Lecture 18: Tests and Confidence Sets for Multiple Coefficients

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, $\widehat{\sigma}^2$, is by

$$\widehat{\sigma}^2 = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\widehat{\beta})^T(\mathbf{Y} - \mathbf{X}\widehat{\beta}). \tag{4}$$

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

$$\frac{n\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-q}. \tag{5}$$

$\widehat{\sigma}^2\frac{n}{n-q}$ is an unbiased estimator of σ^2.

1 Tests for Single Coefficients

Recall that

$$\widehat{\mathrm{se}}_j = \sqrt{\widehat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{j+1,j+1}} \tag{6}$$

and that, if we use the unbiased estimate of σ^2 then

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{\mathrm{se}}\left[\widehat{\beta}_j\right]} \sim t_{n-q} \approx N(0,1). \tag{7}$$

The $1-\alpha$ confidence interval is

$$\widehat{\beta}_j \pm t_{n-q}(\alpha/2)\widehat{\mathrm{se}}\left[\widehat{\beta}_q\right] \approx \widehat{\beta}_j \pm z(\alpha/2)\widehat{\mathrm{se}}\left[\widehat{\beta}_q\right] \tag{8}$$

which can be obtained from the `confint` function, when applied to the output of `lm`.

Here is an example that shows that testing if a coefficient is 0 is not the same as testing if that covariate is important. Suppose that the true model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \tag{9}$$

with all the usual assumptions being met. Suppose we did not know about X_2. So we fit the model

$$Y = \gamma_0 + \gamma_1 X_1 + \eta. \tag{10}$$

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

$$\gamma_1 = \frac{\mathrm{Cov}\left[X_1, Y\right]}{\mathrm{Var}\left[X_1\right]}. \tag{11}$$

Substituting in for Y,

$$\begin{aligned}
\gamma_1 &= \frac{\mathrm{Cov}\left[X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon\right]}{\mathrm{Var}\left[X_1\right]} && (12)\\
&= \frac{\mathrm{Cov}\left[X_1, \beta_0\right] + \mathrm{Cov}\left[X_1, \beta_1 X_1\right] + \mathrm{Cov}\left[X_1, \beta_2 X_2\right] + \mathrm{Cov}\left[X_1, \epsilon\right]}{\mathrm{Var}\left[X_1\right]} && (13)\\
&= \frac{0 + \beta_1 \mathrm{Cov}\left[X_1, X_1\right] + \beta_2 \mathrm{Cov}\left[X_1, X_2\right] + 0}{\mathrm{Var}\left[X_1\right]} && (14)\\
&= \beta_1 + \beta_2 \frac{\mathrm{Cov}\left[X_1, X_2\right]}{\mathrm{Var}\left[X_1\right]}. && (15)
\end{aligned}$$

Thus, even if $\beta_1 = 0$, we can easily have $\gamma_1 \neq 0$, and vice versa. The value of a coefficient depends on what other variables are included (or not included) ion the model.

2 F Tests for Multiple Coefficients Being Zero

Let $S \subset \{1, \ldots, p\}$. Suppose we want to test of all the $(\beta_j : j \in S)$ are zero. Let s be the number of variables in S. For example, if $S = \{1, 6, 17\}$ then we are testing the null hypothesis that $\beta_1 = \beta_6 = \beta_{17} = 0$.

To test this hypothesis, fit the full model (with all the variables) and the null model (with the variables in S omitted). Let $\widehat{\sigma}^2_{\text{full}}$ be thge estinmate of σ^2 from the full model and let $\widehat{\sigma}^2_{\text{null}}$ be thge estinmate of σ^2 from the null model. Note that $\widehat{\sigma}^2_{\text{null}} \geq \widehat{\sigma}^2_{\text{full}}$. We can test the null by comparing these variances.

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

$$\frac{n\widehat{\sigma}^2_{full}}{\sigma^2} \sim \chi^2_{n-q} \tag{16}$$

while, under the null hypothesis,

$$\frac{n(\widehat{\sigma}^2_{null} - \widehat{\sigma}^2_{full})}{\sigma^2} \sim \chi^2_s. \tag{17}$$

Note that s is the difference of the dimensions of the two models. Under the null hypothesis we have that

$$F = \frac{(\widehat{\sigma}^2_{null} - \widehat{\sigma}^2_{full})/s}{\widehat{\sigma}^2_{full}/(n-q)} \sim F_{s,n-q}. \tag{18}$$

We therefore reject the null hypothesis when

$$F > F_{s,n-q}(\alpha). \tag{19}$$

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 $(X_j : \ j \in S)$ be non-zero reduce the MSE more than we would expect just by noise?"

Cautions. The F-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

$$Y = \beta_0 + 0X_1 + \ldots + 0X_p + \epsilon \tag{20}$$

with the alternative being the full model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon \tag{21}$$

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

$$\frac{(s_Y^2 - \widehat{\sigma}^2_{full})/p}{\widehat{\sigma}^2_{full}/(n-q)} \tag{22}$$

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

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.

2.2 F-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:

```
out.full = lm(Mobility ~ Commute + Latitude + longitude,data=mobility)
out.null = lm(Mobility ~ Commute,data=mobility)
anova(out.null,out.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 $n(\widehat{\sigma}^2_{null} - \widehat{\sigma}^2_{full}) = 0.0191114$, 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, assuming all the model assumptions are correct.

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.

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. 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.

3 Confidence Sets

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, where we can say that $\mathbf{P}\Big((\beta_i, \beta_j) \in C\Big) = 1 - \alpha$.

3.1 Confidence Boxes or Rectangles

Suppose that C_i is a $1 - \alpha$ confidence interval for β_i and that C_j is a $1 - \alpha$ confidence interval for β_j. We might guess that the rectangle $R = C_i \times C_j$ is a $1 - \alpha$ confidence set for (β_i, β_j) but this is

not correct. To see this, note that

$$\begin{aligned} P((\beta_i, \beta_j) \notin R) &= P(\beta_i \notin R \text{ or } \beta_j \notin R) \\ &= P(\beta_i \notin R) + P(\beta_j \notin R) - P((\beta_i \notin R) \text{ and } \beta_j \notin R) \\ &= 2\alpha - P((\beta_i \notin R) \text{ and } \beta_j \notin R) \leq 2\alpha. \end{aligned}$$

We can get the correct coverage by using the *Bonferroni* correction. Suppose we want a confidence rectangle for $(\beta_j : \ j \in S)$. Let s be the number of elements in S. Let C_j be a confidence interval for β_j at level $1 - (\alpha/s)$. Now define the rectangle

$$C = \bigotimes_{j \in S} C_j.$$

Recall the fact that, for any events $A_1, \ldots, A_m$ we laywa have $P(A_1 \bigcup \cdots \bigcup A_m) \leq \sum_j P(A_j)$. Let $A_j =$ "β_j is not contained in C_j." Then

$$\begin{aligned} P(\beta_j \notin C \text{ for some } j \in S) = P(A_1 \bigcup \cdots \bigcup A_s) &\leq \sum_j P(\beta_j \notin C_j) \\ &= \sum_j \frac{\alpha}{s} = \alpha. \end{aligned}$$

This trick to building a $1-\alpha$ confidence box for s parameters at once is to use $1-\alpha/s$ confidence intervals for each parameter.

3.2 Confidence Ellipsoids

An alternative to confidence boxes is to try to make confidence *ellipsoid*. Let β_S be the subset of coefficients we are interested in. Let $\Sigma_S = \text{Var}(\widehat{\beta}_S)$. Then

$$(\widehat{\beta}_S - \beta_S)^T \Sigma_S^{-1} (\widehat{\beta}_S - \beta_S) \sim \chi^2_s$$

where s is the length of β_S. The reason why this has a χ^2_s distribution is explained in Section 3.2.2. Let c_α be such that $P(\chi^2_s > c_\alpha) = \alpha$. Then

$$C = \left\{ \beta_S : \ (\widehat{\beta}_S - \beta_S)^T \Sigma_S^{-1} (\widehat{\beta}_S - \beta_S) \leq c_\alpha \right\}$$

is a $1 - \alpha$ confidence ellipsoid for β_S.

Of course, we need to estimate Σ_S. Let $\widehat{\Sigma}_S$ be the corresponding sub-matrix of $\widehat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}$ using the unbiased estimate of σ^2. Then we use

$$C = \left\{ \beta_S : \ (\widehat{\beta}_S - \beta_S)^T \widehat{\Sigma}_S^{-1} (\widehat{\beta}_S - \beta_S) \leq c_\alpha \right\}$$

where, now, c_α satisfies $P(F_{s,n-q} > c_\alpha) = \alpha$.

3.2.1 Confidence Ellipsoids in R

The package `ellipse` 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

```
out = lm(y ~ x1 + x2 + x3)
plot(ellipse(out,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 1. The commands for the figure are:

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

Three-dimensional confidence ellipsoids can be plotted with the `rgl` library. While confidence ellipsoids exist in any number of dimensions, they can't really be visualized when $q > 3$.

3.2.2 Where the χ^2_s Comes From

Here we will explain why $(\widehat{\beta}_S - \beta_S)^T \Sigma_S^{-1} (\widehat{\beta}_S - \beta_S) \sim \chi^2_s$. We know that $\mathbf{\Sigma}_S$ is a square, symmetric, positive-definite matrix. Therefore it can be written as

$$\mathbf{\Sigma}_S = \mathbf{V}\mathbf{U}\mathbf{V}^T \tag{23}$$

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 $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. Note that $\mathbf{V}^T$ is the inverse of $\mathbf{V}$. The square root matrix is $\mathbf{\Sigma}_S^{1/2} = \mathbf{V}\mathbf{U}^{1/2}\mathbf{V}^T$, where $\mathbf{U}^{1/2}$ is the diagonal matrix with the square roots of the eigenvalues. Note that $\mathbf{\Sigma}_S^{1/2}\mathbf{\Sigma}_S^{1/2} = \mathbf{\Sigma}_S$ and $\mathbf{\Sigma}_S^{-1/2} = \mathbf{V}\mathbf{U}^{-1/2}\mathbf{V}^T$. So,

$$\begin{aligned}
\mathrm{Var}\left[\mathbf{\Sigma}_S^{-1/2}(\widehat{\beta}_S - \beta_S)\right] &= \mathbf{\Sigma}_S^{-1/2}\mathrm{Var}(\widehat{\beta}_S - \beta_S)(\mathbf{\Sigma}_S^{-1/2})^T && (24)\\
&= \mathbf{V}\mathbf{U}^{-1/2}\mathbf{V}^T\mathbf{V}\mathbf{U}\mathbf{V}^T && (25)\\
&= \mathbf{V}\mathbf{U}^{-1/2}\mathbf{U}\mathbf{U}^{-1/2}\mathbf{V}^T && (26)\\
&= \mathbf{V}\mathbf{V}^T = \mathbf{I}. && (27)
\end{aligned}$$

So, $\widehat{\beta}_S - \beta_S$ have unequal variances and are correlated with each other, $\mathbf{\Sigma}_S^{-1/2}(\widehat{\beta}_S - \beta_S)$ 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

$$\mathbf{\Sigma}_S^{-1/2}(\widehat{\beta}_S - \beta_S) \sim MVN(\mathbf{0}, \mathbf{I}). \tag{28}$$

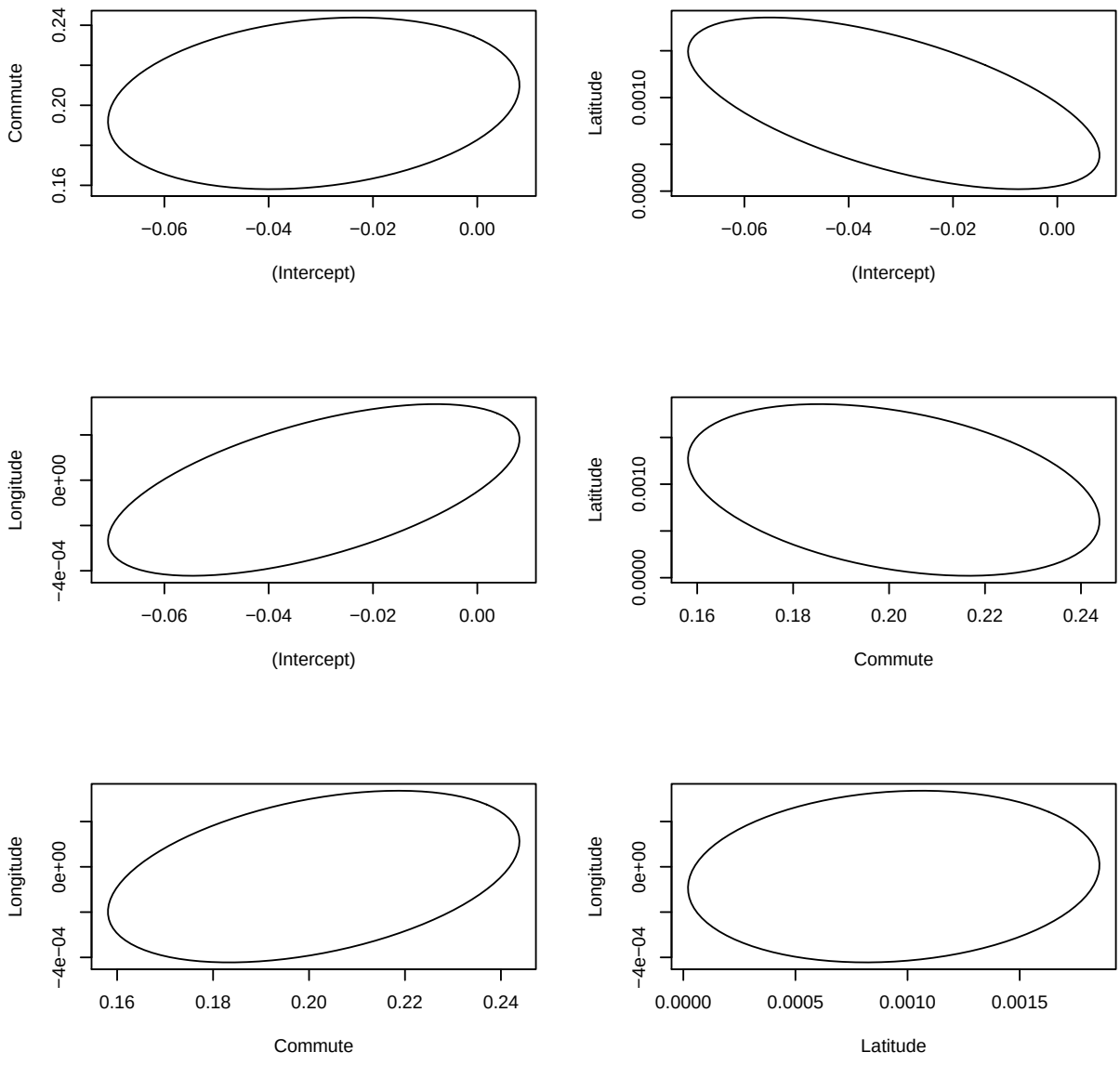

FIGURE 1: *Confidence ellipses for every pair of coefficients in the model where economic mobility is regressed on the prevalence of short commutes, latitude and longitude. (Remember the intercept is the first coefficient.)*

Therefore

$$\left(\mathbf{\Sigma}_S^{-1/2}(\widehat{\beta}_S - \beta_S)\right)^T \mathbf{\Sigma}_S^{-1/2}(\widehat{\beta}_S - \beta_S) \sim \chi_s^2 \tag{29}$$

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

On the other hand,

$$\left(\mathbf{\Sigma}_S^{-1/2}(\widehat{\beta}_S - \beta_S)\right)^T \left(\mathbf{\Sigma}_S^{-1/2}(\widehat{\beta}_S - \beta_S)\right) = (\widehat{\beta}_S - \beta_S)^T \mathbf{\Sigma}_S^{-1}(\widehat{\beta}_S - \beta_S).$$

Hence,

$$(\widehat{\beta}_S - \beta_S)^T \mathbf{\Sigma}_S^{-1}(\widehat{\beta}_S - \beta_S) \sim \chi_s^2. \tag{30}$$