Linear hypotheses in multiple regression

stats545.3 545.3 Hypothesis tests in the linear model, model building and predictions

Gunnar Stefansson

September 5, 2022

Null hypotheses, matrices and geometry

The null hypothesis, $H_i : \beta = 0$ in simple linear regression is a question of whether we can drop the variable x in $E[y_i] = \alpha + \beta x_i$, i.e. whether we can drop a column simplify X to

$$Z = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

or is the projection of y onto $span(Z)$ is "too much" farther away from y than the projection onto $span(X)$.

General null hypotheses are almost always concerned with how one can "reduce" or simplify the model, in this case usually whether one can reduce the number of columns in X or by some other means reduce the number of coefficients in the model.

Null hypothesis as matrices

Have $\underbrace{X}_{n \times p}$ and $\underbrace{Z}_{n \times q}$ s.t. $span(Z) \subseteq span(X)$.

Can estimate models

$$y = X\boldsymbol{\beta} + e_1$$

$$y = Z\boldsymbol{\gamma} + e_2$$

Will derive test for

$$H_0 : X\boldsymbol{\beta} = Z\boldsymbol{\gamma}$$

In simple linear regression, $y_i = \alpha + \beta x_i + e_i$, the most common test is for $\beta = 0$.

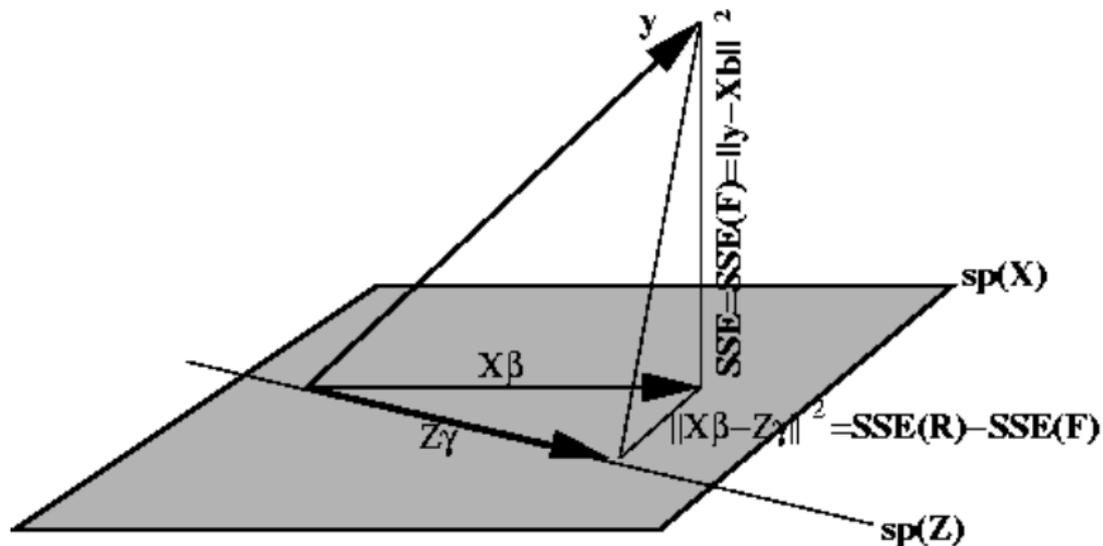# Geometric comparisons of models

Figure: Testing linear hypotheses in linear models corresponds to projecting onto subspaces.

Relationships between sums of squares in two linear models is best viewed geometrically. Starting with a base model as before, $y = X\beta + e$, there is a need to investigate whether this model can be simplified in some manner.

A simpler model can be denoted by $y = Z\gamma + e$ where Z is a matrix, typically with fewer columns than X, and the column vectors of Z span a subspace of that spanned by X.

Example:

Bases for the span of X

Orthonormal basis, $\{u_1, \ldots, u_n\}$ for R^n:

Using Gram-Schmidt, first generate $u_1, \ldots, u_q$ which span $sp\{Z\}$, the next vectors, $u_{q+1}, \ldots, u_r$ are chosen so that $u_1, \ldots, u_r$ span $sp\{X\}$, with $rank\{X\} = r$, and the rest, $u_{r+1}, \ldots, u_n$ are chosen so that the entire set, $u_1, \ldots, u_n$ spans R^n.

$$
\begin{aligned}
Z\hat{\gamma} &= \hat{\zeta}_1 u_1 + \ldots \hat{\zeta}_q u_q \\
X\hat{\beta} &= \hat{\zeta}_1 u_1 + \ldots \hat{\zeta}_q u_q + \hat{\zeta}_{q+1} u_{q+1} + \ldots \hat{\zeta}_r u_r \\
y &= \hat{\zeta}_1 u_1 + \ldots \hat{\zeta}_q u_q + \hat{\zeta}_{q+1} u_{q+1} + \ldots \hat{\zeta}_r u_r + \hat{\zeta}_{r+1} u_{r+1} + \ldots \hat{\zeta}_n u_n
\end{aligned}
$$

Expected values of coefficients

For $i = r + 1, \ldots, n$ we obtain

$$E\left[\hat{\zeta}_i\right] = 0$$

If $H_0 : X\boldsymbol{\beta} = Z\boldsymbol{\gamma}$ is true then for $i = q + 1, \ldots, r$ we obtain

$$E\left[\hat{\zeta}_i\right] = u_i \cdot (Z\boldsymbol{\gamma}) = 0$$

Sums of squares and norms

$$SSE(F) = ||y - X\hat{\boldsymbol{\beta}}||^2 \qquad = \sum_{i=r+1}^{n} \hat{\zeta}_i^2$$

$$SSE(F) - SSE(R) = ||Z\hat{\gamma} - X\hat{\boldsymbol{\beta}}||^2 \quad = \sum_{i=q+1}^{r} \hat{\zeta}_i^2$$

$$SSE(R) = ||y - Z\hat{\gamma}||^2 \qquad = \sum_{i=q+1}^{n} \hat{\zeta}_i^2$$

Each $\hat{\zeta}_i$ is a coordinate in an orthonormal basis, $\hat{\zeta}_i = y \cdot u_i$. When Y_i are independent Gaussian random variables, $\hat{\zeta}_i$ also become independent. As a result, the sums of squares are related to $\sigma^2 \cdot \chi^2$-distributions in a natural way.

Some probability distributions

Matrices, X, Z with $rank(Z) = q < r = rank(X)$
and $sp(Z) \subseteq sp(X)$ (Z may be $n \times q$ and X $n \times p$ w/$p = r$).
$H_0 : E[Y] = Z\gamma$ is a reduction of $E[Y] = X\beta$.
If F = full model and R = reduced then

1) $y - X\hat{\beta} \perp X\hat{\beta} - Z\hat{\gamma}$

2) $||y - X\hat{\beta}||^2$ and $||X\hat{\beta} - Z\hat{\gamma}||^2$ are independent

3) $\frac{||y - X\hat{\beta}||^2}{\sigma^2} \sim \chi^2_{n-r}$ if the model is correct

4) $\frac{||X\hat{\beta} - Z\hat{\gamma}||^2}{\sigma^2} \sim \chi^2_{r-q}$ if H_0 is correct.

3) $\frac{||y - X\hat{\beta}||^2}{\sigma^2} \sim \chi^2_{n-p}$ if the model is correct

4) $\frac{||X\hat{\beta} - Z\hat{\gamma}||^2}{\sigma^2} \sim \chi^2_{p-q}$ if H_0 is correct.

5) $SSE(F) = ||y - X\hat{\beta}||^2$
and
$SSE(R) - SSE(F) = ||X\hat{\beta} - Z\hat{\gamma}||^2$
are independent.

6) $\frac{(SSE(R) - SSE(F))/(p-q)}{SSE(F)/(n-p)} \sim F_{p-q,n-p}$

Here F_{ν_1,ν_2} is the distribution of a ratio

$$F = \frac{U/\nu_1}{V/\nu_2}$$

General F-tests in linear models

In general one can compute the sum of squares from the full model, $SSE(F)$ as above and then compute the sum of squared deviations from the reduced model, $SSE(R) = ||y - Z\hat{\gamma}||^2$. Denote the corresponding degrees of freedom by $df(F)$ and $df(R)$, and assume that both matrices Z and X have full ranks, i.e. $rank(X) = r$ og $rank(Z) = q$.

Then $df(F) = n - r$ and $df(R) = n - q$.

The null hypothesis can then be tested by noting that

$$F = \frac{(SSE(R) - SSE(F))/(r - q)}{SSE(F)/(n - r)}$$

is a realisation of a random variable from an F-distribution with $r - q$ and $n - r$ degrees of freedom under H_0.

References Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. 1996. Applied linear statistical models. McGraw-Hill, Boston. 1408pp.

Copyright 2021, Gunnar Stefansson