A function g of random variable X with probability density function p has expectation

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot p(x) \cdot dx$$

Important rules in probability of random variables in *uppercase* (ex. X) with constants in *lowercase* (ex. a) are

$$\begin{aligned}
\mathbb{V}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\
\mathbb{C}\text{ov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])] = \mathbb{E}[X \cdot Y] + \mathbb{E}[X] \cdot \mathbb{E}[Y] \\
\mathbb{E}[a \cdot X + b \cdot Y] &= a \cdot \mathbb{E}[X] + b \cdot \mathbb{E}[Y] \\
\mathbb{V}(a \cdot X + b \cdot Y) &= a^2 \cdot \mathbb{V}(X) + b^2 \cdot \mathbb{V}(Y) + 2 \cdot a \cdot b \cdot \mathbb{C}\text{ov}(X, Y) \\
\mathbb{E}[Y] &= \mathbb{E}[\mathbb{E}[Y|X]]
\end{aligned}$$

The **bias** and **mean squared error** MSE of estimator $\hat{\theta}$ for parameter θ are

$$\begin{aligned}
\mathbb{B}\text{ias}\left(\hat{\theta}\right) &= \mathbb{E}\left[\hat{\theta}\right] - \theta \\
MSE(\hat{\theta}) &= \mathbb{V}\left(\hat{\theta}\right) + \mathbb{B}\text{ias}\left(\hat{\theta}\right)^2
\end{aligned}$$

The **central limit theorem** says the distribution of $\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ converges to this as $n \to \infty$.

$$\frac{\bar{X} - \mathbb{E}\left[\bar{X}\right]}{\sqrt{\mathbb{V}\left(\bar{X}\right)}} \sim N(0, 1)$$

Linear Predictor

Find **best linear predictor** $m(X) = \beta_0 + \beta_1 \cdot X$ of Y by finding estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the expected squared error

$$\begin{aligned}
&\min_{\beta_0, \beta_1} \mathbb{E}\left[(Y - (\beta_0 + \beta_1 \cdot X))^2\right] \\
&\hat{\beta}_1 = \frac{\mathbb{C}\text{ov}(X, Y)}{\mathbb{V}(X)} \\
&\hat{\beta}_0 = \mathbb{E}[Y] - \hat{\beta}_1 \cdot \mathbb{E}[X]
\end{aligned}$$

Since the true values of $\mathbb{C}\text{ov}(X, Y)$, $\mathbb{V}(X)$, $\mathbb{E}[X]$, and $\mathbb{E}[Y]$ are unknown we use estimates.

$$\begin{aligned}
\hat{\beta}_1 &= \frac{c\hat{o}v_{x,y}}{s_x^2} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \cdot \bar{x} \\
co\hat{v}_{x,y} &= \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - \bar{x} \cdot \bar{y} \\
s_x^2 &= \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2
\end{aligned}$$

Linear Model

Given $X_1, \ldots, X_n \sim F$ and $\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2)$ we assume Y_i is linearly generated

$$(Y_i | X_i = x_i) = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

Least square estimates $\hat{\beta}_1$ and $\hat{\beta}_0$ have properties

$$\hat{\beta}_1 = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \bar{x}) \cdot y_i}{s_x^2} = \beta_1 + \frac{\frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \bar{x}) \cdot \epsilon_i}{s_x^2}$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{n \cdot s_x^2}\right)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1) \cdot \bar{x} + \bar{\epsilon}$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n} \cdot \left(1 + \frac{\bar{x}}{s_x^2}\right)\right)$$

The **predicted value** $\hat{m}(x)$ for Y is

$$\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x = \beta_0 + \beta_1 \cdot x + \frac{1}{n} \cdot \sum_{i=1}^{n}(1 + (x - \bar{x}) \cdot \frac{x_i - \bar{x}}{s_x^2}) \cdot \epsilon_i$$

$$\mathbb{E}\left[\hat{m}(x)\right] = \beta_0 + \beta_1 \cdot x$$

$$\mathbb{V}\left(\hat{m}(x)\right) = \frac{\sigma^2}{n} \cdot (1 + \frac{(x - \bar{x})^2}{s_x^2})$$

The **residuals** e_i are

$$e_i = y_i - \hat{m}(x_i)$$

$$\sum_{i=1}^{n} e_i = 0$$

$$\sum_{i=1}^{n} e_i \cdot x_i = 0$$

The **sum of squared errors** SSE is

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{m}(x_i))^2$$

An **estimate** $\hat{\sigma}^2$ **for** σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} e_i^2 = \frac{1}{n} \cdot \sum_{i=1}^{n}(y_i - \hat{m}(x_i))^2$$

$$\frac{n \cdot \hat{\sigma}^2}{\sigma^2} \sim \chi^2\left(df = n - 2\right)$$

since we know

$$\sigma^2 = \mathbb{E}\left[(Y - (\beta_0 + \beta_1 \cdot X))^2\right]$$

The **estimates for standard error** are

$$\hat{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{s_x \cdot \sqrt{n-2}}$$

$$\hat{se}(\hat{\beta}_0) = \frac{\hat{\sigma}}{s_x \cdot \sqrt{n-2}} \cdot \sqrt{s_x^2 + \bar{x}^2}$$

$$\hat{se}(\hat{m}(x)) = \frac{\hat{\sigma}}{\sqrt{n}} \cdot \sqrt{1 + \frac{(x - \bar{x})^2}{s_x^2}}$$

$$\hat{se}_{pred} = \hat{se}(y - \hat{m}(x)) = \hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{s_x^2}}$$

Confidence interval C for $\hat{\beta}_0$ and $\hat{\beta}_1$ can be made assuming sample size n is large

$$C = \left[\hat{\beta} \pm Z_{\frac{\alpha}{2}} \cdot \hat{se}(\hat{\beta})\right]$$

The confidence interval C for the **actual line** $m(x) = y = \beta_0 + \beta_1 \cdot x$ that produces the data

$$C = [\hat{m}(x) \pm \hat{se}(\hat{m}(x))]$$

is different from the confidence interval C (called prediction interval) for the actual value y generated

$$C = \left[\hat{m}(x) \pm Z_{\frac{\alpha}{2}} \cdot \hat{se}_{pred}\right]$$

The **ANOVA table** gives values for regression sum of squares SS_{reg}, residual sum of squares RSS, and total sum of squares SS_{tot}.

$$\begin{aligned} Y_i - \bar{Y} &= (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \\ \sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \text{ because } \mathbb{C}\text{ov}\left(Y_i - \hat{Y}_i, \hat{Y}_i - \bar{Y}\right) = 0 \\ SS_{tot} &= RSS + SS_{reg} \end{aligned}$$

Matrices

Z is a $n \times 1$ random vector, C is a $m \times n$ constant matrix.

$$\begin{aligned} \mathbb{V}(Z) &= \mathbb{E}\left[ZZ^T\right] - \mathbb{E}[Z]\,\mathbb{E}[Z]^T \\ \mathbb{V}(CZ) &= C \cdot \mathbb{V}(Z) \cdot C^T \end{aligned}$$

The **trace** of square matrices A, B, and C are

$$\begin{aligned} tr(A) &= \sum_i A_{i,i} \\ tr(A + B + C) &= tr(A) + tr(B) + tr(C) \\ tr(ABC) &= tr(BCA) = tr(CAB) \\ \mathbb{E}\left[Z^T C Z\right] &= \mathbb{E}[Z]^T \cdot C \cdot \mathbb{E}[Z] + tr(C \cdot \mathbb{V}(Z)) \end{aligned}$$

Multiple Regression

Y is an $n \times 1$ random vector generated by a $n \times p$ design matrix $\mathbf{X}$, a $p \times 1$ coefficient vector β, and a $n \times 1$ noise vector ϵ.

$$Y = \mathbf{X}\beta + \epsilon$$

The $n \times 1$ **residuals** are

$$e = Y - \mathbf{X}\beta$$

The mean squared error is

$$\begin{aligned} MSE(\beta) &= \frac{1}{n}e^T e \\ &= \frac{1}{n}(Y^T Y - 2\beta^T \mathbf{X}^T Y + \beta^T \mathbf{X}^T \mathbf{X}\beta) \end{aligned}$$

and has gradient

$$\bigtriangledown_\beta MSE(\beta) = \frac{2}{n}(\mathbf{X}^T Y - \mathbf{X}^T \mathbf{X}\beta)$$

making the **score equations**

$$\frac{1}{n}\mathbf{X}^T(Y - \mathbf{X}\beta) = 0$$
$$\frac{1}{n}\mathbf{X}^T e = 0$$

The β that minimizes $MSE(\beta)$ is

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$$

The best linear predictor $\hat{Y}$ for Y is

$$\hat{Y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = \mathbf{H}Y$$

where $n \times n$ **hat matrix** (or influence matrix) $\mathbf{H}$ has properties

$$\mathbf{H} = \mathbf{H}^T$$
$$\mathbf{H} = \mathbf{H}^2$$

and matrix $(\mathbf{I} - \mathbf{H})$ has the same properties. The $n \times 1$ **residuals** are

$$e = Y - \hat{Y} = Y - \mathbf{H}Y = (\mathbf{I} - \mathbf{H})Y$$

Expectations are variances are

$$\begin{aligned}
\mathbb{E}\left[\hat{Y}\right] &= \mathbb{E}\left[\mathbf{H}Y\right] = \mathbf{H}\mathbb{E}\left[\mathbf{X}\beta + \epsilon\right] = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}\beta \\
\mathbb{V}\left(\hat{Y}\right) &= \mathbb{V}\left(\mathbf{H}(\mathbf{X}\beta + \epsilon\right) = \mathbf{H}\mathbb{V}\left(\epsilon\right)\mathbf{H}^T = \mathbf{H}\sigma^2\mathbf{I}\mathbf{H} = \sigma^2\mathbf{H} \\
\mathbb{E}\left[e\right] &= \mathbf{X}\beta - \mathbf{X}\beta = 0 \\
\mathbb{V}\left(e\right) &= \mathbb{V}\left((\mathbf{I} - \mathbf{H})(\mathbf{X}\beta + \epsilon)\right) = (\mathbf{I} - \mathbf{H})\mathbb{V}\left(\epsilon\right)(\mathbf{I} - \mathbf{H})^T = \sigma^2(\mathbf{I} - \mathbf{H}) \\
\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \epsilon) \\
\mathbb{V}\left(\hat{\beta}\right) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right]^T = \ldots = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}$$

A bias estimate $\hat{\sigma}^2$ of σ^2 is

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}e^T e \\
\frac{n\hat{\sigma}^2}{\sigma^2} &\sim \chi^2_{n-(p+1)} \\
\mathbb{E}\left[\hat{\sigma}^2\right] &= \frac{1}{n}\mathbb{E}\left[((\mathbf{I} - \mathbf{H}\epsilon)^T(\mathbf{I} - \mathbf{H})\epsilon\right] = \ldots = \frac{\sigma^2}{n}(n - (p+1))
\end{aligned}$$

meaning an unbiased estimate $\hat{\sigma}^2_{unb}$ of σ^2 is

$$\hat{\sigma}^2_{unb} = \frac{1}{n - (p+1)}e^T e$$

Given $\epsilon_i \sim N(0, \sigma_i^2)$ then

$$\frac{\hat{\beta}_i - \beta_i}{\hat{se}(\hat{\beta}_i)} \sim t_{n-(p+1)}$$
$$\hat{se}(\hat{\beta}) = \sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{i,i}}$$

Multicollinearity

The $(p+1) \times (p+1)$ **gram matrix** $G = (\mathbf{X}^T\mathbf{X})$ has properties

$$\mathbf{G} = \mathbf{G}^T$$
$$a^T\mathbf{G}a \geq 0$$

For any $(p+1) \times 1$ vector a. The $n \times (p+1)$ design matrix $\mathbf{X}$ is **multicollinear** if G is not invertible which happens when $\exists a \neq \vec{0}$ such that

$$a^T\mathbf{G}a = 0$$

$\mathbf{G}$ has an **eigen decomposition** with eigen values $\lambda_1 \geq \ldots \geq \lambda_{p+1}$ and eigen vectors $v_1, \ldots, v_{p+1}$ so

$$\mathbf{G}v_i = \lambda_i v_i$$
$$v_i^T v_j = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$
$$\mathbf{G} = \sum_{i=1}^{p+1} \lambda_i v_i v_i^T = \mathbf{VDV}^T$$

with j^{th} column of $\mathbf{V}$ as v_j and $\mathbf{D}$ is diagonal matrix with $D_{i,i} = \lambda_i$. $\mathbf{G}$ is multicollinear if $\lambda_{p+1} = 0$.

Ridge Regression

A term is added to mean squared error so the new objective is to minimize

$$RR = \frac{1}{n}(Y - \mathbf{X}\beta)^T(Y - \mathbf{X}\beta) + \frac{\lambda}{n}\beta^T\beta$$
$$\nabla_\beta RR = \frac{2}{n}(-\mathbf{X}^TY + \mathbf{X}^T\mathbf{X}\beta + \lambda\beta)$$

to get optimum $\hat{\beta}_\lambda$ as

$$\hat{\beta}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^TY$$

Testing and Confidence Sets

A partial F-test tests if a subset $S \subset \{1, \ldots, p\}$ of β_is are 0 by getting estimates $\hat{\sigma}^2_{full}$ and $\hat{\sigma}^2_{null}$ of σ^2 for the full and null (setting $\beta_i = 0 \; \forall \; i \in S$) so the ratio

$$F^* = \frac{(\hat{\sigma}^2_{null} - \hat{\sigma}^2_{null})/|S|}{\hat{\sigma}^2_{full}(n-(p+1))} \sim F_{|S|,n-(p+1)}$$

and reject the null model at confidence $1 - \alpha$ if

$$F^* > F_{s,n-(p+1)}(\alpha)$$

The complete F-test has $S = \{1, \ldots, p\}$ so $\hat{\sigma}^2_{null} \to s^2_Y$ and $|S| = p$.

To make a $1 - \alpha$ **confidence rectangle** for s parameters use $1 - \frac{\alpha}{s}$ confidence intervals for each parameter. This **Bonferroni correction** accounts for the probability of being outside the rectangle across multiple parameters.

Outliers and Influence

The **standardized residuals** r_i are the residuals e_i normalized to have variance 1.

$$r_i = \frac{e_i}{se(e_i)} = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{i,i}}}$$

The **jackknife residuals** t_i are

$$t_i = \frac{Y_i - \hat{Y}_i}{\hat{\mathbb{V}}(Y_i - \hat{Y}_{i,(-i)})}$$
$$= \frac{e_i}{\hat{\sigma}_{(-i)}\sqrt{1 - h_{i,i}}}$$
$$= r_i\sqrt{\frac{n-p-2}{n-p-1-r_i^2}}$$

where $\hat{Y}_{i,(-i)}$ the prediction for data point i without including data point i while fitting the model. **Hook's Distance** D_i is a measure of the influence a point i has on the regression.

$$D_i = \frac{(Y - \hat{Y}_{(-i)})^T(Y - \hat{Y}_{(-i)})}{(p+1)\hat{\sigma}^2}$$
$$= \left(\frac{r_i^2}{p+1}\right)\left(\frac{h_{i,i}}{1-h_{i,i}}\right)$$

with $D_i > 1$ generally being an influential point. This can also be defined using leave on out on $\hat{\beta}$

$$D_i = \frac{(\hat{\beta}_{(-i)} - \hat{\beta})^T\mathbf{X}^T\mathbf{X}(\hat{\beta}_{(-i)} - \hat{\beta})}{(p+1)\hat{\sigma}^2}$$
$$\hat{\beta}_{(-i)} = \hat{\beta} - \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i^T e_i}{1-h_{i,i}}$$

Model Selection

Expected **training error** T is

$$T = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right]$$

Expected **generalization error** G for a unobserved points Y' is

$$G = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i' - \hat{Y}_i)^2\right]$$

In general $G \geq T$. For the linear model with Gaussian noise

$$G = T + \frac{2}{n}\sum_{i=1}^{n}\mathbb{C}\text{ov}\left(Y_i, \hat{Y}_i\right)$$
$$G = T + \frac{2}{n}\sigma^2(p+1)$$

In **cross validation**, data D is divided into k groups $B_1, \ldots, B_k$ so for $i \in \{1, \ldots, k\}$ estimate $\hat{Y}$ from data $\{B_1 \ldots, B_{i-1}, B_{i+1}, \ldots, B_k\}$ then **generalization error estimate** $\hat{G}$ is

$$\hat{G}_i = \frac{1}{n_i}\sum_{j \in B_i}(Y_j - \hat{Y}_j)^2$$
$$\hat{G} = \frac{1}{k}\sum_{i=1}^{k}\hat{G}_i$$

Leave one out cross validation (LOOCV) is an extreme of this where $k = n - 1$. It has score

$$L = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_{i,(-i)})^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1 - H_{i,i}}\right)^2$$

Mallow's C_p statistic takes $\hat{\sigma}$ from the largest model we consider.

$$C_p = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \frac{2\hat{\sigma}}{n}(p+1)$$