7. Variable selection in regression

- significance test
- variable selection
- step-wise regression

Recap of linear regression

a linear regression model is

$$y = X\beta + u, \quad X \in \mathbf{R}^{N \times n}$$

homoskedasticity assumption: u_i has the same variance for all i, given by σ^2

- prediction (fitted) error: $\hat{u} := \hat{y} - y = X\hat{\beta} - y$
- residual sum of squares: $\mathrm{RSS} = \|\hat{u}\|_2^2$
- a consistent estimate of σ^2: $s^2 = \mathrm{RSS}/(N-n)$
- $(N-n)s^2 \sim \chi^2(N-n)$ (under Gaussian assumption on u)
- square root of s^2 is called **standard error of the regression**
- $\mathbf{Avar}(\hat{\beta}) = s^2(X^T X)^{-1}$ (estimated asymptotic covariance)

Significance tests for linear regression

- testing a hypothesis about a coefficient

$$H_0 : \beta_k = 0 \quad \text{VS} \quad H_1 : \beta_k \neq 0$$

we can use both t and F statistics

- testing using the fit of the regression

$$H_0 : \text{reduced model} \quad \text{VS} \quad H_1 : \text{full model}$$

if H_0 were true, the reduced model $(\beta_k = 0)$ would lead to smaller prediction error than that of the full model $(\beta_k \neq 0)$

Testing a hypothesis about a coefficient

statistics for testing hypotheses:

$$H_0 : \beta_k = 0 \quad \text{VS} \quad H_1 : \beta_k \neq 0$$

- $\frac{\hat{\beta}_k}{\sqrt{s^2((X^T X)^{-1})_{kk}}} \sim t_{N-n}$

- $\frac{(\hat{\beta}_k)^2}{s^2((X^T X)^{-1})_{kk}} \sim F_{1,N-n}$

the above statistics are Wald statistics (see derivations in Greene book)

- the term $\sqrt{s^2((X^T X)^{-1})_{kk}}$ is referred to **standard error of the coefficient**
- the expression of SE can be simplified or derived in many ways (please check)
- e.g. MATLAB, R use t-statistic (two-tail test)

Testing on reduced models

hypotheses are based on the fitting quality of reduced/full models

$$H_0 : \text{reduced model} \quad \mathrm{VS} \quad H_1 : \text{full model}$$

reduced model: $\beta_k = 0$ and full model: $\beta_k \neq 0$

the F-statistic used in this test

$$\frac{(\mathrm{RSS}_R - \mathrm{RSS})}{\mathrm{RSS}/(N-n)} \sim F(1, N-n)$$

- RSS_R and RSS are the residual sum squares of reduced and full models
- RSS_R cannot be smaller than RSS, so if H_0 were true, then the F statistic would be zero
- e.g. `fitlm` in MATLAB use this F statistic, or in ANOVA table

F-test for regression

most statistical softwares assume an *intercept* term in the model

$$y = \beta_0 + X_1\beta_1 + \cdots + X_n\beta_n$$

we ask if all of the regression coefficients are zero (**except** β_0)

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_n = 0, \quad \mathrm{VS} \quad H_1 : \text{at least one } \beta_j \text{ is non-zero}$$

the F-statistic is

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/n}{\mathrm{RSS}/(N-n-1)} \sim F(n, N-n-1)$$

where $\mathrm{TSS} = \sum (y_i - \bar{y})^2 = \|y - \bar{y}\mathbf{1}\|_2^2$ (to test a regression versus a constant model)

- if linear model assumption is correct then we can show $\mathbf{E}[\mathrm{RSS}/(N-n-1)] = \sigma^2$
- $\mathbf{E}[(\mathrm{TSS} - \mathrm{RSS})/n] = \sigma^2$ if H_0 is true and is $> \sigma^2$ if H_1 is true
- if no relationship between predictors and y then F-statistic should be close to 1
- this F statistic is reported in almost all regression software (test model versus constant model)

proof sketch of F statistic

let $(\hat{\beta}_0, \hat{u}_0)$ and $(\hat{\beta}, \hat{u})$ be (solution, error) of restricted and unrestricted models resp.

$$\hat{u}_0 = y - X\hat{\beta}_0 = \hat{u} - X(\hat{\beta}_0 - \hat{\beta}) \quad \Rightarrow \quad \hat{u}_0^T \hat{u}_0 = \hat{u}^T \hat{u} + (\hat{\beta}_0 - \hat{\beta})^T X^T X (\hat{\beta}_0 - \hat{\beta})$$

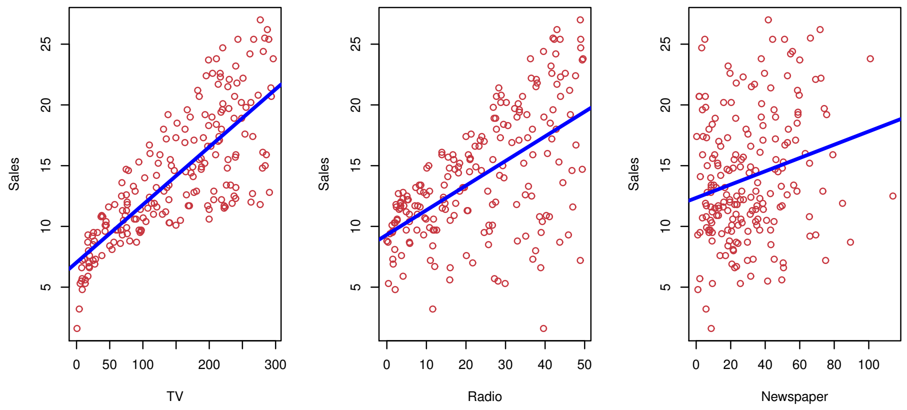
- above equation needs the result that $X^T \hat{u} = 0$ (optimal residual is orthogonal to regressor)
- RSS in restricted model must be greater than that of full model
- the difference between restricted and full RSS is chi-square distributed under H_0

  $$(1/\sigma^2)(\hat{u}_0^T \hat{u}_0 - \hat{u}^T \hat{u}) \triangleq \text{(TSS-RSS)}/\sigma^2 = (\hat{\beta}_0 - \hat{\beta})^T (\sigma^2 (X^T X)^{-1})^{-1} (\hat{\beta}_0 - \hat{\beta}) \sim \mathcal{X}^2(n)$$

  (see more in Wald test statistics)
- RSS$/\sigma^2 = (N - n - 1)s^2/\sigma^2 \sim \mathcal{X}^2(N - n - 1)$
- ratio of two chi-square variables is then an F distribution

Example: single VS multiple regressions

we explore `Advertising` data set[3]



- we regress the units of **sales** (in thousand) on different budgets of advertising channels: **TV, radio and newspaper** (in thousand dollars) for 200 different markets
- performing single and multiple regressions can give different results ?

[3] taken from the book, G.James et al., An Introduction to Statistical Learning, Springer, 2015

regress sales on TV

	Estimate	SE	tStat	pValue
(Intercept)	7.0326	0.45784	15.36	1.4063e-35
x1	0.047537	0.0026906	17.668	1.4674e-42

regress sales on radio

	Estimate	SE	tStat	pValue
(Intercept)	9.3116	0.5629	16.542	3.5611e-39
x1	0.2025	0.020411	9.9208	4.355e-19

regress sales on newspaper

	Estimate	SE	tStat	pValue
(Intercept)	12.351	0.62142	19.876	4.7135e-49
x1	0.054693	0.016576	3.2996	0.0011482

regress sales on TV, radio, newspaper

	Estimate	SE	tStat	pValue
(Intercept)	2.9389	0.31191	9.4223	1.2673e-17
x1	0.045765	0.0013949	32.809	1.51e-81
x2	0.18853	0.0086112	21.893	1.5053e-54
x3	-0.0010375	0.005871	-0.17671	0.85992

- single regression:
  - spending 1000 dollars in each of advertising channels (TV, radio, newspaper) increases the sales around 48, 203, and 55 units respectively
  - with a significance level of $\alpha = 0.001$, all predictors are significant
- multiple regression:
  - regression coef. of TV and radio are almost similar to those in single regression
  - p value of newspaper coefficient is no longer significant, contrary to result in single regression

explanation on advertising data set

- in single regression, the coefficient represents the average affect of the predictor while *ignoring* other predictors
- in multiple regression, a predictor coef. represents the average effect while *holding* other predictors fixed
- correlation matrix of all preditors

```
>> corrcoef([x.TV x.radio x.newspaper])
    1.0000    0.0548    0.0566
    0.0548    1.0000    0.3541
    0.0566    0.3541    1.0000
```

a tendency to spend more on newspaper where more is spent on radio

- newspaper does not actually affect sales but a higher in sales is a result of a tendency of spending more on radio

Variable selection

performing a multiple linear regressing raise questions such as

- is at least one of $X_1, X_2, \ldots, X_n$ useful in predicting Y?
- do all predictors help to explain Y or only a subset is useful?

example: advertisting data set (multiple regression test)

```
Number of observations: 200, Error degrees of freedom: 196
Root Mean Squared Error: 1.69
R-squared: 0.897,  Adjusted R-Squared: 0.896
F-statistic vs. constant model: 570, p-value = 1.58e-96
```

- F is relatively larger than 1 and p-value with F statistic is essentially zero
- at least one of TV, radio, newspaper is associated with the increased sale

the first step in multiple regression is to compute F-statistic and see if at least one predictor is associated with the response

a problem of variable selection involves

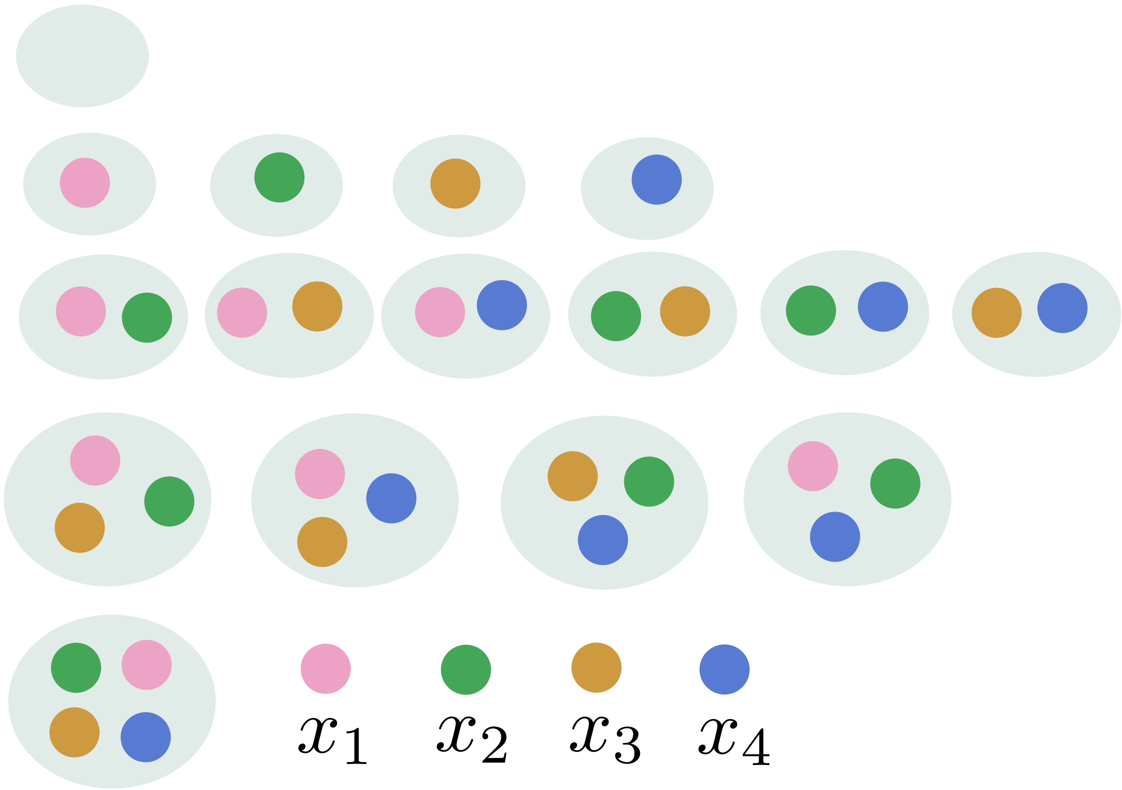
- find a subset of predictors that are associated with the response
- fit a single model consisting of those predictors

these requires model selection criteria such as

- AIC, BIC
- adjusted R^2

Best subset selection

consider $x_1, x_2, \ldots, x_p$ as p predictors



S_k: the model class that each contains k predictors (S_0 has only constant term)
there are $\binom{p}{k}$ sub-models in S_k and no. of all possible sub-models is $\sum_{k=1}^{p} \binom{p}{k} = 2^p$

we would like to pick the 'best' model according to some model selection criterion

steps in variable selection

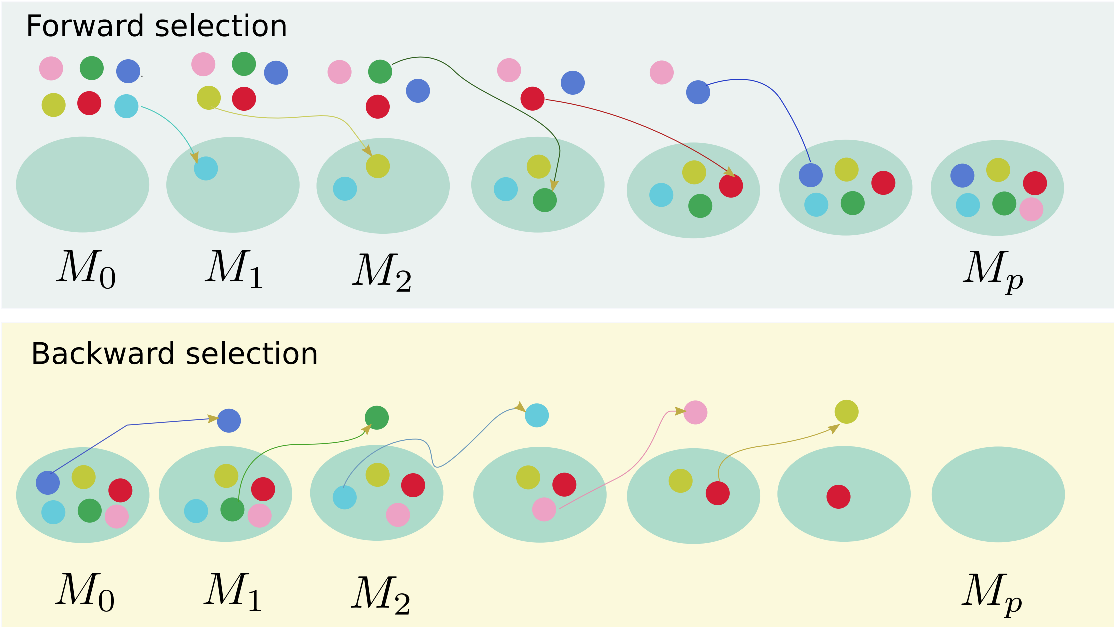
1. for $k = 1, \ldots, p$
2. for $j = 1, \ldots, \binom{p}{k}$
   (a) fit all 'p choose k' sub-models that contain k predictors
   (b) pick the best among $\binom{p}{k}$ models and call it M_j
   (c) here 'best' is defined as having the smallest RSS on training data
3. select a single best model among $M_0, M_1, \ldots, M_p$ using *cross-validated* prediction error, AIC, BIC or adjusted R^2

step 3 is one of the two approaches to obtain the best model having *a low test error*

- *indirectly* estimate test error by *adjusting* training error to account for bias due to overfitting (here, using model selection score instead)
- *directly* estimate the test error, using a validation set/CV approach

Stepwise selection

when p is large, the best subset selection suffers from looking in a large search space



- stepwise selection explores over a a more *restricted* set of models
- forward selection starts from a null model, while backward selection starts from a full model

Forward and backward selection

forward selection

- starts with the **null model** (or model with an intercept)
- sequentially *add* the predictor that *most* improve the fit
- if no predictor improves the model, stop the process and return the model

backward selection

- starts with the **full model**
- sequentially *delete* the predictor that *least* impact on the fit
- stop the process until a stopping rule is satisfied

backward selection can only be used when $N > n$ while forward selection can always be used

mixed selection

- start with no variables in the model and with forward selection
- add the variable that provides the best fit
- continue to add variables one-by-one (p-values for variables can become larger)
- if at any point, the p-value for one of the variables rises above a certain threshold, remove that variable from the model
- perform these forward and backward steps untill all variables have a sufficiently low p-value and all variables outside the model would have a large p-value if added to the model

Step-wise regression in MATLAB

according to MATLAB implementaion

- `stepwiselm` uses forward and backward stepwise regression to determine a final model
- at each step, the function searches for terms to add to the model or remove from the model based on a criterion (specified by the user)
- model specification is given by the user (constant, linear, linear with cross terms, quadratic, etc.)
- criterion functions are sum-squared-error (sse), AIC, BIC, R^2, and adjusted R^2

example: advertising data (y is `sales`, X are `TV`,`radio`,`newspaper`)

```
1. Adding TV, FStat = 312.145, pValue = 1.46739e-42
2. Adding radio, FStat = 546.7388, pValue = 9.776972e-59

Linear regression model:
    sales ~ 1 + TV + radio

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   --------    ---------    ------    ----------

    (Intercept)      2.9211      0.29449    9.9192    4.5656e-19
    TV             0.045755    0.0013904    32.909     5.437e-82
    radio           0.18799      0.00804    23.382     9.777e-59


Number of observations: 200, Error degrees of freedom: 197
Root Mean Squared Error: 1.68
R-squared: 0.897,  Adjusted R-Squared: 0.896
F-statistic vs. constant model: 860, p-value = 4.83e-98
```

MATLAB commands

- common tests are available in many statistical softwares, e.g, `minitab`, `lm` in R, `fitlm` in MATLAB,
- `stepwiselm` perform a step-wise regression

References

G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning: with Application in R*, Springer, 2015

Chapter 4-5 in

W.H. Greene, *Econometric Analysis*, Prentice Hall, 2008

Review of Basic Statistics (online course)

`https://onlinecourses.science.psu.edu/statprogram`

Stat 501 (online course)

https://onlinecourses.science.psu.edu/stat501