Foundations of Data Science I Exam

1 Assignments on part I and II of the course

1.1 Exercise 1: Finite Sample Properties of Ridge Estimator

Given a linear model $Y_i = X_i\beta_0 + \epsilon_i$, with $X_i \in \mathbb{R}^K$, the ridge estimator is defined as follows

$$\hat{\beta}_n^R := \arg\min_{\beta \in \mathbb{R}^k} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \frac{\lambda_n}{2} \|\beta\|_2^2 \right\}, \tag{1}$$

where n is the number of observations and $\lambda_n \in \mathbb{R}_+$ is a penalty parameter.

1. **Monte Carlo simulation of OLS and Ridge estimators:** Simulate observations for $\{X_i, Y_i\}$ following the linear model with specifications given by $X_i \sim N(0, I_K)$, $\epsilon \sim N(0, \sigma^2 I_n)$, $\sigma^2 = .5$, $n = 50$, $K = 5$ and $\beta_0 = [0.1, .05, 0.2, 0.9, 0.5]$.

   - For various simulations compute the OLS and Ridge estimators, $\hat{\beta}_n^{OLS}$ and $\hat{\beta}_n^R$, for penalty parameter $\lambda_n = 0.1 \times n^{1/3}$.
   - Show that the two estimators satisfy the following relation:

     $$\hat{\beta}_n^R = \boldsymbol{W}_n(\lambda_n)\,\hat{\beta}_n^{OLS},$$

     where $\boldsymbol{W}_n(\lambda_n) = (\boldsymbol{I}_K + \lambda_n \boldsymbol{Q}_n^{-1})^{-1}$, $\boldsymbol{Q}_n = \boldsymbol{X}'\boldsymbol{X}/n$ and $\boldsymbol{I}_K$ is a $K \times K$ identity matrix.
   - Plot the histogram of the OLS and the Ridge estimators in your simulation. What can you observe about the bias and variance of the estimators?
   - Show that the Ridge estimator has lower variance than the OLS estimator, i.e.,

     $$\mathbb{V}ar\left(\hat{\beta}_n^{OLS} \mid \boldsymbol{X}\right) \succ \mathbb{V}ar\left(\hat{\beta}_n^R \mid \boldsymbol{X}\right).$$

     Hint: This is an inequality between matrices!

2. **Penalty parameter choice by minimizing the mean square error**

   - Define the parameterized mean square error function $F(\alpha) := \mathbb{E}\left[\left\|\hat{\beta}_n^R - \beta_0\right\|_2^2\right]$, where $\hat{\beta}_n^R$ is the ridge estimator with penalty parameter $\lambda_n = \alpha \times n^{1/3}$. Using similar simulation specifications as above produce a plot of the function F for values of α between 0.01 and 0.1.
   - Let us now define a new function G as follows:

     $$G(\alpha) := \mathbb{E}\left[\text{Trace}\left(\boldsymbol{W}_n(\alpha n^{1/3})\left(\frac{\sigma^2}{n}\boldsymbol{Q}_n^{-1} + \alpha^2 n^{2/3}\boldsymbol{Q}_n^{-1}\beta_0\beta_0{}'\boldsymbol{Q}_n^{-1}\right)\boldsymbol{W}_n(\alpha n^{1/3})\right)\right]$$

     Compute by simulation function G, plot it and show that it is equal to the mean square function F. Explain your finding using the Ridge estimation theory developed in the course.
   - Substitute the parameters β_0 and σ^2 in function G with the OLS estimator $\hat{\beta}_n^{OLS}$ and the error variance estimator $\hat{\sigma}_n^2$, in order to optimize the square error in each simulation by choosing the

optimal penalty parameter. For this purpose, in each simulation compute:

$$\hat{\lambda}_n^{opt} = \arg\min_{\lambda \geq 0} \text{Trace}\left(W_n(\lambda)\left(\frac{\hat{\sigma}_n^2}{n}\boldsymbol{Q}_n^{-1} + \lambda^2 \boldsymbol{Q}_n^{-1}\hat{\beta}_n^{OLS}\hat{\beta}_n^{OLS\prime}\boldsymbol{Q}_n^{-1}\right)W_n(\lambda)\right)$$

where

$$\hat{\sigma}_n^2 = \frac{\left\|\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}_n^{OLS}\right\|_2^2}{n-K}.$$

Using such optimal penalty parameter compute the Ridge estimator and show that it delivers a mean square error similar to the minimum of functions F and G.

1.2 Exercise 2: Nonlinear regression with measurement errors

A researcher considers the (nonlinear) regression model:

$$y_i = h\left(x_i^*, \beta_0\right) + \varepsilon_i,$$

where β_0 is a scalar parameter and h is a given function (specified below). The explanatory variable is observed with a measurement error:

$$x_i = x_i^* + u_i.$$

The available data for the researcher are $(y_i, x_i), i = 1, \cdots, n$. Moreover, we assume:

$$\left(x_i^*, \varepsilon_i, u_i\right)' \sim IIN\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{x^*}^2 & 0 & 0 \\ 0 & \sigma_\epsilon^2 & 0 \\ 0 & 0 & \sigma_u^2 \end{pmatrix}\right]$$

1. **Misspecified estimation**

   Suppose first that the researcher neglects the measurement errors. We want to study the consequences of this choice. To simplify, let us assume in this part of the exercise that $h\left(x_i^*, \beta_0\right) = x_i^*\beta_0$, that is, the regression model is linear. The researcher proposes the estimator:

   $$\hat{\beta} = \arg\min_\beta \sum_{i=1}^n \left(y_i - x_i\beta\right)^2.$$

   - Compute:

     $$\beta_0^* = \arg\min_\beta E_0\left[\left(y_i - x_i\beta\right)^2\right]$$

     where $E_0[\cdot]$ denotes expectation w.r.t. the true distribution of the data. Is the estimator $\hat{\beta}$ consistent for β_0?

   - Derive the asymptotic distribution of the M-estimator $\hat{\beta}$.

2. **NLS estimation**

   Suppose now that the researcher properly accounts for measurement errors. Let us assume that $h\left(x_i^*, \beta_0\right) = \left(x_i^* + \beta_0\right)^2$.

   - Show that:

     $$E_0\left[y_i \mid x_i\right] = \left(x_i + \beta_0\right)^2 + \sigma_u^2$$

- Propose a consistent NLS estimator of β_0, called $\hat{\beta}_{NLS}$. Derive the asymptotic distribution of $\hat{\beta}_{NLS}$.

1.3 Exercise 3: PML estimation in a duration model

The positive variables (durations) y_i are generated by the model

$$y_i = (\beta_0 x_i)\, \varepsilon_i, \quad i = 1, \dots, n,$$

where:

- regressors x_i and errors ε_i are positive and such that $(x_i, \varepsilon_i) \sim i.i.d$, where x_i and ε_i are independent with
$$E\left[\varepsilon_i\right] = 1$$
- $\beta_0 > 0$ is an unknown scalar parameter.

In this exercise, we consider different pseudo-models for a PML estimation of β.

1. Conditional expectation
   - Prove that
   $$E_0\left[y_i \mid x_i\right] = \beta_0 x_i$$
   where $E_0\left[.\right]$ denotes expectation under the true model.
2. PML with exponential density
   - The econometrician considers the following parametric family of pseudodensities
   $$f(y \mid x; \beta) = \frac{1}{\beta x} e^{-\frac{y}{\beta x}}, \quad y \geq 0, \tag{2}$$
   with parameter $\beta > 0$.
   - Compute the pseudo-true value defined by
   $$\beta_0^* = \arg\max_\beta E_0\left[\log f\left(y_i \mid x_i; \beta\right)\right]$$
   and show that $\beta_0^* = \beta_0$.
   - Is this result surprising? Explain it using the general PML theory derived in the course!
   - Compute the PML estimator of β based on family (2):
   $$\hat{\beta} = \arg\max_\beta \sum_{i=1}^{n} \log f\left(y_i \mid x_i; \beta\right)$$
   and show that
   $$\hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i}.$$
   Is estimator $\hat{\beta}$ consistent? Give its asymptotic distribution.

3. PML with Weibull density

   - Let us now consider the density function on $\mathbb{R}_+$:

   $$g(\varepsilon) = \frac{2\varepsilon}{c} e^{-\frac{\varepsilon^2}{c}}, \quad \varepsilon \geq 0,$$

   where $c := 4/\pi$. It is possible to show (you don't have to show this!) that $\int_0^\infty g(\varepsilon) d\varepsilon = 1$ and $\int_0^\infty \varepsilon g(\varepsilon) d\varepsilon = 1$.

   - Explain why

   $$\tilde{f}(y \mid x; \beta) = \frac{2y}{c(\beta x)^2} e^{-\frac{1}{c}\left(\frac{y}{\beta x}\right)^2}, \quad y \geq 0, \tag{3}$$

   with $\beta > 0$, defines a parametric family of conditional densities with correctly specified mean for our problem.

   - The PML estimator of β based on family (3) is

   $$\tilde{\beta} = \arg\max_{\beta} \sum_i \log \tilde{f}\left(y_i \mid x_i; \beta\right).$$

   Is $\tilde{\beta}$ a consistent PML estimator for estimating β_0? Explain your answer using the general PML theory derived in the course.

2 Assignments on part III of the course

2.1 Exercise 1: Jackknife

The aim of this project is to estimate the variance-covariance matrix of the OLSE of the parameters in linear regression by the jackknife method. Let's assume the linear model

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \ldots, n$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$ and ϵ_i are independent identically distributed random variables with some distribution F such that $E(\epsilon_i) = 0$. In matrix notation the model can be written as

$$\begin{array}{rrcl} & \mathbf{y} & = & \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \\ \text{where} & \mathbf{y} & = & (y_1, \ldots, y_n)', \\ & \boldsymbol{\epsilon} & = & (\epsilon_1, \ldots, \epsilon_n)', \\ & \mathbf{X} & = & \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} \in \mathbb{R}^{n \times p}. \end{array}$$

Let $\hat{\boldsymbol{\beta}}$ be the OLSE of $\boldsymbol{\beta}$ and let $\hat{\epsilon}_i = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}$ be the *ith* residual.

1. Show that

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i \hat{\epsilon}_i^m$$

using the equalities

$$\begin{aligned} \mathbf{X}'_{(i)}\mathbf{y}_{(i)} &= \mathbf{X}'\mathbf{y} - \mathbf{x}_i y_i \ , \\ (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1} &= (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1-m_{ii}}, \end{aligned}$$

where $m_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ is the *ith* diagonal element of the projection matrix $\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\hat{\epsilon}_i^m = \frac{\hat{\epsilon}_i}{1-m_{ii}}$.

2. Give the interpretation of the difference $\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}$.

3. Show that the jackknife estimator of the variance-covariance matrix of (the random vector) $\hat{\boldsymbol{\beta}}$ is given by

$$\begin{aligned} \hat{V}_{Jack}(\hat{\boldsymbol{\beta}}) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{\boldsymbol{\beta}}^{*i} - \hat{\boldsymbol{\beta}}^{*\cdot})(\hat{\boldsymbol{\beta}}^{*i} - \hat{\boldsymbol{\beta}}^{*\cdot})' \\ &= \frac{n-1}{n}(\mathbf{X}'\mathbf{X})^{-1}\left[\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'(\hat{\epsilon}_i^m)^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\mathbf{x}_i\hat{\epsilon}_i^m\right)\left(\sum_{i=1}^{n}\mathbf{x}_i'\hat{\epsilon}_i^m\right)\right](\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

   where the $\hat{\boldsymbol{\beta}}^{*i}$ are the pseudo-values and $\hat{\boldsymbol{\beta}}^{*\cdot}$ their mean.

4. Consider an approximation of the jackknife estimator obtained at point 3. by replacing $1 - m_{ii}$ by 1. When is this approximation justified? (Compute the average value of the m_{ii}, *i.e.*$\frac{1}{n}\sum_{i=1}^{n} m_{ii}$.)

5. Give the formula of jackknife estimator of point 3. obtained by replacing $1 - m_{ii}$ by 1 and verify that this estimator is the estimator proposed by H. White (1980), *A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity*, Econometrica, pp. 817-838. Give the exact location in that paper where we find this estimator.

6. Read p. 817, 820 (first half), and p. 821 (first half) of the cited article and summarize in at most one page the problem and the proposed solution.

7. Compare the jacknife estimator derived at point 3. with a bootstrap estimator.

2.2 Exercise 2: Wild bootstrap

The aim of this project is to implement the so-called "wild bootstrap" procedure to assess the uncertainty of OLS estimates of a linear model, and compare it with the uncertainty assessments of "paired bootstrap" and "residual bootstrap" procedures. Data containing information on medical expenses (E), standardized income (I) and smoking habit (S, taking value 1 for a smoker and 0 for a non smoker) can be found in the file "medical.csv" located in the exam folder. Create and precisely document a Python code performing in sequence the following tasks:

1. Read the csv file into your Python code. Obtain basic informations on the data types of each variable, the number of observations, the number of missing values.

   (*Hint:* functions from the Pandas library can ease this task.)

2. Compute summary statistics of each variable, visualize their correlation matrix and plot their histogram and pairwise scatter plots.

3. Create a function computing the OLS estimates, R^2 and residuals of linear regression model

   $$E_i = a + bI_i + cS_i + \varepsilon_i \ , \quad i = 1, \ldots, n, \tag{4}$$

   and obtain these estimates.

4. Produce the pairwise scatter plots of the residuals versus the explanatory variables. What do you notice?

   (*Hint:* do you notice any heteroskedasticity with respect to a specific explanatory variable?)

5. Define three functions computing the bootstrap estimates of the model coefficients and R^2 for the paired bootstrap, the residual bootstrap and the wild bootstrap procedures according to the algorithms reported below.[1]

6. Compute the bootstrap estimates of the coefficients and the R^2 of model (4) for these three bootstrap schemes. Visualize and compare the histograms of the bootstrap distributions of the coefficients and the R^2. What do you notice?

   (*Hint:* The histograms of the wild bootstrap distributions resemble the ones of the paired bootstrap or the ones of the residual bootstrap?)

7. Compute and visualize the confidence intervals for the model coefficients a, b and c based on the three bootstrap schemes. Which one/ones would you trust and why?

Algorithm 1 Paired bootstrap

```
set the random number generation seed and B (the number of bootstrap samples)
for b=1,...,B do
    create a sample $\{(E_i^*, I_i^*, S_i^*)\}$ of size $n$ drawing with replacement from $\{(E_i, I_i, S_i)\}$
    compute the OLS estimates of the model coefficients and the $R^2$ for this bootstrap sample
end for
return all bootstrap coefficients and $R^2$ estimates
```

Algorithm 2 Residual bootstrap

```
set the random number generation seed and B
for b=1,...,B do
    create a sample $\{\hat{\varepsilon}_i^*\}$ of size $n$ drawing with replacement from $\{\hat{\varepsilon}_i\}$, the estimated residuals
    compute the OLS estimates of the coefficients and the $R^2$ for the model
        $E_i^* = a + bI_i + cS_i + \hat{\varepsilon}_i^* \ , \quad i = 1, \ldots, n$
end for
return all bootstrap coefficients and $R^2$ estimates
```

[1] The residual bootstrap assumes that $\mathbb{E}[E_i|I_i, S_i] = a + bI_i + cS_i$ and that the error terms ε_i are IID and homoskedastic. The paired bootstrap only assumes that there exists a joint probability distribution of $\{(E_i, I_i, S_i)\}$ and it makes no assumptions about the properties of the error terms ε_i or about the functional form of $\mathbb{E}[E_i|I_i, S_i]$. The wild bootstrap lies in between these two bootstrap schemes, as it assumes that $\mathbb{E}[E_i|I_i, S_i] = a + bI_i + cS_i$, but it allows for heteroskedasticity by conditioning on the (transformed) residuals.

Algorithm 3 Wild bootstrap

```
set the random number generation seed and B
for b=1,...,B do
    create a sample $\{\hat{\varepsilon}_i^*\}$ of size $n$ where $\hat{\varepsilon}_i^* = f(\hat{\varepsilon}_i)v_i$ with
        $f(\hat{\varepsilon}_i) = \sqrt{n/(n-K)}\hat{\varepsilon}_i$ ,
    $K$ being the number of coefficients, and
        $v_i = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases}$
    compute the OLS estimates of the coefficients and the $R^2$ for the model
        $E_i^* = a + bI_i + cS_i + \hat{\varepsilon}_i^* \ , \quad i = 1, \dots, n$
end for
return all bootstrap coefficients and $R^2$ estimates
```