



Stat 401 B – Lecture 24



Model Selection

- In multiple regression we often have many explanatory variables.
- How do we find the “best” model?


1



Model Selection

- How can we select the set of explanatory variables that will explain the most variation in the response and have each variable adding significantly to the model?

2




Cruising Timber

- Response: Mean Diameter at Breast Height (MDBH) of a tree.
- Explanatory:
 - X_1 = Mean Height of Pines
 - X_2 = Age of Tract times the Number of Pines
 - X_3 = Mean Height of Pines divided by the Number of Pines

3


Stat 401 B – Lecture 24



Forward Selection

- Begin with no variables in the model.
- At each step check to see if you can add a variable to the model.
 - If you can, add the variable.
 - If not, stop.


4



Forward Selection – Step 1

- Select the variable that has the highest correlation with the response.
- If this correlation is statistically significant, add the variable to the model.

5

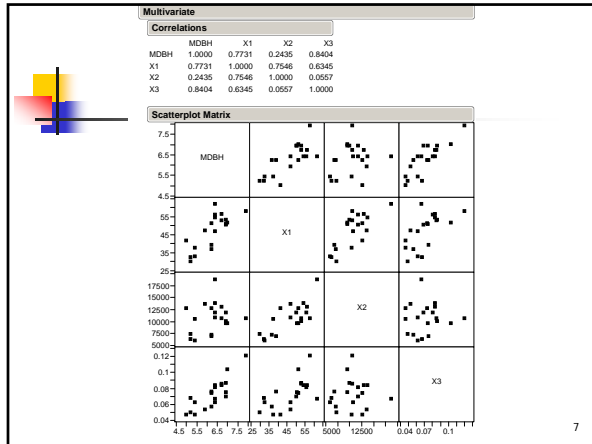


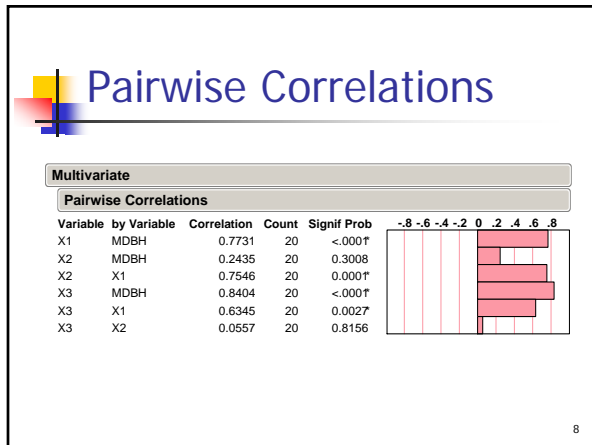
JMP

- Multivariate Methods
- Multivariate
- Put MDBH, X_1 , X_2 , and X_3 in the Y, Columns box.

6

Stat 401 B – Lecture 24





Comment

- The explanatory variable X_3 has the highest correlation with MDBH.
 - $r = 0.8404$
- The correlation between X_3 and MDBH is statistically significant.
 - Signif Prob < 0.0001, small P-value.

Stat 401 B – Lecture 24

Step 1 - Action

- Fit the simple linear regression of MDBH on X_3 .
- Predicted MDBH = $3.896 + 32.937 * X_3$
- $R^2 = 0.7063$
- RMSE = 0.4117

10

SLR of MDBH on X_3

- Test of Model Utility
 - $F = 43.2886$, P-value < 0.0001
- Statistical Significance of X_3
 - $t = 6.58$, P-value < 0.0001
- Exactly the same as the test for significant correlation.


11

Can we do better?

- Can we explain more variation in MDBH by adding one of the other variables to the model with X_3 ?
- Will that addition be statistically significant?

12


Stat 401 B – Lecture 24



Forward Selection – Step 2

- Which variable should we add, X_1 or X_2 ?
- How can we decide?
- Look at partial residual plots.
- Determine statistical significance.


13



Partial Residual Plots

- Look at the residuals from the SLR of Y on X_3 plotted against the other variables once the overlapping information with X_3 has been removed.

14

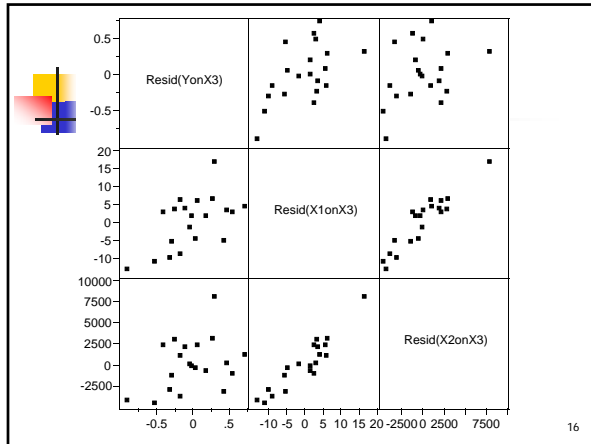


How is this done?

- Fit $MDBH$ versus X_3 and obtain residuals – $\text{Resid}(Y \text{ on } X_3)$
- Fit X_1 versus X_3 and obtain residuals - $\text{Resid}(X_1 \text{ on } X_3)$
- Fit X_2 versus X_3 and obtain residuals - $\text{Resid}(X_2 \text{ on } X_3)$

15

Stat 401 B – Lecture 24




Correlations

| | Resid(YonX3) | Resid(X1onX3) | Resid(X2onX3) |
|---------------|--------------|---------------|---------------|
| Resid(YonX3) | 1.0000 | 0.5726 | 0.3636 |
| Resid(X1onX3) | 0.5726 | 1.0000 | 0.9320 |
| Resid(X2onX3) | 0.3636 | 0.9320 | 1.000 |

Comment

- The residuals (unexplained variation in the response) from the SLR of MDBH on X_3 have the highest correlation with X_1 once we have adjusted for the overlapping information with X_3 .


Stat 401 B – Lecture 24



Statistical Significance

- Does X_1 add significantly to the model that already contains X_3 ?
 - $t = 2.88$, P-value = 0.0104
 - $F = 8.29$, P-value = 0.0104
 - Because the P-value is small, X_1 adds significantly to the model with X_3 .


19



Summary

- Step 1 – add X_3
 - $R^2 = 0.706$
- Step 2 – add X_1 to X_3
 - $R^2 = 0.803$
- Can we do better?

20




Forward Selection – Step 3

- Does X_2 add significantly to the model that already contains X_3 and X_1 ?
 - $t = -2.79$, P-value = 0.0131
 - $F = 7.78$, P-value = 0.0131
 - Because the P-value is small, X_2 adds significantly to the model with X_3 and X_1 .

21


Stat 401 B – Lecture 24



Summary

- Step 1 – add X_3
 - $R^2 = 0.706$
- Step 2 – add X_1 to X_3
 - $R^2 = 0.803$
- Step 3 – add X_2 to X_1 and X_3
 - $R^2 = 0.867$


22



Summary

- At each step the variable being added is statistically significant.
- Has the forward selection procedure found the “best” model?

23



“Best” Model?

- The model with all three variables is useful.
 - $F = 34.83$, $P\text{-value} < 0.0001$
- The variable X_3 does not add significantly to the model with just X_1 and X_2 .
 - $t = 0.41$, $P\text{-value} = 0.6844$

24
