Variable Selection Methods

XI(×) – X's are not independent (are correlated)

Model Dimensionality

- **Dimensionality** in a model increases as: (1) **more variables** or (2) higher **polynomials** (or other functions) are added to the model
- Selecting the dimensions of a model is **not** a **clear-cut** process; it involves a number of **tradeoffs**.
- **Business** and functional domain **criteria** should be the main driver for adding or removing variables.
- **Statistically,** if the predictors are largely **uncorrelated**, each additional variable added to the model increases its R^2
  - ➢ It is **OK to add** variables to the model, but only if they increase **significantly** the **predictive power** of the model
- If variables are **highly correlated**, multi-collinearity will cause some problems → need to **inspect** for **multi-collinearity** and take remedial actions if multi-collinearity is severe

Selecting Variables and Models

- It is safe to **remove** a **non-significant** variable, but you **can't** drop **significant** variables without sound justification
- **Larger** models are **less biased**, but they have **more variance**
- **Larger** models are usually **preferred**, provided that **multi-collinearity** is **OK** and they are not over-identified

Comparing Models

Different Models

- To **compare** two models that are very **different** (e.g., both use different variables) or use a **different modeling approach** (e.g., regression vs. regression trees), the best approach is to use a **cross-validation** method and select the model with **lowest MSE**

Related Models

- To compare **related models** -- i.e., **full** model vs. **reduced** model (i.e., with some **variables removed**) you can conduct an **ANOVA F-Test** and evaluate if the full model provides significantly stronger **predictive power** (i.e., higher R^2 and lower **MSE**)
- If the F-Test shows **no significant improvement** with the additional variable, you can go with the **reduced model** because it is simpler.
- **Otherwise** select the **full model**, but do cross-validation checks to ensure that the model is **not over-identified**.

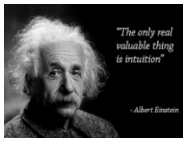
Subset Selection Methods

- **Full vs. Reduced** model testing – use a **business criteria** for variable selection and try a few models that make business sense
- **Best Subset Selection** – with P possible candidate variables, build **P simple** regression models, one for each variable; then build all possible regressions with **2 variables**; then 3 and so on. Use cross-validation to select the best model
  - If P is large → **combinatorial explosion** of models to test
- **Step Methods** – progressively **adding (Forward) or removing (Backwards)** variables, or **both** back and forth **(Stepwise)**

Step Methods

Progressively **adding/removing** variables and **testing**:

- ✓ **Forward method:** start with **no predictors**; then **add** the most significant predictor; then the next one; continue until **none** of the **remaining** variables are **significant** when added to the model
- ✓ **Backwards method:** start with **all predictors**; then **remove** the least significant variable; then the next; and so on until **no more remaining** variables are **significant** when added to the model
- ✓ **Stepwise method:** similar to forward or backwards, except that variables can be **added** or **removed** in any iteration. The **p-value** for inclusion or exclusion can be set separately



Full vs. Reduced Model: Intuition

Suppose you want to test 2 added predictors to a reduced model:

Reduced (or Restricted) Model: $Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \ldots\ldots + \varepsilon$

Full Model: $Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \ldots\ldots + \beta_3(X_3) + \beta_4(X_4) + \varepsilon$

- Predictors can be added **1 at a time** or as a **group** of variables
- If we add only **1 predictor** we can see if its **coefficient** is significant
- If we add **a group** of predictors we need to **test** the added variables **as a whole** to see if the full model is better than the reduced model.
- We need to test if the **Full** model's **SSE** is significantly **lower** than the **Reduced** model's **SSE**, taking into account the **loss of degrees of freedom** caused by adding more variables to the model.
- We can do this with an **ANOVA F-Test**

Full vs. Reduced Model Testing

SSE_R = Sum of Squared Errors of the Reduced model

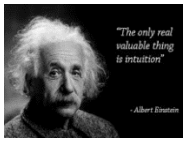
SSE_F = Sum of Squared Errors of the Full model

df_R = Degrees of freedom of the Reduced model

df_F = Degrees of freedom of the Full model

$$F_{(df_R, df_F)} = \frac{SSE\ reduction\ per\ df\ lost}{Relative\ to\ the\ Full\ model} = \frac{\frac{SSE_R - SSE_F}{df_R - df_f}}{\frac{SSE_F}{df_F}}$$

- **If** F test is **not significant** → retain the **reduced model**
- **If** F test is **significant** → retain the **full model**
- But also **inspect** the full model for:
  - ✓ **Multicollinearity** and
  - ✓ **Over-Identification** → **Cross-validation** comparison Full vs. Reduced models



Best Subset Selection: Intuition

Suppose you have P possible predictors → 2 extreme models:

Null Model (no predictors): $Y = \beta_0 + \varepsilon$

Full Model: $Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \ldots\ldots + \beta_P(X_P) + \varepsilon$

- **Start** with the **Null** model, then all **P single-predictor** models, then all possible **2-predictor** models, etc., **ending** with the **Full model**
- You can then compare all resulting models using **cross-validation**
- This method **works** well when **P is small** because you end up testing all possible models
- But if **P is large**, the pool of possible models will grow exponentially **(2^P-1)** and it may not be computationally practical to test all of them.
  - 10 variables → 2^{10}-1 = 1,024 models
  - 20 variables → 2^{20}-1 = 1,048,576 models
- **Step methods** start with 1 (or all) variables and add (or remove) the most (or least) significant variables, one at a time. This is computationally simpler

R Tips

`lm.reduced = lm(y=x1+x2+etc., etc.)` → Reduced model

`lm.full = lm(y=x1+x2+x3+x4+etc., etc.)` → Full model

`anova(lm.reduced, lm.full)` → Tests if the added variables in the full model add significant explanatory power →If the F-Test is significant, most of the coefficients of the added variables should be significant too

`regsubsets(){leaps}` → This function searches for the best subset that minimizes the RSS

```
fit.subsets=regsubsets(y=x1+x2+etc.,
                       data=dataName)
```
→ Works just like `lm()` but it finds the best subset with a default maximum of 8 variables

```
fit.subsets=regsubsets(y=x1+x2+etc.,
                       data=dataName,nvmax=19)
```
→ the `nvmax=` attribute can be used to change the maximum number of variables to be included

KOGOD SCHOOL
of
BUSINESS