   (d) (9 points) You suspect the variance is non-constant, likely increasing with the fitted value

   (e) (9 points) You suspect the relationship between X_1 and Y is non-linear

2. Consider the dataset $\{(x_i, y_i)\}_{i=1}^{2n}$ plotted in the figure below. The dataset is divided in two, with n observations where $x_i < c$ and n observations with $x_i > c$.



To accommodate the change of behavior at $x = c$, you adopt a *change point linear model:*

$$y_i = \alpha_0 + \gamma_0 x_i + (\delta + \eta x_i) Z_i + e_i, \quad i = 1, \ldots, 2n,$$

where

$$Z_i = \begin{cases} 0 & x_i < c \\ 1 & x_i > c, \end{cases}$$

and $e_i \sim N(0, \sigma^2)$.

(a) (3 points) What is the interpretation of δ and η?

Now suppose a previous analyst (naively) adopted separate models for dataset 1, $D_1 = \{(x_i, y_i) : x_i < c\}$, and dataset 2, $D_2 = \{(x_i, y_i) : x_i > c\}$:

$$y_i = \alpha_j + \gamma_j x_i + e_{ij}, \quad e_{ij} | x_i \sim N(0, \sigma_j^2), \quad x_i \in D_j, \quad j = 1, 2$$

and found least squares estimates $\hat{\alpha}_j$, and $\hat{\gamma}_j$, and unbiased estimate $\hat{\sigma}_j$ for $j = 1, 2$.

(b) (5 points) Provide least squares estimates for $\alpha_0, \gamma_0, \delta$, and η and an unbiased estimate of σ^2, as functions of $\hat{\alpha}_j, \hat{\gamma}_j$, and $\hat{\sigma}_j$, $j = 1, 2$.

3. Consider the data in the table below from a study of automobile accidents in Florida in 1988.

	Ejected?	Fatal incident	Nonfatal incident	Total
Seat belt used	Yes	14	1105	1119
	No	483	411,111	411,594
Total		497	412,216	412,713
Seat belt not used	Yes	497	4624	5121
	No	1008	157,342	158,350
Total		1505	161,966	163,471

A logistic binomial GLM was used to model the likelihood of a fatal incident as a function of whether a seat belt was used and whether the driver was ejected. (note: a 'fatality' is when someone dies, and being 'ejected' means falling or flying out of the car in the accident).

The maximum likelihood values for the coefficients and the variance-covariance matrix are given below:

```
Coefficients: (Intercept)          belt         eject
                -5.0436          -1.7173        2.7978

var-cov:       (Intercept)          belt         eject
(Intercept)  0.0009735536 -0.0009350671 -0.0009322571
belt        -0.0009350671  0.0029176633  0.0008062108
eject       -0.0009322571  0.0008062108  0.0030532124

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:  3568  on 3  degrees of freedom
Residual deviance: 2.854  on 1  degrees of freedom
```

(a) (3 points) Estimate the probability that an automobile accident where the driver did not use a seat belt and was not ejected resulted in a fatality.

(b) (3 points) Estimate a 95% confidence interval on the probability from part (a).

(c) (3 points) How much does wearing a seat belt affect the estimated odds of having a fatal incident, holding other variables constant? Is this effect statistically significantly different from 0?

(d) (3 points) Adam argues that the effect in part (c) represents the causal effect of wearing a seat belt on fatalities. Beth argues that this is only part of the causal effect: people who wear seat belts are also less likely to be ejected from the car, and therefore even less likely to die. Who is right? Support your answer, and, if you agree with Beth, provide a numerical argument based on the data given.

4. (10 points) **Data Structures** A tree is a fundamental data structure for many algorithms. A binary search tree is the simplest tree data structure for searching and sorting. Recall that in the binary search tree, the left subtree of a node contains only nodes with keys lesser than the node's key and the right subtree contains only nodes with keys greater than or equal to the node's key. Suppose that you have a balanced binary search tree where each node key is the value of a scalar data point. The balanced binary tree data structure encodes the dataset $x = [0, 1, 2, 3, 4, 5, 6]$.

   (a) (8 points) Draw a picture of the balanced binary search tree and label each node with the value of the data point at the node.

   (b) (8 points) For this data set, what is the maximum number of nodes one has to visit in order to find a particular data point (key) or certify that the data point value (key) does not exist in the data set? And in general, what is an upper bound on the maximum number of nodes that must be visited for a data set of size n? Describe your reasoning for your answer.

   (c) (3 points) How many nodes must be visited in order to obtain the median of the data set and why?

   (d) (3 points) How many nodes must be visited in order to obtain the average of the data set and why?

5. **Simulation** One way to model a coin-flip experiment is with independent draws from a Bernoulli distribution which we denote by the random variable X. The ith sample from the Bernoulli is X_i, a *heads* outcome is associated with $X_i = 1$, and a *tails* outcome is associated with $X_i = 0$.

   (a) (3 points) **Summary Statistics** The outcome of the experiment can be summarized by the count of the number of heads and tails in a table. For example, a sequence `TTHHTT` can be summarized as

H	T
2	4

   The ability to reduce the full sequence to a table of heads and tails depends on a stated assumption of the experiment. What is the assumption and why does it allow the reduction of the sequence to the summary table?

   (b) (10 points) **Compound Sampling** Suppose that the number of coin-flips is itself a random variable with a Poisson distribution. Write a *function* in `python` or `R` called `coinflip` that returns a sample from a coin-flip experiment whose number of trials is Poisson-distributed. The function should return a list of n items where each item is an outcome of the Bernoulli trial. The python function `numpy.random.poisson(lam)` produces a sample from a Poisson distribution where `lam` is the rate parameter. The R function `rpois(1, lam)` produces a sample from a Poisson distribution where `lam` is the rate parameter. The python function `numpy.random.binomial(n,p)` produces a sample from a Binomial distribution where `n` is the number of trials and `p` is the probability of success. The R function `rbinom(1,n,p)` produces a sample from a Binomial distribution where `n` is the number of trials and `p` is the probability of success. Note that the `binomial` functions will return the number of successes in `n` trials. The `coinflip` function should take `lam` and `p` as parameters.