CISC 271 Class 10

Patterns – Linear Regression

Text Correspondence: §4.3

Main Concepts:

- *Independent and dependent variables*
- *Statistical regression*
- *Residual error*
- *Linear regression*
- *Regression of standardized data*

Sample Problem, Machine Inference: How can we find a linear pattern in data?

In empirical studies, it is common to have a theoretical model of a physical or other process but for various reasons the data do not follow the model exactly. This common problem has many names, depending on the field of study in which it arises, and also has many differing solutions that depend on the assumptions that are made.

We will think of this problem as pattern recognition from sparse data. By sparse data we mean that we are given values, such as pairs of numbers; we will not examine "dense" data, such as all of the pixels in a digital image. By pattern recognition, we mean that we have a specific pattern that we expect to recognize. In this course, the pattern will be a relation among the data values. This will be a function and our problem will be to find parameters, or coefficients, of the function that provide a "good" match to the data.

This process has many other names, one of which is functional approximation. Here, an approximation means that the parameterized function may not actually evaluate to exactly match even one set of data values; the idea is that, taking the data as an ensemble, the parameters produce a "good" approximation.

To begin now, we will make two important assumptions about the data. The first is that each set of data is a pair of the form (a, c). The value a is an *independent* value, which means that experimentally that value can be changed by the experimenter. The value c is a *dependent* value, which means that experimentally that value can be measured.

We will examine functional relationships between data pairs. The function has two arguments, which we will separate by a semicolon to clarify that one argument is "fixed" data and one is the

model parameter that we want to "learn". For a data pair, we will write the relationship among the independent data value a, the dependent data value c, and the *model parameter* w as

$$c \approx F(w; a) \tag{10.1}$$

Usually, we have an observation of n variables. The i^{th} observation, is a "row" object that we can write as $\vec{a}_i^T$. A linear model of the independent data would be a vector $\vec{w} \in \mathbb{R}^n$ and our model would be

$$c_i \approx F(\vec{w}; \vec{a}_i^T) \tag{10.2}$$

10.1 Residual Error

A first step towards finding the model parameters in $\vec{w}$ is find the error of the model. This is typically the *residual error*, which is defined as the difference between the dependent value and the model value. The residual error is the difference between the dependent value c_i and the model value $F(\vec{w}; \vec{a}_i^T)$. As is usual in the literature, we will omit the independent variables when we write the residual error for each observation as

$$e_i(\vec{w}) \stackrel{\text{def}}{=} c_i - F(\vec{w}; \vec{a}_i^T) \tag{10.3}$$

We can gather m residual errors, one each for the m observations, as a vector

$$\vec{e}(\vec{w}) \stackrel{\text{def}}{=} \begin{bmatrix} e_1(\vec{w}) \\ e_2(\vec{w}) \\ \vdots \\ e_m(\vec{w}) \end{bmatrix} \tag{10.4}$$

Our regression problem is to minimize the residual error $\vec{e}$ of Equation 10.4, which is a function of our model parameter $\vec{w}$. This is usually approached by using a single number to "measure" the residual error vector. A common choice is to measure the sum of the squares of the individual residual errors, which is

$$\begin{aligned} E_2(\vec{w}) &= \sum_{i=1}^{m} (e_i(\vec{w}))^2 \\ &= \|\vec{e}(\vec{w})\|^2 \end{aligned} \tag{10.5}$$

Equation 10.5 is the squared length of the residual vector. Geometrically, minimizing the sum of the squares of the residual errors is equivalent to minimizing the squared length of the error vector $\vec{e}$.

10.2 Linear Regression

In this course, we will use a linear model of the independent data. That is, the model function $F(\vec{w}\vec{a}_i^T)$ will be a linear function of the independent data. Our linear model will be

$$\begin{aligned} F(\vec{w};\vec{a}_i^T) &\stackrel{\text{def}}{=} a_{i1}w_1 + a_{i2}w_2 + \cdots + a_{in}w_n \\ &= \vec{a}_i^T\vec{w} \end{aligned} \tag{10.6}$$

Our goal is to minimize the residual error between the dependent value c_i and the model value $F(\vec{w};\vec{a}_i^T)$. Substituting Equation 10.6 into Equation 10.3, and gathering the terms, we can write the individual residual errors and the residual error vector as

$$\begin{aligned} e_i(\vec{w}) &= c_i - \vec{a}_i^T\vec{w} \\ \vec{e}(\vec{w}) &= \vec{c} - A\vec{w} \end{aligned} \tag{10.7}$$

We have already solved the problem of minimizing the error vector in Equation 10.7. This is the projection problem that we explored in Class 8. The solution to the sum of squares of the residual error vector $\vec{e}$ in Equation 10.5 is to use the normal equation, which is

$$[A^TA]\vec{w} = A^T\vec{c} \tag{10.8}$$

This gives us a remarkable result:

Linear regression is orthogonal projection to a vector space

10.3 Example: Hooke's Law

An example of a use of Equation 10.6 is that we can use it to write a basic principle of physics, which is that simple mechanical springs can be modeled using a very simple formula. The formula is called Hooke's Law, named after the British polymath Robert Hooke.

The principle is that if a spring is extended or compressed by some scalar distance a from its resting length, then the scalar force c needed to accomplish the distance change is a constant scalar multiple w of the distance change. Mathematically, the relationship between a specific value of the distance change and a specific measurement of the force is given in Equation 10.6. A fictitious example of data for Hooke's Law is shown in Figure 10.1.

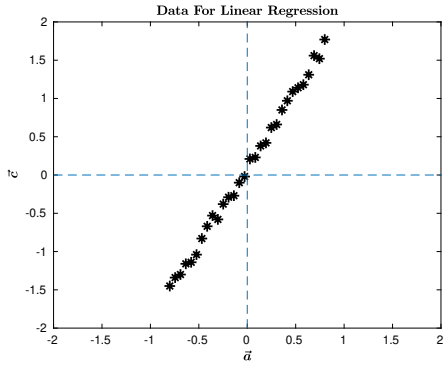

Figure 10.1: Fictitious data for Hooke's Law, where a is the distance change of a mechanical spring and c is the spring force.

The normal equation for Hooke's Law is the parameter $\hat{w}$ that projects the dependent vector $\vec{c}$ to the independent vector $\vec{a}$. The solution to this problem is

$$\begin{aligned} (\vec{a}^T\vec{a})\hat{w} &= \vec{a}^T\vec{c} \\ \Rightarrow \quad \hat{w} &= \frac{\vec{a}^T\vec{c}}{\vec{a}^T\vec{a}} \end{aligned} \tag{10.9}$$

An alternative derivation for Equation 10.9, from basic differential calculus, is provided in the extra notes for this class.

Equation 10.9 is often called the least-squares solution to Hooke's Law. We will think of it as a rule to find a simple pattern in sparse data.

For this linear pattern, least-squares optimization is the same as projection.

10.4 Linear Regression – One Dependent Variable, Plus Intercept

There is an alternative model function $F(w; a_i)$ that is also called "linear regression". Suppose that we are given data such as those shown in Figure 10.2, which seem to have a simple relationship but which do not have $c_i \approx wa_i$.

An appropriate model function for the data in Figure 10.2 might be a first-order polynomial. This would relate a given independent value a_i to a dependent value c_i. Using the `MATLAB` convention for numbering coefficients and powers, we can write the function of a line in 2D as

$$\begin{aligned} P_1(a) &= w_1 a + w_2 \\ c_i &\approx P_1(a_i) = w_1 a_i + w_2 \end{aligned} \tag{10.10}$$

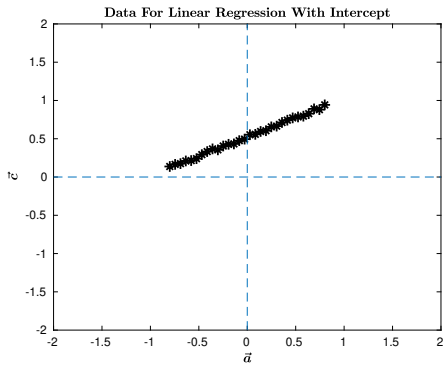

Figure 10.2: Fictitious data for linear regression, where a is the independent variable and c is the dependent variable.

How do we gather the values in Equation 10.10 into vectors? A key insight is to write w_2 as a product of two real numbers, w_2 and 1. When we do so, we can write Equation 10.10 as

$$\begin{bmatrix} a_1 & 1 \\ a_2 & 1 \\ a_3 & 1 \\ \vdots & \vdots \\ a_m & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \approx \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_m \end{bmatrix} \quad \text{or as} \quad \begin{bmatrix} \vec{a} & \vec{1} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \approx \vec{c} \quad \text{or as} \quad A \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \approx \vec{c} \tag{10.11}$$

This kind of linear regression has an *intercept* coefficient w_2. Equation 10.11 is a projection of the dependent vector $\vec{c}$ to the column space of the data matrix A. The solution to this problem of linear regression is Equation 10.8, which is the normal equation for projection.

10.5 Data Matrix For Linear Regression

Earlier in this course, we introduced a *data matrix* that we wrote as the matrix A. We defined the i^{th} observation as the i^{th} row of A. What is the observation for linear regression?

The answer is that we must specify whether or not the regression model includes an intercept. If there is no intercept, then each observation $\vec{t_i}^{\,T}$ will be of size 1, that is, $\vec{t_i}^{\,T}$ will be a scalar. For a single independent variable value a_i, we would define each no-intercept data vector as

$$\vec{t_i}^{\,T} \stackrel{\text{def}}{=} a_i \tag{10.12}$$

If the linear regression is modeled as having an intercept coefficient, then we would define each observation as

$$\vec{t_i}^{\,T} \stackrel{\text{def}}{=} \begin{bmatrix} a_i & 1 \end{bmatrix} \tag{10.13}$$

These concepts generalize to more complicated data. For example, if there are multiple independent variables, we must decide whether or not to include an intercept term. Suppose that we have three independent variables and no intercept term; in this case, the i^{th} observation will have the form

$$\vec{t_i}^{\,T} = \begin{bmatrix} a_{i1} & a_{i2} & a_{i3} \end{bmatrix} \tag{10.14}$$

The data matrix for observations that are described by Equation 10.14 will be

$$A_3 = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & a_{m3} \end{bmatrix} \tag{10.15}$$

Likewise, if we have three independent variables and we include an intercept term, each observation will have the form

$$\vec{t_i}^{\,T} = \begin{bmatrix} a_{i1} & a_{i2} & a_{i3} & 1 \end{bmatrix} \tag{10.16}$$

The data matrix for observations that are described by Equation 10.16 will be

$$A_4 = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 1 \\ a_{21} & a_{22} & a_{23} & 1 \\ a_{31} & a_{32} & a_{33} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & 1 \end{bmatrix} \tag{10.17}$$

The choice of whether to include an intercept term affects the size of the matrix and the resulting regression. The only strict rule is whether we will choose to standardize the data for regression: standardized data will have no intercept term because the process of creating a zero-mean vector for the final column of the data matrix will result in a zero vector, which implies that the standardized data matrix X will not have full rank.

The choice of whether or not to standardize data is often a matter of human judgment. Some algorithms require standardized data, and some data require an intercept term. Methods that reconcile these seemingly conflicting requirements are beyond the scope of this course.

10.6 Assessment Of Linear Regression

How "good" is a solution to a linear regression? The usual way to explore this concept is to assess the residual error vector $\vec{e}$, which was how we solved the regression problem. Let us suppose that we have computed a weight vector $\hat{w}$ that projects the dependent data $\vec{c}$ to the column space of A, that is, $\hat{w}$ solves

$$A\hat{w} \approx \vec{c}$$

Computing the residual error vector requires the data matrix A, the dependent vector $\vec{c}$, and the estimated weight vector $\hat{w}$. Our solution to linear regression minimized the squared norm $\|\vec{e}\|^2$, so this is a reasonable start to our assessment.

We can immediately observe that, if the number of observations m is very large, then even small individual residual errors e_i can add up to a large value. One way that we can compensate is to scale the squared norm by the number of observations, so we might assess a linear regression as

$$\frac{\|\vec{e}\|^2}{m} \tag{10.18}$$

Equation 10.18 assesses a linear regression using a value that is the square of the error. In data analysis, a more common assessment is the square root of the value in Equation 10.18. This is often called the *root mean square* error, variously abbreviated as *RMS error* or as *RMSE*. We will use the former term.

For clarity, we will define the RMS error as depending on the data matrix A, the dependent vector $\vec{c}$, and the estimated weight vector $\hat{w}$. We will define the RMS error as

$$\begin{aligned} \text{RMS}(\hat{w}; A, \vec{c}) &\stackrel{\text{def}}{=} \sqrt{\frac{1}{m}(e_1^2 + e_2^2 + \cdots + e_m^2)} \\ &= \frac{\|\vec{e}(\hat{w})\|}{\sqrt{m}} \end{aligned} \tag{10.19}$$

We can see that, in Equation 10.19, we are using all available data to assess our linear regression. This decision can be altered, which would produce different assessments of our regression.

10.7 Extra Notes For Derivations of Linear Regression

The extra notes provide calculus-based derivations of material from prerequisite courses. These are simple linear regression formulas, first for regression with no intercept and then for regression with an intercept term.

LINEAR REGRESSION – NO INTERCEPT

The residual error for our model of a mechanical spring using Hooke's Law is

$$e_i(w) \stackrel{\text{def}}{=} c_i - wa_i \tag{10.20}$$

In Equation 10.20, we have explicitly made the residual a function of the spring stiffness scalar w. The objective function is the squared error of the residual error vector that we will write as the objective function $E_2(w)$. The objective function for Hooke's Law will be

$$\begin{aligned} E_2(w) &\stackrel{\text{def}}{=} \|\vec{e}(w)\|^2 \\ &\stackrel{\text{def}}{=} \sum_{i=1}^{m}(c_i - wa_i)^2 \end{aligned} \tag{10.21}$$

From basic differential calculus, we know that a necessary condition of a function to be optimized is that the first derivative must be zero. Taking the derivative of Equation 10.21 by distributing the derivative into the summation and applying the Chain Rule, then simplifying, gives

$$\begin{aligned} \frac{dE_2(w)}{dw} &= \sum_{i=1}^{m}\frac{d}{dw}(c_i - wa_i)^2 \\ &= \sum_{i=1}^{m}2(c_i - wa_i)(a_i) \\ &= \sum_{i=1}^{m}(2a_ic_i - 2wa_i^2) \\ &= \sum_{i=1}^{m}2a_ic_i - \sum_{i=1}^{m}2wa_i^2 \end{aligned} \tag{10.22}$$

For the optimal value $\hat{w}$, the condition on Equation 10.22 is

$$\begin{aligned}
\frac{dE_2}{dw}(\hat{w}) &= 0 \\
\Rightarrow \quad \sum_{i=1}^{m} 2\hat{w}a_i^2 &= \sum_{i=1}^{m} 2a_i c_i \\
\Rightarrow \quad \hat{w}\sum_{i=1}^{m} a_i^2 &= \sum_{i=1}^{m} a_i c_i \qquad (10.23) \\
\Rightarrow \quad \hat{w} &= \left(\sum_{i=1}^{m} a_i c_i\right) \Big/ \left(\sum_{i=1}^{m} a_i^2\right) \qquad (10.24)
\end{aligned}$$

Equation 10.24 is our the least-squares solution to Hooke's Law.

Linear Regression – With Intercept

We can formulate the objective function for linear regression with an intercept as

$$\begin{aligned}
E_2(w_1, w_2) &\stackrel{\text{def}}{=} \|\vec{e}(w_1, w_2)\|^2 \\
&\stackrel{\text{def}}{=} \sum_{i=1}^{m} (c_i - (w_1 a_i + w_2))^2 \qquad (10.25)
\end{aligned}$$

Because $E_2(\cdot)$ is a function of two variables, a necessary condition at its minimum is that the partial derivative with respect to each variable – to w_1 and to w_2 – is zero. Using the Chain Rule, and factoring the constant terms out of the summations, the partial derivatives are

$$\begin{aligned}
\frac{\partial E_2}{\partial w_1}(w_1, w_2) &= 2\sum_{i=1}^{m}(-a_i)(c_1 - (w_1 a_i + w_2)) \\
\frac{\partial E_2}{\partial w_2}(w_1, w_2) &= 2\sum_{i=1}^{m}(-1)(c_i - (w_1 a_i + w_2)) \qquad (10.26)
\end{aligned}$$

For the computed values $\hat{w}_1$ and $\hat{w}_2$, the partial derivatives must be equal to zero. Substituting the computed terms into Equation 10.26 produces the two equations

$$\frac{\partial E_2}{\partial w_1}(\hat{w}_1, \hat{w}_2) = 0 \quad \text{and} \quad \frac{\partial E_2}{\partial w_2}(\hat{w}_1, \hat{w}_2) = 0 \qquad (10.27)$$

Dividing each side of Equation 10.27 by -2 gives

$$\begin{aligned}\sum_{i=1}^{m}(a_i)(c_1 - (\hat{w}_1 a_i + \hat{w}_2)) &= 0 \\ \sum_{i=1}^{m}(c_i - (\hat{w}_1 a_i + \hat{w}_2)) &= 0 \end{aligned} \tag{10.28}$$

Gathering terms, and observing that $\sum_{i=1}^{m} \hat{w}_2 = m$, Equations 10.28 can be written as

$$\begin{aligned}\left(\sum_{i=1}^{m} a_i^2\right)\hat{w}_1 + \left(\sum_{i=1}^{m} a_i\right)\hat{w}_2 &= \sum_{i=1}^{m} a_i c_i \\ \left(\sum_{i=1}^{m} a_i\right)\hat{w}_1 + m\hat{w}_2 &= \sum_{i=1}^{m} c_i \end{aligned}$$

In matrix form, this becomes

$$\begin{bmatrix} \sum_{i=1}^{m} a_i^2 & \sum_{i=1}^{m} a_i \\ \sum_{i=1}^{m} a_i & m \end{bmatrix} \begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} a_i c_i \\ \sum_{i=1}^{m} c_i \end{bmatrix} \tag{10.29}$$

In statistics and related fields, Equation 10.29 may be solved explicitly as

$$\begin{aligned}\hat{w}_1 &= \frac{\left(\sum_{i=1}^{m} a_i\right)\left(\sum_{i=1}^{m} c_i\right) - m\left(\sum_{i=1}^{m} a_i c_i\right)}{\left(\sum_{i=1}^{m} a_i\right)^2 - m\sum_{i=1}^{m} a_i^2} \\ \hat{w}_2 &= \frac{\left(\sum_{i=1}^{m} a_i\right)\left(\sum_{i=1}^{m} a_i c_i\right) - \left(\sum_{i=1}^{m} a_i^2\right)\left(\sum_{i=1}^{m} c_i\right)}{\left(\sum_{i=1}^{m} a_i\right)^2 - m\sum_{i=1}^{m} a_i^2} \end{aligned} \tag{10.30}$$