TD for Generalized Linear Model and Model Assessment

TrungTin Nguyen

STATIFY team, Inria centre at the University Grenoble Alpes, France

Statistical Analysis and Document Mining

Complementary Course, Master of Applied Mathematics in Grenoble

Outline

Outline

1. Ordinary least square

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \ldots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.

1. Ordinary least square

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \ldots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- Given any error term ϵ, multiple linear regression function takes the form

$$Y = \beta_0 + \sum_{p=1}^{P} \beta_p X_p + \epsilon \equiv r(\mathbf{X}) + \epsilon, \quad \boldsymbol{\beta} \equiv (\beta_0, \beta_1, \ldots, \beta_P).$$

1. Ordinary least square

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \ldots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- Given any error term ϵ, multiple linear regression function takes the form

$$Y = \beta_0 + \sum_{p=1}^{P} \beta_p X_p + \epsilon \equiv r(\mathbf{X}) + \epsilon, \quad \boldsymbol{\beta} \equiv (\beta_0, \beta_1, \ldots, \beta_P).$$

- **Training error:** using ordinary least squares (OLS), we obtain coefficient estimates $\widehat{\beta}$ and $\widehat{r}_{\mathcal{D}}(\mathbf{x}_n) \equiv \widehat{r}_{\widehat{\beta},\mathcal{D}}(\mathbf{x}_n) \equiv \widehat{\beta}_0 + \sum_{p=1}^{P} \widehat{\beta}_p x_{np}$ such that $\forall n \in [N]$,

$$y_n \approx \widehat{r}_{\mathcal{D}}(\mathbf{x}_n), \text{ or } \mathrm{RSS}(\widehat{\beta}) \equiv \mathcal{L}(\widehat{r}_{\mathcal{D}}, \mathcal{D}) \equiv \frac{1}{N} \sum_{n=1}^{N} \left(y_n - \widehat{r}_{\mathcal{D}}(\mathbf{x}_n)\right)^2 \approx 0. \tag{1}$$

1. Ordinary least square

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \ldots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- Given any error term ϵ, multiple linear regression function takes the form

$$Y = \beta_0 + \sum_{p=1}^{P} \beta_p X_p + \epsilon \equiv r(\mathbf{X}) + \epsilon, \quad \boldsymbol{\beta} \equiv (\beta_0, \beta_1, \ldots, \beta_P).$$

- **Training error:** using ordinary least squares (OLS), we obtain coefficient estimates $\widehat{\beta}$ and $\widehat{r}_{\mathcal{D}}(\mathbf{x}_n) \equiv \widehat{r}_{\widehat{\beta},\mathcal{D}}(\mathbf{x}_n) \equiv \widehat{\beta}_0 + \sum_{p=1}^{P} \widehat{\beta}_p x_{np}$ such that $\forall n \in [N]$,

$$y_n \approx \widehat{r}_{\mathcal{D}}(\mathbf{x}_n), \text{ or } \mathrm{RSS}(\widehat{\beta}) \equiv \mathcal{L}(\widehat{r}_{\mathcal{D}}, \mathcal{D}) \equiv \frac{1}{N} \sum_{n=1}^{N} (y_n - \widehat{r}_{\mathcal{D}}(\mathbf{x}_n))^2 \approx 0. \qquad (1)$$

? **1. Least squares solution:** Prove that the solution for $\operatorname{argmin}_{\beta} \mathrm{RSS}(\beta)$ is given by: $\widehat{\boldsymbol{\beta}} \equiv (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_P) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$, is it the unique solution? Here $\mathbf{X} = (x_{n\times p})_{n\in[N],p\in[P]}$ is an $N \times P$ matrix with each row an input vector, and $\mathbf{y} = (y_n)_{n\in[N]}$ is an N-vector of the outputs in the training set.

Outline

Solution 1. Ordinary least square

Often it is convenient to include the constant variable 1 in $\mathbf{X}$, include β_0 in the vector of coefficients β, and then write the linear model in vector form as an inner product $Y = \mathbf{X}\beta \equiv \mathbf{X}_{N\times(P+1)}\beta_{(P+1)\times 1}$.

Solution 1. Ordinary least square

Often it is convenient to include the constant variable 1 in $\mathbf{X}$, include β_0 in the vector of coefficients β, and then write the linear model in vector form as an inner product $Y = \mathbf{X}\beta \equiv \mathbf{X}_{N\times(P+1)}\beta_{(P+1)\times 1}$. The solution is easiest to characterize in matrix notation.

Solution 1. Ordinary least square

Often it is convenient to include the constant variable 1 in $\mathbf{X}$, include β_0 in the vector of coefficients β, and then write the linear model in vector form as an inner product $Y = \mathbf{X}\beta \equiv \mathbf{X}_{N\times(P+1)}\beta_{(P+1)\times 1}$. The solution is easiest to characterize in matrix notation. We can write RSS in matrix form as $N\,\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)$.

Solution 1. Ordinary least square

Often it is convenient to include the constant variable 1 in $\mathbf{X}$, include β_0 in the vector of coefficients β, and then write the linear model in vector form as an inner product $Y = \mathbf{X}\beta \equiv \mathbf{X}_{N\times(P+1)}\beta_{(P+1)\times 1}$. The solution is easiest to characterize in matrix notation. We can write RSS in matrix form as $N\,\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)$. Now to minimize the function, set the derivative to zero, we obtain,

$$N\frac{\partial\,\mathrm{RSS}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = N\frac{\partial\left(\mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}} = -2N\mathbf{X}^\top\mathbf{y} + 2N\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = 0. \tag{2}$$

Solution 1. Ordinary least square

Often it is convenient to include the constant variable 1 in $\mathbf{X}$, include β_0 in the vector of coefficients β, and then write the linear model in vector form as an inner product $Y = \mathbf{X}\beta \equiv \mathbf{X}_{N\times(P+1)}\beta_{(P+1)\times 1}$. The solution is easiest to characterize in matrix notation. We can write RSS in matrix form as $N\,\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)$. Now to minimize the function, set the derivative to zero, we obtain,

$$N\frac{\partial\,\mathrm{RSS}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = N\frac{\partial\left(\mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}} = -2N\mathbf{X}^\top\mathbf{y} + 2N\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = 0. \tag{2}$$

If $\mathbf{X}$ is full rank, we get the unique solution in $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$.

Outline

Up to now we have made minimal assumptions about the true distribution of the data. In order to pin down the sampling properties of $\widehat{\boldsymbol{\beta}}$, we now assume that the observations y_n are uncorrelated and have constant variance σ^2, and that the $\mathbf{x}_n$ are fixed (non random). We also assume that the deviations of Y around its expectation are additive and Gaussian. That is, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathrm{I}_N)$.

? 2. Unbiased estimates:

a) Prove that $\mathbb{E}\left[\widehat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}$.

b) Calculate $\operatorname{var}\left[\widehat{\boldsymbol{\beta}}\right]$ and deduce an unbiased estimate of σ^2.

Outline

1. TD on Generalized Linear Model
   - 1. Ordinary least square
   - Solution 1. Ordinary least square
   - 2. Unbiased estimates
   - Solution 2. Unbiased estimates
   - 3. Expected squared prediction errors
   - Solution 3. Expected squared prediction errors
   - 4. Expected absolute prediction errors
   - Solution 4. Expected absolute prediction errors
   - 5. Reduced weighted least squares problem
   - Solution 5. Reduced weighted least squares problem

2. Perspectives

Solution 2. Unbiased estimates

a) Prove that $\mathbb{E}\left[\widehat{\beta}\right] = \beta$.

Solution 2. Unbiased estimates

a) Prove that $\mathbb{E}\left[\widehat{\beta}\right] = \beta$. Using the fact that $\mathbb{E}[\mathbf{y}] = \mathbf{X}\beta$,

Solution 2. Unbiased estimates

a) Prove that $\mathbb{E}\left[\widehat{\beta}\right] = \beta$. Using the fact that $\mathbb{E}\left[\mathbf{y}\right] = \mathbf{X}\beta$, we get $\mathbb{E}\left[\widehat{\beta}\right] = \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{X}\beta$. Here, ϵ is a random column vector of dimension N.

Solution 2. Unbiased estimates

a) Prove that $\mathbb{E}\left[\widehat{\beta}\right] = \beta$. Using the fact that $\mathbb{E}[\mathbf{y}] = \mathbf{X}\beta$, we get $\mathbb{E}\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X}\beta$.
Here, ϵ is a random column vector of dimension N.
b) Calculate var $\left[\widehat{\beta}\right]$ and deduce an unbiased estimate of σ^2.

Solution 2. Unbiased estimates

a) Prove that $\mathbb{E}\left[\widehat{\beta}\right] = \beta$. Using the fact that $\mathbb{E}\left[\mathbf{y}\right] = \mathbf{X}\beta$, we get $\mathbb{E}\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X}\beta$.
Here, ϵ is a random column vector of dimension N.
b) Calculate var $\left[\widehat{\beta}\right]$ and deduce an unbiased estimate of σ^2.
$\operatorname{var}\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \operatorname{var}\left[\epsilon\right] \mathbf{X} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$.

Solution 2. Unbiased estimates

a) Prove that $\mathbb{E}\left[\widehat{\beta}\right] = \beta$. Using the fact that $\mathbb{E}\left[\mathbf{y}\right] = \mathbf{X}\beta$, we get $\mathbb{E}\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\beta$.
Here, ϵ is a random column vector of dimension N.
b) Calculate var $\left[\widehat{\beta}\right]$ and deduce an unbiased estimate of σ^2.
var $\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top \mathrm{var}\left[\epsilon\right]\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$. Then var $\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\sigma^2$.

Solution 2. Unbiased estimates

a) Prove that $\mathbb{E}\left[\widehat{\beta}\right] = \beta$. Using the fact that $\mathbb{E}[\mathbf{y}] = \mathbf{X}\beta$, we get $\mathbb{E}\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X}\beta$.
Here, ϵ is a random column vector of dimension N.
b) Calculate var $\left[\widehat{\beta}\right]$ and deduce an unbiased estimate of σ^2.
var $\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathrm{var}[\epsilon] \mathbf{X} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$. Then var $\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \sigma^2$. It remains to specify how to determine σ^2. To that end once β is estimated we can compute

$$\widehat{\sigma}^2 = \frac{1}{N-P-1} \sum_{n=1}^{N} \left(y_n - \mathbf{x}_n \widehat{\beta}\right)^2 \equiv \frac{1}{N-P-1} \|\mathbf{y} - \mathbf{X}\widehat{\beta}\|_2^2. \tag{3}$$

Solution 2. Unbiased estimates

a) Prove that $\mathbb{E}\left[\widehat{\beta}\right] = \beta$. Using the fact that $\mathbb{E}[\mathbf{y}] = \mathbf{X}\beta$, we get $\mathbb{E}\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\beta$.
Here, ϵ is a random column vector of dimension N.
b) Calculate $\operatorname{var}\left[\widehat{\beta}\right]$ and deduce an unbiased estimate of σ^2.
$\operatorname{var}\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\operatorname{var}[\epsilon]\,\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$. Then $\operatorname{var}\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\sigma^2$. It remains to specify how to determine σ^2. To that end once β is estimated we can compute

$$\widehat{\sigma}^2 = \frac{1}{N-P-1}\sum_{n=1}^{N}\left(y_n - \mathbf{x}_n\widehat{\beta}\right)^2 \equiv \frac{1}{N-P-1}\|\mathbf{y} - \mathbf{X}\widehat{\beta}\|_2^2. \tag{3}$$

We have

$$\mathbb{E}\left[\widehat{\sigma}^2\right] = \frac{1}{N-P-1}\sum_{n=1}^{N}\mathbb{E}\left[\left(y_n - \widehat{\mathbf{x}}_n\beta\right)^2\right] = \frac{(N-P-1)\sigma^2}{N-P-1}. \tag{4}$$

Solution 2. Unbiased estimates

a) Prove that $\mathbb{E}\left[\widehat{\beta}\right] = \beta$. Using the fact that $\mathbb{E}\left[\mathbf{y}\right] = \mathbf{X}\beta$, we get $\mathbb{E}\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\beta$.
Here, ϵ is a random column vector of dimension N.
b) Calculate var $\left[\widehat{\beta}\right]$ and deduce an unbiased estimate of σ^2.
var $\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top \mathrm{var}\left[\epsilon\right]\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$. Then var $\left[\widehat{\beta}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\sigma^2$. It remains to specify how to determine σ^2. To that end once β is estimated we can compute

$$\widehat{\sigma}^2 = \frac{1}{N-P-1}\sum_{n=1}^{N}\left(y_n - \mathbf{x}_n\widehat{\beta}\right)^2 \equiv \frac{1}{N-P-1}\|\mathbf{y} - \mathbf{X}\widehat{\beta}\|_2^2. \tag{3}$$

We have

$$\mathbb{E}\left[\widehat{\sigma}^2\right] = \frac{1}{N-P-1}\sum_{n=1}^{N}\mathbb{E}\left[\left(y_n - \widehat{\mathbf{x}}_n\beta\right)^2\right] = \frac{(N-P-1)\sigma^2}{N-P-1}. \tag{4}$$

Here, we used the fact that $\widehat{\beta} \sim \mathcal{N}\left(\beta, \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\sigma^2\right)$, and $(N-P-1)\widehat{\sigma}^2 \sim \sigma^2\chi^2_{N-P-1}$, a chi-squared distribution with $N-P-1$ degrees of freedom.

Solution 2. Unbiased estimates

a) Prove that $\mathbb{E}\left[\widehat{\beta}\right]=\beta$. Using the fact that $\mathbb{E}\left[\mathbf{y}\right]=\mathbf{X}\beta$, we get $\mathbb{E}\left[\widehat{\beta}\right]=\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\beta$. Here, ϵ is a random column vector of dimension N.
b) Calculate $\operatorname{var}\left[\widehat{\beta}\right]$ and deduce an unbiased estimate of σ^2.
$\operatorname{var}\left[\widehat{\beta}\right]=\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\operatorname{var}\left[\epsilon\right]\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$. Then $\operatorname{var}\left[\widehat{\beta}\right]=\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\sigma^2$. It remains to specify how to determine σ^2. To that end once β is estimated we can compute

$$\widehat{\sigma}^2=\frac{1}{N-P-1}\sum_{n=1}^{N}\left(y_n-\mathbf{x}_n\widehat{\beta}\right)^2\equiv\frac{1}{N-P-1}\|\mathbf{y}-\mathbf{X}\widehat{\beta}\|_2^2. \tag{3}$$

We have

$$\mathbb{E}\left[\widehat{\sigma}^2\right]=\frac{1}{N-P-1}\sum_{n=1}^{N}\mathbb{E}\left[\left(y_n-\widehat{\mathbf{x}}_n\beta\right)^2\right]=\frac{(N-P-1)\sigma^2}{N-P-1}. \tag{4}$$

Here, we used the fact that $\widehat{\beta}\sim\mathcal{N}\left(\beta,\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\sigma^2\right)$, and $(N-P-1)\widehat{\sigma}^2\sim\sigma^2\chi^2_{N-P-1}$, a chi-squared distribution with $N-P-1$ degrees of freedom.
Recall that if $\mathbf{X}\sim\mathcal{N}(\mu,\sigma^2\,\mathrm{I}_N)$ then $1/(\sigma^2)\|\mathbf{X}-\boldsymbol{\mu}\|_2^2\sim\chi^2_N$. Furthermore, if $\mathbf{X}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ then $\boldsymbol{A}\mathbf{X}+\boldsymbol{b}\sim\mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{b},\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)$.

Solution 2. Unbiased estimates

$\boxed{\text{Why } (N-P-1)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{N-P-1}?}$

Notice that if $\mathbf{A}$ is an idempotent (i.e. $\mathbf{A}^2 = \mathbf{A}$) and symmetric matrix of rank r and $\mathbf{Z} \sim \mathcal{N}(0, \sigma^2 I)$, then $\mathbf{Z}^\top \mathbf{A}\mathbf{Z}/\sigma^2 \sim \chi^2_r$ (with r is a trace of $\mathbf{A}$).

Solution 2. Unbiased estimates

$\boxed{\text{Why } (N-P-1)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{N-P-1}?}$

Notice that if $\mathbf{A}$ is an idempotent (i.e. $\mathbf{A}^2 = \mathbf{A}$) and symmetric matrix of rank r and $\mathbf{Z} \sim \mathcal{N}(0, \sigma^2 I)$, then $\mathbf{Z}^\top \mathbf{A}\mathbf{Z}/\sigma^2 \sim \chi^2_r$ (with r is a trace of $\mathbf{A}$).

Proof: (hint) Decomposition of symmetric idempotent matrix.

Solution 2. Unbiased estimates

$\boxed{\text{Why } (N-P-1)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{N-P-1}?}$

Notice that if $\mathbf{A}$ is an idempotent (i.e. $\mathbf{A}^2 = \mathbf{A}$) and symmetric matrix of rank r and $\mathbf{Z} \sim \mathcal{N}(0, \sigma^2 I)$, then $\mathbf{Z}^\top \mathbf{A}\mathbf{Z}/\sigma^2 \sim \chi^2_r$ (with r is a trace of $\mathbf{A}$).
Proof: (hint) Decomposition of symmetric idempotent matrix.

We have $(\mathbf{y} - \mathbf{X}\widehat{\beta}) = (\mathbf{y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}) = (I_N - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{y} = H\mathbf{y} = H(\mathbf{X}\beta + \epsilon) = H\epsilon$, where $H = I_N - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ and $H^2 = H$.

Solution 2. Unbiased estimates

$\boxed{\text{Why } (N-P-1)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{N-P-1}?}$

Notice that if $\mathbf{A}$ is an idempotent (i.e. $\mathbf{A}^2 = \mathbf{A}$) and symmetric matrix of rank r and $\mathbf{Z} \sim \mathcal{N}(0, \sigma^2 I)$, then $\mathbf{Z}^\top \mathbf{A}\mathbf{Z}/\sigma^2 \sim \chi^2_r$ (with r is a trace of $\mathbf{A}$).

Proof: (hint) Decomposition of symmetric idempotent matrix.

We have $(\mathbf{y} - \mathbf{X}\widehat{\beta}) = (\mathbf{y} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}) = (I_N - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{y} = H\mathbf{y} = H(\mathbf{X}\beta + \epsilon) = H\epsilon$, where $H = I_N - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$ and $H^2 = H$. Hence, $(N-P-1)\widehat{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\widehat{\beta}\|_2^2 = (\mathbf{y} - \mathbf{X}\widehat{\beta})^\top(\mathbf{y} - \mathbf{X}\widehat{\beta}) = \epsilon^\top H\epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

Solution 2. Unbiased estimates

$\boxed{\text{Why } (N-P-1)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{N-P-1}?}$

Notice that if $\mathbf{A}$ is an idempotent (i.e. $\mathbf{A}^2 = \mathbf{A}$) and symmetric matrix of rank r and $\mathbf{Z} \sim \mathcal{N}(0, \sigma^2 I)$, then $\mathbf{Z}^\top \mathbf{A}\mathbf{Z}/\sigma^2 \sim \chi^2_r$ (with r is a trace of $\mathbf{A}$).
Proof: (hint) Decomposition of symmetric idempotent matrix.

We have $(\mathbf{y} - \mathbf{X}\widehat{\beta}) = (\mathbf{y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}) = (I_N - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{y} = H\mathbf{y} = H(\mathbf{X}\beta + \epsilon) = H\epsilon$, where $H = I_N - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ and $H^2 = H$.
Hence, $(N-P-1)\widehat{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\widehat{\beta}\|_2^2 = (\mathbf{y} - \mathbf{X}\widehat{\beta})^\top(\mathbf{y} - \mathbf{X}\widehat{\beta}) = \epsilon^\top H\epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.
Moreover, we recall that $\mathbf{X}$ is an $N \times (P+1)$ matrix with $(P+1) \leq N$ and we assume that $\mathbf{X}$ has full rank then

Solution 2. Unbiased estimates

$\boxed{\text{Why } (N-P-1)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{N-P-1}?}$

Notice that if $\mathbf{A}$ is an idempotent (i.e. $\mathbf{A}^2 = \mathbf{A}$) and symmetric matrix of rank r and $\mathbf{Z} \sim \mathcal{N}(0, \sigma^2 I)$, then $\mathbf{Z}^\top \mathbf{A}\mathbf{Z}/\sigma^2 \sim \chi^2_r$ (with r is a trace of $\mathbf{A}$).
Proof: (hint) Decomposition of symmetric idempotent matrix.

We have $(\mathbf{y} - \mathbf{X}\widehat{\beta}) = (\mathbf{y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}) = (I_N - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{y} = H\mathbf{y} = H(\mathbf{X}\beta + \epsilon) = H\epsilon$, where $H = I_N - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ and $H^2 = H$.
Hence, $(N-P-1)\widehat{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\widehat{\beta}\|_2^2 = (\mathbf{y} - \mathbf{X}\widehat{\beta})^\top(\mathbf{y} - \mathbf{X}\widehat{\beta}) = \epsilon^\top H\epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.
Moreover, we recall that $\mathbf{X}$ is an $N \times (P+1)$ matrix with $(P+1) \leq N$ and we assume that $\mathbf{X}$ has full rank then
$rank(\mathbf{X}) = \min(N, P+1) = P+1$ so that $(\mathbf{X}^\top\mathbf{X})^{-1}$ exists.

Solution 2. Unbiased estimates

$\boxed{\text{Why } (N-P-1)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{N-P-1}?}$

Notice that if $\mathbf{A}$ is an idempotent (i.e. $\mathbf{A}^2 = \mathbf{A}$) and symmetric matrix of rank r and $\mathbf{Z} \sim \mathcal{N}(0, \sigma^2 I)$, then $\mathbf{Z}^\top \mathbf{A}\mathbf{Z}/\sigma^2 \sim \chi^2_r$ (with r is a trace of $\mathbf{A}$).
Proof: (hint) Decomposition of symmetric idempotent matrix.

We have $(\mathbf{y} - \mathbf{X}\widehat{\beta}) = (\mathbf{y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}) = (I_N - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{y} = H\mathbf{y} = H(\mathbf{X}\beta + \epsilon) = H\epsilon$, where $H = I_N - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ and $H^2 = H$.
Hence, $(N-P-1)\widehat{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\widehat{\beta}\|_2^2 = (\mathbf{y} - \mathbf{X}\widehat{\beta})^\top(\mathbf{y} - \mathbf{X}\widehat{\beta}) = \epsilon^\top H\epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.
Moreover, we recall that $\mathbf{X}$ is an $N \times (P+1)$ matrix with $(P+1) \leq N$ and we assume that $\mathbf{X}$ has full rank then $rank(\mathbf{X}) = \min(N, P+1) = P+1$ so that $(\mathbf{X}^\top\mathbf{X})^{-1}$ exists.
By the commutativity of the trace operator
$tr(H) = tr(I_N) - tr(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) = N - tr(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}) = N - tr(I_{P+1}) = N - (P+1)$.

Outline

3. Expected squared prediction errors

We consider the expect prediction error
$\text{EPSE}(r) = \mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \int (y - r(\mathbf{x}))^2 \Pr(d\mathbf{x}, dy)$. Let $\mathbf{X} \in \mathbb{R}^P$ denote a real valued random input vector, and $Y \in \mathbb{R}$ a real valued random output variable, with Pr is a joint distribution of $(\mathbf{X}, Y)$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

We consider the expect prediction error
$\mathrm{EPSE}(r) = \mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \int (y - r(\mathbf{x}))^2 \Pr(d\mathbf{x}, dy)$. Let $\mathbf{X} \in \mathbb{R}^P$ denote a real valued random input vector, and $Y \in \mathbb{R}$ a real valued random output variable, with Pr is a joint distribution of $(\mathbf{X}, Y)$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 3. Expected squared prediction errors:

a) Prove that we can write EPSE as
$\mathrm{EPSE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[(Y - r(\mathbf{X}))^2 | \mathbf{X}\right]\right]$.

b) Prove that the (pointwise w.r.t. $\mathbf{x}$) minimizer solution of $\operatorname{argmin}_r \mathrm{EPSE}(r)$ is given by $\widehat{r}(\mathbf{x}) = \mathbb{E}\left[Y | \mathbf{X} = \mathbf{x}\right]$.

Outline

Solution 3. Expected squared prediction errors

We consider the expect prediction error $\mathrm{EPSE}(r) = \mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \int (y - r(\mathbf{x}))^2 \Pr(d\mathbf{x}, dy)$. Let $\mathbf{X} \in \mathbb{R}^P$ denote a real valued random input vector, and $Y \in \mathbb{R}$ a real valued random output variable, with Pr is a joint distribution of $(\mathbf{X}, Y)$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

Solution 3. Expected squared prediction errors

We consider the expect prediction error $\text{EPSE}(r) = \mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \int (y - r(\mathbf{x}))^2 \Pr(d\mathbf{x}, dy)$. Let $\mathbf{X} \in \mathbb{R}^P$ denote a real valued random input vector, and $Y \in \mathbb{R}$ a real valued random output variable, with Pr is a joint distribution of $(\mathbf{X}, Y)$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 3. Expected squared prediction errors:

a) Prove that we can write EPE as $\text{EPSE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[(Y - r(\mathbf{X}))^2 | \mathbf{X}\right]\right]$.

Solution 3. Expected squared prediction errors

We consider the expect prediction error
$\mathrm{EPSE}(r) = \mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \int (y - r(\mathbf{x}))^2 \Pr(d\mathbf{x}, dy)$. Let $\mathbf{X} \in \mathbb{R}^P$ denote a real valued random input vector, and $Y \in \mathbb{R}$ a real valued random output variable, with Pr is a joint distribution of $(\mathbf{X}, Y)$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 3. Expected squared prediction errors:

a) Prove that we can write EPE as $\mathrm{EPSE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[(Y - r(\mathbf{X}))^2 | \mathbf{X}\right]\right]$.

Recall that we can factor the joint density as $\Pr(\mathbf{X}, Y) = \Pr(Y|\mathbf{X}) \Pr(\mathbf{X})$.

We consider the expect prediction error
$\mathrm{EPSE}(r) = \mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \int (y - r(\mathbf{x}))^2 \Pr(d\mathbf{x}, dy)$. Let $\mathbf{X} \in \mathbb{R}^P$ denote a real valued random input vector, and $Y \in \mathbb{R}$ a real valued random output variable, with Pr is a joint distribution of $(\mathbf{X}, Y)$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 3. Expected squared prediction errors:

a) Prove that we can write EPE as $\mathrm{EPSE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[(Y - r(\mathbf{X}))^2 | \mathbf{X}\right]\right]$.

Recall that we can factor the joint density as $\Pr(\mathbf{X}, Y) = \Pr(Y|\mathbf{X}) \Pr(\mathbf{X})$. Then, $\mathrm{EPSE}(r) = \mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \int_{\mathbf{x}} \int_y (y - r(\mathbf{x}))^2 \Pr(y|\mathbf{x}) \Pr(\mathbf{x}) dy d\mathbf{x}$.

We consider the expect prediction error
$\mathrm{EPSE}(r) = \mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \int (y - r(\mathbf{x}))^2 \Pr(d\mathbf{x}, dy)$. Let $\mathbf{X} \in \mathbb{R}^P$ denote a real valued random input vector, and $Y \in \mathbb{R}$ a real valued random output variable, with Pr is a joint distribution of $(\mathbf{X}, Y)$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 3. Expected squared prediction errors:

a) Prove that we can write EPE as $\mathrm{EPSE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[(Y - r(\mathbf{X}))^2 | \mathbf{X}\right]\right]$.

Recall that we can factor the joint density as $\Pr(\mathbf{X}, Y) = \Pr(Y|\mathbf{X}) \Pr(\mathbf{X})$. Then, $\mathrm{EPSE}(r) = \mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \int_{\mathbf{x}} \int_y (y - r(\mathbf{x}))^2 \Pr(y|\mathbf{x}) \Pr(\mathbf{x}) dy d\mathbf{x}$.

b) Prove that the (pointwise w.r.t. $\mathbf{x}$) minimizer solution of $\mathrm{argmin}_f \mathrm{EPSE}(f)$ is given by $\widehat{r}(\mathbf{x}) = \mathbb{E}[Y | \mathbf{X} = \mathbf{x}]$.

We consider the expect prediction error $\text{EPSE}(r) = \mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \int (y - r(\mathbf{x}))^2 \Pr(d\mathbf{x}, dy)$. Let $\mathbf{X} \in \mathbb{R}^P$ denote a real valued random input vector, and $Y \in \mathbb{R}$ a real valued random output variable, with Pr is a joint distribution of $(\mathbf{X}, Y)$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 3. Expected squared prediction errors:

a) Prove that we can write EPE as $\text{EPSE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[(Y - r(\mathbf{X}))^2|\mathbf{X}\right]\right]$.
Recall that we can factor the joint density as $\Pr(\mathbf{X}, Y) = \Pr(Y|\mathbf{X})\Pr(\mathbf{X})$. Then, $\text{EPSE}(r) = \mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \int_{\mathbf{x}}\int_{y}(y - r(\mathbf{x}))^2 \Pr(y|\mathbf{x})\Pr(\mathbf{x})dyd\mathbf{x}$.

b) Prove that the (pointwise w.r.t. $\mathbf{x}$) minimizer solution of $\text{argmin}_f \text{EPSE}(f)$ is given by $\widehat{r}(\mathbf{x}) = \mathbb{E}\left[Y|\mathbf{X} = \mathbf{x}\right]$. Notice that by conditioning on $\mathbf{X}$, we have freed the dependency of the function r on $\mathbf{X}$ and since the quantity $(Y - r)^2$ is convex, there is a unique solution.

Solution 3. Expected squared prediction errors

We consider the expect prediction error $\text{EPSE}(r) = \mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \int (y - r(\mathbf{x}))^2 \Pr(d\mathbf{x}, dy)$. Let $\mathbf{X} \in \mathbb{R}^P$ denote a real valued random input vector, and $Y \in \mathbb{R}$ a real valued random output variable, with Pr is a joint distribution of $(\mathbf{X}, Y)$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 3. Expected squared prediction errors:

a) Prove that we can write EPE as $\text{EPSE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[(Y - r(\mathbf{X}))^2|\mathbf{X}\right]\right]$.
Recall that we can factor the joint density as $\Pr(\mathbf{X}, Y) = \Pr(Y|\mathbf{X}) \Pr(\mathbf{X})$. Then, $\text{EPSE}(r) = \mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \int_{\mathbf{x}} \int_y (y - r(\mathbf{x}))^2 \Pr(y|\mathbf{x}) \Pr(\mathbf{x}) dy d\mathbf{x}$.

b) Prove that the (pointwise w.r.t. $\mathbf{x}$) minimizer solution of $\text{argmin}_f \text{EPSE}(f)$ is given by $\widehat{r}(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. Notice that by conditioning on $\mathbf{X}$, we have freed the dependency of the function r on $\mathbf{X}$ and since the quantity $(Y - r)^2$ is convex, there is a unique solution. We see that it suffices to minimize EPE pointwise:

$$r(\mathbf{x}) = \underset{f}{\text{argmin}}\, \mathbb{E}_{Y|\mathbf{X}}\left[(Y - r(\mathbf{X}))^2|\mathbf{X} = \mathbf{x}\right]. \tag{5}$$

Solution 3. Expected squared prediction errors

We consider the expect prediction error $\text{EPSE}(r) = \mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \int (y - r(\mathbf{x}))^2 \Pr(d\mathbf{x}, dy)$. Let $\mathbf{X} \in \mathbb{R}^P$ denote a real valued random input vector, and $Y \in \mathbb{R}$ a real valued random output variable, with Pr is a joint distribution of $(\mathbf{X}, Y)$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 3. Expected squared prediction errors:

a) Prove that we can write EPE as $\text{EPSE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[(Y - r(\mathbf{X}))^2|\mathbf{X}\right]\right]$.

Recall that we can factor the joint density as $\Pr(\mathbf{X}, Y) = \Pr(Y|\mathbf{X})\Pr(\mathbf{X})$. Then, $\text{EPSE}(r) = \mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \int_{\mathbf{x}}\int_{y}(y - r(\mathbf{x}))^2 \Pr(y|\mathbf{x})\Pr(\mathbf{x})dyd\mathbf{x}$.

b) Prove that the (pointwise w.r.t. $\mathbf{x}$) minimizer solution of $\text{argmin}_f \text{EPSE}(f)$ is given by $\widehat{r}(\mathbf{x}) = \mathbb{E}\left[Y|\mathbf{X} = \mathbf{x}\right]$. Notice that by conditioning on $\mathbf{X}$, we have freed the dependency of the function r on $\mathbf{X}$ and since the quantity $(Y - r)^2$ is convex, there is a unique solution. We see that it suffices to minimize EPE pointwise:

$$r(\mathbf{x}) = \underset{f}{\text{argmin}}\, \mathbb{E}_{Y|\mathbf{X}}\left[(Y - r(\mathbf{X}))^2|\mathbf{X} = \mathbf{x}\right]. \tag{5}$$

By taking the derivative of conditional expectation w.r.t. r and set to zero, we obtain $\widehat{r}(\mathbf{x}) = \mathbb{E}\left[Y|\mathbf{X} = \mathbf{x}\right]$.

Outline

4. Expected absolute prediction errors

Are we happy with the expected squared prediction errors criterion? What happens if we replace the L_2 loss function with the $L_1 : \mathbb{E}\left[|\mathbf{y} - \mathbf{X}|\right]$?

Are we happy with the expected squared prediction errors criterion? What happens if we replace the L_2 loss function with the $L_1 : \mathbb{E}\left[|\mathbf{y} - \mathbf{X}|\right]$?

We consider the expect absolute prediction error $\text{EPAE}(r) = \mathbb{E}\left[|Y - r(\mathbf{X})|\right] = \int |y - r(\mathbf{x})| \Pr(d\mathbf{x}, dy)$. Let $\mathbf{X} \in \mathbb{R}^P$ denote a real valued random input vector, and $Y \in \mathbb{R}$ a real valued random output variable, with Pr is a joint distribution of $(\mathbf{X}, Y)$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 4. Expected absolute prediction errors:

a) Prove that we can write EPAE as $\text{EPAE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[|Y - r(\mathbf{X})||\mathbf{X}\right]\right]$.

b) What is the (pointwise w.r.t. $\mathbf{x}$) minimizer solution of $\operatorname{argmin}_r \text{EPAE}(r)$?

Outline

1. TD on Generalized Linear Model
   - 1. Ordinary least square
   - Solution 1. Ordinary least square
   - 2. Unbiased estimates
   - Solution 2. Unbiased estimates
   - 3. Expected squared prediction errors
   - Solution 3. Expected squared prediction errors
   - 4. Expected absolute prediction errors
   - Solution 4. Expected absolute prediction errors
   - 5. Reduced weighted least squares problem
   - Solution 5. Reduced weighted least squares problem
2. Perspectives

Solution 4. Expected absolute prediction errors

Are we happy with the expected squared prediction errors criterion? What happens if we replace the L_2 loss function with the $L_1 : \mathbb{E}\left[|\mathbf{y} - \mathbf{X}|\right]$?

Solution 4. Expected absolute prediction errors

Are we happy with the expected squared prediction errors criterion? What happens if we replace the L_2 loss function with the $L_1 : \mathbb{E}\left[|\mathbf{y} - \mathbf{X}|\right]$? Its estimates are more robust than those for the conditional mean. L_1 criteria have discontinuities in their derivatives, which have hindered their widespread use.

Solution 4. Expected absolute prediction errors

Are we happy with the expected squared prediction errors criterion? What happens if we replace the L_2 loss function with the $L_1 : \mathbb{E}\left[|\mathbf{y} - \mathbf{X}|\right]$? Its estimates are more robust than those for the conditional mean. L_1 criteria have discontinuities in their derivatives, which have hindered their widespread use.

We consider the expect absolute prediction error $\mathrm{EPAE}(r) = \mathbb{E}\left[|Y - r(\mathbf{X})|\right]$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 4. Expected absolute prediction errors:

a) Prove that we can write EPAE as $\mathrm{EPAE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[|Y - r(\mathbf{X})||\mathbf{X}\right]\right]$. The same solution as Question 3.

Solution 4. Expected absolute prediction errors

Are we happy with the expected squared prediction errors criterion? What happens if we replace the L_2 loss function with the $L_1 : \mathbb{E}\left[|\mathbf{y} - \mathbf{X}|\right]$? Its estimates are more robust than those for the conditional mean. L_1 criteria have discontinuities in their derivatives, which have hindered their widespread use.

We consider the expect absolute prediction error $\mathrm{EPAE}(r) = \mathbb{E}\left[|Y - r(\mathbf{X})|\right]$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 4. Expected absolute prediction errors:

a) Prove that we can write EPAE as $\mathrm{EPAE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[|Y - r(\mathbf{X})||\mathbf{X}\right]\right]$. The same solution as Question 3.

b) What is the (pointwise w.r.t. $\mathbf{x}$) minimizer solution of $\mathrm{argmin}_r\, \mathrm{EPAE}(r)$?

Solution 4. Expected absolute prediction errors

Are we happy with the expected squared prediction errors criterion? What happens if we replace the L_2 loss function with the $L_1 : \mathbb{E}\left[|\mathbf{y} - \mathbf{X}|\right]$? Its estimates are more robust than those for the conditional mean. L_1 criteria have discontinuities in their derivatives, which have hindered their widespread use.

We consider the expect absolute prediction error $\mathrm{EPAE}(r) = \mathbb{E}\left[|Y - r(\mathbf{X})|\right]$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 4. Expected absolute prediction errors:

a) Prove that we can write EPAE as $\mathrm{EPAE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[|Y - r(\mathbf{X})||\mathbf{X}\right]\right]$. The same solution as Question 3.

b) What is the (pointwise w.r.t. $\mathbf{x}$) minimizer solution of $\mathrm{argmin}_r\, \mathrm{EPAE}(r)$? The integration is a little sophisticated and requires a branch of analysis known as measure theory.

Solution 4. Expected absolute prediction errors

Are we happy with the expected squared prediction errors criterion? What happens if we replace the L_2 loss function with the $L_1 : \mathbb{E}[|\mathbf{y} - \mathbf{X}|]$? Its estimates are more robust than those for the conditional mean. L_1 criteria have discontinuities in their derivatives, which have hindered their widespread use.

We consider the expect absolute prediction error $\text{EPAE}(r) = \mathbb{E}[|Y - r(\mathbf{X})|]$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 4. Expected absolute prediction errors:

a) Prove that we can write EPAE as $\text{EPAE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}[|Y - r(\mathbf{X})||\mathbf{X}]\right]$. The same solution as Question 3.

b) What is the (pointwise w.r.t. $\mathbf{x}$) minimizer solution of $\text{argmin}_r \text{EPAE}(r)$? The integration is a little sophisticated and requires a branch of analysis known as measure theory. For this reason, we will provide an approximation that does not address the smaller details. By using the law of large numbers, we get the following

$$\int_y |Y - r| \Pr(y|\mathbf{X}) dy = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} |Y_n - r| \approx \frac{1}{N} \sum_{n=1}^{N} |Y_n - r|. \tag{6}$$

Solution 4. Expected absolute prediction errors

Are we happy with the expected squared prediction errors criterion? What happens if we replace the L_2 loss function with the $L_1 : \mathbb{E}\left[|\mathbf{y} - \mathbf{X}|\right]$? Its estimates are more robust than those for the conditional mean. L_1 criteria have discontinuities in their derivatives, which have hindered their widespread use.
We consider the expect absolute prediction error $\mathrm{EPAE}(r) = \mathbb{E}\left[|Y - r(\mathbf{X})|\right]$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 4. Expected absolute prediction errors:

a) Prove that we can write EPAE as $\mathrm{EPAE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[|Y - r(\mathbf{X})||\mathbf{X}\right]\right]$. The same solution as Question 3.

b) What is the (pointwise w.r.t. $\mathbf{x}$) minimizer solution of $\operatorname{argmin}_r \mathrm{EPAE}(r)$? The integration is a little sophisticated and requires a branch of analysis known as measure theory. For this reason, we will provide an approximation that does not address the smaller details. By using the law of large numbers, we get the following

$$\int_y |Y - r| \Pr(y|\mathbf{X}) dy = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} |Y_n - r| \approx \frac{1}{N} \sum_{n=1}^{N} |Y_n - r|. \tag{6}$$

By taking the derivative of conditional expectation w.r.t. r and set to zero, we look for r such that $\sum_{n=1}^{N} \operatorname{sign}(Y_n - r) = 0$.

Solution 4. Expected absolute prediction errors

Are we happy with the expected squared prediction errors criterion? What happens if we replace the L_2 loss function with the $L_1 : \mathbb{E}[|\mathbf{y} - \mathbf{X}|]$? Its estimates are more robust than those for the conditional mean. L_1 criteria have discontinuities in their derivatives, which have hindered their widespread use.

We consider the expect absolute prediction error $\mathrm{EPAE}(r) = \mathbb{E}[|Y - r(\mathbf{X})|]$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 4. Expected absolute prediction errors:

a) Prove that we can write EPAE as $\mathrm{EPAE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}[|Y - r(\mathbf{X})||\mathbf{X}]\right]$. The same solution as Question 3.

b) What is the (pointwise w.r.t. $\mathbf{x}$) minimizer solution of $\mathrm{argmin}_r\, \mathrm{EPAE}(r)$? The integration is a little sophisticated and requires a branch of analysis known as measure theory. For this reason, we will provide an approximation that does not address the smaller details. By using the law of large numbers, we get the following

$$\int_y |Y - r| \Pr(y|\mathbf{X}) dy = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} |Y_n - r| \approx \frac{1}{N} \sum_{n=1}^{N} |Y_n - r|. \tag{6}$$

By taking the derivative of conditional expectation w.r.t. r and set to zero, we look for r such that $\sum_{n=1}^{N} \mathrm{sign}(Y_n - r) = 0$. At what value of r does the above quantity hold?

Solution 4. Expected absolute prediction errors

Are we happy with the expected squared prediction errors criterion? What happens if we replace the L_2 loss function with the $L_1 : \mathbb{E}[|\mathbf{y} - \mathbf{X}|]$? Its estimates are more robust than those for the conditional mean. L_1 criteria have discontinuities in their derivatives, which have hindered their widespread use.

We consider the expect absolute prediction error $\mathrm{EPAE}(r) = \mathbb{E}[|Y - r(\mathbf{X})|]$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 4. Expected absolute prediction errors:

a) Prove that we can write EPAE as $\mathrm{EPAE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}[|Y - r(\mathbf{X})||\mathbf{X}]\right]$. The same solution as Question 3.

b) What is the (pointwise w.r.t. $\mathbf{x}$) minimizer solution of $\mathrm{argmin}_r\, \mathrm{EPAE}(r)$? The integration is a little sophisticated and requires a branch of analysis known as measure theory. For this reason, we will provide an approximation that does not address the smaller details. By using the law of large numbers, we get the following

$$\int_y |Y - r| \Pr(y|\mathbf{X}) dy = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} |Y_n - r| \approx \frac{1}{N} \sum_{n=1}^{N} |Y_n - r|. \tag{6}$$

By taking the derivative of conditional expectation w.r.t. r and set to zero, we look for r such that $\sum_{n=1}^{N} \mathrm{sign}(Y_n - r) = 0$. At what value of r does the above quantity hold? It holds when there is an equal number of positive and negative values; that is, where $\mathrm{card}(Y_n > r) = \mathrm{card}(Y_n < r)$.

Solution 4. Expected absolute prediction errors

Are we happy with the expected squared prediction errors criterion? What happens if we replace the L_2 loss function with the $L_1 : \mathbb{E}\left[|\mathbf{y} - \mathbf{X}|\right]$? Its estimates are more robust than those for the conditional mean. L_1 criteria have discontinuities in their derivatives, which have hindered their widespread use.

We consider the expect absolute prediction error $\mathrm{EPAE}(r) = \mathbb{E}\left[|Y - r(\mathbf{X})|\right]$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 4. Expected absolute prediction errors:

a) Prove that we can write EPAE as $\mathrm{EPAE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[|Y - r(\mathbf{X})||\mathbf{X}\right]\right]$. The same solution as Question 3.

b) What is the (pointwise w.r.t. $\mathbf{x}$) minimizer solution of $\operatorname{argmin}_r \mathrm{EPAE}(r)$? The integration is a little sophisticated and requires a branch of analysis known as measure theory. For this reason, we will provide an approximation that does not address the smaller details. By using the law of large numbers, we get the following

$$\int_y |Y - r| \Pr(y|\mathbf{X}) dy = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} |Y_n - r| \approx \frac{1}{N} \sum_{n=1}^{N} |Y_n - r|. \tag{6}$$

By taking the derivative of conditional expectation w.r.t. r and set to zero, we look for r such that $\sum_{n=1}^{N} \operatorname{sign}(Y_n - r) = 0$. At what value of r does the above quantity hold? It holds when there is an equal number of positive and negative values; that is, where $\operatorname{card}(Y_n > r) = \operatorname{card}(Y_n < r)$. The value of r where that is true is the median.

Solution 4. Expected absolute prediction errors

Are we happy with the expected squared prediction errors criterion? What happens if we replace the L_2 loss function with the $L_1 : \mathbb{E}[|\mathbf{y} - \mathbf{X}|]$? Its estimates are more robust than those for the conditional mean. L_1 criteria have discontinuities in their derivatives, which have hindered their widespread use.
We consider the expect absolute prediction error $\mathrm{EPAE}(r) = \mathbb{E}[|Y - r(\mathbf{X})|]$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 4. Expected absolute prediction errors:

a) Prove that we can write EPAE as $\mathrm{EPAE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}[|Y - r(\mathbf{X})||\mathbf{X}]\right]$. The same solution as Question 3.

b) What is the (pointwise w.r.t. $\mathbf{x}$) minimizer solution of $\mathrm{argmin}_r\, \mathrm{EPAE}(r)$? The integration is a little sophisticated and requires a branch of analysis known as measure theory. For this reason, we will provide an approximation that does not address the smaller details. By using the law of large numbers, we get the following

$$\int_y |Y - r| \Pr(y|\mathbf{X}) dy = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} |Y_n - r| \approx \frac{1}{N} \sum_{n=1}^{N} |Y_n - r|. \tag{6}$$

By taking the derivative of conditional expectation w.r.t. r and set to zero, we look for r such that $\sum_{n=1}^{N} \mathrm{sign}(Y_n - r) = 0$. At what value of r does the above quantity hold? It holds when there is an equal number of positive and negative values; that is, where $\mathrm{card}(Y_n > r) = \mathrm{card}(Y_n < r)$. The value of r where that is true is the median. Recall that the median can be found by sorting a finite list of numbers from lowest value to highest value and picking the middle one.

Solution 4. Expected absolute prediction errors

Are we happy with the expected squared prediction errors criterion? What happens if we replace the L_2 loss function with the $L_1 : \mathbb{E}\left[|\mathbf{y} - \mathbf{X}|\right]$? Its estimates are more robust than those for the conditional mean. L_1 criteria have discontinuities in their derivatives, which have hindered their widespread use.

We consider the expect absolute prediction error $\text{EPAE}(r) = \mathbb{E}\left[|Y - r(\mathbf{X})|\right]$. We seek a function $r(\mathbf{X})$ for predicting Y given values of the input $\mathbf{X}$.

? 4. Expected absolute prediction errors:

a) Prove that we can write EPAE as $\text{EPAE}(r) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[|Y - r(\mathbf{X})||\mathbf{X}\right]\right]$. The same solution as Question 3.

b) What is the (pointwise w.r.t. $\mathbf{x}$) minimizer solution of $\text{argmin}_r \text{EPAE}(r)$? The integration is a little sophisticated and requires a branch of analysis known as measure theory. For this reason, we will provide an approximation that does not address the smaller details. By using the law of large numbers, we get the following

$$\int_y |Y - r| \Pr(y|\mathbf{X}) dy = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} |Y_n - r| \approx \frac{1}{N} \sum_{n=1}^{N} |Y_n - r|. \tag{6}$$

By taking the derivative of conditional expectation w.r.t. r and set to zero, we look for r such that $\sum_{n=1}^{N} \text{sign}(Y_n - r) = 0$. At what value of r does the above quantity hold? It holds when there is an equal number of positive and negative values; that is, where $\text{card}(Y_n > r) = \text{card}(Y_n < r)$. The value of r where that is true is the median. Recall that the median can be found by sorting a finite list of numbers from lowest value to highest value and picking the middle one. In conclusion, we have shown that $\widehat{r}(\mathbf{x}) = \text{median}(Y|\mathbf{X} = xy)$.

Outline

5. Reduced weighted least squares problem

We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \ldots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$. Then, we consider a parametrized model $r_\beta(\mathbf{x})$ to be fit by least squares.

Show that if there are observations with tied or identical values of $\mathbf{x}$, then the fit can be obtained from a reduced weighted least squares problem.

We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \ldots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$. Then, we consider a parametrized model $r_\beta(\mathbf{x})$ to be fit by least squares.

Show that if there are observations with tied or identical values of $\mathbf{x}$, then the fit can be obtained from a reduced weighted least squares problem.

$\boxed{\text{Motivation?}}$

5. Reduced weighted least squares problem

We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \ldots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$. Then, we consider a parametrized model $r_\beta(\mathbf{x})$ to be fit by least squares.

Show that if there are observations with tied or identical values of $\mathbf{x}$, then the fit can be obtained from a reduced weighted least squares problem.

$\boxed{\text{Motivation?}}$ The motivation for this discussion is that often experimentally one would like to get an accurate estimate of the error ϵ in the model $y = r(\mathbf{X}) + \epsilon$.

5. Reduced weighted least squares problem

We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \ldots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$. Then, we consider a parametrized model $r_\beta(\mathbf{x})$ to be fit by least squares.

Show that if there are observations with tied or identical values of $\mathbf{x}$, then the fit can be obtained from a reduced weighted least squares problem.

$\boxed{\text{Motivation?}}$ The motivation for this discussion is that often experimentally one would like to get an accurate estimate of the error ϵ in the model $y = r(\mathbf{X}) + \epsilon$. One way to do this is to perform many experiments, observing the different values of y produced by the data generating process when the same value of $\mathbf{x}$ is produced for each experiment.

5. Reduced weighted least squares problem

We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \ldots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$. Then, we consider a parametrized model $r_\beta(\mathbf{x})$ to be fit by least squares.

Show that if there are observations with tied or identical values of $\mathbf{x}$, then the fit can be obtained from a reduced weighted least squares problem.

$\boxed{\text{Motivation?}}$ The motivation for this discussion is that often experimentally one would like to get an accurate estimate of the error ϵ in the model $y = r(\mathbf{X}) + \epsilon$. One way to do this is to perform many experiments, observing the different values of y produced by the data generating process when the same value of $\mathbf{x}$ is produced for each experiment.

If $\epsilon = 0$?

5. Reduced weighted least squares problem

We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \ldots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$. Then, we consider a parametrized model $r_\beta(\mathbf{x})$ to be fit by least squares.

Show that if there are observations with tied or identical values of $\mathbf{x}$, then the fit can be obtained from a reduced weighted least squares problem.

$\boxed{\text{Motivation?}}$ The motivation for this discussion is that often experimentally one would like to get an accurate estimate of the error ϵ in the model $y = r(\mathbf{X}) + \epsilon$. One way to do this is to perform many experiments, observing the different values of y produced by the data generating process when the same value of $\mathbf{x}$ is produced for each experiment.

If $\epsilon = 0$? we would expect the results of each of these experiments to be the same.

5. Reduced weighted least squares problem

We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \ldots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$. Then, we consider a parametrized model $r_\beta(\mathbf{x})$ to be fit by least squares.

Show that if there are observations with tied or identical values of $\mathbf{x}$, then the fit can be obtained from a reduced weighted least squares problem.

Motivation? The motivation for this discussion is that often experimentally one would like to get an accurate estimate of the error ϵ in the model $y = r(\mathbf{X}) + \epsilon$. One way to do this is to perform many experiments, observing the different values of y produced by the data generating process when the same value of $\mathbf{x}$ is produced for each experiment.

If $\epsilon = 0$? we would expect the results of each of these experiments to be the same.

This problem is known a reduced weighted least square?

5. Reduced weighted least squares problem

We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \ldots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$. Then, we consider a parametrized model $r_\beta(\mathbf{x})$ to be fit by least squares.

Show that if there are observations with tied or identical values of $\mathbf{x}$, then the fit can be obtained from a reduced weighted least squares problem.

$\boxed{\text{Motivation?}}$ The motivation for this discussion is that often experimentally one would like to get an accurate estimate of the error ϵ in the model $y = r(\mathbf{X}) + \epsilon$. One way to do this is to perform many experiments, observing the different values of y produced by the data generating process when the same value of $\mathbf{x}$ is produced for each experiment.

If $\epsilon = 0$? we would expect the results of each of these experiments to be the same.

$\boxed{\text{This problem is known a reduced weighted least square?}}$ Each residual error is weighted by how many times the measurement of $\mathbf{x}_n$ was taken.

5. Reduced weighted least squares problem

We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \ldots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$. Then, we consider a parametrized model $r_\beta(\mathbf{x})$ to be fit by least squares.

Show that if there are observations with tied or identical values of $\mathbf{x}$, then the fit can be obtained from a reduced weighted least squares problem.

Motivation? The motivation for this discussion is that often experimentally one would like to get an accurate estimate of the error ϵ in the model $y = r(\mathbf{X}) + \epsilon$. One way to do this is to perform many experiments, observing the different values of y produced by the data generating process when the same value of $\mathbf{x}$ is produced for each experiment.

If $\epsilon = 0$? we would expect the results of each of these experiments to be the same.

This problem is known a reduced weighted least square? Each residual error is weighted by how many times the measurement of $\mathbf{x}_n$ was taken. It is a reduced problem? since the number of points we are working, see next slide that $N_u < N$. Here, N_u be the number of unique inputs $\mathbf{x}$, that is, the number of distinct inputs after discarding duplicates.

Outline

Solution 5. Reduced weighted least squares problem

Modeling step: Let N_u be the number of unique inputs $\mathbf{x}$, that is, the number of distinct inputs after discarding duplicates. Assume that if the nth unique x value gives rise to N_n potentially different $y_{nm}, m \in [N_n]$ values. Then, we define $\overline{y}_n = \frac{1}{N_n} \sum_{m=1}^{N_n} y_{nm}$, the average of all responses y resulting from the same input $\mathbf{x}_n$.

Solution 5. Reduced weighted least squares problem

Modeling step: Let N_u be the number of unique inputs $\mathbf{x}$, that is, the number of distinct inputs after discarding duplicates. Assume that if the nth unique x value gives rise to N_n potentially different $y_{nm}, m \in [N_n]$ values. Then, we define $\overline{y}_n = \frac{1}{N_n} \sum_{m=1}^{N_n} y_{nm}$, the average of all responses y resulting from the same input $\mathbf{x}_n$. We wish to prove that we can minimize the equivalent optimization problem:

$$\mathrm{RSS}(\beta) = \sum_{n=1}^{N_u} \frac{1}{N_n} \left(\overline{y}_n - r_\beta(\mathbf{x}_n)\right)^2. \tag{7}$$

Solution 5. Reduced weighted least squares problem

Modeling step: Let N_u be the number of unique inputs $\mathbf{x}$, that is, the number of distinct inputs after discarding duplicates. Assume that if the nth unique x value gives rise to N_n potentially different $y_{nm}, m \in [N_n]$ values. Then, we define $\overline{y}_n = \frac{1}{N_n} \sum_{m=1}^{N_n} y_{nm}$, the average of all responses y resulting from the same input $\mathbf{x}_n$. We wish to prove that we can minimize the equivalent optimization problem:

$$\mathrm{RSS}(\beta) = \sum_{n=1}^{N_u} \frac{1}{N_n} \left(\overline{y}_n - r_\beta(\mathbf{x}_n)\right)^2. \tag{7}$$

Indeed, with the previous notation we can write the $\mathrm{RSS}(\beta)$ above as

$$N\,\mathrm{RSS}(\beta) = \sum_{n=1}^{N_u} \sum_{m=1}^{N_n} \left(y_{nm} - r_\beta(\mathbf{x}_n)\right)^2. \tag{8}$$

Solution 5. Reduced weighted least squares problem

Modeling step: Let N_u be the number of unique inputs $\mathbf{x}$, that is, the number of distinct inputs after discarding duplicates. Assume that if the nth unique x value gives rise to N_n potentially different $y_{nm}, m \in [N_n]$ values. Then, we define $\overline{y}_n = \frac{1}{N_n} \sum_{m=1}^{N_n} y_{nm}$, the average of all responses y resulting from the same input $\mathbf{x}_n$. We wish to prove that we can minimize the equivalent optimization problem:

$$\mathrm{RSS}(\beta) = \sum_{n=1}^{N_u} \frac{1}{N_n} \left(\overline{y}_n - r_\beta(\mathbf{x}_n)\right)^2. \tag{7}$$

Indeed, with the previous notation we can write the RSS(β) above as

$$N\,\mathrm{RSS}(\beta) = \sum_{n=1}^{N_u} \sum_{m=1}^{N_n} \left(y_{nm} - r_\beta(\mathbf{x}_n)\right)^2. \tag{8}$$

Expanding the quadratic in the above expression we have

$$N\,\mathrm{RSS}(\beta) = \sum_{n=1}^{N_u} \sum_{m=1}^{N_n} \left(y_{nm}^2 - 2y_{nm} r_\beta(\mathbf{x}_n) + r_\beta^2(\mathbf{x}_n)\right). \tag{9}$$

Solution 5. Reduced weighted least squares problem

Modeling step: Let N_u be the number of unique inputs $\mathbf{x}$, that is, the number of distinct inputs after discarding duplicates. Assume that if the nth unique x value gives rise to N_n potentially different $y_{nm}, m \in [N_n]$ values. Then, we define $\overline{y}_n = \frac{1}{N_n} \sum_{m=1}^{N_n} y_{nm}$, the average of all responses y resulting from the same input $\mathbf{x}_n$. We wish to prove that we can minimize the equivalent optimization problem:

$$\mathrm{RSS}(\beta) = \sum_{n=1}^{N_u} \frac{1}{N_n} \left(\overline{y}_n - r_\beta(\mathbf{x}_n) \right)^2 . \tag{7}$$

Indeed, with the previous notation we can write the $\mathrm{RSS}(\beta)$ above as

$$N\,\mathrm{RSS}(\beta) = \sum_{n=1}^{N_u} \sum_{m=1}^{N_n} \left(y_{nm} - r_\beta(\mathbf{x}_n) \right)^2 . \tag{8}$$

Expanding the quadratic in the above expression we have

$$N\,\mathrm{RSS}(\beta) = \sum_{n=1}^{N_u} \sum_{m=1}^{N_n} \left(y_{nm}^2 - 2 y_{nm} r_\beta(\mathbf{x}_n) + r_\beta^2(\mathbf{x}_n) \right) . \tag{9}$$

Using definition of $\overline{y}_n$ and completing the square, we have

$$N\,\mathrm{RSS}(\beta) = \sum^{N_u} \frac{1}{N_n} \left(\overline{y}_n - r_\beta(\mathbf{x}_n) \right)^2 + \sum^{N_u} \sum^{N_n} y_{nm}^2 - \sum^{N_u} N_n \overline{y}_n^2 . \tag{10}$$

Outline

Perspectives

1. $\boxed{\text{Week 10 (18/04/2023):}}$ **Some Exercises (TD)** for the bonus grade and **send the Final Evaluation CC** (Deadline 02/05/2023).

Perspectives

1. $\boxed{\text{Week 10 (18/04/2023):}}$ **Some Exercises (TD)** for the bonus grade and **send the Final Evaluation CC** (Deadline 02/05/2023).
    - TD on Model Assessment.
    - TD on Model Selection (AIC, BIC, Cross-validation).
2. $\boxed{\text{Week 11 (25/04/2023):}}$ **Last CC with questions**.