Advanced Regression: 2a Introduction to variable ranking and selection

Verena Zuber

Epidemiology and Biostatistics, Imperial College London

24th January 2019

Variable ranking, variable importance and variable selection

Classical variable or model selection

- Variance explained and ANOVA
- Akaike criterion and other likelihood-based measures

Variable importance

- Concept of variable importance
- Measures of variable importance

Variable ranking

- Concept and aim of variable ranking
- Quantitative outcome
- Example: Genome-wide association studies
- 2-groups or binary outcome: *t*-test
- Example: Differential expression

Variable selection, variable importance and variable ranking

- **Variable or model selection**: To select a model (a set of variables, i.e. one or many variables) jointly.
- **Variable importance**: To rank each variable with respect to its association with the outcome accounting for its correlation with other predictor variables.
- **Variable ranking**: To rank each variable marginally (without considering the influence of other variables) with respect to its association with the outcome.

Altough these terms are often used interchangeable, each concept has its own motivation and interpretation.

Variable or model selection

Variable or model selection

To select a model (a set of variables, i.e. one or many variables) jointly.

- Focus is not on a single variable but on a model, i.e. one or a combination of many variables.
- Motivation: **To understand** which combination of variables explains best the outcome and **to predict** future outcomes.

Classical variable or model selection

Measures used to compare models:

- Proportion of variance explained
- $F-$statistic and analysis of variances (ANOVA)
- Likelihood based methods
  - Akaike information criterion (AIC)
  - Bayesian information criterion (BIC)

Proportion of variance explained

Linear model

$$y = x\beta + \epsilon$$

- Variance decomposition:

$$\underbrace{var(y)}_{\text{Total Variance}} = \underbrace{var(x\beta)}_{\text{Explained Variance}} + \underbrace{var(\epsilon)}_{\text{Error Variance}}$$

- R^2 is the proportion of variance explained by a model

$$R^2 = \frac{var(x\beta)}{var(y)} = 1 - \frac{var(\epsilon)}{Var(y)}$$

Proportion of variance explained

- How to compute? Using sum of squares (SS):
  - ◇ Total variance

$$\hat{var}(y) = \frac{1}{n-1} SS_{Total} = \frac{1}{n}\sum_{i=1}^{n}[y_i - \bar{y}]^2$$

  - ◇ Explained variance $\hat{y} = x\beta$

$$\hat{var}(\hat{y}) = \frac{1}{n-1} SS_{Explained} = \frac{1}{n}\sum_{i=1}^{n}[\hat{y}_i - \bar{y}]^2$$

  - ◇ Error variance

$$\hat{var}(\epsilon) = \frac{1}{n-1} SS_{Error} = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{y}_i]^2$$

where the mean is defined as $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Occam's razor

- When comparing two models it is important not only to consider R^2 but also how complex they are, i.e. how many variables they include.
- Occam's razor (law of parsimony): Simpler solutions are more likely to be correct than complex ones (William of Ockham 1287–1347ad).
- Problem: R^2 will always increase when including more variables.
- Question: Is including more variables actually improving the model fit significantly?

Adjusted proportion of variance explained

- Adjusted R^2

$$R^2_{adj} = 1 - (1 - R) \times \frac{n-1}{n-p-1}$$

- Alternative representation with degrees of freedom (df)

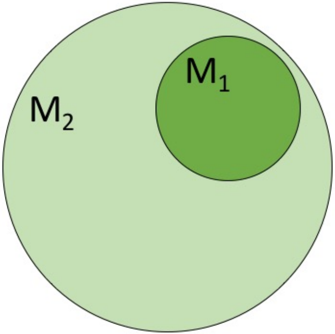
$$R^2_{adj} = 1 - \frac{SS_{Error}/df_e}{SS_{Total}/df_t}$$

where $df_t = n - 1$ and $df_e = n - p - 1$

Analysis of Variances (ANOVA)

What is a nested model?

- Model M_1 is a nested model of model M_2 when model M_2 contains M_1.
- M_1 is a subset of M_2: $M_1 \subset M_2$
- Example: $M_1 = bmi$ and $M_2 = bmi + map$



Analysis of Variances (ANOVA)

- $F-$test, to compare two nested models: a 'full' and a 'reduced' model.
- M_2 'full' model included p_2 predictors.

	df	SS
Regression fit	p_2	$SS_{Explained}(M_2)$
Error	$n - p_2 - 1$	$SS_{Error}(M_2)$
Total	$n - 1$	$SS_{Total}(M_2)$

- M_1 'reduced' model included p_1 predictors, where $p_1 < p_2$.

	df	SS
Regression fit	p_1	$SS_{Explained}(M_1)$
Error	$n - p_1 - 1$	$SS_{Error}(M_1)$
Total	$n - 1$	$SS_{Total}(M_1)$

Analysis of Variances (ANOVA)

- F-test, to compare two nested models: a ‘full’ model (M_2) and a ‘reduced’ model (M_1).
- It will always hold that
  - $R^2(M1) \leq R^2(M2)$
  - $SS_{Error}(M_1) \geq SS_{Error}(M_2)$
  - $p_2 > p_1$
- But is the ‘full’ model (M_2) significantly better than a ‘reduced’ model (M_1)?

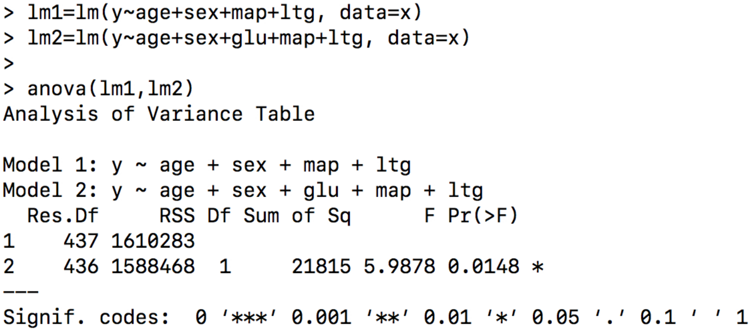
Analysis of Variances (ANOVA)

- H_0: Model M_2 fits the data as good as model M_1.

$$F = \frac{(SS_{Error}(M_1) - SS_{Error}(M_2))/(p_2 - p_1)}{SS_{Error}(M_2)/(n - p_2 - 1)}$$

- Under the Null, the test statistic F follows an F-distribution with $(p_2 - p_1)$ and $(n - p_2 - 1)$ degrees of freedom.
- Interpretation 1: If we reject H_0, M_2 fits the data significantly better than model M_1.
- Interpretation 2: By adding more predictors in the complex model compared to the reduced model we can explain more of the variation in Y.
- `anova(M1,M2)` command in `R`.

ANOVA example: Diabetes data



```
> lm1=lm(y~age+sex+map+ltg, data=x)
> lm2=lm(y~age+sex+glu+map+ltg, data=x)
>
> anova(lm1,lm2)
Analysis of Variance Table

Model 1: y ~ age + sex + map + ltg
Model 2: y ~ age + sex + glu + map + ltg
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    437 1610283
2    436 1588468  1     21815 5.9878 0.0148 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- It improves the model fit to add the variable glu to the model.
- The 'full' model (M_2) is better than the 'reduced' model (M_1).

Akaike information criterion (AIC)

- Akaike information criterion (AIC) combines a measure of model fit with a measure of model complexity.

$$AIC = -2 log L + 2p$$

  - L Maximum likelihood of the model
  - p Model complexity: Number of parameters in the model
- The best model is the one with the minimum AIC value (minimum information loss).
- The AIC can be used for model comparison, not to assess the quality of the model fit.
- `AIC(M1,M2)` command in R.

Bayesian information criterion (BIC)

- Also the Bayes information criterion (BIC) combines a measure of model fit with a measure of model complexity.

$$BIC = -2 logL + \log(n)p$$

  - ◇ L Maximum likelihood of the model
  - ◇ p Model complexity: Number of parameters in the model
- The best model is the one with the minimum BIC value (minimum information loss).
- The BIC can be used for model comparison, not to assess the quality of the model fit.
- `BIC(M1,M2)` command in R.

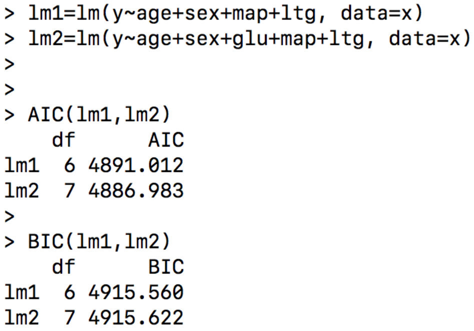
AIC and BIC

- More generally, we can understand information criteria (IC) as a compromise between model fit and model complexity

$$IC = -2logL + k \times p$$

  - L Model fit: Maximum likelihood of the model
  - p Model complexity: Number of parameters in the model
- The best model is the one with the minimum IC value (minimum information loss).
- k defines the penalty of the model complexity
  - AIC: $k = 2$
  - BIC: $k = \log(n)$

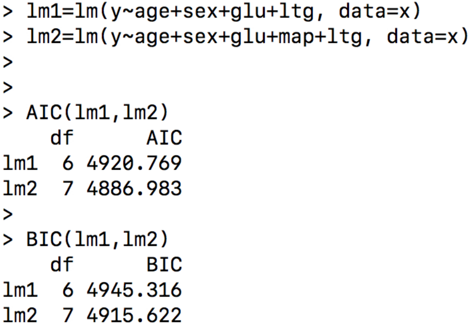
AIC and BIC in R



```
> lm1=lm(y~age+sex+map+ltg, data=x)
> lm2=lm(y~age+sex+glu+map+ltg, data=x)
>
>
> AIC(lm1,lm2)
    df      AIC
lm1  6 4891.012
lm2  7 4886.983
>
> BIC(lm1,lm2)
    df      BIC
lm1  6 4915.560
lm2  7 4915.622
```

- There is no consensus between AIC and BIC.
- Using the AIC we would prefer M_2, but using the BIC we would prefer M_1.
- This is not a strong evidence that adding the variable glu (glucoase) has a lot of benefit.

AIC and BIC in R



```
> lm1=lm(y~age+sex+glu+ltg, data=x)
> lm2=lm(y~age+sex+glu+map+ltg, data=x)
>
>
> AIC(lm1,lm2)
    df      AIC
lm1  6 4920.769
lm2  7 4886.983
>
> BIC(lm1,lm2)
    df      BIC
lm1  6 4945.316
lm2  7 4915.622
```

- In contrast the variable map (blood pressure) greatly improves the model fit.
- Since the variable map (blood pressure) is supported by both methods we have greater confidence that it improves the model fit.

How to decide which models to test?

- Backward selection
  1. Start with the full model and include all p variables available.
  2. Identify the variable with the weakest evidence and remove it.
  3. Evaluate the model with $p-1$ variables.
  4. Identify the variable with the weakest evidence and remove it.
  5. ...
- Forward selection
  1. Start to evaluate all models including just a single predictor variable.
  2. Identify the variable with the strongest univariable impact.
  3. Evaluate all models including the best single predictor variable and one additional variable.
  4. Identify the tuple of two variables with the strongest impact.
  5. ...

Alternatives for model selection

Warning!

Backward and forward selection do rarely agree. They are highly instable and there is no guaranty that they find the optimal model. It is not recommended to use them.

Alternatives for model selection:

- Evaluate all possible models: Becomes computationally infeasible even with a moderate number of variables.
  - $p = 10$ variables have $2^{10} = 1,024$ possible models
  - $p = 20$ variables have $2^{20} = 1,048,576$ possible models
- Penalised regression (Lecture 3b)

Variable importance

Variable importance

To rank each variable with respect to its association with the outcome accounting for its correlation with other predictor variables.

Motivations:

1. Interpretation: Within a given model how important is the contribution of a single variable accounting for its correlation with other predictor variables.
2. Conditional on the model, the focus is on a single variable.
3. Motivation: To understand

How to define a good measure for variable importance ϕ?

Variable importance: Conditions

1. Decomposition of the proportion of variance explained R^2
   The sum over the importance of all p variables should sum up to R^2
   $$\sum_{j=1}^{p} \phi_i = R^2$$
2. Non-negativity:
   A variable should not be negatively important. Negative values in importance or shares are not interpretable.
3. Inclusion:
   If a variable has a true effect on the outcome, its importance should be larger than zero.
4. Exclusion:
   If a variable has no true effect on the outcome, its importance should be zero.

Variable importance: Available measures

- If there is no correlation between the predictors x, then the marginal Pearson sample correlation $\hat{cor}(x, y)$ is a measure for variable importance.
- Pearson sample correlation for variable $j = 1, ..., p$ is defined as

$$\hat{cor}(x_j, y) = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- If there is correlation between the predictors x, then there is no consensus which measure to use.
  - Standardized regression coefficients and partial correlations do not decompose R^2.
  - Relative importance measures like Genizi, Pratt, or CAR score decompose R^2 but are not widely used.
  - Variable importance will be revisited in regression trees and random forrests (Lecture 4c).

Variable ranking

Variable ranking:

- To rank each variable individually (without considering the influence of other variables) with respect to its association with the outcome.

- Rankings based on univariable measures provide an intuitive representation which variable has the strongest marginal association.
- 'Marginal': Considered in isolation, not taking into account correlation with other predictors.
- Interpretation: Which variable is marginally most associated.
- Motivation: To understand, not to predict.

Variable ranking

- Some ranking methods can condition on relevant covariates (e.g. clinical variables or confounder).
- Methods for variable ranking are computationally simple and easy to interpret.
- Robust to multi-collinearity (Lecture 2b).
- Perform p independent tests.
- Cut-off points which variables are considered as significant or Non-Null need to adjust for multiple testing (Lecture 2c).
    1. Quantitative outcome: Massively univariate linear model
    2. 2-groups or binary outcome: t-test

Quantitative outcome: Massively univariate linear model

Massively univariate linear model

- For each variable x_j, where $j, \ldots p$ we fit a univariable linear model

$$y = \alpha + x_j \beta_{UNI}(j) + \epsilon$$

- Interpretation: $\beta_{UNI}(j)$ is the expected change in y for a one-unit change in x_j.
- The effect of variable j is considered marginally, in isolation irrespective of the other $j = 1, \ldots, p$ variables.
- The linear model can include additional covariates that need to be adjusted for.

Quantitative outcome: How to rank?

- Interpretation: β-coefficients from the univariable linear model tell us only about the mechanics of the effect, but not about the actual importance.
- This is in analogy with the β-coefficients from the multivariable linear model.
- Instead for the ranking it is recommended to use:
  - $\rightarrow$ Standardised betas, or effect sizes, or z-scores:

$$z_j = \beta_{UNI}(j)/se(\beta_{UNI}(j))$$

  - $\rightarrow$ $p-$values (2-sided)

$$p_j = 2 * pnorm(-abs(z_j))$$

Quantitative outcome: Massively univariate linear model

Contrasting the ordinary least squares estimate with the massively univariate linear model estimate:

- Ordinary least squares estimate

$$\hat{\beta}_{OLS} = \underbrace{(x^t x)^{(-1)}}_{p \times p} \underbrace{x^t}_{p \times n} \underbrace{y}_{n \times 1}$$

- Massively univariate linear regression estimate

$$\hat{\beta}_{UNI} = \underbrace{diag(x^t x)^{(-1)}}_{p \times p} \underbrace{x^t}_{p \times n} \underbrace{y}_{n \times 1}$$

Example: Genome-wide association studies

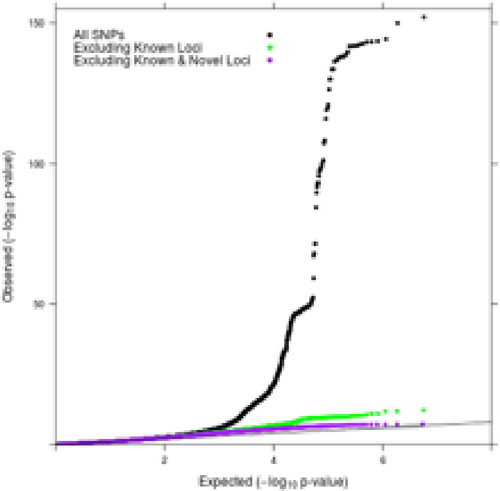
- Variables: All single-nucleotide polymorphisms (SNPs) or genetic variants genome-wide
- SNPs are recoded to have an additive effect:

$$x_{ij} = \begin{cases} 2 & \text{if genotype of individual } i \text{ at SNP } j \text{ is } AA \\ 1 & \text{if genotype of individual } i \text{ at SNP } j \text{ is } Aa \\ 0 & \text{if genotype of individual } i \text{ at SNP } j \text{ is } aa \end{cases}$$

- Assume we have a quantitative outcome Y like for example BMI or blood pressure.
- After accounting for population stratification we fit a linear model

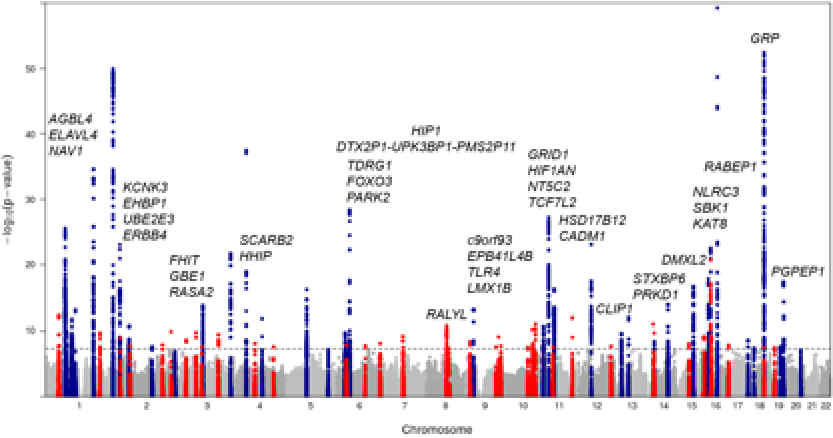
$$y = \alpha + x_j \beta_{UNI}(j) + \epsilon.$$

Example: Genome-wide association studies on BMI



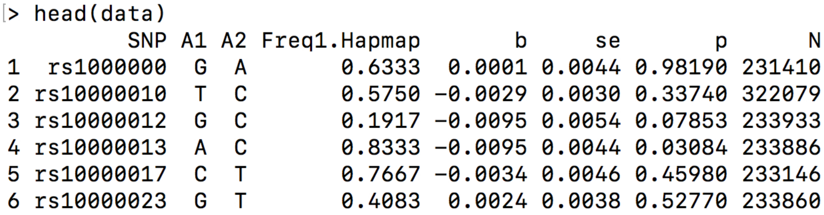
Locke et al, 2015

Example: Genome-wide association studies on BMI



Locke et al, 2015

Example: Genome-wide association studies on BMI



```
> head(data)
         SNP A1 A2 Freq1.Hapmap       b     se       p      N
1  rs1000000  G  A       0.6333  0.0001 0.0044 0.98190 231410
2 rs10000010  T  C       0.5750 -0.0029 0.0030 0.33740 322079
3 rs10000012  G  C       0.1917 -0.0095 0.0054 0.07853 233933
4 rs10000013  A  C       0.8333 -0.0095 0.0044 0.03084 233886
5 rs10000017  C  T       0.7667 -0.0034 0.0046 0.45980 233146
6 rs10000023  G  T       0.4083  0.0024 0.0038 0.52770 233860
```

Publicely available summary data from `https://portals.broadinstitute.org/collaboration/giant`

Locke et al, 2015

2-groups or binary outcome: *t*-test

Motivation:

- Consider we have two groups, where
  - Group 1 has sample size n_1
  - Group 2 has sample size n_2
- Example: Group 1 is cases and group 2 is healthy controls

$$y_i = \begin{cases} 1 & \text{if individual } i \text{ is in group 1} \\ 0 & \text{if individual } i \text{ is in group 2} \end{cases}$$

Differential effects

Is there a differential effect between group 1 and group 2?

2-groups or binary outcome: t-test

Differential effects

For each variable x_j, where $j, ...p$ we fit a t-test t_j

$$t_j = \frac{\mu_j(1) - \mu_j(2)}{s_j^2}$$

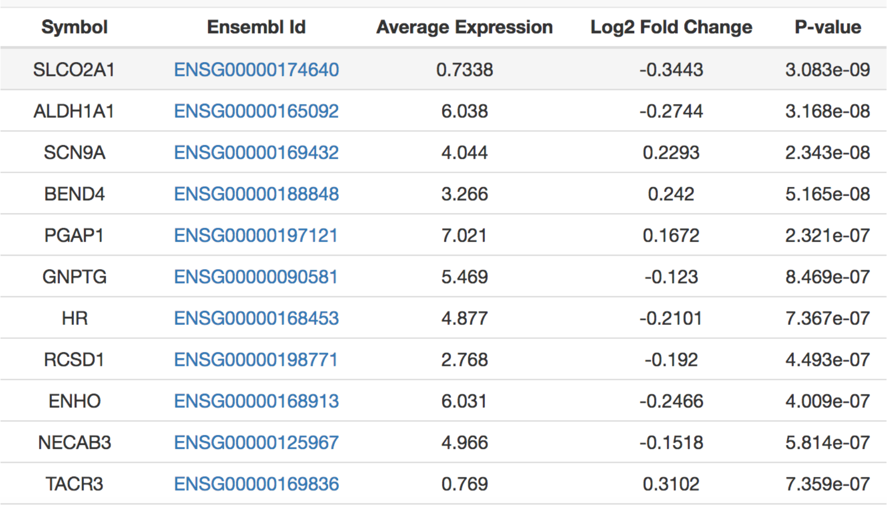
- $\mu_j(1)$: mean of x_j in group 1
- $\mu_j(2)$: mean of x_j in group 2
- $\mu_j(1) - \mu_j(2)$: fold change
- s_j^2 sample variance of the fold change, a weighted mixture of the sample variance in group 1 and group 2

$$s_j^2 = \frac{1}{n_1 + n_2 - 2}((n_1 - 1)var(x_1) + (n_2 - 1)var(x_2))$$

Differential expression: Schizophrenia cases vs controls

- Fromer et al sequenced RNA from dorsolateral prefrontal cortex of schizophrenia cases ($n = 258$) and control subjects ($n = 279$).
- Goal: Detect genes that are differentially expressed between schizophrenia cases and controls.
- Findings: 693 genes show significant case/control differential expression.

Differential expression: Schizophrenia cases vs controls



Symbol	Ensembl Id	Average Expression	Log2 Fold Change	P-value
SLCO2A1	ENSG00000174640	0.7338	-0.3443	3.083e-09
ALDH1A1	ENSG00000165092	6.038	-0.2744	3.168e-08
SCN9A	ENSG00000169432	4.044	0.2293	2.343e-08
BEND4	ENSG00000188848	3.266	0.242	5.165e-08
PGAP1	ENSG00000197121	7.021	0.1672	2.321e-07
GNPTG	ENSG00000090581	5.469	-0.123	8.469e-07
HR	ENSG00000168453	4.877	-0.2101	7.367e-07
RCSD1	ENSG00000198771	2.768	-0.192	4.493e-07
ENHO	ENSG00000168913	6.031	-0.2466	4.009e-07
NECAB3	ENSG00000125967	4.966	-0.1518	5.814e-07
TACR3	ENSG00000169836	0.769	0.3102	7.359e-07

https://www.synapse.org

Take away: Variable ranking, variable importance and variable selection

Variable ranking, variable importance, and variable or model selection are three different concepts.

- Variable ranking considers the marginal effect of a variable.
- Variable importance considers the contribution of a single variable within a model.
- Variable or model selection considers not single variables but models, i.e. sets of variables.

Take away: What is your aim?

To understand

- Variable and model selection: To understand how variability of y can be explained by a model.
  Classical approaches to model selection can help with precise hypothesis, but are impractical for high-dimensional data.
  Lecture 3b on penalised regression
- Variable importance: To understand the importance of a variable within a given model.
  Lecture 2b How to work with correlated predictors
- Variable ranking: To understand the marginal importance of a variable.
  Lecture 2c Multiple testing

To predict

- Variable and model selection