Tommy Norberg
November 9, 2011
30 pages

`lect3.tex`

Part a) Single-sample case

Let $x_1, \ldots, x_n$ be i.i.d observations of some random variable X with mean

$$EX = \mu$$

and variance

$$\operatorname{Var} X = \sigma^2$$

The sample mean $\bar{x}$ and variance s^2 are unbiased estimators of the mean μ and the variance σ^2, respectively,

$$E\bar{x} = \mu$$

$$Es^2 = \sigma^2$$

Recall also the c.l.t, implying

$$\frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{n}} \approx \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \overset{\text{ap}}{\sim} \mathrm{N}(0, 1)$$

where $\hat{\sigma}$ is some natural estimate of σ.

Note that

$$\sqrt{\operatorname{Var} \bar{x}} = \sigma/\sqrt{n}$$

often is referred to as the *standard error of the mean.* Its natural estimator is $\hat{\sigma}/\sqrt{n}$.

The case $X \sim \mathbf{Ber}(p)$.

Here,

$$\mu = p \quad \text{and} \quad \sigma^2 = p(1-p)$$

Moreover, $f = \sum_i x_i$ is the frequency of “positives”, and

$$\hat{p} = \bar{x} = \frac{f}{n}$$

estimates p without bias.

The “natural” estimator of the standard deviation is

$$\hat{\sigma} = \sqrt{\hat{p}(1-\hat{p})}$$

Suppose we want to test

$$H_0 : p \geq p_0 \quad \text{vs} \quad H_1 : p < p_0$$

It is natural to reject H_0 when $f \leq c$, where c is determined from the requirement

$$P(f \leq c | p_0) \leq \alpha \quad \text{and} \quad P(f \leq c+1 | p_0) > \alpha$$

since f has a discrete distribution.

The power of the test is

$$1 - \beta = P(f \leq c | p_1) \quad \text{for} \quad p_1 < p_0$$

In Chapter 15, Montgomery introduces typa-A and type-B calculations of the probability $P(f \leq c|p)$.

We now add a "type-C calculation" based on the normal approximation of the Binomial distribution,

$$P(f \leq c|p) = \Phi\left(\frac{c + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

This is appropriate when n is large and p is neither too small or too large.

The two-sided test

$$H_0 : p = p_0 \quad \text{vs} \quad H_1 : p \neq p_0$$

is treated similarly. Reject H_0 when $f \leq c_1$ or $f \geq c_2$, where

$$P(f \leq c_1|p_0) \approx P(f \geq c_2|p_0) \approx \frac{\alpha}{2}$$

A type-C calculation leads to

$$c_1 \approx np_0 - 0.5 - z_{\alpha/2}\sqrt{np_0(1-p_0)}$$
$$c_2 \approx np_0 + 0.5 + z_{\alpha/2}\sqrt{np_0(1-p_0)}$$

The half-number correction is often omitted.

If so, the rule is to reject $H_0 : p = p_0$ when $|z| > z_{\alpha/2}$, where

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \overset{\text{ap}}{\sim} \mathrm{N}(0,1)$$

A corresponding confidence interval for p consists of all p_0 that are not rejected by the the test. That is, all p_0 such that

$$-z_{\alpha/2} \leq \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \leq z_{\alpha/2}$$

It is somewhat easier to use as starting point the fact that the event

$$-z_{\alpha/2} \leq \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq z_{\alpha/2}$$

occurs with probability $\approx 1 - \alpha$, and rearrange to

$$\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

which not seldom is written

$$p = \hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

The case $X \sim \mathbf{N}(\mu, \sigma^2)$.

Here,

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1)$$

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

It follows from the fact that $\bar{x}$ and s^2 are independent that

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

Suppose we want to test

$$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0$$

It is natural to reject H_0 when $\bar{x} > c$, where c is determined from the requirement

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha \quad \Leftrightarrow \quad \bar{x} > \mu_0 + z_\alpha \sigma/\sqrt{n}$$

if σ is known, and from

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{\alpha,n-1} \quad \Leftrightarrow \quad \bar{x} > \mu_0 + t_{\alpha,n-1} s/\sqrt{n}$$

if σ is unknown and must be estimated.

The corresponding confidence interval consists of all μ_0 that are not rejected by the test. That is, μ is in the confidence interval if

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \leq t_{\alpha,n-1}$$

Clearly, this holds true if, and only if,

$$\mu \geq \bar{x} - t_{\alpha,n-1} s/\sqrt{n}$$

Consider next the test of

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

The rule is to reject if $|t| > t_{\alpha/2,n-1}$, where

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

If σ is known, replace s and $t_{\alpha/2,n-1}$ by σ and $z_{\alpha/2}$, respectively. In this case we write z instead of t.

The corresponding confidence interval consists of all μ_0 that are not rejected by the test. That is, all μ such that

$$-t_{\alpha/2,n-1} \leq \frac{\bar{x} - \mu}{s/\sqrt{n}} \leq t_{\alpha/2,n-1}$$

Rearrange.

The test of

$$H_0 : \sigma \geq \sigma_0 \quad \text{vs} \quad H_1 : \sigma < \sigma_0$$

rejects when

$$\frac{(n-1)s^2}{\sigma_0^2} < \chi^2_{1-\alpha,n-1} \quad \Leftrightarrow \quad s < \sigma_0 \sqrt{\frac{\chi^2_{1-\alpha,n-1}}{n-1}}$$

The corresponding confidence interval consists of all σ_0 that are not rejected by the test, i.e, all σ satisfying

$$\sigma \leq s \sqrt{\frac{n-1}{\chi^2_{1-\alpha,n-1}}}$$

If the alternative is two-sided, reject $H_0 : \sigma = \sigma_0$ when

$$\frac{(n-1)s^2}{\sigma_0^2} < \chi^2_{1-\alpha/2,n-1} \quad \text{or} \quad \frac{(n-1)s^2}{\sigma_0^2} > \chi^2_{\alpha/2,n-1}$$

The corresponding confidence interval consists of all σ satisfying

$$\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}$$

Rearrange.

P-value.

Definition. The P-value is the smallest level of significance that would lead to rejection of the null hypothesis.

Example. Normal data. The P-value of the test of

$$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0$$

is $P(T > t)$, where $T \sim t(n-1)$ and

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Also note that many statisticians prefer to write the null hypothesis as

$$H_0 : \mu = \mu_0 \text{ instead of } H_0 : \mu \leq \mu_0$$

since the level is calculated under the assumption that the true mean is μ_0. This applies of course also to similar tests against one-sided alternatives.

Example. Bernoulli data. The P-value of the test of

$$H_0 : p = p_0 \quad \text{vs} \quad H_1 : p \neq p_0$$

is $2P(Z > |z|)$, where $Z \sim \mathrm{N}(0, 1)$ and

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Part b) The two-sample case

	empirism			theory	
1st sample	n_1	$\bar{x}_1$	s_1^2	μ_1	σ_1^2
2nd sample	n_2	$\bar{x}_2$	s_2^2	μ_2	σ_2^2

The samples must be independent. They may be from the same population, but need not be. Typically, they are from different subpopulations of some population.

If data are normal,

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

Thus, we may reject

$$H_0 : \sigma_1 = \sigma_2 \quad \text{vs} \quad H_1 : \sigma_1 \neq \sigma_2$$

at level α, if

$$\frac{s_1^2}{s_2^2} < f_{1-\alpha/2, n_1-1, n_2-1} \quad \text{or} \quad \frac{s_1^2}{s_2^2} > f_{\alpha/2, n_1-1, n_2-1}$$

Note that the level should be high (typically 10% or even 20%), if the test is used to determine whether one can assume that $\sigma_1 = \sigma_2$ in a subsequent analysis of the difference $\mu_1 - \mu_2$.

Normal data. The case $\sigma_1 = \sigma_2 = \sigma$.

The pooled sample variance,

$$s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

is an unbiased estimator of σ^2. Moreover,

$$\frac{(n_1 + n_2 - 2)s_P^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$

Now note that

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim \mathrm{N}(0, 1)$$

and, since the sample variances are independent of the sample means,

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s\sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2)$$

Tests and confidence intervals are based on theese facts.

Note that there are slightly different procedures depending on whether σ is known or not.

Normal data. The case $\sigma_1 \neq \sigma_2$.

Here,

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathrm{N}(0,1)$$

and it can be shown that

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \overset{\text{ap}}{\sim} t(\nu)$$

where ν is the integer part of

$$\frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1} + \frac{\left(s_2^2/n_2\right)^2}{n_2-1}}$$

(There is a typo in Formula (4.55) on p 134.)

Tests and confidence intervals are based on theese facts.

Bernoulli data.

Here, $f_i = n_i \bar{x}_i$ are the frequencies of positives in the two samples, and

$$\hat{p}_i = \frac{f_i}{n_i}$$

is an unbiased estimate of the proportion p_i of positives in the ith sample.

The type-C test statistic for testing the null hypothesis

$$H_0 : p_1 = p_2 = p$$

is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\,(1/n_1 + 1/n_2)}} \overset{\text{ap}}{\sim} \mathrm{N}(0,1)$$

where

$$\hat{p} = \frac{f_1 + f_2}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

is unbiased for the under H_0 common proportion p.

Two-sided confidence intervals for $p_1 - p_2$ typically use

$$-z_{\alpha/2} \leq \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}} \leq z_{\alpha/2}$$

as starting point. Single-sided confidence statements use only one of the inequalities and $\alpha/2$ should be replaced by α.

Part c) Analysis of variance

The Paper Tensile Strength Experiment

Tensile strength of paper (psi)

		Observation no j							
x	i	1	2	3	4	5	6	$n\bar{y}_i$	$\bar{y}_i$
0.05	1	7	8	15	11	9	10	60	10.00
0.10	2	12	17	13	18	19	15	94	15.67
0.15	3	14	18	19	17	16	18	102	17.00
0.20	4	19	25	22	23	18	20	127	21.17

See Section 4.5.1 on p 140.

This is a so called single-factor experiment.

Data and model:

$$y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij}$$

for $i = 1, \ldots, a$ and $j = 1, \ldots, n$, where

$$\mu = \frac{1}{a} \sum_i \mu_i \quad \text{and} \quad \tau_i = \mu_i - \mu$$

Note

$$\sum_i \tau_i = 0$$

We assume that the ϵ_{ij}'s are i.i.d $\mathrm{N}(0,\, \sigma^2)$, i.e that the y_{ij}'s are independent and

$$y_{ij} \sim \mathrm{N}(\mu + \tau_i,\, \sigma^2)$$

The null hypothesis of no effect

$$H_0 : \tau_1 = \cdots = \tau_a = 0$$

is to be tested vs the alternative

$$H_1 : \tau_i \neq 0 \text{ for at least one } i$$

The total variation in the data is described by the *total sum of squares*

$$\begin{aligned}
\mathrm{SS_{Tot}} &= \sum_i \sum_j (y_{ij} - \bar{y})^2 \\
&= \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + n \sum_i (\bar{y}_i - \bar{y})^2 \\
&= \mathrm{SS_{Err}} + \mathrm{SS_{Treat}}
\end{aligned}$$

This is the *sum of squares identity.*

The associated degrees of freedom are

$$\begin{aligned}
\mathrm{df_{Tot}} &= an - 1 \\
\mathrm{df_{Err}} &= a(n-1) \\
\mathrm{df_{Treat}} &= a - 1
\end{aligned}$$

Note

$$\mathrm{df_{Tot}} = \mathrm{df_{Err}} + \mathrm{df_{Treat}}$$

The idea of an analysis of variance is to relate (or compare) the within groups variation $\mathrm{SS_{Err}}/\mathrm{df_{Err}}$ and the between groups variation $\mathrm{SS_{Treat}}/\mathrm{df_{Treat}}$.

If we divide an SS with its df, we get a mean sum of squares, denoted MS.

It is easy to see that

$$s^2 = \mathrm{MS}_{\mathrm{Err}} = \frac{\mathrm{SS}_{\mathrm{Err}}}{a(n-1)}$$

is unbiased for σ^2. Moreover,

$$\frac{a(n-1)\, s^2}{\sigma^2} \sim \chi^2(a(n-1))$$

It is somewhat harder to see that

$$E\mathrm{SS}_{\mathrm{Treat}} = (a-1)\,\sigma^2 + n \sum_i \tau_i^2$$

Thus, under H_0, $\mathrm{MS}_{\mathrm{Treat}}$ is unbiased for σ^2.

Moreover, $\mathrm{MS}_{\mathrm{Treat}}$ is independent of $\mathrm{MS}_{\mathrm{Err}}$, and

$$\frac{(a-1)\,\mathrm{MS}_{\mathrm{Treat}}}{\sigma^2} \sim \chi^2(a-1)$$

It follows that

$$F_0 = \frac{\mathrm{MS}_{\mathrm{Treat}}}{\mathrm{MS}_{\mathrm{Err}}} \sim F(a-1, a(n-1))$$

and that we may reject the null hypothesis of no effect, when

$$F_0 > F_{\alpha, a-1, a(n-1)}$$

The result of an ANOVA is most often summarized in an ANOVA table.

The Paper Tensile Strength Experiment.

See Exanple 4.12 on p 147.

Tensile strength of paper (psi)

		Observation no j							
x	i	1	2	3	4	5	6	$n\bar{y}_i$	$\bar{y}_i$
0.05	1	7	8	15	11	9	10	60	10.00
0.10	2	12	17	13	18	19	15	94	15.67
0.15	3	14	18	19	17	16	18	102	17.00
0.20	4	19	25	22	23	18	20	127	21.17

Montgomery calculates $\mathrm{SS_{Tot}}$ and $\mathrm{SS_{Treat}}$ by hand, and then finds $\mathrm{SS_{Err}}$ by means of the sum of squares identity,

$$\mathrm{SS_{Tot}} = \mathrm{SS_{Treat}} + \mathrm{SS_{Err}}$$

One-Way ANOVA

Source	df	SS	MS	F	P-value
Factor	3	382.79	127.60	19.61	0.000
Error	20	130.17	6.51		
Total	23	512.96			

Because of the very low P-value, there is strong evidence for the alternative hypothesis that the hardwood concentration affects the strength of the paper.

Suppose next that we want to investigate whether

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

where

$$E\epsilon = 0 \quad \text{and} \quad \operatorname{Var} \epsilon = \sigma^2$$

so that

$$Ey = \beta_0 + \beta_1 x + \beta_2 x^2 \quad \text{and} \quad \operatorname{Var} y = \sigma^2$$

is a reasonable model for the tensile strength data.

Part d) Linear regression models

Data are n independent observations of a $(k+1)$-tuple

$$y, x_1, \ldots, x_k$$

The model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

where

$$E[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

$$\operatorname{Var} y = \operatorname{Var} \epsilon = \sigma^2$$

There are $p = k + 1$ *regression coefficients* in the model.

The *co-variates (regressors, predictor variables)*

$$x_1, \ldots, x_k$$

may be random but need not be.

In the former case, the error ϵ must be independent of the co-variates $x_1, \ldots, x_k$, and the the statistical analysis is conditional on the observed x_i's.

This is typically not the case in quality improvment, where most often the experiments are designed.

There may be all sorts of dependence within the x_i's.

In matrix notation, the model may be written

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \text{ is the column vector of data,}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \text{ is the } \textit{design matrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ represents the regression coefficients,}$$

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \text{ represents the measurement errors.}$$

The least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$$

This is the solution to the normal equations

$$\left(\boldsymbol{X}'\boldsymbol{X}\right)\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$$

The fitted regression model is

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

The $(n \times 1)$ vector of residuals is

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

Estimating σ^2. The residual or *error sum of squares* is

$$\begin{aligned} \mathrm{SS}_{\mathrm{Err}} &= \boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{y} - \hat{\boldsymbol{y}})' (\boldsymbol{y} - \hat{\boldsymbol{y}}) \\ &= \cdots = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} \end{aligned}$$

There are $n-p$ degrees of freedom associated with $\mathrm{SS}_{\mathrm{Err}}$ and

$$s^2 = \mathrm{MS}_{\mathrm{Err}} = \frac{\mathrm{SS}_{\mathrm{Err}}}{n-p}$$

estimates σ^2 without bias.

Under the normal assumption, i.e,

$$y \sim \mathrm{N}\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \sigma^2\right)$$

we have

$$\frac{(n-p)s^2}{\sigma^2} \sim \chi^2(n-p)$$

Properties of the estimators.

$$E\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$$

$$\operatorname{Cov}\hat{\boldsymbol{\beta}} = E\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\right] = \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

Under the normal assumption, $\hat{\boldsymbol{\beta}}$ is multivariate normal.

Further, $\hat{\boldsymbol{\beta}}$ and s^2 are independent.

Denote by c_{jj} the jth diagonal element of $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$.

We have

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{c_{jj}}} \sim \mathrm{N}(0,1)$$

so, by the independence,

$$\frac{\hat{\beta}_j - \beta_j}{s\sqrt{c_{jj}}} \sim t(n-p)$$

Confidence intervals and tests on single regression coefficients are based on this fact.

The test for significance of regression is

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$

vs

$$H_1 : \beta_j \neq 0 \text{ for at least one } j$$

The total sum of squares

$$\mathrm{SS}_{\mathrm{Tot}} = (\boldsymbol{y} - \bar{\boldsymbol{y}})' (\boldsymbol{y} - \bar{\boldsymbol{y}}) = \boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2$$

where

$$\bar{\boldsymbol{y}} = \begin{bmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix} = \bar{y}\,\mathbf{1}$$

may be partitioned into

$$\mathrm{SS}_{\mathrm{Tot}} = \mathrm{SS}_{\mathrm{Reg}} + \mathrm{SS}_{\mathrm{Err}}$$

where $\mathrm{SS}_{\mathrm{Err}}$ is the already defined error sum of squares and

$$\mathrm{SS}_{\mathrm{Reg}} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - n\bar{y}^2$$

is the *regression sum of squares.*

The associated partition of the number of degrees of freedom is

$$n - 1 = k + (n - p)$$

The test statistic

$$F_0 = \frac{\mathrm{MS}_{\mathrm{Reg}}}{\mathrm{MS}_{\mathrm{Err}}} = \frac{\mathrm{SS}_{\mathrm{Reg}}/k}{\mathrm{SS}_{\mathrm{Err}}/(n-p)}$$

has under H_0 an $F(k, n-p)$-distribution.

The rule is to reject H_0 if

$$F_0 > F_{\alpha,k,n-p}$$

Equivalently, if the P-value is less than α.

The *coefficient of multiple determination*

$$R^2 = \frac{\mathrm{SS}_{\mathrm{Reg}}}{\mathrm{SS}_{\mathrm{Tot}}} = 1 - \frac{\mathrm{SS}_{\mathrm{Err}}}{\mathrm{SS}_{\mathrm{Tot}}}$$

measures the amount of reduction in the variability of the data $y_1, \ldots, y_n$ obtained by the regression model.

Adding a variable to the model will allways increase R^2. Therefore R^2 is not a very good measure of the model fit.

The *adjusted R^2 statistic*

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{MS}_{\mathrm{Err}}}{\mathrm{MS}_{\mathrm{Tot}}} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2)$$

is a better measure of the fit of the model, since adding unnecessary variables not seldom decreases its value.

Tests on groups of coefficients.

It is also possible to partition the vector of regression coefficients

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

where $\boldsymbol{\beta}_1$ is $r \times 1$ and $\boldsymbol{\beta}_2$ is $(p-r) \times 1$, and to test

$$H_0 : \boldsymbol{\beta}_1 = \mathbf{0} \quad \text{vs} \quad H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}$$

The model is then written as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

where $\boldsymbol{X}_1$ represents the columns of $\boldsymbol{X}$ associated with $\boldsymbol{\beta}_1$ and similarly for $\boldsymbol{X}_2$.

The regression sum of squares for the non-constant variables $x_1, \ldots, x_k$ is

$$\mathrm{SS}_{\mathrm{Reg}} = \mathrm{SS}_{\mathrm{Tot}} - \mathrm{SS}_{\mathrm{Error}} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - n\bar{y}^2$$

The regression sum of squares for all p variables is

$$\mathrm{SS}_{\mathrm{Reg}}(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$$

The associated number of degrees of freedom is p.

The reduced model, valid under H_0, is

$$\boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

The least squares estimator of $\boldsymbol{\beta}_2$ is

$$\hat{\boldsymbol{\beta}}_2 = \left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'\boldsymbol{y}$$

in the reduced model.

The regression sum of squares for the variables associated with $\boldsymbol{\beta}_2$ is

$$\mathrm{SS}_{\mathrm{Reg}}(\boldsymbol{\beta}_2) = \hat{\boldsymbol{\beta}}_{\boldsymbol{2}}'\boldsymbol{X}_2'\boldsymbol{y}$$

It has $p - r$ degrees of freedom.

The regression sum of squares due to $\boldsymbol{\beta}_1$ given that $\boldsymbol{\beta}_2$ already is in the model is

$$\mathrm{SS}_{\mathrm{Reg}}(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2) = \mathrm{SS}_{\mathrm{Reg}}(\boldsymbol{\beta}) - \mathrm{SS}_{\mathrm{Reg}}(\boldsymbol{\beta}_2)$$

This sum of squares has r degrees of freedom.

Clearly,

$$\mathrm{SS}_{\mathrm{Reg}}(\boldsymbol{\beta}) = \mathrm{SS}_{\mathrm{Reg}}(\boldsymbol{\beta}_2) + \mathrm{SS}_{\mathrm{Reg}}(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)$$

Hence $\mathrm{SS}_{\mathrm{Reg}}(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)$ is the increase in the regression sum of squares due to including the variables associated with $\boldsymbol{\beta}_1$ in the model.

Now,

$$\mathrm{MS_{Reg}}(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2) = \frac{\mathrm{SS_{Reg}}(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)}{r}$$

is independent of $\mathrm{MS_{Error}}$ and unbiased for σ^2 under the null hypothesis $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$.

Thus, H_0 may be tested by the statistic

$$F_0 = \frac{\mathrm{MS_{Reg}}(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)}{\mathrm{MS_{Error}}} \sim F(r, n-p)$$

The rule is to reject H_0 when $F_0 > F_{\alpha,r,n-p}$.

Rejection means that $H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}$ is accepted, which is to say that at least one of the variables associated to $\boldsymbol{\beta}_1$ significantly contributes to the regression model.