STA 2101/442 Assignment Eight[1]

The questions are just practice for the quiz, and are not to be handed in. Use R for Question 5, and bring your printout to the quiz. **Your printout should show *all* R input and output, and *only* R input and output**. Do not write anything on your printouts except your name and student number.

1. This question explores the practice of "centering" quantitative explanatory variables in a regression by subtracting off the mean. Geometrically, this should not alter the configuration of data points in the multi-dimensional scatterplot. All it does is shift the axes. Thus, the intercept of the least squares plane should change, but the slopes should not.

   (a) Consider a simple experimental study with an experimental group, a control group and a single quantitative covariate. Independently for $i = 1, \ldots, n$ let

   $$Y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \epsilon_i,$$

   where x_i is the covariate and d_i is an indicator dummy variable for the experimental group. If the covariate is "centered," the model can be written

   $$Y_i = \beta_0^* + \beta_1^*(x_i - \overline{x}) + \beta_2^* d_i + \epsilon_i,$$

   where $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$. The prime just means another β, not the transpose.

      i. Express the β^* quantities in terms of the original β quantities.
      ii. Let's generalize this. For the general linear model in matrix form suppose $\boldsymbol{\beta}^* = \mathbf{A}\boldsymbol{\beta}$, where $\mathbf{A}$ is a square matrix with an inverse. This makes $\boldsymbol{\beta}^*$ a one-to-one function of $\boldsymbol{\beta}$. Show that $\widehat{\boldsymbol{\beta}}^* = \mathbf{A}\widehat{\boldsymbol{\beta}}$.
      iii. Give the matrix $\mathbf{A}$ for this $p = 3$ model.
      iv. If the data are centered, what is $E(Y|x)$ for the experimental group, and what is $E(Y|x)$ for the control group?

   (b) In the following model, there are $p-1$ quantitative explanatory variables. The un-centered version is

   $$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i,$$

   and the centered version is

   $$Y_i = \beta_0^* + \beta_1^*(x_{i,1} - \overline{x}_1) + \ldots + \beta_{p-1}^*(x_{i,p-1} - \overline{x}_{p-1}) + \epsilon_i,$$

   where $\overline{x}_j = \frac{1}{n}\sum_{i=1}^n x_{i,j}$ for $j = 1, \ldots, p-1$.

      i. What is β_0^* in terms of the β quantities?
      ii. What is β_j^* in terms of the β quantities?
      iii. What is $\widehat{\beta}_0$ in terms of the $\widehat{\beta}^*$ quantities?
      iv. Using $\sum_{i=1}^n \widehat{Y}_i = \sum_{i=1}^n Y_i$, show that $\widehat{\beta}_0^* = \overline{Y}$.

[1] Copyright information is at the end of the last page.

(c) Now consider again the study with an experimental group, a control group and a single covariate. This time the interaction is included.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 x_i d_i + \epsilon_i$$

The centered version is

$$Y_i = \beta_0^* + \beta_1^*(x_i - \overline{x}) + \beta_2^* d_i + \beta_3^*(x_i - \overline{x})d_i + \epsilon_i$$

i. Express the β^* quantities from the centered model in terms of the β quantities from the un-centered model. Is the correspondence one to one?

ii. For the un-centered model, what is the difference between $E(Y|X = \overline{x})$ for the experimental group and $E(Y|X = \overline{x})$ for the control group?

iii. What is the difference between intercepts for the centered model? Compare this to your answer to Question 1(c)ii.

2. One version of the delta method says that if $X_1, \ldots, X_n$ are a random sample from a distribution with mean μ and variance σ^2, and $g(x)$ is a function whose derivative is continuous in a neighbourhood of $x = \mu$, then $\sqrt{n}\left(g(\overline{X}_n) - g(\mu)\right) \overset{d}{\to} T \sim N(0, g'(\mu)^2\sigma^2)$. In many applications, both μ and σ^2 are functions of some parameter θ.

(a) Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter θ. Find the limiting distribution of

$$Z_n = 2\sqrt{n}\left(\sin^{-1}\sqrt{\overline{X}_n} - \sin^{-1}\sqrt{\theta}\right).$$

Hint: $\frac{d}{dx}\sin^{-1}(x) = \frac{1}{\sqrt{1-x^2}}$.

(b) In 2 coffee taste test example, suppose 60 out of 100 consumers prefer a new blend of coffee beans. Using your answer to the first part of this question, test the null hypothesis using a variance-stabilized test statistic. Give the value of the test statistic (a number), and state whether you reject H_0 at the usual $\alpha = 0.05$ significance level. I believe you will need a calculator for this one.

(c) Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with parameter θ, so that $E(X_i) = \theta$ and $Var(X_i) = \theta^2$.

i. Find a variance-stabilizing transformation. That is, find a function $g(x)$ such that the limiting distribution of

$$Y_n = \sqrt{n}\left(g(\overline{X}_n) - g(\theta)\right)$$

does not depend on θ.

ii. According to a Poisson process model for calls answered by a service technician, service times (that is, time intervals between taking 2 successive calls; there is always somebody on hold) are independent exponential random variables with mean θ. In 50 successive calls, one technician's mean service time was 3.4 minutes. Test whether this technician's mean service time differs from the mandated average time of 3 minutes. Use your answer to the first part of this question.

(d) Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on $(0, \theta)$.

i. Find a variance-stabilizing transformation. That is, find a function $g(x)$ such that the limiting distribution of

$$Y_n = \sqrt{n}\left(g(2\overline{X}_n) - g(\theta)\right)$$

does not depend on θ.

ii. To check, find the limiting distribution of Y_n.

(e) The label on the peanut butter jar says peanuts, partially hydrogenated peanut oil, salt and sugar. But we all know there is other stuff in there too. There is very good reason to assume that the number of rat hairs in a 500g jar of peanut butter has a Poisson distribution with mean λ, because it's easy to justify a Poisson process model for how the hairs get into the jars. A sample of 30 jars of Brand A yields $\overline{X} = 6.8$, while an independent sample of 40 jars of Brand B yields $\overline{Y} = 7.275$.

i. State the model for this problem.

ii. What is the parameter space Θ?

iii. State the null hypothesis in symbols.

iv. Find a variance-stabilizing transformation for the Poisson distribution.

v. Using your variance-stabilizing transformation, derive a test statistic that has an approximate standard normal distribution under H_0. Now square it to get a chi-squared test with one degree of freedom.

vi. Calculate the chi-squared statistic for these data. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

vii. In plain, non-statistical language, what do you conclude? Your answer is something about peanut butter and rat hairs.

3. In the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, let $cov(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$, with $\mathbf{V}$ a *known* symmetric positive definite matrix. Derive the weighted least squares estimate $\widehat{\boldsymbol{\beta}}_{wls} = (\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{y}$.

4. For a very simple aggregated data set, our data are a collection of sample means $\overline{Y}_1, \ldots, \overline{Y}_n$. Data values in the *unaggregated* data set come from a distribution with common mean μ and common variance σ^2. Sample mean i is based on m_i observations, so that (approximately by the Central Limit Theorem), $\overline{Y}_i \sim N(\mu, \frac{\sigma^2}{m_i})$.

(a) One could estimate μ with the arithmetic mean of the sample means. Is this estimator unbiased?

(b) Specialize the weighted least squares estimate for this problem. Is it unbiased?

(c) If you had access t the unaggregated data, how would you estimate μ? What is the connection of this statistic to the weighted least squares estimator?

5. Returning to the Chick Weights study (R data set `chickwts`), we seek to compare the weights of chickens fed `linseed`, `meatmeal` and `soybean`. This time we'll do it by discarding the data for the other feed types.

(a) Carry out an ordinary F-test.

(b) Do the same thing with a randomization test, obtaining a randomization p-value. Is it close to what you got from the F-test?

(c) Of course your randomization p-value is merely an *estimate* of the permutation p-value. Give an approximate 99% confidence interval for the permutation p-value.

Please bring your printout to the quiz. **Your printout should show *all* R input and output, and *only* R input and output**. Do not write anything on your printouts except your name and student number.