Applied Regression

Multiple Linear Regression Model (MLR Model)

Adequacy - Part-3

Module 5 Lecture - 5-3

Using graphs to check model adequacy is an important step to build a good model. If the graph shows some inadequacies, then the next step is to fix the issue, but it is not an easy straight forward job. In very many situations, it requires practical knowledge about the situation as well well knowledge of statistical behavior of the random variable. In this lecture, we will learn few of those techniques through example. Namely those are:

1. **Variance Stabilizing Transformation**

2. **Linearizing the Model**

Once again there is no exact immediate answers of the issues but there are some common graphs and techniques which needs to be tried. At times there can be two different models which can performed almost equally. Then other consideration like experience and prior knowledge should be used.

1. Variance Stabilizing Transformation

In basic regression model we assume that mean of Y changes with X's but variance is constant and it is one of the basic requirement of regression analysis. But for many random variables, variance does change with the mean change because variance is functionally related to the mean.

For example, if y is a Poisson random variable in a simple linear regression model, then the variance of y is equal to the mean. Since the mean of y is related to the regressor variable x, the variance of y will be proportional to x. Variance-stabilizing transformations are often useful in these cases. Thus, if the distribution of y is Poisson, we could regress $y' = \sqrt{y}$ against x since the variance of the square root of a Poisson random variable is independent of the mean.

As another example, if the response variable is a proportion $(0 < y_i < 1)$ and the plot of the residuals versus $\hat{y}_i$ has the "double-bow" pattern (vertical width is highest in the middle) then of the arcsin transformation $y' = sin^{-1}(\sqrt{y})$ is appropriate.

Following are some common transformations:

Relationship of σ^2 to E(y)	Transformation
$\sigma^2 \propto$ constant	$y' = y$
$\sigma^2 \propto E(Y)$	$y' = \sqrt{y}$
$\sigma^2 \propto E(Y)(1 - E(Y))$	$y' = sin^{-1}(\sqrt{y})$
$\sigma^2 \propto (E(Y))^2$	$y' = ln(y)$
$\sigma^2 \propto (E(Y))^3$	$y' = y^{-1/2}$
$\sigma^2 \propto (E(Y))^4$	$y' = y^{-1}$

Example 5.2

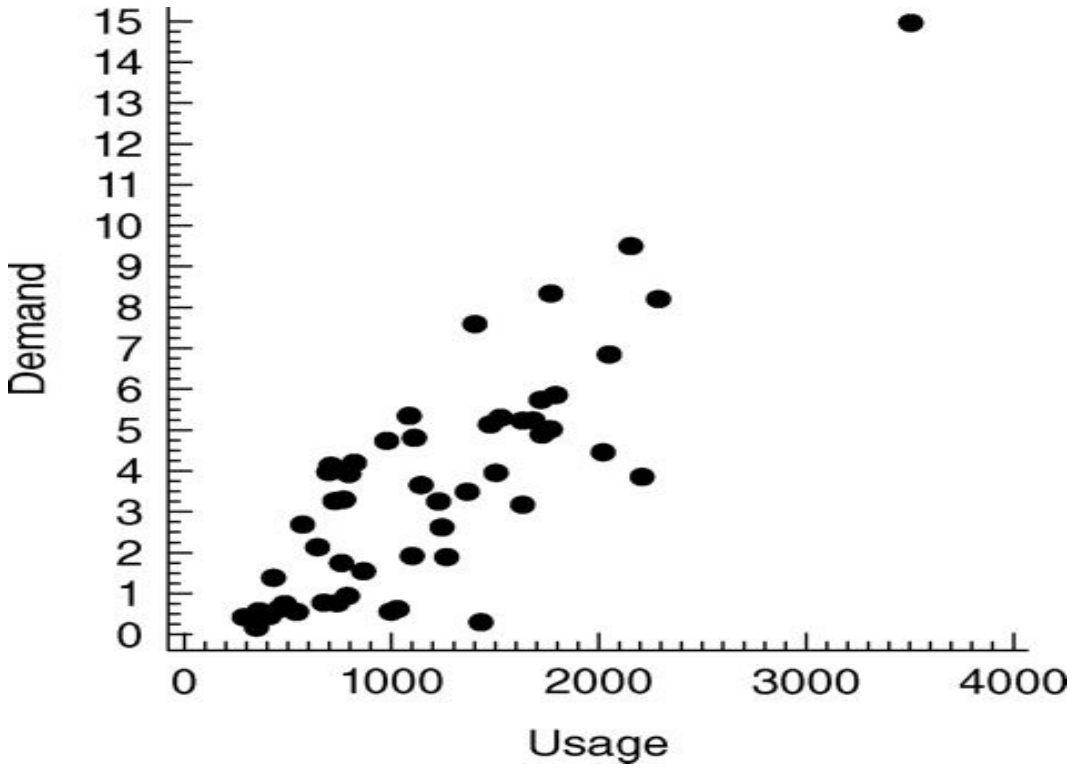
An electric utility is interested in developing a model relating peak-hour demand (y) to total energy usage during the month (x). Data for 53 residential customers for the month of August are given in the book . By looking at the scatter diagram (in next 2 slides), a simple linear regression model is assumed, and the least-squares fit

$$\hat{y} = -0.8313 + 0.00368\,x$$

Data - Example 5.2

Customer	x (kWh)	y (kW)	Customer	x (kWh)	y(kW)
1	679	0.79	27	837	4.20
2	292	0.44	28	1748	4.88
3	1012	0.56	29	1381	3.48
4	493	0.79	30	1428	7.58
5	582	2.70	31	1255	2.63
6	1156	3.64	32	1777	4.99
7	997	4.73	33	370	0.59
8	2189	9.50	34	2316	8.19
9	1097	5.34	35	1130	4.79
10	2078	6.85	36	463	0.51
11	1818	5.84	37	770	1.74
12	1700	5.21	38	724	4.10
13	747	3.25	39	808	3.94
14	2030	4.43	40	790	0.96
15	1643	3.16	41	783	3.29
16	414	0.50	42	406	0.44
17	354	0.17	43	1242	3.24
18	1276	1.88	44	658	2.14
19	745	0.77	45	1746	5.71
20	435	1.39	46	468	0.64
21	540	0.56	47	1114	1.90
22	874	1.56	48	413	0.51
23	1543	5.28	49	1787	8.33
24	1029	0.64	50	3560	14.94
25	710	4.00	51	1495	5.11
26	1434	0.31	52	2221	3.85
			53	1526	3.93

Scatter Plot Example 5.2



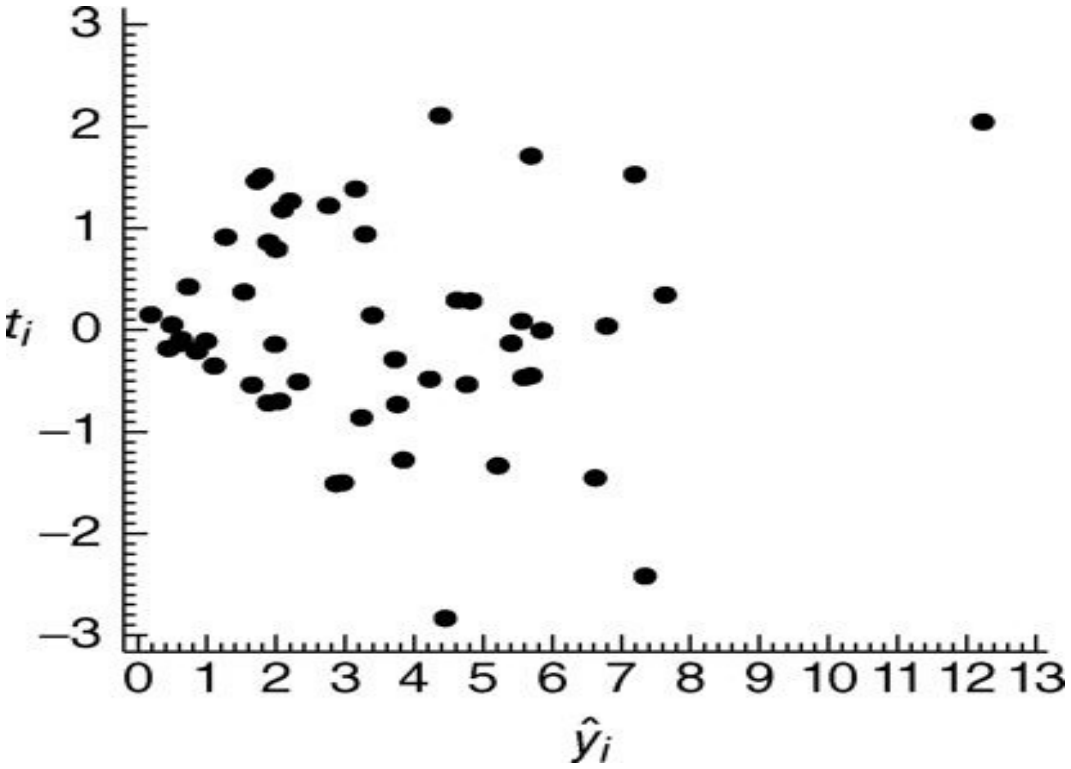
The ANOVA table is

Source	Sum of Squares	DF	Mean Square	F-Stat	p-value
Regression	302.6331	1	302.6331	121.66	< 0.0001
Residual	126.8660	51	2.4876		
Total	429.4991	52			

The analysis of variance Table above shows the model is highly significant.

The model $R^2 = 0.7046$ that is, about 70% of the variability in demand is accounted for by the straight-line fit to energy usage which is apparently reasonable. The summary statistics do not reveal any obvious problems with this model. Also there seems to be one high leverage point (needs to checked).

$\hat{y}$ vs t_i - Example 5.1



A plot of the R-student residuals versus the fitted values shows that error variance may be increasing with the mean. That is to say that the error

variance is increasing as energy consumption increases. A transformation may be helpful in correcting this model inadequacy.

A common transformation to try first is square root transformation. Hence the new model tried is

$$y^{\star} = \sqrt{y} = \beta_0 \; + \; \beta_1 \, x \; + e$$

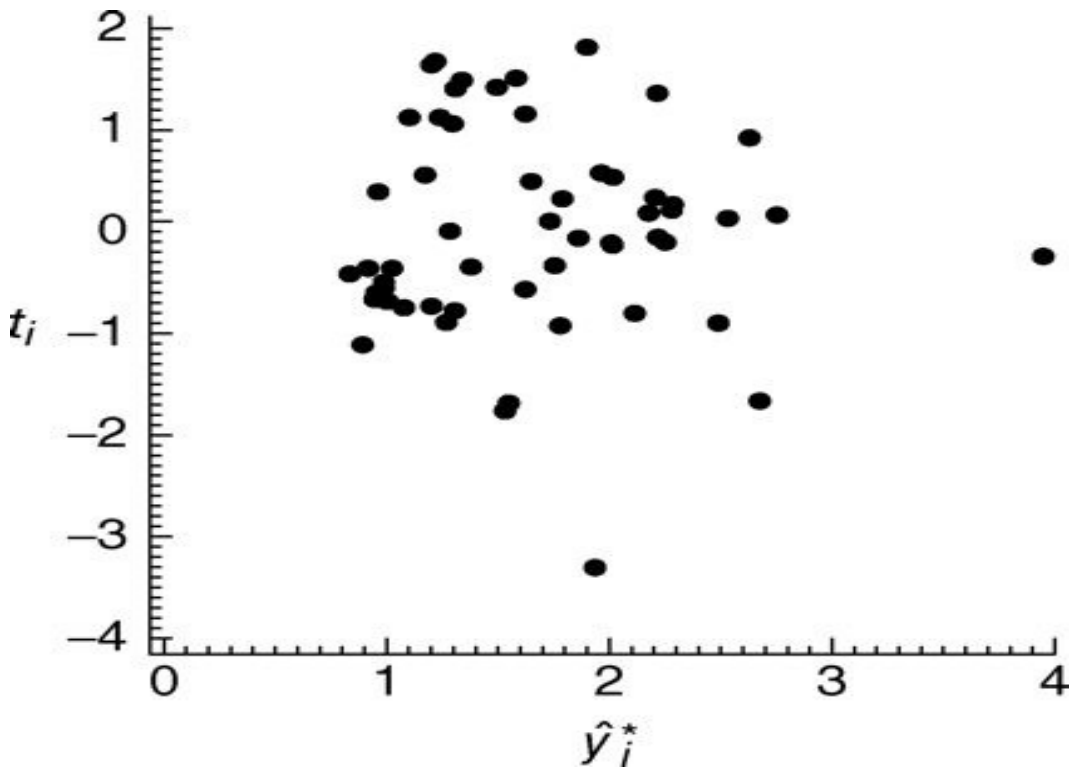
the resulting least square is

$$\hat{y^{\star}} = 0.5822 \; + \; 0.0009529 \, x$$

So the next step is to go through the same plotting process to check whether the variance has stabilized or not.

The R-student values from this new least-squares fit are plotted against again (next slide). The visual inspection of this graph suggests that the variance is stable. Consequently, we conclude that the transformed model is adequate.

$\hat{y}$ vs t_i - Transformed Data Example 5.1



Note that there is one suspiciously large residual and one customer whose energy usage is somewhat large. The effect of these two points on the fit should be studied further before the model is released for use.

2. Linearizing the Model

Even though it is possible to use Non-Linear Regression theory to model relationship, in some cases a nonlinear function can be linearized by using a suitable transformation. The assumption of a linear relationship between y and the regressors is the usual starting point in multiple linear regression analysis. Nonlinearity may be detected via

(1) The lack-of-fit test

(2)The matrix of scatter plots.

(3) Residual plots.

(4) Prior experience or theoretical considerations.

To understand a nonlinear model which can be linearize, consider the exponential function

$$y = \beta_0 \times e^{\beta_1 x} \times \epsilon \quad \text{Multiplicative error}$$

This function is intrinsically linear since it can be transformed to a straight line by a logarithmic transformation

$$ln(y) = ln(\beta_0) + \beta_1 x + ln(\epsilon)$$

This transformation requires that the transformed error terms $ln(\epsilon)$ are normally and independently distributed with mean zero and variance σ^2. This implies that the multiplicative error ϵ in the original model is log-normally distributed. We should look at the residuals from the transformed model to see if the assumptions are valid.

Various types of reciprocal transformations are also useful. For example, the model

$$y = \beta_0 + \beta_1 \frac{1}{x} + \epsilon$$

can be linearized by using the reciprocal transformation as $x' = 1/x$.

Sometimes if the the scatter plot is "Parabola" type then adding a second degree term in the model may be enough instead of linearizing. Then the model is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

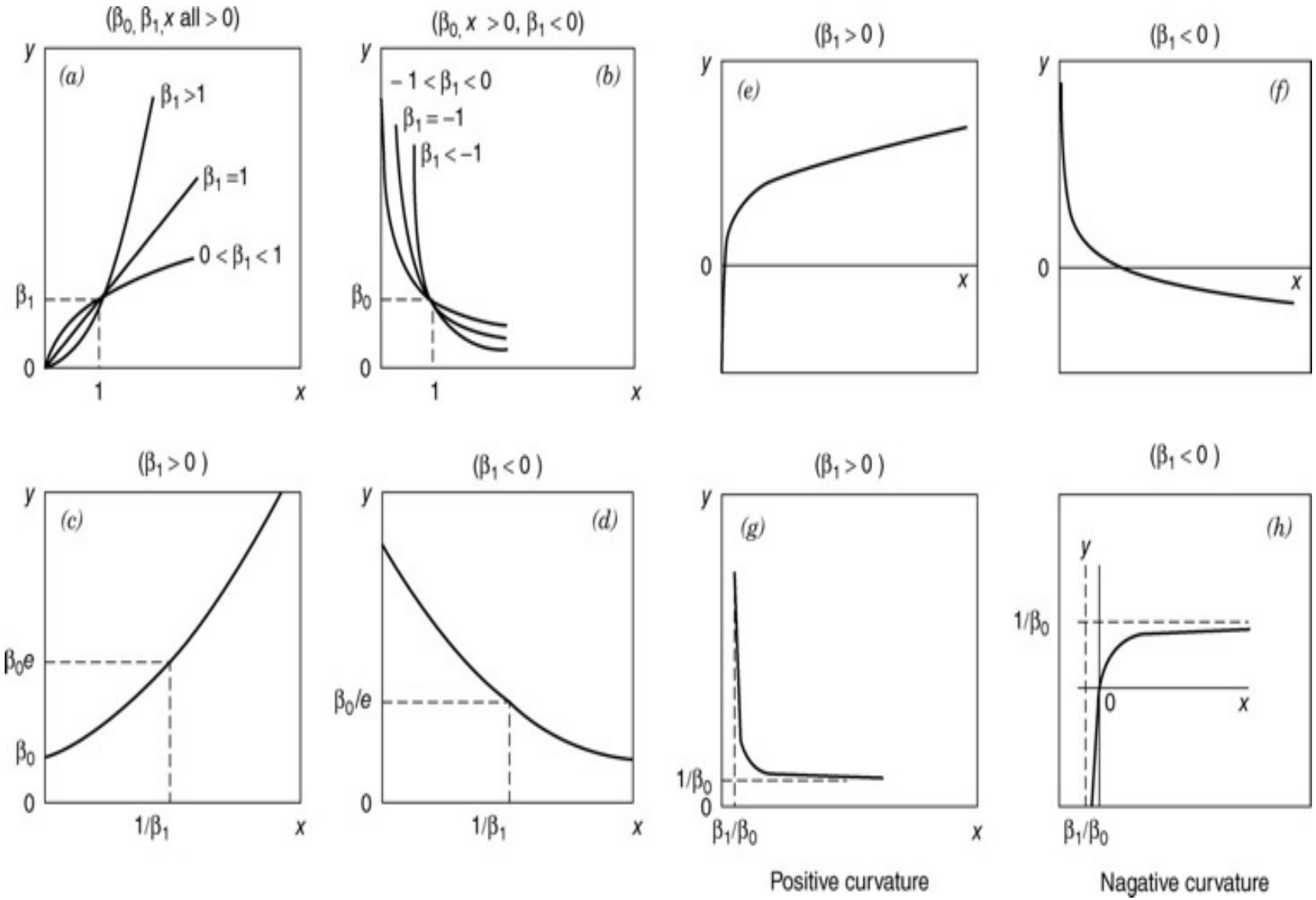
The above model is treated just like other MLR model where $x_1 = x \,\&\, x_2 = x^2$.

Several linearizable functions are shown in graph (next slide) and the corresponding nonlinear functions are shown in the following table. When the scatter diagram of y against x indicates curvature, we may be able to match the observed behavior of the plot to one of the curves in graphs and use the linearized form of the function to represent the data. It is possible that two different models look appropriate and in that situation both can be tried and compared.

Some Common Graph to Linearize

Figure	Linearizable Function	Transformation	Linear Form
(a) (b)	$y = \beta_0 \, . \, x^{\beta_1}$	$y' = log(y),\ x' = log(x)$	$y' = log(\beta_0) + \beta_1 \, x'$
(c) (d)	$y = \beta_0 \, . \, e^{\beta_1 \, x}$	$y' = ln(y)$	$y' = ln(\beta_0) + \beta_1 \, x$
(e) (f)	$y = \beta_0 + \beta_1 \, log(x)$	$x' = log(x)$	$y = \beta_0 + \beta_1 \, x'$
(g) (h)	$y = \frac{x}{\beta_0 \, x - \beta_1}$	$y' = \frac{1}{y} \, x' = \frac{1}{x}$	$y' = \beta_0 - \beta_1 \, x'$

Scatter Plot - Linearize



Positive curvature

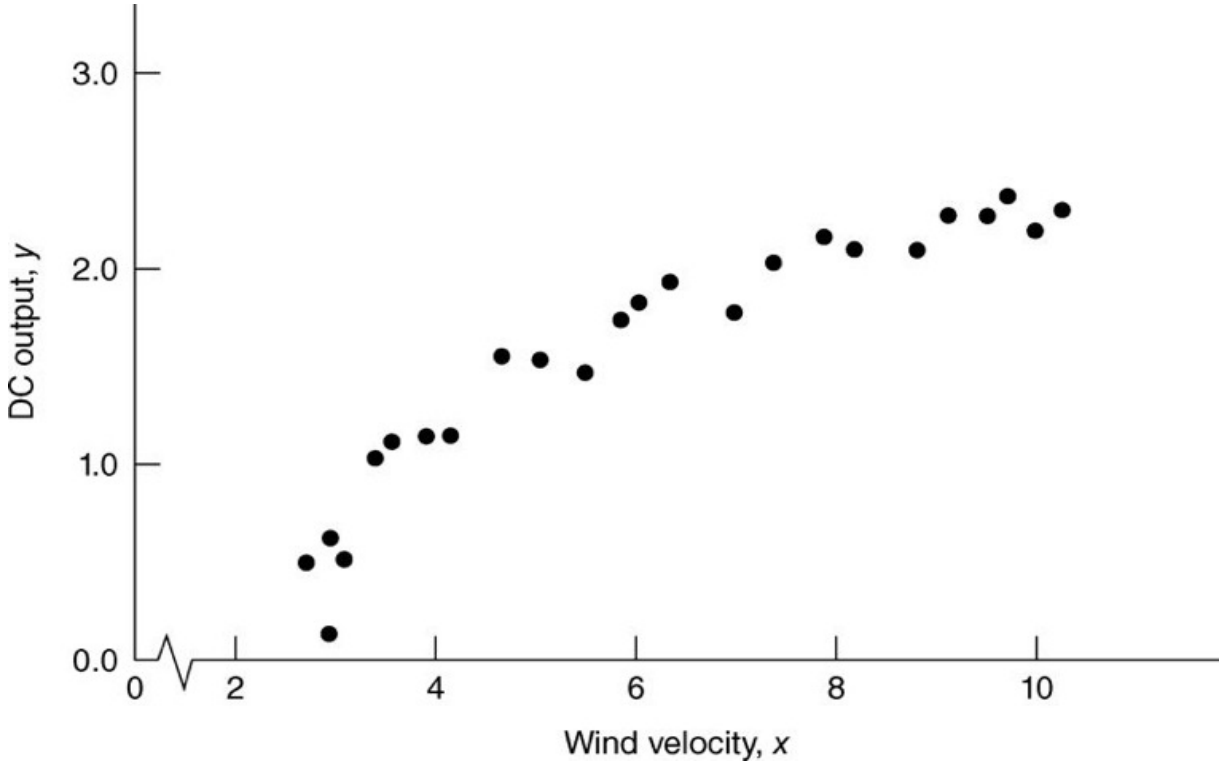
Nagative curvature

Example 5.3 (Original Data set is in the reference book A research engineer is investigating the use of a windmill to generate electricity. He has collected data on the DC output from his windmill and the corresponding wind velocity. Scatter plot is as follows.

Data - Example 5.3

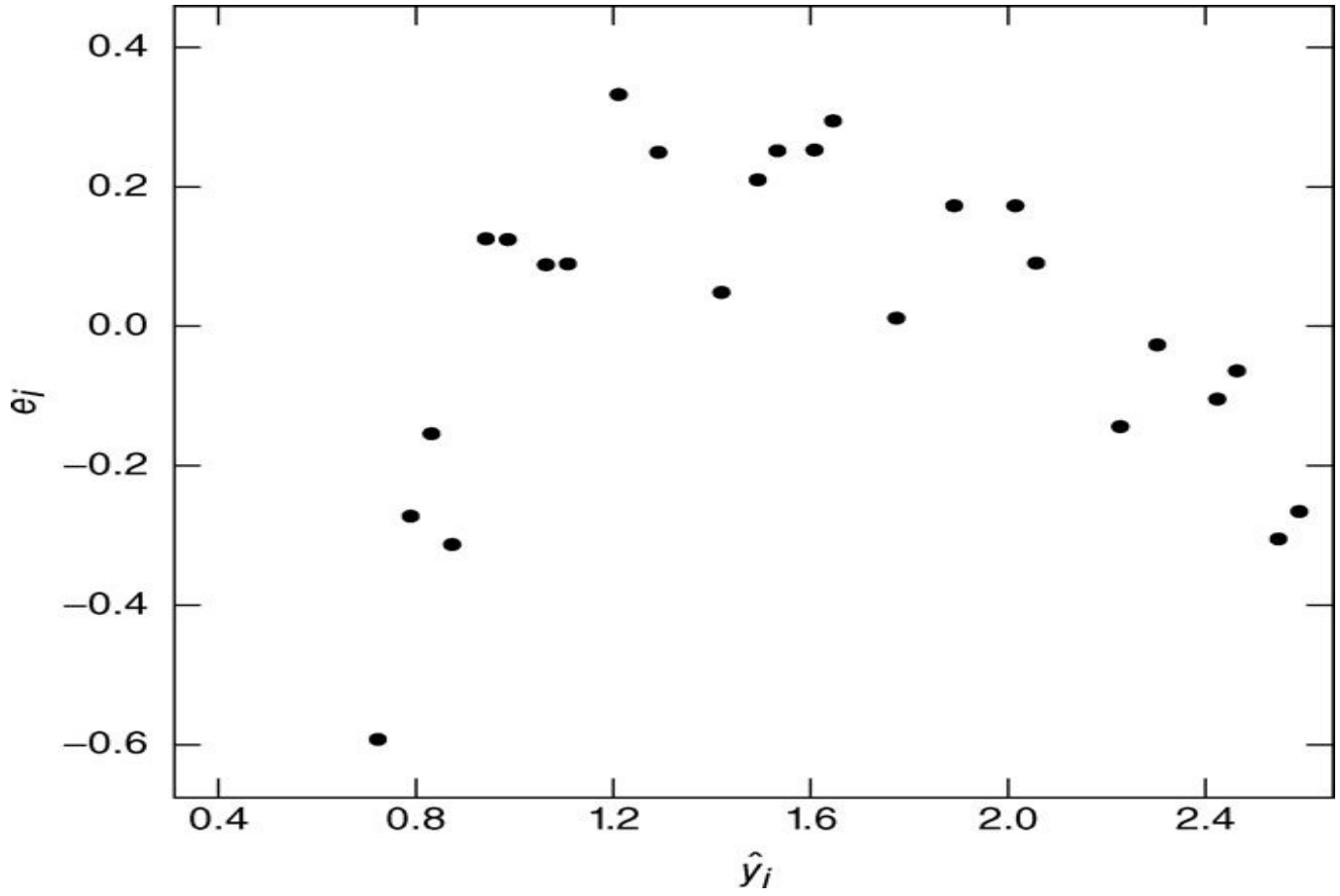
Observation Number, i	Wind Velocity, x_i (mph)	DC Output, y_i
1	5.00	1.582
2	6.00	1.822
3	3.40	1.057
4	2.70	0.500
5	10.00	2.236
6	9.70	2.386
7	9.55	2.294
8	3.05	0.558
9	8.15	2.166
10	6.20	1.866
11	2.90	0.653
12	6.35	1.930
13	4.60	1.562
14	5.80	1.737
15	7.40	2.088
16	3.60	1.137
17	7.85	2.179
18	8.80	2.112
19	7.00	1.800
20	5.45	1.501
21	9.10	2.303
22	10.20	2.310
23	4.10	1.194
24	3.95	1.144
25	2.45	0.123

Scatter Plot - Example 5.3



Inspection of the scatter diagram indicates that the relationship between DC output (y) and wind velocity (x) may be nonlinear. However, we initially fit a straight-line model to the data. The summary statistics for the SLR model are $R^2 = 0.8745$, $MSR = 0.0557$, and $F - Stat = 160.26$ (the P value is less than 0.0001).

Residual vs $\hat{y}$ - Example 5.3



This residual plot indicates model inadequacy and implies that the linear relationship has not captured all of the information in the wind speed variable. Note that the curvature that was apparent in the scatter diagram is greatly amplified in the residual plot.

Clearly some other model form must be considered. There two possibilities:

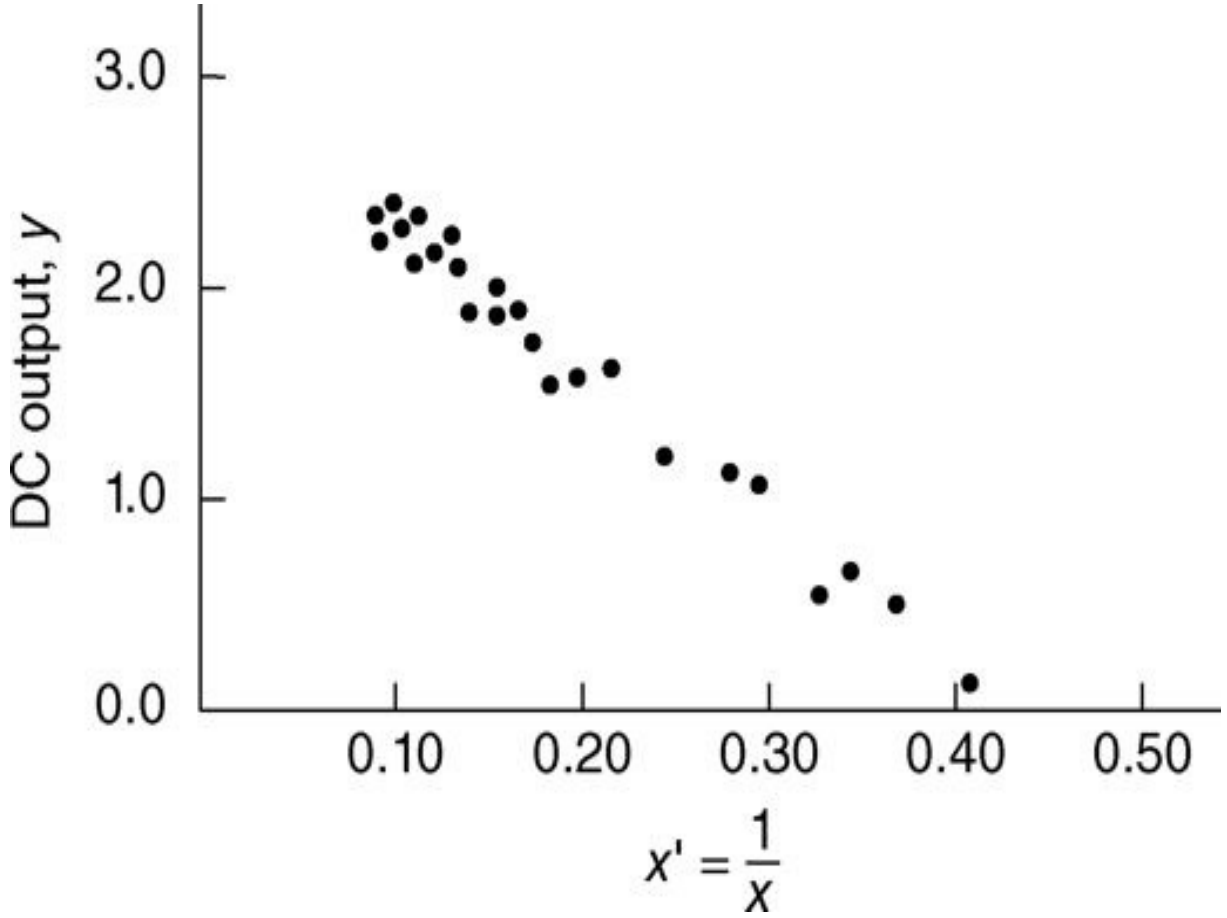
$$\text{(1) - Quadratic Model} \quad y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

or

$$\text{(2) - Reciprocal Model} \quad y = \beta_0 + \beta_1 \frac{1}{x} + \epsilon$$

We might initially consider using a quadratic model such as to account for the apparent curvature. But theory of windmill suggests that as wind speed increases, DC output approaches an upper limit of approximately 2.5. Hence a more reasonable model for the windmill data that incorporates an upper asymptote would be the reciprocal model. Scatter diagram with the transformed variable $x' = 1/x$ shows almost a perfect straight line indicating that the reciprocal transformation is appropriate.

Scatter Plot - Transformed - Example 5.3



The new transformed fitted regression model summary statistics are $R^2 = 0.9800$, $MSR = 0.0089$, $F_0 = 1128.43$ & $P - value < 0.0001$. A plot of R-student against $\hat{y}$ values does not reveal any serious problem with inequality of variance. Other residual plots are satisfactory, and so because there is no strong signal of model inadequacy, we conclude that the transformed

model is satisfactory.

Residual vs $\hat{y}$ - Transformed - Example 5.3