4. Heteroscedastic Regression

For the next three problems, consider the following setting.

We measure characteristics of n individuals, sampled randomly from a population. Let (X_i, y_i) be the observed data of the ith individual, where $y_i \in \mathbb{R}$ is the dependent variable and $X_i \in \mathbb{R}^p$ is the vector of p **deterministic** explanatory variables. Our goal is to estimate the coefficients of $\beta = (\beta_1, \ldots, \beta_p)^T$ in the linear regression:

$$y_i = X_i^T \beta + \epsilon_i, \qquad i = 1, \ldots, n$$

We will consider the case where the model is potentially **heteroscedastic** (i.e. the error terms ϵ_i are **not** i.i.d.).

More specifically, assume that the vector $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is a n-dimensional Gaussian with mean 0 and known **nonsingular** covariance matrix Σ. Denote by $\mathbb{X}$ the matrix in $\mathbb{R}^{n \times p}$ whose rows are $\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T$ and by $\mathbf{Y}$ the vector with coordinates $y_1, \ldots, y_n$.

Instead of the usual Least Squares Estimator, **instead consider the estimator $\hat{\beta}$ that minimizes, over all $\beta \in \mathbb{R}^p$,**

$$(\mathbf{Y} - \mathbb{X}\beta)^T \Sigma^{-1} (\mathbf{Y} - \mathbb{X}\beta).$$

(a) A Generalized Estimator

1/1 point (graded)

Let I_n be the $n \times n$ identity matrix. If $\Sigma = \sigma^2 I_n$ (i.e. homoscedastic ε) for some $\sigma^2 > 0$, then which one of the following statement about $\hat{\beta}$ **must** be true? **Make no assumptions about the rank of $\mathbb{X}$.**

- $\hat{\beta}$ has positive entries.
- $\hat{\beta}$ is the least squares estimator. ✔
- $\hat{\beta}$ is the unique minimizer of the specified loss.
- The components of $\hat{\beta}$ are independent.

Solution:

The second choice is the correct answer. Using the fact that $\|v\|^2 = v^T v$, the specified loss simplifies to

$$\frac{1}{\sigma^2}(\mathbf{Y} - \mathbb{X}\beta)^T I_n (\mathbf{Y} - \mathbb{X}\beta) = \frac{1}{\sigma^2}\|\mathbf{Y} - \mathbb{X}\beta)\|^2$$

which is minimized if and only if $\|\mathbf{Y} - \mathbb{X}\beta\|^2$ is minimized, ignoring the term $1/\sigma^2$. Any β that attains such a minimum is the least squares estimator.

Now, given the knowledge that $\hat{\beta}$ is the LSE, we can understand the other choices to be false, since they are not properties of the LSE. For example, we would actually want **"$\hat{\beta}$ has positive entries"** to be false! Imagine the scenario where $\beta_1 = -100$. If our LSE $\hat{\beta}$ somehow had the property that it was always positive, then it would not be a very good estimator to begin with.

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

(b) The Maximum Likelihood Estimator

1.0/1 point (graded)

In this exercise, we will prove that $\hat{\beta}$ is equal to the Maximum Likelihood Estimator, even for general Σ. Recall the form of the n-dimensional Gaussian density from Lecture 10.

Let Σ be an arbitrary $n \times n$ matrix. The maximum likelihood estimator β_{MLE} is the value of β **maximizes** the log-likelihood function $\ell(\beta) = \ln L(\mathbb{X}, \mathbb{Y}; \beta)$.

Write down the function ℓ, in terms of $\mathbb{X}$, $\mathbf{Y}$, β, Σ, and n.

(Type **X** for $\mathbb{X}$, **Y** for $\mathbf{Y}$, **Sigma** for Σ. Type **trans(X)** for the transpose $\mathbb{X}^T$, **det(X)** for the determinant $\det\mathbb{X}$, and **X^(-1)** for the inverse $\mathbb{X}^{-1}$, of a matrix $\mathbb{X}$.)

$\ell(\beta) = \ln L(\mathbb{X}, \mathbb{Y}; \beta) =$

-1/2*ln((2*pi)^n*det(Sigma))-1/2*trans(Y-X*beta)*Sigma^-1*(Y-X*beta) ✔

Answer: -(1/2)*trans(Y-X*beta)*(Sigma)^(-1)*(Y-X*beta)-(1/2)*ln((2*pi)^n*det(Sigma))

STANDARD NOTATION

Solution:

Recall that the n-dimensional Gaussian has density

$$f(\mathbf{x}; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \det\Sigma}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)}.$$

In our specific setting, the relationship $y_i = X_i^T \beta + \epsilon$ prescribes $\mathbf{Y}$ to be plugged into $\mathbf{x}$ in the above density, and $\mu = \mathbb{X}\beta$. The answer is the logarithm of f:

$$-\frac{1}{2}\ln\left((2\pi)^n \det\Sigma\right) - \frac{1}{2}(\mathbf{Y} - \mathbb{X}\beta)^T \Sigma^{-1} (\mathbf{Y} - \mathbb{X}\beta)$$

Therefore, the MLE maximizes $-\frac{1}{2}(\mathbf{Y} - \mathbb{X}\beta)^T \Sigma^{-1} (\mathbf{Y} - \mathbb{X}\beta)$ over all possible choices of β.

To finish the proof, note that we can flip the sign – this turns the maximization into a minimization, so that the MLE minimizes $(\mathbf{Y} - \mathbb{X}\beta)^T \Sigma^{-1} (\mathbf{Y} - \mathbb{X}\beta)$. This matches the provided definition of $\hat{\beta}$.

Submit You have used 2 of 4 attempts

Answers are displayed within the problem

(c)

2/2 points (graded)

Assume that the rank of $\mathbb{X}$ is p. (As a consequence, for any nonsingular $n \times n$ matrix Q, the $p \times p$ matrix product $X^T Q X$ is also nonsingular.)

Which of the following is a correct formula for $\hat{\beta}$ in terms of $\mathbb{X}$, $\mathbb{Y}$ and Σ?

- $(\mathbb{X}^T \Sigma^{-1} \mathbb{X})^{-1} \mathbb{X}^T \Sigma^{-1} \mathbb{Y}$ ✔

- $(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$
- $(\mathbb{X}^T\Sigma^{-1}\mathbb{X})^{-1}\mathbb{Y}^T\Sigma^{-1}\mathbb{X}$
- $\mathbb{X}^{-1}\mathbb{Y}$

Using the result from the above, which of the following correctly characterizes the distribution of $\hat{\beta}$?

- $\mathcal{N}(0, \Sigma)$
- $\mathcal{N}(0, (\mathbb{X}^T\Sigma^{-1}\mathbb{X})^{-1})$
- $\mathcal{N}(0, (\mathbb{X}^T\mathbb{X})^{-1})$
- $\mathcal{N}(\beta, (\mathbb{X}^T\Sigma^{-1}\mathbb{X})^{-1})$ ✔
- $\mathcal{N}(\beta, (\mathbb{X}^T\mathbb{X})^{-1})$

Solution:

For the first question: Start by expanding the product.

$$\begin{aligned}(\mathbb{Y}-\mathbb{X}\beta)^T\Sigma^{-1}(\mathbb{Y}-\mathbb{X}\beta) &= \mathbb{Y}^T\Sigma^{-1}\mathbb{Y} - \mathbb{Y}^T\Sigma^{-1}\mathbb{X}\beta - \beta^T\mathbb{X}^T\Sigma^{-1}\mathbb{Y} + \beta^T\mathbb{X}^T\Sigma^{-1}\mathbb{X}\beta \\ &= \mathbb{Y}^T\Sigma^{-1}\mathbb{Y} - 2\beta^T\mathbb{X}^T\Sigma^{-1}\mathbb{Y} + \beta^T\mathbb{X}^T\Sigma^{-1}\mathbb{X}\beta.\end{aligned}$$

The second equality uses the fact that $(\mathbb{Y}^T\Sigma^{-1}\mathbb{X}\beta)^T = \mathbb{Y}^T\Sigma^{-1}\mathbb{X}\beta$, since the left hand side is a scalar, and Σ is symmetric. Next, take the gradient with respect to β, which evaluates to $-2\mathbb{X}^T\Sigma^{-1}\mathbb{Y} + 2\mathbb{X}^T\Sigma^{-1}\mathbb{X}\beta$. Set this equal to zero and solve for β to obtain the estimator $\hat{\beta}$:

$$\hat{\beta} = (\mathbb{X}^T\Sigma^{-1}\mathbb{X})^{-1}\mathbb{X}^T\Sigma^{-1}\mathbb{Y}.$$

Thus, the correct answer is the **first choice**.

In the initial step, one needs to be a bit careful. If instead we wrote $2\mathbb{Y}^T\Sigma^{-1}\mathbb{X}\beta$ for the middle term in the second equality, we might be tempted to choose the third choice instead. By computing the gradient more carefully, the answer does consistently turn out to be the first choice. An easy way to reason about this, without explicitly writing out the calculus, is to remember that the dimensions must match up. For instance, the matrix $(\mathbb{X}^T\Sigma^{-1}\mathbb{X})^{-1}$ is $p \times p$, yet $\mathbb{Y}^T$ is a $1 \times p$ matrix - therefore, the third choice is not a valid expression.

For the second question, start with $\hat{\beta} = (\mathbb{X}^T\Sigma^{-1}\mathbb{X})^{-1}\mathbb{X}^T\Sigma^{-1}\mathbb{Y}$ and substitute $\mathbb{Y} = \mathbb{X}\beta + \epsilon$.

$$\begin{aligned}\hat{\beta} &= (\mathbb{X}^T\Sigma^{-1}\mathbb{X})^{-1}\mathbb{X}^T\Sigma^{-1}(\mathbb{X}\beta + \epsilon) \\ &= (\mathbb{X}^T\Sigma^{-1}\mathbb{X})^{-1}\mathbb{X}^T\Sigma^{-1}\mathbb{X}\beta + (\mathbb{X}^T\Sigma^{-1}\mathbb{X})^{-1}\mathbb{X}^T\Sigma^{-1}\epsilon \\ &= \beta + ((\mathbb{X}^T\Sigma^{-1}\mathbb{X})^{-1}\mathbb{X}^T\Sigma^{-1})\,\epsilon.\end{aligned}$$

Next, we use the rule of transforming a gaussian $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$ by a matrix M, which states that $M\mathbf{x} \sim \mathcal{N}(0, M\Sigma M^T)$. In particular,

$$(\hat{\beta} - \beta) \sim \mathcal{N}(0, [(\mathbb{X}^T\Sigma^{-1}\mathbb{X})^{-1}\mathbb{X}^T\Sigma^{-1}]\,\Sigma[(\mathbb{X}^T\Sigma^{-1}\mathbb{X})^{-1}\mathbb{X}^T\Sigma^{-1}]^T)$$

and the covariance simplifies to just $\left(\mathbb{X}^T\Sigma^{-1}\mathbb{X}\right)^{-1}$. Thus, the correct answer is $\mathcal{N}\left(\beta, \left(\mathbb{X}^T\Sigma^{-1}\mathbb{X}\right)^{-1}\right)$ after re-shifting the mean by β.

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Discussion

Show Discussion

Topic: Unit 6 Linear Regression:Homework 10 Linear regression / 4. Heteroscedastic Regression