Multiple Regression Analysis

Recall: In simple linear regression models our objective was to relate one quantitative predictor variable to a quantitative response variable. For example, the high school GPA of a college student may be used to predict that student's success in college. Of course there are many other variables that could also be used to predict a student's college success (e.g., gender, major, financial support etc.)

MULTIPLE REGRESSION ANALYSIS describes models which use two or more predictors (usually quantitative but possibly categorical) to describe the behavior of a quantitative response.

DEFINITION: A GENERAL ADDITIVE MULTIPLE REGRESSION MODEL equation is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

where the ϵ are assumed to be independent, normally distributed,with mean zero and common variance σ^2.

Similarly to the one-dimensional case of simple linear regression we will use the following notation:

- $\beta_0, \beta_1, \ldots, \beta_k$ are the (usually unknown) true population parameters;
- $b_0, b_1, \ldots, b_k$ or $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are their estimates;
- $\mu_{Y|x_1,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k$ is the mean of future observations Y if the predictors are fixed at levels $x_1, \ldots, x_k$, respectively;
- β_k can be interpreted as the average increase in the response y, if the k^{th} predictor x_k is increased by one unit and all other predictors are held fixed.

Categorical Predictors

Suppose you want to model college success (in college GPA) as a function of a student's previous success in high-school (measured by high-school GPA) and gender (male, female). The predictor high-school GPA is a quantitative variable, but gender is categorical. To include categorical variables into regression models we will use a trick:

Create a numerical DUMMY or INDICATOR variable and code it to represent the categories that appear in your data.

Example: Consider

$$Y = \text{college GPA}, \quad X_1 = \text{high school GPA}, \quad X_2 = \begin{cases} 0 & \text{if the student is male} \\ 1 & \text{if the student is female} \end{cases}$$

Fit a multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Interpret the meaning of the coefficients $\beta_0, \beta_1, \beta_2$.

What do you do if your categorical variable has more than two categories (male/female)? For example, let's assume a students major falls into exactly one of the groups: Computer Science, Economics, Nursing. Would it be ok to define a dummy variable

$$X_3 = \begin{cases} 0 & \text{Computer Science} \\ 1 & \text{Economics} \\ 2 & \text{Nursing} \end{cases}$$

Why or why not?

A Better Solution: Define two dummy variables

$$X_3 = \begin{cases} 1 & \text{if the major is CS} \\ 0 & \text{otherwise} \end{cases}, \qquad X_4 = \begin{cases} 1 & \text{if the major is Econ} \\ 0 & \text{otherwise} \end{cases}$$

In this case, how could the coefficients in the multiple regression model be interpreted?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

Parameter Estimates

Statistical computer programs (such as SPSS) will compute estimates of the model parameters $\beta_0, \beta_1, \ldots, \beta_k$ and σ^2 for you based again on the principle of least squares:

$$\text{SSE} = \sum \epsilon_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - (b_0 + b_1 x_{1i} + \cdots + b_k x_{ki}))^2$$

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n - (k+1)} = \text{MSE}$$

$R^2 = 1$ - SSE/SST is the COEFFICIENT OF MULTIPLE DETERMINATION and can be interpreted as the percentage of variation in the response explained by the regression model based on the k predictor variables. As in polynomial regression, R^2 is sometimes adjusted for the number of parameters in the model:

$$\text{adjusted } R^2 = \frac{(n-1)R^2 - k}{n - (k+1)}$$

The Model Utility Test

Before interpreting a multiple regression model, we should make sure that the response really depends on (at least one of) the predictors:

Null hypothesis: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$
Alternative hypothesis: H_a : at least one $\beta_j \neq 0$.
Test statistic: $f = \frac{R^2/k}{(1-R^2)/(n-(k+1))} = \frac{\text{MSR}}{\text{MSE}} \sim F_k, n - (k+1)$.

Inference in Multiple Regression

Especially in multiple regression models where the number of predictors is large, you should carefully check (for each single predictor) whether they really do have a significant influence on the response.

Example: To predict a student's college success we could use not only high-school GPA, gender, and major but also age, family status, hair color, number of piercings, etc. Some of these clearly will be more important predictors than others.

Completely analogously to the simple linear regression scenario we can perform hypothesis tests

$$H_0 : \beta_j = 0, \qquad \text{Test statistic:} \quad T = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}$$

and compute confidence intervals

$$\text{CI } \left[\hat{\beta}_j \pm t_{\alpha/2,n-(k+1)} \cdot se(\hat{\beta}_j)\right].$$

We can also formulate confidence intervals for the response mean

$$\text{CI for } \mu_{Y|x_1,\ldots,x_k} : \left[\hat{Y} \pm t_{\alpha/2,n-(k+1)} s_{\hat{Y}}\right]$$

and prediction intervals for a future observation Y

$$\text{PI for } Y: \ \left[\hat{Y} \pm t_{\alpha/2,n-(k+1)} \cdot \sqrt{s^2 + s^2_{\hat{Y}}}\right].$$

Test for a Group of Predictors

Suppose you have a large number of possible predictors (e.g., high-school GPA, gender, age, hair color, etc.) of which you think that only a subset will be relevant to predict values of a quantitative response (college GPA). To perform a statistical hypothesis test order the predictors in order of relevance

$$x_1, \ldots, x_l \text{ (relevant)}, \qquad x_{l+1}, \ldots, x_k \text{ (not relevant?)}$$

We want to test the null hypothesis

$$H_0 : \beta_{l+1} = \cdots = \beta_k = 0 \text{ (the "reduced" model is correct)}$$

versus

$$H_a : \text{ at least one } \beta_j \neq 0 \quad (j = l+1, \ldots, k)$$

Let SSE_k be the unexplained variation in the full model and SSE_l the unexplained variation in the reduced model. Then the test statistic becomes

$$f = \frac{(\text{SSE}_l - \text{SSE}_k)/(k-l)}{\text{SSE}_k/(n-(k+1))} \sim F_{k-l,n-(k+1)}$$

Diagnostic Plots

The bare minimum of diagnostic plots that should be considered in a multiple regression model are:

- A normal probability plot for the residuals ϵ_i.
- A residual plot of ϵ_i against the predicted response $\hat{y}_i$.

  If this plot shows any dependence of ϵ on $\hat{y}$, then it may also be necessary to plot the residuals ϵ against the individual predictors x_j.
- Scatter plots of the response against each individual predictor may help to notice unusual patterns or outliers.

Variable Selection

If a large number of possible predictor variables is available to predict one quantitative response it may become necessary to select a meaningful subset of predictors. There are two reasons for not using all available predictors: Simpler models (with fewer predictors) are usually preferable over more complicated ones. Some predictors may not be statistically significant for the response.

A "good" subset of predictors should satisfy the following criteria:

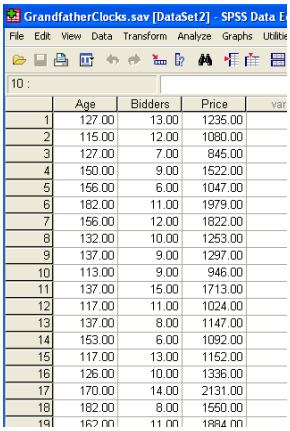
- High adjusted R^2.
- Significant predictors (small p-values for predictors which remain in the model).
- Assumptions on ϵ reasonably satisfied.

If the number m of possible predictors is so large that it becomes unpractical to investigate all 2^m possible subsets of predictors individually, one of two automatic variable selection methods is often used:

Backward Variable Selection: Start with the model that contains all predictors. Fit the multiple regression model. Look at the p-values of all predictors. If the largest p-value is larger than a threshold value (say 0.05) exclude the predictor with the largest p-value from the model. Fit a model with the remaining $m-1$ predictors and repeat...

Forward Variable Selection: Fit simple linear regression models for each of the m possible predictors: Y against X_1, Y against X_2, etc. Determine which predictor variable has the smallest p-value for the $H_0 : \beta_k = 0$ test (provided that the p-value is smaller than a threshold, say 0.05). Include this variable into your model. Now, fit models with two variables - the one selected in the first step and one of the remaining ones. Again, select the variable with the next-smallest p-value (provided it is less than the threshold). Continue until no additional variables with small p-values can be included.

Example: The price (in $) of antique Grandfather clocks at auction is to be predicted by the age of the clock (in years) and the number of bidders. The data entered into SPSS looks like this:


GrandfatherClocks.sav [DataSet2] - SPSS Data E

File Edit View Data Transform Analyze Graphs Utilitie

10 :

	Age	Bidders	Price	var
1	127.00	13.00	1235.00	
2	115.00	12.00	1080.00	
3	127.00	7.00	845.00	
4	150.00	9.00	1522.00	
5	156.00	6.00	1047.00	
6	182.00	11.00	1979.00	
7	156.00	12.00	1822.00	
8	132.00	10.00	1253.00	
9	137.00	9.00	1297.00	
10	113.00	9.00	946.00	
11	137.00	15.00	1713.00	
12	117.00	11.00	1024.00	
13	137.00	8.00	1147.00	
14	153.00	6.00	1092.00	
15	117.00	13.00	1152.00	
16	126.00	10.00	1336.00	
17	170.00	14.00	2131.00	
18	182.00	8.00	1550.00	
19	162.00	11.00	1884.00	

(a) Look at scatter plots of AGE and BIDDERS with PRICE. The correlation is moderate ($\rho = 0.395$ Bidders and Price) to strong ($\rho = 0.730$ Age and Price).

(b) Read off the results of the Model Utility test from the ANOVA table produced by SPSS? Is regression of Price on AGE and BIDDERS meaningful?

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2554859	1	2554859.011	34.273	.000[a]
	Residual	2236335	30	74544.507		
	Total	4791194	31			
2	Regression	4277160	2	2138579.852	120.651	.000[b]
	Residual	514034.5	29	17725.328		
	Total	4791194	31			

a. Predictors: (Constant), Age

b. Predictors: (Constant), Age, Bidders

c. Dependent Variable: Price

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Age	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
2	Bidders	.	Forward (Criterion: Probability-of-F-to-enter <= .050)

a. Dependent Variable: Price

(c) Which variable selection method was used in this problem?

(d) Read off the final model from the SPSS output:

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-191.658	263.887		-.726	.473
	Age	10.479	1.790	.730	5.854	.000
2	(Constant)	-1336.722	173.356		-7.711	.000
	Age	12.736	.902	.888	14.114	.000
	Bidders	85.815	8.706	.620	9.857	.000

a. Dependent Variable: Price

Multicollinearity

Consider the usual multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots \beta_k X_k + \epsilon$$

In many situations, the predictors $X_1, \ldots, X_k$ are interdependent (correlated). If the correlation between two variables is particularly high, then it may not be necessary to include both variables in the model.

It can be shown that

$$V(\hat{\beta}_i) = \frac{\sigma^2}{\sum_{j=1}^{n} (x_{ij} - \hat{x}_{ij})^2}$$

When the predictor x_i values can be predicted very well from the other predictor values, the denominator will be very small. Consequently, the variance of the model parameter estimate $\hat{\beta}_j$ will be very large. If this is the case for at least one predictor, the data is said to exhibit MULTICOLLINEARITY.

An indication for multicollinearity is computer output in which R^2 is large, but the t-test statistics for predictors that (intuitively) seem to be important are small. Consequently, the p-values for those predictors are large and the predictors may (erroneously!) be excluded from the model.

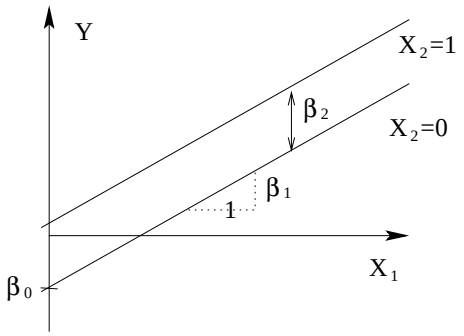
Interaction in Multiple Regression

Situation: Consider a multiple linear regression model with a quantitative response Y, one quantitative predictor X_1 and one categorical predictor X_2 with two levels (coded as 0 and 1). The additive multiple regression model you have seen so far

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

assumes that the two regression lines belonging to the groups specified by the categorical predictor of the response Y against X_1 are parallel.

In this case, β_0 is the y-intercept of the $(x_2 = 0)$-line, and β_1 is the common x_1-slope of both lines. β_2 is the vertical distance between the two regression lines.

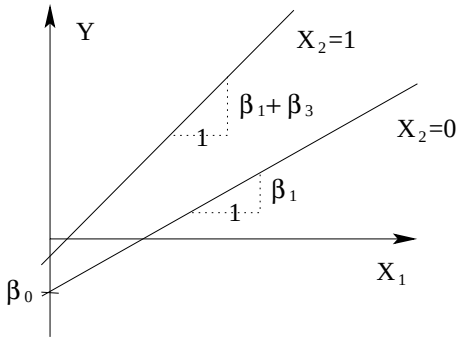


If you want to allow the Y vs. X_1 regression slopes of the two groups specified by the levels of the categorical predictor to be different, an additional interaction term can be included in the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

$$\begin{aligned} X_2 = 0: &\qquad Y = \beta_0 + \beta_1 X_1 + \epsilon \\ X_2 = 1: &\qquad Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + \epsilon \end{aligned}$$

Notice, that $X_1 X_2$ is equal to X_1 if $X_2 = 1$ and 0 otherwise. In this model, β_0 is again the intercept of the $(x_2 = 0)$-line, the intercept of the $(x_2 = 1)$-line is $\beta_0 + \beta_2$. The regression slope of Y vs. X_1 is now β_1 if $X_2 = 0$ and $\beta_1 + \beta_3$ if $X_2 = 1$.

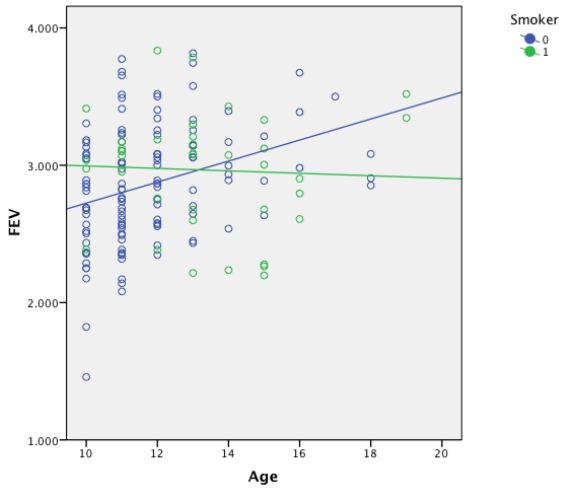


Note: The idea of interaction terms in a multiple linear regression model can be extended to the case of a categorical predictor with more than two levels (include a new slope for each dummy variable) and to interactions of two quantitative variables (compute product of the variables) to obtain a quadratic model. In the quadratic model the slope of one predictor is a linear function of the value of the other predictor.

Example: In a study on respiratory disease the forced expratory volume (FEV) was measured for 163 girls aged 10-19. FEV is an index for pulmonary function that measures the volume of air expelled after one second of constant effort. The children were further classified to their smoking status (0 = nonsmoker, 1 = current smoker) and their age (in years).

(a) To fit a regression model with interaction in SPSS, compute a new variable (TRANSFORM → COMPUTE VARIABLE) that is the product of AGE and SMOKING.

(b) Look at the scatterplot. Phrase several meaningful questions to ask of this regression model.



(c) Answer your questions based on the regression output produced by SPSS. Interpret the slopes from the coefficients table.

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1.958	.245		7.979	.000
	Age	.077	.021	.360	3.714	.000
	Smoker	1.126	.469	1.127	2.401	.018
	Interaction	-.086	.036	-1.155	-2.363	.019

a. Dependent Variable: FEV