Regresní analýza dat
01REAN - Cvičení 06

Jiri Franc

Czech Technical University
Faculty of Nuclear Sciences and Physical Engineering
Department of Mathematics

Outline of todays exercises:

- Model selection in multivariable regression
- Post hoc analysis

Model selection in multivariable regression

How to select the "best"model fitting your data?

Model selection - Helpful metrics and criterion	
STATISTIC	**CRITERION**
Mean squared error (MSE)	Lower the better
R^2	Higher the better
	$>$ 0.60 technical and $>$ 0.20 social sciences
Adj R^2	Higher the better
Fisher-Snedecor F Statistic	Higher the better
Std. Error	Closer to zero the better
t-statistic	Greater better, depends on p-value
AIC	Lower the better
BIC	Lower the better
C_p	Lower the better (close to number of predictors)

Mean squared error (MSE)

Mean squared error (MSE, residual error MS) is the unbiased estimate of σ_ε^2.

$$\text{MSE} = \hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-p-1}$$

The linear regression model

$$Y_i = X_{i1}\beta_1^0 + X_{i2}\beta_2^0 + \cdots + X_{ip}\beta_p^0 + \varepsilon_i = X_i^T\beta^0 + \varepsilon_i, \qquad i = 1 \ldots n$$

In R:

```
model01 <- lm(Y ~ X1*X2 + X3, data=mydata)
    summary(model01) # show results
    summary(model01)$sigma
```

Note: R considers p as a number of predictors, not number of coefficients.
Decomposition of variation same as in ANOVA:

total SS = regression SS + residual error SS

Coefficient of determination

Coefficient of determination R^2

$$R^2 = 1 - \frac{S_R^2}{R_0^2} = 1 - \frac{\text{residual error SS}}{\text{total SS}} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y}_i)^2}$$

We compare our model with the model $Y_i = \bar{Y} + \varepsilon_i$

Adjusted Coefficient of determination R_A^2

$$R_A^2 = 1 - \frac{\text{residual error MS}}{\text{total MS}} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y}_i)^2} \frac{n-1}{n-p-1}$$

We reduce bias in R^2 by replace number of observations n with number of DF, i.e. $R_A^2 < R^2$.

If the model does not include intercept than $R_0^2 = \sum_{i=1}^n Y_i^2$,

$$R^2 = 1 - \frac{S_R^2}{R_0^2} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i)^2}$$

and we compare our model with the model $Y_i = \varepsilon_i$.
Do not compare models with and without intercept on the basis of the coefficient of determination.

In R:

```
summary(model01)$r.squared
summary(model01)$adj.r.squared
```

Fisher-Snedecor F Statistic

If the regression model include the intercept then

$$F = \frac{\frac{R^2}{p-1}}{\frac{1-R^2}{n-p}}$$

If the regression model does not include the intercept then

$$F = \frac{\frac{R^2}{p}}{\frac{1-R^2}{n-p}}$$

Fisher-Snedecor F statistics has F distribution with $p-1$ and $n-p$ DF:

$$\mathcal{L}(F) = F_{p-1,n-1}$$

In R:

```
summary(model01)$fstatistic
```

For forward variable selection (adding one variable per step)

```
add1(model01, Y ~ X1*X2*X3 + X4, test="F")
dropterm( fullmodel, test = "F" )
```

Testing submodel - Model selection

We test the hypothesis that the "true" model is

$$\text{Model I: } Y_i = X_i^T \beta^0 + \varepsilon_i, \quad i = 1, 2, \ldots, n, \quad \text{rank}(X) = p$$

against the alternative that the "true" model is

$$\text{Model II: } Y_i = Z_i^T \beta^0 + e_i, \quad i = 1, 2, \ldots, n, \quad \text{rank}(Z) = q < p$$

under assumption that $\mathcal{M}(Z) \subset \mathcal{M}(X)$, i.e. predictor variables in model II are a subset of those in model I. Then

$$F = \frac{\frac{\text{regression SS for model I - regression SS for model II}}{\text{DF of model I - DF of model II}}}{\frac{\text{residual error SS of model I}}{\text{DF of model I}}} = \frac{\frac{S^2_{R(Z)} - S^2_{R(X)}}{p-q}}{\frac{S^2_{R(X)}}{n-p}}$$

Under the null hypothesis, the F statistics has F-distribution with $(p - q)$ and $(n - p)$ DF.

In R:

```
model01 <- lm(Y ~ X1*X2 + X3, data=mydata)
model02 <- lm(Y ~ X1*X2     , data=mydata)
anova(model01, model02)
```

Information Criterion

F-test is not suitable for step selection, because we don't know joint statistical behavior of all possible F-tests.

Akaike Information Criterion (AIC) is defined as:

$$\mathrm{AIC} = n(1 + \log(2\pi\hat{\sigma}^2)) + 2(p + 1),$$

where $p + 1$ is number of estimated coefficients + estimated variance.
Bayesian information criterion (BIC) for regression model is defined as:

$$\mathrm{AIC} = n(1 + \log(2\pi\hat{\sigma}^2)) + \log(n)(p + 1).$$

In R:
The multiple of the number of DF used for the penalty.

- $k = 2$, gives AIC.
- $k = log(n)$, gives BIC (sometimes called SBC).

```
AIC(model01, k=2)
AIC(model01, k=log(n))
```

Mallows C_p

Another statistics to compare full model `model01` and submodel `model02`.

$$C_p = 2q - n + \frac{S^2_{R(2)}}{\hat{\sigma}^2_{(1)}}$$

The relation with Fisher-Snedecor F Statistic:

$$(p-q)(F-1) = \frac{S^2_{R(2)}}{\hat{\sigma}^2_{(1)}} - n + q = C_p - q.$$

In R:

```
leaps(x=mydata[,c("X1","X2","X3")], y=myddata[,c("Y")],
      names=names(mydata)[1:4], method="Cp")
```

Post hoc analysis - Regression Diagnostics

- Check Normality.
- Check Homoscedasticity.
- Check presence of outliers.
- Check presence of Leverages.

Check presence of outliers

Cook's distance:

Check presence of Leverages

Leverages:

Exercise + Next Lesson:

- Solve problems described at the end of the R code.

Next Lesson:

- Collinearity.
- Multicollinearity.
- Variance Inflation.
- Ridge Regression.