Regression and Linear Models (37252)

Lecture 5 - Multiple Linear Regression II

Lecturer: Joanna Wang
Notes adopted from Scott Alexander

School of Mathematical and Physical Sciences, UTS

Autumn 2024

Acknowledgement

These notes are based, in part, on earlier versions prepared by Dr Ed Lidums and Prof. James Brown.

5. Lecture outline

Topics:

- model selection
  - problem setup
- school performance example
  - setup
  - selection from all possible models
  - forward selection
  - backward elimination
  - stepwise regression

5. Model selection

We have now been introduced to the basics of linear regression and know enough to

1. propose a model
2. build a model
3. analyse in term of fit and satisfaction of assumptions.

Of these we have had a good look at the second and third on the list.

Our exposure to the first has been limited to examples in multiple regression requiring the selection of independent variables from a limited number of alternatives, and to simple linear regression where the choice was even more straightforward.

In this lecture we look more closely at this and develop tools for model selection in more complicated settings.

For more details see Chapter 15 of Draper and Smith (1998).

5. Model selection

In practical situations, the models we design have to satisfy competing principles.

On the one hand, we want them to be as accurate as possible. On the other, we would like them to be a simple as possible.

The need for accuracy is obvious. The need for simplicity, essentially, relates to cost, be it the cost of equipment, staff or even delay.

Even if simplicity was not a desirable feature, we know from the problem of collinearity that the model with all potential independent variables included may not even be viable, let alone the best amongst alternatives.

5. Model selection

The examples of model selection we have encountered involved only a small number of potential independent variables, allowing the analysis of the alternatives to be conducted in an ad-hoc manner.

However, in practical situations we often have to choose the independent variables from a very large set of potential variables.

In these situations, an ad hoc approach is infeasible because of the sheer number of alternative models to consider.

We need a **system**, a set of rules and instructions that when followed lead to the selection of the **best** model from possible alternatives.

Even better would be a system that can be automated.

5. Model selection - problem setup

Consider a task involving the identification of the best-performing model from those that can be constructed with combinations of independent variables selected from a set of size m.

Let $\boldsymbol{x}$ be the m-dimensional vector of possible independent variables.

The best model could be one of many, ranging from the **null model**

$$\hat{Y}|\boldsymbol{x} = \hat{\beta}_0$$

to the **maximal model**

$$\hat{Y}|\boldsymbol{x} = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_j$$

and every simple and multiple regression model in between.

5. Model selection - problem setup

The problem is that there are 2^m possible combinations of independent variables and therefore 2^m possible regression models to test.

The number of possible models grows **exponentially** as the dimension m increases.

This is an example of the **curse of dimensionality** that renders some numerical tools, proven in low dimensions, infeasible in high dimensions.

Each of the model selection systems we consider has a method of navigating, or **iterating**, through the possible models and for comparing one against another.

5. School performance example - setup

The process of model selection relies to a large extent on theory already developed, so we illustrate the main ideas with an example

Consider the problem of finding the best performing regression model for predicting school-average student test performance.

The response (or dependent variable) is

- mean student test score (Y or *score*)

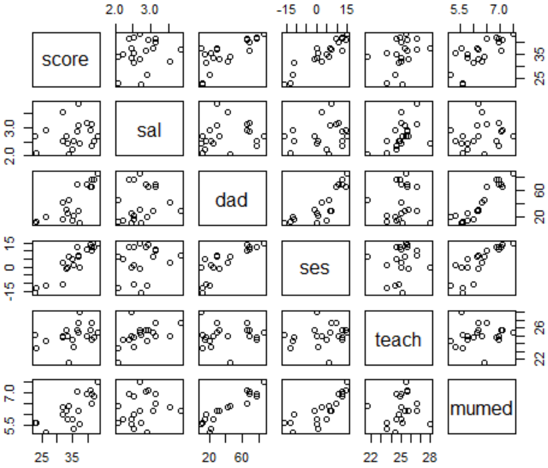
and potential predictors (independent variables)

- staff salaries per pupil (x_1 or *sal*)
- % white-collar fathers (x_2 or *dad*)
- socioeconomic status (x_3 or *ses*)
- teachers' mean score (x_4 or *teach*)
- mothers' mean education level (x_5 or *mumed*).

Data is available in verbal.csv on Canvas.

5. School performance example - setup

After collecting the sample data $(X_{i,1}, \ldots, X_{i,5}, Y_i)$, we look to characterise the relationships between the variables, hoping to spot linear relationships as we seek to fit a plane.



5. School performance example - setup

From the scatter plot we spot seemingly strong positive relationships between the sample data dependent variable Y_i ($score_i$) and sample data independent variables $X_{i,2}$ (dad_i), $X_{1,3}$ (ses_i), and $X_{i,5}$ ($mumed$).

This visual analysis is confirmed by correlation analysis of the sample data

$$\text{corr}(Y_i, X_{i,1}) = \text{corr}(score_i, sal_i) = 0.192$$
$$\text{corr}(Y_i, X_{i,2}) = \text{corr}(score_i, dad_i) = 0.753$$
$$\text{corr}(Y_i, X_{i,3}) = \text{corr}(score_i, ses_i) = 0.927$$
$$\text{corr}(Y_i, X_{i,4}) = \text{corr}(score_i, teach_i) = 0.334$$
$$\text{corr}(Y_i, X_{i,5}) = \text{corr}(score_i, mumed_i) = 0.732.$$

Seeing potential for **collinearity** between x_2 (dad) and x_3 (ses), we look at their sample correlation

$$\text{corr}(X_{i,2}, X_{i,3}) = \text{corr}(dad_i, ses_i) = 0.827$$

which although high, may not be high enough to trigger a VIF (**variance inflation factor**) warning.

5. School performance example - setup

In all there are $2^5 = 32$ possible models, ranging from the null model to the maximal model.

We look at four possible selection methods:

- selection from all possible models
- forward selection
- backward elimination
- stepwise regression.

5. School performance example - selection from all

The obvious way to ensure the best performing model from a set is identified is to test them all. Using the strict interpretation of the **selection from all** method, the best performing model is the one with the best **ranking statistic**.

The ultimate barrier to this method is the curse of dimensionality mentioned earlier, but with only 32 models to choose from this is not an issue here.

Even when feasible, this method still produces many models for analysis with no prescriptive method to do so.

If not applied strictly, many statistical properties of each model must be compared which, given the often conflicting information they provide, can lead to selection being performed in an ad hoc or subjective manner.

In this example we use R^2_{adj} and the p-value from the F-test as ranking statistics.

5. School performance example - selection from all

Data from all models (but null) is summarised below.

```
> library('olsrr')
> mod1 <- lm(score ~ ., data = verbal)
> ols_step_all_possible(mod1)
Index N              Predictors   R-Square Adj. R-Square Mallow's Cp
3      1 1                  ses 0.85962771   0.851829254    5.002821
2      2 1                  dad 0.56761278   0.543591266   48.694760
5      3 1                mumed 0.53571268   0.509918939   53.467726
4      4 1                teach 0.11132200   0.061950995  116.966026
1      5 1                  sal 0.03697606  -0.016525270  128.089834
--------------------------------------------------------------------
13     6 2            ses teach 0.88734850   0.874095385    2.855174
14     7 2            ses mumed 0.86180687   0.845548856    6.676771
10     8 2              dad ses 0.86020753   0.843761359    6.916067
7      9 2              sal ses 0.86007615   0.843614519    6.935725
11    10 2            dad teach 0.65497310   0.614381699   37.623710
15    11 2          teach mumed 0.59599405   0.548463939   46.448290
12    12 2            dad mumed 0.57561917   0.525692012   49.496826
6     13 2              sal dad 0.57083444   0.520344375   50.212728
9     14 2            sal mumed 0.53820829   0.483879856   55.094327
8     15 2            sal teach 0.11213031   0.007675058  118.845084
```

5. School performance example - selection from all

```
19     16 3           sal ses teach 0.90071101   0.882094325    2.855845
25     17 3         ses teach mumed 0.88884972   0.868009037    4.630559
22     18 3           dad ses teach 0.88738875   0.866274139    4.849152
20     19 3           sal ses mumed 0.86222800   0.836395753    8.613760
23     20 3           dad ses mumed 0.86216960   0.836326405    8.622498
16     21 3             sal dad ses 0.86067265   0.834548777    8.846475
17     22 3           sal dad teach 0.66622665   0.603644148   37.939928
24     23 3         dad teach mumed 0.65590099   0.591382431   39.484877
21     24 3         sal teach mumed 0.60260264   0.528090632   47.459498
18     25 3           sal dad mumed 0.57813337   0.499033377   51.120646
-------------------------------------------------------------------------
29     26 4     sal ses teach mumed 0.90185831   0.875687195    4.684183
26     27 4       sal dad ses teach 0.90096690   0.874558069    4.817558
30     28 4     dad ses teach mumed 0.89224918   0.863515622    6.121924
27     29 4       sal dad ses mumed 0.86255877   0.825907778   10.564269
28     30 4     sal dad teach mumed 0.66696870   0.578160357   39.828901
-------------------------------------------------------------------------
31     31 5 sal dad ses teach mumed 0.90643105   0.873013562    6.000000
```

5. School performance example - selection from all

Note:

- You can request the coefficients in each model by using `ols_step_all_possible_betas(mod1)`

  .

- We can also use `regsubsets` function in `leaps` package which gives the best subset of each size according to some selection criteria.

5. School performance example - selection from all

Let's see how far we get selecting the best performing model as the regression with the highest R^2_{adj} with p value less than 0.05 (regression is significant).

1st Choice: $R^2_{\text{adj}} = 0.882$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.11951    9.03643   1.341    0.199
sal         -1.73581    1.18290  -1.467    0.162
ses          0.55321    0.04907  11.273 5.06e-09 ***
teach        1.03582    0.40479   2.559    0.021 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 1.997 on 16 degrees of freedom
Multiple R-squared:  0.9007,Adjusted R-squared:  0.8821
F-statistic: 48.38 on 3 and 16 DF,  p-value: 3.011e-08
```

If not we have to decide what to do with x_1 (*sal*), which is insignificant ($p = 0.162$). Should it be removed or does it contribute to the regression assumptions being satisfied?

5. School performance example - selection from all

2nd Choice: $R^2_{\text{adj}} = 0.876$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.52853   12.34376   1.258   0.2276
sal         -1.71421    1.21571  -1.410   0.1789
ses          0.58246    0.08612   6.763 6.38e-06 ***
teach        1.02494    0.41645   2.461   0.0265 *
mumed       -0.52545    1.25481  -0.419   0.6813
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 2.051 on 15 degrees of freedom
Multiple R-squared:  0.9019,Adjusted R-squared:  0.8757
F-statistic: 34.46 on 4 and 15 DF,  p-value: 2.133e-07
```

This model is next in the ranking, but now we have two variables, x_1 (*sal*) and x_5 (*mumed*), that are insignificant. This is hardly an improvement over the previous model, which had only one insignificant variable and a higher R^2_{adj}.

Note that this model is formed by adding a variable to the 1st Choice model.

5. School performance example - selection from all

3rd Choice: $R^2_{\text{adj}} = 0.875$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.535220   9.781866   1.179    0.257
sal         -1.755821   1.224347  -1.434    0.172
ses          0.538487   0.090308   5.963  2.6e-05 ***
teach        1.052506   0.426037   2.470    0.026 *
dad          0.006525   0.033145   0.197    0.847
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 2.06 on 15 degrees of freedom
Multiple R-squared:  0.901,Adjusted R-squared:  0.8746
F-statistic: 34.12 on 4 and 15 DF,  p-value: 2.281e-07
```

This model is third in the ranking, but again we have two variables, this time x_1 (*sal*) and x_2 (*dad*), that are insignificant. This model is inferior to the 1st Choice.

Note that this model is formed by adding a variable to the 1st Choice model.

5. School performance example - selection from all

4th Choice: $R^2_{\text{adj}} = 0.874$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.58268    9.17541   1.589   0.1304
ses          0.54156    0.05004  10.822 4.81e-09 ***
teach        0.74989    0.36664   2.045   0.0566 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 2.064 on 17 degrees of freedom
Multiple R-squared:  0.8873,Adjusted R-squared:  0.8741
F-statistic: 66.95 on 2 and 17 DF,  p-value: 8.705e-09
```

Now looking at the fourth on the list, once more there is an insignificant variable, x_4 (*teach*), albeit with a p-value not too far from the 0.05 significance level. ($p = 0.057$). But if we are going break rules we might as well stay with the 1st Choice model.

Note that this model is formed by removing a variable from the 1st Choice model.

5. School performance example - selection from all

5th Choice: $R^2_{\text{adj}} = 0.852$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.32280    0.52800   63.11  < 2e-16 ***
ses          0.56033    0.05337   10.50  4.2e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 2.239 on 18 degrees of freedom
Multiple R-squared:  0.8596,Adjusted R-squared:  0.8518
F-statistic: 110.2 on 1 and 18 DF,  p-value: 4.199e-09
```

It has taken until fifth in the ranking before finding a model with no insignificant variables, and this one is a simple linear regression.

Note that this model is formed by removing a variable from the 4th Choice model.

5. School performance example - selection from all

What can be learned from attempting to use a measure like R^2_{adj} to rank the performance of all models in this example?

1. Easy to be in a situation requiring some statistics to be either ignored or treated inconsistently.
2. Moving between models in the ranking involved adding or removing independent variables.

The second point hints at a different approach to model selection, one based on choosing a particular model as a starting point (often the null or maximal model) and adding or deleting independent variables until some condition is met.

The final stop in such an iterative process is the model selected as best performing.

5. School performance example - selection from all

One such approach could be to start with the significant simple regression model with the highest R^2_{adj}. At each iteration the variable associated with the largest increase in R^2_{adj} (keeping the regression significant) is added to the model. The process stops the first time R^2_{adj} falls.

The models considered under this scheme are summarised below.

				R_sq	Ad_R_Sq	p-value (F)
ses (0.000)				**0.860**	**0.852**	**0.000**
dad				0.568	0.544	0.000
mumed				0.536	0.510	0.000
teach				0.111	0.062	0.151
sal				0.037		0.417
ses (0.000)	**teach (0.057)**			**0.887**	**0.874**	**0.000**
ses	mumed			0.862	0.846	0.000
ses	sal			0.860	0.844	0.000
ses	dad			0.860	0.844	0.000
ses (0.000)	**teach (0.021)**	**sal (0.162)**		**0.901**	**0.882**	**0.000**
ses	teach	mumed		0.889	0.868	0.000
ses	teach	dad		0.887	0.866	0.000
ses (0.000)	**teach (0.026)**	**sal (0.179)**	**mumed (0.681)**	**0.902**	**0.876**	**0.000**
ses	teach	sal	dad	0.901	0.875	0.000

Note: t-stats in brackets

This method involves selection from a subset of all regressions.

5. School performance example - selection from all

The principle argument used against the selection from all method becomes weaker as processing power increases. However, in activities where processing power is already fully deployed (e.g. quantitative trading), this argument still stands.

Rather than R^2_{adj}, the selection from all method can use Mallows C_p statistic, Akaike Information Criterion (AIC) or Bayseian Information Criterion (BIC) as the ranking statistic.

Don't use the selection from all method with R^2 as the ranking statistic. This results in the maximal model being selected (recall that R^2 increases with m).

5. School performance example - forward selection

The first method is **forward selection**, where a provisionally optimal model $\hat{Y}^*$ is augmented with the most significant additional predictor at each iteration.

It is defined by the following sequential steps:

1. set $\hat{Y}^*$ as null model (no predictors)
2. construct all possible models by adding one predictor to $\hat{Y}^*$
   - if all models are insignificant (F-test) **GO TO** 3
   - if not set $\hat{Y}^*$ as model with most significant new predictor (T-test) and **REPEAT** 2
3. the optimal model is current iteration of $\hat{Y}^*$. **STOP**.

5. School performance example - forward selection

In R, we can use the command
`ols_step_forward_p(mod1, penter = 0.5, details = T)` to perform forward selection. Using this method, the model selected has independent variable x_3 (*ses*) with $R^2_{\text{adj}} = 0.852$, the 5th Choice using selection from all.

```
Final Model Output
------------------

                              Parameter Estimates
-------------------------------------------------------------------------------------
      model       Beta    Std. Error    Std. Beta        t        Sig      lower      upper
-------------------------------------------------------------------------------------
(Intercept)     33.323         0.528                  63.112    0.000     32.214     34.432
        ses      0.560         0.053        0.927     10.499    0.000      0.448      0.672
-------------------------------------------------------------------------------------
```

5. School performance example - backward selection

Another method is **backward selection**, where a provisionally optimal model $\hat{Y}^*$ is contracted by removing the least significant predictor at each iteration.

It is defined by the following sequential steps:

1. set $\hat{Y}^*$ as maximal model (all predictors)
2. identify least significant predictor (T-test) in $\hat{Y}^*$
   - if all are significant **GO TO** 3
   - if not construct model by removing identified predictor, set $\hat{Y}^*$ as this model and **REPEAT** 2
3. the optimal model is current iteration of $\hat{Y}^*$. **STOP**.

5. School performance example - backward selection

In R, we can use the command
`ols_step_backward_p(mod1, prem = 0.1, details=T)` to perform forward selection. Using this method, the model selected has independent variables x_3 (*ses*) and x_4 (*teach*) with $R^2_{\text{adj}} = 0.874$, the 4th Choice using selection from all.

```
Final Model Output
------------------

                              Parameter Estimates
-------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
-------------------------------------------------------------------------------
(Intercept)    14.583         9.175                 1.589    0.130    -4.776    33.941
        ses     0.542         0.050        0.896   10.822    0.000     0.436     0.647
      teach     0.750         0.367        0.169    2.045    0.057    -0.024     1.523
-------------------------------------------------------------------------------
```

5. School performance example - backward selection

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	dad	0.9019	0.8757	4.6842	91.7366	2.0510
2	mumed	0.9007	0.8821	2.8558	89.9690	1.9974
3	sal	0.8873	0.8741	2.8552	90.4943	2.0641

5. School performance example - backward selection

Note that we remove variables if $p > 0.1$ (`prem=0.1`). Setting this to the more conservative $p > 0.050001$ (removal threshold must be set greater than significance level) results in the selection of the model with independent variable x_3 (*ses*) and $R^2_{\text{adj}} = 0.852$, the 5th Choice using selection from all and the same model using forward selection.

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	dad	0.9019	0.8757	4.6842	91.7366	2.0510
2	mumed	0.9007	0.8821	2.8558	89.9690	1.9974
3	sal	0.8873	0.8741	2.8552	90.4943	2.0641
4	teach	0.8596	0.8518	5.0028	92.8943	2.2392

5. School performance example - stepwise regression

The final iterative method is **stepwise regression**.

It is defined by the following sequential steps:

1. set $\hat{Y}^*$ as chosen initial model (often null or maximal)
2. construct all possible models by adding one predictor to $\hat{Y}^*$
   - if all models are insignificant (F-test) **GO TO** 4
   - if not set $\hat{Y}^*$ as model with most significant new predictor (T-test) and **GO TO** 3
3. identify least significant predictor (T-test) in $\hat{Y}^*$
   - if all are significant **GO TO** 4
   - if not construct model by removing identified predictor, set $\hat{Y}^*$ as this model and **GO TO** 2
4. the optimal model is current iteration of $\hat{Y}^*$. **STOP**.

5. School performance example - stepwise regression

In R, we can use the command
`ols_step_both_p(mod1, pent = 0.05, prem = 0.1, details = T))`
to perform stepwise selection.

Using this method, the model selected has independent variable x_3 (*ses*) with $R^2_{\text{adj}} = 0.852$, the 5th Choice using selection from all, also the model selected using forward selection and also the model using backward selection (if default exit threshold is lowered from 0.1).

```
                              Parameter Estimates
-------------------------------------------------------------------------------------
      model        Beta    Std. Error    Std. Beta      t        Sig      lower     upper
-------------------------------------------------------------------------------------
(Intercept)      33.323         0.528                 63.112    0.000    32.214    34.432
        ses       0.560         0.053        0.927    10.499    0.000     0.448     0.672
-------------------------------------------------------------------------------------
```

References I

Draper, N. R. and Smith, H. (1998). *Applied regression analysis*. Wiley-Interscience, Somerset, US.