Statistics 512: Applied Linear Models

Topic 5a

Topic Overview

This topic will cover

- Ridge Regression

Ridge Regression (Section 11.2)

Some Remedial Measures for Multicollinearity

- Restrict the use of the regression model to infererence on values of predictor variables that follow the same pattern of multicollinearity.

  For example, suppose a model has three predictors: X_1, X_2, X_3. The distribution of (X_1, X_2, X_3) is $N(\mu, \Sigma)$ for some mean vector μ and covariance matrix Σ. If future predictor values come from this distribution, *even if there is serious multicollinearity*, inferences for the predictions using this model are still useful.

- If the model is a polynomial regression model, use centered variables.
- Drop one or more predictor variables (i.e., variable selection).
  - Standard errors on the parameter estimates decrease.
  - *However,* how can we tell if the dropped variable(s) give us any useful information.
  - If the variable is important, the parameter estimates become *biased up.*
- Sometimes, observations can be *designed* to break the multicollinearity.
- Get coefficient estimates from additional data from other contexts.

  For instance, if the model is

  $$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i,$$

  and you have an estimator b_1 (for β_1 based on another data set, you can estimate β_2 by regressing the adjusted variable $Y_i' = Y_i - b_1 X_{i,1}$ on $X_{i,2}$. (Common example: in economics, using cross-sectional data to estimate parameters for a time-dependent model.)

- Use the first few *principal components* (or *factor loadings*) of the predictor variables. (Limitation: may lose interpretability.)
- *Biased Regression* or *Coefficient Shrinkage* (Example: Ridge Regression)

Two Equivalent Formulations of Ridge Regression

Ridge regression shrinks estimators by "penalyzing" their size. (Penalty: $\lambda \sum \beta_j^2$)

Penalized Residual Sum of Squares:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (Y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- λ controls the amount of shrinkage of the parameter estimates
- Large $\lambda \rightarrow$ greater shrinkage (toward zero)

Equivalent Representation:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2,$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq s.$$

- There is a direct relationship between λ and s (although we will usually talk about λ).
- The intercept β_0 is not subject to the shrinkage penalty.

Matrix Representation of Solution

$$\hat{\beta}^{ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

KNNL Example page 256

- SAS code in `ridge.sas`
- 20 healthy female subjects ages 25-34
- Y is fraction body fat
- X_1 is triceps skin fold thickness
- X_2 is thigh circumference
- X_3 is midarm circumference
- Conclusion from previous analysis: could have good model with thigh only or midarm and thickness only.

Input the data

```
data bodyfat;
     infile 'H:\System\Desktop\CH07TA01.dat';
     input skinfold thigh midarm fat;
proc print data = bodyfat;
proc reg data = bodyfat;
     model fat = skinfold thigh midarm;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Root MSE	2.47998	R-Square	0.8014
Dependent Mean	20.19500	Adj R-Sq	0.7641
Coeff Var	12.28017		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	117.08469	99.78240	1.17	0.2578
skinfold	1	4.33409	3.01551	1.44	0.1699
thigh	1	-2.85685	2.58202	-1.11	0.2849
midarm	1	-2.18606	1.59550	-1.37	0.1896

None of the p-values are significant.

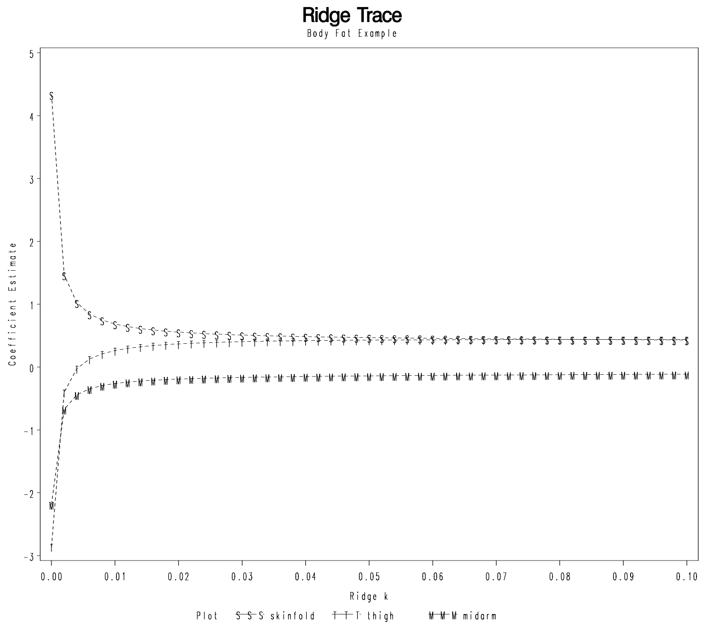
Pearson Correlation Coefficients, N = 20

	skinfold	thigh	midarm	fat
skinfold	1.00000	0.92384	0.45778	0.84327
thigh	0.92384	1.00000	0.08467	0.87809
midarm	0.45778	0.08467	1.00000	0.14244
fat	0.84327	0.87809	0.14244	1.00000

Try Ridge Regression

```
proc reg data = bodyfat
     outest = bfout ridge = 0 to 0.1 by 0.003;
     model fat = skinfold thigh midarm / noprint;
     plot / ridgeplot nomodel nostat;
```

Ridge Trace



Each value of λ (or `Ridge k` in SAS) gives different values of the parameter estimates. (Note the instability of the estimate values for small λ.)

How to Choose λ

Things to look for

- Get the variance inflation factors (VIF) close to 1
- Estimated coefficients should be "stable"
- look for only "modest" change in R^2 or $\hat{\sigma}$.

```
title2 'Variance Inflation Factors';
proc gplot data = bfout;
     plot (skinfold thigh midarm)* _RIDGE_ / overlay;
     where _TYPE_ = 'RIDGEVIF';
run;
```

Graph the VIF's

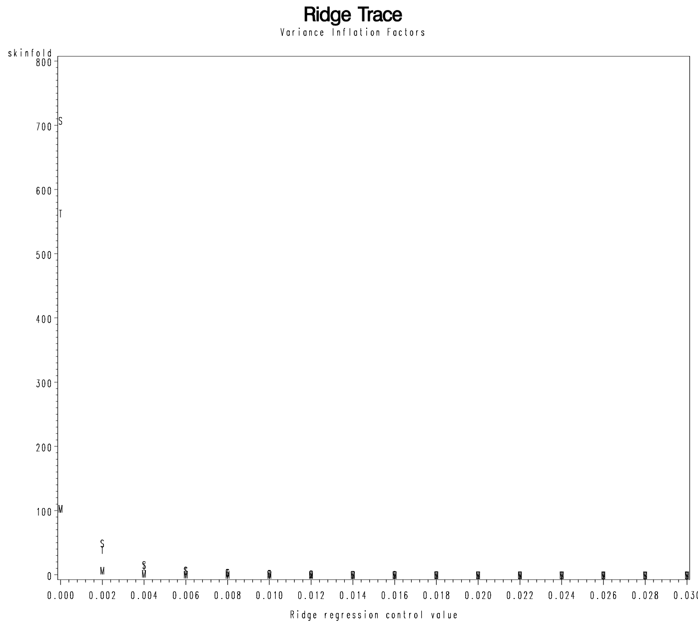


Chart the Estimates and Errors for different λ values

```
proc print data = bfout;
     var _RIDGE_ skinfold thigh midarm;
     where _TYPE_ = 'RIDGEVIF';
proc print data = bfout;
     var _RIDGE_ _RMSE_ Intercept skinfold thigh midarm;
     where _TYPE_ = 'RIDGE';
```

Variance Inflation Factors

Obs	_RIDGE_	skinfold	thigh	midarm
2	0.000	708.843	564.343	104.606
4	0.002	50.559	40.448	8.280
6	0.004	16.982	13.725	3.363
8	0.006	8.503	6.976	2.119
10	0.008	5.147	4.305	1.624
12	0.010	3.486	2.981	1.377
14	0.012	2.543	2.231	1.236
16	0.014	1.958	1.764	1.146
18	0.016	1.570	1.454	1.086
20	0.018	1.299	1.238	1.043
22	0.020	1.103	1.081	1.011
24	0.022	0.956	0.963	0.986
26	0.024	0.843	0.872	0.966
28	0.026	0.754	0.801	0.949
30	0.028	0.683	0.744	0.935
32	0.030	0.626	0.697	0.923

Note that at `RIDGE = 0.020`, the VIF's are close to 1.

Parameter Estimates

Obs	_RIDGE_	_RMSE_	Intercept	skinfold	thigh	midarm

3	0.000	2.47998	117.085	4.33409	-2.85685	-2.18606
5	0.002	2.54921	22.277	1.46445	-0.40119	-0.67381
7	0.004	2.57173	7.725	1.02294	-0.02423	-0.44083
9	0.006	2.58174	1.842	0.84372	0.12820	-0.34604
11	0.008	2.58739	-1.331	0.74645	0.21047	-0.29443
13	0.010	2.59104	-3.312	0.68530	0.26183	-0.26185
15	0.012	2.59360	-4.661	0.64324	0.29685	-0.23934
17	0.014	2.59551	-5.637	0.61249	0.32218	-0.22278
19	0.016	2.59701	-6.373	0.58899	0.34131	-0.21004
21	0.018	2.59822	-6.946	0.57042	0.35623	-0.19991
23	0.020	2.59924	-7.403	0.55535	0.36814	-0.19163
25	0.022	2.60011	-7.776	0.54287	0.37786	-0.18470
27	0.024	2.60087	-8.083	0.53233	0.38590	-0.17881
29	0.026	2.60156	-8.341	0.52331	0.39265	-0.17372
31	0.028	2.60218	-8.559	0.51549	0.39837	-0.16926
33	0.030	2.60276	-8.746	0.50864	0.40327	-0.16531

Note that at `RIDGE = 0.020`, the $RMSE$ is only increased by 5% (so SSE increase by about 10%), and the parameter estimates are closer to making sense.

Conclusion

So the solution at $\lambda = 0.02$ with parameter estimates (-7.4, 0.56, 0.37, -0.19) seems to make the most sense.

Notes

- The book makes a big deal about standardizing the variables... SAS does this for you in the `ridge` option.
- Why ridge regression? Estimates tend to be more stable, particularly outside the region of the predictor variables: less affected by small changes in the data. (Ordinary LS estimates can be highly unstable when there is lots of multicollinearity.)
- Major drawback: ordinary inference procedures d't work so well.
- Other procedures use different penalties, e.g. "Lasso" penalty: $\sum |\beta_j|$.