Linear Models and All the Gory Details

Brian J Vlcek

Introduction

When presented with a data set, $\mathcal{D} = \{(y_1, x_1), (y_2, x_2), ..., (y_{\mathcal{N}}, x_{\mathcal{N}})\}$, where we suspect or would like to investigate if there is any relationship between variables (in the data set we suspect a relationship between the vectors x_i and the scalars y_i), we can use a statistical analysis framework known as regression to determine what, if any, relationships exists. The most widely used and theoretically tractable method is that of linear regression. The concept of linear regression is to assume that there is a true and unknown relationship of the data (x_i) to the dependent variable (y_i) of the form

$$y(x) = \sum_{m=1}^{M} \beta_m f_m(x),$$

where β_m are unknown constants, and $f_m(x)$ a collection of M base functions that are the functional dependencies for $y(x)$. Please note that all x values from here onward can be assumed to be n-tuples (n dimensional vectors) unless otherwise stated. The goal of regression analysis is to estimate the values of β_m given a finite data set. I denote estimators via an over-hat symbol and express the estimated functional relationship as,

$$\hat{y}(x) = \sum_{m=1}^{M} \hat{\beta}_m f_m(x)$$

To determine the optimal estimators $\hat{\beta}_m$ we must assume some theoretical distribution of the data set from which our finite sample $\mathcal{D}$, is drawn. There must exist a probability distribution $p(y, x)$ from which our data set was sampled $\mathcal{N}$ times. The requirement on the estimators will be that if we repeated the measuring a data set of size $\mathcal{N}$ an infinite number of times and repeated our estimation procedure, that the expectation values of $\hat{\beta}_m$ will be the true values β_m

$$E\left[\hat{\beta}_m\right] = \left\langle \hat{\beta}_m \right\rangle = \beta_m,$$

the expectation values of the estimators should unbiased β_m value estimators, however it is we decide to find them.

Signal and Noise

Since we assume that there is a true functional dependence $y(x)$ however when we have our data set, we must assume that the data is noisy, includes errors, or is otherwise marginalized over unknown variables. This concept is known as the signal and noise concept, where the sampled data, y_n, is assumed to be sampled from

$$\text{data} = \text{signal} + \text{noise} \to y_n = y(x_n) + \epsilon_n,$$

Where ϵ_n is a random independent variable that is the source of unknown variables or dependencies for $y(x_n)$. If the noise term is a sum of many independent variables randomly sampled, then the sum of the terms will follow a Gaussian distribution, due to the central limit theorem. This dictates that we may express $y(x_n) - \hat{y}(x_n) \sim \epsilon_n$. From which we can derive the probability of the observed data expressed as the Likelihood function $\mathcal{L}(y, x)$

$$\mathcal{L}(y, x) = \prod_{n=1}^{\mathcal{N}} \frac{1}{\sqrt{2\pi\sigma^2}} \text{Exp}\left[-\frac{(y_n - \hat{y}(x_n))^2}{2\sigma^2}\right] = \frac{1}{(2\pi\sigma^2)^{\mathcal{N}/2}} \text{Exp}\left[-\sum_{n=1}^{\mathcal{N}} \frac{(y_n - \hat{y}(x_n))^2}{2\sigma^2}\right],$$

where σ^2 is the unknown variance of the noise (ϵ) that contaminates our relationship.

Maximum Likelihood Estimators

If we attempt to maximize the likelihood as a function of $\hat{\beta}_m$, then $\text{Log}[\mathcal{L}]$ gives the same optimal parameters, due to Log[x] being a monotonic function (if x increase so does log(x), if x decreases so does log(x)), thus the parameters $\hat{\beta}$ determined in this method are given by

$$\text{Log}[\mathcal{L}] = -\frac{\mathcal{N}}{2}\text{Log}\left[2\pi\sigma^2\right] - \frac{1}{2\sigma^2}\sum_{n=1}^{\mathcal{N}} (y_n - \hat{y}(x_n))^2 = \text{const} + \frac{1}{\sigma^2} E\left(\hat{\beta}\right),$$

where $E\left(\hat{\beta}\right) = \frac{1}{2}\sum_{n=1}^{\mathcal{N}} (y_n - \hat{y}(x_n))^2$ is known as the error function in the form of the least squares. Finding the optima is given by

$$\frac{\partial}{\partial \hat{\beta}_m}\text{Log}[\mathcal{L}] = -\frac{1}{\sigma^2}\sum_{n=1}^{\mathcal{N}} (y_n - \hat{y}(x_n)) \frac{\partial \hat{y}(x_n)}{\partial \hat{\beta}_m} = -\frac{1}{\sigma^2}\sum_{n=1}^{\mathcal{N}} (y_n - \hat{y}(x_n)) f_m(x_n) = 0$$

thus the estimatorfor $\hat{\beta}_j$ satisfies

$$\sum_{n=1}^{\mathcal{N}} \left(y_n - \sum_{m=1}^{M} \hat{\beta}_m f_m(x_n)\right) f_j(x_n) = 0$$

known as the maximum likelihood estimators (MLE), value of $\hat{\beta}_j$. The above equation can be expressed in terms of vectors and matrices via

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_M \end{pmatrix}, y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix},\ F = \begin{pmatrix} f_1(x_1) & f_2(x_1) & ... & f_M(x_1) \\ f_1(x_2) & f_2(x_2) & ... & f_M(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(x_N) & f_2(x_N) & ... & f_M(x_N) \end{pmatrix},\ F_{\mathrm{nm}} = f_m(x_n)$$

the equation above then becomes

$$\textstyle\sum_{n=1}^{\mathcal{N}} \left(y_n - \sum_{k=1}^{M} \hat{\beta}_k F_{\mathrm{nk}}\right) F_{\mathrm{nj}} \to \left(y^\mathsf{T} - \hat{\beta}^\mathsf{T} F^\mathsf{T}\right) F = 0$$

and solving for $\hat{\beta}$ gives the maximum likelihood estimators

$$y^\mathsf{T} F = \hat{\beta}^\mathsf{T}(F^\mathsf{T}F) \to F^\mathsf{T} y = (F^\mathsf{T}F)\hat{\beta}$$

$$\boxed{\hat{\beta} = (F^\mathsf{T}F)^{-1}F^\mathsf{T}y}$$

Confirming Unbiased Estimation

To confirm the MLE are unbiased let us decompose the observations y_n into their true signal components $\tilde{y}$ and noise components as is done below

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \beta_1 f_1(x_1) + \beta_2 f_2(x_1) + ... + \beta_M f_M(x_1) + \epsilon_1 \\ \beta_1 f_1(x_2) + \beta_2 f_2(x_2) + ... + \beta_M f_M(x_2) + \epsilon_2 \\ \vdots \\ \beta_1 f_1(x_N) + \beta_2 f_2(x_N) + ... + \beta_M f_M(x_N) + \epsilon_N \end{pmatrix} = \mathrm{F}\beta + \epsilon = \tilde{y} + \epsilon$$

again, where β are the true values that we do not know and $\tilde{y}$ is the true value of y at each x_n without noise. We can show that MLE are unbiassed estimators of the true values by taking the expectation value

$$\left\langle \hat{\beta} \right\rangle = (F^\mathsf{T}F)^{-1}F^\mathsf{T}\langle y\rangle = (F^\mathsf{T}F)^{-1}F^\mathsf{T}\langle \mathrm{F}\beta + \epsilon\rangle = \beta \to \left\langle \hat{\beta}_n \right\rangle = \beta_n$$

Determining Distribution of $\hat{\beta}$ Values

Great so the MLE are unbiased estimators what is the distribution of possible values of $\hat{\beta}$ under our gaussian noise assumption? To answer that let's define the matrix

$$\Delta = (F^\mathsf{T}F)^{-1}F^\mathsf{T}$$

and use the fact that

$$\Delta^{-1}{}_{\mathrm{mn}}\hat{\beta}_n = y_m,$$

to determine the distribution of $\hat{\beta}_n$ given by

$$P\left(\hat{\beta}_n\right) = \textstyle\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \delta\left(\hat{\beta}_n - \beta_n - \Delta_{\mathrm{nm}}\epsilon_m\right) \frac{1}{(2\pi\sigma^2)^{N/2}} \mathrm{Exp}\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N} {\epsilon_n}^2\right] d\epsilon_1 ... d\epsilon_N$$

$$= \textstyle\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} dk \frac{1}{2\pi} \mathrm{Exp}\left[ik\hat{\beta}_n - ik\beta_n\right] \mathrm{Exp}\left[-ik\Delta_{\mathrm{nm}}\epsilon_m\right] \frac{1}{(2\pi\sigma^2)^{N/2}} \mathrm{Exp}\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N} {\epsilon_n}^2\right] d\epsilon_1 ... d\epsilon_N$$

$$= \textstyle\int_{-\infty}^{\infty} dk \frac{1}{2\pi} \mathrm{Exp}\left[ik\hat{\beta}_n - ik\beta_n\right] \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{N/2}} \mathrm{Exp}\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{N} {\epsilon_k}^2 - ik\Delta_{\mathrm{nm}}\epsilon_m\right] d\epsilon_1 ... d\epsilon_N$$

$$\textstyle\mathrm{Exp}\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{N} {\epsilon_k}^2 - ik\Delta_{\mathrm{nm}}\epsilon_m\right] = \mathrm{Exp}\left[-\frac{1}{2\sigma^2}{\epsilon_1}^2 - ik\Delta_{\mathrm{n1}}\epsilon_1\right] \mathrm{Exp}\left[-\frac{1}{2\sigma^2}{\epsilon_2}^2 - ik\Delta_{\mathrm{n2}}\epsilon_2\right] ...$$

$$= \textstyle\int_{-\infty}^{\infty} dk \frac{1}{2\pi} \mathrm{Exp}\left[-\frac{k^2}{2}\sigma^2\left(\sum_{p=1}^{N} {\Delta_{\mathrm{np}}}^2\right) + ik\left(\hat{\beta}_n - \beta_n\right)\right] = \frac{1}{\sqrt{2\pi\sigma^2\left(\sum_{p=1}^{N} {\Delta_{\mathrm{np}}}^2\right)}} \mathrm{Exp}\left[-\frac{\left(\hat{\beta}_n - \beta_n\right)^2}{2\sigma^2\left(\sum_{p=1}^{N} {\Delta_{\mathrm{np}}}^2\right)}\right]$$

$$\boxed{P\left(\hat{\beta}_n\right) = \frac{1}{\sqrt{2\pi\sigma^2\left(\sum_{p=1}^{N} {\Delta_{\mathrm{np}}}^2\right)}} \mathrm{Exp}\left[-\frac{\left(\hat{\beta}_n - \beta_n\right)^2}{2\sigma^2\left(\sum_{p=1}^{N} {\Delta_{\mathrm{np}}}^2\right)}\right]}$$

thus n-th MLE is distributed via a normal centered around the unknown true β_n value with variance given by

$$\boxed{{\sigma_{\hat{\beta}_n}}^2 = \left\langle \epsilon^2 \right\rangle \textstyle\sum_{p=1}^{N} {\Delta_{\mathrm{np}}}^2,\ \Delta = (F^\mathsf{T}F)^{-1}F^\mathsf{T}}$$

Estimating σ^2 from Residual Error

Notice above that σ^2 appears in the distribution of $\hat{\beta}$ values yet we have no way of determining exactly what $\sigma^2 = \left\langle \epsilon^2 \right\rangle$ is without knowing it a priori. We will have to estimate it. This can be accomplished via

$$\hat{\sigma}^2 = \textstyle\frac{1}{A}\sum_{n=1}^{\mathcal{N}} \left(y_n - \hat{y}(x_n)\right)^2 = \frac{1}{A}\sum_{n=1}^{\mathcal{N}} \left(y_n - \sum_{m=1}^{M} \hat{\beta}_m f_m(x_n)\right)^2,$$

where A is a yet to be determined constant. This is known as the square sum of residuals where the residual errors are given by

$$r_n = y_n - \hat{y}(x_n)$$

We seek to determine A under the requirement of unbiassed estimation. This results in,

$$\left\langle\hat{\sigma}^2\right\rangle = \tfrac{1}{A}\sum_{n=1}^{\mathcal{N}}\left\langle\left(y_n - \sum_{m=1}^{M}\hat{\beta}_m f_m\left(x_n\right)\right)^2\right\rangle$$
$$= \tfrac{1}{A}\sum_{n=1}^{\mathcal{N}}\left\langle {y_n}^2\right\rangle - 2\tfrac{1}{A}\sum_{n=1}^{\mathcal{N}}\sum_{m=1}^{M} F_{\mathrm{nm}}\left\langle y_n\hat{\beta}_m\right\rangle + \sum_{k=1}^{M}\sum_{m=1}^{M} F_{\mathrm{nk}}F_{\mathrm{nm}}\left\langle\hat{\beta}_k\hat{\beta}_m\right\rangle$$
$$= \tfrac{1}{A}\sum_{n=1}^{\mathcal{N}}\left\langle {y_n}^2\right\rangle - 2\tfrac{1}{A}\sum_{n=1}^{\mathcal{N}}\sum_{m=1}^{M} F_{\mathrm{nm}}\left\langle y_n\hat{\beta}_m\right\rangle + \sum_{k=1}^{M}\sum_{m=1}^{M} F_{\mathrm{nk}}F_{\mathrm{nm}}\left\langle\hat{\beta}_k\hat{\beta}_m\right\rangle.$$

Noting that,

$$\Delta\mathrm{F} = (F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}F = 1,$$
$$\Delta\Delta^{\mathsf{T}} = (F^{\mathsf{T}}F)^{-1},$$
$$\dim(\Delta) = M \times \mathcal{N},$$
$$\dim(F) = \mathcal{N} \times M,$$
$$\mathrm{Tr}(\Delta\mathrm{F}) = M$$

we can simplify each term of the above equation

Term 1 (using Einstein summing convention)

$$\left\langle {y_n}^2\right\rangle = \left(F_{\mathrm{nm}}\beta_m\right)^2 + \left\langle\epsilon^2\right\rangle = (\mathrm{F}\beta)_n{}^2 + \left\langle\epsilon^2\right\rangle$$

Term 2

$$\left\langle y_n\hat{\beta}_m\right\rangle = \left\langle F_{\mathrm{nk}}\beta_k\hat{\beta}_m + \epsilon_n\hat{\beta}_m\right\rangle = \left\langle F_{\mathrm{nk}}\beta_k\Delta_{\mathrm{mp}}\left(F_{\mathrm{pq}}\beta_q + \epsilon_p\right) + \epsilon_n\Delta_{\mathrm{mp}}\left(F_{\mathrm{pq}}\beta_q + \epsilon_p\right)\right\rangle$$

$$= F_{\mathrm{nk}}\beta_k\Delta_{\mathrm{mp}}F_{\mathrm{pq}}\beta_q + \Delta_{\mathrm{mp}}\left\langle\epsilon_n\epsilon_p\right\rangle$$
$$= F_{\mathrm{nk}}\beta_k\Delta_{\mathrm{mp}}F_{\mathrm{pq}}\beta_q + \Delta_{\mathrm{mn}}\left\langle\epsilon^2\right\rangle = (\mathrm{F}\beta)_n(\Delta\mathrm{F}\beta)_m + \Delta_{\mathrm{mn}}\left\langle\epsilon^2\right\rangle = (\mathrm{F}\beta)_n\beta_m + \Delta_{\mathrm{mn}}\left\langle\epsilon^2\right\rangle$$

Term 3

$$\left\langle\hat{\beta}_k\hat{\beta}_m\right\rangle = \left\langle(\Delta\mathrm{F}\beta + \Delta\epsilon)_k(\Delta\mathrm{F}\beta + \Delta\epsilon)_m\right\rangle = \left\langle\beta_k\beta_m + (\Delta\epsilon)_k\beta_m + \beta_k(\Delta\epsilon)_m + (\Delta\epsilon)_k(\Delta\epsilon)_m\right\rangle$$
$$= \beta_k\beta_m + (\Delta\Delta^{\mathsf{T}})_{\mathrm{km}}\left\langle\epsilon^2\right\rangle$$

Combining Terms (Term 1 + Term 2 + Term 3)

$$= \mathcal{N}\left\langle\epsilon^2\right\rangle - 2\left\langle\epsilon^2\right\rangle M + \mathrm{Tr}(F^{\mathsf{T}}\Delta^{\mathsf{T}}\Delta\mathrm{F})\left\langle\epsilon^2\right\rangle = (\mathcal{N} - M)\left\langle\epsilon^2\right\rangle$$

Finally giving an unbiassed estimator of the variance as,

$$\boxed{\hat{\sigma}^2 = \tfrac{1}{\mathcal{N}-M}\sum_{n=1}^{\mathcal{N}}\left(y_n - \hat{y}\left(x_n\right)\right)^2 = \tfrac{1}{(\mathcal{N}-M)}\|y - F\hat{\beta}\|^2}$$

where

$$\left\langle\hat{\sigma}^2\right\rangle = \sigma^2$$

The Variance Estimator

Alright now we can estimate σ^2 but what is the distribution of the estimator?

$$P\left(\hat{\sigma}^2\right) =?$$

Let's express the estimated variance in terms of the random variable ϵ

$$\hat{\sigma}^2 = \tfrac{1}{\mathcal{N}-M}\sum_{n=1}^{\mathcal{N}}\left(y_n - \hat{y}\left(x_n\right)\right)^2$$
$$= \tfrac{1}{\mathcal{N}-M}\sum_{n=1}^{\mathcal{N}}\left(\tilde{y}_n + \epsilon_n - \sum_{m=1}^{M} F_{\mathrm{nm}}\hat{\beta}_m\right)^2$$
$$= \tfrac{1}{\mathcal{N}-M}\sum_{n=1}^{\mathcal{N}}\left(\epsilon_n + \sum_{m=1}^{M} F_{\mathrm{nm}}\left(\beta_m - \hat{\beta}_m\right)\right)^2$$

using $\hat{\beta} = (F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}y = (F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}(\mathrm{F}\beta + \epsilon) = \beta + (F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}\epsilon$

$$= \tfrac{1}{\mathcal{N}-M}\sum_{n=1}^{\mathcal{N}}\left(\epsilon_n - \sum_{m=1}^{M} F_{\mathrm{nm}}\left(\left(F^{\mathsf{T}}F\right)^{-1}F^{\mathsf{T}}\epsilon\right)_m\right)^2$$
$$= \tfrac{1}{\mathcal{N}-M}\sum_{n=1}^{\mathcal{N}}\left(\epsilon_n - (\mathrm{F}\Delta\epsilon)_n\right)^2$$
$$= \tfrac{1}{\mathcal{N}-M}\sum_{n=1}^{\mathcal{N}}\sum_{m=1}^{\mathcal{N}}\sum_{p=1}^{\mathcal{N}}\epsilon_m\left(\delta_{\mathrm{nm}} - (\mathrm{F}\Delta)_{\mathrm{nm}}\right)\left(\delta_{\mathrm{np}} - (\mathrm{F}\Delta)_{\mathrm{np}}\right)\epsilon_p$$

using $\sum_{n=1}^{\mathcal{N}}\left(\delta_{\mathrm{nm}} - (\mathrm{F}\Delta)_{\mathrm{nm}}\right)\left(\delta_{\mathrm{np}} - (\mathrm{F}\Delta)_{\mathrm{np}}\right) = \delta_{\mathrm{mp}} - (\mathrm{F}\Delta)_{\mathrm{mp}}$

$$= \tfrac{1}{\mathcal{N}-M}\sum_{m=1}^{\mathcal{N}}\sum_{p=1}^{\mathcal{N}}\epsilon_m\left(\delta_{\mathrm{mp}} - (\mathrm{F}\Delta)_{\mathrm{np}}\right)\epsilon_p$$

or in matrix form

$= \frac{1}{\mathcal{N}-M}\epsilon^{\mathsf{T}} \cdot Q \cdot \epsilon,\ Q = 1 - \mathrm{F}\Delta$

Computing a related distribution will be more valuable soon so lets see how that works,

$P\left(\frac{\mathcal{N}-M}{\sigma^2}\hat{\sigma}^2\right) = \int \delta\left(\frac{\mathcal{N}-M}{\sigma^2}\hat{\sigma}^2 - \frac{1}{\sigma^2}\sum_{p=1}^{\mathcal{N}}\sum_{m=1}^{\mathcal{N}}\epsilon_m Q_{\mathrm{mp}}\epsilon_p\right)\frac{1}{(2\pi\sigma^2)^{\mathcal{N}/2}}\mathrm{Exp}\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{\mathcal{N}}{\epsilon_n}^2\right]d\epsilon_1...d\epsilon_{\mathcal{N}}$

$= \int \frac{dk}{2\pi}\mathrm{Exp}\left[ik\frac{\mathcal{N}-M}{\sigma^2}\hat{\sigma}^2\right]\mathrm{Exp}\left[-\frac{1}{2\sigma^2}\epsilon^{\mathsf{T}}(1+2ikQ)\epsilon\right]\frac{1}{(2\pi\sigma^2)^{\mathcal{N}/2}}d\epsilon_1...d\epsilon_{\mathcal{N}}$

$= \int \frac{dk}{2\pi}\frac{\mathrm{Exp}\left[ik\frac{\mathcal{N}-M}{\sigma^2}\hat{\sigma}^2\right]}{\mathrm{Det}[1+2ikQ]^{1/2}}$

In general the determinate in the denominator will be difficult to compute and contain terms of all powers k^1 up to $k^{\mathcal{N}}$, however if we consider which terms are most important in this integration we can see the fact that the oscillatory nature of the numerator combined with the suppressive nature of the higher powers in the denominator will leave only the smallest powers of k in the denominator as important. Lets expand around the lowest orders of k in the denominator to see how this simplifies,

$\mathrm{Det}[1+2ikQ] = \mathrm{Det}[1+2ik-2ik\mathrm{F}\Delta]$
$= \mathrm{Exp}[\mathrm{Tr}[1+2ik-2ik\mathrm{F}\Delta]]$
$= \mathrm{Det}[1+2ik]\mathrm{Exp}[-2ik\mathrm{Tr}[\mathrm{F}\Delta]]$
$= (1+2ik)^{\mathcal{N}}\mathrm{Exp}[-2ikM]$
$\approx (1+i2\mathcal{N}k)(1-i2kM)$
$= 1+i2(\mathcal{N}-M)k \approx (1+i2k)^{\mathcal{N}-M}$

this approximation is especially accurate as $1 << \mathcal{N}$, and $M << \mathcal{N}$. This replacement makes the integral easy to perform

$\approx \int \frac{dk}{2\pi}\frac{\mathrm{Exp}\left[ik\frac{\mathcal{N}-M}{\sigma^2}\hat{\sigma}^2\right]}{(1+2ik)^{(\mathcal{N}-M)/2}} = \frac{1}{2^{(\mathcal{N}-M)/2}\Gamma\left[\frac{\mathcal{N}-M}{2}\right]}\left(\frac{\mathcal{N}-M}{\sigma^2}\hat{\sigma}^2\right)^{\frac{\mathcal{N}-M}{2}-1}e^{-\frac{1}{2}\frac{\mathcal{N}-M}{\sigma^2}\hat{\sigma}^2}$
$= \chi^2\left(\frac{\mathcal{N}-M}{\sigma^2}\hat{\sigma}^2\middle|\mathcal{N}-M\right)$

Notice we still have σ^2 in this expression! We will find in the next section that this parameter will cancel out with another factor of σ^2.

Estimating MLE Confidence Intervals

If we consider the distribution of the variable, $V = \frac{(\mathcal{N}-M)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2(V|\mathcal{N}-M)$ as seen above, along with the variable, $W_n = \frac{\left(\hat{\beta}_n-\beta_n\right)}{\sigma\sqrt{\sum_{p=1}^{N}{\Delta_{\mathrm{np}}}^2}}, \sim \mathcal{N}\left(W_n|\,0,1\right)$ in the combination

$T_n = \frac{W_n}{\sqrt{V/(\mathcal{N}-M)}} = \frac{\left(\hat{\beta}_n-\beta_n\right)}{\sigma\sqrt{\sum_{p=1}^{N}{\Delta_{\mathrm{np}}}^2}}\frac{1}{\sqrt{(\mathcal{N}-M)\frac{\hat{\sigma}^2}{\sigma^2}\frac{1}{(\mathcal{N}-M)}}} = \frac{\left(\hat{\beta}_n-\beta_n\right)}{\hat{\sigma}\sqrt{\sum_{p=1}^{N}{\Delta_{\mathrm{np}}}^2}},$

we see that we can the unknown true variance σ canceled out! This t-statistic allows us state something about confidence of the true β values so long as we can solve for the distribution of this t-statistic. Let us derive the t-statistic distribution with W $\sim \mathcal{N}(W|0,1)$ and $V \sim \chi^2(V|r)$ the combined pdf is

$f(W,V) = \frac{1}{\sqrt{2\pi}}e^{-W^2/2}\frac{1}{2^{r/2}\Gamma\left[\frac{r}{2}\right]}V^{r/2-1}e^{-V/2}$

and we ask how is the combined variable $T = \frac{W}{\sqrt{V/r}}$ distributed?

$\mathcal{P}(T) = \int_{-\infty}^{\infty}dW\int_0^{\infty}dV\frac{1}{\sqrt{2\pi}}e^{-W^2/2}\frac{1}{2^{r/2}\Gamma\left[\frac{r}{2}\right]}V^{r/2-1}e^{-V/2}\ \delta\left(T-\frac{W}{\sqrt{V/r}}\right)$

a change of variables gives,

$t = W\Big/\sqrt{V/r}\,, u = V \rightarrow\ t\sqrt{\frac{u}{r}} = W, u = V$

performing the Jacobian determinate for the change of variables gives

$|J| = \begin{vmatrix} \partial W/\partial t & \partial W/\partial u \\ \partial V/\partial t & \partial V/\partial u \end{vmatrix} = \begin{vmatrix} \sqrt{u/r} & \frac{t}{2\sqrt{ur}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{u}{r}}$

thus the integral can be expressed as

$= \int_{-\infty}^{\infty}dt\int_0^{\infty}du\sqrt{\frac{u}{r}}\frac{1}{\sqrt{2\pi}}e^{-t^2\frac{u}{2r}}\frac{1}{2^{r/2}\Gamma\left[\frac{r}{2}\right]}u^{r/2-1}e^{-u/2}\ \delta(T-t)$

$= \int_0^{\infty}du\frac{1}{\sqrt{2\pi r}2^{r/2}\Gamma\left[\frac{r}{2}\right]}u^{(r+1)/2-1}e^{-\frac{T^2}{2r}u}e^{-u/2}$

implementing yet another change of variable

$z = \frac{u}{2}\left(1+\frac{T^2}{r}\right), \mathrm{dz} = \frac{1}{2}\left(1+\frac{T^2}{r}\right)\mathrm{du}$

$\mathcal{P}(T) = \frac{1}{\sqrt{\pi r}\Gamma\left[\frac{r}{2}\right]} \frac{1}{\left(1+\frac{T^2}{r}\right)^{\frac{r+1}{2}}} \int_0^\infty e^{-z} z^{\frac{1+r}{2}-1} dz$

This result of this integral is known as student's t-distribution of r degrees of freedom boxed below

$$\boxed{\mathcal{T}(T|r) = \frac{\Gamma\left[\frac{r+1}{2}\right]}{\sqrt{\pi r}\Gamma\left[\frac{r}{2}\right]} \left(1 + \frac{T^2}{r}\right)^{-\frac{(r+1)}{2}}}$$

Making our substitutions for the regression task gives

$T_n = \frac{\left(\hat{\beta}_n - \beta_n\right)}{\hat{\sigma}\sqrt{\sum_{p=1}^{N} \Delta_{\text{np}}{}^2}} \sim \mathcal{T}\left(T_n|\mathcal{N} - M\right)$

The confidence intervals for the estimators $\hat{\beta}_m$ can then be determined, by fixing $T_m = q$, giving

$\hat{\beta}_m = \beta_m \pm q\hat{\sigma}\sqrt{\sum_{n=1}^{\mathcal{N}} \Delta_{\text{mn}}{}^2},$

Thus the true β_m value lies between $\left\{\hat{\beta}_m - q\hat{\sigma}\sqrt{\sum_{n=1}^{\mathcal{N}} \Delta_{\text{mn}}{}^2}, \hat{\beta}_m + q\hat{\sigma}\sqrt{\sum_{n=1}^{\mathcal{N}} \Delta_{\text{mn}}{}^2}\right\}$ with probability of,

$\text{CI} = \int_{-q}^{q} \mathcal{T}\left(T_m|\mathcal{N} - M\right) dT_m = \frac{\Gamma\left[\frac{\mathcal{N}-M+1}{2}\right]}{\sqrt{\pi(\mathcal{N}-M)}\Gamma\left[\frac{\mathcal{N}-M}{2}\right]} \int_{-q}^{q} \left(1 + \frac{T_m{}^2}{\mathcal{N}-M}\right)^{-\frac{(\mathcal{N}-M+1)}{2}} dT_m,$

Note that if $1 \;<< (\mathcal{N} - M)$ the result becomes

$\text{CI} = \frac{1}{\sqrt{2\pi}} \int_{-q}^{q} \text{Exp}\left[-\frac{T_m{}^2}{2}\right] dT_m,$

which are the confidence intervals for that of a Normal distribution.

Prediction Confidence Intervals

Now we know how to express confidence in the MLE the predicted outcome for a point x expressed as

$\hat{y}(x) = \sum_{m=1}^{M} \hat{\beta}_m f_m(x)$

can also have confidence levels associated with it. The idea is that if one repeated this experiment many times one will get varying $\hat{\beta}_m$ values for each data set, and thus varying values of $\hat{y}(x)$. So we ask given the prediction $\hat{y}(x)$ what is the variance of that prediction if repeated experiments were done?

$\langle\hat{y}(x)\rangle = \sum_{m=1}^{M} \left\langle\hat{\beta}_m\right\rangle f_m(x)$
$= \sum_{m=1}^{M} \beta_m f_m(x)$
$= \tilde{y}(x)$

$\langle\hat{y}(x)\hat{y}(w)\rangle = \sum_{m,k=1}^{M} \left\langle\hat{\beta}_m\hat{\beta}_k\right\rangle f_m(x)f_k(w)$
$= \tilde{y}(x)\tilde{y}(w) + \left\langle\epsilon^2\right\rangle \sum_{m,k=1}^{M} (\Delta\Delta^{\mathsf{T}})_{\text{km}} f_m(x)f_k(w)$

$= \tilde{y}(x)\tilde{y}(w) + \left\langle\epsilon^2\right\rangle \sum_{m,k=1}^{M} f_k(w)(F^{\mathsf{T}}F)^{-1}{}_{\text{km}} f_m(x)$
$= \tilde{y}(x)\tilde{y}(w) + \left\langle\epsilon^2\right\rangle \mathcal{C}(w, x)$

$\mathcal{C}(w, x) = \sum_{m,k=1}^{M} f_k(w)(F^{\mathsf{T}}F)^{-1}{}_{\text{km}} f_m(x)$

$\sigma_{\hat{y}(x)}{}^2 = \left\langle\epsilon^2\right\rangle \mathcal{C}(x, x) = \left\langle\epsilon^2\right\rangle \sum_{m,k=1}^{M} f_k(x)(F^{\mathsf{T}}F)^{-1}{}_{\text{km}} f_m(x)$

What is the distribution of $\hat{y}(x)$ from repeated data samples?

$P\left[\hat{y}(x)\right] =?$

$= \int \delta\left(\hat{y}(x) - \sum_{m=1}^{M} \hat{\beta}_m f_m(x)\right) P\left(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_M\right) d\hat{\beta}_1...d\hat{\beta}_M$

$= \int \delta\left(\hat{y}(x) - \sum_{m=1}^{M} \hat{\beta}_m f_m(x)\right) \prod_{n=1}^{M} \frac{1}{\sqrt{2\pi\sigma^2\left(\sum_{p=1}^{N} \Delta_{\text{np}}{}^2\right)}} \text{Exp}\left[-\frac{\left(\hat{\beta}_n - \beta_n\right)^2}{2\sigma^2\left(\sum_{p=1}^{N} \Delta_{\text{np}}{}^2\right)}\right] d\hat{\beta}_n$

$= \int \frac{dk}{2\pi} \text{Exp}\left[ik\hat{y}(x)\right] \prod_{n=1}^{M} \frac{1}{\sqrt{2\pi\sigma^2\left(\sum_{p=1}^{N} \Delta_{\text{np}}{}^2\right)}} \text{Exp}\left[-\frac{\left(\hat{\beta}_n - \beta_n\right)^2}{2\sigma^2\left(\sum_{p=1}^{N} \Delta_{\text{np}}{}^2\right)} - ik\hat{\beta}_n f_n(x)\right] d\hat{\beta}_n$
$= \int \frac{dk}{2\pi} \text{Exp}\left[-\frac{k^2}{2}\sigma^2 \sum_{n=1}^{M} \sum_{p=1}^{N} f_n(x)^2 \Delta_{\text{np}}{}^2 - ik\left(\sum_{n=1}^{M} \beta_n f_n(x) - \hat{y}(x)\right)\right]$

$= \frac{1}{\sqrt{2\pi\sigma^2 \sum_{p=1}^{M} f_p(x)^2 (\Delta\Delta^{\mathsf{T}})_{\text{pp}}}} \text{Exp}\left[-\frac{(\hat{y}(x) - \tilde{y}(x))^2}{2\sigma^2 \sum_{p=1}^{M} f_p(x)^2 (\Delta\Delta^{\mathsf{T}})_{\text{pp}}}\right]$

We can approximate the σ^2 by $\hat{\sigma}^2$ when $1 \;<< (\mathcal{N} - M)$, to give that the true value $\tilde{y}(x)$ confidence interval

$\tilde{y}(x) = \hat{y}(x) \pm q \ \sqrt{\hat{\sigma}^2 \sum_{p=1}^{M} f_p(x)^2 (F^\intercal F)^{-1}{}_{\text{pp}}}$

with probability $p = \int_{-q}^{q} \frac{1}{\sqrt{2\pi}} \text{Exp}\left[-\frac{u^2}{2}\right] du$

$\boxed{\tilde{y}(x) = \hat{y}(x) \pm q \ \sqrt{\hat{\sigma}^2 \sum_{p=1}^{M} f_p(x)^2 (F^\intercal F)^{-1}{}_{\text{pp}}}}$

This result is the confidence internal of the true $\tilde{y}(x)$ value at x, however if we were to perform a measurement at point x we would get both the best guess of the true value $\tilde{y}(x)$ given by $\hat{y}(x)$ and we would get the measurement error on it given by the random variable ϵ, therefore we can predict that a measurement of x would result in

$\hat{y}_p(x) = \sum_{m=1}^{M} \hat{\beta}_m f_m(x) + \epsilon$, where ϵ is a random variable from $\mathcal{N}\left(0, \sigma^2\right)$.

therefore repeated observations of the value of y at x (and only repeating the measurement of y for x, not the whole data set), we would get a normal distribution for $\hat{y}(x)$ like the confidence interval but with the addition of one extra normal variable thus the pdf of the prediction for one value is

$$P\left[\hat{y}_p(x)\right] = \int \delta\left(\hat{y}_p(x) - \hat{y}(x) - \epsilon\right) P\left[\hat{y}(x)\right] \frac{1}{\sqrt{2\pi\sigma^2}} \text{Exp}\left[-\frac{\epsilon^2}{2\sigma^2}\right] d\hat{y}(x) d\epsilon$$

$$= \int \frac{dk}{2\pi} \text{Exp}\left[ik\hat{y}_p(x) - ik\hat{y}(x) - ik\epsilon\right] \frac{1}{\sqrt{2\pi\sigma^2 \sum_{p=1}^{M} f_p(x)^2 (\Delta\Delta^\intercal)_{\text{pp}}}} \text{Exp}\left[-\frac{(\hat{y}(x)-\tilde{y}(x))^2}{2\sigma^2 \sum_{p=1}^{M} f_p(x)^2 (\Delta\Delta^\intercal)_{\text{pp}}}\right] \frac{1}{\sqrt{2\pi\sigma^2}} \text{Exp}\left[-\frac{\epsilon^2}{2\sigma^2}\right] d\hat{y}(x) d\epsilon$$

$$= \int \frac{dk}{2\pi} \text{Exp}\left[-\frac{k^2}{2}\sigma^2 \left(1 + \sum_{p=1}^{M} f_p(x)^2 (\Delta\Delta^\intercal)_{\text{pp}}\right) - ik\left(\tilde{y}(x) - \hat{y}(x)\right)\right]$$

$$= \frac{1}{\sqrt{2\pi\sigma^2\left(1+\sum_{p=1}^{M} f_p(x)^2 (\Delta\Delta^\intercal)_{\text{pp}}\right)}} \text{Exp}\left[-\frac{(\hat{y}(x)-\tilde{y}(x))^2}{2\sigma^2\left(1+\sum_{p=1}^{M} f_p(x)^2 (\Delta\Delta^\intercal)_{\text{pp}}\right)}\right]$$

Again we can replace σ^2 with $\hat{\sigma}^2$ so long as $1 << \mathcal{N}$-M and using $\Delta\Delta^\intercal = (F^\intercal F)^{-1}$, the measured value will have a confidence interval given by

$$\left\{\hat{y}(x) - q\hat{\sigma}\sqrt{1 + \sum_{p=1}^{M} f_p(x)^2 (F^\intercal F)^{-1}{}_{\text{pp}}} \ , \ \hat{y}(x) + q\hat{\sigma}\sqrt{1 + \sum_{p=1}^{M} f_p(x)^2 (F^\intercal F)^{-1}{}_{\text{pp}}}\right\}$$

with q the quantile values.

Regression Quality Metric R^2

Now that we can create best fits and prediction intervals lets compute a single metric that captures how well we're modeling the data. We can do so by looking at the estimator for variance of the observed data set given as the sum of square terms below

$$\text{SST} = \sum_{n=1}^{\mathcal{N}} \left(y_n - \bar{y}\right)^2$$

This variance can be decomposed into two parts, one associated with our estimated signal and one associated to our noise estimate.

$$\text{SST} = \sum_{n=1}^{\mathcal{N}} \left(y_n - \bar{y}\right)^2$$
$$= \sum_{n=1}^{\mathcal{N}} \left(y_n - \hat{y}_n + \hat{y}_n - \bar{y}\right)^2$$
$$= \sum_{n=1}^{\mathcal{N}} \left(y_n - \hat{y}_n\right)^2 + \sum_{n=1}^{\mathcal{N}} \left(\hat{y}_n - \bar{y}\right)^2 + 2\sum_{n=1}^{\mathcal{N}} \left(y_n - \hat{y}_n\right)\left(\hat{y}_n - \bar{y}\right)$$

we can reduce this if we notice the equation for the MLE of $\hat{y}$ to satisfies

$$\sum_{n=1}^{\mathcal{N}} \left(y_n - \hat{y}_n\right) f_j\left(x_n\right) = 0 \rightarrow \sum_{j=1}^{M} \sum_{n=1}^{\mathcal{N}} \left(y_n - \hat{y}_n\right) f_j\left(x_n\right) \beta_j = 0 \rightarrow \sum_{n=1}^{\mathcal{N}} \left(y_n - \hat{y}_n\right) \hat{y}_n = 0$$

as well, if one includes a constant in your fit ie $f_0(x) = 1$ then $j = 0$ above produces

$$\sum_{n=1}^{\mathcal{N}} \left(y_n - \hat{y}_n\right) = 0$$

thus

$$2\sum_{n=1}^{\mathcal{N}} \left(y_n - \hat{y}_n\right)\left(\hat{y}_n - \bar{y}\right) = 0$$

and we can see the SST decomposes into a sum of two parts

$$\text{SST} = \sum_{n=1}^{\mathcal{N}} \left(y_n - \hat{y}_n\right)^2 + \sum_{n=1}^{\mathcal{N}} \left(\hat{y}_n - \bar{y}\right)^2 = \text{SSE} + \text{SSR}$$

the sum of square errors (SSE), and the sum of square residual deviation (SSR)

$$\text{SSE} = \sum_{n=1}^{\mathcal{N}} \left(y_n - \hat{y}_n\right)^2,$$
$$\text{SSR} = \sum_{n=1}^{\mathcal{N}} \left(\hat{y}_n - \bar{y}\right)^2$$

so long as you always include a constant in your fit ($f_0(x) = 1$) we can always do this decomposition . Notice that the SSE is an estimate of the level of noise in they system while the SSR is a measure of how much variance one would get if noise was not included in the system or the inherent variance of the signal. By forming the ratio

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

We measure the amount of deviation we believe to be inherent to the signal, a perfect fit (no noise) would give $R^2 = 1$, while no signal at all or a very weak one would result in $R^2 = 0$. The r-squared ratio can also be expressed in different ways.

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \left(1 + \frac{\text{SSE}}{\text{SSR}}\right)^{-1}$$

Below we demonstrate all the mathematics derived below for a linear fit with Gaussian noise added.

Example Mathematica Code

$\mathcal{N} = 200$; (* number of trial pts*)

$s = 0.25$; (* true variance of the noise – unknown to the analyst*)

(*the true $y(x)$function *)

f[x_]:=0.567x + 0.262

(*range of x to sample FIXED (not drawn from a random distribution)*)

xhigh = 2;

xlow = 1;

(* signal and noise generation *)

sample = Table $\left[\left\{x, f[x] + \sqrt{-2s^2\text{Log[RandomReal[]]}}\,\text{Cos}[2\pi\text{RandomReal[]]}\right\}, \left\{x, \text{xlow}, \text{xhigh}, \frac{\text{xhigh}-\text{xlow}}{\mathcal{N}}\right\}\right]$;

sample = sample[[1;;$\mathcal{N}$]];

(* not known to the analyst *)

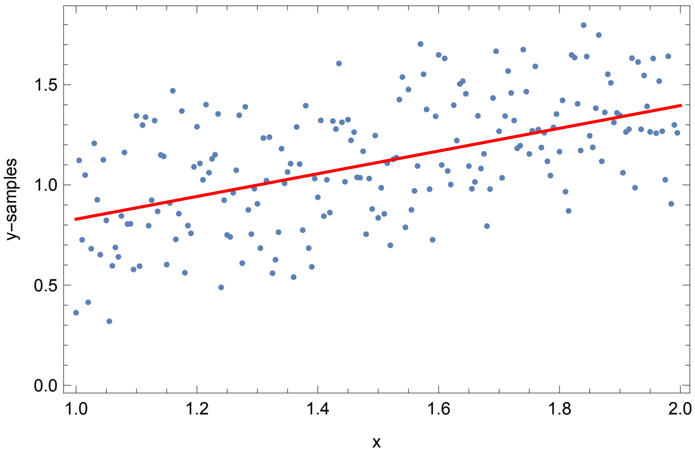
truey = Table $\left[\{x, f[x]\}, \left\{x, \text{xlow}, \text{xhigh}, \frac{\text{xhigh}-\text{xlow}}{\mathcal{N}}\right\}\right]$;

truey = truey[[1;;$\mathcal{N}$]];

(* Visualize the data with a red true curve *)

Truth = Plot[f[x], {x, xlow, xhigh}, PlotStyle → {Bold, Red}, Frame → True];

Show[ListPlot[sample, Frame → True], Truth, FrameLabel → {x, “y-samples” }]



(* Build the F matrix with M parameters to fit *)

$M = 2$;

(*M test functions to fit $f_m(x)$ *)

fm[n_, x_]:=Switch[n,

1, 1,

2, x

];

$F = \text{Table}[N[\text{fm}[m, \text{sample}[[n,1]]]], \{n, 1, \mathcal{N}\}, \{m, 1, M\}];$

(*x, y samples vector*)

$X = \text{sample}[[\text{All}, 1]];$

$y = \text{sample}[[\text{All}, 2]];$

(*constructing the inverse – Delta matrix*)

$\Delta = \text{Inverse}[F^{\mathsf{T}}.F].F^{\mathsf{T}};$

(* MLE β values *)

$\beta\text{hat} = \Delta.y;$

βhat

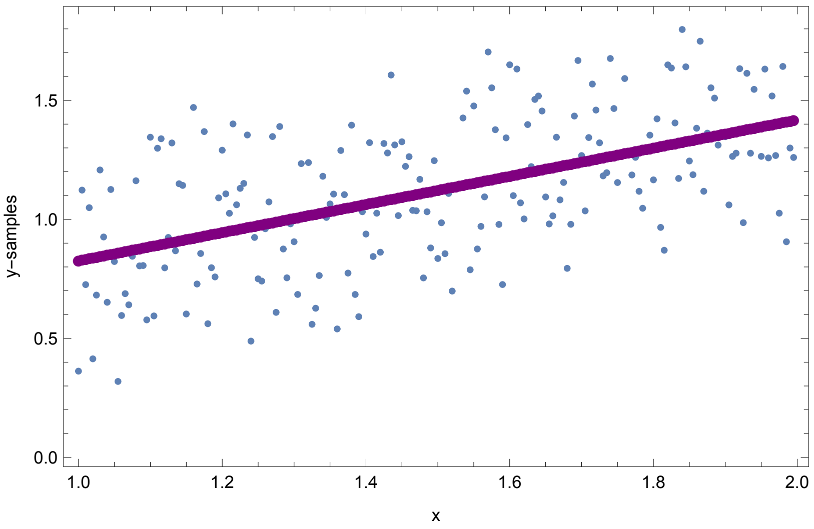
```
{0.237971, 0.591925}
```

(*prediction values $\hat{y}\,(x_n) = F.\beta$ *)
pred = $F.\beta$hat;

(* visualize the predictions in purple *)

predPlot = ListPlot[Thread[{X, pred}], PlotStyle → Purple, Frame → True, Joined → True, PlotMarkers → Automatic];

Show[ListPlot[sample, Frame → True], predPlot, FrameLabel → {x, "y-samples"}]



(*$\hat{\sigma}^2$ sample variance estimator calculation*)

$\sigma\text{hat2} = \frac{1}{\mathcal{N}-M}\text{Total}\left[(y - \text{pred})^2\right]$

```
0.06634
```

(* computing confidence intervals for the MLE *)

$q = 1;$

$\text{proba} = \frac{\text{Gamma}\left[\frac{\mathcal{N}-M+1}{2}\right]}{\sqrt{\pi(\mathcal{N}-M)}\text{Gamma}\left[\frac{\mathcal{N}-M}{2}\right]}\text{NIntegrate}\left[\left(1. + \frac{T^2}{\mathcal{N}-M}\right)^{-\frac{(\mathcal{N}-M+1)}{2}}, \{T, -q, q\}\right];$

$\beta\text{hatCI} = \text{Table}\left[\left\{\text{CI} - \hat{\beta}_m, \beta\text{hat}[[m]] - q\sqrt{\sigma\text{hat2}\sum_{n=1}^{\mathcal{N}}\Delta[[m,n]]^2}, \beta\text{hat}[[m]] + q\sqrt{\sigma\text{hat2}\sum_{n=1}^{\mathcal{N}}\Delta[[m,n]]^2}\right\}, \{m, 1, M\}\right];$
Print["Confidence Level: ", proba];

Grid[βhatCI, Frame → All, Alignment → Right]

```
Confidence Level: 0.681469
  CI − β̂1   0.141753    0.33419
  CI − β̂2   0.528834   0.655016
```

(* Computing The Prediction Intervals *)

$\text{FTFInv} = (F^{\mathsf{T}}.F)^{-1};$

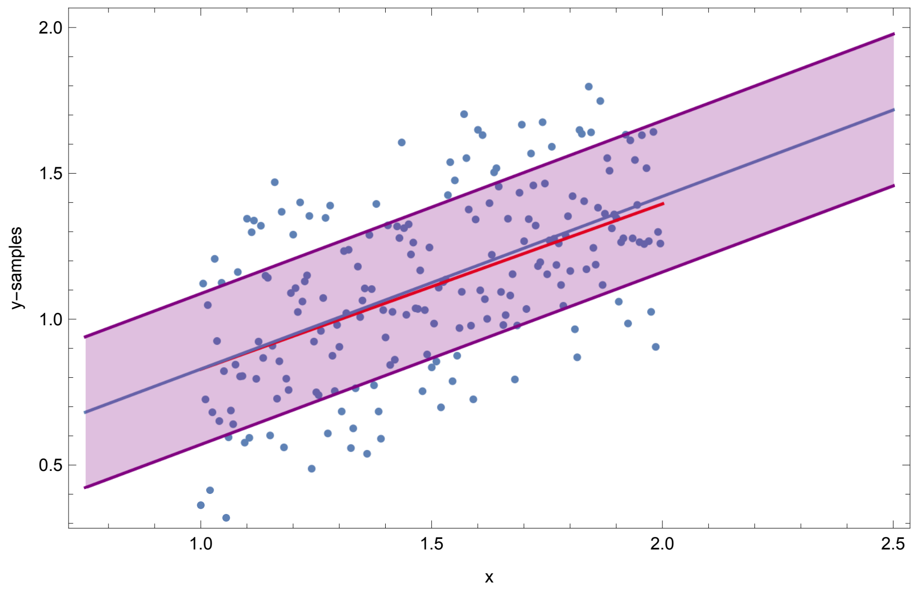
$\text{lowerbound[x\_]} := \sum_{n=1}^{M}(\beta\text{hat}[[n]]\text{fm}[n,x]) - q\sqrt{\sigma\text{hat2}\left(1 + \sum_{m=1}^{M}(\text{FTFInv}[[m,m]]\text{fm}[m,x]^2)\right)}$

$\text{upperbound[x\_]} := \sum_{n=1}^{M}(\beta\text{hat}[[n]]\text{fm}[n,x]) + q\sqrt{\sigma\text{hat2}\left(1 + \sum_{m=1}^{M}(\text{FTFInv}[[m,m]]\text{fm}[m,x]^2)\right)}$

```
(* visualizing the prediction and bounds *)
boundplot = Plot[{lowerbound[x], upperbound[x]}, {x, xlow(0.75), xhigh(1.25)},
Filling → {1 → {2}},
FillingStyle → Directive[Purple, Opacity[0.25]],
PlotStyle → Purple];

smoothexpectation = Plot [∑_{n=1}^{M} (βhat[[n]]fm[n, x]), {x, xlow(0.75), xhigh(1.25)}] ;

Show[ListPlot[sample, Frame → True],
Truth,
smoothexpectation,
boundplot,
FrameLabel → {x, "y-samples"},
PlotRange → All, Axes → False]
```



```
(*Calculating R^2  Metric*)
SST = Total [(y − Mean[y])^2] ;
SSE = σhat2(𝒩 − M);
SSR = Total [(pred − Mean[y])^2] ;

Grid[{{"SST", "SSE", "SSR", "SSE+SSR"}, {SST, SSE, SSR, SSE + SSR}}, Alignment → Left, Frame → All]
Print ["R^2:  ", R2 = 1 − SSE/SST]
```

```
SST      SSE      SSR      SSE+SSR
18.9748  13.1353  5.83944  18.9748
```

R^2: 0.307748