STA 2101/442 Assignment 4[1]

Except for Question 12, the questions on this assignment are practice for the quiz on Friday October 6th, and are not to be handed in. Please do the problems using the formula sheet as necessary. A copy of the formula sheet will be distributed with the quiz.

1. It's easy to say "All these dummy variable coding schemes are equivalent," and the statement is correct — but exactly what does it mean? Consider the example of a 3-category explanatory variable with categories labelled A, B and C, and a single quantitative explanatory variable. This can be extended to cover most cases of interest.

   We have seen two ways of setting up the dummy variables; there are plenty more. For indicator dummy variables with intercept, $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$, where x_1 and x_2 are indicators for categories A and B respectively and x_3 is the quantitative variable. For cell means coding, $Y = \alpha_1 w_1 + \alpha_2 w_2 + \alpha_3 w_3 + \alpha_4 x_3 + \epsilon$, where w_1, w_2 and w_3 are indicators for categories A, B and C respectively and x_3 is the same quantitative explanatory variable. This notation is a reminder that when the dummy variable coding changes, the meaning of some parameters will change too.

   (a) For each of the two coding schemes, make a table showing how the dummy variables are set up. There should be one row for each category, and a column for each dummy variable. Add a another column on the right, showing $E(Y|x)$. I know you have done this before, but it will help. Put the two tables side by side.

   (b) Clearly if you know x_1 and x_2, you know w_1, w_2 and w_3 – and vice versa[2]. The same is true of the regression coefficients. Solve for $\alpha_1, \alpha_2, \alpha_3$ and α_4 in terms of $\beta_0, \beta_1, \beta_2$ and β_3. Your answer consists of four equations.

   (c) Note that the equations are linear, and it would be easy to solve for the β parameters in terms of the α parameters. Thus the re-parameterization of the vector $\boldsymbol{\beta}$ into the vector $\boldsymbol{\alpha}$ is a $1-1$ linear transformation. That is, $\boldsymbol{\alpha} = \mathbf{A}\boldsymbol{\beta}$, where the matrix $\mathbf{A}$ has an inverse. Give the 4×4 matrix $\mathbf{A}$. The answer is a matrix of specific numbers (integers).

   (d) For the general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the one-to-one linear re-parameterization $\boldsymbol{\alpha} = \mathbf{A}\boldsymbol{\beta}$ requires a one-to-one linear transformation of the $\mathbf{X}$ matrix in order not to change what the model says:

   $$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \mathbf{X}\mathbf{A}^{-1}\mathbf{A}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \mathbf{W}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \end{aligned}$$

   where $\mathbf{W} = \mathbf{X}\mathbf{A}^{-1}$ and $\boldsymbol{\alpha} = \mathbf{A}\boldsymbol{\beta}$. This is very general, and is not confined to adopting different dummy variable codings. One can start with a linear transformation of $\mathbf{X}$ or with a linear re-parameterization of $\boldsymbol{\beta}$. One requires the other. Finally, here is the question. Give the 4×4 matrix $\mathbf{A}^{-1}$ for our little dummy variable problem. The answer is a matrix of specific numbers. I actually used R's `solve` function to get it. If you do it this way, don't bother to bring that printout.

[2]For the model with an intercept, there is actually another dummy variable x_0 that always equals one. Thus there are 3 dummy variables in each set.

(e) Give equations for w_1, w_2 and w_3 in terms of x_0, x_1 and x_2. Your answer consists of three scalar equations. This confirms that switching dummy variable coding is a linear transformation for this example.

(f) Now we move to a more general setting in which $\boldsymbol{\alpha} = \mathbf{A}\boldsymbol{\beta}$ is just a 1-1 linear re-parameterization and $\mathbf{W} = \mathbf{X}\mathbf{A}^{-1}$ is the corresponding transformation of the explanatory variables.

(g) Write the least-squares estimate $\widehat{\boldsymbol{\alpha}}$ in terms of $\widehat{\boldsymbol{\beta}}$. Show the calculation. I think it's easiest to start with $\widehat{\boldsymbol{\alpha}}$ and substitute.

(h) Call $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ the *original* model, and $\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$ the *re-parameterized* model. Compare the vector of predicted y values $\widehat{\mathbf{y}}$ from the re-parameterized model to $\widehat{\mathbf{y}}$ from the original model.

(i) Compare the vector of residuals $\mathbf{e}$ from the re-parameterized model to $\mathbf{e}$ from the original model.

(j) Compare the proportion of explained variation R^2 from the original and re-parameterized models.

(k) Consider the null hypothesis $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$ based on the original model. The corresponding (logically equivalent, if and only if) null hypothesis for the re-parameterized model is $H_0 : \mathbf{K}\boldsymbol{\alpha} = \mathbf{h}$. Give a formula for $\mathbf{K}$. Show a little work.

(l) Compare the F statistic for testing $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$ to the F statistic for testing $H_0 : \mathbf{K}\boldsymbol{\alpha} = \mathbf{h}$. Start with the formula for the second one, and then substitute. Show your work. Use the formula sheet.

The overall story is that (all) these dummy variable schemes are equivalent in the sense that they lead to the same predictions and the same conclusions.

2. Suppose $\sqrt{n}(T_n - \theta) \xrightarrow{d} T$. Show $T_n \xrightarrow{p} \theta$. Please use Slutsky lemmas rather than definitions.

3. Let $X_1, \ldots, X_n$ be a random sample from a Binomial distribution with parameters 3 and θ. That is,

$$P(X_i = x_i) = \binom{3}{x_i} \theta^{x_i} (1-\theta)^{3-x_i},$$

for $x_i = 0, 1, 2, 3$. Find the maximum likelihood estimator of θ, and show that it is strongly consistent.

4. Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with density

$$f(x;\tau) = \frac{\tau^{1/2}}{\sqrt{2\pi}} e^{-\frac{\tau x^2}{2}},$$

where the parameter $\tau > 0$. Let

$$\widehat{\tau} = \frac{n}{\sum_{i=1}^{n} X_i^2}.$$

Is $\widehat{\tau}$ a consistent estimator of τ? Answer Yes or No and prove your answer. Hint: You can just write down $E(X^2)$ by inspection. This is a very familiar distribution.

5. Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ. Show that $T_n = \frac{1}{n+400} \sum_{i=1}^{n} X_i$ is a strongly consistent estimator of μ.

6. Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Prove that the sample variance $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ is a strongly consistent estimator of σ^2.

7. Independently for $i = 1, \ldots, n$, let

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

where $E(X_i) = E(\epsilon_i) = 0$, $Var(X_i) = \sigma_X^2$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and ϵ_i is independent of X_i. Let

$$\widehat{\beta}_n = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}.$$

Is $\widehat{\beta}_n$ a consistent estimator of β? Answer Yes or No and prove your answer.

8. In this problem, you'll use (without proof) the *variance rule*, which says that if θ is a real constant and $T_1, T_2, \ldots$ is a sequence of random variables with

$$\lim_{n\to\infty} E(T_n) = \theta \text{ and } \lim_{n\to\infty} Var(T_n) = 0,$$

then $T_n \xrightarrow{P} \theta$.

In Problem 7, the independent variables are random. Here they are fixed constants, which is more standard (though a little strange if you think about it). Accordingly, let

$$Y_i = \beta x_i + \epsilon_i$$

for $i = 1, \ldots, n$, where $\epsilon_1, \ldots, \epsilon_n$ are a random sample from a distribution with expected value zero and variance σ^2, and β and σ^2 are unknown constants.

(a) What is $E(Y_i)$?

(b) What is $Var(Y_i)$?

(c) Use the same estimator as in Problem 7. Is $\widehat{\beta}_n$ unbiased? Answer Yes or No and show your work.

(d) Suppose that the sequence of constants $\sum_{i=1}^n x_i^2 \to \infty$ as $n \to \infty$. Does this guarantee $\widehat{\beta}_n$ will be consistent? Answer Yes or No. Show your work.

(e) Let $\widehat{\beta}_{2,n} = \frac{\overline{Y}_n}{\overline{x}_n}$. Is $\widehat{\beta}_{2,n}$ unbiased? Consistent? Answer Yes or No to each question and show your work. Do you need a condition on the x_i values ?

(f) Prove that $\widehat{\beta}_n$ is a more accurate estimator than $\widehat{\beta}_{2,n}$ in the sense that it has smaller variance. Hint: The sample variance of the explanatory variable values cannot be negative.

9. Let X be a random variable with expected value μ and variance σ^2. Show $\frac{X}{n} \xrightarrow{p} 0$.

10. Let $X_1, \ldots, X_n$ be a random sample from a Gamma distribution with $\alpha = \beta = \theta > 0$. That is, the density is

$$f(x;\theta) = \frac{1}{\theta^\theta \Gamma(\theta)} e^{-x/\theta} x^{\theta-1},$$

for $x > 0$. Let $\widehat{\theta} = \overline{X}_n$. Is $\widehat{\theta}$ a consistent estimator of θ? Answer Yes or No and prove your answer.

11. The ordinary univariate Central Limit Theorem says that if $X_1, \ldots, X_n$ are a random sample (independent and identically distributed) from a distribution with expected value μ and variance σ^2, then

$$Z_n^{(1)} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z \sim N(0,1).$$

An application of some Slutsky theorems (see lecture slides) shows that also,

$$Z_n^{(2)} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\widehat{\sigma}_n} \xrightarrow{d} Z \sim N(0,1),$$

where $\widehat{\sigma}_n$ is any consistent estimator of σ. For this problem, suppose that $X_1, \ldots, X_n$ are Bernoulli(θ).

(a) What is μ?

(b) What is σ^2?

(c) Re-write $Z_n^{(1)}$ for the Bernoulli exanple.

(d) What about $Z_n = \frac{\sqrt{n}(\overline{X}_n - \theta)}{\sqrt{\overline{X}_n(1-\overline{X}_n)}}$? Does Z_n converge in distribution to a standard normal? Why or why not?

(e) What about the t statistic $T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n}$, where S_n is the sample standard deviation? Does T_n converge in distribution to a standard normal? Why or why not?

12. Here is an integral you cannot do in closed form, and numerical integration is challenging. For example, R's `integrate` function fails.

$$\int_0^{1/2} e^{\cos(1/x)}\, dx$$

Using R, approximate the integral with Monte Carlo integration, and give a 99% confidence interval for your answer. You need to produce 3 numbers: the estimate, a lower confidence limit and an upper confidence limit. **Please bring your printout to the quiz**.