Credibility, penalised regression and boosting; let's call the whole thing off

October 2011

Introduction

- There are a variety of contexts in which applying the full effect observed in historical data is undesirable. For example:
  - ✓ Low exposures associated with observation
  - ✓ A large past effect is not expected to continue in the future
  - ✓ The analyst wants to bias the model towards “standardised” predictions, where observations are shrunk back towards the mean
- The first of these is most relevant in the context of today’s talk.
- Three common ways of achieving this desired result are:

Credibility models

Penalised regression models

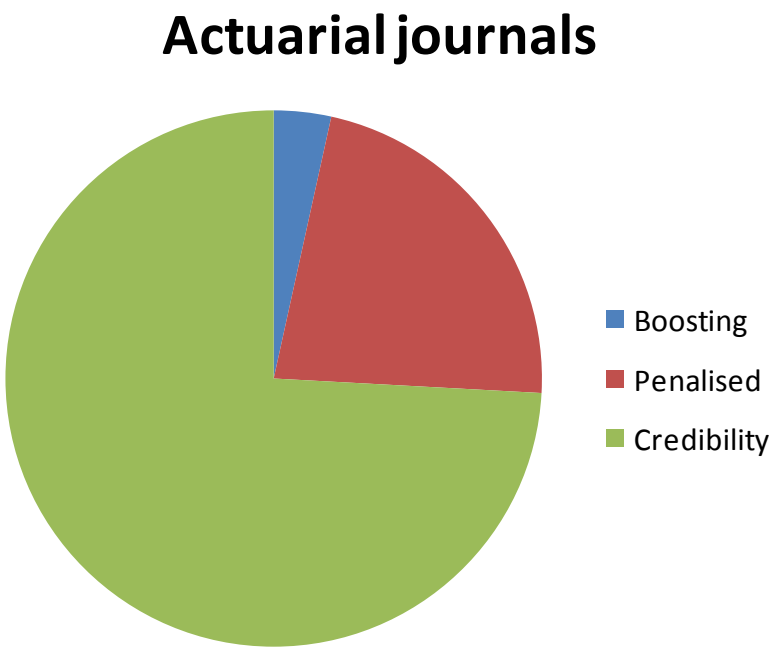
Boosted models

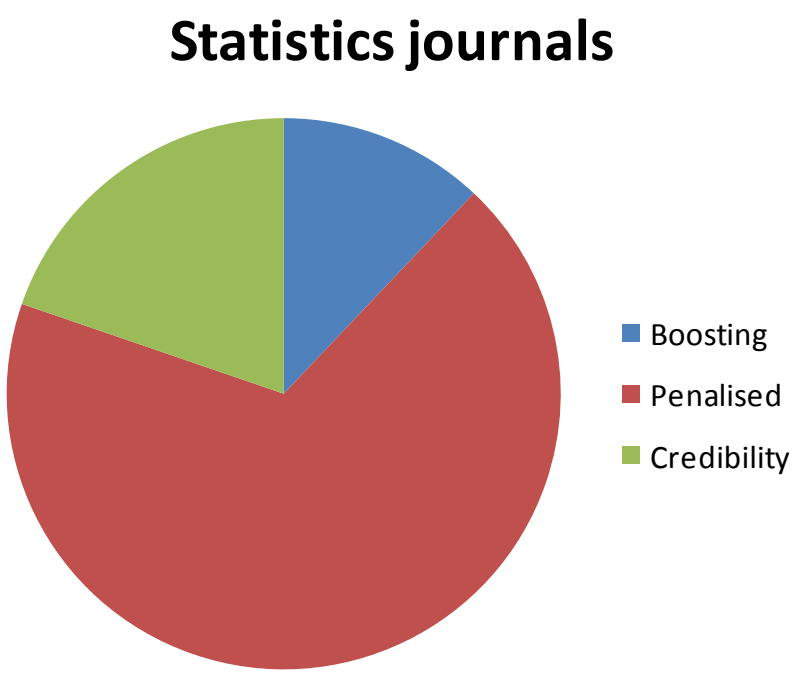
Introduction

- Today's talk will:
  - ✓ Provide a (hopefully gentle) introduction to these topics
  - ✓ Highlight some similarities and differences
  - ✓ Point out how some of these approaches are currently being used

- There will also be a few connoisseur slides, indicated by the picture to the right, as asides for people interested in some more detailed aspects of the talk

Some perspective: proportion of journal papers mentioning various types of models





Basic setup

- Suppose we have a series of responses Y_1, Y_2, Y_3, … and predictor vectors X_1, X_2, X_3, …, with $X_i = \{X_{i1}, X_{i2}, \ldots, X_{ip}\}$, and we observe the first n predictors and responses. We want to build a linear model on the predictors to minimise future predictions of the response. That is, choose $\beta = (\beta_1, \ldots, \beta_p)$ to find $\hat{Y}_i = \mu + \beta^T X_i$ such that the expected value of

$$(Y_i - \hat{Y}_i)^2$$

  is minimised for future observations where we know X_i but not Y_i

- This is the usual ordinary least squares linear model framework. We will not discuss extensions to GLMs and the like today, although they all exist.

Specific example

- Bob has collected some house price data for a suburb:
  - ✓ The suburb mean is $500k
  - ✓ 500 observations
  - ✓ Block of sale also retained – 50 different blocks on dataset (some with only 1 sale, one with 114 sales)
  - ✓ Sales hover around the mean with some random variability (error), and possibly some block-dependent variation
- Bobby wants to know whether his predictions can be improved by allowing for the block of a house. He also has another 500 observations for validating a model.

- Our main measure of model error will be root mean squared error

$$\left\{ \frac{1}{n} \sum_i (Y_i - \hat{Y}_i) \right\}^{1/2}$$

- We'll drop the '000s for the rest of the talk

In terms of our earlier notation, we have 50 binary predictors

Observa-tion	Price ($k)	Block
1	532	#1
2	581	#1
3	543	#1
4	482	#2
…	…	…

- Credibility models
- Penalised regression models
- Boosted models

Credibility models

- Bayesian approach, where prior assumptions are made regarding the distribution of random variability, as well as block variability.
- It is common in the above setup to assume

$$Y_i = \mu + R_{K(i)} + \epsilon_i$$

where *K(i)* is the block that observation *i* belongs to

- ✓ Mean is known to be 500
- ✓ Normally distributed block effects: $R_k \sim N(0, \tau^2)$
- ✓ Normally distributed errors: $\epsilon_i \sim N(0, \sigma^2)$

- For a given block k, we have a **prior** on the block effect:

$$f(r) = (2\pi)^{-0.5} \exp\left\{-r^2/(2\tau^2)\right\}$$

- And a **likelihood** associated with the observations in that block:

$$f(y|r_k) = \prod_{i;K(i)=k} (2\pi)^{-0.5} \exp\left\{-(y_i - \mu - r_k)^2/(2\sigma^2)\right\}$$

- **Bayes Theorem** says that the posterior distribution for r_k is proportional to the prior times likelihood. We can then solve to obtain Bayesian estimates:

$$\hat{r}_k = \frac{n_k}{n_k + \sigma^2/\tau^2}(\bar{Y}_k - \mu)$$

Here $\bar{Y}_k$ is the average for that block, and n_k is the number of sales in that block. We recognise this as a portion of the original observed effect, $(\bar{Y}_k - \mu)$

- We also need to estimate σ and *τ*. There are generally reasonable ways to find these. In this example, the standard formulae are:

$$\hat{\sigma}^2 = \frac{\sum_i (Y_i - \bar{Y}_{K(i)})^2}{n - \kappa}$$

where κ is the number of blocks, and

$$\hat{\tau}^2 = \frac{\sum_k (\bar{Y}_k - \bar{Y})^2 - (\kappa - 1)\hat{\sigma}^2}{n - n^{-1}\sum_k n_k^2}$$

A more modern approach to estimating variances is to consider the problem as one of random effects in a mixed model.

- ✓ Variances estimated simultaneously with block effects via restricted maximum likelihood.
- ✓ Variances guaranteed to be positive
- ✓ Quickest way to a credibility model in SAS, using proc mixed or proc glmmix

How does our credibility model perform?

- Our house price dataset was generated exactly as described in the Bayesian model setup. Here σ was estimated as 29.2 (true answer is 30), and τ was 18.6 (true answer 20).

Model	Train RMSE	Test RMSE
Constant	34.4	34.8
Credibility	28.2	31.6
OLS	27.7	32.0

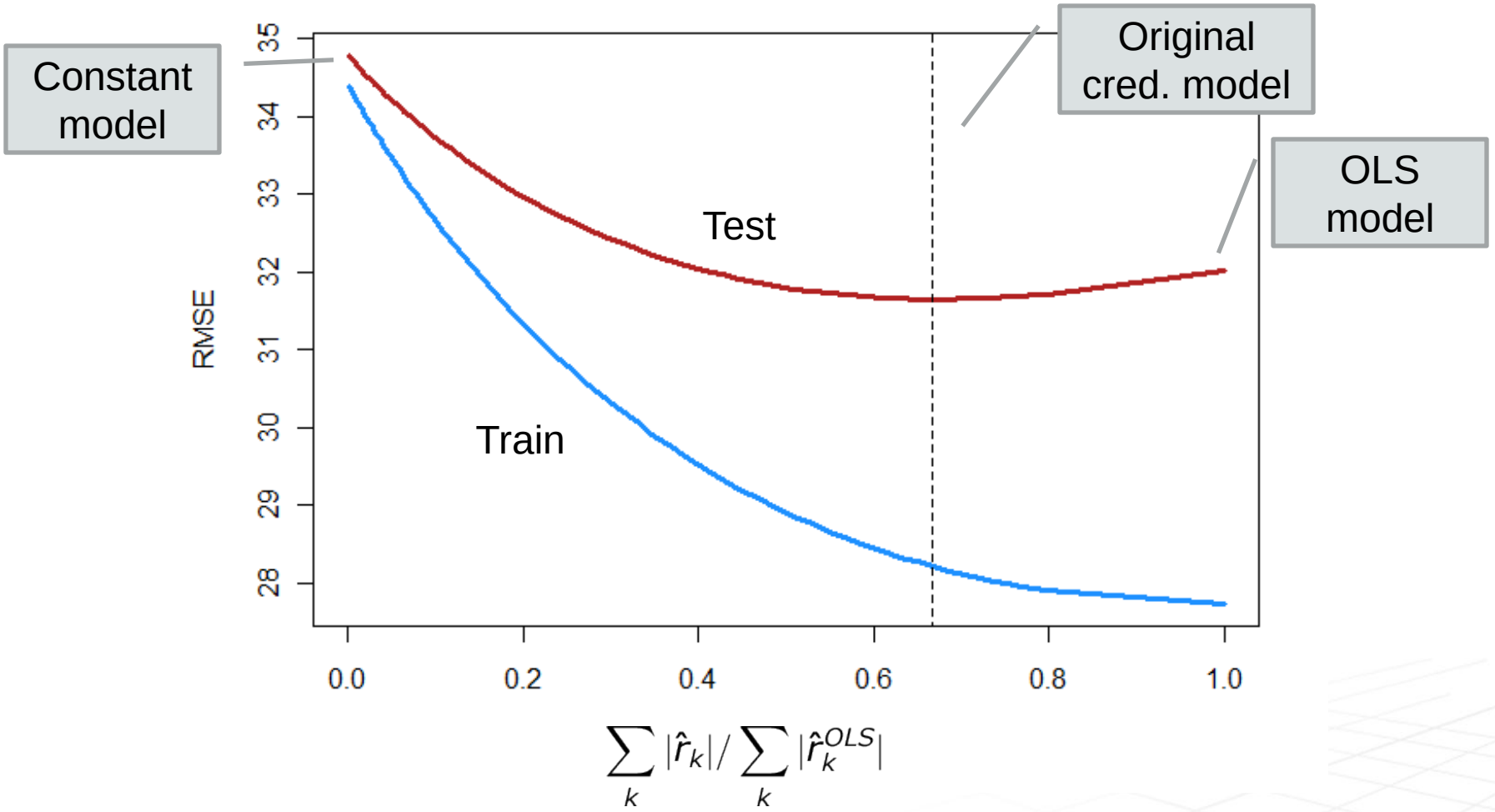
Bayesian thought experiment

- What if instead of estimating *τ,* we just guessed it?
- We could see how it varied in test dataset accuracy for different choices

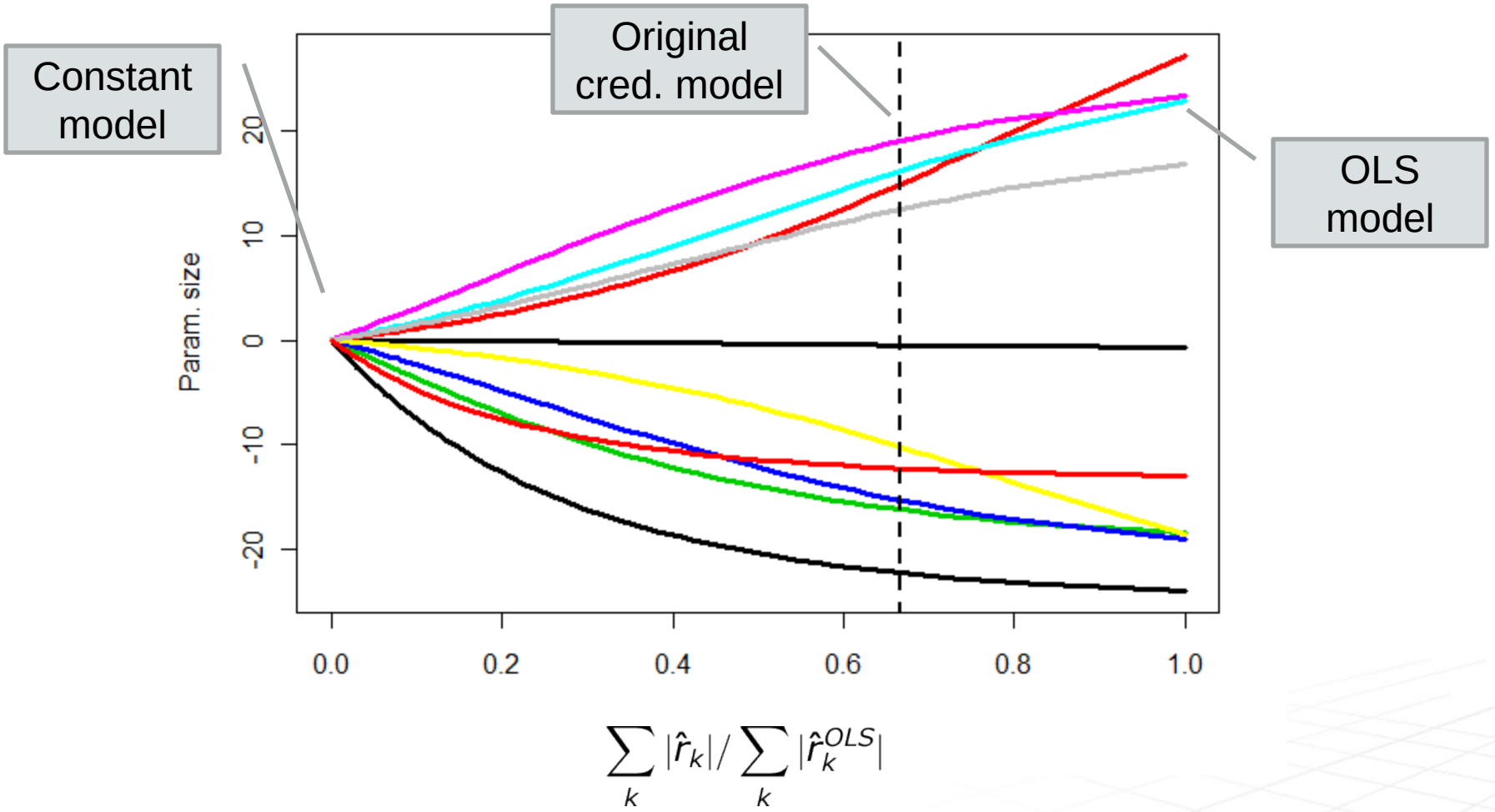
$$\hat{r}_k = \frac{n_k}{n_k + \sigma^2/\tau^2}(\bar{Y}_k - \mu)$$

- For large values, the model would infer that most of the observed variation is genuine signal. $\tau = \infty$ corresponds to the OLS model
- Small values attribute more variation to noise. $\tau = 0$ gives the constant model

Train and test error for various τ



Parameter evolution for various τ, first 10 blocks



Credibility models

Penalised regression models

Boosted models

Penalised regression models

- Recall our initial setup where we seek $\hat{Y}_i = \hat{\mu} + \hat{\beta}^T X_i$, that provides a good approximation for Y_i.
- Applying this notation to our specific example, we have 50 binary predictors:

Observation #	Price ($k) ($Y_i$)	Block 1 flag (X_{i1})	Block 2 flag (X_{i2})	Block 3 flag (X_{i3})	…
1	532	1	0	0	…
2	581	1	0	0	…
3	543	1	0	0	…
4	482	0	1	0	…
…	…	…	…	…	…

- The OLS, or unpenalised solution is to choose parameter values to minimise the objective function; the average sum of squares on the training data.

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i-(\mu+\beta^T X_i)\right\}^2$$

- The penalised regression adds a penalty function that increases the objective function when parameter values are undesirable – most commonly if they are large.

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i-(\mu+\beta^T X_i)\right\}^2+J(\beta)$$

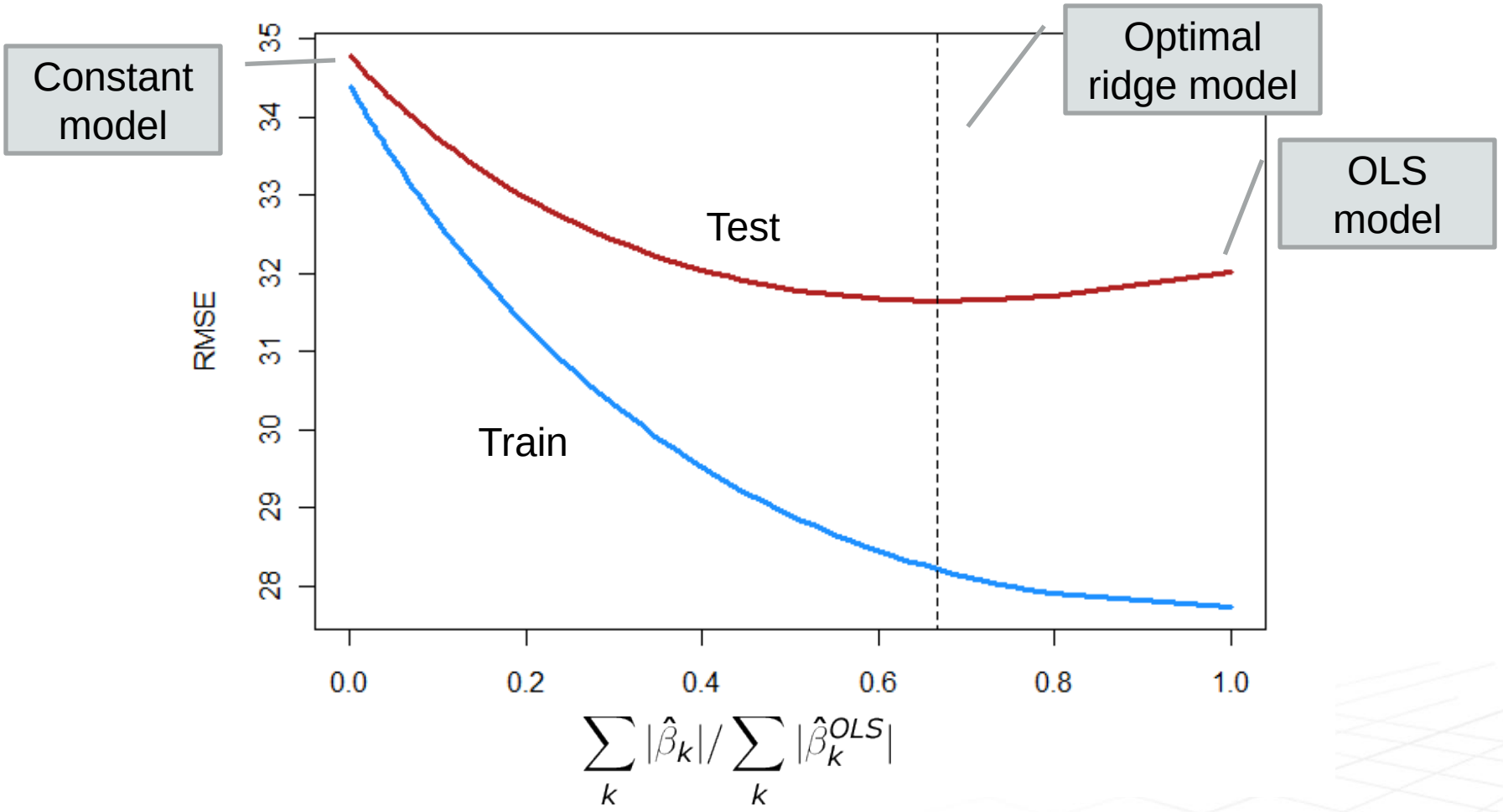
- One of the earliest choices of penalty is the sum of squared parameters, which is called **ridge regression:**

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i-(\mu+\beta^T X_i)\right\}^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

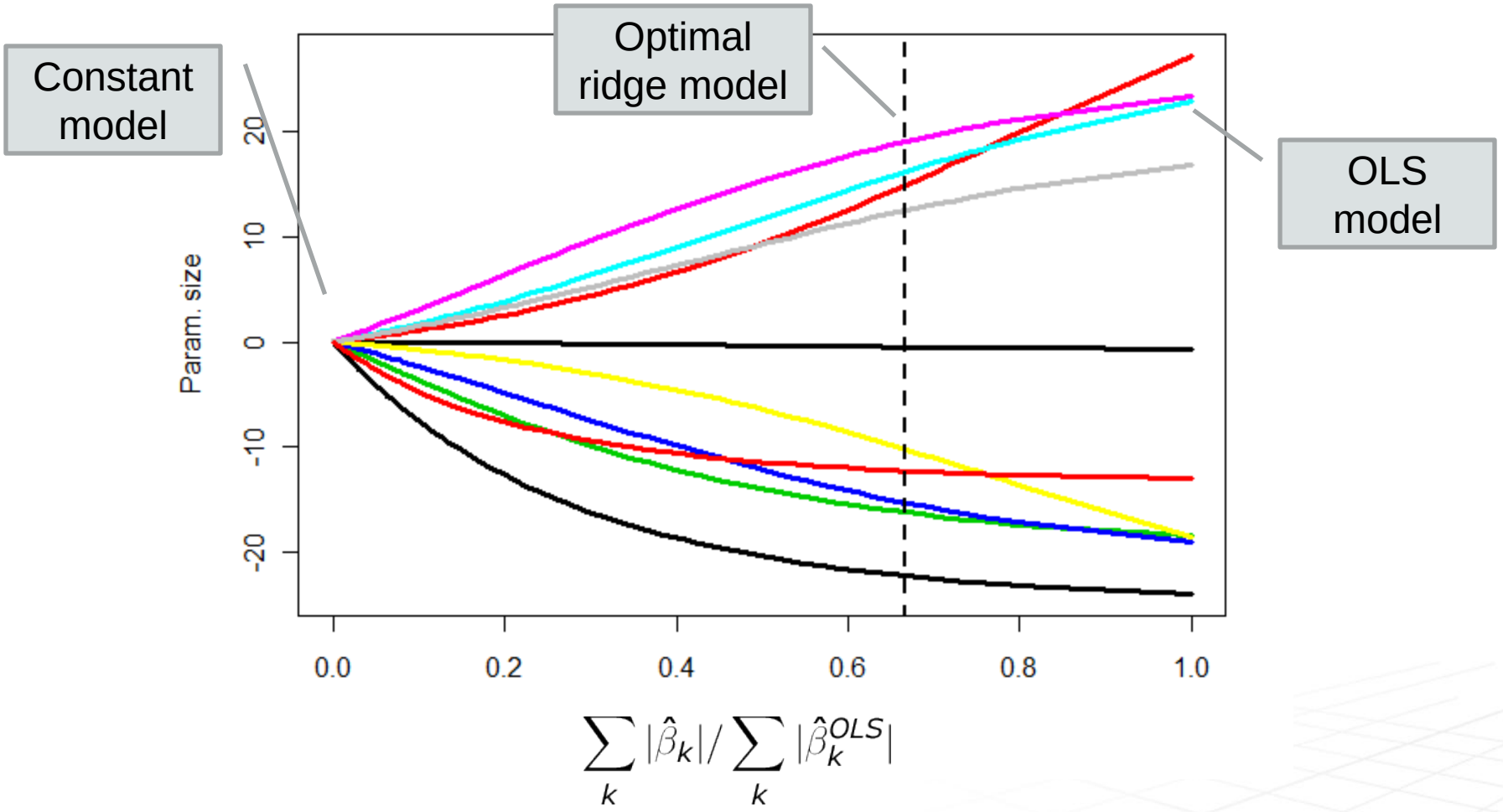
- The parameter λ is called a tuning parameter, and tells the objective function how much penalty is associated with large parameter values.
  - ✓ Usually found via test data performance or cross-validation
  - ✓ Having λ = 0 gives the OLS solution, λ= ∞ is the constant model

- It is usual to scale variables before applying these types of penalisations

Train and test error for various λ, ridge regression

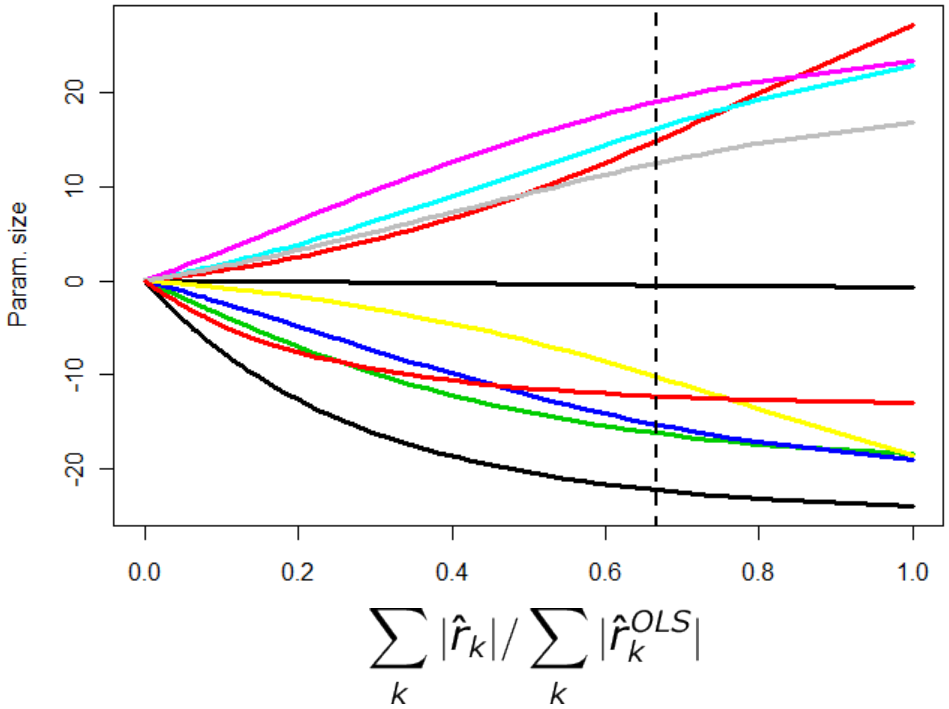


Parameter evolution for various λ, first 10 blocks

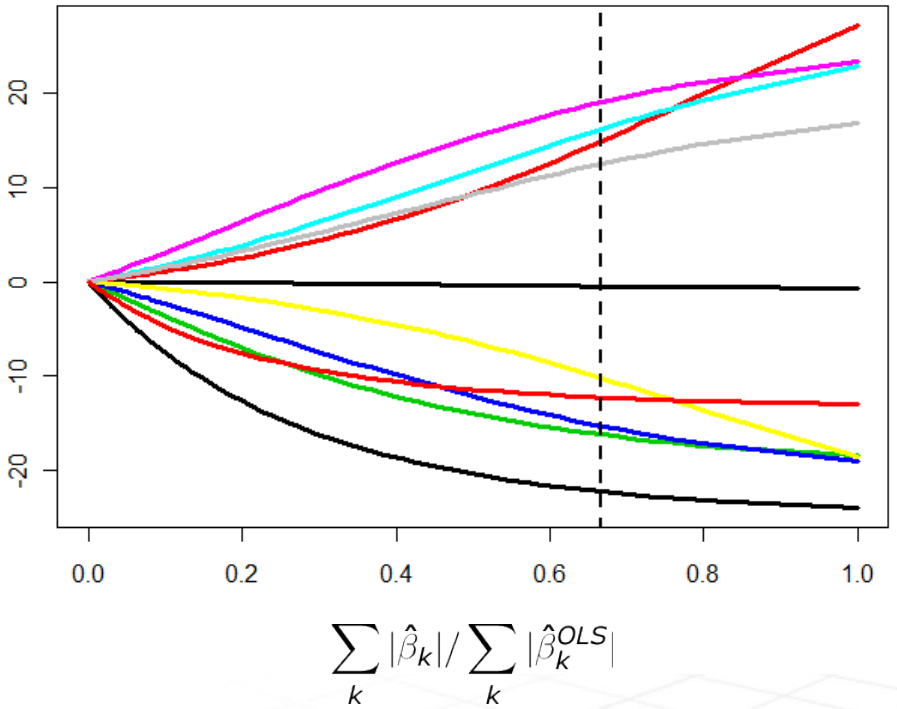


Wait a minute…

Credibility model



Ridge regression model



Result 1

Bayesian credibility with normal priors on error and random effects is equivalent to a penalised ridge regression model

Why is it so?

- Recall we have a posterior for the r_k that looks something like:

$$\prod_i (2\pi)^{-0.5} \exp\left\{-(y_i - \mu - r_{K(i)})^2/(2\sigma^2)\right\} \prod_k (2\pi)^{-0.5} \exp\left\{-(r_k)^2/(2\tau^2)\right\}$$

- We would normally want to find the mean of this distribution. However, we can recognise this as a multivariate normal distribution in the r_k, so finding the mean is equivalent to finding the mode (or maximising the posterior). We can instead maximise after taking logs and adding/multiplying scalars:

$$-\left\{\sum_i \{(y_i - \mu - r_{K(i)})^2/\sigma^2 + \sum_k r_k^2/\tau^2\right\}$$

which is now equivalent to the ridge regression problem

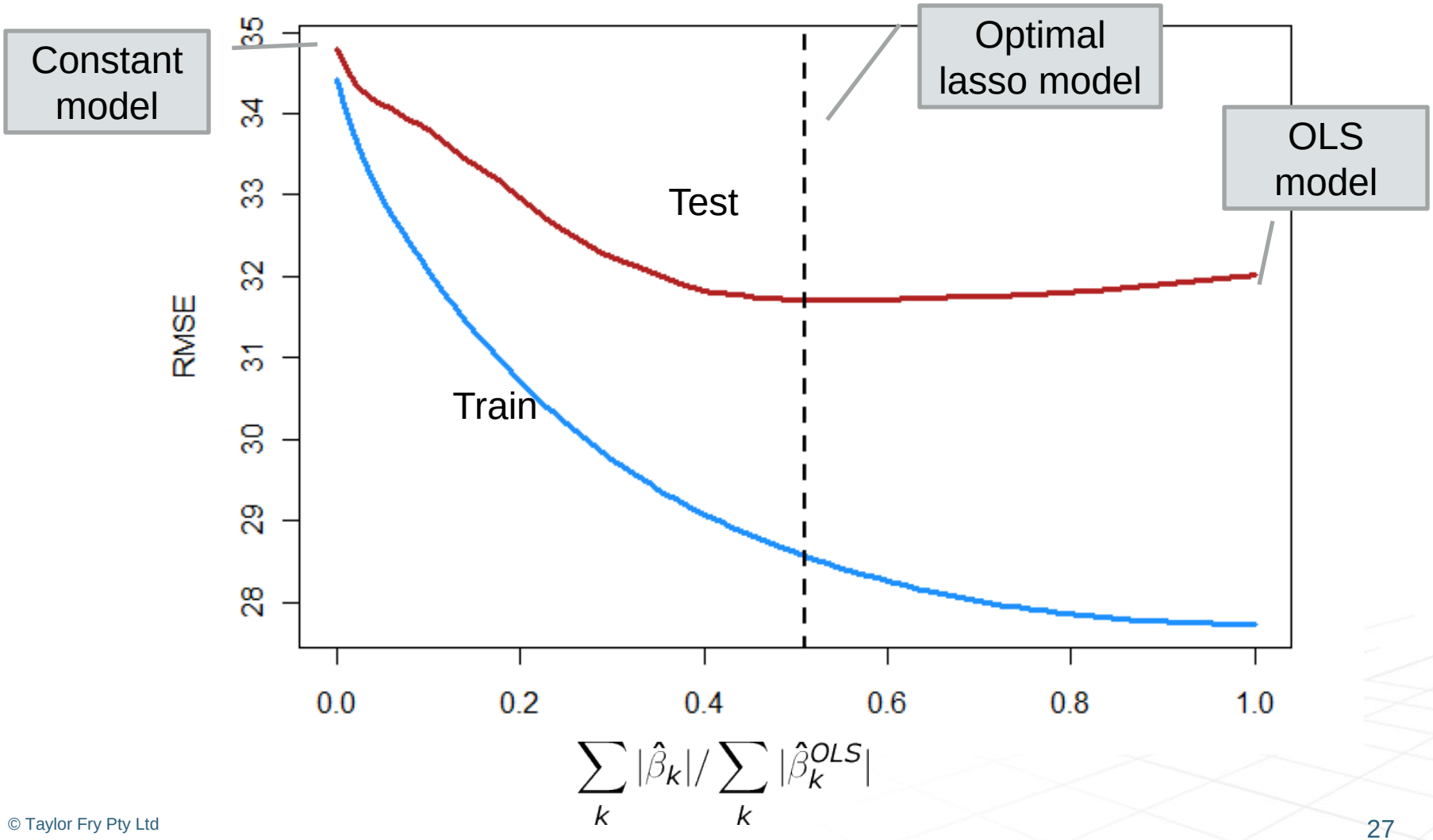
- Another very popular choice of penalty uses absolute values instead of squares, called the **lasso**:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i-(\mu+\beta^T X_i)\right\}^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

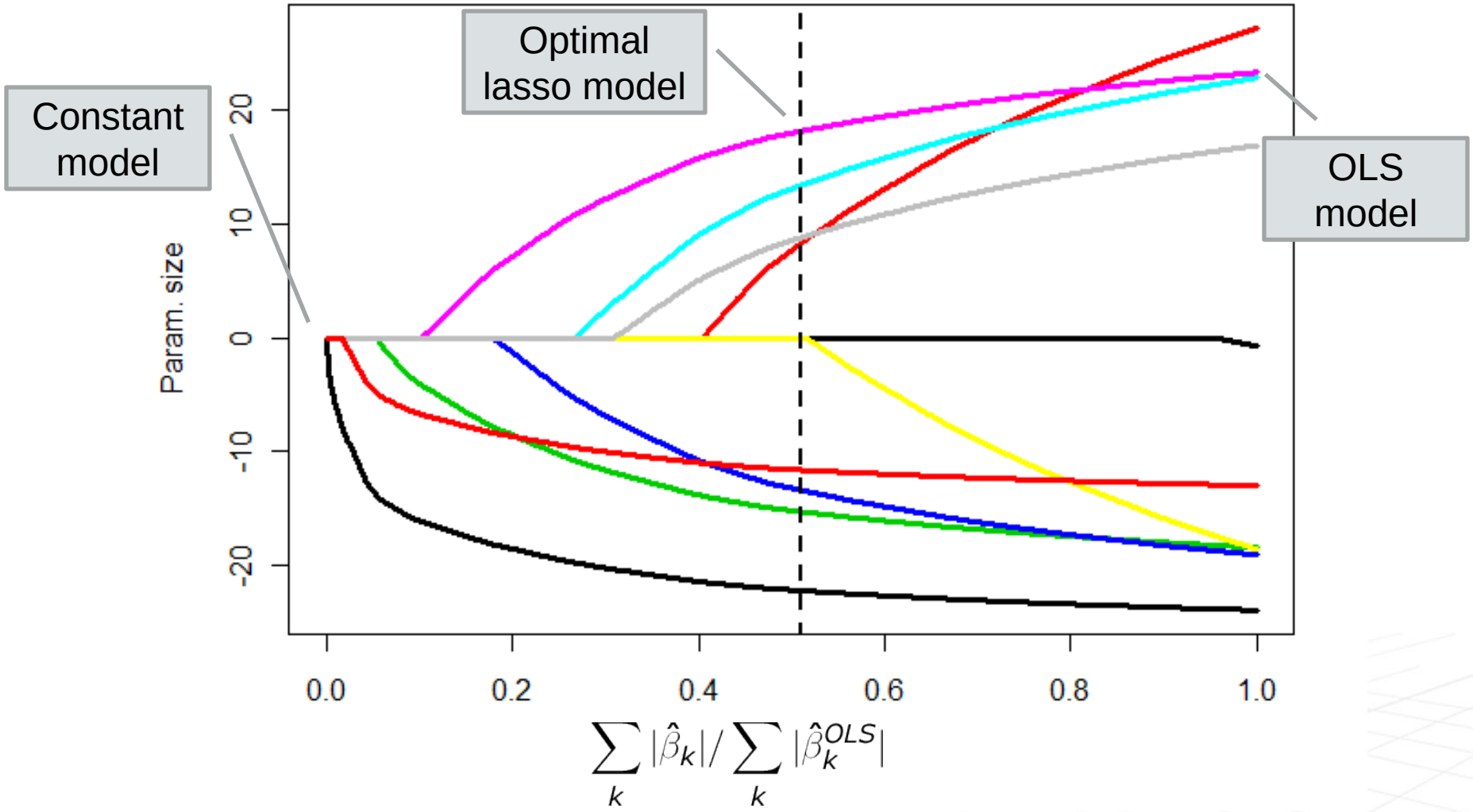
- The parameter λ is called a tuning parameter, and tells the objective function how much penalty is associated with large parameter values.
  - ✓ Usually found via test data performance or cross-validation
  - ✓ Having λ = 0 gives the OLS solution, λ= ∞ is the constant model

- The lasso has the property of producing “sparse” models, where some of the parameters will be exactly zero, while others nonzero. This is in contrast this with ridge regression.

Train and test error for various λ, lasso regression



Parameter evolution for various λ, first 10 blocks



Question

Does the lasso have an equivalent Bayesian formulation?

Result 2

The lasso model is equivalent to a Bayesian model with **normal error**, a **double exponential distribution on parameters**, and where **the mode of the posterior** is used to choose parameters (rather than the mean).

In general, **almost any penalty** you can think of can be reformulated as a Bayesian problem; the two are basically equivalent.

Other considerations

- ✓ Variable scaling
- ✓ Splines, etc
- ✓ Correlations between variables
- ✓ What penalty is best?

Credibility models

Penalised regression models

Boosted models

Boosting

- Relatively new statistical idea, quite popular
- Key idea is to incrementally add weak models to produce a strong final model
- Treenet is one of the most popular approaches; boosting with decision trees

General boosting algorithm

1. Start with $f_0(x) = \mu$
2. For m = 1, 2, … , M
   a. Fit (simple) model $g_m(x)$ on predictors, targeting residuals

$$r_i = Y_i - f_{i-1}(X_i)$$

   b. Set $f_m(x) = f_{m-1}(x) + \delta g_m(x)$

3. Choose best performing $f_m(x)$ (based on test data performance)

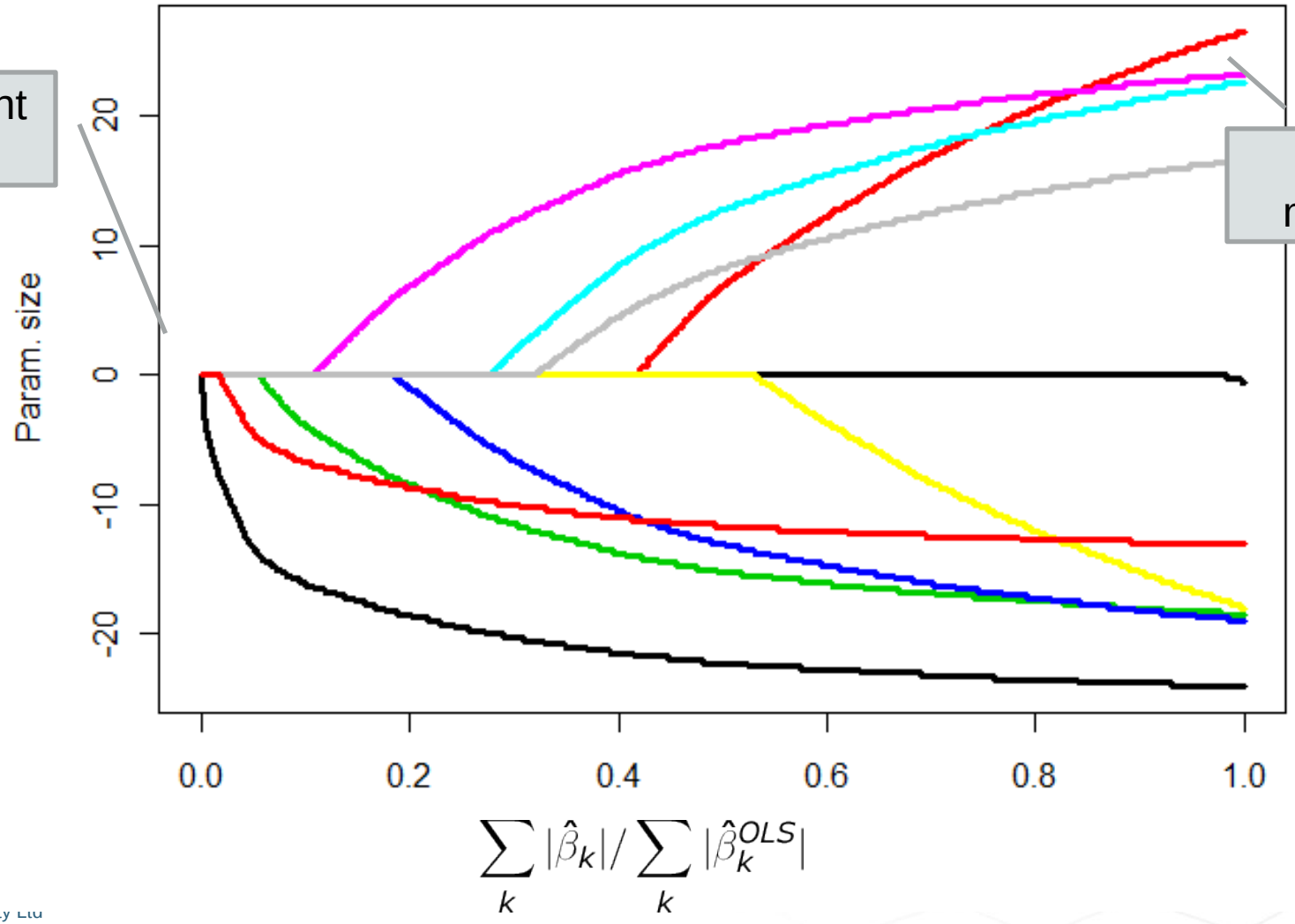
Note that we have not specified how to choose $g_m(x)$. Also, $0 < \delta \leq 1$ is called the learn rate and controls how much of each simple model is added to the main one.

Example – L_0 boosting algorithm

1. Start with $f_0(x) = \mu$
2. For m = 1, 2, … , M
   a. Find j with maximal score $\left|\sum_i x_{ij} r_i\right|$. Set
   $$g_m(X_i) = sign(\sum_i x_{ij} r_i) x_{ij}$$
   b. Set $f_m(x) = f_{m-1}(x) + \delta g_m(x)$
3. Choose best performing $f_m(x)$ (based on test data performance)

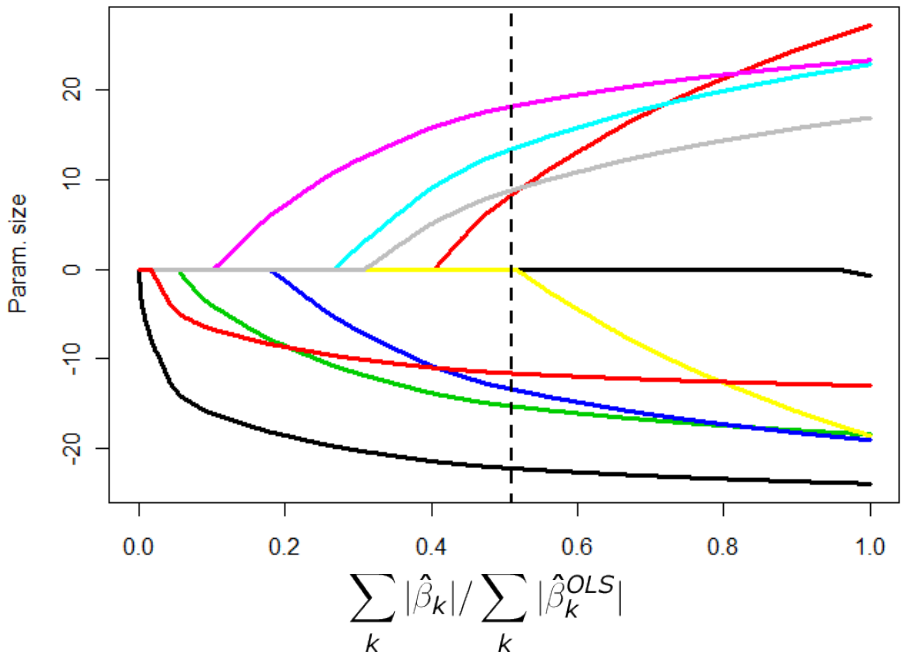
Run this until we get close to the OLS solution.

Parameter evolution for various λ, first 10 blocks

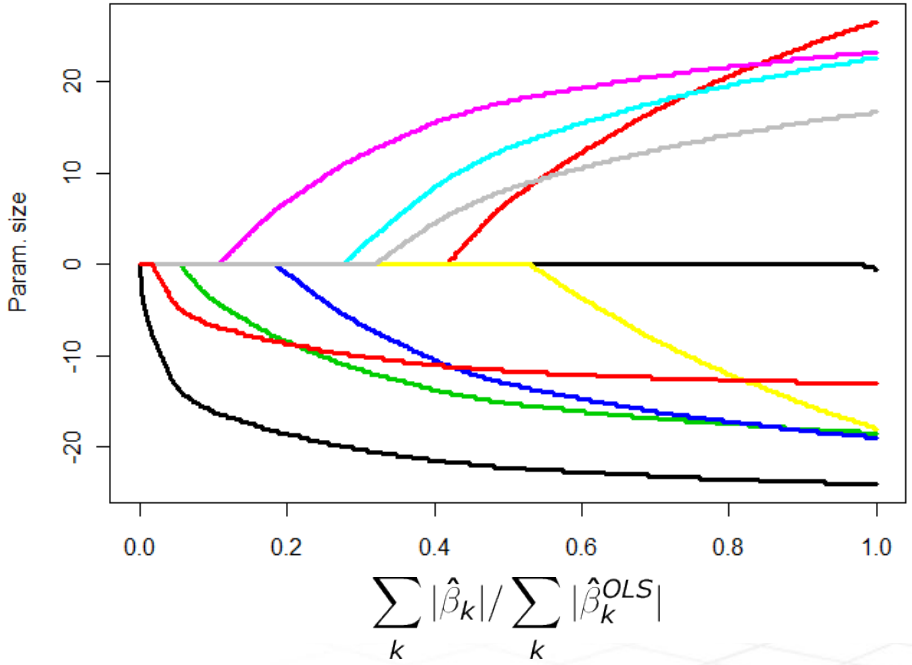


Wait a minute…

Lasso model



Boosted model



Result 3

The lasso model is equivalent to L_0 boosting, when δ is close enough to zero.

In general, there is a boosting algorithm corresponding to **any** chosen penalisation

Summary

Credibility

=

Penalised regression

=

Boosted models

Should we just get rid of two of them?

Probably not - the different frameworks do have relative strengths:

- Credibility produces an explicit model choice
- Credibility allows a user to impose prior beliefs
- Penalisation and boosting produce model spectrums, and are targeted towards predictive accuracy
- Huge variety in penalties exist, often significantly outperforming ridge or lasso
- Boosting allows the possibility from choosing from an infinite array of weak models $g_m(x)$. It can also be simplest to implement.

Some other thoughts

- When the mean is not enough
- Hierarchical models
- Approaches in use at Taylor Fry

When the mean is not enough

Our example today we sought to build a model of the form:

$$\hat{Y}_i = \hat{\mu} + \hat{\beta}^T X_i$$

What if Bob wanted to replace the mean with a standard model (e.g. incorporating house size, number of bathrooms etc).

This is very straight forward; we can build this model first, and then solve a modified version of the above problem:

$$\hat{Y}_i = \hat{\mu}_i + \hat{\beta}^T X_i$$

Hierarchical models

Often multiple effects are to modelled instead of just one. These are often "nested" in a hierarchy

→ For example, if we wanted to model suburb and block on a larger dataset.

These are generally easy to allow for:

- Hierarchical credibility: Have a suburb and block factor to add on, finding credibility expressions for each. Need an extra variance parameter
- Penalised regression: Add a second tuning parameter for suburb

Approaches in use at Taylor Fry

Jobs:

- Crediblity used routinely
- Treenet used routinely
- See other TF presentation at this seminar

Capabilities:

- SAS has some features built in
- Custom-built SAS macros
- Treenet
- R