Chapter 1

The general linear model

Contents

1 Introduction

Regression analysis is a collection of statistical techniques for modeling and investigating the relationship between a response variable of interest and a set of regressors or predictor variables. A very important type of regression model is the linear regression model. Examples of this family of models are:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{simple linear regression} \\ y_i &= \beta_0 + \beta_1 x_{i1} + \ldots \beta_k x_{ik} + \epsilon_i \quad \text{multiple linear regression} \\ y_{ij} &= \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad \text{factorial model} \end{aligned}$$

for $i = 1, \ldots, n$, where we assume that ϵ_i form a random sample from $N(0, \sigma^2)$. Here, y is the dependent variable, $x_1, \ldots, x_k$ are the independent variables, and $\beta_0, ...\beta_k$ are unknown parameters

1.1 Matrix representation of the linear regression model

All models shown above maybe written in matrix from:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \varepsilon \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

where (in the case of multiple regression),

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \ddots & \vdots & \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The assumption $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}) \Rightarrow \boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$ (see results 7.2 and 7.4 of Review of Matrix Algebra).

2 Parameter estimation

The maximum likelihood estimates of the parameters are calculated via the minimization of

$$L(\beta_0, \ldots, \beta_k, \sigma^2 | \underline{x}) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2}{2\sigma^2},$$

minimizing of this function with respect to β_j is equivalent to minimize

$$S = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2$$

or in matrix form:

$$S(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{1.1}$$

This is called the **least squares method**, and it chooses $\boldsymbol{\beta}$ such that the sum of squares of the errors ε_i is minimized.
Taking the partial derivative with respect to $\boldsymbol{\beta}$ (see results in section 6 of Review in Matrix Algebra), and setting it to 0, we get:

$$2\boldsymbol{X}'\boldsymbol{Y} = 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \quad \text{normal equations}$$

if $\boldsymbol{X}$ is of full-rank, the $\boldsymbol{X}'\boldsymbol{X}$ is non-singular, and so the solution to the minimization problem is:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{Y}$$

2.1 Distribution of the parameter estimates

We write

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

Now $E(\boldsymbol{\varepsilon}) = 0$, hence, $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, i.e., $\hat{\boldsymbol{\beta}}$ is unbiased.
Further,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} = \boldsymbol{L}\boldsymbol{\varepsilon}$$

Thus, since $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{L}\boldsymbol{L}'),$$

and $\boldsymbol{L}\boldsymbol{L}' = (\boldsymbol{X}'\boldsymbol{X})^{-1}$, and so we rewrite the expression above as,

$$\hat{\boldsymbol{\beta}} \sim N_n\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right),$$

2.2 Estimation of σ^2

Define

$$S(\hat{\boldsymbol{\beta}}) = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{1.2}$$

as the residual sum of squares (SS_R), then

$$S(\hat{\boldsymbol{\beta}})/\sigma^2 \sim \chi^2_g$$

where $g = n - p$, thus g = number of independent observations minus number of parameters fitted. Furthermore, $S(\hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{\beta}}$ are independent.
Hence,

1. $E(S(\hat{\boldsymbol{\beta}})) = g\sigma^2$, and so $\hat{s}^2 = S(\hat{\boldsymbol{\beta}})/g$ is our unbiased estimator of σ^2.

2. $$\frac{(\hat{\beta}_j - \beta_j)}{\sqrt{v_{jj}s^2}} \sim t_{n-p}$$

   where v_{jj} are the diagonal elements of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$

2.3 Example 1

In an effort to model executive compensation for the year 1979, 33 firms were selected, and data were collected on compensation ($\boldsymbol{y}$), sales ($\boldsymbol{x}_1$), profits ($\boldsymbol{x}_2$) and employment ($\boldsymbol{x}_3$), data are available in the file *reg1.txt*.

```
reg1<-read.table("reg1.txt",header=TRUE)
pairs(reg1)
fit1<-lm(y~x1+x2+x3,data=reg1)
summary(fit1)
Call:
lm(formula = y ~ x1 + x2 + x3, data = reg1)

Residuals:
     Min       1Q   Median       3Q      Max
-152.820  -71.659    9.047   55.077  239.180

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -155.07     170.69  -0.908    0.371
x1             12.37      58.48   0.212    0.834
x2             67.13      44.05   1.524    0.138
x3             12.30      33.40   0.368    0.715

Residual standard error: 92.94 on 29 degrees of freedom
Multiple R-Squared: 0.497,      Adjusted R-squared: 0.4449
F-statistic:  9.55 on 3 and 29 DF,  p-value: 0.0001510
```

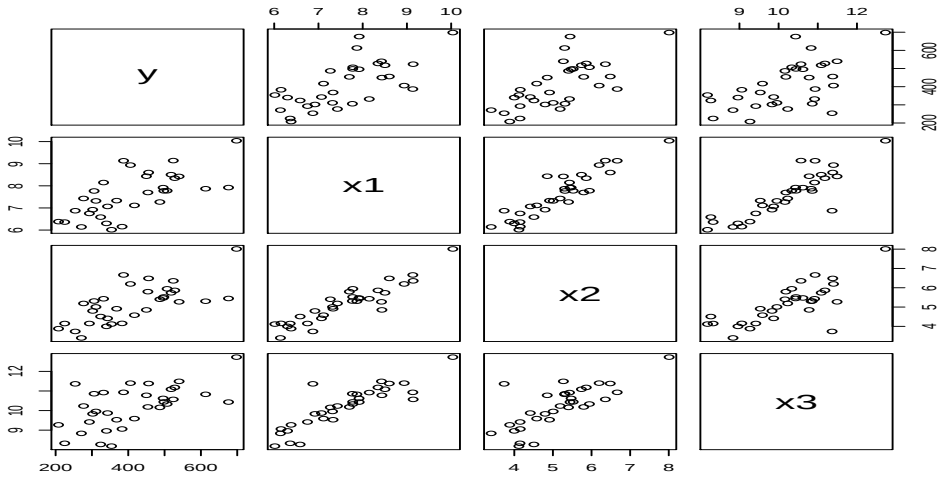


Figure 1: Pairwise scatterplot of variables

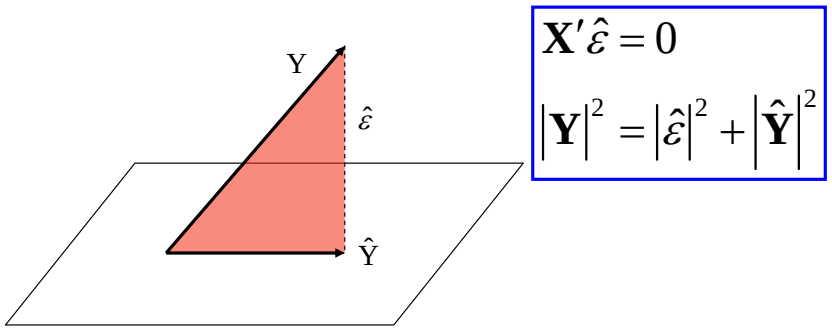
Figure 1 shows pairwise plots between all variables. Note that there is a strong relations among the explanatory variables.
The vector of estimated parameters is $\boldsymbol{\beta} = (-155.07, 12.37, 67.13, 12.30)'$, and $\hat{s}^2 = 92.93^2$.

3 Residuals

Define the *fitted values* as $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and define $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ as the *residuals*. Then we can check that

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y},$$

where $\boldsymbol{H}$ is a projection matrix, i.e., it satisfies $\boldsymbol{H} = \boldsymbol{H}' = \boldsymbol{H}\boldsymbol{H}'$.
Then, $\hat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}$ and so



$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{I}) \Rightarrow \hat{\boldsymbol{\varepsilon}} \sim N_n(\mathbf{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H}))$$

Similarly, $\hat{\boldsymbol{Y}} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{H})$.

3.1 Scaling residuals

In many occasions scaled residuals convey more information than do the ordinary residuals

Standardized residuals

Standardized residuals are defined as

$$d_i = \frac{\hat{\varepsilon}_i}{\sqrt{Var(\hat{\varepsilon}_i)}}$$

As we saw above, the variance of the ith residual is $Var(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$, where h_{ii} is the ith diagonal element of $\boldsymbol{H}$, and it is a measure of the location of ith point in the x-space.

Therefore, the variance of $\hat{\varepsilon}_i$ depends on the position of x_i. Generally, residuals near the center of the x-space will have larger variances than points at more remote locations.

$$d_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{s}^2(1-h_{ii})}}$$

they have mean zero and approximately unit variance; consequently they are useful in looking for **outliers**. An observation with standardized residual outside of $[-3, 3]$ is potentially unusual.

Studentized residuals

Since violation of model assumptions are more likely to occur at remote points, ordinary or standardized residuals may not be useful for detecting these violations. One solution is to define the **studentized residuals**:

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{s}^2_{(i)}(1-h_{ii})}}$$

where $\hat{s}^2_{(i)}$ is the residual mean square obtained by fitting the model without the $i-th$ observation.

3.2 Leverage Points

The hat matrix $\boldsymbol{H}$ is very useful in identifying influential observations (i.e., observations whose presence of absence in the data have an influence in the model fitted).
The elements of $\boldsymbol{H}$ may be interpreted as the amount of leverage exerted by y_i on $\hat{y}_i$, specially the diagonal elements h_{ii}.
$\boldsymbol{H}$ is an idempotent matrix, therefore all its eigenvalues are 0 or 1, and so, $tr(\boldsymbol{H}) = \sum_{i=1}^{n} h_{ii}$ =sum of the eigenvalues of $\boldsymbol{H} = Rank(\boldsymbol{H}) = p$. Therefore the average size of a diagonal element of $\boldsymbol{H}$ is p/n, if a diagonal element h_{ii} is greater that $2p/n$ is a high-leverage point.

Influence on regression coefficients

A measure of the influence of a point in the model fitted is given by the **Cook's Distance** which measures how much the model fitted changes when that observation is present or absent in the data,

$$D_i = \frac{(\hat{y}_i - \hat{y}_{(i)i})^2}{pVar(\hat{y}_i)}$$

where $\hat{y}_{(i)i}$ is the ith fitted value obtained when the parameters of the models are estimated deleting that point from the data. Cook (1977) showed that the D_i statistic may be written as

$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{(1-h_{ii})}$$

D_i consists of the squared studentized residual, which reflects how well the model fits the ith observation y_i and a component that measures how far is the point from the rest of the data,

$\frac{h_{ii}}{(1-h_{ii})}$. In section 4.2 we will see that the expression above is similar equation to (1.3) which is distributed as $F_{p,n-p}$, although D_i is not distributed as an F, points for which $D_i > 1$ are considered influential.

For a complete detailed information on residuals and diagnostic methods see Cook and Weisberg (1982).

Example 1

In the previous example, the different types of residuals, the leverage points and cook's distance are calculated as...left to the reader!

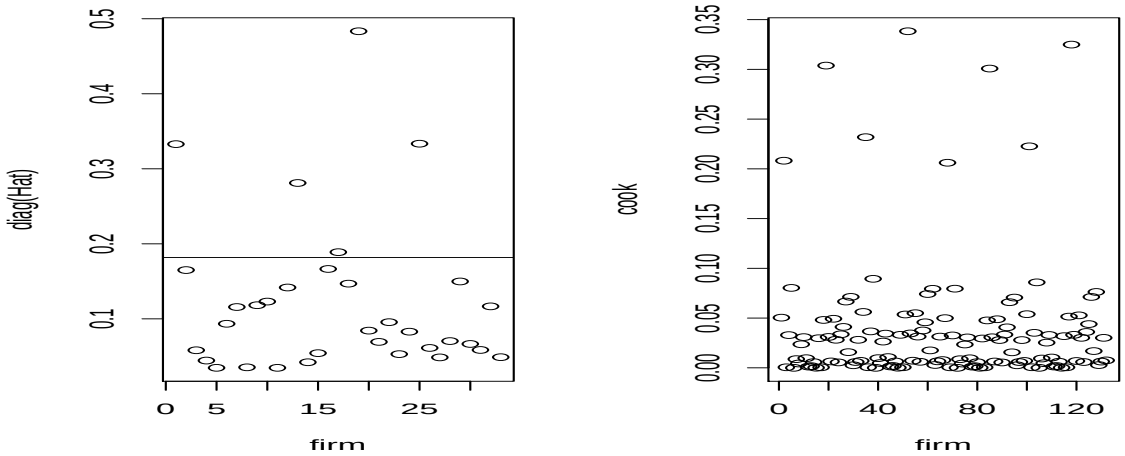


Figure 2: Left: h_{ii} values, the horizontal line correspond to $2p/n$. Right: Cook's distance.

A high-leverage point would have diagonal element $h_{ii} > 6/33$. Figure 2 show that there are 5 high-leverage points

4 Confidence Intervals

4.1 Confidence Intervals on the individual regression coefficients

We showed in section 2.1 that

$$\frac{(\hat{\beta}_j - \beta_j)}{\sqrt{v_{jj}\hat{s}^2}} \sim t_{n-p}$$

Therefore a $100(1-\alpha)\%$ confident interval for β_j is

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{v_{jj}\hat{s}^2} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{v_{jj}\hat{s}^2}$$

In Example 1, we obtain the C.I.s for the parameters as:

```
confint(fit1, level=0.95)
(Intercept) -504.17657 194.04147
x1          -107.23631 131.98495
x2           -22.95738 157.21613
x3           -55.99607  80.60511
```

4.2 Joint confidence region on the regression coefficients

In some cases it is necessary to construct a confidence interval that applies to the entire set of parameters. Such intervals are called **simultaneous confidence intervals**. We use the fact that $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$, then:

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} \sim \chi^2_p$$

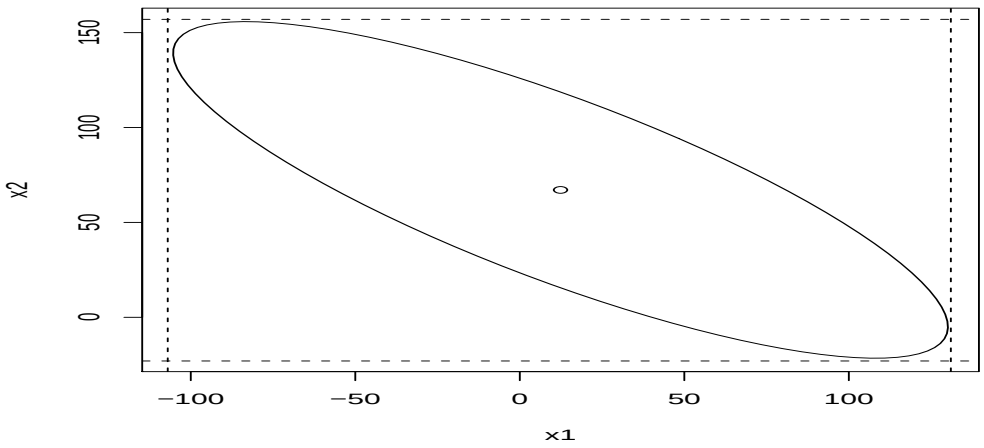
Using the fact that $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ are independent and $\hat{s}^2 = S(\hat{\boldsymbol{\beta}})/n-p$ satisfies $(n-p)\hat{s}^2/\sigma^2 \sim \chi_{n-p}$ (as we saw in section 2.2), then,

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{p\hat{s}^2} \sim F_{p,n-p}. \tag{1.3}$$

Therefore, a $100(1-\alpha)\%$ joint confidence region for all parameters in $\boldsymbol{\beta}$ will contains the values of $\boldsymbol{\beta}$ such that

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{p\hat{s}^2} \leq F_{\alpha,p,n-p}$$

This inequality describes an elliptically shaped region.



The figure in the next page shows the confidence region and confidence intervals for $\hat{\boldsymbol{\beta}} = (\beta_1, \beta_2)'$ in Example 1. Note that all points in the confidence region lie inside the CI's but not vice versa.

5 Inference on the response variable and Prediction

One of the objectives of fitting a regression model is to use the model for prediction. We may have to different aims:

1. Estimate the mean of the distribution $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}_h$: $E[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}_h] = \mu_{y|\boldsymbol{x}_h}$
2. Predict the value of the response variable of an individual from the population for whom we know, $\boldsymbol{X} = \boldsymbol{x}_h$, i.e., we want to predict the value of $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}_h$.

5.1 Estimation of the mean response

Suppose we are interested on the mean response at a particular point $\boldsymbol{x}_h$

$$\boldsymbol{x}_h = \begin{bmatrix} 1 \\ x_{h1} \\ x_{h2} \\ \vdots \\ x_{hk} \end{bmatrix}$$

the mean response at this point is

$$\mu_{y|\boldsymbol{x}_h} = \beta_0 + \beta_1 x_{h1} + \ldots + \beta_k x_{hk} = \boldsymbol{x}_h'\boldsymbol{\beta}$$

An unbiased estimator of the mean response is

$$\hat{y}(\boldsymbol{x}_h) = \boldsymbol{x}_h'\hat{\boldsymbol{\beta}}$$

it is unbiased since $E[\hat{y}(\boldsymbol{x}_h)] = \boldsymbol{x}_h' E[\hat{\boldsymbol{\beta}}] = \boldsymbol{x}_h'\boldsymbol{\beta} = \mu_{y|\boldsymbol{x}_h}$, and the variance is

$$Var[\hat{y}(\boldsymbol{x}_h)] = \sigma^2 \boldsymbol{x}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_h$$

Therefore, a $100(1-\alpha)\%$ confidence interval on the mean response at the points $\boldsymbol{x}_h$ is

$$\hat{y}(\boldsymbol{x}_h) \pm t_{\alpha/2,n-p}\sqrt{\hat{s}^2 \boldsymbol{x}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_h}$$

5.2 Prediction of new response observations

A common use of regression models is to predict the value of the response for given values of the explanatory variables. But, in this case, we have to take into account the randomness of the response variable. Suppose we are interested in predicting the response variable for a given vector $\boldsymbol{x}_h$ of values of the explanatory variables. The value of the response variable $\boldsymbol{y}_h = \boldsymbol{x}_h'\hat{\boldsymbol{\beta}} + \varepsilon$ predicted by the model, is given by

$$\hat{\boldsymbol{y}}_h = \boldsymbol{x}_h'\hat{\boldsymbol{\beta}} = \hat{\mu}_h$$

In order to construct a confidence interval for a predictor (also called *prediction interval*) we need the distribution of $\hat{\boldsymbol{y}}_h$,

$$\begin{aligned} E[\hat{\boldsymbol{y}}_h] &= \boldsymbol{x}_h' E[\hat{\boldsymbol{\beta}}] = \boldsymbol{x}_h' \boldsymbol{\beta} \\ Var[\boldsymbol{y}_h] &= Var[\boldsymbol{x}_h' \hat{\boldsymbol{\beta}}] + Var[\varepsilon] = \sigma^2 \boldsymbol{x}_h' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_h + \sigma^2 = \sigma^2 (1 + \boldsymbol{x}_h' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_h) \end{aligned}$$

the prediction interval is given by:

$$\hat{y}(\boldsymbol{x}_h) \pm t_{\alpha/2, n-p} \sqrt{\hat{s}^2 (1 + \boldsymbol{x}_h' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_h)}$$

The prediction interval is similar to the confidence interval for the mean response but wider. This is not surprising since in the C.I. for the mean response we take into account the variation from the estimators of the regression parameters, and in a prediction C.I. we take into account *both* the variation coming from the estimators of $\boldsymbol{\beta}$, and also the variation from the error term $\boldsymbol{\varepsilon}$ (since we are prediction a value of a random variable)

Example 1

If we wanted to give a confidence interval for the mean compensation in firms with $x1 = 8.43$, $x_2 = 4.85$ y $x_3 = 10.77$:

```
predict.lm(fit1,interval="confidence")[1,]
     fit      lwr      upr
407.6928 298.0220 517.3635
```

and the confidence interval for a new observation with the same values of the predictor variables:

```
predict.lm(fit1,interval="prediction")[1,]
     fit      lwr      upr
407.6928 188.2433 627.1423
```

Note that the projection is the same, but the C.I. is wider in the second case.

6 Hypothesis testing

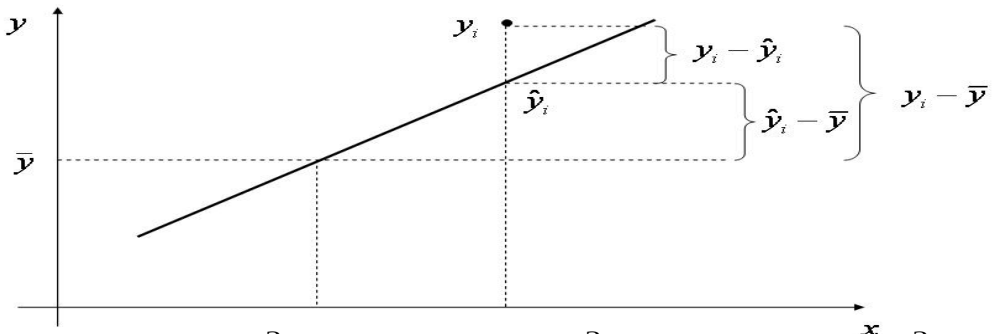
6.1 Test for significance of regression

This test is used to determine if there is a linear relationship between the response variable $\boldsymbol{y}$ and the regressor variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ (see that here $p = k + 1$). The hypotheses are,

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \\ H_1 &: \beta_j \neq 0 \quad \text{at least for one } j \end{aligned}$$

The test is based on the following decomposition of the variability in different sources of variation in the data:

We cannot compare SS_R y SS_E directly, since we do not know their distribution, however we know that:



- $SS_R/\sigma^2 \sim \chi^2_{n-p}$

- if $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$ is true $SS_T/\sigma^2 \sim \chi^2_{n-1}$
- $SS_E/\sigma^2 = \underbrace{SS_T/\sigma^2 - SS_R/\sigma^2}_{independent} \sim \chi^2_{n-1} - \chi_{n-p} \equiv \chi^2_{p-1}$

Therefore,

$$\frac{SS_E/(p-1)}{SS_R/(n-p)} = \frac{MS_E}{\hat{s}^2} \sim F_{p-1,n-p}$$

This can be summarized in an **ANOVA table**:

Source	df	SS	MS	F
Model	$p-1$	$SS_E = \sum_i (\hat{y}_i - \overline{y})^2$	$MSR = SS_E/(p-1)$	$MS_E/\hat{s}^2$
Residual	$n-p$	$SS_R = \sum_i (\hat{y}_i - y)^2$	$\hat{s}^2 = SS_R/(n-p)$	
Total	$n-1$	$SS_T = \sum_i (y_i - \overline{y})^2$	$MS_T = SS_T/(n-1)$	

The null hypothesis is rejected if $\frac{MS_R}{\hat{s}^2} > F_{\alpha,p-1,n-p}$.

Rejecting H_0 means that at least one regression coefficient is non-zero, and hence that at least one of the explanatory variables is useful in predicting the response.

Coefficient of multiple determination

This coefficient gives the proportion of variation in response variable explained by the explanatory variables:

$$R^2 = SS_E/SS_T = 1 - SS_R/SS_T \Rightarrow F = \frac{R^2/(p-1)}{(1-R^2)/(n-p)}$$

We see that $0 \leq R^2 \leq 1$. However, a large value of R^2 does not necessary imply that the model regression is good. Adding a variable to the model will not decrease R^2 regardless of

whether the additional variable is statistically significant or not. A solution is to work with **adjusted** R^2 defined as:

$$R^2_{adj} = 1 - \frac{SS_R/(n-p)}{SS_T/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2)$$

6.2 Tests on individual regression coefficients

The hypotheses test for testing significance of any individual regression coefficient β_j are

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0 \end{aligned}$$

Since $\hat{\beta}_j \sim N(\beta_j, \sigma^2 v_{jj})$ (where v_{jj} are the diagonal elements of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$), if $\beta_j = 0$ is not rejected, then this indicates that x_j can be deleted from the model. The test statistic for this Hypothesis is

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{s^2 v_{jj}}}$$

Therefore, the null hypothesis $H_0 : \beta_j = 0$ is rejected if $|t_0| > t_{\alpha/2,n-p}$.

Example 1

The command

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -155.07     170.69  -0.908    0.371
x1             12.37      58.48   0.212    0.834
x2             67.13      44.05   1.524    0.138
x3             12.30      33.40   0.368    0.715

Residual standard error: 92.94 on 29 degrees of freedom
Multiple R-Squared: 0.497,      Adjusted R-squared: 0.4449
F-statistic:  9.55 on 3 and 29 DF,  p-value: 0.0001510
```

gives the value of the adjusted $R^2 = 0.4449$ (which indicates that the model only explain 44.5% of the variability in the data). The test for significance of regression gives a $p-value = 0.0001510$ which indicates that we reject the hypothesis that all parameters are equal to zero. However, the test on the individual parameters lead to the conclusion that there is no evidence to suppose that they are different from zero, do you see the contradiction?.

6.3 Tests on groups of coefficients

We might be interested on investigating the contribution of a subset of regressor variables to the model. Consider the model con k regressor variables ($p = k + 1$), $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

We would like to determine if the subset of variables $\boldsymbol{x}_{r+1}, \boldsymbol{x}_{r+2}, \ldots, \boldsymbol{x}_k$ $(r < k)$ contribute significantly to the model. Let the vector of regression coefficients be partitioned as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

where $\boldsymbol{\beta}_1$ is $r \times 1$ and $\boldsymbol{\beta}_2$ is $(p-r) \times 1$. We wish to test:

$$\begin{aligned} H_0 &: \boldsymbol{\beta}_2 = \mathbf{0} \\ H_1 &: \boldsymbol{\beta}_2 \neq \mathbf{0} \end{aligned}$$

The regression sum of squares SS_E of the full model (including all variables) is decomposed as:

$$SS_E(\hat{\boldsymbol{\beta}}) = SS_E(\hat{\boldsymbol{\beta}}_1) + SS_E(\hat{\boldsymbol{\beta}}_2|\hat{\boldsymbol{\beta}}_1)$$

where $SS_E(\hat{\boldsymbol{\beta}}_1)$ is the sum of squares explained by the reduced model with r degrees of freedom; and $SS_E(\hat{\boldsymbol{\beta}}_2|\hat{\boldsymbol{\beta}}_1)$ is the sum of squares explained by $\boldsymbol{\beta}_2$ given that $\boldsymbol{\beta}_1$ is already in the model, and it has $p-r$ degrees of freedom. It is the **extra sum of squares** due to including $\boldsymbol{x}_{r+1}, \ldots, \boldsymbol{x}_k$ in the model. Now $SS_E(\hat{\boldsymbol{\beta}}_2|\hat{\boldsymbol{\beta}}_1)$ is independent of $MS_R(\hat{\boldsymbol{\beta}})$, and $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ may be tested by the statistic

$$F_0 = \frac{SS_E(\hat{\boldsymbol{\beta}}_2|\hat{\boldsymbol{\beta}}_1)/(p-r)}{\hat{s}^2}$$

H_0 will be rejected if $F_0 > F_{\alpha,p-r,n-p}$. Some authors call this test, a **partial** F test.

Example 1

Suppose that we want to test $H_0 : \beta_2 = \beta_3 = 0$:

```
fit2<-lm(y~x1,data=reg1)
anova(fit2,fit1)
Analysis of Variance Table

Model 1: y ~ x1
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     31 270552
2     29 250488  2     20064 1.1614 0.3272
```

where the value of the test statistic $F = 1.1614 = \frac{SS_E(\hat{\beta}_2,\hat{\beta}_3|\hat{\beta}_0,\hat{\beta}_1)/r}{\hat{s}^2} = \frac{20064/2}{92.94^2}$, and the conclusion is that there is not enough evidence to include x_2 and x_3 in the model.

Model selection

To determine an appropriate subset of explanatory variables, there are several criteria available:

1. Choose models with high **adjusted** R^2

2. **Mallow's C_p criterion**. The idea is to compare subset models with the full model, and it is a measure of the total mean square error for the regression model. We define the total standardized mean squared error for the regression model as

$$\begin{aligned}\Gamma_p &= \frac{1}{\sigma^2}\sum_{i=1}^{n} E[\hat{y}_i - E(y_i)]^2 \\ &= \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}[E(y_i) - E(\hat{y}_i)]^2 + \sum_{i=1}^{n} V(\hat{y}_i)\right\} \\ &= \frac{1}{\sigma^2}[(\text{bias})^2 + \text{variance}]\end{aligned}$$

Then an estimate of Γ_p is

$$C_p = \frac{SS_R(\hat{\boldsymbol{\beta}}_1)}{\hat{s}^2(\hat{\boldsymbol{\beta}})} - (n - 2r)$$

if the model with p terms has zero bias, $E[C_p|\text{zero bias}] = p$. Therefore, the values of C_p for each model under consideration should be evaluated relative to p and choose model with $C_p \leq p$

3. Other criteria are AIC (Akaike Information Criteria), BIC (Bayesian Information Criteria), etc., (see Akaike (1973) among others).

7 Multicollinearity

It is a frequent problem when there are several explanatory variables, it appears when two or more explanatory variables are highly correlated. The consequence is that it is difficult to separate the effects of the different variables and to measure the individual contribution of each one to the model. It is also a numerical analysis problem, since the dependence among the variables will make the matrix $\boldsymbol{X'X}$ close to singular and it will be difficult to invert it in order to calculate the variance of $\hat{\boldsymbol{\beta}}$. Therefore, the regression coefficients are not well estimated, and might be meaningless, and similarly for the standard errors of these estimates.

One of the effects of multicollinearity is to inflate the estimated variance of $\hat{\beta}_j$. The variance of $\hat{\boldsymbol{\beta}}$ depends on the matrix $(\boldsymbol{X'X})^{-1}$, this is a symmetric matrix with p columns, the singular value decomposition of $\boldsymbol{X'X}$ is given by

$$\boldsymbol{X'X} = \boldsymbol{P'DP} = \sum_{i=1}^{p} \lambda_i p_i p_i'$$

where $\boldsymbol{P} = [p_1 : \ldots : p_p]$ is an orthogonal matrix with columns p_i. If two o more variables are highly correlated there will be one or more λ_i close to zero.
It is easy to show that the singular value decomposition of $(\boldsymbol{X'X})^{-1} = \sum_{i=1}^{p} \lambda_i^{-1} p_i p_i'$, and so one or more values of λ_i^{-1} will be large, thus, some values of $Var(\hat{\beta}_j) = v_{jj}s^2$ (as defined in section 2.2) will be inflated, and so inference on the regression coefficients might be wrong. However, multicollinearity does not affect the fitted values, R^2 or F tests.

7.1 How to identify multicollinearity

1. High correlation between explanatory variables

2. Variables are significant in simple regression, but not in multiple regression

3. Variance Inflation Factor (VIF) (see Silvey (1969)):
   It can be shown that
   $$Var(\hat{\beta}_j) = \sigma^2 \frac{1}{n s_j^2 (1 - R_j^2)}$$
   where R_j^2 is the multiple coefficient of determination that would be calculated by running a regression analysis using the model
   $$x_j = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_{j-1} x_{j-1} + \alpha_{j+1} x_{j+1} + \ldots + \alpha_k x_k$$
   The variance inflation factor is
   $$VIF_j = \frac{1}{1 - R_j^2} \Rightarrow Var(\hat{\beta}_j) = \sigma^2 \frac{VIF_j}{n s_j^2}$$
   If x_j does not depend on the other variables R_j^2 will be close to 0 and VIF_j close to 1, and the stronger the dependence, the close is R_j^2 to and and the larger id FIV_j. Most authors consider that if $VIF > 10$ there is a problem of multicollinearity

4. Condition Index:
   It is defined as
   $$\kappa = \sqrt{\frac{\text{Largest eigenvalue of } \boldsymbol{X'X}}{\text{Smallest eigenvalue of } \boldsymbol{X'X}}}$$
   If variables are dependent, there will be eigenvalues close to zero and the condition index will be large.
   $$\begin{aligned} 10 \leq \kappa \leq 30 &\Rightarrow \text{moderate multicollinearity} \\ \kappa > 30 &\Rightarrow \text{severe multicollinearity} \end{aligned}$$

A possible remedial measure is to use a method of estimation of the parameters less sensible to multicollinearity, an alternative is **ridge regression**, where the parameter estimates are obtained from
$$\boldsymbol{\beta}^*(\lambda) = (\boldsymbol{X}\boldsymbol{X}' + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}'\boldsymbol{y}$$
where $\lambda \geq 0$ (generally $0 \leq \lambda \leq 1$). The ridge estimator $\boldsymbol{\beta}^*(\lambda)$ is not unbiased, but the objective is to find a set of coefficients that are more stable (i.e., they have a small mean square error). For each least squares problem there is an *optimum* value of λ, but generally a value of λ in $0 \leq \lambda \leq 1$ is enough. Usually, the variance of $\boldsymbol{\beta}^*(\lambda)$ is a decreasing function of λ, while the squared bias $[\boldsymbol{\beta} - E(\boldsymbol{\beta}^*(\lambda))]^2$ is an increasing function of λ. Therefore, choosing the value of λ involves trading off these two properties of $\boldsymbol{\beta}^*(\lambda)$.

Example 1

We have seen already, several signs of the existence of multicollinearity in this example: The pairwise plot showed a strong relationship between the explanatory variables, and we also noticed the contradiction between the test for significance regression and the test for individual coefficients. The variance inflation factor for the model parameters are:

```
fv1<-1/(1-summary(lm(reg1$x1~reg1$x2+reg1$x3))$r.squared)
fv2<-1/(1-summary(lm(reg1$x2~reg1$x1+reg1$x3))$r.squared)
fv3<-1/(1-summary(lm(reg1$x3~reg1$x1+reg1$x2))$r.squared)
fv1
[1] 12.61695
fv2
[1] 6.934045
fv3
[1] 4.552285
```

and the condition index:

```
eig.values<-eigen(t(X)%*%X)$values
sqrt(max(eig.values)/min(eig.values))
[1] 147.5221
```

Both criteria indicate multicollinearity. What to do now?, we could drop one or more variables form the model, to select which variables to include we fit all possible models with one and two explanatory variables and choose the simplest model with the highest value of R^2. Another possibility is to use ridge regression, Figure 3 shows a plot of the estimates of $\hat{\beta}_i$ $i = 1, 2, 3$ for different values of λ, which value would you use?

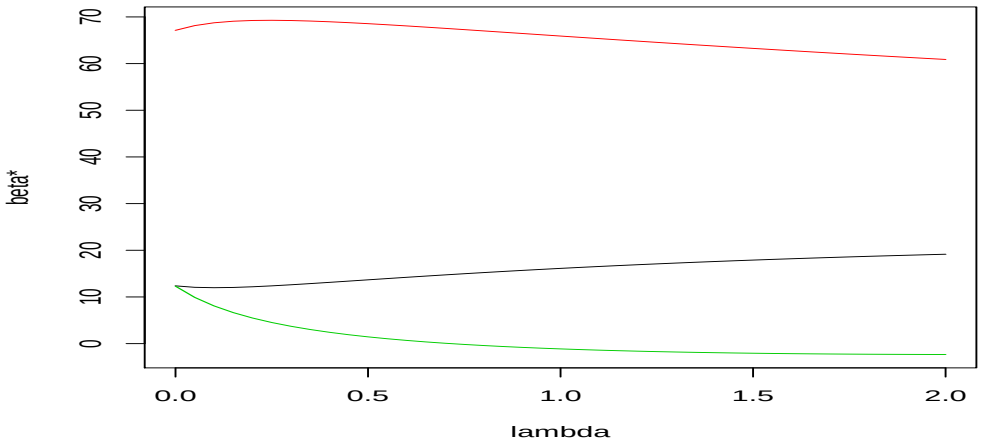


Figure 3: Plot of coefficients versus λ

Bibliography

Akaike, H. (1973). Information theory and an extention of the maximum likelihood principle. In Petrof, B. and Csàki, editors, *Second International Symposium on Information*, pages 267–281, Akademia Kiadó, Budapest.

Cook, R. (1977). Detection of influential observations in linear regression. *Technometrics*, 19:15–18.

Cook, R. and Weisberg, S. (1982). *Residuals and influence in regression.* Monographs on Statistics and Applied Probability. Chapman and Hall.

Silvey, S. (1969). Maulticollinearity and imprecise estimation. *Journal of the Royal Statistical Society, B.*