Multiple Linear Regression

Suppose our model is

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

and we observe $y_1, \ldots, y_n$.

We can summarize all the equations as:

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y = X\beta + \varepsilon$$

The least squares approach would be to minimize $\sum \varepsilon_i^2 = \varepsilon^T \varepsilon$.

Since $\varepsilon = Y - X\beta$, we look to $(y - X\beta)^T (y - X\beta) =$

$$= y^T y - \beta^T X^T y - y^T X \beta + \beta^T X^T X \beta$$

Now $\frac{\partial}{\partial \beta} \varepsilon^T \varepsilon = \frac{\partial}{\partial \beta} \left(y^T y - \beta^T X^T y - y^T X \beta + \beta^T X^T X \beta \right) =$

$$= -1^T X^T y - y^T X \cdot 1 + 1 \cdot X^T X \beta + \beta^T X^T X \cdot 1 = -2 X^T y + 2 X^T X \beta$$

$= 0$ when $X^T X \beta = X^T y$ i.e. when $\hat{\beta} = (X^T X)^{-1} X^T y$.

In the simple linear regression case,

$$X^TX = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix} \qquad (X^TX)^{-1} = \begin{bmatrix} \Sigma x_i^2/nS_{xx} & -\bar{x}/S_{xx} \\ -\bar{x}/S_{xx} & 1/S_{xx} \end{bmatrix} = \begin{bmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{bmatrix}$$

The estimator $\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY$

From this we can calculate the

$$SSE = \Sigma(y_i - \hat{y}_i)^2 = (Y - X\hat{\beta})^T(Y - X\hat{\beta}) =$$
$$= Y^TY - Y^TX\hat{\beta} - \hat{\beta}^TX^TY + \hat{\beta}X^TX\hat{\beta} =$$
$$= Y^TY - 2\hat{\beta}^TX^TY + \hat{\beta}X^TX(X^TX)^{-1}X^TY = Y^TY - 2\hat{\beta}^TX^TY + \hat{\beta}\mathbb{1}\cdot X^TY = Y^TY - \hat{\beta}^TX^TY.$$

All our previous properties carry over for $\hat{\beta}$:

Assume ε_i independent. $\mathbb{E}=0$ $\mathbb{V}=\sigma^2$

1. $\mathbb{E}\hat{\beta}_i = \beta_i$
2. $\mathbb{V}\hat{\beta}_i = c_{ii}\sigma^2$ where $c_{ii} = (X^TX)^{-1}_{ii}$
3. $\text{cov}(\hat{\beta}_i, \hat{\beta}_j) = c_{ij}\sigma^2$ where $c_{ij} = (X^TX)^{-1}_{ij}$
4. $SSE/(n-k-1)$ is unbiased for σ^2.

If ε_i are normally distributed

5. Each $\hat{\beta}_i$ is normal
6. $SSE/\sigma^2 \sim \chi^2(n-k-1)$
7. SSE and $\hat{\beta}_i$ are independent

Inferences on $a^T\beta$

Suppose that we want to make inferences on a linear combination of model parameters.

The linear combination $a^T\beta$ has (by linearity of expectation) an unbiased estimator $a^T\hat{\beta}$. It is normally distributed if each $\hat{\beta}_i$ is.

By expanding $\mathbb{V}(a^T\hat{\beta})$, we can recognize the result as

$$\mathbb{V}(a^T\hat{\beta}) = \left[a^T(X^TX)^{-1}a\right]\sigma^2$$

With a normally distributed estimator with known variance we can construct Wald and T tests and CIs.

We get T-test test statistics

$$T = \frac{a^T\hat{\beta} - (a^T\beta)_0}{S\sqrt{a^T(X^TX)^{-1}a}} \sim t(n-k-1) \quad \text{where } S^2 = \frac{Y^TY - \hat{\beta}^TX^TY}{n-k-1}.$$

$$a^T\beta \in a^T\hat{\beta} \pm t_{\alpha/2}\, S\sqrt{a^T(X^TX)^{-1}a} \quad \text{where } t_\alpha = F^{-1}_{t(n-k-1)}(\alpha).$$

This setup can be used both for inferences on each β_i in isolation, and also for estimation and prediction at a new value $x^* = (x_1^*, \ldots, x_k^*)$.

For estimation, we'd just use the T-statistic as is.
For prediction, we instead (again) estimate the error $y^* - \hat{y}^*$

The same derivation as before leads us to the variance of error $= [1 + a^T (X^T X)^{-1} a] \sigma^2$ and corresponding CIs.

The partial F-test

Suppose we want to figure out whether or not to include all variables when building a model.

This would be of interest eg to avoid overfitting where the model learns more about the data than about generalizing from data

We can do this by fitting linear models both with and without the extra variables and see if sufficiently much variance is explained to motivate the extra data used.

We would have a complete model

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g + \beta_{g+1} x_{g+1} + \cdots + \beta_k x_k + \varepsilon$$

and a reduced model

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g + \varepsilon$$

With data we can calculate SSE for both, producing SSE_R and SSE_C.
We expect $SSE_R \geq SSE_C$, and so we can use their difference as a measure of the importance of $x_{g+1}, \ldots, x_k$.

A hypothesis test would use $H_0: \beta_{g+1} = \cdots = \beta_k = 0$ vs H_A: at least one $\neq 0$.

We have, in the end, three sums of squares — each related to a χ^2 variable — much like in our ANOVA earlier. assuming H_0, the following holds:

SSE_R $\quad X_3 = SSE_R/\sigma^2 \sim \chi^2(n-g-1)$

SSE_C $\quad X_2 = SSE_C/\sigma^2 \sim \chi^2(n-k-1)$

$SSE_R - SSE_C$ $\quad X_1 = (SSE_R - SSE_C)/\sigma^2 \sim \chi^2(k-g)$

X_1 and X_2 are independent random variables.

If the extra variables are important, then the reduction of variance will be large — so X_1 will be large.

With two χ^2 variables it is quite tempting to compute an F-statistic:

$$F = \frac{X_1/(k-g)}{X_2/(n-k-1)} \sim F^{k-g}_{n-k-1}$$

greater values of this F indicate higher importance for the extra variables

By default, R will compute this F-statistic for you with $g=1$ — ie comparing

$Y_i = \beta_0 + \varepsilon$ to

$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$

You get details on the F calculation using anova(model).

For other g, you will need to fit both models yourself and feed them to anova.

Comparison with $Y = \beta_0 + \varepsilon$

In this model, $SSE_R = S_{yy}$.

We can compare the full variation (in S_{yy}) to the variation after accounting for $x_1, \ldots, x_k$ (in SSE_C) by comparing their difference $S_{yy} - SSE_C$ to S_{yy}.

We define the multiple coefficient of determination

$$R^2 = \frac{S_{yy} - SSE_C}{S_{yy}}.$$

R^2 measures proportion of variance explained by all the variables.

ANOVA as a linear model

Recall that ANOVA tests $H_0: \mu_1 = \cdots = \mu_K = 0$ for a set of K groups.

This H_0 is quite similar to the $H_0: \beta_1 = \cdots = \beta_K = 0$.

No coincidence: Let $x_{ij} = 1$ if observation i belongs to group j.
Then $Y = X\beta + \varepsilon$ reduces straight to the ANOVA model.

Linear independence

If the columns in X are not linearly independent, then $(X^TX)^{-1}$ does not exist and the least squares estimation fails.
In an ANOVA, this is guaranteed to be the case because of the presence of β_0.
The usual solution is to just drop a column (eg the last one) and use β_0 instead.