Summary of Model Evaluation and Selection

Model fitting, evaluation and selection is an **iterative** process:

The main purpose of regression modeling:

1. Provide precise predictions of response variable y given a set of predictors/features values
2. Explain the impact of the predictors on the response: how the changes in x impact the changes in y
3. Explain the strengths of associations between predictors and the response.

The general model selection is complex and depends on the main purpose of the study, so people have not agreed on a general rule yet. Here I propose a **flowchart** of model selection process based on three rules of thumbs:

1. The “best” model is only based on the chosen selection criterion.
2. In linear regression, we focus more on “interpretation” than prediction performance. (In time series, we will switch to prediction performance).
3. When it’s too hard to choose, pick the simpler(shorter) model.

Each time we choose an initial model and perform the model selection (basically feature selection), we need to combine the process with model evaluation.

Before fitting the model: pick ALL predictors of interest, and check Multicollinearity

If VIF>10, serious⇒

- Drop one or more highly predictors that are obviously causing the problem
- Combine some highly correlated predictors together if applied
- Scale the data the reduce the impact
- Regularization
- Other regression methods like PCA or Patial LSE

If VIF all <=10
⇓

Fit the initial model 1

Model problem diagnosis
⇓

Check influential points	External residual plot and Cook’s distance > 4/n ⇒	• Don’t simply delete the influential points, report them and present analysis with and without the influential points
Check heteroscedasticity	Breusch- Pagan test or White test with p-value <=0.05 ⇒ Residual plot change of bandwidth with x intervals	• Natual-log transformation or box-cox transformation on y • Use robust standard error instead to perform t tests
Check normality	Jarque- Bera test with p-value <=0.05 ⇒ QQ plot not linear	• Natual-log transformation or box-cox transformation on y • Use weighted least square or generalized

		linear regression • Use robust regression that's not sensitive to assumptions
Check Nonlinearity between x and y (not common in MLR)	Scatter plot ⇒	• Natual-log transformation or power transformation on x

After transformations or obtaining data without influential points
⇓

Fit the initial model 2 with cleaned/transformed data and the same predictors	Perform model problem diagnosis again to check improvements ⇒	• If violations still exit, at this point, you don't need to do more, you can go ahead with the analysis and mark those violations in your discussion

⇓

Model (feature) selection based on different criteria:

Different approaches can give you some candidates, and you can use one standard or combine criteria to choose your final model:

Approach 1: Use t-test for individual coefficient and/or ANOVA(typ=1) to choose the model with shown significant predictors. If more than 1, use the adj-R-squared values to choose a final one.

Approach 2: Best subset with adj-R-squared or Mallow's Cp.

Approach 3: Step-wise selection based on t test p values or AIC/BIC

Optional: when you have several candidate models, you can always use prediction performance to pick the final one. Common criteria: RMSE, MAE, MAPE,...