Advanced Predictive Modeling Workshop

Regression Methods

27 March 2017
San Diego, CA

Mark Jones ACAS, MAAA

mark.j@pwc.com

Agenda

1. Bias-Variance Tradeoff
2. Generalized Additive Models
3. Hierarchical Models

Bias-Variance Tradeoff

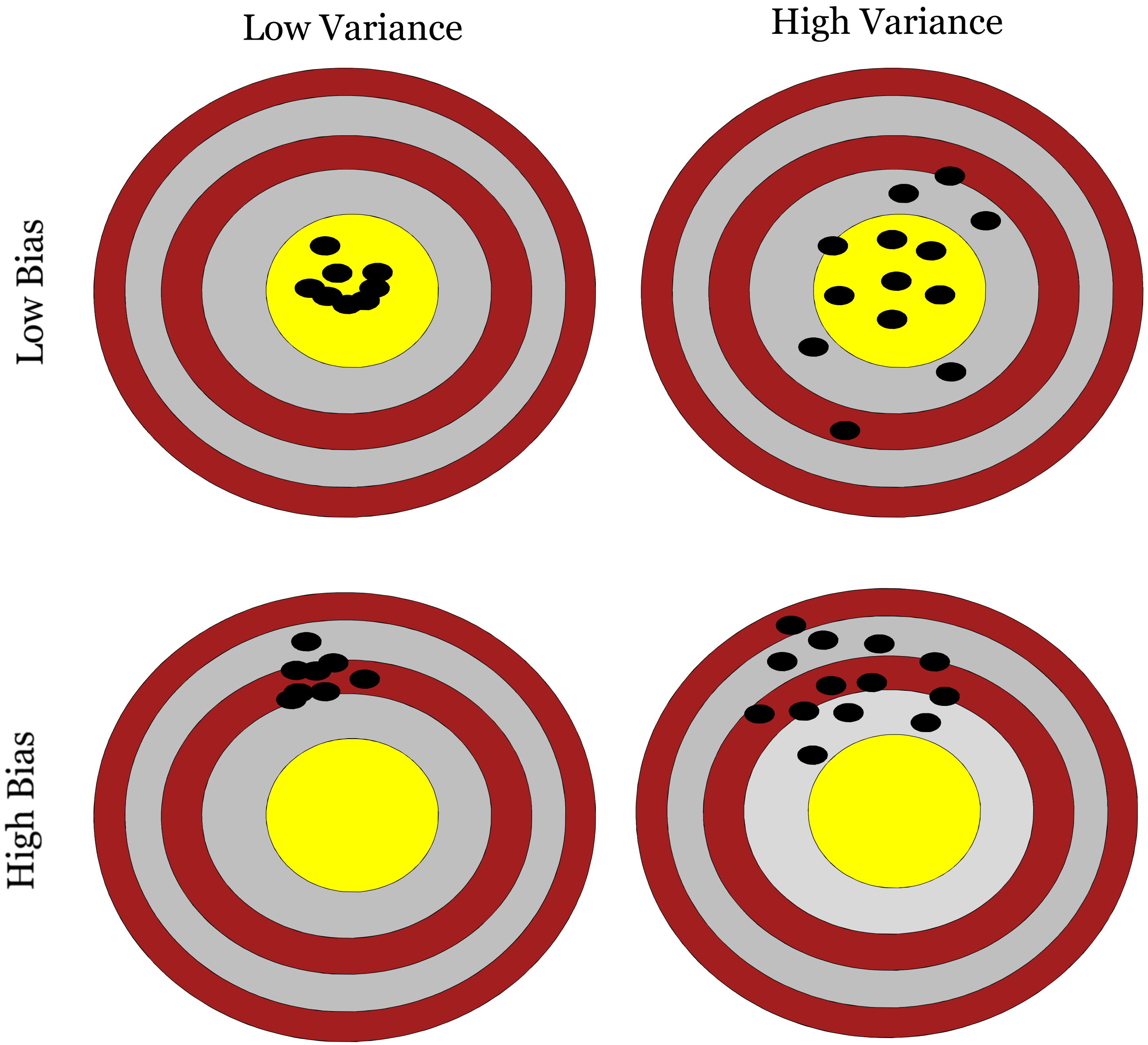
Bias-Variance Tradeoff

- Bias Error – Difference between expected model prediction and true value.

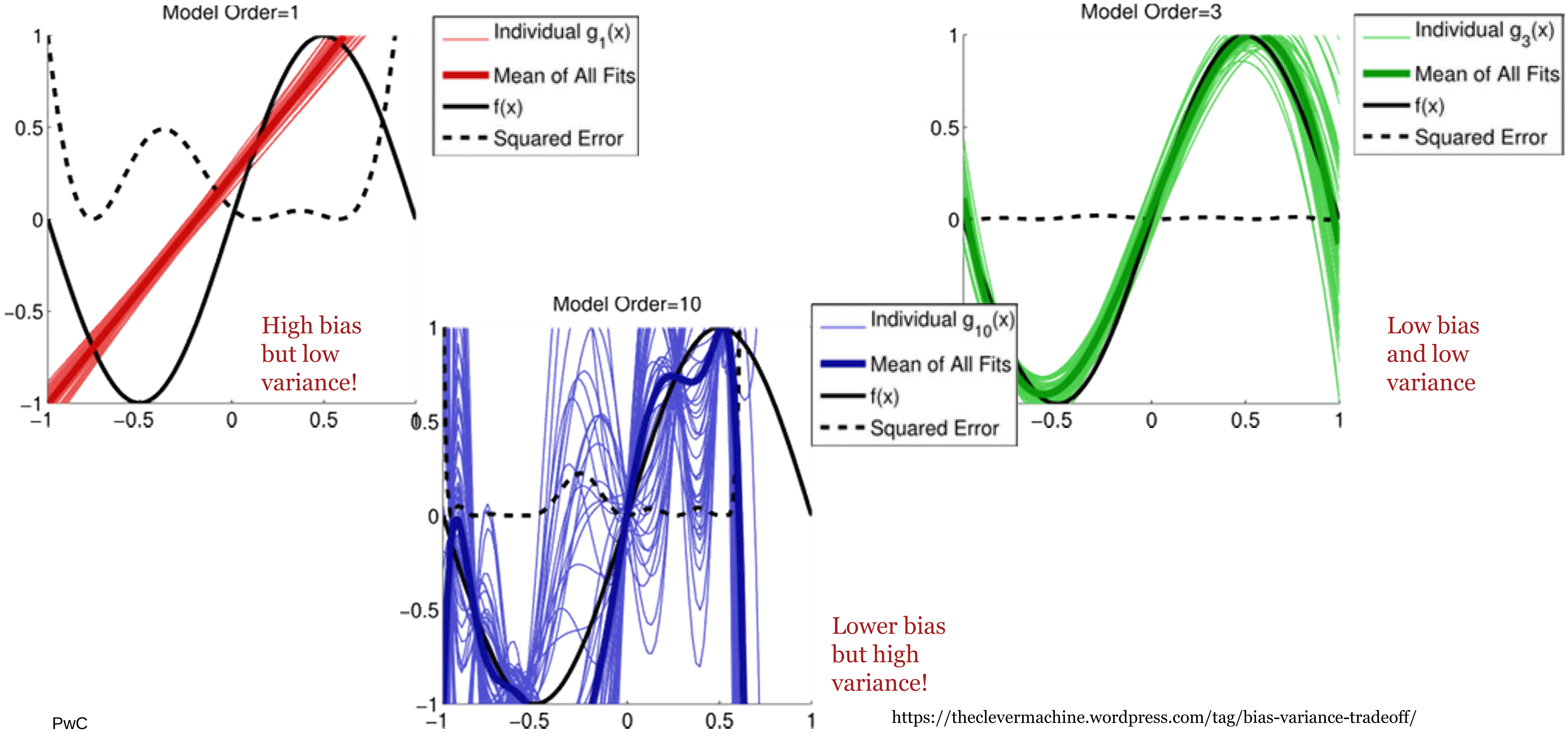
- Variance Error – Difference due to variability in model prediction.

- $Error = Bias^2 + Variance + Noise$

- Noise is that portion of error that cannot be resolved by model.



Bias-Variance Tradeoff



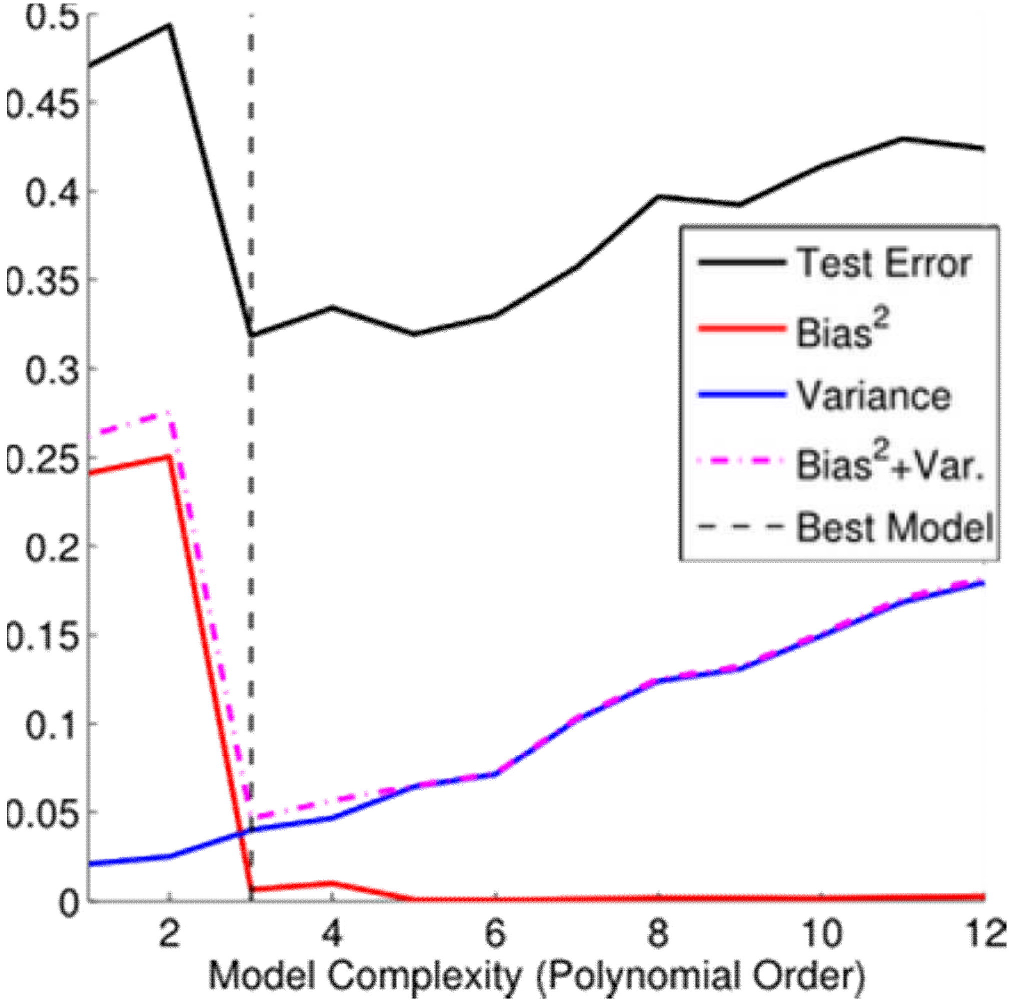
High bias but low variance!

Low bias and low variance

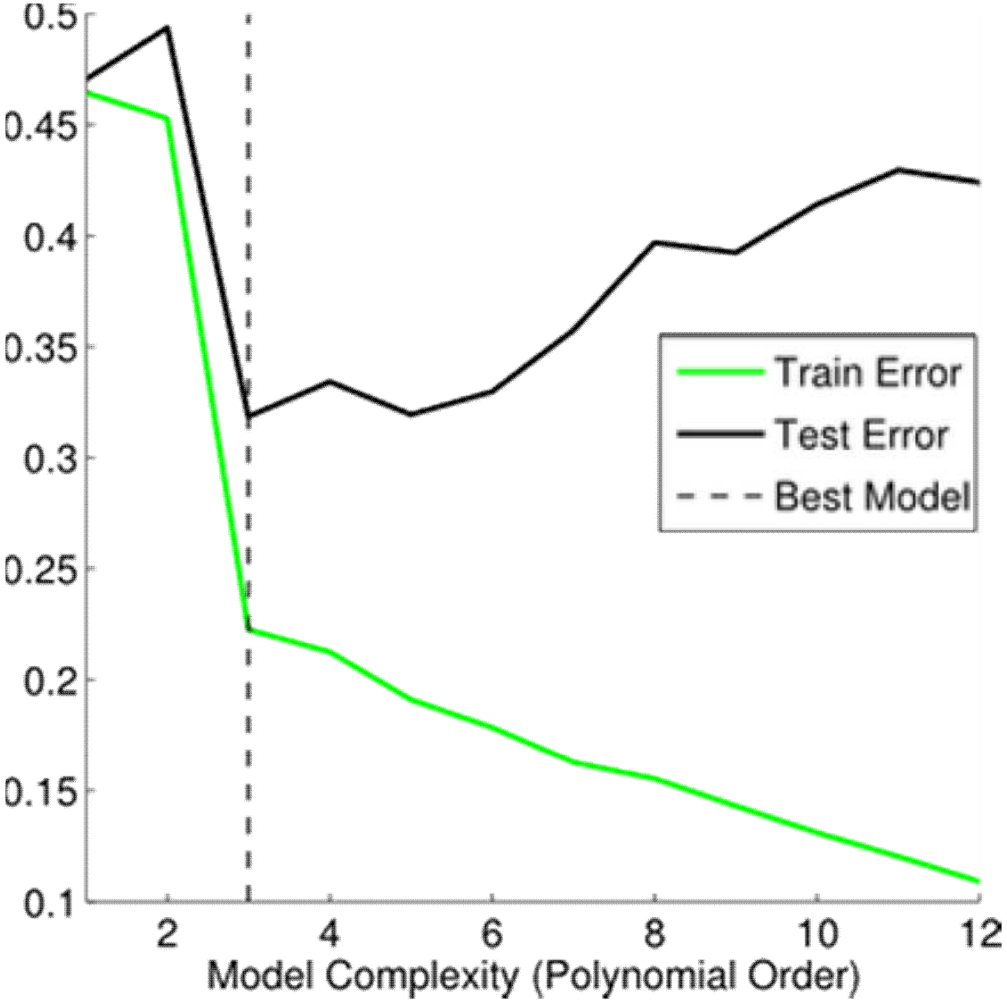
Lower bias but high variance!

https://theclevermachine.wordpress.com/tag/bias-variance-tradeoff/

Bias-Variance Tradeoff



Increasing model complexity leads to lower bias and higher variance.

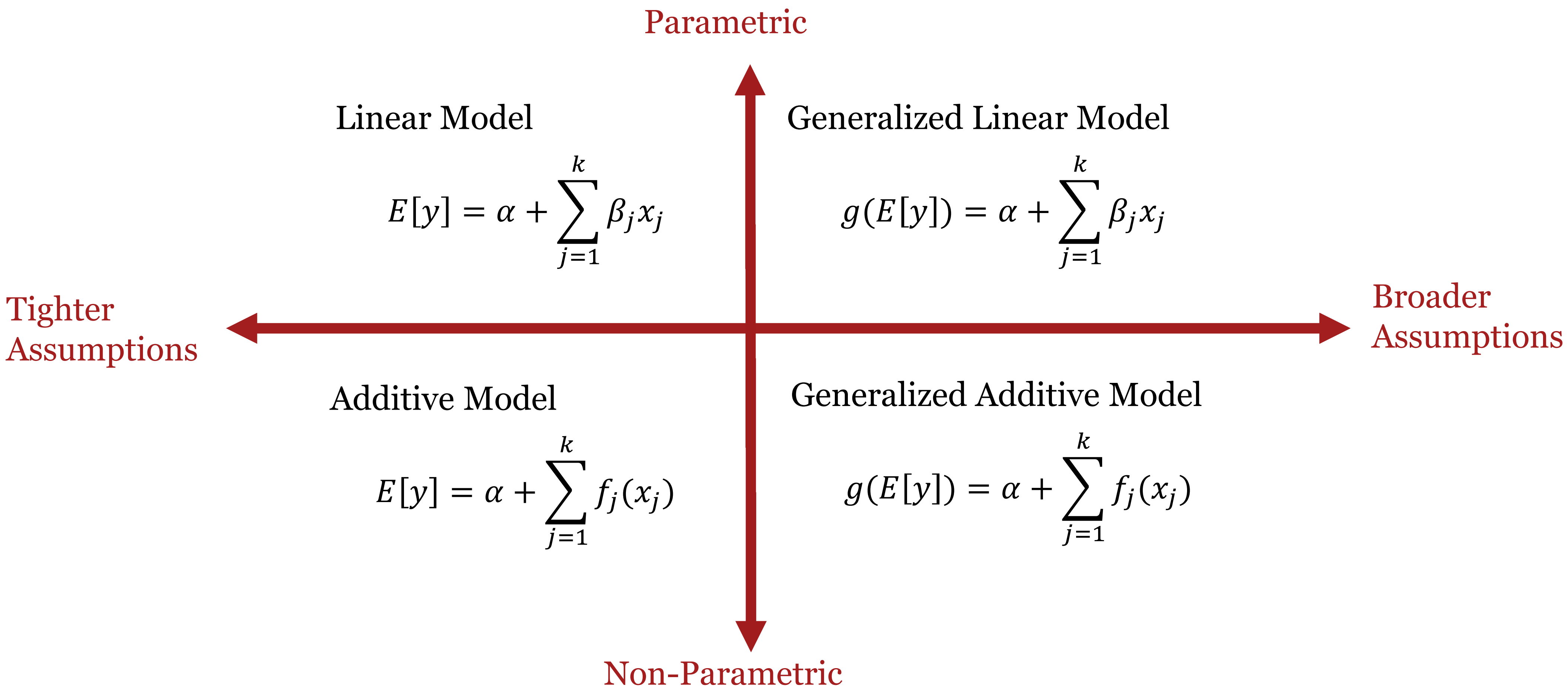


Overfitting is fitting noise so that model fails to generalize to new data.

https://theclevermachine.wordpress.com/tag/bias-variance-tradeoff/

Generalized Additive Models (GAM)

Motivation



Parametric

Linear Model

$$E[y] = \alpha + \sum_{j=1}^{k} \beta_j x_j$$

Generalized Linear Model

$$g(E[y]) = \alpha + \sum_{j=1}^{k} \beta_j x_j$$

Tighter Assumptions

Broader Assumptions

Additive Model

$$E[y] = \alpha + \sum_{j=1}^{k} f_j(x_j)$$

Generalized Additive Model

$$g(E[y]) = \alpha + \sum_{j=1}^{k} f_j(x_j)$$

Non-Parametric

General Ideas

The GAM is similar to a GLM in form and much more flexible.

$$g(E[y]) = \alpha + \sum_{j=1}^{k} f_j(x_j)$$

<u>Model Specification</u>

Modeler specifies smoother type and maximum basis dimension for each variable.

<u>Fitting Process</u>

Through penalization the fitting algorithm identifies the best subspace for each smoother

<u>May Include Parametric Terms</u>

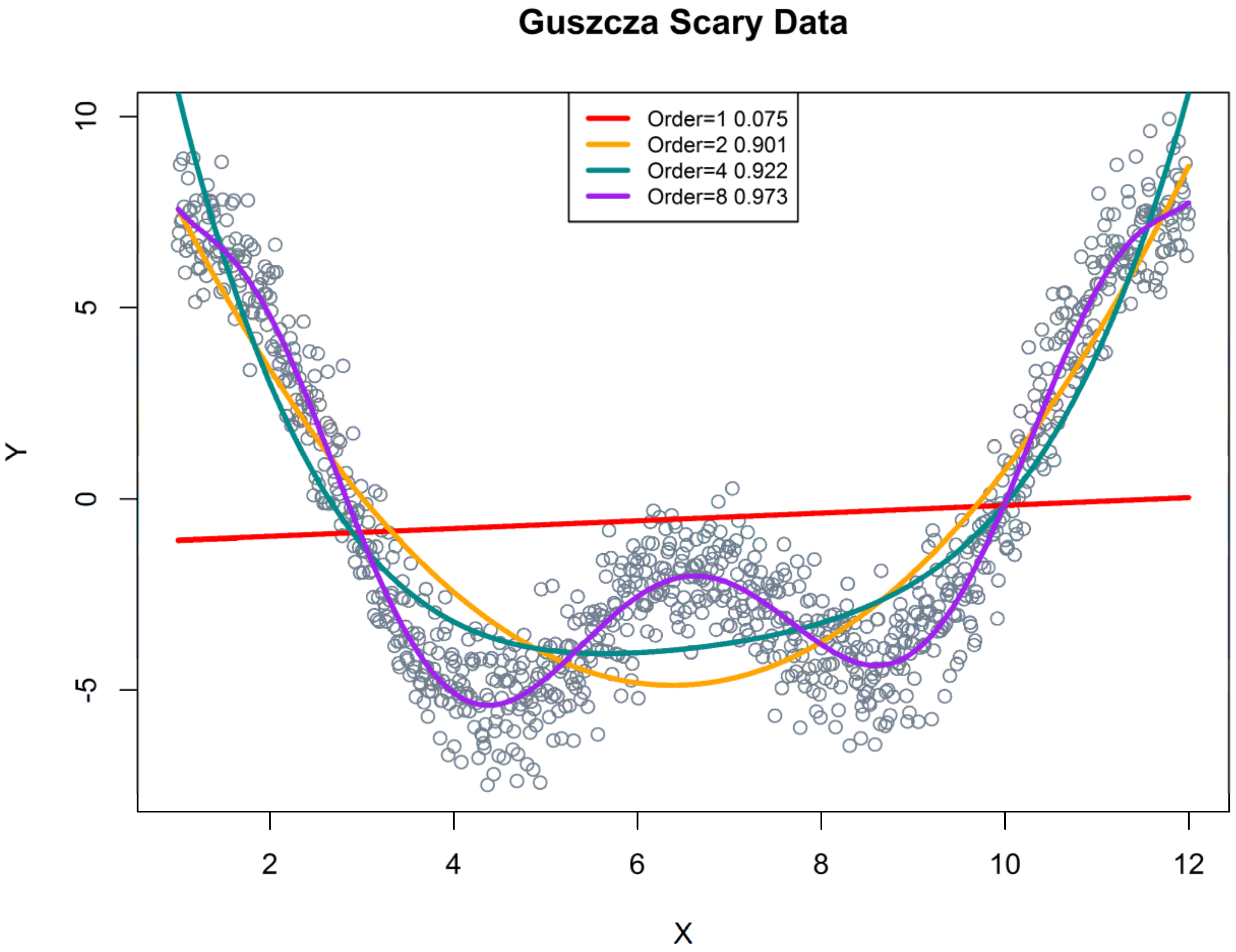
$$E[y] = \alpha + \sum_{i=1}^{n} \beta_i x_i + \sum_{j=1}^{k} f_j(x_j)$$

Polynomial Regression

- Polynomial regression of order n is a familiar example.

$$b_1(x) = 1,$$
$$b_2(x) = x,$$
$$b_3(x) = x^2,$$
$$....$$
$$b_q\ (x) = x^{n-1}$$

- Fitting the data doesn't mean you have created a good predictive model.

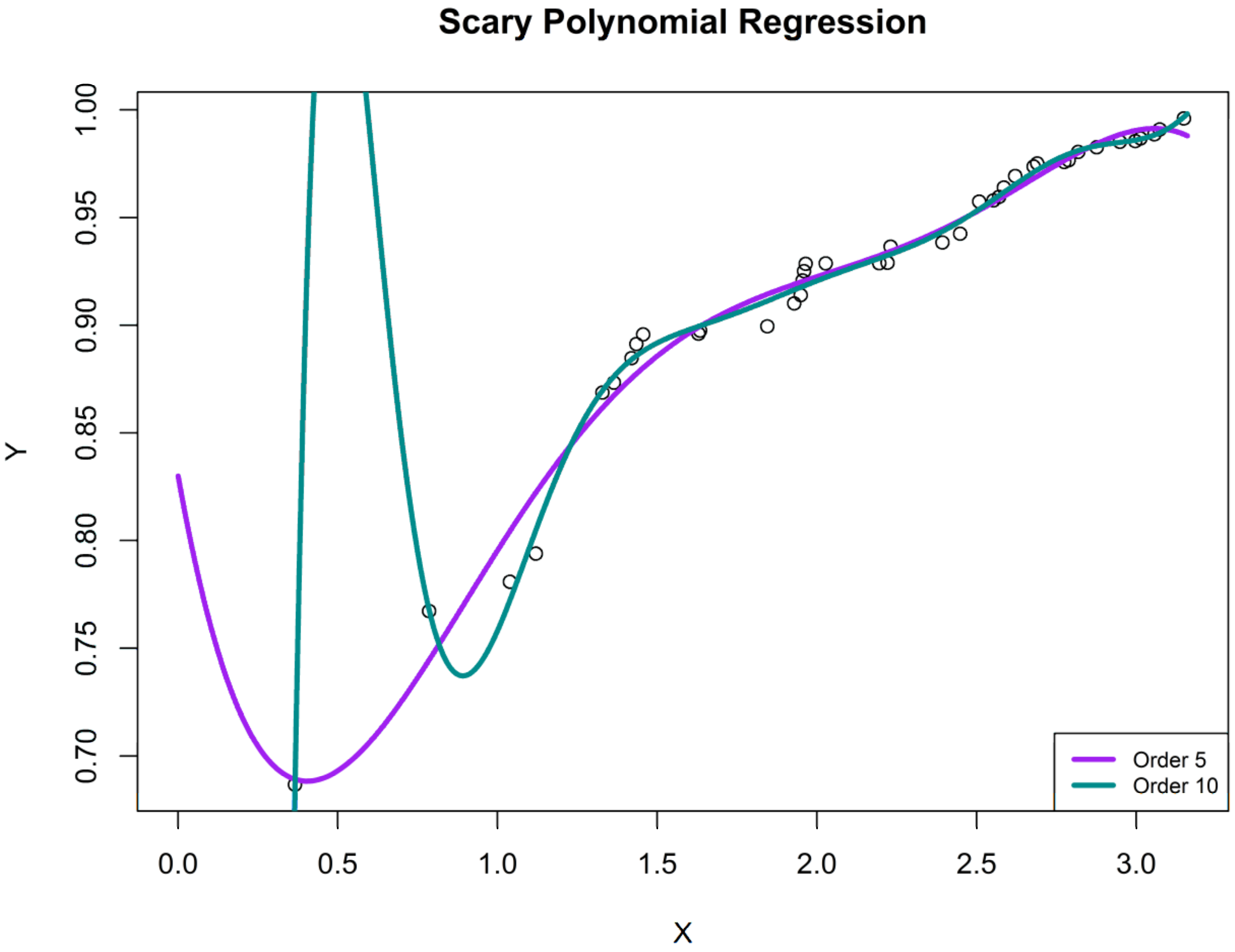


Polynomial Regression

- Polynomial models are reasonable when interested in describing localized behavior

- Sparser data allows too much freedom for the polynomial model to do crazy things.

- Endpoint behavior is another problem. Extrapolation is nonsense.



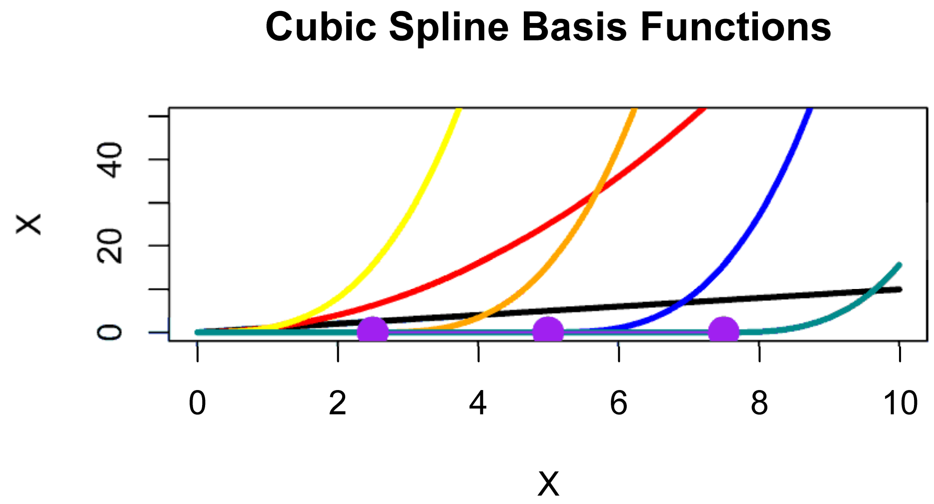
Cubic Regression Splines

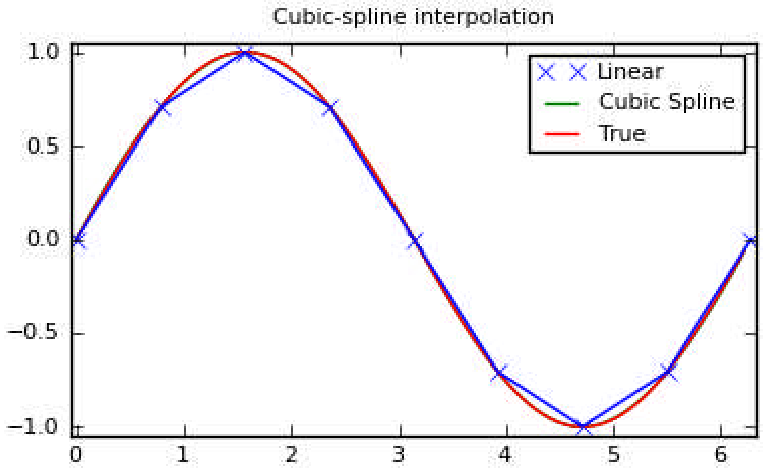
- Cubic polynomials joined together continuously and smoothly at selected “knots” partitioning the domain of application $\xi = \{\xi_1, \xi_2, \ldots, \xi_k\}$
- Choose **basis** functions $\{b_j\}$ flexible enough to approximate any $f(x) = \sum \beta_j b_j$

Cubic Spline Basis

$$b_1(x) = 1,\ b_2(x) = x,\ b_3(x) = x^2,\ \ b_4(x) = x^3,\ \ b_5(x) = (\mathrm{x}{-}\xi_1)_+^3, \ldots,\ b_{k+4}(x) = (\mathrm{x}{-}\xi_k)_+^3$$

$$\text{where } (\mathrm{x}{-}\xi_i)_+^3 = \begin{cases} (x-\xi_i)^3 & ,x > \xi_i \\ 0 & ,x \le \xi_i \end{cases}$$





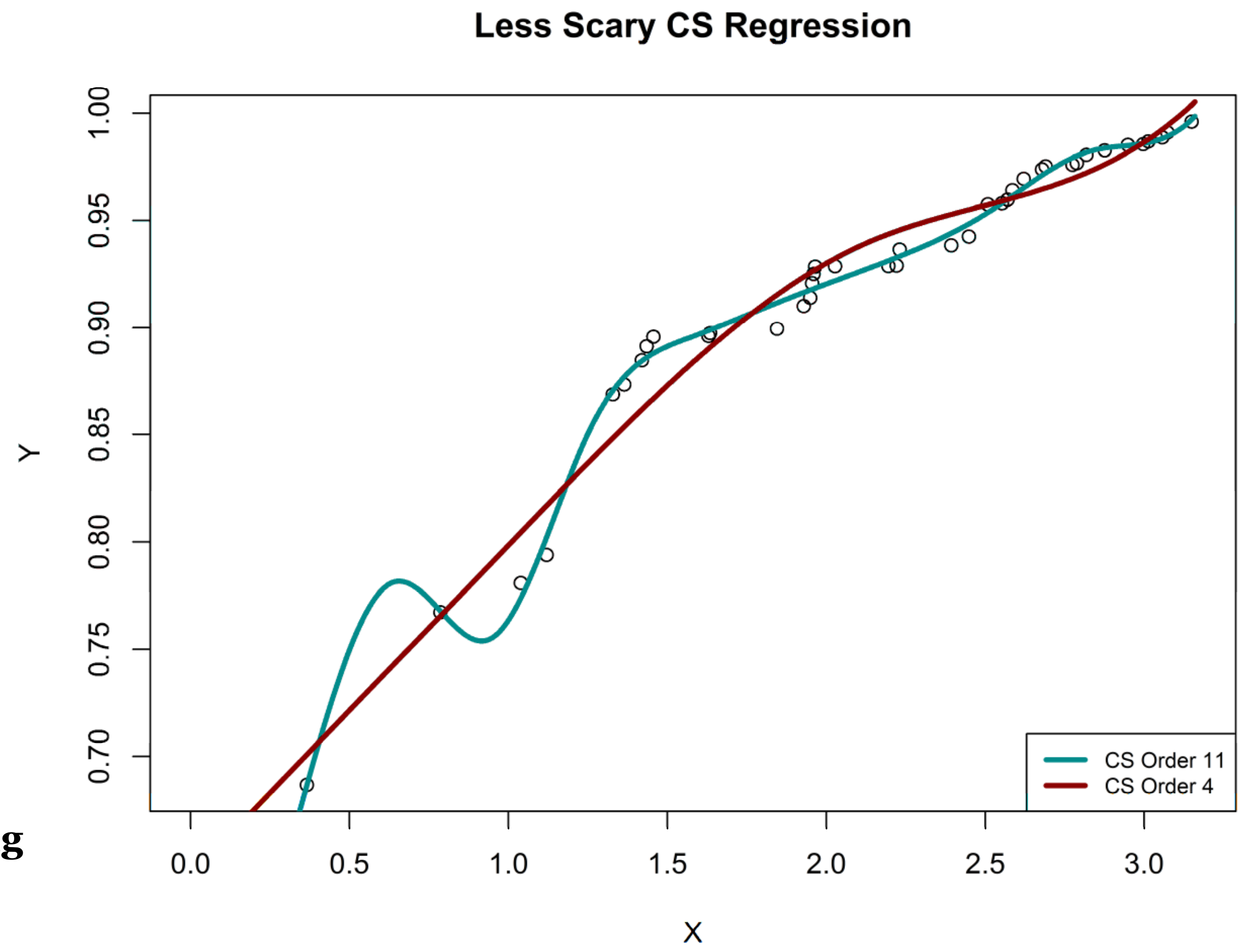
Cubic Regression Splines

- Makes more sense than the polynomials. Endpoint behavior much more reasonable.
- Still a little too “wiggly” in the sparse data region.

[redacted]

→ **Overfitting**

→ **Underfitting**



Spline Penalization

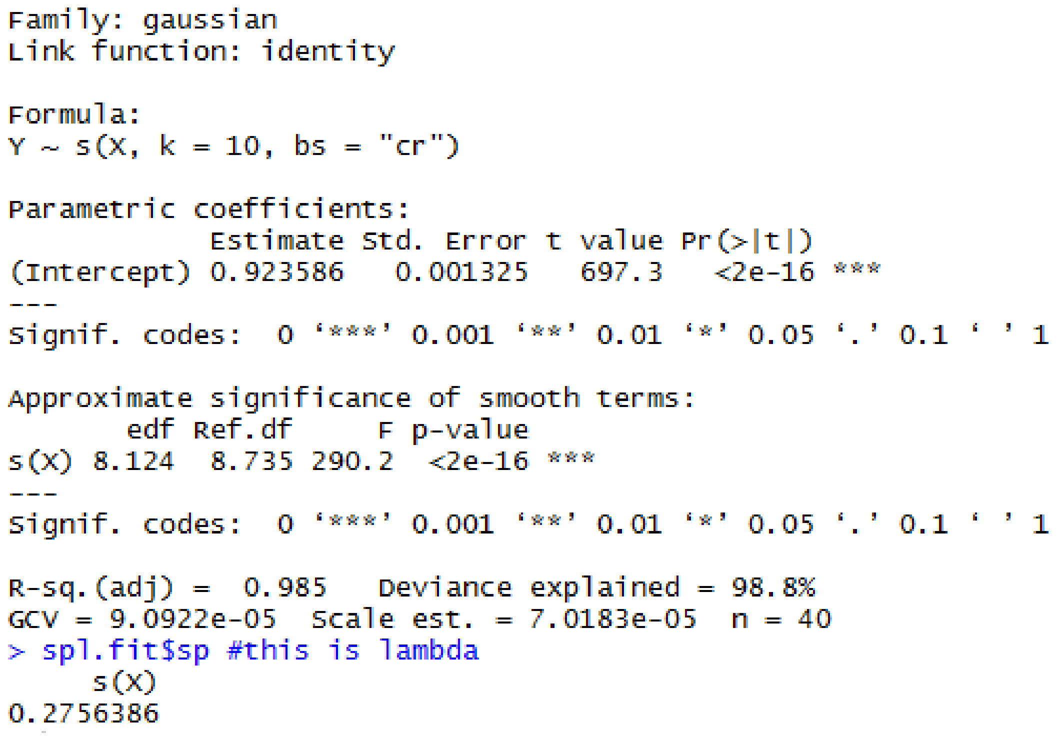
- How can we change the fitting process so that “smoother” solutions are identified?
- Introduce a penalization term to the usual SSE:

$$\sum_{i=1}^{m}[y_i - f(x_i)]^2 + \boxed{\lambda \int_a^b f''(t)^2 dt}$$

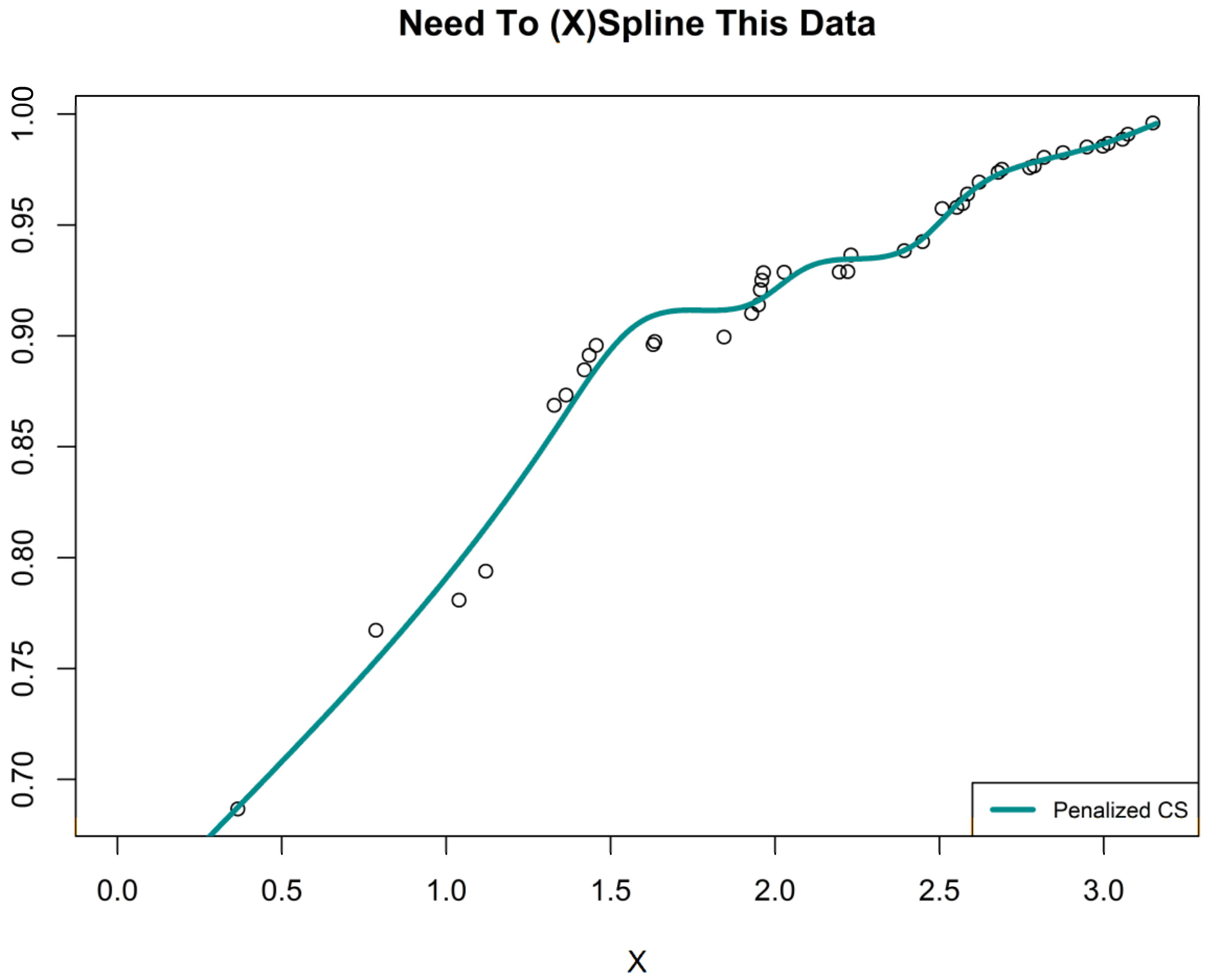
- The integral will be larger the more “wiggly” a potential solution is thus making it less likely that an overfit solution is returned.
- λ , a constant, controls the degree to which the integral penalizes the usual SSE and thus determines the tradeoff between fit and smoothness.
- As $\lambda \to \infty$ the fit approaches a constant slope linear regression.
- As $\lambda \to 0$ the fit approaches an unpenalized regression spline.

Basis Functions	$\sum_{i=1}^{m}[y_i - f(x_i)]^2$	$\lambda \int f''(t)^2 dt$
More	Lower Bias: Term is smaller	Higher Variance: Term is larger
Less	Higher Bias: Term is larger	Lower Variance: Term is smaller

Penalized Cubic Spline



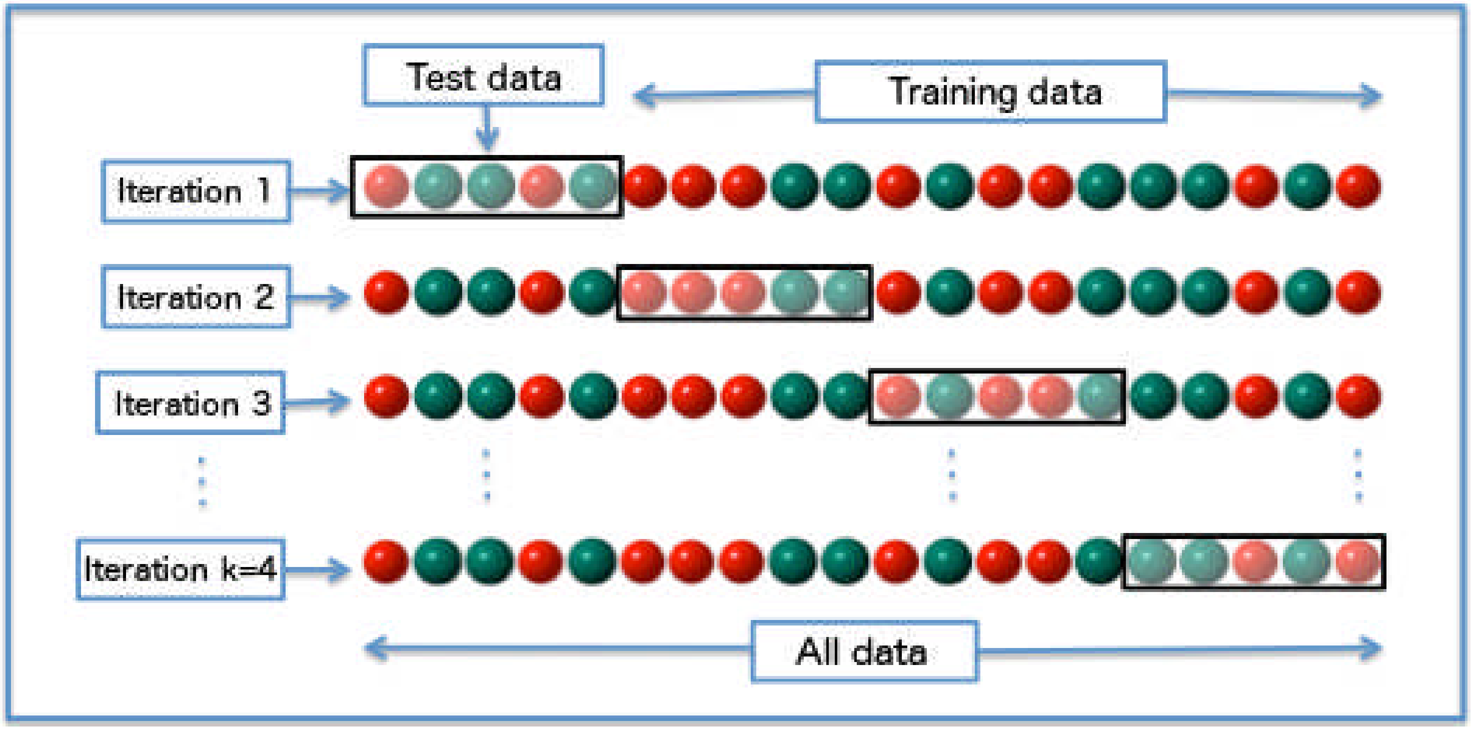
```
Family: gaussian 
Link function: identity 

Formula:
Y ~ s(X, k = 10, bs = "cr")

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.923586   0.001325   697.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
       edf Ref.df     F p-value    
s(X) 8.124  8.735 290.2  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.985   Deviance explained = 98.8%
GCV = 9.0922e-05  Scale est. = 7.0183e-05  n = 40
> spl.fit$sp #this is lambda
     s(X) 
0.2756386 
```



Choosing λ

- Since what we are really interested in when building a predictive model is that the model generalize well to new data (we want to fit signal, not noise), it is reasonable to choose λ based on cross validation.
- Exclude each of the n data points one-by-one, fit the model on the remaining $n-1$ points and then measure how well the model predicts on each of the excluded points (“leave one out CV”). Find λ such that the average of these errors is minimized - **Ordinary Cross Validation Measure (OCV).**

- A much more computationally efficient measure with favorable statistical properties is called the **General Cross Validation Measure (GCV)** but it is basically the same idea.
- You don’t really have to refit the model n times!!!
- Technically, if a scale parameter is known (Binomial/Poisson), minimization of the Un-Biased Risk Estimator (UBRE) (Mallow’s C_p) is used instead.



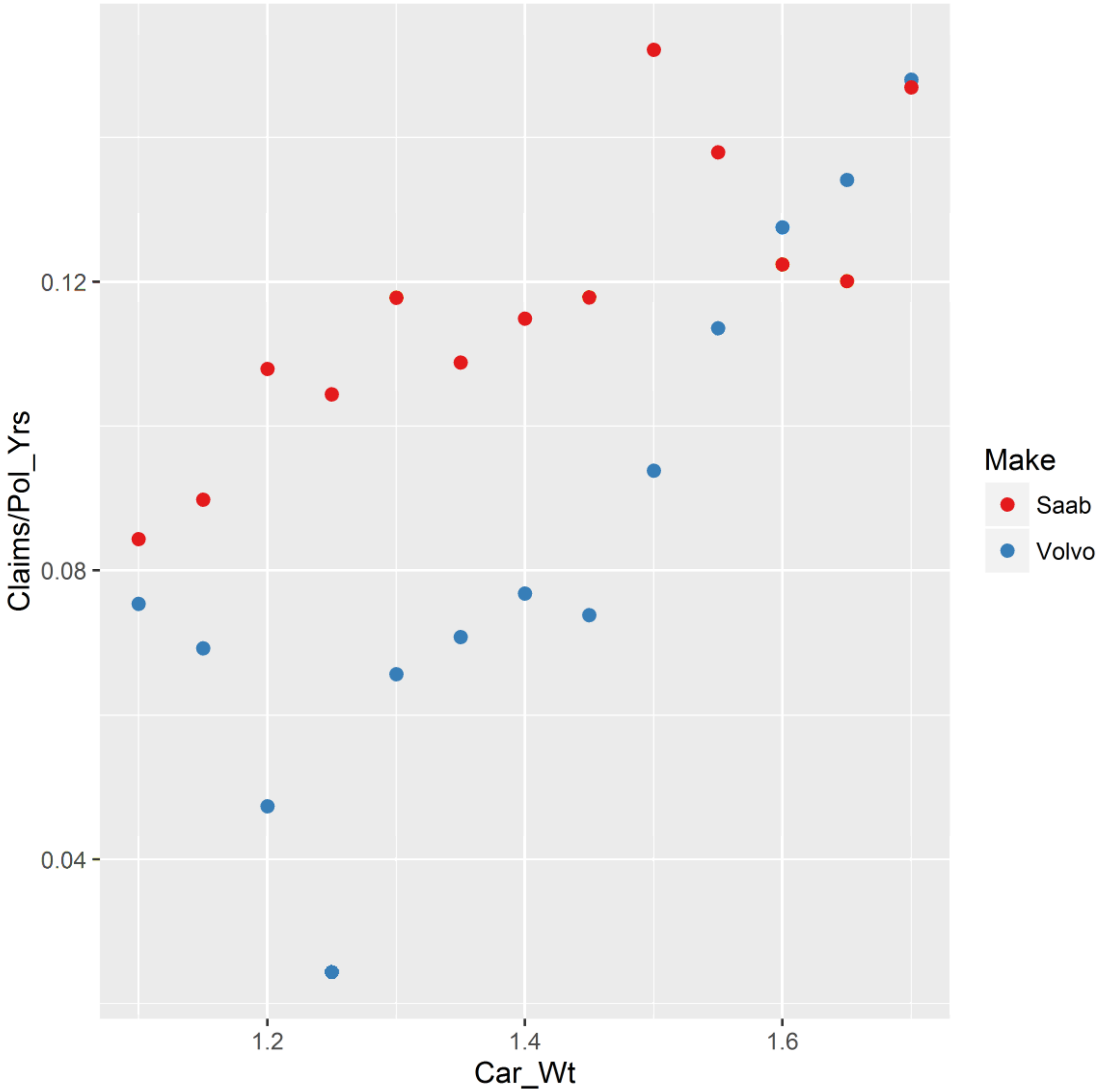
Some More Details

- When you start adding multiple covariates to the model design each smoother will have it's own penalization and thus its own λ.
- Multiple smoothers create an identifiability problem, that is, they are indistinguishable from one another up to an additive constant. A constraint is imposed to eliminate this problem, but it is useful to be aware of this particularly if you are looking at confidence bands around an effect or trying to understand degrees of freedom.
- The estimated degrees of freedoms (EDF) of each smoother is the degrees of freedom after penalization. Roughly, the model starts with the space spanned by the chosen basis with degrees of freedom equal to basis dimension less the identifiability constraint. Penalization will result in a final model with less degrees of freedom than that.
- We will use the mgcv package by Simon Wood to fit GAM's. There are many different kinds of smoothers available including some that can be applied to multiple variables.

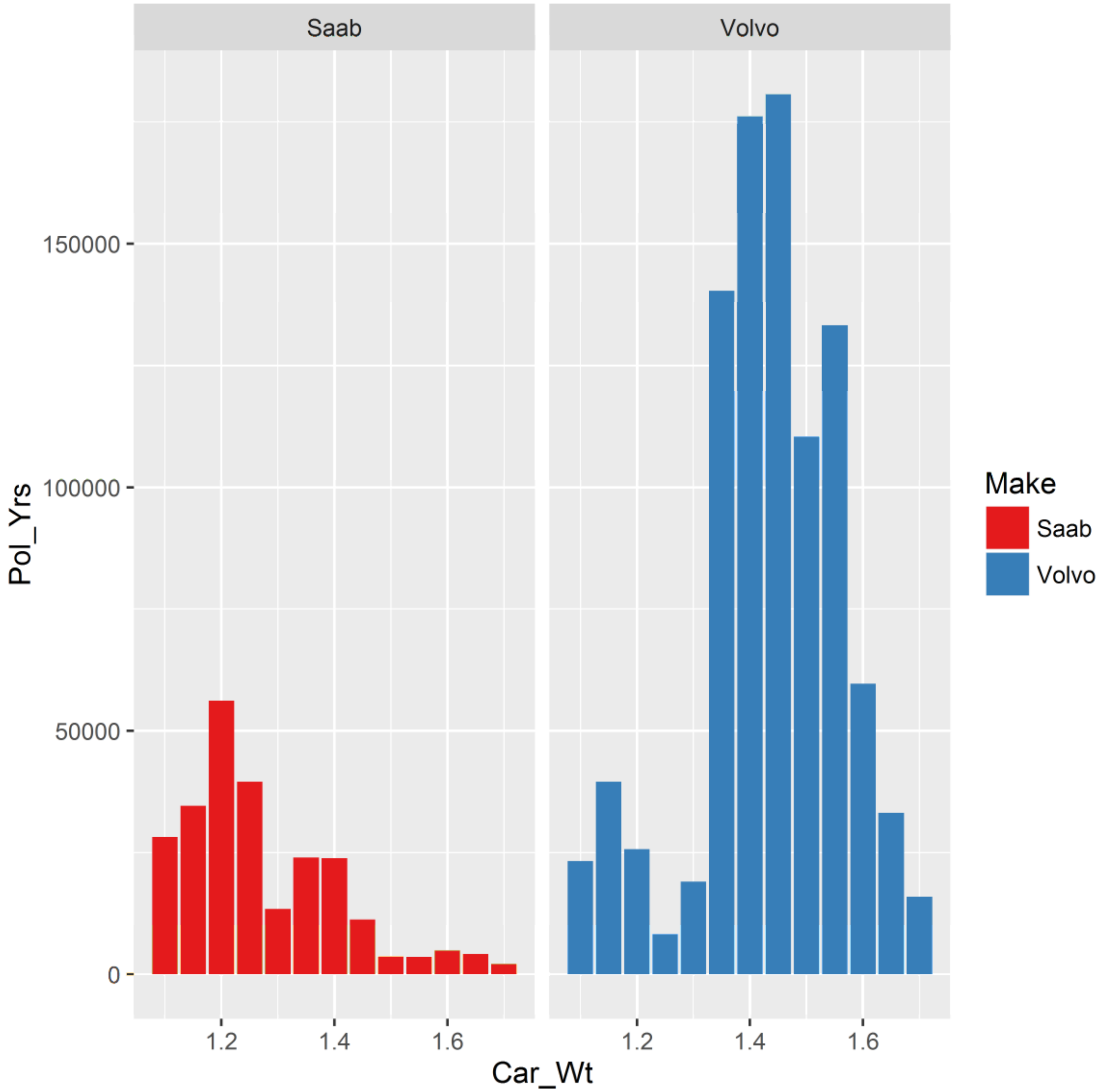
bs=	Description	Advantages	Disadvantages
"tp"	Thin plate regression splines (TPRS)	Any # covariates. Invariant to rotation. Penalty order. No knots.	Computationally costly for large data. Not invariant to covariate rescaling.
"ts"	TPRS with shrinkage	Can zero term completely.	As TPRS.
"cr"	Cubic regression splines (CRS)	Computationally cheap. Interpretable.	Single covariate. Knot based.
"cs"	CRS with shrinkage	Can zero term completely.	As CRS.
"cc"	Cyclic CRS	As CRS, but begin and end same .	As CRS.
"ps"	P-splines	Flexible combination basis and penalty order. Tensor products.	Equally spaced knots.

A Simple Example

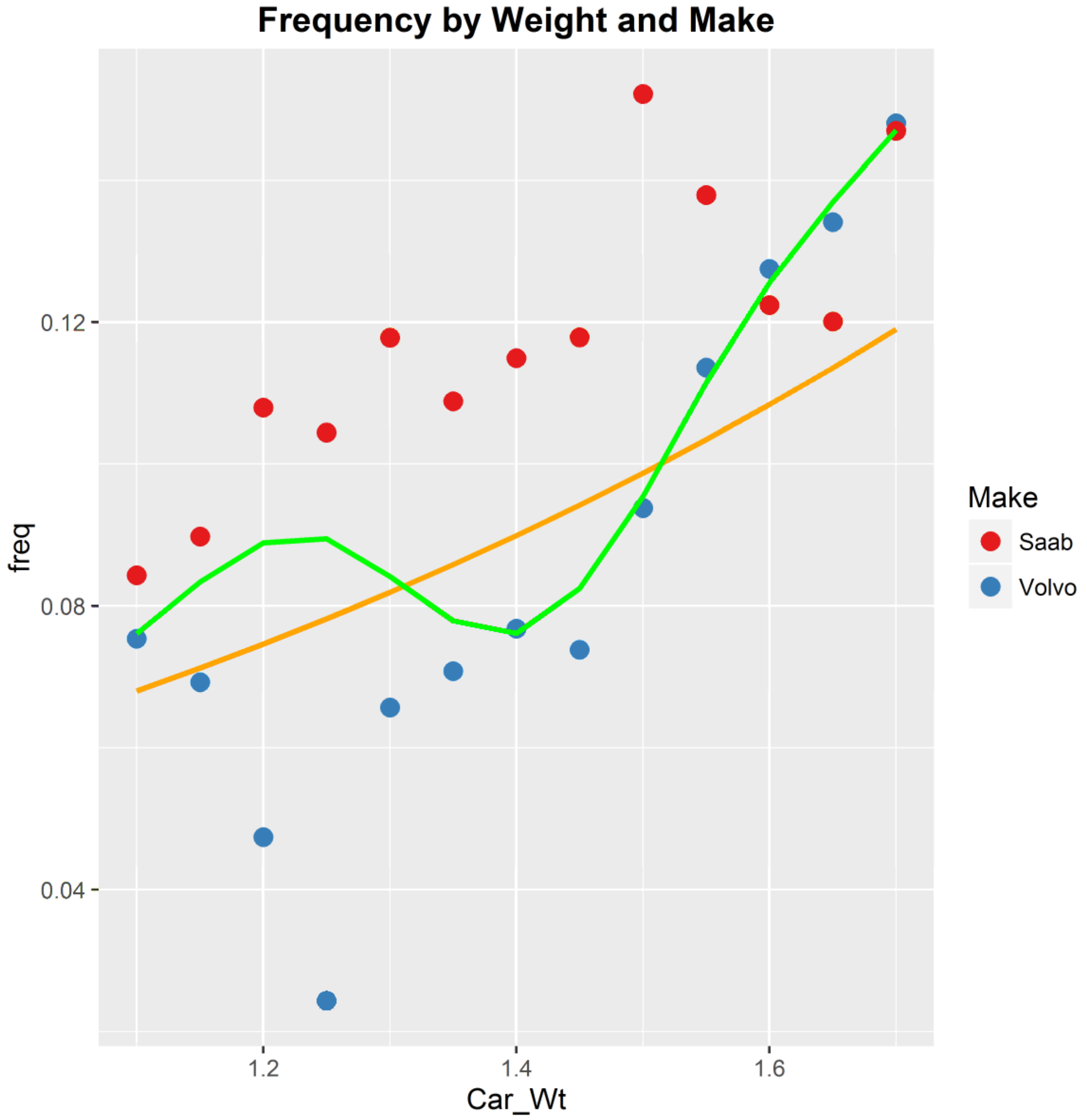
Frequency by Weight and Make

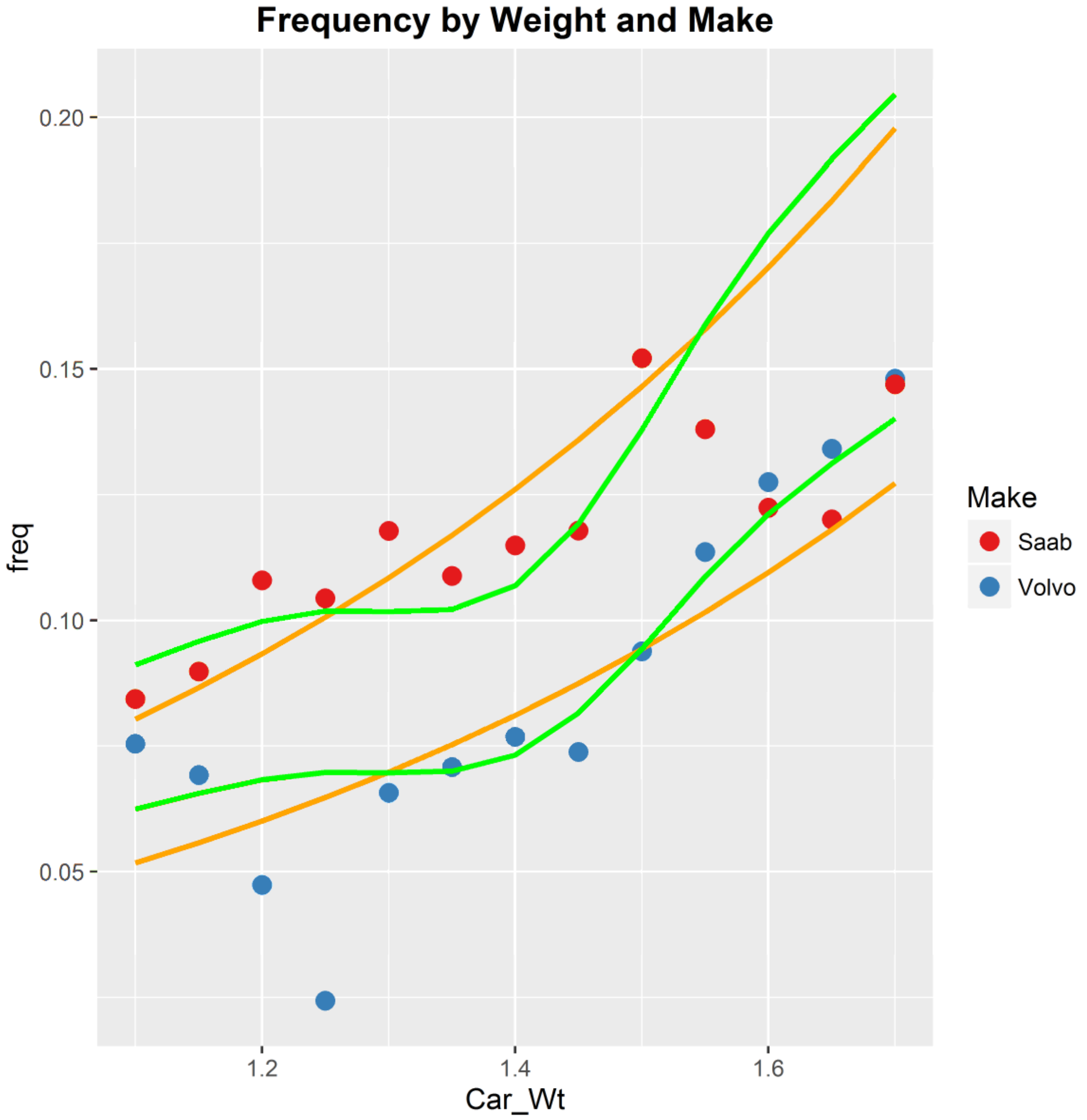


Exposure by Weight and Make

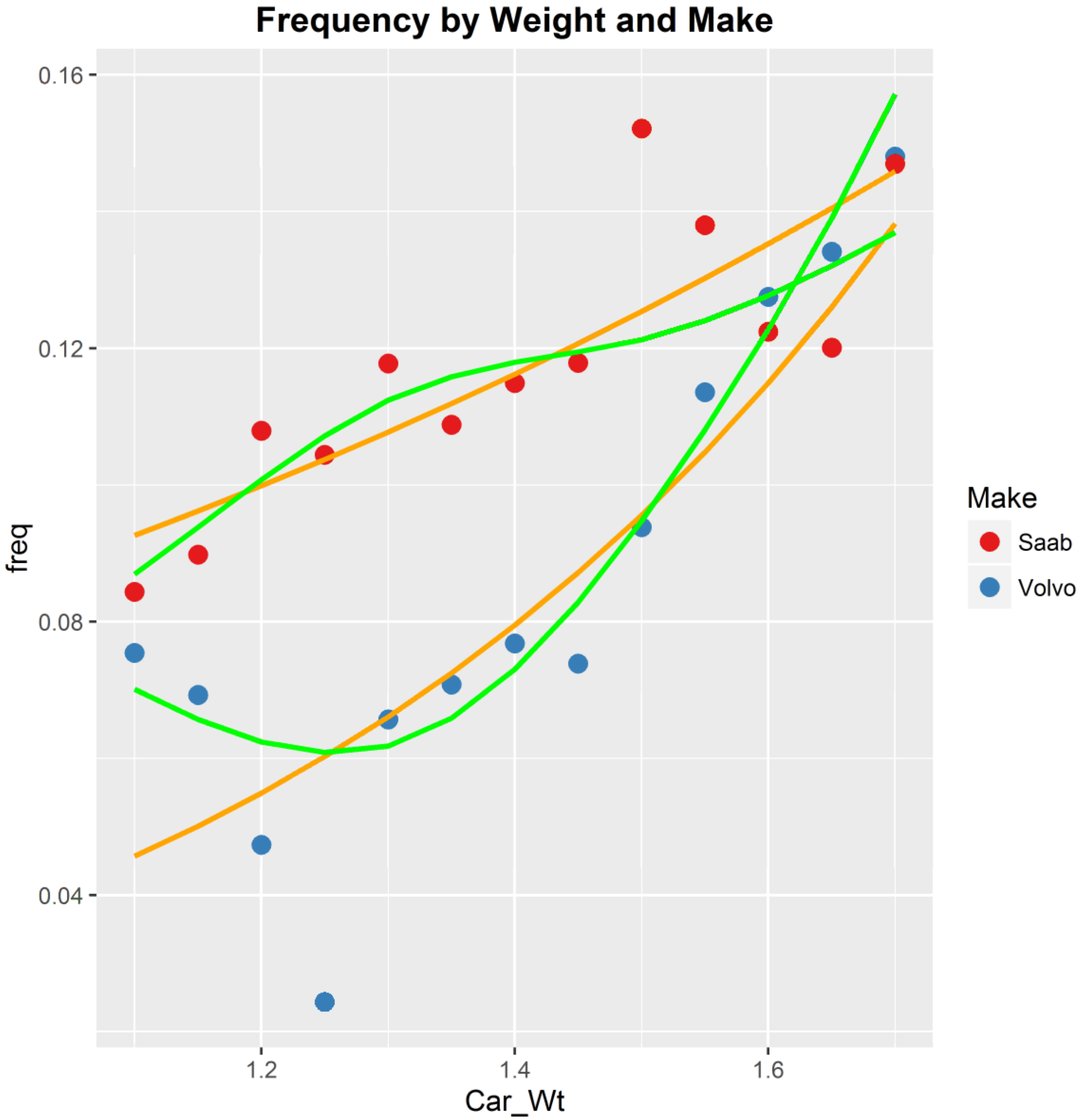


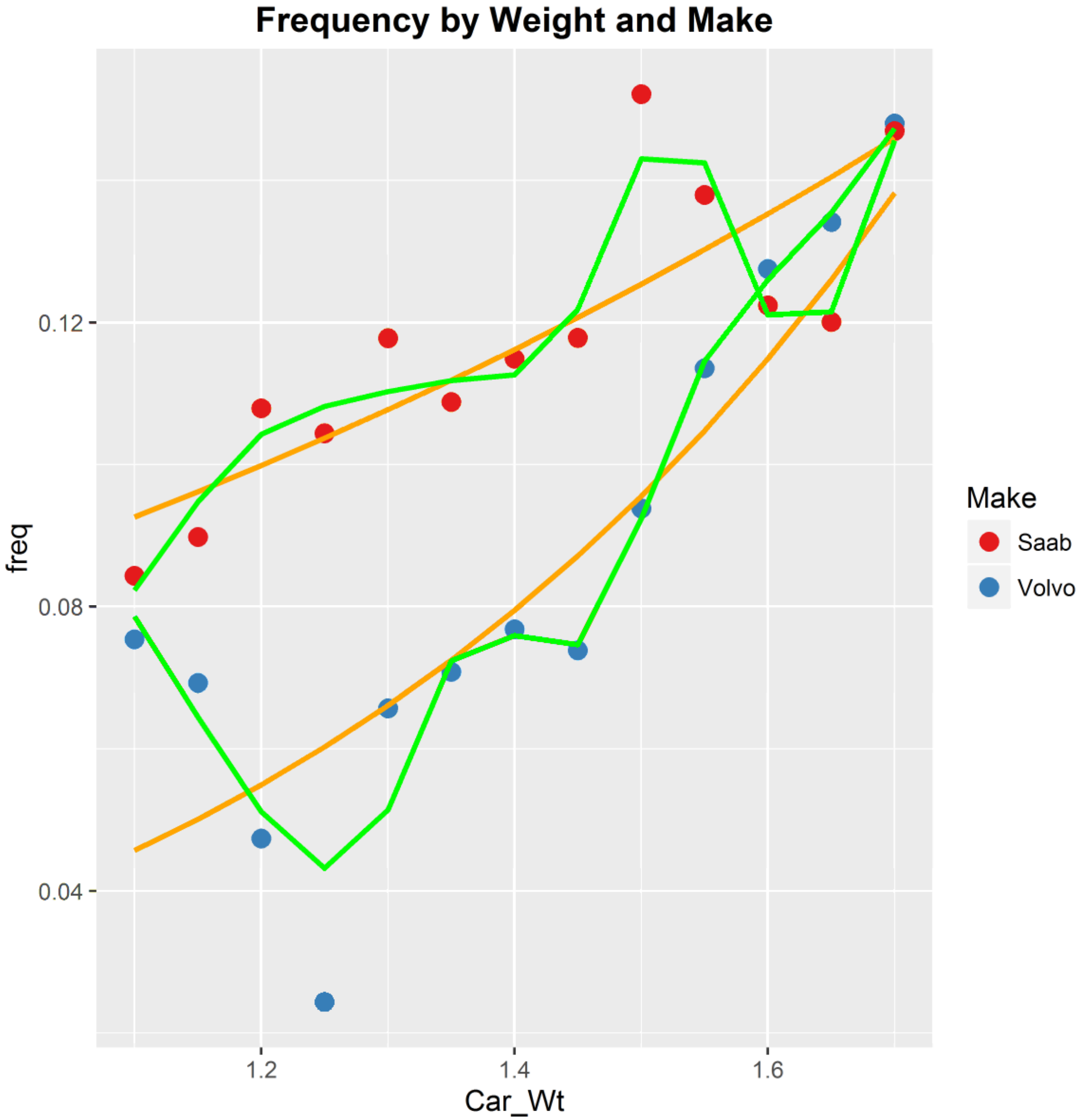
A Simple Example (cont)





A Simple Example (cont)





Geospatial Application of GAM's Case Study

Hierarchical Models

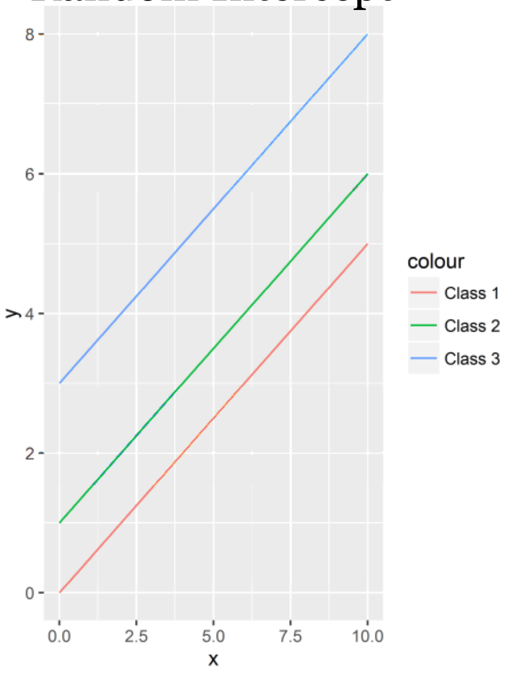
Hierarchical Models

- Hierarchical models are applied to data that is grouped in the predictor variables such that the responses share variance.
- Common examples in the insurance context are:
  1. Claims experience grouped by territory or class
  2. Retention of policies by agency or marketing territory
  3. Longitudinal studies of claim frequency, severity or development
- Hierarchical models are known by many names including:
  1. Random/Mixed effects models
  2. Multilevel models
  3. Longitudinal models
  4. Panel data models
- "The central concept of hierarchical models is that certain model parameters are themselves modeled. In other words, not all of the parameters in a hierarchical model are directly estimated from the data." - James Guszcza – "Hierarchical Growth Curve Models for Loss Reserving"

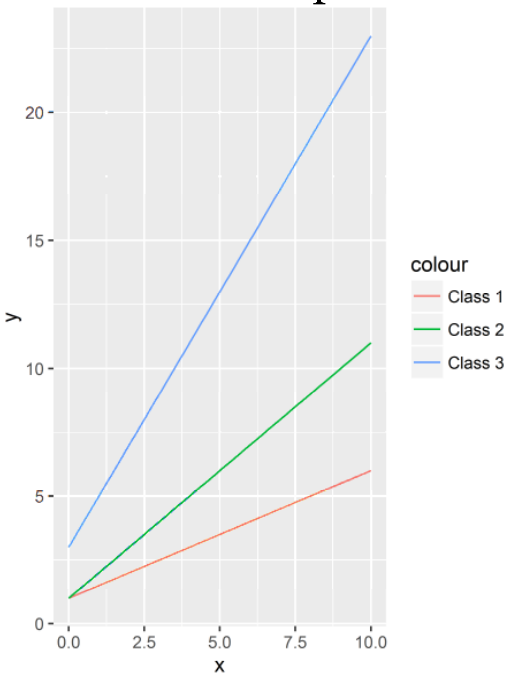
Hierarchical Models

- The term “random effects” refers to the model parameters that are modeled based on “hyperparameters” estimated from the data.
- The term “fixed effects” refers to the model parameters that are estimated directly from the data.
- Assume collection of data $(\boldsymbol{X}_j, Y)_{j=1\ldots N}$ and $j[i]$ means data point i belongs to group j.
- Classical linear model: $\boldsymbol{Y_i = \alpha + \beta X_i + \varepsilon} \longrightarrow$ same α, β for every Y
- Random intercept model: $\boldsymbol{Y_i = \alpha_{j[i]} + \beta X_i + \varepsilon} \longrightarrow$ α varies by group according to $N(\mu_\alpha, \sigma_\alpha^2)$
- Random intercept/slope model: $\boldsymbol{Y_i = \alpha_{j[i]} + \beta_{j[i]} X_i + \varepsilon} \longrightarrow$ α, β vary jointly by group

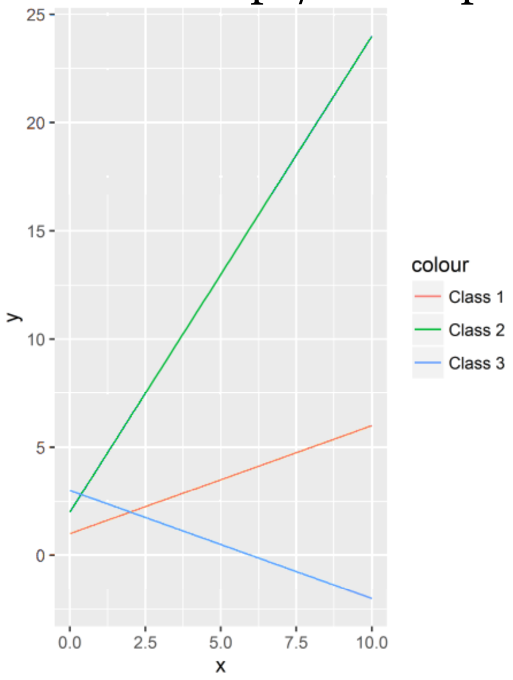
Random Intercept



Random Slope



Random Slope/Intercept



Parameters vs Hyperparameters

- Suppose we wanted to model class code level claim frequency over time with 100 classes
- If we chose to model this with a standard linear regression with a binary variable for each class:
  - $Y_i = \gamma_1 C_1 + \gamma_2 C_2 + \cdots + \gamma_{100} C_{100} + \beta t + \varepsilon$
  - 101 parameters (100 γ_j's and β)
- If we chose to model this with a random intercept model:
  - $Y_i = \alpha_{j[i]} + \beta t + \varepsilon$
  - 4 hyperparameters $(\mu_\alpha, \sigma_\alpha, \beta, \sigma)$
- If we chose to model this with a random intercept/slope model:
  - $Y_i = \alpha_{j[i]} + \beta_{j[i]} t + \varepsilon$
  - 6 hyperparameters $(\mu_\alpha, \sigma_\alpha, \mu_\beta, \sigma_\beta, \sigma_{\alpha\beta}, \sigma)$
- Which of these models is more likely to overfit? How would the parameter/hyperparameter's required change for each model if we now had 200 classes?

Relation to Credibility Theory

- Relationship between hyperparameters and parameters for the random intercept model:

$$\widehat{\alpha_j} = Z_j(\overline{y}_j - \beta\overline{x}_j) + (1\text{-}Z_j)\ \widehat{\mu_\alpha} \text{ where } Z_j = \frac{n_j}{\frac{\sigma^2}{\sigma_\alpha^2}+nj}$$

Hopefully this looks very familiar!

- Each random intercept is a credibility weighted average between the intercept for a model that ignores class entirely (pooled), μ_α, and the intercept for a model fit on each class separately (unpooled), $\overline{y}_j - \beta\overline{x}_j$.
- As σ_α approaches 0, Z_j goes to 0 , so $\widehat{\alpha_j} \longrightarrow$ Unpooled
- As σ_α approaches ∞, Z_j goes to 1 , so $\widehat{\alpha_j} \longrightarrow$ Pooled
- By removing $\overline{x}_j$ from the above expression we have the familiar expression for Buhlmann's credibility model:

$$\widehat{\alpha_j} = Z_j\overline{y}_j + (1\text{-}Z_j)\ \widehat{\mu_\alpha} \text{ where } Z_j = \frac{n_j}{\frac{\sigma^2}{\sigma_\alpha^2}+nj}$$

- Therefore Buhlmann's credibility model is a particular type of hierarchical model and hierarchical models are a means of incorporating credibility theory into the GLM framework.

Case Studies in Hierarchical Modelling

Advanced Predictive Modeling Workshop

Regression Methods

Q&A

Mark Jones ACAS, MAAA

mark.j@pwc.com