Multiple linear Regression

Notation:

- y_i dependent variable / outcome
- x_{ip} p^{th} (linear) dependent variable / predictor / covariate
- β_p vector of unknown parameters
- ε_i stochastic error term
- i denotes individual observation
- P # of parameters

Assumption 1: Linearity

$$y_i = \sum_{p=1}^{P} \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \dots, n$$

If we include a constant, $x_{i1} = 1$ for all i.

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times p)}{\mathbf{X}} \underset{(p\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & & \vdots \\ \vdots & & \\ x_{n1} & & x_{np} \end{pmatrix}$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Quadratic case : only have one predictor $\mathbf{x}_1$, then :

$$X = \begin{pmatrix} x_{11} & x_{11}^2 \\ \vdots & \vdots \\ x_{n1} & x_{n1}^2 \end{pmatrix}$$

Least Squares :

$$SSR(\beta^*) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \beta^*)^2$$

In matrix notation:

$$SSR(\beta^*) = (y - X\beta^*)'(y - X\beta^*)$$

$$\beta := \underset{\beta^* \in \mathbb{R}^p}{\operatorname{argmin}} \; S^{SR}(\beta^*)$$

$$SSR(\beta^*) = y'y - (X\beta^*)'y - y'X\beta^* + \beta^{*\prime} X'X\beta^*$$
$$= y'y - 2y'X\beta^* + \beta^{*\prime} X'X\beta^*$$

$$\Rightarrow \frac{d}{d\beta^*} SSR(\beta^*) = -2X'y + 2X'X\beta^*$$

$$X'X\hat{\beta} = X'y$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

To solve for $\hat{\beta}$: we need to assume that $X'X$ has full rank.

Assumption 2: $\operatorname{rank}(X'X) = p$

Assumption 3: $E(\varepsilon_i \mid X) = 0$

Weak exogeneity: $E(\varepsilon_i) = 0$

This is needed for unbiasedness:

$$E(\hat{\beta}) = \beta \quad , \quad E(\hat{\beta} - \beta) = 0$$

Difference $E(\varepsilon_i | X_i) = 0$ and $E(\varepsilon_i) = 0$

Example: $X_i = 1$ if i is male, 0 if i is female.

$$E(\varepsilon_i | X_i = 1) = 1 \quad , \quad E(\varepsilon_i | X_i = 0) = -1$$

$\Rightarrow$ $E(\varepsilon_i) = 0$ ✓

$E(\varepsilon_i | X) \neq 0$

Assumption: Homoskedasticity

$$E(\varepsilon_i^2 | X) = \sigma^2 > 0$$

Some quantities of interest: $\hat{y}_i = x_i \hat{\beta} = \hat{f}(x_i)$

$\Rightarrow$ OLS fitted values

$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$

$\Rightarrow$ OLS Residuals

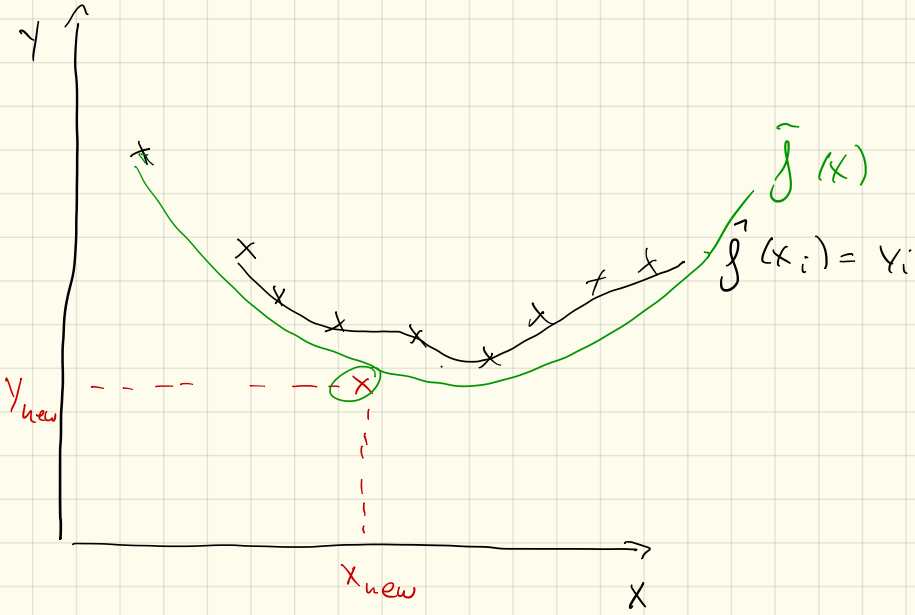
Assessing quality of $\hat{f}$:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}(x_i))^2 \qquad (1)$$

$\hat{f}(x_i)$ is the prediction / fitted value that $\hat{f}$ gives for the i'th observation.

Question: which $\hat{f}(x_i)$ would minimize (1)?

$\hat{f}(x_i) = Y_i \Rightarrow MSE = 0$



What do we want:

1) Assess the structural relationship between Y and X

2) predict y_{new} for a given x_{new}.

=> Related goals, but not the same and not nested.

However: For both MSE is a poor measure of success.

To achieve 2: We want to choose a method that gives the lowest MSE based on a new training observation (x_0, y_0). = test

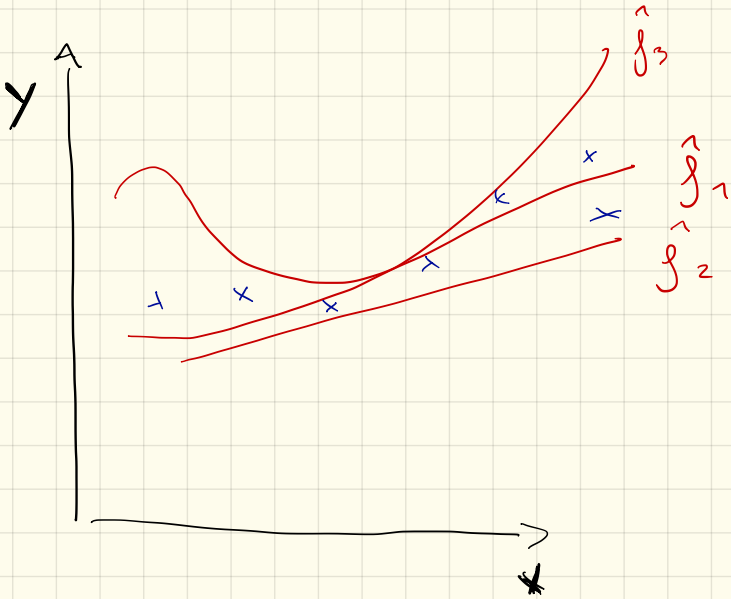
Minimize: $\text{Ave}\left(\hat{f}(x_0) - y_0\right)^2$

where $\hat{f}$ was derived from a training set.

Big question: where to get test observations.

Best: get new data

2. Best: Resampling methods such as Cross-validation.



$\hat{f}_1$: lowest MSE test errors.