TMS-061: Lecture 8
Multiple Regression

Sergei Zuyev

Multiple Regression

Frequently an observed response is influenced by several independent quantitative predictors. The methods by which a straight line can be fitted for a response influenced by one predictor can be extended to several predictors using least squares to fit a model of the form

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon. \qquad (1)$$

The coefficients are estimated by least squares in a closely similar way to that used for straight line regression. The basic assumption is that the error term ε is normally distributed with mean 0 and unknown variance σ^2 that is independent of $X_1, \ldots, X_p$.

Diagnostic Plots for Regression

Before we accept the output from a regression analysis, it is essential that we check whether the assumptions made in the analysis are reasonable for our data set or are clearly violated. Three points that should always be addressed are:

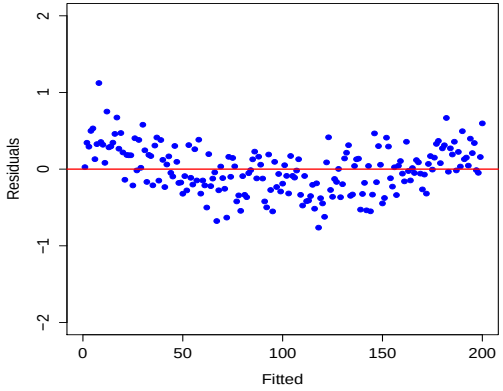
- Is the mean of the residuals always approximately zero for every value of X?
- Is the variance of the residuals approximately constant for all values of X?
- Is the distribution of the residuals approximately Normal?

All three questions will have the answer “yes” if our model is satisfactory.

To address the first two we can save the residuals and the fitted values within Matlab, and examine a scatter-plot of the residuals against the fits with the horizontal zero line drawn in.

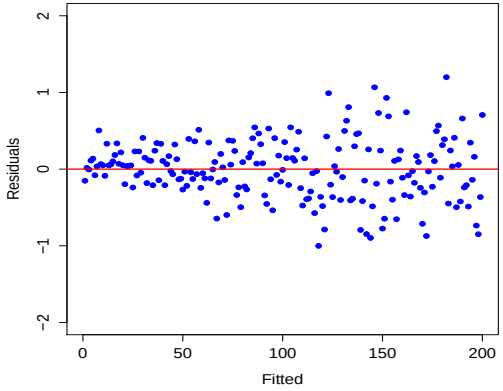
Consider the three figures below. In the first two cases the standard linear model would be unsatisfactory.

Residuals against Fits



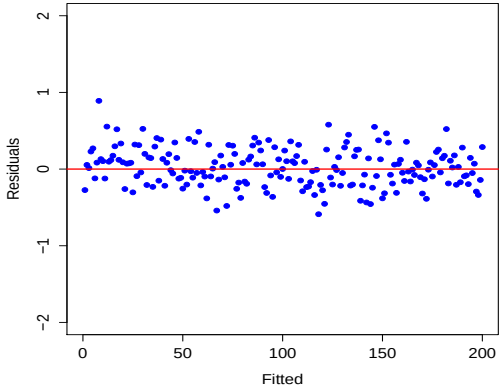
This plot shows evidence of curvature in the scatter-plot.

Residuals against Fits



This plot shows evidence of non-constant variance.

Residuals against Fits

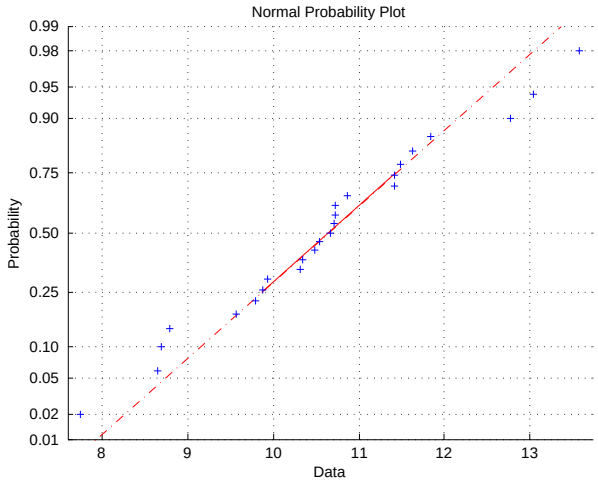


This plot appears to be satisfactory on both these counts.

Checking Normality

To check the normality of the distribution of the residuals the simplest way with Matlab available is construct a Normal Probability plot (`normplot` procedure).

If the distribution is Normal, the points in the plot will lie approximately on a straight line.



Working example

Data are given about 25 patients suffering from cystic fibrosis. The variables measured were:

Subject	Reference number of subject
Sex	0 = male, 1 = female
BMP	Body mass (weight/height^2) as a percentage of the age-specific median in normal individuals
FEV1	Forced expiratory volume in 1 second
RV	Residual volume
FRC	Functional residual capacity
TLC	Total lung capacity
PEmax	Maximal static expiratory pressure (cm H_2O)

The response variable, PEmax, is taken as a measure of malnutrition in these patients. The other variables are measures of lung function. We shall omit categorical variable Sex at present.

Here are the data recorded (available also from VLE):

Age	Sex	Height	Weight	BMP	FEV1	RV	FRC	TLC	PEMax
7	0	109	13.1	68	32	258	183	137	95
7	1	112	12.9	65	19	449	245	134	85
8	0	124	14.1	64	22	441	268	147	100
8	1	125	16.2	67	41	234	146	124	85
8	0	127	21.5	93	52	202	131	104	95
9	0	130	17.5	68	44	308	155	118	80
11	1	139	30.7	89	28	305	179	119	65
12	1	150	28.4	69	18	369	198	103	110
12	0	146	25.1	67	24	312	194	128	70
13	1	155	31.5	68	23	413	225	136	95
13	0	156	39.9	89	39	206	142	95	110
14	1	153	42.1	90	26	253	191	121	90
14	0	160	45.6	93	45	174	139	108	100
15	1	158	51.2	93	45	158	124	90	80
16	1	160	35.9	66	31	302	133	101	134
17	1	153	34.8	70	29	204	118	120	134
17	0	174	44.7	70	49	187	104	103	165
17	1	176	60.1	92	29	188	129	130	120
17	0	171	42.6	69	38	172	130	103	130
19	1	156	37.2	72	21	216	119	81	85
19	0	174	54.6	86	37	184	118	101	85
20	0	178	64.0	86	34	225	148	135	160
23	0	180	73.8	97	57	171	108	98	165
23	0	175	51.1	71	33	224	131	113	95
23	0	179	71.5	95	52	225	127	101	195

Multiple Regression in Matlab

```
S = regstats(PEMax, [Age Height Weight BMP FEV1 RV FRC TLC])
```

calculates the regression equation. `S.tstat.beta` contains the coefficients from which

```
PEMax = 153.039 - 2.115 Age - 0.395 Height + 2.835 Weight
- 1.742 BMP + 1.265 FEV1 + 0.178 RV - 0.248 FRC + 0.208 TLC
```

Value of S.rsquare indicates the regression model was able to explain over 63% of the variation of PEMax.

S.rsquare	S.tstat.beta	S.tstat.pval	S.fstat
0.6359	153.0385	0.4524	sse: 9.7692e+03
	-2.1145	0.6320	dfe: 16
	-0.3948	0.6492	dfr: 8
	2.8349	0.1433	ssr: 1.7063e+04
	-1.7416	0.1397	f: 3.4933
	1.2651	0.1079	pval: 0.0159
	0.1779	0.3226	
	-0.2483	0.5554	
	0.2084	0.6688	

Fix 5% error level. Although the *p*-value for the model as a whole is satisfactory (0.0159), none of the predictor variables is individually significant. This indicates that in fitting this regression we are over-fitting the model.

With only 25 cases (as here) it is not a good idea to use so many predictors. In every case the model that we use for multiple regression should be as *parsimonious* as is consistent with extracting as much information as is practicable from the predictors.

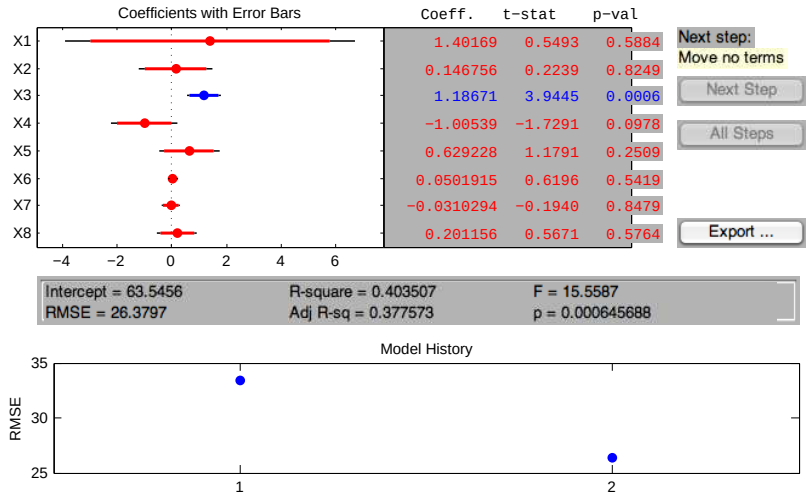
As a rule of thumb, we should hardly ever include more predictors in the model than the square root of the number of cases in the data set. We must therefore start by trying to select the most important predictors for inclusion.

Stepwise Regression

Stepwise regression is to sequentially adds predictors to the model based on their significance until a satisfactory model is found. A measure of significance of the predictors in Matlab is the value of t-statistic of its coefficient.

The more variables we add to the model, the higher is the data fit as expressed by the R^2 value, but we should use other measure which would penalise too many explanatory variables in the model. In Matlab look for RMSE - the square root of the mean variance of residuals which one tries to minimise and R^2 adjusted to the number of degrees of freedom which should be maximised.

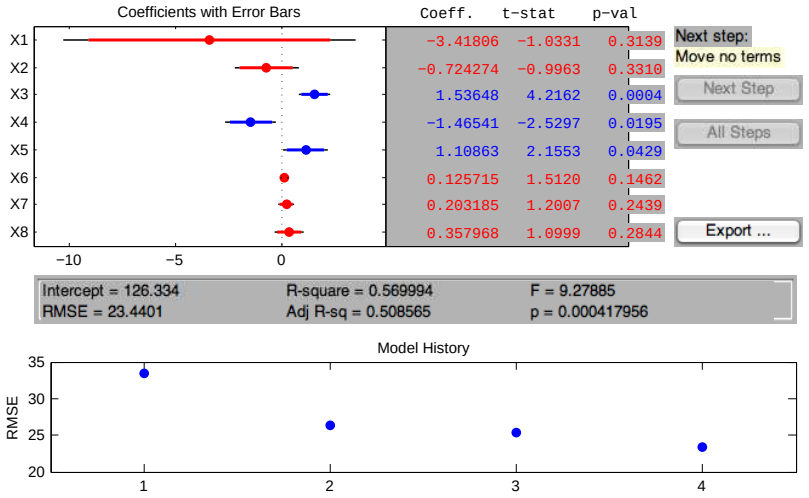

```
stepwise([Age Height Weight BMP FEV1 RV FRC TLC], PEMax)
```



The model suggested by `stepwise` has just one variable: Weight which is highly significant (p-value of 0.0006), but it explains just 40.35% variation in PEMax compared to 64% of the full model. Therefore we proceed with adding variables. The next most significant variable is X4=BMP with p-value 0.0978.

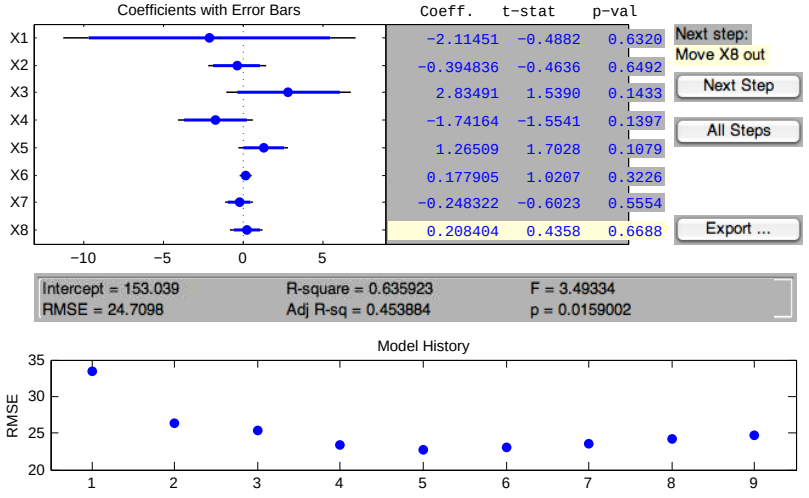
After adding it (click on the corresp. row in the graph) the procedure suggests also to add FEV1 which results in the following model:

PEMax = 126 + 1.54 Weight - 1.47 BMP + 1.11 FEV1



We find now that all predictors contribute significantly to the model, and although R^2 is a little less (57%) than for the full model, R^2-adjusted is greater (0.5086 vs 0.4539), and the overall model has a smaller p-value (0.0004) than the original (0.0159).

We may continue to add variables up to the full model:



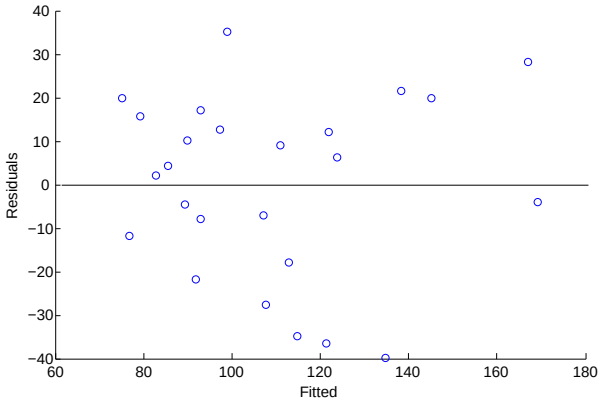
- We see from the bottom frame that the best model in terms of RMSE is actually

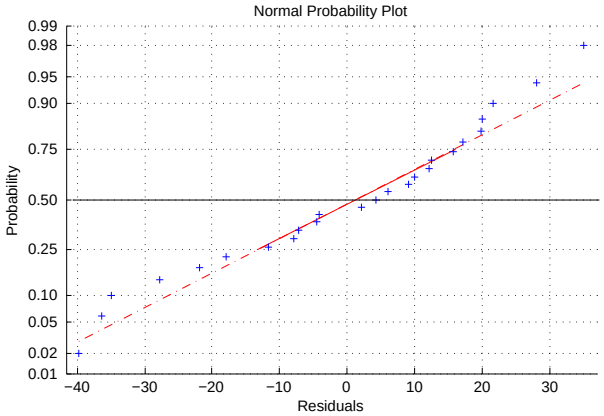
  `PEMax = 63.95 + 1.75 Weight - 1.38 BMP + 1.55 FEV1 + 0.16 RV`

  although the lately added RV is not significant (p-value 0.1462). However, adding it brought R-square from 57% to over 61% which is very close to the maximum of 64% given by the full model, so let's go for it!
- It often appears to the investigator what must be the important predictors to use. In such cases the use of other statistical methods for choosing predictors should be regarded as confirmatory, and if confirmation is not given, then it may be an indication of a need to review the theory underlying the study.

Analysis of the chosen model

Diagnostic plots of the residuals (against fitted values and as a Normal Probability Plot) confirm that the model appears to be satisfactory.





Values Prediction

The model can be used to calculate the fitted value of PEmax for given values of the predictors. Matlab can also calculate a prediction interval for the response at each combination of predictor values. This is no longer so straightforward as for simple linear prediction, so the formula for hand calculation will not be presented. It is however simple to use the quoted standard deviation of each predictor to construct a t-confidence interval for each predictor's coefficient in the model.

Cautionary Note

If a model is *selected* from among many possible models, the fit is quite likely to be better than it should be. This means that if the *same* model is later used with another independent but otherwise identical data set, there may be a much poorer fit. Careful comparison of the fitted models to the two data is then needed.

Inclusion of categorical variables

The variable Sex in the data set is a binary categorical variable. Such a variable can be legitimately included in a multiple regression model. Assuming that the binary variable has no significant interaction with other predictors, we can say that the regression coefficient measures the average shift in the response between the two groups distinguished by the binary variable, after all other effects have been allowed for.

By *interaction* here we mean that the effect of the binary variable does not act independently of the other predictors (for example, there would be an interaction if one of the predictors has an opposite effect on the response for females from what it has for males).

In this case however it turns out that Sex contributes very little to the model, and no sensible model selection scheme would include it.