UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
Basic Exam - Applied Statistics
January, 2018

Work all problems. 60 points needed to pass at the Masters level, 75 to pass at the PhD level.

1. Data from Kelley's Blue Book can be used to model the price of a used car. This problem uses the predictors Make and Mileage and ignores other predictors. Altogether there are 804 lines of data. A few of them are:

| Price | Mileage | Make | MakeName |
|---|---|---|---|
| 17314 | 8221 | 1 | Buick |
| 17542 | 9135 | 1 | Buick |
| 51154 | 2202 | 2 | Cadillac |
| 11096 | 20334 | 3 | Chevrolet |

The codes for Make are 1–Buick, 2–Cadillac, 3–Chevrolet, 4–Pontiac, 5–Saab, and 6–Saturn. This problem ignores makes other than these six. Fitting a model in R yielded the following result.

```
> summary ( lm ( Price $\sim$ Mileage + Make, data=cars ) )

Call:
lm(formula = Price ~ Mileage + Make, data = cars)

Residuals:
   Min     1Q Median     3Q    Max
-14194  -7008  -3753   6563  44852

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.790e+04  1.237e+03  22.549  < 2e-16 ***
Mileage     -1.681e-01  4.184e-02  -4.018 6.42e-05 ***
Make        -9.488e+02  2.579e+02  -3.680 0.000249 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9714 on 801 degrees of freedom
Multiple R-squared:  0.03675,Adjusted R-squared:  0.03434
F-statistic: 15.28 on 2 and 801 DF,  p-value: 3.079e-07
```

(a) **5 pts** What went wrong and how would you attempt to fix it?

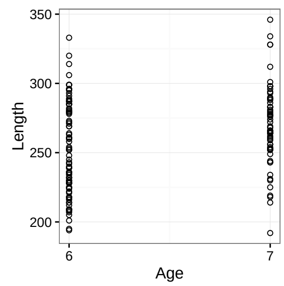
(b) **5 pts** Sketch a scatterplot with Mileage on the x-axis and Price on the y-axis that indicates the presence of an interaction between Make and Mileage. You needn't show all 804 data points, just enough to illustrate an interaction. Explain how your plot shows an interaction.

(c) **5 pts** How would you tell R to fit a model with a Make by Mileage interaction?

(d) **5 pts** We might treat these six makes as a sample of size six from the population of all makes. And we might think that each make in the population has its own linear relationship between Price and Mileage. What kind of model could we use to learn, from these six makes, about the population of linear relationships? Be as specific as you can.

(e) **5 pts** How would you tell R to fit such a model?

2. The scatterplot below shows the Age and Length of smallmouth bass — a fish species — caught in West Bearskin Lake in Minnesota. This problem uses just the six and seven year-old fish; the full data set contains younger and older fish too.




A linear model `lm ( Length ~ Age )` yields

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  148.971     34.101   4.369 2.36e-05 ***
Age           17.271      5.303   3.257   0.0014 **
```

(a) **5 pts** What is the formula for the $t$ statistic for Age?

(b) **10 pts** The $t$ statistic for Age is large and the p-value is small, indicating that Age is a useful predictor of Length. Yet there is lots of overlap between the Lengths of six year-old and seven year-old fish. How can that be?

3. Sometimes we know that a linear regression line goes through the origin so we would adopt a model

$$Y_i = \beta X_i + \epsilon_i.$$

   (a) **10 pts** Under the usual assumptions, derive the least-squares estimator of $\beta$.
   (b) **15 pts** Derive the variance of the estimator.

4. In nondestructive testing of aluminum blocks, an electromagnetic probe is used to detect flaws below the surface. The sensitivity $Y$ of the probe is known to be related to the thickness $X$ of the wire used to construct the coil in the probe. An investigator interested in understanding this relationship collected a random sample of measurements of thickness and sensitivity. For the thickness values considered in the experiment, suppose the assumptions for the nonlinear regression hold with the regression function $Y$ on $X$ given by

$$E(Y|X=x) = \mu_Y(x) = \beta_1(1 - e^{-e^{(\beta_2+\beta_3 x)}}).$$

Use the following R output, answer the following questions:

(a) **10 pts** What are the least squares estimates of the parameters $\beta_1$, $\beta_2$, $\beta_3$ and $\sigma$? How are the corresponding estimators defined?

(b) **10 pts** If $\beta_1$ and $\beta_2$ are known to be positive, then show that no matter how thin the wire used, the average sensitivity can never exceed $\beta_1(1 - e^{-e^{-\beta_2}})$. Denote this upper bound for the sensitivity by $\theta$. Estimate $\theta$.

(c) **15 pts** Obtain one-at-a-time 95% two-sided approximate confidence intervals for $\beta_1$ and $\beta_2$. Using these estimates, obtain an approximate confidence interval for $\theta$ with confidence coefficient greater than or equal to 0.90 (use the Bonferroni inequality which states that $P(A \cap B) \geq 1 - P(A^c) - P(B^c)$).

```
Formula: Sensitivity ~ beta_1 * (1 - exp(-exp(-(beta_2 + beta_3 * Thickness))))

Parameters:
  Estimate Std. Error t value Pr(>|t|)
beta_1   1.9484     0.4725   4.124 0.001199 **
beta_2  -1.2699     0.6809  -1.865 0.084902 .
beta_3  14.3630     2.7038   5.312 0.000141 ***
  ---
  Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1025 on 13 degrees of freedom

Number of iterations to convergence: 6
Achieved convergence tolerance: 8.168e-06
```