Multiple regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

- Least squares estimator:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

- The Hat matrix is $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$
- predicted values are $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$
- residuals are $\boldsymbol{y} - \hat{\boldsymbol{y}}$
- if the assumptions of the regression model hold, then
  - the elements of $\hat{\boldsymbol{\beta}}$ are normally distributed
  - $\boldsymbol{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$
  - $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$
- estimate of σ^2

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-(k+1)} = \frac{SSE}{n-p}$$

- a $100(1-\alpha)\%$ confidence interval for β_j is given by

$$\hat{\beta}_j \pm t_{\alpha/2,n-p}\hat{\sigma}\sqrt{C_{j,j}}$$

  where $C_{j,j}$ is the $j+1$'th diagonal element of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

- a $100(1-\alpha)\%$ confidence interval for the mean of y when $\boldsymbol{x} = \boldsymbol{x}_0$ is

$$\hat{y}_0 \pm t_{\alpha/2,n-p}\hat{\sigma}\sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}$$

- a $100(1-\alpha)\%$ prediction interval for a future value of y when $\boldsymbol{x} = \boldsymbol{x}_0$ is

$$\hat{y}_0 \pm t_{\alpha/2,n-p}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}$$

- Simultaneous confidence intervals: if making m intervals, replace α by α/m.

Sums of squares

- $\boldsymbol{1}$ is a vector of 1's of length n
- $\boldsymbol{J} = \boldsymbol{1}\boldsymbol{1}^T$
- $\boldsymbol{J}/n = \boldsymbol{1}(\boldsymbol{1}^T\boldsymbol{1})^{-1}\boldsymbol{1}^T$
- Residuals are $\boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$.
- The matrices $\boldsymbol{H}, \boldsymbol{I} - \boldsymbol{H}, \boldsymbol{J}/n, \boldsymbol{I} - \boldsymbol{J}/n$, and $\boldsymbol{H} - \boldsymbol{J}/n$ are *projection* matrices.
- Projection matrices are symmetric and idempotent (which means that when you multiply the matrix by itself, you get back the orignal matrix).
- Matrix expressions for the sums of squares.

$$SST = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{J}/n)\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{y} - n\bar{y}^2$$

$$\begin{aligned} SSR &= \boldsymbol{y}^T(\boldsymbol{H} - \boldsymbol{J}/n)\boldsymbol{y} = \hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}} - n\bar{y}^2 \\ &= \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y} - n\bar{y}^2 \end{aligned}$$

$$\begin{aligned} SSE &= \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \\ &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y} \end{aligned}$$

Anova table

- with n observations, k predictors, $p = k+1$, the Anova table is given by:

Source	SS	df	MS	F
Regression	SSR	k	MSR=SSR/k	MSR/MSE
Residual	SSE	n-p	MSE=SSE/(n-p)	
Total	SST	n-1		

- The proportion of the variation in $\boldsymbol{y}$ explained by the k predictor variables is $R^2 = SSR/SST$.

Means, variances and covariances

- Expectation of a linear combination of random variables.

$$E\left[\sum_{i=1}^{n} a_i Y_i\right] \;=\; \sum_{i=1}^{n} a_i E[Y_i].$$

- Variance of a general linear combination of random variables.

$$Var\left[\sum_{i=1}^{n} a_i Y_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov[Y_i,\ Y_j]$$

$$= \sum_{i=1}^{n} a_i^2 Var[Y_i] + 2\sum\sum_{j>i}^{n} a_i a_j Cov[Y_i,\ Y_j].$$

- If the Y_i are uncorrelated, then

$$Var\left[\sum_{i=1}^{n} a_i Y_i\right] = \sum_{i=1}^{n} a_i^2 Var[Y_i].$$

- Covariance of two linear combinations.

$$Cov(\sum_{i=1}^{n} a_i Y_i, \sum_{i=1}^{n} b_i Y_i) = \sum_i \sum_j a_i b_j Cov(Y_i, Y_j)$$

$$= \sum_i a_i b_i Var(Y_i) + \sum_{i\neq j} a_i b_j Cov(Y_i, Y_j).$$

Means and covariances of random vectors

- If $\boldsymbol{A}$ is a matrix of constants and $\boldsymbol{c}$ a vector of constants, and $\boldsymbol{Y}$ is a random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$, then

$$E(\boldsymbol{AY} + \boldsymbol{c}) = \boldsymbol{A\mu} + \boldsymbol{c}$$

$$Var(\boldsymbol{AY} + \boldsymbol{c}) = \boldsymbol{A}Var(\boldsymbol{Y})\boldsymbol{A}^T = \boldsymbol{AVA}^T.$$

- Suppose $Cov(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{C}$. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be non random matrices and $\boldsymbol{c}$ and $\boldsymbol{d}$ non random vectors. Then

$$Cov(\boldsymbol{AX} + \boldsymbol{c}, \boldsymbol{BY} + \boldsymbol{d}) = \boldsymbol{ACB}^T.$$

- The expectation of the *quadratic form* $\boldsymbol{Y}^T\boldsymbol{AY}$ is

$$E(\boldsymbol{Y}^T\boldsymbol{AY}) = \boldsymbol{\mu}^T\boldsymbol{A\mu} + tr(\boldsymbol{VA})$$

- Suppose that $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices of constants, $\boldsymbol{c}$ is a vector of constants, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent random vectors with covariance matrices $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$ respectively. Then the covariance matrix of $\boldsymbol{AX} + \boldsymbol{BY} + \boldsymbol{c}$ is given by $\boldsymbol{A\Sigma_X A^T} + \boldsymbol{B\Sigma_Y B^T}$.

  Indicator of the event A: let $I(A) = 1$ if A occurs, and $I(A) = 0$ otherwise.

Partial F test

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

$$H_1 = X_1(X_1^T X_1)^{-1} X_1^T$$

$$H_{2.1} = (I - H_1)X_2[X_2^T(I - H_1)X_2]^{-1}X_2^T(I - H_1)$$

To test $H_0 : \beta_2 = 0$ against the $\beta_2 \neq 0$, use:

$$F = \frac{MSR(\beta_2|\beta_1)}{MSE} = \frac{(SSE(\beta_1) - SSE(\beta_1, \beta_2))/r}{MSE}$$

ANOVA table

Source	SS	df	MS
X_1	$SSR(\beta_1)$	$k - r$	$SSR(\beta_1)/(k - r)$
$X_2\|X_1$	$SSR(\beta_2\|\beta_1)$	r	$SSR(\beta_2\|\beta_1)/r$
Error	SSE	$n - p$	MSE
Total	SST	$n - 1$	

3 step procedure

Regression on both sets of variables can be thought of as a sequential three step procedure:

1. Regress y on X_1 to get residuals $e_1 = (I - H_1)y$ and estimates $\hat{\theta}$.
2. Regress X_2 on X_1 (each column) to get residuals $e_2 = (I - H_1)X_2$.
3. Regress e_1 on e_2 to get $\hat{\beta}_2$, and then solve for $\hat{\beta}_1$.

General linear hypothesis

to test $H_0 : T\beta = c$, vs $H_A : T\beta \neq c$, use:

$$F = \frac{(T\hat{\beta} - c)' \left[T(X'X)^{-1}T'\right]^{-1} (T\hat{\beta} - c)/r}{SSE_{full}/(n - p)}$$

Variance stabilizing transform

If Y has mean μ_Y and variance σ_Y^2, then a function $h(Y)$ has approximate mean $h(\mu_Y)$ and variance approximately equal to $(h'(\mu_Y))^2\sigma_Y^2$

Generalized least squares

if $y = X\beta + \epsilon$, with $E(\epsilon) = 0$ and $Cov(\epsilon) = \sigma^2 V$, the generalized least squares estimator is given by

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$$

Model selection

For multiple linear regression, a reasonable model selection strategy is to choose that model which maximizes $R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$.

Diagnostics

- **variance inflation factor**: $VIF_j = \frac{1}{1-R_j^2}$
- **leverage** of the i'th case is h_{ii}
- **deleted residual**: $e_{(i)} = \frac{e_i}{1-h_{ii}}$
- **deleted residual sum of squares**: $SSE_{(i)} = SSE - \frac{e_i^2}{1-h_{ii}}$
- **deleted variance estimate**: $s^2_{(i)} = \frac{SSE_{(i)}}{n-k-2}$
- **externally studentized residual**:
$$t_i = \frac{e_i/(1 - h_{ii})}{s_{(i)}/\sqrt{1 - h_{ii}}} = \frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}}$$
- The **deleted estimate of β**:
$$\hat{\beta}_{(i)} = \hat{\beta} - (X^T X)^{-1} x_i \frac{e_i}{1 - h_{ii}}$$
- **Cook's distance**:
$$D_i = \frac{1}{(k + 1)s^2}(\hat{\beta} - \hat{\beta}_{(i)})^T X^T X(\hat{\beta} - \hat{\beta}_{(i)})$$
or
$$D_i = \frac{(\hat{y}_{(i)} - \hat{y})'(\hat{y}_{(i)} - \hat{y})}{(k + 1)s^2}$$
- Flag cases for which $h_{ii} > 2p/n$, $|t_i| > 2$, or $D_i > 1$. Flag variables for which $VIF > 10$.