Lecture 10: F-Tests, ANOVA and R^2

1 ANOVA

We saw that we could test the null hypothesis that $\beta_1 = 0$ using the statistic $(\widehat{\beta}_1 - 0)/\widehat{\text{se}}$. (Although I also mentioned that confidence intervals are generally more important than testing). There is another approach called Analysis of Variance (ANOVA). It's out-dated but it is in the book and you should know how it works.

The idea is to compare two models:

$$Y = \beta_0 + \epsilon \quad \text{versus} \quad Y = \beta_0 + \beta_1 X + \epsilon.$$

If we fit the first model, the least squares estimator is $\widehat{\beta}_0 = \overline{Y}$. The idea is not to create a statistic that measures how much better the second model is than the first model. The residual sums iof squares (RSS) is thus $\sum_i (Y_i - \overline{Y})^2$. This is called the total sums of squares and is denoted by $\text{SS}_{\text{total}} = \sum_i (Y_i - \overline{Y})^2$. If we fit thw second model (the usual linear model) we get a smaller residual sums of squares, $\text{RSS} = \sum_i e_i^2$.

The difference $\text{SS}_{\text{total}} - \text{RSS}$ is called the sums of squares due to regression and is denoted by SS_{reg}. If $\beta_1 = 0$ we expect this to be small. In the olden days, poeple summarized this in a ANOVA table lioke this:

Source	df	SS	MS	F	p-value
Regression	1	SS_{reg}	$\text{MS}_{\text{reg}} = \frac{\text{SS}_{\text{reg}}}{1}$	$F = \frac{\text{MS}_{\text{reg}}}{\text{MS}_{\text{res}}}$	
Residual	n-2	RSS	$\widehat{\sigma}^2 = \frac{\text{RSS}}{n-2}$		
Total	n-1	SS_{total}			

The degrees of freedom (df) are just numbers that are defined, frankly, to make things work right. The mean squared errors (MS) are the sums of squares divided by the df. The F test is

$$F = \frac{\text{MS}_{\text{reg}}}{\text{MS}_{\text{res}}}.$$

Under H_0, the statistic has a known distribution called the F distribution. This distribution depends on two parameters (just as the χ^2 distribution depends on one parameter). These are called the degrees of freedom for the F distribution. We denote the distribution by $F_{1,n-2}$. The p-value is

$$P(F > F_{\text{observed}})$$

where $F \sim F_{1,n-2}$ and F_{observed} is the actual observed value you comoute from the data.

This is equivalent to using our previous test and squaring it.

A little more formally, an F random variable is defined by

$$\frac{\chi_a^2/a}{\chi_b^2/b}$$

when χ_a^2 and χ_b^2 are independent.

Since χ^2 distributions arise from sums of Gaussians, F-distributed random variables tend to arise when we are dealing with ratios of sums of Gaussians. The MS terms in our table are independent χ^2 random variables under the usual assumptions.

2 ANOVA in R

The easiest way to do this in R is to use the `anova` function. This will give you an analysis-of-variance table for the model. The actual object the function returns is an `anova` object, which is a special type of data frame. The columns record, respectively, degrees of freedom, sums of squares, mean squares, the actual F statistic, and the p value of the F statistic. What we'll care about will be the first row of this table, which will give us the test information for the slope on X.

Let's do an example:

```
library(gamair)
out = lm(death ~ tmpd,data=chicago)
anova(out)
```

The output looks like this:

```
## Analysis of Variance Table
##
## Response: death
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## tmpd          1  162473  162473  803.07 < 2.2e-16 ***
## Residuals 5112 1034236     202
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Assumptions In deriving the F distribution, it is absolutely vital that all of the assumptions of the Gaussian-noise simple linear regression model hold: the true model must be linear, the noise must be Gaussian, the noise variance must be constant, the noise must be independent of X and independent across measurements. The *only* hypothesis being tested is whether, maintaining all these assumptions, we must reject the flat model $\beta_1 = 0$ in favor of a line at an angle. In particular, the test never doubts that the right model is a straight line.

ANOVA is an historical relic. In serious applied work from the modern (say, post-1985) era, I have never seen any study where filling out an ANOVA table for a regression, etc., was at all important.

3 What the F Test Really Tests

The textbook (§2.7–2.8) goes into great detail about an F test for whether the simple linear regression model "explains" (really, predicts) a "significant" amount of the variance in the response. What this really does is compare two versions of the simple linear regression model. The null hypothesis is that all of the assumptions of that model hold, *and* the slope, β_1, is exactly 0. (This is sometimes called the "intercept-only" model, for obvious reasons.) The alternative is that all of the simple linear regression assumptions hold with $\beta_1 \in \mathbb{R}$. The alternative, non-zero-slope model will always fit the data better than the null, intercept-only model; the F test asks if the improvement in fit is larger than we'd expect under the null.

There are situations where it is useful to know about this precise quantity, and so run an F test on the regression. It is hardly ever, however, a good way to check whether the simple linear

regression model is correctly specified, because neither retaining nor rejecting the null gives us information about what we really want to know.

Suppose first that we retain the null hypothesis, i.e., we do not find any significant share of variance associated with the regression. This could be because (i) the intercept-only model is right; (iii) $\beta_1 \neq 0$ but the test doesn't have enough power to detect departures from the null. We don't know which it is. There is also possibility that the real relationship is nonlinear, but the best linear approximation to it has slope (nearly) zero, in which case the F test will have no power to detect the nonlinearity.

Suppose instead that we reject the null, intercept-only hypothesis. *This does not mean that the simple linear model is right.* It means that the latter model predicts better than the intercept-only model — too much better to be due to chance. The simple linear regression model can be absolute garbage, with every single one of its assumptions flagrantly violated, and yet better than the model which makes all those assumptions *and* thinks the optimal slope is zero.

Neither the F test of $\beta_1 = 0$ vs. $\beta_1 \neq 0$ nor the Wald/t test of the same hypothesis tell us *anything* about the correctness of the simple linear regression model. All these tests *presume* the simple linear regression model with Gaussian noise is true, and check a special case (flat line) against the general one (titled line). They do not test linearity, constant variance, lack of correlation, or Gaussianity.

4 R^2

Another quantity that gets mentioned a lot in regression (which is also a historical relic) is R^2. It is defined by

$$R^2 = \frac{\mathrm{SS}_{\mathrm{reg}}}{\mathrm{SS}_{\mathrm{total}}}.$$

It is often described as "the fraction of variability explained by the regression." It can be shown that it can be written as $R^2 = r^2$ where

$$r = \frac{\widehat{\mathrm{Cov}(\mathrm{X}, \mathrm{Y})}}{s_X s_Y}$$

in other words, the correlation coefficient squared.

R^2 will be 0 when $\widehat{\beta}_1 = 0$. On the other hand, if all the residuals are zero, then $R^2 = 1$. It is not too hard to show that R^2 can't possible be bigger than 1, so we have marked out the limits: a sample slope of 0 gives an R^2 of 0, the lowest possbile, and all the data points falling exactly on a straight line gives an R^2 of 1, the largest possible.

What does R^2 converge to as $n \to \infty$. The population version is

$$\begin{aligned} R^2 &= \frac{\mathrm{Var}\left[m(X)\right]}{\mathrm{Var}\left[Y\right]} && (1)\\ &= \frac{\mathrm{Var}\left[\beta_0 + \beta_1 X\right]}{\mathrm{Var}\left[\beta_0 + \beta_1 X + \epsilon\right]} && (2)\\ &= \frac{\mathrm{Var}\left[\beta_1 X\right]}{\mathrm{Var}\left[\beta_1 X + \epsilon\right]} && (3)\\ &= \frac{\beta_1^2 \mathrm{Var}\left[X\right]}{\beta_1^2 \mathrm{Var}\left[X\right] + \sigma^2} && (4) \end{aligned}$$

Since all our parameter estimates are consistent, and this formula is continuous in all the parameters, the R^2 we get from our estimate will converge on this limit.

Unfortunately, a lot of myths about R^2 have become endemic in the scientific community, and it is vital at this point to immunize you against them.

1. The most fundamental is that R^2 *does not measure goodness of fit.*
   (a) R^2 *can be arbitrarily low when the model is completely correct.* Look at Eq. 4. By making $\mathrm{Var}[X]$ small, or σ^2 large, we drive R^2 towards 0, even when every assumption of the simple linear regression model is correct in every particular.
   (b) R^2 *can be arbitrarily close to 1 when the model is totally wrong.* There is, indeed, no limit to how high R^2 can get when the true model is nonlinear. All that's needed is for the slope of the best linear approximation to be non-zero, and for $\mathrm{Var}[X]$ to get big.
2. R^2 is also pretty useless as a measure of predictability.
   (a) R^2 *says nothing about prediction error.* R^2 can be anywhere between 0 and 1 just by changing the range of X. Mean squared error is a *much* better measure of how good predictions are; better yet are estimates of out-of-sample error which we'll cover later in the course.
   (b) R^2 *says nothing about interval forecasts.* In particular, it gives us no idea how big prediction intervals, or confidence intervals for $m(x)$, might be.
3. R^2 cannot be compared across data sets.
4. R^2 cannot be compared between a model with untransformed Y and one with transformed Y, or between different transformations of Y.
5. The one situation where R^2 can be compared is when different models are fit to the same data set with the same, untransformed response variable. Then increasing R^2 is the same as decreasing in-sample MSE (by Eq. **??**). In that case, however, you might as well just compare the MSEs.
6. It is very common to say that R^2 is "the fraction of variance explained" by the regression. But it is also extremely easy to devise situations where R^2 is high even though neither one could possible explain the other.

At this point, you might be wondering just what R^2 is good for — what job it does that isn't better done by other tools. The only honest answer I can give you is that I have never found a situation where it helped at all. If I could design the regression curriculum from scratch, I would never mention it. Unfortunately, it lives on as a historical relic, so you need to know what it is, and what mis-understandings about it people suffer from.

5 The Correlation Coefficient

As you know, the correlation coefficient between X and Y is

$$\rho_{XY} = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}}$$

which lies between -1 and 1. It takes its extreme values when Y is a linear function of X.

Recall, from lecture 1, that the slope of the ideal linear predictor β_1 is

$$\frac{\mathrm{Cov}\left[X, Y\right]}{\mathrm{Var}\left[X\right]}$$

so

$$\rho_{XY} = \beta_1 \sqrt{\frac{\mathrm{Var}\left[X\right]}{\mathrm{Var}\left[Y\right]}}.$$

As we saw, R^2 is just $\widehat{\rho}^2_{XY}$.

6 Concluding Comment

The tone I have taken when discussing F tests, R^2 and correlation has been dismissive. This is deliberate, because they are grossly abused and over-used in current practice, especially by non-statisticians, and I want you to be too proud (or too ashamed) to engage in those abuses. In a better world, we'd just skip over them, but you will have to deal with colleagues, and bosses, who learned their statistics in the bad old days, and so have to understand what they're doing wrong.

In all fairness, the people who *came up* with these tools were great scientists, struggling with very hard problems when nothing was clear; they were inventing all the tools and concepts we take for granted in a class like this. Anyone in this class, me included, would be doing very well to come up with *one* idea over the whole of our careers which is as good as R^2. But we best respect our ancestors, and the tradition they left us, when we improve that tradition where we can. Sometimes that means throwing out the broken bits.