UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
Basic Exam - Applied Statistics
Monday, August 26, 2019

Work all problems. 60 points are needed to pass at the Masters Level and 75 to pass at the Ph.D. level. Each part is worth 5 points, except the last, which is worth 10.

1. Consider data $\{\mathbf{X}, \mathbf{Y}\}$, where $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$. We know that $E(Y_i) = \beta_0 + \beta_1 X_i, \quad Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \ \ \forall i$. Suppose that $\mathbf{Y}$ has multivariate normal distribution with mean vector $E(\mathbf{Y})$, and covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma^2(1+k) & \rho & 0 & \ldots & 0 \\ \rho & \sigma^2(1+2k) & \rho & \ldots & 0 \\ 0 & \rho & \sigma^2(1+3k) & \ldots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \ldots & \sigma^2(1+nk) \end{pmatrix},$$

such that Σ_{ij} is the covariance between ϵ_i and ϵ_j. We wish to model Y as a function of X in order to estimate β_0 and β_1.

(a) Which values of ρ and k would result in a simple linear regression model?

(b) Suppose you know $\rho = 0$ and you wish to evaluate whether a simple linear regression model is appropriate. Describe one diagnostic you would conduct to determine this.

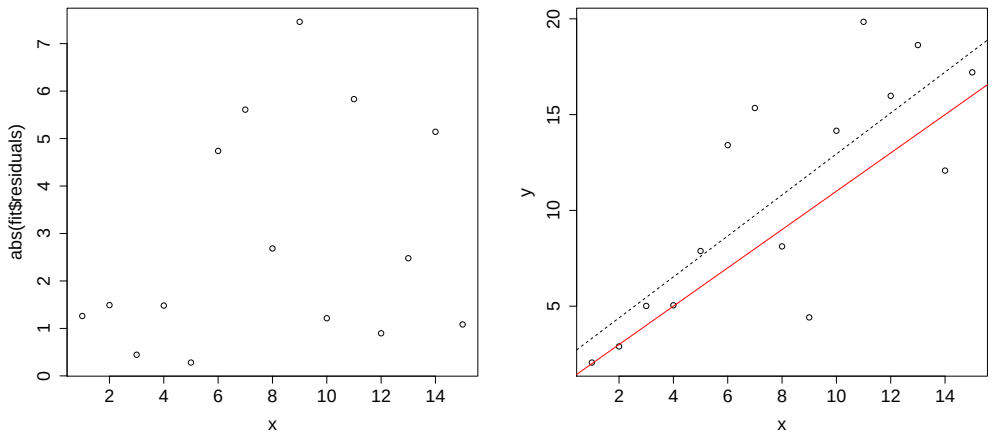
(c) Suppose your analysis in the previous part reveals that a simple linear regression model is not appropriate. Describe how you would model the data.

(d) Suppose $k = 0$ and you suspect $\rho \neq 0$. Describe one diagnostic you would conduct to determine this.

The following plots are generated from the above multivariate normal distribution with $k = 2$, $\rho = 0$, and $\mathbf{X} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$.

(e) Consider the residual plot. What evidence do you see about the value of k?

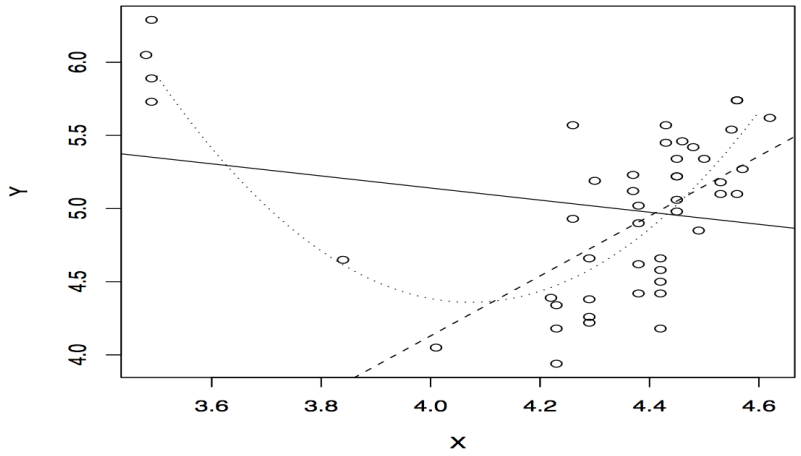
(f) The second plot includes two regression lines: the true regression line, and the fitted regression line. Which line is which, and how can you tell?



2. The figure below shows a plot of

$$Y = \log(\text{Light Intensity}) \text{ verses } X = \log(\text{Surface Temperature}),$$

based on measurements for 47 stars in a certain star cluster. The goal of the study is to characterize the relationship between light intensity and surface temperature.



The solid line shows the OLS fit to a linear model of the form $E(Y|X) = \beta_0 + \beta_1 X$ while the curve shows the OLS fit to a quadratic model of the form $E(Y|X) = \beta_0 + \beta_1 X + \gamma X^2$. The dashed line corresponds to an OLS fit of the same linear model as the solid line, but without the four data points in the upper left hand corner of the figure (i.e. without points for which $X < 3.6$).

(a) Consider the solid line and the curve. Suppose that you wish to test whether a linear model is adequate, or whether a quadratic model is more appropriate. State an appropriate pair of null and alternative hypotheses for this problem. Provide an expression for an appropriate F-statistic, and state the distribution of that statistic under the assumption of normal (i.e.,Gaussian) errors. (NOTE: Be careful to specify precisely each of the components in your F-statistic.)

(b) Comment on the degree of (i) outlying-ness, (ii) leverage, and (iii) influence of the four points in the upper left hand corner, based on evidence in the two lines. Justify your answer through appropriate description of the the likely values of at least two of the statistics t_i, $h_i i$, and D_i. (That is, the outlier t-test value, the hat-matrix entry, and Cook's distance.)

(c) Based on visual inspection of the plot, comment on the appropriateness of the three models shown. Which would you suggest to the astronomers? What question(s) might you have for the astronomers?

3. Consider the problem of modeling sales of 200 companies as a function of the amount of money spent on advertising on youtube, facebook, and in newspapers. Summary statistics are below:

```
    youtube           facebook       newspaper          sales
Min.   :  0.84   Min.   : 0.00   Min.   :  0.36   Min.   : 1.92
1st Qu.: 89.25   1st Qu.:11.97   1st Qu.: 15.30   1st Qu.:12.45
Median :179.70   Median :27.48   Median : 30.90   Median :15.48
Mean   :176.45   Mean   :27.92   Mean   : 36.66   Mean   :16.83
3rd Qu.:262.59   3rd Qu.:43.83   3rd Qu.: 54.12   3rd Qu.:20.88
Max.   :355.68   Max.   :59.52   Max.   :136.80   Max.   :32.40
```

```
> round(cor(marketing),2)
           youtube facebook newspaper sales
youtube       1.00     0.05      0.06  0.78
facebook      0.05     1.00      0.35  0.58
newspaper     0.06     0.35      1.00  0.23
sales         0.78     0.58      0.23  1.00
```

You fit two models:

- Model 1: Main effects only,
- Model 2: All main effects and all interactions (including the three-way interaction) between the three predictors.

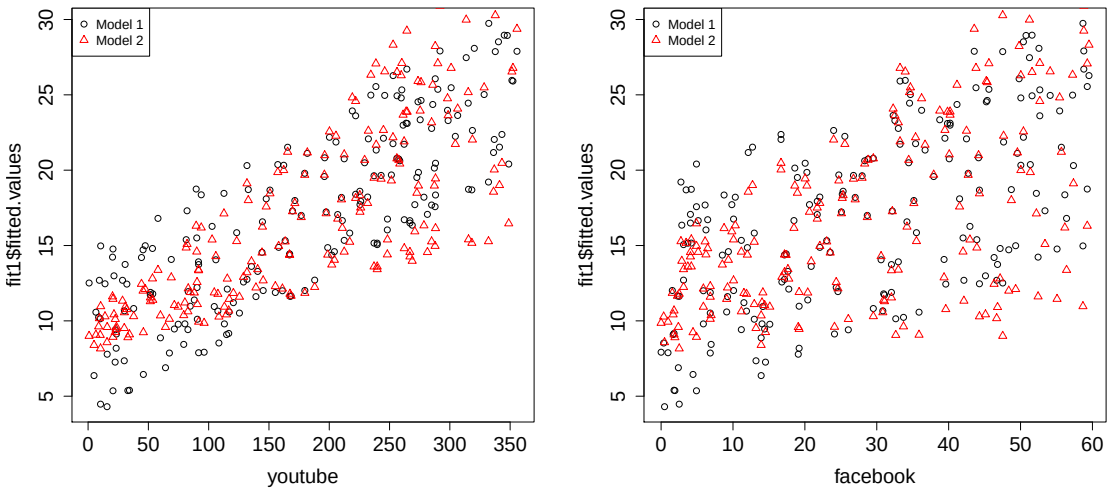
the fits are:

```
Call:
lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5932  -1.0690   0.2902   1.4272   3.3951

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.526667   0.374290   9.422   <2e-16 ***
youtube      0.045765   0.001395  32.809   <2e-16 ***
facebook     0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.023 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = sales ~ youtube * facebook * newspaper, data = marketing)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0746 -0.4660  0.2326  0.7039  1.8288

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 7.867e+00  5.586e-01  14.083  < 2e-16 ***
youtube                     1.971e-02  2.719e-03   7.250 9.95e-12 ***
facebook                    1.962e-02  1.639e-02   1.197    0.233
newspaper                   1.311e-02  1.721e-02   0.761    0.447
youtube:facebook            9.679e-04  8.128e-05  11.909  < 2e-16 ***
youtube:newspaper          -4.621e-05  7.772e-05  -0.595    0.553
facebook:newspaper          7.553e-06  4.026e-04   0.019    0.985
youtube:facebook:newspaper -5.285e-07  1.875e-06  -0.282    0.778
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.129 on 192 degrees of freedom
Multiple R-squared:  0.9686,    Adjusted R-squared:  0.9675
F-statistic: 847.3 on 7 and 192 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Model 1: sales ~ youtube + facebook + newspaper
Model 2: sales ~ youtube * facebook * newspaper
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    196 801.83
2    192 244.60  4    557.23 109.35 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

You may also be interested in plots of the fitted values from the two fits, against two of the predictors:



(a) Consider Model 1. The coefficient of facebook is much larger than the coefficient of youtube. Does this mean facebook is more strongly related to sales than youtube?

(b) The term for facebook in the first model is highly statistically significant, but is not significant in the second model. How do you interpret this change? Should the facebook term be removed from the model?

(c) Are there other analyses or plots you would want to see before evaluating the model fits? If so, describe them.

(d) What do you conclude about the relative importance of the different advertising venues? State any additional assumptions you need to reach these conclusions.

4. **Data Structures** A tree is a fundamental data structure for many algorithms. A binary search tree is a simple 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. The dataset for this problem is $x = [0, 1, 2, 3, 4, 5, 6]$.

(a) Suppose that you have a binary search tree where each node key is the value of a scalar data point. Draw a picture of the binary search tree and label each node with

the value of the data point at the node. Recall that this is *not* a *balanced* binary search tree for this part.

(b) For this data set and an arbitrary binary search tree for these data, 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? 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) Suppose now that the binary search tree is a balanced binary search tree. What is the upper bound on the maximum number of nodes that must be visited in order to find a particular data point (key) or certify that the data point value (key) does not exist for a data set of size n? Describe your reasoning.

(d) How many nodes in a balanced binary search tree must be visited in order to obtain the average of the data set? 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) **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 $\begin{matrix} \text{H} & \text{T} \\ 2 & 4 \end{matrix}$. 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 probability of heads is itself a random variable with a Beta distribution. Write a *function* in `python` or `R` called `coinflip` that returns a sample from a coin-flip experiment where the probability of heads is first drawn from a Beta distribution and then the coin is flipped—the Beta draw is done *for each flip of the coin*. The function should return a list of n items where each item is an outcome of the Bernoulli trial. The python function `numpy.random.beta(a,b)` produces a sample from a Beta distribution where `a` and `b` are the parameters of the beta distribution. The R function `rbeta(1, a, b)` produces a sample from a Beta distribution where `a` and `b` are the parameters of the beta distribution. 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 `a`, `b`, and `n` as parameters.