

Hypothesis Testing and Model Choice

Merlise Clyde

STA721 Linear Models

Duke University

October 30, 2017

Outline

Topics

- ▶ Climate Example
- ▶ t-tests
- ▶ Overall F-test
- ▶ Sequential F-tests
- ▶ Added Variable Plots
- ▶ Summary

Readings: Christensen Chapter 2 (section 7), Chapter 10, Appendix B

Climate Change ?

Scientists are interested in the Earth's temperature change since the last glacial maximum, about 20,000 years ago.

- ▶ The first study to estimate the temperature change was published in 1980
- ▶ Estimated a change of -1.5 degrees C, ± 1.2 degrees C in tropical sea surface temperatures.
- ▶ The negative value means that the Earth was colder then than now.
- ▶ Since 1980 there have been many other studies, which use different measurement techniques, or proxies.
- ▶ Some proxies can be used over land, others over water.

Proxies

The 8 proxies used are

1. "Mg/Ca" 1
2. "alkenone" 2
3. "Faunal" 3
4. "Sr/Ca" 4
5. "del 180" 5
6. "Ice Core" 6
7. "Pollen" 7
8. "Noble Gas" 8

Variables

```
climate =  
read.table("http://www.stat.duke.edu/courses/Fall10/sta290/datasets/climate.dat",  
header=T)
```

Each of the 53 studies reported

- ▶ deltaT an estimate of the temperature change
- ▶ sdev a standard deviation of that estimate
- ▶ proxy the proxy used (coded 1 to 8),
- ▶ T.M whether it was a terrestrial or marine study (T/M), which is coded as 0 for Terrestrial, 1 for Marine,
- ▶ latitude at which data were collected

Questions of Interest

1. Do estimates vary systematically by proxy?
2. Do terrestrial estimates differ systematically from marine estimates?
3. Do estimates vary systematically by latitude?
4. Can we combine the studies to get a better estimate of the overall temperature change?
5. Are temperatures changing?

Build a larger model or series of models to address these questions?

$$E[\Delta T] = f(\text{Proxy, latitude})$$

Model Building

George E. P. Box

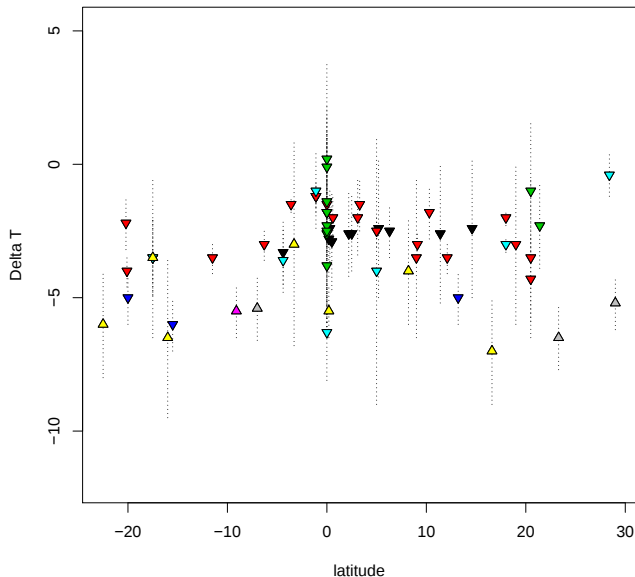
Essentially, all models are wrong, but some are useful. *Empirical Model-Building and Response Surfaces* (1987), co-authored with Norman R. Draper, p. 424

- ▶ “true” model may be a complicated function of latitude, proxy, as well as other (omitted) covariates
- ▶ Assume that for each proxy p , there is a nonlinear relationship between ΔT and latitude l and omitted variables o ; $f(p, l, o)$
- ▶ Taylor’s series expansion about some point l_0 :

$$\begin{aligned} f(p, l, o) &= f(p, l_0, o) + (l - l_0)f'(p, l_0, o) + (l - l_0)^2 \frac{f''(p, l_0, o)}{2} \\ &\quad + R(l, p, o) \\ f(p, l) &\approx \beta_{p0} + \beta_{1p}l + \beta_{2p}l^2 \end{aligned}$$

- ▶ Ignore o and remainder term

A Picture is Worth a Thousand Words



R Model Formula

$\text{DeltaT} \sim \text{proxy} * \text{poly}(\text{latitude}, 2)$

- ▶ Expand out predictors as
`proxy + poly(latitude, 2) + proxy:poly(latitude, 2)`
- ▶ `proxy` is a factor; default coding is to create 8 indicators of each proxy and then drop the column associated with the first level of the factor (MG/Ca in the example)
- ▶ `poly(latitude, 2)` creates an orthonormal basis for a second order polynomial in latitude $[1, l, l^2]$
- ▶ `proxy:poly(latitude, 2)` takes the product of each of the 7 dummy variables for proxy times the linear and quadratic terms for latitude
- ▶ Look at `model.matrix( ~ poly(latitude,2)*proxy, data=climate)`

Estimates

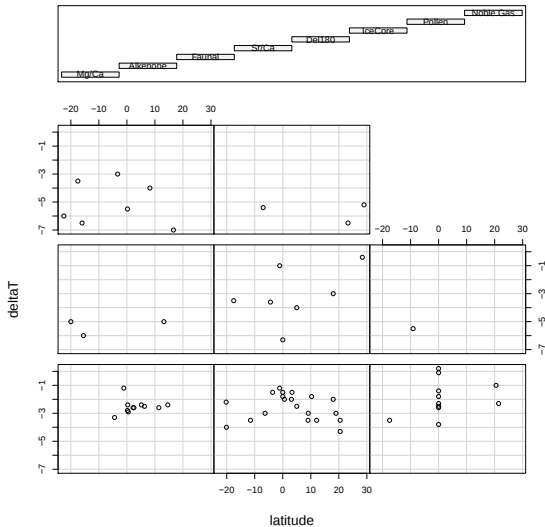
```
> summary(climate.lm)
Coefficients: (2 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.7933     2.3189  -1.205   0.235
Alkenone         0.4463     2.3234   0.192   0.849
Faunal          0.7235     2.4525   0.295   0.769
Sr/Ca          -2.9254     2.5318  -1.155   0.255
Del180         -0.3037     2.4030  -0.126   0.900
IceCore        -3.1407     2.8504  -1.102   0.277
Pollen         -2.6751     2.4528  -1.091   0.282
Noble Gas     -3.2520     2.5698  -1.265   0.213
poly(latitude, 2)1 -3.0092    10.5916  -0.284   0.778
poly(latitude, 2)2 -7.3654    26.6516  -0.276   0.784
Alkenone:poly(latitude, 2)1  3.5493    10.6675   0.333   0.741
Faunal:poly(latitude, 2)1   6.5637    11.7978   0.556   0.581
Sr/Ca:poly(latitude, 2)1  11.8701    15.6097   0.760   0.451
Del180:poly(latitude, 2)1   0.8912    11.7526   0.076   0.940
IceCore:poly(latitude, 2)1      NA         NA      NA      NA
Pollen:poly(latitude, 2)1  -4.0769    13.5600  -0.301   0.765
Noble Gas:poly(latitude, 2)1 -8.7078    17.9962  -0.484   0.631
Alkenone:poly(latitude, 2)2  3.0832    26.6984   0.115   0.909
Faunal:poly(latitude, 2)2   2.8690    27.4056   0.105   0.917
Sr/Ca:poly(latitude, 2)2  19.2753    31.4567   0.613   0.543
Del180:poly(latitude, 2)2  16.1802    26.9623   0.600   0.552
IceCore:poly(latitude, 2)2      NA         NA      NA      NA
Pollen:poly(latitude, 2)2   3.3119    27.6753   0.120   0.905
Noble Gas:poly(latitude, 2)2 18.6612    30.0579   0.621   0.538

Residual standard error: 2.112 on 41 degrees of freedom
Multiple R-squared: 0.682, Adjusted R-squared: 0.5191
F-statistic: 4.187 on 21 and 41 DF, p-value: 4.382e-05
```

Conditional Plot

```
coplot(deltaT ~ latitude | proxy, data=climate)
```

Given : proxy



Estimates and t-statistics

- ▶ MLEs do not depend on the order of the variables in the model
- ▶ regression coefficients are adjusted for the other variables in the model
- ▶ t-statistics

$$\frac{\lambda^T \beta - \lambda^T b_0}{\hat{\sigma} \sqrt{\lambda^T (\mathbf{X}^T \mathbf{X})^{-1} \lambda}} \sim t(n - p, 0, 1)$$

under the hypothesis $\lambda^T b_0 = 0$

- ▶ t-values correspond to test statistic for testing hypothesis $H_o: \beta_j = 0$ versus $H_a: \beta_j \neq 0$ given the other variables are in the model
- ▶ all p-values greater than α does not mean that all coefficients are zero!
- ▶ redundancy
- ▶ with factors use ANOVA for simultaneous testing

Anova and 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			

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			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Order Matters!

Decomposition

Consider a series of nested models:

$$\mathcal{M}_0 : \mathbf{Y} = \mathbf{1}_n \beta_0 + \epsilon$$

$$\mathcal{M}_1 : \mathbf{Y} = \mathbf{1}_n \beta_0 + \mathbf{X}_1 \beta_1 + \epsilon$$

$$\mathcal{M}_2 : \mathbf{Y} = \mathbf{1}_n \beta_0 + \mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \epsilon$$

$$\vdots \quad \quad \vdots$$

$$\mathcal{M}_k : \mathbf{Y} = \mathbf{1}_n \beta_0 + \mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \dots \mathbf{X}_k \beta_k + \epsilon$$

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

$$\|\mathbf{Y}\|^2 = \|\mathbf{P}_0 \mathbf{Y}\|^2 + \|(\mathbf{P}_1 - \mathbf{P}_0) \mathbf{Y}\|^2 + \|(\mathbf{P}_2 - \mathbf{P}_1) \mathbf{Y}\|^2 + \dots \|(\mathbf{P}_k - \mathbf{P}_{k-1}) \mathbf{Y}\|^2 + \|\mathbf{P}_k \mathbf{Y}\|^2$$

F tests

The F statistic

$$F = \frac{\|(\mathbf{P}_k - \mathbf{P}_{k-1})\mathbf{Y}\|^2 / (r(\mathbf{P}_k) - r(\mathbf{P}_{k-1}))}{\hat{\sigma}^2} \sim F(r(\mathbf{P}_k) - r(\mathbf{P}_{k-1}), n - p)$$

under the null hypothesis.

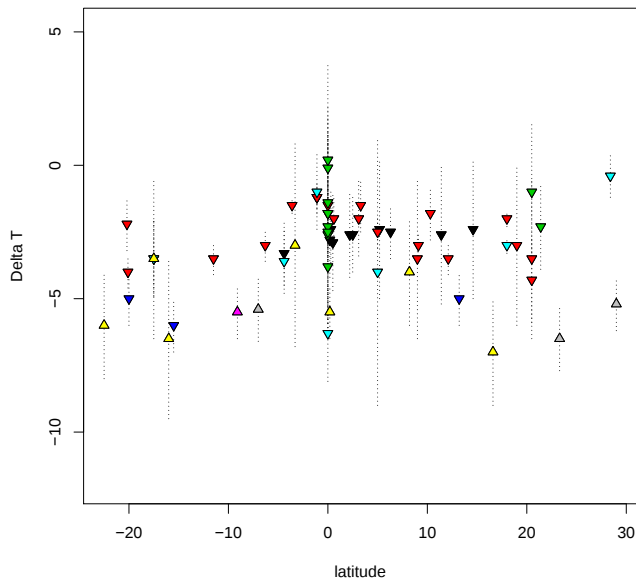
- ▶ Numerator is a χ^2 over df
- ▶ Denominator is a χ^2 over df
- ▶ numerator and denominator are independent
- ▶ Nested models $C(M_k)$ contains $C(M_{k-1})$

Sequential F tests

Hypothesis*	SS	df	F
$\beta_1 = 0$	$\ (\mathbf{P}_1 - \mathbf{P}_0)\mathbf{Y}\ ^2$	$r(\mathbf{P}_1) - r(\mathbf{P}_0)$	$\frac{\frac{\ (\mathbf{P}_1 - \mathbf{P}_0)\mathbf{Y}\ ^2}{r(\mathbf{P}_1) - r(\mathbf{P}_0)}}{\hat{\sigma}^2}$
$\beta_2 = 0$	$\ (\mathbf{P}_2 - \mathbf{P}_1)\mathbf{Y}\ ^2$	$r(\mathbf{P}_2) - r(\mathbf{P}_1)$	$\frac{\frac{\ (\mathbf{P}_2 - \mathbf{P}_1)\mathbf{Y}\ ^2}{r(\mathbf{P}_2) - r(\mathbf{P}_1)}}{\hat{\sigma}^2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\beta_k = 0$	$\ (\mathbf{P}_k - \mathbf{P}_{k-1})\mathbf{Y}\ ^2$	$r(\mathbf{P}_k) - r(\mathbf{P}_{k-1})$	$\frac{\frac{\ (\mathbf{P}_k - \mathbf{P}_{k-1})\mathbf{Y}\ ^2}{r(\mathbf{P}_k) - r(\mathbf{P}_{k-1})}}{\hat{\sigma}^2}$

- ▶ Sequential test $\beta_j = 0$ includes variables from the previous model $\beta_0, \beta_1, \dots, \beta_{j-1}$ but β_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

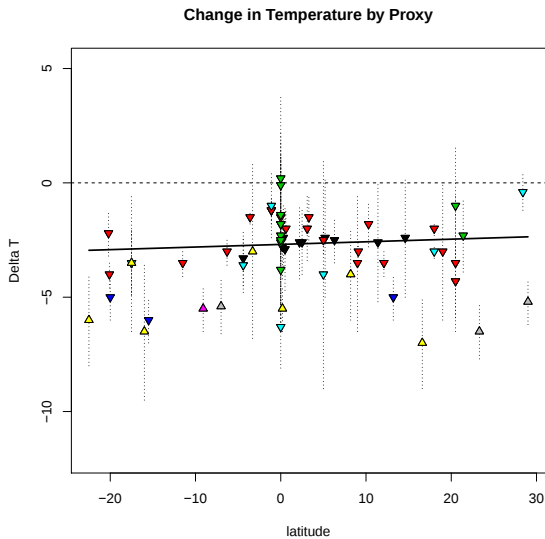
```
>anova(lm(deltaT ~ (poly(latitude,2))* proxy, weights=1/sdev^2,  
        data=climate))
```

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 $C(\mathbf{X}_0, \dots, \mathbf{X}_{j-1}, \mathbf{X}_{j+1}, \dots, \mathbf{X}_k)$ (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
 $\mu \in C(\mathbf{X}_0, \dots, \mathbf{X}_{j-1}, \mathbf{X}_j, \mathbf{X}_{j+1}, \dots, \mathbf{X}_k)$
6. `library(car)`
7. `avPlots(climate1.lm, terms=~.)`

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 ' ' 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

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

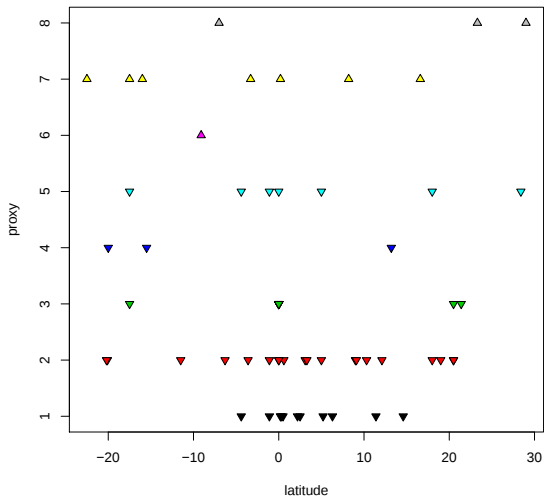
Multiple R-squared: 0.5103, Adjusted R-squared: 0.4939

Model 1: $\text{deltaT} \sim \text{T.M} + \text{I}(\text{proxy} == \text{"Sr/Ca"})$

Model 2: $\text{deltaT} \sim \text{T.M} + \text{proxy} - 1$

	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

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 (potential confounding)
- ▶ 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°C) with Marine sites significantly cooler (3°C)
- ▶ Sr/Ca proxies are significantly cooler than other marine proxies by about 3°C

Uncertainty Measures? Normal Assumptions?