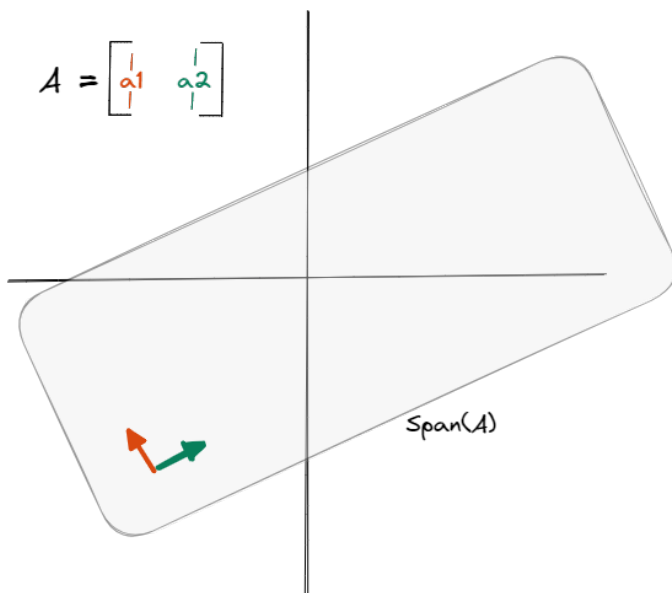


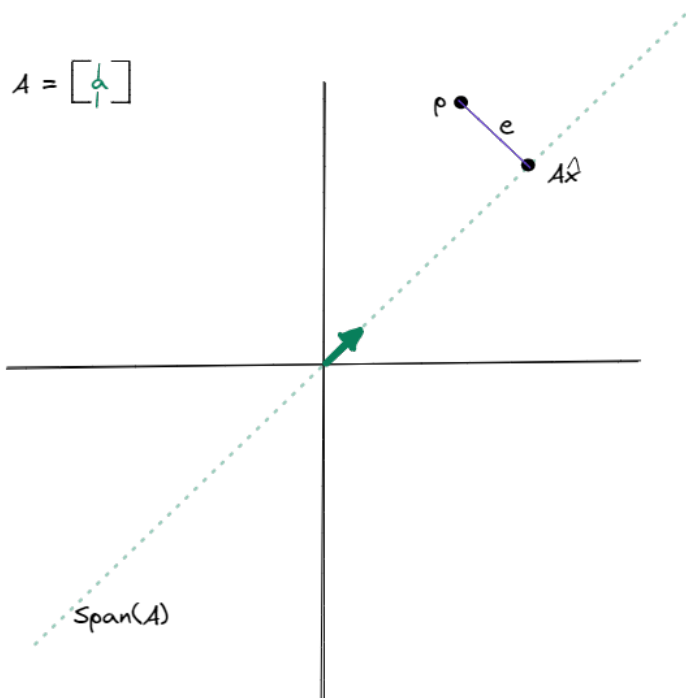
Regression notes

Linear algebra

Before getting into regression, let's warm up by looking at some linear algebra. Think of a matrix A as a collection of column vectors, then Ax can be interpreted as a linear combination of these vectors. Whenever you have a matrix, a natural thing to think about is the *span* - the collection of all vectors which can be written as a linear combination of the columns of A . Or, using less words, all vectors $b = Ax$ for some x . A rough sketch of the geometry is shown below:



If A is an $n \times p$ matrix and $n > p$, then A will not span the entire space - $\dim(\text{span}(A)) < p$. There will be some p -dimensional vectors which can't be written as Ax . It's easiest to see this if you look at a single vector in 2D. The span is just all scalar multiples of the vector - a line. Any point not on the line can't be expressed as a multiple of the vector:



It's impossible to write $p = Ax$, since p isn't in the span of A . A natural question to ask is *which point in the span is closest to p* ? To solve this, you want to find a vector \hat{x} such that the error $e = p - A\hat{x}$ is as small as possible. From the geometry it's clear that, for the closest point, the error is perpendicular to $\text{span}(A)$. In linear algebra, things are perpendicular if their dot product is zero. This gives us a constraint, which we can use to figure out \hat{x} :

$$0 = A^T e \implies \hat{x} = (A^T A)^{-1} A^T p$$

So the point closest to p is

$$A\hat{x} = A(A^T A)^{-1} A^T p = H p$$

H is sometimes called the *hat matrix*. H is an $n \times n$ matrix with the following properties:

- it is idempotent ($H^2 = H$) - to see this just do the multiplication
- it is symmetric ($H = H^T$) - this is because $A^T A$ is symmetric
- $\text{trace}(H) = p$ - this is because trace doesn't care about the order of matrices so $\text{trace}(A(A^T A)^{-1} A^T) = \text{trace}((A^T A)^{-1} A^T A) = \text{trace}(I_p) = p$

Linear regression

What's all this got to do with linear regression? In linear regression you collect some data D , which contains n measurements of p variables. There is also a response vector y which you're interested in estimating. Linear regression says that y can be estimated using a linear predictor. For the i th observation in the data, linear regression says that the estimated response variable \hat{y}_i has the form

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}$$

The data D is an $n \times p$ matrix, and in most cases you have more measurements than variables so $n > p$. To make the linear algebra version of regression, form the *design matrix* $X = [1 \ D]$ (i.e add a column of 1's to the front of the data matrix), and a parameter vector $\beta = [\beta_0 \ \beta_1 \ \dots \ \beta_p]^T$. Then the estimates \hat{y} can be written as

$$\hat{y} = X\beta$$

Since the design matrix has more rows than columns, there will be values of y which can't be written as $X\beta$. We would like to find the parameters $\hat{\beta}$ which gets \hat{y} as close as possible to y . This is exactly the problem from the last section, so we can immediately write the solution as

$$\hat{y} = Hy, \quad H = X(X^T X)^{-1} X$$

This is why H is called the hat matrix - it puts a hat on y .

You can think of regression as projecting the true response variable y onto a lower dimensional space spanned by the model variables. This projection is an estimate of the true y so there will be some error, and the hat matrix H gives the best projection (in terms of lowest possible error). Usually the true distribution of y is unknown, and the lower dimensional model allows us to work with a process which would otherwise be impossible to work with. For example, modelling the detailed chemical processes which happen after you take a medication is extremely difficult or impossible, but we can still measure some variables (e.g dose, strength of response, age, sex, ...) which will allow us to examine the effect of changing the dose has on the response.

Inference

We need to make additional assumptions if we want to estimate the parameters β . In particular we assume a normal model

$$y \sim N(X\beta, \sigma^2 I)$$

Which tells us that $E\hat{y} = y$ and $V\hat{y} = \sigma^2$. Usually σ^2 is unknown so needs to be estimated from the data in the usual way

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \sum_i (y_i - \hat{y}_i)^2$$

The denominator has the form $n_{obs} - n_{df}$, where n_{df} is the number of parameters estimated when fitting the model (p parameters + 1 intercept). Because $\hat{\sigma}^2$ is estimated it is a random variable, and since it is a sum of squared normals it follows that $\hat{\sigma}^2 \sim \chi^2$. This means that the usual test statistics are then ratios of χ^2 variables, so will be F distributed. Occasionally the test statistics will be t distributed, since $F_{1,p} = t_p^2$ - this usually happens if you're testing a fitted coefficient against zero.

Diagnostics

Once you've fit a model, you need to check if it's any good. The main way how model fit is assessed is through *residuals*, defined as

$$r = y - \hat{y} = y - Hy = (I - H)y$$

The variance of the residuals are

$$Vr = V[(I - H)y] = (I - H)^T Vy (I - H) = \sigma^2(I - H)^T(I - H) = \sigma^2(I - H)$$

Where the last equality comes from the fact $I - H$ is symmetric and idempotent (because H is). Once again, $I - H$ is a big $n \times n$ matrix. The diagonals are the variance of the residual, and the off diagonals are the covariances between residuals. The variance of the i th residual is

$$Vr_i = (1 - h_{ii})\sigma^2$$

So residual variance is determined by two things - the diagonal elements of the hat matrix, and the variance in y . The diagonals are called *leverage* - large leverages result in large residual variance. We can use the fact that $\text{trace}(H) = p + 1$ (the number of parameters + intercept) to get a feel for what constitutes a 'large' leverage, because H is $n \times n$ there are n diagonals and we would expect an average diagonal to be around $(p + 1)/n$. If the leverage for observation i is larger than $2(p + 1)/n$ - twice the average expected leverage - then that observation may have a large effect on the model. High leverage points should be examined to see if they are outliers.

Another way to look at residuals is to standardise them. The standardised residuals (also known as 'studentised residuals') are the observed residual compared to the expected residual

$$r_i^{std} = \frac{r_i}{\sqrt{Vr_i}} = \frac{r_i}{\sqrt{1 - h_{ii}}\hat{\sigma}}, \quad r_i \sim t_{n-p-1}$$

Standardised residuals are the (square root of) a ratio of chi-squared distributions, and the numerator has 1 degree of freedom - this means that standardised residuals are t distributed. In most cases the residual degrees of freedom are large enough that t is approximately normal, so you can check for potential outliers by looking for any standardised residuals greater than 2, $|r_i^{std}| > 2$.

It feels a bit weird using $\hat{\sigma}$ when calculating standardised residuals - remember $\hat{\sigma}^2 = \sum_j r_j^2 / (n - p - 1)$, so it depends on r_i . If observation i is an outlier then it will have a large residual r_i , which will lead to an overestimate of $\hat{\sigma}^2$ and so underestimates of r_i^{std} . A solution to this problem is to simply remove r_i when calculating $\hat{\sigma}^2$ (and reduce the number of degrees of freedom by 1), resulting in the *jack-knife* or *prediction* residuals

$$r_i^{pred} = \frac{r_i}{\sqrt{Vr_i}} = \frac{r_i}{\sqrt{1 - h_{ii}}\hat{\sigma}_{(i)}}, \quad \hat{\sigma}_{(i)}^2 = \frac{1}{n - p - 2} \sum_{j \neq i} r_j^2$$

It turns out that removing the i th residual in the variance calculation is exactly identical to fitting a model with the i th row removed from the data, using the model to predict \hat{y}_i , and calculating a residual using $r_i = y_i - \hat{y}_i$.

One last diagnostic measure is *cooks distance*, which measures the change in parameter estimates after removing one observation from the data. Cooks distance can also be thought of as a measure of the changes in predicted values after removing one observation. The formula for the cooks distance for observation i is

$$d_i = \frac{(X\hat{\beta}_{(i)} - X\hat{\beta})^T (X\hat{\beta}_{(i)} - X\hat{\beta})}{(p+1)\hat{\sigma}^2} = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{(p+1)\hat{\sigma}^2}$$

It's a weighted sum of squared differences of the predictions from the two models. The weight is the estimated variance (MSE) and the number of parameters in the model. Looking at the second formula, the numerator is a sum of $p+1$ squared normals so is distributed χ_{p+1}^2 . The MSE in the denominator is distributed χ_{n-p-1}^2 , so $d_i \sim F_{p+1, n-(p+1)}$. Any $d_i > 0.5$ is suspicious and may need further investigation.

ANOVA, ANCOVA

To do

Generalised Linear Models (GLMs)