DATA ANALYTICS

Unit 2:Linear Regression

Mamatha.H.R and Bharathi R
Department of Computer Science and Engineering

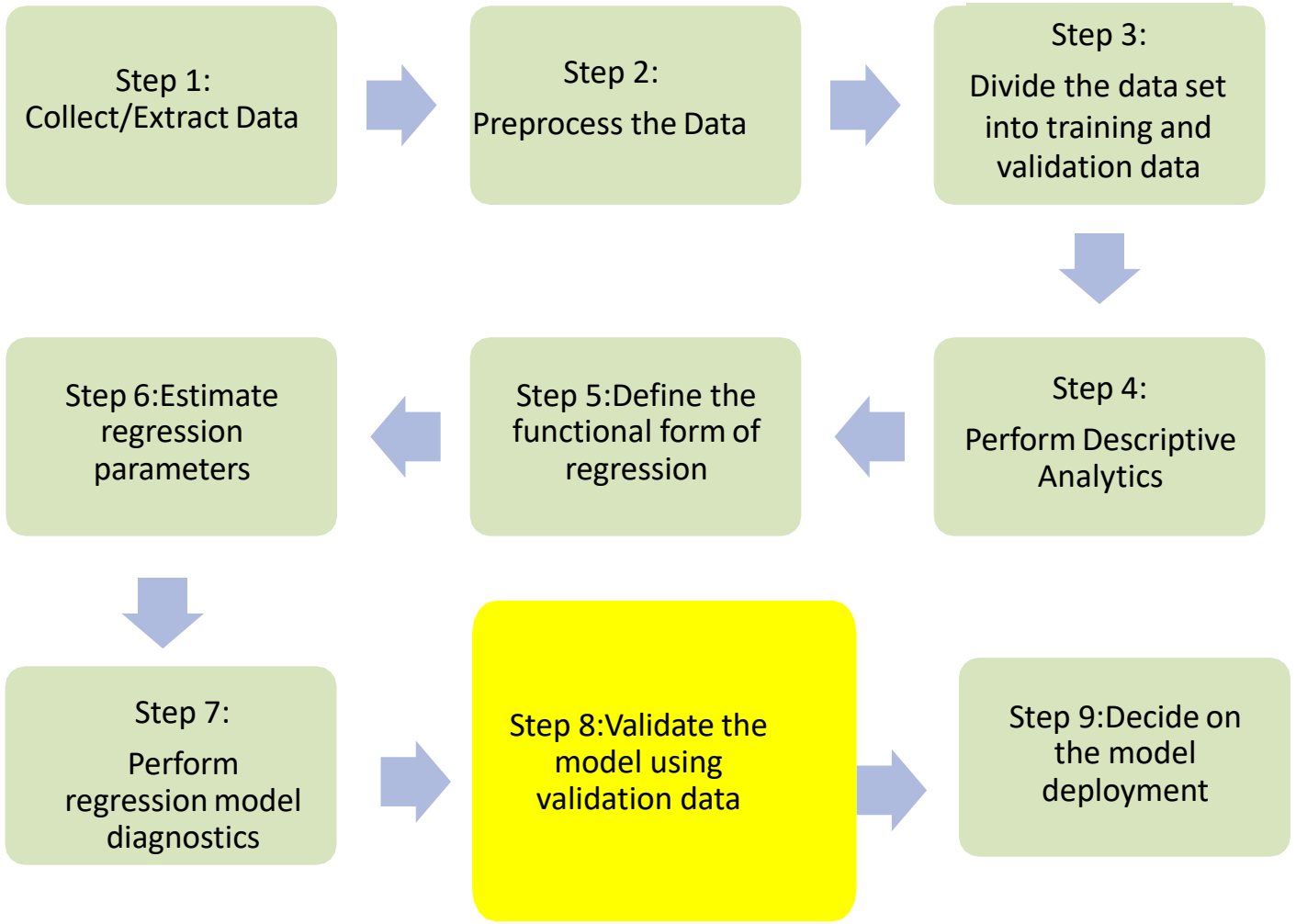
DATA ANALYTICS

Unit 2:Linear Regression Contd.,

Mamatha H R

Department of Computer Science and Engineering

Framework for SLR model development



Validation of the simple linear regression model

It is important to validate the regression model to ensure its validity and goodness of fit before it can be used for practical applications. The following measures are used to validate the simple linear regression models:

1. Co-efficient of determination **(R-square).**
2. Hypothesis test for the regression coefficient **β1**.
3. Analysis of Variance for overall model validity (relevant more for multiple linear regression)- **ANOVA**.
4. Residual analysis to validate the regression model assumptions.
5. Outlier analysis.

The above measures and tests are essential, but not exhaustive.

1.Coefficient of Determination (R-Square or R^2)

- The co-efficient of determination (or *R*-square or R^2) measures the percentage of variation in *Y* explained by the model ($\beta_0 + \beta_1 X$).
- The simple linear regression model can be broken into explained variation and unexplained variation as shown in

$$\underbrace{Y_i}_{\text{Variation in Y}} = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Variation in Y explained by the model}} + \underbrace{\varepsilon_i}_{\text{Variation in Y not explained by the model}}$$

In absence of the predictive model for Y_i, the users will use the mean value of Y_i. Thus, the total variation is measured as the difference between Y_i and mean value of Y_i (i.e., $Y_i - \bar{Y}$).

Description of total variation, explained variation and unexplained variation

Variation Type	Measure	Description
Total Variation (SST)	$(Y_i - \bar{Y})$	Total variation is the difference between the actual value and the mean value.
Variation explained by the model	$(\hat{Y}_i - \bar{Y})$	Variation explained by the model is the difference between the estimated value of Y_i and the mean value of Y
Variation not explained by model	$(Y_i - \hat{Y}_i)$	Variation not explained by the model is the difference between the actual value and the predicted value of Y_i (error in prediction)

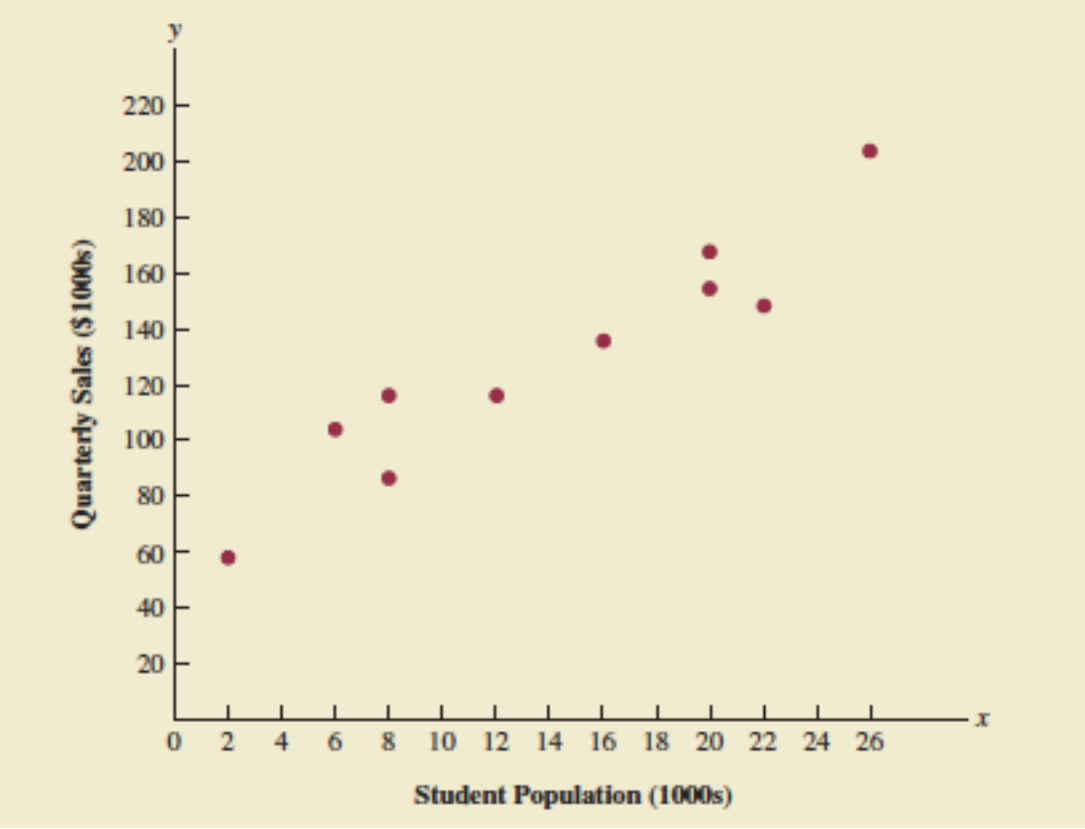
Exercise

STUDENT POPULATION AND QUARTERLY SALES DATA
FOR 10 Restaurants

Restaurant i	Student Population (1000s) x_i	Quarterly Sales ($1000s) y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Exercise

STUDENT POPULATION AND QUARTERLY SALES DATA
FOR 10 Restaurant : **Scatter Plot**



Exercise

STUDENT POPULATION AND QUARTERLY SALES DATA FOR 10 Restaurants

Calculations for the least squares estimated regression Equation

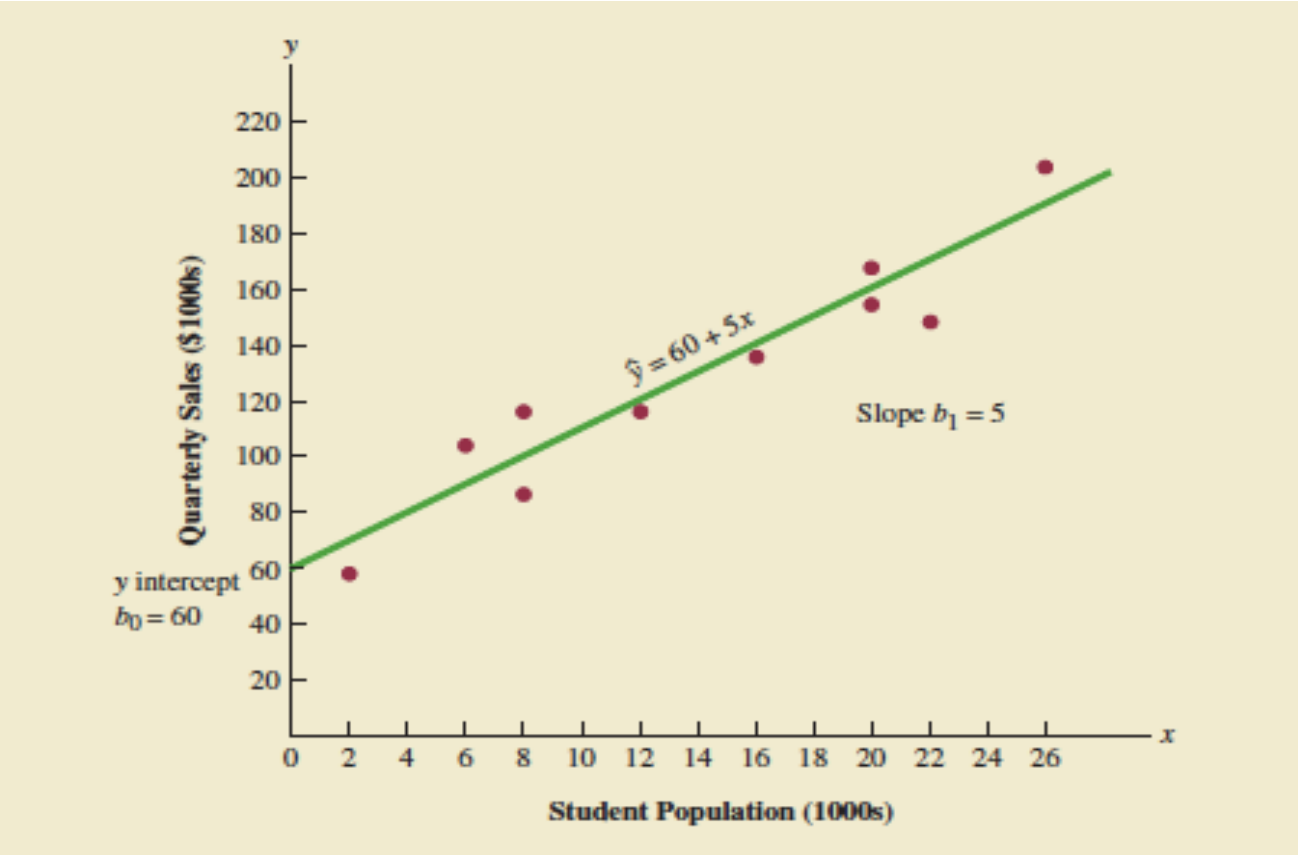
Restaurant i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totals	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

Thus, the estimated regression equation is

$\hat{y} = 60 + 5x$

Exercise STUDENT POPULATION AND QUARTERLY SALES DATA FOR 10 Restaurants

Graph of the estimated regression equation $\hat{y} = 60 + 5x$



Exercise

STUDENT POPULATION AND QUARTERLY SALES DATA FOR 10 Restaurant

Restaurant i	x_i = Student Population (1000s)	y_i = Quarterly Sales ($1000s)	Predicted Sales $\hat{y}_i = 60 + 5x_i$	Error $y_i - \hat{y}_i$	Squared Error $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
				SSE =	1530

CALCULATION OF SSE

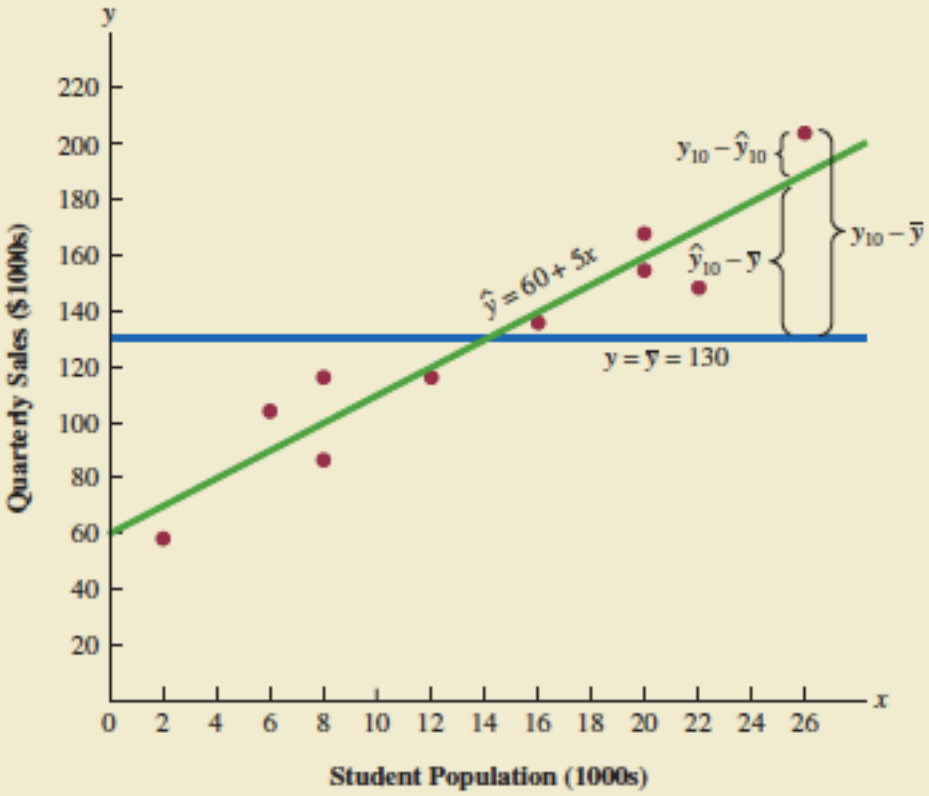
Restaurant i	x_i = Student Population (1000s)	y_i = Quarterly Sales ($1000s)	Deviation $y_i - \bar{y}$	Squared Deviation $(y_i - \bar{y})^2$
1	2	58	−72	5184
2	6	105	−25	625
3	8	88	−42	1764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1521
9	22	149	19	361
10	26	202	72	5184
			SST =	15,730

Computation of the total sum of squares

DATA ANALYTICS

Exercise

DEVIATIONS ABOUT ThE ESTIMATED REGRESSION LINE AND ThE LINE y 5 y FOR ARMAND'S PIZZA PARLORS



Exercise

STUDENT POPULATION AND QUARTERLY SALES DATA FOR 10 Restaurant

Coefficient of Determination: the coefficient of determination provides a measure of the goodness of fit for the estimated regression equation.

SUM OF SQUARES DUE TO ERROR

$$SSE = \Sigma(y_i - \hat{y}_i)^2$$

The value of SSE is a measure of the error in using the estimated regression equation to predict the values of the dependent variable in the sample.

TOTAL SUM OF SQUARES

$$SST = \Sigma(y_i - \bar{y})^2$$

RELATIONSHIP AMONG SST, SSR, AND SSE

$$SST = SSR + SSE$$

where

SST = total sum of squares
SSR = sum of squares due to regression
SSE = sum of squares due to error

SUM OF SQUARES DUE TO REGRESSION

$$SSR = \Sigma(\hat{y}_i - \bar{y})^2$$

COEFFICIENT OF DETERMINATION

$$r^2 = \frac{SSR}{SST}$$

$$r^2 = \frac{SSR}{SST} = \frac{14{,}200}{15{,}730} = .9027$$

DATA ANALYTICS

Exercise

Correlation Coefficient : the correlation coefficient as a descriptive measure of the
strength of linear association between two variables, x and y. Values of the correlation
coefficient are always between −1 and +1.

SAMPLE CORRELATION COEFFICIENT

$$r_{xy} = (\text{sign of } b_1)\sqrt{\text{Coefficient of determination}}$$
$$= (\text{sign of } b_1)\sqrt{r^2}$$

where

b_1 = the slope of the estimated regression equation $\hat{y} = b_0 + b_1 x$

The relationship between the total variation, explained variation and the unexplained variation is given as follows:

$$\underbrace{Y_i - \bar{Y}}_{\text{Total Variation in Y}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\text{Variation in Y explained by the model}} + \underbrace{Y_i - \hat{Y}_i}_{\text{Variation in Y not explained by the model}}$$

It can be proved mathematically that sum of squares of total variation is equal to sum of squares of explained variation plus sum of squares of unexplained variation

$$\underbrace{\sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2}_{SST} = \underbrace{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y} \right)^2}_{SSR} + \underbrace{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2}_{SSE}$$

where SST is the sum of squares of total variation, SSR is the sum of squares of variation explained by the regression model and SSE is the sum of squares of errors or unexplained variation.

Coefficient of Determination or R-Square

The coefficient of determination (R^2) is given by

$$\text{Coefficient of determination} = R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{SSR}{SST} = \frac{\left(\hat{Y}_i - \bar{Y}\right)^2}{\left(Y_i - \bar{Y}\right)^2}$$

Since SSR = SST – SSE, the above Eq. can be written as

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\left(\hat{Y}_i - Y_i\right)^2}{\left(Y_i - \bar{Y}\right)^2}$$

Coefficient of Determination or R-Square

Thus, R^2 is the proportion of variation in response variable Y explained by the regression model. Coefficient of determination (R^2) has the following properties:

- The value of R^2 lies between 0 and 1.
- Higher value of R^2 implies better fit, but one should be aware of spurious regression.
- Mathematically, the square of correlation coefficient is equal to coefficient of determination (i.e., $r^2 = R^2$).
- We do not put any minimum threshold for R^2; higher value of R^2 implies better fit. However, a minimum value of R^2 for a given significance value α can be derived using the relationship between the F-statistic and R^2

Spurious Regression

One of the major problems with coefficient of determination (R^2) is that **two sets of data without any relationship can have a very high coefficient of determination value.**

Spurious Regression

Number of Facebook users and the number of people who died of helium poisoning in UK

Year	Number of Facebook users in millions (X)	Number of people who died of helium poisoning in UK (Y)
2004	1	2
2005	6	2
2006	12	2
2007	58	2
2008	145	11
2009	360	21
2010	608	31
2011	845	40
2012	1056	51

DATA ANALYTICS

Facebook users versus helium poisoning in UK

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.996442					
R Square	**0.992896**					
Standard Error	1.69286					
Observations	9					
ANOVA						
		SS	MS	F	Significance F	
Regression	1	2803.94	2803.94	978.4229	8.82E-09	
Residual	7	20.06042	2.865775			
Total	8	2824				
	Coefficients	Standard Error	t-stat	P-value	Lower 95%	Upper 95%
Intercept	1.9967	0.76169	2.62143	0.034338	0.195607	3.79783
FB	0.0465	0.00149	31.27975	8.82E-09	0.043074	0.050119

A high R-square value is not necessarily a good indicator of the correctness of the model; it could be a spurious relationship.

The *R*-square value for regression model between the number of deaths due to helium poisoning in UK and the number of Facebook users is 0.9928. That is, 99.28% variation in the number of deaths due to helium poisoning in UK is explained by the number of Facebook users.

The regression model is given as $Y = 1.9967 + 0.0465\ X$

2.Hypothesis Test for Regression Co-efficient (t-Test)

- The regression co-efficient (β_1) captures the existence of a linear relationship between the response variable and the explanatory variable.
- If $\beta_1 = 0$, we can conclude that there is no statistically significant linear relationship between the two variables.

➢ The estimate of β_1 using OLS is given by

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} - \frac{\bar{X}\sum_{i=1}^{n}(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\beta_1 = \frac{\sum_{i=1}^{n} K_i Y_i}{\sum_{i=1}^{n} K_i^2} \text{ where } K_i = (X_i - \bar{X})$$

Above eq. can be written as follows:

That is, the value of β_1 is a function of Y_i (K_i is a constant since X_i is assumed to be non-stochastic)

The standard error of β_1 is given by

$$S_e(\hat{\beta}_1) = \frac{S_e}{\sqrt{(X_i - \bar{X})^2}}$$

Contd.,

In above Eq. Se is the standard error of estimate (or standard error of the residuals) that measures the accuracy of prediction and is given by

$$S_e = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-2}}$$

The denominator in above Eq. is (n – 2) since β_0 and β_1 are estimated from the sample in estimating Y_i and thus two degrees of freedom are lost. The standard error of $\hat{\beta}_1$ can be written as

$$S_e(\hat{\beta}_1) = \frac{S_e}{\sqrt{\sum_{t=1}^{n} (X_t - \bar{X})^2}} = \frac{\sqrt{\sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2 \Big/ n-2}}{\sqrt{\sum_{t=1}^{n} (X_t - \bar{X})^2}}$$

Contd.,

The null and alternative hypotheses for the SLR model can be stated as follows:

H_0: There is no relationship between *X* and *Y*

H_A: There is a relationship between *X* and *Y*

- $\beta_1 = 0$ would imply that there is no linear relationship between the response variable *Y* and the explanatory variable *X*. Thus, the null and alternative hypotheses can be restated as follows:

  $H_0: \beta_1 = 0$

  $H_A: \beta_1 \neq 0$

- The corresponding *t*-statistic is given as

$$t = \frac{\hat{\beta}_1 - \beta_1}{Se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{Se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{Se(\hat{\beta}_1)}$$

DATA ANALYTICS

Salary of MBA students versus their grade 10 marks : Table 9.2

S. No.	Percentage in Grade 10	Salary	S. No.	Percentage in Grade 10	Salary
1	62	270000	26	64.6	250000
2	76.33	200000	27	50	180000
3	72	240000	28	74	218000
4	60	250000	29	58	360000
5	61	180000	30	67	150000
6	55	300000	31	75	250000
7	70	260000	32	60	200000
8	68	235000	33	55	300000
9	82.8	425000	34	78	330000
10	59	240000	35	50.08	265000
11	58	250000	36	56	340000
12	60	180000	37	68	177600
13	66	428000	38	52	236000
14	83	450000	39	54	265000
15	68	300000	40	52	200000
16	37.33	240000	41	76	393000
17	79	252000	42	64.8	360000
18	68.4	280000	43	74.4	300000
19	70	231000	44	74.5	250000
20	59	224000	45	73.5	360000
21	63	120000	46	57.58	180000
22	50	260000	47	68	180000
23	69	300000	48	69	270000
24	52	120000	49	66	240000
25	49	120000	50	60.8	300000

Solution

Using Eqs., the estimated values of β_0 and β_1 are given by

$$\hat{\beta}_0 = 61555.3553 \text{ and } \hat{\beta}_1 = 3076.1774$$

The corresponding regression equation is given by

$$\hat{Y}_i = 61555.3553 + 3076.1774 X_i$$

Where $\hat{Y}_i$ is the predicted value of Y for a given value of X_i. .

The equation can be interpreted as follows:

for every one percentage increase in grade 10 marks, the salary of the MBA students will increase at the rate of 3076.1774 on an average. The notations

DATA ANALYTICS

Solution Continued

$\hat{\beta}_0$ and $\hat{\beta}_1$ are used to denote that these are estimated values of the regression coefficients from the sample of 50 students.

Regression coefficient estimates using Microsoft Excel Table 9.6

	Coefficients	Standard Error	t-stat	p-value
Intercept	61555.35534	66701.901	0.9228	0.3607
Percentage in grade 10	3076.177438	1031.5258	2.9821	0.0044

At the .05 level of significance , is there any significant relationship between the two variables?

Exercise

At the .05 level of significance , is there any significant relationship between the two variables?

The t-value for the variable percentage marks grade 10 is given by

$$t = \frac{\hat{\beta}_1}{S_e(\hat{\beta}_1)} = \frac{3076.1774}{1031.526} = 2.9821$$

- The corresponding degrees of freedom **(df) is (n – 2)** which in this case is 50 – 2 = 48.
- Here two degrees of freedom are lost since β_0 and β_1 are estimated from the data.
- This is a two-tailed test, the **critical t-value is 2.01 for a = 0.05 and df = 48**.
- The **p-value** corresponding to **t = 2.9821** with 48 degrees of freedom is **0.0044** .
- Since the **p-value is less than 0.05**, we **reject** the null hypothesis and conclude that there is significant evidence suggesting a **linear relationship between X and Y.**

3. Test for Overall Model: Analysis of Variance (F-test)

The null and alternative hypothesis for *F*-test is given by

H_0: There is no statistically significant relationship between *Y* and any of the explanatory variables (i.e., all regression coefficients are zero).

H_A: Not all regression coefficients are zero

- Alternatively:

H_0: All regression coefficients are equal to zero

H_A: Not all regression coefficients are equal to zero

- The *F*-statistic is given by

$$F = \frac{MSR}{MSE} = \frac{MSR/1}{MSE/n-2}$$

Exercise : Salary of MBA students versus their grade 10

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/n-2} = \frac{4.51\times10^{10}}{5.07\times10^{9}} = 8.8932$$

- The p-value corresponding to F-statistic value of 8.8932 is 0.0044. Since the p-value is less than 0.05 (assume that a = 0.05), the null hypothesis is rejected.
- Note that the p-value of t-test and F-test are same in Table 9.6.
- This is due to the fact that the model has only one independent variable and the null hypothesis for both t-test and F-test are identical (in SLR, F = t2).

The mathematical relationship between F-statistic and R2 in a simple linear regression is given by

$$F = \frac{R^2}{(1-R)^2/(n-2)} = \frac{0.156315}{(1-0.156315)/48} = 8.8932$$

Note. On F test

In a simple linear regression $F = t^2$ and the p-value is same since F-test and t-test are equivalent for SLR.

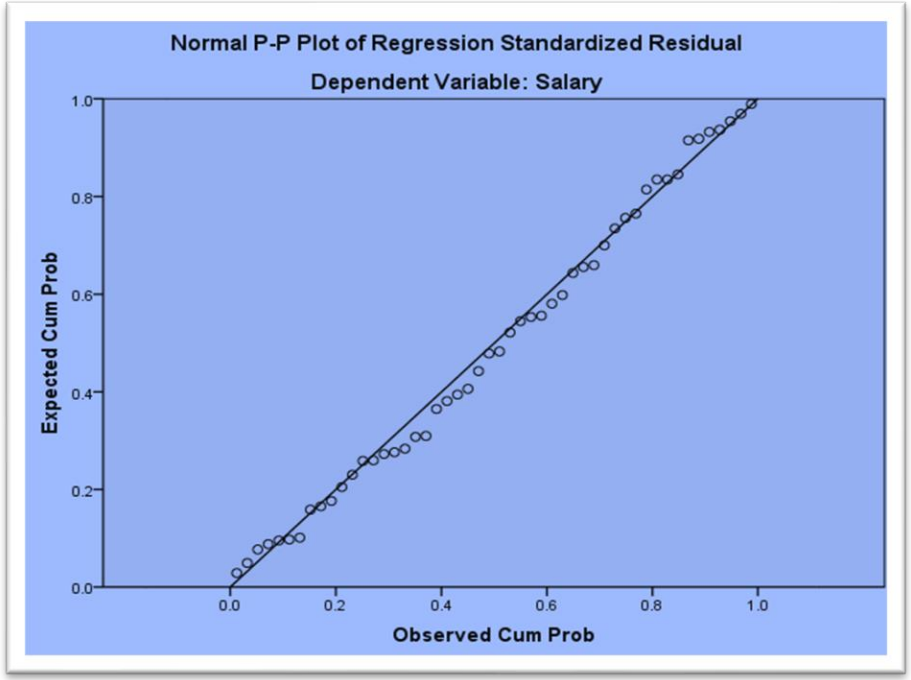
4.Residual Analysis

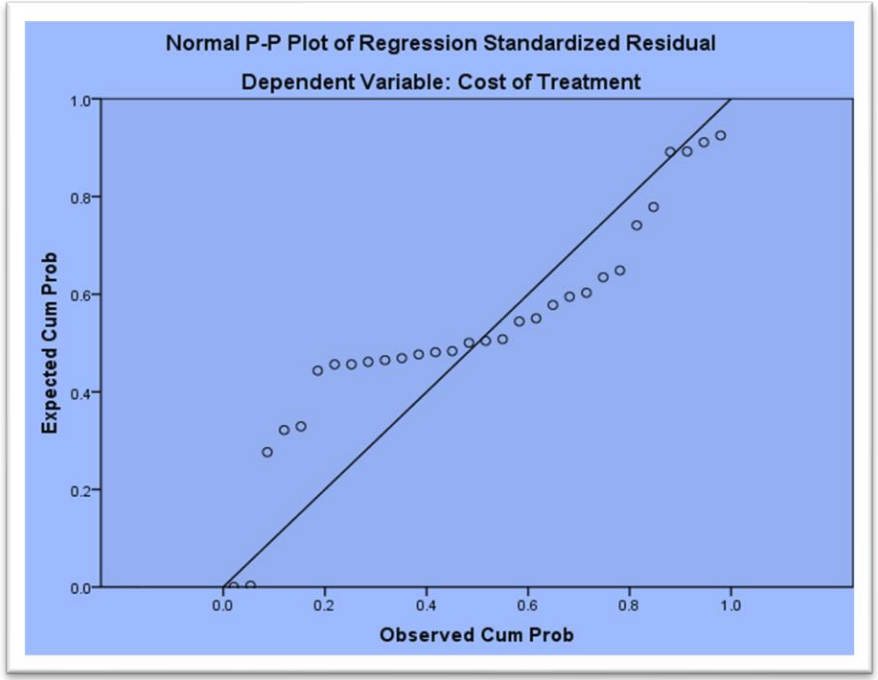
Residual (error) analysis is important to check whether the assumptions of regression models have been satisfied. It is performed to check the following:

1. The residuals $(Y_i - \hat{Y}_i)$ are normally distributed.
2. The variance of residual is constant (homoscedasticity).
3. The functional form of regression is correctly specified.
4. If there are any outliers

4.1.Checking for Normal Distribution of Residuals $(Y_i - \hat{Y}_i)$

- The easiest technique to check whether the residuals follow normal distribution is to use the P-P plot (Probability-Probability plot).
- The P-P plot compares the cumulative distribution function of two probability distributions against each other



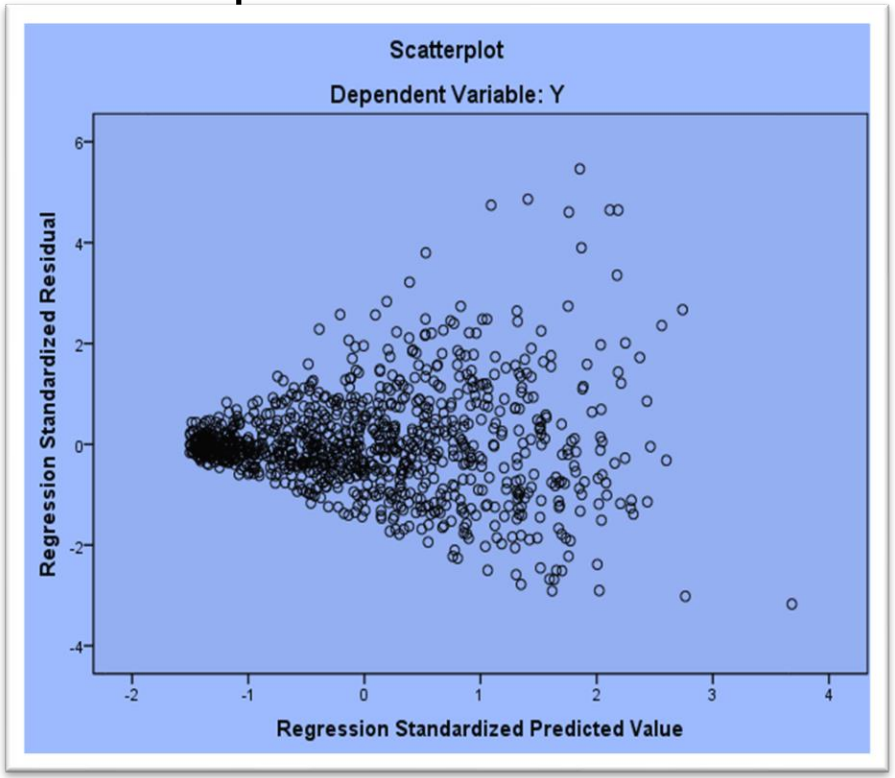


4.2. Test of Homoscedasticity

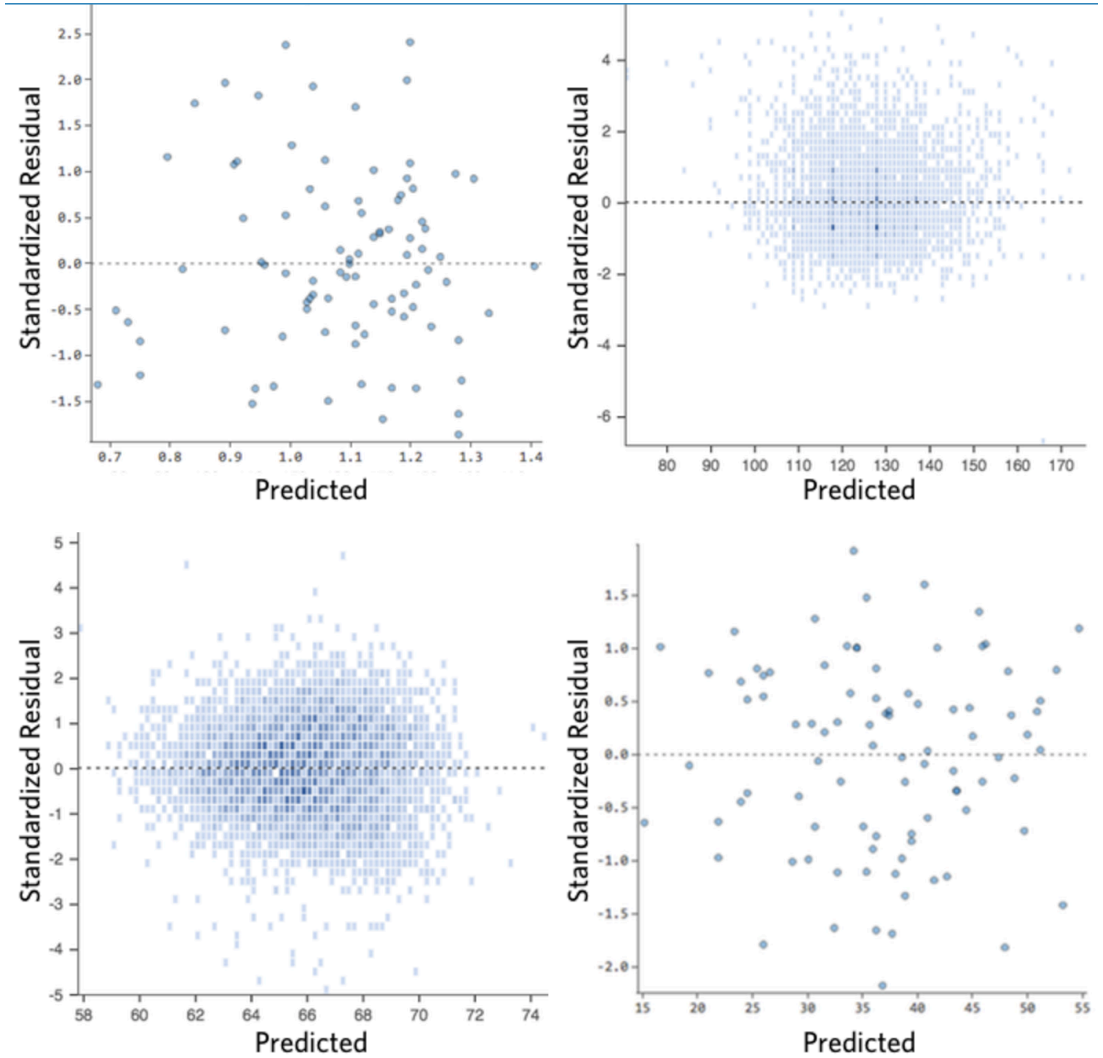
An important assumption of regression model is that the residuals have constant variance (homoscedasticity) across different values of the explanatory variable (*X*).

That is, the variance of residuals is assumed to be independent of variable *X*. Failure to meet this assumption will result in unreliability of the hypothesis tests.

Funnel shape in the standardized residual plot indicates heteroscedasticity.



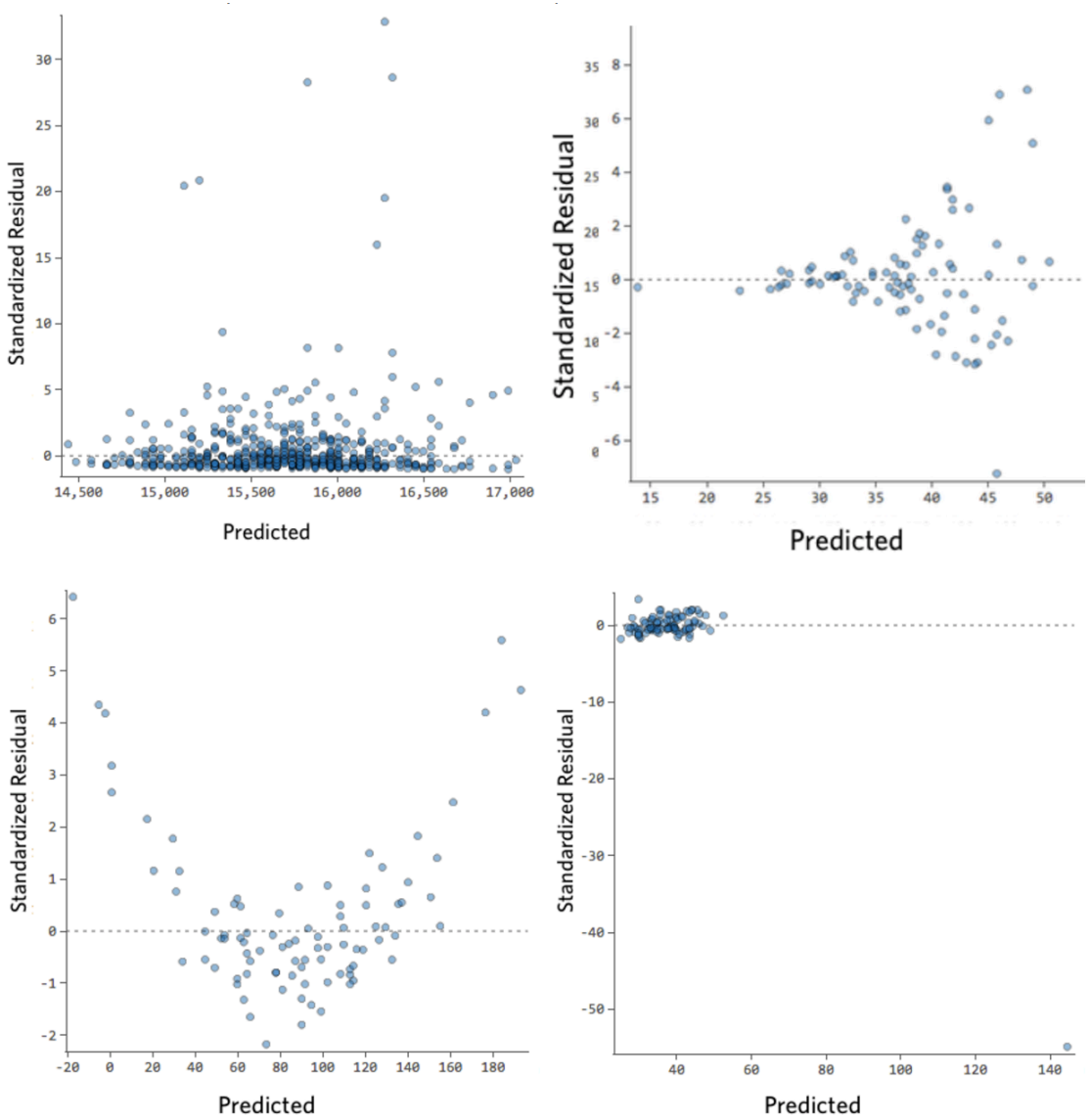
4.2. Test of Homoscedasticity- Examples



(1) they're pretty symmetrically distributed, tending to cluster towards the middle of the plot.
(2) they're clustered around the lower single digits of the y-axis (e.g., 0.5 or 1.5, not 30 or 150).
(3) in general, there aren't any clear patterns.

https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

4.2. Test of Homoscedasticity- Examples

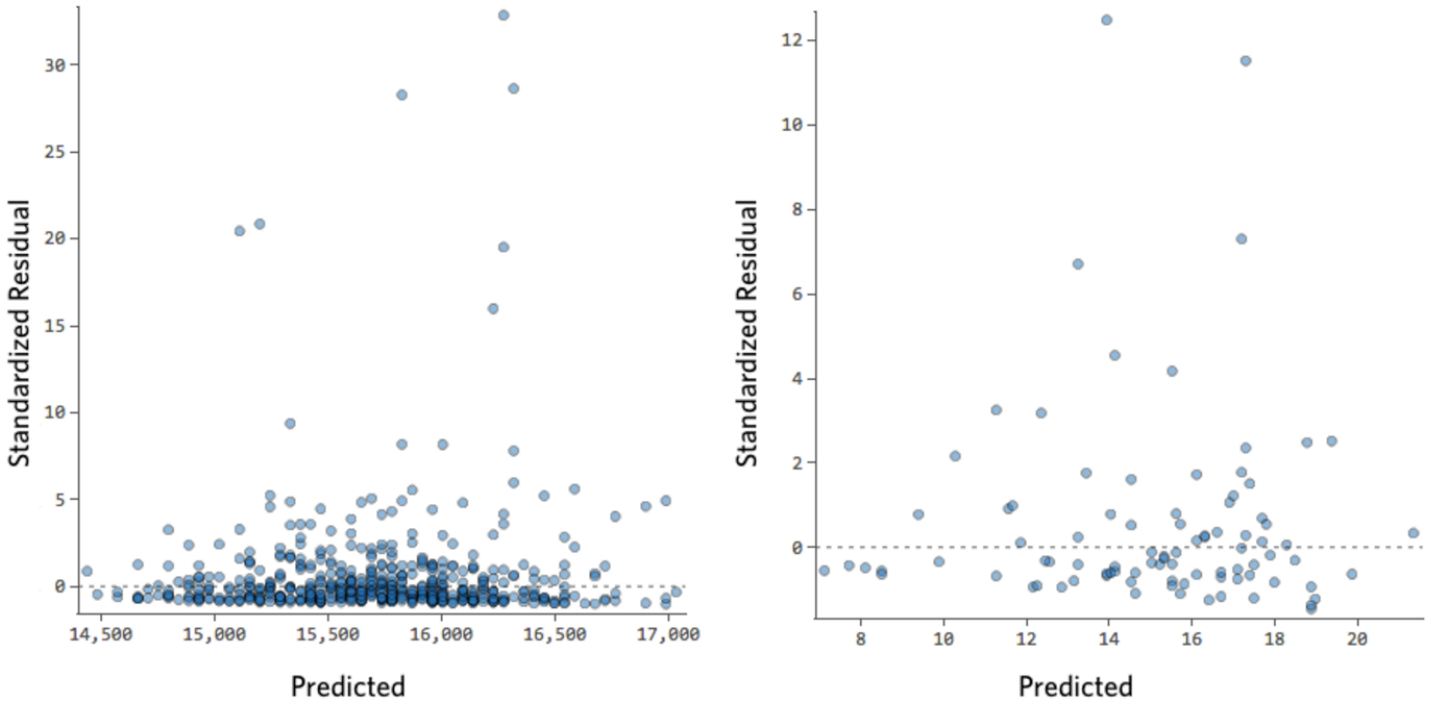


These plots aren't evenly distributed vertically, or they have an outlier, or they have a clear shape to them.

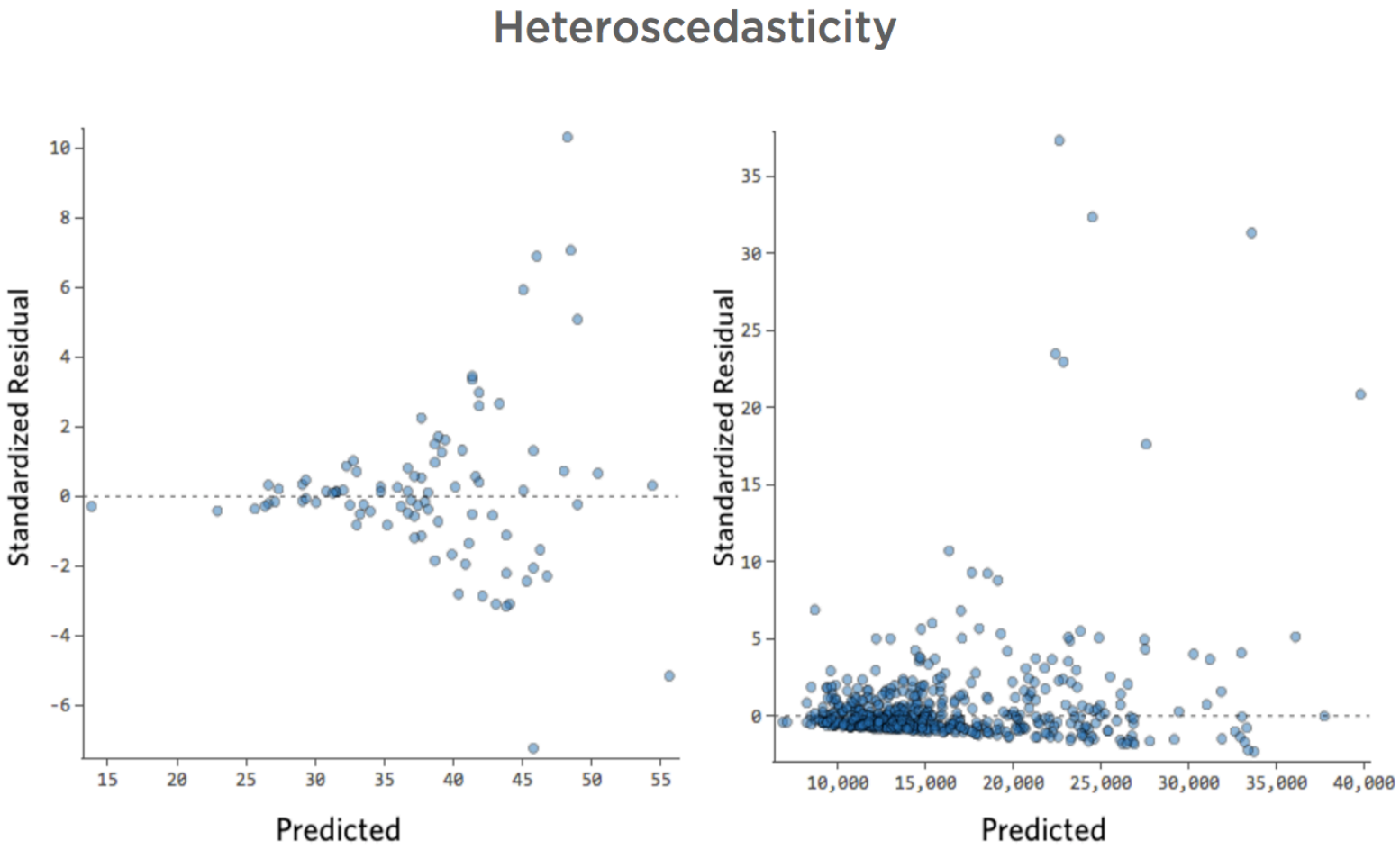
If you can detect a clear pattern or trend in your residuals, then your model has room for improvement.

4.2. Test of Homoscedasticity- Examples

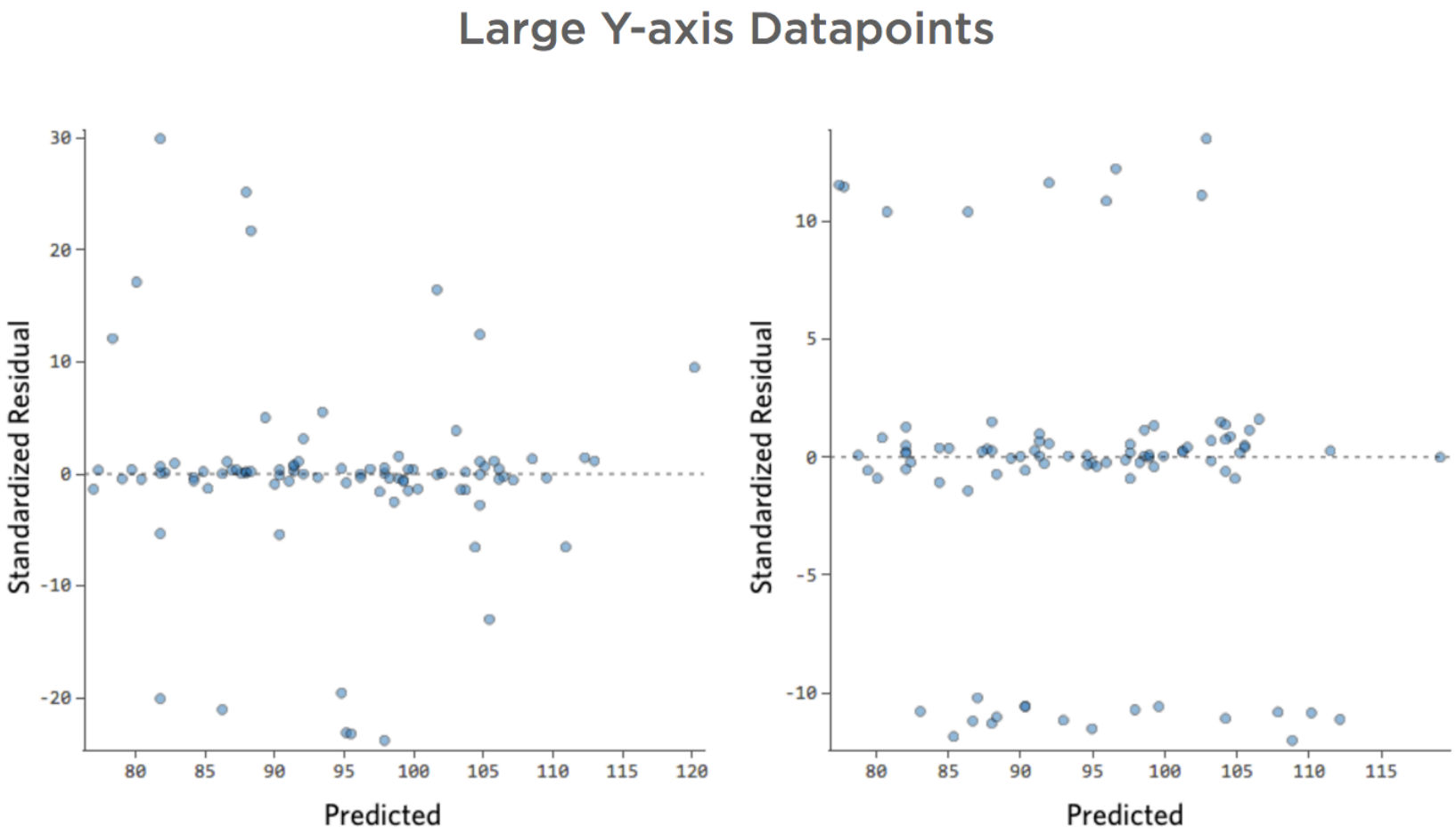
Y-axis Unbalanced



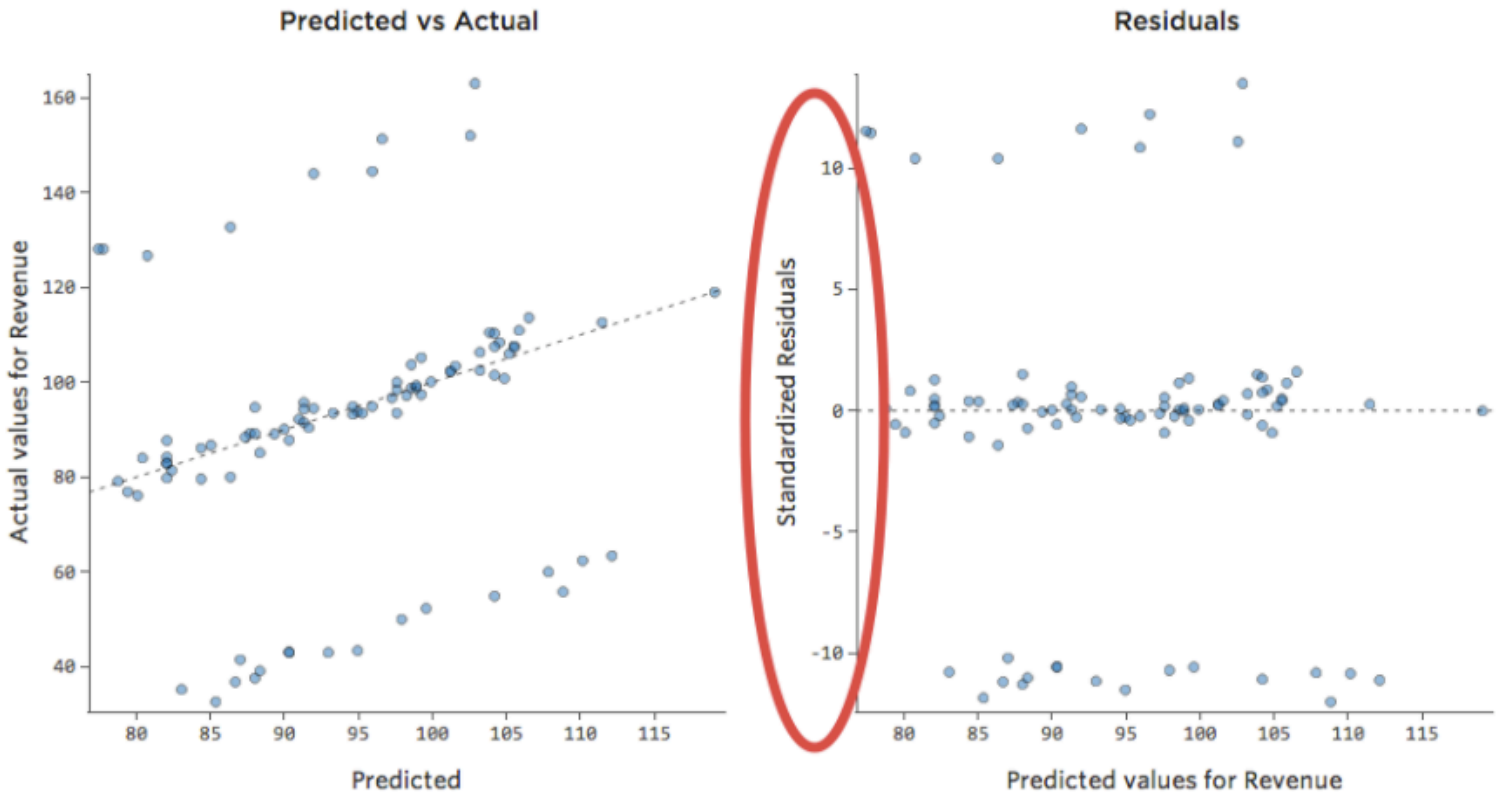
4.2. Test of Homoscedasticity- Examples



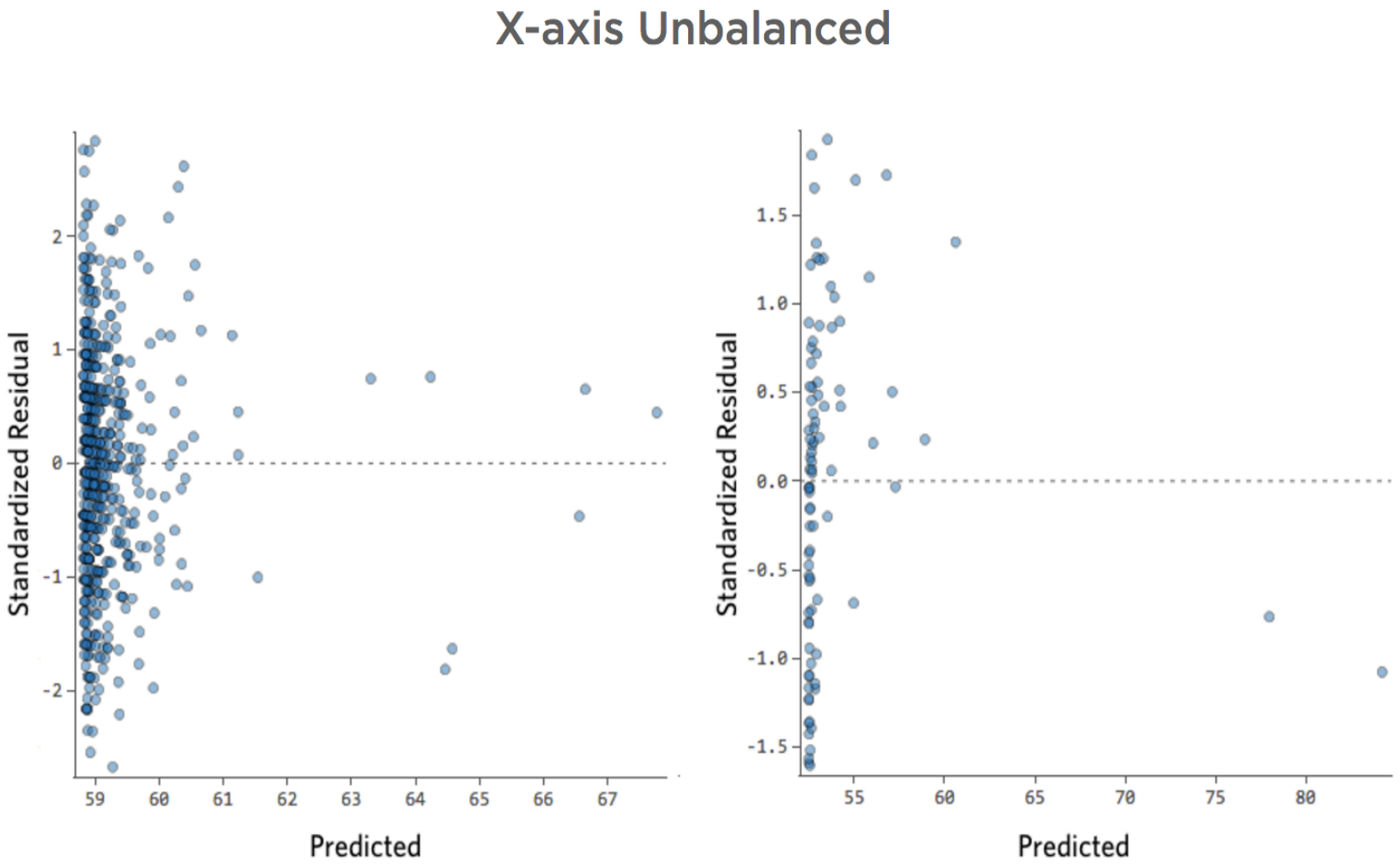
4.2. LARGE Y-AXIS DATAPOINTS- Examples



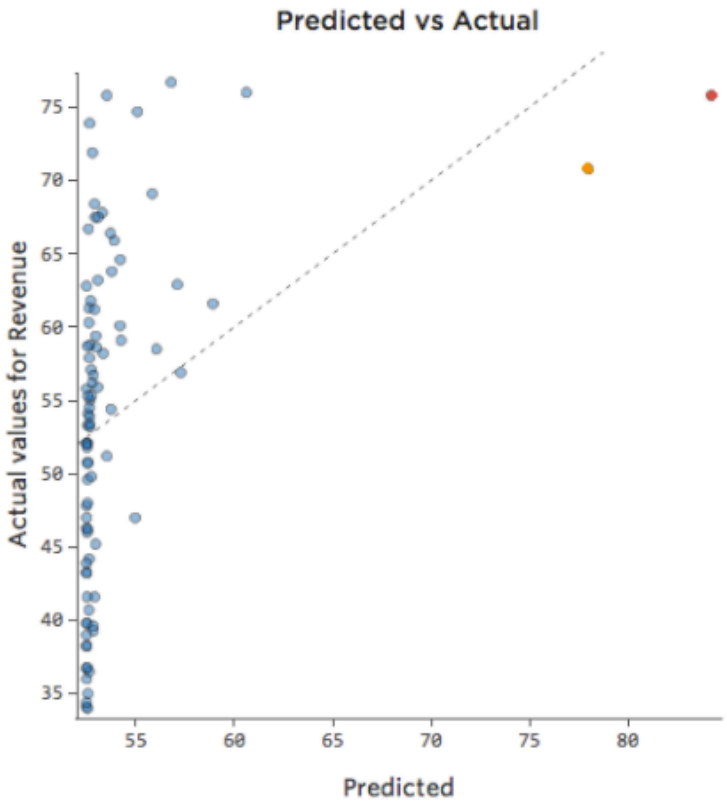
4.2. LARGE Y-AXIS DATAPOINTS- Examples

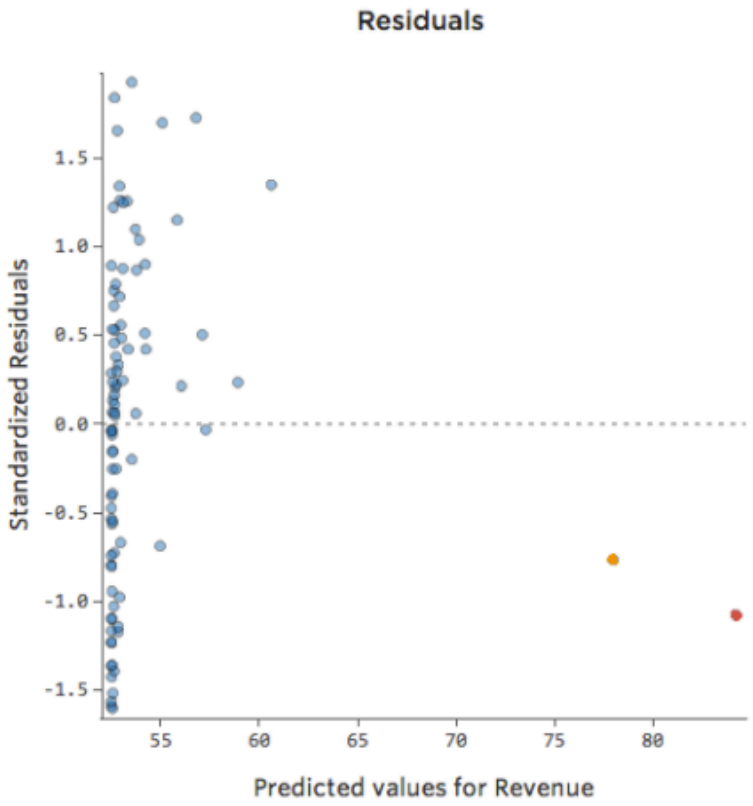


4.2. X-AXIS UNBALANCED - Examples

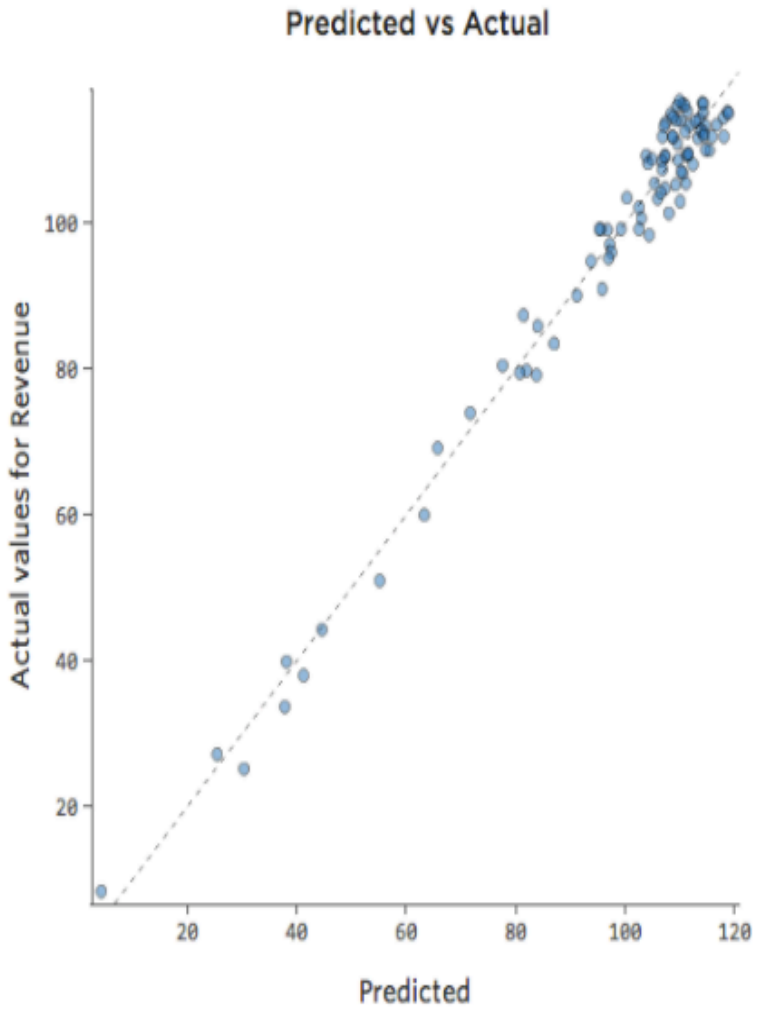


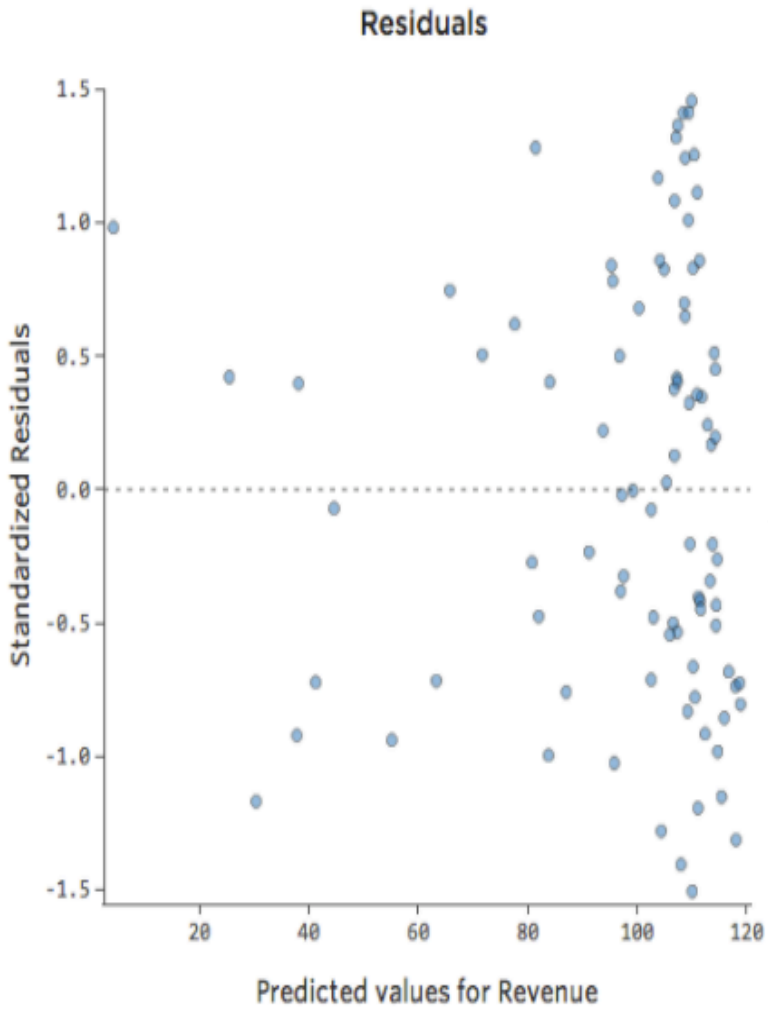
4.2. X-AXIS UNBALANCED - Examples





4.2. X-AXIS UNBALANCED - Examples





Sometimes there's actually nothing wrong with your model.
In the above example, it's quite clear that this isn't a good model,
but sometimes the residual plot is unbalanced and the model is quite good.
The only ways to tell are to a) experiment with transforming your data and see if you can improve it and b) look at the predicted vs. actual plot and see if your prediction is wildly off for a lot of datapoints, as in the above example
(but unlike the below example).

DATA ANALYTICS

4.2. Improving Your Model: Transforming Variables

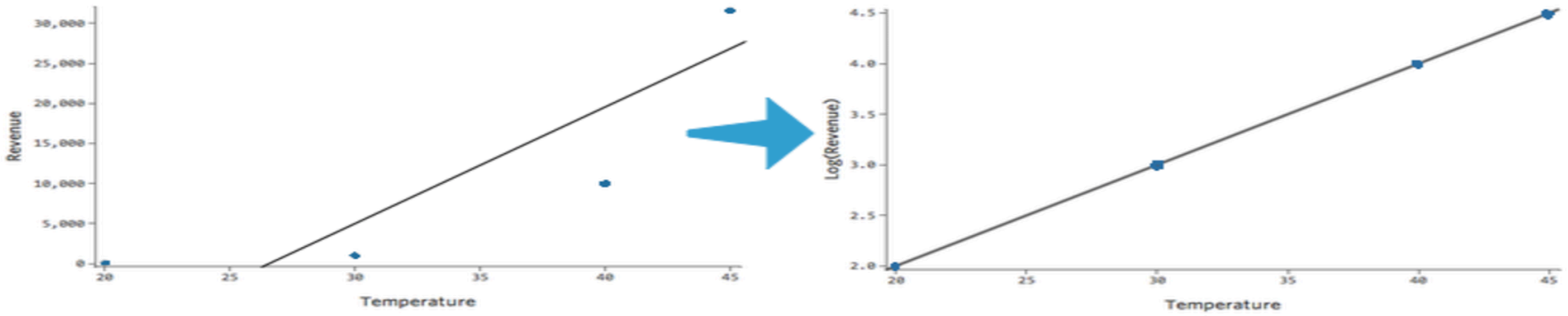
- The most common way to improve a model is to transform one or more variables, usually using a “log” transformation.
- Transforming a variable changes the shape of its distribution. Typically the best place to start is a variable that has an asymmetrical distribution, as opposed to a more symmetrical or bell-shaped distribution.
- After transforming a variable, note how its distribution, the r-squared of the regression, and the patterns of the residual plot change. If those improve (particularly the r-squared and the residuals), it’s probably best to keep the transformation.
- If a transformation is necessary, you should start by taking a “log” transformation because the results of your model will still be easy to understand. Note that you’ll run into issues if the data you’re trying to transform includes zeros or negative values, though.
- To learn why taking a log is so useful, or if you have non-positive numbers you want to transform, or if you just want to get a better understanding of what’s happening when you transform data, read on through the details below.

4.2. Improving Your Model: Transforming Variables

DETAILS

Temperature	Revenue	Log(Revenue)
20	100	2
30	1,000	3
40	10,000	4
45	31,623	4.5

Note that if we plot “Temperature” vs. “Revenue,” and “Temperature” vs. Log(“Revenue”), the latter model fits much better.



Revenue = 1,120 * Temperature + 27,132

r-squared = 0.69

Log(Revenue) = 0.1 * Temperature

r-squared = 1

4.2. Improving Your Model: Transforming Variables

DETAILS

Also note that you can't take the log of 0 or of a negative number (there is no X where $10^X = 0$ or $10^X = -5$), so if you do a log transformation, you'll lose those datapoints from the regression. There's 4 common ways of handling the situation:

1.Take a square root, or a cube root. Those won't change the shape of the curve as dramatically as taking a log, but they allow zeros to remain in the regression.

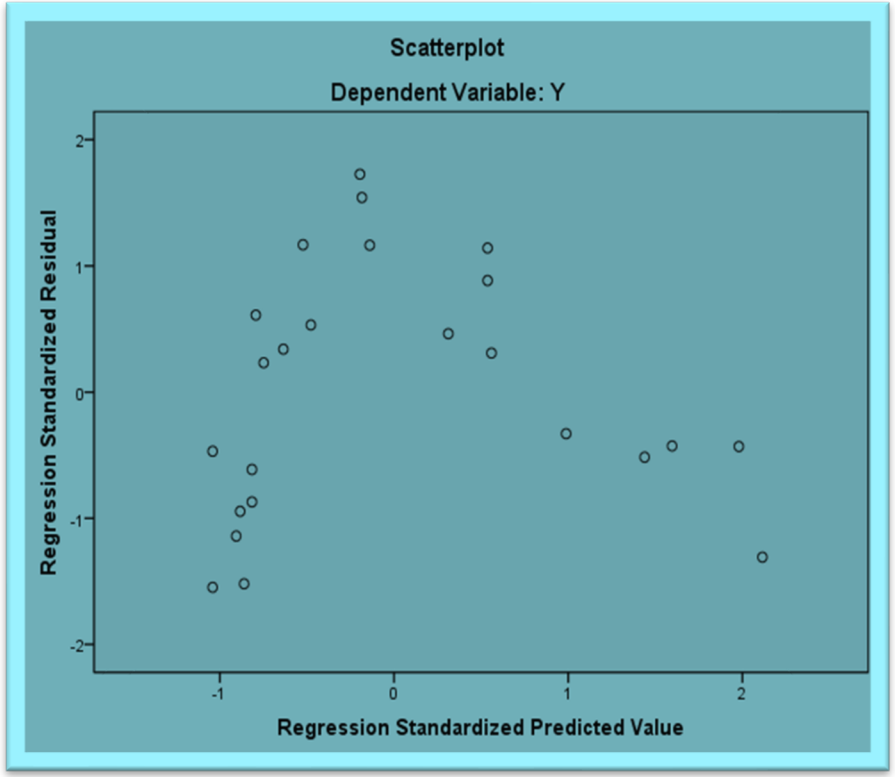
2.If it's not too many rows of data that have a zero, and those rows aren't theoretically important, you can decide to go ahead with the log and lose a few rows from your regression.

3.Instead of taking log(y), take log(y+1), such that zeros become ones and can then be kept in the regression. This biases your model a bit and is somewhat frowned upon, but in practice, its negative side effects are typically pretty minor.

4.3. Testing the Functional Form of Regression Model

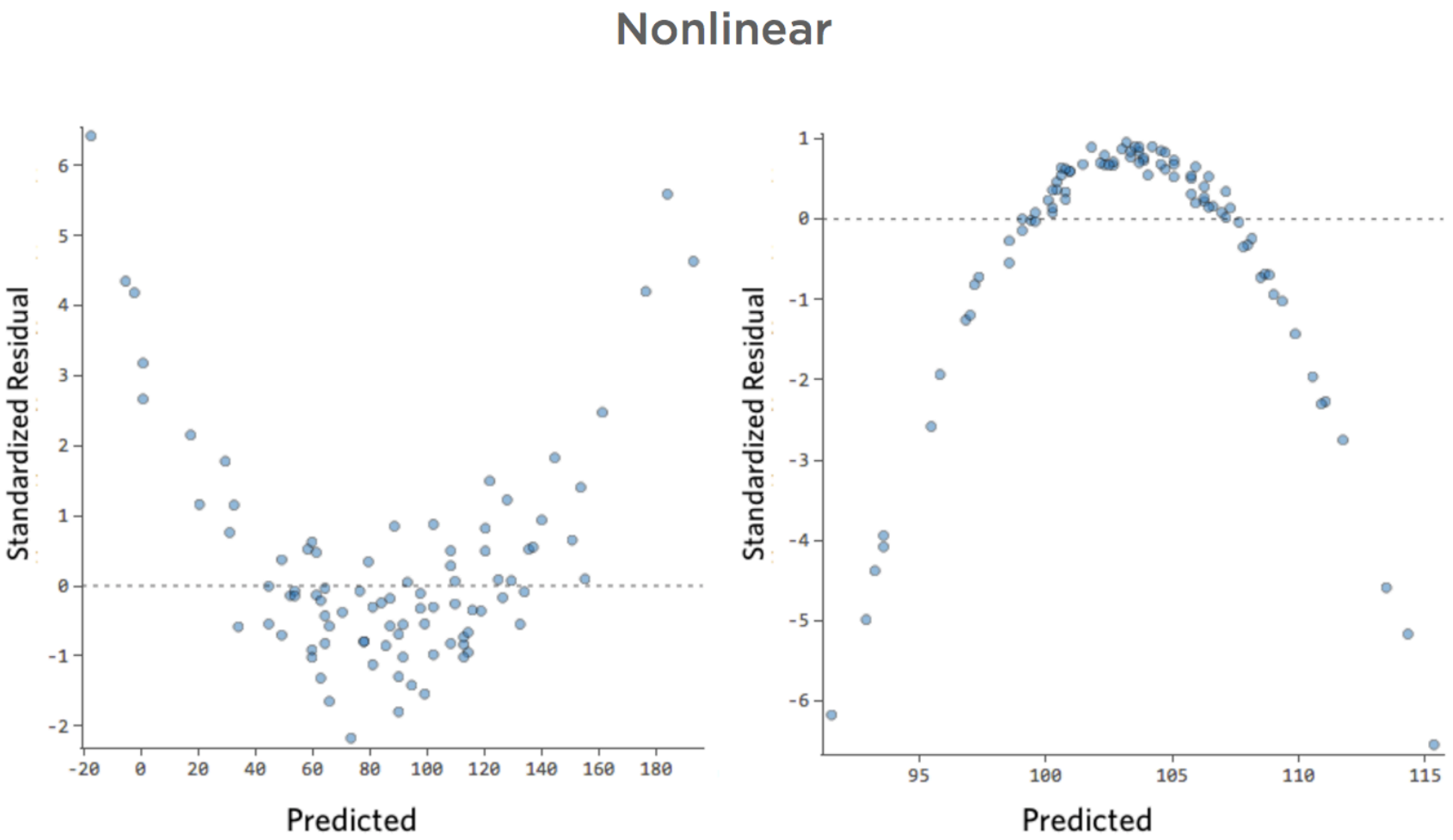
Any pattern in the residual plot would indicate incorrect specification (misspecification) of the model.

Residual plot in Figure shows a parabolic shape indicating the model mis-specification, that is, an incorrect functional form is used.

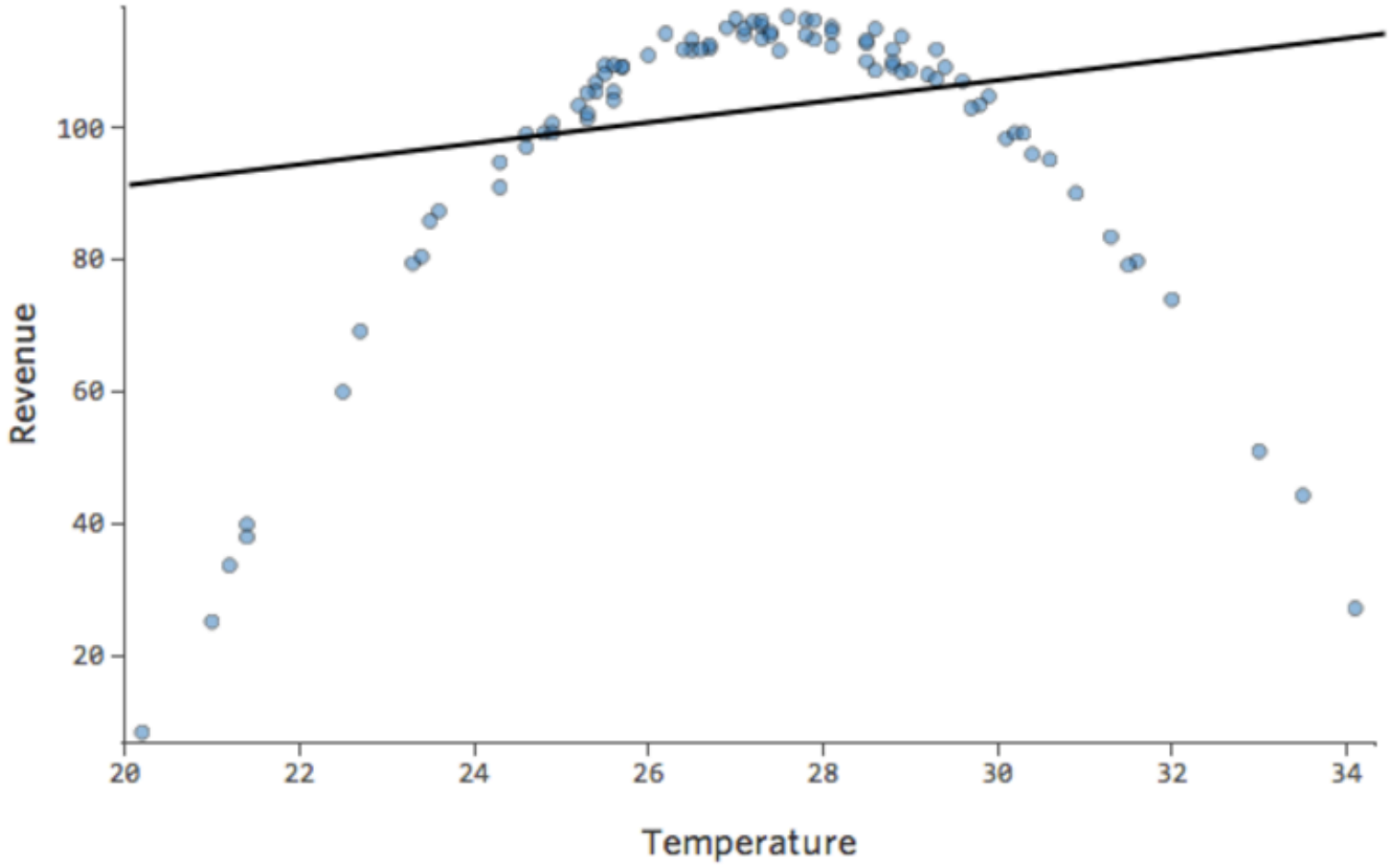


A pattern (parabola) in the residual plot indicates model misspecification

4.2. NonLinear- Examples



4.2. NonLinear- Examples - Examples



The model, represented by the line, is terrible. The predictions would be way off, meaning your model doesn't accurately represent the relationship between "Temperature" and "Revenue."

4.3. Testing the Functional Form of Regression Model

Improving Your Model: Fixing Nonlinearity

You might notice that the shape is that of a parabola, which you might recall is typically associated with formulas that look like this:

$y = x^2 + x + 1$

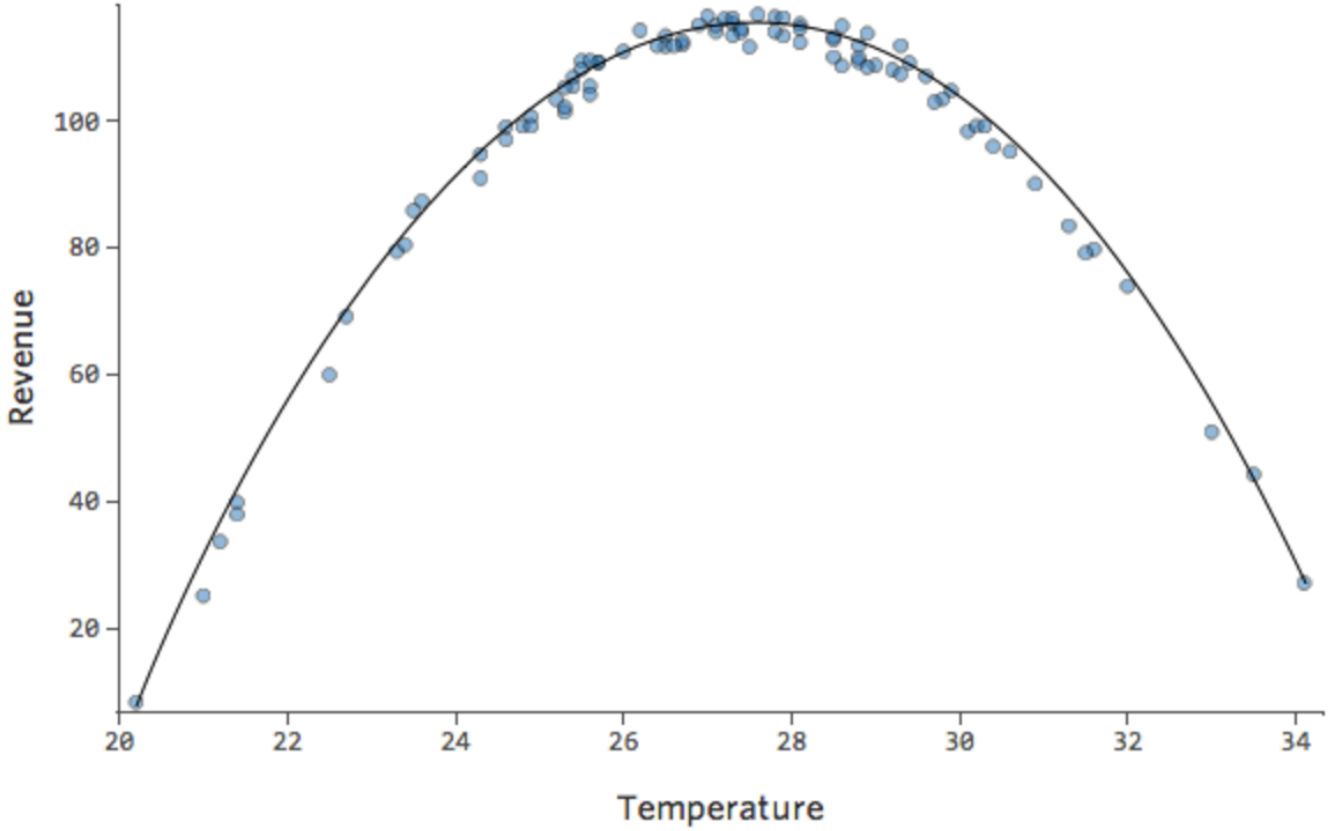
By default, regression uses a linear model that looks like this:

$y = x + 1$

In fact, the line in the plot above has this formula:

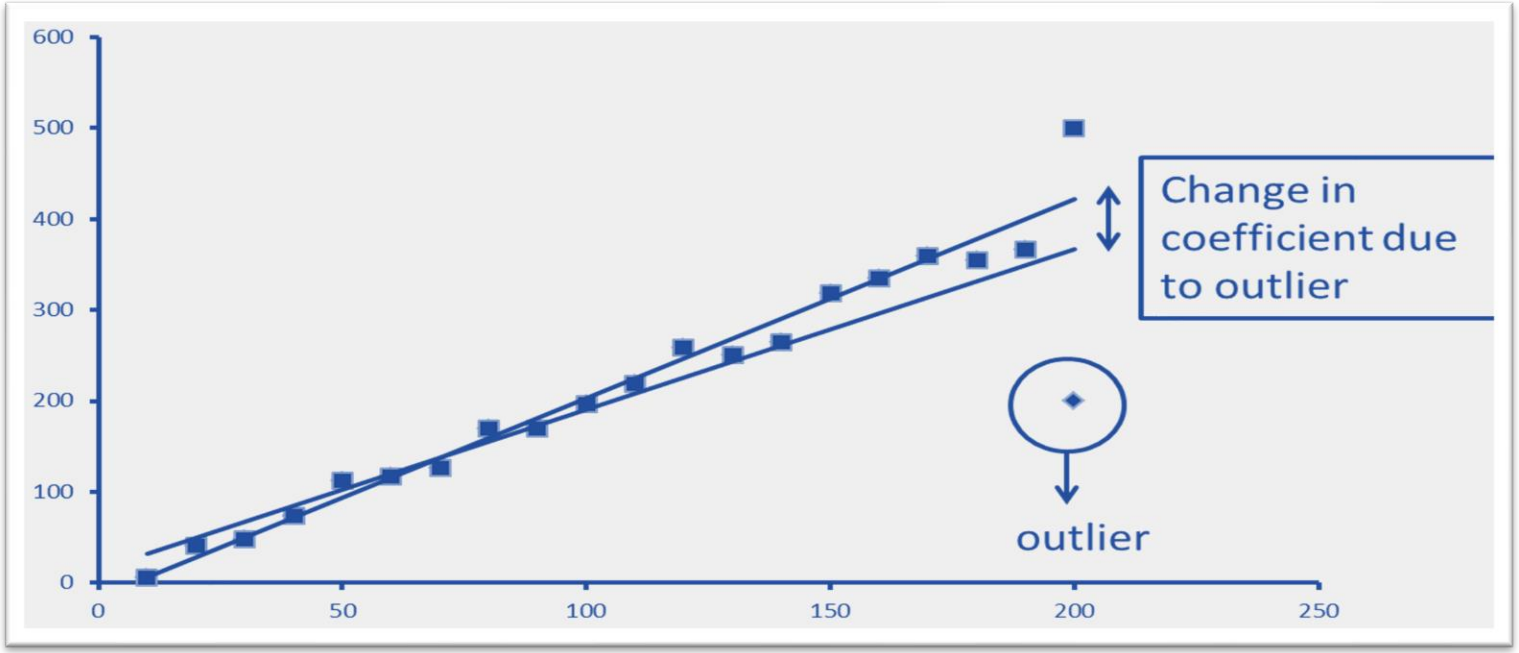
$y = 1.7x + 51$

But it's a terrible fit. So if we add an x^2 term, our model has a better chance of fitting the curve. In fact, it creates this:



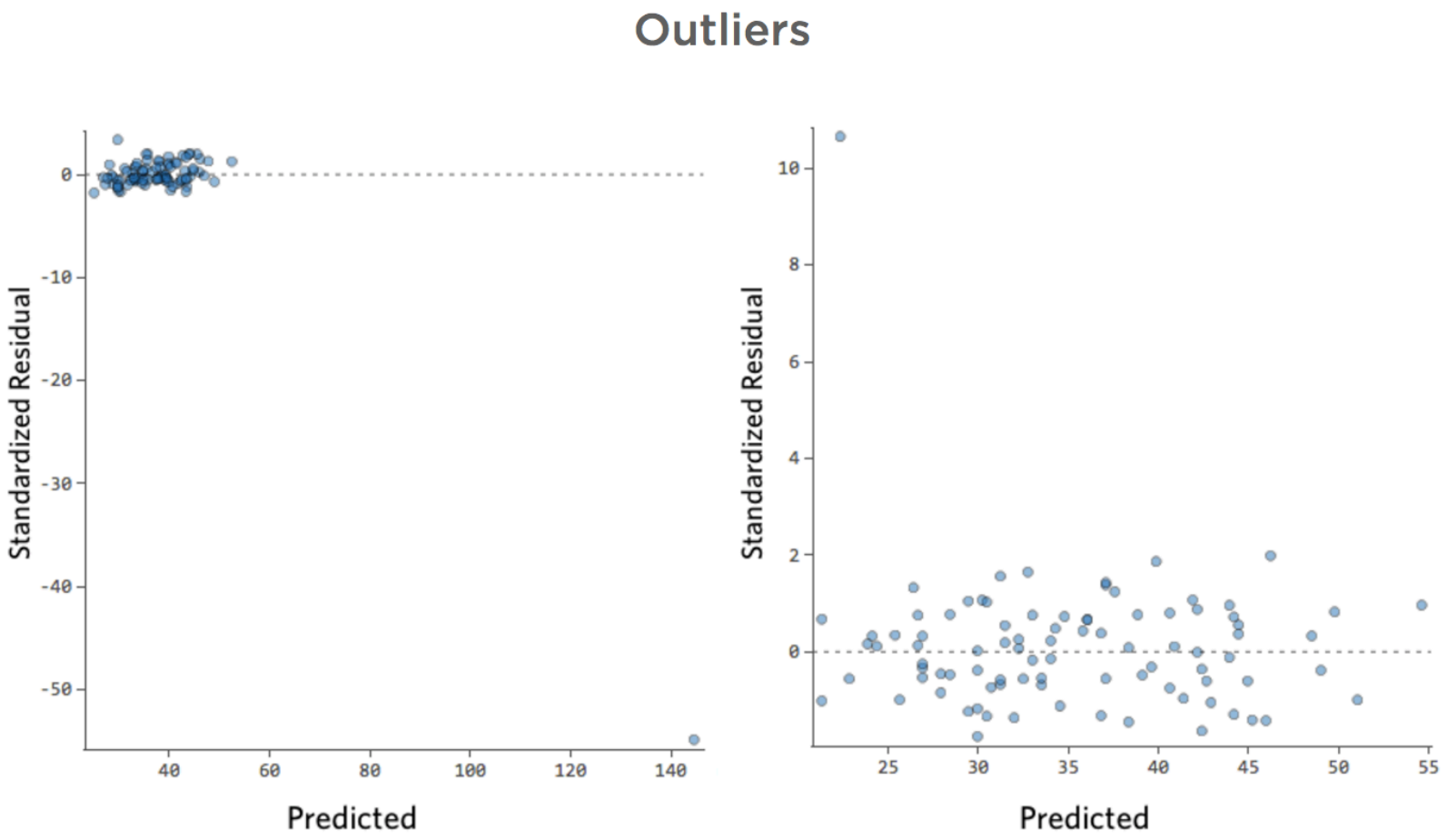
5.Outlier Analysis

- Outliers are observations whose values show a large deviation from mean value, that is ( ) large
- Presence of an outlier can have significant influence on values of regression coefficients. Thus, it is important to identify the existence of outliers in the data

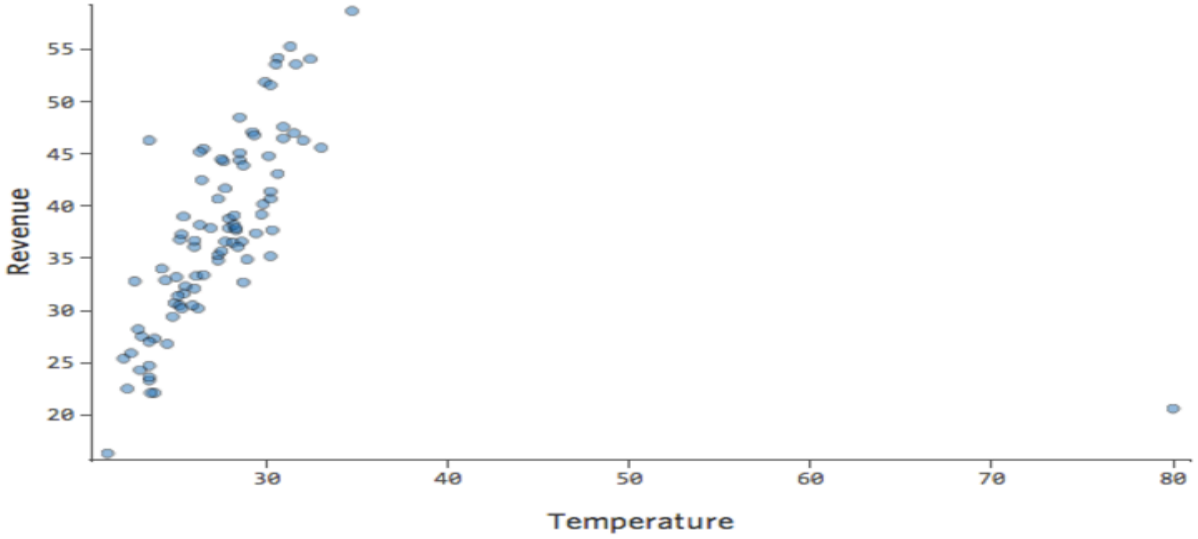


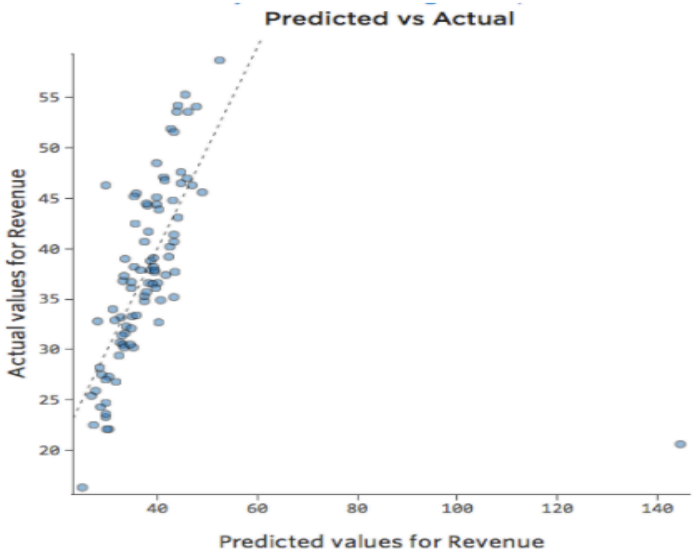
Influence of outliers on regression coefficients

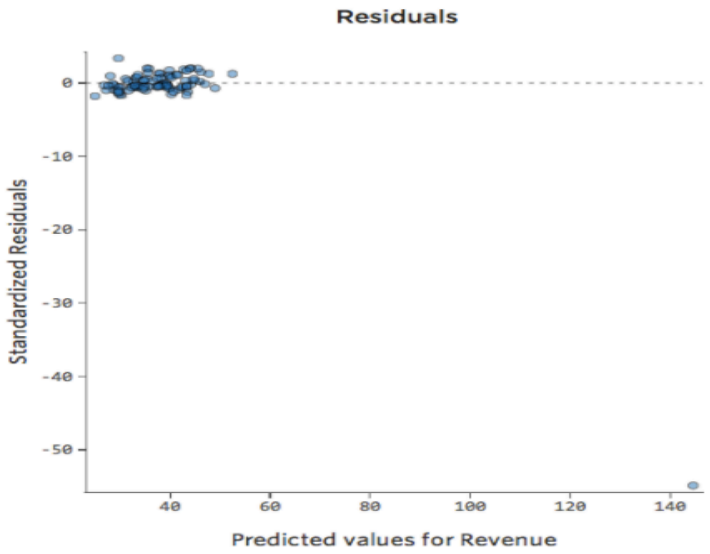
4.2. Outliers- Examples



4.2. Outliers- Examples







This regression has an outlying datapoint on an input variable

5.Outlier Analysis - Methods

The following distance measures are useful in identifying the influential observations:

1. Z-Score
2. Mahalanobis Distance
3. Cook's Distance
4. Leverage Values
5. DFBeta and DFFit values

5.1 Z-Score

Z-score is the standardized distance of an observation from its mean value. For the predicted value of the dependent variable *Y*, the Z-score is given by

$$Z = \left(\frac{\hat{Y}_i - \bar{Y}}{\sigma_Y} \right)$$

Where and are, respectively, the mean and the standard deviation of dependent variable estimated from the sample data.

5.2 Mahalanobis Distance

Mahalanobis distance is the distance between specific values of the independent variable (X_i) to the centroid of all observations of the explanatory variable. Distances value of more than chi-square critical value (with degrees of freedom is equal to the number of explanatory variables) is classified as outliers.

5.3 Cook's Distance

Cook's distance measures how much the predicted value of the dependent variable changes for all the observations in the sample when a particular observation is excluded from sample for the estimation of regression parameters. Cook's distance for simple linear regression is given by

$$D_i = \frac{\sum_j (\hat{Y}_j - \hat{Y}_{j(i)})^2}{(k+1) \times MSE}$$

where D_i is the Cook's distance measure for i^{th} observation,

$\hat{Y}_{j(i)}$ is the predicted value of j^{th} observation including i^{th} observation,

$\hat{Y}_j$ is the predicted value of j^{th} observation after excluding i^{th} observation from the sample, MSE is the Mean–Squared–Error.

5.4 Leverage Value

Leverage value of an observation measures the influence of that observation on the overall fit of the regression function. Leverage value for an observation in SLR is given by

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Leverage value of more than $2/n$ or $3/n$ is treated as highly influential observation. In Eq. the first term $(1/n)$ will tend to zero for large value of n.

Distance measures for first 10 cases in Table 9.2

TABLE 9.9 Distance measures for first 10 cases in Table 9.2

S. No.	Percentage in Grade10	Salary	Mahalanobis Distance	Cook's Distance	Leverage Value
1.0	62.00	270,000	0.03801	0.00067	0.00078
2.0	76.33	200,000	1.58353	0.05336	0.03232
3.0	72.00	240,000	0.67115	0.00659	0.01370
4.0	60.00	250,000	0.15825	0.00004	0.00323
5.0	61.00	180,000	0.08785	0.01076	0.00179
6.0	55.00	300,000	0.81887	0.01872	0.01671
7.0	70.00	260,000	0.37994	0.00083	0.00775
8.0	68.00	235,000	0.17103	0.00310	0.00349
9.0	82.80	425,000	3.66560	0.13495	0.07481
10.0	59.00	240,000.0	0.24923	0.00002	0.00509

5.5 DFFit and DFBeta

- DFFit is the change in the predicted value of Y_i when case *i* is removed from the data set.
- DFBeta is the change in the regression coefficient values when an observation *i* is removed from the data.
- Discussed in Multiple Regression

Confidence Interval for Regression coefficients β0 and β1

The standard error of estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by

$$S_e(\hat{\beta}_0) = \frac{S_e \times \sqrt{\sum_{i=1}^{n} X_i^2}}{\sqrt{n \times SS_X}}$$

$$S_e(\hat{\beta}_1) = \frac{S_e}{\sqrt{SS_X}}$$

where

$$S_e = \sqrt{\frac{\left(Y_i - \hat{Y}_i\right)^2}{n-2}}$$

Where S_e is the standard error of residuals and SSX = $\sum_{i=1}^{n}(X_i - \bar{X})^2$

The interval estimate or (1-α)100% confidence interval for $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by

$$\hat{\beta}_1 \mp t_{\alpha/2,n-2} S_e(\hat{\beta}_1)$$

$$\hat{\beta}_0 \mp t_{\alpha/2,n-2} S_e(\hat{\beta}_0)$$

Confidence Interval for Regression coefficients β0 and β1

PES UNIVERSITY ONLINE

Exercise : Salary of MBA students versus their grade 10

The confidence interval for $\hat{\beta}_0$ and $\hat{\beta}_1$ in Example 9.1 is shown in Table 9.10. The 95% confidence interval for $\hat{\beta}_0$ is (−72557.805, 195668.515) and the 95% confidence interval for $\hat{\beta}_1$ is (1002.156, 5150.199).

TABLE 9.10 Confidence interval for $\hat{\beta}_0$ and $\hat{\beta}_1$ for Example 9.1

Model		Unstandardized Coefficients		*t*	Sig.	95.0% Confidence Interval for *B*	
		B	Std. Error			Lower Bound	Upper Bound
1	(Constant)	61555.355	66701.901	.923	.361	−72557.805	195668.515
	Percentage in grade 10	3076.177	1031.526	2.982	.004	1002.156	5150.199

2) Confidence Interval for the Expected Value of Y for a Given X

- Since the point estimates are subjected to higher levels of error, due to uncertainties around estimation of parameters and natural variation in the data around the predicted line, the user would like to know the interval estimate or the **confidence interval** for the conditional expected value.
- The confidence interval of the expected value of Y_i for a given value of X_i is given by

$$\hat{Y}_i \pm t_{\alpha/2,n-2} \times S_e \times \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

- Where the term $S_e \times \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$ is the standard error of E(Y|X).

3) Prediction Interval for the Value of Y for a Given X

The prediction interval of Y_i for a given value of X_i is given by

$$\hat{Y}_i \pm t_{\alpha/2,n-2} \times S_e \times \sqrt{1+\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}$$

where the term, $S_e \times \sqrt{1+\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}$ is the standard

error of Yi for a given Xi value

Contd.,

For large *n*, the confidence interval of $E(Y/X)$ will converge to

$$\hat{Y}_i \pm t_{\alpha/2,n-2} \times S_e$$

This is because, as n $\rightarrow \infty$, the term

$$\sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

converges to 1

DATA ANALYTICS

Exercise

Reading Exercise:

1. https://www.statisticssolutions.com/assumptions-of-linear-regression/
2. https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

Homework

Hypothesis testing
&
Confidence interval
→ Range of values for population parameter

Uses. When this is used
Difference.

References

1.Text Book:

"Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017

2. "Statistics for Business and Economics"
Anderson, Sweeney, Williams, Cengage Learning.

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834