DS 3010

Spring 2026

Wenting Xu

IOWA STATE UNIVERSITY

Module 1: Multiple Linear Regression

Part 4: Multiple Testing

Wenting Xu

Iowa State University

Copyright & Attribution

Some lecture slides and instructional materials in this course are adapted from the following sources:

- *An Introduction to Statistical Learning: With Applications in R (Second Edition)*
Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
Springer, 2021

- Online course materials developed by Trevor Hastie, Robert Tibshirani, and collaborators.

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 true population parameters.
  - 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 seen before.
- How to carry out inference on our model.
  - Hypothesis testing.
  - Confidence intervals.

Is there a relationship between X's and Y?

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

What do you think of this approach?

- 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.

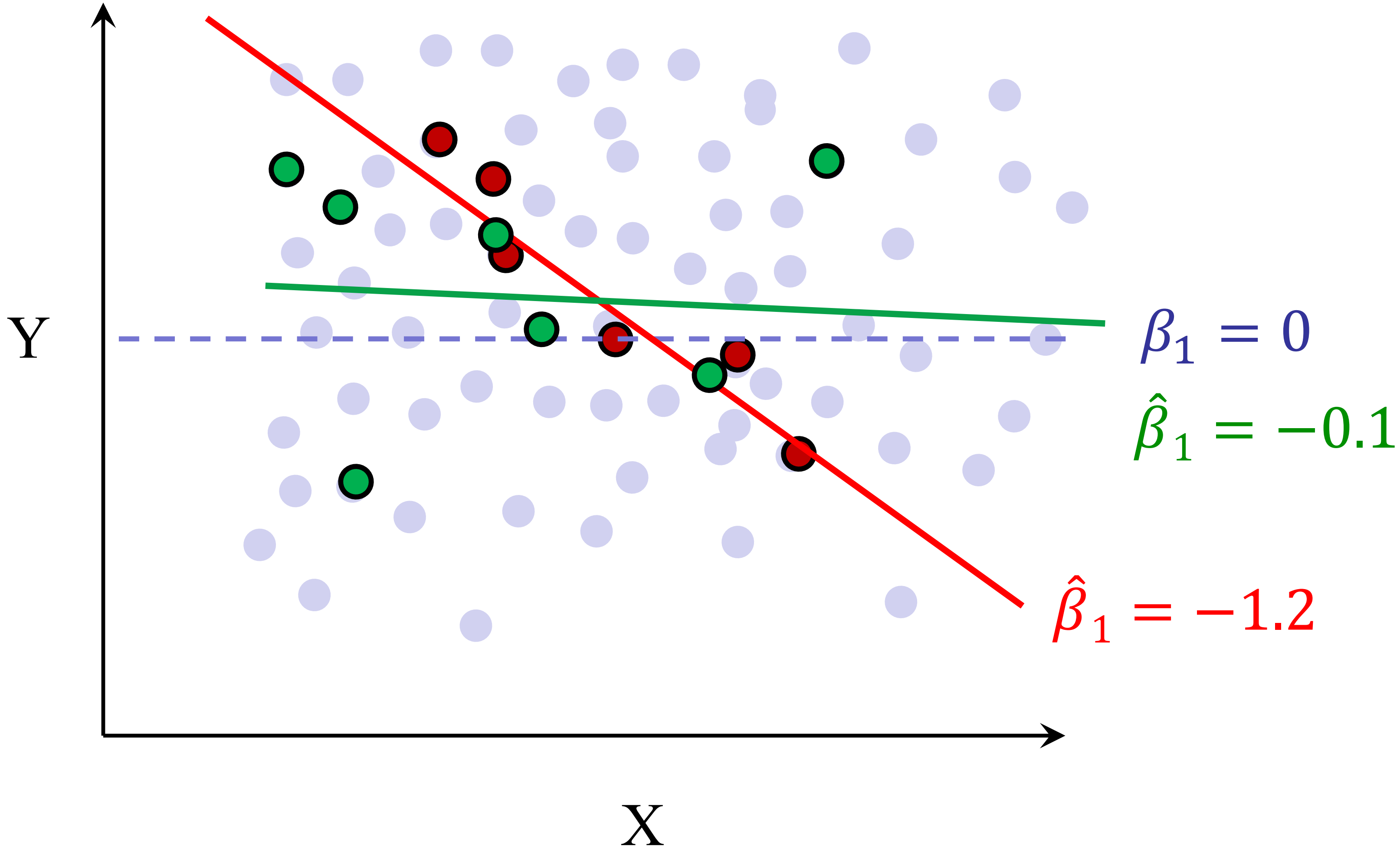
- If any of the individual tests is significant (p-value $< \alpha$), then this means at least one of the predictors is related to Y.

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.
- 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.

Type I error



Type 1 error: we reject *H*0, but we shouldn't have.

Let’s apply this logic to our approach:

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

- 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 results just by chance? $\rightarrow 0, \underbrace{1, 2, 3, \ldots, 100}_{\text{"at least 1"}}$

$\mathbb{P}(\text{at least 1 sig. result})$

$= 1 - \mathbb{P}(0)$

$= 1 - (0.95)^{100}$ ← right decision for all predictors

$\approx 99.5\%$

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

⇒ **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? → $10{,}000 \cdot 0.01 = 100$

R scipts

See R script: multiple testing.R

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 predictors are important in our model.

In the context of linear regression...

1. Does a relationship exists between at least one of the predictors and the response? → is the model useful?
   - Overall *F* -test.

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

Does a relationship exists between at least 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.

- 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).**

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

low → good model

none

full (all)

2. Test statistic:

so for a good model, this is high →

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

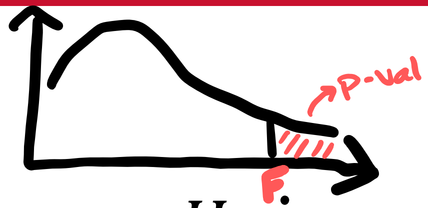
degree of freedom

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.
- RSS_F : RSS of full model.
- RSS_R : RSS of reduced model.
- It is always true that $RSS_F < RSS_R$.

If the full model is good, its RSS will be much smaller, which leads to a **large F-statistic** and a **small p-value**.

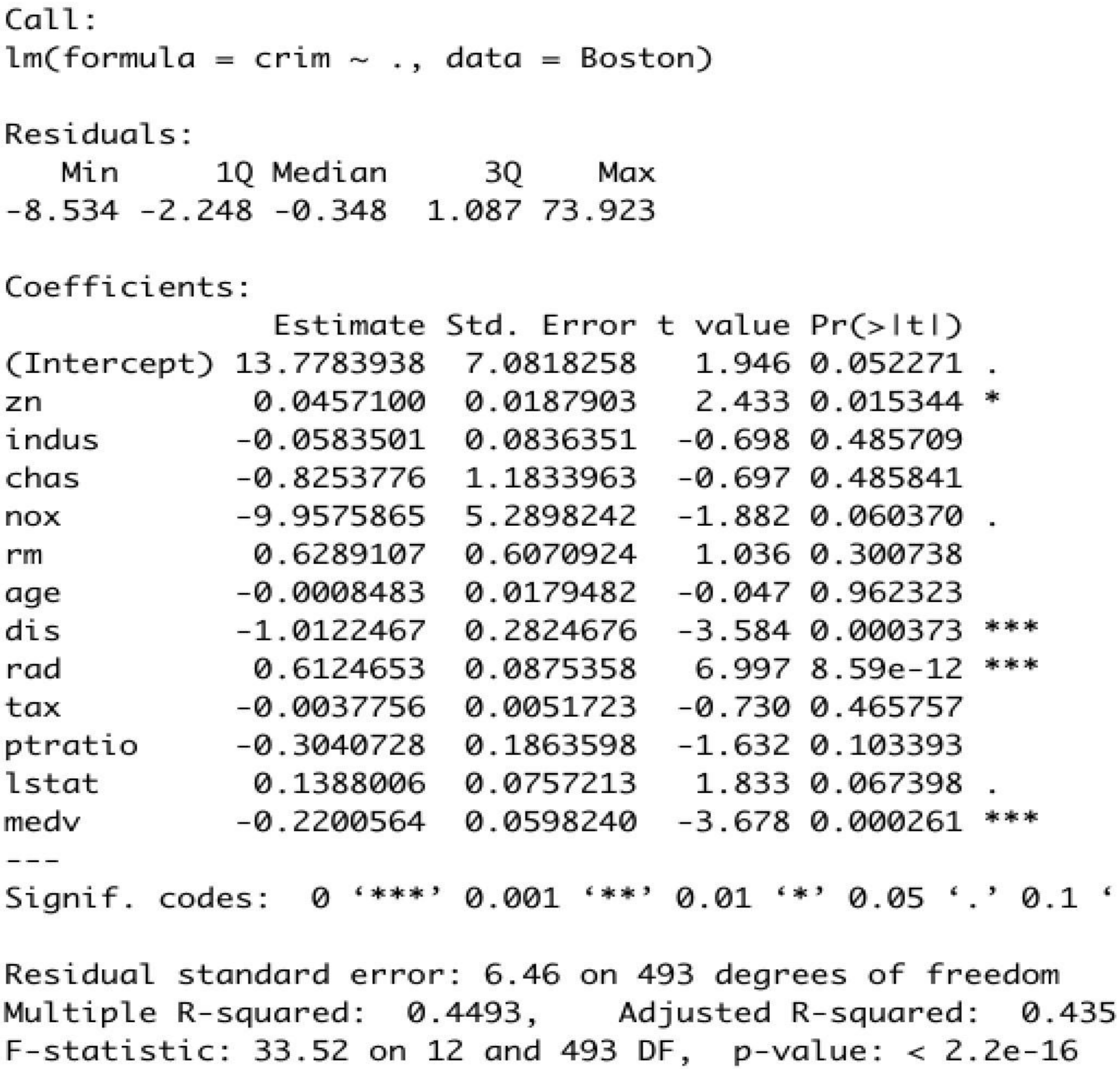
larger F → smaller p-value



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 $F_{p,n-(p+1)}$.

4. p-value given in l m 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).



```
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
```

If p-value > α → don't reject H_0

↳ large p-val => model is no better than a model w/ no predictors

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*.

p-value < 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.