Inference in regression: confidence intervals

ST552 Lecture 9

Charlotte Wickham

2019-01-28

- Another view of F-tests
- Confidence intervals for single parameters
- Confidence intervals for linear combinations of parameters
- Confidence intervals for parameters jointly

Last time. . .

Certain hypotheses of interest can be set up as competing models.
A full model and a simpler model (nested in the full model).
A.K.A testing models.

Identify the models of interest. Fit both. **Check fit of full model.**
Find F-statistic, and answer questions of interest.

Another way to set up F-tests A.K.A testing linear parametric functions

Assuming the regression model:

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)$$

Consider the hypotheses:

$$H_0 : K^T\beta = m$$
$$H_1 : K^T\beta \neq m$$

where $K_{k \times p}$ matrix with rank(K) = k.

Then under the null hypothesis,

$$F = \frac{\left((K^T\beta - m)^T \left(K^T(X^TX)^{-1}K \right)^{-1} (K^T\beta - m) \right) / k}{\text{RSS}/(n-p)} \sim F_{k, n-p}$$

(Don't memorise for ST552, maybe for comps)

4

You get the same answer

This alternative is equivalent to the model testing setup we considered. Every null hypothesis of the form $K^T \beta = m$ is comparing a full and reduced model and vice versa.

For example, consider

$$K = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad m = 0$$

where the 1 in K occurs in the ith row.

What is the null hypothesis being tested?

What are K and m for exercises 1 and 5 from the handout from last time? HW#4

The t-test for an individual parameter can be flipped around to give $100(1 - \alpha)\%$ confidence intervals of the form

$$\hat{\beta}_j \pm t_{n-p}^{(\alpha/2)} \text{SE}\left(\hat{\beta}_j\right)$$

(Remember SE $\left(\hat{\beta}_j\right)$ is coming from the diagonal entry of the estimated variance-covariance matrix.)
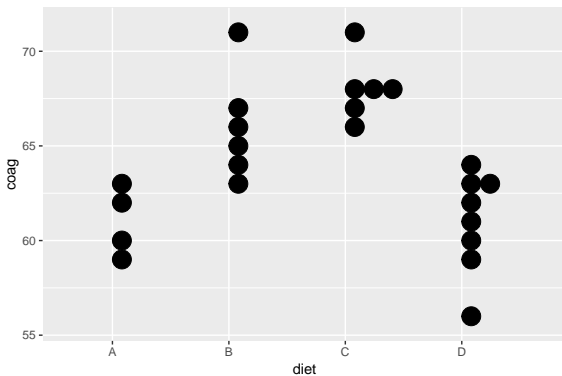
Coagulation times

Dataset comes from a study of blood coagulation times. 24 animals were randomly assigned to four different diets and the samples were taken in a random order.

Consider the model:

$$\text{Coagulation time (s)}_i = \beta_0 + \beta_1 1\{\text{Diet B}\}_i$$
$$+ \beta_2 1\{\text{Diet C}\}_i + \beta_3 1\{\text{Diet D}\}_i + \epsilon_i$$

Coagulation

```
data(coagulation, package = "faraway")
ggplot(coagulation, aes(diet, coag)) +
  geom_dotplot(binaxis = "y",  binwidth = 1)
```

Your turn: cont.

```
fit <- lm(coag ~ diet, data = coagulation)
broom::tidy(fit) %>%
  knitr::kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	61	1.18	51.55	0
dietB	5	1.53	3.27	0
dietC	7	1.53	4.58	0
dietD	0	1.45	0.00	1

Find a 95% CI for β_0?

$t_{n-p}^{(0.975)} = t_{20}^{(0.975)} = 2.09$

Your turn: cont.

In R:

```
broom::tidy(fit, conf.int = TRUE)
# OR
(cis <- confint(fit))
```

Confidence intervals for linear combinations of parameters of β_j

Similarly, confidence intervals for a **linear combination** of the parameters, $c^T\beta$ where $c_{p\times 1}$, can be formed with

$$c^T\hat{\beta} \pm t_{n-p}^{(\alpha/2)}\sqrt{\hat{\sigma}^2 c^T(X^TX)^{-1}c}$$

With the `coagulation` example

$$\text{Coagulation time (s)}_i = \beta_0 + \beta_1 1\{\text{Diet B}\}_i$$
$$+ \beta_2 1\{\text{Diet C}\}_i + \beta_3 1\{\text{Diet D}\}_i + \epsilon_i$$

What is c for the linear combination $\beta_0 - \beta_1$?

Find $c^T (X^T X)^{-1} c$.

```
X <- model.matrix(fit)
round(solve(t(X) %*% X), 2)
```

```
##              (Intercept) dietB dietC dietD
## (Intercept)        0.25 -0.25 -0.25 -0.25
## dietB             -0.25  0.42  0.25  0.25
## dietC             -0.25  0.25  0.42  0.25
## dietD             -0.25  0.25  0.25  0.37
```

Joint confidence regions

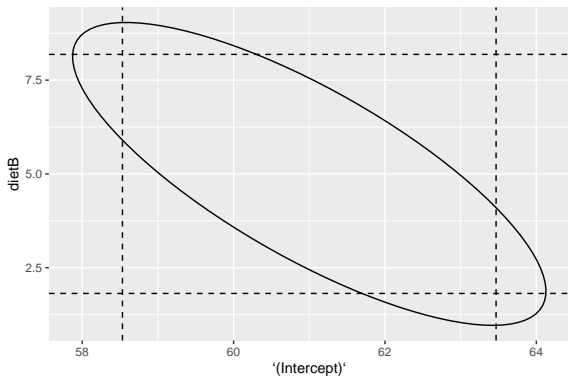A joint $100(1 - \alpha)\%$ confidence for the vector β can be formed using,

$$(\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta) \leq p\hat{\sigma}^2 F_{p,n-p}^{(\alpha)}$$

and results in p-dimensional ellipsoids (very hard to visualise, but essential for communicating joint uncertainty when the parameters are correlated).

2D ellipsoid example: correlated estimates

For example, (β_0, β_1) in

$$\text{Coagulation time (s)}_i = \beta_0 + \beta_1 1\{\text{Diet B}\}_i$$
$$+ \beta_2 1\{\text{Diet C}\}_i + \beta_3 1\{\text{Diet D}\}_i + \epsilon_i$$

2D ellipsoid example: uncorrelated estimates

Compare to γ_0 and γ_1 in this parameterization:

$$\text{Coagulation time (s)}_i = \gamma_0 1\{\text{Diet A}\}_i + \gamma_1 1\{\text{Diet B}\}_i$$
$$+ \gamma_2 1\{\text{Diet C}\}_i + \gamma_3 1\{\text{Diet D}\}_i + \epsilon_i$$

```
fit_nointercept <- lm(coag ~ diet - 1, data = coagulation)
```