Homework 3

Spring 2025

Instructions

- The homework is due on **Mar. 7, 2025**, 11:59 pm.
- The homework should be submited electronically via Canvas.
- Please upload a single PDF containing the solutions in the correct order. If you include scanned images, make sure that they are organized and easy to read.

---

1. (12*6=72 points) Using the 'Baseball Salary Data' in canvas, which includes the data (a `.csv` file) and a `.txt` file with some background information on the data and the variables, do the following:

   (a) Fit a linear regression model with `"salary"` as the response, and the other 16 variables (excluding `"names"`) as the predictors.

   (b) What percentage of the variation in salaries is explained by the linear model above?

   (c) Comment on the coefficient of the predictor `"hits"`. Is this coefficient consistent with what your intuition says should be the relationship between number of `hits` and `salary`? Why or why not?

   (d) Test the null hypothesis (using level of significance $\alpha = 0.05$) that none of the 16 predictors is related to `salary`. What is the proper conclusion about the linear model above and its utility?

   (e) Test the null hypothesis (using level of significance $\alpha = 0.05$) that the variables `"batting average"`, `"on base percentage"`, `"hits"`, `"doubles"` and `"triples"` are **not** needed in the same model with the other 11 predictors. Is the result surprising? Give a possible explanation for the result.

   (f) What percentage of the variation in salaries is explained by the linear model containing the 11 variables not named in part **(e)** above?

   (g) Obtain residuals from the linear model fitted in part **(f)** above, and produce the following three plots: **(i)** the residuals versus the predicted values, **(ii)** a kernel density estimate of the residuals, and **(iii)** a Normal probability (or Q-Q) plot of the *standardized* residuals. Comment on the plots.

   (h) Use the command `'leaps'` in the `R` package `"leaps"` along with the strategy discussed in class to choose a good subset of the 16 predictors to include in a linear model with `"salary"` as the response. **Describe fully the rationale you use** in choosing your model.

   (i) Plot the *standardized* residuals from your chosen model versus an index running from 1 to 337. Identify any players who have standardized residuals that are larger in absolute value than 3. Are these players different in any important way from most of the other players?

   (j) Provide a plot of the standardized residuals versus the predicted values and comment on the plot.

   (k) Provide a Normal probability (or Q-Q) plot of the standardized residuals and comment on the plot.

   (l) Provide a plot of Cook's D values. Do any data points seem to be influential? Why or why not?

2. (Problems 3, on page 359 of textbook, Section 10.2). (9+9+10=28 points) Four different concentrations of ethanol are compared at level $\alpha = 0.05$ for their effect on sleep time. Each concentration was given to a sample of 5 rats and the REM (rapid eye movement) sleep time for each rat was recorded (SleepRem.txt). Do the four concentrations differ in terms of their effect on REM sleep time?

   (a) State the relevant null and alternative hypotheses for answering this question, and use hand calculations to conduct the ANOVA F test at level of significance 0.05. State any assumptions needed for the validity of this test procedure. (Hint. You may use the summary statistics $\overline{X}_1 = 79.28$, $\overline{X}_2 = 61.54$, $\overline{X}_3 = 47.92$, $\overline{X}_4 = 32.76$, and $MSE = S_p^2 = 92.95$.)

(b) Import the data into the R data frame sl, and use the R command anova(aov(sl$values sl$ind)) to conduct the ANOVA F test. Give the ANOVA table, stating the p-value, and the outcome of the test at level of significance 0.05.

(c) Use R commands to test the assumptions of equal variances and normality. Report the p-values from the two tests and the conclusions reached. Next, construct a boxplot and the normal Q-Q plot for the residuals, and comment on the validity of the normality assumption on the basis of these plots.