Advanced Statistics
Part one: Parametrical Statistics

M. Pérez-Casany

Dept. of Statistics and Operations Research and DAMA-UPC
Technical University of Catalunya

First Semester: Fourth Session

3. Linear Models: Regression

Assume that one is interested in, for instance:

1) Describing the **Processor times** behaviour as a function of of the number of I/O's
2) Describing the behaviour of a remote procedure mechanism, denoted by (RPC), by means of the **Total elapsed time** as a function of the *Data size*.
3) Describing the behaviour of a remote procedure mechanism, denoted by (RPC), by means of the **Total elapsed time** as a function of the *Data size* and the *Operation System*.

Linear model:Regression

In the case where the covariables are continuous variables, the linear model is called *Linear Regression model*.

It may be *simple* if just one covariables is considered and multiple if there are more than one explanatori variables.

Linear model: Simple Regression

Example: In a database information system that allows its users to search backward for several days wanted to depvelop a formula to predict the time it would take to search (Y) as a function of the days (X).

$$Y_i = \beta_0 + x_i\beta_1 + e_i, \; i = 1, \cdots, n;$$

in matrix form,

$$\begin{pmatrix} 0.65 \\ 0.79 \\ 1.36 \\ 2.26 \\ 3.59 \\ 5.39 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 8 \\ 1 & 16 \\ 1 & 25 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{pmatrix}$$

Linear model: Simple Regression

Hypothesis:

- $e_i \sim N(0, \sigma^2)$
- Independ. of e_i and e_j
- equality of variances (homocedasticity)

Observation: In this case usually $X^t X$ is a not singular matrix, and hence, its inverse usually exists.

Linear model: Simple Regression

Parameter estimations:

$\hat{\beta}_0$ and $\hat{\beta}_1$ are the ones that minimize the residual sum of squares:

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0 - x_i\, \beta_1)^2$$

and are equal to:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n(\overline{x})^2} = \frac{Cov(X,Y)}{Var(X)}$$

Observation 1: $(\overline{x}, \overline{y})$ belongs to the regression line.

Observation 2: β_1 is the increment of Y increasing one unit the value of X. β_0 is the response when $X = 0$.

Linear model: Simple Regression

Estimation of the variance parameter:

$$S_e^2 = \hat{\sigma}^2 = \frac{SSE}{n-2} \longrightarrow S_e = \hat{\sigma} = \sqrt{\frac{SSE}{n-2}}$$

S_e^2 is known as *mean square error*.

Linear model: Simple Regression

Standard deviation of the parameter estimations

$$S_{\hat{\beta}_0} = S_e \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} \right]^{1/2}$$

and

$$S_{\hat{\beta}_1} = S_e \, \frac{1}{[\sum_{i=1}^n x_i^2 - n\,\overline{x}^2]^{1/2}}$$

Linear model: Simple Regression

Confidence intervals for regresion parameters

$$\hat{\beta}_0 \pm t_{1-\alpha/2,n-2} \cdot S_{\hat{\beta}_0}$$

and

$$\hat{\beta}_1 \pm t_{1-\alpha/2,n-2} \cdot S_{\hat{\beta}_1}$$

Observation: If the interval contains the **zero value**, the corresponding covariate is **not statistically significative.**

Linear model: Simple Regression

Confidence intervals for the predictions

Predicted value: $\hat{y}_i = \hat{\beta}_0 + x_i\hat{\beta}_1$.

The prediction under the i-thm experimental conditions is also a r.v. that we denote by Y_i

The puntual estimation: $\hat{y}_i$ is the **predicted mean value** of the response, i.e, $E(\hat{Y}_i) = \hat{y}_i$

Linear model: Simple Regression

If we get a sample of size m of responses under condition i-thm then,

$$S_{\overline{y}_{i,m}} = S_e \cdot \left[\frac{1}{m} + \frac{1}{n} + \frac{(x_i - \overline{x}^2)}{\sum_{k=1}^{n} x_k^2 - n\overline{x}^2}\right]^{1/2}$$

Observation: As closer is x_i to $\overline{x}$, smaller is the s.d.

Observation: Near $(\overline{x}, \overline{y})$ the predictions are more accurate.

Linear model: Simple Regression

The estandard deviation of a **single future observation** is equal to:

$$S_{\hat{y}_i} = S_e \cdot \left[1 + \frac{1}{n} + \frac{(x_i - \overline{x}^2)}{\sum_{k=1}^{n} x_k^2 - n\overline{x}^2}\right]^{1/2}$$

and, when m tends to infinity one obtains the mean of a large number of future observations at x_i, which is interpreted as the sd **of the mean value** $E(Y_i)$. It is equal to:

$$S_{\hat{y}_i} = S_e \cdot \left[\frac{1}{n} + \frac{(x_i - \overline{x}^2)}{\sum_{k=1}^{n} x_k^2 - n\overline{x}^2}\right]^{1/2}$$

Observation: The sd of the mean of an infinite future sample is smaller than the one of a finite sample.

Linear model: Simple Regression

Coefficient of determination

$$R^2 = \left(\frac{S_{xy}^2}{S_x \cdot S_y}\right)^2 = (r_{yx})^2 = \frac{SS_T - SS_E}{SS_T} = \frac{SS_R}{SS_T}$$

It is the square of the sample correlation coefficient.

It describes the proportion of variability in the data explained by the model.

The higher the value of R^2 better the model is.

Linear model: Simple Regression

Checking the model assumptions

- Scatter plot of y vs x, linearity must be observed.
- Scatter plot of $\hat{e}_i$ vs $\hat{y}_i$, no trend must be observed.
- qq-plot for $\hat{e}_i$, linearity must be observed.
- Sometimes, plot $\hat{e}_i$ vs order in which the observations are conducted.

Linear model: Multiple Regression

Example: Seven programs were monitored to observe their resource demands. In particular, the number of disk I/='s, menory size (in Kilobytes), and CPU time (in milliseconds) were observed. One is interested in modelling CPU time (Y) as a function of the other two.

$$Y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + e_i, \ i = 1, \cdots, n;$$

assuming the general linear model assumptions-

Linear model: Multiple Regression

CPU Time (y_i)	Disk I/0's (x_{1i})	Memory size (X_{2i})
2	14	70
5	16	75
7	27	144
9	42	190
10	39	210
13	50	235
20	83	400

Linear model: Multiple Regression

- Parameter estimation: $\hat{\beta} = (X^t \cdot X)^{-1} X^t y$
- Predicted values: $\hat{y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2$.
- Residual sum of squares: $SSE = \sum_i (y_i - \hat{y}_i)^2$.
- Variance estimation: $S_e^2 = \hat{\sigma}^2 = MSE = \frac{SSE}{n-p-1}$
- Standard deviation for β's: $S_{\beta_j} = S_e \cdot \sqrt{c_{jj}}$, being c_{jj} the diagonal elements of $(X^t \cdot X)^{-1}$
- Conf. int for parameter: $\hat{\beta}_i \pm t_{1-\alpha/2,n-p-1} \cdot S_{\beta_j}$
- Coef. of determination: $R^2 = \frac{SSR}{SST} = \frac{SS_T - SSE}{SST}$

Linear model: Multiple Regression

For the CPU time example,

$\hat{\beta} = (-0.1614, 0.1182, 0.02265)^t$.

The model is:

$$\text{CPU time} = -0.1614 + 0.1182 \cdot \text{Number of disk I/0's} + 0.0265 \cdot \text{Memory size}$$

Residual sum of squares: $SSE = \sum_i e_i^2 = 5.3$

Linear model: Multiple Regression

Coefficient of determination:

$$R^2 = \frac{SST - SSE}{SST} = 0.97$$

Standard deviation of errors:

$$S_e = \sqrt{SSE/(7-3)} = 1.2$$

Linear model: Multiple Regression

Estandard errors and conf. interv. for parameter estimations

	S_{b_i}	Conf. interv. (90%)
β_0	$1.2\sqrt{0.6297} = 0.9131$	$(-2.11, 1.79)$
$\beta_!$	$1.2\sqrt{0.0280} = 0.1925$	$(-0.29, 0.53)$
β_2	$1.2\sqrt{0.0012} = 0.0404$	$(-0.06, 0.11)$

Observation: None of the parameters are significant at 90% confidence level.

Linear model: Multiple Regression

Multicolliniarity: When two predictor variables are linearly dependent, they are called collinear.

Correlations among predictors:

$$R_{x_1x_2} = \frac{\sum_i x_{1i}x_{2i} - n\overline{x}_1\,\overline{x}_2}{\left[\sum_i x_{1i}^2 - n\overline{x}_1^2\right]^{1/2}\left[\sum_i x_{2i}^2 - n\overline{x}_2^2\right]^{1/2}}$$

Observation 1: If a correlation term is relatively high, one of the x's has to be eleminated and redo the analysis.

Observation 2: To choose the better subset of predictors from a set of k, one can perform the $2^k - 1$ analysis and choose the one that gives the best results with a small number of variables.

Linear model: Multiple Regression

In the CPU example:

Computing the correlation coef. between I/O's and memory size we get:

$$R_{x_1x_2} = 0.9947$$

This may be due to large programs (large memory size) doing more I/O's than small programs.

Peforming the two simple linear regressions one can see that both are useful to estimate CPU Time. So, one can consider each one of the regresors but not both.

Curvilinear Regression

Curvilinear regression: The relation of Y with the predictors is not linear, but may be linearized using a suitable transformation.

Examples with one covariate:

Nonlinear	Linear
$y = a + b/x$	$y = a + b(1/x)$
$y = x/(a + bx)$	$(x/y) = a + bx$
$y = a + bx^n$	$y = a + b(x^n)$
$y = 1/(a + bx)$	$(1/y) = a + bx$
$y = bx^a$	$\log y = \log b + a \log x$

Curvilinear Regression

Observation 1: If one covariable appears in more than one transformed predicted variable colliniarity problems may appear.

Observation 2: Parameter interpretation is different from the linear case.

1) $\log y = \beta_0 + \beta_1 x$, one unit of increment of x implies $(e^{\beta_1} - 1) \cdot 100$ % of increment of y
2) $y = \beta_0 + \beta_1 \log x$, 1% of change in x implies $\beta_1 \log(1.01)$ change in y.
3) $\log y = \beta_0 + \beta_1 \log x$, 1% of change in x implies $((1.01)^{\beta_1} - 1) \cdot 100\%$ change in y.

Regression: Outlier

An observation that is atypical from the remaining observations is suitable to be an **Outlier**.

Steeps to follow when there exists outliers:

1) To look the scattered plot to identify the possible outliers.
2) To investigate in the neighborhood to see if the are caused by an experimental error.
3) To perform the analysis with and without the atypical observations, ans state the results for both situations.
4) To divide the operating section in two (or more) subregions and to present a model for each subregion.

Regression: Common mistakes

- Not verifying that the relationship is linear
- Relying on the automated results without visual verification
- Not taking into account that parameter estimations depend upon the units of the predicted and predictor variables.
- Confusing the Coeff. of Determination and the Coeff. of Correlation.
- Using highly correlated variables in the prediction
- Using regression to predict far beyond the measured range. The statistical confidence decreases as we move outside the measured range.

- Using too many predictor variables (overfitting).
- Measuring only a small subset of the complete range of operation of a system.
- Confusing correlation with causality: two variables may be hightly correlated but none of them controls the other one.

  Regression the I/O's variable with respect to CPU time, we can deduce that more CPU time may be used to predict the number of disk I/O's.

  Nevertheless, installing a faster CPU will not imply a reduction of the number of disk I/O's, inspite that the observed CPU time will be smaller.