Project: STK4170-f09: Bootstrapping and resampling.

The data sets you need can be found on the course web page, together with some R-code you may find helpful.

You can use a statistical package at your choice for doing the numerical computations. However, if you use another than R, you must remember that lacking capabilities in the software you use, is no valid excuse for not answering a question.

Explain carefully the calculations and computations. It is recommended to split the report in two. In the first part you concentrate on explaining what you are doing and answering the questions. This part shall not contain more numerical output and computer code than what is strictly necessary for the arguments. In the second part you can collect the output to which you can refer in the main part of the report.

For the report you can use a word processor, but a handwritten report is also acceptable.

Discussing the problem with other students is OK, but the papers you turn in shall be **individually** written. If you cooperate with someone, you must inform in the report with whom.

The deadline is

Wednesday December 9th at 5 pm.

You can email the report to swensen@math.uio.no, deliver it at my office or use my mail box at the 7th floor, N. H. Abels house.

In the morning Thursday December 10th I will put a list on the course web page of the reports I have received. If something is missing, I must be notified immediately. At the same time I will announce the schedule for the oral examination.

I will also post misprints, corrections and clarifications on the web page, so please inform me of any obscurities.

Good luck,

Anders

Problem 1, (based on DH: 2.20.10 and 6.7.3)

The linear regression model with no intercept is defined as

$$Y_j = \beta x_j + \epsilon_j, j = 1, \ldots, n$$

where x_j are known and $\epsilon_j, j = 1, \ldots, n$ are independent, identically distributed variables with $E[\epsilon_j] = 0$ and $var[\epsilon_j] = \sigma^2$.

Suppose $(y_1, x_1), \ldots, (y_n, x_n)$ are observations from this model.

a) Show that the least squares estimate for β is $\hat{\beta} = \sum x_j y_j / \sum x_j^2$.

Define the residuals $e_j = y_j - \hat{\beta} x_j, j = 1, \ldots, n$, and the resampled observations $y_j^* = \hat{\beta} x_j + \epsilon_j^*, j = 1, \ldots, n$, where $\epsilon_j^*, j = 1, \ldots, n$ are sampled with replacement from the residuals $e_j, j = 1, \ldots, n$

b) Show that the estimate $\hat{\beta}^*$ computed from the resampled observations has expectation and variance respectively

$$E^*(\hat{\beta}^*) = \hat{\beta} + \frac{n\bar{x}\bar{e}}{\sum x_j^2} \quad var^*(\hat{\beta}^*) = \frac{\sum (e_j - \bar{e})^2}{n \sum x_j^2}$$

where $\bar{x} = \sum x_j/n$ and $\bar{e} = \sum e_j/n$.

c) What are the corresponding results if the resampled observations $y_j^* = \hat{\beta}x_j + \epsilon_j^*, j = 1, \ldots, n$ are defined with $\epsilon_j^*, j = 1, \ldots, n$ sampled with replacement from the *centered* residuals $e_j - \bar{e}, j = 1, \ldots, n$?

Problem 2, (based on DH: 6.7.1)

In this problem you shall consider a situation with a bivariate random variable having cumulative distribution function F_{XY} where marginal distribution functions are denoted by F_X and F_Y. Let μ_x and μ_y be the expectations of X and Y and σ_x^2, σ_y^2 and σ_{xy} the variances and covariance.

In the first part the marginal distribution of X is treated.

a) Consider the statistic $t_1(F_X) = var(X) = \sigma_X^2$, and find the influence function $L_{t_1}(x; F_X)$.

b) Now consider the statistic $t_2(F_X) = \sqrt{var(X)} = \sigma_X$. What is the influence function of t_2, $L_{t_2}(x; F_X)$?

c) What is the nonparametric delta method variance estimate for $t_2(F_X) = \sigma_X$?

Now, consider the simultaneous distribution of the random variables X and Y.

d) The regression coefficient, β_1, of Y on X is defined as $\beta_1 = \sigma_{xy}/\sigma_x^2$. What is the influence function of β_1, $L_{\beta_1}((x, y); F_{XY})$?

[Hint: It may be useful to write $\beta_1 = \rho\sigma_y/\sigma_x$, where ρ is the correlation coefficient between X and Y. Then you can use the result from part a) together with the expression for the influence function of the correlation coefficient from Example 2.18 in Davison and Hinkley.]

Problem 3

The following 14 pairs of observations describe measurements of hydrogen content and gas porosity for a certain technique for producing aluminium alloy castings.

content	0.18	0.20	0.21	0.21	0.21	0.22	0.23
porosity	0.46	0.70	0.41	0.45	0.55	0.44	0.24
content	0.23	0.24	0.24	0.25	0.28	0.30	0.37
porosity	0.47	0.22	0.80	0.88	0.70	0.72	0.75

In this problem you shall apply bootstrapping methods to investigate the distribution of the empirical correlation coefficient between gas content and porosity.

Recall that if the population correlation is ρ, the empirical correlation coefficient $\hat{\rho}$ has an asymptotic distribution where $\hat{\rho} - \rho$ is approximately $N(0, (1 - \rho^2)^2/n)$.

a) Describe an appropriate algorithm for constructing a bootstrap sample for testing

$$H_0 : \rho = 0 \text{ against } H_A : \rho > 0.$$

What is the P-value based on $R = 999$ replications?

b) Construct a basic bootstrap 95% confidence interval for ρ. Find the upper and lower value, and indicate the value of the quantiles that you use. Use $R = 999$ replications.

c) Do the same as in part b), but for the studentized bootstrap confidence interval.

We shall now consider the Fisher-transformation $\frac{1}{2}\log\frac{1+\hat{\rho}}{1-\hat{\rho}}$.

d) Construct a studentized bootstrap 95% confidence interval for ρ making use of the Fisher-transformation. Find the upper and lower value, and indicate the value of the quantiles that you use, still based on $R = 999$ replications.

e) Make a scatterplot of the replicated empirical correlation coefficients vs. the replicated variance estimates. Compare this scatterplot with a scatterplot of the replicated empirical transformed correlation coefficients vs. the replicated variance estimates of the transformed correlation. Explain the difference.