STAT 135 CONCEPTS OF STATISTICS
HOMEWORK 7

Assigned August 3, 2021, due August 10, 2021

This homework pertains to materials covered in Lecture 18, 19 and 20. The assignment can be typed or handwritten, with your name on the document, and with properly labeled input code and computer output for those problems that require it. I suggest you attach your (R, SAS, or whatver you are using) input code at the end of your file clearly indicating the problem it corresponds to. If you choose to collaborate, the write-up should be your own. Please show your work! Upload the file to the Week 6 Assignment on bCourses.

Problem 1. The dataset `chicks` was obtained from BLSS: The Berkeley Interactive Statistical System by Abrahams and Rizzardi. Each observation corresponds to an egg (and the resulting chick) of a bird called the Snowy Plover. The data were taken at Point Reyes Bird Observatory. Column 1 contains the egg length in millimeters, Column 2 the egg breadth in millimeters, Column 3 the egg weight in grams, and Column 4 the chick weight in grams. The object is to estimate the weight of the chick based on dimensions of the egg.

(1) First you are going to regress chick weight on egg length using the standard regression model ($y_i = \beta_0 + \beta_1 x_i + \epsilon_i$) in terms of these two variables. You may assume that ϵ_i is normally distributed. Find the means and SDs of both variables, as well as the correlation between them. Use the formulas you derived earlier in Lecture 18 to find the slope and the intercept of the regression line. Provide units for the slope and the intercept, and write the equation of the regression line. Draw the regression line on your plot. Do not use `lm()` for this part;
(2) Now use `fit <- lm(`$\cdots$`)` to regress chick weight on egg length. Check that `lm()` produces the same slope and intercept that you got in part (1). Look at the residual plot and the Normal-QQ plot to see whether the model assumptions are reasonable;
(3) For each of the t statistics in `summary(fit)`, state the null and alternative hypotheses that are being tested, and state the conclusion of the test.

Problem 2. (**Continuing Problem 1**) Now we examine the relationship between chick weight and egg weight.

(1) If possible, construct a 95% - confidence interval for the mean weight of Snowy Plover chicks that hatch from eggs weighing 8.5 grams.

(*Hint*: To estimate σ in the formula for the se of the regression line use the residual se output by `lm()`.)

(2) I have a Snowy Plover egg that weighs 8.5 grams. If possible, construct a 95% - prediction interval for the weight of the chick that will hatch from this egg. It's a prediction interval, rather than a confidence interval, because it's trying to predict the value of a random variable instead of estimating a fixed parameter.

(3) Repeat parts (1) and (2) when the egg weight is 12 grams instead of 8.5 grams.
(*Hint*: beware of extrapolation)

Problem 3. (**Continuing Problem 1**) The object is still to find a good way to predict the weight of a chick given measurements on the egg, using linear regression as the only tool. The difference between this problem and the preceding two problems is that now you are going to use a combination of variables to estimate the weights of the chicks.

(1) Regress the weights of the chicks on the lengths and breadths of the eggs. Assess the regression. Compare it with the simple regression analyses you performed in Problem 1 and 2. Is one noticeably better than others? Draw the residual plot. Is there any noticeable heteroscedasticity?
(2) Regress egg weight on egg length and egg breadth. Assess this regression, especially the R^2 value. Report the Variance Inflation Factor (VIF) value.
(3) Now regress the weights of the chicks on all three predictor variables: egg length, egg breadth, and egg weight. How do you reconcile the result of the F test in this regression with the results of the t-tests in `summary(fit)`? Explain why this regression is not as impressive as any of the three you compared in (1), even though it has a higher R^2.
(4) Perform all possible regressions of chick weight using combinations of the three predictor variables used in (2). *Do not turn in all the results.* Are there any that are clearly better than the others? Which ones, and why?

Problem 4. Let $X_1, \dots, X_n \overset{iid}{\sim} N(\beta, \sigma^2)$, and suppose that the prior distribution on θ is $N(\mu, \tau^2)$. (Here we assume that σ^2, μ and τ^2 are all known.

(1) Prove that the posterior distribution of θ is also normal, with mean and variance given by

$$E(\theta|\mathbf{X}_n) = \frac{\tau^2}{\tau^2 + \sigma^2/n}\bar{X}_n + \frac{\sigma^2/n}{\sigma^2/n + \tau^2}\mu,$$
$$\mathrm{Var}(\theta|\mathbf{X}_n) = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}.$$

(2) How will the posterior distribution look like when $n \to \infty$?

(3) Intuitively, as the prior information becomes more vague, the posterior mean as a Bayes estimator for θ should rely more on the sample information. Examine $E(\theta|\mathbf{X}_n)$ as $\tau^2 \to \infty$, and explain why your result is consistent with the intuition.

Problem 5. Let $X_1, \ldots, X_n$ be iid Poisson(λ), and let λ have a Gamma(α, β) distribution.

(1) Find the posterior distribution of λ;
(2) Calculate the posterior mean and variance;
(3) Is posterior mean, again, a linear combination of the prior and sample means?