Stat 206: Linear Models

✓ Lecture 9

October 23, 2019

Recap: Sampling Distributions of Sums of Squares (SS)

Under the Normal error model:

- SSE and SSR are independent.
- $SSE \sim \sigma^2 \chi^2_{(n-p)}$.
- If $\beta_1 = \cdots = \beta_{p-1} = 0$, then $SSR \sim \sigma^2 \chi^2_{(p-1)}$.

Mean squares (MS): **MS** $=$ **SS**/**d.f.**(**SS**).

- MSE:

$$MSE = \frac{SSE}{n-p}, \quad E(MSE) = \sigma^2.$$

MSE is an estimator of the error variance σ^2**.**

- MSR:

$$MSR = \frac{SSR}{p-1}.$$

$$E(MSR) = \begin{cases} & \textit{if} \quad \beta_1 = \cdots = \beta_{p-1} = 0 \\ & \textit{if} \quad \textit{otherwise} \end{cases}$$

Mean squares (MS): $\mathbf{MS} = \mathbf{SS}/\mathbf{d.f.}(\mathbf{SS})$.

- MSE:

$$MSE = \frac{SSE}{n-p}, \quad E(MSE) = \sigma^2.$$

MSE is an unbiased estimator of the error variance σ^2.

- MSR:

$$MSR = \frac{SSR}{p-1}.$$

$$E(MSR) = \begin{cases} \sigma^2 & \text{if} \quad \beta_1 = \cdots = \beta_{p-1} = 0 \\ > \sigma^2 & \text{if} \quad \text{otherwise} \end{cases}$$

Why?

- $MSTO = \frac{SSTO}{n-1}$.

For n cases, up to how many X variables can be included in the model?

F Test of Regression Relation

Under the Normal error model:

- Test whether there is a regression relation between the response variable Y and the set of X variables:

- F ratio and its null distribution:

$$F^* = \qquad , \quad F^* \sim_{H_0} \qquad ,$$

  where $F_{p-1,n-p}$ denotes the F distribution with $(p-1, n-p)$ degrees of freedom.

- Decision rule at level α: reject H_0 if $F^* > \qquad .$

F Test of Regression Relation

Under the Normal error model

- Test whether there is a regression relation between the response variable Y and the set of X variables:

$$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0 \quad vs.$$

$$H_a : \text{not all } \beta_k\text{s equal zero.}$$

- F ratio and its null distribution:

$$F^* = \frac{MSR}{MSE}, \quad F^* \sim_{H_0} F_{p-1,n-p},$$

  where $F_{p-1,n-p}$ denotes the F distribution with $(p-1, n-p)$ degrees of freedom.
- Decision rule at level α: reject H_0 if $F^* > F(1-\alpha; p-1, n-p)$.

ANOVA Table

Source of Variation	SS	d.f.	MS	F^*
Regression	$SSR = \mathbf{Y}'(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)\mathbf{Y}$	$p-1$	$MSR = \frac{SSR}{p-1}$	$F^* = \frac{MSR}{MSE}$
Error	$SSE = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y}$	$n-p$	$MSE = \frac{SSE}{n-p}$	
Total	$SSTO = \mathbf{Y}'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y}$	$n-1$		

Example Model 2: $n = 30, p = 5$.

Source of Variation	SS	d.f.	MS	F^*
Regression	$SSR = 366.4846$	4	$MSR = 91.62116$	$F^* = 87.03703$
Error	$SSE = 26.31672$	25	$MSE = 1.052669$	
Total	$SSTO = 392.8013$	29		

Pvalue $= P(F_{4,25} > 87.037) \approx 0$, so there is a significant regression relation between Y and X_1, X_2, X_3, X_1X_2.

Coefficient of Multiple Determination

$$R^2 := \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}.$$

- R^2 is the of the total variation in Y by using the X variables to explain Y.
- $0 \leq R^2 \leq 1$.

  When $R^2 = 0$? When $R^2 = 1$?
- **Adding more X variables to the model will always** R^2 because:
  (i) $SSTO$.
  (ii) SSE .

Coefficient of Multiple Determination

$$R^2 := \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}.$$

- R^2 is the proportional reduction of the total variation in Y by using the X variables to explain Y.
- $0 \leq R^2 \leq 1$.
  When $R^2 = 0$? When $R^2 = 1$?
- **Adding more X variables to the model will always increase R^2** because:
  (i) $SSTO$ remains the same. *Why?*
  (ii) SSE becomes smaller. *Why?*

Since adding more X variables can only R^2, does this mean we should use as many X variables as possible?

- With more X variables, the model does fit the observed data , indicated by SSE.
- However, a model with many X variables that are unrelated to the response variable and/or are highly correlated with each other tends to
  - the observed data and often do a job for prediction (i.e., generalize poorly for new cases) due to sampling variability.
  - make interpretation .
  - make model maintenance more .

Since adding more X variables can only increase R^2, does this mean we should use as many X variables as possible?

- With more X variables, the model does fit the observed data better, indicated by smaller SSE.
- However, a model with many X variables that are unrelated to the response variable and/or are highly correlated with each other tends to
  - **overfit** the observed data and often do a poor job for prediction (i.e.,) due to increased sampling variability.
  - make interpretation difficult.
  - make model maintenance more costly.

Adjusted Coefficient of Multiple Determination

Adjust for of X variables in the model:

- R_a^2 R^2.
- R_a^2 **may become when adding more X variables into the model** because:
  - the in SSE may be more than offset by the

    in SSE.

Adjusted Coefficient of Multiple Determination

Adjust for the number of X variables in the model:

$$R_a^2 = 1 - \frac{MSE}{MSTO} = 1 - \frac{n-1}{n-p}\frac{SSE}{SSTO}.$$

- $R_a^2 \leq R^2$.
- R_a^2 **may become smaller when adding more X variables into the model** because:
  - the decrease in SSE may be more than offset by the loss of degrees of freedom in SSE.

Example

- Model 1: $Y \sim X_1, X_2, X_3$

$$R^2 = 0.8883, \quad R_a^2 = 0.8754$$

- Model 2 : $Y \sim X_1, X_2, X_3, X_1 X_2$

$$R^2 = 0.933, \quad R_a^2 = 0.9223.$$

- Model 3: $Y \sim X_1, X_2, X_3, X_1 X_2, X_1 X_3, X_2 X_3.$

$$R^2 = 0.937, \quad R_a^2 = 0.9205.$$

(i) For each model, $R^2 > R_a^2$; (ii) Adding more X variable(s) increases R^2. The increase of R^2 is much more from Model 1 to Model 2 than from Model 2 to Model 3; (iii) Model 3 has a smaller R_a^2 than Model 2.

Inferences about Regression Coefficients

LS estimators:

$$\underset{p\times 1}{\hat{\boldsymbol{\beta}}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix} =$$

$$\underset{p\times 1}{\mathbf{E}\{\hat{\boldsymbol{\beta}}\}} = \qquad , \quad \underset{p\times p}{\boldsymbol{\sigma}^2\{\hat{\boldsymbol{\beta}}\}} = \qquad .$$

The standard error of $\hat{\beta}_k$, $s(\hat{\beta}_k)$, is the

of $MSE(\mathbf{X}'\mathbf{X})^{-1}$.

Inferences about Regression Coefficients

LS estimators:

$$\underset{p\times 1}{\hat{\boldsymbol{\beta}}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix} = \underset{p\times p}{(\mathbf{X}'\mathbf{X})^{-1}} \underset{p\times n}{\mathbf{X}'} \underset{n\times 1}{\mathbf{Y}}.$$

$$\underset{p\times 1}{\mathbf{E}\{\hat{\boldsymbol{\beta}}\}} = \boldsymbol{\beta}, \quad \underset{p\times p}{\boldsymbol{\sigma}^{\mathbf{2}}\{\hat{\boldsymbol{\beta}}\}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

The standard error of $\hat{\beta}_k$, $s(\hat{\beta}_k)$, is the positive square-root of the $(k+1)th$ diagonal element of $MSE(\mathbf{X}'\mathbf{X})^{-1}$.

- Studentized pivotal quantity:

$$\frac{\hat{\beta}_k - \beta_k}{s\{\hat{\beta}_k\}} \sim \qquad .$$

- $(1-\alpha)$-Confidence interval for β_k:

- T statistic:

$$T^* = \qquad .$$

- Two-sided T-Test: $H_0 : \beta_k = \beta_k^0$ vs. $H_a : \beta_k \neq \beta_k^0$.
At level α, the decision rule is to reject H_0 if and only if $|T^*| \qquad .$
What are decision rules for one-sided tests?

- Studentized pivotal quantity:

$$\frac{\hat{\beta}_k - \beta_k}{s\{\hat{\beta}_k\}} \sim t_{(n-p)}.$$

- $(1-\alpha)$-Confidence interval for β_k:

$$\hat{\beta}_k \pm t(1-\alpha/2; (n-p))s\{\hat{\beta}_k\}.$$

- T statistic:

$$T^* = \frac{\hat{\beta}_k - \beta_k^0}{s\{\hat{\beta}_k\}} \underset{H_0}{\sim} t_{(n-p)}.$$

- Two-sided T-Test: $H_0 : \beta_k = \beta_k^0$ *vs.* $H_a : \beta_k \neq \beta_k^0$.
  At level α, the decision rule is to reject H_0 if and only if $|T^*| > t(1-\alpha/2; (n-p))$.
  What are decision rules for one-sided tests?

Multiple Regression: Example

$n = 30$ cases, response variable Y and three predictor variables X_1, X_2, X_3.

case	Y	X1	X2	X3
1	3.01	1.06	0.86	-1.23
2	-3.40	-0.30	-0.08	-0.48
3	2.74	1.05	0.22	-0.40
...	...	...	...	...
30	-1.42	2.12	-0.8	-0.62

Example: Model 2

Nonadditive model with interaction between X_1 and X_2:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \epsilon_i, \quad i = 1, \cdots, 30.$$

$(p = 5)$

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X1:X2, data = data)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.8832      0.2153    4.103  0.00038 ***
X1             1.5946      0.2421    6.587 6.69e-07 ***
X2             1.7091      0.2605    6.560 7.16e-07 ***
X3             2.1266      0.2687    7.916 2.85e-08 ***
X1:X2          1.0076      0.2467    4.084  0.00040 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.026 on 25 degrees of freedom
Multiple R-squared: 0.933,      Adjusted R-squared: 0.9223
F-statistic: 87.04 on 4 and 25 DF,  p-value: 2.681e-14
```

Model 3

Test whether there is an interaction between X_1 and X_2. Use $\alpha = 0.01$.

- H_0 : , *vs.*, H_a : .
- $T^* =$
- $n = 30, p = 5,$.
- Since , the null hypothesis and conclude that there is interaction effect between X_1 and X_2.
- Alternatively, pvalue= , so H_0.

Notes: pvalue for the two-sided alternative is in the R output.

What is a 99% confidence interval for β_4? How to test the right-sided alternative?

Test whether there is an interaction between X_1 and X_2. Use $\alpha = 0.01$.

- $H_0 : \beta_4 = 0, \quad vs., \quad H_a : \beta_4 \neq 0.$
- $T^* = \frac{1.0076-0}{0.2467} = 4.084.$
- $n = 30, p = 5, t(0.995; 25) = 2.787.$
- Since $|4.084| > 2.787$, reject the null hypothesis and conclude that there is a significant interaction effect between X_1 and X_2.
- Alternatively, pvalue=$P(|t_{(25)}| > |4.084|) = 0.00040 < 0.01$, so reject H_0.

Notes: pvalue for the two-sided alternative is in the R output.

What is a 99% confidence interval for β_4? How to test the right-sided alternative?

Estimation of the Mean Response

- For a given set of values of the X variables:

$$\underset{p\times 1}{\mathbf{X}_h} = \begin{bmatrix} 1 \\ X_{h1} \\ \vdots \\ X_{h,p-1} \end{bmatrix}$$

- Corresponding mean response:

$$E(Y_h) =$$

Estimation of the Mean Response

- For a given set of values of the X variables:

$$\underset{p\times 1}{\mathbf{X}_h} = \begin{bmatrix} 1 \\ X_{h1} \\ \vdots \\ X_{h,p-1} \end{bmatrix}$$

- Corresponding mean response:

$$E(Y_h) = \mathbf{X}_h'\boldsymbol{\beta} = \beta_0 + \beta_1 X_{h1} + \cdots + \beta_{p-1} X_{h,p-1}.$$

- $\widehat{Y}_h :=$ is an estimator of $E(Y_h)$:

$$E(\widehat{Y}_h) = \qquad .$$

$$\sigma^2(\widehat{Y}_h) = \qquad .$$

- Standard error of $\widehat{Y}_h$:

$$s(\widehat{Y}_h) = \qquad .$$

- $(1-\alpha)$-confidence interval for $E(Y_h)$:

- $\widehat{Y}_h := \mathbf{X}_h'\hat{\beta}$ is an unbiased estimator of $E(Y_h)$:

$$E(\widehat{Y}_h) = E(\mathbf{X}_h'\hat{\beta}) = \mathbf{X}_h'\mathbf{E}\{\hat{\beta}\} = \mathbf{X}_h'\beta = E(Y_h).$$

$$\sigma^2(\widehat{Y}_h) = \mathbf{X}_h'\boldsymbol{\sigma}^2\{\hat{\beta}\}\mathbf{X}_h = \sigma^2\left(\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h\right).$$

- Standard error of $\widehat{Y}_h$:

$$s(\widehat{Y}_h) = \sqrt{MSE\left(\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h\right).}$$

- $(1-\alpha)$-confidence interval for $E(Y_h)$:

$$\widehat{Y}_h \pm t(1-\alpha/2; n-p)s(\widehat{Y}_h).$$

Example

Estimate the mean response when $X_1 = 0.8, X_2 = 0.5, X_3 = -1$ under Model 2.

- $\mathbf{X}_h' =$.
- $n = 30, p = 5$:

$$\widehat{Y}_h := \mathbf{X}_h'\hat{\beta} = 1.290,$$

$$\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h = 0.170, \quad MSE = 1.053,$$

$$s(\widehat{Y}_h) = \qquad .$$

- A 99%-confidence interval for $E(Y_h)$: $t(0.995; 25) = 2.787$

$$1.290 \pm 2.787 \times 0.423 = [0.111, 2.469].$$

Example

Estimate the mean response when $X_1 = 0.8, X_2 = 0.5, X_3 = -1$ under Model 2.

- $\mathbf{X}'_h = \begin{bmatrix} 1 & 0.8 & 0.5 & -1 & 0.8 \times 0.5 \end{bmatrix}$
- $n = 30, p = 5$:

$$\widehat{Y}_h := \mathbf{X}'_h \hat{\beta} = 1.290,$$

$$\mathbf{X}'_h (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_h = 0.170, \quad MSE = 1.053$$

$$s(\widehat{Y}_h) = \sqrt{1.053 \times 0.170} = 0.423.$$

- A 99%-confidence interval for $E(Y_h)$: $t(0.995; 25) = 2.787$

$$1.290 \pm 2.787 \times 0.423 = [0.111, 2.469].$$