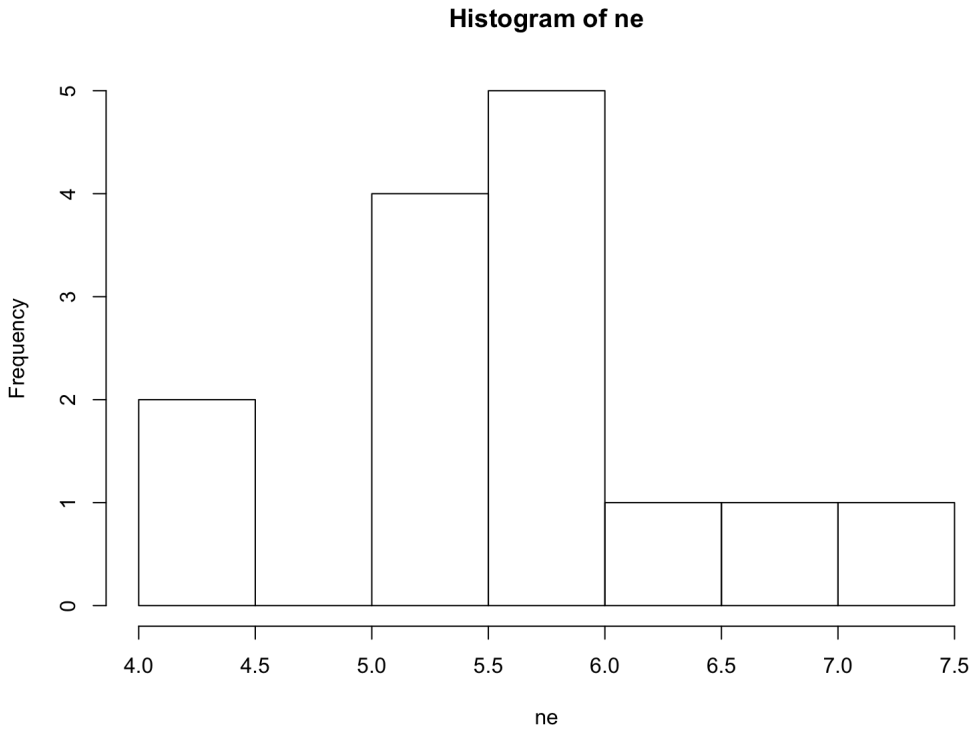
UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
Basic Exam - Applied Statistics
Monday, January 13, 2020

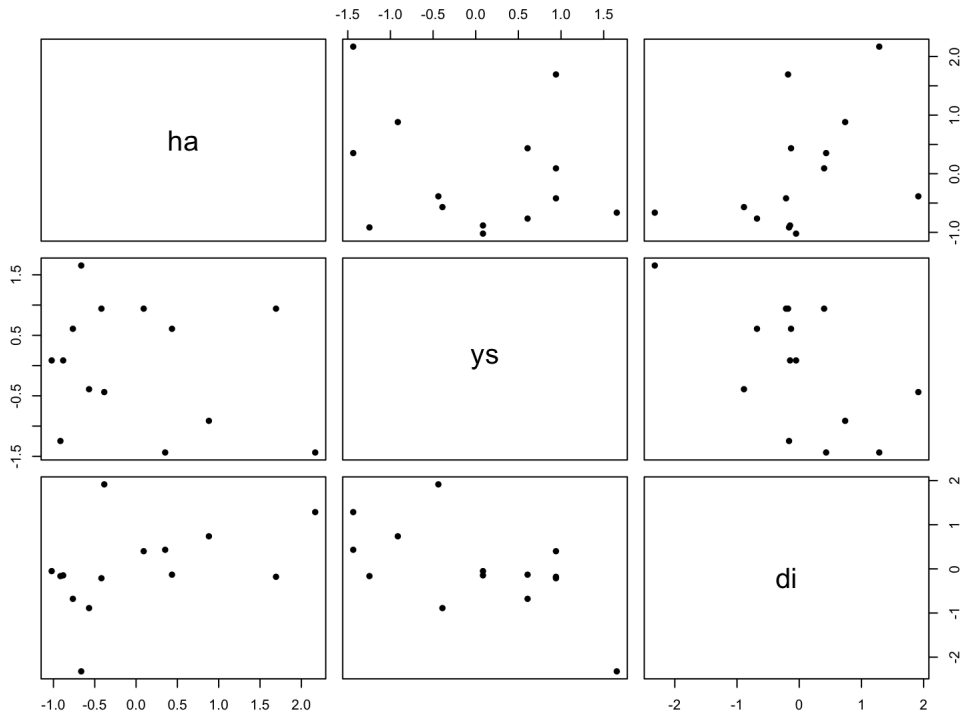
Work all problems. 60 points are needed to pass at the Masters Level and 75 to pass at the Ph.D. level. Each part is worth 5 points, except the last, which is worth 10.

1. Consider the problem of modeling estimates of the effective number of alleles in 14 distinct populations of mice in 14 different New York City parks (`ne`) as a function of the area of the mice habitat in each park (`ha`), number of years since each park's founding (`ys`), and distance from Central Park (`di`). Throughout this problem, consider a (possibly generalized) linear model with identity link for the response `ne` and main effects for the covariates:

$$\mu_i = \mathbb{E}\left[\texttt{ne}_i\right] = \beta_0 + \beta_{\texttt{ha}}\texttt{ha}_i + \beta_{\texttt{ys}}\texttt{ys}_i + \beta_{\texttt{di}}\texttt{di}_i. \tag{1}$$

A histogram of the response `ne` and pairs plots for the covariates `ha`, `ys`, and `di` are printed below. The covariates have been centered and standardized to have zero mean and unit variance.



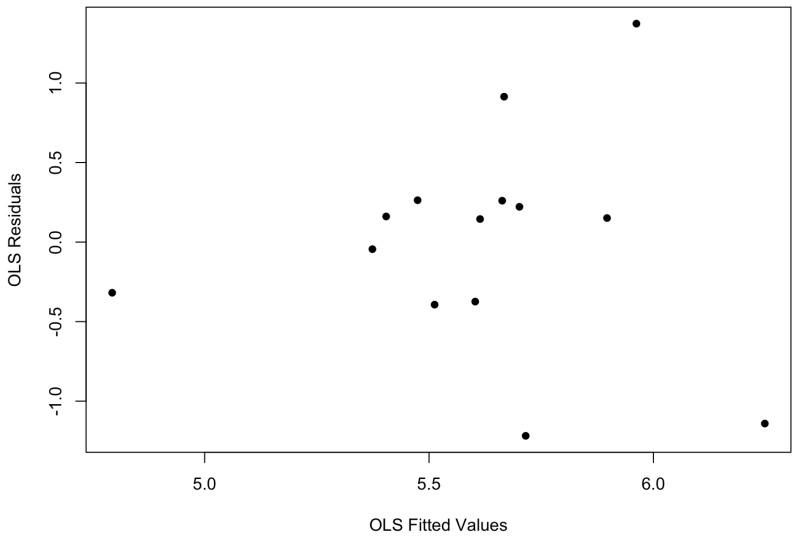


(a) Consider a new quantity $\alpha_{\texttt{ha,di}}$ equal to the expected change in the response `ne` given a simultaneous increase of one unit in `ha` and `di`. Give an expression for $\alpha_{\texttt{ha,di}}$ in terms of the the regression coefficients $\beta_{\texttt{ha}}$, $\beta_{\texttt{ys}}$, and $\beta_{\texttt{di}}$. Based on the pairs plot of the covariates, explain why $\alpha_{\texttt{ha,di}}$ is a more practically meaningful quantity than both $\beta_{\texttt{ha}}$ and $\beta_{\texttt{di}}$.

(b) Given that the response `ne` represents a kind of count and is strictly positive, we might consider a Poisson, gamma, or Gaussian generalized linear model with identity link:

- $\texttt{ne}_i \overset{indep.}{\sim} \text{Poisson}\left(\mu_i\right)$;
- $\texttt{ne}_i \overset{indep.}{\sim} \text{Gamma}\left(\text{shape} = \alpha, \text{rate} = \alpha\mu_i^{-1}\right)$;
- $\texttt{ne}_i \overset{indep.}{\sim} \text{Gaussian}\left(\mu_i, \sigma^2\right)$.

- Write down the variance of the response $\mathbb{V}\left[\texttt{ne}_i\right]$ as a function of the mean of the response $\mathbb{E}\left[\texttt{ne}_i\right]$ and any additional parameters under each of these models. (hint: Recall that the Gamma distribution has two common parameterizations. For one, $E(Y) = \alpha\beta, Var(Y) = \alpha\beta^2$, for the other, $E(Y) = \alpha/\beta, Var(Y) = \alpha/\beta^2$.)
- Based the fitted values and residuals from the ordinary least squares fit to (1) given below, which model is most appropriate for this data? The ordinary least squares fit is the fit that minimizes $\sum_{i=1}^{14}\left(\texttt{ne}_i - \beta_0 - \beta_{\texttt{ha}}\texttt{ha}_i - \beta_{\texttt{ys}}\texttt{ys}_i - \beta_{\texttt{di}}\texttt{di}_i\right)^2$ with respect to β_0, $\beta_{\texttt{ha}}$, $\beta_{\texttt{ys}}$ and $\beta_{\texttt{di}}$.



(c) Parameter estimates and estimated standard errors for each model are given below. Comment on the parameter estimates across the models. Do they differ much? Why or why not? Comment on the standard errors across the models. Are there any overall trends across models? If yes, explain these trends with reference to the mean-variance relationships described above.

	Poisson	Gaussian	Gamma
(Intercept)	5.6164313	5.6164313	5.6179906
ha	0.2167146	0.2190645	0.2122274
ys	-0.1130173	-0.1050783	-0.1174902
di	-0.4415293	-0.4096010	-0.4760347

Estimated standard errors for each model:

	Poisson	Gaussian	Gamma
(Intercept)	0.6333827	0.2097937	0.2011483
ha	0.7234161	0.2405127	0.2275322
ys	0.8176810	0.2716797	0.2582098
di	0.8573525	0.2915651	0.2628532

(d) In practice, we often use a specific link function given an assumed distribution for the data, called a canonical link function. The identity link function is neither the canonical link function for a Poisson generalized linear model, nor the canonical link function for a gamma generalized linear model. The canonical link function for the Poisson distribution is the log-link function, which corresponds to the mean model

$$\log\left(\mathbb{E}\left[\texttt{ne}_i\right]\right) = \beta_0 + \beta_{\texttt{ha}}\texttt{ha}_i + \beta_{\texttt{ys}}\texttt{ys}_i + \beta_{\texttt{di}}\texttt{di}_i.$$

What makes the identity link function a strange choice if we assume a Poisson generalized linear model, compared to a log link? Does it make sense to compare estimates of $\beta_{\texttt{ha}}$ across different models if we use the canonical link functions for each model?

(e) Based on the model fits, what do you conclude about the relationship between the effective number of alleles in a park's mouse population and the park's habitat area, years since founding, and distance from central park?

2. In a study of coins, W. Stanley Jevons weighed 274 gold sovereigns that he collected from circulation in Manchester, England. From each coin, he recorded the weight after cleaning to the nearest .001 gram, and the date of issue. The table below lists the number of coins and the average weight for each age class. The age classes are coded 1 to 5, roughly corresponding to the age of the coin in decades. The standard weight of a gold sovereign was supposed to be 7.9876 grams; the minimum legal weight was 7.9379 grams.

Age, x Decades	Sample Size=n	Average Weight= $\bar{y}$	SD
1	123	7.9725	0.01409
2	78	7.9503	0.02272
3	32	7.9276	0.03426
4	17	7.8962	0.04057
5	24	7.8730	0.05353

(a) Draw a scatter plot of $\bar{y}$ vs x, and comment on the applicability of the usual assumptions of the linear regression model. Also draw a scatter plot of the SDs verses x, and summarize the information in this plot.

(b) Since the numbers of coins n in each age class are all fairly large, it is reasonable to pretend that the variance of coin weight at x is well approximated by SD^2, and hence $var(\bar{y})$ is given by SD^2/n. Consider a linear regression of $\bar{y}$ on x. Write numerical expressions for the weights you would use for each of the 5 points in this regression (you need not evaluate them). Explain, either mathematically or with `R` code, how you would fit this weighted linear model.

The output of such a weighted linear model fit is given here:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.9965218  0.0013220    6049 9.96e-12 ***
x           -0.0237562  0.0008797     -27 0.000111 ***
---
Residual standard error: 0.5549 on 3 degrees of freedom
Multiple R-squared:  0.9959,Adjusted R-squared:  0.9945
F-statistic: 729.2 on 1 and 3 DF,  p-value: 0.0001114
```

(c) Is the fitted regression consistent with the known standard weight for a new coin? Explain.

(d) Explain, in equations or in `R` code, how you would estimate the probability that a new coin of age $x_{new} = 3$ would weigh less than the legal minimum. For this part, you may assume that the SDs given are accurate estimates of the true standard deviations of coins of that age.

(e) Set up the equation(s) and explain how you would solve to determine the age at which, according to this model, the predicted weight of coins is equal to the legal minimum.

3. Consider the simple linear regression model, $Y_i \sim N(\mu_i, \sigma^2)$, $\mu_i = \beta_0 + \beta_1 x_i$.
   (a) Suppose each x_i is replaced with $x_i^* = cx_i$, for constant $c \neq 0$. How are $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma^2}, R^2$, and the t-test of $H_0 : \beta_1 = 0$ affected?
   (b) Now suppose each y_i is replaced with $y_i' = dy_i$, for constant $d \neq 0$. How are $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma^2}, R^2$, and the t-test of $H_0 : \beta_1 = 0$ affected?

4. In many applications, there can be multiple response variables, in which case the linear regression model can be written as
$$Y = XB + E,$$
where $Y = (y_{ij})1 \leq i \leq n,\ 1 \leq j \leq d$ is a matrix in $\mathbb{R}^{n \times d}$, $X = (x_{ij})1 \leq i \leq n,\ 1 \leq j \leq p$ is a matrix in $\mathbb{R}^{n \times p}$, $B = (\beta_{ij}), 1 \leq i \leq p,\ 1 \leq j \leq d$ is a matrix in $\mathbb{R}^{p \times d}$ and $E = (e_{ij})1 \leq i \leq n,\ 1 \leq j \leq d$ is a matrix in $\mathbb{R}^{n \times d}$. Assume that the errors $e_{ij}, 1 \leq i \leq n,\ 1 \leq i \leq d$, are independent normal random variables with mean zero and variance σ^2. In this case, the least squares estimator is defined as
$$\hat{B} = argmin_{B \in \mathbb{R}^{p \times d}} ||Y - XB||^2,$$
where for a matrix $A = (a_{ij})$, the norm is defined as $||A|| = (\Sigma_{i,j} a_{ij}^2)^{1/2}$. Suppose you only observe the response matrix Y and the design matrix X.
   (a) Find the least squares estimate $\hat{B}$.
   (b) Find an unbiased estimate of σ^2: $\hat{\sigma^2}$.
   (c) Construct a test for the null hypothesis that the first column of X is an irrelevant variable.

5. Under complex genetic models, it is sometimes necessary to assess the probability of a given DNA sequence by simulation. Suppose that you want to compute the probability of the DNA sequence `ACTGACTG`. For simplicity here, we use a simple model, and treat it as though probabilities cannot be calculated directly: Suppose each of the 8 base pairs is chosen independently, uniformly at random from $\mathcal{B} = \{A, C, T, G\}$.
   (a) Describe in words or with pseudo-code how you might simulate 1×10^6 random DNA sequences of length 8 and keep track of how many times each unique sequence was generated.
   (b) If you only wanted to use these simulation results to estimate the probability of the given DNA sequence `ACTGACTG`, what data structure would you (or did you) choose to store the simulation results and why you chose that data structure.
   (c) Suppose now that instead of computing the probability of a given sequence, you wanted to use your simulation results to keep track of the most common sequence so far as the simulation proceeds (where there are 1 million steps for generating 1 million sequences). What data structure would you choose and why?

6. Cross-validation is a common tool for model selection based on prediction accuracy. Consider a linear regression model $y \sim \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where ϵ is a Gaussian random variable. You are interested in minimizing the squared error of prediction. You have $n = 1000$ (x, y) measurement pairs. Describe in words how you would use cross-validation to test whether you should include an additional term $\beta_3 x_1 x_2$ in the model and what the meaning of a lower cross-validation error means in this context.