Linear Regression Model:

$$y = X\beta + \epsilon$$

y: n by 1 vector of phenotypes/outcome/responses

X: n by p matrix of covariates/explanatory variables

β: p by 1 vector of coefficients/regression parameters

ε: n by 1 vector of residual errors

Ordinal lease squares (OLS)

To obtain OLS estimates, we minimize the sum of squared residuals (SSR); also called the error sum of squares (ESS) or residual sum of squares (RSS):

SSR=$(y - X\beta)^T(y - X\beta) = y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta$

To minimize this quantity, we simply take the first derivative

And set it to be zero:

$$\frac{\partial SSR}{\partial \beta} = 0 - 2X^Ty + 2X^TX\beta = 0$$

$$\hat{\beta}_{OLS} = (X^TX)^{-1}X^Ty$$

When we make two assumptions: E(ε)=0; V(ε)=$\sigma^2 I_{nxn}$

Then we can derive several key properties of $\hat{\beta}_{OLS}$:

1. Unbiased: $E\left(\hat{\beta}_{OLS}\right) = E((X^TX)^{-1}X^Ty)$= $(X^TX)^{-1}X^TE(y) = (X^TX)^{-1}X^TX\beta = \beta$

2. $V\left(\hat{\beta}_{OLS}\right) = (X^TX)^{-1}X^TV(y)X(X^TX)^{-1} = (X^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$

3. is the best linear unbiased estimator (BLUE):

$\hat{\beta}_{OLS}$ has the smallest variance among all unbiased estimators

that are linear functions of y (i.e. of the form Cy, where C is p by n matrix)

based on the Gauss-Markov theorem

4. asymptotically normal (by CTL):

$$\hat{\beta}_{OLS} = (X^TX)^{-1}X^Ty = (X^TX)^{-1}X^T(X\beta + \epsilon) = \beta + (X^TX)^{-1}X^T\epsilon$$

Prediction: with X_{new}, we obtain $\hat{y}_{new} = X_{new}\hat{\beta}_{OLS}$

With X, we can obtain $\hat{y} = X\hat{\beta}_{OLS} = X(X^TX)^{-1}X^Ty = Hy$

Here: $H = X(X^TX)^{-1}X^T$ is an n by n matrix called the hat/projection matrix,

which projects to the column space of X

Properties of H:

1. H is idempotent: HH=H
2. I-H is also idempotent: (I-H)(I-H)=I-H
3. H(I-H)=0
4. HX=X, (I-H)X=0 (X is invariant under H)

Estimate σ^2:

$$\hat{\sigma}^2_{OLS} = \frac{SSE}{n-p} = \frac{(y-\hat{y})^T(y-\hat{y})}{n-p} = \frac{(y-Hy)^T(y-Hy)}{n-p} = \frac{y^T(I-H)^T(I-H)y}{n-p} = \frac{y^T(I-H)y}{n-p}$$

$$\hat{\sigma}^2_{OLS} = \frac{\epsilon^T(I-H)\epsilon}{n-p}$$

$$E(\hat{\sigma}^2_{OLS}) = E\left(\frac{\epsilon^T(I-H)\epsilon}{n-p}\right) = E\left(\frac{trace(\epsilon^T(I-H)\epsilon)}{n-p}\right) = E\left(\frac{trace\big((I-H)\epsilon\epsilon^T\big)}{n-p}\right)$$
$$= \frac{trace\big((I-H)E(\epsilon\epsilon^T)\big)}{n-p} = \frac{trace\big((I-H)\sigma^2\big)}{n-p} = \sigma^2\frac{trace\big((I-H)\big)}{n-p} = \sigma^2$$

Maximum likelihood estimate (MLE)

Assumption: $\epsilon \sim MVN(0, \sigma^2 I_{nxn})$

Likelihood: $L(\beta, \sigma^2) = P(y|X, \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)}$

Log-likelihood: $l = \log L(\beta, \sigma^2) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)$

To obtain MLE, we obtain first derivatives and set them to 0:

$$\frac{\partial l}{\partial \beta} = -\frac{1}{2\sigma^2}(2X^TX\beta - 2X^Ty) = 0$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y-X\beta)^T(y-X\beta) = 0$$

$$\hat{\beta}_{MLE} = (X^T X)^{-1} X^T y \sim MVN(\beta, \sigma^2 (X^T X)^{-1})$$

$$\hat{\sigma}^2_{MLE} = \frac{\left(y - X\hat{\beta}_{MLE}\right)^T (y - X\hat{\beta}_{MLE})}{n} \sim \frac{\sigma^2}{n} \chi^2_{n-p}$$

REML: restricted/residual maximum likelihood

One approach to derive REML estimator is to integrate out β

$$L_r = \int_\beta L(\beta, \sigma^2) d\beta = \int_\beta \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)} d\beta = \int_\beta \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}(\beta^T X^T X\beta - 2\beta^T X^T y + y^T y)} d\beta$$

$$= \int_\beta \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\left\{(\beta-\hat{\beta})^T X^T X(\beta-\hat{\beta}) + y^T y - \hat{\beta}^T X^T X\hat{\beta}\right\}} d\beta \;\; (\hat{\beta} = (X^T X)^{-1} X^T y)$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} (2\pi\sigma^2)^{\frac{p}{2}} |X^T X|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}\{y^T y - \hat{\beta}^T X^T X\hat{\beta}\}}$$

$$= |X^T X|^{-\frac{1}{2}} \frac{1}{(2\pi\sigma^2)^{\frac{n-p}{2}}} e^{-\frac{1}{2\sigma^2}\{y^T y - \hat{\beta}^T X^T X\hat{\beta}\}} = |X^T X|^{-\frac{1}{2}} \frac{1}{(2\pi\sigma^2)^{\frac{n-p}{2}}} e^{-\frac{1}{2\sigma^2}\{y^T (I-H) y\}}$$

To obtain REML estimate for σ^2, we can take the first derivative on the log(Lr) and we will set derivative to be zero

$$\frac{\partial log L_r}{\partial \sigma^2} = -\frac{n-p}{2\sigma^2} + \frac{y^T (I-H) y}{2(\sigma^2)^2} = 0$$

$$\hat{\sigma}^2_{REML} = \frac{y^T (I-H) y}{n-p}$$

Therefore, $E(\hat{\sigma}^2_{REML}) = \sigma^2$.

For β estimation, we simply plug in the $\hat{\sigma}^2_{REML}$, and obtain $\hat{\beta} = (X^T X)^{-1} X^T y$

Hypothesis Test: $H_0: \theta = \theta_0$, $H_1: \theta \neq \theta_0$. Here, $\theta = L\beta$ is an r by 1 vector of transformed parameters (a.k.a. secondary parameters). θ is defined by L, which is called the contrast matrix. θ_0 is usually a vector of zeros.

Wald Test

If σ^2 is known

$$\hat{\beta} \sim MVN(\beta, \sigma^2 (X^T X)^{-1})$$

Therefore, $\hat{\theta} = L\hat{\beta} \sim MVN(L\beta, \sigma^2 L(X^T X)^{-1} L^T)$

To test $H_0: \theta = 0$, we just need to compute the following Wald statistic that follows exactly a chi-square distribution: $W^2 = \left(L\hat{\beta}\right)^T (\sigma^2 L(X^T X)^{-1} L^T)^{-1} \left(L\hat{\beta}\right) \sim \chi^2_r$

If σ^2 is not known, to test $H_0: \theta = 0$, we just need to compute the following Wald statistic that follows asymptotically a chi-square distribution: $W^2 = \left(L\hat{\beta}\right)^T (\hat{\sigma}^2 L(X^T X)^{-1} L^T)^{-1} \left(L\hat{\beta}\right) \sim \chi_r^2$

Likelihood Ratio Test

$$G^2 = 2(\log\left(L_{full}\right) - \log(L_{null}))$$

L_{full} is the maximized likelihood under unrestricted model

L_{null} is the maximized likelihood under H_0

$G^2 \sim \chi_r^2$ asymptotically

Score Test

Define the score function: $U(\theta) = \frac{\partial \log L}{\partial \theta}$, which is an r-vector

And information matrix: $I(\theta) = E(-\frac{\partial^2 \log L}{\partial \theta^2})$, which is an r by r matrix

$S = U^T(\theta) I(\theta)^{-1} U(\theta)$ is the score statistics evaluated at θ_0 under the null, and $S \sim \chi_r^2$ asymptotically

All three tests are asymptotically equivalent

In linear models, Wald statistics >= Likelihood ratio statistics >= score statistics

The likelihood ratio test often times has the best small-sample properties