Regression

Sofia Olhede

EPFL

November 2, 2020

1. Linear Regression
   - Least squares regression
   - Residuals
   - Confidence intervals for coefficients and variance
   - Confidence intervals for coefficients and variance
   - Regression Diagnostics and Distribution Plots

Set–up

- Consider a set of measurements given by the response variable Y_i and with a corresponding set of predictor variables $x_{i1}, \dots, x_{ip}$. Hence the data set is

$$\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^{n}.$$

- Definition: A linear model is

$$\mathbb{E}\{\mathrm{Y}\} = \mathrm{X}\boldsymbol{\beta},$$

$n \times 1$ $n \times p \cdot p \times 1$

where $\mathrm{Y} = \begin{pmatrix} Y_1 & \dots & Y_n \end{pmatrix}^T$, is the vector of observations, X is the known $n \times p$ design matrix and $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 & \dots & \beta_p \end{pmatrix}^T$ is the $p \times 1$ parameter vector.

- We are trying to quantify the systemic variation in Y due to $\mathrm{X}\boldsymbol{\beta}$.

Linear Regression

- Example: polynomial regression. This can be written as

$$\mathbb{E}\{Y_i\} = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p,$$

where x_i is the ith predictor variable corresponding to Y_i.

- For example we might fit a linear model of the form

$$\mathbb{E}\{Y_i\} = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i3}^2,$$

where x_{ki} is the value of the kth predictor for observation i.

- Note that

$$E(Y_i) = \beta_1 + \beta_2 x^{\beta_3},$$

is not a linear model.

- We will assume $p \leq n$ (full rank).

A matrix is said to have full rank if its rank equals the largest possible for a matrix of the same dimensions, which is the lesser of the number of rows and columns. The maximum number of linearly independent rows (columns) in a matrix A is called the row (column) rank of A.

- The rank of the matrix X is the dimension of the space spanned by the columns of X. Assume rank(X)=p.

Linear Regression

- We can also add further assumptions

  Second-order assumptions (SOA) $\mathrm{var}(\mathrm{Y}) = \sigma^2 I_n$ where σ^2 is unknown. Thus $\mathrm{var}(Y_i) = \sigma^2$ for all i and the Y_is are uncorrelated.

  Normal theory assumptions (NTA) The Y_is are independently and normally distributed with common unknown variance σ^2 so
  $$\mathrm{Y} \sim N(\mathrm{X}\beta, \sigma^2 I_n).$$

- NTA imples SOA but for now we will only assume the weaker SOA.

Linear Regression

- The linear model can be rewritten as

$$\begin{aligned} \mathrm{Y} &= \mathrm{X}\beta + \epsilon \quad \text{noise} \\ \begin{pmatrix} Y_1 \\ \cdots \\ Y_n \end{pmatrix} &= \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \\ &+ \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}. \end{aligned}$$

  where $\underline{E(\epsilon) = 0}$ and $\underline{\mathrm{var}(\epsilon) = \sigma^2 I_n}$.
- Minimise the difference between the observed values and the model fit to it.

Linear Regression

the residual sum of squares (RSS), also known as the sum of squared residuals (SSR) or the sum of squared estimate of errors (SSE), is the sum of the squares of residuals (deviations predicted from actual empirical values of data). It is a measure of the discrepancy between the data and an estimation model. A small RSS indicates a tight fit of the model to the data. It is used as an optimality criterion in parameter selection and model selection

- Find $\widehat{\beta}$ that minimise the residual sum of squares (RSS), i.e. find

$$\widehat{\beta} = \arg\min_{\beta}(\epsilon^T\epsilon = \sum_{i=1}^{n}\epsilon_i^2).$$

from before: $\epsilon = Y - X\beta$

- Write $\theta = \mathrm{X}\beta$. Then $\theta \in R(\mathrm{X}) = \Theta$, (the vector space spanned by the columns of X).
- The lse is the $\widehat{\theta}$ that minimises $||\mathrm{Y} - \theta||^2$, the square of the length of $\mathrm{Y} - \theta$. This is minimised when $\mathrm{Y} - \widehat{\theta}$ is perpendicular to Θ.

least square estimate

- v, is perpendicular to Θ if $\mathrm{X}^T v = 0$. Thus

$$\mathrm{X}^T(\mathrm{Y} - \widehat{\theta}) = 0 \quad so \quad \widehat{\beta} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{Y},$$

closed-form solution

if $\mathrm{X}^T\mathrm{X}$ is invertible.

Linear Regression

- Here, $\widehat{\beta}$ is the **ordinary least squares estimate** of β and is **unique**.
- Or:

$$\begin{aligned}\epsilon^T \epsilon &= (\mathrm{Y} - \mathrm{X}\beta)^T(\mathrm{Y} - \mathrm{X}\beta) \\ &= \mathrm{Y}^T\mathrm{Y} - 2\beta^T\mathrm{X}^T\mathrm{Y} + \beta^T\mathrm{X}^T\mathrm{X}\beta,\end{aligned}$$

- $\beta^T\mathrm{X}^T\mathrm{Y} = \mathrm{Y}^T\mathrm{X}\beta$ (both are scalars).
- Differentiating wrt β and setting to zero we see that

closed form solution
as we know it:

$$-2\mathrm{X}^T\mathrm{Y} + 2\mathrm{X}^T\mathrm{X}\beta = 0$$

$$\widehat{\beta} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{Y},$$

as

$$\frac{\partial}{\partial\beta}(\mathrm{a}^T\beta) = \mathrm{a}, \quad \frac{\partial}{\partial\beta}(\beta^T\mathrm{A}\beta) = 2A\beta.$$

Linear Regression

- $\widehat{\beta}$ is linear in Y, and $\widehat{\beta}$ is unbiased for β:

$$\begin{aligned} E(\widehat{\beta}) &= (X^TX)^{-1}X^TE(\underline{Y}) \\ &= \underbrace{(X^TX)^{-1}X^T(\underline{X}\beta)}_{=I} = \beta, \end{aligned}$$

$\mathbb{E}(Y) = X\beta$

- Let $A = (X^TX)^{-1}X^T$:

$$\begin{aligned} \mathbb{V}\mathrm{ar}(\widehat{\beta}) &= \mathbb{V}\mathrm{ar}(AY) \\ &= A\ \mathbb{V}\mathrm{ar}(Y)\ A^T \\ &= \sigma^2 AA^T \\ &= \sigma^2(X^TX)^{-1}\underbrace{X^TX(X^TX)^{-1}}_{=I} \\ &= \sigma^2(X^TX)^{-1}, \end{aligned}$$

formula 375 Matrix Cookbook (*)

as

(*)

$$\mathbb{V}\mathrm{ar}(AY) = A\ \mathbb{V}\mathrm{ar}(Y)\ A^T.$$

Linear Regression

- Gauss-Markov Theorem Among all unbiased linear estimates of $\boldsymbol{\beta}$ for a full rank linear model satisfying SOA, any linear combination of the least squares estimator $\widehat{\boldsymbol{\beta}}$ has the smaller or equal variance to that of any other, e.g. $\mathbb{V}\mathrm{ar}\{\mathrm{a}^T\widehat{\boldsymbol{\beta}}\} \leq \mathbb{V}\mathrm{ar}\{\mathrm{a}^T\widetilde{\boldsymbol{\beta}}\}$

?

Proof Write another estimator $\tilde{\beta} = \mathrm{BY}$ (linearity). We can calculate the expectation of this estimator to be

$$\begin{aligned} \mathbb{E}\{\tilde{\beta}\} &= \mathrm{B}\,\mathbb{E}\{\mathrm{Y}\} \\ &= \mathrm{BX}\boldsymbol{\beta} = \boldsymbol{\beta}. \end{aligned} \tag{1}$$

This implies that $\mathrm{BX} = \mathrm{I}$. We define

$$\mathrm{C} = \mathrm{B} - (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T \tag{2}$$

$$\tilde{\beta} = (\mathrm{C} + (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T)\mathrm{Y} = \widehat{\beta} + \mathrm{CY}. \tag{3}$$

and $\mathrm{CX} = 0$ to preserve unbiasedness.

Linear Regression

?

- For any constant vector a we note

$$\begin{aligned}\mathbb{V}\mathrm{ar}\{\mathrm{a}^T\widetilde{\boldsymbol{\beta}}\} &= \mathbb{V}\mathrm{ar}\{\mathrm{a}^T\{\widehat{\beta}+\mathrm{CY}\}\} \\ &= \mathrm{a}^T\,\mathbb{V}\mathrm{ar}\{\widehat{\beta}\}\mathrm{a}+\mathrm{a}^T\,\mathbb{V}\mathrm{ar}\{\mathrm{CY}\}\mathrm{a}+2\,\mathbb{C}\mathrm{ov}\{\mathrm{a}^T\widehat{\beta},\mathrm{a}^T\mathrm{CY}\}.\end{aligned} \tag{4}$$

We now only need to show that the covariance term is zero. As

$$\begin{aligned}\mathbb{C}\mathrm{ov}\{\mathrm{a}^T\widehat{\beta},\mathrm{a}^T\mathrm{CY}\} &= \mathrm{a}^T(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\,\mathbb{C}\mathrm{ov}\{\mathrm{Y},\mathrm{Y}\}\mathrm{C}^T\mathrm{a} \\ &= 0,\end{aligned} \tag{5}$$

and so the result follows. $\square$.

Simple Linear Regression

- Let

$(y = b + mx + \text{noise})$

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

- $\mathrm{Y}^T = (Y_1, \ldots, Y_n)$, $\boldsymbol{\beta}^T = (\beta_1, \beta_2)$ and

$$\mathrm{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

- Assume SOA and NO x_is are equal

$$X^TX = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} \sum_i^n 1 = n & \sum_i^n x_i = \bar{x}n \\ \sum_i^n x_i = \bar{x}_n & \sum_i^n x_i^2 \end{bmatrix}$$

$$\mathrm{X}^T\mathrm{X} = \begin{pmatrix} n & n\overline{x} \\ n\overline{x} & \sum x_i^2 \end{pmatrix}$$

inverse of 2x2 has a closed form solution: if $\underset{2\times 2}{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ then $A^{-1} = \frac{1}{\det A}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

$$(\mathrm{X}^T\mathrm{X})^{-1} = \frac{1}{n\sum x_i^2 - n^2\overline{x}^2} \begin{pmatrix} \sum x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{pmatrix}$$

$$\mathrm{X}^T\mathrm{Y} = \begin{pmatrix} n\overline{Y} \\ \sum x_i Y_i \end{pmatrix}.$$

Simple Linear Regression

from before

Now we can find $\widehat{\beta} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{Y}$, hence

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \frac{1}{\sum x_i^2 - n\overline{x}^2} \times \begin{pmatrix} \overline{Y}\sum x_i^2 - \overline{x}\sum x_i Y_i \\ \sum x_i Y_i - n\overline{x}\,\overline{Y} \end{pmatrix}.$$

$$\widehat{\beta}_2 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \overline{x})(Y_i - \overline{Y})}{\sum(x_i - \overline{x})^2}$$

$$\widehat{\beta}_1 = \overline{Y} - \widehat{\beta}_2\overline{x}.$$

$$\mathbb{V}\mathrm{ar}(\widehat{\beta}) = \frac{\sigma^2}{nS_{xx}}\begin{pmatrix} \sum x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{pmatrix}.$$

Simple Linear Regression

- If $\bar{x} = 0$ everything becomes easy: the covariance matrix is diagonal and $\widehat{\beta}_1 = \bar{Y}$.
- To get a diagonal covariance we adopting the alternative linear model

Cool

$$Y_i = \beta_1 + \beta_2(x_i - \bar{x}) + \epsilon_i, \qquad i = 1, \ldots, n.$$

Then we find that $\hat{\beta}_1 = \bar{Y}$, $\hat{\beta}_2 = S_{xy}/S_{xx}$ and

$$var(\widehat{\beta}) = \begin{pmatrix} n^{-1} & 0 \\ 0 & S_{xx}^{-1} \end{pmatrix}.$$

This idea could be generalised to orthogonal polynomials.

?

Linear Regression

- Let $\widehat{Y} = X\widehat{\beta}$. We found $\widehat{\beta}$ by minimising the RSS (Residual Sum of Squares),

$$
\begin{aligned}
e^T e &= \min_{\beta} \epsilon^T \epsilon \\
&= (Y - X\widehat{\beta})^T (Y - X\widehat{\beta}) \\
&= Y^T Y - 2\widehat{\beta}^T X^T Y + \widehat{\beta}^T X^T X \widehat{\beta} \\
&= Y^T Y - \widehat{\beta}^T X^T Y \\
&\quad + \widehat{\beta}^T (X^T X \widehat{\beta} - X^T Y) \\
&= (Y^T - \widehat{\beta}^T X^T) Y \\
&= Y^T (Y - X\widehat{\beta}) \\
&= Y^T Y - \widehat{\beta}^T X^T X \widehat{\beta}.
\end{aligned}
$$

from before

Linear Regression

- Also the RSS is given by

$$RSS = \mathrm{e}^T\mathrm{e} = \mathrm{Y}^T\mathrm{Y} - \widehat{Y}^T\widehat{Y},$$

the difference between the squares of the observed and fitted Y values.

- The **residuals** of the model are given by the difference between the observed and fitted values so that

$$\begin{aligned}\mathrm{e} &= \mathrm{Y} - \widehat{Y}\\ &= \mathrm{Y} - \mathrm{X}\widehat{\beta}\\ &= \{I_n - \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\}\mathrm{Y}\\ &= (I_n - \mathrm{P})\mathrm{Y},\end{aligned}$$

- $\mathrm{P} = \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T$ is known as the "hat" matrix and relates the fitted and observed responses as $\widehat{Y} = \mathrm{PY}$.

Linear Regression

- The hat matrix has a number of known properties:
  1. P is a symmetric $n \times n$ matrix
  2. P is idempotent so that $P^2 = P$
  3. The rank of Pis the same as rank X(i.e. both of rank p). From this note $\text{rank}(I_n - P) = n - rank(P) = n - p$ and that $(I_n - P)$ is also idempotent as

$$(I_n - P)^2 = I_n^2 - 2P + P^2 = I_n - P,$$

as $P^2 = P$.

- Firstly we find the $E(e) = 0$ as

$$E(e) = (I_n - P)E(Y) = (I_n - P)X\beta = 0,$$

$e = (I_n - P)Y$ from prev. slide and $E(Y) = \beta X$

as

$$\begin{aligned} PX &= X\overbrace{(X^TX)^{-1}X^TX}^{=I} \\ &= X \end{aligned}$$

Linear Regression

- More is known about the residuals:

  Theorem The residual sum of squares is an unbiased estimator of $(n-p)\sigma^2$.

- Thus we know that

$$\hat{\sigma}^2 = \frac{RSS}{n-p} = \frac{(\mathrm{Y}-\mathrm{X}\widehat{\beta})^T(\mathrm{Y}-\mathrm{X}\widehat{\beta})}{n-p} = \frac{\mathrm{Y}^T\mathrm{Y}-\widehat{Y}^T\widehat{Y}}{n-p},$$

slide 16

is an unbiased estimator of σ^2.

Linear Regression

- Note that

total sum of squares

fitted sum of squares

$$\begin{aligned}
\mathbb{E}\{RSS\} &= \mathbb{E}\{\mathsf{Y}^T\mathsf{Y} - \widehat{Y}^T\widehat{Y}\} \\
&= \mathbb{E}\{\{(\mathsf{I}-\mathsf{P})\mathsf{Y}\}^T\{(\mathsf{I}-\mathsf{P})\mathsf{Y}\}\} \\
&= \mathbb{E}\{\mathrm{trace}\{(\mathsf{I}-\mathsf{P})\mathsf{Y}\}\{(\mathsf{I}-\mathsf{P})\mathsf{Y}\}^T\} \\
&= \mathbb{E}\{\mathrm{trace}\{(\mathsf{I}-\mathsf{P})\mathsf{Y}\mathsf{Y}^T\{(\mathsf{I}-\mathsf{P})\}^T\}\} \\
&= \sigma^2\mathrm{trace}(\mathsf{I}-\mathsf{P}) \\
&= \sigma^2\{n-p\}.
\end{aligned}$$

In linear algebra, the trace of a square matrix A, is defined to be the sum of elements on the main diagonal (from the upper left to the lower right) of A

The result thus follows.

Maximum likelihood approach

Normal Theory Assumption

- Let $Y \sim N(X\beta, \sigma^2 I_n)$, i.e. NTA.
- The log-likelihood of the data is

$$\begin{aligned} L(\beta, \sigma^2) &= -\frac{n}{2}\log(2\pi\sigma^2) \\ &\quad -\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta). \end{aligned}$$

- maximising L with respect to β is equivalent to minimising $(Y - X\beta)^T(Y - X\beta)$
- The maximum likelihood estimate to σ^2 is RSS/n.

ie same estimator as Least Square

Maximum likelihood approach

- With NTA:

$$
\begin{aligned}
\hat\beta &\sim N(\beta, \overbrace{\sigma^2(\mathrm{X}^T\mathrm{X})^{-1}}^{\text{covariance}}) \\
V = \frac{(n-p)\hat\sigma^2}{\sigma^2} &\sim \chi^2_{n-p}
\end{aligned}
$$

independent

↳ chi-square dist

- $\hat\beta$ and $\hat\sigma^2$ are independent.

Theorem 15 If $\mathrm{A}=\{a_{ij}\} = (\mathrm{X}^T\mathrm{X})^{-1}$ (so $\mathrm{var}(\hat\beta) = \sigma^2\mathrm{A}$), then under NTA, the following are $100(1-\alpha)\%$ confidence intervals for the β_js and σ^2:

1. $(\hat\beta_j - t_{1-\alpha/2}\hat\sigma\sqrt{a_{jj}}, \hat\beta_j + t_{1-\alpha/2}\hat\sigma\sqrt{a_{jj}})$?
2. $\left(\frac{(n-p)\hat\sigma^2}{\chi^2_{1-\alpha/2}}, \frac{(n-p)\hat\sigma^2}{\chi^2_{\alpha/2}}\right)$

Maximum likelihood approach

repeated slide?

- With NTA:

$$\begin{aligned} \hat{\beta} &\sim N(\beta, \sigma^2 (X^T X)^{-1}) \\ V = \frac{(n-p)\hat{\sigma}^2}{\sigma^2} &\sim \chi^2_{n-p} \end{aligned}$$

- $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.

Theorem 15 If A=$\{a_{ij}\} = (X^T X)^{-1}$ (so $\mathrm{var}(\hat{\beta}) = \sigma^2$A), then under NTA, the following are $100(1-\alpha)\%$ confidence intervals for the β_js and σ^2:

1. $(\hat{\beta}_j - t_{1-\alpha/2}\hat{\sigma}\sqrt{a_{jj}}, \hat{\beta}_j + t_{1-\alpha/2}\hat{\sigma}\sqrt{a_{jj}})$
2. $\left(\frac{(n-p)\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}, \frac{(n-p)\hat{\sigma}^2}{\chi^2_{\alpha/2}}\right)$

Residuals

Let

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

but that the analyst incorrectly assumes that

$$Y_i = \beta_0 + \epsilon_i$$

Then

$$\begin{aligned} E\{e_i\} &= E\left\{Y_i - \hat{\beta}_0\right\} \\ &= E\left\{Y_i - \frac{1}{n}\sum Y_i\right\} \\ &= \frac{n-1}{n}(\beta_1 x_i) + \frac{1}{n}\sum_{j\neq i}(\beta_1 x_j) \\ &= \beta_1(x_i - \bar{x}) \end{aligned} \tag{6}$$

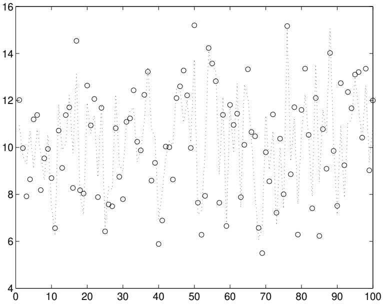

Figure:

Here $Y_i = 10 + 2x_i + 3\epsilon_i$. This is not apparent from the plot, of Y_i (dots) and $E_{\mathrm{Y}|\beta,\sigma^2}(Y_i)$ (dotted line).

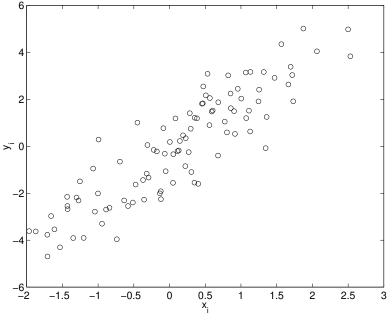

Figure:

Looking at a plot of the residuals against the explanatory variable gives a different opinion.

last slide missing