Statistics 512 Homework 3 Spring 2021

- A reminder: Please submit this and every homework as a single PDF document. If you need assistance preparing a PDF, feel free to ask for help in office hours.
- It is not necessary to provide units unless the question specifically asks for them.
- This homework is officially due on **Thursday, March 4**, at 11:59pm. The unusual due date is because of the wellness day on Friday, March 5. However, assignments will be accepted without penalty until 11:59pm on Friday, March 5. The usual 24-hour grace period (with a lateness penalty) will apply to homework submitted on Saturday, March 6.

Part 1: Non-linear least squares

The data set `lamp.txt` gives the energy radiated from a carbon filament lamp per cm^2 per second, and the absolute temperature of the filament in 1000 degrees Kelvin. These data are given in Daniel and Wood (*Fitting Equations to Data*, 1980), and originally published in E. S. Keeping (*Introduction to Statistical Inference*, 1962). The variables are:

- `temperature` The absolute temperature of the filament in 1000 degrees Kelvin.
- `energy` The energy radiated from a carbon filament lamp per cm^2 per second.

An exponential model is proposed to capture the relationship between temperature (the predictor, x) and energy radiated (the response, y). The proposed model is:

$$y = \theta_1 e^{\theta_2 x} + \varepsilon \tag{1}$$

Use non-linear least squares to fit the model in equation (1) to these data. Use your model fit to answer questions 1 – 2.

Part 2: Interactions

S. Giery conducted a study to determine if water coloration and the presence of a predator affect the number of spots on the fin of a particular species of fish. At each of $n = 21$ ponds, he recorded the following variables:

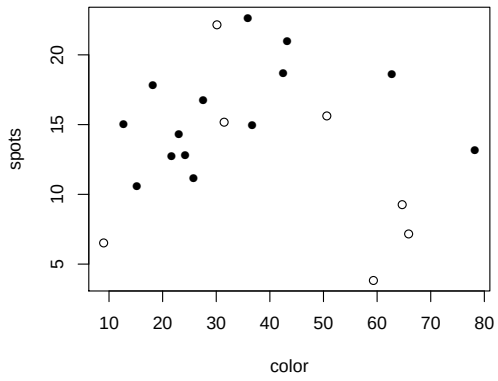
- `predator` A binary categorical variable that indicates whether predators are present or absent from the pond.
- `color` A quantitative measure of the balance between red and blue light transmitted through the pond water.
- `spots` The average number of spots on the fin of a particular species of fish.

To accommodate the categorical predictor `predator`, the following indicator variable is created:

$$\texttt{predator.id} = \begin{cases} 1 & \texttt{predator = present} \\ 0 & \texttt{predator = absent.} \end{cases}$$

The data can be found on the class website in the file `fish-spots.txt`.

The plot below shows the relationship between `spots` and `color`, with filled points showing ponds where predators are present, and open symbols showing ponds where predators are absent.



The scientist is interested in characterizing the effects of the presence of `predator`s and the `color` of the pond water on the number of fin `spots` on these fish. For the questions that follow, use the following notation. Let y denote the number of fin `spots` on this species of fish. Let x_1 denote the indicator variable `predator.id`. Let x_2 denote the predictor `color`.

The plot strongly suggests a non-linear relationship between `spots` and `color`. Fit a model that includes both the effect of `predator`s and a quadratic effect of `color` on `spots`. In other words, fit the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon \tag{2}$$

Use this model to answer question 3.

The scientist is interested in asking if the effect of `predator`s on the number of fin `spots` depends on the `color` of the pond water. In other words, he is interested in asking whether there is an interaction between `predator` presence and pond water `color` with respect to their effect on fin `spots`. Because the effect of `color` is non-linear, a model with an interaction must include the new predictors x_1x_2 and $x_1x_2^2$. In other words, we want to fit the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + +\beta_4 x_1 x_2 + \beta_5 x_1 x_2^2 + \varepsilon \tag{3}$$

and then test

$$H_0 : \beta_4 = \beta_5 = 0.$$

Fit the model in equation 3, and test $H_0 : \beta_4 = \beta_5 = 0$. Use your result to answer questions 4 – 6.

Part 3: Variable selection

Measuring body fat accurately requires immersing people in water to measure their body volume. Obviously, such measurements are costly. Thus, there is an incentive to be able to predict body fat on the basis of more easily measurable quantities. The data set `bodyfat.txt` has records collected for 248 men. Each record contains the following variables:

bodyfat the percent body fat, measured by the immersion method

age in years

weight in lbs

height in inches

neck circumference in cm

chest circumference in cm

abdomen circumference in cm

hip circumference in cm

thigh circumference in cm

knee circumference in cm

ankle circumference in cm

biceps circumference in cm

forearm circumference in cm

The data are available on the course website.

Use stepwise variable selection with AIC to build a regression model that predicts `bodyfat`, and uses the remaining variables as possible predictors. You do not need to consider models with interactions or with polynomial terms. If you are using R, use the `step` function. Use your model to answer questions 7 – 9.

As part of these questions, you will need to find AIC values for various regression models. Note that `step` will show the AIC value for each model that it considers. However, you can also obtain the AIC for a regression model in R by using the `extractAIC` command, as so:

```
> fm1 <- lm(y ~ x1 + x2, data = my.data)  # some regression model here
> extractAIC(fm1)                         # get the AIC value for fm1
```

The AIC value is the second of the two values produced by `extractAIC`.

Confusingly, R has two different functions for finding the AIC value from a fitted model: `extractAIC` and `AIC`. Unfortunately, for regression models, these two functions give different answers. For a particular data set, the answers will always differ by the same amount, so either function can be used to compare models fit to the same data set. For this question, be sure to use `extractAIC`, because this is the function that `step` uses. If you are curious as to why these functions give different behavior, see the help documentation for `extractAIC` for the gory details.

ST 512 Homework 3 Questions.

1. (2 points) Report the least-squares estimates of the parameters θ_1 and θ_2 for the lamp data. Show work or computer code.

   $\hat{\theta_1} =$

   $\hat{\theta_2} =$

2. (2 points) Use your model fit from question 1 to predict the energy output of this lamp when the temperature is $x = 1.4$ (which corresponds to 1400 K). Show work or computer code.

3. (2 points) In the model without an interaction (eq. 2), briefly interpret the estimate of the partial regression coefficient associated with `predator.id`.

4. (3 points) In the model with an interaction (eq. 3), test $H_0 : \beta_4 = \beta_5 = 0$. Report your test statistic and the associated p-value. Show work or computer code.

5. (2 points) Briefly interpret the outcome of your test from question (4).

6. (3 points) Use the model in equation 3 to estimate the effect of `predator` presence on fish `spots` at each of the following values for `color`: `color` = 25, `color` = 37.5, and `color` = 50. (These are roughly the quartiles of the distribution of `color`.) In other words, calculate how the presence of `predator`s changes the predicted value of `spots` when `color` = 25, `color` = 37.5, and `color` = 50. Show some work or computer code.

   Effect of predator presence when `color` = 25:

   Effect of predator presence when `color` = 37.5:

   Effect of predator presence when `color` = 50:

7. (3 points) Use stepwise selection with AIC as the variable-selection criterion to build a regression model for `bodyfat`. Report the predictors included in your final model. (You do not have to give parameter estimates for the partial regression coefficients.) Also report the AIC and adjusted R^2 for the final model. Show the computer code that you used.

   Predictors in final model:

   AIC:

   Adjusted R^2:

8. (2 points) Calculate and report the variance inflation factors (VIFs) for all the predictors in your final model from question 7. Show some work or computer code.

9. (1 point) Do any of the VIFs in question 8 suggest problematic collinearity among the predictors? Answer yes or no, and briefly defend your answer.