1. Simple linear regression

1.1. **The regression line.**

Consider a data set

$$\mathscr{D} = \{(x_i, y_i) : i = 1, \ldots, n\}.$$

If the *mean-squared error* function

$$\mathrm{MSE}(a, b) = \frac{1}{n} \sum_{i=1}^{n} (a x_i + b - y_i)^2$$

achieves its absolute minimum value at

$$(a, b) = (\alpha, \beta)$$

then the line $y = \alpha x + \beta$ is called the *regression line* or *least-squares line* for $\mathscr{D}$.

The *slope*, α, and the *intercept*, β of the regression line (its *coefficients*) can be expressed in terms of basic statistics of $\mathscr{D}$:

means: $$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

variances: $$s_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2, \qquad s_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

covariance: $$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Theorem 1 (Gauss/Legendre)**.** *The coefficients of the regression line of $\mathscr{D}$ are:*

$$a = \frac{s_{xy}}{s_x^2}, \qquad b = \overline{y} = a\overline{x}.$$

Proof. Notice that

$$\min_{(a,b)} \mathrm{MSE}(a, b) = \min_{a} \left(\min_{b} \mathrm{MSE}(a, b) \right).$$

For a given a, the quantity $\mathrm{MSE}(a, b)$ is a quadratic polynomial in b:

$$\mathrm{MSE}(a, b) = b^2 - 2 \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - a x_i) \right) b + \sum_{i=1}^{n} (y_i - a x_i)$$

Since a quadratic polynomial $t^2 - 2qt + r$ achieves its minimum value at $t = q$, $\mathrm{MSE}(a, b)$ achieves its minimum value when

$$b = \frac{1}{n} \sum_{i=1}^{n} (y_i - a x_i) = \overline{y} - a\overline{x}.$$

It remains to determine

$$\min_{a} \mathrm{MSE}(a, \overline{y} - a\overline{x}) = \min_{a} \frac{1}{n} \sum_{i=1}^{n} (a x_i + (\overline{y} - a\overline{x}) - y_i)^2.$$

Expanding and rearranging, we get

$$\frac{1}{n}\sum_{i=1}^{n}(ax_i + (\overline{y} - a\overline{x}) - y_i)^2 = s_x^2 a^2 - 2s_{xy}a + s_y^2.$$

Since a quadratic polynomial $pt^2 - 2qt + r$ achieves its minimum value at $t = q/p$, the function $\mathrm{MSE}(a, \overline{y} - a\overline{x})$ achieves its minimum value when $a = s_{xy}/s_x^2$.

Thus, $\mathrm{MSE}(a, b)$ is minimized when

$$a = \frac{s_{xy}}{s_x^2}, \qquad b = \overline{y} - a\overline{x}. \qquad \square$$

Define $\mathbf{1}$, $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^n$ by

$$\mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

For α, $\beta \in \mathbb{R}$, define the associated *residual vector*, $\boldsymbol{e}(\alpha, \beta)$, by

$$\boldsymbol{e}(\alpha, \beta) = \alpha\boldsymbol{x} + \beta\mathbf{1} - \boldsymbol{y}.$$

Then

$$\mathrm{MSE}(\alpha, \beta) = \frac{1}{n}\|\boldsymbol{e}(\alpha, \beta)\|^2.$$

Let U be the subspace of $\mathbb{R}^n$ spanned by the vectors $\boldsymbol{x}$ and $\mathbf{1}$:

$$U = \{\alpha\boldsymbol{x} + \beta\mathbf{1} : \alpha, \beta \in \mathbb{R}^n\}.$$

Let $d(\boldsymbol{y}, U)$ be the distance from $\boldsymbol{y}$ to U, i.e., the minimal distance from $\boldsymbol{y}$ to an element of U:

$$d(\boldsymbol{y}, U) = \inf_{a,b} \|a\boldsymbol{x} + b\mathbf{1} - \boldsymbol{y}\|.$$

The infimum on the right is achieved by *orthogonal projection of $\boldsymbol{y}$ onto U*, i.e., the unique vector $\widehat{\boldsymbol{y}} \in U$ such that

$$\langle \widehat{\boldsymbol{y}}, \boldsymbol{y} - \widehat{\boldsymbol{y}} \rangle = 0.$$

If $\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ is any orthonormal basis of U, then

$$\widehat{\boldsymbol{y}} = \langle \boldsymbol{u}_1, \boldsymbol{y} \rangle \boldsymbol{u}_1 + \langle \boldsymbol{u}_2, \boldsymbol{y} \rangle \boldsymbol{u}_2.$$

We can construct an orthonormal basis of U be applying the *Gram-Schmidt orthonormalization procedure* to the spanning set $\{\mathbf{1}, \boldsymbol{x}\}$. Let

$$\begin{aligned} \boldsymbol{u}_1 &= \frac{1}{\|\mathbf{1}\|}\mathbf{1} = \frac{1}{\sqrt{n}}\mathbf{1}, \\ \boldsymbol{u}_2' &= \boldsymbol{x} - \langle \boldsymbol{u}_1, \boldsymbol{x} \rangle \boldsymbol{u}_1 \\ &= \boldsymbol{x} - \frac{1}{\sqrt{n}}\langle \mathbf{1}, \boldsymbol{x} \rangle \frac{1}{\sqrt{n}}\mathbf{1} \\ &= \boldsymbol{x} - \overline{x}\mathbf{1}, \end{aligned}$$

Assume that $\boldsymbol{x}$ and $\mathbf{1}$ are linearly independent. Then $\boldsymbol{u}_2' \neq 0$ and we may set

$$\begin{aligned}\boldsymbol{u}_2 &= \frac{1}{\|\boldsymbol{u}_2'\|}\boldsymbol{u}_2' \\ &= \frac{1}{\sqrt{n}s_x}(\boldsymbol{x} - \overline{x}\mathbf{1})\end{aligned}$$

Thus, if $\boldsymbol{x}$ and $\mathbf{1}$ are linearly independent, then

$$\left\{ \frac{1}{\sqrt{n}}\mathbf{1},\ \frac{1}{\sqrt{n}s_x}(\boldsymbol{x} - \overline{x}\mathbf{1}) \right\}.$$

is an orthonormal basis of U. It follows that

$$\widehat{\boldsymbol{y}} = \frac{1}{n}\left\langle \mathbf{1}, \boldsymbol{y}\right\rangle \mathbf{1} + \frac{1}{ns_x^2}\left\langle \boldsymbol{x} - \overline{x}\mathbf{1}, \boldsymbol{y}\right\rangle (\boldsymbol{x} - \overline{x}\mathbf{1})$$

Since $\boldsymbol{x} - \overline{x}\mathbf{1}$ is orthogonal to $\mathbf{1}$,

$$\frac{1}{n}\left\langle \boldsymbol{x} - \overline{x}\mathbf{1}, \boldsymbol{y}\right\rangle = \frac{1}{n}\left\langle \boldsymbol{x} - \overline{x}\mathbf{1}, \boldsymbol{y} - \overline{y}\mathbf{1}\right\rangle = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = s_{xy}.$$

$$\widehat{\boldsymbol{y}} = \overline{y}\mathbf{1} + \frac{s_{xy}}{s_x^2}(\boldsymbol{x} - \overline{x}\mathbf{1}) = \frac{s_{xy}}{s_x^2}\boldsymbol{x} + \left(\overline{y} - \frac{s_{xy}}{s_x^2}\overline{x}\right)\mathbf{1}$$

Theorem 2.

(1) *There is a unique vector* $\widehat{\boldsymbol{y}} \in U$ *such that*

$$d(\boldsymbol{y}, U) = \|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|.$$

(2) *If the vectors* $\mathbf{1}$ *and* $\boldsymbol{x}$ *are linearly independent, then there are unique scalars* $\widehat{a}$ *and* $\widehat{b}$ *such that*

$$\widehat{\boldsymbol{y}} = \widehat{a}\boldsymbol{x} + \widehat{b}\mathbf{1}.$$

$$\|\boldsymbol{y} - \overline{y}\mathbf{1}\|^2 = \|(\boldsymbol{y} - \widehat{\boldsymbol{y}}) + (\widehat{\boldsymbol{y}} - \overline{y}\mathbf{1})\|^2 = \|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2 + \|\widehat{\boldsymbol{y}} - \overline{y}\mathbf{1}\|^2 = \mathrm{SSE} + s_{\widehat{\boldsymbol{y}}}^2$$

1.2. **Sums of squares.** The regression line gives the estimate

$$\widehat{y}_i = ax_i + b$$

for y_i. The $\widehat{y}_i$ and the y_i have the same mean:

$$\overline{\widehat{y}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) = a\overline{x} + b = \overline{y},$$

the final equality following from Theorem 1.

$$
\begin{aligned}
s_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i + \widehat{y}_i - \overline{y})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + 2\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) + \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \\
&= \mathrm{MSE}(a,b) + 2s_{e\widehat{y}} + s_{\widehat{y}}^2.
\end{aligned}
$$

2. The bivariate normal distribution

The bivariate normal density with means μ_1 and μ_2, variances σ_1 and σ_2, and correlation ρ is defined by

$$
f(x_1, x_2) = \frac{1}{2\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}Q(x_1,x_2)},
$$

where

$$
Q(x_1, x_2) = \frac{1}{\sqrt{1-\rho^2}}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]
$$

We write

$$
(X_1, X_2) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)
$$

if (X_1, X_2) has density $f(x_1, x_2)$.

Suppose $X \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Prove:

(1) The marginal density of X_1 is the univariate normal density with mean μ_1 and variance σ_1^2, i.e.,

$$
\int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2 = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)}.
$$

(2) $\mathrm{E}[X_i] = \mu_i$, $\mathrm{E}[(X_i - \mu_i)^2] = \sigma_i^2$, and $\mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] = \sigma_1\sigma_2\rho$.

(3) The conditional density of X_2 given X_1 is given by

$$
f(x_2|x_1) = \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_2} e^{-\frac{1}{2}\left(\frac{x_2 - \left(\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)+\mu_2\right)}{\sqrt{1-\rho^2}\sigma_2}\right)^2}.
$$

(4) The conditional expectation and variance of X_2 given X_1 are given by

$$
\mathrm{E}[X_2|X_1] = \rho\frac{\sigma_2}{\sigma_1}(X_1 - \mu_1) + \mu_2
$$

and

$$
\mathrm{E}[(X_2 - \mathrm{E}[X_2|X_1])^2|X_1] = \sqrt{1-\rho^2}\sigma_2,
$$

respectively. Note that the latter quantity is independent of X_1.

3. Conditional Expectation

Theorem-Definition 3. *Let Ω be a set equipped with a probability measure, P. Given random variables X and Y on Ω, there is a unique function $f : \mathbb{R} \to \mathbb{R}$ such that*

$$\int_{[X \in G]} Y \, dP = \int_{[X \in G]} f(X) \, dP,$$

for every $E \subseteq \mathbb{R}$. The random variable $f(X)$ is called the conditional expectation of Y given X *and denoted* $\mathrm{E}[Y|X]$.

(1) If $Y = f(X)$, then $\mathrm{E}[Y|X] = Y$.

(2) If $X = 1$, then $\mathrm{E}[Y|X] = \mathrm{E}[Y]$:

$$1 \notin G: \qquad \int_{[X \in G]} Y \, dP = \int_{\varnothing} Y \, dP = 0 = \int_{\varnothing} \mathrm{E}[Y] \, dP = \int_{[X \in G]} \mathrm{E}[Y] \, dP$$

$$1 \in G: \qquad \int_{[X \in G]} Y \, dP = \int_{\Omega} Y \, dP = \mathrm{E}[Y] = \int_{\Omega} \mathrm{E}[Y] \, dP = \int_{[X \in G]} \mathrm{E}[Y] \, dP$$

(3) If $\mathrm{E}[Y|X] = f(X)$, then

$$\mathrm{E}[I_H(X) Y | X] = I_H(X) f(X)$$

for all $H \subseteq \mathbb{R}$:

$$\int_{[X \in G]} I_H(X) Y \, dP = \int_{[X \in G \cap H]} Y \, dP = \int_{[X \in G \cap H]} f(X) \, dP = \int_{[X \in G]} I_H(X) f(X) \, dP$$

(4) If $u : \mathbb{R} \to \mathbb{R}$, then

$$\mathrm{E}[u(X) Y | X] = u(X) \, \mathrm{E}[Y|X].$$

(Proof: Exercise?)

(5) If $X = u(Y)$, then

$$\mathrm{E}[\mathrm{E}[Z|Y]|X] = \mathrm{E}[Z|X].$$

$$\begin{aligned} \int_{[u(X) \in G]} \mathrm{E}[Y|X] \, dP &= \int_{[X \in u^{-1}(G)]} \mathrm{E}[Y|X] \, dP \\ &= \int_{[X \in u^{-1}(G)]} Y \, dP \\ &= \int_{[u(X) \in G]} Y \, dP \\ &= \int_{[u(X) \in G]} \mathrm{E}[Y|u(X)] \, dP \end{aligned}$$

Exercise: X has countable range...

Lemma 4. $\mathrm{Cov}(u(X), Y - \mathrm{E}[X]) = 0$.

Proof.

$$\operatorname{Cov}(u(X), Y - E[Y|X]) = \mathrm{E}[u(X)\,\mathrm{E}[Y|X]]$$

□

$$\begin{aligned}
\mathrm{E}[(Y - f(X))^2] &= \mathrm{E}[(Y - \mathrm{E}[Y|X] + \mathrm{E}[Y|X] - f(X))^2] \\
&= \mathrm{E}[(Y - \mathrm{E}[Y|X])^2] + 2\operatorname{Cov}(Y - \mathrm{E}[Y|X], E[Y|X] - f(X)) + \mathrm{E}[f(X)^2]
\end{aligned}$$

Lemma 5. *The following are equivalent:*

(1) $\mathrm{E}[Y|X] = Y$

(2) $Y = f(X)$ *for some* $f : \mathbb{R} \to \mathbb{R}$.

(3) $\operatorname{Cov}(Y, Z - \mathrm{E}[Z|X]) = 0$ *for all random variables* Z.

Proof.

(1) ⇒ (2) $\mathrm{E}[Y|X]$ is, by definition, a function of X.
(2) ⇒ (3) We have:

$$\begin{aligned}
\operatorname{Cov}(f(X), Z - \mathrm{E}[Z|X]) &= \mathrm{E}[f(X)(Z - \mathrm{E}[Z|X])] \\
&= \mathrm{E}[f(X)Z] - \mathrm{E}[f(X)\,\mathrm{E}[Z|X]] \\
&= \mathrm{E}[f(X)Z] - \mathrm{E}[\mathrm{E}[f(X)Z|X]] \\
&= \mathrm{E}[f(X)Z] - \mathrm{E}[f(X)Z] \\
&= 0.
\end{aligned}$$

(3) ⇒ (1)

□

$$\begin{aligned}
\mathrm{E}[u(X)Y] &= \mathrm{E}[\mathrm{E}[u(X)Y|X]] \\
&= \mathrm{E}[u(X)\,\mathrm{E}[Y|X]]
\end{aligned}$$

Let $f(x, y)$ be the empirical density associated to the data set $(x_1, y_1), \ldots, (x_n, y_n)$:

$$f(x, y) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)\delta(y - y_i)$$

Suppose that (X, Y) has joint density $f(x, y)$. The marginal densities $f(x)$ and $f(y)$ of X and Y are

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i) \quad \text{and} \quad f(y) = \frac{1}{n}\sum_{i=1}^{n} \delta(y - y_i).$$

Let's project $Y - \mathrm{E}\,Y$ onto the span of the uncorrelated random variables 1 and $X - \mathrm{E}\,X$. It's easy to show (exercise) that $\mathrm{E}\,X = \overline{x}$ and $\mathrm{E}\,Y = \overline{y}$.

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{and} \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

Therefore,

$$\begin{aligned}\mathrm{E}[(X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)] &= \iint (y - \overline{y})(x - \overline{x}) f(x,y)\,dx\,dy \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \operatorname{cov}(x,y).\end{aligned}$$

Obviously,

$$\mathrm{E}[1(Y - \mathrm{E}\,Y)] = 0.$$

Therefore, the projection of $Y - \mathrm{E}\,Y$ onto the span of 1 and $X - \mathrm{E}\,X$ is

$$\frac{\mathrm{E}[1(Y - \mathrm{E}\,Y)]}{\mathrm{E}[1^2]}1 + \frac{\mathrm{E}[(X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)]}{\mathrm{E}[(X - \mathrm{E}\,X)^2]}(X - \mathrm{E}\,X) = \frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)}(X - \overline{x})$$

It follows that the linear regression of Y on X is

$$\widehat{Y} = \frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)}(X - \overline{x}) + \overline{y}$$

Consider the probability space

$$(\mathbb{R}^2, f(x,y)\,dx\,dy),$$

where $f(x,y)$ is the *empirical density* associated to the data set $(x_1,y_1),\ldots,(x_n,y_n)$:

$$f(x,y) = \frac{1}{n}\sum_{i=1}^{n}\delta(x - x_i)\delta(y - y_i).$$

Let

$$V := L^2(\mathbb{R}^2, f(x,y)\,dx\,dy) = \left\{Z : \mathbb{R}^2 \to \mathbb{R} : \iint |Z(x,y)|^2 f(x,y)\,dx\,dy\right\}$$

- You want to "average away" the noise. Interpolating noisy data gives wiggly graphs.
- large oscillations near left and right endpoints
- Increasing size of training set increases model complexity (degree).

4. Bias-variance decomposition

Let $\widehat{\theta} = \widehat{\theta}(X)$ be an estimator of θ. The *bias of* $\widehat{\theta}$ is defined by

$$\operatorname{Bias}(\widehat{\theta}, \theta) = \mathrm{E}\,\widehat{\theta} - \theta.$$

The variance of the random variable $\widehat{\theta}$ is given, as usual, by

$$\operatorname{Var}\widehat{\theta} = \mathrm{E}\left[(\widehat{\theta} - \mathrm{E}\,\widehat{\theta})^2\right]$$

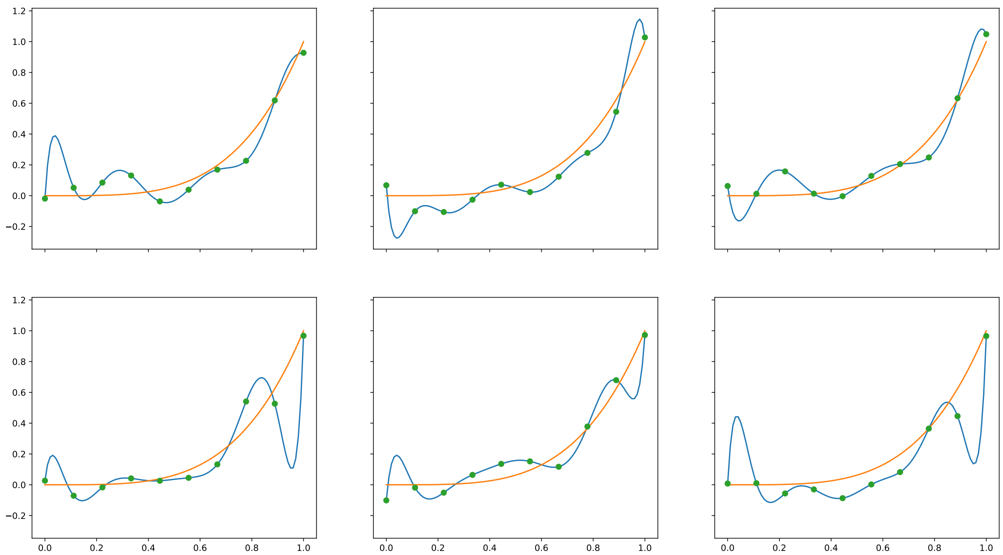

FIGURE 1. —— $y = x^4$, ● $y_i = x_i^4 + \text{noise}$, —— polynomial through (x_i, y_i)

Theorem 6 (Bias-Variance decomposition)**.**

$$\mathrm{E}\left[(\widehat{\theta} - \theta)^2\right] = \mathrm{Bias}(\widehat{\theta}, \theta)^2 + \mathrm{Var}\,\widehat{\theta}$$

Proof.

$$\begin{aligned}
\mathrm{E}\left[(\widehat{\theta} - \theta)^2\right] &= \mathrm{E}\left[(\widehat{\theta} - \mathrm{E}\,\widehat{\theta} + \mathrm{E}\,\widehat{\theta} - \theta)^2\right] \\
&= \mathrm{E}\left[(\widehat{\theta} - \mathrm{E}\,\widehat{\theta})^2\right] + 2\,\mathrm{E}\left[(\widehat{\theta} - \mathrm{E}\,\widehat{\theta})(\mathrm{E}\,\widehat{\theta} - \theta)\right] + \mathrm{E}\left[(\mathrm{E}\,\widehat{\theta} - \theta)^2\right] \\
&= \mathrm{Var}\,\widehat{\theta} + \mathrm{Bias}(\widehat{\theta}, \theta)^2,
\end{aligned}$$

as

$$\mathrm{E}\left[(\widehat{\theta} - \mathrm{E}\,\widehat{\theta})(\mathrm{E}\,\widehat{\theta} - \theta)\right] = (\mathrm{E}\,\widehat{\theta} - \theta)\,\underbrace{\mathrm{E}[\widehat{\theta} - \mathrm{E}\,\widehat{\theta}]}_{=0} = 0. \qquad \square$$

Let $f : \mathbb{R} \to \mathbb{R}$ be an unknown function and let $\widehat{f} : \mathbb{R} \to \mathbb{R}$ be a known approximation to f. Let $x_0 \in \mathbb{R}$ and suppose that

$$Y = f(x_0) + \varepsilon, \quad \text{where} \quad \mathrm{E}[\varepsilon] = 0.$$

The *squared prediction error* is

$$
\begin{aligned}
(f(x_0) - \widehat{f}(x_0))^2 &= \mathrm{E}\left[(\widehat{f}(x_0) - f(x_0))^2\right] \\
&= \mathrm{E}\left[(\widehat{f}(x_0) - Y - \varepsilon)^2\right] \\
&= \mathrm{E}\left[(Y - f + f - \widehat{f})^2\right] \\
&= \mathrm{E}\left[(Y - f)^2\right] + 2\,\mathrm{E}\left[(Y - f)(f - \widehat{f})\right] + \mathrm{E}\left[(f - \widehat{f})^2\right] \\
&= \mathrm{E}[\varepsilon^2] + 2\varepsilon\,\mathrm{E}[f - \widehat{f}] + \mathrm{Bias}(\widehat{f}, f)
\end{aligned}
$$

Let $\theta \in \mathbb{R}$, let ε be a random variable with $\mathrm{E}[\varepsilon] = 0$, and let

$$Y = \theta + \varepsilon.$$

Let $\widehat{\theta}$ be an estimator of θ such that $\widehat{\theta}$ and ε are independent.

$$
\begin{aligned}
\mathrm{E}[(\widehat{\theta} - Y)^2] &= \mathrm{E}[(\widehat{\theta} - \theta - \varepsilon)^2] \\
&= \mathrm{E}[(\widehat{\theta} - \mathrm{E}\,\widehat{\theta} + \mathrm{E}\,\widehat{\theta} - \theta - \varepsilon)^2] \\
&= \mathrm{E}[(\widehat{\theta} - \mathrm{E}\,\widehat{\theta})^2] + \mathrm{E}[(\mathrm{E}\,\widehat{\theta} - \theta)^2] + \mathrm{E}[\varepsilon^2] \\
&\quad + 2\,\mathrm{E}[(\widehat{\theta} - \mathrm{E}\,\widehat{\theta})(\mathrm{E}\,\widehat{\theta} - \theta)] - 2\,\mathrm{E}[(\widehat{\theta} - \mathrm{E}\,\widehat{\theta})\varepsilon] - 2\,\mathrm{E}[(\mathrm{E}\,\widehat{\theta} - \theta)\varepsilon] \\
&= \mathrm{Var}\,\widehat{\theta} + \mathrm{Bias}(\widehat{\theta}, \theta) + \mathrm{Var}\,\varepsilon
\end{aligned}
$$

We have:

- $\mathrm{E}[(\widehat{\theta} - \mathrm{E}\,\widehat{\theta})^2] = \mathrm{Var}\,\widehat{\theta}$
- $\mathrm{E}\,\widehat{\theta} - \theta$ is a constant, so
$$
\begin{aligned}
\mathrm{E}[(\mathrm{E}\,\widehat{\theta} - \theta)^2] &= (\mathrm{E}\,\widehat{\theta} - \theta)^2 = \mathrm{Bias}(\widehat{\theta}, \theta)^2, \\
\mathrm{E}[(\widehat{\theta} - \mathrm{E}\,\widehat{\theta})(\mathrm{E}\,\widehat{\theta} - \theta)] &= \mathrm{E}[\widehat{\theta} - \mathrm{E}\,\widehat{\theta}](\mathrm{E}\,\widehat{\theta} - \theta) = 0 && (\text{as } \mathrm{E}[\widehat{\theta} - \mathrm{E}\,\widehat{\theta}] = 0), \\
\mathrm{E}[(\mathrm{E}\,\widehat{\theta} - \theta)\varepsilon] &= (\mathrm{E}\,\widehat{\theta} - \theta)\,\mathrm{E}\,\varepsilon = 0 && (\text{as } \mathrm{E}\,\varepsilon = 0).
\end{aligned}
$$
- $\mathrm{E}[\varepsilon^2] = \mathrm{Var}\,\varepsilon$
- ε is independent of $\widehat{\theta}$ and, hence, of $\widehat{\theta} - \mathrm{E}\,\widehat{\theta}$. Therefore,
$$\mathrm{E}[(\widehat{\theta} - \mathrm{E}\,\widehat{\theta})\varepsilon] = \underbrace{\mathrm{E}[\widehat{\theta} - \mathrm{E}\,\widehat{\theta}]}_{=0}\,\mathrm{E}\,\varepsilon = 0.$$

If you can sample from a distribution, and you have an unbiased estimator, you can learn the parameters of the distribution. The amount of data you need depends on the variance of the estimator.

5. Notes

Statistics is the science of the *collection*, *analysis*, and *interpretation* of data. [TPE p. 1]

Data analysis: Oraganization and summarization of data. Emphasize main features. Expose underlying structure. Avoid extraneous assumptions.

Statistical inference: We postulate that the data are values realized by random variables obeying a probability distribution belonging to some known class, $\mathscr{P}$. Typically, $\mathscr{P}$ is indexed by some *parameter space,* Θ.

$$\mathscr{P} = \{P_\theta : \theta \in \Theta\}$$

We call the family $\mathscr{P}$ a *parametric* if $\Theta \subseteq \mathbb{R}^n$, for some n, and *nonparametric*, otherwise. In statistical inference, we use data to infer (point estimation) a plausible value of θ or (confidence sets) a subset of Θ that plausibly contains θ

The estimation problem: Given $g : \Theta \to \mathbb{R}$ and an $\mathscr{X}$-valued *random observable* X distributed according to some $P \in \mathscr{P}$, determine $g(\theta(P))$. An *estimator* is a function $\delta : \mathscr{X} \to \mathbb{R}$. We want to find an estimator δ such that $\delta(X) \approx g(\theta(P))$.

A *parametric family of distributions* is one that is naturally indexed by a subset Θ of some Euclidean space $\mathbb{R}^n$. The set Θ is called the *parameter space* of the family.

Suppose we are given a sample space $\mathscr{X} \subseteq \mathbb{R}^p$ and a family $\mathscr{P}$ of distributions on $\mathscr{X}$.

Let X be an $\mathscr{X}$-valued random vector such that $X \sim P$ for some unknown $P \in \mathscr{P}$.

Using data x realizing X to make draw conclusions about P is called *statistical inference.*

Let g be a *functional* (real-valued function) on $\mathscr{P}$.

Using data x realizing X to estimate $g(P)$ is called *point estimation.*

Point estimation is a type of statistical inference.

Let $P_{\mu,\sigma}$ be the distribution with density

$$\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}.$$

It's the distribution of an i.i.d. sample of size p drawn from $N(\mu, \sigma)$.

Using such a sample to estimate μ (resp., σ) is an example of point estimation with

$$\mathscr{X} = \mathbb{R}, \quad \mathscr{P} = \{P_{\mu,\sigma}\} \quad \text{and} \quad g(P_{\mu,\sigma}) = \mu \text{ (resp., } \sigma)$$

A *statistical functional on* $\mathscr{P}$ is a function $g : \mathscr{P} \to \mathbb{R}$. Let $\mathscr{P}$ be the set of all probability distributions on $\mathbb{R}$. For $a \in \mathbb{R}$, define

$$g_a(P) = \int_{-\infty}^{a} dP(x)$$

$$\mu(P) = \int_{-\infty}^{\infty} x\, dP(x)$$

$$m_k(P) = \int_{-\infty}^{\infty} (x - \mu(P))^k\, dP(x)$$

Estimating a functional $g : \mathscr{P} \to \mathbb{R}$ from data means constructing a function $\delta : \mathscr{X} \to \mathbb{R}$ such that for all distributions $P \in \mathscr{P}$ and all $\mathscr{X}$-valued random variables $X \sim P$, the quantity $\delta(X)$ is "close to" $g(P)$. We call g and δ the *estimand* and *estimator*, respectively.

We must make the descriptor "close to" precise if we are to evaluate the quality of an estimator δ of a functional g in any meaningful way. The notion of *bias* is a natural interpretation of closeness. Define

$$\mathrm{Bias}(\delta(X), g(P)) = \mathrm{E}\, \delta(X) - g(P).$$

If $\mathrm{Bias}(\delta(X), g(P)) < 0$ (resp., $\mathrm{Bias}(\delta(X), g(P)) > 0$), then $\delta(X)$ tends to underestimate (resp., overestimate) $g(P)$. We say that $\delta(X)$ is an *biased (resp., unbiased) estimator of* $g(P)$ if $\mathrm{Bias}(\delta(X), g(P)) \neq 0$ (resp., $\mathrm{Bias}(\delta(X), g(P)) = 0$).

Let $X \sim P \in \mathscr{P}$.

$$\mathrm{Bias}(\delta(X), g(P)) = \mathrm{E}[\delta(X) - g(P)]$$

Note that if $X \sim P$, then $\mathrm{E}\, \delta(X)$ depends only on δ and P and not on X:

$$\mathrm{E}[\delta(X)] = \int_{\mathscr{X}} \delta(x)\, dP(x)$$

Define the *(mean) bias* functional associated to the estimator δ of g,

$$\mathrm{Bias}(\delta, g) : \mathscr{P} \longrightarrow \mathbb{R}$$

by

$$P \mapsto \mathrm{Bias}_P(\delta, g) := \mathrm{E}[\delta(X)] - g(P),$$

where X is any $\mathscr{X}$-valued random variable such that $X \sim P$.

δ is an unbiased estimator of g if and only if $\mathrm{Bias}(\delta, g)$ is an unbiased estimator of the zero functional.

Mean bias vs. median bias. Exercise?

6. Logistic regression

Define the *sigmoid function*, also called the *expit function* or the *logistic function*, by

$$\sigma(x) = \frac{1}{1 - e^{-x}}.$$

It maps $\mathbb{R}$ bijectively onto $(0, 1)$. Its inverse is the *logit function*, defined by

$$\mathrm{logit}(x) = \log\left(\frac{x}{1 - x}\right).$$

The logit function maps $(0, 1)$ bijectively onto $\mathbb{R}$.

$$f(t) = y \log \sigma(t) + (1 - y) \log(1 - \sigma(t))$$

$$\begin{aligned}
f' &= y\frac{\sigma'}{\sigma} + (y-1)\frac{\sigma'}{1-\sigma} \\
&= \frac{1}{\sigma(1-\sigma)}\big[y\sigma'(1-\sigma) + (y-1)\sigma'\sigma\big] \\
&= \frac{\sigma'(y-\sigma)}{\sigma(1-\sigma)}
\end{aligned}$$

7. Newton's method

7.1. One variable. To refine an approximation $f(a) \approx 0$, we solve

$$\big(\text{first order approximation to } f(x) \text{ at } x = a\big) = 0.$$

$$f(a) + f'(a)(x-a) = 0$$

$$\text{refined approximation} = x = a - \frac{f(a)}{f'(a)}$$

Hence, Newton's recursion:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

7.2. Two variables. Suppose we want to solve

$$(f(\boldsymbol{x}), g(\boldsymbol{x})) = (0,0) = \mathbf{0}.$$

To refine an approximate solution $(f(a,b), g(a,b)) \approx (0,0)$, we solve the system

$$\begin{aligned}
\big(\text{first order approximation to } f(x,y) \text{ at } (x,y) = (a,b)\big) &= 0 \\
\big(\text{first order approximation to } g(x,y) \text{ at } (x,y) = (a,b)\big) &= 0.
\end{aligned}$$

$$\begin{gathered}
f(a,b) + f_x(a,b)(x-a) + f_y(a,b)(y-b) = 0 \\
g(a,b) + g_x(a,b)(x-a) + g_y(a,b)(y-b) = 0 \\
\begin{pmatrix} f(a,b) \\ g(a,b) \end{pmatrix} + \begin{pmatrix} f_x(a,b) & f_y(a,b) \\ g_x(a,b) & g_y(a,b) \end{pmatrix} \left[\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} a \\ b \end{pmatrix}\right]
\end{gathered}$$

$$\begin{aligned}
\text{refined approximation} = \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} a \\ b \end{pmatrix} - \begin{pmatrix} f_x(a,b) & f_y(a,b) \\ g_x(a,b) & g_y(a,b) \end{pmatrix}^{-1} \begin{pmatrix} f(a,b) \\ g(a,b) \end{pmatrix} \\
&= \begin{pmatrix} a \\ b \end{pmatrix} - \frac{1}{J_{f,g}(a,b)} \begin{pmatrix} g_y(a,b) & -f_y(a,b) \\ -g_x(a,b) & f_x(a,b) \end{pmatrix} \begin{pmatrix} f(a,b) \\ g(a,b) \end{pmatrix},
\end{aligned}$$

where

$$J_{f,g}(a,b) = \begin{vmatrix} f_x(a,b) & f_y(a,b) \\ g_x(a,b) & g_y(a,b) \end{vmatrix} = f_x(a,b)g_y(x,y) - f_y(a,b)g_x(a,b).$$

Hence, Newton's recursion:

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} x_n \\ y_n \end{pmatrix} - \frac{1}{|J_{f,g}(x_n,y_n)|} \begin{pmatrix} g_y(x_n,y_n) & -f_y(x_n,y_n) \\ -g_x(x_n,y_n) & f_x(x_n,y_n) \end{pmatrix} \begin{pmatrix} f(x_n,y_n) \\ g(x_n,y_n). \end{pmatrix}$$

7.3. **Application to simple logistic regression.**

$$\ell(a,b) = \text{log-likelihood}$$

To maximize $\ell(a,b)$, we solve

$$\nabla\ell(a,b) = \begin{pmatrix} \dfrac{\partial \ell}{\partial a} \\[2ex] \dfrac{\partial \ell}{\partial b} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

We apply the above with $f = \dfrac{\partial \ell}{\partial a}$ and $g = \dfrac{\partial \ell}{\partial b}$. In this case, the intervening matrix is the *Hessian matrix* of ℓ:

$$H_\ell = \begin{pmatrix} \dfrac{\partial^2 \ell}{\partial a^2} & \dfrac{\partial^2 \ell}{\partial a \partial b} \\[2ex] \dfrac{\partial^2 \ell}{\partial b \partial a} & \dfrac{\partial^2 \ell}{\partial b^2} \end{pmatrix}.$$

Newton's recursion becomes:

$$\begin{pmatrix} a_{n+1} \\ b_{n+1} \end{pmatrix} = \begin{pmatrix} a_n \\ b_n \end{pmatrix} - H_\ell(a_n,b_n)^{-1}\nabla\ell(a_n,b_n)$$

By equality of mixed partials, the *Hessian (determinant)* of ℓ is given by:

$$J_{\nabla\ell} = \frac{\partial^2 \ell}{\partial a^2}\frac{\partial^2 \ell}{\partial b^2} - \frac{\partial^2 \ell}{\partial a \partial b}$$

7.4. n**-variables.**

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \boldsymbol{f}'(\boldsymbol{x}_n)^{-1}\boldsymbol{f}(\boldsymbol{x}_n), \quad \text{where} \quad \boldsymbol{f}'(\boldsymbol{x}) = \left(\frac{\partial f_i}{\partial x_j}(\boldsymbol{x})\right)_{i,j=1,\ldots,n}.$$

8. Convexity

$$\sigma'(x) = \sigma(x)(1-\sigma(x)) \tag{1}$$

$$\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta),$$

where

$$\ell_i(\theta) = y^{(i)} \log \sigma(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}) + (1-y^{(i)}) \log(1-\sigma(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}))$$

$$\begin{aligned} \frac{\partial \ell_i}{\partial \theta_j} &= y^{(i)} \frac{\sigma'(\boldsymbol{\theta}^T \boldsymbol{x})}{\sigma(\boldsymbol{\theta}^T \boldsymbol{x})} x_j^{(i)} - (1-y^{(i)}) \frac{\sigma'(\boldsymbol{\theta}^T \boldsymbol{x})}{1-\sigma(\boldsymbol{\theta}^T \boldsymbol{x})} x_j^{(i)} \\ &= y^{(i)}(1-\sigma(\boldsymbol{\theta}^T \boldsymbol{x}))x_j^{(i)} - (1-y^{(i)})\sigma(\boldsymbol{\theta}^T \boldsymbol{x})x_j^{(i)} \\ &= (y^{(i)} - \sigma(\boldsymbol{\theta}^T \boldsymbol{x}))x_j^{(i)} \end{aligned}$$

$$\frac{\partial^2 \ell_i}{\partial\theta_j \partial\theta_k} = -\sigma'(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}) x_j^{(i)} x_k^{(i)}$$

$$\frac{\partial^2 \ell}{\partial\theta_j \partial\theta_k} = -\sum_{i=1}^{n} \sigma'(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}) x_j^{(i)} x_k^{(i)}$$

Let

$$X = \left(x_j^{(i)}\right) \in \mathbb{R}^{n\times p}, \quad D = \operatorname{diag}\left(\sigma'(\theta^T \boldsymbol{x}^{(1)}), \dots, \sigma'(\theta^T \boldsymbol{x}^{(n)})\right).$$

Then

$$H_\ell(\theta) = X^T D X.$$

9. Gradient descent

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} - \alpha \nabla f(\boldsymbol{\theta}^{(n)})^T$$

$$f(\boldsymbol{x}) \le f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})(\boldsymbol{x}-\boldsymbol{a}) + (\boldsymbol{x}-\boldsymbol{a})^T \nabla^2 f(\boldsymbol{a})(\boldsymbol{x}-\boldsymbol{a})$$

Descent lemma: Let $g(t) = f(\boldsymbol{x} + t\boldsymbol{y})$

$$\begin{aligned}
f(\boldsymbol{x}+\boldsymbol{y}) - f(\boldsymbol{x}) &= g(1) - g(0) \\
&= \int_0^1 g'(t)\,dt \\
&= \int_0^1 \nabla f(\boldsymbol{x}+t\boldsymbol{y})\boldsymbol{y}\,dt \\
&= \int_0^1 \big(f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}+t\boldsymbol{y}) - f(\boldsymbol{x})\big)\boldsymbol{y}\,dt \\
&= \int_0^1 \nabla f(\boldsymbol{x})\boldsymbol{y}\,dt + \int_0^1 \big(\nabla f(\boldsymbol{x}+t\boldsymbol{y}) - \nabla f(\boldsymbol{x})\big)\boldsymbol{y}\,dt \\
&\le \nabla f(\boldsymbol{x})\boldsymbol{y} + \int_0^1 \|\nabla f(\boldsymbol{x}+t\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|\,\|\boldsymbol{y}\|\,dt \\
&\le \nabla f(\boldsymbol{x})\boldsymbol{y} + \int_0^1 L\,\|t\boldsymbol{y}\|\,\|\boldsymbol{y}\|\,dt \\
&= \nabla f(\boldsymbol{x})\boldsymbol{y} + \frac{L}{2}\|\boldsymbol{y}\|^2
\end{aligned}$$

Replace $\boldsymbol{y}$ by $\boldsymbol{y}-\boldsymbol{x}$:

$$f(\boldsymbol{y}) \le f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2$$

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \frac{1}{L}\nabla f(\boldsymbol{x}_n)^T$$

$$f(x_{n+1}) \leq f(x_n) - \frac{1}{L}\|\nabla f(x_n)\|^2 + \frac{1}{2L}\|\nabla f(x_n)\|^2 = f(x_n) - \frac{1}{2L}\|\nabla f(x_n)\|^2$$

Being a decreasing sequence that is bounded below, $f(x_n)$ converges.

$$f(x_0) - f(x_n) = \sum_{k=0}^{n-1} \big(f(x_k) - f(x_{k+1})\big) \geq \frac{1}{2L}\sum_{k=0}^{n-1} \|\nabla f(x_k)\|^2$$

$$n \min_{k<n} \|\nabla f(x_k)\|^2 \leq \sum_{k=0}^{n-1} \|\nabla f(x_k)\|^2 \leq 2L(f(x_0) - f(x_n)) \leq 2L(f(x_0) - f(x_\infty))$$

$$\min_{k<n} \|\nabla f(x_k)\|^2 \leq \frac{2L(f(x_0) - f(x_\infty))}{n} = O(1/n)$$

$$\liminf_{n\to\infty} \|\nabla f(x_k)\|^2 = 0.$$

9.1. **Lipschitz constant for** MSE**.** Consider the mean square error function for simple linear regression with data set $(x_1, y_1), \ldots, (x_n, y_n)$:

$$f(a, b) = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i - y_i)^2$$

Then

$$\nabla f(a, b) = \frac{1}{n}\sum_{i=1}^{n} 2(a + bx_i - y_i)\begin{pmatrix}1 & x_i\end{pmatrix}$$

$$\nabla f(a, b) - \nabla f(a', b') = \frac{2}{n}\sum_{i=1}^{n}((a - a') + (b - b')x_i)\begin{pmatrix}1 & x_i\end{pmatrix}$$

Therefore,

$$\begin{aligned}
\|\nabla f(a, b) - \nabla f(a', b')\| &\leq \frac{2}{n}\sum_{i=1}^{n} |(a - a') + (b - b')x_i|\sqrt{1 + x_i^2} \\
&\leq \frac{2}{n}\big(|a - a'| + |b - b'|\big)\sum_{i=1}^{n} \max\{1, |x_i|\}\sqrt{1 + x_i^2} \\
&\leq \frac{2\sqrt{2}}{n}\sqrt{(a - a')^2 + (b - b')^2}\sum_{i=1}^{n} \max\{1, |x_i|\}\sqrt{1 + x_i^2} \\
&= L\|(a, b) - (a', b')\|,
\end{aligned}$$

where

$$L = \frac{2\sqrt{2}}{n}\sum_{i=1}^{n} \max\{1, |x_i|\}\sqrt{1 + x_i^2}$$

$$L \leq \frac{4}{n}\sum_{i=1}^{n}(1 + x_i^2) = 4\left(1 + \frac{S_{xx}}{n}\right)$$