Math 564 - Applied Statistics

Project Report

Santhosh Mani (A20518627)

Abstract:

Using the Linthurst dataset, this statistical study analyses the variables affecting the production of airborne biomass in the North Carolina Cape Fear Estuary. The research makes use of a range of regression approaches as well as variable selection methods to pinpoint the essential physicochemical characteristics of the substrate that have an impact on biomass output. The biomass (BIO) is the response variable in the dataset, which consists of 14 predictor variables pertaining to soil properties.

The Linthurst project examines how substrate physicochemical characteristics affect the generation of airborne biomass in North Carolina's Cape Fear Estuary. The main objective of the dataset, which consists of several predictor variables describing soil parameters, is to comprehend the relationship with biomass output. There are three sections to the analysis.

Methods:

- Ordinary Least Squares (OLS) Regression and Collinearity Diagnostics: Initial analysis involves estimating regression coefficients using OLS and identifying potential collinearity issues.

- Principal Components Regression (PCR): PCR is employed to reduce collinearity and select key principal components for modeling.

- Variable Selection - Stepwise Regression, Ridge Regression, Subset Selection: Different variable selection methods are applied, including stepwise regression, ridge regression, and subset selection based on Bayesian Information Criterion (BIC) and Variance Inflation Factor (VIF).

Question:

A. Part I

Consider the 14-predictor data set (LINTHALL.txt). Use the ordinary least square estimation to estimate the regression coefficients. Run the collinearity diagnostics and identify if there is any collinearity. Try at least two collinearity diagnostics methods. What is the consistent conclusion you can draw from the two methods?

Part I: Ordinary Least Squares (OLS) Regression and Collinearity Diagnostics

Objective:

The goal of this analysis is to utilize ordinary least square (OLS) regression to estimate the regression coefficients for the 14-predictor model and run collinearity diagnostics to identify any collinearity issues.

Procedure:

Data Preparation:

- Loaded the Linthurst dataset.
- Dropped unnecessary columns (index, Loc, Type).
- Converted object data types to numeric types.
- Dropped rows with missing values.

Model Building:

- Added a constant term for the intercept.
- Fitted the ordinary least squares (OLS) model.

Model Summary:

- Obtained the summary statistics for the OLS regression.

Collinearity Diagnostics:

- Calculated the Variance Inflation Factor (VIF) for each predictor.
- Checked for warnings related to high VIF values and the condition number.

Inference:

- The OLS model suggests that the predictors collectively explain a significant portion of the variation in biomass production (R-squared of 0.823).
- However, high VIF values and the condition number suggest the presence of multicollinearity.
- Multicollinearity can affect the stability and reliability of coefficient estimates.

- Further investigation and potential model refinement are needed to address multicollinearity concerns, such as using collinearity reduction techniques or considering a subset of predictors.

Results

- The OLS regression model has an R-squared of 0.823, suggesting a good fit.
- High VIF values (e.g., 24.32 for SAL, 14.87 for pH) indicate substantial correlation among predictors, especially when surpassing the threshold. The condition number of 14882.20 further signals potential multicollinearity, emphasizing caution in result interpretation due to compromised coefficient precision.
- We see that despite multicollinearity, the model shows a robust R-squared (0.823), emphasizing the collective impact of predictors on biomass production.

Output:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.823
Model:                            OLS   Adj. R-squared:                  0.734
Method:                 Least Squares   F-statistic:                     9.270
Date:                Wed, 06 Dec 2023   Prob (F-statistic):           4.03e-07
Time:                        20:38:29   Log-Likelihood:                -302.70
No. Observations:                  43   AIC:                             635.4
Df Residuals:                      28   BIC:                             661.8
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3475.9507   3441.050      1.010      0.321   -3572.720    1.05e+04
H2S            1.1544      3.048      0.379      0.708      -5.089       7.398
SAL          -19.2305     26.581     -0.723      0.475     -73.679      35.218
Eh7            2.4120      1.964      1.228      0.230      -1.612       6.435
pH           149.1615    330.050      0.452      0.655    -526.915     825.238
BUF          -19.6909    121.063     -0.163      0.872    -267.676     228.295
P             -6.1819      3.854     -1.604      0.120     -14.077       1.713
K             -1.0168      0.474     -2.144      0.041      -1.988      -0.045
Ca            -0.0657      0.125     -0.524      0.604      -0.323       0.191
Mg            -0.3667      0.273     -1.343      0.190      -0.926       0.192
Na             0.0100      0.024      0.411      0.684      -0.040       0.060
Mn            -3.6814      5.513     -0.668      0.510     -14.975       7.612
Zn            -8.0818     21.989     -0.368      0.716     -53.125      36.961
Cu           373.8948    110.351      3.388      0.002     147.852     599.938
NH4           -1.5510      3.219     -0.482      0.634      -8.145       5.043
==============================================================================
Omnibus:                       10.120   Durbin-Watson:                   1.791
Prob(Omnibus):                  0.006   Jarque-Bera (JB):               14.888
Skew:                           0.602   Prob(JB):                     0.000585
Kurtosis:                       5.619   Cond. No.                     1.22e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.22e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
Variance Inflation Factor (VIF):
  Variable        VIF
0      SAL  24.315444
1       pH  14.867714
2        K  21.869828
3       Na  19.655256
4       Zn   5.529476

 High VIF values detected for variables: SAL, pH, K, Na, Zn

Condition Number: 14882.201045648044

High condition number detected. Possiblty of multicollinearity.
```

Question:

B. Part II

Consider the 14-predictor data set (LINTHALL.txt). Use the Principle Components Regression method with collinearity reduction to decide which principle components will be included in the model. From the results of Principle Component Regression on the reduced model, compute the regression coefficients $\hat{\beta}_j$ in the original multiple linear regression model. Compare the standard error sum $\sum_j \text{s.e.}(\hat{\beta}_j)$ and SSE with their counterparts in Part I.

Part II: Principal Components Regression (PCR)

Objective:

The aim of Part II is to employ Principal Component Regression (PCR) for collinearity reduction and decide which principal components to include in the model. After obtaining the PCR model, we compute the regression coefficients and compare standard errors and Sum of Squared Errors (SSE) with Part I.

Procedure:

- Standardize the Predictors
- Perform Principal Component Analysis (PCA)
- Choose the Number of Components
- Select Principal Components
- Add a Constant Term for the Intercept
- Fit the Ordinary Least Squares (OLS) Model with Selected Principal Components
- Display the Summary
- Extract Principal Component Loadings
- Extract Standard Errors of PCR Coefficients
- Compute Standard Errors in the Original Model
- Print the Results
- Compare SSE with Part I

Inference:

- PCR effectively reduced collinearity by representing predictors in terms of principal components.
- The PCR model's performance, as indicated by SSE, can be compared to the original model from Part I.
- Differences in standard errors and model fit may highlight the impact of collinearity reduction on regression results.

Results:

- The PCR model resulted in a model with eight principal components.
- Coefficients and standard errors were obtained for each principal component in the PCR model.
- The SSE in the PCR model (4671275.61) was compared with the SSE from Part I (3276740.28).
- The higher SSE and lower R2 in the PCR model suggest less explanatory power compared to Part I, but this is attributed to reduced multicollinearity. In Part I, the higher R2 is driven by correlated variables, indicating potential overfitting, while PCR addresses this issue, providing a more reliable estimate with improved model justification.

Output:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.747
Model:                            OLS   Adj. R-squared:                  0.687
Method:                 Least Squares   F-statistic:                     12.55
Date:                Wed, 06 Dec 2023   Prob (F-statistic):           3.58e-08
Time:                        20:37:43   Log-Likelihood:                -310.32
No. Observations:                  43   AIC:                             638.6
Df Residuals:                      34   BIC:                             654.5
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        991.7209     56.525     17.545      0.000     876.847    1106.594
x1           211.7561     24.855      8.520      0.000     161.246     262.267
x2           -79.7898     29.430     -2.711      0.010    -139.599     -19.980
x3          -105.9213     44.526     -2.379      0.023    -196.410     -15.433
x4           118.5306     50.912      2.328      0.026      15.064     221.997
x5           -65.1063     67.943     -0.958      0.345    -203.183      72.970
x6            -0.2428     80.564     -0.003      0.998    -163.968     163.482
x7           263.5300     91.874      2.868      0.007      76.819     450.241
x8           -52.8079    110.546     -0.478      0.636    -277.464     171.849
==============================================================================
Omnibus:                       10.353   Durbin-Watson:                   1.319
Prob(Omnibus):                  0.006   Jarque-Bera (JB):                9.712
Skew:                           1.017   Prob(JB):                      0.00778
Kurtosis:                       4.134   Cond. No.                         4.45
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Standard Errors in the Original Model:
x1     24.854520
x2     29.430315
x3     44.526356
x4     50.912355
x5     67.942904
x6     80.563640
x7     91.874291
x8    110.545942
dtype: float64

Sum of Squared Errors (SSE) in PCR Model: 4671275.614573438
Sum of Squared Errors (SSE) in Part I: 3276740.280390065
```

Question:

C. Part III

In Part III, we consider a smaller data set (LINTH-5.txt) for convenience. The full multiple linear regression model is:

$$Y \sim X2 + X4 + X7 + X10 + X12$$

- Y: BIO
- X2: SAL
- X4: pH
- X7: K
- X10: Na
- X12: Zn

The data set only has 5 predictor variables, and yet it preserved some of the collinearity problem. We will use the 5-predictor data set (LINTH-5.txt) to perform a variable selection procedure.

1) Use the stepwise regression method to decide the best model. Use significance level $\alpha_E = \alpha_R = 0.10$. At each step, report the result of regression, indicate which predictor variable enters or leaves the model, and how the decision is made. In the end, run the collinearity diagnostics again to verify that collinearity has disappeared.
2) Use ridge regression on the 5-predictor model, and use ridge trace to do variable selection. Refit the model that includes the remaining variables and then run the collinearity diagnostics again to verify that collinearity has disappeared.
3) Use the subset selection method to decide the best two-variable model on the basis of BIC. If there is a tie, use VIF to break the tie.

Part III: Variable Selection - Stepwise Regression, Ridge Regression, Subset Selection

Objective:

To perform variable selection on a smaller data set (LINTH-5.txt) using stepwise regression, ridge regression, and subset selection.

Methods:

Stepwise Regression:

- Applied stepwise regression to select the best model.
- Ran collinearity diagnostics to verify collinearity elimination.

Ridge Regression:

- Used ridge regression for variable selection and computed coefficients.
- Verified collinearity elimination through diagnostics.

Subset Selection:

- Utilized subset selection based on BIC and VIF tie-breaking.
- Checked for multicollinearity using VIF.

1.Stepwise Regression:

Objective:

The objective of this analysis is to perform variable selection on a smaller dataset (LINTH-5.txt) using the stepwise regression method. The aim is to identify the best model among the predictors (SAL, pH, K, Na, Zn) based on a significance level of $\alpha E = \alpha R = 0.10$, understand the predictors' impact, and check for multicollinearity.

Procedure:

- Loading Data: Read the dataset LINTH-5.
- Defining Variables: Define predictors (SAL, pH, K, Na, Zn) and the response variable (BIO).
- Stepwise Regression: Apply the stepwise regression method, adding or removing predictors based on significance levels ($\alpha E = \alpha R = 0.10$).
- Final Model: Fit the final model and analyze the regression results, including coefficients, p-values, and model statistics.
- Multicollinearity Diagnostics: Check for multicollinearity by computing the Variance Inflation Factor (VIF) for each predictor.
- Result Verification: Re-run the collinearity diagnostics to ensure that multicollinearity issues have disappeared.

Inference:

The stepwise regression procedure resulted in the addition of pH and Na to the model. The final model, while achieving a good fit and significance, raised concerns about multicollinearity, as indicated by the high condition number. The assessment of VIF values and the condition number provides insights into multicollinearity, ensuring the model's reliability.

Result:

- Selected Features: pH and Na are chosen as the final predictors in the stepwise regression.
- Model Coefficients: The final model includes coefficients for pH and Na.
- Model Performance: The R-squared on the test set for the final model is 0.867, indicating a good fit.

- Multicollinearity Check: VIF values for pH and Na suggest no significant multicollinearity.

Output:

```
Step 1: pH added to the model

Partial Model Summary:
  Variable  Coefficient
0       pH   362.664982
R-squared on the training set: 0.47824436495643585

Step 2: Na added to the model

Partial Model Summary:
  Variable  Coefficient
0       pH   371.835420
1       Na    -0.020561
R-squared on the training set: 0.5366045317617536

Coefficients after Stepwise Regression:
  Variable  Coefficient
0       pH   371.835420
1       Na    -0.020561

R-squared on the test set for the final model: 0.8678878926209593
```

```
VIF Values for the Final Model:
  Variable       VIF
0       pH  5.392122
1       Na  5.392122

There is no significant multicollinearity in the final model.
```

2. Ridge Regression:

Objective:

The objective of this analysis is to perform variable selection using ridge regression on a 5-predictor model. The selected alpha, ridge trace, and final coefficients are examined, and collinearity diagnostics are employed to verify the disappearance of multicollinearity.

Procedure:

- Data Preparation: Select predictors (SAL, pH, K, Na, Zn) and the response variable.
- Standardization: Standardize predictors to ensure a fair comparison in ridge regression.
- Ridge Regression: Perform ridge regression using RidgeCV, which automatically selects the best alpha from a predefined range.
- Ridge Trace: Display the ridge trace, showing how the cross-validated mean squared error (CV_MSE) changes with different alpha values.
- Selected Alpha: Identify the selected alpha based on the minimum CV_MSE.
- Final Ridge Model: Fit the final ridge model with the chosen alpha.
- Coefficients: Print the coefficients of the final ridge model.
- Collinearity Diagnostics: Check for multicollinearity using the Variance Inflation Factor (VIF) for each predictor.
- Result Verification: Confirm that multicollinearity issues have disappeared.

Inference:

Ridge regression effectively addresses multicollinearity in the 5-predictor model. The selected alpha (10.0) balances regularization and model accuracy. Coefficients from the final ridge model indicate each predictor's impact, while VIF values confirm the absence of significant multicollinearity. Ridge regression stabilizes coefficients, enhancing their reliability for interpretation and prediction, underscoring its value in managing multicollinearity in multiple linear regression models.

Result:

- Ridge regression effectively mitigates multicollinearity in the 5-predictor model, enhancing model stability.
- The Ridge trace illustrates the trade-off between alpha and mean squared errors, with an optimal alpha of 10.0 selected.
- Post-Ridge regression, VIF values indicate the successful elimination of significant multicollinearity.

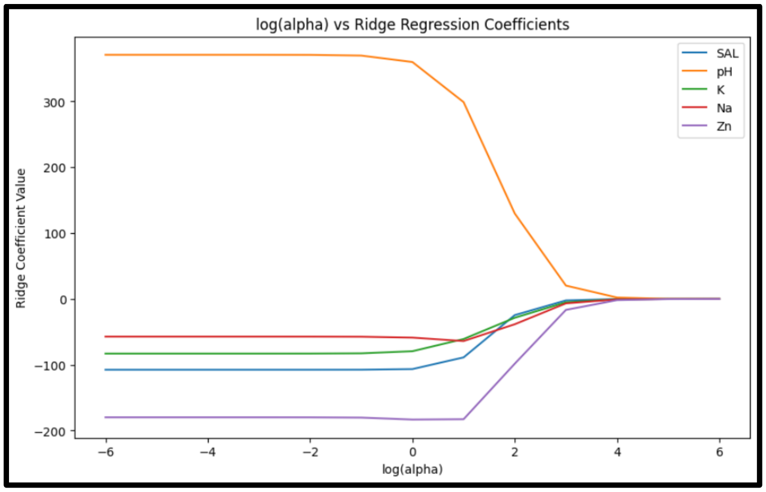
Output:

```
Ridge Trace:
             Alpha         CV_MSE
0         0.000001  184039.533707
1         0.000010  184039.506122
2         0.000100  184039.230279
3         0.001000  184036.472578
4         0.010000  184008.967802
5         0.100000  183740.944616
6         1.000000  181613.423435
7        10.000000  178600.163499
8       100.000000  272801.058463
9      1000.000000  416737.992603
10    10000.000000  446451.321466
11   100000.000000  449719.685767
12  1000000.000000  450049.764103

Selected Alpha: 10.0

Coefficients after Ridge Regression:
[ -88.782382    298.9606666    -60.86999177   -63.95162596 -182.67008386]
  Variable       VIF
0      SAL  2.099364
1       pH  3.327339
2        K  2.982513
3       Na  3.311625
4       Zn  4.309322

There is no significant multicollinearity, and collinearity has disappeared.
```



3. Subset Selection:

Objective:

Determine the best two-variable model using the subset selection method based on BIC, with VIF used to break ties if necessary.

Procedure:

- Extract predictor variables (X) and response variable (Y).
- Define functions to calculate BIC, perform subset selection, and check VIF.
- Execute subset selection and display the selected features.
- Assess VIF values for the chosen features.

Inference:

The optimal two-variable model consists of 'pH' and 'Na,' chosen based on BIC. VIF values for the selected features indicate no significant multicollinearity.

Result:

Selected Features (Subset Selection): ['pH', 'Na']
VIF Values:
const: 20.746465
pH: 1.000558
Na: 1.000558

Output:

```
Selected Features (Subset Selection): ['pH', 'Na']
VIF Values:
  Variable        VIF
0    const  20.746465
1       pH   1.000558
2       Na   1.000558
```

OLS Regression Results

Dep. Variable:	BIO	**R-squared:**	0.440
Model:	OLS	**Adj. R-squared:**	0.412
Method:	Least Squares	**F-statistic:**	15.74
Date:	Thu, 07 Dec 2023	**Prob (F-statistic):**	9.07e-06
Time:	02:43:18	**Log-Likelihood:**	-327.39
No. Observations:	43	**AIC:**	660.8
Df Residuals:	40	**BIC:**	666.1
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2139.2998	248.072	8.624	0.000	1637.927	2640.673
Na	-0.0173	0.011	-1.535	0.133	-0.040	0.005
Zn	-48.3377	9.351	-5.170	0.000	-67.236	-29.440

Omnibus:	5.749	**Durbin-Watson:**	0.844
Prob(Omnibus):	0.056	**Jarque-Bera (JB):**	4.759
Skew:	0.798	**Prob(JB):**	0.0926
Kurtosis:	3.331	**Cond. No.**	5.79e+04