Linear regression: ordinary least squares

Regression can be defined as a statistical technique for investigating and modeling statistical relation between the *mean value* of a dependent variable(s) $\mathbf{y}$ (***response, target***) and corresponding values of independent variable(s) $\mathbf{x}$ – ***predictors***, ***covariates***:

$$\boxed{E\{\mathbf{y}|\mathbf{x}\} = f(x)} \tag{1}$$

The regression equation *is only an approximation* to the true functional relationship between the variables of interest. Thus, ***regression does not mean causation***.

Generally, ***regression equations are valid only over the region of the regressor variables contained*** in the observed data.

Why classical LS linear regression:

- Central topic in machine learning – can be generalized to other popular ML methods – classification problems and logistic regression, neural networks;
- Has a closed-form expression for solution;
- Easy to introduce basic concepts like ***bias-variance tradeoff***, ***cross-validation***, ***resampling***, ***regularization*** techniques and many other ML topics.

1. Multiple linear regression formulation

Assume the following form of functional relation between the variables:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2}$$

With $\varepsilon_i \sim N(0,\ \sigma^2)$ it follows that:

- $E\{y_i|x\} = \boldsymbol{x}_i^T \boldsymbol{\beta} \Rightarrow$ ***regression model is a line (or hyperplane in general)***;
- $\mathrm{Var}\{y|x\} = \sigma^2_{y|x} = \mathrm{Var}\left(\beta_0 + \beta_1 x + \varepsilon\right) = \sigma^2$.

If to concatenate in columns all the observations i, we can re-formulate (2) in the matrix form:

$$\boxed{\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}} \tag{3}$$

where $\mathbf{X}$ is the ***design matrix*** n – observations $\times$ p – features:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{12} & \ldots & x_{1p} \\ 1 & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \ldots & x_{np} \end{pmatrix} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p) \tag{4}$$

Often $\boldsymbol{x}_1 = \mathbf{1}$ so that β_0 is a ***free term***.

$\boldsymbol{\beta}$ is the $p \times 1$ vector of model parameters;
$\mathbf{y}$ is the $1 \times n$ vector of target observations.

In classical linear regression model apart from relation (3) also imposed the following assumptions on stochastic structure of the model – ***5 basic assumptions of classical model***:

- Model is sufficient i.e. $E\{y|x\}$ is ***linear***, ***no grouping***. (*Solution:* transform data, ln or x/y^{-1});
- $E\{\varepsilon_i\} = 0$ + random scatter is symmetric (*Solution:* data transformation);
- $E\{\varepsilon_i^2\} = \sigma^2$ – const error variance;
- $E\{\varepsilon_i \varepsilon_j\} = 0, \quad \forall\ i \neq j$ – uncorrelated errors;
- $rk(\mathbf{X}) = p < n$ (rank is the maximum number of linearly independent column vectors);

Also:

- Linear regression is ***sensitive to outliers*** (*Solution:* transform data or remove them);
- $\varepsilon \sim N(0,\ \sigma^2)$ – explicit assumption on the error form ***for hypothesis testing and interval estimation***.

2. Least squares estimation for multiple linear regression

Under the described above assumptions, the main method of model parameters ($\boldsymbol{\beta}$) estimation is the ***Ordinary Least Squares (OLS)***:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \arg\min_{\beta} \sum_{i=1}^{n} \left(\mathbf{y_i} - \mathbf{x_i}^T \beta\right)^2 \tag{5}$$

Which has a general solution in matrix form as:

$$\boxed{\hat{\boldsymbol{\beta}}_{\text{OLS}} = \underbrace{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T}_{\text{MP pseudoinverse}}\mathbf{y}} \tag{6}$$

As a result of this minimization problem, we get ***fitted values*** $\hat{y}_i = x_i^T\hat{\beta}$ and ***residuals*** $e_i = y_i - \hat{y}_i$.

2.1 LSE derivation and solution

We wish to find a vector of OLS model parameters that minimizes:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{7}$$

Open the brackets:

$$\begin{aligned} S(\boldsymbol{\beta}) &= \mathbf{y}^T\mathbf{y} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \end{aligned} \tag{8}$$

$\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}$ is a scalar which transpose $(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y})^T = \boldsymbol{\beta}\mathbf{X}\mathbf{y}^T$ is the same scalar!

The optimal values of parameters $\hat{\boldsymbol{\beta}}$ should minimize $S(\boldsymbol{\beta})$, thus:

$$\left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = (\text{simple lin. alg. from above}) = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0} \tag{9}$$

Simplify it to the ***LS normal equation***:

$$\boxed{\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}} \tag{10}$$

Normal equation in details:

$$\begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ip} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i_1}x_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^n x_{ip} & \sum_{i=1}^n x_{ip}x_{i1} & \sum_{i=1}^n x_{ip}x_{i2} & \cdots & \sum_{i=1}^n x_{ip}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \vdots \\ \sum_{i=1}^n x_{ip}y_i \end{bmatrix} \tag{11}$$

Solution of the normal equation is (6), provided that $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ exist which is ***always true if regressors are linearly independent*** – no column of $\mathbf{X}$ is a linear combination of other columns.

2.2 Hat (projection) matrix H

The ***fitted values*** from the LSE are then:

$$\boxed{\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}} \tag{12}$$

where

$$\mathbf{H}_{n\times n} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \tag{13}$$

Is $n \times n$ squared ***hat matrix***. ***It maps the vector of observed target values onto a vector of fitted values.***

2.3 Residuals e

The difference between the observed and corresponding fitted values is the residuals: $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$.

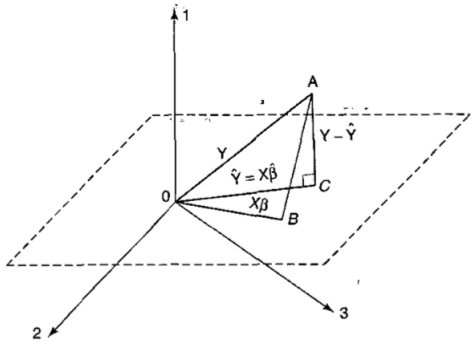
We can express them using (12) also as:

$$\boxed{\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{y}} \tag{14}$$

Residuals play an important role in in investigating ***model adequacy*** and in detecting departures from the underlying assumptions.

2.4 Geometrical interpretation of the least squares

In the following figure the ***estimation space*** is $p = 2$ which is a subspace of a 3D ***sample space*** $n = 3$.



We may think of the vector of observations $\mathbf{y} = [y_1, y_2, ...y_n]$ as defining the vector in the sample space from the origin to the point A.

The $\mathbf{X}$ matrix consists of p $(n \times 1)$ vectors defined from the origin in the sample space. These vectors forms a p-dimensional estimation space. In the figure we have only two vectors ($p = 2$). We may represent any point in the estimation space as a linear combination of vectors $\mathbf{1}, \mathbf{x_1}, ..., \mathbf{x_p}$ so that any point on the estimation space is in the form of $\mathbf{X}\boldsymbol{\beta}$ determing the point B.

What we do in OLS is minimization of the distance between OB and OA. The solution in the estimation space satisfying this is the vector OC ($\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$) which is the projection of the vector $\mathbf{y}$ onto the sample space $\mathbf{1}, \mathbf{x_1}, ..., \mathbf{x_p}$.

We see that the vector of residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is orthogonal to the sample space $\mathbf{X}$. Therefore, the following is true:

$$\boxed{\mathbf{X}^T\mathbf{e} = \mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0}, \tag{15}$$

which we recognize as the LS ***normal equation***:

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}. \tag{16}$$

Normal equation in details:

$$\begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ip} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^n x_{ip} & \sum_{i=1}^n x_{ip}x_{i1} & \sum_{i=1}^n x_{ip}x_{i2} & \cdots & \sum_{i=1}^n x_{ip}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \vdots \\ \sum_{i=1}^n x_{ip}y_i \end{bmatrix} \tag{17}$$

Solution of the normal equation is (6), provided that $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ exist which is ***always true if regressors are linearly independent*** – no column of $\mathbf{X}$ is a linear combination of other columns.

3. Simple linear regression model

In general, OLS optimal model parameters:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{18}$$

Specify this solution for the simple linear regression of a singe predictor with a single free term:

$$y_i = x_i\beta_1 + \beta_0 + \varepsilon_i, \quad i = 1, \ldots, n. \tag{19}$$

3.1 Derivation of LS estimators $\widehat{\beta}_0$, $\widehat{\beta}_1$

As a first step, introduce normalizing factors $1/n$ to the both factors of (18) for the convenience:

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}^T\mathbf{y}\right) \tag{20}$$

Then let's look at two factors separately:

$$\frac{1}{n}\mathbf{X}^T\mathbf{X} = \frac{1}{n}\begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix} = \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{bmatrix} \tag{21}$$

Now, we need to take the inverse:

$$\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1} = \frac{1}{\overline{x^2} - \bar{x}^2}\begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} = \frac{1}{\sigma_x^2}\begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \tag{22}$$

Another factor:

$$\frac{1}{n}\mathbf{X}^T\mathbf{y} = \frac{1}{n}\begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \frac{1}{n}\begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \overline{xy} \end{bmatrix} \tag{23}$$

Now, multiply the two factors:

$$\begin{aligned} \frac{1}{s_X^2}\begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}\begin{bmatrix} \bar{y} \\ \overline{xy} \end{bmatrix} &= \frac{1}{\sigma_x^2}\begin{bmatrix} \overline{x^2}\bar{y} - \overline{xxy} \\ -\overline{xy} + \overline{xy} \end{bmatrix} = \frac{1}{\sigma_x^2}\begin{bmatrix} \left(s_X^2 + \bar{x}^2\right)\bar{y} - \bar{x}\left(c_{XY} + \bar{x}\bar{y}\right) \\ c_{XY} \end{bmatrix} \\ &= \frac{1}{\sigma_x^2}\begin{bmatrix} \sigma_x^2\bar{y} + \bar{x}^2\bar{y} - \bar{x}c_{XY} - \bar{x}^2\bar{y} \\ c_{XY} \end{bmatrix} = \begin{bmatrix} \bar{y} - \frac{C_{XY}}{\sigma_x^2}\bar{x} \\ \frac{C_{XY}}{\sigma_x^2} \end{bmatrix} \end{aligned} \tag{24}$$

Finally, the model parameters of the simple regression ($i = 1, ..., n$):

$$\boxed{\hat{\beta}_1 = \frac{C_{xy}}{SS_x} = \frac{\cancel{\frac{1}{n-1}}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\cancel{\frac{1}{n-1}}\sum_{i=1}^n (x_i - \bar{x})^2}} \quad - \textbf{\textit{slope}} \tag{25}$$

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}} \quad - \textbf{\textit{intercept}} \tag{26}$$

Conclusion: regression is a linear relationship summary (that's what the correlation measures).

Gauss-Markov theorem. The OLS estimator of $\boldsymbol{\beta}$ $\hat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ ***where errors are uncorrelated***, ***have equal variances and zero expectations*** is ***BLUE*** i.e. $\hat{\beta}$ has the smallest variance among the class of all unbiased estimators that are the linear combination of data. Errors do not need to be normally distributed nor iid.

3.4 Estimation of σ^2

Ideally, we want to have a prior information about σ^2. If this is not possible, the ***Residual Sum of Squares*** is used

$$\boxed{RSS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2} \tag{27}$$

The residual sum of squares has $n-2$ degrees of freedom, because ***2 degrees of freedom are already associated with estimates of*** $\hat{\beta}_0$ ***and*** $\hat{\beta}_1$ ***involved in obtaining*** $\hat{y}$.

It can be shown that the expected value of RSS is $E(RSS) = (n-p)\sigma^2$ (the degrees of freedom comes from the rank of $\mathbf{I}-\mathbf{H}$ which is trace $tr(\mathbf{I}-\mathbf{H}) = n-p$). Thus, an unbiased estimator of σ^2 is:

$$\boxed{\hat{\sigma}^2 = \frac{RSS}{n-p} = MS_{\text{Res}}} \quad - \textbf{\textit{Residual Mean Square}}. \tag{28}$$

The quantity $\sqrt{\hat{\sigma}^2}$ is sometimes called ***Standard Error of Regression***.

Since $\hat{\sigma}^2$ depend on the RSS, any violation of the assumptions on the model errors or any misspecification of the model form may seriously damage the usefulness of its estimate given in (36) $\Rightarrow$ we say that it is ***a model dependent estimate*** of σ^2.

4. Coefficient of determination, R^2

Let $\mathbf{SS_R} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ and $\mathbf{SS_T} = \sum_{i=1}^{n} (y_i - \bar{y})^2$. Then the quantity

$$\boxed{R^2 = \frac{SS_{\text{R}}}{SS_{\text{T}}} = 1 - \frac{RSS}{SS_{\text{T}}}} \tag{29}$$

is the ***coefficient of determination*** i.e. ***proportion of variation explained by the regressor*** x. Since $0 \leq \text{RSS} \leq \text{SS}_{\text{T}}$, it follows that:

$$\boxed{0 \leq R^2 \leq 1} \tag{30}$$

Values of R^2 ***close to*** 1 ***imply that most of the variability in*** y ***is explained by the regression model*** e.g. if $R^2 = 0.9018$ that is, 90.18% of the variability in strength is accounted for by the regression model. R^2 ***should be used with caution***, since it is always possible to make R^2 large by adding enough terms to the model i.e. it is ***sensitive to overfitting***.

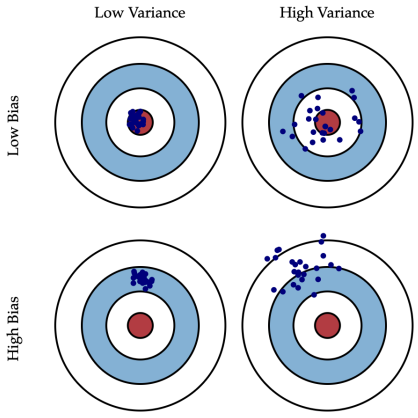
R^2 ***does not measure the appropriateness of the linear model***, for R^2 will often be large even though y and x are *nonlinearly related*. Remember that ***large*** R^2 ***does not necessarily imply that the regression model will be an accurate predictor***.

5. Bias-variance trade-off

Conceptual definition

- **Error due to bias:** Due to the randomness of the dataset, the underlying model will make predictions covering a certain range. Difference of the average outcome of our model and the correct values we are aiming to predict is a model bias – how far are the model predictions in general.
- **Error due to variance:** For a given data we want to predict, how wide is the spread of the model predictions w.r.t. these data.

Graphical definition



Mathematical definition

Say that exist some true *deterministic* function that we want to approximate with our ML model

$$f = f(\mathbf{x})$$

Although, we cannot observe outcomes of this unknown function directly but with some additive stochastic component, say being centered $\mathrm{E}\{\varepsilon\} = 0$:

$$t = f + \varepsilon$$

Consider a dataset:

$$\mathbf{D} = \{(t_i, \mathbf{x}_i)\,,\ i = 1 \dots N\}$$

Having this dataset $\mathbf{D}$, we train a parametric model g (e.g. neural network or linear OLS regression) to approximate the function of interest f based on the 'noised' observations t:

$$y_i = g(\mathbf{x}, \mathbf{w})$$

Assume that the measure of quality of our model is MSE:

$$\mathrm{MSE} = \mathrm{E}\left\{(t_i - y_i)^2\right\}$$

Do some math over this expectation:

$$E\left\{(t_i - y_i)^2\right\} = /\text{add and substract hypothetical true values then associate}/ =$$
$$= E\left\{((t_i - f_i) + (f_i - y_i))^2\right\} = E\left\{(t_i - f_i)^2\right\} + E\left\{(f_i - y_i)^2\right\} + 2E\left\{(f_i - y_i)\,(t_i - f_i)\right\}$$
$$= E\left\{\varepsilon^2\right\} + E\left\{(f_i - y_i)^2\right\} + 2\left(E\left\{f_i t_i\right\} - E\left\{f_i^2\right\} - E\left\{y_i t_i\right\} + E\left\{y_i f_i\right\}\right) = E\left\{\varepsilon^2\right\} + E\left\{(f_i - y_i)^2\right\}$$

The last term goes to zero because:

- $E\left\{f_i t_i\right\} = f_i^2 \Leftarrow E\left\{f_i(f_i + \varepsilon)\right\} = E\{f_i^2 + f_i\varepsilon\} = f_i$ is determenistic $= E\{f_i^2\} + 0 = f_i^2$;
- $E\left\{y_i t_i\right\} = E\left\{y_i\,(f_i + \varepsilon)\right\} = E\left\{y_i f_i + y_i\varepsilon\right\} = E\left\{y_i f_i\right\} + E\left\{y_i\right\} E\left\{\varepsilon\right\} = E\left\{y_i f_i\right\} + 0$.

So now we have

$$E\left\{(t_i - y_i)^2\right\} = \sigma_\varepsilon^2 + E\left\{(f_i - y_i)^2\right\}$$

where we can apply the same trick as before to the second term but adding and subtracting the mean of the model outcomes:

$$E\left\{(f_i - y_i)^2\right\} = E\left\{(f_i - \bar{y}_i + \bar{y}_i - y_i)^2\right\} =$$
$$E\left\{(f_i - \bar{y}_i)^2\right\} + E\left\{(\bar{y}_i - y_i)^2\right\} + 2E\left\{(\bar{y}_i - y_i)\,(f_i - \bar{y}_i)\right\} =$$
$$E\left\{(f_i - \bar{y}_i)^2\right\} + E\left\{(\bar{y}_i - y_i)^2\right\} + 2\left(E\left\{f_i\bar{y}_i\right\} - E\left\{\bar{y}_i{}^2\right\} - E\left\{y_i f_i\right\} + E\left\{y_i\bar{y}_i\right\}\right)$$

The last term also cancels to zero because:

- $E\left\{f_i\bar{y}_i\right\} = f_i$ is determenistic $= f_i E\left\{\bar{y}_i\right\} = f_i\bar{y}_i$;
- $E\left\{\bar{y}_i{}^2\right\} =$ mean is determenistic $= \bar{y}_i{}^2$;
- $E\left\{y_i f_i\right\} = f_i\bar{y}_i$
- $E\left\{y_i\bar{y}_i\right\} = \bar{y}_i{}^2$;

Then, if to sum up we have the following:

$$\boxed{\mathrm{MSE} = E\left\{(t_i - y_i)^2\right\} = E\left\{(f_i - \bar{y}_i)^2\right\} + E\left\{(\bar{y}_i - y_i)^2\right\} + \sigma_\varepsilon^2 = \mathrm{Bias} + \mathrm{Var}_y + \sigma_\varepsilon^2}$$

Thus to minimize the MSE of the model, we have to minimize both the bias of the model as well its variance although it is not trivial!

The following is commonly true about the components of the bias-variance equation:

- **A high bias model** has a relatively large error, mostly due to *wrong assumptions about the data features*;
- **A high variance model** is *quite sensitive to small variations in the train data*;
- **Irreducible error** is due to the intrinsic *noisiness of the data itself.*

6. Gradient descent optimization methods

TBA