HOMEWORK 2

1. 证明题

(1) Given conditions:

(A1) The relationship between response ($\mathbf{y}$) and covariates ($\mathbf{X}$) is linear;

(A2) $\mathbf{X}$ is a non-stochastic matrix and $\text{rank}(\mathbf{X}) = p$;

(A3) $E(\boldsymbol{\varepsilon}) = \mathbf{0}$. This implies $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$;

(A4) $\text{cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \boldsymbol{I}_N$; (Homoscedasticity);

(A5) $\boldsymbol{\varepsilon}$ follows multivariate normal distribution $N(\mathbf{0}, \sigma^2 \boldsymbol{I}_N)$ (Normality).

Prove the following results:

(1.1) Prove that the OLS estimator $\widehat{\boldsymbol{\beta}}$ is the same as the maximum likelihood estimator.

(1.2) Prove

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}) \tag{1}$$

$$(N-p)\widehat{\sigma}^2 \sim \sigma^2\chi^2_{N-p} \tag{2}$$

(2) Suppose y follows the log-linear regression relationship with non-stochastic $x \in \mathbb{R}^p$, i.e.,

$$\log(y) = x^\top\beta + \epsilon, \tag{3}$$

where ϵ follows normal distribution $N(0, \sigma^2)$. Please calculate $E(y)$.

(3) Let y_i be the dependent variable, $\boldsymbol{x}_i$ be the vector of independent variables including an intercept term, and $\widehat{\boldsymbol{\beta}}$ be the vector of regression coefficients estimated

by OLS. Define $\widehat{y}_i = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}$. Define the total sum of squares (TSS), explained sum of squares (ESS) and residual sum of squares (RSS) as follows

$$\mathrm{TSS} = \sum_i (y_i - \overline{y})^2, \quad \mathrm{ESS} = \sum_i (\widehat{y}_i - \overline{y})^2, \quad \mathrm{RSS} = \sum_i (y_i - \widehat{y}_i)^2.$$

Please prove: TSS = ESS + RSS.

2. 岭回归分析： 在实际问题中, 我们常常会遇到样本容量相对较小, 而特征很多的场景. 在这类情况中如果直接求解线性回归模型, 较小的样本无法获得唯一的模型参数, 会具有多个模型能够“完美”拟合训练集中的所有数据点. 此外, 模型很容易过拟合. 为缓解这些问题, 常在线性回归的损失函数中引入正则化项 $p(\beta)$, 通常形式如下:

$$\widehat{\beta}^p = \arg\min_\beta \left\{ \sum_{i=1}^N \left(y_i - \sum_j x_{ij}\beta_j \right)^2 + \lambda p(\beta) \right\} \tag{4}$$

其中, $\lambda > 0$ 为正则化参数. 正则化表示了对模型的一种偏好, 例如 $p(\beta)$ 一般对模型的复杂度进行约束, 它在保持良好预测性能的同时，倾向于选择较为简单的模型，从而帮助防止过拟合并提高模型的泛化能力。考虑岭回归 (ridge regression) 问题, 即设置公式(4)中正则项 $p(\beta) = \sum_j \beta_j^2$. 本题中将对岭回归的显式解以及正则化的影响进行探讨.

(1) 请证明岭回归的最优解 $\widehat{\beta}^{\text{ridge}}$ 的显示解表达式具有以下两种等价形式

$$\widehat{\beta}^{\text{ridge}} = \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_N\right)^{-1} \boldsymbol{y}.$$

请分析以上两种最优解分别在何种情况下计算速度更快?

(2) 分析岭回归的最优解 $\widehat{\beta}^{\text{ridge}}$ 和最小二乘估计 $\widehat{\beta}^{\text{LS}}$ 的区别;

(3) 针对附录中描述的北京租房数据，完成以下任务

(3.1) 完成数据读入与汇总统计，绘制训练集数据中月租金（rent）的直方图，观察月租金的大致分布，并进行简要解读；绘制训练集数据中月租金（rent）-城区（region）分组箱线图，分析不同城区的房价差异，并给出简要解读

(3.2) 利用训练集建立以月租金（rent）为因变量，其余为自变量的线性回归模型，编程实现最小二乘估计（不调用回归分析的包），写出拟合得到的模型并计算测试集上的均方误差 (Mean Square Error, MSE)

(3.3) 编程实现岭回归估计（不调用回归分析的包），在训练集上使用十折交叉验证，画出验证集上平均均方误差 (Mean Square Error, MSE) 与 λ 的折线图，选出合适的 λ

(3.4) 用选出的 λ 在训练集拟合最终模型，写出拟合得到的模型并计算测试集上的均方误差.

提示：将属性变量转化为 onehot 编码后再利用显式解得到参数估计

提交时间：10 月 10 日 18:30 之前。请预留一定的时间，迟交作业扣 3 分，作业抄袭 0 分。

附：北京租房数据集介绍

本案例的数据来源于某租房平台，数据已被划分为训练集和测试集，分别对应文件“train_data.csv”和“test_data.csv”。数据集共采集了北京市某年某月 5149 条合租房源的信息。本案例针对合租房间进行分析，若同一套房中有多个待租的房间，这些房间在本案例的数据中会对应多条数据，每一条数据对应其中一个合租房间，并且这些房间的数据中房源整体的信息相同（如房屋结构、地理位置等），但租赁面积、月租金不同。具体数据说明表如下

<table>
<tr><th colspan="2">变量类型</th><th colspan="2">变量名</th><th>详细说明</th><th>取值范围</th></tr>
<tr><td colspan="2">因变量</td><td>rent</td><td>季均销量</td><td>定量变量，单位：元</td><td>1150~6460</td></tr>
<tr><td rowspan="9">自变量</td><td rowspan="6">内部结构</td><td>area</td><td>租赁房间面积</td><td>定量变量，单位：平方米</td><td>5~30</td></tr>
<tr><td>room</td><td>租赁房间类型</td><td>定性变量，2 个水平</td><td>主卧、次卧</td></tr>
<tr><td>bedroom</td><td>卧室数</td><td>定量变量，单位：个</td><td>2~5</td></tr>
<tr><td>livingroom</td><td>厅数</td><td>定量变量，单位：个</td><td>1~2</td></tr>
<tr><td>bathroom</td><td>卫生间数</td><td>定量变量，单位：个</td><td>1~2</td></tr>
<tr><td>heating</td><td>供暖方式</td><td>定性变量，2 个水平</td><td>集中供暖、自采暖</td></tr>
<tr><td rowspan="3">外部条件</td><td>floor_grp</td><td>所在楼层</td><td>定性变量，3 个水平</td><td>高楼层、中楼层、低楼层</td></tr>
<tr><td>subway</td><td>邻近地铁</td><td>定性变量，2 个水平</td><td>是、否</td></tr>
<tr><td>region</td><td>所在城区</td><td>定性变量，11 水平</td><td>朝阳、海淀、东城、西城、昌平、大兴、通州、石景山、丰台、顺义、房山</td></tr>
</table>