Statistics for the Sciences

Model Selection for Multiple Linear Regression Models

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- Detecting Influential Points
- Detecting Multicollinearity
- Partial correlation
- Model Selection methods
- Criteria for Model Selection
- Cross-Validation (CV)
- Lab

Detecting Influential Points

- Frequently in regression analysis applications, the data set contains some cases which are called **influential points**: points that are strongly inconsistent with the regression model.
- Methods for detecting influential points are related to the concept of PRESS residuals.
- **PRESS residuals** are computed as follows:
  - First, observation (X_i, Y_i) is omitted from the data and the least squares line fit to the remaining data, giving the parameter estimates $\hat{\beta}_0^{(i)}, \hat{\beta}_1^{(i)}, \ldots, \hat{\beta}_k^{(i)}$.
  - Next, the **deleted fitted value**, $\hat{Y}_{(i)} = \hat{\beta}_0^{(i)} + \hat{\beta}_1^{(i)} X_{1i} + \cdots \hat{\beta}_k^{(i)} X_{ki}$ is computed, $i = 1, \ldots, n$.
  - Then, the **deleted residual** or **PRESS residuals** $e_{(i)} = Y_i - \hat{Y}_{(i)}$ is computed, $i = 1, \ldots, n$.

Detecting Influential Points

- (1) **DFFITS** is used to identify influential data points that have a substantial impact on the n fitted values.

$$|(\text{DFFITS})_i| \geq 1, i = 1, \ldots, n$$

is considered extreme for small to mediate data sets and $|(\text{DFFITS})_i| \geq 2\sqrt{(k+1)/n}, i = 1, \ldots, n$ is considered extreme for large data sets.

Detecting Influential Points

- ② Cook's Distance - Influence on all fitted values:

In contrast to the DFFITS measure which considers the influence of the ith case on the fitted value Y_i for this case, Cook's distance measure considers the influence of the ith case on all n fitted values.

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{(k+1)MSE}, i = 1, \ldots, n$$

- A D_i is considered large if

$$D_i \geq F_{0.25,k+1,n-1-k} \quad \text{(upper quartile of the F distribution).}$$

- A common rule of thumb is that an observation is considered influential if $D_i > 4/n$.

Detecting Influential Points

- 3. DFBETAS - Influence on the Regression Coefficients: DFBETAS is used to assess the influence of each individual data point on the estimated regression coefficients.
- A DFBETA displays the (studentized) change in the jth estimated slope when the ith data point is deleted from the data. It is extreme if

$$|\text{DEBETAS}_{j(i)}| \geq 1$$

for small/medium sized data sets and $|\text{DEBETAS}_{j(i)}| \geq 2/\sqrt{n}$ is considered extreme for large data sets.

Detecting Multicollinearity

- A very desirable condition in a set of regression data is to have predictors that are not "moving with each other" in the data set. Linear dependencies render it more difficult to sort out the impact of each predictor on the response.
  - If two predictors are highly (linearly) correlated, one should be removed.
- **Multicollinearity** simply occurs when there are near linear dependencies among the predictors.
  - Variances of the ls estimates of the regression parameters will be inflated due to collinearity.

Detecting Multicollinearity

- ① $(VIF)_j$ is called the **variance inflation factor** (VIF) for $\widehat{\beta}_j^*, j = 1, 2, \ldots, k$.
- It can be shown that

$$(VIF)_j = (1 - R_j^2)^{-1},\ j = 1, 2, \ldots, k$$

  where R_j^2 is the coefficient of multiple determination when X_j is regressed on the $k - 1$ other X variables in the model.
- The largest VIF value among all X variables is often used as an indicator of the severity of multicollinearity. It is generally believed that if any VIF exceeds 10, there is a reason for at least some concern.
- ② $tolerance_j = 1/(VIF)_j, j = 1, 2, \ldots, k$.

Partial correlation

- SSR is the variation explained by a multiple linear regression model
  - When more predictor variables are added to the model, SSR is always increased
- We want to check the marginal reduction in SSE when more predictor variables are added to the model

For example,

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

and a reduced model

$$E(Y) = \beta_0 + \beta_1 x_1.$$

- $SSE(X_1, X_2)$ measures the variation in Y when both X_1 and X_2 are included in the model.
- $SSE(X_1)$ measures the variation in Y when X_1 is included in the model.
  - Which is larger? $SSE(X_1)$.
- A natural question is if the predictor variable X_2 can be eliminated from the full model.
  - The relative marginal reduction in the variation in Y associated with X_2 when X_1 is already in the model is:

$$\frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)}.$$

Partial correlation

- This measure is the **coefficient of partial determination** between Y and X_2, given that X_1 is in the model.

$$R^2_{Y2|1} = \frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)}.$$

- $R^2_{Y2|1}$ thus measures the proportionate reduction in the variation of Y remaining gained by including X_2 in the model when X_1 is already in the model.

Partial correlation

- The generalization of coefficients of partial determination to three or more X variables in the model is immediate. For instance:

$$
\begin{aligned}
R^2_{Y1|23} &= \frac{SSR(X_1|X_2, X_3)}{SSE(X_2, X_3)} \\
R^2_{Y2|13} &= \frac{SSR(X_2|X_1, X_3)}{SSE(X_1, X_3)} \\
R^2_{Y3|12} &= \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)} \\
R^2_{Y4|123} &= \frac{SSR(X_4|X_1, X_2, X_3)}{SSE(X_1, X_2, X_3)}
\end{aligned}
$$

- The entries to the left of the vertical bar show X variable(s) being **added**.
- The entries to the right of the vertical bar show the X variables **already in the model**.

Partial correlation

- **Coefficients of Partial Correlation** The square root of a coefficient of partial determination is called a **coefficient of partial correlation**. It is given the same **sign** as that of the corresponding **regression coefficient** in the fitted regression function.
  - For example, $r_{Y3|12}$ has the same sign as $\widehat{\beta}_3$
- Coefficients of partial correlation are frequently used in practice, although they do not have as clear a meaning as coefficients of partial determination.
- It is used to find the best predictor **variable to be selected next for inclusion** in the regression model.

Model Selection

- Model Selection
  - Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response? (F-test in ANOVA)
  - Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- Variable selection
  - The most direct approach is called all subsets or best subsets regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.
  - However we often can't examine all possible models, since they are 2^p of them; for example when $p = 40$ there are over a billion models! Instead we need an automated approach that searches through a subset of them.

Some Model Selection Methods

- **Forward Stepwise Selection** involves starting with no predictors in the model and adding significant ones one at a time, testing each addition for statistical significance. At each step, if a predictor's p-value is less than the significance level for entry (sle), it is added, and the model is then checked to see if any included predictors should be dropped based on the significance level for staying (sls). This process continues until no further predictors can be added or removed, ensuring the model is optimized for predictive accuracy.
- **Backward Elimination** starts with all variables in the model and iteratively removes the least statistically significant variable, identified by the largest p-value. This process continues, refitting the model each time, until all remaining variables have p-values below a specified significance threshold.
- **Stepwise Selection** is a combination of Forward Selection and Backward Elimination. It adds and removes predictors simultaneously based on which action improves the model the most according to some criterion (e.g., AIC).

Criteria for Model Selection

- 1. **Adjusted** R^2: Since R^2 does not take account of the number of parameters in the regression model, the adjusted coefficient of multiple determination R_a^2 has been suggested as an alternative criterion:

$$R_a^2 = 1 - \frac{\mathrm{SSE}/(n-1-k)}{SS_{total}/(n-1)} = 1 - \left(\frac{n-1}{n-1-k}\right)\frac{SSE}{SST}.$$

Criteria for Model Selection

- (2) Mallows' C_p Criterion: Mallows' C_p criterion is a measure used to assess the fit of a regression model. It focuses on the total mean squared error (MSE) of the predicted values for each subset regression model.

$$C_p = \frac{SSE_p}{MSE(X_1, \ldots, X_k)} - (n - 2p)$$

  where $p = 1 + k$, where SSE_p is the error sum of squares for the fitted subset regression model with $p = k + 1$ parameters (k X variables).
- The model with the smallest C_p value is preferred, and $C_p = p$ suggests no estimated bias. A C_p value much larger than p indicates a heavily biased model.

Criteria for Model Selection

- (3) *AIC* **and** *BIC* Criteria: Two popular alternatives that also provide penalties for adding predictors are Akaike's Information Criterion (AIC) and Schwarz' Bayesian Criterion (SBC, also known as BIC). We search for models that have **small values** of AIC or BIC, where these criteria are given by:

$$AIC_p = n \ln SSE_p - n \ln n + 2p$$
$$SBC_p = n \ln SSE_p - n \ln n + (\ln n)p$$

- *AIC* is derived from the principle of **maximum likelihood estimation** with an added penalty term $2p$ to account for the number of parameters, balancing fit and complexity.
- *BIC* is derived from an approximation of the Bayes factor, which compares the posterior probabilities of two models, incorporating a likelihood-based penalty $(\ln n)p$ adjusted by the sample size,
- Both criteria are used to select models that provide a good fit to the data while penalizing excessive complexity to avoid overfitting.

Criteria for Model Selection

- (4) $PRESS_p$ Criterion
- Recall the definition of PRESS residuals.
- The $PRESS_p$ Criterion is the sum of the squared PRESS residuals over all n cases

$$PRESS_p = \sum_{i=1}^{n} (Y_i - \hat{Y}_{(i)})^2.$$

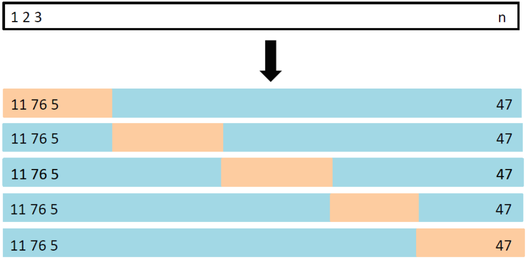
- Models with small $PRESS_p$ values are considered good candidate models.

Cross Validation

- As we fit more complex models - e.g. models with more variables - the fitting error will always decrease.
  - But the model could be very bad at predictions. We should try avoid overfitting.
- In machine learning, a technique Cross-Validation (CV) is used: Split the observations into training set and validation set.
  - Fit the model using the training set.
  - Check the performance of the model using the validation set.
- A popular CV method is K-fold cross-validation (K is chosen to be 5 or 10)

Cross Validation

- 5-Fold Cross-Validation: Split the observations into 5 sets. Repeatedly train the model on 4 sets and evaluate its performance on the 5th.



Cross Validation

A generalization of K-fold cross-validation:

- 1. Split the n observations into K equally-sized folds.
- 2. For $k = 1, \ldots, K$:
  - a. Fit the model using the observations not in the kth fold.
  - b. Let e_k denote the test error (MSE or square root of the MSE) for the observations in the kth fold.
- 3. Calculate $\sum_{k=1}^{K} e_k$, the total CV error.

- We compare available models, and select the model with least test error. Also, it can give an idea of the test error of the final chosen model.

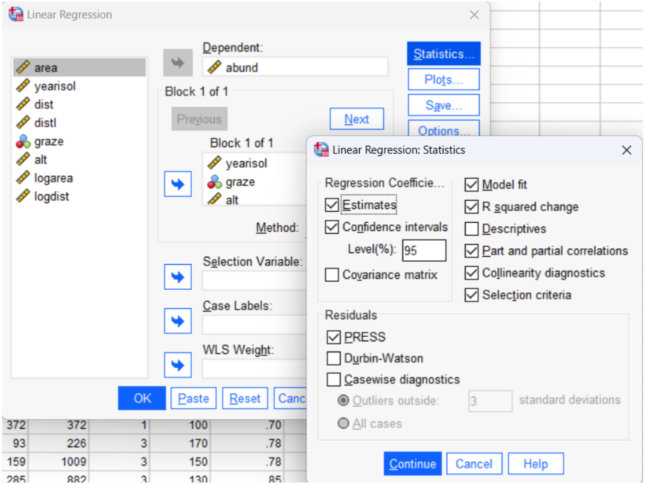
Lab

- Consider the data in the last lab `loyn.csv`: Loyn (1987) selected 56 forest patches in southeastern Victoria, Australia, and related the abundance of forest birds in each patch to six predictor variables: patch area (ha), distance to nearest patch (km), distance to nearest larger patch (km), grazing stock (1 to 5 indicating light to heavy), altitude (m) and years since isolation (years).
  - Add log10 transformation of `area` and `dist`

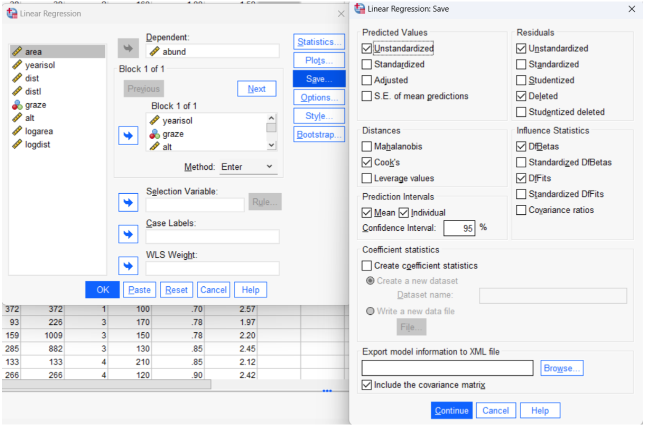
```
##    abund area yearisol dist distl graze alt   logarea  logdist
## 1    5.3  0.1     1968   39    39     2 160 -1.00000 1.591065
## 2    2.0  0.5     1920  234   234     5  60 -0.30103 2.369216
## 3    1.5  0.5     1900  104   311     5 140 -0.30103 2.017033
## 4   17.1  1.0     1966   66    66     3 160  0.00000 1.819544
## 5   13.8  1.0     1918  246   246     5 140  0.00000 2.390935
## 6   14.1  1.0     1965  234   285     3 130  0.00000 2.369216
## 7    3.8  1.0     1955  467   467     5  90  0.00000 2.669317
## 8    2.2  1.0     1920  284  1829     5  60  0.00000 2.453318
## 9    3.3  1.0     1965  156   156     4 130  0.00000 2.193125
## 10   3.0  1.0     1900  311   571     5 130  0.00000 2.492760
```

Lab

- Click on `Analyze` → `Regression` → `Linear ...`



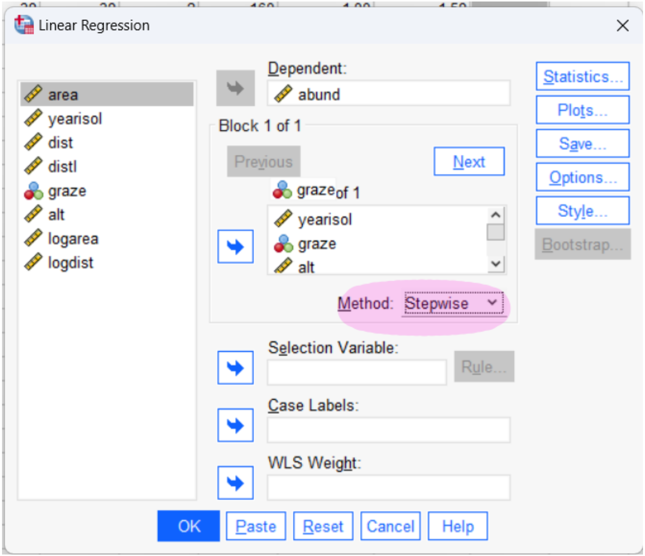
Lab



Linear Regression
area
yearisol
dist
distl
graze
alt
logarea
logdist
Dependent:
abund
Block 1 of 1
Previous
Next
Block 1 of 1
yearisol
graze
alt
Method: Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Style...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Linear Regression: Save
Predicted Values
Unstandardized
Standardized
Adjusted
S.E. of mean predictions
Residuals
Unstandardized
Standardized
Studentized
Deleted
Studentized deleted
Distances
Mahalanobis
Cook's
Leverage values
Influence Statistics
DfBetas
Standardized DfBetas
DfFits
Standardized DfFits
Covariance ratios
Prediction Intervals
Mean
Individual
Confidence Interval: 95 %
Coefficient statistics
Create coefficient statistics
Create a new dataset
Dataset name:
Write a new data file
File...
Export model information to XML file
Browse...
Include the covariance matrix
Continue
Cancel
Help

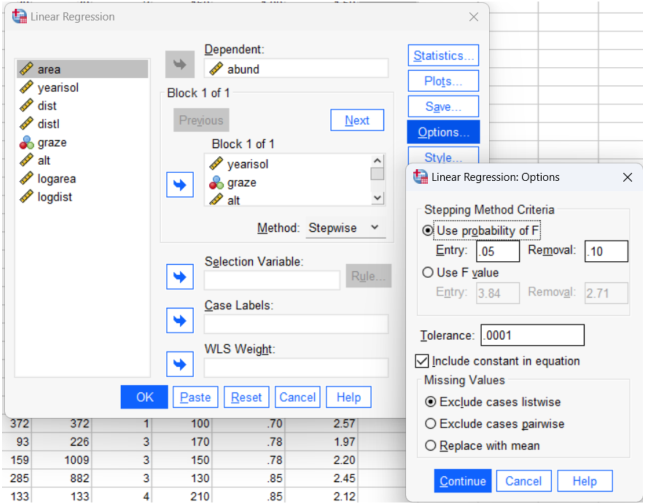
Lab

- Conduct stepwise model selection



Lab

- You can decide the significance level of entry/removal



Lab

- stepwise regression results

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	logarea	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	graze	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: abund

Model Summary[c]

					Change Statistics					Selection Criteria				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Akaike Information Criterion	Amemiya Prediction Criterion	Mallows' Prediction Criterion	Schwarz Bayesian Criterion	PRESS
1	.740[a]	.548	.539	7.2864	.548	65.377	1	54	<.001	224.396	.486	19.650	228.447	
2	.808[b]	.653	.640	6.4442	.105	16.038	1	53	<.001	211.592	.387	5.006	217.668	2453.887

a. Predictors: (Constant), logarea

b. Predictors: (Constant), logarea, graze

c. Dependent Variable: abund

Lab

- ANOVA table

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3470.985	1	3470.985	65.377	<.001[b]
	Residual	2866.943	54	53.092		
	Total	6337.929	55			
2	Regression	4136.984	2	2068.492	49.810	<.001[c]
	Residual	2200.945	53	41.527		
	Total	6337.929	55			

a. Dependent Variable: abund

b. Predictors: (Constant), logarea

c. Predictors: (Constant), logarea, graze

Lab

- LS estimates and statistical inferences

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound	Correlations Zero-order	Correlations Partial	Correlations Part	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	10.401	1.489		6.984	<.001	7.415	13.387					
	logarea	9.778	1.209	.740	8.086	<.001	7.354	12.203	.740	.740	.740	1.000	1.000
2	(Constant)	21.603	3.092		6.987	<.001	15.402	27.804					
	logarea	6.890	1.290	.521	5.341	<.001	4.303	9.478	.740	.592	.432	.687	1.455
	graze	-2.854	.713	-.391	-4.005	<.001	-4.283	-1.424	-.683	-.482	-.324	.687	1.455

a. Dependent Variable: abund

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance	Collinearity Statistics VIF	Collinearity Statistics Minimum Tolerance
1	yearisol	.322[b]	3.774	<.001	.460	.922	1.084	.922
	graze	-.391[b]	-4.005	<.001	-.482	.687	1.455	.687
	alt	.197[b]	2.138	.037	.282	.924	1.082	.924
	logdist	-.107[b]	-1.113	.271	-.151	.909	1.100	.909
2	yearisol	.187[c]	1.805	.077	.243	.587	1.702	.438
	alt	.100[c]	1.130	.264	.155	.831	1.203	.618
	logdist	-.095[c]	-1.126	.265	-.154	.908	1.102	.637

a. Dependent Variable: abund

b. Predictors in the Model: (Constant), logarea

c. Predictors in the Model: (Constant), logarea, graze

Lab

- Collinearity diagnostics

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	logarea	graze
1	1	1.757	1.000	.12	.12	
	2	.243	2.687	.88	.88	
2	1	2.460	1.000	.01	.03	.02
	2	.493	2.234	.00	.37	.10
	3	.047	7.257	.99	.59	.89

a. Dependent Variable: abund

License