Multiple Testing, F-Test, & Prediction Intervals

DS 301

Iowa State University

Recap

So far, we know:

- How to fit a linear regression model and obtain the least square estimates.
  - We know these least square estimates are unbiased estimates of the truth.
  - We can also quantify the uncertainty surrounding these estimates (standard error).
- How to obtain a realistic estimate of our model's prediction error on data it has never see.
- How to carry out inference on our model.
  - Hypothesis testing.
  - Confidence intervals.
- Assumptions needed for our model to be valid.

R output

1.sided:

pt(|ts|, df, lower.tail = FALSE)

2.sided:

" " x 2.

confint( )

β_0

β_1

⋮

β_p.

```
> summary(lm(crim~.,data=Boston))

Call:
lm(formula = crim ~ ., data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-9.924 -2.120 -0.353  1.019 75.051

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.033228   7.234903   2.354 0.018949 *
zn            0.044855   0.018734   2.394 0.017025 *
indus        -0.063855   0.083407  -0.766 0.444294
chas         -0.749134   1.180147  -0.635 0.525867
nox         -10.313535   5.275536  -1.955 0.051152 .
rm            0.430131   0.612830   0.702 0.483089
age           0.001452   0.017925   0.081 0.935488
dis          -0.987176   0.281817  -3.503 0.000502 ***
rad           0.588209   0.088049   6.680 6.46e-11 ***
tax          -0.003780   0.005156  -0.733 0.463793
ptratio      -0.271081   0.186450  -1.454 0.146611
black        -0.007538   0.003673  -2.052 0.040702 *
lstat         0.126211   0.075725   1.667 0.096208 .
medv         -0.198887   0.060516  -3.287 0.001087 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared:  0.454,     Adjusted R-squared:  0.4396
F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

$\boxed{H_0: \beta_1 = 0 \quad H_1: \beta_1 > 0.}$

$H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0.$

predict( ).

See R script `MLR_Inference.R`

Is there a relationship between X's and Y?

More precisely: is there at least one β_j, ($j = 1, \ldots, p$) that is non-zero?

What do you think of this approach?

$Y \sim X_1 + X_2 + \cdots + X_p$

$\hat{\beta}_1$ $se(\hat{\beta}_1)$

$\hat{\beta}_2$

$\vdots$ $\vdots$

$\hat{\beta}_p$ $se(\hat{\beta}_p)$

- Test each β_j separately:
  - $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$
  - $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$
  - ...
  - ...
  - $H_0 : \beta_p = 0$ versus $H_1 : \beta_p \neq 0$
- Carry out p hypothesis tests.
- If any of the individual tests is significant (p-value $< \alpha$), then this means at least one of the predictors is related to Y.

0.05

This approach is problematic..

... especially when the number of predictors p is large.

- Every time we carry out a test, there is always a chance we make a mistake.
- One type of mistake is called type 1 error: we reject H_0, but we shouldn't have. (false discovery)
- We control how large of a type 1 error we are willing to accept: α (significance level)
- For example, if we set $\alpha = 0.05$, we are willing to accept a 5% chance of making a type 1 error.

Let's apply this logic to our approach:

Suppose you have 100 predictors ($p = 100$).

$p \approx 20$

$p \approx 100$ . . .

- Carry out 100 individual tests at $\alpha = 0.05$.
- Suppose we know that H_0 is true (there is really no relationship between X's and Y).

  What is the probability we will see at least one significant result just by chance?

  P(at least one significant result) ?

= 1 - P(no significant result)

$= 1 - (0.95)^{100}$

≈ 0.994

Therefore, even when H_0 is true, we are almost guaranteed to see at least one significant result by chance.

$\Rightarrow$ Multiple testing problem

- When we carry out a large number of hypothesis tests, we are bound to get some very small p-values by chance.
- If we make a decision about whether or not to reject each hypothesis test, without taking into account the fact that we have performed a large number of tests, we may end up making a large number of type 1 errors.
- Suppose we have 10,000 tests and we set $\alpha = 0.01$. How many type 1 errors can we expect to make?

$10{,}000 \times 0.01 = 100$ false discoveries

In the context of linear regression...

... the multiple testing problem is why we cannot fully depend on individual *p*-values to tell us

1. Whether or not a relationship exists between at least of the predictors and the response,

2. Which variables are important in our model.

In the context of linear regression...

1. Does a relationship exists between at least of the predictors and the response?
   - Overall F-test.

2. Which subset of predictors are important in our model?
   - Model selection techniques: subset, forward, backward, stepwise selection.

See R script: `multiple_testing.R`

Does a relationship exists between at least one of the predictors and the response?

Overall F-test: this is a single test and it takes into account the number of predictors in our model. $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- Idea: compare the residual sum of squares (RSS) from the full model (with all predictors of interest) versus the residual sum of squares from the null model (model with no predictors).

full model: $Y \sim X_1 + X_2 + X_3 + X_4$

Null model: $Y \sim 1$

$RSS_F = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$RSS_R = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$\hat{y}_i = \bar{y}$

1. $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$
   H_1 : at least one β_j is non-zero.
2. Test statistic:

$$F^\star = \frac{(RSS_R - RSS_F)/(df_R - df_F)}{RSS_F/df_F}$$

Reject if F^* is relatively large

Details: RSS $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

- Measures fit of a model: a smaller RSS indicates a model fits data well.
- RSS_F versus RSS_R $\begin{cases} RSS_R : Y \sim 1 \\ RSS_F : Y \sim X_1 + X_2 + \cdots + X_p \end{cases}$
- It is always true that $RSS_F < RSS_R$.

question: is the difference large enough to provide evidence that the full model is a significantly better fit than the reduced model?

$df_F = (n-(p+1))$ $\qquad$ $df_R = (n-1)$

t_{df} F_{df_1, df_2}

3. Null distribution: When $\epsilon_i \sim N(0, \sigma^2)$ and we assume H_0 is true, $F^\star$ has a null distribution of $F_{p, n-(p+1)}$.

   # of predictors in full model

4. p-value given in `lm` output.

   F-tests are inherently one-sided tests (even though H_1 is two-sided). This is because we only care if our test statistic is large (not small).

R output

$H_0: \beta_1 = \beta_2 = \dots = \beta_{12} = 0$, H_1: at least one β_j, $j=1,\dots,12$ is non-zero

```
Call:
lm(formula = crim ~ ., data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-8.534 -2.248 -0.348  1.087 73.923

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.7783938  7.0818258   1.946 0.052271 .
zn           0.0457100  0.0187903   2.433 0.015344 *
indus       -0.0583501  0.0836351  -0.698 0.485709
chas        -0.8253776  1.1833963  -0.697 0.485841
nox         -9.9575865  5.2898242  -1.882 0.060370 .
rm           0.6289107  0.6070924   1.036 0.300738
age         -0.0008483  0.0179482  -0.047 0.962323
dis         -1.0122467  0.2824676  -3.584 0.000373 ***
rad          0.6124653  0.0875358   6.997 8.59e-12 ***
tax         -0.0037756  0.0051723  -0.730 0.465757
ptratio     -0.3040728  0.1863598  -1.632 0.103393
lstat        0.1388006  0.0757213   1.833 0.067398 .
medv        -0.2200564  0.0598240  -3.678 0.000261 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 6.46 on 493 degrees of freedom
Multiple R-squared:  0.4493,    Adjusted R-squared:  0.435
F-statistic: 33.52 on 12 and 493 DF,  p-value: < 2.2e-16
```

$F^* = 33.52$

Null distr:

if $\varepsilon_i \sim N(0, \sigma^2)$, then

$F^* \overset{H_0}{\sim} F_{12,493}$

pvalue: < 0.001

conclusion: reject H_0

5. Conclusion:
    - If we do not reject H_0: we do not find evidence of any significant relationship between Y and at least one of the predictors, at significant level α.
    - If we reject H_0: we find evidence of a relationship between Y and at least one of the predictors, at significance level α.

F-test limitations

Let's say we reject H_0:

- This does not mean a linear regression model is right for this data.
- It only means that the linear regression model does better than the model with no predictors, too much better to be due to chance.
- It does not tell us which predictors are useful.

Let's say we do not reject H_0:

- This could be because we made a mistake (type 2 error).
- Could be because we don't have enough power to detect departures from H_0.
- Could be because the relationship between X's and Y is non-linear.

Some Important Questions

When we perform MLR, we are usually interested in answering a few important questions.

1. What is a realistic estimate of prediction error for our model on data it has not seen before?
    - Test MSE
2. Is at least of the predictors $X_1, \ldots, X_p$ useful in predicting the response?
    - Overall F-test
3. Which subset of predictors are most useful in explaining Y?
    - Model selection (next week)
4. How well does the model fit the data?
5. Given a set of predictors, how accurate is our prediction of Y for specific values of X_1, $X_2, \ldots, X_p$?

Model Fit

R^2: coefficient of determination.

- Unit-less (does not depend on units of Y).
- Reported as a percentage (or proportion); always takes on a value between 0 and 1.
- $R^2 = 1 - \frac{RSS}{TSS}$. RSSR : Y ~ 1
  - $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$: total sum of squares
  - RSS is the residual sum of squares. $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- R^2 measures the proportion of variability in Y that can be explained by the model.