Block 10

Testing and Analysis of Variance

Testing and Analysis of Variance

So far, we have seen testing of a single parameter (β_j) or a contrast. This boils down to:
$H_0 : \theta = \theta_0 \quad (\theta \equiv \text{parameter of interest})$
$H_1 : \theta \neq \theta_0$
We have used the *t-statistic*:

$$t = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})} \overset{H_0}{\sim} t_{\text{df}} \quad \text{df depends on } \hat{\sigma}^2$$

Here we drill down more into comparing the fit of mean functions rather than comparing parameter estimates.

F-tests (also called ANOVA or LR Tests)

Set up: Suppose we have: Y and $X'_{p+1} = (X'_1, X'_2)$
where: X'_1 is (p!q) regressors, X'_2 is q regressors
Consider Testing

$$H_0 : E(Y \mid X_1 = x_1, X_2 = x_2) = x'_1\beta_1$$

$$H_1 : E(Y \mid X_1 = x_1, X_2 = x_2) = x'_1\beta_1 + x'_2\beta_2$$

Different since H_0 and H_1 refer to mean function specifications, [not restrictions on the parameters]
Also require H_0 to be a part of H_1, [ie. H_0 is obtained by setting $\beta_2 = 0$]
Idea:
- Recall that **RSS** (Residual Sum of Squares) measures the amount of variation in the response not explained by the regressors.
$\Rightarrow$ If H_0 is false, then $RSS_{H_1} < RSS_{H_0}$

$\Rightarrow$ Test Statistic

$$F = \frac{\frac{RSS_{H_0} - RSS_{H_1}}{df_{H_0} - df_{H_1}}}{\frac{RSS_{H_1}}{df_{H_1}}} = \frac{\frac{SSReg}{df_{\text{reg}}}}{\hat{\sigma}^2}$$

F: due to R.A. Fisher . $\frac{SSReg}{df_{\text{reg}}}$also called mean square for regression

Note:

$SSReg = RSS_{H_0} - RSS_{H_1}$
$\hat{\sigma}^2 = \frac{RSS_{H_1}}{df_{H_1}}$ [under H_1] $\hat{\sigma}^2$: other choices available
If assume $e|X \sim N(0, \sigma^2 I)$, then under H_0,

$$F \overset{H_0}{\sim} F(df_{\text{reg}}, df_{H_1})$$

Large observed values of (F_{obs}) indicate evidence against H_0.

Special Cases

Overall Test, Simple Regression

$$H_0 : E(Y \mid X = x) = \beta_0 \quad \text{(called "null" model)}$$

$$H_1 : E(Y \mid X = x) = \beta_0 + \beta_1 x$$

(observed value) $RSS_{H_0} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SYY, df_{H_0} = n - 1$

$$\text{(observed value) } RSS_{H_1} = RSS \quad \text{(from SLR)}, df_{H_1} = n - 2$$

$$F = \frac{(SYY - RSS)/\left[(n-1) - (n-2)\right]}{\hat{\sigma}^2}$$

$$= \frac{SSReg}{\hat{\sigma}^2} = \frac{SYY - RSS}{\hat{\sigma}^2}$$

Under H_0, $F \sim F(1, n-2)$
If F_{obs} is large with respect to the null sampling distribution, this provides evidence against H_0.

Example: Forbes Data

Scottish physicist **James Forbes** ran experiments examining the relationship between atmospheric pressure and the boiling point of water.

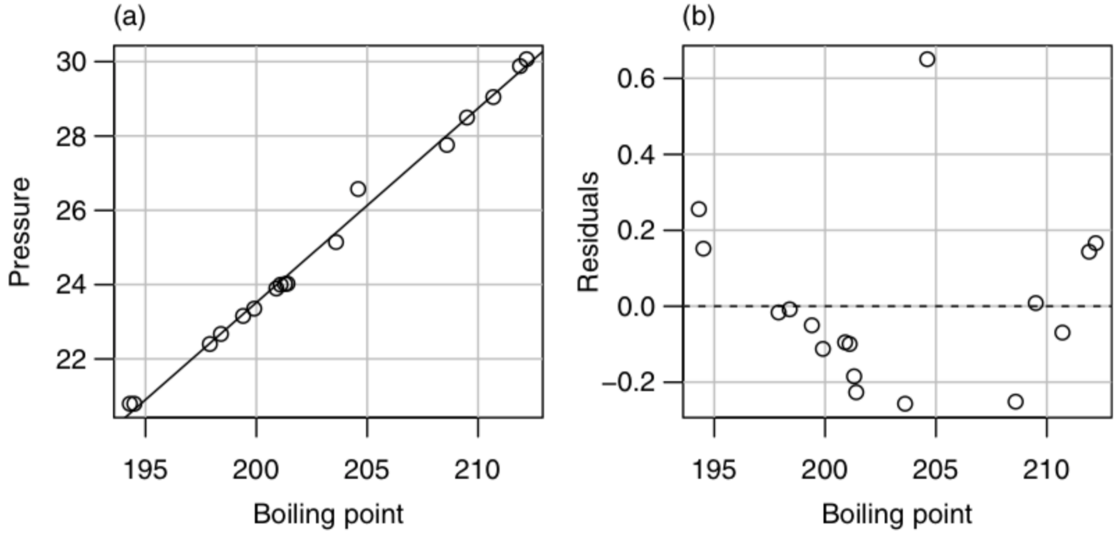


Figure 1.3 Forbes data: (a) `pres` versus `bp`; (b) residuals versus `bp`.
n = 17

$$SYY = 427.794, RSS = 2.155, \hat{\sigma}^2 = 0.144$$

$$\Rightarrow F_{\text{obs}} = \frac{427.794 - 2.155}{0.144} = 2962.79$$

$$F \overset{H_0}{\sim} F(1, 15) \Rightarrow p < 0.0001 \text{ (very strong evidence against } H_0)$$

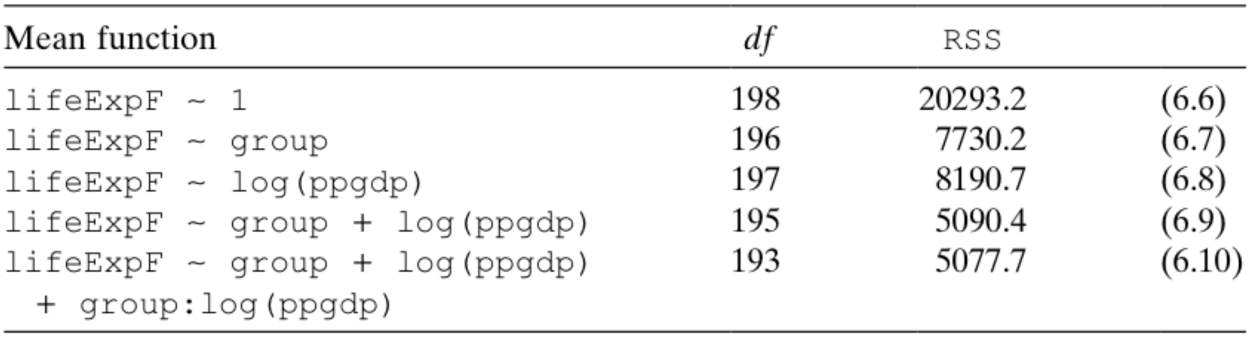
Overall Test, Simple Regression

$H_0 : E(Y \mid X = x) = \beta_0$
$H_1 : E(Y \mid X = x) = \beta_0 + \beta_1 x$(where x is a vector of length p)
Fit Under H_0 $\Rightarrow \hat{\beta}_0 = \bar{y}$ $\Rightarrow RSS_{H_0} = SYY$, $df_{H_0} = n - 1$
Under H_1, RSS_{H_1} and df_{H_1} are from the full model fit.

Fuel consumption data example

$F_{\text{obs}} = 11.99, F \overset{H_0}{\sim} F(4, 46) \Rightarrow p < 0.0001 \therefore$ strong evidence against H_0.

Example: UN data



Mean function	*df*	RSS	
`lifeExpF ~ 1`	198	20293.2	(6.6)
`lifeExpF ~ group`	196	7730.2	(6.7)
`lifeExpF ~ log(ppgdp)`	197	8190.7	(6.8)
`lifeExpF ~ group + log(ppgdp)`	195	5090.4	(6.9)

`lifeExpF ~ group + log(ppgdp)`
`+ group:log(ppgdp)`
	193	5077.7	(6.10)

Mean function (6.6) is the null model, so: $RSS_{\text{null}} = SYY, df_{\text{null}} = n - 1$

Mean function (6.7) has a separate mean for each level of group.

Mean function (6.8) has a common slope and intercept for each level of group.

Mean function (6.9) has a separate intercept but a common slope.

Mean function (6.10) has separate slopes and intercepts.

Different comparisons can be made.

Example1

H_0 : mean function (6.9)

H_1 : mean function (6.10)

$$F_{\text{obs}} = \frac{(5090.4 - 5077.7)/(195 - 193)}{5077.7/193} = 0.24$$

$$F \overset{H_0}{\sim} F(2, 193) \Rightarrow p = 0.79 \text{ , no evidence against } H_0$$

Example2

H_0 : Mean function (6.8)

H_1 : Mean function (6.9)

$$F_{\text{obs}} = \frac{(8190.7 - 5090.4)/(197 - 195)}{5090.4/195} = 59.38$$

$$F \overset{H_0}{\sim} F(2, 195) \Rightarrow p < 0.0001 \therefore \text{strong evidence against } H_0$$

Note:
Two tests just illustrated use different denominators for F-tests.
However, when testing summarized in ANOVA table, the largest model (6.10) would be used for all tests.
When doing that for the second test: $F_{\text{obs}} = 58.92 \quad$ (little change)

General Likelihood Ratio Tests

Fact: F-tests described are the same as Likelihood Ratio Tests for linear models with normal errors.

The Analysis of Variance

Suppose we fit the following model:

$$Y \sim A + B + C + A : B + A : C + B : C + A : B : C,$$

where each of A, B, or C could represent a continuous predictor with a single degree of freedom, or a factor, polynomial, or spline basis with more than one degree of freedom. An interaction like $A : B$ can have many degrees of freedom.

The approach to testing we adopt follows from the *marginality principle*. A lower-order term, such as the A main effect, is never tested in models that include any of its higher-order relatives like $A : B$, $A : C$, or $A : B : C$. All regressors that are not higher-order relatives of the regressor of interest, such as B, C, and $B : C$, are always included in both H_0 and H_1.

Based on the marginality principle, testing should begin with the highest-order interaction first:

$$H_0 : Y \sim A + B + C + A : B + A : C + B : C,$$

$$H_1 : Y \sim A + B + C + A : B + A : C + B : C + A : B : C.$$

If the $A : B : C$ interaction is judged to be nonzero, no further testing is required, since $A : B : C$ is a higher-order relative of all remaining regressors in the mean function.

If the $A : B : C$ interaction is judged nonsignificant, then proceed to examine the two-factor interactions, such as:

$$H_0 : Y \sim A + B + C + A : C + B : C,$$

$$H_1 : Y \sim A + B + C + A : B + A : C + B : C,$$

which tests the interaction $A : B$.

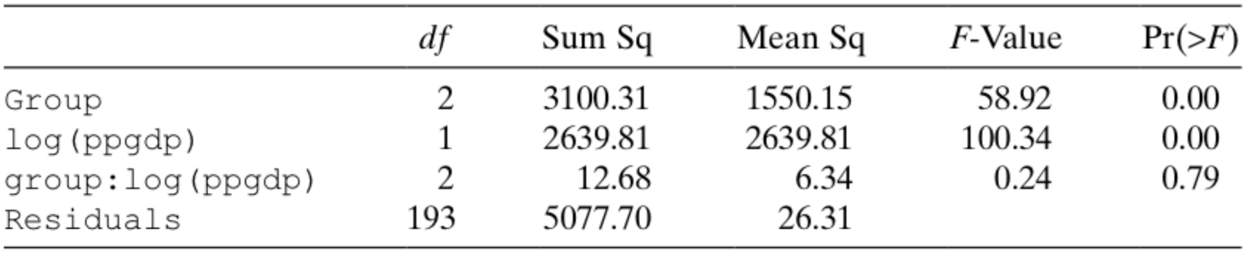
Tests for a main effect like A would be carried out only if all its higher-order relatives, $A : B : C$, $A : B$, and $A : C$, are judged to be unimportant. One would then test:

$$H_0 : Y \sim B + C + B : C,$$

$$H_1 : Y \sim A + B + C + B : C.$$

where $B : C$ is included in both the H_0 and the H_1. Table shows the analysis of variance for the UN data.

Table 6.1 Analysis of Variance for the UN Data



	df	Sum Sq	Mean Sq	*F*-Value	Pr(>*F*)
`Group`	2	3100.31	1550.15	58.92	0.00
`log(ppgdp)`	1	2639.81	2639.81	100.34	0.00
`group:log(ppgdp)`	2	12.68	6.34	0.24	0.79
`Residuals`	193	5077.70	26.31		

An analysis of variable table derived under the marginality principle has the unfortunate name of Type II analysis of variance. At least two other types of ANOVA are commonly available in software packages:

Type I ANOVA, also called sequential ANOVA, fits the model according to the order that the regressors are entered into the mean function. For example, if we fit the model:

$Y \sim A + B + C + A : B + A : C + B : C + A : B : C$,

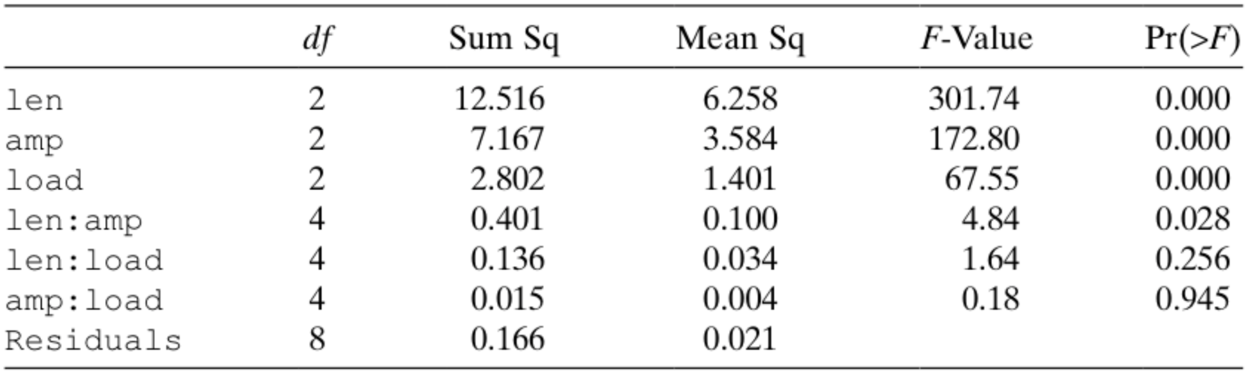
then the sequence of models that would be represented in the ANOVA table would have regressors: $\{A\}, \{A, B\}, \{A, B, C\}, \{A, B, C, A : B\}, \{A, B, C, A : B, A : C\}, \{A, B, C, A : B, A : C, B : C\}$ and $\{A, B, C, A : B, A : C, B : C, A : B : C\}$.

If the terms were written in a different order, then the analysis would have different conditioning. Type I ANOVA generally has only pedagogical interest and should not be used.

Type III ANOVA violates the marginality principle. It computes the test for every regressor adjusted for every other regressor. For example, the test for the A main effect would include the interactions $A : B$, $A : C$, and $A : B : C$ in both H_0 and H_1. There is a justification for this testing paradigm, called the marginal means method, but some of these tests depend on the parameterization used for the regressors and so they are not recommended for general use.

The wool data is from a designed experiment in which all the factors are orthogonal to each other. Table shows the ANOVA table for the full second-order model. Because the regressors are orthogonal, Type I, Type II, and Type III tests are identical.

Table 6.2 Analysis of Variance for the Second-Order Model for the Wool Data



	df	Sum Sq	Mean Sq	*F*-Value	Pr(>*F*)
len	2	12.516	6.258	301.74	0.000
amp	2	7.167	3.584	172.80	0.000
load	2	2.802	1.401	67.55	0.000
len:amp	4	0.401	0.100	4.84	0.028
len:load	4	0.136	0.034	1.64	0.256
amp:load	4	0.015	0.004	0.18	0.945
Residuals	8	0.166	0.021		

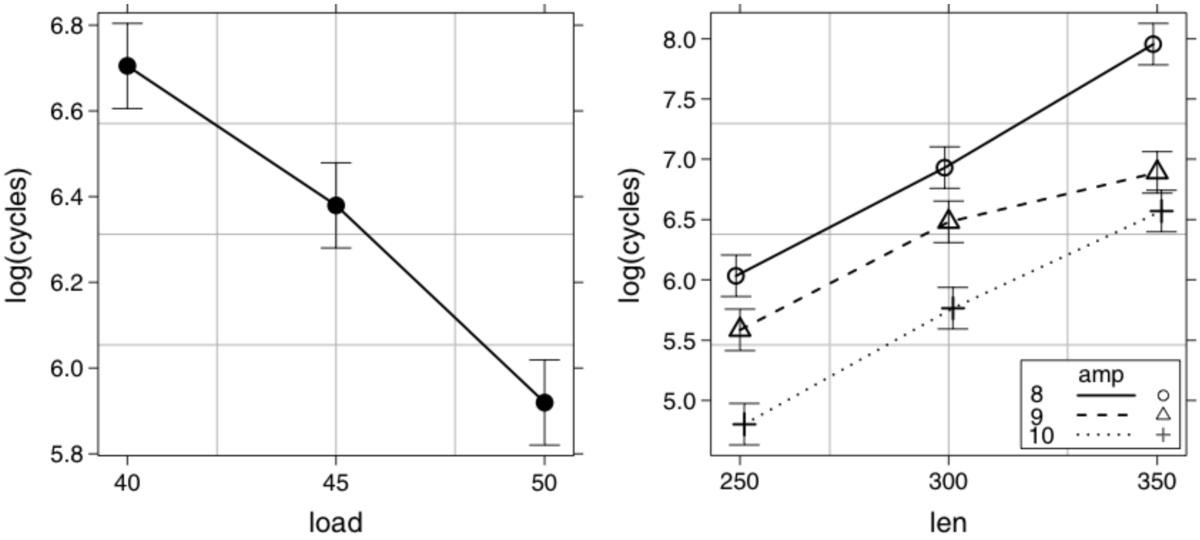


Figure 6.1 Effects plots for the wool data after deleting unimportant interactions.

Power and Non-Null Distributions

For the full-reduced model test, the test statistic is

$$F = \frac{(RSS_{H_0} - RSS_{H_1})/(df_{H_0} - df_{H_1})}{RSS_{H_1}/df_{H_1}}$$

For a fixed significance level, the probability of rejecting a false H_0 is called the power of the test:

$$\textit{power} = \text{Prob(detect a false } H_0) = \text{Prob}(F > f^* \mid H_1 \text{ is true}),$$

where f^* denotes the critical value of the test.

When H_1 is true, the numerator and denominator of the test statistic remain independent. The denominator estimates σ^2 under both the H_0 and the H_1. The distribution of the numerator sum of squares is different under the H_0 and the H_1. Apart from *df*, the numerator under the H_1 is distributed as σ^2 times a noncentral χ^2.

The expected value of the numerator will be

$$\sigma^2(1+\lambda),$$

where λ is called the noncentrality parameter and can be expressed as

$$\lambda = \frac{\beta_2' X_2' \left(I - X_1 \left(X_1' X_1 \right)^{-1} X_1' \right) X_2 \beta_2}{q\sigma^2}.$$

In the case that X_2 consists of a single variable, i.e., $q = 1$, β_2 is a scalar, and

$$\lambda = (n-1) \left(\frac{\beta_2}{\sigma} \right)^2 SD_2^2 \left(1 - R^2_{x_2,x_1} \right),$$

where SD_2 is the standard deviation of X_2, and $R^2_{x_2,x_1}$ is the value of R^2 for the OLS regression with response X_2 and regressor X_1.

Power increases with λ, so it increases with sample size n, the size of the parameter relative to the error standard deviation $\left(\frac{\beta_2}{\sigma} \right)^2$, and SD_2^2.

In most designed experiments, interesting tests concern effects that are orthogonal. Then

$$\lambda = (n-1) \frac{\beta_2' S_2 \beta_2}{q\sigma^2},$$

where S_2 is the sample covariance matrix for X_2.

Wald Tests

Wald tests about regression coefficients are based on the distribution of the estimate $\hat{\beta}$. In most regression problems, the estimator is at least approximately normally distributed,

$$\hat{\beta} \sim N(\beta, V).$$

Generally, V is unknown, but an estimate $\hat{V}$ is available. For OLS estimators, we have

$$\hat{V} = \hat{\sigma}^2(X'X)^{-1}.$$

One Coefficient

To test a hypothesis, say $H_0 : \beta_j = \beta_{j_0}$ versus $H_1 : \beta_j \neq \beta_{j_0}$, compute

$$t = \frac{\hat{\beta}_j - \beta_{j_0}}{\sqrt{\hat{v}_{jj}}},$$

where $\hat{v}_{jj}$ is the (j, j)th element of $\hat{V}$. This test is compared with the t-distribution with degrees of freedom equal to the degrees of freedom in estimating σ^2 to get p-values.

In problems like logistic regression, where there is no σ^2 to estimate, the Wald test is compared with the standard normal distribution.

One Linear Combination

Suppose $\boldsymbol{a}$ is a vector. Then the linear combination $l = \boldsymbol{a}'\boldsymbol{\beta}$ has the estimate $\hat{l} = \boldsymbol{a}'\hat{\boldsymbol{\beta}}$ and

$$\hat{l} \sim N(l, \boldsymbol{a}'V\boldsymbol{a}).$$

Therefore, for $H_0 : l = l_0$, the statistic is

$$t = \frac{\hat{l} - l_0}{\sqrt{\boldsymbol{a}'\hat{V}\boldsymbol{a}}},$$

which is compared with the t-distribution with degrees of freedom given by the degrees of freedom for $\hat{\sigma}^2$.

General Linear Hypothesis

Suppose we wish to test $H_0 : L\boldsymbol{\beta} = \boldsymbol{c}$ versus $H_1 : L\boldsymbol{\beta} \neq \boldsymbol{c}$, where L is a $q \times p'$ matrix of constants.

The test statistic is

$$F = \frac{(L\hat{\boldsymbol{\beta}} - \boldsymbol{c})'(L\hat{V}L')^{-1}(L\hat{\boldsymbol{\beta}} - \boldsymbol{c})}{q}.$$

Under H_0 and normality, this statistic can be compared with an $F(q, n-p')$ distribution to get significance levels.

Equivalence of Wald and Likelihood-Ratio Tests

For linear regression, the Wald tests and the likelihood ratio tests give the same answer for any fixed hypothesis test. This equality does not carry over to other regression settings like logistic regression. Wald and likelihood ratio tests for logistic regression are equivalent, in the sense that for large enough samples they will give the same inference, but not equal, as the computed statistics generally have different values. Likelihood ratio tests are generally preferable.

Interpreting Tests

Interpreting *p*-values

Under the appropriate assumptions, the p-value is the conditional probability of observing a value of the computed statistic as extreme or more extreme than the observed value, given that the H_0 is true. A small p-value provides evidence against the H_0.

In many research areas it has become traditional to adopt a fixed significance level when examining p-values. The most common choice for the significance level is $\alpha = 0.05$, which would mean that, were the H_0 to be true, we would incorrectly find evidence against it about 5% of the time, or about one test in 20.

There is an important distinction between statistical significance, the observation of a sufficiently small p-value, and scientific significance, observing an effect of sufficient magnitude to be meaningful. Judgment of the latter usually will require examination of more than just the p-value.

Why Most Published Research Findings Are False

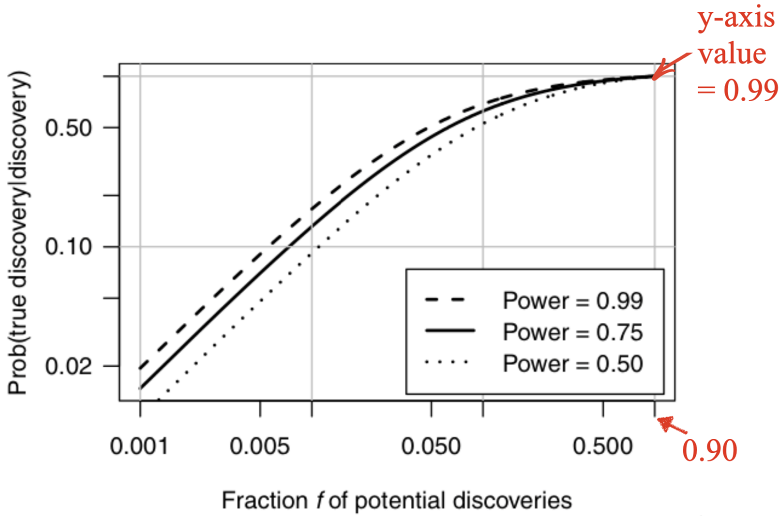
Following Ioannidis (2005):

Set up:

- Multiple tests all at level α
- All with the same power to reject H_0 (say γ)

1. Suppose fraction f are potential discoveries.
2. True discovery happens with probability $f\gamma$.
3. False discovery happens with probability $(1-f)\alpha$.
4. $P(\text{discovery}) = f\gamma + (1-f)\alpha$

$$\therefore P(\text{true discovery} \mid \text{discovery}) = \frac{f\gamma}{f\gamma + (1-f)\alpha}$$



y-axis value = 0.99

0.90

Figure 6.2: The probability of a true discovery as a function of the fraction f of false null hypotheses (NH) and the power of the test.

considered $\gamma = [0.50,0.75,0.99]$, $\alpha = 0.05$

Multiple Testing

Multiple testing is a significant issue in interpreting tests. If 100 independent tests are conducted at $\alpha = 0.05$, and if H_0 is true in all cases, about $100 \times 0.05 = 5$ tests are expected to show significance purely by chance.

Traditional methods of surviving multiple testing are to control the *family-wise error rate* rather than the *per-test error rate*, but recent methodology is based on controlling the *false discovery rate.* Except for testing for outliers, we leave discussion and application of multiple testing methods to other sources.