Hypothesis Testing and Model Choice

Merlise Clyde

STA721 Linear Models

Duke University

October 21, 2015

Decomposition

Consider a series of nested models:

$$
\begin{aligned}
\mathcal{M}_0 : \mathbf{Y} &= \mathbf{1}_n\beta_0 + \boldsymbol{\epsilon} \\
\mathcal{M}_1 : \mathbf{Y} &= \mathbf{1}_n\beta_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon} \\
\mathcal{M}_2 : \mathbf{Y} &= \mathbf{1}_n\beta_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \\
\vdots \quad &\quad \vdots \\
\mathcal{M}_k : \mathbf{Y} &= \mathbf{1}_n\beta_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \ldots \mathbf{X}_k\boldsymbol{\beta}_k + \boldsymbol{\epsilon}
\end{aligned}
$$

Let $\mathbf{P}_j$ denote the projection on the column space in each of the models $\mathcal{M}_j$: $C(\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_j)$

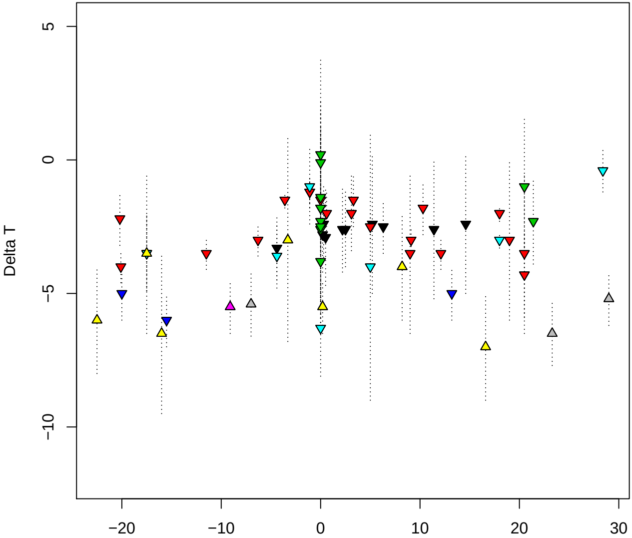
$$
\begin{aligned}
\|\mathbf{Y}^T\mathbf{Y}\|^2 =& \|\mathbf{P}_0\mathbf{Y}\|^2 + \|(\mathbf{P}_1 - \mathbf{P}_0)\mathbf{Y}\|^2 + \|(\mathbf{P}_2 - \mathbf{P}_1)\mathbf{Y}\|^2 + \ldots \|(\mathbf{P}_k - \mathbf{P}_{k-1})\mathbf{Y}\|^2 + \\
& \|(\mathbf{I}_n - \mathbf{P}_k)\mathbf{Y}\|^2
\end{aligned}
$$

Sequential F tests

Hypothesis*	SS	df	F
$\boldsymbol{\beta}_1 = 0$	$\|(\mathbf{P}_1 - \mathbf{P}_0)\mathbf{Y}\|^2$	$r(\mathbf{P}_1) - r(\mathbf{P}_0)$	$\frac{\frac{\|(\mathbf{P}_1 - \mathbf{P}_0)\mathbf{Y}\|^2}{r(\mathbf{P}_1) - r(\mathbf{P}_0)}}{\hat{\sigma}^2}$
$\boldsymbol{\beta}_2 = 0$	$\|(\mathbf{P}_2 - \mathbf{P}_1)\mathbf{Y}\|^2$	$r(\mathbf{P}_2) - r(\mathbf{P}_1)$	$\frac{\frac{\|(\mathbf{P}_2 - \mathbf{P}_1)\mathbf{Y}\|^2}{r(\mathbf{P}_2) - r(\mathbf{P}_1)}}{\hat{\sigma}^2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{\beta}_k = 0$	$\|(\mathbf{P}_k - \mathbf{P}_{k-1})\mathbf{Y}\|^2$	$r(\mathbf{P}_k) - r(\mathbf{P}_{k-1})$	$\frac{\frac{\|(\mathbf{P}_k - \mathbf{P}_{k-1})\mathbf{Y}\|^2}{r(\mathbf{P}_k) - r(\mathbf{P}_{k-1})}}{\hat{\sigma}^2}$

- Sequential test $\boldsymbol{\beta}_j = 0$ includes variables from the previous model $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{j-1}$ but $\boldsymbol{\beta}_i$ for $i > j$ are all set to 0
- All use estimate of $\hat{\sigma}^2 = \|(\mathbf{I}_n - \mathbf{P}_k)\mathbf{Y}\|^2/(n - r(\mathbf{P}_k))$ under largest model
- Unless $\mathbf{P}_j\mathbf{P}_i = \mathbf{0}$ for $i \neq j$, decomposition will depend on the order of $\mathbf{X}_j$ in the model
- If last $\mathbf{X}_k$ is $n \times 1$, then $t^2 = F$ for testing H_0: $\beta_k = 0$

Data



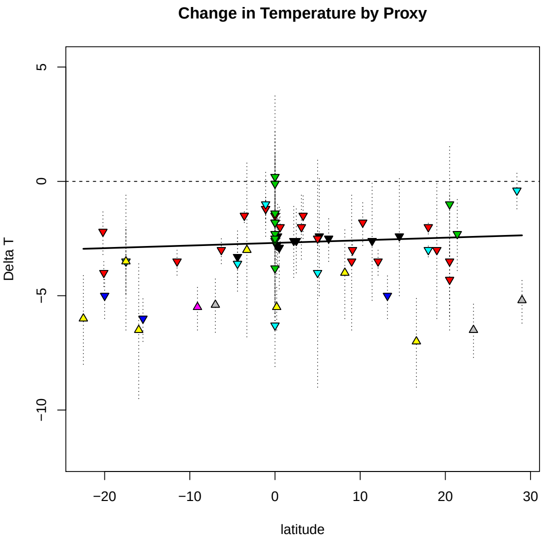
Order 1: Sequential Sum of Squares

```
climate.lm = lm(deltaT ~ proxy *(poly(latitude,2)),
                weights=(1/sdev^2),
                data=climate)
anova(climate.lm)
Response: deltaT
                        Df  Sum Sq Mean Sq F value    Pr(>F)
proxy                    7 307.598  43.943  9.8541 3.848e-07 ***
poly(latitude, 2)        2  10.457   5.228  1.1725    0.3198
proxy:poly(latitude, 2) 12  74.065   6.172  1.3841    0.2126
Residuals               41 182.833   4.459
```

Order 2: Sequential Sum of Squares

```
>anova(lm(deltaT ~ (poly(latitude,2))* proxy, weights=1/sdev^2,
         data=climate))
 Analysis of Variance Table

Response: deltaT
                        Df  Sum Sq Mean Sq F value    Pr(>F)
poly(latitude, 2)        2  79.869  39.935  8.9553 0.0005931 ***
proxy                    7 238.185  34.026  7.6304  6.93e-06 ***
poly(latitude, 2):proxy 12  74.065   6.172  1.3841 0.2125512
Residuals               41 182.833   4.459
```

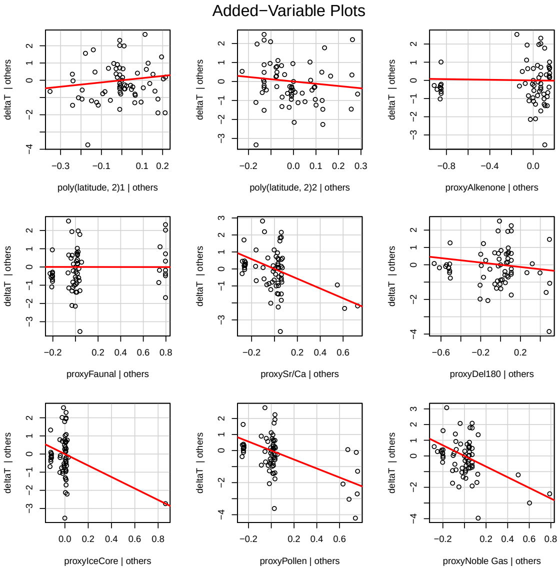
Prediction with Latitude



Added Variable Plots

1. Let $\mathbf{P}_{(-j)}$ denote the projection on the space spanned by $C(\mathbf{X}_0, \ldots, \mathbf{X}_{j-1}, \mathbf{X}_{j+1}, \ldots \mathbf{X}_k)$ (omit variable j)
2. Find residuals $\mathbf{e}_{\mathbf{Y}|\mathbf{X}_{(-j)}} = (\mathbf{I} - \mathbf{P}_{(-j)})\mathbf{Y}$ from regressing $\mathbf{Y}$ on all variables except $\mathbf{X}_j$
3. Remove the effect of other explanatory variables from $\mathbf{X}_j$ by taking residuals $\mathbf{e}_{\mathbf{X}_j|\mathbf{X}_{(-j)}} = (\mathbf{I} - \mathbf{P}_{(-j)})\mathbf{X}_j$
4. Plot $\mathbf{e}_{\mathbf{Y}|\mathbf{X}_{(-j)}}$ versus $\mathbf{e}_{\mathbf{X}_j|\mathbf{X}_{(-j)}}$
5. Slope is adjusted regression coefficient in full model $\boldsymbol{\mu} \in C(\mathbf{X}_0, \ldots, \mathbf{X}_{j-1}, \mathbf{X}_j, \mathbf{X}_{j+1}, \ldots \mathbf{X}_k)$
6. `library(car)`
7. `avPlots(climate1.lm, terms=~ .)`

avPlots



Added−Variable Plots
deltaT | others
poly(latitude, 2)1 | others
poly(latitude, 2)2 | others
proxyAlkenone | others
proxyFaunal | others
proxySr/Ca | others
proxyDel180 | others
proxyIceCore | others
proxyPollen | others
proxyNoble Gas | others

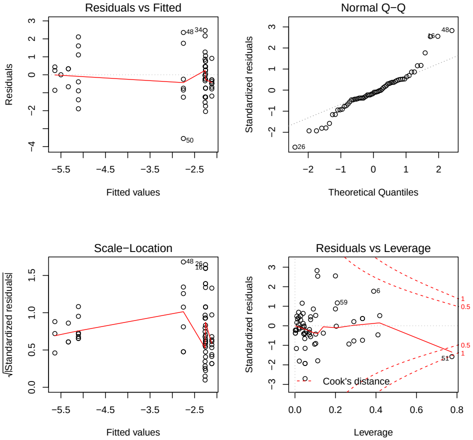
Multiple Model Objects and Anova in R

```
> anova(climate3.lm,climate2.lm,climate1.lm, climate.lm)
Analysis of Variance Table

Model 1: deltaT ~ T.M
Model 2: deltaT ~ poly(latitude, 2) + T.M
Model 3: deltaT ~ poly(latitude, 2) + proxy
Model 4: deltaT ~ proxy * (poly(latitude, 2))
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     61 385.66
2     59 347.11  2    38.542 4.3215 0.019814 *
3     53 256.90  6    90.215 3.3718 0.008552 **
4     41 182.83 12    74.065 1.3841 0.212551
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Other order

```
> anova(climate3.lm,climate2.lm,climate1.lm, climate.lm)
Analysis of Variance Table

Model 1: deltaT ~ T.M
Model 2: deltaT ~ proxy
Model 3: deltaT ~ poly(latitude, 2) + proxy
Model 4: deltaT ~ proxy * (poly(latitude, 2))
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     61 385.66
2     55 267.35  6   118.301 4.4215 0.001555 **
3     53 256.90  2    10.457 1.1725 0.319767
4     41 182.83 12    74.065 1.3841 0.212551
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual Plots



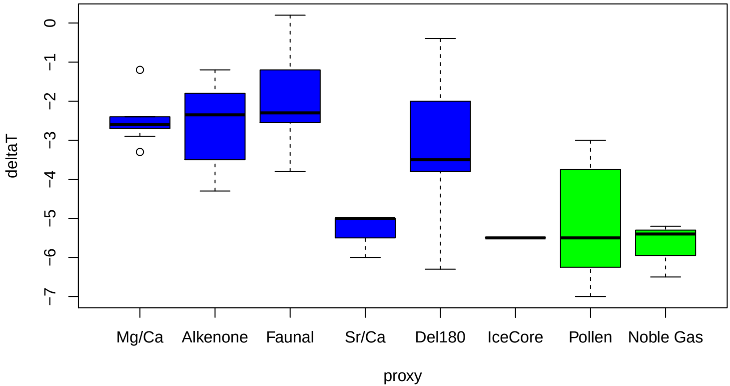
Terrestrial versus Marine

```
climate.final = lm(deltaT ~ T.M + proxy -1, weights=(1/sdev^2))

                Estimate Std. Error t value Pr(>|t|)
T.MT             -5.6360     0.7132  -7.902 1.26e-10 ***
T.MM             -2.1145     0.4124  -5.127 3.93e-06 ***
proxyAlkenone    -0.1408     0.4381  -0.321    0.749
proxyFaunal      -0.1507     0.8971  -0.168    0.867
proxySr/Ca       -3.2188     0.7584  -4.244 8.49e-05 ***
proxyDel180      -0.6378     0.5048  -1.263    0.212
proxyIceCore      0.1360     1.3130   0.104    0.918
proxyPollen       0.5283     1.0033   0.527    0.601
proxyNoble Gas        NA         NA      NA       NA

Multiple R-squared: 0.9115,Adjusted R-squared: 0.8986

          Df  Sum Sq Mean Sq  F value   Pr(>F)
T.M        2 2635.27 1317.63 271.0625 < 2e-16 ***
proxy      6  118.30   19.72   4.0561 0.00195 **
Residuals 55  267.35    4.86
```

Even Simpler ?

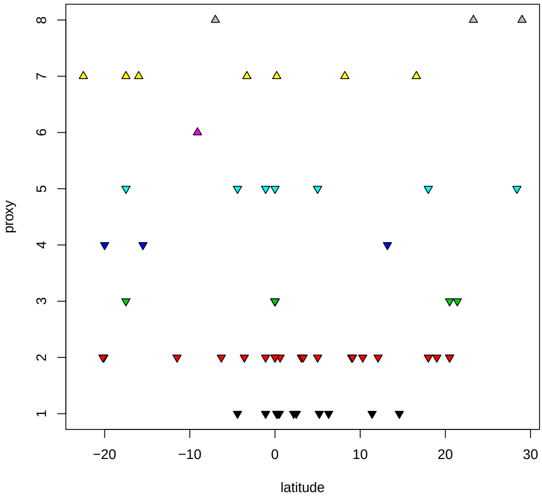
```
lm(formula = deltaT ~ T.M + I(proxy == "Sr/Ca"), weights = (1/sd

                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               -5.3915     0.4486 -12.018  < 2e-16 ***
T.MM                       3.0585     0.4649   6.579 1.30e-08 ***
I(proxy == "Sr/Ca")TRUE  -3.0003     0.6371  -4.709 1.52e-05 ***

Residual standard error: 2.166 on 60 degrees of freedom
Multiple R-squared: 0.5103,Adjusted R-squared: 0.4939

Model 1: deltaT ~ T.M + I(proxy == "Sr/Ca")
Model 2: deltaT ~ T.M + proxy - 1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     60 281.58
2     55 267.36  5    14.228 0.5854  0.711
```

Boxplots



deltaT
0
−1
−2
−3
−4
−5
−6
−7
Mg/Ca
Alkenone
Faunal
Sr/Ca
Del180
IceCore
Pollen
Noble Gas
proxy

Design



Summary

- Ignoring proxies, there are systematic trends with latitude.
- Difference among proxies, even after adjusting for latitude
- Weak evidence of a latitude effect, after taking into account proxies
- Terrestrial sites differ from Marine sites, however there are significant difference among proxies within the Marine group driven by the `Sr/Ca` proxy which indicates a significantly greater increases in temperatures
- Significant warming for Terrestrial ($5.4^\circ C$) with Marine sites significantly cooler ($3^\circ C$)
- `Sr/Ca` proxies are significantly cooler than other marine proxies by about $3^\circ C$

Uncertainty Measures? Normal Assumptions?