ML01 – Introduction to Machine Learning

Linear Regression

Thierry Denœux

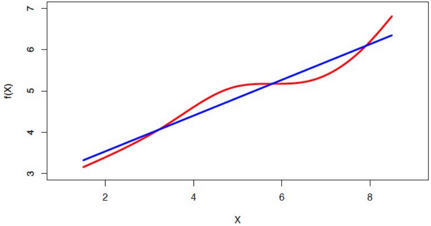
tdenoeux@utc.fr
https://www.hds.utc.fr/~tdenoeux

Université de technologie de Compiègne

Spring 2021

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.
- True regression functions are never linear!



- 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
- 11 predictors:
  - MPAA (Motion Picture Association of America) rating (G, PG, PG13)
  - 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?
- Is there synergy among the factors that influence success (e.g., between genre and budget)?

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 log, square-root or square
  3. Basis expansions, such as $X_2 = X^2$, $X_3 = X^3$, leading to a polynomial representation
  4. Interactions between variables, for example, $X_3 = X_1 \cdot X_2$. This allows us to model synergy (interaction) between variables
  5. 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 have to use only 3 out of the 4 dummy variables.

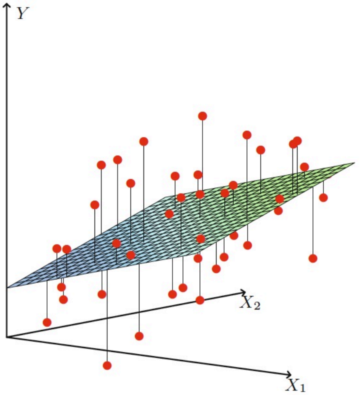
Overview

1. The method of least squares
   - LS estimates
   - Analysis of variance
   - Application in R and interpretation of the coefficients

2. Inference
   - Additional assumptions and properties of the estimates
   - Tests of significance
   - Prediction

Estimation

- Typically we have a set of training data $(x_1, y_1), \ldots, (x_n, y_n)$ from which to estimate the parameters β. Each $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ is a vector of predictor measurements for the ith case.



- The most popular estimation method is least squares, in which we minimize the sum of squared residuals (differences between y_i and $f(x_i)$).

The RSS criterion

- The mean squared error or residual sum of squares (RSS) is

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

- To find the vector β that minimizes $\text{RSS}(\beta)$, it is convenient to use matrix notation.

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 $\text{RSS}(\beta)$, 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 **A** *be a constant matrix and* β *is a vector. 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 **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

$$\frac{\partial \text{RSS}}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta\right) \tag{2a}$$

$$= -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta \tag{2b}$$

- Setting the gradient to zero, we get

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

- Assuming that $\mathbf{X}$ has full column rank, $\mathbf{X}^T\mathbf{X}$ 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

- Let $\widehat{\mathbf{y}} = (\widehat{y}_1, \ldots, \widehat{y}_n)^T$ be the vector of fitted values at the training inputs,

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

- It can be computed as

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

- Matrix $\mathbf{H}$ is sometimes called the hat matrix because it puts the hat on $\mathbf{y}$.

Overview

1. The method of least squares
   - LS estimates
   - Analysis of variance
   - Application in R and interpretation of the coefficients
2. Inference
   - Additional assumptions and properties of the estimates
   - Tests of significance
   - 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 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. $\mathbf{y} = \widehat{\mathbf{y}} \in \mathcal{S}$: 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

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
```

Interpretation of the coefficients

- We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed.
- This interpretation may be delicate when the predictors are correlated!
- Example:
    - Y total amount of change in your pocket
    - $X_1 = \#$ of coins
    - $X_2 = \#$ of 10 cts and 20 cts.

  By itself, regression coefficient of Y on X_2 will be > 0. But how about with X_1 in model?

Overview

Statistical significance

- From the result of the regression analysis, can we infer that, say, there is a relation between the budget of a movie and box office receipts?
- Can we infer which factors contribute to box office receipts?
- It is difficult to answer these questions because, even if $\beta_j = 0$ for some predictor X_j, i.e., if there is no relation between X_j and the response variable Y, the estimated coefficient $\widehat{\beta}_j$ will not be exactly equal to 0.
- Which values of $\widehat{\beta}_j$ can be considered statistically significant?
- To answer this question, we need to make additional assumptions.

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 assume that
  1. The observations Y_i are uncorrelated and have constant variance σ^2:
  $$\text{Var}(Y_i) = \sigma^2, \quad \text{Cov}(Y_i, Y_i) = 0, \forall i \neq j,$$
  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 $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\beta$).
  2. The x_i are fixed (nonrandom), so $\mathbf{X}$ is a constant matrix.

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

Proposition

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

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

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

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

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

Variance estimation

- Typically one estimates the variance σ^2 by

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

- The $n-p-1$ rather than n in the denominator makes $\widehat{\sigma}^2$ an unbiased estimate of σ^2:

$$\mathbb{E}(\widehat{\sigma}^2) = \sigma^2.$$

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

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

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,

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

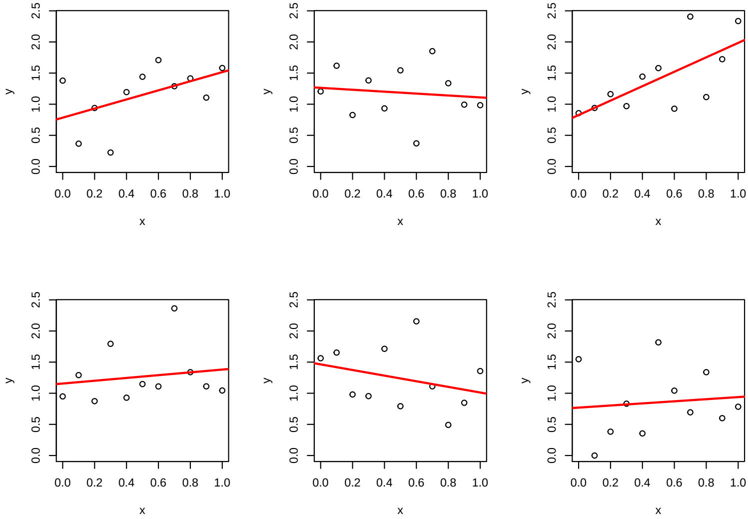
Simulation example

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

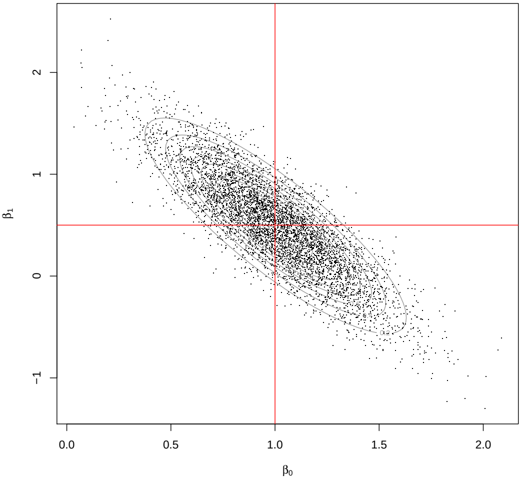
$$Y_i = 1 + 0.5x_i + \epsilon_i$$

  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

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

Overview

Significance of a coefficient

- Assume that we have observed the estimate $\widehat{\beta}_j$ for predictor X_j.
- We always have $\widehat{\beta}_j \neq 0$. Can we deduce that $\beta_j \neq 0$?
- Let H_{j0} denote the hypothesis $\beta_j = 0$ (the "null hypothesis")
- Method of approach: compute the distribution of $\widehat{\beta}_j$ assuming that H_{j0} is true.
- If it is unlikely that the observed value of $\widehat{\beta}_j$ was drawn from this distribution, then we can reject hypothesis H_{j0}.

Significance of a coefficient

- We have seen that $\widehat{\beta} \sim \mathcal{N}(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$. Let v_j be the jth diagonal element of matrix $(\mathbf{X}^T\mathbf{X})^{-1}$.
- Assuming $\beta_j = 0$, we have

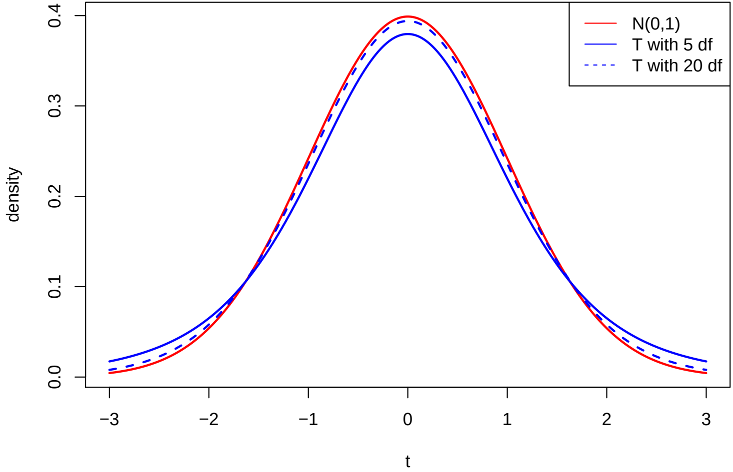
$$\widehat{\beta}_i \sim \mathcal{N}(0, v_j\,\sigma^2), \quad \text{i.e.,} \quad \frac{\widehat{\beta}_j}{\sigma\sqrt{v_j}} \sim \mathcal{N}(0,1).$$

- We don't know σ, but we can replace it by $\widehat{\sigma}$. We can show that

$$\boxed{T_j = \frac{\widehat{\beta}_j}{\widehat{\sigma}\sqrt{v_j}} \sim \mathcal{T}_{n-p-1}}$$

where $\mathcal{T}_{n-p-1}$ denotes Student distribution with $n-p-1$ degrees of freedom. Variable T_j is called the standardized coefficient.

Student distribution



Significance of a coefficient

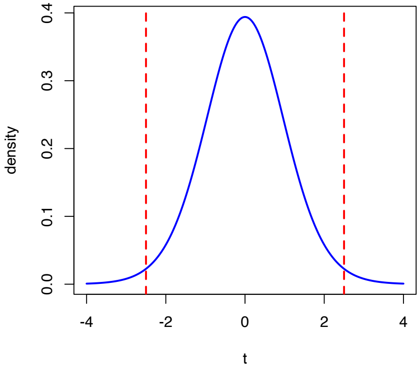
- As the distribution of T_j under hypothesis H_{j0} is centered around 0, if we observe t_j far from 0, it makes H_{j0} unlikely.
- How far from 0 should t_j be to make us reject H_{j0}?
- We define the p-value as the probability, if H_{j0} is true, that $|T_j|$ is at least as large as the observed value $|t_j|$:

$$p_j = \mathbb{P}_{H_0}(|T_j| \geq |t_j|) = \mathbb{P}_{H_0}(T_j \geq |t_j|) + \mathbb{P}_{H_0}(T_j \leq -|t_j|)$$

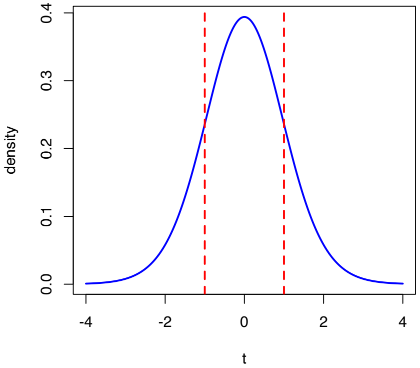
  (see example on next slide)
- The smaller this value, the more we doubt that $\beta_j = 0$.
- Usual reference values: ≤ 0.1 (weakly significant), ≤ 0.05 (significant), ≤ 0.01 (very significant).
- For $t_j = 2$ we have $p \approx 0.05$.

Examples

$t = 2.5, \ p = 0.021$



$t = 1, \ p = 0.33$



Example (Movies dataset)

```
> 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
```

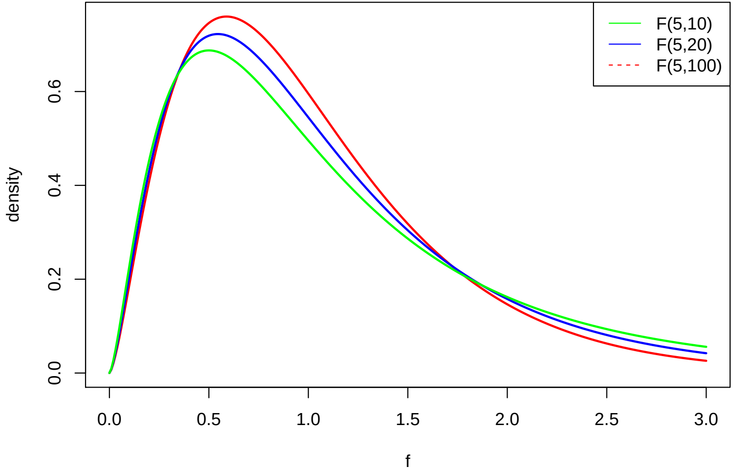
Test of overall significance

- Assume we want to test hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$, meaning that no predictor can explain the variability of Y.
- We cannot simply consider all the previous p-values, because if there are many predictors, there is a high chance that some $|t_j|$ will be large even if H_0 is true.
- We can show that, if H_0 is true, the statistic

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

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

Fisher distribution



Test of overall significance

- When $R^2 \to 1$, $F \to +\infty$: a large value of f corresponds to a value of R^2 close to 1, and it is evidence against H_0.
- The p-value of the test of overall significance of the regression is

$$p = \mathbb{P}_{H_0}(F \geq f).$$

- It is the probability, if H_0 is true (no relation between the X_j's and Y), of observing a value of statistics F at least as high as the value f that we did observe.

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 $f(x_0) = \beta^T x_0$?
  - 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

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

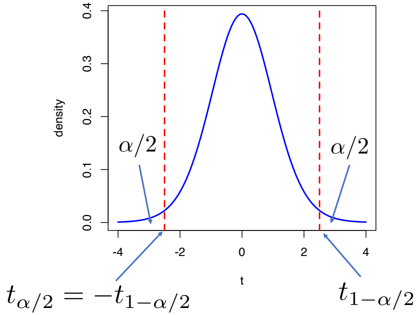
- Hence,

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

- After replacing σ by $\widehat{\sigma}$, we can show that

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

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



Consequently, we have

$$\mathbb{P}\left(-t_{1-\frac{\alpha}{2}} \leq \frac{\widehat{f}(x_0) - f(x_0)}{\widehat{\sigma}\sqrt{x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}} \leq t_{1-\frac{\alpha}{2}}\right) = 1 - \alpha,$$

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

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

- Equivalently,

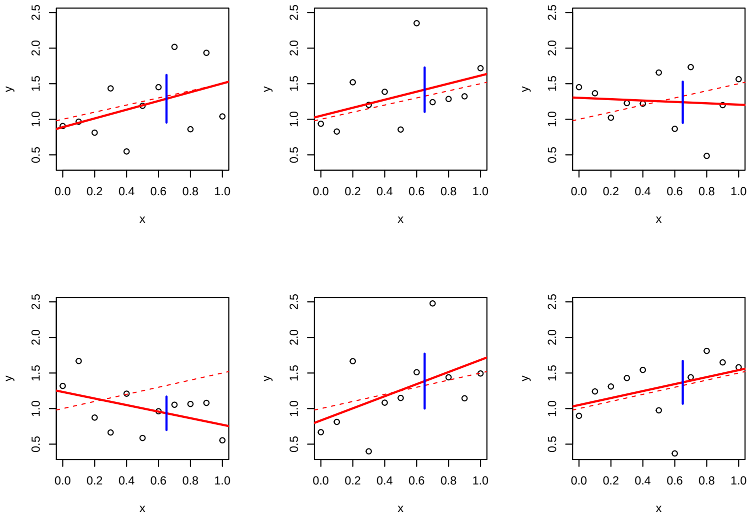
$$\mathbb{P}\left(\widehat{f}(x_0) - t_{1-\frac{\alpha}{2}}\widehat{\sigma}\sqrt{x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0} \leq f(x_0) \leq \widehat{f}(x_0) + t_{1-\frac{\alpha}{2}}\widehat{\sigma}\sqrt{x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}\right) = 1-\alpha$$

- We thus have the following CI:

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

- 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,

$$Y_0 - \widehat{f}(x_0) \sim \mathcal{N}\left(0, \sigma^2[1 + x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0]\right)$$

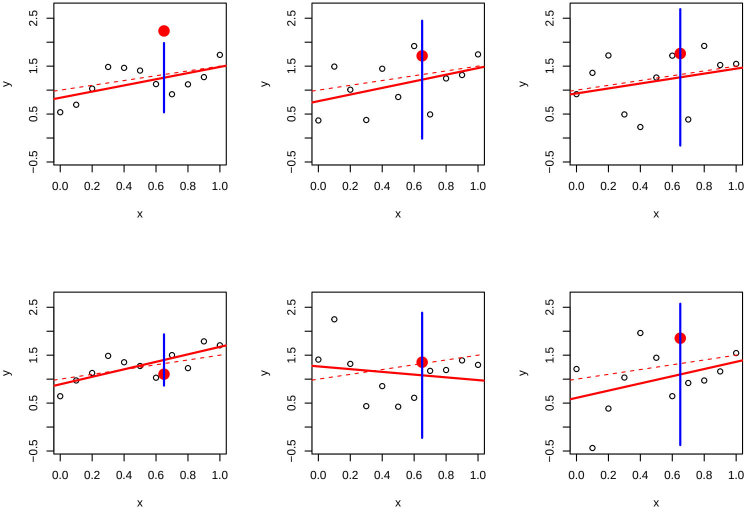
- Hence,

$$\frac{Y_0 - \widehat{f}(x_0)}{\widehat{\sigma}\sqrt{1 + x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}} \sim \mathcal{T}_{n-p-1}$$

- Prediction interval:

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

Example



Example in R

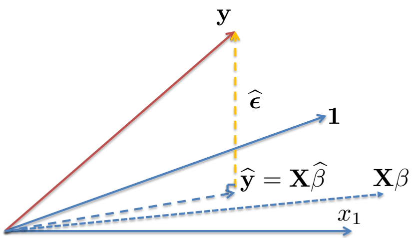
```
> x0 <- data.frame(MPRATING='PG13',BUDGET=5,STARPOWR=20,
SEQUEL=0, ACTION=1,COMEDY=0,ANIMATED=0, HORROR=0,BUZZ=1)

> 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

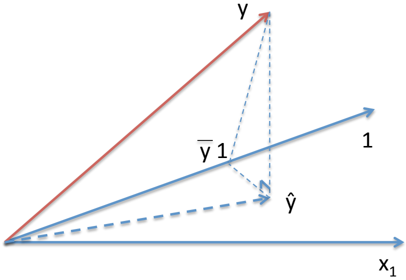
- The vectors $\mathbf{1}, \mathbf{x}_1, \ldots \mathbf{x}_p$ span a subspace $\mathcal{S}$ of $\mathbb{R}^n$, also referred to as the column space of $\mathbf{X}$. We have $\mathbf{X}\beta = \beta_0\mathbf{1} + \beta_1\mathbf{x}_1 + \ldots + \beta_p\mathbf{x}_p \in \mathcal{S}$.



- 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, and hence it is also known as a projection matrix.

- 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