STA 2101/442 Assignment 1 (Review)[1]

The questions on this assignment are not to be handed in. They are practice for the midterm test and final exam.

1. In a political poll, a random sample of n registered voters are to indicate which of two candidates they prefer. State a reasonable model for these data, in which the population proportion of registered voters favouring Candidate A is denoted by θ. Denote the observations $Y_1, \ldots, Y_n$.

2. A medical researcher conducts a study using twenty-seven litters of cancer-prone mice. Two members are randomly selected from each litter, and all mice are subjected to daily doses of cigarette smoke. For each pair of mice, one is randomly assigned to Drug A and one to Drug B. Time (in weeks) until the first clinical sign of cancer is recorded.

   (a) State a reasonable model for these data. Remember, a statistical model is a set of assertions that partly specify the probability distribution of the observable data. For simplicity, you may assume that the study continues until all the mice get cancer, and that log time until cancer has a normal distribution.

   (b) What is the parameter space for your model?

3. Suppose that volunteer patients undergoing elective surgery at a large hospital are randomly assigned to one of three different pain killing drugs, and one week after surgery they rate the amount of pain they have experienced on a scale from zero (no pain) to 100 (extreme pain).

   (a) State a reasonable model for these data. For simplicity, you may assume normality.

   (b) What is the parameter space?

4. Let $X_1, \ldots, X_n$ be a random sample (meaning independent and identically distributed) from a distribution with density $f(x) = \frac{\theta}{x^{\theta+1}}$ for $x > 1$, where $\theta > 0$.

   (a) Find the maximum likelihood estimator of θ. Show your work. The answer is a formula involving $X_1, \ldots, X_n$.

   (b) Suppose you observe these data: `1.37, 2.89, 1.52, 1.77, 1.04, 2.71, 1.19, 1.13, 15.66, 1.43`. Calculate the maximum likelihood estimate. My answer is 1.469102.

5. Label each statement below True or False. Write "T" or "F" beside each statement. Assume the $\alpha = 0.05$ significance level.

   (a) _____ The p-value is the probability that the null hypothesis is true.

   (b) _____ The p-value is the probability that the null hypothesis is false.

   (c) _____ In a study comparing a new drug to the current standard treatment, the null hypothesis is rejected. We conclude that the new drug is ineffective.

[1] This assignment was prepared by Jerry Brunner, Department of Statistics, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/~brunner/oldclass/appliedf18

(d) _____ If $p > .05$ we reject the null hypothesis at the .05 level.

(e) _____ If $p < .05$ we reject the null hypothesis at the .05 level.

(f) _____ The greater the p-value, the stronger the evidence against the null hypothesis.

(g) _____ In a study comparing a new drug to the current standard treatment, $p > .05$. We conclude that the new drug and the existing treatment are not equally effective.

(h) _____ The 95% confidence interval for β_3 is from -0.26 to 3.12. This means $P\{-0.26 < \beta_3 < 3.12\} = 0.95$.

6. Let $Y_1, \ldots, Y_n$ be a random sample from a normal distribution with mean μ and variance σ^2, so that $T = \frac{\sqrt{n}(\overline{Y}-\mu)}{S} \sim t(n-1)$. This is something you don't need to prove, for now.

   (a) Derive a $(1-\alpha)100\%$ confidence interval for μ. "Derive" means show all the high school algebra. Use the symbol $t_{\alpha/2}$ for the number satisfying $Pr(T > t_{\alpha/2}) = \alpha/2$.

   (b) A random sample with $n = 23$ yields $\overline{Y} = 2.57$ and a sample variance of $S^2 = 5.85$. Using the critical value $t_{0.025} = 2.07$, give a 95% confidence interval for μ. The answer is a pair of numbers.

   (c) Test $H_0 : \mu = 3$ at $\alpha = 0.05$.

      i. Give the value of the T statistic. The answer is a number.

      ii. State whether you reject H_0, Yes or No.

      iii. Can you conclude that μ is different from 3? Answer Yes or No.

      iv. If the answer is Yes, state whether $\mu > 3$ or $\mu < 3$. Pick one.

   (d) Show that using a t-test, $H_0 : \mu = \mu_0$ is rejected at significance level α if and only the $(1-\alpha)100\%$ confidence interval for μ does not include μ_0. The problem is easier if you start by writing the set of T values for which H_0 is *not* rejected.

   (e) In Question 6b, does this mean $Pr\{1.53 < \mu < 3.61\} = 0.95$? Answer Yes or No and briefly explain.

7. Let $Y_1, \ldots, Y_n$ be a random sample from a distribution with mean μ and standard deviation σ.

   (a) Show that the sample variance $S^2 = \frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{n-1}$ is an unbiased estimator of σ^2.

   (b) Denote the sample standard deviation by $S = \sqrt{S^2}$. Assume that the data come from a continuous distribution, so that $Var(S) > 0$. Using this fact, show that S is a *biased* estimator of σ.

8. In the *centered* linear regression model, sample means are subtracted from the explanatory variables, so that values above average are positive and values below average are negative. Here is a version with one explanatory variable. For $i = 1, \ldots, n$, let $y_i = \beta_0 + \beta_1(x_i - \overline{x}) + \epsilon_i$, where

   β_0 and β_1 are unknown constants (parameters).

   x_i are known, observed constants.

   $\epsilon_1, \ldots, \epsilon_n$ are unobservable random variables with $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$ and $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.

   σ^2 is an unknown constant (parameter).

   $y_1, \ldots, y_n$ are observable random variables.

   (a) What is $E(y_i)$? $Var(y_i)$?

   (b) Prove that $Cov(y_i, y_j) = 0$. Use the definition
   $Cov(U, V) = E\{(U - E(U))(V - E(V))\}$.

   (c) If ϵ_i and ϵ_j are independent (not just uncorrelated), then so are y_i and y_j, because functions of independent random variables are independent. Proving this in full generality requires advanced definitions, but in this case the functions are so simple that we can get away with an elementary definition. Let X_1 and X_2 be independent random variables, meaning $P\{X_1 \leq x_1, X_2 \leq x_2\} = P\{X_1 \leq x_1\}P\{X_2 \leq x_2\}$ for all real x_1 and x_2. Let $Y_1 = X_1 + a$ and $Y_2 = X_2 + b$, where a and b are constants. Prove that Y_1 and Y_2 are independent.

   (d) In *least squares estimation*, we observe random variables $y_1, \ldots, y_n$ whose distributions depend on a parameter θ, which could be a vector. To estimate θ, write the expected value of y_i as a function of θ, say $E_\theta(y_i)$, and then estimate θ by the value that gets the observed data values as close as possible to their expected values. To do this, minimize

   $$Q = \sum_{i=1}^{n} \left(y_i - E_\theta(y_i)\right)^2.$$

   The value of θ that makes Q as small as possible is the least squares estimate.

   Using this framework, find the least squares estimates of β_0 and β_1 for the centered regression model. The answer is a pair of formulas. Show your work.

   (e) Because of the centering, it is possible to verify that the solution actually *minimizes* the sum of squares Q, using only single-variable second derivative tests. Do this part too.

   (f) How about a least squares estimate of σ^2?

   (g) You know that the least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ must be unbiased, but show it by calculating their expected values for this particular case.

   (h) Calculate $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for the following data. Your answer is a pair of numbers.

x	8	7	7	9	4
y	9	13	9	8	6

   I get $\widehat{\beta}_1 = \frac{1}{2}$.

(i) Going back to the general setting (not just the numerical example with $n = 5$), suppose the ϵ_i are normally distributed.

   i. What is the distribution of y_i?

   ii. Write the log likelihood function.

   iii. Obtain the maximum likelihood estimates of β_0 and β_1; don't bother with σ^2. The answer is a pair of formulas. *Don't do more work than you have to!* As soon as you realize that you have already solved this problem, stop and write down the answer.

(j) Still for this centered model with a single explanatory variable, suppose we centered the y_i values too. In this case what is the least squares estimate of β_0? Show your work.

9. Consider the centered *multiple* regression model

$$y_i = \beta_0 + \beta_1(x_{i,1} - \overline{x}_1) + \cdots + \beta_{p-1}(x_{i,p-1} - \overline{x}_{p-1}) + \epsilon_i$$

with the usual details.

   (a) What is $E_{\boldsymbol{\beta}}(y_i)$?

   (b) What is the least squares estimate of β_0? Show your work.

   (c) For an ordinary uncentered regression model, what is the height of the least squares plane at the point where all x variables are equal to their sample mean values?

10. Suppose that volunteer patients undergoing elective surgery at a large hospital are randomly assigned to one of three different pain killing drugs, and one week after surgery they rate the amount of pain they have experienced on a scale from zero (no pain) to 100 (extreme pain). Write a multiple regression model for these data; specify how the explanatory variables are defined.

11. Let $\mathbf{A} = \left(\begin{array}{cc} 1 & 2 \\ 2 & 4 \end{array} \right)$ $\mathbf{B} = \left(\begin{array}{cc} 0 & 2 \\ 2 & 1 \end{array} \right)$ $\mathbf{C} = \left(\begin{array}{cc} 2 & 0 \\ 1 & 2 \end{array} \right)$

   (a) Calculate $\mathbf{AB}$ and $\mathbf{AC}$

   (b) Do we have $\mathbf{AB} = \mathbf{AC}$? Answer Yes or No.

   (c) Prove $\mathbf{B} = \mathbf{C}$. Show your work.

12. Let $\mathbf{X}$ be an n by p matrix with $n \neq p$. Why is it incorrect to say that $(\mathbf{X}^\top\mathbf{X})^{-1} = \mathbf{X}^{-1}\mathbf{X}^{\top -1}$?

13. Let $\mathbf{a}$ be an $n \times 1$ matrix of real constants. How do you know $\mathbf{a}^\top\mathbf{a} \geq 0$?

14. The $p \times p$ matrix $\boldsymbol{\Sigma}$ is said to be *positive definite* if $\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a} > 0$ for all $p \times 1$ vectors $\mathbf{a} \neq \mathbf{0}$. Show that the eigenvalues of a positive definite matrix are all strictly positive. Hint: start with the definition of an eigenvalue and the corresponding eigenvalue: $\boldsymbol{\Sigma}\mathbf{v} = \lambda\mathbf{v}$. Eigenvectors are typically scaled to have length one, so you may assume $\mathbf{v}^\top\mathbf{v} = 1$.

15. Recall the *spectral decomposition* of a symmetric matrix (for example, a variance-covariance matrix). Any such matrix $\mathbf{\Sigma}$ can be written as $\mathbf{\Sigma} = \mathbf{P\Lambda P}^\top$, where $\mathbf{P}$ is a matrix whose columns are the (orthonormal) eigenvectors of $\mathbf{\Sigma}$, $\mathbf{\Lambda}$ is a diagonal matrix of the corresponding eigenvalues, and $\mathbf{P}^\top\mathbf{P} = \mathbf{PP}^\top = \mathbf{I}$. If $\mathbf{\Sigma}$ is real, the eigenvalues are real as well.

    (a) Let $\mathbf{\Sigma}$ be a square symmetric matrix with eigenvalues that are all strictly positive.

    i. What is $\mathbf{\Lambda}^{-1}$?

    ii. Show $\mathbf{\Sigma}^{-1} = \mathbf{P\Lambda}^{-1}\mathbf{P}^\top$

    (b) Let $\mathbf{\Sigma}$ be a square symmetric matrix, and this time the eigenvalues are non-negative.

    i. What do you think $\mathbf{\Lambda}^{1/2}$ might be?

    ii. Define $\mathbf{\Sigma}^{1/2}$ as $\mathbf{P\Lambda}^{1/2}\mathbf{P}^\top$. Show $\mathbf{\Sigma}^{1/2}$ is symmetric.

    iii. Show $\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2} = \mathbf{\Sigma}$, justifying the notation.

    (c) Now return to the situation where the eigenvalues of the square symmetric matrix $\mathbf{\Sigma}$ are all strictly positive. Define $\mathbf{\Sigma}^{-1/2}$ as $\mathbf{P\Lambda}^{-1/2}\mathbf{P}^\top$, where the elements of the diagonal matrix $\mathbf{\Lambda}^{-1/2}$ are the reciprocals of the corresponding elements of $\mathbf{\Lambda}^{1/2}$.

    i. Show that the inverse of $\mathbf{\Sigma}^{1/2}$ is $\mathbf{\Sigma}^{-1/2}$, justifying the notation.

    ii. Show $\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{-1/2} = \mathbf{\Sigma}^{-1}$.

    (d) Let $\mathbf{\Sigma}$ be a symmetric, positive definite matrix. How do you know that $\mathbf{\Sigma}^{-1}$ exists?

16. Let $\mathbf{X}$ be an $n \times p$ matrix of constants. The idea is that $\mathbf{X}$ is the "design matrix" in the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, so this problem is really about linear regression.

    (a) Recall the definition of linear independence. The columns of $\mathbf{A}$ are said to be *linearly dependent* if there exists a column vector $\mathbf{v} \neq \mathbf{0}$ with $\mathbf{Av} = \mathbf{0}$. If $\mathbf{Av} = \mathbf{0}$ implies $\mathbf{v} = \mathbf{0}$, the columns of $\mathbf{A}$ are said to be linearly *independent*. Show that if the columns of $\mathbf{X}$ are linearly independent, then $\mathbf{X}^\top\mathbf{X}$ is positive definite.

    (b) Show that if $\mathbf{X}^\top\mathbf{X}$ is positive definite then $(\mathbf{X}^\top\mathbf{X})^{-1}$ exists.

    (c) Show that if $(\mathbf{X}^\top\mathbf{X})^{-1}$ exists then the columns of $\mathbf{X}$ are linearly independent.

    This is a good problem because it establishes that the least squares estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ exists if and only if the columns of $\mathbf{X}$ are linearly independent.