CS 760: Machine Learning Fall 2020

Topic 5: Linear Regression

Instructor: Daniel L. Pimentel-Alarcón

In memory of my friend John Brady, who taught me, among many other things, how to do figures right.

GO GREEN. AVOID PRINTING, OR PRINT 2-SIDED OR MULTIPAGE.

5.1 Introduction

One of the most elemental problems in machine learning can be summarized as predicting the value of a *response* y as a function of other *features* $\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{D}$. For example:

- Predicting my glucose level (response) as a function of my height, weight, age, and gender (features).
- Predicting stock prices (response) as a function of the market state (features).
- Predicting the activation level of a gene that determines a disease, like cancer (response) as a function of other genes' activation levels (features); this is often known as genomics wide association studies (GWAS).
- Predicting magnitude of solar flares (response) as a function of solar images (features).

The main idea behind linear regression is to approximate y as a linear combination of $\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{D}$, i.e.

$$\mathrm{y} \;\approx\; \theta_0 + \theta_1 \mathrm{x}_1 + \theta_2 \mathrm{x}_2 + \cdots + \theta_\mathrm{D} \mathrm{x}_\mathrm{D}, \tag{5.1}$$

where θ_0 is essentially an *offset*, and $\theta_1, \theta_2, \ldots, \theta_\mathrm{D}$ are the weights of each feature. For instance, in our glucose example, (5.1) is essentially saying:

$$\text{glucose level} \;\approx\; \theta_0 + \theta_1 \text{height} + \theta_2 \text{weight} + \theta_3 \text{age} + \theta_4 \text{gender}.$$

Notice that by letting $\mathbf{x} = [\,1 \;\; \mathrm{x}_1 \;\; \mathrm{x}_2 \;\; \cdots \;\; \mathrm{x}_\mathrm{D}]^\mathsf{T}$ and $\boldsymbol{\theta} = [\theta_0 \;\; \theta_1 \;\; \theta_2 \;\; \cdots \;\; \theta_\mathrm{D}]^\mathsf{T}$ we can rewrite (5.1) in vector form as

$$\mathrm{y} \;\approx\; \mathbf{x}^\mathsf{T} \boldsymbol{\theta}. \tag{5.2}$$

The goal is to determine the weights vector $\boldsymbol{\theta}$ that best explains y as a function of $\mathbf{x}$. Effectively, this equates to finding the function f in the family of functions of the form $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\boldsymbol{\theta}$ that best predicts y.

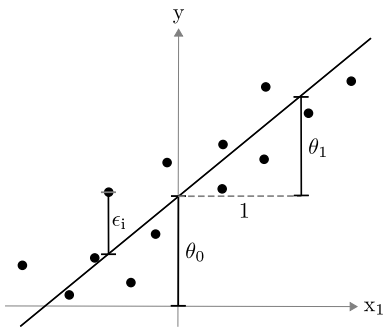
5.2 Learning θ by MSE Minimization

The main idea is to find the vector $\boldsymbol{\theta}$ that *best explains* the *training labels* $\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{N}$ as a function of the *training feature vectors* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\mathrm{N}$. There are several ways to interpret what it means to *best explain*.

One of the most common ones is as minimizing the *mean squared error* (MSE), i.e.,

$$\hat{\boldsymbol{\theta}} \;=\; \underset{\boldsymbol{\theta}\in\mathbb{R}^{D+1}}{\arg\min}\; \frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2. \tag{5.3}$$

Intuitively, $\hat{\boldsymbol{\theta}}$ is the *line* that best explains the y_i's as function of the $\mathbf{x}_i$'s. This is illustrated in the following figure, where each point represents a pair $(\mathbf{x}_i, y_i)$, and ϵ_i represents the *error* on the i^{th} sample:



To solve (5.3), rewrite it as

$$\hat{\boldsymbol{\theta}} \;=\; \underset{\boldsymbol{\theta}\in\mathbb{R}^{D+1}}{\arg\min}\; \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2, \tag{5.4}$$

where $\mathbf{y} = [y_1 \;\; y_2 \;\; \cdots \;\; y_N]^T$, $\mathbf{X} = [\mathbf{x}_1 \;\; \mathbf{x}_2 \;\; \cdots \;\; \mathbf{x}_N]^T$, and the $1/N$ factor is removed because it does not affect the minimization. Then notice that:

$$\begin{aligned}
\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 \;&=\; (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \\
&=\; \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} \\
&=\; \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}.
\end{aligned}$$

To obtain the derivative with respect to $\boldsymbol{\theta}$ (to learn more about how to take derivatives w.r.t. vectors and matrices see *Old and new matrix algebra useful for statistics* by Thomas P. Minka), first compute the differential:

$$\begin{aligned}
\mathrm{d}\,\mathsf{tr}\left(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}\right) \;&=\; -2\mathsf{tr}\left((\mathrm{d}\boldsymbol{\theta})^T\mathbf{X}^T\mathbf{y}\right) + \mathsf{tr}\left((\mathrm{d}\boldsymbol{\theta})^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}\right) + \mathsf{tr}\left(\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}(\mathrm{d}\boldsymbol{\theta})\right) \\
&=\; -2\mathsf{tr}\left((\mathrm{d}\boldsymbol{\theta})^T\mathbf{X}^T\mathbf{y}\right) + 2\mathsf{tr}\left((\mathrm{d}\boldsymbol{\theta})^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}\right) \\
&=\; 2\mathsf{tr}\left((\mathrm{d}\boldsymbol{\theta})^T(\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - \mathbf{X}^T\mathbf{y})\right).
\end{aligned}$$

It follows that

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} \;=\; 2\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - 2\mathbf{X}^T\mathbf{y}.$$

Setting this derivative to zero and solving for $\boldsymbol{\theta}$ we conclude that the solution to (5.3) is

$$\hat{\boldsymbol{\theta}} \;=\; (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{5.5}$$

which is essentially the coefficient of the projection of $\mathbf{y}$ onto $\mathrm{span}\{\mathbf{X}\}$.

5.3 Learning θ by Likelihood Maximization

A Note on Likelihood

Recall that a probability distribution $\mathbb{P}(y|\boldsymbol{\theta})$ determines the frequency with which that a random variable y takes each value, given some parameter $\boldsymbol{\theta}$. For example, if $y \sim \text{Bernoulli}(\theta)$, with $\theta = 1/2$, then the probability that y takes the value 1 is $\mathbb{P}(y = 1|\theta) = \theta = 1/2$.

Conversely, the *likelihood* $\mathbb{P}(\mathrm{y}|\boldsymbol{\theta})$ determines the probability that a parameter $\boldsymbol{\theta}$ was the one that generated a sample y. We emphasize this distinction using y instead of y, to indicate that y is already known, i.e., observed data that has already taken a specific value. Under the same Bernoulli example, if we observe $\mathrm{y} = 1$, then the likelihood of the parameter θ is $\mathbb{P}(\mathrm{y} = 1|\theta) \;=\; \theta$.

The probability and the likelihood may *look* a lot alike. The difference is very subtle, and mainly conceptually: the probability $\mathbb{P}(y|\boldsymbol{\theta})$ is a function where y is the variable, and $\boldsymbol{\theta}$ is fixed. In contrast, the likelihood $\mathbb{P}(\mathrm{y}|\boldsymbol{\theta})$ is a function where $\boldsymbol{\theta}$ is the variable, and y is fixed. We use $\mathbb{P}(y|\boldsymbol{\theta})$ when we know $\boldsymbol{\theta}$ and want to guess y; we use $\mathbb{P}(\mathrm{y}|\boldsymbol{\theta})$ when we have already observed data with the specific value y, and we want to guess the parameter $\boldsymbol{\theta}$ that generated it.

Example 5.1. Suppose $y_1, \ldots, y_6$ are *independently and identically distributed* (i.i.d.) according to a Bernoulli($1/4$) distribution. Then the probability that $y_1 = y_2 = y_3 = 1$, and $y_4 = y_5 = y_6 = 0$ is:

$$\begin{aligned}\mathbb{P}(y_1 = y_2 = y_3 = 1, y_4 = y_5 = y_6 = 0|\theta) \;&=\; \prod_{\mathrm{i}=1}^{3} \mathbb{P}(y_\mathrm{i} = 1|\theta) \cdot \prod_{\mathrm{i}=4}^{6} \mathbb{P}(y_\mathrm{i} = 0|\theta) \\ &=\; \theta^3(1-\theta)^3 \;=\; (1/4)^3\,(3/4)^3.\end{aligned}$$

Instead, suppose that we observe $\mathrm{y}_1 = \mathrm{y}_2 = \mathrm{y}_3 = 1$, and $\mathrm{y}_4 = \mathrm{y}_5 = \mathrm{y}_6 = 0$. Then the likelihood of θ under this sample is:

$$\begin{aligned}\mathbb{P}(\mathrm{y}_1 = \mathrm{y}_2 = \mathrm{y}_3 = 1, \mathrm{y}_4 = \mathrm{y}_5 = \mathrm{y}_6 = 0|\theta) \;&=\; \prod_{\mathrm{i}=1}^{3} \mathbb{P}(\mathrm{y}_\mathrm{i} = 1|\theta) \cdot \prod_{\mathrm{i}=4}^{6} \mathbb{P}(\mathrm{y}_\mathrm{i} = 0|\theta) \\ &=\; \theta^3(1-\theta)^3.\end{aligned}$$

Based on this sample, which would be your intuitive best guess at the value of θ? Is this the same value that maximizes the likelihood $\mathbb{P}(\mathrm{y}_1, \ldots, \mathrm{y}_6|\theta)$?

Likelihood in Linear Regression

Recall that our main goal is to find the vector $\boldsymbol{\theta}$ that *best explains* the training pairs $\{\mathbf{x}_\mathrm{i}, \mathrm{y}_\mathrm{i}\}_{\mathrm{i}=1}^{\mathrm{N}}$. Another common interpretation of *best explaining* is as maximizing the likelihood of the sample, under the probabilistic model

$$y_\mathrm{i} \;=\; \mathbf{x}_\mathrm{i}^\mathsf{T}\boldsymbol{\theta}^\star + \epsilon_\mathrm{i} \qquad\qquad \text{for every i,} \tag{5.6}$$

where $\epsilon_\mathrm{i} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ models a random error, and $\boldsymbol{\theta}^\star$ denotes the unknown *true* parameter that would perfectly describe y in terms of $\mathbf{x}$. Our goal is to estimate $\boldsymbol{\theta}^\star$. Letting again $\boldsymbol{y} = [y_1 \;\; y_2 \;\; \cdots \;\; y_\mathrm{N}]^\mathsf{T}$ and $\mathbf{X} = [\mathbf{x}_1 \;\; \mathbf{x}_2 \;\; \cdots \;\; \mathbf{x}_\mathrm{N}]^\mathsf{T}$, and defining $\boldsymbol{\epsilon} = [\epsilon_1 \;\; \epsilon_2 \;\; \cdots \;\; \epsilon_\mathrm{N}]^\mathsf{T}$, we can rewrite (5.6) as:

$$\boldsymbol{y} \;=\; \mathbf{X}\boldsymbol{\theta}^\star + \boldsymbol{\epsilon}. \tag{5.7}$$

It follows that $\boldsymbol{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta}^\star, \sigma^2\mathbf{I})$; make sure to know why. Then the likelihood of our sample is

$$\mathbb{P}(\mathbf{y}, \mathbf{X}|\boldsymbol{\theta}, \sigma) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}^{\mathrm{N}}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})},$$

and our goal is to find the parameter $\boldsymbol{\theta}$ that maximizes this likelihood, i.e.,

$$\begin{aligned}
\hat{\boldsymbol{\theta}} \;&=\; \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\mathbb{R}^{\mathrm{D}+1}} \; \mathbb{P}(\mathbf{y}, \mathbf{X}|\boldsymbol{\theta}, \sigma) \\
&=\; \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\mathbb{R}^{\mathrm{D}+1}} \; \frac{1}{\sqrt{2\pi\sigma^2}^{\mathrm{N}}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})} \\
&=\; \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\mathbb{R}^{\mathrm{D}+1}} \; -(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}) \\
&=\; \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^{\mathrm{D}+1}} \; \|\mathbf{y}-\mathbf{X}\boldsymbol{\theta}\|_2^2,
\end{aligned} \tag{5.8}$$

which we recognize as the same problem as (5.4). It follows that the solution to (5.8) is given by (5.5). In other words, the *maximum likelihood estimator* (MLE) is the same as the minimizer of the MSE from Section 5.2. We point out that in general, different optimality criteria (e.g., minimizing MSE and maximizing likelihood) may produce different optimal parameters.

5.4 Confidence

Recall that our ultimate goal is to predict y as a function of $\mathbf{x}$. At this point we have identified the *line* determined by $\hat{\boldsymbol{\theta}}$ that best explains the training labels $\mathbf{y}$ as a function of the training features $\mathbf{X}$. Given a new $\mathbf{x}$, the linear prediction of its corresponding y is given by

$$\hat{\mathrm{y}} \;=\; \mathbf{x}^\mathsf{T}\hat{\boldsymbol{\theta}} = \mathbf{x}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}.$$

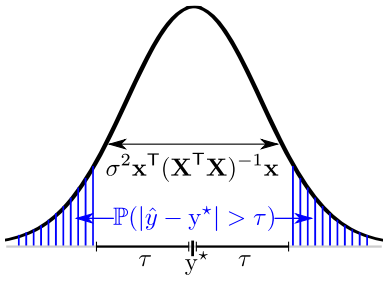
The next question we should be asking is whether we can trust our prediction $\hat{\mathrm{y}}$. To discover this, observe that under the probabilistic model in (5.7) (make sure you understand the difference between the *estimator* $\hat{y}$ and the *estimate* $\hat{\mathrm{y}}$),

$$\hat{y} \;\sim\; \mathcal{N}(\mathbf{x}^\mathsf{T}\boldsymbol{\theta}^\star, \sigma^2\mathbf{x}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}).$$

In words, this means that $\hat{y}$ is a Normal random variable centered around the *true* $\mathrm{y}^\star := \mathbf{x}^\mathsf{T}\boldsymbol{\theta}^\star$, and has variance $\sigma^2\mathbf{x}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}$. Hence, the probability that $\hat{y}$ is τ-away from the *true* $\mathrm{y}^\star$ is

$$\mathbb{P}(|\hat{y}-\mathrm{y}^\star| > \tau) \;=\; 2\,\Phi\big(\,\tau \;\big|\; 0, \sigma^2\mathbf{x}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}\big),$$

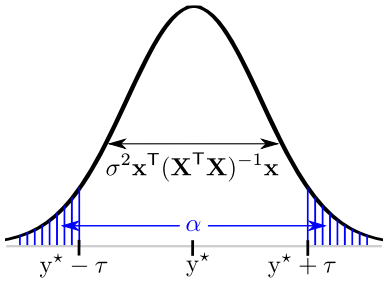
where $\Phi(\,\cdot \mid \mu, \nu)$ is the tail function of the $\mathcal{N}(\mu, \nu)$ distribution:



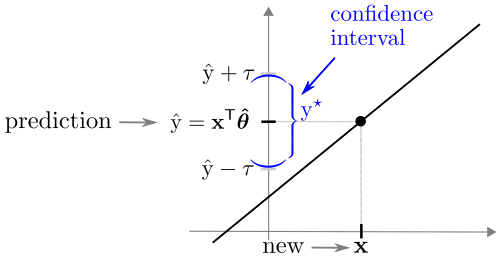
Conversely, given a desired significance level α (typically set to 0.05), we can find a threshold

$$\tau \;=\; \Phi^{-1}\big(\,\alpha/2 \;\big|\; 0, \sigma^2 \mathbf{x}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}\big)$$

such that $\mathbb{P}(|\hat{y} - \mathrm{y}^\star| \le \tau) = 1 - \alpha$:



Equivalently, we conclude that with probability $1 - \alpha$ (typically set to 0.95), the *true* (but unknown) $\mathrm{y}^\star$ is in the *confidence interval* $(\hat{\mathrm{y}} - \tau, \hat{\mathrm{y}} + \tau)$:



5.5 Learning Significant Features

Another question we may ask is which features are relevant, and which are not. This can be determined by the coefficients in $\boldsymbol{\theta}^\star$. If $\theta^\star_\mathrm{j} = 0$, we can conclude that the j^{th} feature is irrelevant, and if $\theta^\star_\mathrm{j} \neq 0$ we can conclude that the j^{th} feature is significant. Unfortunately, we do not know $\boldsymbol{\theta}^\star$. However, we do know that its MLE:

$$\hat{\boldsymbol{\theta}} \;\sim\; \mathcal{N}(\boldsymbol{\theta}^\star, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}).$$

Letting ν^2_j denote the $(\mathrm{j},\mathrm{j})^{\mathrm{th}}$ entry in the covariance matrix $\sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$, we can pose our question as a hypothesis test:

$$\begin{aligned} H_0:\ \hat{\theta}_\mathrm{j} \;&\sim\; \mathcal{N}(\theta^\star_\mathrm{j}, \nu^2_\mathrm{j}), \quad \theta^\star_\mathrm{j} = 0 && \Rightarrow \mathrm{j}^{\mathrm{th}} \text{ feature is irrelevant,} \\ H_1:\ \hat{\theta}_\mathrm{j} \;&\sim\; \mathcal{N}(\theta^\star_\mathrm{j}, \nu^2_\mathrm{j}), \quad \theta^\star_\mathrm{j} \neq 0 && \Rightarrow \mathrm{j}^{\mathrm{th}} \text{ feature is significant.} \end{aligned}$$

Here our goal is to decide between H_0 and H_1. If your hunch is to simply pick which ever is more likely, then your intuition is correct. That is the main idea of the *likelihood ratio test* (LRT):

$$\frac{\mathbb{P}(\hat{\theta}_\mathrm{j} | \theta^\star_\mathrm{j} \neq 0)}{\mathbb{P}(\hat{\theta}_\mathrm{j} | \theta^\star_\mathrm{j} = 0)} \underset{H_0}{\overset{H_1}{\gtrless}} 1. \tag{5.9}$$

Here $\mathbb{P}(\hat{\theta}_j|\theta_j^\star \neq 0)$ denotes the likelihood of $\hat{\theta}_j$ under H_1, and similarly $\mathbb{P}(\hat{\theta}_j|\theta_j^\star = 0)$ denotes the likelihood of $\hat{\theta}_j$ under H_0. In words, (5.9) decides H_1 if the *likelihood ratio* $\Lambda(\hat{\theta}_j) := \frac{\mathbb{P}(\hat{\theta}_j|\theta_j^\star \neq 0)}{\mathbb{P}(\hat{\theta}_j|\theta_j^\star \neq 0)}$ is larger than 1 (meaning the likelihood under H_1 is larger than under H_0), and decides H_0 otherwise. However, since we do not know the specific value of $\theta_j^\star$ under H_1, we cannot compute (5.9) directly. Instead we have to use a *generalized likelihood ratio test* (GLRT):

$$\frac{\max\limits_{\theta_j^\star \neq 0} \mathbb{P}(\hat{\theta}_j|\theta_j^\star)}{\mathbb{P}(\hat{\theta}_j|\theta_j^\star = 0)} \underset{H_0}{\overset{H_1}{\gtrless}} \tau, \tag{5.10}$$

where τ can be chosen to bound the probability of a certain type of error (e.g., deciding H_1 when H_0 is true, often called Type 1 error) or guarantee the probability of a correct decision (e.g., correctly rejecting H_0). The main idea behind (5.10) is to substitute $\theta_j^\star$ with its MLE, which we already know from before to be $\hat{\theta}_j$. Then our test becomes:

$$\Lambda(\hat{\theta}_j) = \frac{\mathbb{P}(\hat{\theta}_j|\theta_j^\star = \hat{\theta}_j)}{\mathbb{P}(\hat{\theta}_j|\theta_j^\star = 0)} = \frac{\frac{1}{\sqrt{2\pi}\nu_j} e^{-\frac{1}{2}\left(\frac{\hat{\theta}_j - \hat{\theta}_j}{\nu_j}\right)^2}}{\frac{1}{\sqrt{2\pi}\nu_j} e^{-\frac{1}{2}\left(\frac{\hat{\theta}_j - 0}{\nu_j}\right)^2}} = e^{\frac{\hat{\theta}_j^2}{2\nu_j^2}} \underset{H_0}{\overset{H_1}{\gtrless}} \tau.$$

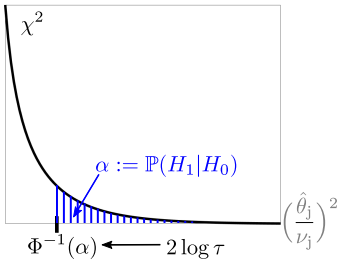
Taking log and with minor algebra manipulations we can further simplify our test into:

$$\left(\frac{\hat{\theta}_j}{\nu_j}\right)^2 \underset{H_0}{\overset{H_1}{\gtrless}} 2\log\tau.$$

Under H_0, $\hat{\theta}_j \sim \mathcal{N}(0, \nu_j^2)$, so $\hat{\theta}_j/\nu_j \sim \mathcal{N}(0,1)$, which implies $(\hat{\theta}_j/\nu_j)^2 \sim \chi^2$. Given a desired significance level α (probability of deciding H_1 given that H_0 is true, typically set to 0.05), we can select τ as:

$$\tau = e^{\frac{1}{2}\Phi^{-1}(\alpha)},$$

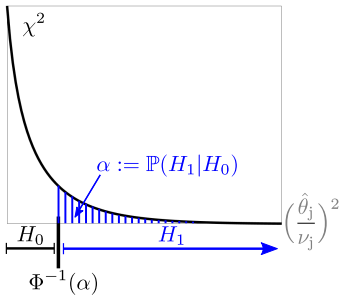
where Φ is the tail function of the χ^2 distribution:



With this, our test further simplifies into:

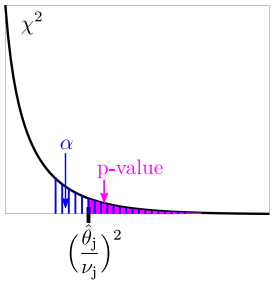
$$\left(\frac{\hat{\theta}_j}{\nu_j}\right)^2 \underset{H_0}{\overset{H_1}{\gtrless}} \Phi^{-1}(\alpha) \tag{5.11}$$

In words, (5.11) says: decide H_1 if $\hat{\theta}_j^2 > \nu_j^2 \Phi^{-1}(\alpha)$, and decide H_0 otherwise, which matches our intuition, essentially saying: if $|\hat{\theta}_j|$ is large enough, conclude that $\theta_j^\star \neq 0$, and that the j^{th} feature is significant; conversely, if $|\hat{\theta}_j|$ is too small, conclude that $\theta_j^\star = 0$, and that the j^{th} feature is irrelevant:



Finally, given an instance of the test statistic $(\hat{\theta}_j/\nu_j)^2$, its p-value (indicating the probability of observing a larger test statistic under H_0) is

$$\Phi\Big(\frac{\hat{\theta}_j^2}{\nu_j^2}\Big).$$



5.6 Linear Regression Recipe

To summarize, here is how we would use linear regression in a typical scenario:

- Collect *labels* $\mathbf{y} \in \mathbb{R}^N$ and *features* $\mathbf{X} \in \mathbb{R}^{N \times D}$ from N samples.
- Compute $\hat{\boldsymbol{\theta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$.
- Given a new sample $\mathbf{x}$, predict $\hat{y} = \mathbf{x}^\mathsf{T}\hat{\boldsymbol{\theta}} = \mathbf{x}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$.
- Given a significance level α (typically set to 0.05), we know that with probability $1-\alpha$, the *true* $y^\star$ lies in the confidence interval $(\hat{y}-\tau, \hat{y}+\tau)$, where $\tau = \Phi^{-1}\big(\,\alpha/2 \;\big|\; 0, \sigma^2\mathbf{x}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}\big)$, and $\Phi(\,\cdot \mid \mu, \nu)$ is the tail function of the $\mathcal{N}(\mu, \nu)$ distribution.
- With p-value $\Phi(\hat{\theta}_j^2/\nu_j^2)$, conclude that the j^{th} feature is significant if $\hat{\theta}_j^2 > \nu_j^2\Phi^{-1}(\alpha)$, and decide it is irrelevant otherwise; here ν_j^2 is the $(j,j)^{\text{th}}$ entry in $\mathsf{cov}(\hat{\boldsymbol{\theta}}) = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$.

5.7 Estimating σ

Notice that our confidence interval and test both depend on σ, which in general may be unknown. To estimate it, recall that by definition, $\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta}^\star \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Then

$$\begin{aligned}\hat{\sigma} &:= \arg\max_{\sigma\in\mathbb{R}} \ \mathbb{P}(\mathbf{y}, \mathbf{X}|\boldsymbol{\theta}^\star, \sigma) \\ &= \arg\max_{\sigma\in\mathbb{R}} \ \frac{1}{\sqrt{2\pi\sigma^2}^{\mathrm{N}}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}^\star)^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}^\star)} \\ &= \arg\max_{\sigma\in\mathbb{R}} \ -\frac{\mathrm{N}}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}^\star)^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}^\star).\end{aligned}$$

Taking derivative we have:

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}\log\mathbb{P}(\mathbf{y}, \mathbf{X}|\boldsymbol{\theta}^\star, \sigma) \ = \ -\frac{\mathrm{N}}{\sigma} + \frac{1}{\sigma^3}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}^\star)^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}^\star).$$

Setting to zero and solving for σ^2 we conclude:

$$\hat{\sigma}^2 \ = \ \frac{1}{\mathrm{N}}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}^\star)^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}^\star) \ = \ \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{y}_\mathrm{i} - \mathbf{x}_\mathrm{i}^\mathsf{T}\boldsymbol{\theta}^\star)^2, \tag{5.12}$$

which is the classic MLE of σ^2 for the i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables $\epsilon_\mathrm{i} = \mathrm{y}_\mathrm{i} - \mathbf{x}_\mathrm{i}^\mathsf{T}\boldsymbol{\theta}^\star$. Notice, however, that (5.12) depends on $\boldsymbol{\theta}^\star$, which we do not know and had to estimate. Luckily, the following theorem shows us that the MLE of a function is the function of the MLE:

Theorem 5.1 (Invariance of the MLE)**.** Let $\epsilon_1, \ldots, \epsilon_\mathrm{N} \overset{iid}{\sim} \mathbb{P}(\epsilon|\boldsymbol{\theta}^\star)$. Let $\sigma^\star = g(\boldsymbol{\theta}^\star)$ for some surjective (a.k.a. onto) function $g : \mathcal{B} \twoheadrightarrow \Gamma$. Then the MLE of $\sigma^\star$, defined as

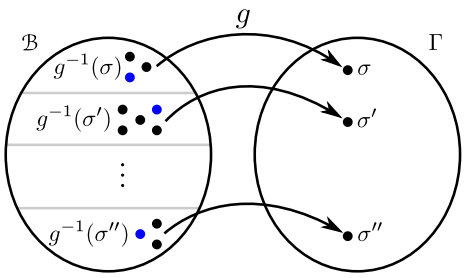
$$\hat{\sigma} \ := \ \arg\max_{\sigma\in\Gamma}\Big(\max_{\boldsymbol{\theta}\in g^{-1}(\sigma)}\prod_{\mathrm{i}=1}^{\mathrm{N}}\mathbb{P}(\epsilon_\mathrm{i}, \boldsymbol{\theta})\Big)$$

is given by $\hat{\sigma} = g(\hat{\boldsymbol{\theta}})$, where

$$\hat{\boldsymbol{\theta}} \ := \ g\Big(\arg\max_{\boldsymbol{\theta}\in\mathcal{B}}\prod_{\mathrm{i}=1}^{\mathrm{N}}\mathbb{P}(\epsilon_\mathrm{i}|\boldsymbol{\theta})\Big)$$

is the MLE of $\boldsymbol{\theta}^\star$. Here g^{-1} denotes the inverse image of g, i.e., $g^{-1}(\sigma) = \{\boldsymbol{\theta} \in \mathcal{B} : g(\boldsymbol{\theta}) = \sigma\}$.

Proof. Since g is onto, even if g is not injective (a.k.a. one-to-one), the sets $\{g^{-1}(\sigma)\}_{\sigma\in\Gamma}$ form a partition of $\mathcal{B}$:



Therefore,

$$\bigcup_{\sigma\in\Gamma} g^{-1}(\sigma) \;=\; \mathcal{B},$$

which in turns implies

$$\max_{\sigma\in\Gamma}\Big(\max_{\boldsymbol{\theta}\in g^{-1}(\sigma)}\prod_{\mathrm{i}=1}^{\mathrm{N}}\mathbb{P}(\epsilon_{\mathrm{i}}|\boldsymbol{\theta})\Big) \;=\; \max_{\theta\in\mathcal{B}}\prod_{\mathrm{i}=1}^{\mathrm{N}}\mathbb{P}(\epsilon_{\mathrm{i}}|\boldsymbol{\theta}).$$

□

It follows directly that the MLE of σ is obtained by replacing $\boldsymbol{\theta}^\star$ with $\hat{\boldsymbol{\theta}}$ in (5.12), i.e.:

$$\hat{\sigma}^2 \;=\; \frac{1}{\mathrm{N}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\theta}})^{\mathsf{T}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\theta}}) \;=\; \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{y}_{\mathrm{i}}-\mathbf{x}_{\mathrm{i}}^{\mathsf{T}}\hat{\boldsymbol{\theta}})^2,$$