– LINEAR REGRESSION MODELS – Likelihood and Reference Bayesian Analysis

Mike West

May 3, 1999

1 Straight Line Regression Models

We begin with the simple straight line regression model

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where the design points x_i are fixed in advance, and the measurement/sampling errors ϵ_i are independent and normally distributed, $\epsilon_i \sim N(0, \sigma^2)$ for each $i = 1, \ldots, n$. In this context, we have looked at general modelling questions, data and the fitting of least squares estimates of α and β. Now we turn to more formal likelihood and Bayesian inference.

1.1 Likelihood and MLEs

The formal parametric inference problem is a multi-parameter problem: we require inferences on the three parameters $(\alpha, \beta, \sigma^2)$. The likelihood function has a simple enough form, as we now show. Throughout, we do not indicate the design points in conditioning statements, though they are implicitly conditioned upon. Write $Y = \{y_1, \ldots, y_n\}$ and $X = \{x_1, \ldots, x_n\}$. Given X and the model parameters, each y_i is the corresponding zero-mean normal random quantity ϵ_i plus the term $\alpha + \beta x_i$, so that y_i is normal with this term as its mean and variance σ^2. Also, since the ϵ_i are independent, so are the y_i. Thus

$$(y_i | \alpha, \beta, \sigma^2) \sim N(\alpha + \beta x_i, \sigma^2)$$

with (conditional) density function

$$p(y_i | \alpha, \beta, \sigma^2) = \exp\{-(y_i - \alpha - \beta x_i)^2 / 2\sigma^2\} / (2\pi\sigma^2)^{1/2}$$

for each i. Also, by independence, the joint density function is

$$p(Y | \alpha, \beta, \sigma^2) = \prod_{i=1}^{n} p(y_i | \alpha, \beta, \sigma^2).$$

Given the observed response values Y this provides the likelihood function for the three parameters: the joint density above evaluated at the observed and hence now fixed values of Y, now viewed as a function of α, β, σ^2 as they vary across possible parameter values. It is clear that this likelihood function is given by

$$p(Y|\alpha, \beta, \sigma^2) \propto \exp(-Q(\alpha, \beta)/2\sigma^2)/\sigma^n$$

where

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

and where a constant term $1/(2\pi)^{n/2}$ has been dropped.

Let us look at computing the *joint maximum likelihood estimates (MLEs)* of the three parameters. This involves finding the values $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ such that $p(Y|\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2) > p(Y|\alpha, \beta, \sigma^2)$ for any other parameter values $(\alpha, \beta, \sigma^2)$. We do this as follows.

- For any fixed value of σ^2, the likelihood function is a strictly increasing function of $Q(\alpha, \beta)$. Hence, changing (α, β) to decrease the value of Q implies that the value of the likelihood function increases. Clearly, choosing (α, β) to minimise Q implies that we maximise the likelihood function. As a result, the MLEs of (α, β) for any specific value of σ^2 are simply the LSEs.

- From the earlier discussion of least square estimation, we know that the LSEs $(\hat{\alpha}, \hat{\beta})$ do not, in fact, depend on the value of the variance σ^2. As a result, the full three-dimensional maximisation is solved at the LSE values $(\hat{\alpha}, \hat{\beta})$ and by choosing $\hat{\sigma}^2$ to maximise $p(Y|\hat{\alpha}, \hat{\beta}, \sigma^2)$ as a function of just σ^2. This trivially leads to

  $$\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\epsilon}_i^2/n$$

  where $\hat{\epsilon}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$ for each i; this MLE of σ^2 is the usual residual sum of squares with divisor n.

1.2 Reference Bayesian Analyses

1.2.1 Parametrisation and Reference Prior

To develop the reference Bayesian posterior distribution for $(\alpha, \beta, \sigma^2)$ it is traditional to *reparameterise* from σ^2 to the *precision* parameter $\phi = 1/\sigma^2$. This is done simply for clarity of exposition and ease of development. Plugging $\sigma^2 = 1/\phi$ in the likelihood function leads simply to

$$p(Y|\alpha, \beta, \phi) \propto \phi^{n/2} \exp(-\phi Q(\alpha, \beta)/2)$$

– defining the joint likelihood function for the three parameters (α, β, ϕ).

Bayesian inference requires a prior $p(\alpha, \beta, \phi)$ – a trivariate prior density defined on the model parameter space. Here we use the traditional reference prior in which:

- The three parameters are independent, so $p(\alpha, \beta, \phi) = p(\alpha)p(\beta)p(\phi)$.
- As a function of α, the log-likelihood function is quadratic, indicating that the likelihood function will contribute a normal form in α. Hence a normal prior for α would be conjugate, leading to a normal posterior. On this basis, a normal prior with an extremely large variance (as in reference analysis of normal models) will represent a vague or uninformative prior position. Taking the formal limit of a normal prior with a variance tending to infinity provides the traditional reference prior $p(\alpha) \propto constant$.
- The same reasoning applies to β, leading to the traditional reference prior $p(\beta) \propto constant$.
- As a function of ϕ alone, the likelihood function has the same form as that of a gamma density function in ϕ. Hence a gamma prior for ϕ would be conjugate, leading to a gamma posterior. On this basis, a gamma prior with very small defining parameters (as in reference analysis of Poisson or exponential models) will represent a vague or uninformative prior position. Taking the formal limit of the $\phi \sim Gamma(a, b)$ prior at $a = b = 0$ provides the traditional reference prior $p(\phi) \propto \phi^{-1}$.

Combing these components produces the standard non-informative/reference prior $p(\alpha, \beta, \phi) \propto \phi^{-1}$. Then Bayes' theorem leads to the reference posterior

$$p(\alpha, \beta, \phi|Y) \propto p(\alpha, \beta, \phi)p(Y|\alpha, \beta, \phi) \propto \phi^{n/2-1} \exp(-\phi Q(\alpha, \beta)/2),$$

over real-valued α and β and $\phi > 0$. This is a joint density for the three quantities, and reference inference follows by exploring and summarising its properties. Notice that the posterior is almost the normalised likelihood function – the only difference is in the prior term ϕ^{-1}. We quote key features of this reference posterior in the following sections.

1.2.2 Marginal Reference Posterior for ϕ

The marginal posterior for ϕ is available by integrating (α, β), i.e.,

$$p(\phi|Y) = \int p(\alpha, \beta, \phi|Y) d\alpha d\beta$$

where the range of integration is $-\infty < \alpha, \beta < \infty$. It can be shown that this yields the simple form

$$p(\phi|Y) \propto \phi^{a-1} \exp\{-b\phi\}$$

where $a = (n-2)/2$ and $b = \sum_{i=1}^{n} \hat{\epsilon}_i^2/2$. As a result, the posterior for ϕ is simply $Gamma(a, b)$ with these values of (a, b). In particular, the posterior mean is

$$E(\phi|Y) = 1/s^2$$

where

$$s^2 = \sum_{i=1}^{n} \hat{\epsilon}_i^2/(n-2).$$

Since $E(\phi|Y)$ is a point estimate of $\phi = 1/\sigma^2$, then s^2 is a corresponding point estimate of σ^2. It is referred to as the residual variance estimate, as it is a sample variance computed from the fitted residuals $\hat{\epsilon}_i$. Note that, unlike the MLE $\hat{\sigma}^2$, the estimate s^2 has a divisor $n-2$. The common-sense interpretation of this is that the effective number of observations is the actual total n reduced by the number of fitted parameters, here just two. The term $n-2$ is called the residual degrees of freedom, reflecting this adjustment from n, the initial degrees of freedom. One implication is that $s^2 > \hat{\sigma}^2$, reflecting a more conservative estimate of variance after accounting for the estimation of the two parameters.

1.2.3 Marginal Reference Posterior for (α, β)

The marginal posterior for (α, β) is obtained as

$$p(\alpha, \beta|Y) = \int p(\alpha, \beta, \phi|Y) d\phi$$

and turns out to be a bivariate T distribution. Of key practical relevance are the implied univariate margins for posterior for β and linear functions of (α, β) alone, and these are all univariate T distributions (see Appendix for details of T distributions). Specific univariate margins are as follows.

- Define $v_\beta^2 = 1/S_{xx}$. The univariate posterior for β has a density function

  $$p(\beta|Y) \propto \{(n-2) + (\beta - \hat{\beta})^2/(s^2 v_\beta^2)\}^{-(n-1)/2},$$

  and this is the density of a T distribution with $n-2$ degrees of freedom, mode $\hat{\beta}$ and scale sv_β. By way of notation we have

  $$(\beta|Y) \sim T_{n-2}(\hat{\beta}, s^2 v_\beta^2).$$

  As long as $n > 4$, it is also true that $E(\beta|Y) = \hat{\beta}$ and $V(\beta|Y) = cs^2 v_\beta^2$ with $c = (n-2)/(n-4)$. The posterior is symmetric and normal shaped about the mode, though has heavier tails than a normal posterior. We can write

  $$\beta = \hat{\beta} + (sv_\beta)t \qquad \text{and} \qquad t = (\beta - \hat{\beta})/(sv_\beta)$$

  where the random quantity $t \sim T_{n-2}(0,1)$. Posterior probabilities and intervals for β follow from those of the Student T distribution: if t_p is the $100p\%$ quantile of t, then that for β is simply $\hat{\beta} + (sv_\beta)t_p$. The term sv_β is called the posterior *standard error* of the β coefficient.

  For large degrees of freedom, the Student T distribution approaches the standard normal, in which case we have the approximation $t \sim N(0,1)$ and so

  $$(\beta|Y) \sim N(\hat{\beta}, s^2 v_\beta^2).$$

Otherwise, we can view the distribution informally as "like the normal but with a little bit of additional uncertainty."

- The univariate margin for α is similarly a T distribution with $n-2$ degrees of freedom, mode $\hat{\alpha}$ and scale sv_α where $v_\alpha^2 = n^{-1} + \bar{x}^2/S_{xx}$, i.e.,
$$(\alpha|Y) \sim T_{n-2}(\hat{\alpha}, s^2v_\alpha^2).$$
- Under $p(\alpha, \beta|Y)$ the two parameters are generally correlated. Assuming $n > 4$ so that second moments of the T distribution exist, the posterior covariance is $s^2c_{\alpha,\beta}(n-2)/(n-4)$ where $c_{\alpha,\beta} = -\bar{x}/S_{xx}$. Note that $(n-2)/(n-4) \approx 1$ when n is large, when the posterior is approximately normal; in that case, the posterior covariance is just the term $s^2c_{\alpha,\beta}$ above. Note also that the covariance, hence the correlation, is zero if and only if $\bar{x} = 0$. This correlation is relevant when considering posterior inferences and predictions involving linear functions of the two parameters, as now follows.

1.2.4 Intervals for β and Significance of the Regression

The standard T test of the significance of the regression fit can be understood as an assessment of the support for the value $\beta = 0$ under the marginal posterior distribution. Note that, if the assumption of a linear regression model is really appropriate, then $\beta = 0$ would imply no relationship between the x and y variables, whereas large values of $|\beta|$ imply a strong relationship.

Under the symmetric posterior distribution $(\beta|Y) \sim T_{n-2}(\hat{\beta}, s^2v_\beta^2)$, HPD intervals coincide with equal-tails intervals. Specifically, the $100(1-p)\%$ posterior interval is
$$\hat{\beta} \pm (sv_\beta)t_{p/2}$$
for any probability p and where $t_{p/2}$ is the $100(p/2)\%$ quantile of $T_{n-2}(0,1)$. For example, $p = 0.05$ means that $t_{0.025}$ is the (lower) 2.5% point of the standard T distribution, and the above interval is a 95% (equal tails, HPD) interval for β. If the value $\beta = 0$ lies outside this interval, then we are assured that the regression is significant at (at least) the 95% level; that is, $\beta = 0$ is among the 5% least likely values of the parameter.

The standard $p-$value for the test of $\beta = 0$ can be interpreted as the posterior probability of β values that have lower posterior density than $\beta = 0$. To do this we need to find the probability outside the HPD interval defined with one end-point at $\beta = 0$. From the form of the posterior density function, it easily follows (as detailed in the Appendix) that this exclude all β values such that $|t| > |\hat{\beta}/sv_\beta|$ where $t = (\beta - \hat{\beta})/sv_\beta$. The resulting "tail-area" $p-$value is therefore simply
$$p - \text{value} = 2Pr(t > |\hat{\beta}/sv_\beta|)$$
where $t \sim T_{n-2}(0,1)$. The observed quantity $\hat{\beta}/sv_\beta$ here is called the *standardised T test statistic* – it is simply the size of the estimated coefficent relative to

its standard error. It is tempting to refer to the $p-$value as the "probability of $\beta = 0$" although it is obviously not quite that; it is, however, a standard measure of the significance of the fit of the model, with a low $p-$value commensurate with a significant fit.

1.2.5 F Tests, ANOVA and Deviances

It is traditional to summarise the statistical significance of the regression fit with a summary *F test* and the associated *analysis of deviance*, historically called the *analysis of variance*, or ANOVA. This adds nothing new methodologically; rather, the F test and anova presentation is simply another way of summarising the goodness of fit in terms of R^2 and the $p-$value based on the T test for $\beta = 0$, as discussed in the previous section. One reason for restating the conclusions in these different terms is simply historical precendent. A second, more important reason is that the extension of significance assessment to multiple regression models – models with more than one predictor variable – cannot be done with T tests alone, and inherently involves ANOVA ideas. Hence it is useful to explain the connection in this simplest of cases.

Begin by noting that the $p-$value from the T test of the significance of the regression is equivalent to

$$p - \text{value} = Pr(F > (\hat{\beta}/sv_\beta)^2)$$

where $F = t^2$, a random variable obtained as the square of $t \sim T_{n-2}(0,1)$. The name of the distribution of F is the *F distribution with* $(1, n-2)$ *degrees of freedom*; we write $F \sim F_{1,n-2}$. Such distributions are well known, tabulated and available in computer software. Hence we could use the upper tail-area of the $F_{1,n-2}$ to compute the $p-$value, as an alternative to the $T_{n-2}(0,1)$.

Write

$$f_{obs} = (\hat{\beta}/sv_\beta)^2 = \frac{\hat{\beta}^2/v_\beta^2}{s^2}.$$

It is easily shown (see exercises) that the numerator term in f_{obs} reduces to

$$\frac{\hat{\beta}^2}{v_\beta^2} = S_{yy} - (n-2)s^2$$

where $(n-2)s^2 = \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ is the residual sum of squares from the model fit. As a result, it is trivial to compute

$$f_{obs} = (S_{yy} - (n-2)s^2)/s^2$$

and the implied $p-$value.

The above expression is intuitively reasonable. The term $S_{yy} - (n-2)s^2$ is called the *fitted sum of squares* or the *variation explained* by the regression of y on x. This reflects the fact that it is the difference between the total variation in the response data (S_{yy}) and the residual variation remaining after the model has been fitted. If the variation explained is large, then f_{obs} will be large and the $p-$value small; just how large the reduction in variation due to the model must be is relative to the scale of the data, so that the variation explained is assessed relative to s^2 as is clear in the formula for f_{obs}.

In terms of the quadratic notation $Q(\alpha,\beta)$, recall that $S_{yy} = Q(\bar{y},0)$ and the residual sum of squares is $(n-2)s^2 = Q(\hat{\alpha},\hat{\beta})$. So, in this notation, the explained variation is simply $Q(\bar{y},0)-Q(\hat{\alpha},\hat{\beta})$. In some areas of modern regression analysis, these quadratic, sums of squares measures are referred to as *deviances.* Thus $Q(\bar{y},0)$ is the *total deviance*, and $Q(\hat{\alpha},\hat{\beta})$ is the *residual deviance* from the fitted model; the difference is the *deviance explained* by the regression on x, or the *reduction in deviance* due to the model. ANOVA, the analysis of variance, is simply a name for the representation of total variability in the response data using the above components. In modern times the term *analysis of deviance* is more apt, as it generalises to non-normal and non-linear regression models. The summary is simple: *Total deviance* = *Deviance explained* + *Residual deviance,* and the F test measures the significance of the model fit by assessing just how large the "deviance explained" here is.

1.2.6 Honest Prediction

Consider predicting a new case y_{n+1} at a further, or future, design point x_{n+1}. Formally, this requires the evaluation of the *posterior predictive distribution* $p(y_{n+1}|Y)$, i.e., simply the distribution for the new value conditional on the data observed so far (as above, the x_i values are implicit and ignored in the notation). It can be shown that this leads to a T distribution, and we denote this by

$$(y_{n+1}|Y) \sim T_{n-2}(\hat{y}, s^2 v_y^2)$$

where

- $\hat{y} = \hat{\alpha} + \hat{\beta} x_{n+1}$, just the fitted line at x_{n+1}, and
- $v_y^2 = 1 + w^2$ with

  $$w^2 = v_\alpha^2 + x_{n+1}^2 v_\beta^2 + 2x_{n+1} c_{\alpha,\beta},$$

  and where $c_{\alpha,\beta} = -\bar{x}/S_{xx}$ was given above in the posterior covariance between α and β.

It is easy to motivate this by considering the model directly, that is

$$y_{n+1} = \alpha + \beta x_{n+1} + \epsilon_{n+1}$$

where ϵ_{n+1} is the deviation from the line for the new case and so is independent of past data. Taking expectations across this equation gives $E(y_{n+1}|Y) = \hat{y}$, using linearity of expectations. The scale factor sv_y is similarly computed; note that the posterior correlation between α and β enters into its computation.

Note the following:

- For large n, the predictive distribution is close to normal, in which case $(y_{n+1}|Y) \approx N(\hat{y}, s^2 v_y^2)$.

- Since $v_y^2 = 1 + w^2$ we have $s^2 v_y^2 = s^2 + s^2 w^2$. Here the first s^2 is the estimate of the variance σ^2 of ϵ_{n+1}, whereas the term $s^2 w^2$ measures the uncertainty about the line $\alpha + \beta x_{n+1}$ at the new design point.

- With very large data sets the posterior for (α, β) will be quite precise, and w will be small, in which case the posterior is almost $N(\hat{y}, s^2)$. Otherwise, the predictions using the above results will be "honest" in the sense that the additional term $s^2 w^2$ appropriately reflects the additional uncertainty in predicting due to uncertainty about the regression line parameters.

- It can be shown that the above formula for w^2 can be reduced to

  $$w^2 = n^{-1} + (x_{n+1} - \bar{x})^2 / S_{xx}.$$

  As a result, w^2 will be small when x_{n+1} is close to $\bar{x}$, but grows larger for larger values of $|x_{n+1} - \bar{x}|$. This means that the spread of the predictive distribution is greater for new design points that are further away from the past design points; in this sense, predictions also appropriately reflect increased uncertainty in extrapolating outside the range of past experience. Nevertheless, extrapolation must always be cautioned, unless substantive theory exists to indicate the validity of extending the straight line regression assumption into regions where no data are available.

In practice, computer programs work with matrix formulations of linear models so that the explicit details of the calculations above are never needed for actual numerical work. They are important, however, for understanding and interpretation.

2 Sampling Theoretic Inference

The standard sampling-theoretic inference framework leads to essentially the same methods – the same point estimates of $(\alpha, \beta, \sigma^2)$, the same intervals and tests – in terms of numerical values, although, of course, their interpretations and rationale are quite different to those in the Bayesian framework. The sampling theory results are summarised here for comparison.

Under the assumption that the straight line model generates the data, consider the *statistics* $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ as functions of the random quantities Y prior to their being observed. These have a joint sampling distribution $p(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2 | \alpha, \beta, \sigma^2)$ that is well-known (e.g., De Groot, chapter 10). Remember that this is the repeat-sampling distribution of the estimators based on the model assuming the parameter fixed at their "true" values. This trivariate sampling distribution has the following features.

- The estimators are unbiased; that is, $E(\hat{\alpha}|\alpha) = \alpha$, $E(\hat{\beta}|\beta) = \beta$ and $E(\hat{\sigma}^2|\sigma^2) = \sigma^2$. They are also consistent, in that they converge (in probability) to their respective parameters as n tends to infinity.
- The statistic $(\hat{\alpha} - \alpha)/sv_\alpha$ is distributed as $T_{n-2}(0, 1)$. Note that this describes sampling variability in both $\hat{\alpha}$ in the numerator and s in the denominator. As a result, confidence intervals for α have the form

  $$\hat{\alpha} \pm (sv_\alpha)t_{p/2}$$

  for any probability p and where $t_{p/2}$ is the $100(p/2)\%$ quantile of $T_{n-2}(0, 1)$.
- The statistic $(\hat{\beta} - \beta)/sv_\beta$ is also distributed as $T_{n-2}(0, 1)$. Note that this describes sampling variability in both $\hat{\beta}$ in the numerator and s in the denominator. As a result, confidence intervals for β have the form

  $$\hat{\beta} \pm (sv_\beta)t_{p/2}$$

  for any probability p and where $t_{p/2}$ is the $100(p/2)\%$ quantile of $T_{n-2}(0, 1)$.
- The statistic s^2 has a sampling distribution that depends only on σ^2, and is given by $(s^2|\sigma^2) \sim Gamma((n-2)/2, (n-2)/2\sigma^2)$. Equivalently, the quantity $s^2(n-2)/\sigma^2$ is distributed as $Gamma((n-2)/2, 1/2)$ which is the chi-squared distribution with $n-2$ degrees of freedom.

Note that the point estimates and confidence intervals for the regression parameters coincide numerically with those in the reference Bayesian analysis, and that the same point estimate of σ^2 is generated too. These numerical equivalences extend to predictions.

On the issue of testing the significance of the regression model fit, the sampling-theoretic interpretation of the $p-$value may be based on the following reasoning. Condition on the hypothesis that $\beta = 0$, so that the sampling distribution theory above implies $\hat{\beta}/sv_\beta \sim T_{n-2}(0, 1)$. Write $t = \hat{\beta}/sv_\beta$. Then large values of the statistic $|t|$ are rare if the hypothesis that $\beta = 0$ is actually true, and just how rare can be measured by probabilities with respect to this sampling distribution. Write t_{obs} for the observed value of $\hat{\beta}/sv_\beta$ in this data set. Then the probability of observing a future, hypothetical data set Y that gives rise to values of t that is at least as extreme as t_{obs} is simply

$$Pr(|t| > |t_{obs}|) = 2Pr(t > |t_{obs}|),$$

and this is easily computed from the standard $T_{n-2}(0,1)$ distribution of t. Notice that, though the conceptual basis for this calculation is radically different from the conditional Bayesian approach, the numerical results are again the same. The above is the standard sampling theory, or frequentist, definition of the *observed significance level* or $p-$value of the test of the hypothesis that $\beta = 0$. It coincides with the Bayesian $p-$value based on tail areas under the posterior density for β. The numerical correspondence of the associated F test follows immediately.

It turns out that this test has other optimality properties from a sampling theoretic viewpoint. In particular, it is a *likelihood ratio test* of the hypothesis $\beta = 0$ compared to all other values of β, and a *uniformly most powerful test* in addition (De Groot, chapter 10).

3 Multiple Linear Regression Models: Summary

The course slides provide coverage of the multiple linear regression model, notation, mathematical structure, examples, and theory. This section is a supplement on the basic theoretical results (no proofs) related to the reference posterior distribution and its uses.

The model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots \beta_p x_{ip} + \epsilon_i$$

for observations $i = 1, \dots, n$. In vector/matrix notation we have

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i$$

and

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

where:

- the $n \times 1$ response vector $\mathbf{y}$ has elements y_i;
- the $n \times k$ design matrix $\mathbf{X}$ has rows $\mathbf{x}_i'$ where

  $$\mathbf{x}_i' = (1, x_{i1}, \dots, x_{ip})$$

  and, of course, $k = p + 1$;
- the $k \times 1$ regression parameter vector $\boldsymbol{\beta}$ has elements β_i.
- the $n \times 1$ error (or deviation) vector $\boldsymbol{\epsilon}$ has elements ϵ_i.

In detail,

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

and

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}.$$

Row i of $\mathbf{X}$ is just $\mathbf{x}_i'$, containing the values of all predictor variables (in order) for observation i. Column j of $\mathbf{X}$ contains the n values of predictor variable j, except for when $j = 1$ when it is just a column with all entries 1, corresponding to the intercept term of the model.

4 Multiple Linear Regression Models: Reference Posterior Distribution

The reference posterior $p(\boldsymbol{\beta}|Y)$ is a multivariate T distribution. The parameter vector $\boldsymbol{\beta}$ contains k parameters, so the posterior is a joint distribution for these k parameters, or a $k-$variate distribution. The multivariate T distribution has $n-k$ degrees of freedom, and we write it as

$$(\boldsymbol{\beta}|Y) \sim T_{n-k}(\hat{\boldsymbol{\beta}}, s^2\mathbf{V})$$

with the following ingredients:

- The dimension k is implicit, and not made explicit in the notation;
- $\mathbf{V}$ is the $k \times k$ matrix defined by

  $$\mathbf{V} = (\mathbf{X}'\mathbf{X})^{-1},$$

  and arises in the formula for
- $\hat{\boldsymbol{\beta}}$, the LSE of $\boldsymbol{\beta}$, defined by

  $$\hat{\boldsymbol{\beta}} = \mathbf{V}\mathbf{X}'\mathbf{y};$$
- s^2 is the residual estimate of the variance σ^2, given by

  $$s^2 = Q(\hat{\boldsymbol{\beta}})/(n-k)$$

  where

  $$Q(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

  is the residual sum of squares of the model fit, or the residual deviance, based on fitted residuals

  $$\hat{\epsilon}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$$

  for each i. As in the straight line model, s^2 is the usual estimate of σ^2, computed as a sample variance in which the sample values are the fitted residuals, and the denominator $n-k$ corrects the sample size n by subtracting the number of parameters estimated in $\boldsymbol{\beta}$.

4.1 Univariate marginal posteriors

One of the key properties of the joint posterior distribution above is that the implied univariate marginal posterior for any element β_i of $\boldsymbol{\beta}$ is a Student T distribution. Specifically, for $i = 1, \dots, k$,

$$(\beta_i|Y) \sim T_{n-k}(\hat{\beta}_i, s^2 v_i^2)$$

where $\hat{\beta}_i$ is the corresponding estimate from $\hat{\boldsymbol{\beta}}$ and v_i^2 is the corresponding diagonal element of $\mathbf{V}$, or $v_i^2 = (\mathbf{V})_{ii}$.

Inference on any one parameter individually involves summarising this T distribution, in the usual ways.

4.2 Other marginal posteriors

A second key property of the joint posterior distribution above is that the implied multivariate marginal posterior for any subset of r elements is also a multivariate T distribution. This is of most interest in connection with developing measures of importance of subsets of predictor variables in explaining variation observed in the response variable. We return to this below in discussing *subset F tests* of such questions.

4.3 Honest prediction

A further key implication of the model and its analysis is that predictive distributions for new/future response variables are also T distributions. Specifically, consider a future predictor vector $\mathbf{x}_{n+1}$. The posterior predictive distribution for the associated future response outcome

$$y_{n+1} = \mathbf{x}'_{n+1}\boldsymbol{\beta} + \epsilon_{n+1}$$

is given by

$$(y_{n+1}|Y) \sim T_{n-k}(\hat{y}_{n+1}, s^2 v_y^2)$$

where

- the point prediction is simply

  $$\hat{y}_{n+1} = \mathbf{x}'_{n+1}\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1,1} + \ldots + \hat{\beta}_p x_{n+1,p},$$

  or just the value of the fitted regression function at the new design point;
- $v_y^2 = 1 + w^2$ where $w^2 = \mathbf{x}'_{n+1}\mathbf{V}\mathbf{x}_{n+1}$.

The term sv_y is called the predictive standard error. Note that w, and hence v_y, depend on $\mathbf{x}_{n+1}$ (though the notation does not make this explicit), and this dependence is important in reflecting differing degrees of uncertainty about y_{n+1} at different design points. In particular, as in the simple straight line regression model, predictions at new points $\mathbf{x}_{n+1}$ that are far from the region of previous experience will have higher standard errors since w, and hence v_y, will be larger. Further, note that the term w appears to reflect the estimation uncertainty – the uncertainty due to lack of precise knowledge of $\boldsymbol{\beta}$. With large sample sizes, the posterior for $\boldsymbol{\beta}$ will be very concentrated about its mean of $\hat{\boldsymbol{\beta}}$, $\mathbf{V}$ will have small elements, and hence w will be small. In such cases, two effects arise: first, $s_y \approx 1$ so that the predictive standard error will be approximately s^2; secondly, the T distribution, with a very large degrees of freedom, will be approximately normal. In such cases, therefore, we may use the simpler approximate predictive distribution

$$(y_{n+1}|Y) \approx N(\hat{y}_{n+1}, s^2).$$

5 Assessing subsets of predictors

5.1 Posterior assessment of parameters

Standard *subset F tests* provide numerical measures of the contributions of subsets of predictor variables to the overall fit of the model. Consider a specific subset of $r < p$ predictors to be assessed. Suppose these predictors are those numbered $j_1, \ldots, j_r$, so that the corresponding regression parameter are

$$\boldsymbol{\gamma} = \begin{pmatrix} \beta_{j_1} \\ \vdots \\ \beta_{j_r} \end{pmatrix}.$$

Under the reference posterior distribution, the $r-$vector parameter $\boldsymbol{\gamma}$ has a marginal multivariate T_{n-k} distribution. One way to assess how important the r predictors in question are is to assess how much support this posterior gives to values near $\boldsymbol{\gamma} = \mathbf{0}$. The reasoning is parallel to that used in univariate cases: if the point $\boldsymbol{\gamma} = \mathbf{0}$ is unsupported by the data, and so well out "in the tails" of the posterior, then the corresponding predictors play a meaningful role in the model fit. If, on the other hand, the point $\boldsymbol{\gamma} = \mathbf{0}$ is close to the mean of the posterior and so is a likely value, the predictors have less relevance. Notice that this is explored with no reference to the other predictors – it is therefore to be understood that this assessment of $\boldsymbol{\gamma}$ is conditional upon the other covariates being in the model.

To formally assess the support for $\boldsymbol{\gamma} = \mathbf{0}$ we

- identify all values of $\boldsymbol{\gamma}$ that have lower posterior density than $\boldsymbol{\gamma} = 0$, and then
- find the posterior probability of those values to deliver a $p-$value to assess just how extreme $\boldsymbol{\gamma} = \mathbf{0}$ is.

The formal interpretation of such a $p-$value is that it is the posterior probability on $\boldsymbol{\gamma}$ values that lie outside the *highest posterior density region* defined by $\boldsymbol{\gamma} = \mathbf{0}$; the common-sense terminology is that the $p-$value is the probability on values of $\boldsymbol{\gamma}$ less well-supported by the data than $\boldsymbol{\gamma} = \mathbf{0}$. A small $p-$value indicates that $\boldsymbol{\gamma}$ is most probably different to $\mathbf{0}$ and so indicates that the r predictors in question contribute significantly to the model fit.

Standard theory tells us that the resulting $p-$value may be easily computed as a tail-area under a univariate F distribution. The details and derivation are not given here, but the formula of the threshold for the F distribution is both easy to remember and important to interpret. The result is that

$$p - value = Pr(F > f_{obs}),$$

where the random variable F has an F distribution on r and $n - k$ degrees of freedom, or $F \sim F_{r,n-k}$, and where f_{obs} is the threshold value f_{obs} computed from the posterior. The formula for f_{obs} is discussed in the following subsection. Note that the $p-$value is trivially computed from software packages with cumulative distribution functions of F distributions – it is just $1 - P(f_{obs})$ where P here stands for the cdf of $F_{r,n-k}$.

5.2 F test statistics

Call the model above Model A. In Model A, we have p predictor variables and are interested in whether or not a specific subset of r of them are really relevant in terms of statistical measures of fit. Consider a different linear model, *Model B*, that is just Model A with the r predictor variables in question removed. We can trivially fit Model B as well, compute the usual posterior distributions and numerical summaries, and use these to make comparisons with Model A. Model A has more predictor variables than Model B: Model A has p predictors, but Model B has only $p - r$. In addition, we say that Model B is *nested within* Model A, since it arises as a special case of Model A when $\gamma = 0$. If the r predictors in question are really not relevant in describing observed variations in the response data, then the two models will produce similar fits. If, on the other hand, these specific predictors are really related to the response, Model A will produce a different and better fit, better in the sense of explaining more observed variation. To this end, we compare the residual sums of squares, or deviances, from the two models.

Write Q_A and Q_B for the residual deviances in Models A and B, respectively. We can interpret $Q_B - Q_A$ as the decrease in deviance in moving from the simpler Model B to the more elaborate Model A. A large difference is indicative of a significantly improved fit, suggesting the r predictors are relevant. To measure how large the decrease in deviance is, we first note that the reduced deviance under Model A is achieved at the cost of r extra parameters, so that we might better consider the change in deviance per parameter, or $(Q_B - Q_A)/r$. Second, deviances are measured on a scale that is the square of the response, or that of σ^2 and its estimates; to standardise to a dimensional measure we divide by the best estimate of σ^2 available, namely s_A^2, the residual estimate of variance under the larger Model A. This leads us to the standardised, per parameter "reduction in deviance" measure $(Q_B - Q_A)/rs_A^2$. as a natural summary of just how much the r predictors in question improve model fit in the context of the other predictors already in Model B.

The F distribution theory mentioned in the previous section identifies just this standardised deviance measure as the critical threshold value, i.e.,

$$f_{obs} = (Q_B - Q_A)/rs_A^2$$

or

$$f_{obs} = \frac{difference\ in\ deviance/difference\ in\ number\ of\ parameters}{residual\ estimate\ of\ variance}$$

where the "differences" in deviances and number of parameters compare elaborating the data description from Model B to Model A, "residual estimate of variance" is that in the more elaborate Model A. The resulting $p-$value simply converts this common-sense measure of "difference" between the two models to a probability scale.

Finally, note the special case when we consider $r = p$ and assess all predictor variables. In this case, the F test is a test of overall regression fit, with the simpler "baseline" Model B being the random sample model with no predictors. Many computer packages routinely generate the f_{obs} threshold, and its corresponding $p-$value under the $F_{p,n-p-1}$ distribution, as a summary of overall model fit.

5.3 Cautionary note

As usual, it is important to be aware that $p-$values will tend to be smaller for larger sample sizes, the F test being more "powerful" in identifying increasingly small deviations of $\boldsymbol{\gamma}$ from $\mathbf{0}$ when n is very large. Hence it is important to be aware that, if dealing with very large samples, most subsets of potential predictors may be found to be statistically significant even though their inclusion in a model may lead to only minor changes in inferences and predictions. Generally, and again as in all statistical models, choice of candidate predictor variables should be based as much as possible on underlying scientific relevance and validity, and on empirical performance in out-of-sample predictions.

Appendix: T distributions

A real-valued random quantity t has a standard Student T distribution with $k > 0$ degrees of freedom if the density function is

$$p(t) \propto \{k + t^2\}^{-(k+1)/2}$$

with normalising constant $a = k^{k/2}\Gamma((k+1)/2)/\{\sqrt{\pi}\Gamma(k/2)\}$. We write $t \sim T_k(0,1)$. The density is symmetric about zero with a shape similar to the normal density curve. For high values of k, say exceeding 20, the density is very close to normal. For lower values of k, the density is heavier-tailed than normal, giving higher probability to more extreme values. If $k > 1$ the mean exists and is $E(t) = 0$. If $k > 2$ the variance exists and is $V(t) = k/(k-2)$; for large k, the variance approaches 1 as the density approaches the standard normal.

For any constants m and v, with $v > 0$, define the random quantity $x = m + vt$. Then x has a T distribution with location m (the mode of $p(t)$ and the mean if $k > 1$), and scale v. We write $x \sim T_k(m, v^2)$. The variance is $V(x) = v^2 k/(k-2)$, so that v^2 is the variance for large k. The density is, by transformation,

$$p(x) \propto \{k + (x-m)^2/v^2\}^{-(k+1)/2}.$$

For large k, the distribution of x is approximately $N(m, v^2)$. Otherwise, we can view the distribution informally as "$N(m, v^2)$ with a little bit of additional uncertainty."

Quantiles of the Student T distributions are tabulated and available in computer software for any value of k. Suppose that t_p is the $100p\%$ quantile of the distribution, i.e., such that $Pr(t \le t_p) = p$. Then the corresponding quantile of x is simply $m + vt_p$.

Appendix: T test in straight line regression

Under the reference posterior $(\beta|Y) \sim T_{n-2}(\hat{\beta}, s^2 v_\beta^2)$ the values of β having lower density than $\beta = 0$ simply satisfy

$$p(\beta|Y) < p(0|Y)$$

where $p(\beta|Y) = c\{n - 2 + (\beta - \hat{\beta})^2/s^2 v_\beta^2\}^{-(n-1)/2}$ is the posterior density. The value at zero is just

$$p(0|Y) = c\{n - 2 + (0 - \hat{\beta})^2/s^2 v_\beta^2\}^{-(n-1)/2} = c\{n - 2 + T^2\}^{-(n-1)/2}$$

where $T = \hat{\beta}/sv_\beta$ is the standardised T statistic. Hence $p(\beta|Y) < p(0|Y)$ if any only if

$$\{n - 2 + (\beta - \hat{\beta})^2/s^2 v_\beta^2\} > \{n - 2 + T^2\}$$

which reduces to

$$t^2 > T^2 \qquad \text{or} \qquad |t| > |T|$$

where $t = (\beta - \hat{\beta})/sv_\beta$ is a random quantity with the standard $T_{n-2}(0,1)$ distribution.

As a result, the posterior probability of β values having lower density than $\beta = 0$ is just $Pr(|t| > |T|) = 2Pr(t > |T|)$ since the T standard distribution is symmetric about zero. This is the standard $p-$value for the test of $\beta = 0$.

6 Exercises

1. Verify the formulæ for the LSEs $(\hat{\alpha}, \hat{\beta})$ in the straight line regression model.

2. In the straight line regression model, consider the the residual sum of squares $Q(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$. By substituting the identity $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ show that $Q(\hat{\alpha}, \hat{\beta}) = S_{yy} - \hat{\beta}^2/v_\beta^2$. Deduce that $\hat{\beta}^2/v_\beta^2$ is the fitted sum of squares, or deviance explained, by the regression model.

3.

4.

5.