Applied Regression

Multiple Linear Regression Model (MLR Model)

Adequacy

Module 5 Lecture - 5-2

MLR Model In the last lecture we have learned how to judge whether a point should be considered an outlier and whether to delete it or not. But the measures are model dependent, meaning given the model we calculate the residuals and then test. Here we are focusing on the model itself and trying to judge whether the model itself should be changed or not. These are done usually by plots and it is a very effective way of updating the model. Following are the plots we discuss:

1. Plot and Residual

(a) Normal Probability - Plot

(b) Residual vs Fitted Values

(c) Residual vs Each Regressors Values

(d) Residual vs time (if over time).

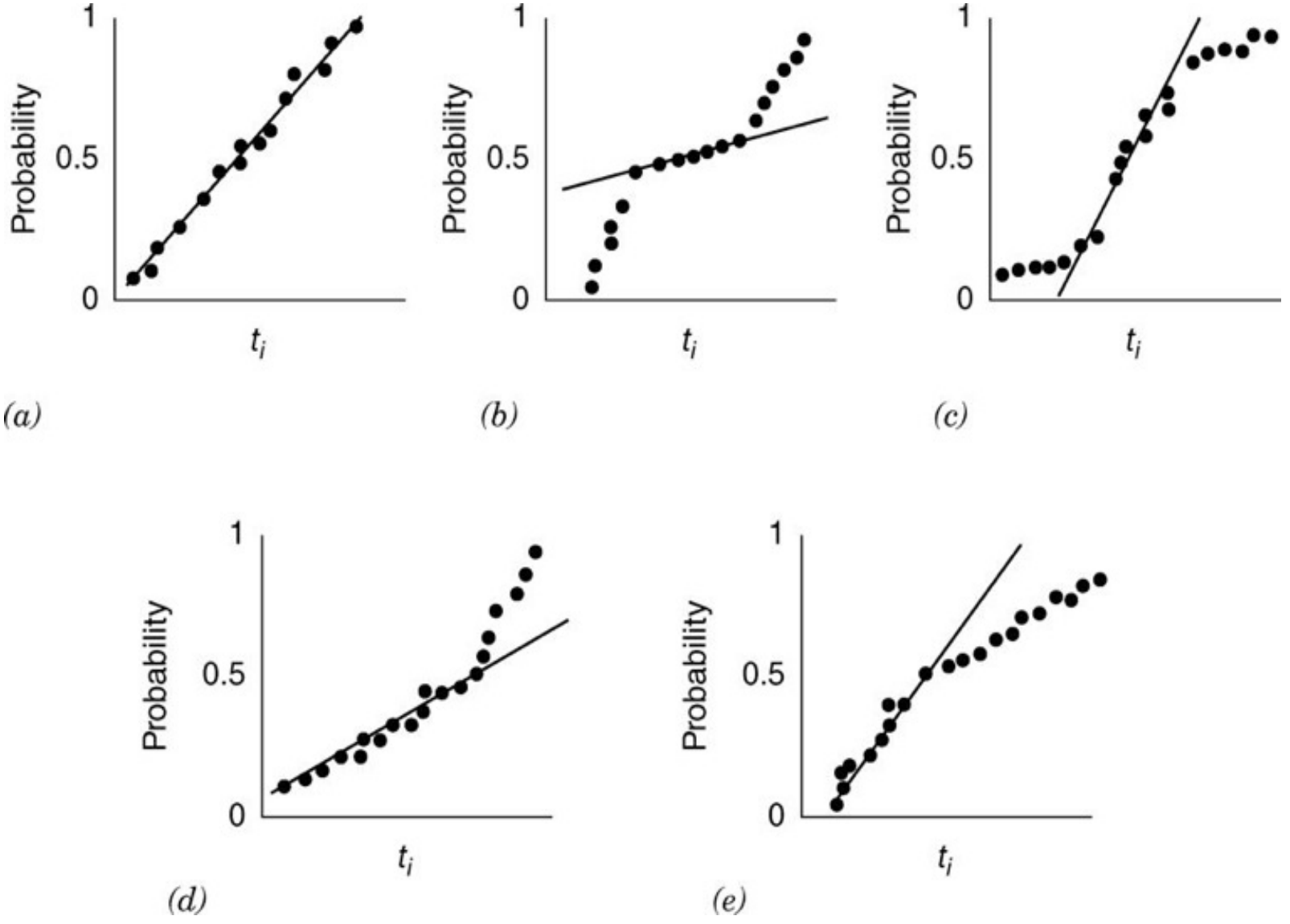
Remark: Besides the above plot we are also going to discuss, why it is necessary to look at the basic multi-plots (all y vs x_j's and all x_i vs x_j). It reveals many other things like dependency among the x_j's.

(a) Normal Probability Plot

Little (no exact numbers) departure from the normality assumption do not affect the model specially if sample size is large, but gross non-normality is the violation of assumption which can potentially effect the distribution of the statistics. Normal probability plot of the residuals is simple method of checking whether the normality assumption is correct or not. This is a graph designed in such a way that we expect the point to approximately on a straight line.

Let $t_{[1]} < t_{[2]} < \ldots < t_{[n]}$ be the values of the R-students ranked in increasing order. Now we plot $t_{[i]}$ against $P_i = (i - \frac{1}{2})/n$ (the cumulative probability) and visually try to examine. In an ideal situation we expect the graph to be a straight line (mainly within the range of 0.33 to 0.67) rather than the extremes. Those it is possible to formally test against normality but visual inspection is done more often than not. Because expected value of the ith order statistic from normal distribution is $E(t_{[i]}) = \Phi^{-1}\left[(i-\frac{1}{2})/n\right]$, sometimes $t_{[i]}$'s are plotted against its expected value $E(t_{[i]})$. Now the departures from the straight line (mainly at the extremes) indicates the analysis in certain directions. Let's look at the following graphs.

Normal Probability Plots



Consider the 5 previous normal probability plots, following are the conclusions we can draw:

(a) Is the ideal - Follows Normal

(b) light-tailed distribution - Below the line in lower percentile and above the line in higher percentile

(c) heavy-tailed distribution - Above the line in lower percentile and below the line in higher percentile

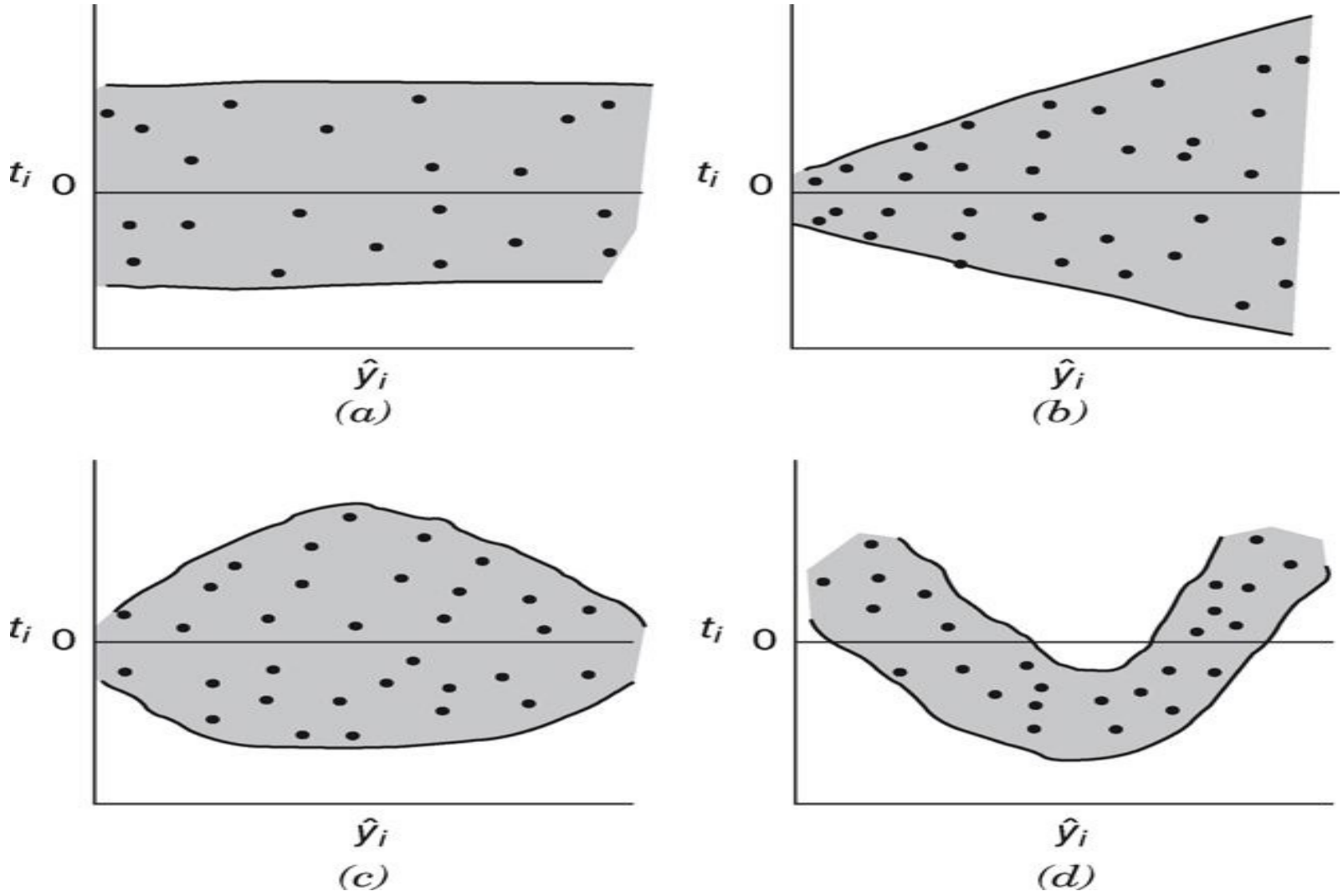
(d) positive (or right) skewed distribution - Only above the line in higher percentile

(e) negative (or left) skewed distribution - Only below the line in higher percentile

The graph with smaller sample size ($n \leq 16$) often deviate from the ideal plots but plots with reasonably larger sample size ($n \geq 32$) are well behaved. It is most common to see that it is close to the line in the middle. Also extreme deviation for one point signals the potential outlier.

(b) <u>Residual vs Fitted Values - Plot</u> - A plot of the $(\hat{y}_i, t_i)$ is useful for detecting several common types of model inadequacies.

<u>Predicted $\hat{Y}_i$ against t_i - Plot</u>



Here are the conclusions from some common situations:

Graph (a) - It is the ideal graph where points are in a horizontal band and departure from that indicates model defects.

Graph (b)- The the vertical width of the band is increasing with the value of $\hat{y}$ (outward-opening funnel pattern) indicates that the variance is an increasing function of y. It also could happen other way meaning vertical width of the band is decreasing with the value of $\hat{y}$. Such situation is usually dealt with some transformations on Y's or/and X's. The other way of dealing with such situation is called weighted least squares where covariance matrix is not identity.

Graph (c)- In this graph, the vertical width of the band is increasing with the value of $\hat{y}$ first and then decreasing. Hence the maximum occurs in the middle of the range of $\hat{y}$. This situation also requires some transformation which may be different from the previous one. This situation is common when dependent variable is proportion type. Because the variance of a binomial random variable is highest when the proportion of success (p) is 0.5 and less as p approaches 0 or 1. Hence the transformation is done accordingly (discussed later).

Graph (d)- This graph is little harder to deal with as it suggests that relationship is not linear. It may mean that more x variables are needed or functional form is not correct.

(c) Plot of Residuals against the Regressor

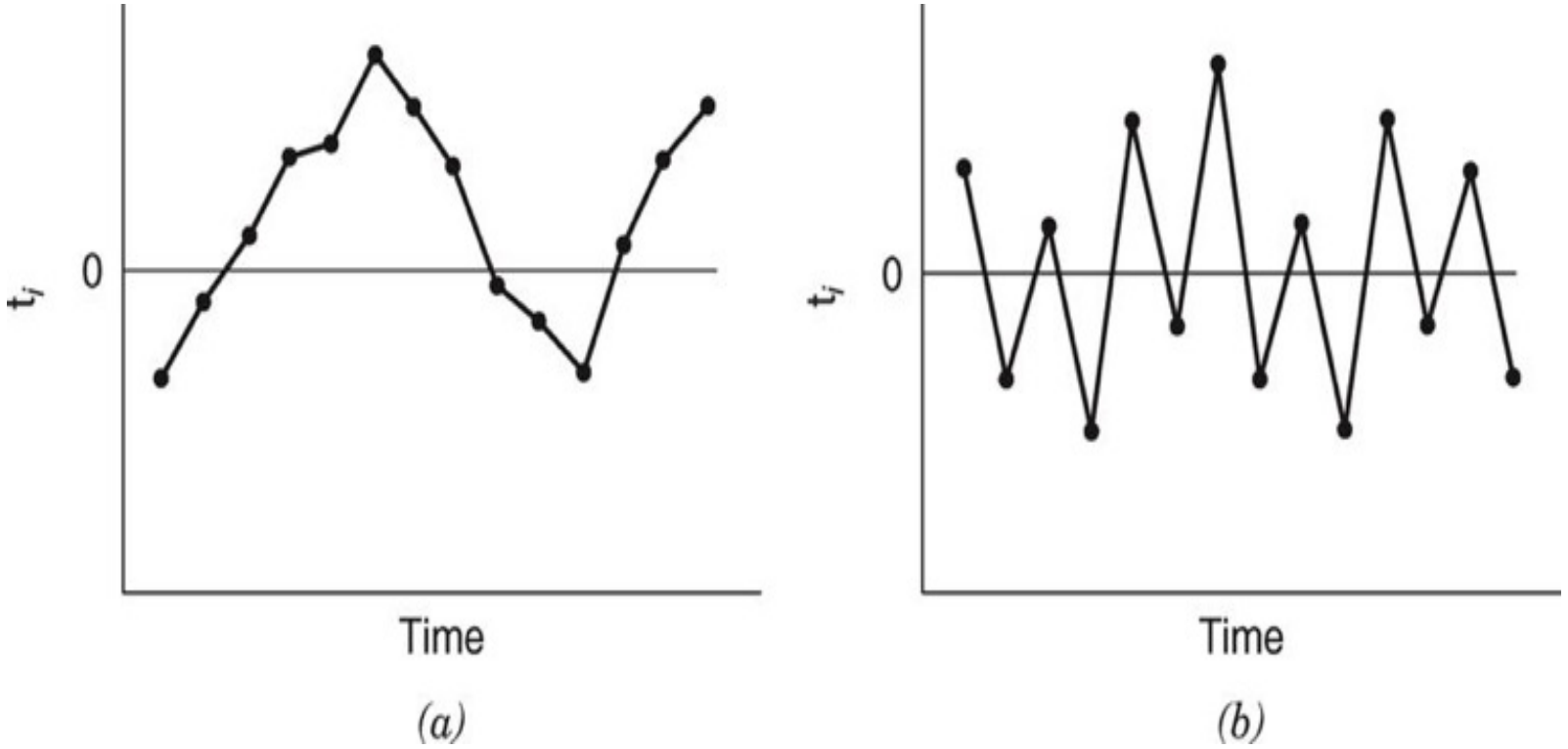
Plotting t_i against each of the X_j is also an important plot. First thing to note that $\hat{Y}$ is a linear combination of X_i's. Hence these plots often exhibit patterns similar to the plots vs $\hat{Y}$, but it may be very specific to some X_j or not to others. Sometime it simply requires and addition of square terms (X_j^2) in the model.

Note that in simple linear regressor case, it is not necessary to plot residuals versus both $\hat{Y}$ and the regressor variable. The reason is that the fitted values are linear combinations of the regressor values xi, so the plots would only differ in the scale for the abscissa.

In a multiple linear regression situation, it is also helpful to plot residuals against potential regressor variables that are not currently in the model. Any pattern in the plot of residuals versus an omitted variable indicates that incorporation of that variable could improve the model.

(d) Plot of Residuals in Time Sequence

Predicted $\hat{Y}_i$ against t_i - Plot



Sometime the regressor variable is time (as in Time Series models), it is a good idea to plot the residuals against time order. Though plot may look like anything but certain pattern indicates something. This kind of plot are routinely done to investigate autocorrelation. The correlation between model errors at different time periods is called autocorrelation.

If the plot resembles Figure (a), that is, a horizontal band will enclose all of the residuals, it indicates positive autocorrelation. However, if this plot resembles the patterns in Figures (b) this may indicate typical of negative autocorrelation. There are many other indications as well and we will discuss those in Time Series Model.

2. PRESS Statistic:

Previously, we defined the PRESS residuals as $e_{(i)} = y_i - \hat{y}_{(i)}$ where $\hat{y}_{(i)}$is the predicted value of the ith observed response based on a model fit to the remaining (n-1) sample points. We noted that large PRESS residuals are potentially useful in identifying observations where the model does not fit the data well or observations for which the model is likely to provide poor future predictions.

The sum of squares of these residuals gives the PRESS statistic which is used as a measure of model quality. Hence PRESS staistic is

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

PRESS is generally regarded as a measure of how well a regression model

will perform in predicting new data. It is expected that models with smaller PRESS value will predict better future values.

As R^2 is one of the most commonly used statistic in practice to judge the model, a similar R^2 for prediction is suggested based on PRESS statistic. This statistic gives some indication of the predictive capability of the regression model. It is defined as:

$$R^2_{prediction} = 1 - \frac{PRESS}{SST}$$

3. Lack of Fit (LOF)

A famous quote from Prof. George Box is "All models are wrong; some models are useful."

This comment has ties to the fact why "lack-of-fit" is important. The formal statistical test for the lack of fit of a regression model assumes that the normality, independence, and constant-variance requirements are met. The lack-of-fit test requires that we have replicate observations on the response y for at least one level of x. We emphasize that these should be true replications, not just duplicate readings or measurements of y. All the formulas here are

based on the fact that we are in a SLR set up (meaning there is only one x variable). Suppose there are n_i replications at $x = x_i$.

The error variance σ^2 includes two types of errors

(1) The Pure variation from observing many y's at the same x_i's which is model independent. These replicated observations are used to obtain a model-independent estimate of σ^2.

(2) The error from wrong model

Suppose there are n_i observations of y at the ith level of the regressor $x = x_i$. Let y_{ij} denote the jth observation on the response at x_i for i = 1, 2, .., m and j = 1, 2, .., ni. The test procedure involves partitioning the residual sum of squares into two components,

(1) Sum of squares due to pure error = $SS_{PE} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$

Note that in the SS_{PE} formula there is no $\hat{y}$. As a result it is independent of model and it is not going to change even you change the model.

(2) Sum of squares due to lack of fit = $SS_{LOF} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{y}_i - \hat{y}_i)^2 = \sum_{i=1}^{m} n_i \, (\bar{y}_i - \hat{y}_i)^2$

Observe that sum over for each fixed i summing over j has the fixed quantity and that's why it is multiplied by n_i.

To develop this partitioning of SSE, note that the predicted value for the same value of i are all same and (ij)th residual can be written as

$$y_{ij} - \hat{y}_i = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)$$

Also it can be shown that (since the cross product is 0).

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_i - \hat{y}_i)^2$$

$$SSE = SS_{PE} + SS_{LOF}$$

To test the lack of fit, we use the F-statistic,

$$F_0 = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m} = \frac{MS_{LOF}}{MS_{PE}} \quad \sim \; F_{(m-2),(n-m)} \text{ if the model is correct}$$

Hence we reject the model is $F_0 > F_{\alpha,(m-2),(n-m)}$.

Example 5-1: The kinematic viscosity of a certain solvent system depends on the ratio of the two solvents and the temperature. Table B.10 summarizes a set of experimental results.

a. Fit a multiple linear regression model relating the viscosity to the two regressors.

b. Construct a normality plot of the residuals from the full model. Does there seem to be any problem with the normality assumption?

c. Construct and interpret a plot of the residuals versus the predicted response.

d. Fit a SLR model using temperature as the only regressor. Compute the PRESS statistic for both models. Based on this statistic, which model is most likely to provide better predictions of new data?

SAS Output - Example 5-1

Solution:

Model 1

Root MSE	0.25934	**R-Square**	0.8220
Dependent Mean	1.04829	**Adj R-Sq**	0.8124
Coeff Var	24.73913		

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8.06061	8.06061	51.71	<.0001
Error	38	5.92338	0.15588		
Corrected Total	39	13.98399			

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.67944	0.14353	4.73	<.0001
x1	x1	1	1.40733	0.19693	7.15	<.0001
x2	x2	1	-0.01563	0.00143	-10.95	<.0001

Model 2

Root MSE	0.39481	**R-Square**	0.5764
Dependent Mean	1.04829	**Adj R-Sq**	0.5653
Coeff Var	37.66288		

Analysis of Variance

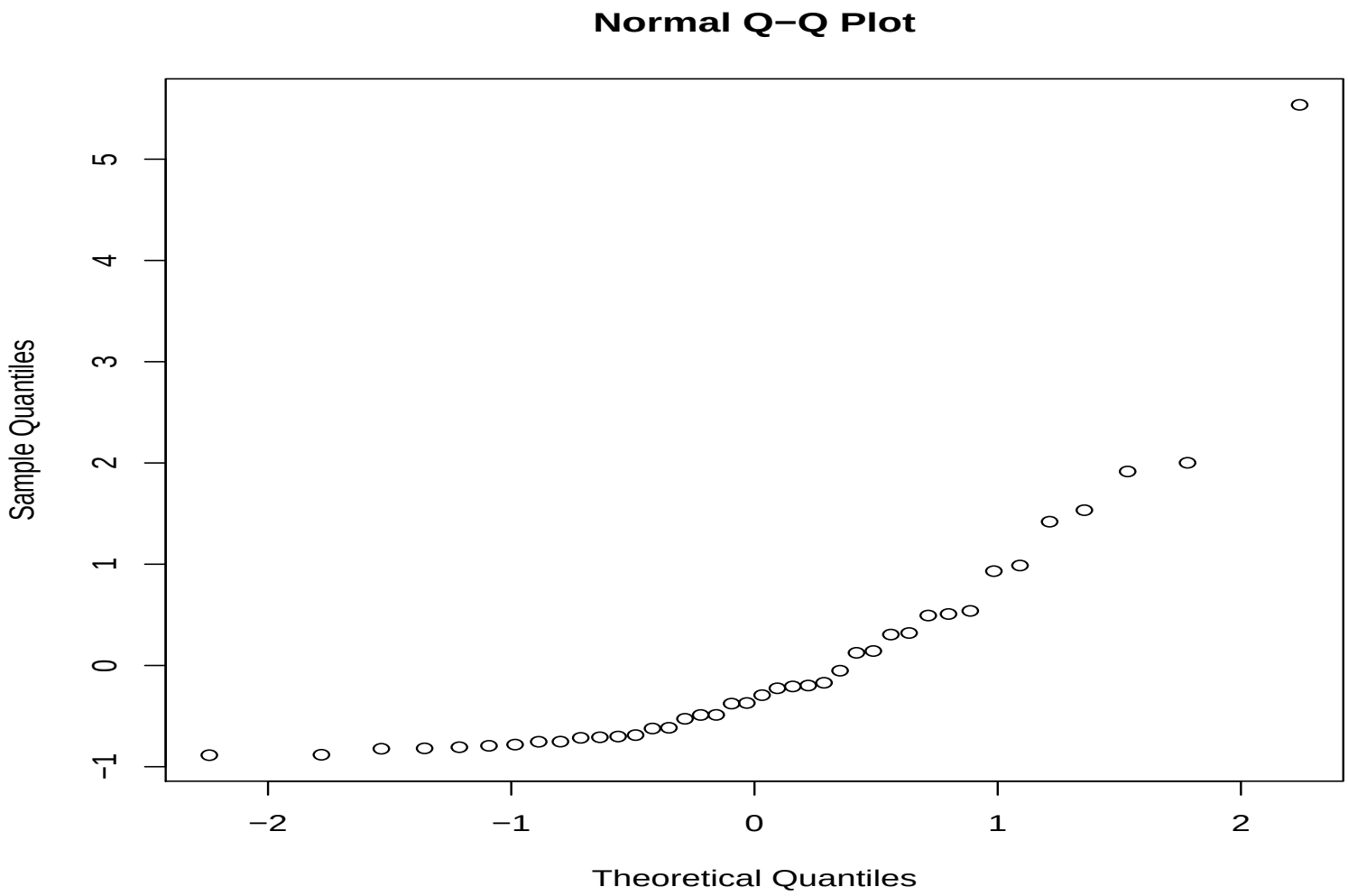
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	11.49553	5.74777	85.46	<.0001
Error	37	2.48845	0.06726		
Corrected Total	39	13.98399			

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.59529	0.09840	16.21	<.0001
x2	x2	1	-0.01563	0.00217	-7.19	<.0001

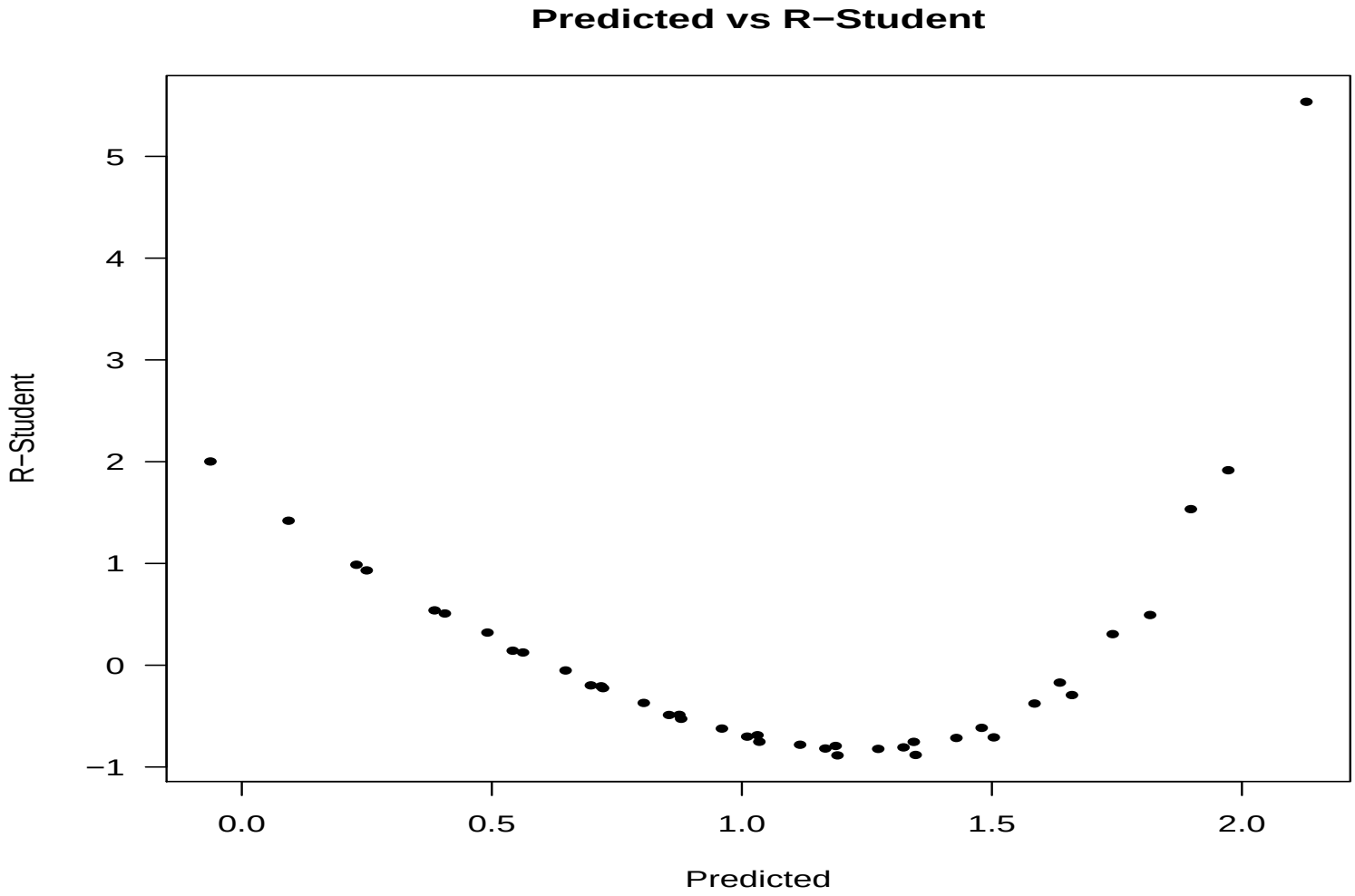
(a) Above are the outputs from two models. Model-1 has two variables (X_1 & X_2) and Model-2 has only X_2.

QQ - Plot Example 5-1



(b) As it is not at all a straight line there is some problem with the normality assumption.

$\hat{y}$ vs R-Student Plot - Example 5-1



(c) As discussed before, this does not show a linear relationship. Hence some transformation is necessary. (done in next chapter).

(d) As we see in the SAS output below that PRESS statistic is much higher for Model-2 and hence if one out of these two needs to be chosen then Model-1 has a better prediction power. Also not that LOF has been computed for both models but Model-1 does not have any repeat values and as a result LOF stat could not be calculated.

PRESS Stat & LOF - Example 5-1

Model – 1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	11.49553	5.74777	85.46	<.0001
Error	37	2.48845	0.06726		
Lack of Fit	37	2.48845	0.06726	.	.
Pure Error	0	0	.		
Corrected Total	39	13.98399			

Model - 2

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8.06061	8.06061	51.71	<.0001
Error	38	5.92338	0.15588		
Lack of Fit	8	0.80297	0.10037	0.59	0.7797
Pure Error	30	5.12041	0.17068		
Corrected Total	39	13.98399			

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	_PRESS_	Intercept	x1	x2	y
1	MODEL1	PARMS	y	0.25934	3.11213	0.67944	1.40733	-0.015629	-1
2	MODEL2	PARMS	y	0.39481	6.77664	1.59529	.	-0.015629	-1