Model building

(Statistical Modelling I)

Dr Lubna Shaheen

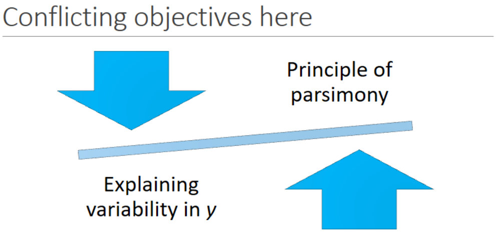
Week 8, Lecture 2

Outline

Model building

In building a multiple regression model we have two objectives which seem to be in conflict:

1. having a model that describes the data as well as possible



2. having a model that is as simple as possible (the principle of parsimony)

We need to select multiple linear regression model that gives a balance between these 2 objectives.

How we will be using the F test to delete variables

1. **Step 1**: Lets say we start with a model of $p-1$ explanatory variables and p parameters.
2. **Step 2**: With an ANOVA table we can carry out a test of the overall model and see that not all of the β_i parameters are zero and hence the multiple linear regression model has some significance and some explanatory power.
3. **Step 3**: But perhaps we could delete some of the explanatory variables to leave a simpler model that still contains explanatory power.
4. **Step 4**: We do this with a Subset test. We are looking to see whether the p parameter model could be reduced to a q parameter model $(q < p)$.
5. **Step 5**: We are looking to see whether we can keep $x_1, x_2, \cdots, x_{q-1}$ but remove $x_q, \cdots, x_{p-1}$.

We call this process a **Subset Test**. We will cover today how to identify that which variable to be consider for deletion.

Extra sum of squares principle

More specifically we are interested in whether these variables under consideration for deletion significantly

1. increase the sum of squares due to regression **or**
2. significantly reduce the sum of squares due to residuals compared with the simpler model that does not include them.
3. The idea is that we seek models that maximise the proportion of sums of squares that are due to regression and minimise the proportion due to residuals.

This idea is referred to as the "Extra sum of squares principle".

Approaches to delete variables

We seek the **extra sum of squares** due to $x_q, \cdots, x_{p-1}$ given that $x_1, \cdots, x_{q-1}$ are already in the model. This can be written

$$\mathbf{SS}(\mathbf{x_q}, \cdots, \mathbf{x_{p-1}} | \mathbf{x_1}, \cdots, \mathbf{x_{q-1}})$$

Extra SS = Regression SS under the full model – Regression SS under the reduced model
and

Extra SS = Residual SS under the reduced model – Residual SS under the full model

Using F tests to delete variables

Extra SS:

If we calculate these sums of squares and call them

SS_R^{Full} and SS_E^{Full} for the full model
SS_R^{Red} and SS_E^{Red} for the reduced model
Then Extra SS is

$$\begin{aligned} SS_{extra} &= SS_R^{Full} - SS_R^{Red} \\ &= SS_E^{Full} - SS_E^{Red} \end{aligned} \tag{1}$$

Using F tests to delete variables

Matrix form for Extra SS:

We can split the parameter vector β into a vector for the reduced model and a second vector of the parameters we are considering deleting

Let $\beta_{\mathbf{1}}^{\mathbf{T}} = (\beta_{\mathbf{0}}, \beta_{\mathbf{1}}, \cdots, \beta_{\mathbf{q-1}})$ and $\beta_{\mathbf{2}}^{\mathbf{T}} = (\beta_{\mathbf{q}}, \cdots, \beta_{\mathbf{p-1}})$

so that

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

Similarly we can split X into

- $\mathbf{X_1}$: a columns of $1's$ and then $q - 1$ columns
- $\mathbf{X_2}$: p-q columns relating to the explanatory variables we may delete

Approaches to delete variables

Full and Reduced Models

The full model is

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\,\beta + \epsilon \\ \mathbf{Y} &= \mathbf{X_1}\,\beta_{\mathbf{1}} + \mathbf{X_2}\,\beta_{\mathbf{2}} + \epsilon \end{aligned} \tag{2}$$

The reduced model is

$$\mathbf{Y} = \mathbf{X_1}\beta_{\mathbf{1}} + \epsilon \tag{3}$$

Approaches to delete variables

Extra SS in Matrix form:

In matrix form

$$\mathbf{SS_{extra}} = \widehat{\beta}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{Y} - \widehat{\beta}^{\mathbf{T}}\mathbf{X}_{\mathbf{1}}^{\mathbf{T}}\mathbf{Y} \tag{4}$$

We now need to test whether the amount the Extra SS is [statistically] significant or not. If it is significant we should keep the full model and not delete variables down to the reduced model.

Approaches to delete variables

Subset test of hypothesis:

Our test is

$$\begin{aligned} H_0 &= \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} \\ H_1 &= \text{ at least one of these parameters is not zero} \end{aligned} \tag{5}$$

Reject $H_0 \rightarrow$ evidence at least some of the variables $x_q, \cdots, x_{p-1}$ are significant and should be included in the model.

Cannot reject $H_0 \rightarrow$ delete the variables $x_q, \cdots, x_{p-1}$ and use reduced model.

Approaches to delete variables

Test Statistics:

$$\mathbf{F}^* = \frac{\frac{\mathbf{SS_{extra}}}{\mathbf{p-q}}}{\mathbf{s^2}}$$

where $s^2 = MS_E$ in the full model

Under H_0 $F^* \sim F^{p-q}_{n-p}$ so we reject H_0 at α significance level if $F^* > F^{p-q}_{n-p}(\alpha)$
If we reject H_0 at least some of the additional $p-q$ variables should be retained.

ANOVA table presentation

Source	d.f.	SS	MS	VR = F*
$x_1, \dots, x_{q-1}$ $x_q, \dots, x_{p-1} \mid x_1, \dots, x_{q-1}$	q − 1 p − q	$SS(x_1, \dots, x_{q-1})$ SS_{extra}	 $\frac{SS_{extra}}{p-q}$	 $\frac{\left(\frac{SS_{extra}}{p-q}\right)}{s^2}$
Overall Regression Residual	p − 1 n − p	SS_R SS_E	 s²	
Total	n − 1	SS_T		

Other Approaches

There are a number of techniques to help decide which explanatory variables to keep in a multiple linear regression model:

1. Using F tests to delete variables
2. Considering All subsets Regression
3. Backward Elimination
4. Stepwise Regression or Modified Forward Regression
5. Akaike's Information Criterion (AIC)

Mortgage Repossessions Example

Mortgage Repossessions example

```
>   View(Mortgage_Repossessions_Data)
> y = Mortgage_Repossessions_Data$Repossessions
> x1 = Mortgage_Repossessions_Data$Affordability
> x2 = Mortgage_Repossessions_Data$Unemployed
> x3 = Mortgage_Repossessions_Data$StartFTSE
> x4 = Mortgage_Repossessions_Data$DebtIncome
```

Full Model

```
> full_model = lm(y~x1+x2+x3+x4)
> summary(full_model)
Call:
lm(formula = y ~ x1 + x2 + x3 + x4)
Coefficients:
              Estimate Std. Error t value
(Intercept)   -672.312  16639.845  -0.040
x1          -19882.473   3151.763  -6.308
x2              12.677      3.930   3.226
x3              -2.128      1.406  -1.514
x4             929.521    168.320   5.522
```

Full model ANOVA

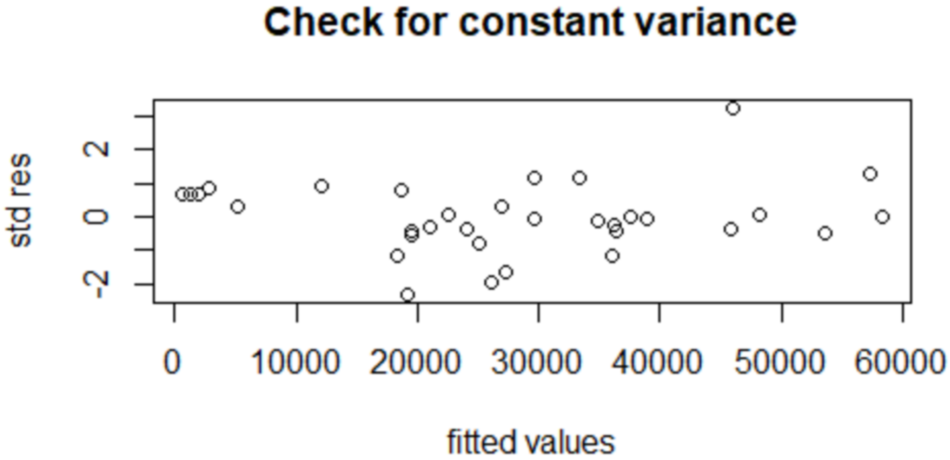
```
> anova(full_model)
Analysis of Variance Table

Response: y
          Df     Sum Sq    Mean Sq F value
x1         1 4844479363 4844479363 51.8941
x2         1 1000678815 1000678815 10.7193
x3         1   13941608   13941608  0.1493
x4         1 2846927652 2846927652 30.4963
Residuals 29 2707244719   93353266
```

Check the model assumptions

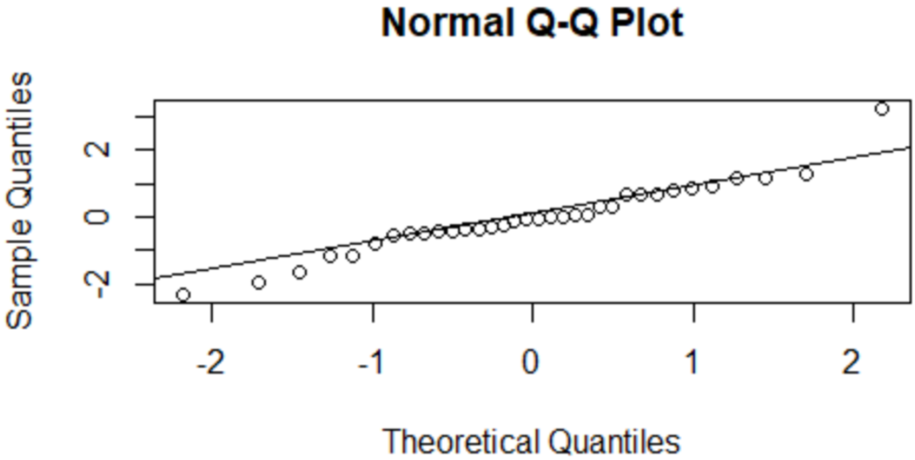
For the full model

```
> d = rstandard(full_model)
> yhat = fitted(full_model)
> plot(yhat, d, main = "Check for constant variance",
xlab = "fitted values", ylab = "std res")
> qqnorm(d)
> qqline(d)
```

Mortgage Repossessions Example



Mortgage Repossessions Example

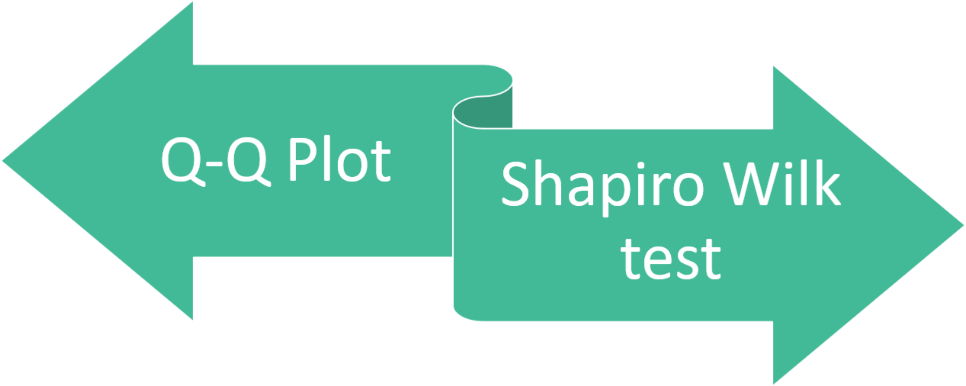


Alternative check of normality assumption

```
> shapiro.test(d)

        Shapiro-Wilk normality test

data:  d
W = 0.94901, p-value = 0.1147
```

Mortgage Repossessions Example

Which do you prefer?



Reduced Model

```
> reduced_model = lm(y~x1+x2+x4)
> summary(reduced_model)

Call:
lm(formula = y ~ x1 + x2 + x4)
Coefficients:
              Estimate Std. Error t value
(Intercept)  -8801.979  16085.005  -0.547
x1          -19926.222   3218.771  -6.191
x2              15.560      3.511   4.432
x4             877.117    168.231   5.214
```

Reduced Model ANOVA

```
> anova(reduced_model)
Analysis of Variance Table

Response: y
          Df     Sum Sq    Mean Sq F value
x1         1 4844479363 4844479363  49.752
x2         1 1000678815 1000678815  10.277
x4         1 2646920869 2646920869  27.183
Residuals 30 2921193109   97373104
```

Extra Sum of Squares

Full model $p = 5$, Reduced model $q = 4$, $p - q = 1$

From Full ANOVA $SS_E^{Full} = 2{,}707{,}244{,}719$

From Reduced ANOVA $SS_E^{Red} = 2{,}921{,}193{,}109$

ExtraSS = $SS_E^{Red} - SS_E^{Full} = 213{,}948{,}390$

From Full Model, $s^2 = MS_E = 93{,}353{,}266$

Subset Test of Hypothesis

H_0: $\beta_3 = 0$

H_1: $\beta_3 \neq 0$

Test Statistic is $F^* = \frac{\left(\frac{SS_{extra}}{p-q}\right)}{s^2} = \frac{\frac{213{,}948{,}390}{5-4}}{93{,}353{,}266} = 2.2918$

under H_0 $F^* \sim F^{p-q}_{n-p} = F^{5-4}_{34-5} = F^{1}_{29}$

We will consider an F test at 95% significance

F test

```
> qf(0.05, 1, 29, lower.tail = FALSE)
[1] 4.182964
```

$F^* = 2.2918 < F_{29}^{1}(0.05)$

Therefore, we cannot reject H_0: $\beta_3 = 0$ at 95% significance level

Meaning we are able to delete x3 and work with the reduced model

Two special cases

There are two cases where the Subset F test can be replaced by a t test

1. Where $p - q = 1$
2. Where there is a natural ordering to the explanatory variables

Deleting one explanatory variable

If p – q = 1 so we are only considering one variable for deletion

The reduced model has just one parameter less than the full model

Then $F^* = t^2$

Where $t = {\widehat{\beta}_{p-1}}\big/{s.e.(\widehat{\beta}_{p-1})}$

And we compare t with $t_{n-p}\left(\frac{\alpha}{2}\right)$ for a 2-sided test of $H_0\colon\ \beta_{p-1} = 0$

Natural ordering of the X_i's

If there is a natural order to the explanatory variables we can consider their significance one at a time via t tests.

The full model with $p-1$ variables and whose multiple regression has $p-1$ df can be thought of as the sum of $p-1$ one variable models

X_1

$X_2|X_1$

...

$X_{p-1}|X_1, \dots, X_{p-2}$ where each has 1 d.f.

Model with natural order

Model with natural order

With this special construction (which will only apply if there is one natural order in which to consider the x's)

We can test successive β_i parameters $i = 1, 2, \ldots, p-1$ with t tests on the parameters divided by their respective standard errors

Model with natural order

If there is a natural order to the X_i's

- Deleting explanatory variables one at a time can be considered by t tests
- This only works if there is a natural order to the explanatory variables
  - which will not typically be the case
- Without a natural order we need some other methods to evaluate which explanatory variables to keep (by Subset test or other means)
- again we have a few alternatives and there is no one correct answer

Exams Style Question

Based on the Boston dataset available on the library MASS, relative to Housing Values in Suburbs of Boston. The variables of interest are:

- Y equal to *medv* is median value of owner-occupied homes in \$1000.
- X_1 equal to *lstat* is the lower status of the population (percent)
- X_2 equal to *rm* is the average number of rooms per dwelling
- X_3 equal to *age* is the proportion of owner-occupied units built prior to 1940

For Model 1: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$, where $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$:

(a) test the hypothesis regarding the overall regression by using the F-test
(b) test the hypothesis regarding the parameters β_j for $j = 0, 1, 2, 3$ by using the t-test

For Model 2: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, where $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$:

(c) test the hypothesis regarding overall regression and the parameters
(d) Which is the best model?

Exams Style Question

We fit the Model 1 to the data:

```
Call:
lm(formula = medv ~ lstat + rm + age)

Residuals:
    Min      1Q  Median      3Q     Max
-18.210  -3.467  -1.053   1.957  27.500

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.175311   3.181924  -0.369    0.712
lstat       -0.668513   0.054357 -12.298   <2e-16 ***
rm           5.019133   0.454306  11.048   <2e-16 ***
age          0.009091   0.011215   0.811    0.418
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.542 on 502 degrees of freedom
Multiple R-squared:  0.639,Adjusted R-squared:  0.6369
F-statistic: 296.2 on 3 and 502 DF,  p-value: < 2.2e-16
```

Exams Style Question

We fit the Model 2 to the data:

```
Call:
lm(formula = medv ~ lstat + rm)

Residuals:
    Min      1Q  Median      3Q     Max
-18.076  -3.516  -1.010   1.909  28.131

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.35827    3.17283  -0.428    0.669
lstat       -0.64236    0.04373 -14.689   <2e-16 ***
rm           5.09479    0.44447  11.463   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.54 on 503 degrees of freedom
Multiple R-squared:  0.6386,Adjusted R-squared:  0.6371
F-statistic: 444.3 on 2 and 503 DF,  p-value: < 2.2e-16
```

Exams Style Question

Solutions:

Looking at the Summary of Model 1

(a) Looking at the last line of the command summary, we find that the F-Test is equal to 296.2 and there is strong evidence against the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ and the R^2 is equal to 63% similar to the adjusted R^2

(b) Moving to the parameters of interest, we look at the summary described above. In this case, we have that there is evidence to reject the null hypothesis $H_0 : \beta_j = 0$ against the alternative $H_1 : \beta_j \neq 0$ for β_1 and β_2, thus the coefficients for *lstat* and *rm* are statistically significant. On the other hand, the intercept and the parameter related to *age* could not reject the null hypothesis, thus the two coefficients are not statistically significant.

Exams Style Question

Solutions:

Looking at the Summary of Model 2

(c) As previously described, we have that the F-statistic is 444.3, thus the overall regression is statistically significant and there is strong evidence against the null hypothesis. Moving to the parameters, in this scenario the parameter of *lstat* and *rm* are statistically significant, while the intercept continuously remains not statistically significant.

(d) Regarding the best model, we compare the adjusted R^2 for both the models. For Model 1, $adj(R^2) = 0.6369$, while for Model 2, $adj(R^2) = 0.6371$, thus the Model 2 is the best model and in this case also all the parameters except the intercept are statistically significant.