Εφαρμοσμένη Στατιστική

Δημήτριος Μπάγκαβος

Τμήμα Μαθηματικών και Εφαρμοσμένων Μαθηματικών
Πανεπιστήμιο Κρήτης

17 Απριλίου 2018

Πολλαπλή Γραμμική Παλινδρόμηση.

- Σε αρκετά προβλήματα η μεταβλητή απόκρισης Y μπορεί επηρεάζεται από περισσότερες από μια ερμηνευτικές μεταβλητές, έστω $X_1, X_2, \ldots, X_p$.
- Μπορούμε να γενικεύσουμε το απλό γραμμικό μοντέλο για να διερευνήσουμε την εξάρτηση της Y από τις $X_1, X_2, \ldots, X_p$ με μία σχέση της μορφής

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

δεδομένου δείγματος μεγέθους n από την εξαρτημένη και τις ανεξάρτητες μεταβλητές.

- Αυτό σημαίνει ότι για την $i-$οστή παρατήρηση έχουμε μετρήσεις $(Y_i, X_{i1}, X_{i2}, \ldots, X_{ip})$ που θεωρούνται γνωστοί αριθμοί.
- Τα σφάλματα ε_i θεωρούνται ανεξάρτητες τ.μ. με $\varepsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n$
- Όπως και στο απλό γραμμικό μοντέλο οι $X_1, X_2, \ldots, X_p$ δεν θεωρούνται τ.μ.
- Για ευκολία, το πολλαπλό γραμμικό μοντέλο γράφεται με τη βοήθεια πινάκων ως

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

Πολλαπλή Γραμμική Παλινδρόμηση.

όπου

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1p} \\ 1 & X_{21} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

- Ο πίνακας $\mathbf{X}$ καλείται και πίνακας σχεδιασμού,
- Στο πολλαπλό μοντέλο θεωρούμε ότι τα σημεία δεν βρίσκονται «κοντά» σε μια ευθεία αλλά «κοντά» σε ένα (υπερ)επίπεδο p διαστάσεων.
- Εφόσον το τ.δ. ε αποτελείται από n ανεξάρτητες $N(0, \sigma^2)$ τ.μ., κατά συνέπεια η από κοινού κατανομή του ε είναι $N(0, \sigma^2 I_n)$ όπου I_n είναι ο $n \times n$ μοναδιαίος πίνακας.
- Άρα,

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n)$$

δηλαδή ακολουθεί την πολυδιάστατη κανονική κατανομή.

Πολυδιάστατη κανονική κατανομή.

- Παρατήρηση: Η πολυδιάστατη κανονική κατανομή ορίζεται ως

$$f(x_1,\ldots,x_k) = \frac{1}{(2\pi)^{k/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

όπου

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}, \ \mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \ldots & \sigma_{kk} \end{pmatrix}.$$

- όπου A^T συμβολίζει ανάστροφο πίνακα, ο $\mathbf{\Sigma}$ είναι $k \times k$ πίνακας με $\mathbf{x^T\Sigma x} > 0$ για κάθε $x \neq 0$, $|\Sigma|$ η ορίζουσα του.
- Αποδεικνύεται ότι αν $Z = (Z_1, Z_2, \ldots, Z_k) \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$ τότε $\mathrm{Cov}(Z_i, Z_j) = \sigma_{ij}$.
- Δηλαδή ο $\mathbf{\Sigma}$ είναι ο πίνακας διασπορών - συνδιασπορών του Z - γράφουμε $V(Z) = \mathbf{\Sigma}$.
- Επίσης, $Z_i \sim N(\mu_i, \sigma_{ii})$ δηλ. στη διαγώνιο βρίσκονται οι διασπορές των Z_i.
- Όταν $\mathbf{\Sigma} = \sigma^2 I_k$ τότε οι τ.μ. $Z_1, Z_2, \ldots, Z_k$ είναι ανεξάρτητες κανονικές τ.μ. με διασπορά σ^2.

Πολλαπλή Παλινδρόμηση: Εκτίμηση παραμέτρων.

Θεώρημα 1:

Με δεδομένο ότι $\mathsf{Rank}(\mathbf{X}) = p$, οι εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων $\boldsymbol{\beta}$ του μοντέλου (1) δίνονται από την $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y}$

Απόδειξη: Η συνάρτηση πιθανοφάνειας της $\mathbf{Y}$ δεδομένων n παρατηρήσεων $y_1, \ldots, y_n$ είναι η

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}\sigma} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}} \qquad (2)$$

και μεγιστοποιείται ως προς $\boldsymbol{\beta}$ όταν ελαχιστοποιείται το

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \sum_{i=1}^{n} \varepsilon_i^2.$$

Παραγωγίζοντας ως προς $\boldsymbol{\beta}$ ($\frac{d}{d\boldsymbol{\beta}} = (\frac{\partial}{\partial\beta_1}, \frac{\partial}{\partial\beta_2}, \ldots, \frac{\partial}{\partial\beta_n})$) έχουμε ότι

$$\frac{d}{d\boldsymbol{\beta}}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = \frac{d}{d\boldsymbol{\beta}}(\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}) = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

Η παραπάνω παράγωγος (δηλ. το διάνυσμα των μερικών παραγώγων) είναι ίση με 0 όταν

Πολλαπλή Παλινδρόμηση: Εκτίμηση παραμέτρων.

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{Y}^T\mathbf{Y}. \quad \text{(κανονικές εξισώσεις)} \tag{3}$$

Στην περίπτωση του μοντέλου (1) η (3) είναι ένα $p \times p$ σύστημα που έχει μοναδική λύση όταν υπάρχει ο αντίστροφος του $\mathbf{X^TX}$.

Για να υπάρχει ο $(\mathbf{X^TX})^{-1}$ θα πρέπει να ισχύει $\text{Rank}(\mathbf{X^TX}) = \min(n, p) = p$ το οποίο όμως ικανοποιείται γιατί από την υπόθεση του θεωρήματος, $\text{Rank}(\mathbf{X}) = p$ και ξέρουμε από τη γραμμική άλγεβρα ότι αυτό συνεπάγεται και ότι $\text{Rank}(\mathbf{X^TX}) = p$.

Άρα ο αντίστροφος υπάρχει οπότε, λύνοντας την (3) ως προς $\boldsymbol{\beta}$ παίρνουμε $\hat{\boldsymbol{\beta}} = (\mathbf{X^TX})^{-1}\mathbf{X^TY}$ το οποίο τελειώνει την απόδειξη.
Σημείωση: Τέτοια μοντέλα λέγονται μοντέλα πλήρους βαθμίδας.

- Όταν ο $(\mathbf{X^TX})^{-1}$ δεν υπάρχει (δηλ. $\text{Rank}(\mathbf{X^TX}) = \min(n, p) = n$ (τι σημαίνει αυτό;) τότε το μοντέλο λέγεται μη πλήρους βαθμίδας και η έννοια της παλινδρόμησης δεν ισχύει.
- Τέτοια μοντέλα αντιστοιχούν σε μοντέλα ανάλυσης διακύμανσης.

Εκτίμηση των Y_i και ορισμός υπολοίπων.

- Προβλέψεις των Y_i ή προσαρμοσμένες τιμές (πάνω στο εκτιμημένο επίπεδο γραμμικής παλινδρόμησης) των Y_i ονομάζουμε τις εκτιμήσεις της μέσης τιμής της Y_i, έστω $\mathbb{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$, από τις

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}, i = 1, \ldots, n.$$

- Συνοπτικά, γράφουμε

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y} = \mathbf{P}\mathbf{Y}$$

όπου θέσαμε $\mathbf{P} = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}$.

- Όπως στο απλό γραμμικό μοντέλο έτσι και εδώ, οι διαφορές των προσαρμοσμένων $\hat{Y}_i$ από τις παρατηρούμενες Y_i καλούνται υπόλοιπα ή εκτιμημένα σφάλματα και συμβολίζονται με

$$\varepsilon_i = Y_i - \hat{Y}_i, i = 1, 2, \ldots, n.$$

- Συνοπτικά, γράφουμε

$$\boldsymbol{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = (I_n - \mathbf{P})\mathbf{Y}.$$

Πολλαπλή Παλινδρόμηση: Εκτίμηση παραμέτρων.

Θεώρημα 2:

Στο μοντέλο (1) με $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$ ο εκτιμητής ε.μ.π. του σ^2 είναι $\tilde{\sigma}^2 = n^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$

Απόδειξη: Λογαριθμίζοντας και παραγωγίζοντας την (2) παίρνουμε

$$\frac{\partial L(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

από όπου εξισώνοντας την παραπάνω εξίσωση με 0 και λύνοντας ως προς σ^2 παίρνουμε το ζητούμενο.

- Παρατηρούμε ότι ισοδύναμα ο ε.μ.π. της σ^2 γράφεται

$$\tilde{\sigma}^2 = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_p X_{ip})^2 \equiv n^{-1}\mathrm{SSR}. \quad (4)$$

- Επίσης, βλέπουμε ότι $\mathbb{E}\tilde{\sigma}^2 \neq \sigma^2$ δηλ. ο $\tilde{\sigma}^2$ δεν είναι αμερόληπτος εκτιμητής του σ^2.

Πολλαπλή Παλινδρόμηση: Ιδιότητες εκτιμητών.

Θεώρημα 3:

Ένας αμερόληπτος εκτιμητής του σ^2 είναι ο $\hat{\sigma}^2 = (n-p)^{-1}\text{SSR}$

Απόδειξη: Γράφουμε $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y}$ οπότε $\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (I_n - \mathbf{P})\mathbf{Y}$. Οπότε

$$(n-p)\text{SSR} = \mathbf{Y}^\mathbf{T}(I_n - \mathbf{P})(I_n - \mathbf{P})\mathbf{Y} = \mathbf{Y}^\mathbf{T}(I_n - \mathbf{P})\mathbf{Y}$$

όπου χρησιμοποιήσαμε ότι $I_n - \mathbf{P} = (I_n - \mathbf{P})^T = (I_n - \mathbf{P})^2$ δηλαδή ότι ο πίνακας είναι ταυτοδύναμος (ο πίνακας $\mathbf{A}$ είναι ταυτοδύναμος (idempodent) όταν $\mathbf{A}^2 = \mathbf{A}$). Ξέρουμε ότι για ταυτοδύναμους πίνακες ισχύει $\text{tr}(\mathbf{A}) = \text{rank}(\mathbf{A})$. Επίσης χρησιμοποιούμε το ότι για $n \times n$ συμμετρικό πίνακα $\mathbf{A}$ και $n \times 1$ τυχαίο δ/μα $\mathbf{x}$ με $\mathbb{E}\mathbf{x} = \mu$ και $\mathbb{V}(\mathbf{x}) = \boldsymbol{\Sigma}$ ισχύει

$$\mathbb{E}(\mathbf{x}^\mathbf{T}\mathbf{A}\mathbf{x}) = \text{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \mu^\mathbf{T}\mathbf{A}\mu.$$

Κατά συνέπεια,

$$\begin{aligned}(n-p)\mathbb{E}\{\text{SSR}\} = \mathbb{E}\left\{\mathbf{Y}^\mathbf{T}(I_n - \mathbf{P})\mathbf{Y}\right\} &= \sigma^2\text{tr}(I_n - \mathbf{P}) + \boldsymbol{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}(I_n - \mathbf{P})\mathbf{X}\boldsymbol{\beta} \\ &= \sigma^2\text{tr}(I_n - \mathbf{P}) = \sigma^2(n-p)\end{aligned}$$

από όπου προκύπτει αμέσως το ζητούμενο.

Πολλαπλή Παλινδρόμηση: Ιδιότητες εκτιμητών.

Θεώρημα 4:

Δεδομένου ότι για το μοντέλο (1) με $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$ ισχύει $\mathbb{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ και $\mathbb{V}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$

Απόδειξη: Χρησιμοποιούμε ότι από την (1), $\mathbb{E}\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbb{E}\boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta}$ οπότε,

$$\mathbb{E}\hat{\boldsymbol{\beta}} = \mathbb{E}\left\{(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y}\right\} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbb{E}\mathbf{Y} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

Σημείωση 1: Από την παραπάνω εξίσωση παίρνουμε ότι για οποιαδήποτε διάνυσμα σταθερών τιμών $\boldsymbol{\lambda}$, $\mathbb{E}\{\boldsymbol{\lambda}\hat{\boldsymbol{\beta}}\} = \boldsymbol{\lambda}\boldsymbol{\beta}$.
Ακόμα,

$$\begin{aligned}\mathbb{V}(\hat{\boldsymbol{\beta}}) = \mathbb{V}\left((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y}\right) &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbb{V}(\mathbf{Y})\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \\ &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}^\mathsf{T}\sigma^2 I_n\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\end{aligned}$$

Σημείωση 2: Από την παραπάνω εξίσωση παίρνουμε ότι για οποιαδήποτε διάνυσμα σταθερών τιμών $\boldsymbol{\lambda}$, $\mathbb{V}\{\boldsymbol{\lambda}\hat{\boldsymbol{\beta}}\} = \sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{\lambda}$.

Κατανομή των παραμέτρων του μοντέλου.

Θεώρημα 4:

Με δεδομένο ότι $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I_n)$, έχουμε ότι

1. $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})$,
2. $\sigma^{-2}(n-p)\hat{\sigma}^2 \sim \chi^2_{n-p}$

όπου N_p συμβολίζει πολυδιάστατη κανονική κατανομή, ενώ $\sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ είναι ο πίνακας διακύμανσης / συνδιακύμανσης της.

Απόδειξη: Για το 1, πρώτα γράφουμε

$$\hat{\boldsymbol{\beta}} = \underbrace{(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}}_{\mathbf{C}}\mathbf{Y} \equiv \mathbf{C}\mathbf{Y} \tag{5}$$

όπου ο $\mathbf{C}$ είναι $p \times n$ πίνακας με $\text{rank}(\mathbf{C}) = \text{rank}(\mathbf{X}^\mathsf{T}) = \text{rank}(\mathbf{X}) = p$.

> Όταν $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ τότε $\mathbf{C}\mathbf{x} \sim N_p(\mathbf{C}\mathbf{x} - \boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\mathsf{T})$

Εφαρμόζοντας αυτήν την ιδιότητα στην (5) με βάση το Θεώρημα 4, από όπου $\boldsymbol{\mu} = \mathbb{E}\hat{\beta}, \mathbb{V}(\hat{\beta}) = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ το 1 ακολουθεί αμέσως.

Κατανομή των παραμέτρων του μοντέλου.

Συνέχεια της απόδειξης.

Για το 2, πρώτα γράφουμε

$$(n-p)\hat{\sigma}^2 = \mathbf{Y}^{\mathbf{T}}(\mathbf{I_n}-\mathbf{P})\mathbf{Y} = (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{I_n}-\mathbf{P})(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})$$
$$= \boldsymbol{\varepsilon}^{\boldsymbol{T}}(\mathbf{I_n}-\mathbf{P})\boldsymbol{\varepsilon}. \quad (6)$$

Ξέρουμε ότι

> Όταν $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I_n})$ και $\mathbf{K}, n\times n$ συμμετρικός πίνακας με $\mathrm{rank}(\mathbf{K}) = r$ τότε $\sigma^{-2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{K}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi^2_r$ ανν $\mathbf{K^2} = \mathbf{K}$ $(*)$

- Η (6) θα προκύψει από την $(*)$ θέτοντας $\boldsymbol{\varepsilon} = \mathbf{x}-\boldsymbol{\mu}$ και $\mathbf{K} = \mathbf{I_n}-\mathbf{P}$.
  1. Ο $\mathbf{I_n}-\mathbf{P}$ είναι $n\times n$ συμμετρικός πίνακας.
  2. Ο $\mathbf{I_n}-\mathbf{P}$ είναι ταυτοδύναμος οπότε $\mathrm{rank}(\mathbf{I_n}-\mathbf{P}) = n-p$.
  3. Επίσης ικανοποιεί και την $(I_n-\mathbf{P})^2 = \mathbf{I_n}-\mathbf{P}$.
  4. Από την υπόθεση του θεωρήματος, $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2\mathbf{I_n})$
- Χρησιμοποιώντας τις $1-4$ στην $(*)$ βλέπουμε ότι η (6) γίνεται

$$\boldsymbol{\varepsilon}^{\boldsymbol{T}}(\mathbf{I_n}-\mathbf{P})\boldsymbol{\varepsilon} = \sigma^{-2}(n-p)\hat{\sigma}^2 \sim \chi^2_{n-p} \quad (7)$$

που είναι αυτό που θέλουμε.

Έλεγχοι υποθέσεων για τις παραμέτρους του μοντέλου.

Έστω $c_{ii}, i = 1, \ldots, p$ τα διαγώνια στοιχεία του $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$. Τότε,

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii}), i = 1, \ldots, p$$

οπότε και πάλι,

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{c_{i+1i+1}}} \sim t_{n-p-1}, i = 1, 2, \ldots, p$$

Έτσι, ένα δ/μα εμπιστοσύνης για το $\beta_i, i = 1, \ldots, p$ είναι το

$$\left(\hat{\beta}_i - \hat{\sigma}\sqrt{c_{i+1i+1}}t_{n-p-1,\alpha/2}, \hat{\beta}_i + \hat{\sigma}\sqrt{c_{i+1i+1}}t_{n-p-1,\alpha/2}\right) \qquad (8)$$

Η $H_0 : \beta_i = 0$ έναντι της $H_0 : \beta_i \neq 0$ θα απορρίπτεται όταν

$$\left|\frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{c_{i+1i+1}}}\right| > t_{n-p-1,\alpha/2}, i = 1, 2, \ldots, p.$$

Η απόρριψη της $H_0 : \beta_i = 0$ για κάποιο i συνεπάγεται ότι η μεταβλητή Y εξαρτάται από την X_i.

Ερμηνεύοντας τη συνολική μεταβλητότητα του μοντέλου.

- Όπως ακριβώς και στο απλό γραμμικό μοντέλο, η δειγματική διασπορά των παρατηρήσεων Y_i αποδεικνύεται ότι χωρίζεται σε δύο αθροίσματα:

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}}$$

- Το SST εκφράζει τη συνολική παρατηρούμενη μεταβλητότητα των Y_i, το SSR εκφράζει τη μεταβλητότητα των προσαρμοσμένων τιμών ενώ το SSE εκφράζει τη μεταβλητότητα των Y_i σε σχέση με τις αντίστοιχες προσαρμοσμένες τιμές.
- Η μεταβλητότητα του SSR ερμηνεύεται από το μοντέλο ενώ του SSE όχι.
- Το πηλίκο (συντελεστής προσδιορισμού)

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}}$$

είναι το ποσοστό της μεταβλητότητας των Y_i που ερμηνεύεται από το μοντέλο.

- Από την (7) έχουμε ότι

Ερμηνεύοντας τη συνολική μεταβλητότητα του μοντέλου.

$$\frac{\text{SSE}}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n} \underbrace{(Y_i - \hat{Y}_i)^2}_{=\varepsilon_i^2 \sim \{N(0,\sigma^2)\}^2 \sim \chi^2_1} \sim \chi^2_{n-p}.$$

- Επίσης, αν $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ τότε

$$\frac{\text{SSR}}{\sigma^2} \sim \chi^2_{p-1}, \quad \frac{\text{SST}}{\sigma^2} \sim \chi^2_n.$$

Σε αυτήν την περίπτωση,

$$\frac{\frac{\text{SSR}}{\sigma^2}/(p-1)}{\frac{\text{SSE}}{\sigma^2}/(n-p)} = \frac{\text{SSR}/(p-1)}{\text{SSE}/(n-p)} \sim F_{p-1,n-p}. \tag{9}$$

γιατί οι SSR και SSE είναι ανεξάρτητες. Η (9) μπορεί να χρησιμοποιηθεί για τον έλεγχο της $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ δηλαδή ότι η Y δεν εξαρτάται από καμία από τις $X_1, X_2, \ldots, X_p$. Το στατιστικό αυτό προκύπτει γιατί υπό την H_0: $\hat{Y}_i = \bar{Y}$ και $Y_i = \bar{Y}$

- Συγκεκριμένα η H_0 θα απορρίπτεται όταν

$$F = \frac{\text{SSR}/(p-1)}{\text{SSE}/(n-p)} > F_{p-1,n-p,\alpha}$$

Ατομική και μέση πρόβλεψη της Y.

- Είναι σημαντικό να διαχωρίσουμε εδώ τους εξής ελέγχους:
  1. Ταυτόχρονος έλεγχος αν όλες οι παράμετροι του μοντέλου είναι ίσες με το 0 (F test)
  2. Έλεγχος αν κάποιο β_i (μόνο ένα) είναι ίσο με το 0 (t test)
- Είναι φανερό ότι η κατανομή του στατιστικού είναι διαφορετική στις περιπτώσεις 1 και 2.

- Από την (9) προκύπτει ένα δ.ε. για το διάνυσμα $\hat{\boldsymbol{\beta}}$:

$$(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\mathbf{X}^\mathsf{T}\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \leq (p+1)\frac{\mathrm{SSR}/(p-1)}{\mathrm{SSE}/(n-p)}F_{p-1,n-p,\alpha}$$

- και πάλι, η διαφορά αυτού του διαστήματος σε σχέση με τα διαστήματα για τα μεμονωμένα β_i είναι ότι εδώ έχουμε συνολική (ταυτόχρονη) εκτίμηση όλων των τιμών του διανύσματος των παραμέτρων.
- Έχοντας εκτιμήσει τους συντελεστές $\beta_1, \beta_2, \ldots, \beta_p$ μπορούμε να κάνουμε πρόβλεψη της μεταβλητής Y για οποιεσδήποτε τιμές των μεταβλητών $X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p$

Ατομική και μέση πρόβλεψη της Y.

- Η πρόβλεψη δίνεται από την

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \tag{10}$$

- Όπως και στο απλό γραμμικό μοντέλο μπορούμε να προβλέψουμε την $\mathbb{E}Y$ και μέσω ενός διαστήματος.
- Έστω $X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p$. Ένα δ.ε. συντ. $1-\alpha$ για την εκτίμηση της μέσης τιμής $\mathbb{E}(Y)$ είναι:

$$\left[\mathbf{x}^\mathsf{T}\hat{\boldsymbol{\beta}} - \hat{\sigma}\sqrt{\mathbf{x}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}^\mathsf{T}}t_{n-p-1,\alpha/2}, \mathbf{x}^\mathsf{T}\hat{\boldsymbol{\beta}} + \hat{\sigma}\sqrt{\mathbf{x}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}^\mathsf{T}}t_{n-p-1,\alpha/2}\right]$$

- Αν πάρουμε έναν μεγάλο αριθμό παρατηρήσεων με $X_1 = x_1, \ldots, X_p = x_p$ τότε η μέση τιμή της Y θα βρίσκεται μέσα στο διάστημα μέσης πρόβλεψης με σ.ε. $1-\alpha$.
- Το διάστημα ατομικής πρόβλεψης του $\hat{Y}$ η οποία προκύπτει από αντικατάσταση στη (10) των $X_1 = x_1, \ldots, X_p = x_p$, είναι το διάστημα με σ.ε. $1-\alpha$:

$$\left[\mathbf{x}^\mathsf{T}\hat{\boldsymbol{\beta}} \pm \hat{\sigma}\sqrt{1 + \mathbf{x}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}^\mathsf{T}}t_{n-p-1,\frac{\alpha}{2}}\right] \tag{11}$$

Εξέταση της ορθότητας του μοντέλου.

- Όπως και στο απλό γραμμικό μοντέλο θα πρέπει να βεβαιωθούμε ότι οι παρατηρήσεις μας προσαρμόζονται ικανοποιητικά στο πολλαπλό μοντέλο ώστε τα συμπεράσματα που προκύπτουν να θεωρούνται αξιόπιστα.
  1. Ένας πρώτος έλεγχος για την ορθότητα του μοντέλου είναι το F test με το οποίο ελέγχουμε την καλή προσαρμογή του μοντέλου.
  2. Μια δεύτερη ένδειξη είναι η τιμή του συντελεστή R^2 - όσο μεγαλύτερο τόσο μεγαλύτερο ποσοστό της μεταβλητότητας της Y εξηγείται από τις επεξηγηματικές μεταβλητές.
  3. Επίσης είναι η τιμή του $\hat{\sigma}$ - όσο μικρότερο, τόσο το καλύτερο.
- Ο έλεγχος της ορθότητας του πολλαπλού μοντέλου γίνεται και με τη βοήθεια των τυποποιημένων υπόλοιπων,

$$\hat{\varepsilon}_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-p_{ii}}}, i = 1, \ldots, n \tag{12}$$

όπου τα p_{ii} είναι τα διαγώνια στοιχεία του πίνακα $\mathbf{P}$.

Εφόσον θεωρούμε ότι το μοντέλο είναι σωστό, μπορούμε να προχωρήσουμε σε συμπεράσματα: Στο πολλαπλό μοντέλο, κάθε εκτιμητής β_i εκφράζει την αυξηση στην Y_i όταν το αντίστοιχο X_i αυξάνεται κατά 1 δεδομένου ότι οι υπόλοιπες μεταβλητές μένουν σταθερές.

Σύγκριση μοντέλων.

- Αρκετές φορές συμβαίνει να έχουμε καταγράψει τις τιμές αρκετών ανεξάρτητων μεταβλητών $X_1, X_2, \ldots, X_p$ και θέλουμε να εξετάσουμε ποιες από αυτές επηρεάζουν την μεταβλητή απόκρισης Y.
- Ένας τρόπος (που δεν είναι πάντα ο συντομότερος) είναι να εφαρμόσουμε όλα τα δυνατά μοντέλα και να επιλέξουμε αυτό που δίνει τα «καλύτερα» αποτελέσματα. Αν π.χ. έχουμε καταγράψει τρεις μεταβλητές X_1, X_2, X_3 τότε μπορούμε να εξετάσουμε τα μοντέλα:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon, Y = \beta_0 + \beta_2 X_2 + \varepsilon, Y = \beta_0 + \beta_3 X_3 + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon,$$
$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

- και να θεωρήσουμε καλύτερο αυτό που δίνει το μεγαλύτερο συντελεστή προσδιορισμού R^2.
- Γρήγορα όμως διαπιστώνουμε ότι το μοντέλο με το μεγαλύτερο $R^2 = 1 - \text{SSE}/\text{SST}$ δεν είναι πάντοτε το καλύτερο.
- Αυτό συμβαίνει διότι όσο προσθέτουμε ανεξάρτητες μεταβλητές στο μοντέλο (όποιες και αν είναι αυτές) το R^2 αυξάνεται (ή παραμένει σταθερό).

Σύγκριση μοντέλων.

- Πράγματι, όταν προσθέτουμε ανεξάρτητες μεταβλητές το SSE μειώνεται ή μένει σταθερό αφού

$$\text{SSE} = \min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})^2$$

- (ενώ το SST παραμένει πάντοτε σταθερό) και η παραπάνω ελαχιστοποίηση γίνεται σε μεγαλύτερο χώρο (περισσότερα β_i).
- Π.χ. έστω ότι η Y επηρεάζεται από τις X_1, X_2 ενώ δεν επηρεάζεται καθόλου από την X_3 οπότε καλύτερο μοντέλο θα έπρεπε να είναι το $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.
- Προσθέτοντας όμως την X_3 σε αυτό το μοντέλο προκύπτει το μοντέλο $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, το οποίο, σύμφωνα με τα παραπάνω, θα έχει μεγαλύτερο ή ίσο R^2 από το $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.
- Επομένως το R^2 δεν δείχνει πάντοτε το «καλύτερο» μοντέλο.
- Αντί του R^2 προτείνεται η χρήση του «προσαρμοσμένου» R^2 (R^2 adjusted).

Σύγκριση μοντέλων.

- Έτσι, καλύτερο θα θεωρείται το μοντέλο με το μεγαλύτερο

$$R^2_{adj} = 1 - \frac{\mathrm{SSE}(n-p-1)}{\mathrm{SST}(n-1)} \tag{13}$$

- Το προσαρμοσμένο R^2 «δείχνει» ως καλύτερο το μοντέλο που έχει το μικρότερο $S^2 = \mathrm{SSE}/(n-p)$ (το SST είναι σταθερό σε όλα τα μοντέλα).
- Ο δείκτης αυτός δεν αυξάνεται πάντοτε όταν αυξάνονται οι ανεξάρτητες μεταβλητές.
- Αξίζει να αναφερθεί ότι για την εύρεση του καλύτερου μοντέλου έχουν προταθεί και άλλοι δείκτες (με διάφορες αιτιολογήσεις).

Πολυσυγγραμμικότητα.

- Στην πολλαπλή παλινδρόμηση είναι δυνατό κάποιες από τις ανεξάρτητες μεταβλητές $X_1, X_2, \ldots, X_p$ να είναι γραμμικά εξαρτημένες με συνέπεια ο πίνακας πληροφορίας $\mathbf{X}^T\mathbf{X}$ να μην αντιστρέφεται (η ορίζουσα του είναι 0).
- Αυτό το πρόβλημα είναι γνωστό ως πρόβλημα πολυσυγγραμικότητας.
- Ένας απλός τρόπος αντιμετώπισης του είναι η αφαίρεση κάποιων ανεξάρτητων μεταβλητών από το μοντέλο (χάνοντας όμως «πληροφορία»).
- Ακόμη και όταν η ορίζουσα του $\mathbf{X}^T\mathbf{X}$ δεν είναι ακριβώς 0 αλλά «κοντά» στο 0 (ασθενής πολυσυγγραμικότητα) παρουσιάζεται πρόβλημα.
- Σε αυτή την περίπτωση μπορεί να εμφανιστούν σφάλματα στρογγύλευσης κατά την αντιστροφή του $\mathbf{X}^T\mathbf{X}$ (με συνέπεια οι εκτιμήσεις που παίρνουμε να μην είναι αξιόπιστες).
- Αυτό συνήθως αντιμετωπίζεται ως ένα βαθμό με τυποποίηση όλων των μεταβλητών (π.χ. ώστε να παίρνουν τιμές στο $(-1, 1)$) και κατά την εμφάνιση των αποτελεσμάτων τις επαναφέρει στην αρχική κλίμακα.

Πολυσυγγραμμικότητα.

- Η ορίζουσα του $\mathbf{X^TX}$ είναι κοντά στο 0 όταν υπάρχει ισχυρή συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών.
- Μία ακόμη «παρενέργεια» της συγκεκριμένης κατάστασης είναι ότι μπορεί ορισμένες μεταβλητές να φαίνονται «σημαντικές» (με αντίστοιχο p-value στο t-τεστ κοντά στο 0) σε κάποιο μοντέλο, ενώ παύουν να είναι σημαντικές όταν στο μοντέλο προσθέσουμε και άλλες ανεξάρτητες μεταβλητές.
- Για παράδειγμα μπορεί στο μοντέλο $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, η X_2 να είναι σημαντική (απορρίπτουμε ότι $\beta_2 \neq 0$) ενώ στο μεγαλύτερο μοντέλο $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, να μην είναι πια σημαντική (ενώ η X_3 που προσθέσαμε να είναι).
- Χοντρικά, αυτό μπορεί να συμβαίνει διότι η X_3 είναι αυτή που επηρεάζει την Y αλλά όταν εφαρμόζουμε το πρώτο μοντέλο (στο οποίο απουσιάζει η X_3), η X_2 φαίνεται σημαντική διότι «μοιάζει» στην X_3.
- Όταν εντοπίσουμε ομάδα ή ομάδες από ισχυρά συσχετισμένες μεταβλητές (π.χ. από τον πίνακα συσχετίσεων των X_i) τότε θα πρέπει στο βέλτιστο μοντέλο να κρατήσουμε μία από κάθε ομάδα.
- Σε αρκετές περιπτώσεις αυτό δεν είναι εύκολο οπότε είναι ανάγκη να χρησιμοποιήσουμε άλλες μεθόδους.

Πρακτική άσκηση 1.

Πρακτική άσκηση 1

Η γεύση ενός τυριού εξαρτάται από τη χημική του σύνθεση η οποία μπορεί να αλλάζει με την πάροδο του χρόνου. Για να διερευνηθούν ποιες χημικές ουσίες καθορίζουν σε μεγάλο βαθμό την γεύση ενός τυριού Cheddar ελήφθη δείγμα $n = 30$ τέτοιων τυριών (σε διάφορα στάδια ωρίμανσης).
Σε καθένα από αυτά έγινε εκτίμηση της γεύσης (από ένα σύνολο δοκιμαστών) και μετρήθηκε η περιεκτικότητα σε τρεις χημικές ουσίες:

1. Acetic: Λογάριθμος της περιεκτικότητας σε οξεικό οξύ (acetic acid)
2. H2S: Λογάριθμος της περιεκτικότητας σε υδρόθειο (hydrogen sulfide)
3. Lactic: περιεκτικότητα σε γαλακτικό οξύ (lactic acid)

Στην συγκεκριμένη περίπτωση επιθυμούμε να διερευνήσουμε την επιρροή της (εξαρτημένης) μεταβλητής taste (Y) από τις (ανεξάρτητες) μεταβλητές Acetic (X_1), H2S (X_2) και Lactic (X_3).

1. Να εφαρμόσετε το μοντέλο $Y = \beta_0 + \beta_1 X_1 + \varepsilon$. Από το μοντέλο αυτό, μπορούμε να πούμε ότι η περιεκτικότητα σε οξεικό οξύ επηρεάζει τη γεύση του τυριού;

Πρακτική άσκηση 1.

- Καταρχήν φορτώνουμε τα δεδομένα και εφαρμόζουμε το ζητούμενο μοντέλο:

```
data.use<-read.csv("data.csv", header=T)
lm.1<-lm(taste~Acetic, data=data.use)
summary.lm(lm.1)
```

```
Call:
lm(formula = taste ~ Acetic, data = data.use)

Residuals:
    Min      1Q  Median      3Q     Max
-29.642  -7.443   2.082   6.597  26.581

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -61.499     24.846  -2.475  0.01964 *
Acetic        15.648      4.496   3.481  0.00166 **
---
Signif. codes:  0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ ” 1

Residual standard error: 13.82 on 28 degrees of freedom
Multiple R-squared:  0.302,        Adjusted R-squared:  0.2771
F-statistic: 12.11 on 1 and 28 DF,  p-value: 0.001658
```

Το πρώτο που κοιτάμε είναι αν ο εκτιμώμενος συντελεσής της μεταβλητής είναι στατιστικά σημαντικός.

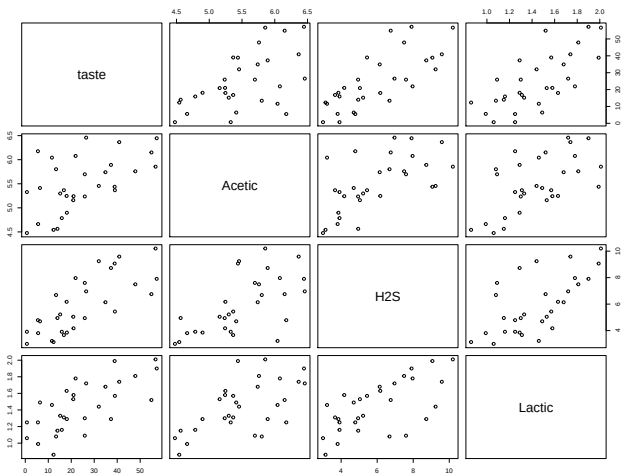
- Το p-value της μεταβλητής είναι 0.00166 άρα η $H_0 : \beta_1 = 0$ έναντι της $H_1 : \beta_1 \neq 0$ απορρίπτεται σε επίπεδο σημαντικότητας 1%
- Αυτό που μας δείχνει επίσης το μοντέλο είναι ότι η μεταβλητή από μόνη της εξηγεί περίπου το 27% της μεταβλητότητας της γεύσης.

Πρακτική άσκηση 1: συνέχεια.

2. Να κατασκευάσετε όλα τα (ανά δύο) γραφήματα διασποράς των μεταβλητών Y, X_1, X_2, X_3. Υπάρχει σχέση μεταξύ τους;
3. Να εφαρμόσετε το μοντέλο $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$.
4. Να εκτιμήσετε τα β_i σημειακά και με δ.ε. συντελεστού 95%.
5. Ποιες ανεξάρτητες μεταβλητές επηρεάζουν την γεύση ενός τυριού;
6. Ελέγξτε την $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ μέσω του F-τεστ.

- Μπορούμε να κάνουμε το ζητούμενο γράφημα με μόλις μία εντολή

```
pairs(~taste+Acetic+H2S+Lactic, data=data.use)
```



Παρατηρούμε ότι (ανά δύο) όλες οι μεταβλητές παρουσιάζουν κάποια θετική συσχέτιση μεταξύ τους. Ιδιαίτερα η Taste φαίνεται να εξαρτάται και από τις τρεις μεταβλητές (3 σχήματα 1ης γραμμής).

- Για την ερώτηση 3 εφαρμόζουμε την

```
lm.al<-lm(taste~Acetic+H2S+Lactic, data=data.use)
```

Πρακτική άσκηση 1: συνέχεια.

- Για την ερ. 4, καταρχήν οι σημειακές εκτιμήσεις δίνονται από την

```
summary.lm(lm.al)

Call:
lm(formula = taste ~ Acetic + H2S + Lactic, data = data.use)

Residuals:
    Min      1Q  Median      3Q     Max
-17.390  -6.612  -1.009   4.908  25.449

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
Acetic        0.3277     4.4598   0.073  0.94198
H2S           3.9118     1.2484   3.133  0.00425 **
Lactic       19.6705     8.6291   2.280  0.03108 *
---
Signif. codes:  0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared:  0.6518,        Adjusted R-squared:  0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

Από τα p-value των παραμέτρων βλέπουμε ότι η Acetic δεν είναι σημαντική μεταβλητή, ενώ σε επίπεδο σημαντικότητας 5% είναι τόσο η H2S όσο και η Lactic

- Ισοδύναμα, αν θέλουμε να απομονώσουμε μόνο τις τιμές των συντελεστών:

```
> lm.al$coefficients
(Intercept)      Acetic          H2S      Lactic
-28.8767696   0.3277413   3.9118411  19.6705434
```

Πρακτική άσκηση 1: συνέχεια.

- Για τα ζητούμενα διαστήματα εμπιστοσύνης των εκτιμητών η (8) υπολογίζεται αυτόματα από την

```
confint(lm.al)
```

```
                2.5 %    97.5 %
(Intercept) -69.443503 11.689964
Acetic       -8.839420  9.494902
H2S           1.345656  6.478026
Lactic        1.933267 37.407820
```

Το δ/μα της Acetic περιλαμβάνει το 0, άρα δεν μπορούμε να αποκλείσουμε την περίπτωση η παράμετρος αυτής της μεταβλητής να είναι 0.

- Μπορούμε να προσδιορίσουμε διαφορετικό επίπεδο εμπιστοσύνης για το ζητούμενο διάστημα ως εξής:

```
confint(lm.al, level=0.90); confint(lm.al, level=0.99)
```

- Για την ερώτηση 5, σύμφωνα με το αποτέλεσμα της παλινδρόμησης μόνο οι H2S και Lactic φαίνεται να επηρεάζουν την γεύση του τυριού.
- Για την ερώτηση 6, κοιτάμε την τελευταία γραμμή από όπου φαίνεται ότι για τον έλεγχο της $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ το p-value είναι πολύ κοντά στο 0 οπότε απορρίπτουμε την υπόθεση οτι η γευση δεν εξαρτάται από κανέναν από τους τρεις παράγοντες.

Πρακτική άσκηση 1: συνέχεια.

7. Ποια είναι η εκτίμηση της διασποράς των σφαλμάτων;
8. Τι % της μεταβλητότητας των Y_i ερμηνεύεται από το μοντέλο;

- Η ερώτηση 7 βασικά ζητάει εκτίμηση του $\mathrm{SSE}/(n-p)$ το οποίο υπολογίζεται από την

```
anova(lm.al)
 Analysis of Variance Table

Response: taste
          Df  Sum Sq Mean Sq F value    Pr(>F)
Acetic     1 2314.14 2314.14 22.5481 6.528e-05 ***
H2S        1 2147.02 2147.02 20.9197 0.0001035 ***
Lactic     1  533.32  533.32  5.1964 0.0310795 *
Residuals 26 2668.41  102.63
---
Signif. codes:  0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ ” 1
```

$$\mathrm{SSE}/(n-p) = 102.63$$

- Για την ερώτηση 8, οι R^2 και R^2_{adj} (ορισμός στη (13)) δίνονται από τη γραμμή

```
Multiple R-squared:  0.6518,        Adjusted R-squared:  0.6116
```

- Στην πράξη χρησιμοποιούμε το R^2_{adj} άρα το μοντέλο ερμηνεύει περίπου το 61% της μεταβλητότητας της Y.

Πρακτική άσκηση 1: συνέχεια.

9. Ποιο από τα δύο μοντέλα θεωρείτε «καλύτερο» $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \varepsilon$, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ Παρατηρείτε κάτι αντιφατικό σε σχέση με το ερώτημα 1;
10. Εφαρμόστε το μοντέλο $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \varepsilon$: εκτιμήσετε τα β_i σημειακά και με δ.ε. συντελεστού 95%. Πώς επηρεάζεται η γεύση όταν μεταβάλλονται το Γαλακτικό οξύ και το υδρόθειο;
11. Ποιας μεταβλητής η μεταβολή επηρεάζει περισσότερο την μεταβολή της γεύσης;

- Το μοντέλο που εφαρμόσαμε στο ερώτημα 1 βρήκαμε την μεταβλητή Acetic ήταν σημαντική.
- Στο πλήρες μοντέλο όμως βρήκαμε το αντίθετο αποτέλεσμα ότι δηλαδή δεν είναι σημαντική.
- Όταν έχουμε τέτοια αντίθεση, κοιτάμε αν υπάρχει πολυσυγραμμικότητα μεταξύ των επεξηγηματικών μεταβλητών
- Στο συγκεκριμένο παράδειγμα υπάρχει υψηλή συσχέτιση μεταξύ της Acetic και των H2S και Lactic(πως θα το διαπιστώσετε);
- Αυτό σημαίνει ότι ένα πιο αποδοτικό μοντέλο περιμένουμε να προκύψει αν βγάλουμε την Acetic.

Πρακτική άσκηση 1: συνέχεια.

```
lm.a2<-lm(taste~ H2S+Lactic, data=data.use)
summary.lm(lm.a2)
```

```
Call:
lm(formula = taste ~ H2S + Lactic, data = data.use)

Residuals:
    Min      1Q  Median      3Q     Max
-17.343  -6.530  -1.164   4.844  25.618

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -27.592      8.982  -3.072  0.00481 **
H2S            3.946      1.136   3.475  0.00174 **
Lactic        19.887      7.959   2.499  0.01885 *
---
Signif. codes:  0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ ” 1

Residual standard error: 9.942 on 27 degrees of freedom
Multiple R-squared:  0.6517,        Adjusted R-squared:  0.6259
F-statistic: 25.26 on 2 and 27 DF,  p-value: 6.551e-07
```

- Βλέπουμε ότι τώρα που βγάλαμε τη μεταβλητή Acetic οι δείκτες του μοντέλου βελτιώθηκαν:
  - Καλύτερο Adjusted R-squared
  - Καλύτερο F test (p-value πιο κοντά στο 0)
- Οπότε όλες οι ενδείξεις δείχνουν ότι το $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \varepsilon$ είναι καλύτερο από το $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$.

Πρακτική άσκηση 1: συνέχεια.

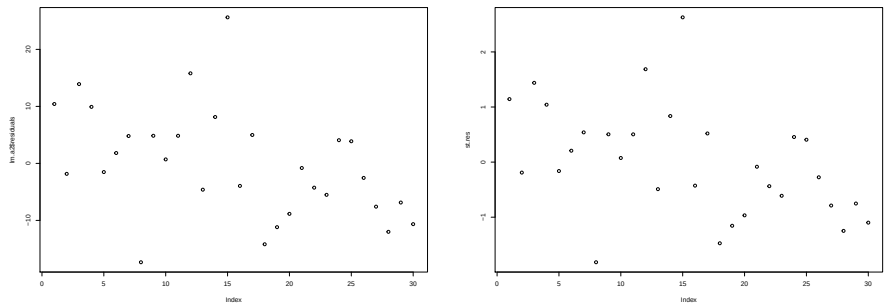
- Για να απαντήσουμε στην ερώτηση: Πώς επηρεάζεται η γεύση όταν μεταβάλλονται το Γαλακτικό οξύ και το υδρόθειο;΄
- Από τα αποτελέσματα του πίνακα της προηγούμενης σελίδας:
  - Παρατηρούμε ότι όσο αυξάνεται η περιεκτικότητα σε υδρόθειο ή Γαλακτικό οξύ, βελτιώνεται και η γεύση του τυριού (τα αντίστοιχα β_i είναι θετικά, άρα μοναδιαία αύξηση στην τιμή κάθε μεταβλητής συνεισφέρει θετικά στη γεύση).
  - Επίσης βλέπουμε ότι η αύξηση της μεταβλητής H2S κατά μία μονάδα αυξάνει την μεταβλητή taste κατά 3.9 μονάδες περίπου (δεδομένου ότι η Lactic παραμένει σταθερή) ενώ αντίστοιχα η αύξηση της μεταβλητής Lactic κατά μία μονάδα αυξάνει την μεταβλητή taste κατά 19.9 μονάδες περίπου (δεδομένου ότι η H2S παραμένει σταθερή).
  - Επομένως η μεταβολή της περιεκτικότητας σε γαλακτικό οξύ επηρεάζει περισσότερο την μεταβολή της γεύσης (Ερώτημα 11).
- Προσοχή, τα παραπάνω συμβαίνουν τουλάχιστον μέσα στο εύρος των τιμών που παίρνουν οι συγκεκριμένες μεταβλητές στο δείγμα (η H2S παίρνει τιμές μεταξύ του 3 και του 10.2 ενώ η Lactic μεταξύ του 0.86 έως του 2.01).
- Έξω από αυτά τα όρια μπορεί το μοντέλο να είναι διαφορετικό.

Πρακτική άσκηση 1: συνέχεια.

12. Να δοθούν οι προσαρμοσμένες τιμές των Y_i (προβλέψεις των Y_i) και τα τυποποιημένα υπόλοιπα.

- Ερώτηση 12: Τα τυποποιημένα υπόλοιπα (12) όπως και τα κανονικά υπόλοιπα, μαζί με τις αντίστοιχες γραφικές παραστάσεις μπορούν να παραχθούν με τις εντολές

```
st.res<-rstandard(lm.a2)
plot(st.res)
plot(lm.a2$residuals)
plot(st.res)
```



- Και οι 2 εκδόσεις των υπολοίπων δείχνουν την ίδια εικόνα (το περιμέναμε;
- Τα υπόλοιπα είναι τυχαία κατανεμημένα στο χώρο και διάσπαρτα γύρω από τον οριζόντιο άξονα - αυτό συνηγορεί στην αποδοχή του μοντέλου.
- Μπορούμε να πάρυμε τις προσαρμοσμένες τιμές των Y_i με την `lm.a2$fitted.values`

Πρακτική άσκηση 1: συνέχεια.

13. Εάν ένα τυρί έχει περιεκτικότητα 3 και 2 σε (log) υδρόθειο και Γαλακτικό οξύ αντίστοιχα, δώστε μια πρόβλεψη του δείκτη της γεύσης του (δώστε σημειακή πρόβλεψη και κατάλληλο διάστημα πρόβλεψης).

- Η ερώτηση είναι εφαρμογή της (10) για ένα ζεύγος καινούργιων παρατηρήσεων.
- Η εισαγωγή των καινούργιων παρατηρήσεων, ο υπολογισμός της $\hat{Y}_i$ και το αντίστοιχο δ/μα εμπιστοσύνης (11) γίνεται με τις εντολές

```
new <- data.frame(H2S=3, Lactic=2)
predict.lm(lm.a2, new, interval = "
    confidence", level=0.95)
```

```
       fit      lwr      upr
1 24.0214 9.039467 39.00332
```

- Παρατηρούμε ότι το δ.ε ουσιαστικά δεν προσφέρει κάποια πληροφορία διότι είναι πολύ ευρύ (σε σχέση με τις τιμές που παίρνει η taste).
- Αυτό συμβαίνει διότι το συγκεκριμένο μοντέλο ερμηνεύει μόνο το 65.2% της παρατηρούμενης μεταβλητότητας της taste.
- Είναι χρήσιμο σε αυτό το σημείο να υπογραμμίσουμε ότι γενικά δεν είναι ασφαλές να ζητάμε πρόβλεψη του Y για τιμές των ανεξάρτητων μεταβλητών εκτός των ορίων των τιμών τους στα δεδομένα.

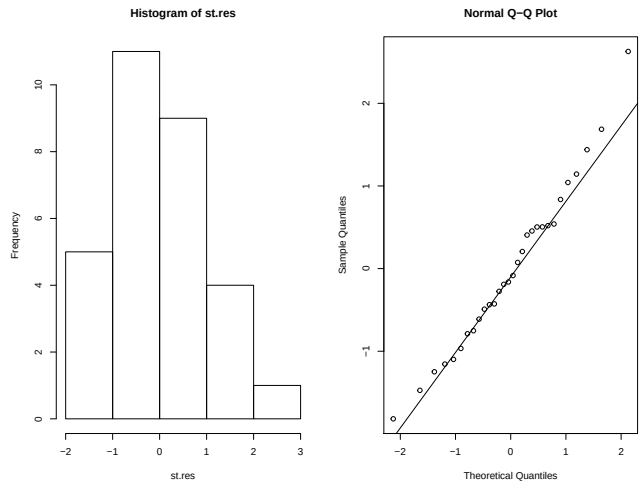
Πρακτική άσκηση 1: έλεγχος ορθότητας του μοντέλου

14. Εξετάστε αν τα τυποποιημένα κατάλοιπα προέρχονται πράγματι από κανονική κατανομή (ιστόγραμμα, Q-Q ή P-P plots και K-S τεστ).

- Οι γραφικοί έλεγχοι (ιστόγραμμα, Q-Q plot) μπορούν να παραχθούν με τις εντολές

```
hist(st.res)
qqnorm(st.res)
qqline(st.res)
```

Οι γραφικές παραστάσεις φαίνεται να συνιστούν κανονική κατανομή υπολοίπων.




```
> ks.test(st.res, "pnorm")

  One-sample Kolmogorov-Smirnov test

data:  st.res
D = 0.094688, p-value = 0.9275
alternative hypothesis: two-sided
```

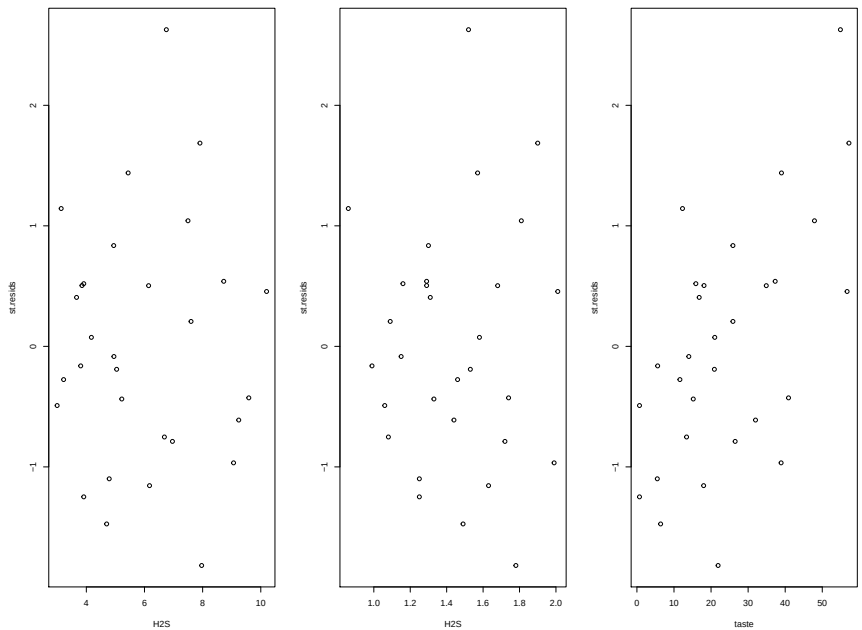
Με βάση ότι p-value = 0.9275 δεν μπορούμε να απορρίψουμε ότι τα τυποποιημένα υπόλοιπα προέρχονται από την κανονική κατανομή.

Πρακτική άσκηση 1: έλεγχος ορθότητας του μοντέλου

15. Εξετάστε αν υπάρχει σχέση μεταξύ των τυποποιημένων υπολοίπων και των μεταβλητών X_2, X_3, Y.

- Οι γραφικές παραστάσεις των τριών μεταβλητών με τα υπόλοιπα είναι

```
par(mfrow=c(1,3))
plot(data.use[,4], st.res,
     xlab="H2S", ylab="st.
     resids")
plot(data.use[,5], st.res,
     xlab="H2S", ylab="st.
     resids")
plot(data.use[,2], st.res,
     xlab="taste", ylab="st.
     resids")
```



- Οι παρατηρήσεις φαίνεται ότι βρίσκονται τυχαία στο επίπεδο και στα τρία γραφήματα και επομένως δεν πρέπει να υπάρχει κάποια σχέση μεταξύ των μεταβλητών αυτών και των υπολοίπων.