SY19 – Machine Learning
Chapter 2: Linear Regression

Thierry Denœux

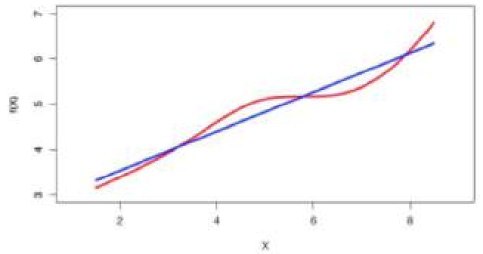
Université de technologie de Compiègne

https://www.hds.utc.fr/~tdenoeux
email: tdenoeux@utc.fr

Automne 2024

Linear regression

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \ldots, X_p$ is linear.
- This is is only an approximation of reality.



- Although it may seem overly simplistic, linear regression is very useful both conceptually and practically.

"Essentially, all models are wrong, but some are useful" *(George E. P. Box)*

Movie Box Office data

- Data about 62 movies released in 2009 (from *Econometric Analysis*, Greene, 2012)
- Response: Box Office receipts
- Predictors:
  - MPAA (Motion Picture Association of America) rating (G, PG, PG13,R)
  - Budget
  - Star power
  - Sequel (yes or no)
  - Genre (action, comedy, animated, horror)
  - Internet buzz

Questions we might ask

- Is there a relationship between the budget of a movie and its commercial success?
- How strong is the relationship between internet buzz and the commercial success of a movie?
- Which factors influence the commercial success of a movie?
- Can we predict the box-office success before the movie has been released?

Overview

The model

- We have an vector $X = (X_1, \ldots, X_p)^T$ of predictors and we want to predict a real-valued response Y. The linear regression model has the form
$$Y = \underbrace{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}_{f(X)=\mathbb{E}(Y|X)} + \epsilon,$$
with $\mathbb{E}(\epsilon) = 0$.
- The linear model either assumes that the regression function $f(X)$ is linear, or that the linear model is a reasonable approximation.
- The β_j's are unknown parameters or coefficients.

Choice of the predictors

- The predictor variables X_j can come from different sources:
  1. Quantitative inputs
  2. Transformations of quantitative inputs, such as power, log, square-root or square
  3. Interactions between variables, for example, $X_3 = X_1 \cdot X_2$. This allows us to model synergy (interaction) between variables
  4. Dummy coding of the levels of qualitative inputs (see next slide).
- In cases 2-4, the relationship between Y and the inputs is actually nonlinear. Yet, the method is still called linear regression, because $f(X)$ is linear in the coefficients β_j.

Representation of a nominal variable (factor)

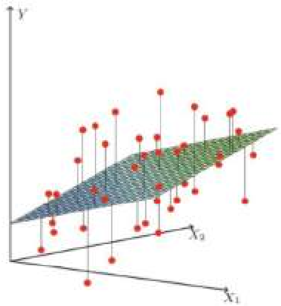
- Let G be a qualitative (nominal) variable with K levels.
- For example, let G be the genre of a movie, with four levels: action, comedy, animated, horror.
- We can encode G as 4 dummy variables:
  - $X_1 = I(G = \text{action})$
  - $X_2 = I(G = \text{comedy})$
  - $X_3 = I(G = \text{animated})$
  - $X_4 = I(G = \text{horror})$
- Since $\sum_{j=1}^{4} X_j = 1$, we need to use only 3 out of the 4 dummy variables.
- Assume we use X_1, X_2, X_3. Then the 4th level is the baseline.

Interpretation of the coefficients

- If X_j is quantitative, we interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed.
- If X_j is a dummy variable encoding a level of a qualitative predictor, β_j is the mean in increase or decrease of Y when $X_j = 1$, as compared to the baseline, holding all other predictors fixed.
- The interpretation of coefficients may be delicate when the predictors are correlated.

Overview

Method of least squares



- We have seen that the regression function minimizes the mean squared error $\text{MSE} = \mathbb{E}_{X,Y}[(Y - f(X))^2]$.
- Here, we have a training set $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^n$.
- To estimate f, we will find the coefficients β minimizing the empirical mean squared error $\widehat{\text{MSE}} = \text{RSS}/n$, where RSS is the residual sum of squares (RSS) defined as

$$\text{RSS}(\beta) = \sum_{i=1}^n \underbrace{(y_i - f(x_i))}_{\text{residuals}}^2 = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2$$

Matrix notation

- Denote by $\mathbf{X}$ the $n \times (p+1)$ design matrix with each row an input vector (with a 1 in the first position). Similarly let $\mathbf{y}$ be the n-vector of outputs in the training set:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}$$

- Let $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ be the $(p+1)$-vector of coefficients.
- The vector of predicted values $(f(x_1), \ldots, f(x_n))^T$ can be written as $\mathbf{X}\beta$.

Reformulation of the RSS criterion

- With this notation, we can rewrite the RSS as

$$\begin{aligned}\text{RSS}(\beta) &= (\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)\\ &= \mathbf{y}^T\mathbf{y}\underbrace{-\mathbf{y}^T\mathbf{X}\beta-\beta^T\mathbf{X}^T\mathbf{y}}_{-2\beta^T\mathbf{X}^T\mathbf{y}}+\beta^T\mathbf{X}^T\mathbf{X}\beta\end{aligned}$$

- This is a quadratic function in the $p+1$ parameters. To minimize RSS(β), we need to solve the equation

$$\frac{\partial \text{RSS}}{\partial \beta} = 0,$$

where $\frac{\partial \text{RSS}}{\partial \beta} = \left(\frac{\partial \text{RSS}}{\partial \beta_0}, \ldots, \frac{\partial \text{RSS}}{\partial \beta_p}\right)^T$ is the gradient of RSS with respect to β.

Reminder

Proposition

Let $\mathbf{A}$ *be a matrix and let* β *and* γ *be vectors. We have*

$$\frac{\partial \beta^T\mathbf{A}\beta}{\partial\beta} = (\mathbf{A}+\mathbf{A}^T)\beta \tag{1a}$$

$$\frac{\partial \beta^T\mathbf{A}\gamma}{\partial\beta} = \mathbf{A}\gamma \tag{1b}$$

If $\mathbf{A}$ *is symmetric,* (1a) *becomes*

$$\frac{\partial \beta^T\mathbf{A}\beta}{\partial\beta} = 2\mathbf{A}\beta \tag{1c}$$

Least-squares estimate

- Differentiating RSS(β) with respect to β we obtain

$$\begin{aligned}\frac{\partial \text{RSS}}{\partial \beta} &= \frac{\partial}{\partial\beta}\left(\mathbf{y}^T\mathbf{y}-2\beta^T\mathbf{X}^T\mathbf{y}+\beta^T\mathbf{X}^T\mathbf{X}\beta\right) &\text{(2a)}\\ &= -2\mathbf{X}^T\mathbf{y}+2\mathbf{X}^T\mathbf{X}\beta &\text{(2b)}\end{aligned}$$

- Setting the gradient to zero, we get

$$-\mathbf{X}^T\mathbf{y}+\mathbf{X}^T\mathbf{X}\beta = 0 \tag{3}$$

- Assume that $\mathbf{X}$ has full column rank; then, $\mathbf{X}^T\mathbf{X}$ has full rank and is nonsingular. Then we get the unique solution:

$$\boxed{\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}$$

- $\widehat{\beta}$ is called the Least-Squares Estimate (LSE) of β.

Fitted values

- The fitted values at the training inputs are the estimates of $f(x_i)$. They can be computed as

$$\widehat{y}_i = \widehat{\beta}_0 + \sum_{j=1}^{p}\widehat{\beta}_j x_{ij}.$$

- Using matrix notation,

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta} = \underbrace{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{\mathbf{H}}\,\mathbf{y}$$

where $\widehat{\mathbf{y}} = (\widehat{y}_1, \ldots, \widehat{y}_n)^T$.

- Matrix $\mathbf{H}$ is sometimes called the projection matrix or the hat matrix.

Overview

1. The method of least squares
   - LS estimates
   - Analysis of variance
   - Application in R and interpretation of the coefficients
2. Theoretical analysis and statistical inference
   - Properties of the LSE
   - Additional assumptions and distribution of $\widehat{\beta}$
   - Hypothesis tests
   - Prediction

Variance decomposition formula

Proposition (Analysis of variance equation)

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{TSS} = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2}_{ESS} + \underbrace{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}_{RSS}$$

Interpretation:

- TSS: Total sum of squares, measures the variability of the y_i
- ESS: Explained sum of squares, measures the variability of the $\widehat{y}_i$ (the variability explained by the predictors)
- RSS: Residual sum of squares, measures the variability of the residuals (the variability not explained by the predictors).

Proof.

R-squared

- The fraction of the total variance explained by the regression is

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\text{TSS}-\text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}},$$

- Properties:
  - $0 \leq R^2 \leq 1$
  - $R^2 = 1$ iff RSS $= 0$, i.e. $\mathbf{y} = \widehat{\mathbf{y}}$: all the variability of the y_i's is explained by the predictors.
  - $R^2 = 0$ iff TSS $=$ RSS, i.e., the predictors play no role in explaining the variability of the y_i's.

Overview

1. The method of least squares
   - LS estimates
   - Analysis of variance
   - Application in R and interpretation of the coefficients
2. Theoretical analysis and statistical inference
   - Properties of the LSE
   - Additional assumptions and distribution of $\widehat{\beta}$
   - Hypothesis tests
   - Prediction

Application en R

```
> fit<- lm(BOX ~ ., data=movie)
> fit

Call: lm(formula = BOX ~ ., data = movie)

Coefficients:
 (Intercept)   MPRATINGPG   MPRATINGPG13   MPRATINGR     BUDGET   STARPOWR
   15.172989     0.069498      -0.273367   -0.443641   0.409218   0.006427
      SEQUEL       ACTION         COMEDY    ANIMATED     HORROR       BUZZ
    0.337876    -0.654258       0.035994   -0.826735   0.685153   0.337698
```


Example

```
> summary(fit)

Call:
lm(formula = BOX ~ ., data = movie)

Residuals:
Min 1Q Median 3Q Max
-2.22095 -0.36924 0.05168 0.41682 1.41499

⋮

Residual standard error: 0.7183 on 50 degrees of freedom
Multiple R-squared: 0.5244, Adjusted R-squared: 0.4198
F-statistic: 5.013 on 11 and 50 DF, p-value: 3.26e-05
```


Overview

Overview

Additional assumptions

- Up to now we have made minimal assumptions about the true distribution of the data.
- In order to study the sampling properties of $\widehat{\beta}$, we now make the following assumptions:
  1. Analysis will be done conditionally on the observed $\mathbf{X}$, considered as a constant matrix. (Whether the elements in $\mathbf{X}$ are fixed constants or random draws from a stochastic process will not influence the results).
  2. The observations Y_i are uncorrelated and have constant variance σ^2:

$$\text{Var}(Y_i) = \sigma^2, \forall i \quad \text{and} \quad \text{Cov}(Y_i, Y_j) = 0, \forall i \neq j,$$

which we can write as

$$\text{Var}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n,$$

where $\mathbf{Y}$ is the random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$. (The expectation of $\mathbf{Y}$ is $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\beta$).

Mean and variance of $\widehat{\beta}$

Proposition

If $\mathbf{A}$ is a constant matrix and $\mathbf{Y}$ is a random vector, then

$$\mathbb{E}(\mathbf{AY}) = \mathbf{A}\,\mathbb{E}(\mathbf{Y}) \quad \textit{and} \quad \text{Var}(\mathbf{AY}) = \mathbf{A}\,\text{Var}(\mathbf{Y})\,\mathbf{A}^T$$

Here, from $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, we get

$$\mathbb{E}(\widehat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \underbrace{\mathbb{E}(\mathbf{Y})}_{\mathbf{X}\beta} = \beta,$$

so $\widehat{\beta}$ is an unbiased estimate of β, and

$$\text{Var}(\widehat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \underbrace{\text{Var}(\mathbf{Y})}_{\sigma^2\mathbf{I}_n} \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2.$$

Variance estimation

- Typically one uses the following unbiased estimate of σ^2:

$$\widehat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{\text{RSS}}{n-p-1}$$

- The variance of $\widehat{\beta}$ can be estimated by

$$\widehat{\text{Var}(\widehat{\beta})} = (\mathbf{X}^T\mathbf{X})^{-1}\widehat{\sigma}^2.$$

Gauss-Markov theorem

Theorem

The LSE $\widehat{\beta}$ is the minimum-variance linear unbiased estimator of β

- What does that mean? Let $\theta = a^T\beta$ be any linear combination of the coefficients. For instance, expectations $f(x_0) = x_0^T\beta$ are of this form.
- The LSE of θ is

$$\widehat{\theta} = a^T\widehat{\beta} = a^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

- We have $\mathbb{E}(a^T\widehat{\beta}) = a^T\beta = \theta$, so $\widehat{\theta}$ is unbiased.
- The Gauss-Markov theorem states that if we have any other estimator $\widetilde{\beta}$ of β that is linear ($\widetilde{\beta} = \mathbf{CY}$) and unbiased ($\mathbb{E}(\widetilde{\beta}) = \beta$), then $a^T\widetilde{\beta}$ is an unbiased estimate of θ, but

$$\text{Var}(a^T\widetilde{\beta}) \geq \text{Var}(a^T\widehat{\beta}).$$

Proof.

Overview

Gaussian errors

- To draw inferences about the parameters and the model, additional assumptions are needed.
- We now assume that the deviations of Y around its expectation are Gaussian. Hence

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

  with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
- Consequently,

$$\boxed{Y \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)}$$

Reminder on the multivariate normal distribution

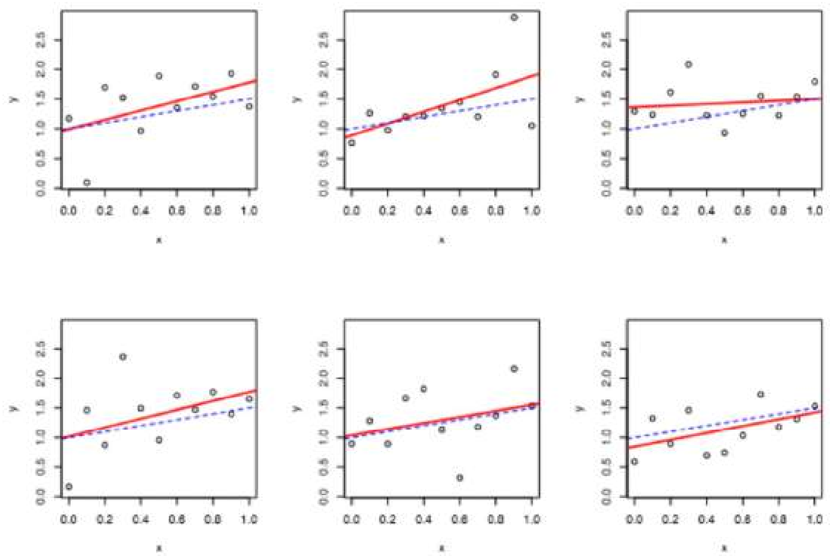
Simulation example

- Assume $p = 1$, $n = 11$, $x_i \in \{0, 0.1, 0.2, \ldots, 0.9, 1\}$, and

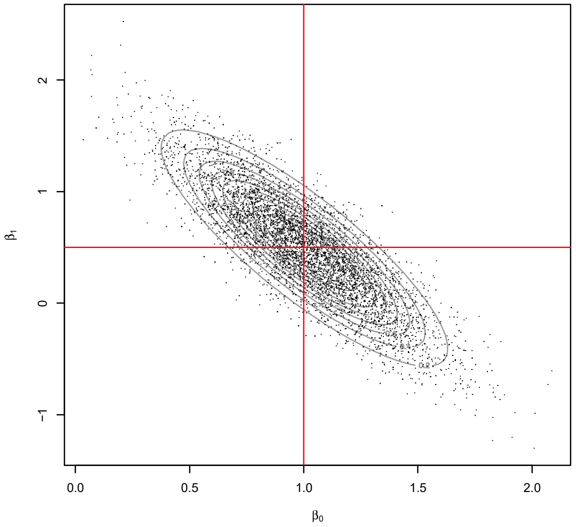
$$Y_i = 1 + 0.5x_i + \epsilon_i$$

  with $\epsilon_i \sim \mathcal{N}(0, (0.5)^2)$.
- So, $\beta_0 = 1$ and $\beta_1 = 0.5$.
- We generated $N = 5000$ datasets $(y_1, \ldots, y_n)$, for the same values of x_i.

Some datasets with the LS line



Empirical distribution of $\widehat{\beta}$



Distribution of the estimates

Proposition

If $\mathbf{Y}$ *has a normal distribution and* $\mathbf{A}$ *is a constant matrix, then* $\mathbf{AY}$ *has a normal distribution.*

- Consequently, from $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ and $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$, we can deduce that
$$\boxed{\widehat{\beta} \sim \mathcal{N}(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)}$$
- Also, we can show that
$$\frac{(n-p-1)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$$
a chi-squared distribution with $n-p-1$ degrees of freedom (df).
- In addition, $\widehat{\beta}$ and $\widehat{\sigma}^2$ are statistically independent.

Overview

Test on one coefficient

- To test the hypothesis that a particular coefficient $\beta_j = 0$, we form the standardized coefficient
$$T_j = \frac{\widehat{\beta}_j}{\widehat{\sigma}\sqrt{v_j}}$$
where v_j is the jth diagonal element of matrix $(\mathbf{X}^T\mathbf{X})^{-1}$.
- Under the null hypothesis $H_{j0} : \beta_j = 0$, T_j has a Student distribution $\mathcal{T}_{n-p-1}$ with $n-p-1$ df.
- Hence a large value of $|T_j|$ will lead to rejection of H_{j0}. Having observed the realization t_j of T_j, the p-value is
$$\begin{aligned} p &= \mathbb{P}_{H_{j0}}(|T_j| > |t_j|) \\ &= 2\left[1 - \mathbb{P}_{H_{j0}}(T_j \le |t_j|)\right] \\ &= 2\left[1 - F_{\mathcal{T}_{n-p-1}}(|t_j|)\right] \end{aligned}$$
- For $t_j = 2$, we have $p \approx 5\%$.

Example

```
> summary(fit)

:
Coefficients:
               Estimate   Std. Error  t value  Pr(>|t|)
 (Intercept)   15.172989  0.890296    17.043   < 2e-16   ***
 MPRATINGPG    0.069498   0.554641    0.125    0.9008
 MPRATINGPG13  -0.273367  0.591322    -0.462   0.6459
 MPRATINGR     -0.443641  0.595927    -0.744   0.4601
 BUDGET        0.409218   0.191454    2.137    0.0375    *
 STARPOWR      0.006427   0.013812    0.465    0.6437
 SEQUEL        0.337876   0.293126    1.153    0.2545
 ACTION        -0.654258  0.305963    -2.138   0.0374    *
 COMEDY        0.035994   0.275897    0.130    0.8967
 ANIMATED      -0.826735  0.462680    -1.787   0.0800    .
 HORROR        0.685153   0.385951    1.775    0.0819    .
 BUZZ          0.337698   0.077204    4.374    6.19e-05  ***
--
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Test of overall significance

- Assume we want to test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0.$$

- Under this hypothesis, we can show that

$$F = \frac{R^2}{1-R^2}\,\frac{n-p-1}{p}$$

  has a Fisher distribution $\mathcal{F}_{p,n-p-1}$ with p and $n-p-1$ df.

- The p-value is

$$p = \mathbb{P}_{H_0}(F \geq f) = \mathbb{P}(\mathcal{F}_{p,n-p-1} \geq f).$$

Example

```
> summary(fit)

:
Coefficients:
               Estimate   Std. Error  t value  Pr(>|t|)
 (Intercept)   15.172989  0.890296    17.043   < 2e-16   ***
 MPRATINGPG    0.069498   0.554641    0.125    0.9008
 MPRATINGPG13  -0.273367  0.591322    -0.462   0.6459
 MPRATINGR     -0.443641  0.595927    -0.744   0.4601
 BUDGET        0.409218   0.191454    2.137    0.0375    *
 STARPOWR      0.006427   0.013812    0.465    0.6437
 SEQUEL        0.337876   0.293126    1.153    0.2545
 ACTION        -0.654258  0.305963    -2.138   0.0374    *
 COMEDY        0.035994   0.275897    0.130    0.8967
 ANIMATED      -0.826735  0.462680    -1.787   0.0800    .
 HORROR        0.685153   0.385951    1.775    0.0819    .
 BUZZ          0.337698   0.077204    4.374    6.19e-05  ***
Residual standard error: 0.7183 on 50 degrees of freedom
Multiple R-squared: 0.5244, Adjusted R-squared: 0.4198
F-statistic: 5.013 on 11 and 50 DF, p-value: 3.26e-05
```


Overview

Exploiting the fitted regression model

- Let $x_0 = (1, x_{10}, \ldots, x_{p0})^T$ be the vector of predictors for a new observation, and Y_0 the corresponding unknown value of the response variable.
- We assume that our previous model is still valid for this new data, i.e., $Y_0 = \beta^T x_0 + \epsilon_0$ with $\epsilon_0 \sim \mathcal{N}(0, \sigma^2)$, and Y_0 is independent from the other observations.
- What can we say
  - About $f(x_0)$?
  - About Y_0?

Estimation of $f(x_0)$

- Point estimation: let $\widehat{f}(x_0) = \widehat{\beta}^T x_0$. It is an unbiased estimate of $f(x_0) = \mathbb{E}(Y_0 \mid x_0) = \beta^T x_0$, as

$$\mathbb{E}(\widehat{\beta}^T x_0) = \mathbb{E}(\widehat{\beta})^T x_0 = \beta^T x_0.$$

- To take into account the uncertainty of this estimation, we often prefer to compute a confidence interval.

Definition

A confidence interval (CI) on $f(x_0)$ at level $1-\alpha$ is a random interval $[L, U]$ that contains the true value of $f(x_0)$ for a proportion $1-\alpha$ of the training data (with fixed x_i's), i.e.,

$$\mathbb{P}_{\mathbf{Y}}(L \leq f(x_0) \leq U) = 1 - \alpha$$

Confidence interval on $f(x_0)$

- From $\widehat{\beta} \sim \mathcal{N}(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$, we get

$$\widehat{f}(x_0) = x_0^T \widehat{\beta} \sim \mathcal{N}(f(x_0), x_0^T (\mathbf{X}^T\mathbf{X})^{-1} x_0 \sigma^2)$$

- Hence,

$$\frac{\widehat{f}(x_0) - f(x_0)}{\sigma\sqrt{x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}} \sim \mathcal{N}(0,1).$$

- After replacing σ by $\widehat{\sigma}$ and using $\frac{(n-p-1)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$, we have

$$\frac{\widehat{f}(x_0) - f(x_0)}{\widehat{\sigma}\sqrt{x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}} \sim \mathcal{T}_{n-p-1}$$

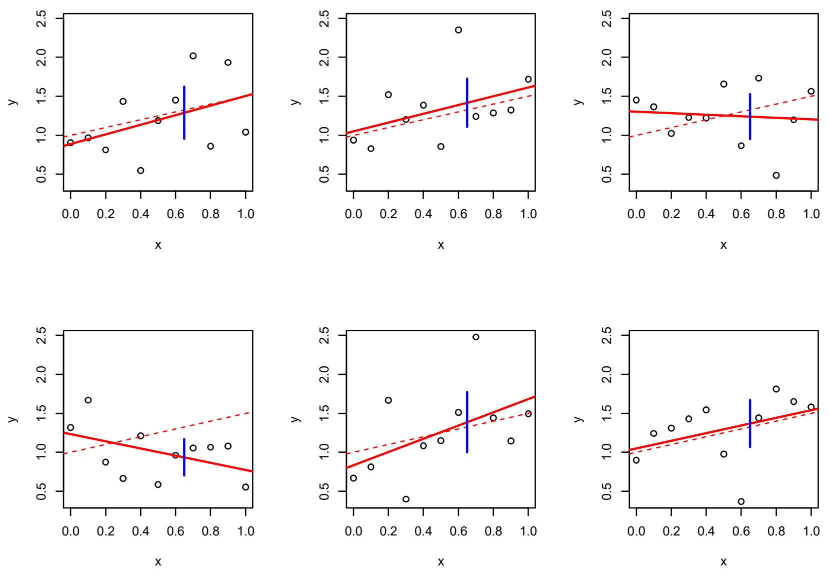
Confidence interval on $f(x_0)$ (continued)

- Given the previous results, it is easy to derive the following CI:

$$\boxed{\widehat{f}(x_0) \pm t_{1-\frac{\alpha}{2}} \widehat{\sigma} \sqrt{x_0^T (\mathbf{X}^T\mathbf{X})^{-1} x_0}}$$

where $t_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the Student distribution $\mathcal{T}_{n-p-1}$.
- For $1-\alpha = 0.95$, $t_{1-\frac{\alpha}{2}} \approx 2$.

Example



Prediction of Y_0

We now turn to the problem of predicting the random variable Y_0.

Definition

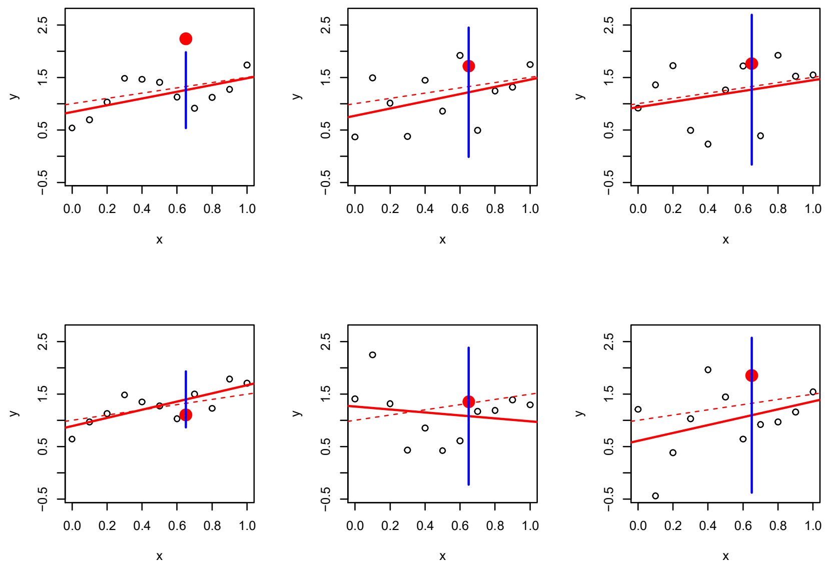
A prediction interval (PI) for Y_0 at level $1-\alpha$ is a random interval $[L, U]$ that contains Y_0 for a proportion $1-\alpha$ of the training data (with fixed x_i's), i.e.,

$$\mathbb{P}_{\mathbf{Y},Y_0}(L \leq Y_0 \leq U) = 1-\alpha$$

Prediction interval

- We have
$$Y_0 \sim \mathcal{N}(f(x_0), \sigma^2) \text{ and } \widehat{f}(x_0) \sim \mathcal{N}(f(x_0), x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\sigma^2).$$
- As Y_0 and $\widehat{f}(x_0)$ are independent,
$$Y_0 - \widehat{f}(x_0) \sim \mathcal{N}\left(0, \sigma^2[1 + x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0]\right)$$
- Hence,
$$\frac{Y_0 - \widehat{f}(x_0)}{\widehat{\sigma}\sqrt{1 + x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}} \sim \mathcal{T}_{n-p-1}$$
- Prediction interval:
$$\boxed{\widehat{f}(x_0) \pm t_{n-p-1;1-\frac{\alpha}{2}}\widehat{\sigma}\sqrt{1 + x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}}$$

Example



Example in R

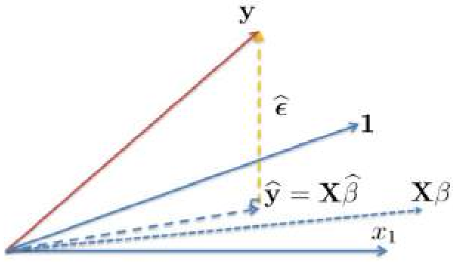
```
> x0 <- data.frame(MPRATING='PG13',BUDGET=5,STARPOWR=20,
SEQUEL=0, ACTION=1,COMEDY=0,ANIMATED=0, HORROR=0,BUZZ=1)

> predict(fit,int="c",newdata=x0)
          fit      lwr      upr
 [1,] 16.75769 16.18435 17.33104

> predict(fit,int="p",newdata=x0)
          fit      lwr      upr
 [1,] 16.75769 15.20528 18.31011
```

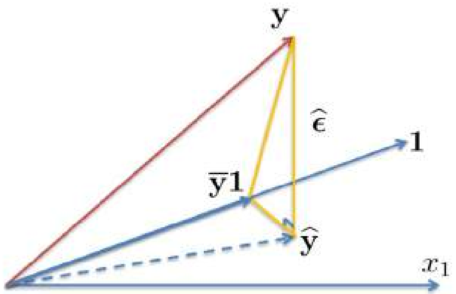

Geometric interpretation of linear regression

- The vectors $\mathbf{1}, \mathbf{x}_1, \ldots \mathbf{x}_p$ span a subspace $\mathcal{S}$ of $\mathbb{R}^n$, also referred to as the column space of $\mathbf{X}$. We have $\mathbf{X}\beta = \beta_0\mathbf{1} + \beta_1\mathbf{x}_1 + \ldots + \beta_p\mathbf{x}_p \in \mathcal{S}$.



- We chose $\widehat{\beta}$ by minimizing the distance between $\mathbf{X}\beta$ and $\mathbf{y}$. The solution is the orthogonal projection $\widehat{\mathbf{y}}$ of $\mathbf{y}$ onto $\mathcal{S}$.
- The hat matrix $\mathbf{H}$ computes the orthogonal projection.
- The residual vector $\widehat{\boldsymbol{\epsilon}} = \mathbf{y} - \widehat{\mathbf{y}}$ is orthogonal to $\mathcal{S}$.

Analysis of variance



- From $\widehat{\epsilon} \perp \mathcal{S}$, we have $\widehat{\epsilon} \perp \mathbf{1}$.
- Hence,

$$\langle \widehat{\epsilon}, \mathbf{1} \rangle = \sum_{i=1}^{n} \widehat{\epsilon}_i = 0,$$

and $\sum_{i=1}^n y_i = \sum_{i=1}^n \widehat{y}_i$.

- The projection of $\mathbf{y}$ on $\mathbf{1}$ is $\frac{\langle \mathbf{y}, \mathbf{1}\rangle}{\|\mathbf{1}\|^2}\mathbf{1} = \overline{y}\mathbf{1}$. Similarly for $\widehat{\mathbf{y}}$.
- Applying the Pythagorean theorem in the triangle $(\mathbf{y}, \widehat{\mathbf{y}}, \overline{y}\mathbf{1})$, we get

$$\|\mathbf{y} - \overline{y}\mathbf{1}\|^2 = \|\widehat{\mathbf{y}} - \overline{y}\mathbf{1}\|^2 + \|\mathbf{y} - \widehat{\mathbf{y}}\|^2,$$

which is the analysis of variance equation.

Back

Proof of the Gauss-Markov theorem I

- Let $\widetilde{\beta} = \mathbf{CY}$ be an linear unbiased estimate of β ($\mathbf{C}$ is a constant matrix).
- $\mathbb{E}(\widetilde{\beta}) = \mathbf{CX}\beta = \beta$, so $\mathbf{CX} = \mathbf{I}$.
- $\text{Var}(\widetilde{\beta}) = \mathbf{C}(\sigma^2\mathbf{I}_n)\mathbf{C}^T = \sigma^2\mathbf{CC}^T$.
- Let $\boldsymbol{D} = \mathbf{C} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, so $\mathbf{C} = \boldsymbol{D} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.
- We have

$$\text{Var}(\widetilde{\beta}) = \sigma^2\left[(\boldsymbol{D} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\boldsymbol{D} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T\right] \quad (4)$$

- Now, from $\mathbf{CX} = \mathbf{I}$, we get $\boldsymbol{D}\mathbf{X} + \underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}}_{\mathbf{I}} = \mathbf{I}$, so $\boldsymbol{D}\mathbf{X} = 0$.

Proof of the Gauss-Markov theorem II

- Developing (4), we get

$$\begin{aligned}\mathrm{Var}(\widetilde{\beta}) &= \sigma^2\left[(\boldsymbol{D}+(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\boldsymbol{D}+(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T\right]\\ &= \sigma^2\left[\boldsymbol{D}\boldsymbol{D}^T+(\mathbf{X}^T\mathbf{X})^{-1}+\underbrace{\boldsymbol{D}\mathbf{X}}_{0}(\mathbf{X}^T\mathbf{X})^{-1}+(\mathbf{X}^T\mathbf{X})^{-1}\underbrace{\mathbf{X}^T\boldsymbol{D}}_{0}\right]\\ &= \underbrace{\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}}_{\mathrm{Var}(\widehat{\beta})}+\sigma^2\boldsymbol{D}\boldsymbol{D}^T\end{aligned}$$

- So, the variance matrix of $\widetilde{\beta}$ equals that of $\widehat{\beta}$ plus nonnegative definite matrix $\sigma^2\boldsymbol{D}\boldsymbol{D}^T$.

Proof of the Gauss-Markov theorem III

- Consequently,

$$\begin{aligned}\mathrm{Var}(a^T\widetilde{\beta}) &= a^T\,\mathrm{Var}(\widetilde{\beta})a\\ &= a^T(\mathrm{Var}(\widehat{\beta})+\sigma^2\boldsymbol{D}\boldsymbol{D}^T)a\\ &= \underbrace{a^T\,\mathrm{Var}(\widehat{\beta})a}_{\mathrm{Var}(a^T\widehat{\beta})}+\underbrace{\sigma^2a^T\boldsymbol{D}\boldsymbol{D}^Ta}_{\geq 0}\\ &\geq \mathrm{Var}(a^T\widehat{\beta})\end{aligned}$$

Back

Definition of the multivariate normal distribution

Définition

Way say that random vector $\mathbf{X}$ *has a multivariate normal distribution if it has the following density function:*

$$p(\mathrm{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathrm{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathrm{x}-\boldsymbol{\mu})\right).$$

Notation: $\mathbf{X}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

Property:

$$\mathbb{E}(\mathbf{X})=\boldsymbol{\mu},\quad \mathrm{Var}(\mathbf{X})=\boldsymbol{\Sigma}.$$

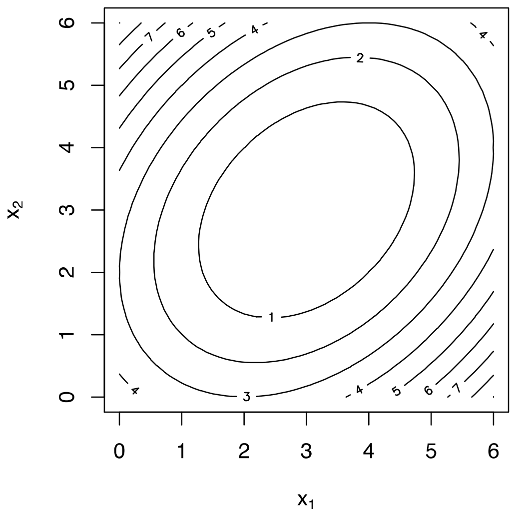
Properties of the multivariate normal distribution

- When $p=1$, we have the univariate normal distribution with $\sigma^2=\boldsymbol{\Sigma}$.
- Matrix Σ is diagonal iff r.v.'s $X_1,\ldots,X_p$ are independent.
- Any sub-vector of $\mathbf{X}$ has a normal distribution. In particular, the components X_i have normal distributions $\mathcal{N}(\mu_i,\sigma_i^2)$ with $\sigma_i^2=(\boldsymbol{\Sigma})_{ii}$.
- The multivariate normal distribution has constant density on ellipses or ellipsoids of the form

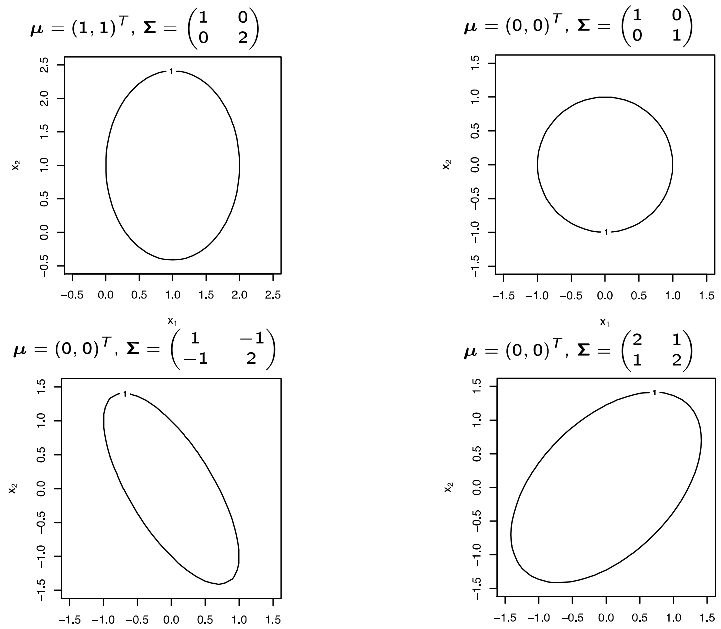
$$(\mathrm{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathrm{x}-\boldsymbol{\mu})=c^2$$

c being a constant. These ellipsoids are called the contours the distribution. For $\boldsymbol{\mu}=0$ these contours are centered at the origin. When $\boldsymbol{\Sigma}=a\mathbf{I}$ the contours are circles or, in higher dimensions, spheres or hyperspheres.

Example with $\boldsymbol{\mu} = (3,3)^T$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$

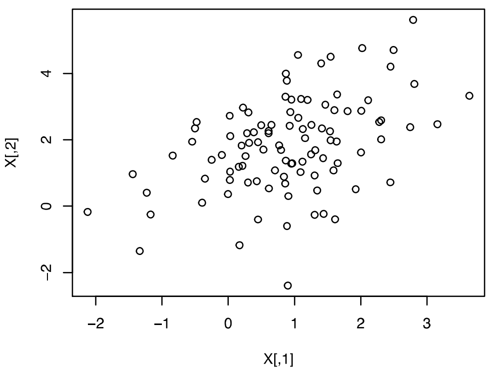


More examples



Multivariate normal random vector generation in R

```
library(mvtnorm)
mu<-c(1,2)
Sigma<-matrix(c(1,0.5,0.5,2),2,2)
X<-rmvnorm(100,mu,Sigma)
plot(X)
```



Back