Statistics and Data Analysis in MATLAB
Kendrick Kay, kendrick.kay@wustl.edu
February 21, 2014

Lecture 3: Model specification

1. Correlation is a simple case of model building

- When computing the correlation between two variables x and y, we are implicitly fitting the linear model $y = ax + b$ where a and b are free parameters. We can say that x is an *input variable*, y is the *output variable*, and the model attempts to predict the output variable based on the input variable. Alternatively, we can say that x is a *regressor*, y is the *data*, and the model attempts to explain the data using the regressor.
- The link between correlation and linear models can be understood as follows: after z-scoring each variable, the slope of the line that best predicts y from x is equal to the correlation value r.
- Thus, correlation is a simple case of model building in which we use a linear model to predict one variable based on another. But what if we have more than one input variable? Or what if the phenomenon we are trying to model is nonlinear? To tackle these cases, we must obtain a more explicit understanding of model building.

2. Overview of model building

- To keep things simple, it is useful to break down model building into four distinct issues.
- *Model specification* refers to choosing the specific type of model to apply to the data. For example, do we use a simple linear model or a complicated nonlinear model? (In the case of correlation, a linear model is implicitly being applied.)
- *Model fitting* refers to estimating the free parameters of a given model based on the observed data. (In the case of correlation, the parameter of interest, r, is computed through a simple sequence of mathematical operations.)
- *Model accuracy* refers to quantifying how well a fitted model describes the data. The tricky issue here is the potential for a model to overfit the data. (In the case of correlation, the analogue of model accuracy is r^2 as it indicates the amount of variance in one variable that can be explained by the other.)
- *Model reliability* refers to quantifying the reliability of the parameters of a fitted model. In other words, how confident are we with regards to the parameters we have estimated from the data? (In the case of correlation, the analogue of model reliability is the sampling error on r.)
- In this lecture, we consider the issue of model specification.

3. Supervised vs. unsupervised learning

- Statistical models can be broadly divided into models that can be used to perform *supervised learning* and models that can be used to perform *unsupervised learning*. The idea behind learning is that by applying a model to data, we learn something about those data.
- In supervised learning (e.g. linear regression), we are trying to learn the mapping between one or more input variables and an output variable. The problem is supervised in the sense that we observe both the input and the output and the goal is clearly defined.
- In unsupervised learning (e.g. PCA, cluster analysis), we are trying to learn the structure in a given set of data. The problem is unsupervised in the sense that we observe only a single set of data and there is no clearly defined, explicit goal.

- For those fluent in probability theory, supervised learning can be viewed as characterizing $p(y \mid x)$, that is, the probability of output *y* given input *x*, whereas unsupervised learning can be viewed as characterizing $p(x)$, that is, the probability distribution underlying a set of inputs *x*.
- We will be focusing on supervised learning in this class.

4. Regression vs. classification
- Both regression and classification involve using one or more input variables to predict an output variable. In terms of the problem specification, the only difference between regression and classification is the nature of the output variable. If the output variable is continuous, the problem is known as *regression*; if the output variable is discrete, the problem is known as *classification*.
- There are, nevertheless, complex technical details that arise when getting into the specifics of classification models; thus, regression and classification are quite different in the details.
- We will be focusing on regression models as we learn about model building, and we will address classification models later in the course. The rest of this lecture describes different types of regression models.

5. Linear models
- *Linear models* are models in which the prediction is a weighted sum of the regressors. For example, suppose we have regressors x_i where *i* ranges from 1 to *n*. The prediction of a linear model is $y = \sum_{i=1}^{n} w_i x_i$ where w_i is the weight on the *i*th regressor.
- To see why $y = ax + b$ counts as a linear model, notice that we can re-write the model as $y = ax + b\mathbf{1}$ where $\mathbf{1}$ is a constant regressor consisting of all ones. Then, if we set $x_1 = x$ and $x_2 = \mathbf{1}$, the model can be re-written as $y = w_1x_1 + w_2x_2$ where w_1 and w_2 are free parameters.
- Linear models are easy to understand and easy to fit. However, not all phenomena are linear!

6. Nonlinear models
- *Nonlinear models* are models in which the prediction cannot be expressed as a weighted sum of the regressors. We can divide nonlinear models into three types.

7. Linearized models
- One type of nonlinear model is a *linearized model*. Linearized models are the same as linear models except that the input space has been expanded using nonlinear functions (e.g. polynomials, Gaussians, sinusoids).
- For example, suppose we start with the linear model $y = ax + b$ where *a* and *b* are free parameters. If we expand the input space to include a new regressor x^2, we obtain the model $y = ax^2 + bx + c$ where *a*, *b*, and *c* are free parameters. This new model is nonlinear because the prediction of the model is not linear with respect to *x*. However, the new model can be viewed as a linear model with respect to an input space consisting of three regressors, $x_1 = x^2$, $x_2 = x$, and $x_3 = \mathbf{1}$.
- Linearized models are as easy to understand and fit as linear models, and has the additional benefit of being able to characterize nonlinear phenomena. However, for any given problem, it is not clear *a priori* exactly what type of nonlinear functions to add into the input space.

8. Parametric nonlinear models

- Another type of nonlinear model is a *parametric nonlinear model*. Parametric nonlinear models can be thought of as any model that can be written down using mathematical operations but which cannot be expressed as a linearized model.
- The key characteristic of linearized models is that they are linear with respect to the free parameters in the models. Parametric nonlinear models do not have this feature.
- For example, consider the model $y = x^n$ where n is a free parameter. The output of this model is not linear with respect to n (there is no way to express x^n as $an + b$ where a and b are weights). Thus, the model does not count as a linearized model.
- Parametric nonlinear models are easy to understand and can characterize nonlinear phenomena. However, they are tricky to fit given that they are nonlinear with respect to the free parameters (e.g. risk of local minima).

9. Nonparametric nonlinear models

- A third type of nonlinear model is a *nonparametric nonlinear model*. Nonparametric nonlinear models are essentially very flexible nonlinear models that can be generically applied to arbitrary datasets.
- The distinction between nonparametric nonlinear models and the other types of nonlinear models can be hazy, but we can make some generalities: (1) Nonparametric nonlinear models are sometimes nonlinear with respect to the free parameters; thus, we cannot categorize them as linearized models. (2) Nonparametric nonlinear models make few assumptions on the form of the nonlinearity of the model, unlike parametric nonlinear models.
- Examples of nonparametric nonlinear models include nearest-neighbor methods (predict the output based on the nearest input example), local regression (fit a simple model at each point in the input space based on nearby data points), and neural networks (use generic basis functions such as sigmoidal or radial basis functions to model the output).
- Nonparametric nonlinear models are flexible and powerful and, in some cases, can also be simple and elegant. However, the fitted models may be difficult to interpret and may suffer from overfitting and local minima. Also, the computational complexity of nonparametric nonlinear models often grows with the size of the dataset and may therefore pose practical problems.

10. Summary of model types

- Let us review the differences between the various types of models. One difference ("Linear?") is whether a given model is linear with respect to the original regressors. A second difference ("Parametric?") is whether a given model makes strong assumptions on the form of the relationship between input and output. A third difference ("Linear in parameters?") is whether a given model is linear with respect to its free parameters.

	Linear?	*Parametric?*	*Linear in parameters?*
Linear models	yes	yes	yes
Linearized models	no	yes	yes
Parametric nonlinear models	no	yes	no
Nonparametric nonlinear models	no	no	sometimes

11. Matrix representation of linear models

- All linear (and linearized) models can be expressed as a weighted sum of one or more fixed regressors. We can formally represent this using matrix notation:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{n}$$

where **y** is a set of data points ($n \times 1$), **X** is a set of regressors ($n \times p$), **w** is a set of weights ($p \times 1$), and **n** is a set of residuals ($n \times 1$). Note that compared to previous formulations of linear models, we are now explicitly including a term for the residuals. The residuals are simply the difference between the data (**y**) and the model fit (**Xw**).