Week_4_Problem_Set

Nguyen Tien Anh Quach

2024-02-08

#Questions - 20 points per regression (40 points total) Your assignment is to conduct two different linear regressions on the TreePlots.csv data: (1) mean tree diameter (mDBH.cm) versus mean height (mH.m), and (2) mean height (mH.m) versus mean wood density (mWD.g.m3). In the first regression, diameter should be your dependent variable. In the second regression, mean height should be your dependent variable. Consider the effect of outliers and examine your qqplots when checking the assumptions of your model. For each regression, write a 1-page description of your analysis, results, and inference. The write-up should include the following information:

- Null and alternative hypotheses - 2.5 1. Mean tree diameter vs. mean height

H_0: There is no significant linear relationship between mean height and mean tree diameter. H_a: There is a significant linear relationship between mean height and mean tree diameter.

2. Linear model result:

```
##
## Call:
## lm(formula = mDBH.cm ~ mH.m, data = tree)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5324 -0.5222 -0.1646  0.4230  2.7311
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.07896    0.88402  -9.139  1.3e-13 ***
## mH.m         1.86315    0.05195  35.863  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7612 on 71 degrees of freedom
## Multiple R-squared:  0.9477, Adjusted R-squared:  0.9469
## F-statistic:  1286 on 1 and 71 DF,  p-value: < 2.2e-16
```

Mean tree height in forest plots has a significantly positive linear relationship with mean tree diameter, with a 1 m increase in mean tree height resulting in a 1.86 cm increase in mean tree diameter ((R^2) = 0.95, Adj. (R^2) = 0.95, F(1,71) = 46.34, $p < 0.001$).

- Results of your statistical model(s), interpreting your model in 2-3 sentences that include the appropriate reporting of the statistics - 2.5 • An interpretation of the regression model (equation) from the analysis (e.g. how mean height varies with different levels of wood density). 5 • Include an appropriate figure with the equation. Plot your confidence intervals and prediction intervals. 5 • A description of how you checked the assumptions of your statistical test, and if you decided to re-run your model after analyzing your diagnostic figures - 2.5 • An interpretation of diagnostic figures (in an appendix) - 2.5

Again, upload your .Rmd separately, and make sure to add comments to your code.

#Normality comment Against better judgment, in the past we have used the shapiro.test() to assess normality. Remember that no test will show that your data has a normal distribution. Normality statistics show when your data is sufficiently inconsistent with a normal distribution that you would reject the null hypothesis of “no difference from a normal distribution”. However, when the sample size is small, even big departures from normality are not detected, and when the sample size is large, even the smallest deviation from normality will lead to a rejected null. In other words, if we have enough data to fail a normality test, we always will because real-world data won’t be clean enough. See (http://www.r-bloggers.com/normality-and-testing-for-normality/) for an example with simulated data. So, where does that leave us? Explore your data for large deviations from normality and make sure to assess heteroscedasticity and outliers. But, don’t get hung up on whether your data are normally distributed or not. As the author of the above link suggests: “When evaluating and summarizing data, rely mainly on your brain and use statistics to catch really big errors in judgment.”