Multiple Regression

Inference for Multiple Regression and A Case Study

IPS Chapters 11.1 and 11.2

Objectives (IPS Chapters 11.1 and 11.2)

Multiple regression

- Data for multiple regression
- Multiple linear regression model
- Estimation of the parameters
- Confidence interval for β_j
- Significance test for β_j
- ANOVA table for multiple regression
- Squared multiple correlation R^2

Data for multiple regression

- Up to this point we have considered, in detail, the linear regression model in one explanatory variable x.

$$\hat{y} = b_0 + b_1 x$$

- Usually more complex linear models are needed in practical situations.

- There are many problems in which a knowledge of more than one explanatory variable is necessary in order to obtain a better understanding and better prediction of a particular response.

- In general, we have data on *n* cases and *p* explanatory variables.

Multiple linear regression model

For "*p*" number of explanatory variables, we can express the population mean response (μ_y) as a linear equation:

$$\mu_y = \beta_0 + \beta_1 x_1 \ldots + \beta_p x_p$$

The statistical model for *n* sample data (i = 1, 2, … n) is then:

Data =	fit	+	residual
y_i =	$(\beta_0 + \beta_1 x_{1i} \ \ldots + \beta_p x_{pi})$	+	(ε_i)

Where the ε_i are independent and normally distributed $N(0, \sigma)$.

Multiple linear regression assumes equal variance σ^2 of y. The parameters of the model are $\beta_0, \beta_1 \ldots \beta_p$.

Estimation of the parameters

We selected a random sample of n individuals for which $p + 1$ variables were measured ($x_1 \ldots , x_p, y$). The least-squares regression method minimizes the sum of squared deviations e_i (= $y_i - \hat{y}_i$) to express y as a linear function of the p explanatory variables:

$$\hat{y}_i = b_0 + b_1 x_{1i} \ldots + b_k x_{pi}$$

As with simple linear regression, the constant b_0 is the y intercept.

- The regression coefficients ($b_1 - b_p$) reflect the unique association of each independent variable with the y variable. They are analogous to the slope in simple regression.

$$\left.\begin{array}{l} \hat{y} \\ b_0 \\ b_p \end{array}\right\} \text{ are unbiased estimates of population parameters } \left\{\begin{array}{l} \mu_y \\ \beta_0 \\ \beta_p \end{array}\right.$$

Confidence interval for β_j

Estimating the regression parameters $\beta_0, \ldots \beta_j \ldots \beta_p$ is a case of one-sample inference with unknown population variance.

➔ We rely on the *t* distribution, with ***n – p – 1* degrees of freedom.**

A level *C* **confidence interval for β_j** is:

$$\boxed{b_j \pm t^* \, SE_{bj}}$$

- SE_{bj} is the standard error of b_j—we rely on software to obtain SE_{bj} .

- t is the t critical for the t (n – 2) distribution with area C between –t* and +t*.*

Significance test for β_j

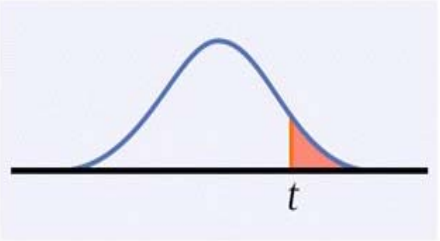
To test the hypothesis H_0: $\beta_j = 0$ versus a 1 or 2 sided alternative.

We calculate the *t* statistic

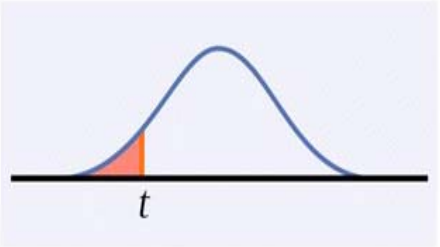
$$\boxed{t = b_j / SE_{bj}}$$

which has the ***t* (*n* – *p* – 1) distribution** to find the p-value of the test.

H_a: $\beta_j > 0$ is $P(T \geq t)$

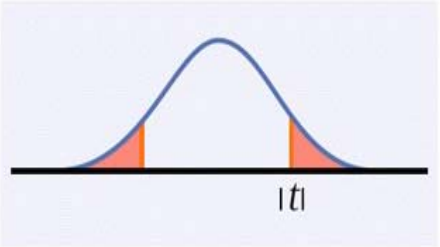


H_a: $\beta_j < 0$ is $P(T \leq t)$



<u>Note</u>: Software typically provides two-sided p-values.

H_a: $\beta_j \neq 0$ is $2P(T \geq |t|)$



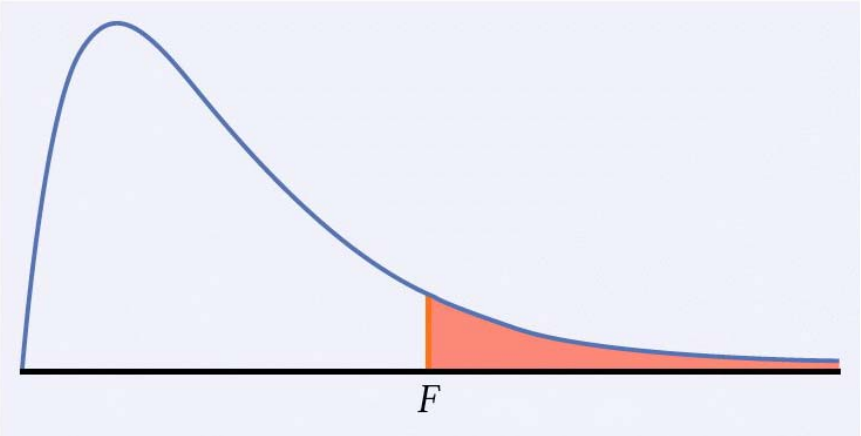
ANOVA F-test for multiple regression

For a multiple linear relationship the ANOVA tests the hypotheses

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad \text{versus } H_a: H_0 \text{ not true}$$

by computing the F statistic: F = MSM / MSE

When H_0 is true, F follows the $F(1, n - p - 1)$ distribution. The p-value is $P(F \geq f)$.



A significant p-value doesn't mean that all *p* explanatory variables have a significant influence on *y*—only that at least one does.

ANOVA table for multiple regression

Source	Sum of squares SS	df	Mean square MS	*F*	P-value
Model	$\sum(\hat{y}_i - \bar{y})^2$	*p*	SSM/DFM	MSM/MSE	Tail area above F
Error	$\sum(y_i - \hat{y}_i)^2$	*n* − *p* − *1*	SSE/DFE		
Total	$\sum(y_i - \bar{y})^2$	*n* − 1			

SST = SSM + SSE

DFT = DFM + DFE

The **standard deviation of the sampling distribution, *s*,** for *n* sample data points is calculated from the residuals $e_i = y_i - \hat{y}_i$

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

s is an unbiased estimate of the regression standard deviation ***σ*.**

Squared multiple correlation R^2

Just as with simple linear regression, ***R^2*, the squared multiple correlation**, is the proportion of the variation in the response variable *y* that is explained by the model.

In the particular case of multiple linear regression, the model is <u>all *p*</u> explanatory variables taken <u>together</u>.

$$\boxed{R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SSModel}{SSTotal}}$$

We have data on 224 first-year computer science majors at a large university in a given year. The data for each student include:

* Cumulative GPA after 2 semesters at the university (*y,* response variable)
* SAT math score (SATM, x1, explanatory variable)
* SAT verbal score (SATV, x2, explanatory variable)
* Average high school grade in math (HSM, x3, explanatory variable)
* Average high school grade in science (HSS, x4, explanatory variable)
* Average high school grade in English (HSE, x5, explanatory variable)

Here are the summary statistics for these data given by software **SAS**:

Variable	N	Mean	Std Dev	Minimum	Maximum
GPA	224	2.6352232	0.7793949	0.1200000	4.0000000
SATM	224	595.2857143	86.4014437	300.0000000	800.0000000
SATV	224	504.5491071	92.6104591	285.0000000	760.0000000
HSM	224	8.3214286	1.6387367	2.0000000	10.0000000
HSS	224	8.0892857	1.6996627	3.0000000	10.0000000
HSE	224	8.0937500	1.5078736	3.0000000	10.0000000

The first step in multiple linear regression is to study all pair-wise relationships between the $p + 1$ variables. Here is the SAS output for all pair-wise correlation analyses (value of r and 2 sided p-value of H_0: $\rho = 0$).

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 224

	GPA	SATM	SATV	HSM	HSS	HSE
GPA	1.00000	0.25171	0.11449	0.43650	0.32943	0.28900
	0.0	0.0001	0.0873	0.0001	0.0001	0.0001
SATM	0.25171	1.00000	0.46394	0.45351	0.24048	0.10828
	0.0001	0.0	0.0001	0.0001	0.0003	0.1060
SATV	0.11449	0.46394	1.00000	0.22112	0.26170	0.24371
	0.0873	0.0001	0.0	0.0009	0.0001	0.0002
HSM	0.43650	0.45351	0.22112	1.00000	0.57569	0.44689
	0.0001	0.0001	0.0009	0.0	0.0001	0.0001
HSS	0.32943	0.24048	0.26170	0.57569	1.00000	0.57937
	0.0001	0.0003	0.0001	0.0001	0.0	0.0001
HSE	0.28900	0.10828	0.24371	0.44689	0.57937	1.00000
	0.0001	0.1060	0.0002	0.0001	0.0001	0.0

Scatterplots for all 15 pair-wise relationships are also necessary to understand the data.

For simplicity, let's first run a multiple linear regression using **only the three high school grade averages:**

Dependent Variable: GPA

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	27.71233	9.23744	18.861	0.0001
Error	220	107.75046	0.48977		
C Total	223	135.46279			

P-value very significant

Root MSE	0.69984	R-Square	0.2046
Dep Mean	2.63522	Adj R-sq	0.1937
C.V.	26.55711		

R^2 is fairly small (20%)

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	0.589877	0.29424324	2.005	0.0462
HSM	1	0.168567	0.03549214	4.749	0.0001
HSS	1	0.034316	0.03755888	0.914	0.3619
HSE	1	0.045102	0.03869585	1.166	0.2451

HSM significant

HSS, HSE not

```
Dependent Variable: GPA
                          Analysis of Variance

                          Sum of          Mean
Source            DF     Squares        Square      F Value      Prob>F

Model              3    27.71233       9.23744       18.861      0.0001
Error            220   107.75046       0.48977
C Total          223   135.46279

    Root MSE        0.69984     R-Square       0.2046
    Dep Mean        2.63522     Adj R-sq       0.1937
    C.V.           26.55711
```

P-value very significant

***R^2* is fairly small (20%)**

The ANOVA for the multiple linear regression using only HSM, HSS, and HSE is very significant ➔ at least one of the regression coefficients is significantly different from zero.

But R^2 is fairly small (0.205) ➔ only about 20% of the variations in cumulative GPA can be explained by these high school scores.

(Remember, a small p-value does not imply a large effect.)

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	0.589877	0.29424324	2.005	0.0462
HSM	1	0.168567	0.03549214	4.749	0.0001
HSS	1	0.034316	0.03755888	0.914	0.3619
HSE	1	0.045102	0.03869585	1.166	0.2451

HSM significant

HSS, HSE not

The tests of hypotheses for each *b* within the multiple linear regression reach significance for HSM only.

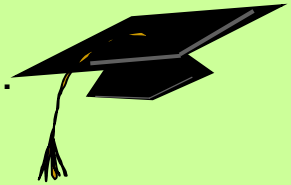
We found a significant correlation between HSS and GPA when analyzed by themselves, so why is b_{HSS} not significant in the multiple regression equation?
Well, HHS and HHM are also significantly correlated.

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 224

	GPA	SATM	SATV	HSM	HSS	HSE
GPA	1.00000 0.0	0.25171 0.0001	0.11449 0.0873	0.43650 0.0001	0.32943 0.0001	0.28900 0.0001
SATM	0.25171 0.0001	1.00000 0.0	0.46394 0.0001	0.45351 0.0001	0.24048 0.0003	0.10828 0.1060
SATV	0.11449 0.0873	0.46394 0.0001	1.00000 0.0	0.22112 0.0009	0.26170 0.0001	0.24371 0.0002
HSM	0.43650 0.0001	0.45351 0.0001	0.22112 0.0009	1.00000 0.0	0.57569 0.0001	0.44689 0.0001
HSS	0.32943 0.0001	0.24048 0.0003	0.26170 0.0001	0.57569 0.0001	1.00000 0.0	0.57937 0.0001
HSE	0.28900 0.0001	0.10828 0.1060	0.24371 0.0002	0.44689 0.0001	0.57937 0.0001	1.00000 0.0

When all three high school averages are used together in the multiple regression analysis, only HSM contributes significantly to our ability to predict GPA.

We now **drop** the least significant variable from the previous model: HSS.



Dependent Variable: GPA

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	27.30349	13.65175	27.894	0.0001
Error	221	108.15930	0.48941		
C Total	223	135.46279			

P-value very significant

Root MSE	0.69958	R-Square	0.2016
Dep Mean	2.63522	Adj R-sq	0.1943
C.V.	26.54718		

R^2 is small (20%)

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	0.624228	0.29172204	2.140	0.0335
HSM	1	0.182654	0.03195581	5.716	0.0001
HSE	1	0.060670	0.03472914	1.747	0.0820

HSM significant

HSE not

The conclusions are about the same. But notice that the actual regression coefficients have changed.

predicted GPA=.590+.169HSM+.045HSE+.034HSS

predicted GPA=.624+.183HSM+.061HSE

Let's run a multiple linear regression with the **two SAT scores only.**

Dependent Variable: GPA

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	8.58384	4.29192	7.476	0.0007
Error	221	126.87895	0.57411		
C Total	223	135.46279			

P-value very significant

Root MSE	0.75770	R-Square	0.0634
Dep Mean	2.63522	Adj R-sq	0.0549
C.V.	28.75287		

R^2 is very small (6%)

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	1.288677	0.37603684	3.427	0.0007
SATM	1	0.002283	0.00066291	3.444	0.0007
SATV	1	-0.000024562	0.00061847	-0.040	0.9684

SATM significant
SATV not

The ANOVA test for β_{SATM} and β_{SATV} is very significant ➔ at least one is not zero.

R^2 is really small (0.06) ➔ only 6% of GPA variations are explained by these tests.

When taken together, only SATM is a significant predictor of GPA (P 0.0007).

We finally run a multiple regression model with **all the variables together**

Dependent Variable: GPA

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	5	28.64364	5.72873	11.691	0.0001
Error	218	106.81914	0.49000		
C Total	223	135.46279			

P-value very significant

Root MSE	0.70000	R-Square	0.2115
Dep Mean	2.63522	Adj R-sq	0.1934
C.V.	26.56311		

R^2 fairly small (21%)

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	0.326719	0.39999643	0.817	0.4149
SATM	1	0.000944	0.00068566	1.376	0.1702
SATV	1	-0.000408	0.00059189	-0.689	0.4915
HSM	1	0.145961	0.03926097	3.718	0.0003
HSS	1	0.035905	0.03779841	0.950	0.3432
HSE	1	0.055293	0.03956869	1.397	0.1637

HSM significant

The overall test is significant, but only the average high school math score (HSM) makes a significant contribution in this model to predicting the cumulative GPA. This conclusion applies to computer majors at this large university.

Regression Statistics	
Multiple R	0.459837234
R Square	0.211450282
Adjusted R Square	0.193364279
Standard Error	0.699997195
Observations	224

ANOVA

	df	SS	MS	F	Signifcance F
Regression	5	28.64364489	5.728729	11.69138	5.06E-10
Residual	218	106.8191439	0.489996		
Total	223	135.4627888			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	upper 95%
intercept	0.326718739	0.399996431	0.816804	0.414932	-0.461636967	1.115074446
HSM	0.14596108	0.039260974	3.717714	0.000256	0.068581358	0.223340801
HSS	0.03590532	0.037798412	0.949916	0.343207	-0.03859183	0.11040247
HSE	0.055292581	0.039568691	1.397382	0.163719	-0.022693622	0.133278785
SATM	0.000843593	0.000685657	1.376187	0.170176	-0.000407774	0.002294959
SATV	0.00040785	0.000591893	-0.68906	0.491518	-0.001574415	0.00075816

Excel

Minitab

```
The regression equation is
GPA = 0.327 + 0.146 HSM + 0.0359 HSS + 0.0553 HSE +0.000944 SATM -
0.000408 SATV

Predictor       Coef       StDev         T        P
Constant      0.3267      0.4000      0.82    0.415
HSM          0.14596     0.03926      3.72    0.000
HSS          0.03591     0.03780      0.95    0.343
HSE          0.05529     0.03957      1.40    0.164
SATM       0.0009436   0.0006857      1.38    0.170
SATV      -0.0004078   0.0005919     -0.69    0.492

S = 0.7000      R-Sq = 21.1%     R-Sq(adj)= 19.3%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    5     28.6436      5.7287     11.69    0.000
Error       218    106.8191      0.4900
Total       223    135.4628
```