

Chapter 6

Model selection

Multicollinearity and Residual
analysis

Selecting A Model

- Automatic Variable Selection Procedures
- Residual Analysis
- Leverage
- Collinearity
 - Variance Inflation Factors, Collinearity and Leverage,
 - Suppressor Variables
- Selection Criteria
- C_p Statistic, Model Validation, PRESS Statistic
- Handling Heteroscedasticity - Transformations

All possible regressions method

- Suppose that there are only 4 variables, x_1 , x_2 , x_3 and x_4 .
- How many possible models are there?
- $E y = \beta_0$ 1 model with no independent variables
- $E y = \beta_0 + \beta_1 x_i$, $i = 1, 2, 3, 4$
 - 4 models with one explanatory variable
- $E y = \beta_0 + \beta_1 x_i + \beta_2 x_j$,
- $(i, j) = (1,2), (1,3), (1,4), (2,3), (2,4), (3,4)$
 - 6 models with two explanatory variables
- $E y = \beta_0 + \beta_1 x_i + \beta_2 x_j + \beta_3 x_k$,
- $(i, j, k) = (1,2,3), (1,2,4), (1,3,4), (2,3,4)$
 - 4 models with three explanatory variables
- $E y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
 - 1 model with all independent variables.
- There are $2^4 = 16$ possible model

Criteria of choice

- R^2
- R^2_a
- Typical error s

Example

- Data Meddicorp
- Dependent variable: sales
- X_1 (ADV)
- X_2 (BONUS)
- X_3 (MKTSHR)
- X_4 (COMPET)

Variants of Stepwise Regression

- 1. Forward selection. Add one variable at a time without trying to delete variables.
- 2. Backwards selection. Start with the full model and delete one variable at a time without trying to add variables.
- 3. Best Regressions.

Forward Selection

- Begins with no variables
- Examine the independent variables to include in the model (P-value $< \alpha$)
- Add the variable with the largest contribution (among significant ones)
- A new regression is run with one less variable.
- Repeat this process until no other variable can be removed.

Backward Selection

- Begins with a regression on all variables
- Examine the independent variables to remove in the model (P-value $> \alpha$)
- Remove the variable with the smallest P-value (among significant ones)
- A new regression is run with one less variable
- Repeat this process until no other variable can be removed

Stepwise Regression Algorithm

- 1. Consider all possible regressions using one explanatory variable. **Retain** that variable with the largest t-ratio. If the t-ratio does not exceed a prespecified t-value (such as 2), do not choose any variables and halt the procedure.
- 2. The next variable to enter is the one that makes the largest significant contribution. To enter, the t-ratio must exceed a specified t-value in absolute value.
- 3. Now, delete a variable that provides the smallest contribution. To be deleted, the t-ratio must be smaller than a specified t-value in absolute value.
- 4. Repeat steps #2 and #3 until all possible additions and deletions are performed.
- Recall, when only one variable is being considered, that $(t\text{-ratio})^2 = F\text{-ratio}$.

Drawbacks of Stepwise Regression

- 1. The algorithm does not consider **nonlinear models**. It also ignores the presence of **outliers** and high **leverage** points.
- 3. The algorithm does not even search all 2^k possible linear regressions.
- 4. The algorithm uses one criterion, a **t-ratio**, and does not consider other criteria such as **s**, **R^2** , **R**, and so on.
- 5. There is a sequence of significance tests involved. Thus, the significance level that determines the **t-value may not be meaningful**.
- 6. By considering each variable separately, the algorithm does not take into account the **joint effect of independent variables**.
- 7. Purely automatic procedures may not take into account an investigator's **special knowledge**.

Health Data, (Data: C₆–World_Health.JMP)

- Life_Exp: life expectancy at birth (Y)
- Temp: Temperature, degree fahrenheit (X₁)
- Urban%: Percent of population living in urban areas (X₂)
- Populn: Total population , in millions(X₃)
- Beds_pop: number of beds per thousand population(X₄)
- Hosp_pop: number of hospitals per thousand population(X₅)
- RN_pop: number of nurses per thousand population(X₆)
- Pharmpop: number of pharmacists per thousand population(X₇)
- MD_pop: number of medical doctors per thousand population(X₈)
- Popdwell: population per dwelling(X₉)
- GNP: Gross national product (in thousand) per population(X₁₀)

Stepwise Fit

Response: LIFE_EXP

Stepwise Regression Control

Prob to Enter 0.250

Prob to Leave 0.100

Direction:

18 rows not used due to missing values.

Current Estimates

	SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
	2342.7823	21	111.56106	0.0000	0.0000	91.624481	104.6971
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	65.6045458	1	0	0.000	1.0000
<input type="checkbox"/>	<input type="checkbox"/>	TEMP	0	1	248.5135	2.373	0.1391
<input type="checkbox"/>	<input type="checkbox"/>	URBAN%	0	1	957.2678	13.818	0.0014
<input type="checkbox"/>	<input type="checkbox"/>	POPLN	0	1	229.9054	2.176	0.1557
<input type="checkbox"/>	<input type="checkbox"/>	BEDS_POP	0	1	1247.885	22.795	0.0001
<input type="checkbox"/>	<input type="checkbox"/>	HOSP_POP	0	1	321.5614	3.182	0.0896
<input type="checkbox"/>	<input type="checkbox"/>	RN_POP	0	1	181.6757	1.681	0.2095
<input type="checkbox"/>	<input type="checkbox"/>	PHARMPOP	0	1	809.7768	10.565	0.0040
<input type="checkbox"/>	<input type="checkbox"/>	MD_POP	0	1	1366.562	27.997	0.0000
<input type="checkbox"/>	<input type="checkbox"/>	POPDWELL	0	1	1694.264	52.250	0.0000
<input type="checkbox"/>	<input type="checkbox"/>	GNP	0	1	1626.473	45.413	0.0000

Step History

Stepwise Fit

Response: LIFE_EXP

Stepwise Regression Control

Prob to Enter 0.250

Prob to Leave 0.100

Direction: Forward 

18 rows not used due to missing values.

Current Estimates

	SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
	287.83807	14	20.559862	0.8771	0.8157	7.7143663	72.56982
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	46.5033791	1	0	0.000	1.0000
<input type="checkbox"/>	<input checked="" type="checkbox"/>	TEMP	0.18867148	1	76.51456	3.722	0.0742
<input type="checkbox"/>	<input checked="" type="checkbox"/>	URBAN%	0.13376252	1	68.43093	3.328	0.0895
<input type="checkbox"/>	<input checked="" type="checkbox"/>	POPLN	-0.0122289	1	70.8873	3.448	0.0845
<input type="checkbox"/>	<input checked="" type="checkbox"/>	BEDS_POP	-0.9176298	1	45.84128	2.230	0.1576
<input type="checkbox"/>	<input type="checkbox"/>	HOSP_POP	0	1	1.666273	0.076	0.7875
<input type="checkbox"/>	<input checked="" type="checkbox"/>	RN_POP	1.1400707	1	18.38255	0.894	0.3604
<input type="checkbox"/>	<input type="checkbox"/>	PHARMPOP	0	1	5.319577	0.245	0.6290
<input type="checkbox"/>	<input type="checkbox"/>	MD_POP	0	1	27.16946	1.355	0.2653
<input type="checkbox"/>	<input checked="" type="checkbox"/>	POPDWELL	-1.4057395	1	9.684543	0.471	0.5037
<input type="checkbox"/>	<input checked="" type="checkbox"/>	GNP	1.09089152	1	248.2011	12.072	0.0037

Model Building

- Goal is to Develop a Good Model with the Fewest Explanatory Variables
 - Easier to interpret
 - Lower probability of collinearity
- Stepwise Regression Procedure
 - Provide limited evaluation of alternative models

Collinearity

- Collinearity, or multicollinearity, occurs when one explanatory variable is, or nearly is, a linear combination of the other explanatory variables.
- Think of the independent variables as being highly correlated with one another.

Facts about Collinearity:

- 1. High correlation (among independent variables) does not preclude us from getting good fits nor making confidence interval type statements about the prediction of new observations.
- 2. Estimates of error variances and, therefore, tests of model adequacy, are still reliable.
- 3. However, estimated regression coefficients tend to have **large sampling variability**, other things equal.

How can we detect collinearity?

- 1. Matrix of correlation coefficients of explanatory variables
- 2. Scatterplot matrix of explanatory variables
- Advantages - easy to create and interpret.
- Disadvantage - only checks for relationships between pairs of explanatory variables.

Detect Collinearity (Variance Inflationary Factor)

- VIF_j Used to Measure Collinearity

$$VIF_j = \frac{1}{(1 - R_j^2)}$$

R_j^2 = coefficient of multiple determination from the regression of X_j on all the other explanatory variables

- If $VIF_j > 5$, X_j is Highly Correlated with the Other Explanatory Variables

More on VIF's

- It turns out that

$$se(b_j) = s \frac{\sqrt{VIF_j}}{\sqrt{n-1} s_{x_j}}$$

- Thus, the larger the VIF, the larger is the standard error associated with the *j*th slope, b_j .
- When collinearity occurs, certain aspects of $X'X$ are near zero.
 - When we attempt to calculate the inverse of $X'X$, this is analogous to dividing by zero.
 - Thus, $(X'X)^{-1}$ is very large.
 - Recall that the $(j+1)$ st diagonal of $(X'X)^{-1}$ is $se(b_j) / s$.

JMP - impgdp- Fit Least Squares

File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help

impgdp

Fit Model

impgdp- Fit Least Squares

Response LOGIMP

Whole Model GDP LOGGDP

Actual by Predicted Plot

Summary of Fit

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-1.057352			
GDP	0.0000182			
LOGGDP	0.8396079			

Effect Tests

Residual by Predicted

Table Style

- Columns
 - ✓ Term
 - ✓ ~Bias
 - ✓ Estimate
 - ✓ Std Error
 - ✓ t Ratio
 - ✓ Prob> |t|
 - Lower 95%
 - Upper 95%
 - Std Beta
 - VIF
- Sort by Column...
- Make into Data Table
- Make Into Matrix

Degree 2

Attributes

No Intercept

Selected	0	22	Panama	2.95	18	
Excluded	1	23	Samoa	0.1	0.45	-2.3025851 -0.7985077

Watch out the Multicollinearity

When interaction term is included in the regression model, Multicollinearity needs to be checked

Response Sales

Whole Model

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	-2.3497	0.688281	-3.41	0.0026	.
Radio Tv	2.3611	0.207524	11.38	<.0001	5.5
Paper	4.1831	0.207524	20.16	<.0001	5.5
Radio*Paper	-0.3489	0.062571	-5.58	<.0001	10

What can we do in the presence of collinearity?

- 1. Recode the variables.
- 2. Ignore it in the analysis but comment on it in the interpretation.
- 3. Remove one or more variables. Easy. Which One? is hard.
 - a. Use interpretation. Which variable(s) do you feel most comfortable with?
 - b. Replace a variable by a proxy variable.
 - c. Use automatic variable selection procedures to suggest a model.

Mallow's C_p Criteria

- Select the model that maximizes R^2 and R_a^2 and that minimizes \underline{s} .
- Another criterion, the \underline{C}_p statistic, defined by
- $$C_p = \{(\text{Error SS})_p / s^2\} - (n-2p)$$
- p = number of regression coefficients s^2 is the MSE for the full model.

- For example, in the full model case, $p=k+1$
- $$C_{k+1} = \{(\text{Error SS})_{k+1} / s^2\} - (n-2(k+1))$$
$$= (n-(k+1)) - (n-2(k+1)) = k+1.$$

More on the C_p Statistic

- In general, if the model is correct, we expect C_p to be close to p .
- $$C_p = \{(n-p) (\text{Error MS})_p / s^2\} - (n-2p)$$
- $$\approx \{(n-p) s^2 / s^2\} - (n-2p) = p.$$
- If the model is not correct, we expect $C_p > p$.
- For a fixed p , choose a model with the smallest C_p .
- Choose the model with C_p closest to p , this implies little bias.

SSPE Statistic

- Step 1:
Begin with a sample size of 'n' and divide this into two sub-samples. $i=1, \dots, n_1$ - obs from 1st subsample, $i=n_1+1, \dots, n_1+n_2 = n$ obs from 2nd subsample.
- Step 2.
For the first sample, fit a candidate model to the data set $i=1, \dots, n_1$.
- Step 3. Use the model created in Step 2 and the explanatory variables from the second sample to "predict" the dependent variables, \hat{y} , where $i=n_1+1, \dots, n_1+n_2$.
- Step 4. Compute the sum of squared prediction errors
- $$SSPE = \sum_i (y_i - \hat{y}_i)^2$$

where the sum is over $i=n_1+1, \dots, n_1+n_2$.
- Repeat Steps 2-4 for various candidate models. Choose the model with the smallest SSPE.

- Drawbacks of SSPE statistic:
 - Labor intensive
 - Choice of sample size splits is not clear ($n_1 = k$ should be % of n ?).
- Useful if you have a lot of data, say \Rightarrow 200 points.

So we introduce PRESS:

- *PRESS* - prediction residual sum of squares
 - A far less computationally intensive version of SSPE.

Procedure for Calculating PRESS:

- Omit the i th point. Use the remaining $n-1$ observations to compute regression coefficients. Use these regression coefficients and the predictor variables for the i th observation to compute $\hat{y}_{(i)}$. Define

$$PRESS = \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2$$

- .
- Some algebra (not done here) leads to

$$PRESS = \sum_{i=1}^n \left(\frac{\hat{e}_i}{1 - h_{ii}} \right)^2$$

- Where \hat{e}_i is the residual point and h_{ii} is the leverage point
- a much easier computational formula.

Transformations-Handling Heteroscedasticity

- An important assumption - common variability among observations.
- Recall "homoscedasticity" for "same scatter."
"heteroscedasticity" for "different scatter".
- Least squares
 - 1. assumes expected variability of each observation is equal
 - 2. gives the same weight in the minimization of the sum of squares procedure.
- To detect heteroscedasticity - Plots residuals versus fitted values, predictors
- Action - When variability is monotone, use transformation of dependent variables to reduce the heteroscedasticity

Residual Analysis

- Role of residuals: If the model formulation is correct, then residuals \approx random errors.
- Four types of Patterns:
 - Unusually large residuals
 - Residuals related to explanatory variables
 - Heteroscedastic Residuals
 - Time Patterns in Residuals (Start in Chapter 9)
- Method of attack: Look for patterns in the residuals.
- Use this information to improve the model specification.

Standardized Residuals

- Standardized residual = residual divided by it's estimated standard error.
- Use standardized residuals because:
 - we can focus on relationships of interest
 - achieve carry-over of experience from one data set to another.
- Three ways to define a standardized residual:
- \hat{e}_i / s
- $\hat{e}_i / \{ s (1 - h_{ii})^{1/2} \}$, $s = (\text{MSE})^{1/2}$ and
- $\hat{e}_i / \{ s_{(i)} (1 - h_{ii})^{1/2} \}$ $s_{(i)} = (\text{MSE})^{1/2}$ when omitting obs (i)
- First choice is simple
- Second choice, from theory, $\text{Var}(e_i) = \sigma^2 (1 - h_{ii})$.
- Third choice is termed "studentized residuals". Idea: numerator is independent of the denominator.

Role of Residuals

- Define an outlier - unusually large residual - possibly due to a special cause which we should locate.
- Unusual means what? When $|\text{standardized residual}| > 2$.
- What to do with outliers?
- Ignore them in the analysis but be sure to discuss their effects.
- Delete them from the data set (but be sure to discuss their effects).
- Flag them with an indicator variable.

Use Residuals to Detect Relationships with Explanatory Variables

- Ways of plotting residuals.
 - 1. Calculate summary statistics and a dotplot of (standardized) residuals to identify outliers.
 - 2. Calculate the correlation between the (standardized) residuals and additional explanatory variables to search for linear relationships.
 - 3. Create scatter plots between the (standardized) residuals and additional explanatory variables to search for nonlinear relationships.
 - 4. Plot standardized residuals vs x 's to determine effects of x 's not in the model and the nonlinear effects of x 's in the model.

Leverage

- A high leverage point is an observation containing an unusual set of explanatory variables.
- This observation has a disproportionate influence in the overall regression fit.
- Regression estimates can be shown to be weighted averages. For example, $b_1 = \sum_i \text{weight}_i \text{slope}_i / \sum_i \text{weight}_i$.
- Here, $\text{weight}_i = (x_i - \bar{x})^2$ and $\text{slope}_i = (y_i - \bar{y}) / (x_i - \bar{x})$.
- To quantify "unusual," recall $\mathbf{H} = \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}$.
- Define h_{ij} , the number in the i th row and j th column of \mathbf{H} .
- The i th row of $\hat{\mathbf{y}} = \mathbf{H} \mathbf{y}$ is:
- $\hat{y}_i = h_{i1} y_1 + h_{i2} y_2 + \dots + h_{ii} y_i + \dots + h_{in} y_n$.
- Define h_{ii} to be the leverage for the i th point.
- The larger is h_{ii} , the larger is the effect of the i th observation on its fitted value.

More on Leverage

- How large is large? From matrix algebra,
- $1/n \leq h_{ii} \leq 1$ and $\bar{h} = (1/n) \sum_i h_{ii} = (k+1) / n$.
- The i th leverage point is "large" if $h_{ii} > 3 (k+1) / n$.
- JMP uses this cut-off.
- What to do with influential points? Options are similar to those available for outliers:
- Ignore them in the analysis but be sure to discuss their effects.
- Delete them from the data set (but be sure to discuss their effects).
- Choose another variable to represent this information (e.g. bedroom versus square footage).

Example 1

- X 1 2 3 4 5
- Y 1.1 2.2 2.4 2.8 3.3

- And
- X 1 2 3 4 5
- Y 1.1 2.2 5 2.8 3.3

Example

- Study the effect of observation 10 for the following:

- | | | | | | | |
|---|-----|-----|-----|-----|-----|----|
| X | 1 | 2 | 3 | 4 | 5 | 10 |
| Y | 1.1 | 2.2 | 2.4 | 2.8 | 3.3 | 6 |

- And

- | | | | | | | |
|---|-----|-----|-----|-----|-----|----|
| X | 1 | 2 | 3 | 4 | 5 | 10 |
| Y | 1.1 | 2.2 | 2.4 | 2.8 | 3.3 | 2 |

Cook's Distance

Combining measures to detect outliers and leverage

- Another measure of "influence." This measure considers both the predictor and response variables ...
-

$$D_i = \frac{\sum_{j=1}^n (y_j - y_{j(i)})^2}{(k+1)s^2} = \left(\frac{e_i}{s\sqrt{1-h_{ii}}} \right)^2 \frac{h_{ii}}{(k+1)(1-h_{ii})}$$

- Here $y_{j(i)}$ is the prediction of the j th observation computed leaving the i th observation out of the regression fit.
- How large should D_i be? Compare D_i to an F-distribution with $df_1 = k+1$ and $df_2 = n-(k+1)$.

Power Family of Transforms

- In lieu of using the response y , use $y^* = y^\lambda$.
- λ 1 1/2 0 -1
- $y^* = y^\lambda$ y \sqrt{y} $\log(y)$ $1/y$
- In general, no accepted theory on the choice of a transform. Values of $\lambda < 1$ serve to "shrink" spread out data.
- Some Rules of thumb:
 - 1. Use square root transform for count data
 - 2. Use logs for multiplicative models, e.g., gravity model

More on Transformations

- We use transforms to :
 - 1. Symmetrize the distribution of the responses/residuals
 - 2. Address the heteroscedasticity problem.
- We often use logs because of the interpretation of the coefficients. Here, b_j 's can be interpreted as proportional changes in expected response per unit change in x_j .
- Example: $\text{LOGSALARY} = 12 + .0861 \text{ YRS EXP}$.
Suppose we are interested in the incremental effect of an 1 year increase in YRS EXP. Then,
- Thus, the proportional change is about 8.6%.