Chapter 11: Multiple Regression

MATH 560-01
Statistical Data Analysis

April 5, 2021

These slides are based on material from *Introduction to the Practice of Statistics* by David S. Moore, George P. McCabe, and Bruce A. Craig, 9th edition.

Sections

11.1 Inference for Multiple Regression

After completing this section, students should be able to:

- compute *confidence intervals* for the regression coefficients of a multiple linear regression model with normal errors

11.1 Inference for Multiple Regression

After completing this section, students should be able to:

- compute *confidence intervals* for the regression coefficients of a multiple linear regression model with normal errors
- perform *one-sided* or *two-sided tests* for significance of the regression coefficients in a multiple linear regression model with normal errors

11.1 Inference for Multiple Regression

After completing this section, students should be able to:

- compute *confidence intervals* for the regression coefficients of a multiple linear regression model with normal errors
- perform *one-sided* or *two-sided tests* for significance of the regression coefficients in a multiple linear regression model with normal errors
- perform an *overall F test* in a multiple linear regression model with normal errors

11.1 Inference for Multiple Regression

- In Chapter 10, we discussed the simple linear regression model which modeled a response variable y using a linear function of a *single* explanatory variable x.

11.1 Inference for Multiple Regression

- In Chapter 10, we discussed the simple linear regression model which modeled a response variable y using a linear function of a *single* explanatory variable x.
- In this chapter, we use more than one explanatory variable to explain or predict the response variable.

11.1 Inference for Multiple Regression

- In Chapter 10, we discussed the simple linear regression model which modeled a response variable y using a linear function of a *single* explanatory variable x.
- In this chapter, we use more than one explanatory variable to explain or predict the response variable.
- Many of the same ideas that we used in making inferences for simple linear regression models also apply for multiple linear regression models.

11.1 Inference for Multiple Regression

- In Chapter 10, we discussed the simple linear regression model which modeled a response variable y using a linear function of a *single* explanatory variable x.
- In this chapter, we use more than one explanatory variable to explain or predict the response variable.
- Many of the same ideas that we used in making inferences for simple linear regression models also apply for multiple linear regression models.
- However, there are some more complicated additional issues which arise. In this chapter, we only discuss some basic facts for making inferences for multiple regression models.

11.1 Inference for Multiple Regression

Population multiple regression equation

- We assume the population follows a **multiple regression model**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

where y is the response variable, $x_1, x_2, \ldots, x_p$ are the p explanatory variables, and ε follows a $N(0, \sigma)$ distribution.

11.1 Inference for Multiple Regression

- We assume the observations in the sample

$$
\begin{aligned}
\text{Case } 1 &: \quad (x_{11}, x_{12}, \ldots, x_{1p}, y_1) \\
\text{Case } 2 &: \quad (x_{21}, x_{22}, \ldots, x_{2p}, y_2) \\
&\vdots \\
\text{Case } n &: \quad (x_{n1}, x_{n2}, \ldots, x_{np}, y_p)
\end{aligned}
$$

are independent observations from the multiple regression model following the equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

where the errors ε_i are independent and Normally distributed with mean 0 and standard deviation σ.

11.1 Inference for Multiple Regression

Parameter estimation

- Estimation of the model parameters is more complicated in the multiple regression model.

11.1 Inference for Multiple Regression

<u>Parameter estimation</u>

- Estimation of the model parameters is more complicated in the multiple regression model.
- Formulas for the estimates are not discussed in the book, but we will discuss them in these slides.

11.1 Inference for Multiple Regression

Parameter estimation

- Estimation of the model parameters is more complicated in the multiple regression model.
- Formulas for the estimates are not discussed in the book, but we will discuss them in these slides.
- To obtain the estimates, we need to put the model in matrix form.

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ & & \vdots & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

11.1 Inference for Multiple Regression

Parameter estimation

- Then the model can be written in matrix form as

  $$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}.$$

11.1 Inference for Multiple Regression

Parameter estimation

- Then the model can be written in matrix form as
  $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$.
- The method of least squares chooses the values $b_0, b_1, b_2, \ldots, b_p$ which minimize
  $$\sum_{i=1}^{n}(y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \ldots - b_px_{ip})^2 = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|^2$$
  where $\boldsymbol{b}=\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix}$.

11.1 Inference for Multiple Regression

Parameter estimation

- Matrix algebra can be used to show that the minimizer is

$$\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

11.1 Inference for Multiple Regression

Parameter estimation

- Matrix algebra can be used to show that the minimizer is

$$\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

- The predicted response when $x_1 = x_{i1}$, $x_2 = x_{i2}$, $\ldots$, $x_p = x_{ip}$ is $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$.

11.1 Inference for Multiple Regression

Parameter estimation

- Matrix algebra can be used to show that the minimizer is

$$\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

- The predicted response when $x_1 = x_{i1}$, $x_2 = x_{i2}$, …, $x_p = x_{ip}$ is $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$.
- The residual for the ith case is $e_i = y_i - \hat{y}_i$.

11.1 Inference for Multiple Regression

Parameter estimation

- Matrix algebra can be used to show that the minimizer is

$$\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

- The predicted response when $x_1 = x_{i1}$, $x_2 = x_{i2}$, ..., $x_p = x_{ip}$ is $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$.
- The residual for the ith case is $e_i = y_i - \hat{y}_i$.
- The estimate of σ^2 for the multiple regression model is

$$s^2 = \frac{\sum e_i^2}{n-p-1} = \frac{\sum (y_i - \hat{y}_i)^2}{n-p-1}.$$

Then $s = \sqrt{s^2}$ estimates σ.

11.1 Inference for Multiple Regression

Parameter estimation

- Matrix algebra can be used to show that the minimizer is

$$\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

- The predicted response when $x_1 = x_{i1}$, $x_2 = x_{i2}$, ..., $x_p = x_{ip}$ is $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$.
- The residual for the ith case is $e_i = y_i - \hat{y}_i$.
- The estimate of σ^2 for the multiple regression model is

$$s^2 = \frac{\sum e_i^2}{n-p-1} = \frac{\sum (y_i - \hat{y}_i)^2}{n-p-1}.$$

  Then $s = \sqrt{s^2}$ estimates σ.
- The quantity $n - p - 1$ is the degrees of freedom for a multiple regression model with p explanatory variables.

11.1 Inference for Multiple Regression

Standard errors for parameter estimates

- Computer software typically includes built-in functions which report the parameter estimates as well as standard errors for these estimates.

11.1 Inference for Multiple Regression

Standard errors for parameter estimates

- Computer software typically includes built-in functions which report the parameter estimates as well as standard errors for these estimates.
- Formulas for standard errors of the estimates are not discussed in the book, but we will discuss them in these slides.

11.1 Inference for Multiple Regression

Standard errors for parameter estimates

- Computer software typically includes built-in functions which report the parameter estimates as well as standard errors for these estimates.
- Formulas for standard errors of the estimates are not discussed in the book, but we will discuss them in these slides.
- If the ε's are normally distributed, then b_i follows a Normal distribution with mean β_i and standard error SE_i which is the square root of the $(i+1)$th diagonal element of $\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.

11.1 Inference for Multiple Regression

Standard errors for parameter estimates

- Computer software typically includes built-in functions which report the parameter estimates as well as standard errors for these estimates.
- Formulas for standard errors of the estimates are not discussed in the book, but we will discuss them in these slides.
- If the ε's are normally distributed, then b_i follows a Normal distribution with mean β_i and standard error SE_i which is the square root of the $(i+1)$th diagonal element of $\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$.
- Since σ is unknown, it is replaced by s and $t = \frac{b_i - \beta_i}{SE_i}$ follows a t-distribution with $n-p-1$ degrees of freedom.

11.1 Inference for Multiple Regression

Confidence interval for β_i in the multiple linear regression model

A level C confidence interval for β_i is

$$b_i \pm t^* SE_{b_i}$$

where t^* is the value for the $t(n-p-1)$ density curve with area C between $-C$ and C.

11.1 Inference for Multiple Regression

Example 11A: A study at a large university examined GPA (on a 4-point scale) for computer science majors based on high school grades in mathematics (HSM), science (HSS), and english (HSE) (on a 10-point scale: A=10, A-=9, B+=8, etc.). The following table gives the least squares estimates and standard errors for the multiple linear regression model based on a sample of 224 students.

Variable	**Estimate**	**SE**
Intercept	0.5899	0.2942
HSM	0.1686	0.0355
HSS	0.0343	0.0376
HSE	0.0451	0.0387

(a) Predict the GPA for a student who got an A in Mathematics, B in Science, and C+ in English.
(b) Give a 95% confidence interval for $\beta_0, \beta_1, \beta_2,$ and β_3.

11.1 Inference for Multiple Regression

Answer:

11.1 Inference for Multiple Regression

Two-sided significance test for the slope in the multiple linear regression model

1. Test $H_0 : \beta_i = 0$ vs. $H_a : \beta_i \neq 0$
2. Test statistic: $t = \dfrac{b_i}{SE_{b_i}}$
3. Use the t-table to find the value t^* such that $P(T > t^*) = \frac{\alpha}{2}$ where T follows a t distribution with $n - p - 1$ degrees of freedom
4. Reject H_0 is $|t| > t^*$

11.1 Inference for Multiple Regression

One-sided (left-sided) significance test for the slope in the multiple linear regression model

1. Test $H_0 : \beta_i = 0$ vs. $H_a : \beta_i < 0$
2. Test statistic: $t = \dfrac{b_i}{SE_{b_i}}$
3. Use the t-table to find the value t^* such that $P(T < -t^*) = P(T > t^*) = \alpha$ where T follows a t distribution with $n - p - 1$ degrees of freedom
4. Reject H_0 is $t < -t^*$

One-sided (right-sided) significance test for the slope in the multiple linear regression model

1. Test $H_0 : \beta_i = 0$ vs. $H_a : \beta_i > 0$
2. Test statistic: $t = \dfrac{b_i}{SE_{b_i}}$
3. Use the t-table to find the value t^* such that $P(T > t^*) = \alpha$ where T follows a t distribution with $n - p - 1$ degrees of freedom
4. Reject H_0 is $t > t^*$

11.1 Inference for Multiple Regression

Example 11B: A study at a large university examined GPA (on a 4-point scale) for computer science majors based on high school grades in mathematics (HSM), science (HSS), and english (HSE) (on a 10-point scale: A=10, A-=9, B+=8, etc.). The following table gives the least squares estimates and standard errors for the multiple linear regression model based on a sample of 224 students.

Variable	**Estimate**	**SE**
Intercept	0.5899	0.2942
HSM	0.1686	0.0355
HSS	0.0343	0.0376
HSE	0.0451	0.0387

Perform a two-sided test of significance for each explanatory variable in the multiple regression model at level .05.

11.1 Inference for Multiple Regression

Answer:

ANOVA F test

- In simple linear regression, the ANOVA F test compares the model with $H_0 : \beta_1 = 0$ against the alternative model with $H_a : \beta_1 \neq 0$.

11.1 Inference for Multiple Regression

ANOVA F test

- In simple linear regression, the ANOVA F test compares the model with $H_0 : \beta_1 = 0$ against the alternative model with $H_a : \beta_1 \neq 0$.
- In multiple linear regression, the ANOVA F test compares the model where all coefficients for the explanatory variables (with the exception of the intercept) are 0 against the alternative that at least one is not 0.

11.1 Inference for Multiple Regression

ANOVA F test

- In simple linear regression, the ANOVA F test compares the model with $H_0 : \beta_1 = 0$ against the alternative model with $H_a : \beta_1 \neq 0$.
- In multiple linear regression, the ANOVA F test compares the model where all coefficients for the explanatory variables (with the exception of the intercept) are 0 against the alternative that at least one is not 0.
- When the null hypothesis is true, the F statistic follows an F distribution with p degrees of freedom in the numerator and $n - p - 1$ degrees of freedom in the denominator.

ANOVA table for multiple regression

- The computations for the F statistic can be summarized in the following table.

Source	**DF**	**SS**	**MS**	**F**
Model	p	$\sum(\hat{y}_i - \bar{y})^2$	SSM/DFM	MSM/MSE
Error	$n - p - 1$	$\sum(y_i - \hat{y}_i)^2$	SSE/DFE	
Total	$n - 1$	$\sum(y_i - \bar{y})^2$	SST/DFT	

11.1 Inference for Multiple Regression

ANOVA table for multiple regression

- The computations for the F statistic can be summarized in the following table.

Source	**DF**	**SS**	**MS**	**F**
Model	p	$\sum(\hat{y}_i - \bar{y})^2$	SSM/DFM	MSM/MSE
Error	$n - p - 1$	$\sum(y_i - \hat{y}_i)^2$	SSE/DFE	
Total	$n - 1$	$\sum(y_i - \bar{y})^2$	SST/DFT	

- The **squared multiple correlation coefficient** $R^2 = \dfrac{SSM}{SST} = \dfrac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$ is the proportion of variation of the response variable that is explained by the explanatory variables $x_1, x_2, \ldots, x_p$ in a multiple linear regression.

11.1 Inference for Multiple Regression

F distribution

- The density curve of the F distribution with $df1$ degrees of freedom in the numerator and $df2$ degrees of freedom in the denominator has the form

$$f(x) = \frac{\Gamma(\frac{df1+df2}{2})}{\Gamma(\frac{df1}{2})\Gamma(\frac{df2}{2})} \left(\frac{df1}{df2}\right)^{\frac{df1}{2}} x^{\frac{df1}{2}-1} \left(1 + \frac{df1}{df2}x\right)^{-\frac{df1+df2}{2}},$$

for $x > 0$.

11.1 Inference for Multiple Regression

F distribution

- The density curve of the F distribution with $df1$ degrees of freedom in the numerator and $df2$ degrees of freedom in the denominator has the form

$$f(x) = \frac{\Gamma(\frac{df1+df2}{2})}{\Gamma(\frac{df1}{2})\Gamma(\frac{df2}{2})} \left(\frac{df1}{df2}\right)^{\frac{df1}{2}} x^{\frac{df1}{2}-1} \left(1 + \frac{df1}{df2}x\right)^{-\frac{df1+df2}{2}},$$

  for $x > 0$.
- This distribution is skewed to the right.

11.1 Inference for Multiple Regression

<u>F distribution</u>

- The density curve of the F distribution with $df1$ degrees of freedom in the numerator and $df2$ degrees of freedom in the denominator has the form

$$f(x) = \frac{\Gamma(\frac{df1+df2}{2})}{\Gamma(\frac{df1}{2})\Gamma(\frac{df2}{2})} \left(\frac{df1}{df2}\right)^{\frac{df1}{2}} x^{\frac{df1}{2}-1} \left(1 + \frac{df1}{df2}x\right)^{-\frac{df1+df2}{2}},$$

  for $x > 0$.
- This distribution is skewed to the right.
- Sometimes, we abbreviate the F distribution with $df1$ and $df2$ degrees of freedom by $F(df1, df2)$.

11.1 Inference for Multiple Regression

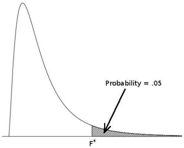

Table for F distribution

F distribution critical values

	df1 = degrees of freedom in the numerator							
		1	2	3	⋯	40	50	100
	1	161.45	199.50	215.71	⋯	251.14	251.77	253.04
df2 =	2	18.51	19.00	19.16	⋯	19.47	19.48	19.49
degrees	3	10.13	9.55	9.28	⋯	8.59	8.58	8.55
					⋮			
of freedom								
in the	50	4.03	3.18	2.79	⋯	1.63	1.60	1.52
denominator	100	3.94	3.09	2.70	⋯	1.52	1.48	1.39
	1000	3.85	3.00	2.61	⋯	1.41	1.36	1.26

11.1 Inference for Multiple Regression

ANOVA F test in the multiple linear regression model

1. Test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ vs.
   H_a : at least one β_i is not 0
2. Test statistic: $F = \dfrac{\text{MSM}}{\text{MSE}}$
3. Use the F-table to find the value f^* such that $P(\mathcal{F} > f^*) = \alpha$ where $\mathcal{F}$ follows a F distribution with p degrees of freedom in the numerator and $n - p - 1$ degrees of freedom in the denominator
4. Reject H_0 if $F > f^*$

Example 11C: A study at a large university examined GPA (on a 4-point scale) for computer science majors based on high school grades in mathematics (HSM), science (HSS), and english (HSE) (on a 10-point scale: A=10, A-=9, B+=8, etc.). Use the following information based on a sample of 224 students to test the null hypothesis that the coefficients of HSM, HSS, and HSE are 0 versus the alternative that at least one of these coefficients is not 0 at level .05.

Source	**Sum of Squares**
Model	27.712
Error	107.750
Total	135.462

11.1 Inference for Multiple Regression

Answer: