## 4.7 Confidence and Prediction Intervals

Instead of conducting tests we could find confidence intervals for a regression coefficient, or a set of regression coefficient, or for the mean of the response given values for the regressors, and sometimes we are interested in the range of possible values for the response given values for the regressors, leading to prediction intervals.

### 4.7.1 Confidence Intervals for regression coefficients

We already proved that in the multiple linear regression model $\hat{\beta}$ is multivariate normal with mean $\vec{\beta}$ and covariance matrix $\sigma^2(X'X)^{-1}$, then is is also true that

**Lemma 1.**
For $0 \leq i \leq k$ $\hat{\beta}_i$ is normally distributed with mean $\beta_i$ and variance $\sigma\sqrt{C_{ii}}$, if $C_{ii}$ is the $i-th$ diagonal entry of $(X'X)^{-1}$.

Then confidence interval can be easily found as

**Lemma 2.**
For $0 \leq i \leq k$ a $(1-\alpha) \times 100\%$ confidence interval for $\beta_i$ is given by

$$\hat{\beta}_i \pm t^{n-p}_{\alpha/2}\hat{\sigma}\sqrt{C_{ii}}$$

**Continue Example. 3.1**
To find a 95% confidence interval for the regression coefficient first find $t^4_{0.025} = 2.776$, then

$$2.38 \pm 2.776(1.388)\sqrt{0.018} \leftrightarrow 2.38 \pm 0.517$$

The 95% confidence interval for $\beta_2$ is [1.863, 2.897]. We are 95% confident that on average the blood pressure increases between 1.86 and 2.90 for every extra kilogram in weight after correcting for age.

**Joint confidence intervals for a set of regression coefficients.**
Instead of calculating individual confidence intervals for a number of regression coefficients one can find a joint confidence region for two or more regression coefficients at once. The result is stronger since in this case the confidence level applies to all coefficients at once instead individually.

**Lemma 3.**
A $(1-\alpha)\%$ confidence region for the $\vec{\beta}_1 \in \mathbb{R}^r$ from the partitioned regression coefficient vector $\vec{\beta} = (\vec{\beta}_1{}', \vec{\beta}_2{}')'$ is given by

$$\frac{(\hat{\beta}_1 - \vec{\beta}_1)'(X_1'X_1)(\hat{\beta}_1 - \vec{\beta}_1)}{r\ MS_{Res}} \leq F_\alpha(r, n-p)$$

**Proof:** The result is due to the fact that

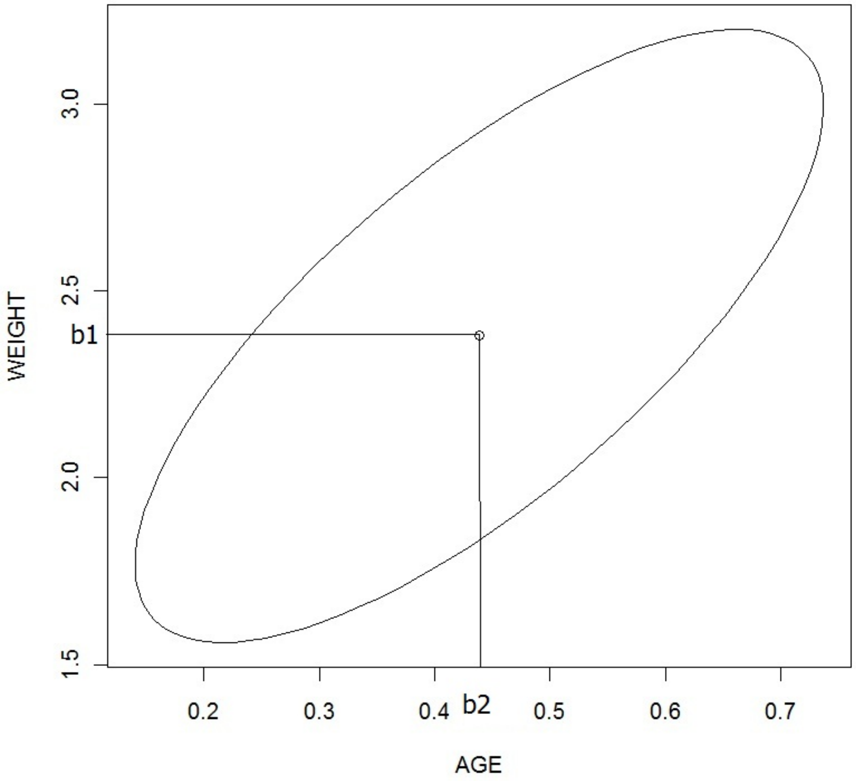
$$\frac{(\hat{\beta}_1 - \vec{\beta}_1)'(X_1'X_1)(\hat{\beta}_1 - \vec{\beta}_1)}{r\ MS_{Res}}$$

is F distributed with $df_n = r$ and $df_d = n - p$.

For two dimensions this is an ellipse, but for three parameters it would be a 3-dimensional ellipsoid.

**Continue Example. 3.1** The following graph shows the 95% confidence region for $\vec{\beta} = (\beta_1, \beta_2)'$ from the model

$$bp = \beta_0 + \beta_1(weight) + \beta_2(age) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$



The graph was created with R using the following commands.

```
library(ellipse)
BP.data<-read.table("BP.txt",header=T)
attach(BP.data)

fit<-lm(BP~AGE+WEIGHT)
plot(ellipse(fit, which = c('AGE', 'WEIGHT'), level = 0.95), type = 'l')
points(fit$coefficients['AGE'], fit$coefficients['WEIGHT'])
```

### 4.7.2 Confidence interval for the mean response and prediction interval

To find a confidence interval for $E(Y|\vec{x}_0)$ the mean response given the regressors variables have values $\vec{x}_0$ observe that a point estimate is given by

$$\hat{Y}_0 = \vec{x}_0{}'\hat{\beta}$$

It is unbiased and has variance $\sigma^2\vec{x}_0{}'(X'X)^{-1}\vec{x}_0$ (Proof!!). Because it is also normal a confidence interval is derived as

**Lemma 4.**
In the MLRM a $(1-\alpha)\times 100\%$ confidence interval for $E(Y|\vec{x}_0)$ is given by

$$\hat{Y}_0 \pm t^{n-p}_{\alpha/2}\ \hat{\sigma}\ \sqrt{\vec{x}_0{}'(X'X)^{-1}\vec{x}_0}$$

**Continue Example. 3.1**
Find a 95% confidence interval for the mean blood pressure of people who are 35 years old and weigh 70kg, i.e. $\vec{x}_0 = (1, 70, 35)'$. $t^4_{0.025} = 2.776$, and

$$\hat{y}_0 = (1, 70, 35)\begin{pmatrix} -66.13 \\ 2.38 \\ 0.44 \end{pmatrix} = 115.79 \text{ and}$$

$$\vec{x}_0{}'(X'X)^{-1}\vec{x}_0 = (1, 70, 35)\begin{pmatrix} 129.777 & -1.534 & -0.452 \\ -1.534 & 0.018 & 0.005 \\ -0.452 & 0.005 & 0.002 \end{pmatrix}\begin{pmatrix} 1 \\ 70 \\ 35 \end{pmatrix} = 0.495$$

then the confidence interval is given by

$$115.79 \pm 2.776(1.388)\sqrt{0.495} \leftrightarrow 115.79 \pm 2.711 \leftrightarrow [113.08, 118.50]$$

We are 95% confident that the mean blood pressure of people who are 35 years old and weigh 70kg falls between 113 mmHg and 118.5 mmHg.

To find a prediction interval for $Y_0 = Y|\vec{x} = \vec{x}_0$ the distribution of

$$\hat{Y}_0 - Y_0$$

has to be considered. Compare to the approach taken for the SLRM. The mean of $\hat{Y}_0 - Y_0$ is 0, and the variance is $\sigma^2\ (1 + \vec{x}_0{}'(X'X)^{-1}\vec{x}_0)$. In addition it is normal as a linear combination of normal random variables. Therefore:

**Lemma 5.**
A $(1-\alpha)\times 100\%$ prediction interval for $Y_0 = Y|\vec{x} = \vec{x}_0$ is given by

$$\hat{Y}_0 \pm t^{n-p}_{\alpha/2}\ \hat{\sigma}\ \sqrt{1 + \vec{x}_0{}'(X'X)^{-1}\vec{x}_0}$$

**Continue Example. 3.1**
Find a 95% prediction interval for the blood pressure of people who are 35 years old and weigh 70 kg, i.e. $\vec{x}_0 = (1, 70, 35)'$. Using the results from above

$$115.79 \pm 2.776(1.388)\sqrt{1 + 0.495} \leftrightarrow 115.79 \pm 4.711 \leftrightarrow [111.08, 120.50]$$

We predict that 95% of people who are 35 years old and weigh 70 kg mean have a blood pressure between 111 and 120.5.

## 4.8 Standardized Regression

In some studies the strength of influence of different regressors on the mean response shall be compared. Since the regressors are usually measured on different scales, this is impossible from the MLRM.
In order to make coefficients comparable the standardized regression coefficients can be used.
The standardized regression coefficients are the least squares estimators in the model, fitting response variable

$$Y_i^* = \frac{Y_i - \bar{Y}}{s_Y},\ 1 \leq i \leq n$$

with regressors

$$z_{ij} = \frac{x_{ij} - \bar{x_j}}{s_j},\ 1 \leq i \leq n, 1 \leq j \leq k$$

where $s_j^2$ is the sample variance of $x_{ij},\ 1 \leq i \leq n$ and $s_Y^2$ is the sample variance of $Y_i,\ 1 \leq i \leq n$.
The model is

$$\vec{Y}^* = b_0 + b_1\vec{z}_1 + b_2\vec{z}_2 + \cdots + b_k\vec{z}_k\varepsilon,\quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

With $b_0 = 0$, because it gives the mean of $Y^*$, if all the $z$ are 0. Since the mean of $Y^* = 0$, we get $b_0 = 0$.
The least squares estimates $\hat{b}$ are

$$\hat{b} = (Z'Z)^{-1}Z'Y^*$$

**Theorem 4.1.**
The least squares estimate $\hat{\beta}$ can be found from the standardized regression coefficients as

$$\hat{\beta}_j = \hat{b}_j\ \frac{s_Y}{s_j} = \hat{b}_j \left(\frac{SS_T}{SS_{jj}}\right)^2,\ 1 \leq j \leq k$$

and

$$\hat{\beta}_0 = \bar{Y} - \sum_{i=1}^{k} \hat{b}_j \bar{x}_j$$

where $s_Y$ and $s_j$ are the standard deviations of $Y$ and $x_j$, respectively, and

$$SS_{jj} = SS_{x_j} = \sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2 = \sum_{i=1}^{n} x_{ij}^2 - \frac{(\sum_{i=1}^{n} x_{ij})^2}{n} = \vec{x}_j{}'\vec{x}_j - \vec{x}_j{}'J_n\vec{x}_j$$

The least squares estimators for this model are called the standardized regression coefficients or beta coefficients and are usually denoted by $\beta$. They measure the change in the response (measured in standard deviations) for an increase in $x_j$ by one standard deviation when leaving the other regressors unchanged. Because the slopes are now measuring the change in $y$ based on standardized increases in the regressors they are comparable and the larger absolute value of the standardized regression coefficient the stronger its effect on the response.
$\hat{b}_0$ for this model is 0, because of the centering of both response and regressors.

**Continue Example. 3.1**
Using the following R-code the standardized regression coefficients are

```
y<-BP.data$BP
y<-(y-mean(y))/sd(y)
(y<-t(t(y)))

x1<-BP.data$WEIGHT
x1<-(x1-mean(x1))/sd(x1)
(x1<-t(t(x1)))
x2<-BP.data$AGE
x2<-(x2-mean(x2))/sd(x2)
(x2<-t(t(x2)))
(X<-cbind(x1,x2))

(beta<-solve(t(X)%*%X)%*%t(X)%*%y)
```

The result

```
          [,1]
[1,] 1.4029059
[2,] 0.7115687
```

So that $\hat{b}_1 = 1.40, \hat{b}_2 = 0.71$ and $\hat{b}_0 = 0$.
Interpretation: On average the blood pressure increases 1.4 standard deviations, when increasing the weight by one standard deviation and keeping the age unchanged, and on average the blood pressure increases 0.7 standard deviations, when increasing the age by one standard deviation and keeping the weight unchanged. According to this data changes in weight have a stronger effect on the blood pressure than changes in age.

## 4.9 Multicollinearity

When the columns of $X$ are linearly dependent then $X$ does not have full rank, therefore $X'X$ does not have full rank and its inverse does not exist and the determinant is 0, which indicates that the normal equation has infinite many solutions.
Should this occur in a practical situation, it indicates that at least one of the regressors is redundant and should be removed from the model.

Multicollinearity refers to the problem when the columns are near-linear dependent, or the determinant of $X'X$ is close to zero. The consequence is that the entries of the covariance matrix of $\hat{\beta}$, which is $\sigma^2(X'X)^{-1}$ are large, resulting in wide confidence intervals for the regression coefficients, indicating unprecise estimates for the regression coefficient resulting poor confidence intervals for the mean response and prediction intervals.

A measure to detect multicollinearity is the determinant of the $(X'X)^{-1}$, or the diagonal elements of the inverse of the correlation matrix, which is standardized and easier to judge than the covariance matrix. If the diagonal entries, called variance inflation factors, exceed 10, they are considered critical. They can also be calculated by

$$VIF_j = \frac{1}{1 - R_j^2}$$

where $R_j^2$ is the coefficient of determination when regressing $x_j$ on the other regressor variables. This definition makes intuitive sense, since $VIF_j$ is large if $R_j^2$ is close to one, which indicate that the variance in $x_j$ can be explained by the other regressors.

**Continue Example.** For our example the inverse of the correlation matrix is

$$\begin{pmatrix} 2.26 & 1.69 \\ 1.69 & 2.26 \end{pmatrix}$$

Indicating that we do not have to be concerned about multicollinearity.

## 4.10 Wrong sign of regression coefficient estimates

Sometimes from theory the sign of a regressor can be deduced (think age and hight of children), but a when estimating the slopes the "wrong" sign arises.
Possible reasons

1. The range of the regressor is too small (only include children of age 5 and 6)
2. Important regressors have not been included in the model (sex is omitted and only older girls and younger boys are observed resulting in a negative slope)
3. multicollinearity (as discussed the standard errors are large and the estimates will not be precise)
4. computational errors (well. . . )