Theory of Linear Models

Steven Gilmour
King's College London

January – February 2020

Sampling Distributions of Estimators

In order to construct confidence intervals and hypothesis tests, it is necessary to know the sampling distributions of the relevant estimators.

In order to make progress we make the distributional assumption $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$.

Theorem

Let $\mathbf{C}$ *be a* $p \times q$ *matrix with columns* $\mathbf{c}_1, \ldots, \mathbf{c}_q$ *such that* $\mathbf{c}_s'\beta$ *are simultaneously estimable for all* $s = 1, \ldots, q$. *Then*

$$\mathbf{C}'\hat{\beta} \sim N_q\left(\mathbf{C}'\beta, \sigma^2\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right)$$

This follows immediately from the fact that $\hat{\beta}$ is linear in $\mathbf{Y}$.

Theorem

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-r}.$$

Theorem

$\mathbf{C}'\hat{\boldsymbol{\beta}}$ *and* SS_R *are independently distributed.*

The following theorem follows immediately from the previous three, using some results on distribution theory.

Theorem

For any estimable function $\mathbf{c}'\boldsymbol{\beta}$,

$$\frac{\mathbf{c}'\hat{\boldsymbol{\beta}} - \mathbf{c}'\boldsymbol{\beta}}{\sqrt{S^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}}} \sim t_{n-r}.$$

Inference for a Single Function

Tests and confidence intervals for a single function of the parameters follow immediately.

To test $H_0 : \mathbf{c}'\boldsymbol{\beta} = \theta_0$, use the test statistic

$$T = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}} - \theta_0}{\sqrt{S^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}}}.$$

Under H_0, $T \sim t_{n-r}$.

A $100(1-\alpha)\%$ confidence interval for $\mathbf{c}'\boldsymbol{\beta}$ is given by

$$\mathbf{c}'\hat{\boldsymbol{\beta}} \pm t_{n-r;1-\alpha/2}\sqrt{S^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}}.$$

Note that this is a generalization of what is given by standard packages:

- It applies whether or not $\mathbf{X}$ is of full rank.
- It applies to any estimable function of the parameters and not just the parameters themselves.
- It applies to any null value θ_0 and not just zero.

Inference for Several Functions

Theorem

Let SS_{R0} be the minimum of S subject to $\mathbf{C}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$. Then

- *SS_R and $SS_{R0} - SS_R$ are independently distributed;*
- *$SS_{R0} - SS_R$ has a noncentral χ^2 distribution with q degrees of freedom;*
- *if $\mathbf{C}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$, then $\frac{SS_{R0} - SS_R}{\sigma^2} \sim \chi^2_q$.*

An immediate corollary is that if $\mathbf{C}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$, then

$$\frac{(SS_{R0} - SS_R)/q}{SS_R/(n-r)} \sim F_{q,n-r}.$$

Tests and confidence regions for several functions of the parameters follow immediately.

To test $H_0 : \mathbf{C}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$, use the test statistic

$$F = \frac{(SS_{R0} - SS_R)/q}{SS_R/(n-r)}.$$

Under H_0, $F \sim F_{q,n-r}$.

These include (but are not restricted to) the usual global and partial F-tests in analysis of variance, which have $\boldsymbol{\theta}_0 = 0$.

These results can be extended further.

If $\mathbf{C}$ is partitioned as $[\mathbf{C}_1\ \mathbf{C}_2]$, with $\mathbf{C}_1$ having dimensions $n \times q_1$, $\boldsymbol{\theta}_0$ is partitioned as $\boldsymbol{\theta}_0' = [\boldsymbol{\theta}_1'\ \boldsymbol{\theta}_2']$, with $\boldsymbol{\theta}_1$ having dimensions $q_1 \times 1$, and SS_{R1} is the minimum of S subject to $\mathbf{C}_1'\boldsymbol{\beta} = \boldsymbol{\theta}_1$, then, using similar arguments to Theorem 10, under $H_0 :\ \mathbf{C}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$,

$$\frac{(SS_{R0} - SS_{R1})/(q - q_1)}{SS_R/(n - r)} \sim F_{q-q_1,n-r}.$$

This provides a basis for using hypothesis tests in model building.

Note that when comparing two reduced models, we should still use the residual sum of squares from the full model (although this might be more or less well defined).

Large Sample Inference

We can use the Central Limit Theorem to justify our inferences for large samples without having to make distributional assumptions initially.

We do, however, have to be a bit careful. The following theorem tells us what we can do.

Theorem

If $E(\mathbf{Y}) = \mathbf{X}\beta$, $V(\mathbf{Y}) = \sigma^2\mathbf{I}$, $\mathbf{X}$ *is of full rank and* $\max_i h_{ii} \to 0$ *as* $n \to \infty$, *where* h_{ii} *are the diagonal elements of* $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, *then*

$$\mathbf{C}'\hat{\beta} \to N_q\left(\mathbf{C}'\beta, \sigma^2\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}\right)$$

as $n \to \infty$.

H is the hat matrix and h_{ii} is the leverage of the ith observation.

The leverage measures the remoteness of a point in the **x**-space.

This is one more reason why study design is important.

How small $\max_i h_{ii}$ has to be before we can invoke CLT depends on how good an approximation we require and how far from normality the distribution is.

A rule of thumb is that if the distribution is not too heavy-tailed or multimodal, then normal theory inference is acceptable if $\max_i h_{ii} < 0.2$.

Even though we are using asymptotic results, we still usually use the t and F distributions, rather than standard normal and χ^2.

Small Sample Inference

In small samples, if it is not considered appropriate to make distributional assumptions, suggested methods of inference include:

- permutation tests, especially in experiments in which the permutations can be defined as the set of possible outcomes of the randomization;
- bootstrapping the residuals, $\mathbf{y} - \hat{\boldsymbol{\mu}}$;
- Bayesian methods with uncertainty about the distributional form expressed using priors.

Ratios of Parameters

It is possible to obtain confidence intervals for ratios of parameters using Fieller's method.

For example, assume that

$$\mu_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2$$

and we wish to estimate the location of the stationary point on the curve

$$x_0 = -\frac{\beta_1}{2\beta_{11}}.$$

Then

$$\hat{\beta}_1 + 2\hat{\beta}_{11} x_0 \sim N(0, \sigma^2 \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}),$$

where $\mathbf{c}' = [0\ 1\ 2x_0]$.

It follows that a $100(1-\alpha)\%$ confidence interval for x_0 is given by the set of all x such that

$$\left| \frac{\hat{\beta}_1 + 2\hat{\beta}_{11}x}{\sqrt{S^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}}} \right| \leq t_{n-r;1-\alpha/2}.$$

Note:

- This idea extends to multiple dimensions.
- This confidence interval will sometimes have infinite area, i.e. it might be $(-\infty, \infty)$.
- This can be avoided by, for example, using (parametric) bootstrap confidence intervals, which always have finite length.
- It can be shown, however, that any confidence interval which always has finite length must have minimum coverage probability of zero.

We are probably more interested in the location of the maximum or minimum (within a particular range of x), rather than the stationary point.

There is probably no satisfactory method for obtaining frequentist confidence intervals in such cases.

Bootstrapping can be used, but will again give intervals with minimum coverage zero.

My recommendation would be to employ Bayesian methods. Samples from the posterior distribution of x_0 are easily obtained from samples from the posterior distribution of $\boldsymbol{\beta}$, e.g. by Gibbs sampling.

Methods for Comparing Models

There are many ways in which we select a particular linear model from a set of candidate models, e.g.

- Fitting polynomial or factorial models of increasing orders.
- Variable selection methods in which we compare all possible models using a criterion such as Mallows' C_p (or the adjusted $\bar{C}_p$), Akaike's information criterion (AIC), Schwarz's Bayesian information criterion (BIC), etc.
- Sequential variable selection methods, such as stepwise selection or stochastic search variable selection.
- Use of a carefully chosen subset, e.g. a fractional replicate of the candidate models.
- Bayesian model choice methods.

These have become even more important with increasingly large data sets.

Variable selection methods can be thought of as an alternative to biased estimation methods, e.g. whereas ridge regression shrinks all the estimates towards zero, variable selection forces some to zero and leaves others alone.

Other methods, such as the lasso, represent a compromise, forcing some estimates to zero and shrinking others.

From this viewpoint, PLS has a the unusual property of stretching some estimates and shrinking others.

The following comments apply to these methods as well as to variable selection, but often in unknown ways.

Inference After Variable Selection

All the frequentist inference we have discussed so far is valid only for the use of a single model.

If the model was selected from among a set of candidates, there is a selection bias. If the number of candidate models is very large, the bias might be enormous.

The idea is best understood through a simple example.

Example: Assume $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. Say the true model is $\mu_i = \beta_0$.

If we fit the single model $\mu_i = \beta_0 + \beta_1 x_{1i}$, then the probability that $\hat{\beta}_1$ is significantly different from zero at the 5% level is 0.05.

If we fit 100 models $\mu_i = \beta_0 + \beta_1 x_{1i}, \quad \cdots, \quad \mu_i = \beta_0 + \beta_1 x_{100i}$, where the explanatory variables are orthogonal, and select the one with the largest value of the F statistic, then the probability that $\hat{\beta}_1$ is significantly different from zero is approximately

$$1 - 0.95^{100} = 0.9941,$$

so we are almost certain to find the regression significant.

The existence of selection bias is well understood and widely recognised, but there is no clear agreement on what to do about it.

Opinion: In a well-designed confirmatory study, we should not use model selection, but should test pre-planned functions of the parameters. We can adjust for multiple testing (or use false discovery rates, etc.), as appropriate.

In an exploratory study we should use model selection, but not follow it up with formal inference.

Residual Analysis

The assumptions of constant variance and normality should, if possible, be checked.

This is most often done using plots of residuals, e.g. against fitted values and normal probability plots.

More formal methods for testing the assumptions have been suggested, but I would not recommend them (because they are not true hypothesis tests). Stick to graphical methods.

The only exception to this is in large and informative data sets, for which we can nest our model in a more general model, e.g. one which does not assume constant variance.

Weighting

We have seen that if $E(\mathbf{Y}) = \mathbf{X}\beta$ and $V(\mathbf{Y}) = \sigma^2\mathbf{G}$, where $\mathbf{G}$ is known, the model can be rewritten as a linear model. This is known as generalized least squares or, if $\mathbf{G}$ is diagonal, with $trace(\mathbf{G}) = n$, as weighted least squares.

Now consider the case where $E(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{X}\beta$ and $V(\mathbf{Y}) = \sigma^2\mathbf{G}$, where $\mathbf{G}$ is diagonal, but unknown.

We typically assume a parametric form relating g_i to μ_i or $\mathbf{x}_i$, where g_i is the ith diagonal element of $\mathbf{G}$.

For example, we might assume that $g_i = x_i^\theta$, where θ is an unknown parameter.

We usually fit such models using iterated weighted least squares, i.e.

1. start with the least squares estimates;
2. use the residuals to estimate θ and hence g_i;
3. use these estimates in a weighted least squares fit to re-estimate $\boldsymbol{\beta}$;
4. iterate steps 2 and 3 until convergence to $\hat{\boldsymbol{\beta}}_G$.

It can be shown that, asymptotically,

$$\hat{\boldsymbol{\beta}}_G \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{G}^{-1}\mathbf{X})^{-1}).$$

Note, however, that the variance depends on $\mathbf{G}$, which is a function of θ and not $\hat{\mathbf{G}}$, which is a function of $\hat{\theta}$.

The quality of the asymptotic approximation depends on how well we estimate θ. In particular, standard errors are usually underestimated if a simple plug-in estimator is used.

Transformations

The famous Box-Cox family of transformations is given by

$$Y^{(\lambda)} = \begin{cases} (Y^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0; \\ \log Y & \text{if } \lambda = 0. \end{cases}$$

The assumed model is

$$Y_i^{(\lambda)} \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2),$$

with Y_is independent.

The main purpose of this transformation is to ensure symmetry (normality) in the distribution of Y.

In a randomization framework, it ensures additivity of unit and treatment effects.

The best fitting λ will be chosen by maximum likelihood and should stabilise the variance.

The MLEs are easily obtained by using an iterative maximisation over λ and finding the least squares estimates of $\boldsymbol{\beta}$, at each λ, i.e. we minimise

$$S(\lambda) = \sum_{i=1}^{n} \left\{ Y_i^{(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta} \right\}^2 .$$

It can be shown that, for a fixed λ,

$$\{S(\lambda) - S(\hat{\lambda})\} \rightarrow \chi_1^2.$$

This means that we can conduct all the usual inference, but with one degree of freedom "lost" for estimating λ.

Both weighting and transformation rely on the assumption that the variance increases (or decreases) with μ or with x. Whereas weighting assumes a symmetrical distribution, transformation assumes that the distribution becomes more asymmetrical the smaller the variance is.

It is also possible to combine transformations and weighting (if the data are sufficiently rich), to simultaneously correct for heteroskedasticity and asymmetry.

Other families of transformations are possible.

It is also possible to transform both sides, i.e.

$$Y_i^{(\lambda)} \sim N\left(\left(\mathbf{x}_i'\boldsymbol{\beta}\right)^{(\lambda)}, \sigma^2\right).$$

This corrects for asymmetry, while maintaining the functional relationship between μ and the explanatory variables.

This makes most sense when the functional form of μ is known (and so is much more common with nonlinear functional forms).