Linear Models and Linear Regression

APCOMP209a: Introduction to Data Science — Wed/Thurs 2:30-3:30 & Wed 5:30-6:30
Nick Hoernle — nhoernle@g.harvard.edu

1 Recap

Recall that we have an unknown function (f) that relates the response variable (y_i) to the input vector ($\mathbf{x_i}$). Our goal is to find a model ($\hat{f}$) (i.e. we are approximating f) such that a *loss function* is minimised. We may want to use this model for **prediction** and/or for **inference**.

We can say we have a training dataset with N *i.i.d* training datapoints $(y_i, \mathbf{x_i})$, $i = 1, \ldots, N$, which each consist of a one dimensional response variable and a p dimensional input vector ($y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$).

An **assumption** in linear regression is that the predictor function that we are approximating is linear. We can then write this relationship as:

$$y_i = f(x_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$$
$$Y = f(\mathbf{X}) = \mathbf{X}\beta + \epsilon$$

Where $\mathbf{X}$ now refers to an $N \times (p+1)$ dimensional matrix.

For now, let us assume that we know 'p' in advance (we will deal with this assumption more under the **model selection** topic in this class). We are now able to construct our *model*:

$$\hat{Y} = \hat{f}(\mathbf{X}) = \mathbf{X}\hat{\beta}$$

Recall, that there is a portion of variance in the data that can be explained by the model and there is a portion of the variance in the data that is purely statistical noise and cannot be explained by the model. Heuristically, we aim to minimise some distance metric between our predictions $\hat{Y}$ and the true training data Y.

For linear regression, we make the **assumption** that the noise (ϵ) is distributed as a Normal random variable with mean 0, variance σ^2. I.e. $\epsilon \sim \mathcal{N}(0, \sigma^2)$. You can then follow that $Y \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2)$. It is therefore common to denote the linear regression problem as finding the **expected value** of Y given the input variables $\mathbf{X}$.

$$E[Y|\mathbf{X}] = \mathbf{X}\hat{\beta}$$

More on this topic in upcoming classes.

2 Matrix Algebra Recap

Please refer to the really useful Matrix Cookbook for a more detailed recap on matrix operations: (`http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf`)

We'll be using the following results (although I highly recommend you download a copy of the cookbook [1] and keep it handy):

1. $(AB)^{-1} = B^{-1}A^{-1}$
2. $(A^T)^{-1} = (A^{-1})^T$
3. $\|x\|_2^2 = x^H x$ ... (note that 'H' refers to the Hermitian vector (transposed, complex conjugated) and thus for most of our purposes (i.e. Real domain), the transposed (T) vector is sufficient).
4. $\frac{\partial}{\partial x}[(b-Ax)^T(b-Ax)] = -2A^T(b-Ax)$
5. The density of $x \sim \mathcal{N}(\mu, \Sigma)$ is $p(x) = \frac{1}{\sqrt{det(2\pi\Sigma)}} exp[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)]$

The above assumes A and B are matrices, x and b are vectors.

3 Minimising the Loss Function

We have a system where the data Y and our model $\hat{Y}(= \mathbf{X}\hat{\beta})$ differ by some **residual** amounts. Our goal is to find the unknown parameters $\hat{\beta}$ such that the residuals are minimised. Since we are trying to minimise the error of the model over all of the datapoints, it makes sense to minimise a sum of all square magnitudes of errors:

$$SSE = \sum_{i=0}^{N} |residual_i|^2 = \sum_{i=0}^{N} |y_i - \beta \mathbf{x}_i^T|^2 = \|Y - \mathbf{X}\hat{\beta}\|_2^2$$

Choosing to minimise the *sum of square errors (SSE)*:

$$\hat{\beta} = \min_{\hat{\beta}}(SSE) = \min_{\hat{\beta}} \|Y - \mathbf{X}\hat{\beta}\|_2^2 = \min_{\hat{\beta}}((Y - \mathbf{X}\hat{\beta})^T(Y - \mathbf{X}\hat{\beta}))$$

Finding the gradient and setting it to zero, we can obtain:

$$\begin{aligned} \frac{\partial SSE}{\partial \hat{\beta}} &= -2\mathbf{X}^T(Y - \mathbf{X}\hat{\beta}) = 0 \\ \mathbf{X}^T\mathbf{X}\hat{\beta} &= \mathbf{X}^T Y \\ \hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y \end{aligned}$$

4 Linear Regression as a Projection

Our predictions $\hat{Y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$ can be condensed into the following equation:

$$\hat{Y} = \mathbf{H}Y$$

Where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. This matrix is often referred to as "the hat matrix as it puts the hat on the y" [2]. Note that the columns of the $\mathbf{X}$ matrix form a subspace of $\mathbb{R}^N$ that is referred to the column space of $\mathbf{X}$.

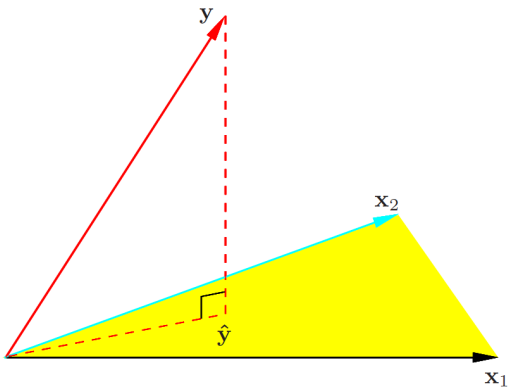

Figure 1: Diagram showing the vector y projected onto the subspace spanned by the matrix $\mathbf{X}$ in this case with two linearly independent dimensions[2]

When we minimise the error between the solution Y and the vector projection $\hat{Y}$ (see Figure 1), the result is that the error must be orthogonal to the column space (i.e. the solution to the least squares problem is the orthogonal projection of the vector Y onto the subspace that is spanned by the columns of $\mathbf{X}$). *Why is this useful to know?* It is useful to visualise the prediction vector $\hat{Y}$ as a linear combination of the columns of $\mathbf{X}$ and this $\hat{Y}$ vector is the 'closest' in $\mathbb{R}^N$ that the prediction can get to the real solution (try to visualise this in terms of the reducible and irreducible errors discussed in class).

5 Statistical Inference and Hypothesis Testing

Remember that our ultimate goal is to model the linear relationship between various predictors and the response variable. If our model is 'good', not only can we use it to make **predictions** about future or unknown events, but we can also use it to make **inferences** about the underlying structure of the system. Up until now we have assumed that we *knew* the true number of predictors and that we *knew* they had a *linear* relationship with the response variable. This is often not the case and therefore in statistical inference, assumptions of the model have to be validated before any inference is done.

To begin, let us tackle the idea of having a linear relation. Given data $(y_i, \mathbf{x_i})$, $i = 1, \ldots, N$, we can ask the question: *is there a true linear relationship between the predictor variable x and the response variable y?* We

need to answer this question with *statistical evidence*. For example, consider the plots below where there are 10 datapoints sampled from four different linear relationships. We need a robust method for analysing which relations are *statistically significant* and which are not (as it is clear that in all cases, due to the noise in the system, there is not one linear function relating the predictor and the response variables). We thus turn to statistical 't-' (5.2) and 'F-' (5.3) tests to make conclusions about the underlying system given the sample of data we have observed.

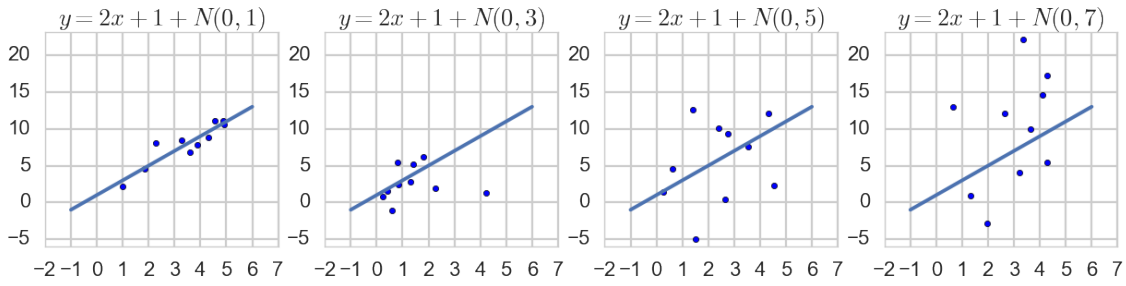

Figure 2: Example of a linear function with varying amounts of noise. We need a robust way of determining if our samples actually have a relationship or if we are just observing noise.

The idea of hypothesis testing is to make some assumptions about the nature of the true system, given the sample that you are observing, and **if those assumptions hold**, you can conclude whether or not a certain null hypothesis is probabilistically reasonable or not.

Examples of assumptions for linear regression include:

- There is a *linear* relationship between the predictor and response variables.
- The noise is *Gaussian* (with mean 0) around $E[Y|X]$.
- The noise has a constant variance around the line of regression. We say this is an assumption of *homoskedasticity*.
- There is little or no *multicollinearity* among the predictor variables.

5.1 p-value in Hypothesis Testing

We are making statements about the statistical likelihood of sampling a certain subset of data given the underlying truth. This is all wrapped into the concept of the p-value which literally translates to *the probability of observing the sampled data or more extreme samples, given the null hypothesis.* It is worth noting that, under the null hypothesis, the p-value follows a uniform distribution (can you connect this to the CDF and inverse CDF of the given distribution?).

A simple example helps: Imagine that we are sampling data from *what we believe is a standard normal distribution* ($\mathcal{N}(0, 1)$). If we have a sample of data ≈ 0 we would agree that our observation coincides with our *null hypothesis* that the system is standard normal. However, now imagine that the observation is ≈ 5. For a standard normal distribution, the probability of observing a sample of 5 (or more) is $p = 2.87 \times 10^{-7}$. That's a REALLY small probability. So with just one sample, we don't have too much to say, but getting

more samples (x) that are $|x| >> 0$ provides substantial statistical evidence that our *null hypothesis* (that the data is sampled from a standard normal distribution) is incorrect.

Therefore, in hypothesis testing, we map the observed data to a *specific* distribution that would hold under a null hypothesis. If the observed data is 'unlikely' enough given the assumptions and the null hypothesis, i.e. if the p-value is very small, we say that we reject the null hypothesis and rather accept an alternative hypothesis.

5.2 t-test

Given data that we believe has a linear relationship mapping the response variables to the predictor, we can test whether a particular predictor has a relationship with the response variable by testing whether the true coefficient $\beta_j \neq 0$ for some j (noting that the observed parameter $\hat{\beta}_j$ is based on a random small sample of data).

We therefore wish to test the following hypotheses (in general we can test whether the coefficient is equal to any value (μ_0), however, in practice we usually want to test $\beta_j = \mu_0 = 0$):

Null and Alternative Hypotheses:

$$H_0 : \beta_j = \mu_0$$
$$H_A : \beta_j \neq \mu_0$$

Under the null hypothesis we have the following t-statistic:

$$t = \frac{\hat{\beta}_j - \mu_0}{SE(\hat{\beta}_j)} \tag{1}$$

Using equation 1, we can calculate a number from the data that we have sampled. This number needs to be compared to a reference distribution to determine how extreme the sample would be under the null hypothesis. Returning to our β estimator from earlier:

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + \epsilon) \\ &= \beta + (X^T X)^{-1} X^T \epsilon \\ \hat{\beta} &= \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})\end{aligned}$$

With $Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$ and so $Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{N} (x_{ij} - \bar{x}_j)^2} = \frac{\sigma^2}{SSX_j}$

Moreover, the variance σ^2 is approximated by $\hat{\sigma}^2 = \frac{SSE}{n-p-1}$, where $SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$

$$
\begin{aligned}
t &= \frac{\hat{\beta}_j - \mu_0}{\sqrt{\frac{\frac{SSE}{n-p-1}}{SSX_j}}} \\
&= \frac{\frac{\hat{\beta}_j - \mu_0}{\sqrt{\frac{\sigma^2}{SSX_j}}}}{\sqrt{\frac{SSE}{\sigma^2(n-p-1)}}} \\
&= \frac{z}{\sqrt{\frac{\sum_{i=1}^{N}(\frac{y_i - \hat{y}_i}{\sigma})^2}{(n-p-1)}}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2_{n-p-1}}{n-p-1}}} \sim t_{n-p-1}
\end{aligned}
$$

As $z = \frac{\hat{\beta}_j - \mu_0}{\sqrt{\frac{\sigma^2}{SSX_j}}}$ follows a standard normal distribution, $V = \sum_{i=1}^{N}(\frac{y_i - \hat{y}_i}{\sigma})^2$ follows a chi-squared distribution with $n-p-1$ degrees of freedom ($V \sim \chi^2_{n-p-1}$), under the null hypothesis, the t-statistic derived above must then follow a t distribution with $n-p-1$ degrees of freedom.

We are now able to calculate the **probability** of observing the **specific sample** (or a more extreme value) under these conditions. This probability is referred to as the p-value (5.1).

Consider the following data generated by $y = 2x + 1 + \epsilon$:

x	3.15	0.77	2.44	2.88	4.93	3.77	4.73	3.00	0.28	1.72
y	5.88	6.26	5.16	2.20	5.24	3.66	9.89	5.91	2.77	5.47

The least squares solution to this problem yields: $\hat{\beta}_0 = 3.74$, $\hat{\beta}_1 = 0.544$. Note how we were unable to recover the original parameters of $\beta_0 = 1$ and $\beta_1 = 2$ from the noisy sample. Regardless, we are trying to determine, given the sample of data, if there **actually is** a linear relationship in the system. Under the hypotheses $H_0 : \beta_1 = 0$ and $H_A : \beta_1 \neq 0$, we can obtain the t-statistic: $t = \frac{\hat{\beta}_j - \mu_0}{SE(\hat{\beta}_j)} = \frac{0.544}{0.456} = 1.193$. The corresponding two sided *p-value* $= 0.267$. See Figure 3 for a graphical representation of this result. Using the usual test for significance $\alpha = 0.05$, we would be unable to reject the null hypothesis in this case. In otherwords, the sample that we have obtained *could* be consistent with a system parameter of $\beta_1 = 0$.

Note that as our linear model complexity increases, we may not want to test relations in isolation and moreover, despite our model assumptions, there may be some *correlation* among predictor variables that affect the inferences that we make from a given model. We can then run a statistical test that accounts for multiple variables at once, this is the F-test (5.3) derived from the ANOVA test.

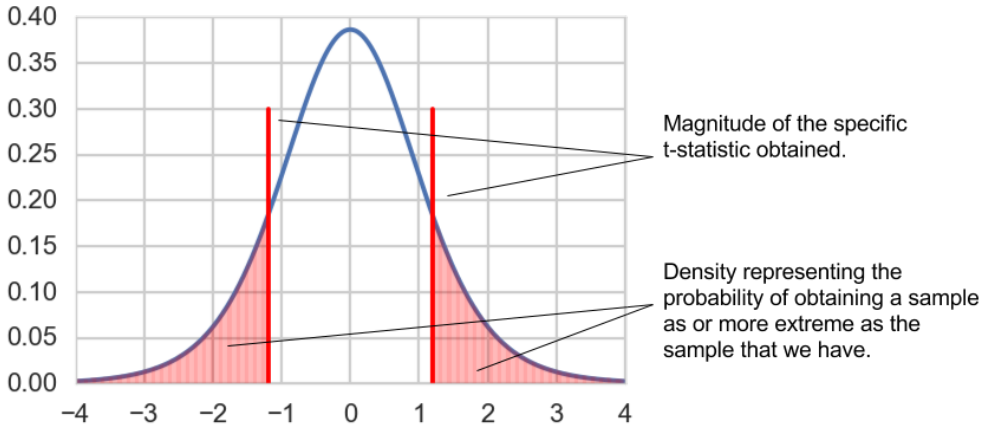

Figure 3: Figure showing the reference t-distribution (with 8 degrees of freedom) and the magnitude of the test statistic achieved in this example

5.3 F-test

Assuming the above assumptions are met, for two nested models (i.e. consider model 1 (M_1) and model 2 (M_2) where M_2 contains all of the predictors of M_1 and more), we can test whether the additional parameters of M_2 form a significantly better model than those found in M_1.

Null and Alternative Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$
$$H_A : \beta_j \neq 0, \textit{for at least one value of } j$$

To start, we consider the R^2 which is the proportion of the overall variability of Y that is explained by the model $\hat{Y}$.

$$R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{TSS - RSS}{TSS}$$

with $MSS = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$, $TSS = \sum_{i=1}^{N}(y_i - \bar{y})^2$, $RSS = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

The F-test then tests if all of the β_j coefficients are 0. Essentially, if the model explains 'most' of the variance in the system, there is evidence that at *least one of the predictors* must be linearly associated with the response. We are comparing the ratio of the average amount of variance that each predictor explains to the average amount of variance that is left unexplained in the model.

$$f = \frac{\textit{Explained Variance}}{\textit{Unexplained Variance}} = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \tag{2}$$

This statistic can be shown to follow a F-distribution:

$$\begin{aligned} f &= \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \\ &= \frac{(\sum_{i=1}^{N}(\frac{\hat{y}_i - \bar{y}}{\sigma})^2)/p}{(\sum_{i=1}^{N}(\frac{y_i - \bar{y}}{\sigma})^2)/(n-p-1)} \sim \frac{\chi_p^2/p}{\chi_{n-p-1}^2/(n-p-1)} \sim F_{p,n-p-1} \end{aligned}$$

Be careful: Violations of the assumptions (linearity/homoskedasticity/Gaussian noise) can lead to incorrect results.

References

[1] K. B. Petersen, M. S. Pedersen, *et al.*, "The matrix cookbook," *Technical University of Denmark*, vol. 7, 2008.

[2] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics New York, 2001.

[3] Ryan Lee's class notes

[4] CS109a Lecture Notes