Life Sciences and
Facility Management

Institute of
Computational Life Sciences

Linear regression
Part II

Machine Learning

Norman Juchler

Quiz SW05: Modelling aspect

Which of the following statements about regression models are true?

In the equations below, ε models the noise in the data. A common assumption is that $\epsilon \sim \mathcal{N}(0, \sigma)$.

True	False	
○	○	The following model can be solved using linear regression. $y(x_1, x_2) = \beta_0 + \beta_1 \sin(x_1) + \beta_2 \cos(x_2) + \epsilon$
○	○	The following equation represents a linear model $y(x_1, x_2) = \beta_1 x_1 + \beta_2 x_1 x_2 + \beta_3 x_2 + \epsilon$
○	○	It was noted above that noise is commonly modeled as $\epsilon \sim \mathcal{N}(0, \sigma)$. This is equivalent to saying that the residuals $r = y - \hat{y}$ (i.e. the difference between true and predicted values) are normally distributed with a mean of zero and a constant variance.
○	○	The following equation represents a quadratic regression model. It can be computed in Python with the help of scikit-learn and its transformer `PolynomialFeatures`. $y(x_1, x_2) = \beta_{10} x_1 + \beta_{20} x_1^2 + \beta_{11} x_1 x_2 + \beta_{02} x_2^2 + \beta_{00} + \epsilon$

Quiz SW05: Modelling aspect

Which of the following statements about regression models are true?

In the equations below, ε models the noise in the data. A common assumption is that $\epsilon \sim \mathcal{N}(0, \sigma)$.

True	False	
○	○	The following model can be solved using linear regression. $y(x_1, x_2) = \beta_0 + \beta_1 \sin(x_1) + \beta_2 \cos(x_2) + \epsilon$
○	○	The following equation represents a linear model $y(x_1, x_2) = \beta_1 x_1 + \beta_2 x_1 x_2 + \beta_3 x_2 + \epsilon$
○	○	It was noted above that noise is commonly modeled as $\epsilon \sim \mathcal{N}(0, \sigma)$. This is equivalent to saying that the residuals $r = y - \hat{y}$ (i.e. the difference between true and predicted values) are normally distributed with a mean of zero and a constant variance.
○	○	The following equation represents a quadratic regression model. It can be computed in Python with the help of scikit-learn and its transformer `PolynomialFeatures`. $y(x_1, x_2) = \beta_{10} x_1 + \beta_{20} x_1^2 + \beta_{11} x_1 x_2 + \beta_{02} x_2^2 + \beta_{00} + \epsilon$

Recapitulation

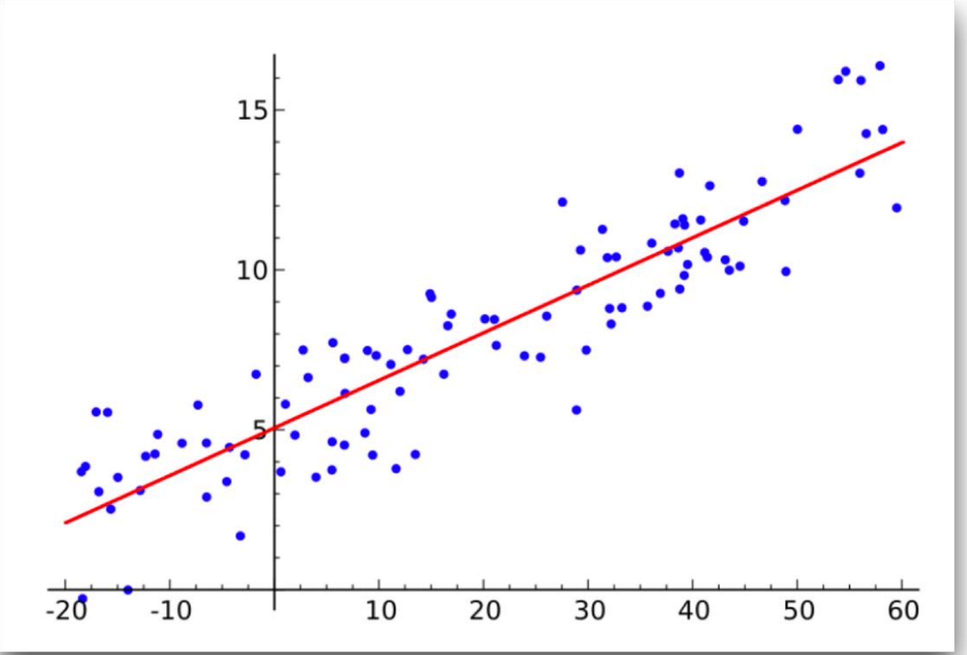
Linear model

- Assume we have a set of p features: $\boldsymbol{x}_i := (x_{i1}, x_{i2}, \ldots, x_{ip})$
- We want to use them to predict a target variable y
- The simple assumption we can make is (model ansatz):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

$$f_i(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}\boldsymbol{x}$$

- Where the β_i are unknown parameters that we want to determine from the data

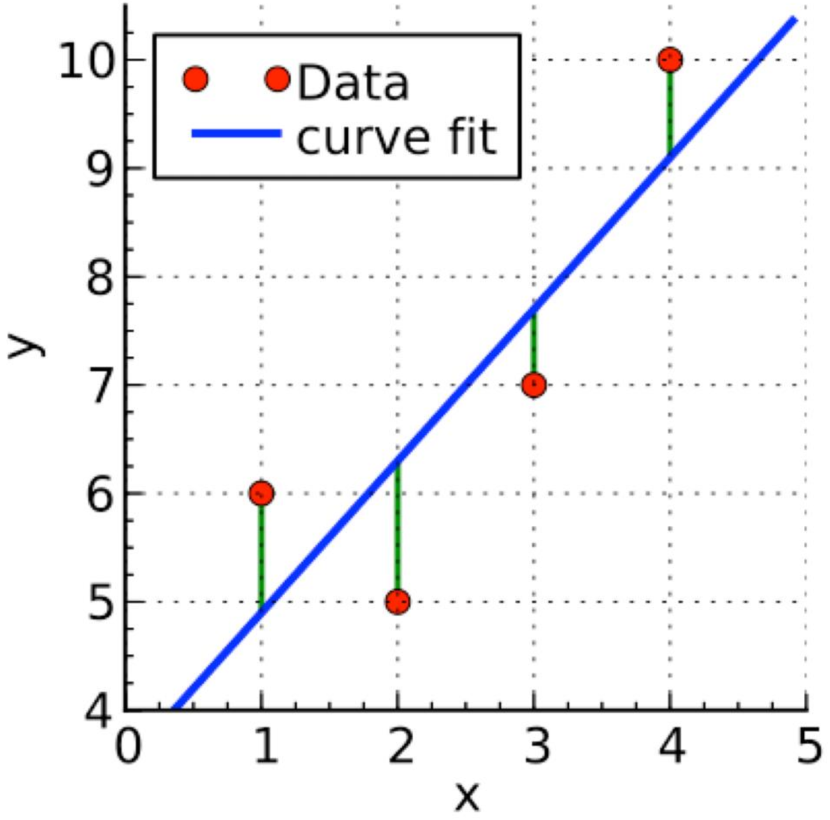


Ordinary least squares (OLS) regression

- Optimization objective for linear regression:

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - f(x_i, \boldsymbol{\beta}))^2 = \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta})$$

- **In words**: Find parameters **β** such that the sum of squared residuals (green lines) is minimized.
- Using **calculus**, we can derive an analytical solution.
- For a general solution of the OLS regression problem, as well as proof, see here:
  - Definition: Simple linear regression
  - Proof: Ordinary least squares / optimal parameters β



Loss functions

- A loss function (a.k.a. cost function) in the context of machine learning usually measures how well the predicted values match the true target values.

$$\mathcal{L}(y - \hat{y})$$

- For regression, we used the residual sum of squares (RSS) as loss function.

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

- Sometimes, it is more meaningful to compute the **mean squared error** (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n} RSS$$

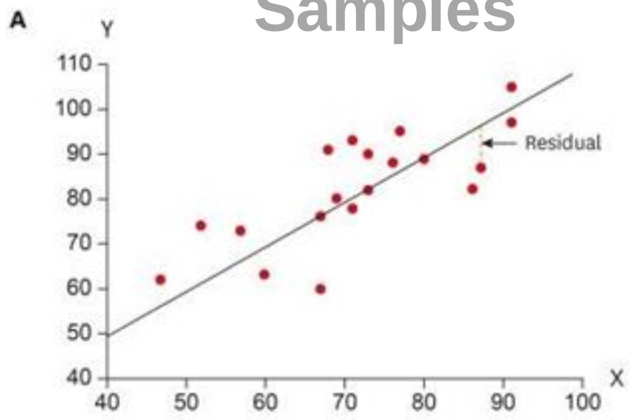
- (Note: both yield the same optimal solution!)

Residuals

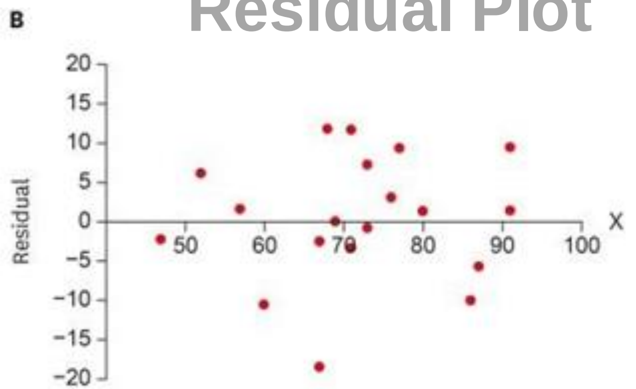
- The **residual** for a sample m is given by $\epsilon_m = y_m - \hat{y}_m$
- It represents the modelling error

- The **residual plot** is a scatter plot of input or predicted values vs. residuals.

Samples



Residual Plot

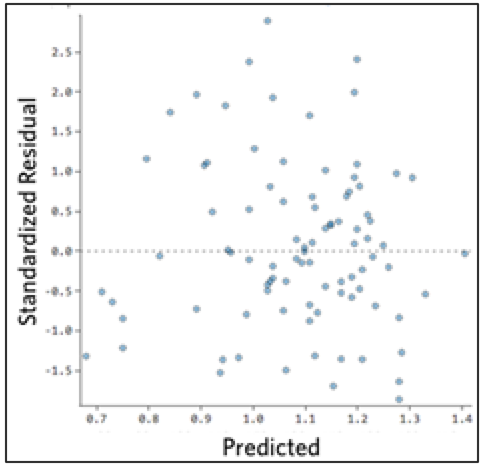


- Residual plots are useful for assessing and “debugging” linear regression

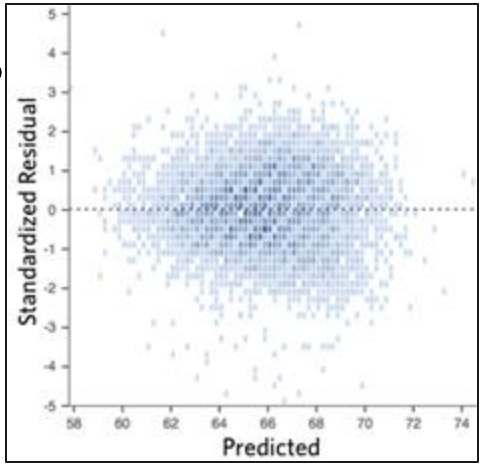
Residuals

Which of the following residual plots show an appropriate setting for linear regression?

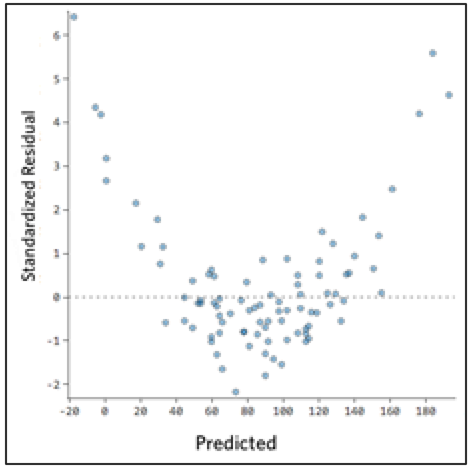
A



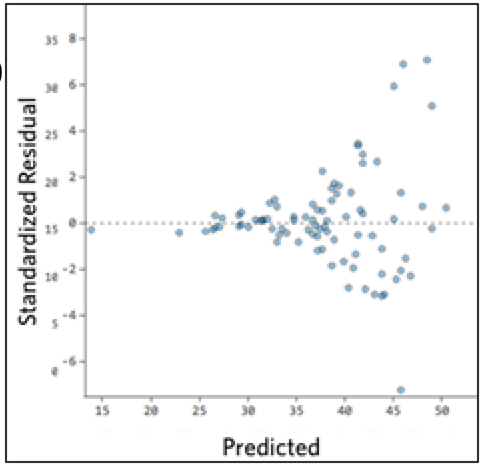
B



C

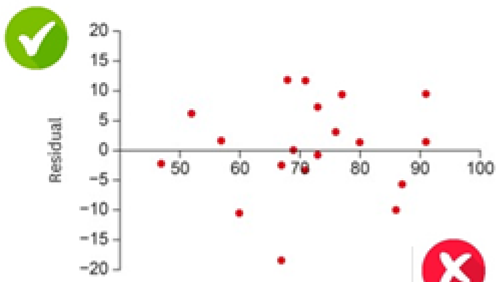


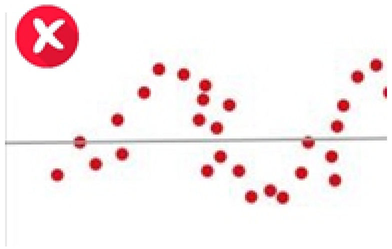
D

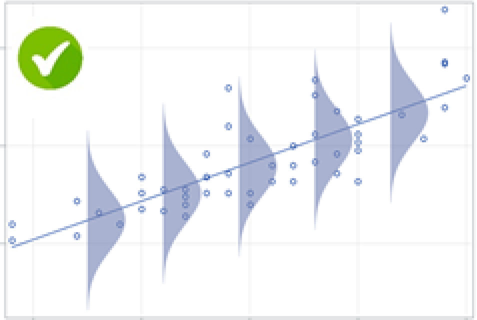


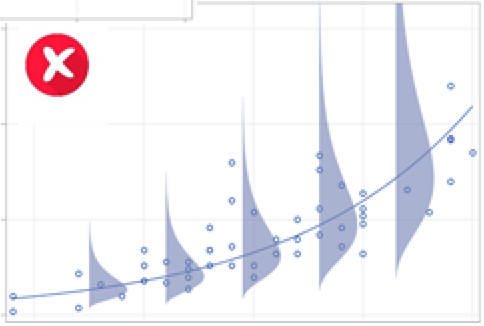
Assumptions of linear regression

- **Linearity**: The input and output values have a linear relationship
- **Independence**: The outcome of one sample does not affect the others
- **Normality**: Errors should be normally distributed, i.e. larger deviations from mean should be less likely
- **Equality of Variance ("Homoscedasticity")**: Error distribution should be the same for all input values









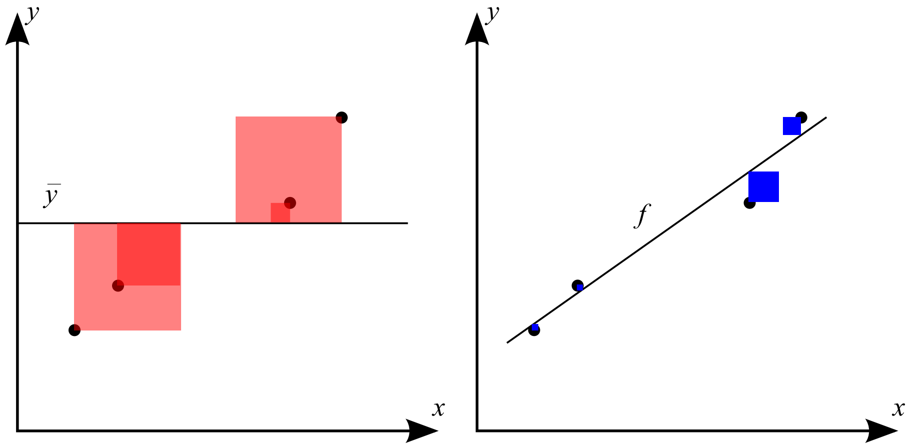
Coefficient of determination / R squared

- **Goodness of fit**:
  R^2 is a statistical measure that indicates how well the predicted values from a regression model fit the observed data.
- **Explained variance**:
  It represents the proportion of the variance in the target variable that is explained by the features in the model.
- **Value range**:
  R^2 values range from 0 to 1, where 1 means the model perfectly predicts the target variable, and 0 means it does not explain any of the variability.
- **Model Comparison**:
  It is often used to compare the fit of different models, but it does not necessarily indicate whether a model is correct or good, especially for non-linear models.

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

$$SS_{\text{res}} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$$

$$SS_{\text{tot}} = \sum_i (y_i - \mu_y)^2$$



Multivariate linear regression

- We just include multiple variables ☺

$$\hat{y}^{(m)} = h_\theta(x^{(m)}) = \theta_0 x_0^{(m)} + \theta_1 x_1^{(m)} + \theta_2 x_2^{(m)} + \ldots + \theta_N x_N^{(m)} = \boldsymbol{\theta}^T \boldsymbol{X}, \qquad \text{Eq. (1)}$$

$x_0^{(m)} := 1$ for all $m = 1, \ldots, M$

$$\epsilon^{(m)} = y^{(m)} - \hat{y}^{(m)} \qquad \text{residuals}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \qquad \text{M equations of the form of Eq. (1)}$$

$$\boldsymbol{X} : M \times (N+1), \boldsymbol{\theta} : (N+1) \times 1, \boldsymbol{y} : M \times 1 \qquad \text{dimensions}$$

Multivariate regression

Regularization / Lasso