Linear Regression and EDA

1 Theory

1. We consider the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad 1 \leq i \leq n,$$

where the X_i's are fixed, $\epsilon_i \sim_{i.i.d.} N(0, \sigma^2)$.

   (a) Compute the Maximum Likelihood Estimator (MLE) $\hat{\beta}_0$, $\hat{\beta}_1$ of β_0 and β_1. Describe these estimators.

   (b) Show that the MLE of β_0, β_1 are unbiased.

   (c) Compute the covariance matrix of $(\hat{\beta}_0, \hat{\beta}_1)$. How does the variance of these estimators vary with σ^2 and the X_i's.

   (d) We set $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ and $e_i = Y_i - \hat{Y}_i$. Show that $\sum_{i=1}^n e_i = 0$.

   (e) Show that $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$ is an unbiased estimator of σ^2.

   (f) We obtain a new covariate X_{n+1}. We define $\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{n+1}$. Compute the variance of this prediction.

   (g) We set $Y_{n+1} = \beta_0 + \beta_1 X_{n+1} + \epsilon_{n+1}$. Compute the variance of $\hat{\epsilon}_{n+1} = Y_{n+1} - \hat{Y}_{n+1}$. Compare this value to the variance of ϵ_i for $1 \leq i \leq n$.

   (h) Gauss-Markov Theorem.

      i. Write $\hat{\beta}_1$ as a linear combination of the observations Y_i's.

      ii. We consider another unbiased linear estimator $\tilde{\beta}_1 = \sum_{i=1}^n \lambda_i Y_i$. Show that

$$\sum_{i=1}^n \lambda_i = 0, \quad \sum_{i=1}^n \lambda_i x_i = 0.$$

      iii. Deduce that $\mathrm{Var}(\tilde{\beta}_1) \geq \mathrm{Var}(\hat{\beta}_1)$.

   (i) What is the distribution of $\hat{\beta}_1$ and $\hat{\sigma}^2$?

2. Generalized least square estimator

   Let us consider the following regression linear model

$$Y = X\beta + \epsilon, \tag{1}$$

   with the unknown parameter $\beta \in \mathbb{R}^p$, $\mathbb{E}[\epsilon] = 0_{\mathbb{R}^n}$ and $\mathsf{Var}\epsilon = \sigma^2\Sigma$. The known symetric matrix of size n Σ has rank n. We suppose here that Σ is not diagonal and the matrix X of size $n \times p$ is full rank.

continued on page 2

(a) Consider the ordinary least square estimator $\widehat{\beta} = (X^T X)^{-1} X^T Y$. Calculate the expectation and the matrice of variance-covariance of $\widehat{\beta}$ under the given assumptions in this exercice.

(b) Justify the existance of a matrix Ω of size $n \times n$ inversible such that $\Sigma = \Omega^T \Omega$.

(c) Prove that $\Omega^{-1} X$ is a full rank matrix.

(d) Set $Y^* = \Omega^{-1} Y$, $X^* = \Omega^{-1} X$ and $\epsilon^* = \Omega^{-1} \epsilon$, prove that we obtain a new model which statisfies the postulats [**P1**–**P3**] of a classical linear regression model.

(e) Deduce a "better" estimator function of X, Y and Σ. We denote by $\widehat{\beta}_G$ this estimator.

(f) Calculate (or deduce) $\mathbb{E}[\widehat{\beta}_G]$ and $\mathsf{Var}\widehat{\beta}_G$.

(g) Prove that $\widehat{\beta}_G$ is optimal among all linear unbiased estimators (say T) in the setting of the model1 i.e.

$$\mathsf{Var}T - \mathsf{Var}\widehat{\beta}_G$$

is positive definite. définie positive. (*see the proof of Gauss-Markov theorem*)

(h) Conclude by a comparaison with the ordinary least square estimator.

(i) Give an unbiased estimator of σ^2 in this case.

2 Exploratory Data Analysis (EDA).

1. The data file `bea-2006.csv` contains information about the economies of the 366 metropolitan statistical areas" (cities) of the US in 2006. In particular, it lists, for each city, the population, the total value of all goods and services produced for sale in the city that year per person (per capita gross metropolitan product", pcgmp), and the share of economic output coming from four selected industries.

   Formulate in clear and concise sentences your hypothesis, reasoning and motivation for any statistical analysis you will implement in R and interpretations of the R outputs. You will present your analysis in a Rmarkdown report that will contain reproducible codes, all necessary graphs and outputs and your complete reasoning. All figures should be clearly labeled and readable.

   (a) Load the data file and verify that it has 366 rows and 7 columns. Why should it have seven columns, when the paragraph above described only six variables?

   (b) Calculate summary statistics for the six numerical-valued columns.

   (c) Make univariate EDA plots for population and for per-capita GMP, and de- scribe their distributions in words. (Use the commands `hist` and `boxplot`.)

   (d) Make a bivariate EDA plot for per-capita GMP as a function of population. Describe the relationship in words.

   (e) Using only the functions `mean, var, cov, sum` and `arithmetic`, calculate the slope and intercept of the least-squares regression line.

continued on page 3

(f) What are the slope and intercept returned by the function lm? Does it agree with your answer in the previous part? Should it?

(g) Add both lines to the bivariate EDA plot. (Add only one line, of course, if you think they are the same.) Comment on the fit. Do the assumptions of the simple linear regression model appear to hold? Are they any places where the fit seems better than others?

(h) Find Pittsburgh in the data set. What is the population? The per-capita GMP? The per-capita GMP predicted by the model? The residual for Pittsburgh?

(i) What is the mean squared error of the regression? That is, what is $n^{-1}\sum_{i=1}^{n} e_i^2$ where $e_i = Y_i - \hat{Y}_i$ is the residual.

(j) Is the residual for Pittsburgh large, small, or typical compared to the mean squared error?

(k) Make a plot of residuals (vertical axis) against population (horizontal axis). What should this look like if the assumptions of the simple linear regression model hold? Is the actual plot compatible with those assumptions? Explain.

(l) Make a plot of squared residuals (vertical axis) against population (horizontal axis). What should this look like if the assumptions of the simple linear regression model hold? Is the actual plot compatible with those assumptions? Explain.

(m) State, carefully, the interpretation of the estimated slope; be sure to refer to the actual variables of the problem, not abstract ones like "the predictor variable" or X.

(n) What per-capita GMP does the model predict for a city with 10^5 more people than Pittsburgh?

(o) What does the model predict would happen to Pittsburgh's per- capita GMP if, by a policy intervention, we added 10^5 people to the population?