Linear Models. List 4

Problems to be solved by hand

1. You use the data to estimate parameters of the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \ .$$

Your estimators are $b_0 = 1, b_1 = 4, b_2 = 3, s = 3$.

a) Predict the value of Y for $X_1 = 2$ and $X_2 = 6$.

b) The estimated standard deviation of the estimator of the expected value of Y for $X_1 = 2$ and $X_2 = 6$ is equal to 2. Estimate the variance of the error of your prediction $\sigma^2(pred)$.

c) The above model was fitted using 20 observations and the estimated standard deviation of b_1, $s(b_1)$, is equal to 1. Construct 95 % confidence interval for β_1.

2. We analyze the data using the following multiple regression model :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \ .$$

Below you have a table with Type I and Type II sums of squares.

	$TypeI$	$TypeII$
X_1	300	30
X_2	40	25
X_3	20	?

Corrected total sum of squares (SST) is equal to 760 and $n = 24$.

(a) What is Type II sum of squares for X_3 equal to ?

(b) Test the hypothesis that $\beta_1 = 0$ (in the full model).

(c) Test the hypothesis that $\beta_2 = \beta_3 = 0$.

(d) Test the hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$.

(e) You decided to drop X_2 and X_3 from your model. Test the hypothesis that $\beta_1 = 0$ in the simple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \ .$$

(Be careful - SSE and dfE are different than for the model with three variables.)

(f) Compute the sample correlation coefficient between Y and X_1.

Simulation study

3. **Influence of correlation**.

   a) Generate the matrix $X_{100\times 2}$ such that its rows are iid random vectors from the multivariate normal distribution $N(0, \Sigma/100)$, where

   $$\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix} .$$

   Then generate the vector of response variable as $Y = \beta_1 X_1 + \epsilon$, where $\beta_1 = 3$, X_1 is the first column of X and $\epsilon \sim N(0, I)$.

   b) Construct the 95% confidence interval for β_1 and perform the 0.05 significance level t-test for the hypothesis $\beta_1 = 0$ using the model of the simple linear regression $Y = \beta_0 + \beta_1 X_1 + \epsilon$ and using the model with both explanatory variables $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$. Compare and explain the results.

   c) Calculate by hand the standard deviations of the estimate of β_1 and the power of identification of X_1 in both models.

   d) Generate 1000 independent copies of the vector of errors ϵ and 1000 related copies of the vector of response variable. For each of such data sets estimate β_1 and perform the test for the significance of β_1 in both models (with one and two explanatory variables). Estimate the standard deviation of β_1 and the power of the test and compare these values to the theoretical results obtained in point c).

4. **Influence of dimension.** Generate the design matrix $X_{1000\times 950}$ such that its elements are iid random variables from $N(0, \sigma = 0.1)$. Then generate the vector of the response variable according to the model

   $$Y = X\beta + \epsilon$$

   , where $\beta = (3, 3, 3, 3, 3, 0, ..., 0)^T$.

   a) Estimate values of regression coefficients and use t-tests at the significance level 0.05 to identify true regressors when the model is build using first

      i) 1
      ii) 2
      iii) 5
      iv) 10
      v) 50
      vi) 100
      vii) 500
      viii) 950

      columns of the design matrix. For each of these models report residual sum of squares $||Y - \hat{Y}||^2$, mean square error of estimating the expected value of Y, $MSE = ||X(\hat{\beta} - \beta)||^2$, value of AIC criterion, p-values corresponding to the first two explanatory variables and the number of false discoveries. Which model would be selected based on AIC ?

   b) Repeat point a) when the models are build using variables with the largest (not the first) estimated regression coefficients. Compare values of calculated statistics with those obtained in point a). Which model would be selected based on AIC ?

c) Using the fact that the random matrix $X_s^T X_s$ has a Wishart distribution calculate the expected values of the elements on the diagonal of $(X_s^T X_s)^{-1}$, where X_s contains first s columns of X, and s takes values specified in point a). Compute also the power of identification of X_1 and the expected number of false discoveries produced by 0.05 significance t-tests, where the expected values are taken with respect to the distribution of ϵ and X.

d) Repeat the generation of X and Y and the steps a) and b) 1000 times. For each of the sub-problems estimate the power of identification of X_1 and the expected number of the false discoveries. Compare these results to the theoretical values calculated in point c). Additionally, compare the average model size selected by AIC for points a) and b).

Real data analysis

The first data set is the patient satisfaction data from the file CH06PR15.txt. The consecutive columns contain: patient's age (first column) and scores describing the severity of illnes, the level of patient's anxiety and the level of patient's satisfaction.

5. . Run the linear regression with age, severity of illness and anxiety level as the explanatory variables and satisfaction as the response variable. Summarize the results by giving the fitted regression equation, the value of R^2, and the result of significance test for the null hypothesis that the three explanatory variables are not associated with the response (give null and alternative hypotheses, test statistic with degrees of freedom, P-value, and a brief conclusion in words).
6. Give separate 95% confidence intervals for regression coefficients of age, severity of illness and anxiety level. Describe the results of the hypothesis tests for the individual regression coefficients (give null and alternative hypotheses, test statistic with degrees of freedom, P-value, and a brief conclusion in words). What is the relationship between these results and the confidence intervals?
7. Plot the residuals versus the predicted satisfaction and each of explana- tory variables. Do you see any unusual patterns or outliers?
8. Are residuals approximately normal ? Use the Shapiro-Wilk test (R function - shapiro.test) and qqplot.

The second data set is the computer science data that we discussed in class. The file name is csdata.dat. The variables are: id, a numerical identifier for each student; GPA, the grade point average after three semesters; HSM; HSS; HSE; SATM; SATV; and SEX, coded as 1 for men and 2 for women.

9. a) Run the following two regressions:
   (i) predict GPA using HSM, HSS and HSE;
   (ii) predict GPA using SATM, SATV, HSM, HSS and HSE.
   Take the difference between SSEs for the two analyses and construct the F statistic for testing the null hypothesis that the coeficients by the two SAT variables are both equal to zero (in the model with all five predictors) .
   b) Use the anova function to calculate the same test statistic. Give the statistic, degrees of freedom, P value and conclusion.
10. Run the regression to predict GPA using SATM, SATV, HSM, HSE and HSS. Put the variables in the order given above on the model statement. Calculate type I and type II sums of squares.

a) Verify (by running additional regressions) that the TYPE I sum of squares for the variable HSM is the difference in the model sum of squares for the two analyses:
(i) regression of GPA on SATM, SATV and HSM;
(ii) regression of GPA on SATM and SATV.

b) Are there any predictors for which SS1 and SS2 are the same ? Explain why.

11. Create a new variable (name it SAT) that is the sum of the two SAT scores. Run the regression to predict GPA using three variables: SATM, SATV and SAT. Describe the output you obtain and explain.

12. Run the regression of GPA on the explanatory variables HSM, HSS, HSE, SATM, SATV and SEX. Examine the partial regression plots. Give a short explanation of what they are and what kind of information they convey. Are there any interesting patterns or observations ?

13. Examine the studentized deleted residuals. Do there appear to be any outliers ?

14. Explain DFFITS. What do you conclude about this data set from an examination of this statistic?

15. What is the tolerance and how is it used ? What do you conclude about this data set from an examination of this statistic ?

16. Select the best regression model using BIC and AIC.

Małgorzata Bogdan