Useful Notes on Statistics

Moment generating functions: The moment generating function (mgf) is defined as:

$$M(t) = E[e^{tX}] = 1 + t\,X + \frac{1}{2!}\,t^2\,X^2 + \frac{1}{3!}\,t^3\,X^3 + \dots \tag{1}$$

For two independent random variables X and Y, with mgfs $M_1(t)$ and $M_2(t)$, mgf of their sum is given by

$$E[X+Y] = E[e^{t\,(X+Y)}] = E[e^{t\,X}]\,E[e^{t\,Y}] = M_1(t)\,M_2(t) \tag{2}$$

The mgf of the linear function $a + b\,X$ is given by

$$E[a + b\,X] = E[e^{a\,t + b\,t\,X}] = e^{a\,t}\,M(b\,t) \tag{3}$$

Binomial Distribution: The probability density function is given by

$$P(x) = \begin{pmatrix} n \\ x \end{pmatrix}\,P^x\,Q^{1-x} \tag{4}$$

where P is the success probability, $Q = 1 - P$ and x is the number of observations. The mgf of the binomial distribution is given by

$$M(t) = E[e^{tX}] = \sum_x \frac{n!}{x!\,(n-x)!}\,e^{tx}\,P^x\,Q^{1-x} = (e^t\,P + Q)^n \tag{5}$$

The moment generating function is useful for computing mean and variance:

$$M'(0) = E[X] = nP \;\;,\;\; M''(0) = E[X^2] = nP + n(n-1)P^2 \tag{6}$$

thus, $\mu = E[X] = nP$ and $\sigma^2 = \mathrm{Var}[X] = E[X^2] - E[X]^2 = nP(1-P) = nPQ$.

Normal Distribution: The probability density function of the normal distribution is given by

$$P(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left[-\frac{(x-\mu)^2}{2\,\sigma^2}\right] \tag{7}$$

Then, the mgf can be computed as

$$\begin{aligned} M(t) = E[e^{tX}] &= \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{\infty} dx\,\exp\left[-\frac{1}{2\sigma^2}\,\left((x-\mu)^2 - 2\,\sigma^2\,tx\right)\right] \\ &= \exp\left[\mu t + \frac{1}{2}\,\sigma^2\,t^2\right] \end{aligned} \tag{8}$$

Using the mgf of the normal distribution, we can show that the varaible $Z = (X - \mu)/\sigma$ is a normal variate, i.e normal distribution with mean 0 and standard deviation 1 ($N(0,1)$). The mgf for Z is given by

$$M_Z(t) = e^{-\mu t/\sigma}\, M(t/\sigma) = e^{\frac{1}{2}\,t^2} \tag{9}$$

which is the mgf of the normal variate.

Central Limit Theorem: The sum of a large number of independent random variables will be approximately normally distributed. Let us prove: Consider the sum of n independent random variables $Y = X_1 + X_2 + \cdots + X_n$. Let μ and σ^2 be the mean and variance of Y and let $M_i(t)$ be the mgf of $X_i - \mu_i$. Then, mgf of $\sum_i (X_i - \mu_i)$ is

$$E[\exp(t(X_1 - \mu_1) + t(X_2 - \mu_2)\ldots)] = \prod_i M_i(t) \tag{10}$$

The mgf of the variate $(Y - \mu)/\sigma$ is then given by

$$M^*(t) = \prod_i M_i(t/\sigma) = \prod_i \left(1 + \frac{\sigma_i^2}{2}\,\frac{t^2}{\sigma^2} + \frac{\mu_{3i}}{3!}\,\frac{t^3}{\sigma^3} + \ldots\right) \tag{11}$$

since $\mu = \mu_1 + \mu_2 + \cdots + \mu_n$. Taking the log of both sides

$$\begin{aligned} \log M^*(t) &= \sum_i^n \log M_i(t/\sigma) \simeq \sum_i^n \log\left(1 + \frac{1}{2}\,\frac{\sigma_i^2 t^2}{\sigma^2}\right) \\ &\simeq \sum_i^n \frac{1}{2}\,\frac{\sigma_i^2\, t^2}{\sigma^2} = \frac{1}{2}\,t^2 \end{aligned} \tag{12}$$

since for larger n, σ^2 will be large as well ($\sigma^2 = \sum_i \sigma_i^2$), so the series expansion in σ_i/σ is convergent. Thus, the mgf of the variate $(Y - \mu)/\sigma$ is $e^{\frac{1}{2}\,t^2}$ which is the mgf of the standard normal variate.

A corlollary of the central limit theorem is that the distributions of the sample means is approximately normally distributed: Let X_i be a sample from a population. The sample mean of n samples is given by

$$\bar{X}_n = \frac{1}{n}\sum_i^n X_i \tag{13}$$

The expected value of $\bar{X}_n$ is computed as

$$E[\bar{X}_n] = \frac{1}{n}\sum_i^n E[X_i] = \mu \tag{14}$$

where μ is the population mean. Here, we assume that all of the X_is are identical and independently distributed (iid). The variance of $\bar{X}_n$ is computed as

$$\begin{aligned} E[\bar{X}_n^2] &= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right] \\ &= \frac{1}{n^2}\,E\left[\sum_{i=1}^{n} X_i^2 + 2\sum_{i=1}^{n}\sum_{j>i} X_i\,X_j\right] \\ &= \frac{1}{n}\,E[X_i^2] + \frac{2}{n}\,\frac{n(n-1)}{2}\,\mu^2 \\ &= \frac{1}{n}\,(\sigma^2+\mu^2) + \frac{n-1}{n}\,\mu^2 \end{aligned} \tag{15}$$

where we have used $E[X_i\,X_j] = E[X_i]\,E[X_j]$ when $j \neq i$. Then, $\mathrm{Var}[\bar{X}_n] = E[\bar{X}_n^2] - E[\bar{X}_n]^2$, so

$$\mathrm{Var}[\bar{X}_n] = \frac{\sigma^2}{n} \tag{16}$$

Thus the central limit theorem on the variate $(\bar{X}_n - \mu)/(\sigma/\sqrt{n})$ gives us:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{17}$$

Poisson distribution: The Poisson distribution is a limit of binomial distribution when the probablity of success $P = \mu/n$ is low but the number of trials n is very large. To arrive at Possion distribution, we rewrite the probability density function of the binomial distribution as follows:

$$\begin{aligned} P(x) &= \frac{n!}{x!\,(n-x)!}\,\left(\frac{\mu}{n}\right)^x\,\left(1-\frac{\mu}{n}\right)^{n-x} \\ &= \left[\frac{n}{n}\cdot\frac{n-1}{n}\,\ldots\,\frac{n-x+1}{n}\right]\,\frac{\mu^x}{x!}\,\left(1-\frac{\mu}{n}\right)^{n-x} \end{aligned} \tag{18}$$

The term in the brackets tend to 1 as $n \to \infty$, while $(1-\mu/n)^n \to e^{-\mu}$ and $(1-\mu/n)^x \to 1$. Then, the above term reduces to

$$P(x) = \frac{e^{-\mu}\,\mu^x}{x!} \tag{19}$$

It is straightforward to show that the mgf of the Poisson distribution is given by $e^{-\mu}\,e^{\mu\,t}$.

χ^2 **distribution:** The χ^2 distribution with f degrees of freedom (dom) is defined as a sum of f iid normal variates Z_i^2:

$$\Upsilon = Z_1^2 + Z_2^2 + \cdots + Z_f^2 \tag{20}$$

For 1 dof, $E[\Upsilon] = E[Z^2] = 1$ and $E[\Upsilon^2] = E[Z^4] = 3$. The $\chi^2_{[f]}$ distribution then satisfies:

$$E[\Upsilon] = f \ \ , \ \ \mathrm{Var}[\Upsilon] = 2\, f \tag{21}$$

Let's compute the mgf of $\chi^2_{[f]}$. Fisrt, for the $\chi^2_{[1]}$ variate

$$M(t) = E[e^{tZ^2}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dz\, e^{-\frac{1}{2}\, z^2 (1-2t)} = (1-2t)^{-1/2} \tag{22}$$

For $\chi^2_{[f]}$, the mgf will be the product of f mgfs which is

$$M_\Upsilon(t) = (1-2t)^{-\frac{1}{2} f} \tag{23}$$

Now, we can show that this mgf can be obtained from the density function $f(y) = \frac{1}{A(f)}\, y^{\frac{1}{2} f - 1}\, e^{-\frac{1}{2} y}$ where $0 \le y < \infty$ and $A(f) = 2^{\frac{1}{2} f}\, \Gamma(\frac{1}{2} f)$.

$$M_f(t) = E[e^{tY}] = \int_0^\infty e^{ty}\, f(y)\, dy = \frac{1}{A(f)} \int_0^\infty e^{ty - \frac{1}{2} y}\, y^{\frac{1}{2} f - 1}\, dy \tag{24}$$

Using the substitution $w = (1-2t)\, y$ we get

$$M_f(t) = (1-2t)^{-\frac{1}{2} f}\, \frac{1}{A(f)} \int_0^\infty dw\, w^{\frac{1}{2} f - 1}\, e^{-\frac{1}{2} w} = (1-2t)^{-\frac{1}{2} f} \tag{25}$$

Using the properties of the Γ-function, one can show that (using partial integration)

$$A(f) = \begin{cases} 1 \cdot 3 \cdot 5 \cdot \cdots \cdot (f-2)\, \sqrt{2\pi}, & \text{odd } f \\ 2 \cdot 4 \cdot 6 \cdot \cdots \cdot (f-2) \cdot 2, & \text{even } f \end{cases} \tag{26}$$

Let us consider an important application of the χ^2 distribution. Let $Z_1, Z_2, \ldots, Z_n$ be iid normal variates and $\Upsilon_1, \Upsilon_2, \ldots, \Upsilon_n$ be linear functions of them:

$$\begin{aligned} \Upsilon_1 &= a_1\, Z_1 + a_2\, Z_2 + \cdots + a_n\, Z_n \\ \Upsilon_2 &= b_1\, Z_1 + b_2\, Z_2 + \cdots + b_n\, Z_n \\ &\cdots\cdots \\ \Upsilon_n &= u_1\, Z_1 + u_2\, Z_2 + \cdots + u_n\, Z_n \end{aligned} \tag{27}$$

with $a_i\,, b_i\,, \ldots, u_i$ being a set of **orthonormal vectors**. Then, it is clear that

$$\sum_i \Upsilon_i^2 = \sum_i Z_i^2 \tag{28}$$

The orthonormality also implies that all the Υ_i's are independently distributed; for example

$$\begin{aligned}
\mathrm{Cov}[\Upsilon_1, \Upsilon_2] &= E[(a_1 Z_1 + \cdots + a_n Z_n) \cdot (b_1 Z_1 + \cdots + b_n Z_n)] \\
&= \sum_{i,j} (a_i b_j) E[Z_i Z_j] \\
&= \sum_i a_i b_i E[Z_i^2] = \sum_i a_i b_i = 0
\end{aligned} \tag{29}$$

since for $i \neq j$, $E[Z_i Z_j] = E[Z_i] E[Z_j] = 0$. Using the identity, it is possible to work out the sampling distribution of the variance.

Sampling distribution of variance: Consider the sum of squared deviations from the mean

$$S^2 = \sum_i (x_i - \bar{x})^2 \tag{30}$$

where x_i's are sample distributions and $\bar{x}$ is the mean of the sample distribution, from a population with mean μ and variance σ^2. Then, the following identity holds:

$$\begin{aligned}
\sum_i (x_i - \bar{x})^2 &= \sum_i \left((x_i - \mu)^2 - (\bar{x} - \mu)\right)^2 \\
&= \sum_i (x_i - \mu)^2 - n(\bar{x} - \mu)^2
\end{aligned} \tag{31}$$

Then, the expectation value of S^2 is given by

$$\begin{aligned}
E[S^2] &= \sum_i E\left[(x_i - \mu)^2\right] - n\, E\left[(\bar{x} - \mu)^2\right] \\
&= \sum_i \mathrm{Var}[x_i] - n\, \mathrm{Var}[\bar{x}] \\
&= n\, \sigma^2 - n \frac{\sigma^2}{n} = (n-1)\, \sigma^2
\end{aligned} \tag{32}$$

Thus, the **unbiased** estimator for the variance is given by

$$s^2 = \frac{1}{n-1} S^2 \tag{33}$$

Now, we have from the above equations

$$\frac{\sum_i (x_i - \mu)^2}{\sigma^2} = \frac{S^2}{\sigma^2} + \frac{n(\bar{x} - \mu)^2}{\sigma^2} \tag{34}$$

Assuming that x_i being **normally** distributed, notice that $\left(\frac{\bar{x}-\mu}{\sigma/n}\right)^2$ becomes a $\chi^2_{[1]}$ variate. At the same time, $\sum_i \left(\frac{x_i-\mu}{\sigma}\right)^2$ becomes a $\chi^2_{[n]}$ variate. Thus, the above equation implies that S^2/σ^2 to be a $\chi^2_{[n-1]}$ variate, as long as S^2 and $\bar{x}$ are independently distributed.

We establish this independence by constructing an orthonormal transformation from a set of normal variates Z_i to their linear combinations Υ_i:

$$Z_i = \frac{x_i - \mu}{\sigma} \quad , \quad i = 1, 2, \ldots, n$$
$$\Upsilon_1 = \frac{1}{\sqrt{n}}\, Z_1 + \cdots + \frac{1}{\sqrt{n}}\, Z_n \tag{35}$$

Notice that

$$\Upsilon_1 = \frac{1}{\sqrt{n}} \sum_i \left(\frac{x_i - \mu}{\sigma} \right) = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$
$$\rightarrow\ \Upsilon_1^2 = \frac{n(\bar{x} - \mu)^2}{\sigma^2} \sim \chi^2_{[1]}$$
$$\sum_i Z_i^2 = \frac{\sum_i (x_i - \mu)^2}{\sigma^2} \tag{36}$$

Now, complete the transformation to from Z_i to Υ_i by adding $\Upsilon_2, \ldots \Upsilon_n$ on Υ_1, such that $\mathrm{Cov}[\Upsilon_i, \Upsilon_j] = \delta_{ij}$. Otrhonormality requires

$$\sum_{i=1}^{n} Z_i^2 = \Upsilon_1^2 + \sum_{i=2}^{n} \Upsilon_i^2$$
$$\rightarrow \sum_{i=2}^{n} \Upsilon_i^2 = \frac{\sum_i (x_i - \mu)^2}{\sigma^2} - \frac{n(\bar{x} - \mu)^2}{\sigma^2} = \frac{S^2}{\sigma^2}\ \chi^2_{[n-1]} \tag{37}$$

where each Υ_i is iid and $\chi^2_{[1]}$ by orthonormality. Thus, S^2 is $\chi^2_{[n-1]}$ variate.

t-Distribution: The Student t-distribution is defined through the random variable T as follows:

$$T = \frac{Z}{\sqrt{\Upsilon/f}} \tag{38}$$

where Υ is a χ^2 variate with f dof and Z is a normal variate. This distribution is useful when the population variance σ^2 is unknown, but estimated by s^2. Since $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ is a normal variate (assuming x_i's are iid and normally distributed) and S^2/σ^2 is a $\chi^2_{[n-1]}$, then

$$\frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}} \Bigg/ \sqrt{\frac{S^2}{(n-1)\sigma^2}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{39}$$

will follow the t-distribution with $n - 1$ dof.

Let us know calculate the probability density function of the t-distribution. First consider the distribution $V = \sqrt{\Upsilon/f}$ so that $T = Z/V$. The cumulative density function for V is given by

$$F_V(v) = P[V \leq v] = P[\sqrt{\Upsilon/f} \leq v] = P[\Upsilon \leq f\, v^2] = F_\Upsilon(f\, v^2) \tag{40}$$

Then, the probability density function is given by

$$
\begin{aligned}
f_V(v) &= \frac{dF_V(v)}{dv} = \frac{dF_\Upsilon(f\,v^2)}{dv} = 2f\,v\,f_\Upsilon(f\,v^2) \\
&= \frac{1}{A(f)}\,2f\,v\,\left(f\,v^2\right)^{\frac{1}{2}f-1}\,e^{-\frac{1}{2}f\,v^2}
\end{aligned}
\tag{41}
$$

Now, for the T distribution, we have

$$
\begin{aligned}
F_T(t) &= P[Z/V \le t] = \sum_v \left(P[Z \le vt] \cup P[v \le V \le v+dv]\right) \\
&= \int F_Z(vt)\,f_V(v)\,dv \\
f_T(t) &= \int f_Z(vt)\,v\,f_v(v)\,dv \\
&= \frac{2\,f^{f/2}}{\sqrt{2\pi}\,A(f)}\int_0^\infty e^{-\frac{1}{2}\,v^2\,t^2-\frac{1}{2}f\,v^2}\,v^f\,dv
\end{aligned}
\tag{42}
$$

The integral can be evaluated by the substitution $\xi = v^2\,(f+t^2)$, and after some algebra

$$
f_T(t) = \frac{A(f+1)}{\sqrt{2\pi\,f}\,A(f)}\left(1+\frac{t^2}{f}\right)^{-\frac{1}{2}\,(f+1)} \tag{43}
$$

t-distribution can be applied to test whether the means of two distributions are the same. Suppose that we have m observations on a random variable X $(x_1,\ldots,x_m)$ and n observations on another random variable Y $(y_1,\ldots,y_n)$. Assuming that X and Y are normally distributed with the same variance σ^2, but different means μ_1 and μ_2, we test the hypotheses $\mu_1 = \mu_2$. We define

$$
\begin{aligned}
&S_1^2 = \sum_{i=1}^m (x_i-\bar{x})^2 \qquad S_2^2 = \sum_{i=1}^n (y_i-\bar{y})^2 \\
&S^2 = S_1^2 + S_2^2 \qquad s^2 = \frac{S^2}{m+n-2}
\end{aligned}
\tag{44}
$$

Then, $\bar{x}-\bar{y}$ will be normally distributed with mean $\mu_1-\mu_2$ and variance $(1/n+1/m)\,\sigma^2$, and S^2 will follow a χ^2 distribution with $m+n-2$ dof. Then,

$$
t = \frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n}+\frac{1}{m}}} \tag{45}
$$

will follow a t-distribution with $m+n-2$ dof.

Linear Regression

We first consider the case of two variables, mutilvariable extension is straightforward. The observations Y_i are modeled by the following liner function:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{46}$$

where $\epsilon_i \sim N(0, \sigma^2)$ are assumed to be iid normal variates. The precition is denoted by hatted symbols, i.e.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{47}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained through least-squares, by minimizing the sum of squared errors:

$$S^2 = \sum_i (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{48}$$

Minimizing with respect to β_0 and β_1, we obtain

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \beta_1 + \frac{\sum_i \epsilon_i (X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} \quad , \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{49}$$

Equivalently, it is easy to show that

$$\begin{aligned} \hat{\beta}_1 = \mathrm{Cor}(Y, X)\,\frac{S_Y}{S_X} \quad &, \quad \mathrm{Cor}(Y, X) = \frac{\mathrm{Cov}(Y, X)}{S_X\, S_Y} \\ \mathrm{Cov}(X, Y) &= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \\ S_X^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \end{aligned} \tag{50}$$

Prediction and confidence intervals: For simplifying the algebra, let us subtract $\bar{X}$ from X_i to normalize:

$$\hat{Y}_i = \beta_0 + \beta_1 (X_i - \bar{X}) + \epsilon_i \tag{51}$$

this is equivalent to redefining $\beta_0 \to \beta_0 - \beta_1 \bar{X}$, which results in $\hat{\beta}_0 = \bar{Y}$. Now it is easy to see the following apply:

$$\begin{aligned} &\bar{X} = \beta_0 + \frac{1}{n} \sum_i \epsilon_i \quad , \quad \bar{Y} = \hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{n}) \\ &E[Y_i] = \beta_0 + \beta_1 (X_i - \bar{X}) \quad , \quad \mathrm{Var}[Y_i] = \mathrm{Var}[\epsilon_i] = \sigma^2 \quad , \quad Y_i \sim N\left(\beta_0 + \beta_1 (X_i - \bar{X}), \sigma^2\right) \end{aligned} \tag{52}$$

now,

$$\begin{aligned}\mathrm{Var}[\hat{\beta}_1] &= \frac{1}{\left[\sum_i (X_i - \bar{X})^2\right]^2} \sum_i (X_i - \bar{X})^2 \,\mathrm{Var}[\epsilon_i] \\ &= \frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}\end{aligned} \tag{53}$$

thus, in summary:

$$\boxed{\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n}\right) \quad , \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}\right)} \tag{54}$$

If we were to add the $\bar{X}$ term back, $\beta_0 \to \beta_0 + \beta_1 \bar{X}$ so $\mathrm{Var}[\hat{\beta}_0] \to \bar{X}^2\,\mathrm{Var}[\hat{\beta}_1] + \mathrm{Var}[\hat{\beta}_0]$, which is more widely used. Now, consider a **new prediction** $\hat{Y}_{n+1}$ from a **new predictor** X_{n+1}, i.e.

$$Y_{n+1} = \beta_0 + \beta_1 (X_{n+1} - \bar{X}) + \epsilon_{n+1} \quad , \quad \hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 (X_{n+1} - \bar{X}) \tag{55}$$

consider the quantity $W = Y_{n+1} - \hat{Y}_{n+1}$. It is easy to see that $E[W] = 0$. Now, we compute the variance of W, which will be related to the **prediction interval**. To proceed, let us first note that $\mathrm{Cov}[Y_{n+1}, \hat{\beta}_0 + \hat{\beta}_1 (X_{n+1} - \bar{X})] = 0$, since Y_{n+1} is a new data uncorrelated to Y_i from which $\hat{\beta}_{0,1}$ are determined. In addition, $\hat{\beta}_0$ and $\hat{\beta}_1$ are uncorrelated as well:

$$\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_1] = \mathrm{Cov}\left[\hat{\beta}_0, \beta_1 + \frac{\sum_i \epsilon_i (X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2}\right] \tag{56}$$

since $E[\hat{\beta}_0] = E[\bar{Y}] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$, we get

$$\begin{aligned}\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_1] &= E\left[\frac{1}{n}\sum_i \epsilon_i \times \frac{\sum_i \epsilon_i (X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2}\right] \\ &= \frac{1}{n\sum_i (X_i - \bar{X})^2} \sum_{i,j} (X_i - \bar{X})\, E[\epsilon_i\, \epsilon_j] \\ &= \frac{1}{n\sum_i (X_i - \bar{X})^2} \sum_i (X_i - \bar{X})\, \sigma^2 = 0\end{aligned} \tag{57}$$

now we have shown that Y_{n+1}, $\hat{\beta}_0$, $\hat{\beta}_1$ are uncorrelated, it is straightforward to compute $\mathrm{Var}[W]$:

$$\begin{aligned}\mathrm{Var}[W] &= \mathrm{Var}[Y_{n+1}] + \mathrm{Var}[\hat{\beta}_0] + (X_{n+1} - \bar{X})^2\,\mathrm{Var}[\hat{\beta}_1] \\ &= \sigma^2 + \frac{\sigma^2}{n} + (X_{n+1} - \bar{X})^2 \frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}\end{aligned} \tag{58}$$

$$\boxed{\mathrm{Var}[W] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right]} \tag{59}$$

the equantity $\mathrm{Var}[W]$ is a measure of the error in prediction, so it determines the **prediction interval** via:

$$\hat{Y}_{n+1} \pm t_{\alpha/2,n-2}\, \hat{\sigma} \left[1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right]^{1/2} \tag{60}$$

where $t_{\alpha/2,n-2}$ is the αth t-quantile with $n-2$ dof, and $\hat{\sigma}$ is the unbaised estimate of σ:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \tag{61}$$

instead, the instrinsic error we make in $\hat{Y}_{n+1}$ is simply $\mathrm{Var}[\hat{Y}_{n+1}]$, which determines the **confidence interval**. the confidence interval is always smaller than the prediction interval, and is characterized by:

$$\boxed{\mathrm{Var}[\hat{Y}_{n+1}] = \sigma^2 \left[\frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right]} \tag{62}$$

Some properties of the residuals: The residuals are defined by $e_i = Y_i - \hat{Y}_i$, and are used in the definition of the unbiased estimate of $\mathrm{Var}[\epsilon_i] = \sigma^2$. The residuals satisfy the following properties:

$$\boxed{\sum_i e_i = 0 \quad , \quad \sum_i e_i\, X_i = 0} \tag{63}$$

Here are the proofs:

$$\begin{aligned}
\sum_i e_i &= \sum_i \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1\, X_i\right) \\
&= \sum_i \left(\beta_0 + \beta_1\, X_i + \epsilon_i - \bar{Y} + \hat{\beta}_1\, X_i - \hat{\beta}_1\, X_i\right) \\
&= \sum_i \left(\beta_0 + \beta_1\, X_i + \epsilon_i - \beta_0 - \beta_1\, \bar{X} - \frac{1}{n} \sum_j \epsilon_j + \hat{\beta}_1\, \bar{X} - \hat{\beta}_1\, X_i\right) \\
&= \left(\beta_1 - \hat{\beta}_1\right) \sum_i (X_i - \bar{X}) = 0 \\
\sum_i e_i\, X_i &= \sum_i e_i\, (X_i - \bar{X}) = \sum_i \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1\, X_i\right) (X_i - \bar{X}) \\
&= \sum_i \epsilon_i\, (X_i - \bar{X}) + (\beta_0 - \hat{\beta}_0) \sum_i (X_i - \bar{X}) + (\beta_1 - \hat{\beta}_1) \sum_i X_i\, (X_i - \bar{X}) \\
&= \sum_i \epsilon_i\, (X_i - \bar{X}) + (\beta_1 - \hat{\beta}_1) \sum_i (X_i - \bar{X})^2 \\
&= \sum_i \epsilon_i\, (X_i - \bar{X}) - \sum_i \epsilon_i\, (X_i - \bar{X}) = 0
\end{aligned} \tag{64}$$

where we used the least squares fit for $\hat{\beta}_1$ in the last line.

Now, let's compute the **variance of the resudials** (using the normalized $X_i \to X_i - \bar{X}$):

$$\begin{aligned}\text{Var}[e_i] &= \text{Var}[Y_i - \hat{\beta}_0 - \hat{\beta}_1\,(X_i - \bar{X})] \\ &= \text{Var}[Y_i] + \text{Var}[\hat{\beta}_0] + (X_i - \bar{X})^2\,\text{Var}[\hat{\beta}_1] - 2\,\text{Cov}\left[Y_i,\,\hat{\beta}_0 + \hat{\beta}_1\,(X_i - \bar{X})\right]\end{aligned} \tag{65}$$

where we have used the fact that $\hat{\beta}_0$ and $\hat{\beta}_1$ are uncorrelated and $\text{Var}[A \pm B] = \text{Var}[A] + \text{Var}[B] \pm 2\,\text{Cov}[A, B]$. The covariance term was not present when computing $\text{Var}[W]$ since there, Y_{n+1} was new data and not correlated to $\hat{\beta}_0$ and $\hat{\beta}_1$. Now, notive that

$$\begin{aligned}&Y_i - E[Y_i] = \epsilon_i \\ &\hat{\beta}_0 + \hat{\beta}_1\,(X_i - \bar{X}) - E\left[\hat{\beta}_0 + \hat{\beta}_1\,(X_i - \bar{X})\right] = (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)\,(X_i - \bar{X})\end{aligned} \tag{66}$$

Thus, the covariance term becomes

$$\begin{aligned}\text{Cov}\left[Y_i,\,\hat{\beta}_0 + \hat{\beta}_1\,(X_i - \bar{X})\right] &= E\left[\epsilon_i \times \left((\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)\,(X_i - \bar{X})\right)\right] \\ &= E\left[(\hat{\beta}_0 - \beta_0)\,\epsilon_i\right] + (X_i - \bar{X})\,E\left[(\hat{\beta}_1 - \beta_1)\,\epsilon_i\right]\end{aligned} \tag{67}$$

Now, $E[(\hat{\beta}_0 - \beta_0)\,\epsilon_i] = E[(\bar{Y} - \beta_0)\,\epsilon_i] = E[\frac{1}{n}\sum_j \epsilon_j\,\epsilon_i] = \sigma^2/n$. Similarly,

$$E\left[(\hat{\beta}_1 - \beta_1)\,\epsilon_i\right] = \frac{1}{\sum_j (X_j - \bar{X})^2}\,E\left[\sum_j (X_j - \bar{X})\,\epsilon_j\,\epsilon_i\right] = \frac{(X_i - \bar{X})\,\sigma^2}{\sum_j (X_j - \bar{X})^2} \tag{68}$$

Thus,

$$\text{Cov}\left[Y_i,\,\hat{\beta}_0 + \hat{\beta}_1\,(X_i - \bar{X})\right] = \sigma^2\,\left[1 + \frac{(X_i - \bar{X})^2}{\sum_j (X_j - \bar{X})^2}\right] \tag{69}$$

The calculation of $\text{Var}[Y_i] + \text{Var}[\hat{\beta}_0] + (X_i - \bar{X})^2\,\text{Var}[\hat{\beta}_1]$ follows similarly to $\text{Var}[W]$, so finally combining with the covariance term we get

$$\boxed{\text{Var}[e_i] = \sigma^2\,\left[1 - \frac{1}{n} - \frac{(X_i - \bar{X})^2}{\sum_j (X_j - \bar{X})^2}\right]} \tag{70}$$

Multiple features: Consider the case of two features (can easily be generalized to more). Assume that $\beta_0 = 0$ (i.e. $\to Y - \beta_0$ is being fitted) then, the least-squares minimization is performed on

$$S^2 = \sum_{i=1}^{n} \left(Y_i - \beta_1\,X_{1i} - \beta_2\,X_{2i}\right)^2 \tag{71}$$

Evaluating partial derivatives with respect to β_1, β_2 and setting them to zero, we obtain

$$\begin{bmatrix} \sum_i X_{1i}^2 & \sum_i X_{1i} X_{2i} \\ \sum_i X_{1i} X_{2i} & \sum_i X_{2i}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \sum_i Y_i X_{1i} \\ \sum_i Y_i X_{2i} \end{bmatrix} \tag{72}$$

This equation is easily solved for $\hat{\beta}_1$ and $\hat{\beta}_2$. The solution is given by

$$\begin{aligned} \Delta &= \left(\sum_i X_{1i}^2\right)\left(\sum_i X_{2i}^2\right) - \left(\sum_i X_{1i} X_{2i}\right)^2 \\ \hat{\beta}_1 &= \left[\left(\sum_i X_{2i}^2\right)\left(\sum_i Y_i X_{1i}\right) - \left(\sum_i X_{1i} X_{2i}\right)\left(\sum_i Y_i X_{2i}\right)\right] / \Delta \\ \hat{\beta}_2 &= \left[\left(\sum_i X_{1i}^2\right)\left(\sum_i Y_i X_{2i}\right) - \left(\sum_i X_{1i} X_{2i}\right)\left(\sum_i Y_i X_{1i}\right)\right] / \Delta \end{aligned} \tag{73}$$

However, we would like to express the solutions in terms of residuals. We define:

$$e_{i,\, X_1|X_2} = X_{1i} - \left(\frac{\sum_j X_{2j} X_{1j}}{\sum_j X_{2j}^2}\right) X_{2i} \tag{74}$$

which is **the residual having fit X_2 on X_1**. The term in the paranthesis is the regression coefficient (via least squares) if we were to fit $X_{1i} = \beta X_{2i}$ and the reisual is $X_{1i} - \hat{\beta} X_{2i}$. The residuals $e_{i,\, Y|X_2}$, $e_{i,\, Y|X_1}$ are similarly defined.

After some algebra, we get

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_i e_{i,\, Y|X_2} e_{i,\, X_1|X_2}}{\sum_i e_{i,\, X_1|X_2}^2} \\ \hat{\beta}_2 &= \frac{\sum_i e_{i,\, Y|X_1} e_{i,\, X_2|X_1}}{\sum_i e_{i,\, X_2|X_1}^2} \end{aligned} \tag{75}$$

In other words, the regression estimate for β_1 is the regression through the origin estimate having regressed X_2 out of both the reponse and the predictors (similar for β_2). Moreover, if we were to fit the regression through the origin between the residuals $e_{i,\, Y|X_2}$ and $e_{i,\, X_1|X_2}$:

$$e_{i,\, Y|X_2} = \gamma\, e_{i,\, X_1|X_2} \tag{76}$$

we would get

$$\hat{\gamma} = \frac{\sum_i e_{i,\, Y|X_2} e_{i,\, X_1|X_2}}{\sum_i e_{i,\, X_1|X_2}^2} \tag{77}$$

which is equivalent to $\hat{\beta}_1$. So the residuals having regressed out X_2 contains the information on the X_1 depenence. This is the reason why we see left over variability in residual plots if there is a feature which is not fitted.

Bias-Variance trade-off

Suppose we know the true model $f(X)$ that characterizes the response Y, i.e.

$$Y = f(X) + \epsilon \tag{78}$$

Here, ϵ represents the irreducible error, i.e. the error that is still there even if we knew the exact model $f(X)$. Let's assume that we have a estimate of the true model by some function $\hat{f}(X)$. Then, we can compute the total error we make (residual sums squared) as

$$\begin{aligned} E\left[(Y - \hat{f}(X))^2\right] &= E\left[\epsilon + (f(X) - \hat{f}(X))^2\right] \\ &= \mathrm{Var}(\epsilon) + E\left[(f(X) - \hat{f}(X))^2\right] \end{aligned} \tag{79}$$

where we have used the fact that $E[\epsilon] = 0$ so $\mathrm{Var}(\epsilon) = E[\epsilon^2]$, and that ϵ and $f(X) - \hat{f}(X)$ are uncorrelated. The first term above is the irreducible error. Let us look into the second term:

$$\begin{aligned} E\left[(f(X) - \hat{f}(X))^2\right] &= E[f(X)^2] - E[\hat{f}(X)]^2 + E[\hat{f}(X)]^2 - 2\, f(X)\, E[\hat{f}(X)] + f(X)^2 \\ &= \mathrm{Var}(\hat{f}(X)) + \Big(E[\hat{f}(X)] - f(X)\Big)^2 \\ &= \mathrm{Var}(\hat{f}(X)) + \mathrm{Bias}(\hat{f}(X)) \end{aligned} \tag{80}$$

The first term above is the variance of $\hat{f}(X)$ and the second one is bias, which measures how much $\hat{f}(X)$ deviates from the true $f(X)$. By including more flexibility in model $\hat{f}(X)$ we can reduce the bias, but the variance increases (the bias-variance trade-off).

Maximum Likelihoods: