窃以为从**回归分析**的角度能够得到比较直观的解释——投影矩阵的**秩**。作为例子，样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ 可以视为 $(X_1, \ldots, X_n)'$ 对 $1_n = (1, \cdots, 1)'$ 回归得到的残差 (residual) 平方和除以自由度(n-1)。

考虑标准线性模型 $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$，其中 $\mathbf{Y} = (y_1, \ldots, y_n)'$，$\mathbf{X} = (x_{ik}) \in \mathbb{R}^{n\times p}$，$\boldsymbol{\beta} = (\beta_k) \in \mathbb{R}^p$，$\boldsymbol{\varepsilon} = (\varepsilon_i) \sim \mathcal{N}(0_n, \sigma^2 I_n)$。最小二乘法得到**投影** $\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{Y}} = \mathrm{proj}_{\mathbf{range}(\mathbf{X})}(\mathbf{Y}) = \arg\min_{\boldsymbol{\theta}\in\mathbf{range}(\mathbf{X})} \|\mathbf{Y} - \boldsymbol{\theta}\|_2$，适合正规方程 $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$；无妨设 $\mathbf{rank}(\mathbf{X}) = p$，则 $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$。

帽子矩阵 (hat matrix) $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 是 $\mathbb{R}^n$ 到 $\mathbf{range}(\mathbf{X}) \cong \mathbb{R}^p$ 的**正交投影**，作为幂等矩阵有 $\mathbf{rank}(\mathbf{H}) = \mathbf{tr}(\mathbf{H}) = \mathbf{tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) = \mathbf{tr}(I_p) = p$，于是**预测值** $\left\{\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} : \mathbf{Y} \in \mathbb{R}^n\right\}$ 对应的自由度是 $\mathbf{rank}(\mathbf{H}) = p$。注意 $I_n - \mathbf{H}$ 是 $\mathbb{R}^n$ 到 $\mathbf{range}(\mathbf{X})^\perp$ 的**正交投影**，作为幂等矩阵有 $\mathbf{rank}(I_n - \mathbf{H}) = n - p$，于是**残差** $\left\{\mathbf{Y} - \hat{\mathbf{Y}} = (I_n - \mathbf{H})\mathbf{Y} : \mathbf{Y} \in \mathbb{R}^n\right\}$ 对应的自由度是 $\mathbf{rank}(I_n - \mathbf{H}) = n - p$。

平方和之所以好用，是因为本质上反映了 (Frobenius/Hilbert-Schmidt) 内积结构 $\langle -, - \rangle : (A, B) \in \mathbb{R}^{r\times s} \times \mathbb{R}^{r\times s} \mapsto \mathbf{tr}(A'B) \in \mathbb{R}$，注意向量是矩阵的特殊形式——只有一列。残差平方和的期望为

$$\begin{aligned}
\mathbb{E}\langle \mathbf{Y} - \hat{\mathbf{Y}}, \mathbf{Y} - \hat{\mathbf{Y}} \rangle &= \mathbb{E}\langle (I_n - \mathbf{H})\boldsymbol{\varepsilon}, (I_n - \mathbf{H})\boldsymbol{\varepsilon} \rangle \\
&= \mathbb{E}\,\mathbf{tr}(\boldsymbol{\varepsilon}'(I_n - \mathbf{H})\boldsymbol{\varepsilon}) \\
&= \mathbb{E}\,\mathbf{tr}((I_n - \mathbf{H})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\
&= \mathbf{tr}((I_n - \mathbf{H})\,\mathbf{Var}(\boldsymbol{\varepsilon})) \\
&= \mathbf{tr}((I_n - \mathbf{H})\sigma^2 I_n) \\
&= (n - p)\sigma^2.
\end{aligned}$$

所以 σ^2 的一个**无偏估计**就是残差平方和除以自由度。

推广到非参数回归需要一定的修饰, 有兴趣的话可以参看

https://www.stat.cmu.edu/~ryantibs/advmethods/notes/df.pdfwww.stat.cmu.edu

---

首先，最严格、最不会产生歧义的定义，就是在卡方分布 $f(x) = \frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}$ 中，定义参数 n 为自由度。但是这种定义完全无法体现自由度的内在概念，我们最多就知道它是 n 个正态随机变量的平方和。我想大多数人都是在学习后继课程的时候才慢慢明白自由度的统计意义的。

第二种方法即为以朴素的限制个数来定义自由度，这也是自由度的雏形，它可以追溯到高斯的时代 - 1821 年。但其早期的定义是由 Gosset 给出，就是 1908 年以'student'署名的、提出 t 分布的那篇发在*生物测量学*期刊的论文 [1]。但是这篇文章中并未提出自由度（degree of freedom）这个名字。（以上来自维基百科 [2]）

'自由度'这个名称的普及，应归功于生物统计学家 Fisher 在 1922 年阐述卡方检验的论文 [3]。在这篇论文中，Fisher 提到：由于在中间过程中，我们用了四个均值，因此自由度降低了四。这个较为初级的定义，最终被扩充为：样本容量减去限制等式的个数。用高级点的语言，就是线性子空间的维数 [4]。

自由度的第三种定义是二次型的秩。这种定义的最初来源是 Cramer 在其 1946 年的著作 Mathematical Methods of Statistics[5] 中提到的（P381）：

> The number $n-p$ is the *rank* of the form Q (cf 11.6), i.e. the smallest number of independent variables on which the form may be brought by a non-singular linear transformation. In statistical applications, this number of free variables entering into a problem is usually, in accordance with the terminology introduced by R. A. Fisher, denoted as the number of *degrees of freedom* (abbreviated *d. of fr.*) of the problem, or of the distribution of the random variables attached to the problem.

很明显，用矩阵的秩定义自由度，相比子空间维数，更偏重代数一些。但还不止于此，其更深刻的意义在于检验。为此，首先介绍 Cochran 定理 [6]（这个 version 相对简单）：

设 $Y \sim N(0, \sigma^2 I_n)$，矩阵 A 是幂等阵，$Y^T A Y = \sum_{i=1}^{k} Y^T A_i Y$，且 A_i 均为对称幂等阵。则有：

$Y^T A_i Y / \sigma^2$ 是相互独立的卡方分布，自由度为 $rank(A_i)$。

$rank(A) = \sum_{i=1}^{k} rank(A_i)$。

我们看到，在这个定理中，二次型的秩被证明为与自由度相同。这或许也是 Cramer 秩定义的灵感来源。

我们知道，F 分布定义为 [卡方分布 / 自由度] 的比值，因此在已知卡方统计量和自由度的情况下，可以直接得到 F 统计量。因此，一旦方差（可写成二次型的形式）可以写成如上的分解式，我们就可以直接做 F 检验了。

例如，设参数个数为 p（不算截距项），有线性模型 $y = \mu + X_0 \beta_0 + \epsilon$，或 $y = X\beta + \epsilon$，其中 $X = [1; X_0]$，$\beta^T = [\mu; \beta_0^T]$，误差项为独立同分布的正态项，此时最小二乘或极大似然估计为：$\hat{\beta} = (X^T X)^{-1} X^T y$（截距项包含在里边了）。

则有 $\hat{y} = X\beta = X(X^T X)^{-1} X^T y := Hy$。可得到残差平方和 $SSR = (y - \hat{y})^T (y - \hat{y}) = y^T (I_n - H) y$。

如果原假设是 $H_0 : \beta_0 = 0$，则在原假设之下，记 $z = (y - \mu) \sim N(0, \sigma^2 I_n)$。有拆分：

$$z^T I_n z = \bar{z}^2 + (y-\mu)^T (I_n - H)(y-\mu) + [(y-\mu)^T H (y-\mu) + z^T I_n z - \bar{z}^2]$$

通过正则方程组及均值的矩阵表达式，上式可化简为：

$$z^T I_n z = z^T \frac{11^T}{n} z + z^T (I_n - H) z + z^T (H - \frac{11^T}{n}) z$$
$$= z^T \frac{11^T}{n} z + y^T (I_n - H) y + y^T (H - \frac{11^T}{n}) y$$

此时，按照矩阵理论，两边的秩分别为 $n = 1 + (n-p-1) + p$，且易证每个二次型都是幂等阵。后边的两项，分别是 SSE 和 SSR。则按照 Cochran 定理，可直接由二次型和秩进行 F 检验。

我们知道幂等阵的特征值只能是 0 或 1，而二次型经变换后可以换成特征值与特征向量结合的形式。此时，秩与自由度便产生了一一对应的关系：秩等于特征值中'1'的个数。而 F 检验及卡方检验也可释义为，每一个自由度，或每一个特征值'1'，给予二次型的平均贡献。

最后，说一下非整数自由度。按照以上的定义方式，第二种定义 - 子空间维数则必为整数，第一种定义并不局限于整数自由度，而第三种定义可以拓展到非整数自由度：幂等阵中特征值中'1'的个数可以等价定义为特征值的和，由矩阵论可知即为二次型的迹，而迹可以是非整数的。

1. Welch 两样本 t 检验中，可以出现非整数的自由度：

```
> t.test(c(1,2,3,4,5),c(2,3,4,5,18))

        Welch Two Sample t-test

data:  c(1, 2, 3, 4, 5) and c(2, 3, 4, 5, 18)
t = -1.1234, df = 4.4604, p-value = 0.3181
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.471813   4.671813
sample estimates:
mean of x mean of y
      3.0       6.4
```


如图，这里的自由度是 4.4604。

大家可能想不到，这里的非整数自由度是以第一种方法定义的，即卡方分布的参数。Welch 的原始论文 [7] 中，他是以分布函数 + Taylor 展开推导出来这个自由度近似公式。

2. 岭回归（Ridge Regression）。

起初为了应对共线性的问题，Tikhonov 提出了以下正则化的线性回归参数估计式：$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$。这时，通过第三种定义，我们仍然能获得二次型的自由度 $trace(H) = trace(X(X^T X + \lambda I)^{-1} X^T y)$。这时，在模型间的比较中，我们可以将该迹替代参数个数 p，代入信息准则 AIC 或 BIC 的计算公式中。

但需要注明的是，虽然整数自由度的三种定义是等价的，非整数自由度却并不是等价的，而仅是近似关系。例如 $(y_1, y_2) \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = 0.5(y_1^2 + y_2^2)$，按矩阵迹，自由度应为 1。但 $(y_1^2 + y_2^2)$ 满足自由度为 2 的卡方分布，计算可知其实际上是指数分布，而不是自由度为 1 的卡方分布。

参考

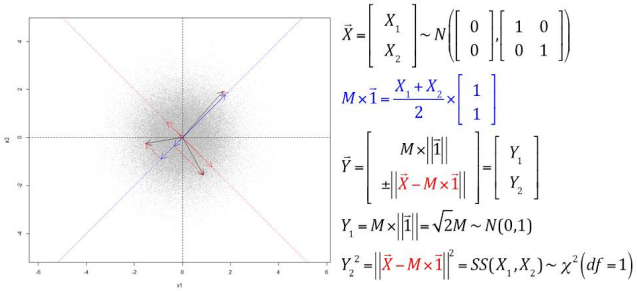
1. ^ Student. (1908). The Probable Error of a Mean. Biometrika. 6 (1): 1–25. doi:10.2307/2331554
2. ^https://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics)#cite_note-5

3. ^Fisher, R. A. (1922). On the Interpretation of χ2 from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society. 85 (1): 87–94. doi:10.2307/2340521
4. ^Walker, H. W. (1940). Degrees of freedom. Journal of Educational Psychology, 31, 253-269.
5. ^Cramer, H. (1946). Mathematical Methods of Statistics. Princeton Univ. Press
6. ^ Cochran, W. G. (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. Mathematical Proceedings of the Cambridge Philosophical Society. 30 (2): 178–191. doi:10.1017/S0305004100016595
7. ^Welch, B. L. (1947). The generalization of "student's" problem when several different population variances are involved. Biometrika, 34: 28–35, doi:10.2307/2332510

课堂上的解释：自由度（*df*）这个词更好的翻译是 自由维度，否则容易误解为自由程度。下面试从*n* = 2 的情形给出平面可视化的解释，再推及 *n* = 3 的空间可脑补化解释，最后推广到一般的子空间正交分解情形。

在下图中，样本量为 2 的标准正态分布理论情形，横轴为X_1，纵轴为X_2。$M=(X_1+X_2)/2$，把图旋转 45 度，新横轴为蓝色虚线Y_1，新纵轴为红色虚线Y_2。可以理解$Y_1=\sqrt{2}M$为啥服从标准正态分布，$Y_2^2=\sum_i (X_i-M)^2=SS(X_1,X_2)$为啥服从$\chi^2_{df=1}$分布。自由度为 1——因为它只在红色虚线方向上（只在一个维度上）随机波动。

Dev (& *SS*) 与 *M* 的正交关系
n=2的简化情形（代码详备注）



$$\vec{X}=\begin{bmatrix}X_1\\X_2\end{bmatrix}\sim N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$$

$$M\times\vec{1}=\frac{X_1+X_2}{2}\times\begin{bmatrix}1\\1\end{bmatrix}$$

$$\vec{Y}=\begin{bmatrix}M\times\left\|\vec{1}\right\|\\\pm\left\|\vec{X}-M\times\vec{1}\right\|\end{bmatrix}=\begin{bmatrix}Y_1\\Y_2\end{bmatrix}$$

$$Y_1=M\times\left\|\vec{1}\right\|=\sqrt{2}M\sim N(0,1)$$

$$Y_2^2=\left\|\vec{X}-M\times\vec{1}\right\|^2=SS(X_1,X_2)\sim\chi^2(df=1)$$

```r
set.seed(1997)
x1 <- rnorm(10^5);
x2 <- rnorm(10^5);

plot(x1,x2,pch='.',asp=1,col='grey')
abline(v=0,lty=2);
abline(h=0,lty=2);
abline(c(0,1),lty=3,col='blue');
abline(c(0,-1),lty=3,col='red');

#以下代码每次运行代表抽一个个案
(i = sample(1:10^5,1))
arrows(0,0,x1[i],x2[i])
xm <- mean(c(x1[i],x2[i]))
arrows(xm,xm,x1[i],x2[i],col='red')
arrows(0,0,x1[i]-xm,x2[i]-xm,col='red')
arrows(0,0,xm,xm,col='blue')
```

如果理解了上面这幅图（不需要理解生成图的 *R* 代码），你可能会发现关键在于两个独立的正态分布相乘得到的二元正态分布可以随便旋转而不改变密度。这是正态分布本质的特性，其它分布是不行的。（其实，从正圆密度等高线甚至还可以反推出正态分布的密度表达式: Developing normal pdf from symmetry & independence）。

看懂上面那幅图，接着可以看下面这张图，样本量从 2 变到 3，μ还是 0、σ还是 1，新坐标架的第一个轴是旧坐标架单位向量所指向的方向。SS是自由分布在二维（=*N*-1）日晷盘面红向量的长度平方。看懂之后，再让μ和σ等于一般情形的取值。再看懂之后，让N推广到具身认知无法体验的更高维空间。

X 向量在任意新直角坐标架的三维投影
分量Y_1, Y_2, Y_3服从三维联合正态分布

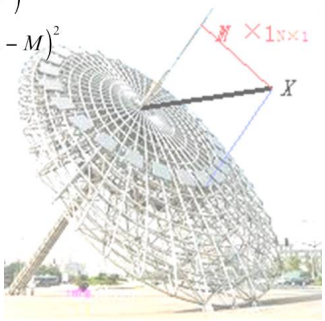
$$\sqrt{N}\left(M-\mu\right)=Y_1\sim N\left(0,\sigma^2\right)$$

$$SS=\left(X_1-M\right)^2+\cdots+\left(X_N-M\right)^2$$

$$=\left\|\vec{X}_{N\times1}-M\times\vec{1}_{N\times1}\right\|^2$$

$$=Y_2^{\ 2}+Y_3^{\ 2}+\cdots+Y_N^{\ 2}$$

$$\frac{SS}{\sigma^2}\sim\chi^2\left(df=N-1\right)$$



以上这两幅图的例子都是回归方程只有截距项参数的特例情形。随着回归方程模型自变量个数增加到 *p*，参数个数（拟合值的自由维度）增加到 1 + *p* 。用 观测值 = 模型拟合值 + 残差 的正交分解，*N* 维度的观测值投影在 (1+*p*) 和 (*N*-1-*p*) 两个正交的子空间。后者的维度就是残差的自由维度。文献教材所谓模型的自由度，大部分情形都是指「模型的残差自由维度」，极少数情形是指「模型的拟合值空间自由维度」