STATS 191: Homework Assignment 7

Dr. Pratheepa Jeganathan

11/15/2019

You may discuss homework problems with other students, but you have to prepare the written assignments yourself.

Please combine all your answers, the computer code and the figures into one PDF file, and submit a copy to gradescope.

Please use **newpage** to write solution for each part of a question.

Please specify the page number for each part of a question in gradescope.

Grading scheme: $\{0, 1, 2\}$ **points per question, total of 30. (Questions 1 and 2)**

Question 3 is for BONUS 12 points. This will be bonus of 3.1 points in the final weighted average.

The maximum of this homework 42/30*100 = 140 so the maximum of all 7 homework is (500+127.27+140) = 767.27. Then, the maximum of weighted average of homework is 767.27/700*55 = 60.2 (55+2.1+3.1).

Due date: 11:59 PM November 22, 2019 (Friday evening).

Question 1

The data set http://stats191.stanford.edu/data/asthma.table contains data on the number of admittances, Y to an emergency room for asthma-related problems in a hospital for several days. On each day, researchers also recorded the daily high temperature T, and the level of some atmospheric pollutants P.

1. Fit a linear regression model to the observed counts Y as a linear function of T and P. 2
2. Looking at the usual diagnostics plots, does the constant variance assumption seem justified? 2
3. The outcomes are counts, for which a common model is the so-called Poisson model which says that $\text{Var}(Y) = E(Y)$. In words, this says that the variance of the outcome is equal to the expected value of the outcome. Using a two-stage procedure, fit a weighted least squares regression to Y as a function of T and P with weights being inversely proportional to the fitted values of the initial model in 1. 2
4. Looking at the usual diagnostics plots of this model (which takes the weights into account), does the constant variance assumption seem more reasonable? (The change may not be astonishing – the point of the problem is to try using weighted least squares.) 2
5. Using the weighted least squares fit, test the hypotheses at level $\alpha = 0.05$ that
   (a) the number of asthma cases is uncorrelated to the temperature allowing for pollutants; 2
   (b) the number of asthma cases is uncorrelated to the atmospheric pollutants allowing for temperature. 2

Question 2

We revisit Question 3 from Homework Assignment 3, using weighted least squares to estimate standard error of the estimate. When generating data below, set X to be sampled as 100 points drawn randomly on the interval (0,1): `runif(100)` and use the regression function

$$f(X) = 1 + 2 \cdot X.$$

(1) Write a function with the same regression function but errors that are not normally distributed using, say, $\epsilon \sim$ exponential $(1) - 1$: `rexp(n, rate = 1)-1`, where n is the number of observations. This means that Var $(\epsilon) = 1$. Use weighted least squares and construct the Z statistic (i.e. ignore degrees of freedom, pretending they are `Inf`) to test $H_0 : \beta_1 = 2$. Do the simulation 1000 times to compute 1000's of Z statistic values. 2

   (a) Does the Z-statistic have close to a standard normal $N(0, 1)$ distribution? (Plot the density of Z-statistic and the standard normal variable.) 2

   (b) How often is your Z-statistic larger than the usual 5% threshold (from the standard normal distribution)? 2

```
generateTstat = function(){
  #n = number of observations
  #X = simulate from runif(n)
  #error = simulate error from rexp(n)-1
  #Y = write Y as a function of X and error
  #fit = fit regression line using lm (what can be the weight?)
  # beta1hat = compute weighted least squares estimate of slope using summary(fit)$coefficient[2,1]
  # se_beta1hat = compute standard error of slope estimate using summary(fit)$coefficients[2, 2]
  # Zstat = Compute the Z-statistic using the appropriate formula (for testing H0: beta_1 = 2)
  # return(Tstat)
}
```

(2) Write a new function with the same regression function but multiply the errors `rexp(n)-1` by $\sqrt{(1+X)}$ so the i-th error variance (given X_i) is proportional to $1 + X_i$. Use weighted least squares and construct the Z statistic (i.e. ignore degrees of freedom, pretending they are `Inf`) to test $H_0 : \beta_1 = 2$. Do the simulation 1000 times to compute 1000's of Z statistic values. 2

   (a) Does the Z-statistic have close to a standard normal $N(0, 1)$ distribution? (Plot the density of Z-statistic and the standard normal variable.) 2

   (b) How often is your Z-statistic larger than the usual 5% threshold (from the standard normal distribution)? 2

(3) Write a new function with the same regression function but multiply the errors `rexp(n)-1` by $\exp(0.5 * (1 + 5 * X))$ so the i-th error variance (given X_i) is proportional to $\exp(1 + 5 * X_i)$. Use weighted least squares and construct the Z statistic (i.e. ignore degrees of freedom, pretending they are `Inf`) to test $H_0 : \beta_1 = 2$. Do the simulation 1000 times to compute 1000's of Z statistic values. 2

   (a) Does the Z-statistic have close to a standard normal $N(0, 1)$ distribution? (Plot the density of Z-statistic and the standard normal variable.) 2

   (b) How often is your Z-statistic larger than the usual 5% threshold (from the standard normal distribution)? 2

Question 3 (Based on CH Chapter 8.4-8.6)

The file http://www1.aucegypt.edu/faculty/hadi/RABE5/Data5/P229-30.txt contains the values of the daily DJIA (Dow Jones Industrial Average) for all the trading days in 1996. The variable `Time` denotes the trading day of the year. There were 262 trading days in 1996.

(1) Fit a linear regression model connecting `DJIA` with `Time` using all 262 trading days in 1996.
   (a) Is the linear trend model adequate?
   (b) Examine the residuals for time dependencies, including a plot of the autocorrelation function.

(2) Regress `DJIA[t]` (start from the second observation) against its lagged by one version `DJIA[t-1]` (there is n-1 observations).
   (a) Is this an adequate model?
   (b) Are there any evidences of autocorrelation in the residuals?

(3) The variability (volatility) of the daily DJIA is large, and to accomodate this phenomenon the analysis is crried out on the logarithm of the DJIA. Repeat part (2) above using `log(DJIA)` instead of `DJIA`.
   (a) Is this an adequate model?
   (b) Are there any evidences of autocorrelation in the residuals?