Subject: Re: [36-617] Degrees of Freedom
From: Brian Junker <bj20@andrew.cmu.edu>
Date: 9/21/2018, 1:20 PM
To: Alan Mishler <amishler@andrew.cmu.edu>, Jun Hee Kim <junheek1@andrew.cmu.edu>
BCC: bj20@andrew.cmu.edu

Alan's response is great. Here's another way of looking at it, perhaps less satisfying but "correct". It is similar to Alan's last paragraph.

Linear models are all about the equation y-hat = H* y, where y is an n-dimensional data vector, H* is a projection matrix onto some lower-dimensional space; let's call that space L. Then y-hat is the projection of y in L. In case H* is symmetric and idempotent, we know from linear algebra that dim(L) = rank(H*) = tr(H*). In fact, df is just theoretical statistics' name for dim(L): df=dim(L).

So, for a well behaved smoother S, H* = S, y-hat = S y, df = tr(S). This is also true for the standard linear model y = Xb + eps, where now H* = H, our usual hat matrix; in this case, y-hat = H y, and df=tr(H)=p+1. Same result as for the general linear smoother.

The sums of squares in an anova table are based on different projections into different-dimensional subspaces. Each sum of squares is just the squared length of the projection of y into each subspace. The cool thing from a statistics pov is that when y is normally distributed, this squared length is distributed as chi-squared rv's with df equal to the dimension of the subspace.

  SST is based on the projection y-hat(1) = (I-H1) y, where H1 is the hat matrix from the intercept only model y = b0 + eps onto a subspace L(1).

  SSreg is based on the projection y-hat(2) = (H-H1) y onto a subspace L(2).

  RSS is based on the projection y-hat(3) = (I - H) y onto a subspace L(3).

it's easy to see that each of (I-H1), (H-H1) and (I-H) are symmetric, idempotent projection matrices (as long as X has a column of ones!), and hence

```
  dim(L(1)) = tr(I-H1) = n-1 = df(SST)
  dim(L(2)) = tr(H-H1) = (p+1)-1 = p = df(SSreg)
  dim(L(3)) = tr(I-H)  = n - p - 1 = df(RSS)
```

So SSreg has different df from the full model (p vs p+1) since it is based on projection into a different, smaller, subspace: the one defined by H-H1, rather than the one defined by H.

If you're interested in a summary of the theory of applying Cochran's theorem in order to get that the SS's actually are independent chi-squared rv's with the df listed above, see for example http://www.stat.columbia.edu/~fwood/Teaching/w4315/Fall2009/lecture_cochran.pdf

hope this helps!

-BJ

On 9/20/2018 11:05 PM, Alan Mishler wrote:

> Hi Jun Hee,
>
> The first line of the ANOVA table here is the degrees of freedom of SSReg, not of the model as a whole. When you're estimating the model, you're estimating a parameter vector of length p + 1, so

> you know your estimate has to lie in a (p + 1)-dimensional subspace. A priori, before you've chosen an estimation method, you could pick any vector in that space (though of course that would be silly). That's the sense (or at least a sense) in which the model degrees of freedom are p + 1.
>
> When it comes to SSreg, the "Source of variation" is the model, but SSreg isn't quite the same as the model itself. SSreg = $\sum_{i=1}^n (\hat{y}_i - \bar{y})$ which is a function of the p + 1 parameters, but also a function of $\bar{y}$. The estimation procedure we've chosen enforces that $(\hat{y}_i - \bar{y})$ has to sum to 0 (since the residuals have to sum to 0), so in a sense, we've lost a degree of freedom by insisting on this condition.
>
> That's probably not a totally satisfying explanation; maybe Professor Junker has a better way of looking at it.
>
> For myself, there are cases like the first where I find it helpful to think of degrees of freedom in terms of subspaces, and there are cases like the second where I simply think of them as parameters of distributions that let us test hypotheses. As in, it's a nice convenient fact that we have this statistic F on the right side of the ANOVA table that lets us test a null hypothesis that we're interested in; and I'm content to call p and $n - p - 1$ the "degrees of freedom" of the distribution even if I can't easily think about them in linear algebraic terms.
>
> I hope that helps,
> Alan
>
> On Wed, Sep 19, 2018 at 6:00 PM Jun Hee Kim <junheek1@andrew.cmu.edu> wrote:
>
> > Dear Professor Junker and Alan:
> >
> > Hello! This is Jun Hee, one of the MSP students.
> >
> > I have a question about degrees of freedom (df) in linear regression. (I asked Professor Junker about this at the end of this Monday class, but we didn't have much time to discuss it.)
> >
> > Let ***p*** denote the number of predictors (so there are total ***p+1*** parameters including the intercept parameter). I remember in 36-402 that for any linear smoother (including linear regression), the df of the model is the trace of the weight matrix, which is ***p+1*** in linear regression. But in the ANOVA table (attached), it says the df for the regression is ***p***.
> >
> > Why/how exactly do these two quantities differ so that one is ***p+1*** and the other is ***p***?
> >
> > Thank you!
> >
> > Sincerely,
> > Jun Hee Kim
>
> --
> Alan Mishler
> PhD Student
> Department of Statistics and Data Science
> Carnegie Mellon University (Pittsburgh)
> www.linkedin.com/in/alanmishler

```
--
Brian Junker                    (412) 268 - 2718
Department of Statistics        brian@stat.cmu.edu
232 Baker Hall                  FAX: (412) CMU-STAT
Carnegie Mellon University        or (412) 268-7828
Pittsburgh PA 15213 USA

WWW: http://www.stat.cmu.edu/~brian/
```