Multiple linear regression: (with K predictors)

Dependent data set: $\left\{ \begin{matrix} y_i & i=1,...n \\ x_{i,k} & i=1,...n, \quad k=1...,K \end{matrix} \right\}$ one predictand, K predictors

The forecast equation is $\hat{y}_i = b_0 + \sum_{k=1}^{K} b_k x_{ik}$

Use matrix notation:

$$Y = \begin{bmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1K} \\ 1 & x_{21} & x_{22} & ... & x_{2K} \\ ... & ... & x_{ik} & ... & ... \\ 1 & x_{n1} & x_{n2} & ... & x_{nK} \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & x_{12} & ... & x_{1K} \\ x_{20} & x_{21} & x_{22} & ... & x_{2K} \\ ... & ... & x_{ik} & ... & ... \\ x_{n0} & x_{n1} & x_{n2} & ... & x_{nK} \end{bmatrix}$$

The regression coefficients are $B = \begin{bmatrix} b_0 \\ . \\ . \\ . \\ b_K \end{bmatrix}$, the forecast equation is $\hat{Y} = XB$, and

the forecast error vector $E = \begin{bmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$ is given by $E = Y - XB$.

Least squares approach: choose B to minimize $|E|^2 = E^T E = SSE$.

The total sum of squares (variance) of Y is $SST = \sum_{i=1}^{n} y_i'^2 = Y'^T Y'$.

The residual (forecast errors) sum of squares is

$$SSE = E^T E = (Y - XB)^T (Y - XB) = Y^T Y - Y^T XB - B^T X^T Y + B^T X^T XB$$

(all terms are scalars).The minimization is clearer in sum notation:

$$SSE = |E|^2 = \sum_{i=1}^{n}\left(y_i - \sum_{l=0}^{K} b_l x_{il} \right)^2$$

The minimization gives the "normal equations":

$$\frac{\partial SSE}{\partial b_k} = 0 = -2\sum_{i=1}^{n}\left(y_i - \sum_{l=0}^{K} b_l x_{il} \right) x_{ik} =$$

$$= -2\sum_{i=1}^{n}\left(\underbrace{x_{ki}^T y_i}_{X^T Y} - \underbrace{x_{ki}^T \sum_{l=0}^{K} b_l x_{il}}_{X^T XB} \right) = 0, \qquad k = 0,1,...,K$$

So, in matrix form, the normal equations for the coefficients are

$$X^T XB = X^T Y \quad \text{or} \quad \boxed{B = (X^T X)^{-1} X^T Y}$$

Again, we separate the total sum of squares SST into the regression (explained) sum of squares (SSR) and the error sum of squares (SSE).

$$SST = Y'^T Y' \qquad Y' = Y - \bar{Y}$$

$$SSE = (Y - XB)^T (Y - XB) = Y^T Y - B^T X^T Y - Y^T XB + B^T X^T XB =$$

$$= Y^T Y - Y^T XB \qquad \text{since } X^T XB = X^T Y$$

Since these are all scalars, they are the same as their transpose:

$$Y^T XB = (Y^T XB)^T = B^T X^T Y$$

so that

$$SSE = Y^T Y - Y^T XB = (Y^T - B^T X^T)Y = E^T Y$$

(the scalar product of the error and the predictand).

Note (prove) that $E^T \hat{Y} = 0$, the error is orthogonal to the forecast).

$$SSR = SST - SSE = R^2 SST$$

$R^2 = \dfrac{SST - SSE}{SST}$ is the square of the generalized correlation coefficient or explained variance.

This is an estimate for the dependent sample used for training the regression, so that it is overoptimistic! In an independent sample, the explained variance is smaller.

Note that the "naïve" forecast error variance $\overline{\varepsilon^2} = \frac{1}{n-1}\sum_{i=1}^{n}\varepsilon_i^2 = \frac{SSE}{n-1}$ is seriously biased (overoptimistic). This is for two reasons: a) we are using K predictors, and b) it is the estimate for the dependent (training) sample.

The unbiased estimate for the forecast error variance for the dependent sample is

$$s_\varepsilon^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-K-1} = \frac{SSE}{n-K-1}$$

This is because the SSE is estimated with *n-K-1* d.o.f.
SST is estimated with n-1 d.o.f (one d.o.f. was used to compute $\bar{y}$).
SSR uses K d.o.f. for the regression coefficients b_k, so that
SSE is left with *n-K-1* d.o.f.

It is clear that is we use too many predictors, i.e., if $K \sim O(n)$ we can have over fitting. If $K=n-1$, we can fit perfectly the dependent sample, so that the naïve dependent sum of errors squared is $SSE = 0$. However, the estimate of the dependent forecast error squared is in that case

$$s_\varepsilon^2 = \frac{SSE}{n-K-1} = \frac{0}{0}$$

So we should never over fit, and should keep a number of predictors such that $K<<n$.
Moreover, the regression coefficients (trained on the dependent sample) also have sampling errors (they are only estimates of the true regression coefficients).

Therefore when we apply the regression formula to new predictors x_0 (an independent data set), the standard deviation of the error is given by

$$\sqrt{\underbrace{\frac{SSE}{n-K-1}}\underbrace{\left(1+x_0^T(X^TX)^{-1}x_0\right)}}$$

uncertainty increases with the number of predictors *K*

uncertainty increases due to errors in the sampling of B when used with independent data

For *K=1* this correction for independent data is $1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \sim 1+\frac{2}{n}$ if x_0 is not far from $\bar{x}$.

Therefore the forecast error for independent data is given by the t distribution

$$\frac{x_0^TB-E(x_0^TB)}{\sqrt{\frac{SSE}{n-K-1}\left(1+x_0^T(X^TX)^{-1}x_0\right)}} \sim t_{n-K-1}$$

We can estimate a 100(1-a)% confidence interval for predicting $Y(x_0)$ as

$$Y(x_0)=x_0^TB \pm \sqrt{\frac{SSE}{n-K-1}\left(1+x_0^T(X^TX)^{-1}x_0\right)}t_{\frac{a}{2},n-K-1}$$

If a=5%, we look for values of $t_{2.5\%,n-K-1}$.

These prediction error estimates are only estimates. A better estimation of the error is to reserve part of the data for independent data testing (cross-validation).
For example, we use 90% of the data to obtain the regression coefficients, and test the forecast on the remaining 10%. In that case this can be repeated 10 times for different 10% subsets ("jackknifing"). This will give a good estimate of the forecast errors to be expected with an independent data set, as well as the error variance of the regression coefficients.

Statistical packages provide information not only about regression coefficients but also about their error estimates (using the formulas above) and about how significantly smaller is the forecast error. This is given in analysis of variance (ANOVA) and parameter error tables.

ANOVA

Source of variance	d.o.f.	Sum of Squares (SS)	Mean Square MS=SS/dof	F_{ratio} test statistic
Total	n-1	SST	SST/(n-1)	
Regression	K	SSR	SSR/K	MSR/MSE
Error (residual)	n-K-1	SSE	SSE(n-K-1)	

Regression Summary

Predictor	Coefficient	Standard error	t-ratio (n-K-1)
Constant	b_0	$\sqrt{s^2_{b_0}}$	$b_0 / \sqrt{s^2_{b_0}}$
x_k	b_k	$\sqrt{s^2_{b_k}}$	$b_k / \sqrt{s^2_{b_k}}$

If the F_{ratio} is large compared to $F_{.05,K,n-K-1}$ then we reject the null hypothesis that the coefficients b_k are really zero for the population and that $b_k \neq 0$ due to sampling.