STAT 500 - Homework 1

Due in class Thursday, January 21

The dataset `teengamb` concerns a study of teenage gambling in Britain. More details about the study can be found in Ide-Smith & Lea (1988) Journal of Gambling Behavior, 4, 110-118. It contains five variables:

- `sex`: $0 = \text{male}, 1 = \text{female}$
- `status`: socioeconomic status score based on parents' occupation
- `income`: in pounds per week
- `verbal`: verbal score in words out of 12 correctly defined
- `gamble`: expenditure on gambling in pounds per year

Make a numerical and graphical summary of the data, commenting on any features that you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of the data. For the variable "sex", assign labels "male" and "female" and make sure R is treating it as a categorical variable before you compute the summary.

Also answer the following questions, **incorporating the answers into your report**:

1. In your summary, report the means and medians of variables "income" and "gamble". Explain why the means are larger than the medians.
2. How many different values are there for the variable "verbal"? Based on the boxplot and/or other summaries of the variable "verbal", what values of the verbal score would you consider to be outliers? Give the row numbers of the outlier observations.

Hints: After you install R (if it is not already installed on the system you are using) and the `faraway` package, you can type `library(faraway)` to load the package and `data(teengamb)` to load the data. Useful R functions for this homework: data(), summary(), hist(), plot(), boxplot(). You can always type `help(subject)` to get detailed help on the `subject`, e.g. `help(plot)`. Or you can type `help.start()` to get interactive help with a search engine.

Solutions to this homework should be no longer than 3 pages.

STATS 500 - Homework 2

Due in class Tuesday, February 2

Part A (Maximum 2 pages). The dataset `uswages` is drawn as a sample from the Current Population Survey in 1988.

1. Fit a regression model with weekly wages as the response and years of education and experience as predictors. Present the output.
2. What percentage of variation in the response is explained by these predictors? (Percentage variance explained is the same as coefficient of determination).
3. Which observation has the largest (positive) residual? Give the case number.
4. Compute the mean and median of the residuals. Explain what the difference between the mean and the median indicates.
5. For two people with the same education and one year difference in experience, what would be the difference in predicted weekly wages?
6. Compute the correlation of the residuals with the fitted values. Plot residuals against fitted values. Explain the value of this correlation using the geometric (projection) interpretation of least squares.

Hints: Useful R functions: data(), lm(), summary(), residuals(), fitted(), which.max(), mean(), median(), cor(), plot(). Note that the experience variable has some negative values which most likely indicate missing data. Those observations should be removed from the analysis.

Part B (Maximum 2 pages). Using R, create a 10×3 matrix X:

$$X = \begin{pmatrix} 1 & 2 & -2 \\ 1 & -1 & -2 \\ 1 & 3 & -2 \\ 1 & 3 & 3 \\ 1 & 2 & 3 \\ 1 & 1 & 3 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

Now create a 3×1 matrix β whose entries are 1, -1, and 2. Next create a 10×1 matrix ϵ whose entries are IID standard normal (useful command: "rnorm"). Finally, set $Y = X\beta + \epsilon$.

1. Calculate $(X'X)^{-1}X'Y$ to estimate β. What do you get? (Don't use the "lm" command. Do the computation directly. You can use the "solve" command to compute a matrix inverse.)

2. What is the true variance of $\hat{\beta}$? (Remember that the variance of $\hat{\beta}$ is a 3×3 matrix.) (I say the "true" variance because, in this example, we know the true value of σ^2, and so don't need to estimate it using the residuals.)

3. Use the residuals to estimate σ^2. What do you get?

4. Now create a new ϵ and re-estimate β. Do this 1,000 times, and save all the answers in memory. Make a histogram of the $1,000$ values of $\hat{\beta}_1$. Do the same for $\hat{\beta}_2$ and $\hat{\beta}_3$. Also calculate the variance for each of these. Do your answers match with question 2?

5. Once again, re-create ϵ 1,000 times. Each time estimate β, and also estimate σ^2, too. Make a histogram of your 1,000 values of $\hat{\sigma}^2$. Based on the histogram, does it look like $\hat{\sigma}^2$ provides a reliable estimate of σ^2? Why do you think this is?

6. Repeat (4) and (5), but instead of using a normal distribution for ϵ use some other distribution that also has expectation 0 and variance 1. Do your answers change much? Explain. You might want to experiment with a few different distributions.

STATS 500 – Homework 3

Due in class Thursday, February 11

Part A: Using the `sat` data (see `help(sat)` for the description of variables):

1. Fit a model with `total` sat score as the response and `takers, ratio` and `salary` as predictors. Comment on the goodness of fit. Using this model, answer the questions 2-6:
2. Suppose you wish to claim that teachers' salary has a positive effect on the SAT scores. State the appropriate null and alternative hypothesis, the test statistic, the p-value, and your conclusion at significance level $\alpha = 0.01$.
3. Test the hypothesis that the variable `ratio` has an effect on the SAT scores in the full model.
4. Test the hypothesis $\beta_{takers} = \beta_{ratio} = \beta_{salary} = 0$. Explain in words what this hypothesis means.
5. Compute the 95% and 99% CIs for the parameter associated with `salary`. Using just these intervals, what can we deduce about the p-value for `salary` in the regression summary?
6. Compute and display a 95% joint confidence region for the parameters associated with `ratio` and `salary`. Add the origin to the plot. The location of the origin on the plot tells us the outcome of a certain hypothesis test. State that test and its outcome.
7. Now add `expend` (current expenditure per pupil) to the model and comment on the coefficients, their significance and the goodness of fit as compared to the model in question 1.
8. In the model of question 7, test the hypothesis $\beta_{salary} = \beta_{expend} = \beta_{ratio} = 0$. Based on your entire analysis, do you feel any of these predictors have an effect on the response?

   Solutions to Part A should be no longer than 3 pages.

Part B: Using the `teengamb` dataset:

Fit a model to predict gambling expenditure from all other available variables. Perform regression diagnostics on this model to answer the following questions. Display **only** those plots that are relevant to the questions below. Present your diagnostics in a logical order.

- Check the constant variance assumption for the errors. Modify the model if necessary (see below).
- Check the normality assumption.
- Check for large leverage points.
- Check for outliers.
- Check for influential points.
- Check the structure of the relationship between the predictors and the response.

Hints: You should start with a linear regression of `gamble` on `sex, status, income` and `verbal`. A diagnostic plot will reveal heteroscedasticity in residuals. A standard solution for the type of heteroscedasticity that you will see is to take the log or the square root of the response. Since in this case the `gamble` variable has some zero values, take the square root of `gamble` and fit a new model for this response. Then do all your diagnostic analysis for the new model.
Solutions to Part B should be no longer than 6 pages.

Part C: Confidence regions:

Consider a regression in which we regress Y on two predictor variables, A and B. For simplicity, assume that A and B are both standardized (mean 0, variance 1). Suppose we draw a joint confidence region for β_A and β_B. Explain, in your own words, why the confidence region will "lean to the left" if A and B are positively correlated, and will "lean to the right" if A and B are anti-correlated.
Hint: In each case (correlated and anti-correlated), think about whether it is easier to estimate $\beta_A + \beta_B$ or $\beta_A - \beta_B$. It might help to think about the case in which A and B are *perfectly* correlated or anti-correlated. Then think about what a confidence region would like like if $\beta_A + \beta_B$ is easy to estimate but $\beta_A - \beta_B$ is difficult to estimate, and vice versa.

STATS 500 - Homework 4

Due Thursday, March 10

For the `longley` data, fit a model with `Employed` as the response and the other variables as predictors.

1. Compute and comment on the condition numbers.
2. Compute and comment on the correlations between the predictors.
3. Compute and comment on the variance inflation factors.
4. Choose a reduced set of predictors that does not exhibit as much collinearity as the full set, fit a new linear model with this reduced set, and comment on the differences between the reduced model and the full model.

Solutions to this homework should not exceed 3 pages.

STATS 500 - Homework 5

Due **Tuesday**, March 22

1. Using the `sat` data, fit a model with `total` as the response and `takers, ratio, salary` and `expend` as predictors using the following methods:

   (a) Ordinary least squares

   (b) Least absolute deviations

   (c) Huber's robust regression

   Compare the results. In each case, comment on the significance of predictors.

2. Use the `prostate` data with `lpsa` as the response and the other variables as predictors. Implement the following variable selection methods to determine the "best" model:

   (a) Backward Elimination

   (b) Adjusted R^2

   (c) Mallows' C_p

   Comment on the models selected (similarities and/or differences). Compare the fits of the full model and those selected by the methods above.