STAT 420: Methods of Applied Statistics

Inference of Linear Regressions, II

Ruoqing Zhu, Ph.D. <rqzhu@illinois.edu>

Course website: https://sites.google.com/site/teazrq/teaching/STAT420

University of Illinois at Urbana-Champaign
February 23, 2017

Linear Regressions

- We are going to perform tests on MLR.
- Many of the results follow the previous derivation of the distribution of $\widehat{\boldsymbol{\beta}}$
- In addition to testing individual variables, we introduce
  - The F-test for joint test of multiple parameters.
  - Connection between F-test and R^2 for testing the entire model.
  - Testing a linear constraint of parameters.

Testing individual predictors

- We already derived the distribution of $\widehat{\boldsymbol{\beta}}$, in a general form:

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})$$

- Hence, to test an individual parameter, we only need the corresponding entry in the variance-covariance matrix.
- For example, the gala dataset in the faraway package: we want to model the number of species on an island. 5 predictors are used.

Example: gala dataset

```
> fit = lm(Species~ Area + Elevation + Nearest + Scruz +
    Adjacent, data = gala)
> summary(fit)
Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
>
> sigma2 = deviance(fit) / df.residual(fit)
> X = as.matrix(cbind(1, gala[, 3:7]))
> round(sqrt(diag(solve(t(X) %*% X)*sigma2)), 6)
        1      Area Elevation   Nearest     Scruz  Adjacent
19.154198  0.022422  0.053663  1.054136  0.215402  0.017700
```

Example: gala dataset

The $\widehat{\sigma}^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ matrix:

```
> round(solve(t(X) %*% X)*sigma2, 4)
                   1    Area Elevation Nearest   Scruz Adjacent
1           366.8833  0.1405   -0.5807 -0.8696 -1.3981   0.0859
Area          0.1405  0.0005   -0.0010  0.0048 -0.0002   0.0002
Elevation    -0.5807 -0.0010    0.0029 -0.0132  0.0011  -0.0006
Nearest      -0.8696  0.0048   -0.0132  1.1112 -0.1421   0.0053
Scruz        -1.3981 -0.0002    0.0011 -0.1421  0.0464  -0.0007
Adjacent      0.0859  0.0002   -0.0006  0.0053 -0.0007   0.0003
```

Testing individual predictors

- To calculate the standard errors for each parameter, we invert the matrix $\mathbf{X}^\intercal\mathbf{X}$, and multiple the $\widehat{\sigma}^2$
- Again, when replacing σ^2 with $\widehat{\sigma}^2$, the distribution changed from Normal distribution to t distribution.
- The corresponding corresponding p-values and confidence intervals are all derived from t distribution with $n - p - 1$ degrees of freedom:

$$\text{df} = \text{sample size} - \text{number of parameters}$$

- The testing procedures

Testing multiple parameters

- Sometimes we are also interested in testing multiple parameters. For example, in the previous gala data fitting, our model is

$$\begin{aligned}\text{Species} = \beta_0 + \beta_1 \text{Area} + \beta_2 \text{Elevation} + \beta_3 \text{Nearest} \\ + \beta_4 \text{Scruz} + \beta_5 \text{Adjacent} + \epsilon\end{aligned}$$

- What if we want to test jointly:

$$\text{H}_0 : \beta_4 = \beta_5 = 0 \quad \text{vs.} \quad \text{H}_1 : \text{any of } \beta_4 \text{ and } \beta_5 \text{ is nonzero}$$

Full model vs. Reduced model

- The first model is referred to as the "full model" — the model that contains all predictors.
- The second model, with $\beta_4 = \beta_5 = 0$, is referred to as the "reduced model".
- How to decide if including the two predictors Scruz and Adjacent is necessary? We need to quantify the variations explained by them.
  1). Fit the full model with all predictors, and obtain the sum of squared errors: SSE_F
  1). Fit the reduced model, and obtain the sum of squared errors: SSE_R
- How to draw conclusion by comparing the two? What distribution to use?

The F-distribution

- The F test statistic for testing multiple parameters is given by

$$F = \frac{(\mathsf{SSE}_\mathsf{R} - \mathsf{SSE}_\mathsf{F})/q}{\mathsf{SSE}_\mathsf{F}/(n-p-1)}$$

  where q is the number of restrictions in the hypothesis test (in the previous example, its 2), and $(n-p-1)$ is the degrees of freedom of the residuals in the full model.
- Intuition: do the additional parameters explain a “significant portion” of the variation?

The F-distribution

- The F distribution has a density function

$$f(x) = \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}x\right)^{-\frac{d_1+d_2}{2}}$$

  where d_1 and d_2 are two degrees of freedoms.
- It can be viewed as the ratio of two independent χ^2 distributed variables (if we divide both by the true variance σ^2), scaled by their d.f. respectively:

$$\frac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2)$$

  where $X_1 \sim \chi^2(d_1)$ and $X_2 \sim \chi^2(d_2)$.
- Reject the hypothesis when the test statistic is greater than $F_{1-\alpha}(d_1, d_2)$
- This is a one-tailed test.

Example: gala dataset

In the gala model

$$\text{Species} = \beta_0 + \beta_1\text{Area} + \beta_2\text{Elevation} + \beta_3\text{Nearest} + \beta_4\text{Scruz} + \beta_5\text{Adjacent} + \epsilon$$

test the hypothesis that

$$\text{H}_0 : \beta_4 = \beta_5 = 0 \quad \text{vs.} \quad \text{H}_1 : \text{any of } \beta_4 \text{ and } \beta_5 \text{ is nonzero}$$

```
> fit = lm(Species~ Area + Elevation + Nearest + Scruz +
    Adjacent, data = gala)
> SSEF = deviance(fit)
> fit2 = lm(Species~ Area + Elevation + Nearest, data = gala)
> SSER = deviance(fit2)
> Fstat = ((SSER - SSEF) / 2) / (SSEF / df.residual(fit))
> 1 - pf(Fstat, 2, df.residual(fit))
[1] 0.0004398959
```

Example: gala dataset

- Consider the same full mode, test the hypothesis

$$\mathsf{H}_0 : \beta_4 = 0 \text{ (for Scruz)} \quad \text{vs.} \quad \mathsf{H}_1 : \beta_4 \neq 0$$

  using F test.

- What is your conclusion at 95% confidence?

$$\mathsf{A} : \text{reject} \quad \mathsf{B} : \text{do not reject}$$

Relationship between t and F

- What is the p-value of the previous test? Isn't it the same as the t-test that we learned previously?
- There is a connection between F distribution and t distribution:
- Since t is

$$\frac{\mathcal{N}(0,1)}{\sqrt{\chi_v^2/v}}$$

  if we square this, the numerator becomes χ_1^2, so

$$\frac{\chi_1^2/1}{\chi_v^2/v}$$

  is exactly $F(1, v)$ distribution.
- t test is two-sided, and if we square that, the F test is one-sided.
- This relationship is more complicated when we test more than one parameter.

Joint test of all predictors

- In the lm function summary output, the "F-statistic" is a test of all predictors:

$$\mathsf{H}_0 : \beta_i = 0, \text{for all } 1 \le i \le p \quad \text{vs.} \quad \mathsf{H}_1 : \text{any of the } \beta_i\text{'s is nonzero}$$

- In this case, the reduced model is the "intercept" model, and SSE_R is the sum of squares total (SST). The different between the full and reduced model is $\mathsf{SSE}_\mathsf{R} - \mathsf{SSE}_\mathsf{F} = \mathsf{SSR}$, sum of squares for regression.
- Hence this F-statistic is essentially

$$\frac{\mathsf{SSR}/p}{\mathsf{SSE}/(n-p-1)} = \frac{R^2}{1-R^2}\frac{n-p-1}{p}$$

- Now validate this in our lm output.

Example: gala dataset

```
> fit = lm(Species~ Area + Elevation + Nearest + Scruz +
    Adjacent, data = gala)
> summary(fit)
Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658, Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07

> 0.7658 / (1-0.7658) *(24 / 5)
[1] 15.6953
```

What about the degrees of freedom in the F test?

Connection between F and the distribution of $\widehat{\beta}$

- We know that $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})$, so how to derive the F distribution for testing multiple parameters?
- Let's still test the parameters β_4 and β_5, what is the distribution of $(\widehat{\beta}_4, \widehat{\beta}_5)^\mathsf{T}$? Define $\mathbf{A}$ such that $\mathbf{A}\boldsymbol{\beta} = (\widehat{\beta}_4, \widehat{\beta}_5)^\mathsf{T}$, then we have

$$(\widehat{\beta}_4, \widehat{\beta}_5)^\mathsf{T} = \mathbf{A}\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\beta}, \sigma^2\mathbf{A}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{A})$$

- If we multiple $\left(\sigma^2\mathbf{A}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{A}\right)^{-1/2}$ to $(\widehat{\beta}_4, \widehat{\beta}_5)^\mathsf{T}$, we have independent standard normal distributions under the Null, i.e. when $\beta_4 = \beta_5 = 0$.
- We can then square each term and sum up to obtain a χ^2_2 distribution, which corresponds to the numerator (divided by σ^2) in the F test statistic.
- The demonstrator (MSE) is, up to the same factor σ^2, a χ^2_{n-p-1}, which we already know.

Testing a linear constraint

- Sometimes we are interested in testing a linear constraint:

$$\mathsf{H}_0 : \mathbf{A}\boldsymbol{\beta} = c$$

  where $\mathbf{A}$ is a linear combination matrix.
- Examples: we want to test if the effect of two parameters are the same:

$$\mathsf{H}_0 : \beta_4 - \boldsymbol{\beta}_5 = 0$$

  or their sum is 1:

$$\mathsf{H}_0 : \beta_4 + \boldsymbol{\beta}_5 = -1$$

- We need to construct the $\mathbf{A}$ matrix correspondingly

Testing a linear constraint

- The corresponding $\mathbf{A}$ matrix can be constructed as :

$$\mathbf{A}\boldsymbol{\beta} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \boldsymbol{\beta} = 0$$

and

$$\mathbf{A}\boldsymbol{\beta} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \boldsymbol{\beta} = -1$$

Testing a linear constraint

- We know that

$$\mathbf{A}\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\beta}, \sigma^2 \mathbf{A}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{A})$$

- Under the Null, we have $\mathbf{A}\boldsymbol{\beta} = c$, so a t-statistic for testing $\mathbf{A}\boldsymbol{\beta} = c$ is

$$\frac{\mathbf{A}\widehat{\boldsymbol{\beta}} - c}{\sqrt{\widehat{\sigma}^2 \mathbf{A}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{A}}}$$

- Lets try the first example.
- Use this test on testing $\mathrm{H}_0 : \beta_4 + \boldsymbol{\beta}_5 = -1$, do you reject at 95% confidence?

A : reject   B : do not reject

Example: testing a linear constraint

Testing the hypothesis

$$\mathsf{H}_0 : \beta_4 - \beta_5 = 0$$

```
# the linear combination matrix
A = c(0, 0, 0, 0, 1, -1)
c = 0

# the variance of this linear combination:

VA = t(A) %*% solve(t(X) %*% X) %*% A

sigma2 = deviance(fit) / df.residual(fit)

tstat = (t(A) %*% fit$coefficients - c)/ sqrt(VA * sigma2)

> 2*(1- pt(abs(tstat), df.residual(fit)))
          [,1]
[1,] 0.4575448
```

Confidence Ellipses

- This part is not required in your exam.
- Confidence ellipses of $(\widehat{\beta}_4, \widehat{\beta}_5)^\mathsf{T}$ can be constructed by looking at their joint normal distribution:
- We already derived their theoretical distribution, now lets look at a graphical representation
- Essentially, we can plot the (3D) density function of $(\widehat{\beta}_4, \widehat{\beta}_5)^\mathsf{T}$ (there is an example in the Intro), and try to find a cut off point of the elevation such that the densities within the cutting line is 95% (or whatever confidence level we prefer).

Example: Confidence Ellipses

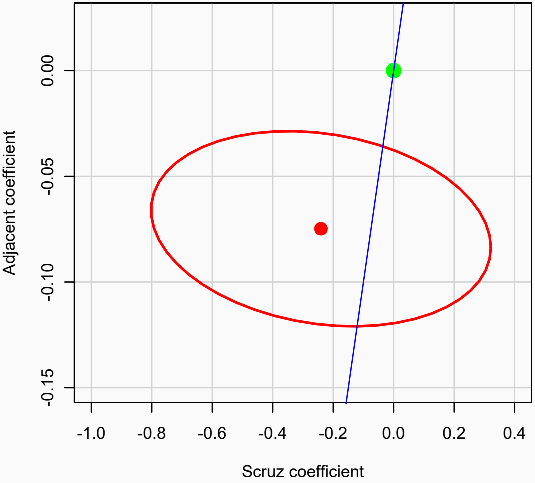
```r
# Confidence Ellipses for beta4 and beta5
library(car)

confidenceEllipse(fit, which.coef = c(5, 6),
                  levels = 0.95, xlim = c(-1, 0.4), ylim = c
                     (-0.15, 0.025))
# the test of beta4 = beta5 = 0
points(0, 0, pch = 16, cex = 2, col = "green")

# the test of beta4 - beta5 = 0
abline(a = 0, b = -1, col = "blue")
```

Example: Confidence Ellipses

How to draw conclusions of the previous tests?



Confidence Intervals for new observation

- Similar to the SLR case, we can predict the mean μ_{new} and outcome Y_{new} for a new subject with covariate x_{new} in a MLR.
- Keep in mind that in a MLR, this x_{new} has multiple dimensions, i.e.,

$$x_{\text{new}} = (x_{\text{new},1},\ x_{\text{new},2},\ \ldots,\ x_{\text{new},p})^{\mathsf{T}}$$

- If we want to get the distribution of μ_{new}, it is essentially another linear combination of $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\mu}_{\text{new}} = (1, x_{\text{new}}^{\mathsf{T}})\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\mu_{\text{new}}, (1, x_{\text{new}}^{\mathsf{T}})\sigma^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}(1, x_{\text{new}}^{\mathsf{T}})^{\mathsf{T}}\right)$$

- Again, if we want to construct the CI for μ_{new}, we replace the σ^2 with $\widehat{\sigma}^2$ as the variance and normal distribution becomes t distribution, with degrees of freedom equal to the df of the residual. The same old trick...

Confidence Intervals for new observation

- The distribution of Y_{new} is nothing but adding a variance of the error into the previous formula
- If we let the variance part be

$$V_{\text{new}} = (1, x_{\text{new}}^{\mathsf{T}})\widehat{\sigma}^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}(1, x_{\text{new}}^{\mathsf{T}})^{\mathsf{T}}$$

  we have the CI for μ_{new}

$$\widehat{\mu}_{\text{new}} \pm t_{1-\alpha/2}(n-p-1)\sqrt{V_{\text{new}}}$$

  and the CI for Y_{new}

$$\widehat{\mu}_{\text{new}} \pm t_{1-\alpha/2}(n-p-1)\sqrt{V_{\text{new}} + \widehat{\sigma}^2}$$

- These can be easily done in R.

Example

```
# specify the values of the new subject
> xnew = data.frame(Area = 260, Elevation = 360, Nearest = 10,
    Scruz = 60, Adjacent = 260)
# CI for mu_new
> predict.lm(fit, xnew, interval = c("confidence"), level =
    0.90)
      fit     lwr      upr
1 82.0623 62.9721 101.1525
# CI for Y_new
> predict.lm(fit, xnew, interval = c("prediction"), level =
    0.90)
      fit       lwr     upr
1 82.0623 -23.99138 188.116
```