TigerSTAT: Simple Linear Regression Model Lab

Introduction to TigerSTAT

The Bol'shaya Koshka (Russian for big cat) Reserve is a newly created animal reserve that was uniquely developed to help endangered species prosper. This 10,000 acre wild animal reservation was selected because an abundance of Siberian tigers have been found in the area. The diverse terrain of the reserve provides a wide variety of habitats for many different species of animals.

Since the tigers in this area are much more abundant than any other area in the world, they are starting to draw a significant number of researchers to the region. Your primary responsibility will be to help these researchers as they study the tigers and then incorporate the results of their research into a system to identify the best management practices for this reserve.

An important component of monitoring endangered species is to understand the age distribution of the population. Shifts in the distribution could indicate potential issues in sustaining the population.

While the exact age is not known for most of the tigers in your reserve, the age of some tigers are known. To estimate the age of a tiger that is captured on your reserve, you will need to compare characteristics of the captured tiger to the ones that live on the research zone (whose ages are known).

When data is collected as an indirect measure for the variable of interest, it is often called ***proxy data***. For example, in their 2004 paper, Whitman, et. al. describe how the color of a lion's nose can be used to estimate it's age. Your mission is to go into the Bol'shaya Koshka reserve and gather sample data on tigers. Then, using your sample data, you are to establish a simple linear regression model to estimate the age of a tiger based on the available ***proxy*** variables.

Play the tutorial for the TigerSTAT game briefly so you are familiar with the game controls. The game is found at the web site http://statgames.tietronix.com/tigerstat/tigerstat_webgl/index.html. Enter a PlayerName and GroupName (The "PlayerName" is a secret name, any combination of letters and numbers with no spaces. Do not use your name or a term that will identify you or your group. All group members should use the same "PlayerName"). The "GroupName" will be provided by your instructor. You can choose either the **Casual** or **Hard** version, select **Continue** and **Load Tutorial**. If you forget commands anytime during game play, you can hit the "p" key to pause the game and see game instructions.

Collect Tiger Data using TigerSTAT. Go to http://statgames.tietronix.com/tigerstat/tigerstat_webgl/index.html and enter your PlayerName and the GroupName provided by your instructor. Use the Full Screen option to see the entire game on your computer screen. Select **Load Mission 1** and then **DataSet1**. Use the Full Screen option to see the entire game on your computer screen. You can choose either the **Casual** or **Hard** game option. You can type "**p**" to **pause** anytime while playing the game. This will allow you to review all the controls, exit the game and save your data.

TASK #1: Preliminary data analysis

For this task we will examine one model developed for lions and see how well it extends to our tigers. We will use the simple linear regression model:

$$Y = \beta_0 + \beta_1 x \qquad (1)$$

In this case, *Y* is the age of the tiger and *x* the proxy variable. For Questions 1-4 you may need to first explore the data collected to determine what proxy variables to consider.

Since this is your first task, you will only be required to collect a minimum of data from five tigers. You have the option to collect more data. Recall that a larger sample size will improve the accuracy of your test results.

1. Calculate the mean and standard deviation of the potential *proxy* variables of the tigers in your sample.
2. Calculate the mean and standard deviation of the Age of the tigers in your sample.
3. Produce a graph of the Age against each potential *proxy* variable for your sample – describe the relationships you observe. Would a linear model be appropriate for these variables?
4. Are there any reasons to suspect your data may be biased? If you could, how would you ensure these issues were addressed in collecting tiger data?
5. In the Whitman et. al. (2004) article the authors used the proportion of nose blackness as their proxy to develop the model. Does this seem like the best choice for the tigers in your sample? What additional work would you want to do to choose the best proxy?

TASK #2: Preliminary model estimation

Use your software package to regress NoseBlackProportion on Age in order to estimate the parameters in equation (1). Report the estimated slope value to the instructor, then answer questions #6 - #9 in preparation for classroom discussion.

Before making any inferences or predictions on the mean values of the response variable, we generally first determine if there is a significant relationship between the predictor and response. If there is no relationship, the slope would be zero hence we desire to test the null hypothesis that $\beta_1 = 0$ versus the alternative hypothesis $\beta_1 \neq 0$. The test statistic (*t*) for this hypothesis is

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\beta_1}} \qquad (2)$$

and the test statistic has a *t*-distribution with *n-2* degrees of freedom when the null hypothesis is true.

6. Compute the test statistic for the null hypothesis Ho: $\beta_1 = 0$. Do we accept or reject the null hypothesis?
7. Interpret the hypothesis test in the context of the study.
8. What is the interpretation of the estimated slope parameter (be specific and be sure the answer is in the context of the tiger age)?
9. Compare your answers in questions 6 through 8 to that of one or two other groups. What issues should you consider in using this model?

TASK #3: Model performance and assessment

Performance: A statically significant relationship is important, but we must assess the model performance and fit before using it. One measure of performance commonly used is the coefficient of determination, R^2. This is the proportion of variability in the data set that is accounted for by the statistical model and gives us insight as to how well future outcomes are likely to be predicted by the model. We compute R^2 using equation (3) below.

$$R^2 = 1 - \frac{SSE}{SST} \quad \text{where} \quad SSE = \sum_i (y_i - \hat{y}_i)^2 \quad \text{and} \quad SST = \sum_i (y_i - \bar{y})^2 \quad . \qquad (3)$$

SSE is the sum of squared error, a measure of the unexplained variance or variability not captured by the model. SST, the sum of squares total, is a measure of the overall sample variance.

10. Compute R^2 for the preliminary model developed in Task #2. Based on this value, how well do you expect your model will perform?
11. Based on your estimate of the model, what age is an average tiger with 10% NoseBlackProportion? 50%? 90%?
12. Compare your estimates from #11 to that of one or two other groups? Comment on the results in terms of the R^2 value of the model.
13. Comment on any strengths/weaknesses you see in your model. What do you think our model might be good for? Is the coefficient of determination found for your model the best way to determine the goodness for this model?

Assessing the model: Checking assumptions for any statistical model is imperative before making inferences. For our simple regression model, we assume that the errors are randomly distributed, following a normal distribution, with mean zero and a constant variance for all values of the predictor. Let's check the validity of these assumptions.

14. Using the parameter values estimated in Task #2, produce the model based estimates of the age of the tigers in your sample, $\hat{y}$.
15. For each estimated value, compute the associated residual or difference between actual and predicted: $e_i = (y_i - \hat{y}_i)$.
16. Create an appropriate plot you have learned about in class (histogram, qq-plot) for assessing the normality assumption for the set of residual values computed. Does the assumption of normality hold?
17. Plot the residuals against the NoseBlackProportion. Does the assumption that the errors are random appear reasonable? Mean of zero? Constant variance?
18. How appropriate to the model seem for the data in the sample? Would you recommend using it to determine the tiger ages in the preserve?

TASK #4: Model revision (optional)

It takes a careful reading of the Whitman et. al. (2004) article to see what model the authors actually used. It is found in the caption for Table 1. Relook at this caption to confirm that they modeled the age of lions using by first computing the arcsin of percentage of nose blackness (NoseBlackProportion), or the model:

$$AGE = \beta_0 + \beta_1 \arcsin(\text{NoseBlackProportion}) \qquad (4)$$

The use of the arcsin is what is known as a "transformation". In statistical modeling when the assumptions of the model do not hold for a data set this is often a means of solving the problem. The choice of transformation is a more advanced topic. In fact, we could choose to transform the response, age, instead of the predictor. Our interest at this point is not to become experts in transforming data. The choice of the arcsin is actually not uncommon in certain fields when the predictor variable is a percentage or proportion. Our interest is whether

the choice, which was used for the lion data, appears reasonable in modeling tiger ages.

19. Create a new variable ANoseBlackProportion by computing the Arcsin of NoseBlackProportion for tigers in your sample (note that most software packages have the arcsin function available – if not, one can first compute this in Excel). Graph AGE as a function of this new variable. Do you think the assumed linear relationship is reasonable? Why or why not?
20. Use your software package to regress ANoseBlackProportion on AGE in order to estimate the parameters in equation (4). Then repeat the key steps used in the model without the transformed data (i.e. answer questions #6-12 for the model with the transformation). Did the transformation improve the model? Do you believe the model using the transformed variable is reasonable for use for tiger age data?
21. To use the transformed data, what is the interpretation of the slope coefficient? How do you then use the model to make age predictions/estimates? Use this new model to produce the estimates in question #14.
22. How do your estimates of the tigers ages compare to those from other groups? What is your advice to the research team about the use of the model/data in predicting ages of Tigers?
23. What can you do to improve the model?