Model Selection

Tools so far:
- t-test on individual coefficients
- F-test between particular subset models
- Backward / Forward search

Issues → selection bias

Ultimate validation of a model = prediction.
Does the model generalize to future data?

Usually, parsimonious models / simple models work best for prediction. Why?

Easier to estimate coefficients of a simple model with a moderate sample size. Limit risk of including a spurious relationship that doesn't generalize.

Go for simplicity - also easier to interpret!

Occam's razor

Flexible models	Simple models
- many parameters	- few parameters
- data adaptive	- rigid
- little (or no) bias	- possible bias
- high estimation variance	- smaller variance
- risk of overfitting	- risk of underfitting
bias ↓	bias ↑
variance ↑	variance ↓

Always — Need to check the overall fit of the model

A simple model won't cut it if

- patterns in residuals
- no better than an intercept model
- R^2 too small
- outliers, leverage, ... 5 basic assumptions

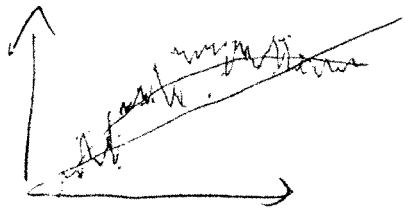
How do we identify a good prediction model?

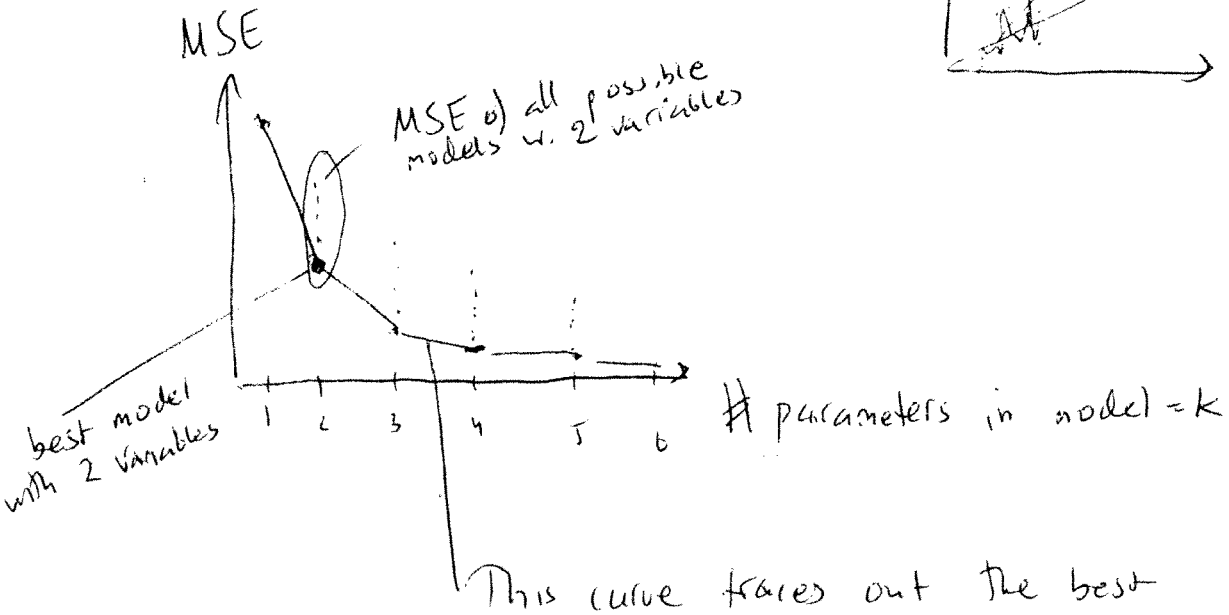
RSS does not work

Why? $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, $MSE = \frac{1}{n} RSS$

always decreases with the complexity of the model.

Note, in extreme case with n variables in the model, RSS = 0 ! Perfect fit





This curve traces out the best models of any size model (k = 1, .., 6)

So, if we used RSS to choose a model ⇒ we end up with the full / largest model.

(4) If we had a separate test data set we could choose model that minimizes the prediction MSE

$$pMSE = \frac{1}{n}\sum_{i=1}^{n}(y_i^{new} - \hat{y}_i)^2$$

Strategy { Training data (Tr) $(x_i, y_i)_{i=1}^n$; Test data (Te) $(x_i, y_i^{new})_{i=1}^n$ } — In practice (could be different x_i's & sample sizes.

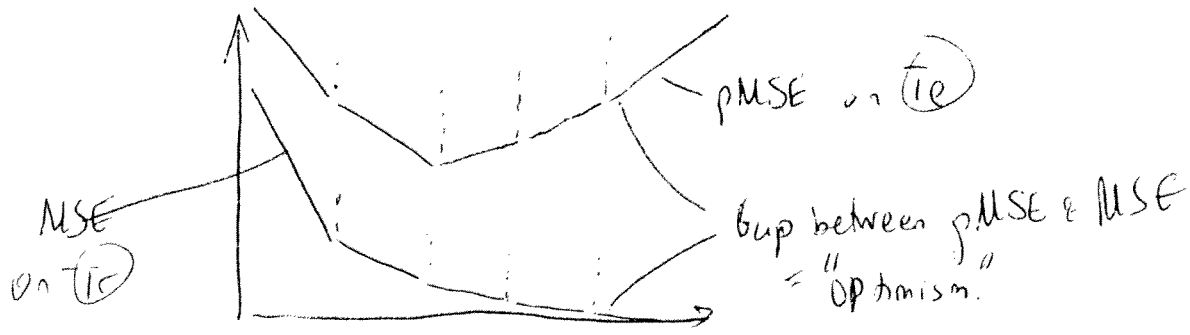
(1) On (Tr) — fit all models of interest; $M_1, M_2, \ldots, M_J$ with coefficients $(\hat{\beta}^1, \hat{\beta}^2, \ldots, \hat{\beta}^J)$

(2) On Te — compute $pMSE^j$ → model M_j predictions

$$pMSE^j = \frac{1}{n}\sum_{i=1}^{n}(y_i^{new} - \hat{y}_i(M_j))^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i^{new} - X_i\hat{\beta}^j)^2$$

(3) Choose the model j^* that minimizes pMSE



(*) $pMSE^j - MSE^j$ = optimism for model j.

That is, because we fit the model on (T) using the LS criterion $\sum_{i=1}^{n}(y_i - \hat{y}_i) = RSS$

So the RSS^j or MSE^j is an underestimate of how well the model would perform on future data.

Now, in practice we don't have an independent test data

$\Rightarrow$ Alternative 1 We use part of the data for training, part for testing

(Cross-validation)

Alternative 2 Estimate the expected size of the gap $pMSE^j - MSE^j$

- or at least an approximation of the gap

(Model selection criteria like AIC, BIC

Let's take a closer look at the gap

Prediction error of model M_j $= PE^j = E_{new}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i^{new} - \hat{y}_i(M_j))^2\right)$

$$= \frac{1}{n}\sum_{i=1}^{n} \underbrace{E_{new}\left(y_i - \hat{y}_i(M_j)\right)^2}_{(*)}$$

$\hat{y}_i(M_j)$: prediction using model M_j where coefficients are estimated from training data $(x_i, y_i)_{i=1}^n$

Notation $\hat{y}_i(M_j) = \hat{y}_i^j$

Expand (*) into 6 terms, adding & subtracting $E(y_i^{new})$ & $E(\hat{y}_i^j)$

$$(*) = E\left(y_i^{new} - E(y_i^{new}) + E(y_i^{new}) - E(\hat{y}_i^j) + E(\hat{y}_i^j) - \hat{y}_i^j\right)^2$$

$$= E\left(y_i^{new} - E(y_i^{new})\right)^2 + E\left(E(y_i^{new}) - E(\hat{y}_i^j)\right)^2 + E\left(E(\hat{y}_i^j) - \hat{y}_i^j\right)^2$$

(1) (2) (3)

$$+2E\left((y_i^{new} - E(y_i^{new}))(E(y_i^{new}) - E(\hat{y}_i^j))\right) + 2E\left((y_i^{new} - E(y_i^{new}))(E(\hat{y}_i^j) - \hat{y}_i^j)\right)$$

(4) (5)

$$+2E\left((E(y_i^{new}) - E(\hat{y}_i^j))(E(\hat{y}_i^j) - \hat{y}_i^j)\right)$$

(6)

Meaning of terms ① – ⑥

① $E(y_i^{new} - E(y_i^{new}))^2 = \sigma^2$ – The irreducible error
– noise around the model

② $E(E(y_i^{new}) - E(\hat{y}_i^j))^2 = bias^2$ of model j at obs. i

($E(y_i^{new})$: true model)

– if model j is adequate the bias is 0

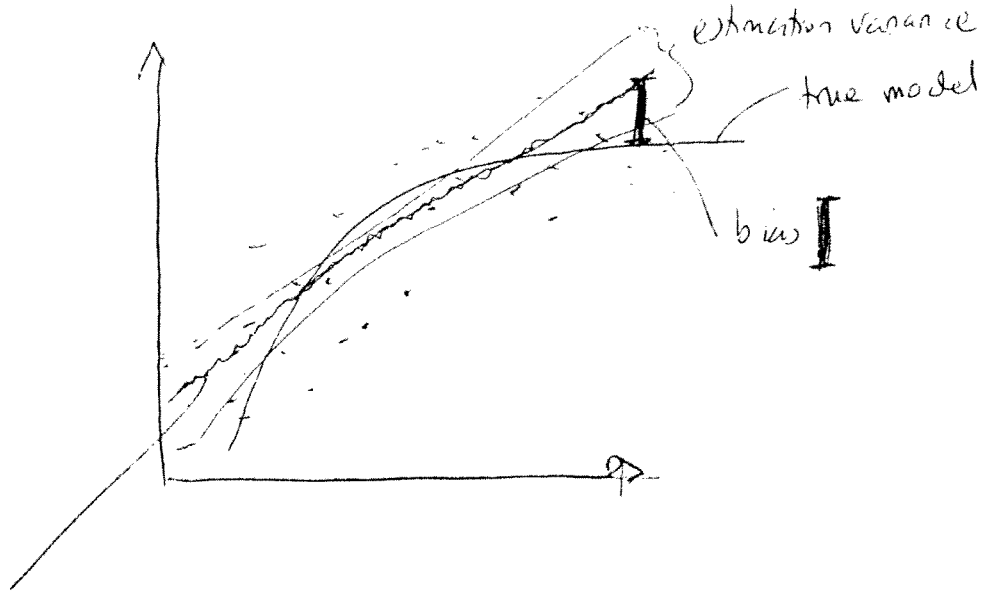
③ $E(E(\hat{y}_i^j) - \hat{y}_i^j)^2 = V(\hat{y}_i^j)$ – estimation variance of model j
– increases with complexity of model.

④ = 0
⑥ = 0
} can pull constant outside

⑤ = 0 since y_i^{new} & y_i are independent data sets

⑥ $PE^j = \frac{1}{n} \sum_{i=1}^{n} \left(\sigma^2 + bias^2(i, M_j) + V(\hat{y}_i^j) \right)$

$= \sigma^2 + \overline{bias^2}(M_j) + \frac{1}{n} \sum_{i=1}^{n} V(\hat{y}_i^j)$



Example: • LS estimate $-\hat{\beta} = (X'X)^{-1}X'y$

- X = $n \times j$ design matrix
- a model with j variables

If model j is adequate → no bias

- $V(\hat{y}) = V(Hy) = \sigma^2 H$
- $V(\hat{y}_i) = \sigma^2 h_{ii}$ → $\frac{1}{n}\sum V(\hat{y}_i) = \frac{\sigma^2}{n}\sum h_{ii}$

↑ leverage

$= \frac{\sigma^2}{n} Tr\{H\}$

- Now, $H = X(X'X)^{-1}X'$

$$Tr\{H\} = Tr\{X(X'X)^{-1}X'\} = Tr\{(X'X)(X'X)^{-1}\} = Tr\{I_j\} = j$$

So $PE_j = \sigma^2\left(1 + \frac{j}{n}\right)$ so increases with j!

parameters in model j

What about the training error?

$$TE^j = E\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^j)^2\right) = \frac{1}{n}\sum_{i=1}^{n} E(y_i - \hat{y}_i^j)^2$$

Like before, we expand $E(y_i - \hat{y}_i^j)^2$ into 6 terms, but now (5) is not 0

$$(5) = 2E\left((y_i - E(y_i))(E(\hat{y}_i^j) - \hat{y}_i^j)\right) = 2E\left(y_i(E(\hat{y}_i^j) - \hat{y}_i^j)\right)$$

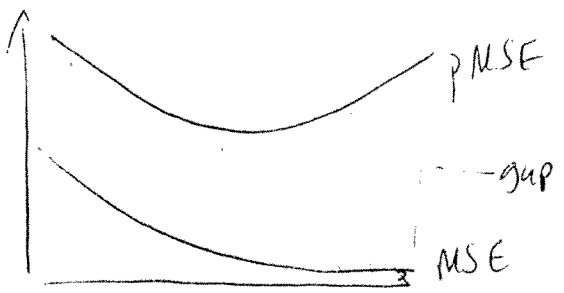
$$= 2E(y_i)E(\hat{y}_i^j) - 2E(y_i\hat{y}_i^j) = -2\,Cov(y_i, \hat{y}_i^j)$$

(So) $$TE^j = PE^j - \frac{2}{n}\sum Cov(y_i, \hat{y}_i^j)$$

What did we learn? Well, $TE < PE$ always

(and) TE much smaller than PE if y_i & $\hat{y}_i$ are highly correlated, which happens for complex models.

That is, the gap between the MSE & pMSE increases with the size of the model.



Example for LS fits $\hat{y} = Hy$ & $Cov(y_i, \hat{y}_i) = \sigma^2 h_{ii}$

$$\frac{2}{n}\sum Cov(y_i, \hat{y}_i) = 2\frac{\sigma^2}{n} \cdot j$$

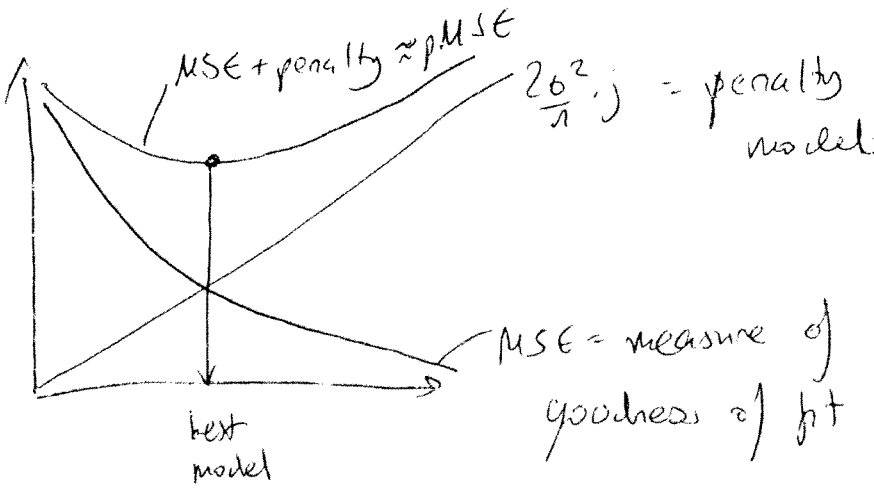
# of parameters in model j.

ie. gap grows linearly with model size.

What's the point?

Alternative 2

- We don't have an independent test set so we can't estimate pMSE, but we need it for selection a good prediction model
- but we have MSE for each model and we know the expected size of the gap between pMSE & MSE $\Rightarrow 2\frac{\sigma^2}{n} \cdot j$
- Approximate $pMSE^j$ by $\boxed{MSE^j + 2\frac{\sigma^2}{n} \cdot j}$



Many model selection criteria have this form

Goodness-of-fit measure + penalty on model size

The one based on the expected gap size

$$MSE + 2\frac{\sigma^2}{n} * (\text{number of parameters})$$

is called **Mallow's C_p**, which you also see as $RSS + 2\sigma^2 \cdot (\text{number of parameters})$ in the texts

To use we need to plug in $\hat{\sigma}^2$ in the penalty. We use the same $\hat{\sigma}^2$ for all models — the $\hat{\sigma}^2$ from an unbiased model (usually the largest model we fit)

Strategy

(1) Enumerate all models $M_1, \ldots, M_J$ $(j = 1, \ldots, J)$

(2) Evaluate MSE_j, RSS_j for all models

(3) Compute $C_p^j = RSS_j + 2\hat{\sigma}^2 \cdot j$, where

$\hat{\sigma}^2 = MSE^J$ (largest model)

(4) Pick model with smallest C_p.

Other model selection criteria

$$\begin{cases} AIC(M_j) = n \log(RSS^j) + 2 \cdot j \\ BIC(M_j) = n \cdot \log(RSS^j) + j \cdot \log(n) \end{cases}$$

Goodness of fit measure

penalty on model size

Note, AIC & BIC measure goodness of fit on the likelihood scale (normal error distribution)

Also, BIC penalty is larger than AIC or C_p so BIC tends to pick smaller models.

In practice, AIC & C_p tend to <u>overfit</u> the data.

A note on PE; $PE = \sigma^2 + \overline{bias^2} + Var(\text{model estimate})$

Irreducible

Trade-off

Perhaps better to sacrifice some bias if we can lower variance?

$\Rightarrow$ Ridge regression ...