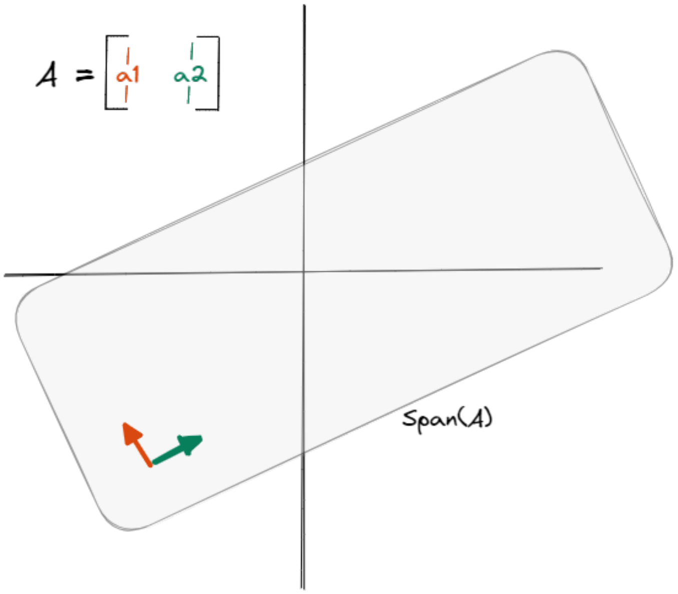
Regression notes

Linear algebra

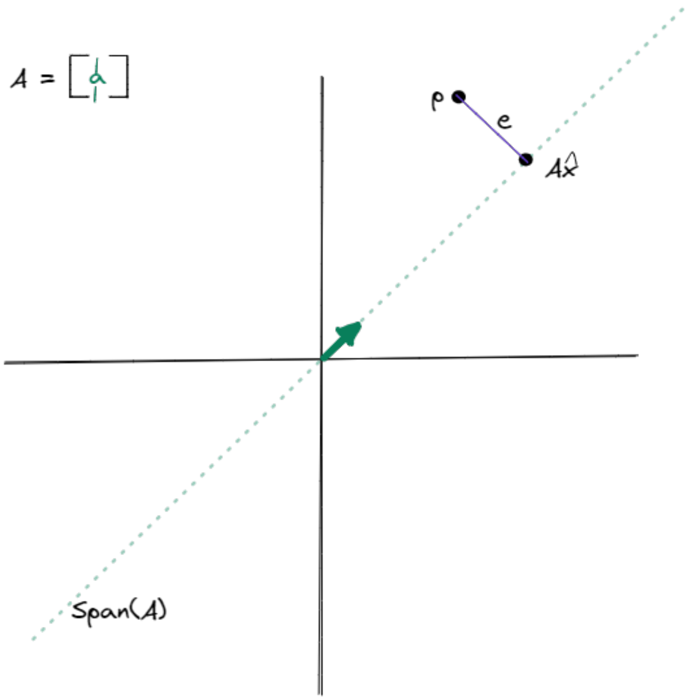
Before getting into regression, lets warm up by looking at some linear algebra. Think of a matrix A as a collection of column vectors, then Ax can be interpreted as a linear combination of these vectors. Whenever you have a matrix, a natural thing to think about is the *span* - the collection of all 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 $\rho = Ax$, since ρ isn't in the span of A. A natural question to ask is *which point in the span is closest to* ρ? To solve this, you want to find a vector $\hat{x}$ such that the error $e = \rho - A\hat{x}$ is as small as possible. From the geometry it's clear that, for the closest point, the error is perpendicular to 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 $\hat{x}$:

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



So the point closest to ρ is

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

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^TA is symmetric
- trace$(H) = p$ - this is because trace doesn't care about the order of matrices so trace$(A(A^TA)^{-1}A^T) =$ trace$((A^TA)^{-1}A^TA) =$ 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 ith observation in the data, linear regression says that the estimated response variable $\hat{y}_i$ has the form

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

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 $\beta = [\beta_0\ \beta_1\ \ldots\ \beta_p]^T$. 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

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

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

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

We observe the data to make the design matrix X, and we try to estimate $\hat{\beta}$. Using the formula for $\hat{\beta}$ from the previous section, $\hat{y} = Hy = X(X^TX)^{-1}X^Ty = X\hat{\beta}$, so $\hat{\beta} = (X^TX)^{-1}X^Ty$. We can use this to figure out the expected value & variance of the parameter estimates. Assuming the model is correctly specified, we have $E\hat{\beta} = E[(X^TX)^{-1}X^Ty] = E[(X^TX)^{-1}X^TX\beta] = \beta$ and $V\hat{\beta} = V[(X^TX)^{-1}X^Ty] = (X^TX)^{-1}X^T\sigma^2 IX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$.

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 $F_{1,p} = t_p$ - this usually happens if you're testing a fitted coefficient against zero.

We've motivated the formulas for $\hat{\beta}$ using just linear algebra so far, but you can go through the usual maximum likelihood arguments and reach the same result. This means that $\hat{\beta}$ is the MLE for β, so the standard asymptotic results apply. In particular

$$\hat{\beta} \sim N\left(E\hat{\beta},\, V\hat{\beta}\right) = N\left(\beta,\, \sigma^2(X^TX)^{-1}\right)$$

This means that 95% confidence intervals for the parameters have the form $\hat{\beta} \pm 1.96\sqrt{V\hat{\beta}}$. Strictly, since the variance depends on the unknown parameter σ^2, we should use a t-distribution here and write $\hat{\beta} \pm t_{n,0.975}\sqrt{V\hat{\beta}}$

where n is the number of residual degrees of freedom (the number of data points less the number of parameters). But usually n is sufficiently large that the difference between the t and normal distributions becomes negligible.

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 ith 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 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

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

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

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

It turns out that removing the ith residual in the variance calculation is exactly identical to fitting a model with the ith 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

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

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^2_{p+1}. The MSE in the denominator is distributed χ^2_{n-p-1}, so $d_i \sim F_{p+1,\, n-(p+1)}$. Any $d_i > 0.5$ is suspicious and may need further investigation.

ANOVA, ANCOVA

For historical reasons people call a regression involving a single categorical variable *ANOVA*, and a regression involving categorical and continuous variables *ANCOVA*. There's two main things to mention about ANOVA - the ANOVA formulation of models, and testing nested models.

ANOVA formulation

Imagine you've got a single categorical variable with two levels x and an outcome variable y. There are two models you could fit - the usual regression model

$$y_i \sim \beta_0 + \beta_1 x_i$$

Or an equivalent ANOVA model

$$y_i \sim \mu_k$$

In the first model β_0 is the mean of y in group 0, and β_1 is the difference in group means between group 1 and group 0. The second model says that the expected value for person i is their group mean. Both of these are equivalent - $\beta_0 = \mu_1$, $\beta_1 = \mu_2 - \mu_1$. If you wanted to test for a difference between the groups you could test the null hypothesis $\beta_1 = 0$, and in the second model you would test to see if the group means are equal to the overall mean of all participants. Both models capture the same information, they just phrase the analysis differently. The ANOVA formulation is quite natural if you're doing a lot of experimental design. For example imagine you're going to give a placebo to one group and an active drug to the other group - it's natural to think in terms of the groups (ANOVA formulation) rather than in terms of the difference between the groups (regression formulation). Use whichever one you're comfortable with though, and make sure you're able to convert between the different formulations!

Differences start to show if you have more than 2 groups. In the regression setup you would test the hypotheses $\beta_1 = 0, \beta_2 = 0, \dots, \beta_m = 0$, and if the number of groups m is large you would run into issues around multiple testing and false positives. The null for the ANOVA setup is unchanged - it still tests the null of 'all group means are similar to the overall mean'. So if you're just looking for evidence that at least one of the groups is different to the others, ANOVA is a better formulation.

Testing nested models - F tests

Regression models the expected value of y, conditional on the predictor variables - $E[y \mid x_1, x_2, \ldots, x_p]$. This is a conditional expectation, so we can use the partition of variance formula $VY = E[V[Y \mid X]] + V[E[Y \mid X]]$. Any regression model partitions the variance in the data into two terms - the variance explained by the model $V[E[Y \mid X]]$ and the residual variance $E[V[Y \mid X]]$. Both of these terms are sums of squared residuals, so will be distributed chi squared (after dividing by the number of residual degrees of freedom). Most stats packages will report the mean sum of squares MSS, which is the scaled sum of residuals we're after. The global F test looks to see if the variance explained by the model is significantly larger than the residual variance:

$$F = \frac{MSS_{model}}{MSS_{resid}}$$

Your model should hopefully explain most of the variance, so large F statistics are good. Compare against the upper tail of the appropriate F distribution to get p-values. This tests that at least one of the model parameters (except the intercept) is non-zero, because the intercept-only model would just capture the variation about the mean of the data ie MSS_{resid}.

Now suppose you have two models m_1, m_2, and m_2 has k extra parameters compared to m_1. We say that $m1$ is *nested* inside $m2$, because if you remove the k extra parameters from m_2 you would recover m_1. Both of these models will partition the variance into explained & residual components, and we want to quantify if m_2 is a better model than m_1. Looking at the difference in the variance explained is a natural place to start, and we can make the difference into a chi-square variable by dividing by the number of extra parameters in model 2: $(SS_2 - SS_1)/k$. This is guaranteed to be ≥ 0 since adding more terms will never reduce the amount of variance the model can explain. Comparing this to the residual variance in model 2 gives another F statistic:

$$F = \frac{(SS_2 - SS_1)/k}{MSS_{resid\ m_2}}$$

This is a *partial* F test, and it tests the hypothesis that all of the k extra parameter coefficients are zero, i.e. that adding the extra variables had no significant effect on improving the variance explained by the model. Similar to the glabal F test, larger values are better and will give smaller p-values.