# Cutting Edge Regression Methods: RIDGE, LASSO, LOESS, and GAM

**David J Corliss, PhD**
**Michigan SAS User Group**
**June 9, 2022**

# Penalized Regression: Ridge Regression

- Regression Type: linear, penalized regression
- Contributions from predictor variables reduced by the square of the magnitude of the coefficients, favoring models with many strong contributors
- Supports measurement and analysis of the amount of collinearity and parameterization of the ridge factor
- SAS = REG with RIDGE option, R = glmnet

# Special Model Types: Ridge Regression Source Code and Options

```
proc reg data=rm.homelessstudents ridge=0 to .04 by .005;
   outvif outest=ridgests plots(only)=ridge(unpack VIFaxis=log);
   model hs_pct_10 = GINI_10 GINI_pct_change Pov_Change_09_11
                     aa_pct hisp_latino_pct indian_alaskan_pct
                     high_school_grad_pct_10  mhi_09 mfi_acs_10;
run;
```

## REG Statement Options

- ridge – ridge parameter limits and step size
- outvif – output variance inflation factor => severity of multicol.
- outseb – output standard errors and parameter estimates

## PLOT Statement Options

- all – lots of plots but many are seldom used
- ridge – shrinking by ridge parameter as the model converges

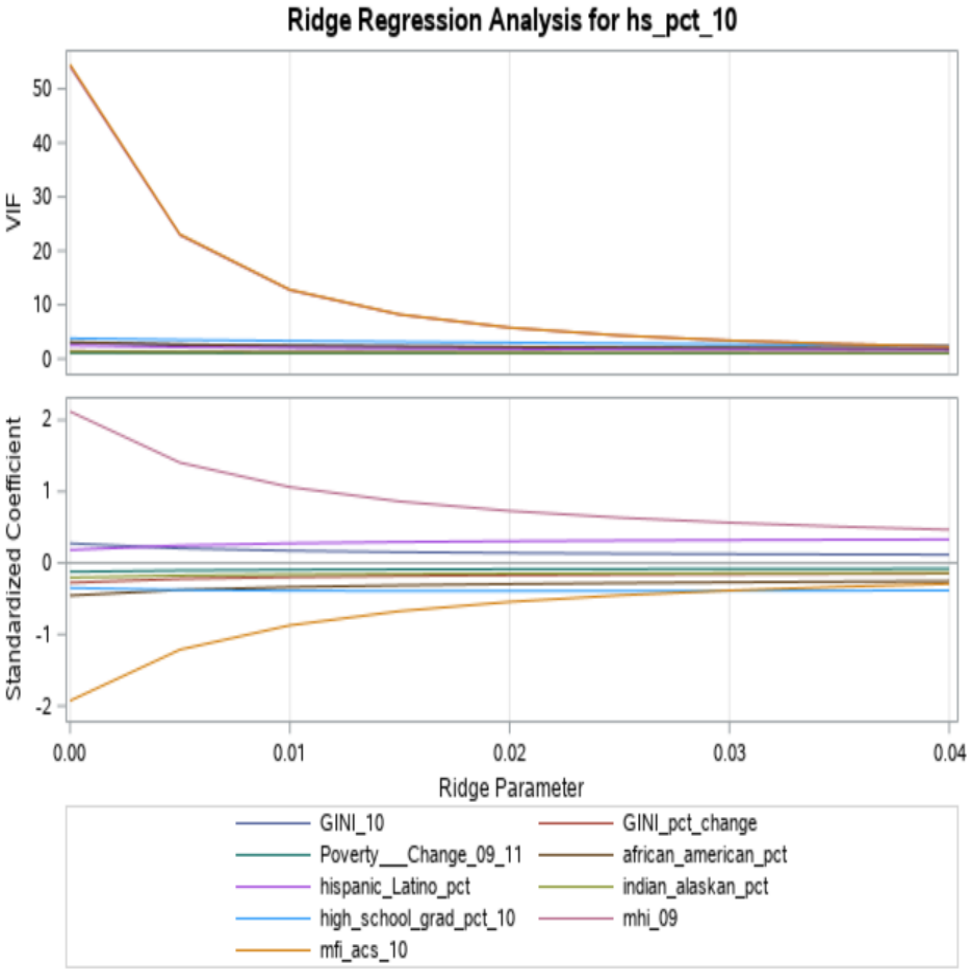
# Special Model Types: Ridge Regression Output and Plots

The REG Procedure
Model: MODEL1
Dependent Variable: hs_pct_10 hs_pct_10

| | |
|---|---|
| Number of Observations Read | 52 |
| Number of Observations Used | 51 |
| Number of Observations with Missing Values | 1 |

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 9 | 0.02665 | 0.00296 | 4.99 | 0.0002 |
| Error | 41 | 0.02435 | 0.00059384 | | |
| Corrected Total | 50 | 0.05100 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.02437 | R-Square | 0.5226 |
| Dependent Mean | 0.01949 | Adj R-Sq | 0.4178 |
| Coeff Var | 125.05331 | | |



# Special Model Types: Ridge Regression
# Example 2: OUTSEB for model parameters

```
proc reg data=rm.homelessstudents ridge=0 to .04 by .005 outseb;
   outvif outest=ridgests plots(only)=ridge(unpack VIFaxis=log);
   model hs_pct_10 = GINI_pct_change african_american_pct
                     high_school_grad_pct_10  mhi_09 mfi_acs_10;
run;
```

| | | | |
|---|---|---|---|
| Root MSE | 0.02518 | R-Square | 0.4405 |
| Dependent Mean | 0.01949 | Adj R-Sq | 0.3783 |
| Coeff Var | 129.22766 | | |

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | Intercept | 1 | 0.56092 | 0.11549 | 4.86 | <.0001 |
| GINI_pct_change | GINI_pct_change | 1 | -0.51430 | 0.26674 | -1.93 | 0.0602 |
| african_american_pct | african_american_pct | 1 | -0.00106 | 0.00038824 | -2.73 | 0.0090 |
| high_school_grad_pct_10 | high_school_grad_pct_10 | 1 | -0.66107 | 0.14763 | -4.48 | <.0001 |
| mhi_09 | mhi_09 | 1 | 0.00000663 | 0.00000252 | 2.63 | 0.0116 |
| mfi_acs_10 | mfi_acs_10 | 1 | -0.00000464 | 0.00000230 | -2.02 | 0.0493 |

**In this example, a final run with the strongest variables from earlier runs performs well**

# Special Model Types: LASSO Regression

- Regression Type: linear, penalized regression
- Contributions from predictors reduced by the sum of the absolute values of the magnitude of the coeffi-cients, favoring models with a few strong contributors
- Supports multiple cross validation, a variety of methods for choosing variables for the model with the CHOOSE statement, and Bayesian analysis
- SAS = GLMSELECT with LASSO option, R = glmnet

# Special Model Types: LASSO Regression Source Code and Options

```
proc glmselect data=rm.sas_baseball plots=all;
   partition fraction(validate=.3);
   model logSalary = nAtBat nHits nHome nRuns nRBI nBB
                     yrMajor crAtBat crHits crHome crRuns
                     crRbi crBB nOuts nAssts nError
         / selection=lasso(stop=none choose=validate);
run;
```

## MODEL Statement Options

- selection – must = lasso to use this method
- stop – sets the criteria for when to stop variable selection, stop=none examines all variables in the MODEL statement
- choose – sets criteria for choosing the model; default is AICC, sbc is Schwarz Bayesian information criterion

## PARTITION Statement Options

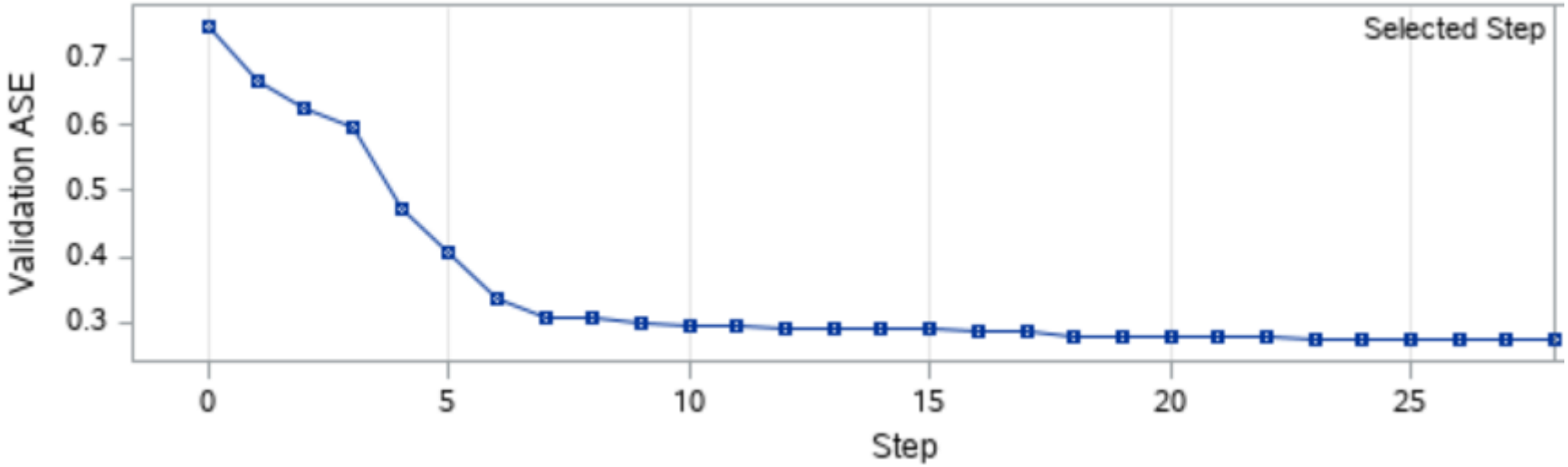
- validate – states the fraction for the validation sample

# Special Model Types: LASSO Regression Output and Plots

| Analysis of Variance | | | | |
|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value |
| Model | 16 | 84.14448 | 5.25903 | 14.28 |
| Error | 161 | 59.30511 | 0.36835 | |
| Corrected Total | 177 | 143.44959 | | |

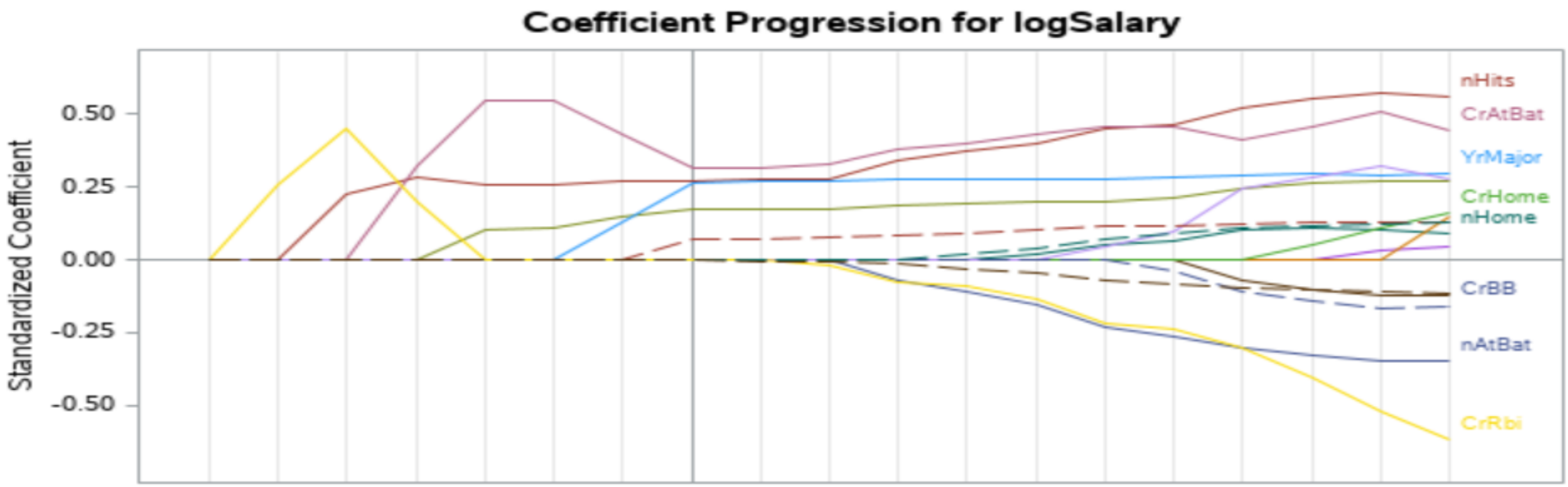
| | |
|---|---|
| Root MSE | 0.60692 |
| Dependent Mean | 5.92151 |
| R-Square | 0.5866 |
| Adj R-Sq | 0.5455 |
| AIC | 18.36231 |
| AICC | 22.66420 |
| SBC | -107.54737 |
| ASE (Train) | 0.33317 |
| ASE (Validate) | 0.27405 |

| LASSO Selection Summary | | | | | |
|---|---|---|---|---|---|
| Step | Effect Entered | Effect Removed | Number Effects In | ASE | Validation ASE |
| 0 | Intercept | | 1 | 0.8059 | 0.7496 |
| 1 | CrRuns | | 2 | 0.7087 | 0.6667 |
| 2 | CrHits | | 3 | 0.6648 | 0.6268 |
| 3 | CrRbi | | 4 | 0.6312 | 0.5963 |
| 4 | nHits | | 5 | 0.5155 | 0.4746 |
| 5 | nBB | | 6 | 0.4549 | 0.4082 |
| 6 | YrMajor | | 7 | 0.3866 | 0.3343 |
| 7 | nRBI | | 8 | 0.3650 | 0.3090 |
| 8 | nRuns | | 9 | 0.3632 | 0.3067 |
| 9 | nError | | 10 | 0.3514 | 0.2989 |
| 10 | nOuts | | 11 | 0.3472 | 0.2953 |



# Special Model Types: LASSO Regression Example 2: Adaptive LASSO w/ Bayesian

```
proc glmselect data=rm.sas_baseball plots=all;
   partition fraction(validate=.3);
   model logSalary = nAtBat nHits nHome nRuns nRBI nBB
                     yrMajor crAtBat crHits crHome crRuns
                     crRbi crBB nOuts nAssts nError
/ selection=lasso(adaptive stop=none choose=sbc);
run;
```



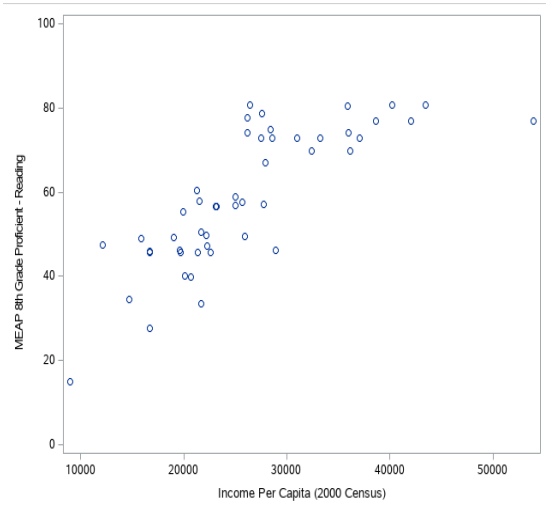

**Adaptive LASSO improves the ability to select just the strongest predictors and Bayesian analysis is specified using choose=sbc**

# Non-Parametric Models: Localized Local Regression

- Regression Type: linear, non-parametric
- Develops a model using non-parametric regression to segments of data and calculates confidence limits for
- the outcome; computationally intensive
- Supports multiple dependent variables, multidimensional predictors and interpolation using kd trees
- SAS = LOESS, R = locfit

# Non-Parametric Models: Local Regression Source Code and Options

```
ods graphics on;
proc loess data=rm.sem_eduction;
   ods output OutputStatistics=GasFit
              FitSummary=Summary;
   model MEAP8_Read = PCI_2000;
run;
ods graphics off;
```




## MODEL Statement Options

- degree – degree of the local polynomials (either 1 or 2)
- select= – specifies a smoothing method: AICC, AICC1, GCV, DF1, DF2, or DF3
- direct – requires direct fitting at every point
- std – outputs the standard of the mean predicted values

## SCORE Statement Options

- clm – output confidence limits with the score

# Non-Parametric Models: Local Regression Output and Plots

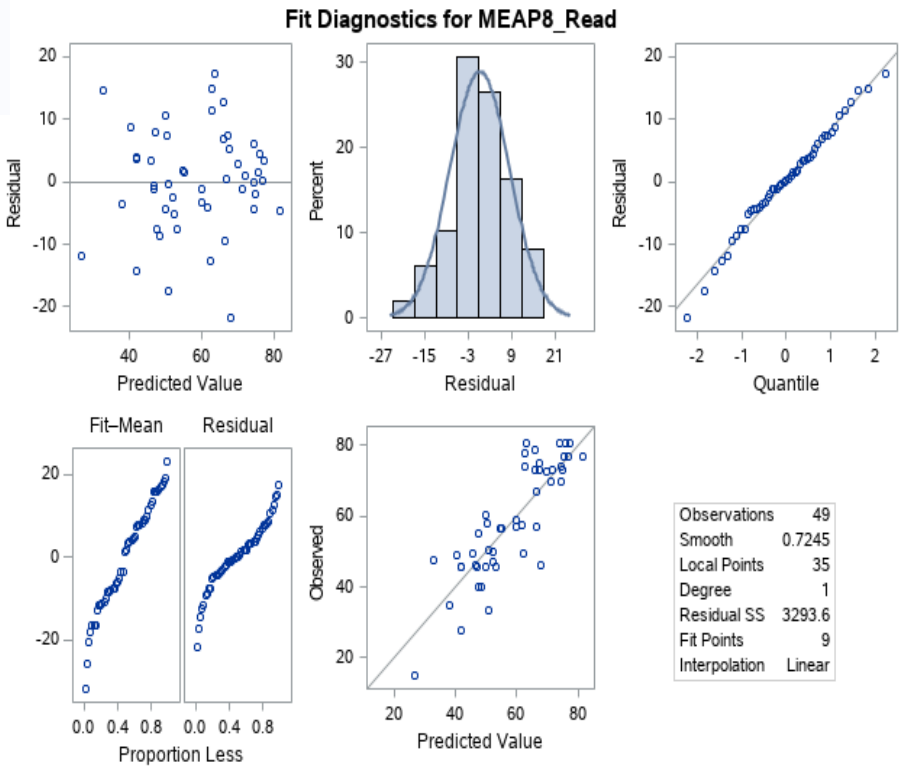
The LOESS Procedure
Selected Smoothing Parameter: 0.724
Dependent Variable: MEAP8_Read

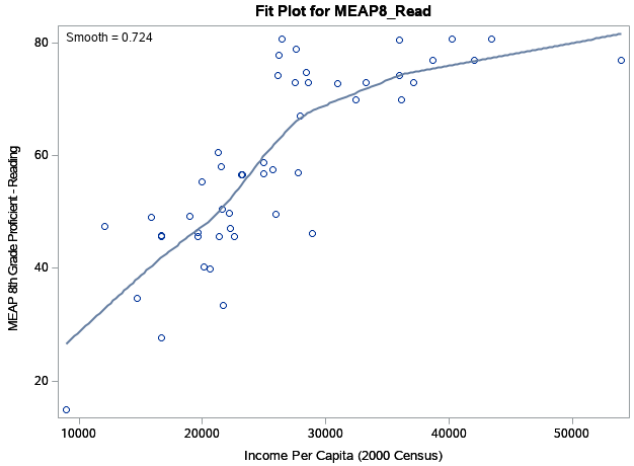
| Fit Summary | |
|---|---|
| Fit Method | kd Tree |
| Blending | Linear |
| Number of Observations | 49 |
| Number of Fitting Points | 9 |
| kd Tree Bucket Size | 7 |
| Degree of Local Polynomials | 1 |
| Smoothing Parameter | 0.72449 |
| Points in Local Neighborhood | 35 |
| Residual Sum of Squares | 3293.62441 |
| Trace[L] | 4.26356 |
| GCV | 1.64570 |
| AICC | 5.45425 |

| Optimal Smoothing Criterion | |
|---|---|
| AICC | Smoothing Parameter |
| 5.45425 | 0.72449 |

The LOESS Procedure

| Independent Variable Scaling | |
|---|---|
| Scaling applied: None | |
| Statistic | Income Per Capita (2000 Census) |
| Minimum Value | 8965 |
| Maximum Value | 53942 |

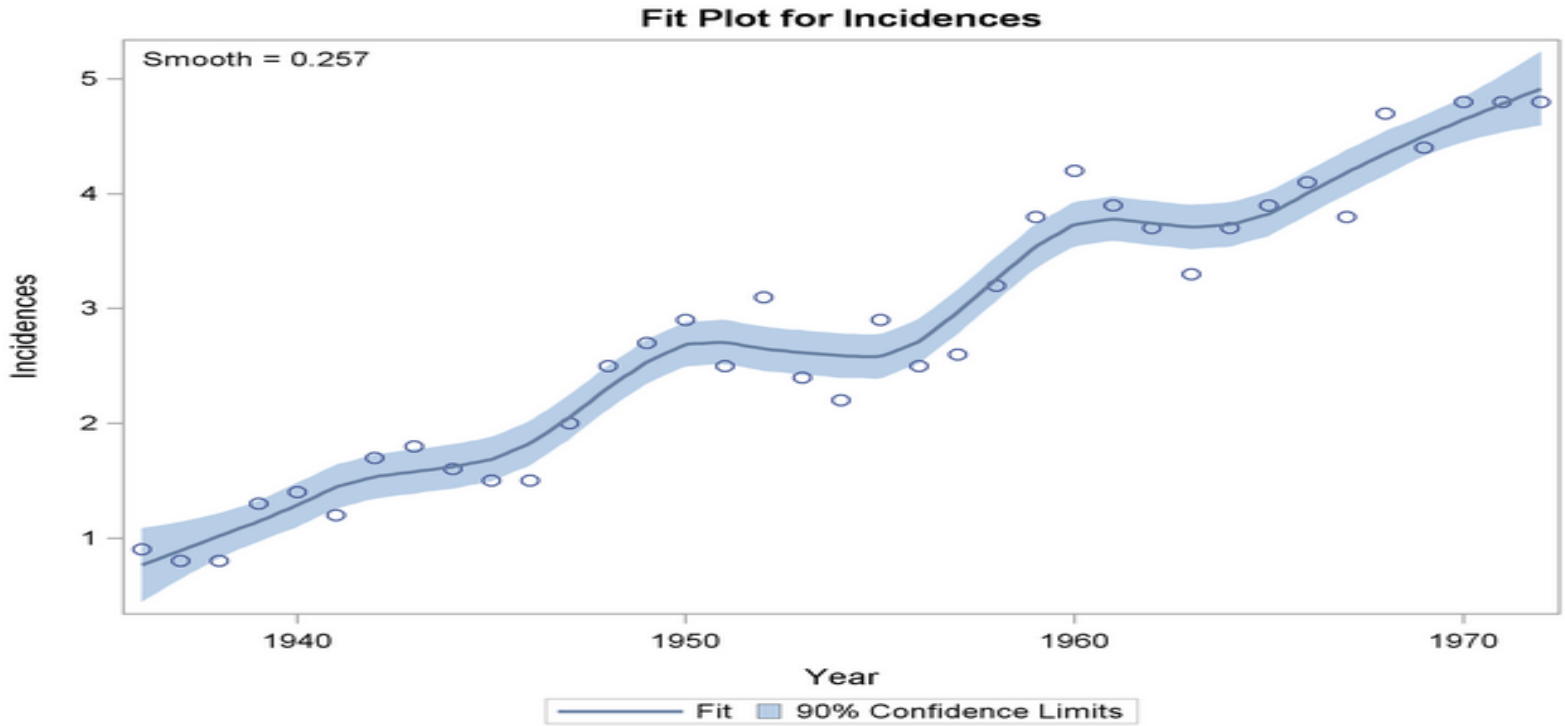



**Local Regression is an option for complex functional forms – here it fits the same data as one example for PROC NLIN**

# Local Regression
# Example 2: Smoothing Data with PROC LOESS

```
proc loess data=Melanoma;
   model Incidences=Year/clm alpha=0.1
run;
```




SAS Figure 50.13 Loess Fit of Melanoma Data with 90% Confidence Limits

**In this example, Local Regression is used with a smoothing factor to model a complex form without overfitting**

# Non-Parametric Models: Additive Generalized Additive Models

- Regression Type: linear, non-parametric
- Generalized Additive Models, with multiple independent non-parametric predictors; univariate smoothing provides finer details than is possible with the piece-wise LOESS procedure
- Supports non-parametric and semi-parametric models, multidimensional predictors
- SAS = GAM, R = gam

# Non-Parametric Models: Generalized Additive Models Source Code and Options

```
proc gam data=rm.baseball plots(unpack)=all;
   model term_2017 = spline(RateApril) spline(RateMay)
                     spline(RateJune / method=gcv;
   output out=PredGAM p=Gam_p_;
run;
```

## GAM Statement Options

- descending – reverses the sort order of the class variable
- plot= – plotting options: all, unpack

## MODEL Statement Options

- anodev – smoothing options: refit, norefit, none
- maxiter – maximum number of estimation iterations
- method=gcv – smoothing parameter uses the generalized cross validation method

# Non-Parametric Models: Generalized Additive Models Output and Plots

The GAM Procedure
Dependent Variable: FinalPlace
Smoothing Model Component(s): spline(RateApril) spline(RateMay) spline(RateJune)

| Summary of Input Data Set | |
|---|---|
| Number of Observations | 30 |
| Number of Missing Observations | 0 |
| Distribution | Gaussian |
| Link Function | Identity |

| Iteration Summary and Fit Statistics | |
|---|---|
| Final Number of Backfitting Iterations | 7 |
| Final Backfitting Criterion | 4.3096977E-9 |
| The Deviance of the Final Estimate | 14.372176478 |

The backfitting algorithm converged.

Regression Model Analysis
Parameter Estimates

| Parameter | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 10.07191 | 1.33201 | 7.56 | <.0001 |
| Linear(RateApril) | -5.57092 | 1.66769 | -3.34 | 0.0039 |
| Linear(RateMay) | -4.06923 | 1.42487 | -2.86 | 0.0109 |
| Linear(RateJune) | -4.63003 | 2.08142 | -2.22 | 0.0400 |

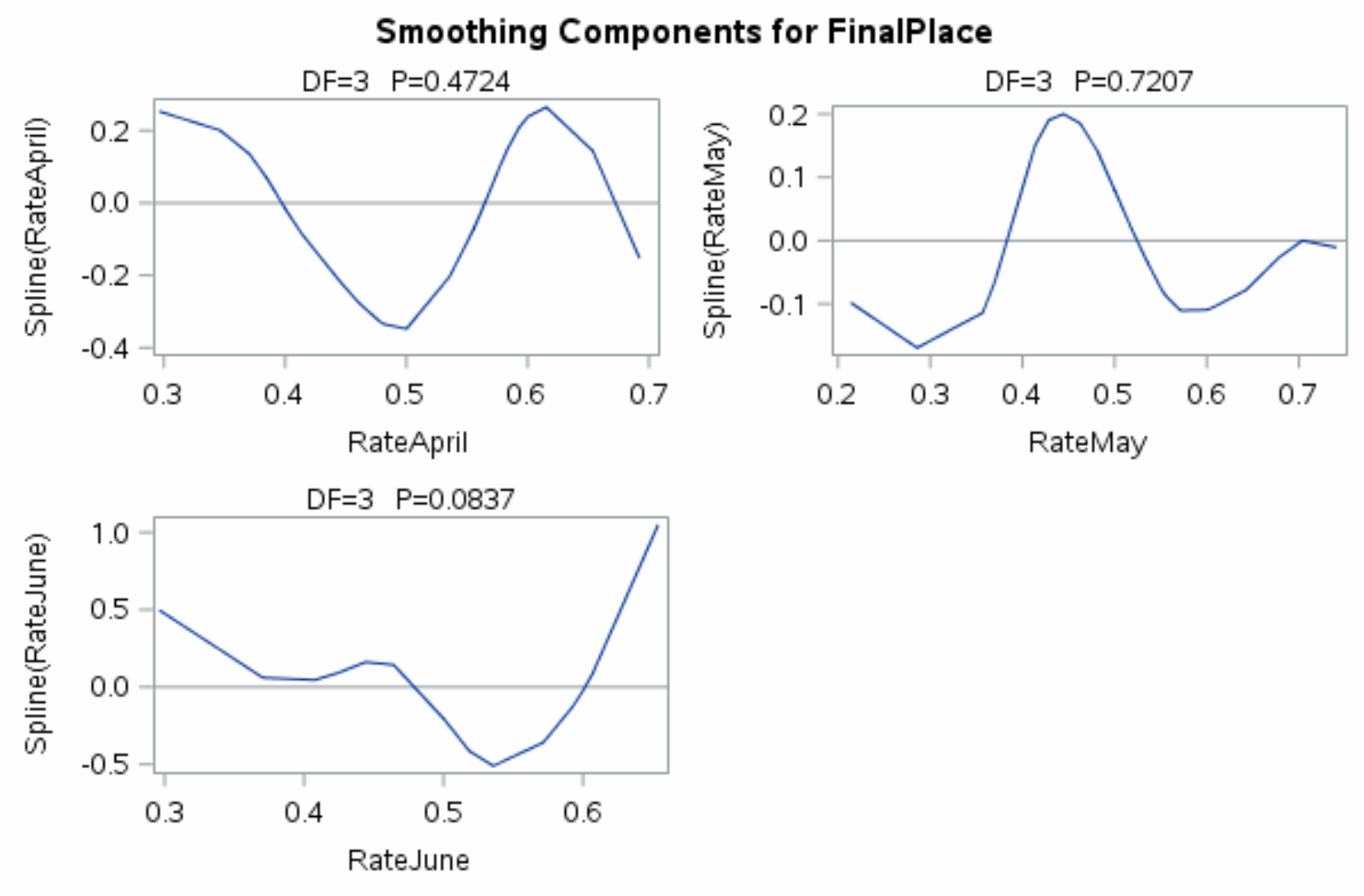
Smoothing Model Analysis
Fit Summary for Smoothing Components

| Component | Smoothing Parameter | DF | GCV | Num Unique Obs |
|---|---|---|---|---|
| Spline(RateApril) | 0.972962 | 3.000000 | 0.791218 | 23 |
| Spline(RateMay) | 0.966501 | 3.000000 | 0.775891 | 22 |
| Spline(RateJune) | 0.841046 | 3.000000 | 0.586709 | 14 |

Smoothing Model Analysis
Approximate Analysis of Deviance

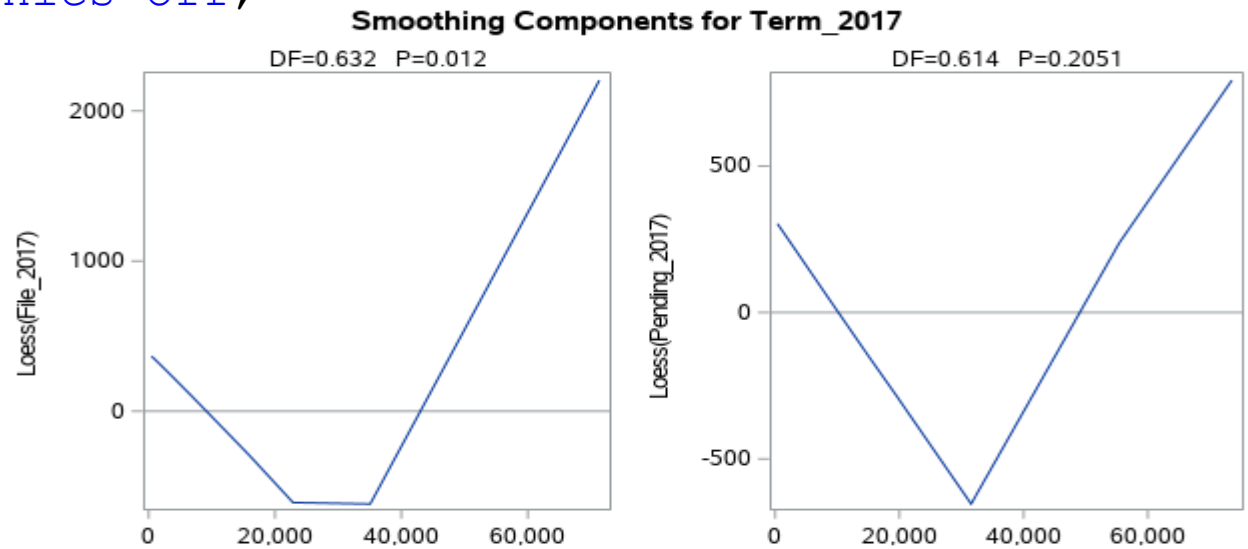
| Source | DF | F Value | Pr > F |
|---|---|---|---|
| Spline(RateApril) | 3.00000 | 0.88 | 0.4724 |
| Spline(RateMay) | 3.00000 | 0.45 | 0.7207 |
| Spline(RateJune) | 3.00000 | 2.63 | 0.0837 |

# Non-Parametric Models: Generalized Additive Models Output and Plots



# Additive Model
# Example 2: Segmented Response Surface

```
ods graphics on;
proc gam data=rm.bankruptcy plots(unpack)=all;
   model term_2017 = loess(file_2017) loess(pending_2017) /
   method=gcv;
   output out=PredGAM p=Gam_p_;
run;
ods graphics off;
```




**An Additive Model used to fit a complex response surface without loss of detail to due piece-wise fitting in local regression**

# Cutting Edge Regression Methods: LASSO, RIDGE, LOESS, GAM

**David J Corliss, PhD**
**Grafham Limited**

**davidjcorliss@gmail.com**