Home assignment 1

Deadline: 2024-09-16, 21:00.

1. Each day Elon Musk solves econometrics problems and creates posts in X. Let y_i be the number of solved problems and x_i be the number of posts in X. You have 3 observations: $x_1 = 2$, $y_1 = 5$, $x_2 = 1$, $y_2 = 10$, $x_3 = 3$, $y_3 = 4$.

   (a) Find $\hat{\beta}$ if fitted values are given by $y_i = \hat{\beta} x_i$.

   (b) Find $\hat{\beta}_0$ and $\hat{\beta}_1$ if fitted values are given by $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

   (c) Find $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ if fitted values are given by $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$.

   Note: you can use any programming language to calculate the 3×3 matrix inverse but you should provide the code :)

2. Simplify as much as possible the following expressions:

$$A = \sum_{i=1}^{n} (x_i - \bar{x})\bar{x}, \quad B = \sum_{i=1}^{n} (x_i - \bar{x})\bar{y}, \quad C = \sum (x_i - \bar{x})^2 + n\bar{x}^2.$$

3. Consider simple regression model with $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. You have n observations (x_1, y_1), ..., (x_n, y_n) and you estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ using OLS.

   What will happen with $\hat{\beta}_0$ and $\hat{\beta}_1$ in each of the following cases?

   (a) You copy every observation from the original dataset twice.

   (b) You add one new observation $(y_{n+1} = \bar{y}, x_{n+1} = \bar{x})$ to the original dataset.

   (c) You add n more observations given by $(x_{n+i} = -x_i, y_{n+i} = y_i)$ for $i = 1$, 2, ..., n to the original dataset.

   Hint: you may start by guessing the answer with an experiment, but the proof is required :)

Home assignment 2

Deadline: 2024-09-23, 21:00.

1. Each day Elon Musk solves econometrics problems and creates posts in X. Let y_i be the number of solved problems and x_i be the number of posts in X. You have 3 observations: $x_1 = 2$, $y_1 = 5$, $x_2 = 1$, $y_2 = 10$, $x_3 = 3$, $y_3 = 4$.

   (a) Calculate SST, SSE, SSR and R^2 if we regress y on x with constant, ie $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

   (b) Calculate SST, SSE, SSR and R^2 if we regress x on y with constant, ie $\hat{x}_i = \hat{\gamma}_0 + \hat{\gamma}_1 y_i$.

   (c) Calculate the hat-matrix H if we regress y on x with constant.

   Note: this exercises uses toy dataset from the previous HA, you may reuse old results provided that you state them explicitely.

2. Kamala Harris removes one observation from the initial set of n observations and reestimates the model $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ using OLS.

   (a) Prove that the total sum of squares SST can't increase.

   (b) Provide an example of a dataset where explained sum of squares SSE will decrease and a second example where it will increase.

3. Consider the dataset of diamond prices,

   `https://github.com/vincentarelbundock/Rdatasets/raw/master/csv/ggplot2/diamonds.csv`.

   Here `price` is the price of diamond in $ and `carat` is the weight of a diamond in carats. Let y_i be the log of diamond price in 1000$ and x_i be the log of diamond weight in carats.

   (a) Estimate the model $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$ using `LinearRegression` from `sklearn.linear_model`.

   (b) Estimate the model $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$ using `ols` from `statsmodels.formula.api`.

   (c) What is your point forecast of a price of a diamond with 2 carats weight?

   Note: the first approach is faster and more stable while the second one gives you much more statistical information.

Home assignment 3

Deadline: 2024-09-30, 21:00.

1. Consider the framework of simple regression model, $y_i = \beta_0 + \beta_1 x_i + u_i$, $\mathbb{E}(u_i \mid x) = 0$, independent observations, $\mathbb{V}\text{ar}(u_i \mid x) = \sigma^2$, $\mathbb{C}\text{ov}(u_i, u_j \mid x) = 0$ for $i \neq j$. We estimate regression $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

   We have $n = 3$ observations with $x_i = i$.

   (a) Find $\mathbb{E}(2\hat{\beta}_0 + 3\hat{\beta}_1 \mid x)$, $\mathbb{V}\text{ar}(2\hat{\beta}_0 + 3\hat{\beta}_1 \mid x)$.

   (b) Find $\mathbb{E}(\hat{y}_1 \mid x)$, $\mathbb{V}\text{ar}(\hat{y}_1 \mid x)$, $\mathbb{E}(\hat{u}_1 \mid x)$, $\mathbb{V}\text{ar}(\hat{u}_1 \mid x)$.

2. Consider the framework of simple regression model, $y_i = \beta_0 + \beta_1 x_i + u_i$, $\mathbb{E}(u_i \mid x) = 0$, independent observations, $\mathbb{V}\text{ar}(u_i \mid x) = \sigma^2$, $\mathbb{C}\text{ov}(u_i, u_j \mid x) = 0$ for $i \neq j$. We estimate regression $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

   We have n observations with $\sum (x_i - \bar{x})^2 > 0$.

   (a) Find $\mathbb{E}(y_i - \bar{y} \mid x)$, $\mathbb{E}((y_i - \bar{y})^2 \mid x)$.

   (b) It possible find the value of γ such that the estimator $s^2 = \gamma \sum_{i=1}^n (y_i - \bar{y})^2$ for σ^2 is unbiased conditional on x.

3. Consider the framework of simple regression model, $y_i = \beta_0 + u_i$, $\beta_0 = 2$, $\mathbb{E}(u_i \mid x) = 0$, independent observations, $\mathbb{V}\text{ar}(u_i \mid x) = \sigma^2 = 4$, $\mathbb{C}\text{ov}(u_i, u_j \mid x) = 0$ for $i \neq j$. Random error is conditionally normally distributed, $(u_i \mid x) \sim \mathcal{N}(0; 4)$. We estimate regression $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. This setup means that we wrongly belive that y_i depends on x_i.

   We have $n = 10$ observations with $x_i \sim \mathcal{N}(0; 1)$.

(a) Generate the dataset and estimate the misspecified regression $B = 10000$ times. Draw the histogram of $\hat{\beta}_0$, the histogram of $\hat{\beta}_1$. Compare these histograms with true values of β_0 and β_1. What can you conclude based on two histogram?

(b) Draw the histogram of R^2 for simulations in point (a). Now repeat $B = 10000$ simulations for regression $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \hat{\beta}_3 x_i^3$. Draw the new histogram of R^2. Describe how this new histogram for R^2 is different from the first histogram for R^2. Can you say that the quality of your new regression is higher?

Home assignment 4

Deadline: 2024-10-07, 23:59.

1. Consider the simple regression model $y_i = \beta_0 + \beta_1 x_i + u_i$ with fitted values given by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and $\hat{u}_i = y_i - \hat{y}_i$. We have $n = 3$ observations, $x_i = i$, all Gauss-Markov assumptions are satisfied, We use ordinary least squares.

   (a) Write $\hat{\beta}_1$ explicitely as a linear function of (y_i), $\hat{\beta}_1 = w_1 y_1 + w_2 y_2 + w_3 y_3$.

   (b) Propose different coefficients w_1', w_2', w_3' such that the estimator $\hat{\beta}_1' = w_1' y_1 + w_2' y_2 + w_3' y_3$ is unbiased for $\hat{\beta}_1'$.

   (c) Check that the variance of alternative estimator $\hat{\beta}_1'$ is larger than the variance of OLS-estimator $\hat{\beta}_1$.

   (d) Find all the diagonal elements of the hat-matrix H_{ii}. Which actual value y_i has more influence on the forecasted value $\hat{y}_i$?

2. Consider the multivariate regression model in a matrix form, $y = X\beta + u$ with fitted values given by $\hat{y} = X\hat{\beta}$ and $\hat{u} = y - \hat{y}$. We have n observations, all Gauss-Markov assumptions are satisfied, We use ordinary least squares.

   (a) Find $\mathbb{E}(\hat{u} \mid X)$, $\mathbb{E}(\hat{y} \mid X)$.

   (b) Find $\mathbb{V}\mathrm{ar}(\hat{u} \mid X)$, $\mathbb{C}\mathrm{ov}(\hat{y}, \hat{\beta} \mid X)$, $\mathbb{C}\mathrm{ov}(\hat{u}, \hat{\beta} \mid X)$.

3. Consider the simple regression model $y_i = \beta_0 + \beta_1 x_i + u_i$ with fitted values given by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and $\hat{u}_i = y_i - \hat{y}_i$. We have n observations, all Gauss-Markov assumptions are satisfied. Yusuf Dikeç copies every observation twice and estimates regression using OLS for $2n$ observations.

   (a) Which Gauss-Markov assumptions are violated for the doubled dataset of $2n$ observations?

   (b) Find the true variance of $\hat{\beta}_1$ in the regression on $2n$ observations assuming Gauss-Markov assumptions for the original dataset.

   (c) Find the variance of $\hat{\beta}_1$ in the regression on $2n$ observations wrongly assuming Gauss-Markov assumptions for the doubled dataset.

Home assignment 5

Deadline: No deadline.
If you wish to upload something somewhere then you are free to submit econometrics memes to the chat.

Home assignment 6

Deadline: 2024-10-24 (updated 2024-10-21), 23:59.

1. Consider the model $y_i = \beta_x x_i + \beta_w w_i + u_i$ with

$$\begin{pmatrix} x_i \\ w_i \\ u_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}; \begin{pmatrix} 5 & 3 & 0 \\ 3 & 10 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right).$$

   Observations are independent.

   (a) Find the probability limit of $\hat{\gamma}_x$ in regression $\hat{y}_i = \hat{\gamma}_x x_i$.

   (b) Is $\hat{\gamma}_x$ consistent estimator of β_x?

   (c) Find the conditional expected value $\mathbb{E}(y_i \mid x_i)$.
   Hint: it should be of the form αx_i, where α is a function of β_x and β_w.

   (d) Is $\hat{\gamma}_x$ consistent estimator of α?

2. Consider the simple regression model $y_i = \beta_0 + \beta_1 x_i + u_i$. We do not observe x_i. Instead we observe two independent measurements of x_i: x_i' and x_i''. Here $x_i' = x_i + v_i$ and $x_i'' = x_i + w_i$, where v_i and w_i are measurement errors.

   Observations are independent, random variables x_i, u_i, v_i and w_i are independent. Let's denote their variance by $\mathbb{V}\mathrm{ar}(x_i) = \sigma_x^2$, $\mathbb{V}\mathrm{ar}(u_i) = \sigma_u^2$, $\mathbb{V}\mathrm{ar}(v_i) = \sigma_v^2$, $\mathbb{V}\mathrm{ar}(w_i) = \sigma_w^2$.

   (a) Check whether the estimator $\hat{\beta}_1^A$ is consistent for β_1:

$$\hat{\beta}_1^A = \frac{\sum (y_i - \bar{y})(x_i' - \bar{x}')}{\sum (x_i' - \bar{x}')(x_i'' - \bar{x}'')}.$$

   (b) Check whether the estimator $\hat{\beta}_1^B$ is consistent for β_1:

$$\hat{\beta}_1^B = \frac{\sum (y_i - \bar{y})(x_i' - \bar{x}')}{\sum (x_i'' - \bar{x}'')^2}.$$

3. Consider again the dataset of diamond prices,
   `https://github.com/vincentarelbundock/Rdatasets/raw/master/csv/ggplot2/diamonds.csv`.

   Here `price` is the price of diamond in \$ and `carat` is the weight of a diamond in carats. Let y_i be the log of diamond price in 1000\$ and x_i be the log of diamond weight in carats.

   (a) Reestimate the model $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$ using `ols` from `statsmodels.formula.api`.

   Let's believe in Gauss — Markov assumptions for this case.

   (b) Extract $SSRes$, SST, $SSExpl$, R^2 and $\hat{\sigma}^2$.

   (c) Write down true $\mathbb{V}\mathrm{ar}(\hat{\beta} \mid X)$ matrix.
   Hint: your matrix should contain unknown σ^2.

(d) Write down the estimate of $\mathbb{V}\mathrm{ar}(\hat{\beta} \mid X)$ matrix.
Hint: no unknown parameters here.

(e) Calculate all diagonal entries H_{ii} and select the most influential observation with highest $\partial \hat{y}_i / \partial y_i$.
Hint: the whole matrix H is really HUGE here, please do not try to calculate it, you need only diagonal elements.

(f) Draw the scatterplot of $|\hat{u}_i|$ against x_i.
Does this plot suggests that Gauss — Markov assumptions are satisfied?

Home assignment 7

Deadline: 2024-11-02, 23:59.

1. Consider a simple regression model with Gauss — Markov assumptions, $y_i = \beta_0 + \beta_1 x_i + u_i$. Random errors are jointly normal $u \mid X \sim \mathcal{N}(0; \sigma^2 I)$.

   You know that
$$X^T X = \begin{pmatrix} 100 & 200 \\ ? & 600 \end{pmatrix}, \quad X^T y = \begin{pmatrix} 0 \\ 300 \end{pmatrix}, \quad y^T y = 1000.$$

   (a) By looking at $X^T X$ recover the number of observations.
   (b) Estimate $\hat{\beta}_0$, $\hat{\beta}_1$, $\widehat{\mathbb{V}\mathrm{ar}}(\hat{\beta} \mid X)$.
   (c) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at significance level $\alpha = 0.05$.
   (d) Construct 99% confidence interval for β_0 and β_1.
   (e) Esimate $\mathbb{E}(y_{101} \mid x_{101} = 5)$ and construct 99% confidence interval for $\mathbb{E}(y_{101} \mid x_{101} = 5)$.

2. You estimated the vector $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ in the model
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \beta_3 h_i + u_i$$
   using OLS with 200 observations. Gauss — Markov assumptions are satisfied. Random errors are jointly normal $u \mid X \sim \mathcal{N}(0; \sigma^2 I)$.

   You know that
$$\hat{\beta} = \begin{pmatrix} 0.2 \\ 0.5 \\ 0.3 \\ 0.6 \end{pmatrix}, \quad \widehat{\mathbb{V}\mathrm{ar}}(\hat{\beta} \mid X) = 0.001 \cdot \begin{pmatrix} 63.14 & -14.78 & 15.56 & 0.335 \\ ? & 7.912 & -3.943 & -1.065 \\ ? & ? & 6.939 & -1.375 \\ ? & ? & ? & 1.178 \end{pmatrix}.$$

   (a) Test H_0: $\beta_1 = \beta_2$ using significance level $\alpha = 0.05$ against H_1: $\beta_1 \neq \beta_2$.
   (b) Test H_0: $\beta_1 + \beta_2 = 1$ using significance level $\alpha = 0.05$ against H_1: $\beta_1 + \beta_2 \neq 1$.
   (c) Construct 99% confidence interval for $\beta_1 + 2\beta_2 + 3\beta_3$.

3. Researches suspect that «College students often have poor sleep habits, staying up late and sleeping short hours, and a great deal of research suggests that lack of sleep can harm cognitive performance».

   Let's build a model where dependent variable is `term_gpa` and other variables below as predictors:

- `demo_race`: binary label for underrepresented and non-underrepresented students;
- `demo_gender`: Gender of the subject (male = 0, female = 1);
- `bedtime_mssd`: Mean successive squared difference of bedtime;
- `TotalSleepTime`: Average time in bed in minutes;
- `cum_gpa`: Cumulative GPA (out of 4.0), for semesters before the one being studied;
- `term_gpa`: End-of-term GPA (out of 4.0) for the semester being studied;
- `units_score`: Standardized number of course units carried in the term;

More info can be found at:

`https://cmustatistics.github.io/data-repository/psychology/cmu-sleep.html`.

(a) Check that the dataset is imported correctly! Remove missing observations.

(b) Estimate the model using OLS.

(c) Test the hypothesis that the effect of `TotalSleepTime` is zero.

(d) Test the hypothesis that the sum of effects for `demo_gender`, `bedtime_mssd` and `units_score` is nonzero.

(e) Test the hypothesis that additional two hours of sleep every day gives additional 0.3 gpa point from 5 in average.

(f) Test the hypothesis that the parameters `demo_gender`, `bedtime_mssd` and `units_score` are jointly insignificant.