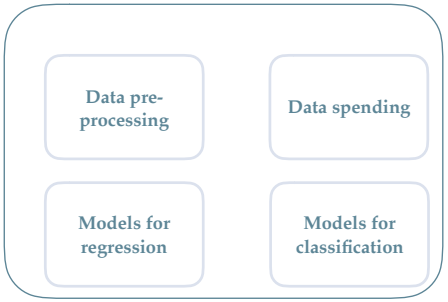
Linear models for regression

Michela Mulas

A quick recap

During the last lectures, we did...

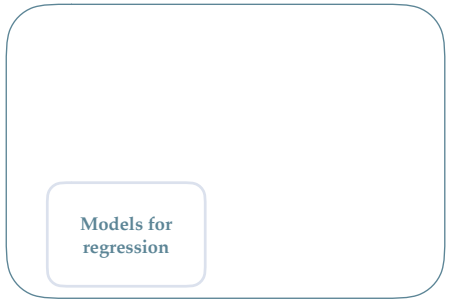
- Focus on **predictive modeling**.



Today's goal

Today, we going to ...

- Focus on **linear model regression**.



Today's goal

Today, we going to ...

- Define the terms.
  - Bias-variance trade-off.
- Define linear regression models.
- Introduce ordinary least square models.
- Examples in R.

Reference

Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*, Springer (2014)

Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Ed., Springer (2017)[1]

[1] The book is available for download at:
https://web.stanford.edu/~hastie/ElemStatLearn/

Defining the terms

We have an **outcome measurement, y**

- Quantitative.
- Qualitative (also called categorical or factors or discrete variables).

For both types of outputs, we want to **use the inputs to predict the output**.

- Given some specific atmospheric measurements today and yesterday, we want to predict the ozone level tomorrow $\rightsquigarrow$ **regression**
- Given the grayscale values for the pixels of the digitized image of the handwritten digit, we want to predict its class label $\rightsquigarrow$ **classification**.

Defining the terms

Focus on regression...

- We have a training set of data.
- We observe the outcome and feature measurements for a set of objects.
- Using this data we build a prediction model, or learner.
- This model will enable us to predict the outcome for new unseen objects.
- A good learner is one that accurately predicts such an outcome.

Defining the terms
Supervised learning

This is the case of a **supervised learning problem**.

- It is called “supervised” because of the presence of the outcome variable to guide the learning process.
- In the **unsupervised learning problem**, we observe only the features and have no measurements of the outcome. Our task is rather to describe how the data are organized or clustered.

Focus on supervised learning...

Defining the terms
Supervised learning

We use a model that in a general form can be written as

$$y = f(X) + \varepsilon \quad \text{where } \varepsilon \text{ is the error}$$

Supervised learning attempts to **learn** *f* **by example** through a *teacher*.

- This process is known as learning by example.
- Upon completion of the learning process the hope is that the artificial and real outputs will be close enough to be useful for all sets of inputs likely to be encountered in practice.

Defining the terms

Measuring performance in regression models

- For models predicting a numeric outcome, some **measure of accuracy** is typically used to evaluate the effectiveness of the model.
- There are different ways to measure accuracy, each with its own nuance.
- To understand the strengths and weaknesses of a particular model, relying solely on a single metric is problematic.
- Visualizations of the model fit, particularly residual plots, are critical to understanding whether the model is fit for purpose.

Defining the terms

Measuring performance in regression models

Qualitative measure of performance: RMSE

- When the output is a number, root mean squared error is the most common method for characterizing a model's predictive capabilities.
- The mean squared error (MSE) is calculated by squaring the residuals, summing them and dividing by the number of samples.
- The value is usually interpreted as either how far (on average) the residuals are from zero or as the average distance between the observed values and the model predictions.

Defining the terms

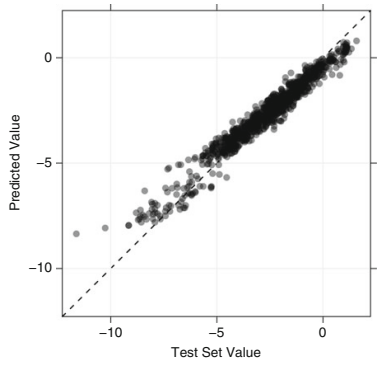
Measuring performance in regression models

Qualitative measure of performance: R^2

- Another common metric is the **coefficient of determination**.
- This value can be interpreted as the proportion of the information in the data that is explained by the model.
- Thus, an R^2 value of 0.75 implies that the model can explain three-quarters of the variation in the outcome.
- There are multiple formulas for calculating this quantity, although the simplest version finds the correlation coefficient between the observed and predicted values (usually denoted by R) and squares it.
- **Note and remember** that R^2 is a measure of correlation, not accuracy.

Defining the terms

Measuring performance in regression models



In this example, R^2 between the observed and predicted values is high (51%) but predictions are not uniformly accurate.

R^2 is dependent on the variation in the outcome.

Defining the terms

Measuring performance in regression models

Formally, the mean squared error of a model is:

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

- y_i is the outcome.
- $\hat{y}_i$ is the model prediction of that sample's outcome.

Assuming the data points that are statistically independent and that the residuals have a theoretical mean of zero and constant variance of σ^2, the **expected value of MSE** is:

$$\text{E[MSE]} = \sigma^2 + \Big(\text{Model bias}\Big)^2 + \text{Model variance}$$

1 The first part (σ^2) is usually called **irreducible noise**.
It cannot be eliminated by modeling.

Defining the terms

Measuring performance in regression models

Formally, the mean squared error of a model is:

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

- y_i is the outcome.
- $\hat{y}_i$ is the model prediction of that sample's outcome.

Assuming the data points that are statistically independent and that the residuals have a theoretical mean of zero and constant variance of σ^2, the **expected value of MSE** is:

$$\text{E[MSE]} = \sigma^2 + \Big(\text{Model bias}\Big)^2 + \text{Model variance}$$

2 The second term is the **squared bias** of the model.
It reflects how close the functional form of the model can get to the true relationship between the predictors and the outcome.

Defining the terms

Measuring performance in regression models

Formally, the mean squared error of a model is:

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

- y_i is the outcome.
- $\hat{y}_i$ is the model prediction of that sample's outcome.

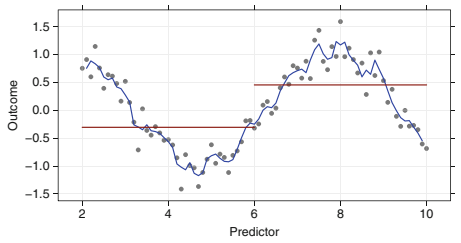
Assuming the data points that are statistically independent and that the residuals have a theoretical mean of zero and constant variance of σ^2, the **expected value of MSE** is:

$$\text{E[MSE]} = \sigma^2 + \Big(\text{Model bias}\Big)^2 + \text{Model variance}$$

3 The last term is the model variance.
It reflects how much the functional form of the model "varies" around the mean.

Defining the terms

Measuring performance in regression models

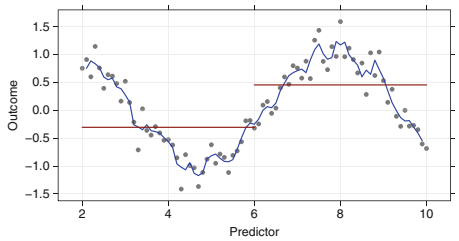


The models in **red** and **blue** are extreme examples either high bias or high variance.

- **Red model** splits the data in half and predicts each half with a simple average.
- This model has low variance since it would not substantially change if another set of data points were generated the same way.
- However, it is ineffective at modeling the data since, due to its simplicity and for this reason, it has high bias.

Defining the terms

Measuring performance in regression models

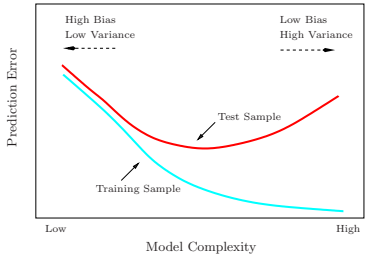


The models in **red** and **blue** are extreme examples either high bias or high variance.

- **Blue model** is a three-point moving average.
- It is flexible enough to model the sin wave (i.e., low bias), but small perturbations in the data will significantly change the model fit. Because of this, it has high variance.

Defining the terms

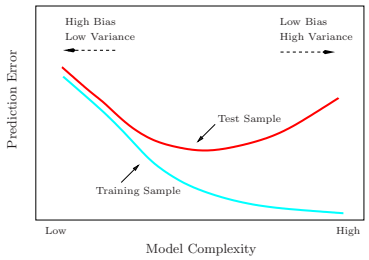
Measuring performance in regression models



- The training error tends to decrease whenever we increase the model complexity, that is, whenever we fit the data harder.
- However with too much fitting, the model adapts itself too closely to the training data, and will not generalize well (i.e., have large test error).

Defining the terms

Measuring performance in regression models



- In contrast, if the model is not complex enough, it will underfit and may have large bias, again resulting in poor generalization.

Defining the terms

Measuring performance in regression models

It is generally true that more complex models can have very high variance, which leads to over-fitting.

On the other hand, simple models tend not to over-fit, but under-fit if they are not flexible enough to model the true relationship (thus high bias).

Also, highly correlated predictors can lead to collinearity issues and this can greatly increase the model variance.

This is referred to as the **variance-bias trade-off**.

Linear methods for regression

Linear regression models can all be directly or indirectly be written in the form:

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_P x_{iP} + \varepsilon_i$$

- y_i epresents the numeric response for the ith sample.
- β_0 represents the estimated intercept.
- β_j represents the estimated coefficient for the jth predictor.
- x_{ij} represents the value of the jth predictor for the ith sample.
- ε_i represents random error that cannot be explained by the model.

In addition to ordinary linear regression, these types of models include **partial least squares** (PLS) and penalized models such as **ridge regression**, the **lasso**, and the **elastic net**.

Linear methods for regression

Linear models were largely developed in the precomputer age of statistics, but even in today's computer era there are still good reasons to study and use them.

- They are simple and often provide an adequate and interpretable description of how the inputs affect the output.
- For prediction purposes they can sometimes outperform fancier nonlinear models, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data.
- Finally, linear methods can be applied to transformations of the inputs and this considerably expands their scope.

Linear models and least squares

We have a vector of inputs $X^T = \left(X_1, X_2, \ldots, X_P\right)$.

We want to predict a real-valued output Y. The linear regression model has the form:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

- β_0 is the intercept.
- β_j's are unknown parameters or coefficients.

The variables X_j can come from different sources:

- Quantitative inputs.
- Transformation of quantitative inputs (log, square-roots or squares).
- Basis expansions (such as, $X_2 = X_1^2, X_3 = X_1^3$), leading to polynomial representation.
- Numeric or "dummy" coding of the levels of quantitative inputs.
- Interactions between variables, for example, $X_3 = X_1 \cdot x_2$.

Linear models and least squares

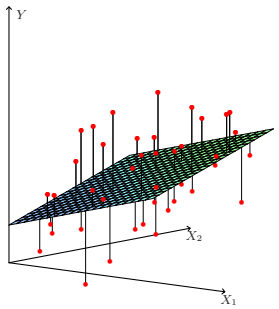
No matter the source of the X_j, the model is linear in the parameters.

- Typically we have a set of training data $(x_1, y_1) \ldots (x_N, y_N)$ from which to estimate the parameters β.
- Each $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ is a vector of feature measurements for the ith.

We use the **least squares method** to pick the coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ to **minimize the residual sum of squares**

$$RSS(\beta) = \sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2 =$$
$$= \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$$

Linear models and least squares



- In the $(p+1)$-dimensional input-output space, $(X, \hat{Y})$ represents a hyperplane.
- $RSS(\beta)$ makes no assumptions about the validity of the model, it simply finds the best linear fit to the data.
- Least squares fitting is intuitively satisfying no matter how the data arise; the criterion measures the average lack of fit.

Linear models and least squares

How do we minimize the residual sum of squares?

- $\mathbf{X}$ is the $N \times (p+1)$ matrix with each row an input vector (with 1 in the 1st position).
- $\mathbf{y}$ is the N-vector of outputs in the training set.

Then, the residual sum of squares can be written as:

$$RSS(\beta) = \left(\mathbf{y} - \mathbf{X}\beta\right)^T \left(\mathbf{y} - \mathbf{X}\beta\right) \quad \text{a quadratic function in the } p+1 \text{ parameters}$$

Differentiating with respect to β we obtain:

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\beta\right)$$
$$\frac{\partial^2 RSS}{\partial \beta \partial \beta^2} = -2\mathbf{X}^T\mathbf{X}$$

Linear models and least squares

Assuming (for the moment) that $\mathbf{X}$ has full column rank, and hence $\mathbf{X}^T\mathbf{X}$ is positive definite, we set the first derivative to zero

$$\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\beta\right) = 0$$

to obtain the unique solution:

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

The predicted values at an input vector x_0 are given by $(\hat{f})(x_0) = (1 : x_0)^T\hat{\beta}$.

The fitted values at the training inputs are:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \underbrace{\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T}_{\text{“hat” matrix}}\mathbf{y}$$

The parameter estimates that minimize the sum of squares are the ones that have the least bias of all possible parameter estimates. Hence, these estimates **minimize the bias component of the bias-variance trade-off**.