**Linear Design Matrices**

So in the Hypothesis test file, we saw how to calculate a p-value for comparing the means of two different groups. And in the Regression file, we saw how to calculate a p-value for a regression curve. Now we’re going to kind of combine the two concepts together. For instance, we’d like to be able take a regression curve for one set of data, say group A, and a regression curve for another set of data, say group B, and calculate a p-value for whether the regressions are meaningfully different. Or another way to say it, is there any predictive power in classifying data into one of those two groups, or should we just treat them as all the same group? So this would be like doing a regression fit over independent variables, some of which are continuous, and some of which are discreet (the group index).

**Linear Regression**

Now we’ll look at comparing linear regressions between groups using a design matrix.

Chart, scatter chart

Description automatically generated

We could write categorical regression equation like this:



where f is modeling size, s, and mA,B = best fit slope for group A,B data, bA,B = best fit intercept for group A,B data, and A = 1,0, B = 0,1 for whether it’s group A,B, just like in previous file. Another, more common I think, way to write this is,



Not sure why this is preferred. The blue guys are the fit parameters. w and B are the data. Accounting for random variations about this fit, we could write this in matrix form,



where,



and the ε’s are normally distributed variables with mean 0 and std σ. And X would be our **design matrix**. If we minimize the total square error of our fit,



then we should get the usual expressions for mA,B and bA,B found in the Regression file. But in general, as we saw in the multiple regression file, we will find the fit parameters are given by:



**Point Estimators for vi**

As before, we’ll recognize that **v** is really a random variable **V**, given by:



where **Y** is the random variable at the top of the page. And in the multiple regression file, we found as well that:



And finally, we found a good point estimator for σ2 was:



where n is the number of data points, and f is the number of d.o.f. of our model. And, so,



So if we know know σ2 for sure, then we can say the ZVi guy,



follows a unit normal distribution. And if we don’t know it for sure (not sure how we would), then we can say,



follows a Student’s T distribution with ν = n-f d.o.f. If ν > 30 or so, then this is just a normal distribution for all intents.

**Hypothesis Testing of Parameters**

If we are running an experiment, and get some particular values, <**V**i> = vi, for our linear regression, we may wish to see whether our results invalidate someone else’s linear regression values, or the standard/accepted linear regression values, vi\*. We can do a hypothesis test for this. We know how Vi are distributed, according to our experiments, and so I guess we could form the Z-statistics,



and then calculate the p-values, the probabilities that values at least as extreme as vi\* occur. So we’d calculate,



p(z) would be the aforementioned Student’s T distribution of course.

**Confidence Intervals for Parameters**

We can also calculate confidence intervals for the statistics Vi. Just like we’ve done with other statistics, we’d say, at the 1-α confidence level,



And the zα value would be for a Student’s T distribution.

**Goodness of Fit: R2 value**

We can define a goodness of fit just as before.



where fi are the regression curve data points, which one could say is a random variable since it depends in known way (if have formula) on the random variable Yi, the data points we’re trying to fit. SSEf = sum of squares about f, and SSEm = sum of squares about mean. This is actually a general formula that applies to all curve fits. Apropos the fraction, the numerator is the ‘variation around the fit, or f’, and the denominator is the ‘variation around the mean’). Can see R2 = 1 if there is no variation around the fit, i.e., if the curve/fit fits the data exactly. On the other hand, if the curve fits the data no better than the mean, then we get R2 = 0. If you have an R2 = 0.75, then you can say that 75% of the variation of the data is explained/predicted by the regression curve f. As usual, we can interpret R2 as:



And last,

**Hypothesis Testing Different Regression Models (ANOVA)**

There are lots of different models we could test now. Presently, we are splitting our data into groups: A, B, C, …, G, and doing a separate regression on each one. But we could treat them all in aggregate, as just one group and do a single regression. Or we could keep them all in their groups and just fit a bunch of means to the data (basically the T-test in the previous file), or we could split into a different set of groups and do a different regression curve, etc. So we want a way to compare the models to see which is better. We’ll use a Hypothesis test framework for this. So say we have a model with f0 d.o.f. **v**(f0) = (v0, v1, …, vf) and another model with f > f0 d.o.f. **v**(f) = (v0, v1, v2, v3, …, vf). Now let’s define a Null Hypothesis.

H0 = assumption that the data is described by model **Y** = X**v**(f0) + ε, which has f degrees of freedom.

And let SSEf0 be the sum of the square errors for this model.Then let’s compare to another model with f> degrees of freedom, fitting **Y** = X**v**(f) + ε. And let SSEf be its sum of squared errors. We would anticipate this to be smaller of course, i.e., SSEf0 > SSEf. The alternative hypothesis would be:

HA = assumption that at least one of the extra f – f0 parameters in the new model Y = Xv(f) + ε is non-zero.

Turns out SSEf follows a known probability distribution, presuming the truth of H­­0. We can form a test statistic,



and n = number of data points. We might recall SSEf/(n-f) is just the point estimator for σ2 in the f-model. Might interpret Z\* as:



Turns out this follows an F-distribution (B is the β function).



which is the probability density of getting an Z-value of x, given the null hypothesis is true. So we can calculate a p-value,



which would be the probability that we’d get an Z-statistic Z\* or higher, out of the new model, if the Null hypothesis were true. So if the f model has true explanatory power, then we should find Z\* >> 0 and the p-value should be small (less than 0.05 at the 95% significance level). But note for instance that if f = n, i.e., if the number of fit parameters equals the number of data points (allowing an exact fit of f to our data points), then Z\* = 0, and so our p-value would be 1 I think. And this would mean that our extra parameters are meaningless.