13:27 Sunday 1st November, 2015
See updates and corrections at http://www.stat.cmu.edu/~cshalizi/mreg/

Lecture 18: Tests and Confidence Sets for Multiple Coefficients

36-401, Fall 2015, Section B

27 October 2015

Contents

Throughout, we'll assume that the Gaussian-noise multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon \tag{1}$$

with $\epsilon \sim N(0, \sigma^2)$ independent of the X_is and independent across observations, is completely correct. We will also use the least squares or maximum likelihood estimates of the slopes,

$$\widehat{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} \tag{2}$$

Under these assumptions, the estimator has a multivariate Gaussian distribution,

$$\widehat{\beta} \sim MVN(\beta, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}) \tag{3}$$

The maximum likelihood estimate of σ^2, $\hat{\sigma}^2$, is by

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{x}\widehat{\beta})^T(\mathbf{y} - \mathbf{x}\widehat{\beta}) \tag{4}$$

This is slightly negatively biased, $\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{n-p-1}{n}\sigma^2$, and has the sampling distribution

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1} \tag{5}$$

$\hat{\sigma}^2 \frac{n}{n-p-1}$ is an unbiased estimator of σ^2.

1 $z-$ and $t-$ Tests for Single Coefficients

Let's write the true standard error of the estimator $\hat{\beta}_i$ as $\text{se}\left[\hat{\beta}_i\right]$. From the general theory about the variance of $\widehat{\beta}$,

$$\text{se}\left[\hat{\beta}_i\right] = \sqrt{\sigma^2(\mathbf{x}^T\mathbf{x})^{-1}_{i+1,i+1}} \tag{6}$$

(Why $i+1$?) Further, from the Gaussian distribution of $\widehat{\beta}$,

$$\frac{\hat{\beta}_i - \beta_i}{\text{se}\left[\hat{\beta}_i\right]} \sim N(0,1) \tag{7}$$

If we know σ^2, so that we can compute $\text{se}\left[\hat{\beta}_i\right]$, we can use this to either test hypotheses about the exact value of β_i, or to form confidence intervals. Specifically, a $1-\alpha$ CI would be

$$\hat{\beta}_i \pm z(\alpha/2)\text{se}\left[\hat{\beta}_i\right] \tag{8}$$

with z_p being the p^{th} quantile of the standard Gaussian distribution.

If we use instead the unbiased estimate of σ^2, $\hat{\sigma}^2\frac{n}{n-p-1}$, to obtain an estimate $\widehat{\text{se}}\left[\hat{\beta}_i\right]$, we find rather

$$\frac{\hat{\beta}_i - \beta_i}{\widehat{\text{se}}\left[\hat{\beta}_i\right]} \sim t_{n-p-1} \tag{9}$$

The reasoning for this is exactly parallel to why we got t_{n-2} distributions for simple linear regression, so I won't rehearse it here. It follows that

$$\hat{\beta}_i \pm t_{n-p-1}(\alpha/2)\widehat{\text{se}}\left[\hat{\beta}_i\right] \tag{10}$$

is a $1-\alpha$ confidence interval for β_i. This is implemented in the `confint` function, when applied to the output of `lm`.

As $n \to \infty$, this becomes

$$\hat{\beta}_i \pm z(\alpha/2)\hat{\sigma}\sqrt{(\mathbf{x}^T\mathbf{x})^{-1}_{i+1,i+1}} \tag{11}$$

which is often a quite practical alternative to the t-based interval.

1.1 What, Exactly, Is `summary` Testing?

When you run `summary` on the output of `lm`, part of what it delivers is a table containing estimated coefficients and standard errors, along with a t-statistic and a p-value for each one. It is important to be very clear about the hypothesis being tested here. There is in fact a different null hypothesis for each row of the table. The null hypothesis for β_i is that

$$Y = \beta_0 + \beta_1 X_1 + \ldots \beta_{i-1} X_{i-} + 0 X_i + \beta_{i+1} X_i + \ldots + \beta_p X_p + \epsilon \tag{12}$$

with ϵ being mean-zero, constant-variance independent Gaussian noise. The alternative hypothesis is that

$$Y = \beta_0 + \beta_1 X_1 + \ldots \beta_{i-1} X_{i-} + \beta_i X_i + \beta_{i+1} X_i + \ldots + \beta_p X_p + \epsilon \tag{13}$$

with $\beta_i \neq 0$, and the same assumptions about ϵ. This matters because *whether the null hypothesis is true or not depends on what other variables are included in the model*. The optimal coefficient on X_i might be zero with one set of covariates and non-zero with another. The t test is, by its nature, incapable of saying whether X_i should be included in the model or not.

(This is in addition to the usual cautions about whether testing $\beta_i = 0$ is really informative, about not mistaking "detectably different from zero" for "important", and about how any $\beta_i \neq 0$ will eventually have a p-value arbitrarily close to 0.)

1.2 No, Really, Whether Coefficients Are Zero Changes with the Covariates

Here is the simplest situation I know of which illustrates that the true (optimal or population-level) coefficient of a given predictor variable changes with the other variables included in the model. Suppose that the true model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \tag{14}$$

with all the usual assumptions being met. Without knowing this, we instead estimate the model

$$Y = \gamma_0 + \gamma_1 X_1 + \eta \tag{15}$$

We know, from our study of the simple linear model, that the (optimal or population) value of γ_1 is

$$\gamma_1 = \frac{\mathrm{Cov}\left[X_1, Y\right]}{\mathrm{Var}\left[X_1\right]} \tag{16}$$

Substituting in for Y,

$$\begin{aligned}
\gamma_1 &= \frac{\mathrm{Cov}\left[X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon\right]}{\mathrm{Var}\left[X_1\right]} && (17)\\
&= \frac{\mathrm{Cov}\left[X_1, \beta_0\right] + \mathrm{Cov}\left[X_1, \beta_1 X_1\right] + \mathrm{Cov}\left[X_1, \beta_2 X_2\right] + \mathrm{Cov}\left[X_1, \epsilon\right]}{\mathrm{Var}\left[X_1\right]} && (18)\\
&= \frac{0 + \beta_1 \mathrm{Cov}\left[X_1, X_1\right] + \beta_2 \mathrm{Cov}\left[X_1, X_2\right] + 0}{\mathrm{Var}\left[X_1\right]} && (19)\\
&= \beta_1 + \beta_2 \frac{\mathrm{Cov}\left[X_1, X_2\right]}{\mathrm{Var}\left[X_1\right]} && (20)
\end{aligned}$$

Thus, even if $\beta_1 = 0$, we can easily have $\gamma_1 \neq 0$, and vice versa. (See also Exercise 1.)

2 Variance Ratio (F) Tests for Multiple Coefficients Being Zero

If we want to test whether a group of multiple coefficients are all simultaneously zero, the traditional approach is a variance ratio or F test. To lay everything out, the null hypothesis is that

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_q X_q + 0 X_{q+1} + \ldots + 0 X_p + \epsilon \tag{21}$$

while the alternative is

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_q X_q + \beta_{q+1} X_{q+1} + \ldots + \beta_p X_p + \epsilon \tag{22}$$

with at least one of the coefficients $\beta_{q+1}, \ldots \beta_p \neq 0$. The null hypothesis, then, is that in a linear model which includes all the predictors $X_1, \ldots X_p$, the optimal coefficients for the last $p - q$ variables are all zero.

For both models, we get an estimate of σ^2, say $\hat{\sigma}^2_{null}$ for the null model (with coefficients fixed at zero) and $\hat{\sigma}^2_{full}$ for the full model. Because the null model is a special case of the full model, and we estimate parameters in each case by minimizing the MSE, $\hat{\sigma}^2_{null} \geq \hat{\sigma}^2_{full}$.

Following reasoning exactly parallel to the way we got the F test for the simple linear regression model,

$$\frac{n\hat{\sigma}^2_{full}}{\sigma^2} \sim \chi^2_{n-p-1} \tag{23}$$

while, under the null hypothesis,

$$\frac{n(\hat{\sigma}^2_{null} - \hat{\sigma}^2_{full})}{\sigma^2} \sim \chi^2_{p-q} \tag{24}$$

and so (again under the null hypothesis)

$$\frac{(\hat{\sigma}^2_{null} - \hat{\sigma}^2_{full})/(p-q)}{\hat{\sigma}^2_{full}/(n-p-1)} \sim F_{p-q,n-p-1} \tag{25}$$

We therefore reject the null hypothesis when the test statistic

$$F = \frac{(\hat{\sigma}^2_{null} - \hat{\sigma}^2_{full})/(p-q)}{\hat{\sigma}^2_{full}/(n-p-1)} \tag{26}$$

is too large compared to the $F_{p-q,n-p-1}$ distribution. This is why this is called an F test for this set of regression coefficients. If we're not testing all the coefficients at once, this is a **partial** F test.

The proper interpretation of this test is "Does letting the slopes for $X_{q+1}, \ldots X_p$ be non-zero reduce the MSE more than we would expect just by noise?" As n grows, increasingly small improvements will become clearly detectable as not-noise, so increasingly small but non-zero sets of coefficients will be detected as significant by the F test.

Cautions The variance ratio test does not test any of the following:

- Whether some variable not among $X_1, \ldots X_p$ ought to be included in the model.
- Whether the relationship between Y and the X_i is linear.
- Whether the Gaussian noise assumption holds.
- Whether any of the other modeling assumptions hold.

2.1 All Slopes at Once

An obvious special case is the hypothesis that all the coefficients are zero. That is, the null hypothesis is

$$Y = \beta_0 + 0X_1 + \ldots + 0X_p + \epsilon \tag{27}$$

with the alternative being the full model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon \tag{28}$$

The estimate of σ^2 under the null is the sample variance of Y, s^2_Y, so the test statistic becomes

$$\frac{(s^2_Y - \hat{\sigma}^2_{full})/p}{\hat{\sigma}^2_{full}/(n-p-1)} \tag{29}$$

whose distribution under the null is $F_{p,n-p-1}$.

This **full** F test is often called a test of the significance of the whole regression. This is true, but has to be understood in a very specific sense. We are testing whether, if Y is linearly regressed on $X_1, \ldots X_p$ and only on those variables, the reduction in the MSE from actually estimating slopes over just using a flat regression surface is bigger than we'd expect from pure noise. Once again, the test has no power to detect violations of any of the modeling assumptions. (See the discussion of the F test for simple linear regression in Lecture 10.)

2.2 Variance Ratio Tests in R

This is most easily done through the `anova` function. We fit the null model and the full model, both with `lm`, and then pass them to the `anova` function:

```
mobility <- read.csv("http://www.stat.cmu.edu/~cshalizi/mreg/15/dap/1/mobility.csv")
mob.null <- lm(Mobility ~ Commute, data=mobility)
mob.full <- lm(Mobility ~ Commute + Latitude + Longitude, data=mobility)
anova(mob.null, mob.full)
```

```
## Analysis of Variance Table
##
## Model 1: Mobility ~ Commute
## Model 2: Mobility ~ Commute + Latitude + Longitude
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    727 1.3143
## 2    725 1.2952  2  0.019111 5.3491 0.004942
```

The second row tells us that the full model has two more parameters than the null, that $n(\hat{\sigma}^2_{null} - \hat{\sigma}^2_{full}) = 0.0191114$, and then what the variance ratio or F statistic and the corresponding p-value are. Here, we learn that the decrease in the root-MSE which comes from adding latitude and longitude as predictors, while very small (0.51 percentage points) is large enough that it is unlikely to have arisen by capitalizing on noise[1].

2.3 Variable Deletion via F Tests

It's not uncommon to use F tests for variable deletion: pick your least favorite set of predictors, test whether all of their βs are zero, and, if so, delete them from the model (and re-estimate). Presuming that we can trust the modeling assumptions, there are still a few points about this procedure which are slightly dubious, or at least call for much more caution than is often exercised.

Statistical power The test controls the probability of rejecting when the null is true — it guarantees that if $\beta_q = \mathbf{0}$, we have a low probability of rejecting that null hypothesis. For deletion to be reliable, however, we'd want a low probability of *missing* variables with non-zero coefficients, i.e., a low probability of retaining the null hypothesis when it's wrong, or high power to detect departures from the null. Power cannot be read off from the p-value, and grows with the magnitude of the departure from the null. One way to get at this is, as usual, to complement the hypothesis test with a confidence set for the coefficients in question. Ignoring variables whose coefficients are *precisely* estimated to be close to zero is much more sensible than ignoring variables because their coefficients can only be estimated very loosely.

[1] Once again, this presumes that the only two possibilities in the world are a completely-correct linear-Gaussian model with just commuting time as a predictor, and a completely-correct linear-Gaussian model with commuting time, latitude and longitude as predictors.

Non-transivitiy The variance ratio test checks whether the MSE of the smaller model is significantly or detectably worse than the MSE of the full model. One drawback to this is that a series of insignificant, undetectably-small steps can add up to a significant, detectably-big change. In mathematical jargon: "is equal to" is a transitive relation, so that if $A = B$ and $B = C$, $A = C$. But "insignificantly different from" is not a transitive relation, so if $A \approx B$ and $B \approx C$, we can't conclude $A \approx C$.

Concretely: a group of variables might show up as significant in a partial F test, even though none of them was individually significant on a t test in the full model[2]. Also, if we delete variables in stages, we can have a situation where at each stage the increase in MSE is insignificant, but the difference between the full model and the final model is highly significant.

2.4 Likelihood Ratio Tests

As with the F test for simple linear models, there is an alternative based on the likelihood ratio. As with the simple model, the log-likelihood of the model, at the maximum likelihood estimate, is

$$-\frac{n}{2}(1+\log 2\pi) - \frac{n}{2}\log \hat{\sigma}^2 \tag{30}$$

Hence the difference in log-likelihoods between the full model, with all p slopes estimated, and the null model, with only q slopes estimated and the other $p-q$ fixed, is

$$\Lambda = -\frac{n}{2}\log \hat{\sigma}^2_{full} + \frac{n}{2}\log \hat{\sigma}^2_{null} = \frac{n}{2}\log \frac{\hat{\sigma}^2_{null}}{\hat{\sigma}^2_{full}} \tag{31}$$

This is the log of the ratio of likelihoods (not the ratio of log likelihoods!) Under the null hypothesis[3],

$$2\Lambda \sim \chi^2_{p-q} \tag{32}$$

The same cautions apply to the likelihood ratio test as to the F test: it does not check modeling assumptions.

One advantage of likelihood ratio tests is that exactly the same procedure can be used to test the hypothesis that $\beta_q = \mathbf{0}$ and to test $\beta_q = \beta_q^*$, for any other particular vector of parameters. For that matter, it can be used to test $\mathbf{c}\beta = \mathbf{r}$, where $\mathbf{c}$ is any non-random $q \times (p+1)$ matrix, and $\mathbf{r}$ is any non-random $q \times 1$ vector. Thus, for example, it can be used to test the hypothesis that two slopes are *equal*, or that all slopes are equal, etc.

[2]This is yet another reason not to pay so much attention to the p-values reported by `summary`.

[3]Strictly speaking, this only becomes exact as $n \to \infty$. This issue is that deriving the χ^2 distribution for Λ presumes every parameter's maximum likelihood estimate has a Gaussian distribution around its true value (see Lecture 10), and while this is true for the $\hat{\beta}_i$s, it is only approximately true for $\hat{\sigma}^2$. See Exercise 4.

Likelihood Ratio vs. F Tests For linear-Gaussian models, both the likelihood ratio and the F statistic are functions of the ratio $\hat{\sigma}^2_{null}/\hat{\sigma}^2_{full}$ (Exercise 2). For fixed p and q, as $n \to \infty$, the two tests deliver the same p-values when $\hat{\sigma}^2_{null}/\hat{\sigma}^2_{full}$ is the same. At finite n, they are somewhat different, with the F test usually giving a somewhat higher p value than the χ^2 test, particularly if p is close to n. Which test is more *accurate* is another question. The likelihood ratio test can actually work for large n when the model is mis-specified, in the sense of telling us which wrong model is closer to the truth (Vuong, 1989), while the F test's refinements over the χ^2 very much depend on all the modeling assumptions being right.

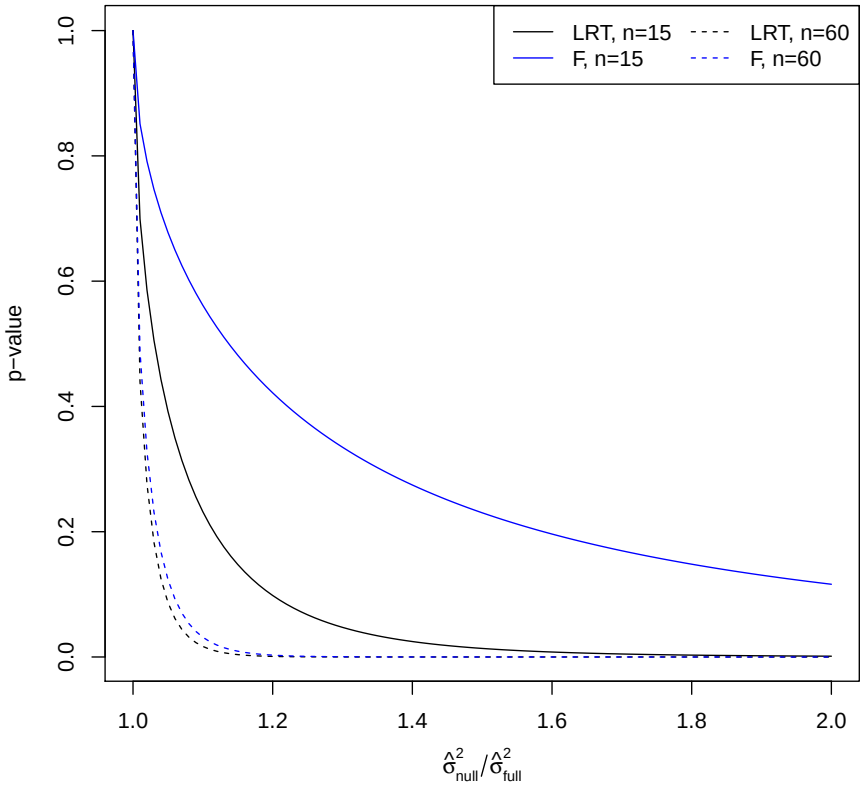


FIGURE 1: *Difference in p-values obtained from using a likelihood ratio test (black) and an F test (blue) on the same data, with $p = 10$, $q = 9$, and n either 15 (solid) or 60 (dotted). In general, the difference between the two tests goes to zero as $n - p$ grows. (See source file for code.)*

3 Confidence Sets for Multiple Coefficients

Suppose we want to do inference on two coefficients, say β_i and β_j, at once. That means we need to come up with a two-dimensional confidence region $C(\alpha)$, where we can say that $\mathbb{P}\left((\beta_i, \beta_j) \in C(\alpha)\right) = 1 - \alpha$. This would involve the same sort of trilemma as confidence intervals for single coefficients. That is, one of three things must be true:

1. Both β_i and β_j are in $C(\alpha)$; or
2. We got data which was very ($\leq \alpha$) improbable under all possible values of the parameters; or
3. Our model is wrong.

If we trust our model, then, we can indeed be confident that both β_i and β_j are simultaneously in $C(\alpha)$.

Clearly, nothing depends on wanting to do inference on just two coefficients at once; we could consider any subset of them we like, up to all $p+1$ of them.

With one parameter, intervals are the most natural confidence sets to work with. With more than one parameter, we have choices to make about the *shape* of the confidence set. The two easiest ones to work with are rectangular boxes, and ellipsoids.

3.1 Confidence Boxes or Rectangles

The natural thing to want to do is to take a confidence interval for each coefficient and put them together into a confidence box or rectangle. For instance, using the t-distribution CI for β_i and β_j, the box would be

$$(\hat{\beta}_i \pm t_{n-p-1}(\alpha/2)\widehat{\mathrm{se}}\left[\hat{\beta}_i\right]) \times (\hat{\beta}_j \pm t_{n-p-1}(\alpha/2)\widehat{\mathrm{se}}\left[\hat{\beta}_j\right]) \tag{33}$$

(And similarly for three or more parameters.) This is, however, not quite right as I've written it. The problem is that while each interval covers its true coefficient with high probability, both intervals *simultaneously* cover the pair of parameters is a different story. Let me abbreviate the interval for β_i as $C_i(\alpha)$, likewise the interval for β_j is $C_j(\alpha)$. We have

$$\mathbb{P}\left(\beta_i \in C_i(\alpha)\right) = 1 - \alpha \ , \ \mathbb{P}\left(\beta_j \in C_j(\alpha)\right) = 1 - \alpha \tag{34}$$

but from this it does not follow that

$$\mathbb{P}\left(\beta_i \in C_i(\alpha), \beta_j \in C_j(\alpha)\right) = 1 - \alpha \tag{35}$$

To see this, let's consider the complementary event: it's

$$\beta_i \notin C_i(\alpha) \vee \beta_j \notin C_j(\alpha) \tag{36}$$

writing $\vee$ for logical-or[4] By basic probability,

$$\mathbb{P}\left(\beta_i \notin C_i(\alpha) \vee \beta_j \notin C_j(\alpha)\right) = \mathbb{P}\left(\beta_i \notin C_i(\alpha)\right) + \mathbb{P}\left(\beta_j \notin C_j(\alpha)\right) - \mathbb{P}\left(\beta_i \notin C_i(\alpha), \beta_j \notin C_j(\alpha)\right) \tag{37}$$

Since C_i and C_j are $1-\alpha$-confidence sets,

$$\mathbb{P}\left(\beta_i \notin C_i(\alpha) \vee \beta_j \notin C_j(\alpha)\right) = 2\alpha - \mathbb{P}\left(\beta_i \notin C_i(\alpha), \beta_j \notin C_j(\alpha)\right) \leq 2\alpha \tag{38}$$

So $C_i(\alpha) \times C_j(\alpha)$ isn't itself a $1-\alpha$ confidence set; its real confidence level could be as little as $1-2\alpha$. If we had been looking at q coefficients at once, the confidence level might have been as low as $1-q\alpha$.

This suggests, however, a very simple, if sometimes over-cautious, way of building a confidence box. If we want the final box to have a $1-\alpha$ confidence level, and we're dealing with q coefficients at once, we calculate $1-\alpha/q$ confidence levels for each coefficient, say $C_i(\alpha/q)$, and then set

$$C(\alpha) = C_1(\alpha/q) \times C_2(\alpha/q) \times \ldots \times C_q(\alpha/q) \tag{39}$$

By our reasoning above, this final $C(\alpha)$ will cover all q parameters at once with probability at least $1-\alpha$.

This trick of building a $1-\alpha$ confidence box for q parameters at once from q $1-\alpha/q$ confidence intervals is completely generic; it doesn't just work on regression coefficients, but for any parameters of any statistical model at all. For more on it, see §4 below.

[4]That is, $A \vee B$ means in ordinary English "A is true or B is true or both are true".

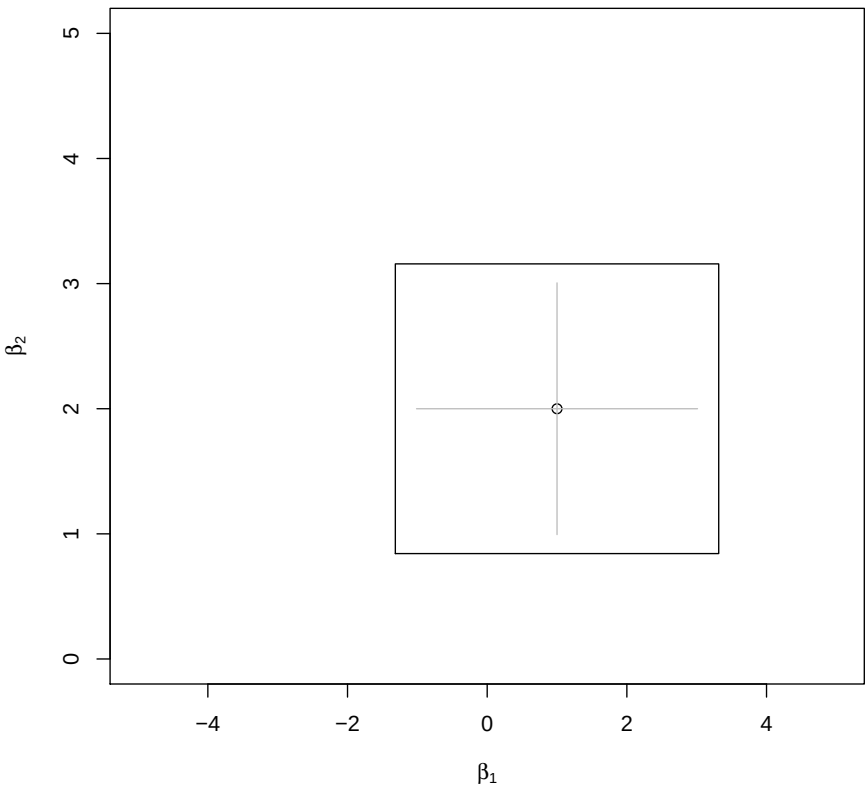


FIGURE 2: *Grey lines: 95% confidence intervals for two coefficients, based on inverting t tests, and so centered at the point estimate (dot). Black box: a 95% confidence rectangle for both coefficients simultaneously. Notice that the grey lines do not touch the sides of the rectangle; the latter correspond to 97.5% CIs for each coefficient. If we did draw the rectangle corresponding to the grey lines, its actual confidence level could be as low as 90%. (See source file for code.)*

3.2 Confidence Balls or Ellipsoids

An alternative to confidence boxes is to try to make confidence *balls*. To see how this could work, suppose first that $\hat{\beta}_i$ and $\hat{\beta}_j$ were uncorrelated. Since

$$\frac{\hat{\beta}_i - \beta_i}{\text{se}\left[\hat{\beta}_i\right]} \sim N(0,1) \tag{40}$$

(and likewise for β_j), we would have[5]

$$\left(\frac{\hat{\beta}_i - \beta_i}{\text{se}\left[\hat{\beta}_i\right]}\right)^2 + \left(\frac{\hat{\beta}_j - \beta_j}{\text{se}\left[\hat{\beta}_j\right]}\right)^2 \sim \chi^2_2 \tag{41}$$

Therefore, a simultaneous $1-\alpha$ confidence region for (β_i, β_j) would be the region where

$$\left(\frac{\hat{\beta}_i - \beta_i}{\text{se}\left[\hat{\beta}_i\right]}\right)^2 + \left(\frac{\hat{\beta}_j - \beta_j}{\text{se}\left[\hat{\beta}_j\right]}\right)^2 \leq \chi^2_2(1-\alpha) \tag{42}$$

A little geometry shows that this region is an ellipse, its axes parallel to the coordinate axis with the length from end to end along one axis being $2\text{se}\left[\hat{\beta}_i\right]\chi^2_2(1-\alpha)$, and its length along the other axis being $2\text{se}\left[\hat{\beta}_j\right]\chi^2_2(1-\alpha)$.

If we had q different uncorrelated coefficients, the confidence region would be the set $(\beta_1, \beta_2, \ldots \beta_q)$ where

$$\sum_{i=1}^{q}\left(\frac{\hat{\beta}_i - \beta_i}{\text{se}\left[\hat{\beta}_i\right]}\right)^2 \leq \chi^2_q(1-\alpha) \tag{43}$$

When $q > 2$, we call this region an "ellipsoid" rather than an "ellipse", but it's the same idea.

Usually, of course, the different coefficient estimates are correlated with each other, so we need to do something a bit different. If we write β_q for the vector of coefficients we're interested in, and $\mathbf{\Sigma}_q$ for its variance-covariance matrix, then the confidence region is the set of all β_q where

$$(\widehat{\beta}_q - \beta_q)^T \mathbf{\Sigma}_q^{-1} (\widehat{\beta}_q - \beta_q) \leq \chi^2_q(1-\alpha) \tag{44}$$

This, too, is an ellipsoid, only now the axes point in the directions given by the eigenvectors of $\mathbf{\Sigma}_q$, and the length along each axis is proportional to the square root of the corresponding eigenvalue. (See §3.2.2 for a derivation.)

Since $\mathbf{\Sigma}_q$ is a $q \times q$ sub-matrix of $\sigma^2(\mathbf{x}^T\mathbf{x})^{-1}$, we can't actually use this. We can, however, use the appropriate sub-matrix of $\hat{\sigma}^2(\mathbf{x}^T\mathbf{x})^{-1}$ as an approximation, which becomes exact as $n \to \infty$. Similarly, if we use the unbiased estimate of σ^2, we replace $\chi^2_q(1-\alpha)$ with $F_{q,n-p-1}(1-\alpha)$.

[5] Because when $Z_1, \ldots Z_d$ are independent $N(0,1)$ variables, $\sum_i Z_i^2 \sim \chi^2_d$.

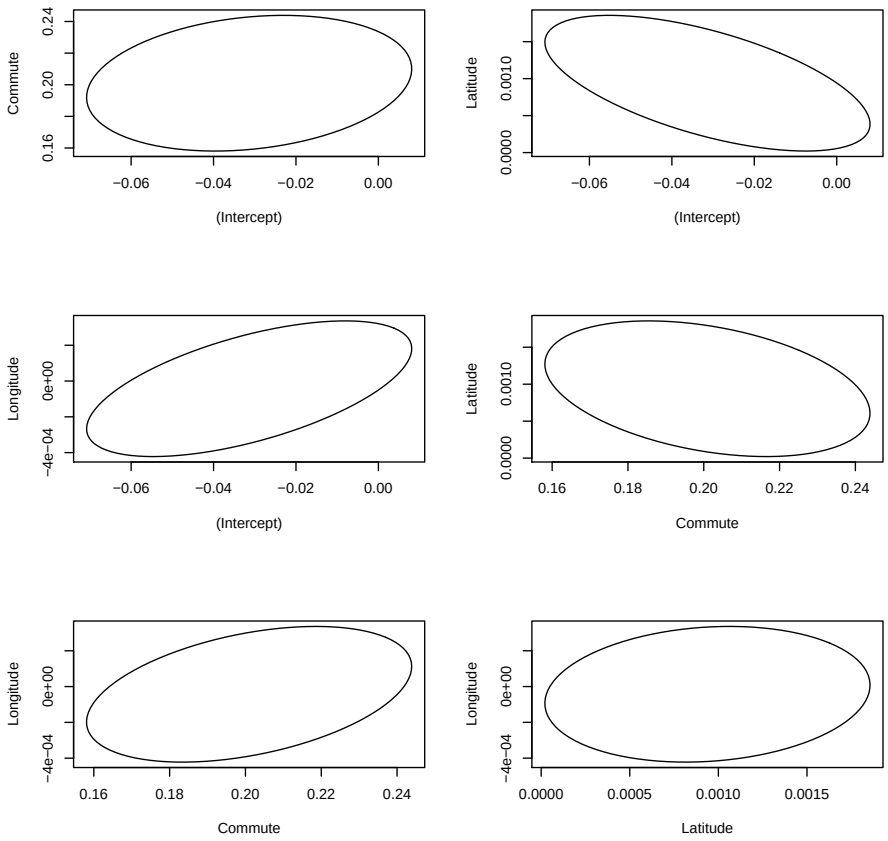
3.2.1 Confidence Ellipsoids in R

The package `ellipse` (Murdoch and Chow, 2013) contains functions for plotting 2D confidence ellipses. The main function is also called `ellipse`, which happens to have a specialized method for `lm` models. The usage is

```
my.model <- lm(y ~ x1+x2+x3)
plot(ellipse(my.model, which=c(1,2), level=0.95))
```

Here `which` is the vector of coefficient indices (it can only be of length 2) and `level` is the confidence level. Notice that what `ellipse` actually returns is a two-column array of coordinates, which can be plotted, or passed along to other graphics functions (like `points` or `lines`). See Figure 3.

Three-dimensional confidence ellipsoids can be made with the `rgl` library (Adler *et al.*, 2014). While confidence ellipsoids exist in any number of dimensions, they can't really be visualized when $q > 3$.




```
library(ellipse)
par(mfrow=c(3,2))
plot(ellipse(mob.full, which=c(1,2), level=1-0.05/6), type="l")
plot(ellipse(mob.full, which=c(1,3), level=1-0.05/6), type="l")
plot(ellipse(mob.full, which=c(1,4), level=1-0.05/6), type="l")
plot(ellipse(mob.full, which=c(2,3), level=1-0.05/6), type="l")
plot(ellipse(mob.full, which=c(2,4), level=1-0.05/6), type="l")
plot(ellipse(mob.full, which=c(3,4), level=1-0.05/6), type="l")
```

FIGURE 3: *Confidence ellipses for every pair of coefficients in the model where economic mobility is regressed on the prevalence of short commutes, latitude and longitude. (Remember the intercept is the first coefficient.) Why do I use this odd-looking confidence level?*

3.2.2 Where the χ^2_q Comes From

To see why this should be so, we need to do some linear algebra, to turn a Gaussian random vector with correlations and unequal variances into a Gaussian random vector where the coordinates are all $\sim N(0,1)$ and independent of each other. The starting point is the fact that $\mathbf{\Sigma}_q$ is a square, symmetric, positive-definite matrix. Therefore it can be written as follows:

$$\mathbf{\Sigma}_q = \mathbf{V}\mathbf{U}\mathbf{V}^T \tag{45}$$

where $\mathbf{U}$ is the diagonal matrix of eigenvalues, and $\mathbf{V}$ is the matrix whose columns are the eigenvectors; $\mathbf{V}^T$ is its transpose, and $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. If we define $\mathbf{\Sigma}_q^{1/2} = \mathbf{V}\mathbf{U}^{1/2}$, where $\mathbf{U}^{1/2}$ is the diagonal matrix with the square roots of the eigenvalues, then

$$\begin{aligned}
\mathrm{Var}\left[\mathbf{\Sigma}_q^{-1/2}(\widehat{\beta}_q - \beta_q)\right] &= \mathbf{\Sigma}_q^{-1/2}\mathrm{Var}\left[\widehat{\beta}_q - \beta_q)\right](\mathbf{\Sigma}_q^{-1/2})^T && (46)\\
&= \mathbf{U}^{-1/2}\mathbf{V}^{-1}\mathbf{V}\mathbf{U}\mathbf{V}^T\mathbf{V}\mathbf{U}^{-1/2} && (47)\\
&= \mathbf{U}^{-1/2}\mathbf{U}\mathbf{U}^{-1/2} && (48)\\
&= \mathbf{I} && (49)
\end{aligned}$$

where the last step works because $\mathbf{U}$ and $\mathbf{U}^{-1/2}$ are both diagonal matrices. In other words, while the coordinates of $\widehat{\beta}_q - \beta_q$ have unequal variances and are correlated with each other, $\mathbf{\Sigma}_q^{-1/2}(\widehat{\beta}_q - \beta_q)$ is a random vector where each coordinate has variance 1 and is uncorrelated with the others. Since the initial vector was Gaussian, this too is Gaussian, hence

$$\mathbf{\Sigma}_q^{-1/2}(\widehat{\beta}_q - \beta_q) \sim MVN(\mathbf{0}, \mathbf{I}) \tag{50}$$

Therefore

$$\left(\mathbf{\Sigma}_q^{-1/2}(\widehat{\beta}_q - \beta_q)\right)^T \mathbf{\Sigma}_q^{-1/2}(\widehat{\beta}_q - \beta_q) \sim \chi^2_q \tag{51}$$

since it's a sum of q squared, independent $N(0,1)$ variables.

On the other hand,

$$\begin{aligned}
&\left(\mathbf{\Sigma}_q^{-1/2}(\widehat{\beta}_q - \beta_q)\right)^T \left(\mathbf{\Sigma}_q^{-1/2}(\widehat{\beta}_q - \beta_q)\right) && (52)\\
&= (\widehat{\beta}_q - \beta_q)^T \left(\mathbf{\Sigma}_q^{-1/2}\right)^T \mathbf{\Sigma}_q^{-1/2}(\widehat{\beta}_q - \beta_q) &&\\
&= (\widehat{\beta}_q - \beta_q)^T\mathbf{V}\mathbf{U}^{-1/2}\mathbf{U}^{-1/2}\mathbf{V}^{-1}(\widehat{\beta}_q - \beta_q) && (53)\\
&= (\widehat{\beta}_q - \beta_q)^T\mathbf{V}\mathbf{U}^{-1}\mathbf{V}^{-1}(\widehat{\beta}_q - \beta_q) && (54)\\
&= (\widehat{\beta}_q - \beta_q)^T\mathbf{\Sigma}_q^{-1}(\widehat{\beta}_q - \beta_q) && (55)
\end{aligned}$$

Combining Eqs. 50 and 54,

$$(\widehat{\beta}_q - \beta_q)^T\mathbf{\Sigma}_q^{-1}(\widehat{\beta}_q - \beta_q) \sim \chi^2_q \tag{56}$$

as was to be shown.

4 Further Reading

Variance and likelihood ratio tests go back to the period of the 1910s–1930s; see references in Lecture 10. Further exposition can be found in any textbook on regression, or general mathematical statistics.

The trick in §3.1, of getting an over-all confidence level of $1-\alpha$ for q parameters simultaneously, by demanding the higher confidence level of $1-\alpha/q$ for each one separately, is one use of an important tool called **Bonferroni correction** or **Bonferroni adjustment**[6]. For an account of the role of this general idea in probability theory, see Galambos and Simonelli (1996). Bonferroni correction is also often used for hypothesis testing: if we test q distinct hypotheses, and we want to have the probability of making *no* false rejections be α, we can achieve that by having each test be of size α/q. Indeed, we could give each test whatever size we like, so long as the sum of the tests is α.

One sometimes encounters the mis-understanding that Bonferroni correction requires the test statistics or confidence intervals to be statistically independent (e.g., Ashby 2011); as you can see from the argument above, this is just wrong. What is true is that Bonferroni correction is very cautious, and that one can sometimes come up with less conservative ways of doing multiple inference if one either uses more detailed information about how the statistics relate to each other (as in §3.2), or one is willing to tolerate a certain number of false positives. The latter idea leads to important work on multiple testing and "false discovery control", which is outside the scope of this course, but see Benjamini and Hochberg (1995); Genovese and Wasserman (2004), and, for an unforgettable demonstration of how ignoring multiple testing issues leads to nonsense, Bennett *et al.* (2010).

5 Exercises

To think through or to practice on, not to hand in.

1. In the scenario of §1.2, is it possible for both ϵ and η to obey the Gaussian noise assumption? That is, it is possible to have $\epsilon \sim N(0, \sigma^2_\epsilon)$, independent of X_1 and X_2, and to have $\eta \sim N(0, \sigma^2_\eta)$, independent of X_1? *Hint:* Suppose X_1 and X_2 are jointly Gaussian, and, for simplicity, that both have mean 0.

2. (a) Show that the variance ratio test statistic (Eq. **??**) depends on the data only through the ratio $\hat{\sigma}^2_{null}/\hat{\sigma}^2_{full}$.

   (b) Show that as $\hat{\sigma}^2_{null} \to \hat{\sigma}^2_{full}$,

$$\log \frac{\hat{\sigma}^2_{null}}{\hat{\sigma}^2_{full}} \to \frac{\hat{\sigma}^2_{null} - \hat{\sigma}^2_{full}}{\hat{\sigma}^2_{full}} \tag{57}$$

[6] Computer scientists, and some mathematicians, call it a "union bound" — can you explain why?

3. Lecture 8 argued that every confidence set comes from inverting a hypothesis test. What is the hypothesis test corresponding to the confidence boxes of §3.1? That is, find an explicit form of the test statistic and of the rejection region.

4. Let $X_n \sim \chi^2_{n-p}$, with fixed p.

   (a) Show that X_n/n approaches a constant a, and find a.

   (b) Show that $(X_n - a)/\sqrt{n}$ approaches a Gaussian distribution, and find the expectation and variance. *Hint:* show that the moment generating functions converge.

   (c) Combine the previous results to write the limiting distribution of X_n/n as a Gaussian, whose parameters (may) change with n.

References

Adler, Daniel, Duncan Murdoch and others (2014). *rgl: 3D visualization device system (OpenGL)*. URL http://CRAN.R-project.org/package=rgl. R package version 0.95.1201.

Ashby, F. Gregory (2011). *Statistical Analysis of fMRI Data*. Cambridge, Massachusetts: MIT Press.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society B*, **57**: 289–300. URL http://www.math.tau.ac.il/~ybenja/MyPapers/benjamini_hochberg1995.pdf.

Bennett, Craig M., Abigail A. Baird, Michael B. Miller and George L. Wolford (2010). "Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction." *Journal of Serendipitous and Unexpected Results*, **1**: 1–5. URL http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf.

Galambos, Janos and Italo Simonelli (1996). *Bonferroni-type Inequalities with Applications*. Berlin: Springer-Verlag.

Genovese, Christopher and Larry Wasserman (2004). "A Stochastic Process Approach to False Discovery Control." *Annals of Statistics*, **32**: 1035–1061. URL http://projecteuclid.org/euclid.aos/1085408494. doi:10.1214/009053604000000283.

Murdoch, Duncan and E. D. Chow (2013). *ellipse: Functions for drawing ellipses and ellipse-like confidence regions*. URL http://CRAN.R-project.org/package=ellipse. R package version 0.3-8.

Vuong, Quang H. (1989). "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica*, **57**: 307–333. URL http://www.jstor.org/pss/1912557.