Week 9.

This week we shall look at <u>Regression analysis</u>.

<u>Refs</u>: • Wikipedia
- James, Witten, Hastie, Tibshirani "An intro to stat. learning" Chap 3.
- Anderson (2003).
- Bai, Jiang, Yao, Zheng (2013) - Testing linear hypotheses.
- Xie, Xiao (2016?) - likelihood ratio test for high-dim linear reg.

Regression is a massive body of litterature. I will only focus on a few topics.

"Method of least squares" Legendre (1805), Gauss (1809)

"regression" term coined by Galton to describe biological phenomenon: "regression toward the mean".

Given data $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$ and we want to determine model

$$y \simeq f(x; \beta)$$

where β is a parameter (vector or scalar). This is the simple univariate case as y is scalar.

We need to specify model f.

In linear regression, the model specification is that the dependent variable y_i is a linear combination of the parameters.

E.g. linear model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i=1,\dots,n.$

parabola: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \quad i=1,2,\dots,n.$

Regression is a supervised learning problem: we are given outputs and inputs and we have to learn the model.

Stats. v. Machine/Statistical learning.

$\Rightarrow$ Stats: learn something about phenomena (Science) and the focus is on supporting and rejecting hypothesis.

$\Rightarrow$ ML: Focus is on optimal out-of-sample prediction. No need or desire to understand the model. "Black box".

More generally in the univariate case, we have observations

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)$$

where $\mathbf{x}_i$ is p-dimensional. We assume model of the form

$$y_i \approx \beta' \mathbf{x}_i \qquad \beta := (\beta_0, \beta_1, \dots, \beta_{p-1})'$$

for $i = 1, \dots, n$ and $n > p$. $\qquad \mathbf{x}_i := (1, x_{i2}, x_{i3}, \dots, x_{ip})$

In other words our model is $f(\mathbf{x}; \beta) := \beta' \mathbf{x}$. This gives us a number of decisions $\mathcal{D} := \{f(\mathbf{x}_1; \beta), f(\mathbf{x}_2; \beta), \dots, f(\mathbf{x}_n; \beta)\}$ and our outcome space $\mathcal{Y} = \{y_1, y_2, \dots, y_n\}$.

We score our model by a loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$.

One example is the least squares loss

$$\ell(\mathcal{D}, \mathcal{Y}) = \sum_{i=1}^{n} |y_i - \beta' \mathbf{x}_i|^2 =: S(\beta)$$ (as only depends on choice of β)

We want to minimise our loss by varying β. Then the optimal choice $\hat{\beta}$ is

$$\hat{\beta} := \arg\min_{\beta} S(\beta).$$

Our problem can be written in matrix notation.

$$\mathbb{X}\beta = \mathbb{Y}$$

where

$$\mathbb{X} = \begin{bmatrix} \mathbb{x}_1' \\ \mathbb{x}_2' \\ \vdots \\ \mathbb{x}_n' \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \mathbb{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

So that $S(\beta) = \sum_{i=1}^{n} |y_i - \beta' \mathbb{x}_i|^2 = \| \mathbb{Y} - \mathbb{X}\beta \|^2$

and the solution of the minimisation is given by the normal equations

$$\mathbb{X}'\mathbb{X}\hat{\beta} = \mathbb{X}'\mathbb{Y}$$

and the solution is $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}$.

Defining $e_i := |y_i - \beta' \mathbb{x}_i|$ then $S(\beta) = e_1^2 + e_2^2 + \cdots + e_n^2$ which gives the name Residual sum of squares (RSS) or sum of squares error (SSE).

We assume true relationship $Y = \beta' x + \varepsilon$ where ε is a mean-zero random error term.

Multivariate case

observations $(\mathbb{Y}_1, \mathbb{X}_1), \dots (\mathbb{Y}_n, \mathbb{X}_n)$

$\mathbb{Y}_i$: p-dimensional $\quad \mathbb{X}_i$: q-dimensional.

Case $p=1$ gives back univariate case.

Assume true relationship between $\mathbb{Y}$ and $\mathbb{X}$ is given by

$$\mathbb{Y}_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{iq} + \varepsilon_i, \qquad i = 1, 2, \dots, n.$$

$$\mathbb{Y}_i = (Y_{i1}, Y_{i2}, \dots Y_{ip})', \qquad \mathbb{X}_i = (x_{i1}, x_{i2}, \dots, x_{iq})'$$

where $\beta_0, \beta_1, \dots, \beta_q$ are p-dimensional parameter vectors to be determined. The error terms $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are p-dimensional random vectors. We assume they are mutually uncorrelated, and that

$$E[\varepsilon_i] = \underset{\sim}{0} \qquad Var[\boldsymbol{\varepsilon}_i] = \Sigma \qquad i = 1, 2, \dots, n.$$

In matrix form

$$\mathbb{Y} = \begin{bmatrix} \mathbb{Y}_1' \\ \vdots \\ \mathbb{Y}_n \end{bmatrix}, \qquad \mathbb{X} = \begin{bmatrix} 1 & \mathbb{X}_1' \\ \vdots & \vdots \\ 1 & \mathbb{X}_n' \end{bmatrix}, \qquad \varepsilon = \begin{bmatrix} \varepsilon_1' \\ \vdots \\ \varepsilon_n' \end{bmatrix}$$

giving

$$\mathbb{Y} = \mathbb{X}\mathcal{B} + \mathcal{E}.$$

$$\mathcal{B} := (\beta_0, \beta_1, \cdots, \beta_q)'.$$

$$\mathbb{E}[\mathbb{Y}] = \mathbb{X}\mathcal{B}, \quad Var[\mathbb{Y}_i] = \Sigma \text{ for all } i = 1, \cdots, n.$$

$$Cov(\mathbb{Y}_i, \mathbb{Y}_j) = 0 \text{ for all } i \neq j.$$

The RSE is $\sum_{i=1}^{n} \sum_{j=1}^{p} \varepsilon_{ij}^2 = tr\,(\mathbb{Y} - \mathbb{X}\mathcal{B})'(\mathbb{Y} - \mathbb{X}\mathcal{B}) = S(\mathcal{B})$

Similar to the univariate case, we seek $\hat{\mathcal{B}} = \underset{\mathcal{B}}{\operatorname{argmin}}\, S(\mathcal{B})$ and the solution is given by

$$\hat{\mathcal{B}} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}.$$

Assume ε_i are normally distributed, then $\hat{\mathcal{B}}$ is the MLE of $\mathcal{B}$ and the MLE of Σ is

$$\hat{\Sigma} = \mathbb{S} = \frac{1}{n}(\mathbb{Y} - \mathbb{X}\hat{\mathcal{B}})'(\mathbb{Y} - \mathbb{X}\hat{\mathcal{B}}).$$

Thm: For the estimators $\hat{\mathcal{B}}$ and $\hat{\Sigma}$ it holds that:

(1) $E[\hat{\mathcal{B}}] = \mathcal{B}$ $\quad E[\hat{\Sigma}] = \frac{1}{n(n-k+1)}\Sigma$ $\quad Cov(\hat{\beta}_i, \hat{\beta}_j) = v_{ij}\Sigma$

where $V = (\mathbb{X}'\mathbb{X})^{-1} = (v_{ij})$

(2) if $\mathbb{Y}$ is normally distributed then $\hat{\mathcal{B}} \sim N_{(k+1)\times p}(\mathcal{B}, \Sigma \otimes V)$
$n\hat{\Sigma} \sim W_p(n-q+1, \Sigma)$ and $\hat{\mathcal{B}}$ and $n\hat{\Sigma}$ are independent

Here, $\otimes$ is Kroneker product $[A \doteq (a_{ij})$ and B then $A \otimes B = (a_{ij}B)]$.

#

Note: $\hat{\Sigma}$ is not unbiased. A common unbiased estimator is

$$\tilde{\Sigma} = \frac{1}{n-k}(\mathbb{Y} - \mathbb{X}\hat{\mathcal{B}})'(\mathbb{Y} - \mathbb{X}\hat{\mathcal{B}})$$

Following Anderson (2003), under normality assumption, we can discuss the equivalent problem: $\mathbb{X}_1, \mathbb{X}_2, \cdots \mathbb{X}_n$ set of n independent observations

$$\mathbb{X}_k \sim N_p(B\mathbb{Z}_k, \Sigma)$$

where the vectors $\mathbb{Z}_k$ (q-dimensional) are called design vectors.

Caution: In other words, my Y_i's become X_i's and my independent variables become Z_i's (instead of X_i).

Inference on the parameters B.

Assume $n \geq p+q$ and rank of $Z = (Z_1, Z_2, \ldots Z_n)$ is q.

Aim: We may want to test that a subset of the inputs $\{Z_i\}_{i=1,\ldots,n}$ play no role in predicting the X_i.

Partition the parameters $B = (B_1, B_2)$ so that B_1 has q_1 columns and B_2 has q_2 columns.

We are going to look at the likelihood ratio criterion for testing the hypothesis

$$H_0 : B_1 = B_{01}$$

where B_{01} is a given matrix (eg. $B_{01} = \underset{\sim}{0}$ matrix)

The maximum likelihood $L = f_{B,\Sigma}(\mathbb{X})$ of a sample $\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_n$ is

$$\max_{B,\Sigma} L = (2\pi)^{-\frac{1}{2}pn} |\hat{\Sigma}_\Omega|^{-\frac{1}{2}n} e^{-\frac{1}{2}pn}$$

where $\Sigma_\Omega = \hat{\Sigma}$.

We need to restrict the MLE for the parameters to the subspace ω induced by the null hypothesis.

Set $\quad \mathbb{Y}_k = \mathbb{X}_k - B_{01} \mathbb{Z}_{1k} \quad k=1, \ldots, n.$

where $\mathbb{Z}_k = \begin{pmatrix} \mathbb{Z}_{1k} \\ \mathbb{Z}_{2k} \end{pmatrix}$ for $k=1,2,\ldots,n$, is partitioned in same way as B.

Then $\mathbb{Y}_k$ can be considered as an observation from $N(B_2 \mathbb{Z}_{2k}, \Sigma)$.

Using our new notation. (Anderson-style)

$$\hat{B} = CA^{-1}$$

$$C = \sum_{k=1}^{n} \mathbb{X}_k \mathbb{Z}_k' \qquad A = \sum_{k=1}^{n} \mathbb{Z}_k \mathbb{Z}_k'$$

old notation:

$$\hat{\beta} = \underbrace{(\mathbb{X}'\mathbb{X})^{-1}}_{A^{-1}} \underbrace{\mathbb{X}'\mathbb{Y}}_{C}$$

By partitioning

$$\hat{B}_{2\omega} = \sum_{k=1}^{n} Y_k Z_{2k}' A_{22}^{-1} = \sum (X_k - B_{0k} Z_{1k}) Z_{2k}' A_{22}^{-1}$$

$$= (C_2 - B_{01} A_{12}) A_{22}^{-1}$$

with $C = (C_1 \; C_2)$ and $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$

The estimator of Σ is

$$n\hat{\Sigma}_\omega = \sum_{k=1}^{n} (Y_k - \hat{B}_{2\omega} Z_{2k})(Y_k - \hat{B}_{2\omega} Z_{2k})'$$

$$= \sum_{k=1}^{n} Y_k Y_k' - \hat{B}_{2\omega} A_{22} \hat{B}_{2\omega}'$$

$$= \sum_{k=1}^{n} (X_k - B_{01} Z_{1k})(X_k - B_{01} Z_{1k})' - \hat{B}_{2\omega} A_{22} \hat{B}_{2\omega}$$

This gives the MLE over ω:

$$\max_{B_2, \Sigma} L = (2\pi)^{-\frac{1}{2}pn} |\hat{\Sigma}_\omega|^{-\frac{1}{2}n} e^{-\frac{1}{2}pn}.$$

The likelihood ratio criterion for testing H_0 is

$$\lambda = \frac{|\hat{\Sigma}_\Omega|^{\frac{1}{2}n}}{|\hat{\Sigma}_\omega|^{\frac{1}{2}n}} \quad \text{reject for } \lambda < \lambda_0.$$

Distribution of λ when H_0 is true

Likelihood ratio criterion λ can be written in terms of

$$U = \lambda^{2/n} = \frac{|\hat{\Sigma}_\Omega|}{|\hat{\Sigma}_\omega|}$$

and (p 296, Anderson 2003):

$$n\hat{\Sigma}_\omega = n\hat{\Sigma}_\Omega + (\hat{B}_{1\Omega} - B_{01}) A_{11\cdot 2} (\hat{B}_{1\omega} - B_{01})'$$

where $A_{11\cdot 2} := A_{11} - A_{12} A_{22}^{-1} A_{21}$.

So we can write

$$U = \frac{|n\hat{\Sigma}_\Omega|}{|n\hat{\Sigma}_\Omega + (\hat{B}_{1\Omega} - B_{01}) A_{11\cdot 2} (\hat{B}_{1\Omega} - B_{01})'|}$$

Lemma: Set $G := n\hat{\Sigma}_\Omega$ $\quad H := (\hat{B}_{1\Omega} - B_{01}) A_{11\cdot 2} (\hat{B}_{1\Omega} - B_{01})'$

Then $G \sim W(n-q, \Sigma)$, $H \sim W(q_1, \Sigma)$, and they are independent. #

The distribution of U can be characterised in terms of a product of beta variables.

$$U = V_1 V_2 \cdots V_p.$$

where $V_1 = g_{11}/(g_{11}+h_{11})$ and $V_i = \dfrac{|G_i|}{|G_{i-1}|} \Big/ \dfrac{|G_i + H_i|}{|G_{i-1} + H_{i-1}|}$

and G_i and H_i are submatrices of G and H, respectively, of the first i rows and columns.

<u>Theorem</u>: When H_0 true, $U = \prod_{i=1}^{p} V_i$ where $V_1, V_2 \cdots, V_p$ independent and V_i has the beta density

$$\beta\left[v; \tfrac{1}{2}(n-q+1-i), \tfrac{1}{2}q_1\right]$$

$$= \frac{\Gamma\left(\frac{1}{2}(n-q+q_1+1-i)\right)}{\Gamma\left(\frac{1}{2}(n-q+1-i)\right)\Gamma\left(\frac{1}{2}q_1\right)} v^{\frac{1}{2}(n-q+1-i)-1}$$

$$\times (1-v)^{\frac{1}{2}q_1 - 1}.$$

$\#$

Typically this is too difficult to work with.

One can develop an asymptotic approximation.

Let $\mathcal{U} =: \mathcal{U}_{p,q_1,n-q}$ where $q = q_1 + q_2$ and let

$u_{p,q_1,n}(\alpha)$ be the significance point for $\mathcal{U}_{p,q_1,n-q}$

that is,

$$P\left(\mathcal{U}_{p,q_1,n-q} \leq u_{p,q_1,n-q}(\alpha) \mid H_0\right) = \alpha .$$

It can be shown (Section 8.5, Anderson 2003) that

$$-(n-q-\tfrac{1}{2}(p-q_1+1))\log\left[\mathcal{U}_{p,q_1,n-q}\right]$$

has a limiting χ^2-distribution with pq_1 degrees of freedom (There is also a Normal and F distribution approx).

Let $\chi^2_{pq_1}(\alpha)$ denote the α significance point of $\chi^2_{pq_1}$ and let

$$C_{p,q_1,n-q-p+1}(\alpha) := \frac{-(n-q-\frac{1}{2}(p-q_1+1))\log(u_{p,q_1,n}(\alpha))}{\chi^2_{pq_1}(\alpha)} .$$

Reject hypothesis if

$$-(n-q-\tfrac{1}{2}(p-q_1+1))\log \mathcal{U}_{p,q_1,n-q} > C_{p,q_1,n-q-p+1}(\alpha)\, \chi^2_{pq_1}(\alpha)$$

The values of $C_{p,q_1,n-q-p+1}(\alpha)$ are usually tabulated somewhere (eg. Anderson, Appendix B, Table 1), and serve as a correction factor wrt. asymptotic χ^2 quantile.

This χ^2 approximation is not very good and a number of researches have tried to correct it.

<u>Theorem</u> (Box and Bartlett). With $k = n - \frac{1}{2}(p - q_1 + 1)$ the CDF of $-k\log(U_{p,q_1,n-q})$ has the following expansion

$$\mathbb{P}(-k\log(U_{p,q_1,n-q}) \leq z) = \varphi_{pq_1}(z) + \frac{\gamma_2}{k^2}\{\varphi_{pq_1+4}(z) - \varphi_{pq_1}(z)\}$$

$$+ \frac{1}{k^4}\left[\gamma_4\{\varphi_{pq_1+8}(z) - \varphi_{pq_1}(z)\} - \gamma_2^2\{\varphi_{pq_1+4}(z) - \varphi_{pq_1}(z)\}\right] + R_n$$

where $\varphi_m(z) := \mathbb{P}(\chi_m^2 \leq z)$ and

$$\gamma_2 = \frac{pq_1(p^2 + q_1^2 - 5)}{48}$$

$$\gamma_4 = \frac{\gamma_2^2}{2} + \frac{pq_1}{1920}\left[3p^4 + 3q_1^4 + 10p^2q_1^2 - 50(p^2 + q_1^2) + 159\right]$$

R_n order $O(n^{-6})$.

We have considered some techniques for testing the general linear hypothesis:

- Distribution of U as product of beta variables; See page 12. (too difficult to implement?)
- χ^2 approximation (poor performance $p > 2$).
- Box and Bartlett correction (still poor for large p).

We now look at the high-dimensional regime using RMT, following Bai et al. (2013).

Recall $U \overset{d}{=} \dfrac{|G|}{|H|} = |GH^{-1}|$ $\quad H \sim W(q_1, \Sigma)$, $\quad G \sim W(n-q, \Sigma)$

So $T_n = -n \log U = n \sum_{j=1}^{p} \log(1 + \ell_j)$

where ℓ_j are eigenvalues of HG^{-1}.

We can assume without loss of generality

$$\Sigma = I_p.$$

For fixed p and q (# params fixed) the eigenvalues behave like $l_j \xrightarrow{P} 0$ at a rate $1/n$. (as $n \to \infty$)

This gives the approximation

$$T_n = n \sum_{j=1}^{n} l_j + O_p(1/n).$$

and T_n is distributed according to $\chi^2_{pq_1}$.

When p, q, n become large together this is not the expected behaviour of the eigenvalues

Assume

$$\frac{p}{q_1} \to y_1 \qquad \frac{p}{n-q} \to y_2 \in (0,1)$$

Define

$$F = \frac{n-q}{q_1} H G^{-1}$$

If populations are normal distributed then we have seen that F is distributed to a random Fisher matrix with $(q_1, n-q)$ degrees of freedom.

Let ℓ_j be eigenvalues of F and define the finite horizon proxies

$$\frac{p}{q_1} =: y_{n_1}, \quad \frac{p}{n-q} =: y_{n_2}.$$

The statistic can be rewritten

$$T_n = n\sum_{j=1}^{p}\left[\log(y_{n_1} + y_{n_2}\ell_j) - \log y_{n_1}\right] =: n\sum_{j=1}^{p} f(\ell_j)$$

with function $f(x) = \log\left(1 + \frac{y_{n_2}}{y_{n_1}} x\right)$.

A CLT can be derived; see Bai et al (2013).

<u>Theorem:</u> Under Ho true,

$$\frac{1}{n} T_n - \mu_n \xrightarrow{\mathcal{D}} N(\eta, \sigma^2).$$

Where

$$\mu_n = -(n-q-p)\log c_n - (q_1-p)\frac{y_{n_1}-1}{y_{n_1}}\log(c_n - d_n h_n)$$
$$+(n-q+q_1)\log\left(\frac{c_n h_n - d_n y_{n_2}}{h_n}\right)$$

$$\eta = \frac{1}{2}\log(y_1+y_2-y_1y_2)$$

$$\sigma^2 = 2\log\left(\frac{y_1+y_2-y_1y_2}{(y_1+y_2)(1-y_2)}\right)$$

$$c_n := \frac{h_n}{\sqrt{y_{n_1}}(1-y_{n_2})} \qquad d_n := \frac{y_{n_2}}{\sqrt{y_{n_1}}(1-y_{n_2})}$$

$$h_n := y_{n_1} + y_{n_2} - y_{n_1}y_{n_2}.$$