3.6 Linear Regression Analysis

Suppose that we have n measurements of two variables X and Y, that is, a sample of pairs of observations

$$\{(x_i, y_i) : i = 1, ..., n\}$$

and it is believed that there is a **linear** relationship between X and Y. Suppose that we regard X as a **controlled** variable, that is, we can control the values of X at which Y is measured. Our aim is to try and **predict** Y for a given value of X, and thus we have to build a probability model for Y conditional on $X = x$ that incorporates the linear dependence.

3.6.1 Terminology

Y is the **response** or **dependent** variable

X is the **covariate** or **independent** variable

A simple relationship between Y and X is the **linear regression model**, where

$$\mathrm{E}[Y|X = x] = \alpha + \beta x,$$

that is, conditional on $X = x$, the expected or "predicted" value of Y is given by $\alpha + \beta x$, where α and β are unknown parameters; in other words, we model the relationship between Y and X as a straight line with **intercept** α and **slope** β.

For data $\{(x_i, y_i) : i = 1, ..., n\}$, the objective is to estimate the unknown parameters α and β. A simple estimation technique, is **least-squares estimation**.

3.6.2 Least-Squares Estimation

Suppose that a sample, $\{(x_i, y_i) : i = 1, ..., n\}$, is believed to follow a linear regression model, $\mathrm{E}[Y|X = x] = \alpha+\beta x$. For fixed values of α and β, let $y_i^{(P)}$ denote the expected value of Y conditional on $X = x_i$, that is

$$y_i^{(P)} = \alpha + \beta x_i$$

Now define error terms e_i, $i = 1, ..., n$ by

$$e_i = y_i - y_i^{(P)} = y_i - \alpha - \beta x_i$$

that is, e_i is the vertical discrepancy between the **observed** and **expected** values of Y.

The objective in least-squares estimation is find a "line of best fit", and this is achieved by inspecting the squares of the error terms e_i, and choosing α and β such that the sum of the squared errors is **minimized**; we aim to find the straight line model for which the total error is smallest.

Let $S(\alpha, \beta)$ denote the total error in fitting a linear regression model with parameters α and β. Then

$$S(\alpha, \beta) \;=\; \sum_{i=1}^{n} e_i^2 \;=\; \sum_{i=1}^{n} (y_i - y_i^{(P)})^2 \;=\; \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

To calculate the least-squares estimates, we have to minimize $S(\alpha, \beta)$ as a function of α and β. This can be achieved in the usual way by taking partial derivatives with respect to the two parameters, and equating the partial derivatives to zero simultaneously.

$$(1)\quad \frac{\partial}{\partial \alpha}\{S(\alpha, \beta)\} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0 \qquad (2)\quad \frac{\partial}{\partial \beta}\{S(\alpha, \beta)\} = -2\sum_{i=1}^{n}x_i(y_i - \alpha - \beta x_i) = 0$$

Solving (1), we obtain an equation for the least-squares estimates $\hat{\alpha}$ and $\hat{\beta}$

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}\,\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{y} - \hat{\beta}\bar{x}.$$

Solving (2) in the same way, and combining the last two equations, and solving for $\hat{\beta}$ gives

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} x_i y_i - \hat{\beta} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i} \qquad \hat{\beta} = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left\{ \sum_{i=1}^{n} x_i \right\}^2}$$

Thus the least-squares estimates of α and β are given by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \hat{\beta} = \frac{nS_{xy} - S_x S_y}{nS_{xx} - \{S_x\}^2}$$

where

$$S_x = \sum_{i=1}^{n} x_i \qquad S_y = \sum_{i=1}^{n} y_i \qquad S_{xx} = \sum_{i=1}^{n} x_i^2 \qquad S_{xy} = \sum_{i=1}^{n} x_i y_i$$

Therefore it is possible to produce estimates of parameters in a linear regression model using least-squares, without any specific reference to probability models. In fact, the least-squares approach is very closely related to maximum likelihood estimation for a specific probability model.

The **correlation coefficient**, r, measures the degree of association between X and Y variables and is given by

$$r = \frac{nS_{xy} - S_x S_y}{\sqrt{(nS_{xx} - S_x^2)(nS_{yy} - S_y^2)}}$$

and therefore is quite closely related to $\hat{\beta}$.

3.6.3 Relationship between least-squares and maximum likelihood

Suppose that X and Y follow a linear regression model,

$$\mathrm{E}[Y|X = x] = \alpha + \beta x,$$

and recall that the error terms e_i were defined

$$e_i = y_i - \alpha - \beta x_i.$$

Now, e_i is the vertical discrepancy between observed and expected behaviour, and thus e_i could be interpreted as the observed version of a **random variable**, say ϵ_i, which represents the random uncertainty involved in measuring Y for a given X. A plausible probability model might therefore be that the random variables ϵ_i, $i = 1, ...n$, were independent and identically distributed, and

$$\epsilon_i \sim N(0, \sigma^2),$$

for some error variance parameter σ^2. Implicit in this assumption is that the distribution of the random error in measuring Y does not depend on the value of X at which the measurement is made. This distributional assumption about the error terms leads to a probability model for the variable Y. As we can write

$$Y = \alpha + \beta X + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$, then given on $X = x_i$, we have the conditional distribution Y_i as

$$Y_i|X = x_i \sim N(\alpha + \beta x_i, \sigma^2),$$

where random variables Y_i and Y_j are **independent** (as ϵ_i and ϵ_j are independent). On the basis of this probability model, we can derive a likelihood function, and hence derive maximum likelihood estimates. For example, we have the likelihood $L(\theta) = L(\alpha, \beta, \sigma^2)$ defined as the product of the n conditional density terms derived as the conditional density of the observed y_i given x_i,

$$L(\theta) = \prod_{i=1}^{n} f(y_i; x_i, \theta) \; = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right\} = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right\}$$

The maximum likelihood estimates of α and β, and error variance σ^2, are obtained as the values at which $L(\alpha, \beta, \sigma^2)$ is maximized. But, $L(\alpha, \beta, \sigma^2)$ is maximized when the term in the exponent, that is,

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

is minimized. But this is **precisely** the least-squares criterion described above, and thus the m.l.e s of α and β assuming a Normal error model are **exactly equivalent** to the least-squares estimates.

3.6.4 Estimates of Error Variance and Residuals

In addition to the estimates of α and β, we can also obtain the maximum likelihood estimate of σ^2,

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta} x_i)^2 = S^2$$

Often, a **corrected** estimate, s^2, of the error variance is used, defined by

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \; = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

where $\hat{y_i} = \hat{\alpha} + \hat{\beta} x_i$ is the **fitted value** of Y at $X = x_i$. Note also that, having fitted a model with parameters $\hat{\alpha}$ and $\hat{\beta}$, we can calculate the error in fit at each data point, or **residual**, denoted $e_i, i = 1, ..., n$, where $e_i = y_i - \hat{y_i} = y_i - \hat{\alpha} - \hat{\beta} x_i$.

3.6.5 Prediction for a new covariate value

Suppose that, having fitted a model, and obtained estimates $\hat{\alpha}$ and $\hat{\beta}$ using maximum likelihood or least-squares, we want to predict the Y value for a new value x^* of covariate X. By considering the nature of the regression model, we obtain the predicted value y^* as

$$y^* = \hat{\alpha} + \hat{\beta} x^*$$

3.6.6 Standard Errors of Estimators and t-statistics

We need to be able to understand how the estimators corresponding to $\hat{\alpha}$ and $\hat{\beta}$ behave, and by how much the estimate is likely to vary. This can be partially achieved by inspection of the **standard errors** of estimates, that is, the square-root of the variance in the sampling distribution of the corresponding estimator. It can be shown that

$$s.e.(\hat{\alpha}) = s\sqrt{\frac{S_{xx}}{nS_{xx} - \{S_x\}^2}} \qquad s.e.(\hat{\beta}) = s\sqrt{\frac{n}{nS_{xx} - \{S_x\}^2}}$$

where s is the square-root of the corrected estimate of the error variance. It is good statistical practice to report standard errors whenever estimates are reported. The standard error of a parameter also allows a test of the hypothesis "parameter is equal to zero". The test is carried out by calculation of the **t-statistic**, that is, the ratio of a parameter estimate to its standard error. The t-statistic must be compared with the 0.025 and 0.975 percentiles of a Student-t distribution with $n - 2$ degrees of freedom as described below.

3.6.7 Hypothesis Tests and Confidence Intervals for Parameters

We may carry out hypothesis tests for the parameters in a linear regression model; as usual we need to be able to understand the sampling distributions of the corresponding estimators. In the linear regression model, the sampling distributions of the estimators of α and β have **Student-t distributions** with $n-2$ degrees of freedom, hence we use the test statistics

$$t_\alpha = \frac{\hat{\alpha}-c}{s\sqrt{\dfrac{S_{xx}}{nS_{xx}-\{S_x\}^2}}} = \frac{\hat{\alpha}-c}{s.e.(\hat{\alpha})} \qquad t_\beta = \frac{\hat{\beta}-c}{s\sqrt{\dfrac{n}{nS_{xx}-\{S_x\}^2}}} = \frac{\hat{\beta}-c}{s.e.(\hat{\beta})}$$

to test the null hypothesis that the parameter is equal to c. Typically, we use a test at the 5 % significance level, so the appropriate critical values are the 0.025 and 0.975 quantiles of a $St(n-2)$ distribution.

It is also useful to report, for each parameter, a confidence interval in which we think the **true** parameter value (that we have estimated by $\hat{\alpha}$ or $\hat{\beta}$) lies with high probability. It can be shown that the 95% confidence intervals are given by

$$\alpha \;:\; \hat{\alpha} \;\pm\; t_{n-2}(0.975)\; s\,\sqrt{\frac{S_{xx}}{nS_{xx}-\{S_x\}^2}} \qquad \beta \;:\; \hat{\beta} \;\pm\; t_{n-2}(0.975)\; s\,\sqrt{\frac{n}{nS_{xx}-\{S_x\}^2}}$$

where $t_{n-2}(0.975)$ is the 97.5th percentile of a Student-t distribution with $n-2$ degrees of freedom.

The confidence intervals are useful because they provide an alternative method for carrying out hypothesis tests. For example, if we want to test the hypothesis that $\alpha = c$, say, we simply note whether the 95% confidence interval contains c. If it does, the hypothesis can be accepted; if not the hypothesis should be rejected, as the confidence interval provides evidence that $\alpha \neq c$.

We may carry out a hypothesis test to carry out whether there is significant correlation between two variables. We denote by ρ the true correlation; then to test the hypothesis

$$\begin{array}{ll} H_0 & : \rho = 0 \\ H_1 & : \rho \neq 0 \end{array}$$

we use the test statistic

$$t_r = r\sqrt{\frac{n-2}{1-r^2}}$$

which we compare with the null distribution which is Student-t with $n-2$ degrees of freedom. If $|t_r| > t_{n-2}(0.975)$, then we can conclude that the true correlation ρ is significantly different from zero.

3.6.8 Multiple Linear Regression

In everything that is described above, we have used a model in which we predicted a response Y from a single covariate X. This simple model can be extended to the case where Y is modelled as a function of p covariates $X_1, ..., X_p$, that is, we have the conditional expectation of Y given by

$$\mathrm{E}[Y|X_1 = x_1, ..., X_p = x_p] = \alpha + \beta_1 x_1 + ... + \beta_p x_p,$$

so that the observation model is given by

$$Y_i|X_1 = x_{i1}, ..., X_p = x_{ip} \sim N(\alpha + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \sigma^2).$$

Again, we can use maximum likelihood estimation to obtain estimates of the parameters in the model, that is, parameter vector $(\alpha, \beta_1, ..., \beta_p, \sigma^2)$, but the details are slightly more complex, as we have to solve $p+1$ equations simultaneously. The procedure is simplified if we write the parameters as a single vector, and perform matrix manipulation and calculus to obtain the estimates.

3.6.9 Worked Example

The following data are believed to follow a linear regression model;

x	0.54	2.03	3.15	3.96	6.25	8.17	11.08	12.44	14.04	14.34	18.71	19.90
y	11.37	11.21	11.61	8.26	14.08	16.25	11.00	14.94	16.91	15.78	21.26	20.25

We want to calculate estimates of α and β from these data. First, we calculate the summary statistics;

$$S_x = \sum_{i=1}^{n} x_i = 118.63 \qquad S_y = \sum_{i=1}^{n} y_i = 172.92 \qquad S_{xx} = \sum_{i=1}^{n} x_i^2 = 1598.6 \qquad S_{xy} = \sum_{i=1}^{n} x_i y_i = 1930.9$$

with $n = 12$ which leads to parameter estimates

$$\hat{\beta} = \frac{nS_{xy} - S_x S_y}{nS_{xx} - \{S_x\}^2} = \frac{12 \times 1930.9 - 118.63 \times 172.92}{12 \times 1598.6 - (118.63)^2} = 0.5201 \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 14.410 - 0.5201 \times 9.8842 = 9.269$$

This fit leads to the following fitted values and residuals;

x	0.54	2.03	3.15	3.96	6.25	8.17	11.08	12.44	14.04	14.34	18.71	19.90
y	11.37	11.21	11.61	8.26	14.08	16.25	11.00	14.94	16.91	15.78	21.26	20.25
$\hat{y}$	9.55	10.33	11.95	12.37	12.52	13.52	15.03	15.73	16.57	16.73	19.00	19.62
e	1.82	0.88	−0.34	−4.11	1.56	2.73	−4.03	−0.80	0.34	−0.95	2.26	0.63

The **corrected variance estimate**, s^2, is given by

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 3.438 \qquad \Longrightarrow \qquad s = 2.332$$

The **standard errors** for the two parameters are given by

$$s.e.(\hat{\alpha}) = s\sqrt{\frac{S_{xx}}{nS_{xx} - \{S_x\}^2}} = 1.304 \qquad s.e.(\hat{\beta}) = s\sqrt{\frac{n}{nS_{xx} - \{S_x\}^2}} = 0.113$$

The **t-statistics** for the two parameters are given by

$$t_\alpha = \frac{\hat{\alpha}}{s.e.(\hat{\alpha})} = \frac{9.269}{1.304} = 7.109 \qquad t_\beta = \frac{\hat{\beta}}{s.e.(\hat{\beta})} = \frac{0.520}{0.113} = 4.604.$$

The 0.975 percentile of a Student-t distribution with $n - 2 = 10$ degrees of freedom is found from tables to be 2.228. Both t-statistics are more extreme than this critical value, and hence it can be concluded that both parameters are significantly different from zero.

To calculate the **confidence intervals** for the two parameters. we need to use the 0.975 percentile of a $St(10)$ distribution. From above, we have that $St(10)(0.975) = 2.228$, and so the confidence intervals are given by

$$\alpha \ : \ \hat{\alpha} \pm t_{n-2}(0.975)\, s\sqrt{\frac{S_{xx}}{nS_{xx} - \{S_x\}^2}} = 9.269 \pm 2.228 \times 1.304 = (6.364 : 12.174)$$

$$\beta \ : \ \hat{\beta} \pm t_{n-2}(0.975)\, s\sqrt{\frac{n}{nS_{xx} - \{S_x\}^2}} = 0.5201 \pm 2.228 \times 0.113 = (0.268 : 0.772)$$

so that, informally, we are 95% certain that the true value of α lies in the interval (6.724 : 12.174), and that the true value of β lies in the interval (0.268 : 0.772). This amounts to evidence that, for example, $\alpha \neq 0$ (as the confidence interval for α does not contain 0), and evidence that $\beta \neq 1$ (as the confidence interval for β does not contain 1).

3.7 Model Validation

Techniques used for estimation and hypothesis testing allow specific and quantitative questions about the parameters in a probability model to be posed and resolved on the basis of a collection of sample data $x_1, ..., x_n$. However, the question as to the validity of the assumed probability model (for example, Binomial, Poisson, Exponential, Normal etc.) has yet to be addressed.

3.7.1 Probability Plots

The probability plotting technique involves comparing *predicted* and *observed* behaviour by comparing quantiles of the proposed probability distribution with sample quantiles. Suppose that a sample of data of size n are to be modelled using a proposed probability model with cdf F_X which possibly depends on unknown parameter(s) θ. The sample data are first sorted into ascending order, and then the ith datum, x_i, corresponds to the $100i/(n+1)$th quantile of the sample. Now, the equivalent *hypothetical* quantile of the distribution, q_i is found as the solution of

$$F_X(q_i) = \frac{i}{n+1} \qquad i = 1, ..., n.$$

If the model encapsulated in F_X is an acceptable model for the sample data, then for large n, $x_i \approx q_i$, so a plot of $\{(q_i, x_i) : i = 1, ..., n\}$ should be a straight line through the origin with slope 1. Hence the validity of F_X as a model for the sample data can be assessed through such a plot.

EXAMPLE For the *Exponential*(1) model, F_X is given by

$$F_X(x) = 1 - e^{-x} \qquad x \geq 0$$

so the probability plot consists of examining $\{(q_i, x_i) : i = 1, ..., n\}$ where

$$1 - e^{-q_i} = \frac{i}{n+1} \Longrightarrow q_i = -\log\left\{1 - \frac{i}{n+1}\right\}$$

EXAMPLE For the $N(0,1)$ model, $F_X \equiv \Phi$ is only available numerically (for example via statistical tables). Here the probability plot consists of examining $\{(q_i, x_i) : i = 1, ..., n\}$ where

$$\Phi(q_i) = \frac{i}{n+1} \Longrightarrow q_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$$

EXAMPLE For the *Exponential*(λ) model, we plot $\{(q_i, x_i) : i = 1, ..., n\}$ where

$$F_X(q_i) = 1 - e^{-\lambda q_i} = \frac{i}{n+1} \Longrightarrow q_i = -\frac{1}{\lambda}\log\left\{1 - \frac{i}{n+1}\right\}.$$

Hence, if we define q_i^* by

$$q_i^* = -\log\left\{1 - \frac{i}{n+1}\right\}$$

then if the model is correct, a plot of $\{(q_i^*, x_i) : i = 1, ..., n\}$ should be approximately a straight line through the origin with slope $1/\lambda$; hence λ can be estimated from this plot by using linear regression.

EXAMPLE For the $N(\mu, \sigma^2)$ model, is again only available numerically (for example via statistical tables). Here the probability plot consists of examining $\{(q_i, x_i) : i = 1, ..., n\}$ where

$$F_X(q_i) = \Phi\left(\frac{q_i - \mu}{\sigma}\right) = \frac{i}{n+1} \Longrightarrow q_i = \mu + \sigma\Phi^{-1}\left(\frac{i}{n+1}\right).$$

Hence, if we define q_i^* by

$$q_i^* = \Phi^{-1}\left(\frac{i}{n+1}\right)$$

then if the model is correct, a plot of $\{(q_i^*, x_i) : i = 1, ..., n\}$ should be approximately a straight line with intercept μ and slope σ; hence μ, σ can again be estimated from this plot by using linear regression.

3.7.2 The Chi-Squared Goodness-of-Fit Test

The problem of testing a hypothesis as to whether a data sample $x_1, ..., x_n$ is well-modelled by a specified probability distribution can be approached from a "goodness-of-fit" perspective.

Suppose that the data are recorded as the number of observations, O_i, say in a sample of size n that fall into each of k categories or "bins". Suppose that under the hypothesized model with mass/density function f_X or cdf F_X, the data follow a specific probability distribution specified by probabilities $\{p_i : i = 1, ..., k\}$. These probabilities can be calculated directly from f_X or F_X, possibly after parameters in the model have been estimated using maximum likelihood. Then, if the hypothesized model is correct, $E_i = np_i$ observations would be expected to fall into category i. An intuitively sensible measure of the **goodness-of-fit** of the data to the hypothesized distribution is given by the **chi-squared statistic**

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

A formal hypothesis test of model adequacy can be carried out in the usual framework; here the chi-squared statistic is the test statistic, and the null distribution (the distribution of the test statistic if the hypothesis is TRUE) is approximately a **chi-squared distribution** with $k - d - 1$ degrees of freedom, where d is the number of parameters in f_X or F_X that were estimated in order calculate the probabilities $p_1, ..., p_k$.

EXAMPLE : Testing the fit of a Poisson distribution

An early experiment into the properties of radioactive materials involved counting the number of alpha particles emitted from a radioactive source in 2612 consecutive 7.5 second intervals. A total of 10126 particles were counted, and the observed frequencies for each of the numbers of counts (per 7.5s) from 0 to 12 were recorded.

Count	O_i	p_i	E_i	$(O_i - E_i)^2/E_i$
0	57	0.021	54	0.167
1	204	0.080	210	0.171
2	383	0.156	407	1.415
3	525	0.201	525	0.000
4	532	0.195	510	0.949
5	408	0.151	395	0.428
6	273	0.098	255	1.271
7	139	0.054	141	0.028
8	49	0.026	68	5.309
9	27	0.011	30	0.300
10	10	0.004	11	0.091
11	4	0.002	4	0.000
12	2	0.000	1	1.000
>12	0	0.001	1	1.000
Total	2612	1.000	2612	12.129

To test the hypothesis that the data follow a Poisson distribution, a chi-squared test can be performed. First, we estimate Poisson parameter λ by its m.l.e., which is $\hat{\lambda} = \bar{x} = 10126/2612 = 3.877$. Secondly, we calculate probabilities p_i using the Poisson formula. Thirdly, we calculate the theoretical (expected) frequencies $E_i = np_i$ for each category. Finally, we calculate the χ^2 statistic as the sum of the (standardized) squared differences between observed and expected frequencies.

In this case, $\chi^2 = 12.129$. To complete the test we find that the 95th percentile of a Chi-squared distribution with $k - 1 - 1 = 12$ degrees of freedom is 21.03. This implies that the χ^2 statistic would only be surprising at a significance level of 0.05 if it was larger than 21.03. Here, as $\chi^2 = 12.129$, and therefore not surprising. Hence there is no evidence to indicate that the data are not from a Poisson distribution.

Clearly, the categorization is arbitrary, and several of the categories in example 1 could be combined. As a general rule, the categories should be chosen so that there is at least **five** observed counts in each.

EXAMPLE : Contingency Tables

One common use of the Chi-Squared Goodness-of-Fit test is in the context of **contingency tables**; a sample of data of size n are classified according to D factors, with each factor having k_d levels or categories, for $d = 1, ..., D$. When the classification is complete, the result can be represented by a D-way table of $k_1 \times k_2 \times ... \times k_D$ "cells", with each cell containing a fraction of the original data. For example, if $D = 2$, the table consists of k_1 rows and k_2 columns, and the number data in cell (i, j) is denoted n_{ij} for $i = 1, ..., k_1$ and $j = 1, ..., k_2$, where

$$\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} n_{ij} = n$$

It is often of interest to test whether row classification is **independent** of column classification, as this would indicate independence between row and column factors. This test is readily carried out using a Chi-Squared Goodness of Fit test; it is easy to show that, if the independence model is correct, the expected cell frequencies E_{ij} can be calculated as

$$E_{ij} = \frac{n_{i.} n_{.j}}{n} \qquad i = 1, ..., k_1, \; j = 1, ..., k_2$$

where $n_{i.}$ is the *total* of cell counts in row i and $n_{.j}$ is the *total* of cell counts in column j, and that, under independence, the χ^2 test statistic has an approximate chi-squared distribution with $(k_1 - 1)(k_2 - 1)$ degrees of freedom.

Summary

If a given hypothesis is true, it can be shown that the chi-squared statistic χ^2 for a sample of data has a particular Chi-squared distribution. If χ^2 takes a value that is surprising or unlikely under that probability distribution (for example if its value lies in the extreme right-hand tail and is larger, say, than the 95th percentile of the distribution) it is very likely that the hypothesis is false and should be be rejected.

In general, to test a hypothesis, consider a statistic calculated from the sample data. Derive mathematically the probability distribution of the statistic when the hypothesis is true, and compare the actual value of the statistic with the hypothetical probability distribution. Ask the question "Is the value a likely observation from this probability distribution ?". If the answer is "No", then reject the hypothesis.

3.7.3 Model Validation in Linear Regression

The adequacy of the fit of a linear regression model can be assessed on a **global** (overall) level using the R^2 (R-squared) statistic, which is calculated as the square of the correlation coefficient, r; R^2 quantifies the proportion of the total variation observed in the data that is explained by the regression (i.e. systematic) component of the model, as compared to the random variation. If the covariate X is a good predictor, then the majority of the total variation Y will be explained by variation in X, and so R^2 will be near 1.

The adequacy of the fit of a linear regression model can be assessed on a **local** (point-by-point) level by inspection of the *residuals*; recall that, if $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ is the **fitted value** of Y at $X = x_i$ for parameter estimates $\hat{\alpha}$ and $\hat{\beta}$, the residual e_i is the error in fit at each data point, so $e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$. If the regression model is adequate, then because of the assumptions underlying the model, the residuals $e_i, i = 1, ..., n$ should form an i.i.d. sample from a $N(0, \sigma^2)$ distribution, whose magnitude should not vary systematically with x_i, y_i or the fitted value $\hat{y}_i$; this could be checked using simple scatterplots. The normality of the residuals can be checked using, for example, a probability plot. Clearly, in practice, the random variation in the model, as encapsulated in variance σ^2, is unknown, but can be estimated using the corrected error variance estimate

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \; = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Using this estimate, it is possible to compute **standardized residuals**

$$e_i^* = \frac{y_i - \hat{y}_i}{s}$$

which can be shown to follow a Student-t distribution with $n - 2$ degrees of freedom. The advantage of using standardized residuals is that they have, approximately, variance equal to 1.