

Lecture 11: Multiple linear regression

Pratheepa Jeganathan

10/16/2019

Recap

- ▶ What is a regression model?
- ▶ Descriptive statistics – graphical
- ▶ Descriptive statistics – numerical
- ▶ Inference about a population mean
- ▶ Difference between two population means
- ▶ Some tips on R

Recap

- ▶ Simple linear regression (covariance, correlation, estimation, geometry of least squares)
 - ▶ Inference on simple linear regression model
 - ▶ Goodness of fit of regression: analysis of variance.
 - ▶ F -statistics.
 - ▶ Residuals.
 - ▶ Diagnostic plots for simple linear regression (graphical methods).

Recap

- ▶ Multiple linear regression
 - ▶ Specifying the model.
 - ▶ Fitting the model: least squares.
 - ▶ Interpretation of the coefficients.

Multiple linear regression

Outline

- ▶ More on F -statistics.
- ▶ Matrix approach to linear regression.
- ▶ T -statistics revisited.
- ▶ More F statistics.

Goodness of fit for multiple regression

- ▶ After fitting the multiple regression model, we can define the sum of squares as follows:

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

$$SSR = \sum_{i=1}^n (\bar{Y} - \hat{Y}_i)^2$$

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2 = SSE + SSR.$$

- ▶ R^2 is called the *multiple correlation coefficient* of the model, or the *coefficient of multiple determination*.
- ▶ The sums of squares and R^2 are defined analogously to those in simple linear regression.

$$R^2 = \frac{SSR}{SST}.$$

Example (sum of squares and mean squares)

```
Y = prostate$lpsa
n = length(Y)
SST = sum((Y - mean(Y))^2)
# degrees of freedom for SST is (n-1)
MST = SST / (n - 1)
SSE = sum(resid(prostate.lm)^2)
# degrees of freedom for SSE (n-p-1)
# for an intercept model
MSE = SSE / prostate.lm$df.residual
SSR = SST - SSE
# degrees of freedom for SSR =
# degrees of freedom for SST -
# degrees of freedom for SSE =
# p (if we consider an intercept model)
MSR = SSR / (n - 1 - prostate.lm$df.residual)
```


Example (sum of squares and mean squares)

```
print(c(SST, SSE, SSR))
```

```
## [1] 127.91766 43.10756 84.81010
```

```
print(c(MST,MSE,MSR))
```

```
## [1] 1.3324756 0.4843546 12.1157287
```

Adjusted R^2

- ▶ As we add more and more variables to the model – even random ones, R^2 will increase to 1.
- ▶ Adjusted R^2 tries to take this into account by replacing sums of squares by *mean squares*

$$R_a^2 = 1 - \frac{SSE/(n - p - 1)}{SST/(n - 1)} = 1 - \frac{MSE}{MST}.$$

Example (Adjusted R^2)

```
round(1 - MSE/MST, digits = 3)
```

```
## [1] 0.637
```

```
round(summary(prostate.lm)$adj.r.squared, digits = 3)
```

```
## [1] 0.637
```

Goodness of fit test

- For the intercept model with p predictors, the analysis of variance (ANOVA) table is as follows:

Source	df	SS	MS	F-statistic
Regression	p	$SSR = \sum_{i=1}^n (\bar{Y} - \hat{Y}_i)^2$	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
Error	$n - p - 1$	$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n - p - 1}$	
Total	$n - 1$	$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$	$MST = \frac{SST}{n - 1}$	

Goodness of fit test

- ▶ As in simple linear regression, we measure the goodness of fit of the regression model by

$$F = \frac{MSR}{MSE}.$$

- ▶ Under $H_0 : \beta_1 = \cdots = \beta_p = 0$,

$$F \sim F_{p, n-p-1}$$

so reject H_0 at level α if $F > F_{p, n-p-1, 1-\alpha}$.

Call:

```
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +  
    pgg45, data = prostate)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.76395	-0.35764	-0.02143	0.37762	1.58178

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.494155	0.873567	0.566	0.57304	
lcavol	0.569546	0.085847	6.634	2.46e-09	***
lweight	0.614420	0.198449	3.096	0.00262	**
age	-0.020913	0.010978	-1.905	0.06000	.
lbph	0.097353	0.057584	1.691	0.09441	.
svi	0.752397	0.238180	3.159	0.00216	**
lcp	-0.104959	0.089347	-1.175	0.24323	
pgg45	0.005324	0.003385	1.573	0.11923	

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.696 on 89 degrees of freedom

Multiple R-squared: 0.663, Adjusted R-squared: 0.6365

F-statistic: 25.01 on 7 and 89 DF, p-value: < 2.2e-16

F -test revisited

The F test can be thought of as comparing two models:

- ▶ *Full (bigger) model :*

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots \beta_p X_{ip} + \varepsilon_i$$

- ▶ *Reduced (smaller) model:*

$$Y_i = \beta_0 + \varepsilon_i$$

- ▶ The F -statistic has the form

$$F = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F}.$$

- ▶ **Note: the smaller model should be nested within the bigger model.**

Example

```
prostate.lm.reduced = lm(lpsa ~ 1, data = prostate)
anova(prostate.lm.reduced, prostate.lm)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Model 1: lpsa ~ 1
```

```
## Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + l
```

```
##   Res.Df      RSS Df Sum of Sq      F      Pr(>F)
```

```
## 1      96 127.918
```

```
## 2      89  43.108  7      84.81 25.014 < 2.2e-16 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```


Matrix formulation



$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

- ▶ In matrix form:

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times (p+1)} \boldsymbol{\beta}_{(p+1) \times 1} + \boldsymbol{\epsilon}_{n \times 1}$$

- ▶ \mathbf{X} is called the *design matrix* of the model
- ▶ $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_{n \times n})$ is multivariate normal

SSE in matrix form

$$SSE(\beta) = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2$$

Design matrix

- ▶ The **design matrix** is the $n \times (p + 1)$ matrix with entries

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix}.$$

Example (design matrix)

```
n = nrow(prostate)
attach(prostate)
X = cbind(rep(1,n), lcavol,
          lweight, age, lbph, svi, lcp, pgg45)
detach(prostate)
colnames(X)[1] = '(Intercept)'
```

Example (design matrix)

```
print(xtable(head(X)))
```

% latex table generated in R 3.6.0 by xtable 1.8-4 package % Tue
Oct 15 17:27:05 2019

	(Intercept)	lcavol	lweight	age	lbph	svi	lcp	pgg45
1	1.00	-0.58	2.77	50.00	-1.39	0.00	-1.39	0.00
2	1.00	-0.99	3.32	58.00	-1.39	0.00	-1.39	0.00
3	1.00	-0.51	2.69	74.00	-1.39	0.00	-1.39	20.00
4	1.00	-1.20	3.28	58.00	-1.39	0.00	-1.39	0.00
5	1.00	0.75	3.43	62.00	-1.39	0.00	-1.39	0.00
6	1.00	-1.05	3.23	50.00	-1.39	0.00	-1.39	0.00

Example (design matrix)

- ▶ The matrix **X** is the same as formed by R

```
print(xtable(head(model.matrix(prostate.lm))))
```

% latex table generated in R 3.6.0 by xtable 1.8-4 package % Tue
Oct 15 17:28:42 2019

	(Intercept)	lcavol	lweight	age	lbph	svi	lcp	pgg45
1	1.00	-0.58	2.77	50.00	-1.39	0.00	-1.39	0.00
2	1.00	-0.99	3.32	58.00	-1.39	0.00	-1.39	0.00
3	1.00	-0.51	2.69	74.00	-1.39	0.00	-1.39	20.00
4	1.00	-1.20	3.28	58.00	-1.39	0.00	-1.39	0.00
5	1.00	0.75	3.43	62.00	-1.39	0.00	-1.39	0.00
6	1.00	-1.05	3.23	50.00	-1.39	0.00	-1.39	0.00