

Linear Regression

HW 3

Due 9/14 at 11:59pm

Directions: Submit a .pdf file containing your responses. The .pdf can be converted from a Latex file, pictures of your handwritten solutions, word files, markdown files, etc. If there are coding problems, upload a separate notebook for Python code.

Written Questions:

- Set up the \mathbf{X} matrix and $\boldsymbol{\beta}$ vector for each of the following regression models (assume $i = 1, \dots, 5$):

- $Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \varepsilon_i$
- $\sqrt{Y_i} = \beta_0 + \beta_1 X_{i1} + \beta_2 \log_{10} X_{i2} + \varepsilon_i$

- A student stated: “Adding predictor variables to a regression model can never reduce R^2 , so we should include all available predictor variables in the model.” Comment.
- Consider the multiple regression model:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad i = 1, \dots, n$$

where the ε_i are uncorrelated, with $E\{\varepsilon_i\} = 0$ and $\sigma^2\{\varepsilon_i\} = \sigma^2$.

- State the least squares criterion and derive the least squares estimators of β_1 and β_2 .
 - Assuming that the ε_i are independent normal random variables, state the likelihood function and obtain the maximum likelihood estimators of β_1 and β_2 . Are these the same as the least squares estimators?
- An analyst wanted to fit the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i, \quad i = 1, \dots, n,$$

by the method of least squares when it is known that $\beta_2 = 4$. How can the analyst obtain the desired fit by using a multiple regression computer program?

- In a small-scale regression study, the following data were obtained:

i	1	2	3	4	5	6
X_{i1}	7	4	16	3	21	8
X_{i2}	33	41	7	49	5	31
Y_i	42	33	75	28	91	55

Assume that regression model with independent normal error terms is appropriate. Using matrix methods, obtain (a) $\hat{\beta}$; (b) \mathbf{e} ; (c) \mathbf{H} ; (d) SSR ; (e) $\widehat{SE}(\hat{\beta})$; (f) \hat{Y}_h when $X_{h1} = 10$, $X_{h2} = 30$; and (g) $\widehat{SE}(\hat{Y}_h)$ when $X_{h1} = 10$, $X_{h2} = 30$.

Write down matrix X and vector y , and the formula to calculate each value before giving the answers. You can do the matrix calculation using Python matrix multiplication or any calculator. Do NOT use Python packages to fit the model and read the output directly.

6. Consider the multiple linear regression model in the matrix form $Y = X\beta + \epsilon$ with $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2 I_n$. Let $\hat{\beta}$ be the least squares estimator of β , $\hat{Y} = X\hat{\beta}$ be the fitted values, and $e = Y - \hat{Y}$ be the residuals. Express $\text{Var}(\hat{\beta})$, $\text{Var}(\hat{Y})$, and $\text{Var}(e)$ in terms of σ^2 and X . Show your steps to obtain the expressions in detail.

Coding questions:

1. The dataset “insurance.csv” contains information on the individuals’ medical costs billed by health insurance with the variables defined as below:

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary’s residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

Regress “charges” against “age”, “bmi” and “children” (use the order: $\text{charges} \sim \text{age} + \text{bmi} + \text{children}$) in Python.

- (a) ONLY based on the model summary table (`model.summary()`), we have discussed several important things to look at when it comes to the model interpretation: Global F-test p-value, R^2 and R_a^2 , and t-test for individual coefficient. Comment on each values above from your model summary, what kind of conclusion you have about your model fit and significance of predictors so far?
- (b) Run `sm.stats.anova_lm(model, typ=1)` to obtain `sum_sq` of “bmi” in the ANOVA is a sequential sum of squares of “bmi” which is defined as $\text{SeqSS}(\text{bmi}) = \text{SSE}_{\text{reduced}} - \text{SSE}_{\text{full}}$. What is the reduced model and full model here? How do we interpret the p-value of “bmi” in this F-test?
- (c) Run `sm.stats.anova_lm(model, typ=2)` to obtain `sum_sq` of “bmi” in the anova is a partial sum of squares of “bmi” defined by $\text{PartialSS}(\text{bmi}) = \text{SSE}_{\text{reduced}} - \text{SSE}_{\text{full}}$. What is the reduced model and full model here? How do we interpret the p-value of “bmi” in this F-test?
2. A commercial real estate company evaluates age (X_1), operating expenses (X_2 , in thousands of dollars), vacancy rate (X_3), total square footage (X_4) and rental rates (Y , in thousands of dollars) for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data are taken from 81 suburban commercial properties. (The data can be found in `property.txt`; The first column is Y , followed by X_1 , X_2 , X_3 , X_4 .) Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \epsilon_i, \quad i = 1, \dots, n, \quad \text{with } \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2).$$

- (a) Fit the model in Python and produce the summary table.
- (b) Compute the fitted values and residuals. Print the first 6 cases. Based on the residuals, give an estimate of σ^2 .
- (c) Test whether $\beta_2 = 0$ or not at 0.01 significance level. Interpret this coefficient in the real life terms of the problem.

- (d) Test $H_0 : \beta_1 = \beta_2 = 0$. vs. H_a : either β_1 or β_2 not equal to 0.
- (e) Consider a property with the following characteristics: $X_1 = 4, X_2 = 10, X_3 = 0.1, X_4 = 80,000$.
Give a 90% prediction interval for the rental rate.