ORIGIN ≡ 0

General F-test for Model Comparison

The General Linear Models (GLM) F-Test using formal comparison between "full" and "reduced" statistical models *that internest* ask whether fewer variables of a reduced model (RM) are sufficient to describe the dependent variable Y in a dataset, or whether a more complex full model (FM) is instead required. A FM may always be identified by having smaller Sum of Squares Error than than the corresponding RM. One can also compare the total number of parameters estimated for each model. A FM always has more estimated parameters than the corresponding RM. The GLM F-test involves H_0 that sets one or more coefficients of the FM to zero. Unlike marginal tests that must treat one independent variable at a time, this procedure allows any comparison between models as long as one is a subset of the other. Worked example is drawn from Ch. 9 in Kuter et al. (KNNL) A*pplied Linear Statistical Models* 5th Edition.

Surgical Unit Example KNNL Table 9.1

	X1	X2	X3	X4	X5	X6	Y
1	6.7	62	81	2.59	50	0	6.544
2	5.1	59	66	1.70	39	0	5.999
3	7.4	57	83	2.16	55	0	6.565
4	6.5	73	41	2.01	48	0	5.854
5	7.8	65	115	4.30	45	0	7.759
6	5.8	38	72	1.42	65	1	5.852
7	5.7	46	63	1.91	49	1	6.250
8	3.7	68	81	2.57	69	1	6.619
9	6.0	67	93	2.50	58	0	6.962
10	3.7	76	94	2.40	48	0	6.875
11	6.3	84	83	4.13	37	0	6.613
12	6.7	51	43	1.86	57	0	5.549
13	5.8	96	114	3.95	63	1	7.361
14	5.8	83	88	3.95	52	1	6.754
15	7.7	62	67	3.40	58	0	6.554
16	7.4	74	68	2.40	64	1	6.695
17	6.0	85	28	2.98	36	1	6.526
18	3.7	51	41	1.55	39	0	5.321
19	7.3	68	74	3.56	59	1	6.309
20	5.6	57	87	3.02	63	0	6.731
21	5.2	52	76	2.85	39	0	5.883
22	3.4	83	53	1.12	67	1	5.866
23	6.7	26	68	2.10	30	0	6.395
24	5.8	67	86	3.40	49	1	6.332
25	6.3	59	100	2.95	36	1	6.478
26	5.8	61	73	3.50	62	1	6.621
27	5.2	52	86	2.45	70	0	6.302
28	11.2	76	90	5.59	58	1	7.583
29	5.2	54	56	2.71	44	1	6.167
30	5.8	76	59	2.58	61	1	6.396
31	3.2	64	65	0.74	53	0	6.094
32	8.7	45	23	2.52	68	0	5.198
33	5.0	59	73	3.50	57	0	6.019
34	5.8	72	93	3.30	39	1	6.944
35	5.4	58	70	2.64	31	1	6.179
36	5.3	51	99	2.60	48	0	6.453
37	2.6	74	86	2.05	45	0	6.519
38	4.3	8	119	2.85	65	1	5.893
39	4.8	61	76	2.45	51	1	6.457
40	5.4	52	88	1.81	40	1	6.558
41	5.2	49	72	1.84	46	0	6.283
42	3.6	28	99	1.30	55	0	6.366
43	8.8	86	88	6.40	30	1	7.147
44	6.5	56	77	2.85	41	0	6.288
45	3.4	77	93	1.48	69	0	6.178
46	6.5	40	84	3.00	54	1	6.416
47	4.5	73	106	3.05	47	1	6.867
48	4.8	86	101	4.10	35	1	7.170
49	5.1	67	77	2.86	66	1	6.365
50	3.9	82	103	4.55	50	0	6.983
51	6.6	77	46	1.95	50	0	6.005
52	6.4	85	40	1.21	58	0	6.361
53	6.4	59	85	2.33	63	0	6.310
54	8.8	78	72	3.20	56	0	6.478

Example in R:

```
#GLM F-TEST OF FM VS RM
K=read.table("c"/DATA/Biostatistics/KNNLCh9SurgicalUnit.txt")
K
attach(K)
options(digits=6)
#FITTING THE FULL LINEAR MODEL
FM=lm(Y~X1+X2+X3+X4+X5+factor(X6))
#FITTING A REDUCED LINEAR MODEL
RM=lm(Y~X1+X2+X3+X5)
#COMPARING MODELS
anova(FM)
anova(RM)
```

FM: **compare RSS below:**

```
Analysis of Variance Table
Response: Y
           Df Sum Sq Mean Sq F value   Pr(>F)
X1          1  0.776   0.776  12.558 0.000904 ***
X2          1  2.589   2.589  41.880 5.19e-08 ***
X3          1  6.334   6.334 102.470 2.16e-13 ***
X4          1  0.025   0.025   0.398 0.531382
X5          1  0.126   0.126   2.046 0.159218
factor(X6)  1  0.052   0.052   0.845 0.362735
Residuals  47  2.905   0.062
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

RM:

```
Analysis of Variance Table
Response: Y
          Df  Sum Sq Mean Sq F value   Pr(>F)
X1         1  0.776   0.776  12.849 0.000776 ***
X2         1  2.589   2.589  42.853 3.35e-08 ***
X3         1  6.334   6.334 104.850 9.12e-14 ***
X5         1  0.148   0.148   2.456 0.123503
Residuals 49  2.960   0.060
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GLM F-Test:

Assumptions:

- Standard Linear Regression depends on specifying in advance which variable is to be considered 'dependent' and which 'independent'. This decision matters as changing roles for Y & X usually produces a different result.

- Y (dependent variable) is a single vector random sample ~ $N(\mu,\sigma^2)$.
- $X_1, X_2, X_3, \ldots, X_j$ (independent variables) multiple vectors with each value of X_i matched to Y_i

Within this setup, two models for the relationship between X and Y variables are explicitly compared:

Full Model:

$$Y_i = \beta_0 + \Sigma\beta_j X_i + \varepsilon_i$$

Reduced Model:

$$Y_i = \beta_0 + \Sigma\beta_k X_i + \varepsilon_i$$

where: Y_i and $[X_1,X_2, \ldots X_j]$ are matched dependent and independent variables, and

β_0 is the y intercept of the regression line (translation)

β_j are slope coefficients for the *full set* of independent variables $X_1,X_2, \ldots X_j$

β_k are slope coefficients for a smaller *set* of independent variables *within* X_j

ε_i is the error factor in prediction of Y_i and a random variable ~$N(0,\sigma^2)$.

Hypotheses:

H_0: coefficients in j but NOT INCLUDED in k = 0

H_1: at least some of these coeficients not 0

Note that Null Hypotheses are, in general, a formal statement of parsimony (i.e., "simplicity" of explanation). The null hypothesis says that the simpler of two alternatives is to be preferred unless the data require us to reject it. In a one population t-test of mean, for example, we ask whether an observed mean value X_{bar} is statistically indistinguishable from some specified value μ_0. We normally interpret the Null Hypothesis H_0 to say "the differences we observe between X_{bar} and μ_0 are the expected result of random behavior". However, random must always be defined in light of some model of what we expect for random, such as ~$N(\mu,\sigma)$. We might more accurately claim instead that H_0 says "unless compelled to do so, prefer the simpler hypothesis about difference between X_{bar} and μ_0", namely that there is nothing more to explain about the relationship between X_{bar} and $\mu_0$than ~$N(\mu,\sigma)$. The Null Hypothesis for GLM above works exactly this way.

Degrees of Freedom:

```
#DEGREES OF FREEDOM
n=length(Y)
n
dfF = summary.lm(FM)$df[1]
dfF
dfR = summary.lm(RM)$df[1]
dfR
```

```
[1] 54
[1] 47
[1] 49
```

< n = number of matched observations in dataset

< degrees of freedom for FM

< degrees of freedom for RM

Sums of Squares:

```
#SUMS OF SQUARES ERROR
SSEF = summary(FM)$sigma^2*summary(FM)$df[2]
SSEF
SSER = summary(RM)$sigma^2*summary(RM)$df[2]
SSER
```

```
[1] 2.90527
[1] 2.96016
```

< SSE_F

< SSE_R

GLM Test Statistic:

$SSE_R := 2.96016$ $df_R := 49$

$SSE_F := 2.90527$ $df_F := 47$

$$F := \frac{\frac{SSE_R - SSE_F}{df_R - df_F}}{\frac{SSE_F}{df_F}}$$

$F = 0.444$

#GLM TEST F-STATISTIC
F=((SSER-SSEF)/(dfR-dfF))/(SSEF/dfF)
F

```
[1] 0.444
```

Critical Value of the Test:

$\alpha := 0.01$ **< Probability of Type I error must be explicitly set**

$CV := qF(1 - \alpha, df_R - df_F, df_F)$ $CV = 5.0874$ **< note degrees of freedom here**

Decision Rule:

IF F > CV, THEN REJECT H_0 OTHERWISE ACCEPT H_0

$F = 0.444$ $CV = 5.0874$

Probability Value:

$P := 1 - pF(F, df_R - df_F, df_F)$ $P = 0.6441$

IMPORTANT NOTE:

FALURE to reject H_0 in this test means that the MORE PARSIMONIOUS model RM is PREFERRED!

#ANOVA GLM F-TEST
anova(RM,FM)

```
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X3 + X5
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + factor(X6)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     49 2.960
2     47 2.905  2   0.05489 0.444  0.644
```

^ difference in degrees of freedom