

ST3131 Regression Analysis

AY2021/22 Semester 2

1. Simple Linear Regression

1.1. Introduction

- Pearson's Correlation: Let X and Y be two random variables, the theoretical Pearson's Correlation is defined as $\rho_{xy} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$.
- $\rho_{xy} > 0 \rightarrow$ positive correlation
- $\rho_{xy} < 0 \rightarrow$ negative correlation
- $\rho_{xy} = 0 \rightarrow$ no linear relationship
- Simple Regression Model: $Y = \beta_0 + \beta_1 X + \epsilon$, where β_0, β_1 are regression coefficients and ϵ is a random error.
- Regression Function: $EY = \beta_0 + \beta_1 X$.
- General assumptions:
 - 1) x_i and ϵ_i are independent.
 - 2) ϵ_i 's have mean zero.
 - 3) ϵ_i 's are pairwise uncorrelated ($\text{Cov}(\epsilon_i, \epsilon_j) = 0$).
 - 4) (Homogeneity) ϵ_i 's have common variance σ^2 .
 - 5) (Normality) ϵ_i 's have a normal distribution.

1.2. The Estimation of LRM

- Least Square Estimation of β_0 and β_1 :

(β_0, β_1) minimizes $E(Y - b_0 - b_1 X)^2$

$$= \frac{1}{n} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Let $Q = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$. By minimizing Q , we can estimate the parameters:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{X})^2}$$

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\bar{Y} = \frac{1}{n} \sum_{i=1}^n y_i$.

- Least Square Estimation of σ^2 :

Since $\epsilon \sim N(0, \sigma^2)$, $\sigma^2 = E(\epsilon^2) - E(\epsilon)^2 = E(\epsilon^2)$.

Hence it can be estimated by:

$$\widehat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} \rightarrow \text{degree of freedom}$$

where $\widehat{\sigma}$ is called residue standard error.

- Log Likelihood Function:

$$L(\sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

The maximum likelihood estimate (MLE) of σ^2 is given by:

$$\widehat{\sigma}_{MLE}^2 = \frac{n-2}{n} s^2$$

- Fitted Regression Function: $\bar{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{X}$

- Fitted values: $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$

- Residues: $e_i = y_i - \widehat{y}_i$
- Sum of Squares
 - Total sum of squares (SST) estimates the variance of Y :
$$\text{SST} = \sum_{i=1}^n (y_i - \bar{Y})^2$$
- Regression sum of squares (SSR) is the variation of Y explained by X :
$$\text{SSR} = \sum_{i=1}^n (\widehat{y}_i - \bar{Y})^2$$
- Residue sum of squares (SSE) is the variation of Y caused by random errors:
$$\text{SSE} = \sum_{i=1}^n (y_i - \widehat{y}_i)^2$$
- $\text{SST} \equiv \text{SSR} + \text{SSE}$

Source	d.f.	SS	MS	F
Regression	1	SSR	MSR	MSR/MSE
Error	$n-2$	SSE	MSE	
Total	$n-1$	SST		

- Coefficient of Determination: $R^2 = \frac{\text{SSR}}{\text{SST}}$
 - R^2 is the proportion of the variation in Y explained by X . It measures the strength of correlation between Y and X .
 - $R^2 = \text{corr}(Y, X)^2 = \text{corr}(Y, \hat{Y})^2$
 - Adjusted R^2 : A less biased estimate.

$$R_a^2 = \frac{n-1}{n-p-1} R^2 - \frac{p}{n-p-1}$$

1.3. Theories of Normal Distribution

- Normal Distribution: $f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$
 - Denotation: $N(\mu, \sigma^2)$
 - If $\mu = 0, \sigma^2 = 1$, it is called standard normal distribution.
- Multivariate Normal Distribution:

$$f(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi|\boldsymbol{\Sigma}|)^{\frac{q}{2}}} e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})}$$
 - Denotation: $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
 - If \mathbf{Z} is a multivariate normal vector, then for any constant matrix \mathbf{B} , $\mathbf{BZ} \sim N(\mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T)$.
 - For any constant vector \mathbf{c} , the linear combination $\mathbf{c}^T \mathbf{Z}$ has a univariate normal distribution $N(\mathbf{c}^T \boldsymbol{\mu}, \mathbf{c}^T \boldsymbol{\Sigma} \mathbf{c})$.
- χ^2 -distribution: Suppose $\mathbf{Z} = (Z_1, \dots, Z_m)^T$ is a multivariate normal vector with $N(0, \mathbf{I})$. The distribution of $\mathbf{Z}^T \mathbf{Z} = \sum_{i=1}^m Z_i^2$ is called the χ^2 -distribution with d.f. m and is denoted by χ^2_m .

- t -distribution: Suppose $Z \sim N(0,1)$, $U \sim \chi_m^2$, Z and U are independent. The distribution of $\frac{Z}{\sqrt{U/m}}$ is called the t -distribution with d.f. m and is denoted by t_m .
- F -distribution: Suppose $U \sim \chi_m^2$, $V \sim \chi_n^2$, U and V are independent. The distribution of $\frac{U/m}{V/n}$ is called the F -distribution of d.f. m and n , denoted by $F_{m,n}$.

1.4. Properties of Least Square Estimation

- Unbiasedness: $E(\hat{\beta}_1) = \beta_1$, $E(\hat{\beta}_2) = \beta_2$, $E(s^2) = \sigma^2$
- Properties of fitted parameters:
 - $\hat{\beta}_0 \sim N\left(\beta_0, \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right] \sigma^2\right)$
 - $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$
 - $(n-2)\frac{s^2}{\sigma^2} \sim \chi_{n-2}^2$
- Properties of fitted value, \hat{Y} :
 - $E\hat{Y} = EY$
 - $\text{Var}(\hat{Y}) = \text{Var}(\hat{\beta}_0) + 2X\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) + X^2\text{Var}(\hat{\beta}_1)$
 $= \left[\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right] \sigma^2$
 - $E(Y - \hat{Y})^2 = \text{Var}(\hat{Y}) - \text{Var}(\epsilon)$
 $= \left[1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right] \sigma^2$

1.5. Statistical Inference for Simple Regression Model

- Significant Test: To test whether or not there is a linear regression relationship between the response variable and the covariate.
 - Null hypothesis: There is no such relationship.
 - Test statistics: $F = \frac{MSR}{MSE}$
 - Under the null hypothesis, $F \sim F_{1,n-2}$.
 - If $F > f_{1,n-2}(\alpha)$, the null hypothesis is rejected at level α . Otherwise, it is not rejected.
- Two-sided Significant Test for β_1 :
 - Null hypothesis: $H_0: \beta_1 = 0$
 - Alternative hypothesis: $H_1: \beta \neq 0$
 - Test statistics: $T_{0\beta_1} = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)}$. Under H_0 , $T_{0\beta_1} \sim t_{n-2}$.
 Here, $s^2(\hat{\beta}_1) = \frac{s^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$.
 - p -value: $p = P(|T_{0\beta_1}| > |T_{0\beta_1}|)$
 - Decision rule (with significance level α):
 - If $|T_{0\beta_1}| > t_{n-2} \left(\frac{\alpha}{2}\right)$, reject H_0 ; otherwise, do not reject H_0 .
 - If $p < \alpha$, reject H_0 ; otherwise, do not reject H_0 .
 - $100(1 - \alpha)\%$ confidence interval for β_1 :
 $\left[\hat{\beta}_1 - t_{n-2} \left(\frac{\alpha}{2}\right) s(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2} \left(\frac{\alpha}{2}\right) s(\hat{\beta}_1)\right]$
- One-sided Significant Test for β_1 :
 - Hypotheses:
 - $H_0: \beta_1 \leq 0$ vs $H_1: \beta_1 > 0$;
 - $H_0: \beta_1 \geq 0$ vs $H_1: \beta_1 < 0$.

Note that the equal sign must with H_0 .

Decision rule:

- If $T_{0\beta_1} > t_{n-2}(\alpha)$, reject H_0 . Otherwise, do not reject H_0 .
- If $T_{0\beta_1} < -t_{n-2}(\alpha)$, reject H_0 . Otherwise, do not reject H_0 .

p -value:

- $p = P(t_{n-2} > T_{0\beta_1})$
- $p = P(t_{n-2} < T_{0\beta_1})$

$100(1 - \alpha)\%$ lower confidence bound for β_1 :

$$\beta_1 \geq \hat{\beta}_1 - t_{n-2}(\alpha) s(\hat{\beta}_1)$$

This corresponds to the one-sided test (1).

$100(1 - \alpha)\%$ lower confidence bound for β_1 :

$$\beta_1 \leq \hat{\beta}_1 + t_{n-2}(\alpha) s(\hat{\beta}_1)$$

This corresponds to the one-sided test (2).

Significant Test for β_0 :

- Test statistics: $T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_0}{s(\hat{\beta}_0)}$. Under H_0 , $T_{\beta_0} \sim t_{n-2}$.
 Here, $s^2(\hat{\beta}_0) = s^2 \left[\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]$.

Confidence Interval for $E(Y)$:

$$\hat{y}_h \pm t_{n-2} \left(\frac{\alpha}{2}\right) \sqrt{s^2 \left(\frac{1}{n} + \frac{(x_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)}$$

Prediction Interval for Y_{new} :

$$\hat{y}_h \pm t_{n-2} \left(\frac{\alpha}{2}\right) \sqrt{s^2 \left(1 + \frac{1}{n} + \frac{(x_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)}$$

1.6. R Commands and Interpretation of Results

Example – Simple Linear Regression Model

Input the data into a data frame:

```

1 impre adver
2 32.1 50.1 q1.dat=read.table(PATH_TO_FILE,header=TRUE)
3 99.6 74.1
4 11.7 19.3
5 21.9 22.9
6 60.8 82.4
7 78.6 40.1
8 92.4 185.9
9 50.7 26.9
10 21.4 20.4
11 48.1 166.2
12 48.8 27.0
13 10.4 45.6
14 88.9 154.9
15 12.0 5.0
16 29.2 49.7
17 38.0 26.9
18 10.0 5.7
19 12.3 7.6
20 23.4 9.2
21 71.1 32.4
22 4.4 6.1
23 29.2 49.7
24 38.0 26.9
25 10.0 5.7
26 12.3 7.6
27 23.4 9.2
28 71.1 32.4
29 4.4 6.1
30 29.2 49.7
31 38.0 26.9
32 10.0 5.7
33 12.3 7.6
34 23.4 9.2
35 71.1 32.4
36 4.4 6.1
37 29.2 49.7
38 38.0 26.9
39 10.0 5.7
40 12.3 7.6
41 23.4 9.2
42 71.1 32.4
43 4.4 6.1
44 29.2 49.7
45 38.0 26.9
46 10.0 5.7
47 12.3 7.6
48 23.4 9.2
49 71.1 32.4
50 4.4 6.1
51 29.2 49.7
52 38.0 26.9
53 10.0 5.7
54 12.3 7.6
55 23.4 9.2
56 71.1 32.4
57 4.4 6.1
58 29.2 49.7
59 38.0 26.9
60 10.0 5.7
61 12.3 7.6
62 23.4 9.2
63 71.1 32.4
64 4.4 6.1
65 29.2 49.7
66 38.0 26.9
67 10.0 5.7
68 12.3 7.6
69 23.4 9.2
70 71.1 32.4
71 4.4 6.1
72 29.2 49.7
73 38.0 26.9
74 10.0 5.7
75 12.3 7.6
76 23.4 9.2
77 71.1 32.4
78 4.4 6.1
79 29.2 49.7
80 38.0 26.9
81 10.0 5.7
82 12.3 7.6
83 23.4 9.2
84 71.1 32.4
85 4.4 6.1
86 29.2 49.7
87 38.0 26.9
88 10.0 5.7
89 12.3 7.6
90 23.4 9.2
91 71.1 32.4
92 4.4 6.1
93 29.2 49.7
94 38.0 26.9
95 10.0 5.7
96 12.3 7.6
97 23.4 9.2
98 71.1 32.4
99 4.4 6.1
100 29.2 49.7
101 38.0 26.9
102 10.0 5.7
103 12.3 7.6
104 23.4 9.2
105 71.1 32.4
106 4.4 6.1
107 29.2 49.7
108 38.0 26.9
109 10.0 5.7
110 12.3 7.6
111 23.4 9.2
112 71.1 32.4
113 4.4 6.1
114 29.2 49.7
115 38.0 26.9
116 10.0 5.7
117 12.3 7.6
118 23.4 9.2
119 71.1 32.4
120 4.4 6.1
121 29.2 49.7
122 38.0 26.9
123 10.0 5.7
124 12.3 7.6
125 23.4 9.2
126 71.1 32.4
127 4.4 6.1
128 29.2 49.7
129 38.0 26.9
130 10.0 5.7
131 12.3 7.6
132 23.4 9.2
133 71.1 32.4
134 4.4 6.1
135 29.2 49.7
136 38.0 26.9
137 10.0 5.7
138 12.3 7.6
139 23.4 9.2
140 71.1 32.4
141 4.4 6.1
142 29.2 49.7
143 38.0 26.9
144 10.0 5.7
145 12.3 7.6
146 23.4 9.2
147 71.1 32.4
148 4.4 6.1
149 29.2 49.7
150 38.0 26.9
151 10.0 5.7
152 12.3 7.6
153 23.4 9.2
154 71.1 32.4
155 4.4 6.1
156 29.2 49.7
157 38.0 26.9
158 10.0 5.7
159 12.3 7.6
160 23.4 9.2
161 71.1 32.4
162 4.4 6.1
163 29.2 49.7
164 38.0 26.9
165 10.0 5.7
166 12.3 7.6
167 23.4 9.2
168 71.1 32.4
169 4.4 6.1
170 29.2 49.7
171 38.0 26.9
172 10.0 5.7
173 12.3 7.6
174 23.4 9.2
175 71.1 32.4
176 4.4 6.1
177 29.2 49.7
178 38.0 26.9
179 10.0 5.7
180 12.3 7.6
181 23.4 9.2
182 71.1 32.4
183 4.4 6.1
184 29.2 49.7
185 38.0 26.9
186 10.0 5.7
187 12.3 7.6
188 23.4 9.2
189 71.1 32.4
190 4.4 6.1
191 29.2 49.7
192 38.0 26.9
193 10.0 5.7
194 12.3 7.6
195 23.4 9.2
196 71.1 32.4
197 4.4 6.1
198 29.2 49.7
199 38.0 26.9
200 10.0 5.7
201 12.3 7.6
202 23.4 9.2
203 71.1 32.4
204 4.4 6.1
205 29.2 49.7
206 38.0 26.9
207 10.0 5.7
208 12.3 7.6
209 23.4 9.2
210 71.1 32.4
211 4.4 6.1
212 29.2 49.7
213 38.0 26.9
214 10.0 5.7
215 12.3 7.6
216 23.4 9.2
217 71.1 32.4
218 4.4 6.1
219 29.2 49.7
220 38.0 26.9
221 10.0 5.7
222 12.3 7.6
223 23.4 9.2
224 71.1 32.4
225 4.4 6.1
226 29.2 49.7
227 38.0 26.9
228 10.0 5.7
229 12.3 7.6
230 23.4 9.2
231 71.1 32.4
232 4.4 6.1
233 29.2 49.7
234 38.0 26.9
235 10.0 5.7
236 12.3 7.6
237 23.4 9.2
238 71.1 32.4
239 4.4 6.1
240 29.2 49.7
241 38.0 26.9
242 10.0 5.7
243 12.3 7.6
244 23.4 9.2
245 71.1 32.4
246 4.4 6.1
247 29.2 49.7
248 38.0 26.9
249 10.0 5.7
250 12.3 7.6
251 23.4 9.2
252 71.1 32.4
253 4.4 6.1
254 29.2 49.7
255 38.0 26.9
256 10.0 5.7
257 12.3 7.6
258 23.4 9.2
259 71.1 32.4
260 4.4 6.1
261 29.2 49.7
262 38.0 26.9
263 10.0 5.7
264 12.3 7.6
265 23.4 9.2
266 71.1 32.4
267 4.4 6.1
268 29.2 49.7
269 38.0 26.9
270 10.0 5.7
271 12.3 7.6
272 23.4 9.2
273 71.1 32.4
274 4.4 6.1
275 29.2 49.7
276 38.0 26.9
277 10.0 5.7
278 12.3 7.6
279 23.4 9.2
280 71.1 32.4
281 4.4 6.1
282 29.2 49.7
283 38.0 26.9
284 10.0 5.7
285 12.3 7.6
286 23.4 9.2
287 71.1 32.4
288 4.4 6.1
289 29.2 49.7
290 38.0 26.9
291 10.0 5.7
292 12.3 7.6
293 23.4 9.2
294 71.1 32.4
295 4.4 6.1
296 29.2 49.7
297 38.0 26.9
298 10.0 5.7
299 12.3 7.6
300 23.4 9.2
301 71.1 32.4
302 4.4 6.1
303 29.2 49.7
304 38.0 26.9
305 10.0 5.7
306 12.3 7.6
307 23.4 9.2
308 71.1 32.4
309 4.4 6.1
310 29.2 49.7
311 38.0 26.9
312 10.0 5.7
313 12.3 7.6
314 23.4 9.2
315 71.1 32.4
316 4.4 6.1
317 29.2 49.7
318 38.0 26.9
319 10.0 5.7
320 12.3 7.6
321 23.4 9.2
322 71.1 32.4
323 4.4 6.1
324 29.2 49.7
325 38.0 26.9
326 10.0 5.7
327 12.3 7.6
328 23.4 9.2
329 71.1 32.4
330 4.4 6.1
331 29.2 49.7
332 38.0 26.9
333 10.0 5.7
334 12.3 7.6
335 23.4 9.2
336 71.1 32.4
337 4.4 6.1
338 29.2 49.7
339 38.0 26.9
340 10.0 5.7
341 12.3 7.6
342 23.4 9.2
343 71.1 32.4
344 4.4 6.1
345 29.2 49.7
346 38.0 26.9
347 10.0 5.7
348 12.3 7.6
349 23.4 9.2
350 71.1 32.4
351 4.4 6.1
352 29.2 49.7
353 38.0 26.9
354 10.0 5.7
355 12.3 7.6
356 23.4 9.2
357 71.1 32.4
358 4.4 6.1
359 29.2 49.7
360 38.0 26.9
361 10.0 5.7
362 12.3 7.6
363 23.4 9.2
364 71.1 32.4
365 4.4 6.1
366 29.2 49.7
367 38.0 26.9
368 10.0 5.7
369 12.3 7.6
370 23.4 9.2
371 71.1 32.4
372 4.4 6.1
373 29.2 49.7
374 38.0 26.9
375 10.0 5.7
376 12.3 7.6
377 23.4 9.2
378 71.1 32.4
379 4.4 6.1
380 29.2 49.7
381 38.0 26.9
382 10.0 5.7
383 12.3 7.6
384 23.4 9.2
385 71.1 32.4
386 4.4 6.1
387 29.2 49.7
388 38.0 26.9
389 10.0 5.7
390 12.3 7.6
391 23.4 9.2
392 71.1 32.4
393 4.4 6.1
394 29.2 49.7
395 38.0 26.9
396 10.0 5.7
397 12.3 7.6
398 23.4 9.2
399 71.1 32.4
400 4.4 6.1
401 29.2 49.7
402 38.0 26.9
403 10.0 5.7
404 12.3 7.6
405 23.4 9.2
406 71.1 32.4
407 4.4 6.1
408 29.2 49.7
409 38.0 26.9
410 10.0 5.7
411 12.3 7.6
412 23.4 9.2
413 71.1 32.4
414 4.4 6.1
415 29.2 49.7
416 38.0 26.9
417 10.0 5.7
418 12.3 7.6
419 23.4 9.2
420 71.1 32.4
421 4.4 6.1
422 29.2 49.7
423 38.0 26.9
424 10.0 5.7
425 12.3 7.6
426 23.4 9.2
427 71.1 32.4
428 4.4 6.1
429 29.2 49.7
430 38.0 26.9
431 10.0 5.7
432 12.3 7.6
433 23.4 9.2
434 71.1 32.4
435 4.4 6.1
436 29.2 49.7
437 38.0 26.9
438 10.0 5.7
439 12.3 7.6
440 23.4 9.2
441 71.1 32.4
442 4.4 6.1
443 29.2 49.7
444 38.0 26.9
445 10.0 5.7
446 12.3 7.6
447 23.4 9.2
448 71.1 32.4
449 4.4 6.1
450 29.2 49.7
451 38.0 26.9
452 10.0 5.7
453 12.3 7.6
454 23.4 9.2
455 71.1 32.4
456 4.4 6.1
457 29.2 49.7
458 38.0 26.9
459 10.0 5.7
460 12.3 7.6
461 23.4 9.2
462 71.1 32.4
463 4.4 6.1
464 29.2 49.7
465 38.0 26.9
466 10.0 5.7
467 12.3 7.6
468 23.4 9.2
469 71.1 32.4
470 4.4 6.1
471 29.2 49.7
472 38.0 26.9
473 10.0 5.7
474 12.3 7.6
475 23.4 9.2
476 71.1 32.4
477 4.4 6.1
478 29.2 49.7
479 38.0 26.9
480 10.0 5.7
481 12.3 7.6
482 23.4 9.2
483 71.1 32.4
484 4.4 6.1
485 29.2 49.7
486 38.0 26.9
487 10.0 5.7
488 12.3 7.6
489 23.4 9.2
490 71.1 32.4
491 4.4 6.1
492 29.2 49.7
493 38.0 26.9
494 10.0 5.7
495 12.3 7.6
496 23.4 9.2
497 71.1 32.4
498 4.4 6.1
499 29.2 49.7
500 38.0 26.9
501 10.0 5.7
502 12.3 7.6
503 23.4 9.2
504 71.1 32.4
505 4.4 6.1
506 29.2 49.7
507 38.0 26.9
508 10.0 5.7
509 12.3 7.6
510 23.4 9.2
511 71.1 32.4
512 4.4 6.1
513 29.2 49.7
514 38.0 26.9
515 10.0 5.7
516 12.3 7.6
517 23.4 9.2
518 71.1 32.4
519 4.4 6.1
520 29.2 49.7
521 38.0 26.9
522 10.0 5.7
523 12.3 7.6
524 23.4 9.2
525 71.1 32.4
526 4.4 6.1
527 29.2 49.7
528 38.0 26.9
529 10.0 5.7
530 12.3 7.6
531 23.4 9.2
532 71.1 32.4
533 4.4 6.1
534 29.2 49.7
535 38.0 26.9
536 10.0 5.7
537 12.3 7.6
538 23.4 9.2
539 71.1 32.4
540 4.4 6.1
541 29.2 49.7
542 38.0 26.9
543 10.0 5.7
544 12.3 7.6
545 23.4 9.2
546 71.1 32.4
547 4.4 6.1
548 29.2 49.7
549 38.0 26.9
550 10.0 5.7
551 12.3 7.6
552 23.4 9.2
553 71.1 32.4
554 4.4 6.1
555 29.2 49.7
556 38.0 26.9
557 10.0 5.7
558 12.3 7.6
559 23.4 9.2
560 71.1 32.4
561 4.4 6.1
562 29.2 49.7
563 38.0 26.9
564 10.0 5.7
565 12.3 7.6
566 23.4 9.2
567 71.1 32.4
568 4.4 6.1
569 29.2 49.7
570 38.0 26.9
571 10.0 5.7
572 12.3 7.6
573 23.4 9.2
574 71.1 32.4
575 4.4 6.1
576 29.2 49.7
577 38.0 26.9
578 10.0 5.7
579 12.3 7.6
580 23.4 9.2
581 71.1 32.4
582 4.4 6.1
583 29.2 49.7
584 38.0 26.9
585 10.0 5.7
586 12.3 7.6
587 23.4 9.2
588 71.1 32.4
589 4.4 6.1
590 29.2 49.7
591 38.0 26.9
592 10.0 5.7
593 12.3 7.6
594 23.4 9.2
595 71.1 32.4
596 4.4 6.1
597 29.2 49.7
598 38.0 26.9
599 10.0 5.7
600 12.3 7.6
601 23.4 9.2
602 71.1 32.4
603 4.4 6.1
604 29.2 49.7
605 38.0 26.9
606 10.0 5.7
607 12.3 7.6
608 23.4 9.2
609 71.1 32.4
610 4.4 6.1
611 29.2 49.7
612 38.0 26
```

Interpretation:

- ▶ Estimate: (Intercept) — $\hat{\beta}_0$, X — $\hat{\beta}_1$.
- ▶ Std. Error: $sd(\hat{\beta}_0)$, $sd(\hat{\beta}_1)$.
- ▶ t value: $\hat{\beta}_0/sd(\hat{\beta}_0)$, $\hat{\beta}_1/sd(\hat{\beta}_1)$.
- ▶ Pr(>|t|): p-values for two-sided tests on β_0 and β_1 .
- ▶ Residual standard error: $\hat{\sigma}$.
- ▶ Multiple R-squared: R^2 .
- ▶ Adjusted R-squared: R_a^2 .
- ▶ F-statistic: MSR/MSE .
- ▶ p-value: p-value of the significant F test.

2. Multiple Linear Regression

2.1. Multiple LSM and its Estimation

- Multiple Linear Regression Model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- Multiple Linear Regression Function:

$$EY = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

EY can also be written as $E(Y|X)$.

- $\mathbf{y} = X\boldsymbol{\beta} + \epsilon$

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p} \\ 1 & x_{21} & \cdots & x_{2,p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- Least Square Estimation of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$$

- Least Square Estimation of σ^2 :

$$\sigma^2 = \frac{\|\mathbf{y} - X(X^T X)^{-1} X^T \mathbf{y}\|^2}{n - p - 1}$$

- Properties of Hat Matrix $H = (X^T X)^{-1} X^T$:

- $HX = X$
- $H^2 = H$
- $(I - H)^2 = I - H$
- Residue vector: $e = \mathbf{y} - \hat{\mathbf{y}} = (I - H)\mathbf{y}$

- Sum of Squares

- SST: Let $\mathbf{y}_t = \begin{pmatrix} y_1 - \bar{y} \\ \dots \\ y_n - \bar{y} \end{pmatrix} = \mathbf{y} - \mathbf{1}\bar{y} \equiv H_t \mathbf{y}$.
 $SST = \mathbf{y}_t^T \mathbf{y}_t = \mathbf{y}^T H_t^T H_t \mathbf{y}$

- SSR:

$$SSR = \mathbf{y}_r^T \mathbf{y}_r = \mathbf{y}^T H_r^T H_r \mathbf{y}$$

- SSE:

$$SSE = \mathbf{e}^T \mathbf{e} = \mathbf{y}^T H_e^T H_e \mathbf{y}$$

- SST = SSR + SSE

- d.f. of SST, SSR, SSE: $n - 1, p, n - p - 1$

- Distributions of Sum of Squares:

- $\frac{SSE}{\sigma^2} \sim \chi^2_{n-p-1}$

- $\frac{SSR}{\sigma^2} \sim \chi^2_p$

Source of Variation	SS	df	MS	F-statistic
Regression	SSR	p	$MSR = \frac{SSR}{p}$	MSR/MSE
Error	SSE	$n - p - 1$	$MSE = \frac{SSE}{n-p-1}$	
Total	SST	$n - 1$		

- Coefficient of Multiple Determination:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{SSR}{SST}$$

- $R^2 = \text{corr}(\mathbf{y}, \hat{\mathbf{y}})^2$

- Adjusted R^2 : $R_a^2 = R^2 - \frac{p}{n-p-1}(1-R^2)$

2.2. Inference on Least Square Estimation

- Properties of Least Square Estimation:

- $\hat{\boldsymbol{\beta}} \sim N(E(\hat{\boldsymbol{\beta}}), \text{Var}(\hat{\boldsymbol{\beta}}))$

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (X^T X)^{-1}$$

- $(n - p - 1) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$

- $\frac{MSR}{MSE} \sim F_{p,n-p-1}$

- Significance Test: Does the regression function have a significant effect on the response variable?

- $H_0: \beta_1 = \cdots = \beta_p = 0$

- $H_1: \text{At least one of } \beta_1, \dots, \beta_p \neq 0$

- Test statistics: $F = \frac{MSR}{MSE}$. Under H_0 , $F \sim F_{p,n-p-1}$.

- Wald test statistics: $W = \hat{\boldsymbol{\beta}}^T \hat{\Sigma}_{\boldsymbol{\beta}}^{-1} \hat{\boldsymbol{\beta}}$. $F = \frac{W}{p}$.

- Decision rule: If $f > f_{p,n-p-1}(\alpha)$, reject H_0 ; otherwise, do not reject H_0 .

- Individual t-test:

- Hypothesis: $H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$.

- Test statistics: $t = \frac{\hat{\beta}_j}{\hat{\sigma}_j}$. Under H_0 , $t \sim t_{n-p-1}$.

- p-value: $p = 2P(t_{n-p-1} \geq |t|)$

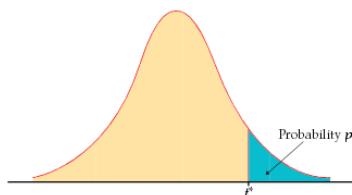
- Decision rule: If $|t| \geq t_{n-p-1} \left(\frac{\alpha}{2} \right)$, reject H_0 ; otherwise, do not reject H_0 .

- Confidence interval:

$$[\hat{\beta}_j - \hat{\sigma}_j t_{n-p-1} \left(\frac{\alpha}{2} \right), \hat{\beta}_j + \hat{\sigma}_j t_{n-p-1} \left(\frac{\alpha}{2} \right)]$$

t-distribution table

Areas in the upper tail are given along the top of the table. Critical t* values are given in the table.



df	0.1	0.05	0.025	0.02	0.01	0.005
1	3.078	6.314	12.706	15.895	31.821	63.657
2	1.886	2.920	4.303	4.843	6.965	9.925
3	1.638	2.353	3.182	3.482	4.541	5.841
4	1.533	2.132	2.776	2.993	3.747	4.604
5	1.476	2.015	2.571	2.757	3.365	4.032
6	1.440	1.943	2.447	2.612	3.143	3.707
7	1.415	1.895	2.365	2.517	2.998	3.499
8	1.397	1.860	2.306	2.449	2.896	3.355
9	1.383	1.833	2.262	2.398	2.821	3.250
10	1.372	1.812	2.228	2.353	2.764	3.169
11	1.363	1.796	2.201	2.328	2.718	3.106
12	1.356	1.782	2.179	2.303	2.681	3.055
13	1.350	1.771	2.160	2.282	2.650	3.012
14	1.345	1.761	2.145	2.264	2.624	2.977
15	1.341	1.753	2.131	2.243	2.602	2.947
16	1.337	1.746	2.120	2.235	2.583	2.921
17	1.333	1.740	2.110	2.224	2.567	2.898
18	1.330	1.734	2.101	2.214	2.552	2.878
19	1.328	1.729	2.093	2.205	2.539	2.861
20	1.325	1.725	2.086	2.197	2.528	2.845
21	1.323	1.721	2.080	2.183	2.518	2.831
22	1.321	1.717	2.074	2.183	2.508	2.819
23	1.319	1.714	2.069	2.177	2.500	2.807
24	1.318	1.711	2.064	2.172	2.492	2.797
25	1.316	1.708	2.060	2.167	2.485	2.787
26	1.315	1.706	2.056	2.162	2.479	2.779
27	1.314	1.703	2.052	2.158	2.473	2.771
28	1.313	1.701	2.048	2.154	2.467	2.763
29	1.311	1.699	2.045	2.150	2.462	2.756
30	1.310	1.697	2.042	2.147	2.457	2.750
31	1.309	1.696	2.040	2.144	2.453	2.744
32	1.309	1.694	2.037	2.141	2.449	2.738
33	1.308	1.692	2.035	2.138	2.445	2.733
34	1.307	1.691	2.032	2.136	2.441	2.728
35	1.306	1.690	2.030	2.133	2.438	2.724
36	1.306	1.688	2.028	2.131	2.434	2.719
37	1.305	1.687	2.026	2.129	2.431	2.715
38	1.304	1.686	2.024	2.127	2.429	2.712
39	1.304	1.685	2.023	2.125	2.426	2.708
40	1.303	1.684	2.021	2.123	2.423	2.704
41	1.303	1.683	2.020	2.121	2.421	2.701
42	1.302	1.682	2.018	2.120	2.418	2.698
43	1.302	1.681	2.017	2.118	2.416	2.695
44	1.301	1.680	2.015	2.116	2.414	2.692
45	1.301	1.679	2.014	2.115	2.412	2.690
46	1.300	1.679	2.013	2.114	2.410	2.687
47	1.300	1.678	2.012	2.112	2.408	2.685
48	1.299	1.677	2.011	2.111	2.407	2.682
49	1.299	1.677	2.010	2.110	2.405	2.680
50	1.299	1.676	2.009	2.109	2.403	2.678

df	0.1	0.05	0.025	0.02	0.01	0.005
51	1.298	1.675	2.008	2.108	2.402	2.676
52	1.298	1.675	2.007	2.107	2.400	2.674
53	1.298	1.674	2.006	2.106	2.399	2.672
54	1.297	1.674	2.005	2.105	2.397	2.670
55	1.297	1.673	2.004	2.104	2.396	2.668
56	1.297	1.673	2.003	2.103	2.395	2.667
57	1.297	1.672	2.002	2.102	2.394	2.665
58	1.296	1.672	2.002	2.101	2.392	2.663
59	1.296	1.671	2.001	2.100	2.391	2.662
60	1.296	1.671	2.000	2.099	2.390	2.660
61	1.296	1.670	2.000	2.099	2.389	2.659
62	1.295	1.670	1.999	2.098	2.388	2.657
63	1.295	1.669	1.998	2.097	2.387	2.656
64	1.295	1.669	1.998	2.096	2.386	2.655
65	1.295	1.669	1.997	2.096	2.385	2.654
66	1.295	1.668	1.997	2.095	2.384	2.652
67	1.294	1.668	1.996	2.095	2.383	2.651
68	1.294	1.668	1.995	2.094	2.382	2.650
69	1.294	1.667	1.995	2.093	2.382	2.649
70	1.294	1.667	1.994	2.093	2.381	2.648
71	1.294	1.667	1.994	2.092	2.380	2.647
72	1.293	1.666	1.993	2.092	2.379	2.646
73	1.293	1.666	1.993	2.091	2.379	2.645
74	1.293	1.666	1.993	2.091	2.378	2.644
75	1.293	1.665	1.992	2.090	2.377	2.643
76	1.293	1.665	1.992	2.090	2.376	2.642
77	1.293	1.665	1.991	2.089	2.376	2.641
78	1.292	1.665	1.991	2.089	2.375	2.640
79	1.292	1.664	1.990	2.088	2.374	2.640
80	1.292	1.664	1.990	2.088	2.374	2.639
81	1.292	1.664	1.990	2.087	2.373	2.638
82	1.292	1.664	1.989	2.087	2.373	2.637
83	1.292	1.663	1.989	2.087	2.372	2.636
84	1.292	1.663	1.989	2.086	2.372	2.636
85	1.292	1.663	1.988	2.086	2.371	2.635
86	1.291	1.663	1.988	2.085	2.370	2.634
87	1.291	1.663	1.988	2.085	2.370	2.634
88	1.291	1.662	1.987	2.085	2.369	2.633
89	1.291	1.662	1.987	2.084	2.369	2.632
90	1.291	1.662	1.987	2.084	2.368	2.632
91	1.291	1.662	1.986	2.084	2.368	2.631
92	1.291	1.662	1.986	2.083	2.368	2.630
93	1.291	1.661	1.986	2.083	2.367	2.630
94	1.291	1.661	1.986	2.083	2.367	2.629
95	1.291	1.661	1.985	2.082	2.366	2.629
96	1.290	1.661	1.985	2.082	2.366	2.628
97	1.290	1.661	1.985	2.082	2.365	2.627
98	1.290	1.661	1.984	2.081	2.365	2.627
99	1.290	1.660	1.984	2.081	2.365	2.626
100	1.290	1.660	1.984	2.081	2.364	2.626