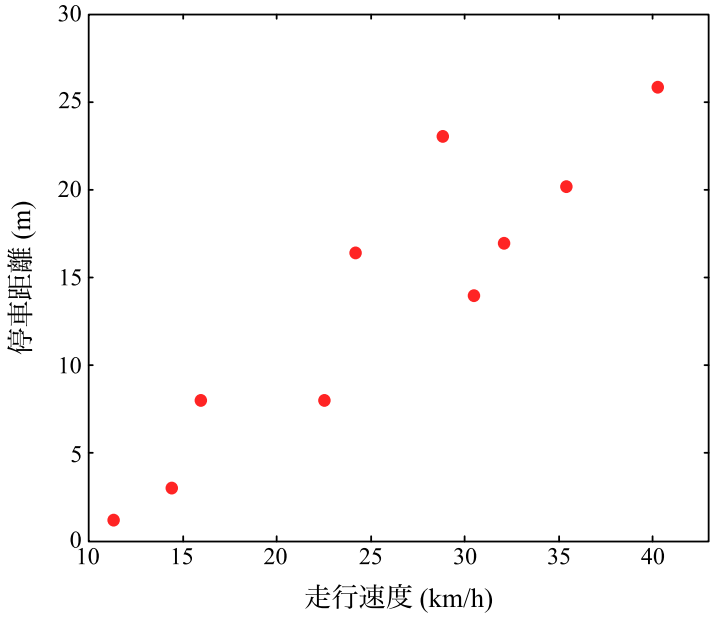
条件付き対数尤度と平均二乗誤差

ここでは、最尤法 (または最尤推定) と呼ばれる方法について説明します。この方法は様々な場面で使える汎用性の高い考え方ですが、ここでは最小二乗法に対して統計的な観点から根拠を与えることができることを説明します。

♣ 「走行速度」と「停車距離」

ここでは自動車の「走行速度」と「停車距離」の関係を例に、最小二乗法と最尤法の関係を説明していきます。次のようなデータがあったとしましょう。



自動車の走行速度と、それぞれの走行速度での停車距離 (「車を停止させよう」と感じてからブレーキを踏み、そこから車が実際に停止するまでの距離) に関する 10 組のデータ

$$\begin{aligned}\left(x^{(1)}, y^{(1)}\right) &= (11.31,\ 1.21)\\ &\vdots\\ \left(x^{(10)}, y^{(10)}\right) &= (40.28,\ 25.86)\end{aligned}$$

をプロットしています。後で使うので、次のような量を用意しておきます[1]。

$$\boldsymbol{X} = \left(x^{(1)}, x^{(2)}, \cdots, x^{(10)}\right)$$

$$\boldsymbol{Y} = \left(y^{(1)}, y^{(2)}, \cdots, y^{(10)}\right)$$

[1] 「ベクトル表記」と呼ばれる表記法です。太文字にして、ベクトルであること表しています。

$\boldsymbol{X}$ は自動車の走行速度、$\boldsymbol{Y}$ は停車距離の一覧となっています。プログラミング風に言えば、$\boldsymbol{X}$ と $\boldsymbol{Y}$ はそれぞれ、走行速度と停車距離の数値を格納した長さが 10 の 1 次元「配列」です。

以下では、走行速度 $\hat{x}$ と停車距離 $\hat{y}$ の間にある関係を、データ $\boldsymbol{X}$、$\boldsymbol{Y}$ から見出していきます。ここで、$\hat{x}$ と $\hat{y}$ は走行速度と停車距離を表す「変数」です。実際に測定されたデータは 10 点なので、現時点では 10 個の走行速度に対する停車距離しか分かっていません。しかし、より一般に様々な走行速度に対する停車距離が分かると便利ですよね。もちろん何回も測定すれば良いですが、現実的には費用や時間の問題で測定できる回数は限られていることが多いででしょう。そこで、走行速度と停車距離を $\hat{x}$ と $\hat{y}$ という変数で表し、測定データ $\boldsymbol{X}$、$\boldsymbol{Y}$ から $\hat{x}$ と $\hat{y}$ の関係を予想、任意の速度 $\hat{x}$ に対する停車距離 $\hat{y}$ を推定する[2]、ということを行います。

まず、データをプロットしたグラフを見ると、走行速度 $\hat{x}$ が増加すると停車距離 $\hat{y}$ も増加する傾向にあります[3]。そこで、$\hat{x}$ と $\hat{y}$ との間に次の関係が成り立つと仮説を立ててみます。

$$\hat{y} = \theta_0 + \theta_1 \hat{x}$$

このように $\hat{x}$ と $\hat{y}$ の間の関係性について仮説を立てて数式化することを、モデリング (modeling) と呼びます。どのようなモデルを設定するかは、データに合わせて解析者が適切に判断する必要があります。ここでは、シンプルな 1 次関数のモデル[4]を取り敢えず採用しています[5]。θ_0 と θ_1 はこのモデルを特徴づける量でパラメータ (parameter) と呼ばれます。今の場合、θ_0 と θ_1 はそれぞれ直線の切片と傾きに対応しています。

実際に停車距離を測定した走行速度 $x^{(i)}$ $(i = 0, \cdots, 10)$ に対するモデルの「推定値」は、次のように表すことができます[6]。

$$\hat{y}^{(i)} = \theta_0 + \theta_1 x^{(i)}$$

これは、次の 10 本の式をまとめた書き方です。

$$\begin{aligned}
\hat{y}^{(1)} &= \theta_0 + \theta_1 x^{(1)} \\
\hat{y}^{(2)} &= \theta_0 + \theta_1 x^{(2)} \\
&\vdots \\
\hat{y}^{(10)} &= \theta_0 + \theta_1 x^{(10)}
\end{aligned}$$

[2]自由に値を決められる変数 $\hat{x}$ を「独立変数」、$\hat{x}$ の値に応じて値が決定する $\hat{y}$ を「従属変数」と言います。
[3]正の相関がある、ということです。
[4]正確には、線形回帰モデル (linear regression model) と呼ばれます
[5]実は、走行速度と停車距離に対しては、2 次関数的な曲線のモデルの方が良いことが知られています。
[6]実際のデータには ^ (ハット) が付いていない文字、モデルの推定値には ^ の付いた文字を使っています。

i は「添え字」と呼ばれる変数で、今の場合、1 から 10 まで変化します。上のような 10 本の式をいちいち書くのがめんどくさいので、添え字を使ってまとめて表現してしまっています。$\hat{y}^{(i)} = \theta_0 + \theta_1 x^{(i)}$ という式は、i が $1, \cdots, 10$ まで、どの場合でも成り立つ式となっています。

モデルが完璧ならば、この 10 個の推定値 $\hat{y}^{(i)}$ は 10 個の実際の測定データ $y^{(i)}$ と一致します[7]。しかし現実にはそのようなケースは殆どなく、両者にはズレがあります。

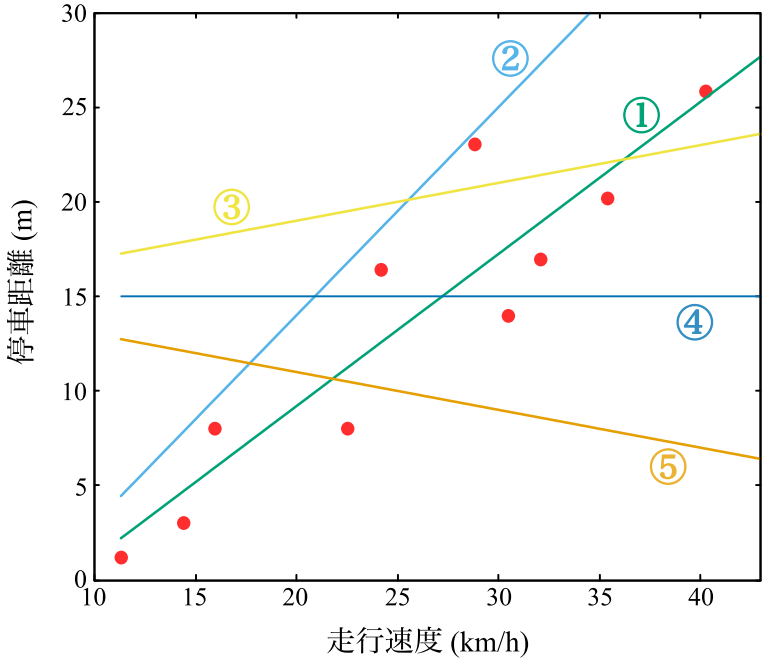
$$e^{(i)} = y^{(i)} - \hat{y}^{(i)} = y^{(i)} - (\theta_0 + \theta_1 x^{(i)})$$

各データ点での推定値 $\hat{y}^{(i)}$ と実際のデータ $y^{(i)}$ の誤差 (モデルでは説明できない量) を $e^{(i)}$ と表すことにします。

以上でモデルの構築が終了しました。ただし、このままでは θ_0、θ_1 の具体的な値が決まっておらず、不完全なモデルとなっています。以下では、このパラメータ θ_0、θ_1 を決めてモデルを完成させていきましょう。

♣ 最小二乗法によるフィッテイング

試しに、θ_0、θ_1 を色々変えて、傾きと切片の異なる 5 本のモデル直線を引いてみました。



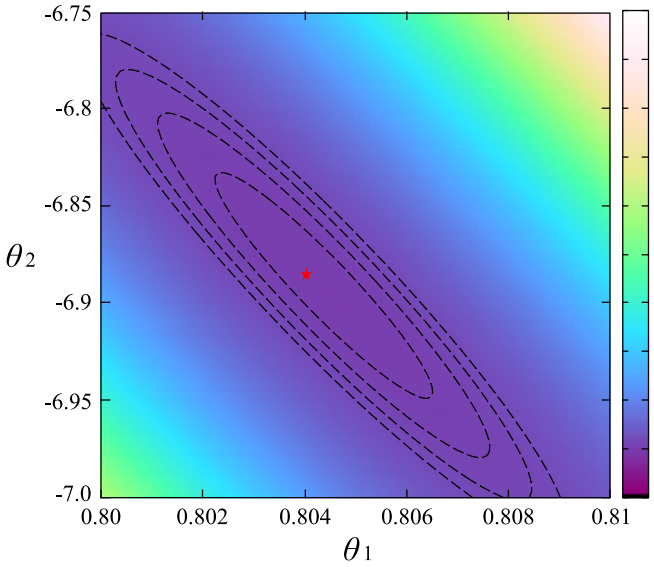
どの直線が最も相応しいモデルとなっているでしょうか？ 最も相応しいパラメータ θ_0、θ_1 の値を決めるには、何か方針が必要です。最も自然な方針は、誤差 $e^{(i)}$ がなるべく小さくなるようにパラメータの値を決める、というものです。$e^{(i)}$ はモデルと実際のデータのズレを表す量です

[7] $\hat{y}^{(i)}$ のことを「予測ラベル」、$y^{(i)}$ のことを「正解ラベル」と呼ぶこともあります。

から、なるべく小さい方が「良いモデル」と言えます。このような方針でパラメータを決定する方法として、**最小二乗法 (least squares method)** が一般的な方法として知られています。最小二乗法では、次の量を考えます[8]。

$$\mathrm{MSE} = \frac{1}{10}\sum_{i=1}^{10}\left(e^{(i)}\right)^2 = \frac{1}{10}\sum_{i=1}^{10}\left[y^{(i)} - (\theta_0 + \theta_1 x^{(i)})\right]^2$$

式に出てくる $\sum$ は総和記号です[9]。この誤差の 2 乗を全て足し合わせた量を、**平均二乗誤差** (MSE: mean squared error) と呼びます。最小二乗法は、この平均二乗誤差が最も小さくなるようにパラメータ θ_0、θ_1 の値を決定する方法です。下の図は、θ_0、θ_1 の値を変化させた時の平均二乗誤差の大きさを表しています。



色が濃く紫に近いほど、平均二乗誤差の大きさが小さくなっています。黒い点線は、最小点 (赤い星) 周辺の “等高線” を表しています。

色が濃いほど、値が小さくなっています。適切なアルゴリズムで平均二乗誤差が最小となる点を探すと、$(\theta_0, \theta_1) = (0.804, -6.891)$ というパラメータの組が得られます (上の図で赤い星で示したところです)。このパラメータの直線は、試しに引いた 5 本の直線のうち、① の緑の直線となっています。確かに最もらしい直線を引くことができています。

以上のようにしてパラメータを無事に決定し、それらしいモデルを得ることができましたが、1 つ疑問が残ります。平均二乗誤差という量が突然出てきて、その量を最小化することでパラメータの値を決定しましたが、果たしてこれは正しいのでしょうか？ 確かにこの方法で、最もらしい直線を引くことができました。しかし、平均二乗誤差の代わりに誤差の絶対値の和や四乗の和を使う方法や、あるいは全く別の方法を使っても良さそうに思えます。最小二乗法が

[8] $\frac{1}{m}$ (m はデータ数) は単なる定数倍で重要でないため、つけない場合もあります。
[9] 総和記号については、softmax 関数のところで説明しています。

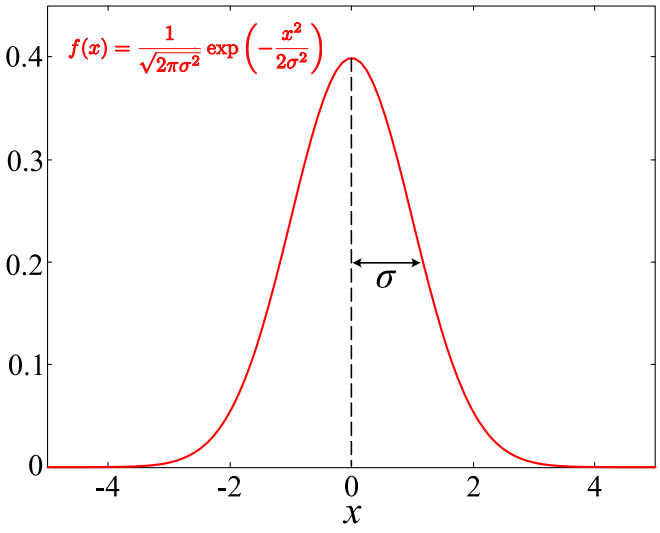
「正しい方法」である根拠はあるのでしょうか？この問題を解決するために、モデルに確率の考え方を導入します。そして、確率の考え方により裏付けされた方法である「最尤法[10]」を用いて、パラメータ θ_0、θ_1 の値を決定していきます。

♣ モデルに確率を導入

前のセクションで誤差 $e^{(i)}$ を導入しました。仮にモデルが完璧だとすると、この誤差は装置の持つ測定誤差のように理論では予想できないランダムな要素であると考えられます。こういったランダムな要素と確率の考え方は、相性が良いです。

ここまで誤差は、各データ点に対応した値であったため、i という添え字を必ず付けて $e^{(i)}$ と表現していました。以下ではこの誤差も $\hat{x}$ と $\hat{y}$ と同じような変数とみなして、単に e と表記することにします[11]。ランダムに変動する e のうち、実際に測定された値として $e^{(i)}$ $(i = 1, \cdots, 10)$ があると考えます。

e はランダムな値をとる変数ですが、どのような確率分布[12]に従うのでしょうか？ この確率分布はこちらで仮定するしかありません。よく使われるのは、Gauss (ガウス) 分布と呼ばれる分布です[13]。今回も変数 e の確率分布は、Gauss 分布で与えられるものとします。また、Gauss 分布自体にも平均と分散の 2 つのパラメータが含まれています。今回は、平均は 0 として、分散は未知の値 σ^2 として扱うことにします[14]。



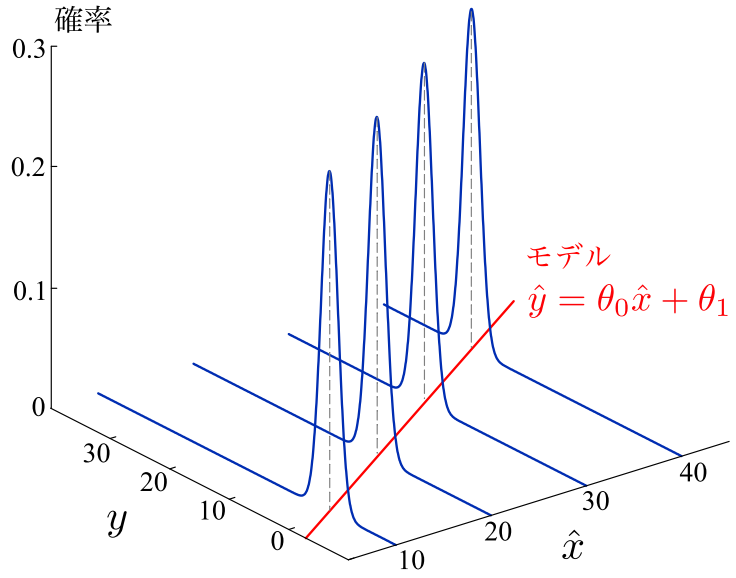
誤差 e の確率分布 (平均 0、分散 σ^2 の Gauss 分布 $\mathcal{N}(x; 0, \sigma^2)$)

[10]「さいゆうほう」と読みます
[11]ネイピア数と同じ文字となってしまいますが、混乱はないかと思います。
[12]e に対して、各々の値をとる確率を表したもの。Bernoulli 分布のところで詳しく解説しています。
[13]Gauss 分布については、また別のプリントで詳しく説明しています。
[14]e はモデルの推定値と実際のデータとの「誤差」であるため、平均すれば 0 というのは自然な仮定です。

このように誤差 e を設定すると、測定値 y は次のように表現されます。

$$y = \hat{y} + e = \left(\theta_0 + \theta_1 \hat{x}\right) + e$$

実際の停車距離の測定値は 10 点しかないので、ここまで測定値を表す場合には $y^{(i)}$ としてきました。しかしながら、以下ではこの測定値も変数として考え、y と表すことにします。誤差 e が Gauss 分布に従う変数であるため、y も Gauss 分布に従う変数となります。ただし、e の平均が 0 であったのに対し、y は上の式から分かるように、$\theta_0 + \theta_1 \hat{x}$ が平均となります。変数 y の値は、モデルの推定値 $\theta_0 + \theta_1 \hat{x}$ の値と一致する確率が最も高くなっていますが、誤差によって多少幅を持つようになっています (下図参照)。誤差 e と同様に、ランダムに変動する y のうち、実際に測定できた値として $y^{(i)}$ $(i = 1, \cdots, 10)$ があると考えることにします。



走行速度 $\hat{x}$ に対する停車距離 y の確率分布

少しややこしくなってしまいましたが、以上で準備が終わりました。次のセクションで、「最尤法」の考え方を説明していきます。

♣ 最尤法 (method of maximum likelihood) の考え方

前のセクションで、変数 y は Gauss 分布に従う変数で、この変数が 10 個のデータ $y^{(i)}$ を生成するものと考えることにしました。変数 y の確率分布は、元々のモデルのパラメータ (θ_0 と θ_1) と Gauss 分布の分散 σ^2 の 3 つのパラメータ $(\theta_0, \theta_1, \sigma^2)$ を変化させると変化します。最尤法とは、10 個の観測データ $y^{(i)}$ が生成される確率が最も大きくなるように、これらのパラメータ $(\theta_0, \theta_1, \sigma^2)$ の値を決定する、という考え方です[15]。この考え方は、1912 年から 1922 年にかけてロナルド・フィッシャーにより開発されました。

[15]最尤推定 (maximum likelihood estimation) とも呼ばれます

この最尤法の考え方を数式を用いて表現していきましょう。1 番目のデータ $y^{(1)}$ が生成される確率 P を次のように表すことにします。

$$P\big(y^{(1)}|x^{(1)};\theta_0,\theta_1,\sigma^2\big)$$

これは、**「条件付き確率」**と呼ばれる量です。$P(A|B)$ と書くと「B が与えられた時の A の確率」を意味します。単なる A の確率とは異なり、B という条件下での A の確率を意味します。今の場合、1 番目の停車距離のデータ $y^{(1)}$ が生成される確率は、パラメータ $(\theta_0,\theta_1,\sigma^2)$ と走行速度 $x^{(1)}$ によって変化します (前ページの図参照)。これらの値が与えられた条件下で $y^{(1)}$ が生成される確率を考えるため、上の式のように条件付き確率を使って表現されています。

さて、1 番目のデータ $y^{(1)}$ が生成される確率が $P\big(y^{(1)}|x^{(1)};\theta_0,\theta_1,\sigma^2\big)$ と表現されることが分かりましたが、ここで考えたいのは 10 個全てのデータが生成される確率です。この確率は、数式では次のように表現されます。

$$P\big(y^{(1)},\cdots,y^{(10)}|x^{(1)},\cdots,x^{(10)};\theta_0,\theta_1,\sigma^2\big)$$

モデルのパラメータ $(\theta_0,\theta_1,\sigma^2)$ と、10 個の走行速度のデータ $x^{(1)},\cdots,x^{(10)}$ が与えられた条件の下で、10 個の停車距離のデータ $y^{(1)},\cdots,y^{(10)}$ が生成される条件付き確率、という意味です。この確率を計算するにあたっては、確率の「積の法則」と呼ばれる法則を用いることになります。例えば、大小 2 つのサイコロを振るという例を考えましょう。両方のサイコロの目が 6 となる確率はいくつになるでしょうか？ 目の出方の全ての場合の数は、$6\times6=36$ 通りあります。そのため、両方のサイコロの目が 6 となる確率は、$\dfrac{1}{36}$ です。これは、次のように計算することもできます。大きいサイコロで 6 の目が出る確率は $\dfrac{1}{6}$、小さいサイコロで 6 の目が出る確率も $\dfrac{1}{6}$ です。両方とも 6 の目が出る確率は、この 2 つが同時に起こる確率であるため、次のような掛け算で求めることができます。

$$\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}$$

つまり、A が起こる確率を p_{A}、B が起こる確率を p_{B} とすると、A と B が同時に起こる確率は $p_{\mathrm{A}}\times p_{\mathrm{B}}$ と表すことができるのです。これを確率の積の法則と呼びます。この法則は便利なのですが、1 つ注意しなくてはいけない点があります。それは、A と B が無関係な事柄である場合にのみに適用できる法則であるという点です[16]。例えば、A の結果が B の結果に影響するような場合には使えません。このサイコロの例の場合は、大きいサイコロの出る目は、小さいサイコロの出る目に何も影響しません (大きい方で 6 が出たからといって、小さい方で 6 が出やすくなるようなことはありませんよね)。そのため、積の法則が成り立っているのです。

[16]正確には、事象 A と事象 B が「独立」である場合に使える法則となっています。

この「積の法則」を用いると、10 個全てのデータ $y^{(i)}$ $(i = 1, \cdots, 10)$ が生成される確率は、次のように表せることが分かります。

$$P\big(y^{(1)}, \cdots, y^{(10)} | x^{(1)}, \cdots, x^{(10)}; \theta_0, \theta_1, \sigma^2\big) = P\big(y^{(1)} | x^{(1)}; \theta_0, \theta_1, \sigma^2\big) \times \cdots \times P\big(y^{(10)} | x^{(10)}; \theta_0, \theta_1, \sigma^2\big)$$

ここでは、1 番目から 10 番目までのデータは同じ確率分布に従って生成されますが、生成されるデータはお互いに影響することはないと仮定して[17]、積の法則を使っています。上の式は、なんか文字がたくさんあって分かりにくいですね。最初に説明した $\boldsymbol{X}$、$\boldsymbol{Y}$ という記号を使うと、次のようにシンプルにまとめることができます。

$$P(\boldsymbol{Y} | \boldsymbol{X}; \boldsymbol{\theta}) = \prod_{i=1}^{10} P\big(y^{(i)} | x^{(i)}; \boldsymbol{\theta}\big)$$

$\boldsymbol{X}$、$\boldsymbol{Y}$ のように、モデルの 3 つのパラメータを格納した配列 $\boldsymbol{\theta}$ を導入しています。

$$\boldsymbol{\theta} = \big(\theta_0, \theta_1, \sigma^2\big)$$

また、$\prod_{i=1}^{10}$ は総乗記号と呼ばれる記号です[18]。次のように複数の変数の積を簡潔に表現できる、便利な記号となっています。

$$\prod_{i=1}^{10} a_i = a_1 \times a_2 \times \cdots a_{10}$$

今の場合は、1 番目のデータが生成される確率から 10 番目のデータが生成される確率までを全て掛け合わせるのを表現するために用いています。以上のようにして、10 個の観測データ $y^{(i)}$ $(i = 1, \cdots, 10)$ が生成される確率 $P(\boldsymbol{Y} | \boldsymbol{X}; \boldsymbol{\theta})$ を表現することができました。この $P(\boldsymbol{Y} | \boldsymbol{X}; \boldsymbol{\theta})$ は、**「条件付き尤度 (conditional likelihood)」**と呼ばれます。この量は、モデルがどのくらい実際の観測データに即しているかを表した量と解釈できます。最尤法の考えは、この量を最大化するように (つまりモデルをできる限り観測データに即しているように) パラメータ $\boldsymbol{\theta}$ を決定する、という非常に自然なものです。数式では、次のように表現されます。

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \ P(\boldsymbol{Y} | \boldsymbol{X}; \boldsymbol{\theta})$$

難しい数式に見えますが、この数式が言っているのは「 $P(\boldsymbol{Y} | \boldsymbol{X}; \boldsymbol{\theta})$ が最大となる時の $\boldsymbol{\theta}$ を $\boldsymbol{\theta}_{\mathrm{ML}}$ とする」ということです[19]。この尤度を最大化するパラメータ $\boldsymbol{\theta}_{\mathrm{ML}}$ のことを**「最尤推定量 (maximum likelihood estimator)」**と呼び、この値が求まればモデルが完成することになります。

[17] このような仮定は英語では independently and identically distributed といい、よく「i.i.d.」と略されます。
[18] 記号 $\prod$ は、ギリシャ文字「パイ」の大文字です。小文字は円周率でお馴染みの π です。
[19] arg max は、argument of the maximum (最大値を与える引数) の略です。

♣ 条件付き対数尤度の導入

では、具体的に最初にプロットした「走行速度」と「停車距離」のデータでの条件付き尤度 $P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta})$ を計算し、この量を最大化するパラメータ $\boldsymbol{\theta}_{\mathrm{ML}}$ を見つけましょう。条件付き尤度は、次の式で与えられました。

$$P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta}) = \prod_{i=1}^{10} P\big(y^{(i)}|x^{(i)};\boldsymbol{\theta}\big)$$

つまり、10 個のデータ、それぞれが生成される条件付き確率 $P\big(y^{(i)}|x^{(i)};\boldsymbol{\theta}\big)$ を計算して、それらを掛け合わせれば、条件付き尤度を計算することができます。この計算をモデルのパラメータ $\boldsymbol{\theta} = \big(\theta_0, \theta_1, \sigma^2\big)$ を変えながら繰り返し、最も大きな値となるようにパラメータを調整すればいいだけです。しかし、この作業をそのままコンピューターで行おうとすると問題が生じます。確率は必ず 1 より小さい数であるため、確率を掛け合わせた尤度は、データ数が増えるとどんどん小さな数になってしまいます。このような小さすぎる量をコンピューターで計算すると、「アンダーフロー」と呼ばれる問題が生じ、正確に計算することができなくなってしまいます。当然、機械学習ではコンピューターで計算を行うため、これは大問題です。そこで、条件付き尤度を直接計算するのではなく、次の**「条件付き対数尤度 (conditional Log-likelihood)」**と呼ばれる量を計算するのが一般的となっています。

$$\log P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta}) = \log\left(\prod_{i=1}^{10} P\big(y^{(i)}|x^{(i)};\boldsymbol{\theta}\big)\right) = \sum_{i=1}^{10} \log P\big(y^{(i)}|x^{(i)};\boldsymbol{\theta}\big)$$

log という記号は**「対数」**を表します。いきなり何のことか分からないかもしれませんが、対数については次のセクションで丁寧に解説しますので、一先ず置いておきましょう。ここで重要なのは、条件付き尤度では各データが生成される条件付き確率の「積」となっていましたが、条件付き対数尤度では各データが生成される条件付き確率の「和」となる点です。これにより、条件付き尤度の計算で生じるアンダーフローの問題を回避することができるのです。後ほど説明しますが、この条件付き対数尤度を最大化するパラメータ $\boldsymbol{\theta}$ は、 条件付き尤度を最大化するパラメータ $\boldsymbol{\theta}$ (つまり最尤推定量 $\boldsymbol{\theta}_{\mathrm{ML}}$) と同じになります。 つまり、コンピューターで計算しやすい条件付き対数尤度を最大化するパラメータ $\boldsymbol{\theta}$ を探せば、それが最尤推定量 $\boldsymbol{\theta}_{\mathrm{ML}}$ となっているのです。

♣ log (対数) とは？

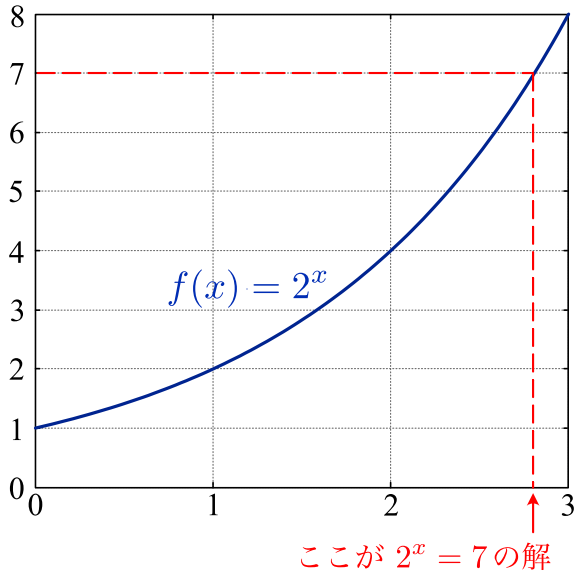
さて、少し話は逸れますが、条件付き対数尤度で出てきた「対数」について、基本的なところから説明します。先ず、次の x に関する方程式を考えましょう。

$$2^x = 8$$

この方程式の解は、$2^3 = 2 \times 2 \times 2 = 8$ ですから、$x = 3$ であることが直ぐに分かります。では、少し値を変えた次の方程式はどうでしょうか？

$$2^x = 7$$

$2^2 = 4$、$2^3 = 8$ ですから、整数の解はありません。しかし、softmax 関数のところで説明したように、累乗の概念は指数が小数の場合にも拡張できます。そのため、x が整数以外の場合も考慮すれば、上の方程式の解は $2 < x < 3$ の範囲にありそうです。$f(x) = 2^x$ のグラフを利用すれば、視覚的にもこの事が理解できます。



では、この小数の具体的な値はいくつなのでしょうか？ 残念ながら、この値は簡単に計算することはできません。そこで、この値を次のよう記号を用いて表現することとします[20]。

$$x = \log_2 7$$

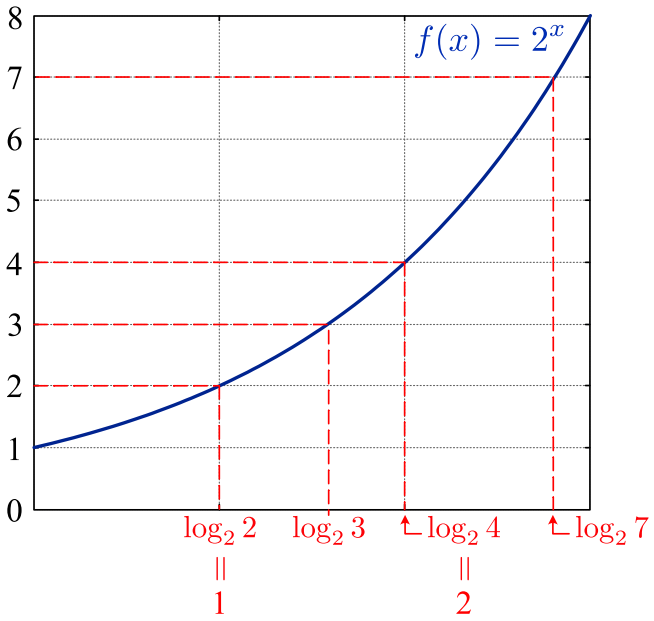
この $\log_2 7$ のような数のことを「対数 (logarithm)」と言います。記号 log (ログ) は、logarithm の頭文字となっています。一般に、対数

$$\log_a b$$

は、「a を何乗すれば b になるかを表す数」のことです。x に関する方程式 $a^x = b$ の解が $x = \log_a b$ と言うこともできます。尚、a のことを底 (てい)、b のことを真数と言います。

さて、再び指数関数 $f(x) = 2^x$ のグラフを見てみると、$f(x) = 2$、$f(x) = 3$、$\cdots$ となる点の x 座標は、対数を使って表せることが分かります。

[20] $x^2 = 3$ を満たす x の値が簡単には分からないから、$x = \pm\sqrt{3}$ とルートが導入されたのと同じことです。



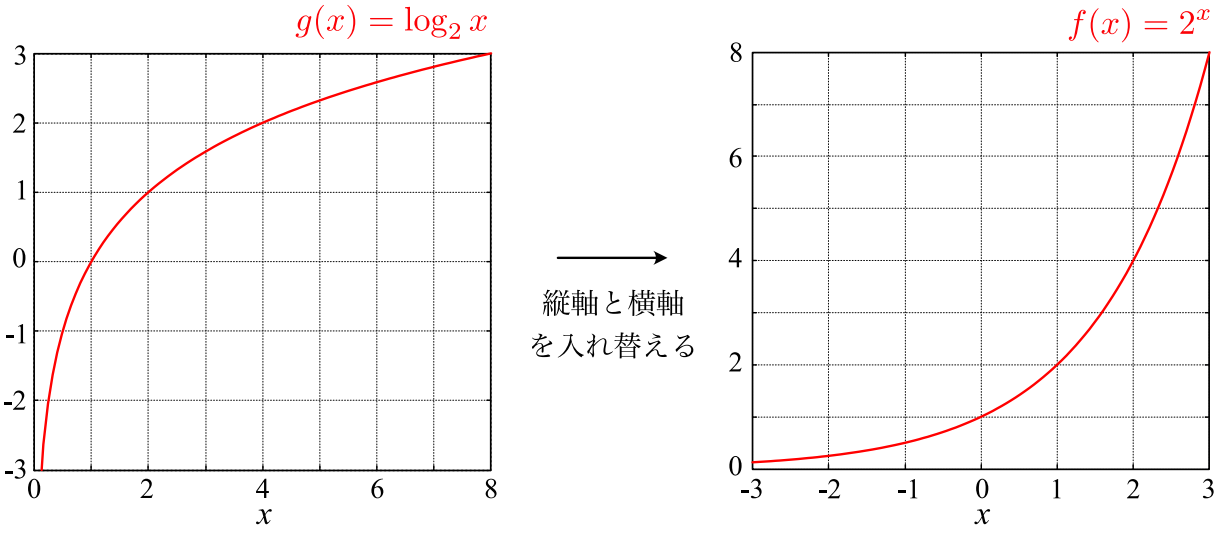
(例として、$f(x) = 2, 3, 4, 7$ となる x 座標を対数を使って表しています)

一般には、$f(x) = x$ となる点の x 座標が $\log_2 x$ となります。なお、上の図で、次のような対数の値は簡単に分かります。

$$\log_2 2 = 1$$

$$\log_2 4 = 2$$

2 は 1 乗すれば 2 になるので、$\log_2 2$ というのは 1 のことです。同じように、2 は 2 乗すれば 4 になるので $\log_2 4$ は 2 のことです[21]。対数の導入の最後に、対数関数 $g(x) = \log_2 x$ を考えます。この関数は、入力値 x に対して、「2 を何乗すれば x になるかを表す数」を出力値として返します。グラフを描くと次のようになります。



[21] もちろんこれらは特殊な場合であって、一般の log の値は簡単には分かりません。

例えば、$2^0 = 1$ ですから、$g(1) = \log_2 1 = 0$ となっています。他にも、$2^1 = 2$ ですから、$g(2) = \log_2 2 = 1$ といった具合です。この対数関数 $g(x)$ のグラフと指数関数 $f(x)$ のグラフ、曲線の形状は全く同じなのが分かるでしょうか？ 2つは単に縦軸と横軸をひっくり返しただけの関係です[22]。このグラフからも分かりますが、対数というのは単に指数 (累乗) の見方を変えただけの量なのです。

対数のグラフ $g(x) = \log_2 x$ の特徴で以下で重要となるのは、単調増加である (x が大きくなると関数の値も大きくなる) 点です。つまり、$x_1 < x_2$ であれば必ず $\log_2 x_1 < \log_2 x_2$ が成り立ちます。このように対数を取っても大小関係が変わらないため、条件付き尤度を最大化することと条件付き対数尤度を最大化することが等価となるのです。

♣ 対数の足し算は真数の掛け算

前のセクションで対数と呼ばれる数を導入しました。ここでは、この対数に関する計算ルールを確認していきます。対数に関する計算では、次の式を最もよく使います。

$$\log_a M + \log_a N = \log_a MN$$

つまり「対数の足し算は、真数の掛け算」ということです。具体例をいくつか挙げると、

$$\log_2 3 + \log_2 4 = \log_2(3 \times 4) = \log_2 12$$

$$\log_{10} 4 + \log_{10} 25 = \log_{10}(4 \times 25) = \log_{10} 100 = 2$$

といった具合です。ただし、底が違う場合 ($\log_2 3 + \log_3 4$ など) にはこのような計算はできないので注意して下さい。また逆に、真数の掛け算を対数の足し算に変換することもできます。

$$\begin{aligned}\log_2 12 = &= \log_2(3 \times 4) = \log_2 3 + \log_2 4 \\ &= \log_2(2 \times 6) = \log_2 2 + \log_2 6 = 1 + \log_2 6\end{aligned}$$

この対数の計算は非常に便利で、この後の説明でも出てくるので、よく覚えておいて下さい。因みに、なぜこのような計算ができるのかは、簡単に確かめることができます。M と N を次のように a の累乗の形で表したとしましょう。

$$M = a^p, \quad N = a^q \quad (p, q \text{ は実数})$$

[22] 対数関数は指数関数のいわゆる「逆関数」になっています。

すると、M と N の積は、

$$MN = a^p a^q = a^{p+q}$$

となります。いわゆる「指数法則」と呼ばれる計算です。MN は、a の $p+q$ 乗であることが分かるので、対数の定義より、

$$\log_a MN = p + q$$

となります。ところで、$M = a^p$、$N = a^q$ なので、対数を考えると、

$$\log_a M = p, \quad \log_a N = q$$

です。以上より、

$$\log_a MN = p + q = \log_a M + \log_a N$$

が導かれます。この導出でわかったと思いますが、この対数の計算に関する規則は、本質的には指数法則なのです。指数法則は「累乗の掛け算は、指数の足し算」という法則ですが、これの見方を変えたものが「対数の足し算は、真数の掛け算」なのです。

また、この対数の性質を応用すると、次の公式も得られます。

$$\log_a M^n = n \log_a M$$

これは単純に、「真数の掛け算は、対数の足し算」を使っただけです。

$$\log_a M^n = \log_a \big(\underbrace{M \times M \times \cdots M}_{n \text{ 個}} \big) = \underbrace{\log_a M + \log_a M + \cdots \log_a M}_{n \text{ 個}}$$

n 個の $\log_a M$ を足すため、$\log_a M^n = n \log_a M$ となる訳です。この公式も対数の計算では非常によく使うので、覚えておくと便利でしょう。

♣ もう一度、条件付き対数尤度を見てみる

さて、対数について確認したところで、再び条件付き対数尤度を見てみましょう。条件付き対数尤度は、条件付き尤度に関して対数を取った量でした。

$$\log P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta}) = \log \left(\prod_{i=1}^{10} P\big(y^{(i)}|x^{(i)};\boldsymbol{\theta}\big) \right)$$

この対数の底は何でも良いのですが、一般的にはネイピア数 e を底とします[23]。いちいち底を書くのは面倒なので、底の値がそれほど重要でない場合にはよく省略してしまいます。前のセ

[23] このような対数は「自然対数」と呼ばれ、log ではなく ln と表記することもあります。

クションで説明したように、log の中の真数の掛け算は、対数の足し算に変換することができます。この性質を用いることで、条件付き対数尤度は次のように書き換えることができるのです。

$$\log P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta}) = \sum_{i=1}^{10} \log P\big(y^{(i)}|x^{(i)};\boldsymbol{\theta}\big)$$

観測値 $y = \theta_0 + \theta_1 x + e$ は、誤差 e が Gauss 分布に従うため、y 自身も Gauss 分布に従いました。モデルのパラメータ $\boldsymbol{\theta} = \big(\theta_0, \theta_1, \sigma^2\big)$、走行速度の観測値 $x^{(i)}$ が与えられた時、y の確率分布は、平均が $\theta_0 + \theta_1 x^{(i)}$、分散が σ^2 の Gauss 分布となります。そのため、i 番目のデータ $y^{(i)}$ が生成される条件付き確率は、次のように与えられます[24]。

$$P\big(y^{(i)}|x^{(i)};\boldsymbol{\theta}\big) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left[\frac{y^{(i)} - [\theta_0 + \theta_1 x^{(i)}]}{\sigma}\right]^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left[\frac{y^{(i)} - \hat{y}^{(i)}}{\sigma}\right]^2\right)$$

これより、条件付き対数尤度は、次のように計算することができます。

$$\begin{aligned}
\log P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta}) &= \sum_{i=1}^{10} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left[\frac{y^{(i)} - \hat{y}^{(i)}}{\sigma}\right]^2\right)\right) \\
&= \sum_{i=1}^{10} \left[\log\left(\frac{1}{\sigma}\right) + \log\left(\frac{1}{\sqrt{2\pi}}\right) + \log\left(\exp\left(-\frac{1}{2}\left[\frac{y^{(i)} - \hat{y}^{(i)}}{\sigma}\right]^2\right)\right)\right] \\
&= \sum_{i=1}^{10} \log\left(\frac{1}{\sigma}\right) + \sum_{i=1}^{10} \log\left(\frac{1}{\sqrt{2\pi}}\right) + \sum_{i=1}^{10} \log\left(\exp\left(-\frac{1}{2}\left[\frac{y^{(i)} - \hat{y}^{(i)}}{\sigma}\right]^2\right)\right)
\end{aligned}$$

この式変形では、「真数の掛け算は対数の足し算」という対数の性質を再び使っています。この式は、まだまだ変形することができます。分数に関しては、次のように累乗を用いて表現することができました[25]。

$$\frac{1}{\sigma} = \sigma^{-1}, \qquad \frac{1}{\sqrt{2\pi}} = (2\pi)^{-\frac{1}{2}}$$

-1 乗は逆数を、$\dfrac{1}{2}$ 乗は平方根を表しました。これより、$\log_a M^n = n \log_a M$ の公式を用いると、次のように計算することができます。

$$\log\left(\frac{1}{\sigma}\right) = \log \sigma^{-1} = -\log \sigma$$

$$\log\left(\frac{1}{\sqrt{2\pi}}\right) = \log (2\pi)^{-1/2} = -\frac{1}{2} \log (2\pi)$$

[24]指数が複雑な場合、e^x のことを $\exp(x)$ と表すことが多いです。
[25]softmax 関数の解説を参照して下さい。

これより、

$$\sum_{i=1}^{10} \log\left(\frac{1}{\sigma}\right) = 10 \times \big(-\log\sigma\big)$$

$$\sum_{i=1}^{10} \log\left(\frac{1}{\sqrt{2\pi}}\right) = 10 \times \left(-\frac{1}{2}\log\big(2\pi\big)\right)$$

となります。また、再び $\log_a M^n = n\log_a M$ の公式より、

$$\log\left(\exp\left(-\frac{1}{2}\left[\frac{y^{(i)}-\hat{y}^{(i)}}{\sigma}\right]^2\right)\right) = -\frac{\big[y^{(i)}-\hat{y}^{(i)}\big]^2}{2\sigma^2}\log e = -\frac{\big[y^{(i)}-\hat{y}^{(i)}\big]^2}{2\sigma^2}$$

ここでの対数の底はネイピア数としているので、$\log e = \log_e e = 1$ となります。これらの結果を用いると、条件付き対数尤度は、最終的に次のような式で与えられることが分かります。

$$\log P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta}) = -m\log\sigma - \frac{m}{2}\log\big(2\pi\big) - \sum_{i=1}^{m}\frac{\big[y^{(i)}-\hat{y}^{(i)}\big]^2}{2\sigma^2}$$

ここでは少し一般化して、データ点が m 点あるとしています (車の停車距離に関する具体例では、データ点は 10 点であるため、上の式で $m = 10$ となります)。後は、この条件付き対数尤度を最大化するようにパラメータを決定していきます。上の式で、モデルの直線の切片と傾きを与えるパラメータ θ_0、θ_1 は、$\hat{y}^{(i)} = \theta_0 + \theta_1 x^{(i)}$ のみに含まれています。つまり、条件付き対数尤度を θ_0、θ_1 に関して最大化するためには、この $\hat{y}^{(i)}$ が含まれている

$$\frac{1}{2\sigma^2}\sum_{i=1}^{m}\big[y^{(i)}-\hat{y}^{(i)}\big]^2$$

の部分を最小化すれば良いことが分かります (この部分は必ず正で、$\log P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta})$ には引き算の形で入っているので、条件付き対数尤度が最大となる時、この部分は最小となります)。この部分を最小化するということは、最小二乗法に他なりません。最小二乗法では平均二乗誤差

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}\big[y^{(i)}-\hat{y}^{(i)}\big]^2$$

を最小化するようにモデルのパラメータ θ_0、θ_1 を決定しました。この MSE と上の条件付き対数尤度に含まれる部分は単なる定数倍の違いしかなく、両者の最小値を与える θ_0、θ_1 は同じとなるのです。つまり、最小二乗法は、誤差に平均 0 の Gauss 分布を仮定した場合の最尤推定となっていることが分かりました。尚、条件付き対数尤度を σ^2 に関しても最小化すると、この分散の最尤推定値 σ^2_{ML} も決定することができます。この量と、計測機器の測定誤差や実験誤差などを比較することにより、モデルの妥当性を評価できます。

♣ 最尤法と最小二乗法の関係について

ここまでの話を整理しましょう。最小二乗法は、誤差の 2 乗の和 (平均二乗誤差) を最小化することでモデルのパラメータを決定する方法です。なぜ、誤差の絶対値や 4 乗の和を考えないかというと、2 乗の和を最小化することが、誤差が平均 0 の Gauss 分布となっている場合には最尤推定となっているからです。このようにして、確率の考え方を用いて、最小二乗法を統計的な視点から正当化することが出来ました。

では、同じパラメータが得られるのならば、今後はシンプルな最小二乗法のみを使えばよく、確率の考えを使ってややこしい最尤法をわざわざ使う必要は一切ないのでしょうか？実はそうではなく、モデルを拡張する際に、最尤法の考え方が必要となるのです。今回は、自動車の速度と停車距離の関係が直線的な関係にあるとしてモデルを構築しましたが、データによっては 2 次関数や指数関数のモデルの方が精度よくデータを説明できることがあります。このようにモデルが複雑になると、もはや最小二乗法ではモデルのパラメータを決定することはできなくなります。一方、最尤法を用いると、このような複雑なモデルの場合でも、確率に基づいた適切な方法でパラメータを推定することができるのです。