Part I: Linear Models

Chapter I.7

Random Measurement Points in Regression Models

Topics

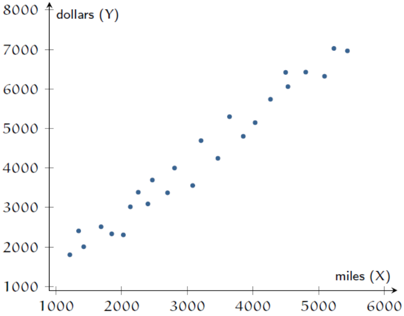
To be discussed...

- Random measurement points
- Maximum likelihood estimation
- Connection to LMs
- Prediction problems

Some Data – X and Y are both random!

I.7.1 Example

Consider the following data where X denotes the number of miles traveled by a credit card holder and Y denotes the charges (in US$). The credit card company suspects that the charges increase with the number of traveled miles.



Passenger	Miles (X)	Dollars (Y)
1	1211	1802
2	1345	2405
3	1422	2005
4	1687	2511
5	1849	2332
6	2026	2305
7	2133	3016
8	2253	3385
9	2400	3090
10	2468	3694
11	2699	3371
12	2806	3998
13	3082	3555
14	3209	4692
15	3466	4244
16	3643	5298
17	3852	4801
18	4033	5147
19	4267	5738
20	4498	6420
21	4533	6059
22	4804	6426
23	5090	6321
24	5233	7026
25	5439	6964

I.7.2 Remark

- So far, we have discussed regression models with fixed effects, that is, the measurement points are supposed known and fixed in advance.
- In a linear multiple regression model, the data is given by $\mathbf{Y}$ assuming a LM $\mathbf{Y} = B\boldsymbol{\beta} + \varepsilon$ with

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad B = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} \in \mathbb{R}^{n \times m}.$$

- Now, it is assumed that the measurement points are also realisations of a random vector. Therefore, it is assumed that we have random vectors $\mathbf{X}_1, \dots, \mathbf{X}_n$ (as rows of the design matrix B) where $\mathbf{x}_i' = (x_{i1}, \dots, x_{im})$, $1 \leqslant i \leqslant n$, are the corresponding realizations.
- Thus, we have a sample $\boldsymbol{W}_i = (Y_i, \mathbf{X}_i'), 1 \leqslant i \leqslant n \overset{\text{iid}}{\sim} P_{\boldsymbol{\beta}}^{Y, \mathbf{X}'}, \boldsymbol{\beta} \in \Theta$.

I.7.3 Model with random measurement points

Consider random vectors $\boldsymbol{W}_i = (Y_i, \mathbf{X}_i')$, $1 \leqslant i \leqslant n$, as in Remark I.7.2. Assume that $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n \overset{\text{iid}}{\sim} N_{m+1}(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} \in \mathbb{R}^{m+1}, \Sigma \in \mathbb{R}_{>0}^{(m+1)\times(m+1)}$ where

$$\boldsymbol{\mu} = E\boldsymbol{W}_i = \begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \quad \Sigma = \text{Cov}\,(\boldsymbol{W}_i) = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx}' \\ \Sigma_{yx} & \Sigma_{xx} \end{bmatrix},$$

$$\mu_y = EY_i, \boldsymbol{\mu}_x = E\mathbf{X}_i, \quad \Sigma_{yy} = \text{Var}\,(Y_i), \Sigma_{yx} = \text{Cov}\,(Y_i, \mathbf{X}_i), \Sigma_{xx} = \text{Cov}\,(\mathbf{X}_i).$$

From Theorem I.2.8, we get directly the conditional distribution of Y_i given $\mathbf{X}_i = \boldsymbol{x}_i$, $1 \leqslant i \leqslant n$.

I.7.4 Theorem

Consider Model I.7.3 and let $\Sigma_{yy|x} = \Sigma_{yy} - \Sigma_{yx}'\Sigma_{xx}^{-1}\Sigma_{yx}$. Then, we have for $1 \leqslant i \leqslant n$:

1. $Y_i \mid \mathbf{X}_i = \boldsymbol{x}_i \sim N_1(\mu_y + \Sigma_{yx}'\Sigma_{xx}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_x), \Sigma_{yy|x})$
2. $E(Y_i \mid \mathbf{X}_i = \boldsymbol{x}_i) = \mu_y + \Sigma_{yx}'\Sigma_{xx}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_x)$
3. $\text{Cov}\,(Y_i \mid \mathbf{X}_i = \boldsymbol{x}_i) = \Sigma_{yy|x}$

Moreover, conditionally on $\mathbf{X}_i = \boldsymbol{x}_i, 1 \leqslant i \leqslant n$, $Y_1, \ldots, Y_n$ are independent random variables.

I.7.5 Remark

- The conditional expectation $E(Y \mid \mathbf{X} = \boldsymbol{x}) = \mu_y + \Sigma'_{yx}\Sigma_{xx}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_x)$ is a linear function of $\boldsymbol{x}$. Thus, we can write

$$E(Y \mid \mathbf{X} = \boldsymbol{x}) = \beta_0 + \boldsymbol{x}'\boldsymbol{\beta}_1 = [1 \mid \boldsymbol{x}']\,\boldsymbol{\beta}. \tag{I.14}$$

  This yields $\mathbf{Y} \mid [\mathbf{X}_i = \boldsymbol{x}_i, 1 \leqslant i \leqslant n] \sim N_n(B\boldsymbol{\beta}, \Sigma_{yy|x} I_n)$ with $B = \begin{bmatrix} 1 & \boldsymbol{x}'_1 \\ \vdots & \vdots \\ 1 & \boldsymbol{x}'_n \end{bmatrix}$ and $\boldsymbol{\beta}_1 = \boldsymbol{\mu}_x$,

  $\beta_0 = \mu_y - \Sigma'_{yx}\Sigma_{xx}^{-1}\boldsymbol{\mu}_x$.

  Therefore, conditionally on $\mathbf{X}_i = \boldsymbol{x}_i, 1 \leqslant i \leqslant n$, we get a LM as in Definition I.4.2.
- $\sigma^2 = \text{Var}\,(Y \mid \mathbf{X} = \boldsymbol{x}) = \Sigma_{yy|x} > 0$ is independent of $\boldsymbol{x}$.
- Notice that the conditional mean $E(Y \mid \mathbf{X} = \boldsymbol{x})$ is also linear in $\boldsymbol{x}$.
- The matrix $\Sigma'_{yx}\Sigma_{xx}^{-1}$ is called **regression matrix**.

I.7.6 Maximum-Likelihood-Estimation

Let $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$ as in Model I.7.3. Then, the likelihood function (with given realisations $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$) reads

$$
\begin{aligned}
L(\boldsymbol{\mu}, \Sigma \mid \boldsymbol{w}_1, \ldots, \boldsymbol{w}_n) &= \prod_{j=1}^{n} f_{\boldsymbol{\mu},\Sigma}(\boldsymbol{w}_i) \\
&= (2\pi)^{-\frac{n(m+1)}{2}} (\det \Sigma)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{j=1}^{n} \|\boldsymbol{w}_j - \boldsymbol{\mu}\|_{\Sigma}^2\right) \quad \to \max_{\boldsymbol{\mu}\in\mathbb{R}^{m+1}, \Sigma\in\mathbb{R}_{>0}^{(m+1)\times(m+1)}} .
\end{aligned}
$$

I.7.7 Theorem

Consider Model I.7.3 with $n \geqslant m+1$ and $\mathcal{W} = \left[\boldsymbol{W}_1 \mid \cdots \mid \boldsymbol{W}_n\right]$. Let $\overline{\boldsymbol{W}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{W}_i = \frac{1}{n}\mathcal{W}'\mathbb{1}_n$ be the sample mean and $S = \frac{1}{n-1}\mathcal{W}'E_n\mathcal{W}$ be the (sample) scatter matrix.

Then, $(\overline{\boldsymbol{W}}, \frac{n-1}{n}S)$ is the MLE of $(\boldsymbol{\mu}, \Sigma)$.

A proof can be found in, e.g., Rencher, Schaalje (2008).

I.7.8 Theorem

In the situation of Theorem I.7.7, $\overline{\boldsymbol{W}}$ and $S = \frac{1}{n-1}\mathcal{W}'E_n\mathcal{W}$ are independent and unbiased estimators, that is,

$$E\overline{\boldsymbol{W}} = \boldsymbol{\mu}, \quad ES = \Sigma.$$

Furthermore, $\overline{\boldsymbol{W}} \sim N_{m+1}(\boldsymbol{\mu}, \frac{1}{n}\Sigma)$. The distribution of $(n-1)S$ is a Wishart distribution with $n-1$ degrees of freedom and parameter Σ.

I.7.9 Remark

- We are interested in the MLEs of $E(Y \mid \mathbf{X} = \boldsymbol{x}) = [1 \mid \boldsymbol{x}']\,\boldsymbol{\beta}$ as in (I.14) and of $\sigma^2 = \text{Var}\,(Y \mid \mathbf{X} = \boldsymbol{x}) = \Sigma_{yy|\boldsymbol{x}}$.
- We partition the sample scatter matrix as the covariance matrix Σ, that is, $S = \begin{bmatrix} S_{yy} & S'_{yx} \\ S_{yx} & S_{xx} \end{bmatrix}$.
- We get the desired MLEs using the *invariance property* of MLEs (see Hogg and Craig (1995), p. 265[a]):

  Let $\widehat{\theta}$ be the MLE of θ and g be a known 'smooth' function. Then, $g(\widehat{\theta})$ is the MLE of $g(\theta)$.

[a]Hogg, R. V. and A. T. Craig (1995). Introduction to Mathematical Statistics (5th ed.). Englewood Cliffs, NJ: Prentice-Hall

I.7.10 Theorem

Consider Model I.7.3 with $n \geqslant m+1$ and $\mathcal{W} = \left[\boldsymbol{W}_1 \mid \cdots \mid \boldsymbol{W}_n\right]$. Let $\overline{\boldsymbol{W}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{W}_i = \frac{1}{n}\mathcal{W}'\mathbb{1}_n$ be the sample mean and $S = \frac{1}{n}\mathcal{W}'E_n\mathcal{W}$ be the (sample) scatter matrix.

Then, the MLEs of $\boldsymbol{\beta}$ and σ^2 are given by

$$\widehat{\beta}_0 = \overline{Y} - S_{yx}' S_{xx}^{-1}\overline{X}, \quad \widehat{\boldsymbol{\beta}}_1 = S_{xx}^{-1} S_{yx}, \quad \widehat{\sigma}^2 = \frac{n-1}{n} s^2$$

where $s^2 = s_{yy} - S_{yx}' S_{xx}^{-1} S_{yx}$ is an unbiased estimator of $\Sigma_{yy|x}$.

The MLE of the squared population multiple correlation $\rho_{y|x}^2 = \dfrac{\Sigma_{yx}'\Sigma_{xx}^{-1}\Sigma_{yx}}{\Sigma_{yy}}$ is given by the squared sample multiple correlation coefficient

$$R^2 = \frac{S_{yx}' S_{xx}^{-1} S_{yx}}{S_{yy}}.$$

I.7.11 Remark

- The MLEs $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the same algebraic functions of the observations as the LSEs given in Corollary I.5.4. Therefore, its reasonable to use the estimators in both models.
- However, under a normal assumption, $\widehat{\beta}_1$ does not have a normal distribution as it is true for fixed effects.
- Nevertheless, the F-test can be used although the LSE has not an normal distribution!

  In order to see this, notice that the test statistic
  $$\frac{\mathsf{SSR}/m}{\mathsf{SSE}/(n-m-1)} \text{ conditionally on } \mathbf{X}_1 = \boldsymbol{x}_1, \ldots, \mathbf{X}_n = \boldsymbol{x}_n$$
  has an $F(m, n-m-1)$- distribution under H_0 which is independent (!) of the observed values $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Thus, using the law of total probability, we get (under H_0) that $\frac{\mathsf{SSR}/m}{\mathsf{SSE}/(n-m-1)} \sim F(m, n-m-1)$.

  A similar comment applies to the construction of confidence regions (see Theorem I.5.27).

Regression & Prediction

I.7.12 Problem

Consider Model I.7.3 without assuming a normal distribution. Then, we are looking for a solution of the following problem: We wish to find a (measurable) function $g : \mathbb{R}^m \to \mathbb{R}$ of $\mathbf{X}$ predicting Y best in the sense

$$E(Y - g(\mathbf{X}))^2 \to \min .$$

Therefore, we can interpret regression as the problem to find a function g that predicts the random variable Y in some way based on the random vector $\mathbf{X}$ or the random variables $X_1, \ldots, X_m$.

I.7.13 Theorem

The solution of Problem I.7.12 is given by the conditional expectation $E(Y \mid \mathbf{X})$.

$\widehat{\pi}(\mathbf{X}) = E(Y \mid \mathbf{X})$ is called the **best predictor** (BP) for Y based on $\mathbf{X}$.

Regression & Prediction

I.7.14 Remark

- For a proof of Theorem I.7.13, see Christensen (2011).
- Given a normal assumption, the best predictor for Y based on $\mathbf{X}$ is given by

$$\widehat{\pi}_L(\mathbf{X}) = \mu_y + \Sigma'_{yx}\Sigma_{xx}^{-1}(\mathbf{X} - \boldsymbol{\mu}_x) = [1 \mid \mathbf{X}']\,\boldsymbol{\beta}$$

  provided that Σ and $\boldsymbol{\mu}$ are known (see eq. (I.14)).
- Similar, one may study the **best linear predictor** (BLP) for Y based on $\mathbf{X}$, which is the result of the minimization problem

$$E(Y - B\,\boldsymbol{\beta})^2 \to \min_{\boldsymbol{\beta}},$$

  where $B = [1 \mid \mathbf{X}']$. Notice that the solution of this least squares problem is given by

$$\widehat{\pi}_L(\mathbf{X}) = \mu_y + \Sigma'_{yx}\Sigma_{xx}^{-1}(\mathbf{X} - \boldsymbol{\mu}_x)$$

  provided that Σ and $\boldsymbol{\mu}$ are known. For a normal distribution assumption, $\widehat{\pi}(\mathbf{X}) = \widehat{\pi}_L(\mathbf{X})$.