LECTURE - Regression

The Simple Linear Regression (SLR) Model - Part 1

What is Regression?

Regression analysis is a statistical technique for investigating and modeling the relationship between variables based on data. Applications of regression are numerous and occur in almost every field, including engineering, the physical and chemical sciences, economics, management, life and biological sciences, and the social sciences. In fact, regression analysis is one the most widely used statistical techniques in the world.

Example: For any real estate agent, it is important to know the relationship between "Listing Price" and "Selling Price" for any real estate specially house. Typically, after a real estate agent list the house for sell, they would like to predict what will be the selling price. Following are the data for some houses sold in Philadelphia county in recent past. The first step to establish the relationship between two variables is to plot them in X and Y axis. This is called scatter plot which helps to guess the relationship in terms of equation. Before we go ahead and guess the equation we must identify the role of these variables as there are not treated equally in regression. The variable which comes first and impacts the other is called independent variable and denoted by X. On the other hand the variable we would like to predict and is impacted by X is called dependent variable and denoted by Y. In this example, it is

clear that X = Listing price and Y = Selling price. Here we are considering the houses listed below $400,000 (Blue high-lighted data are not considered at this point.

Philadelphia County ZIP codes Date range: Jul-Oct '16			Philadelphia County ZIP codes Date range: Jul-Oct '16		
ZIP	X = Listing price	Y = Sales price	ZIP	X = Listing price	Y = Sales price
19102	10.38	6.03	19111	1.72	1.50
19103	10.15	4.10	19150	1.67	1.35
19118	9.50	4.78	19126	1.63	1.60
19106	6.82	3.86	19121	1.46	1.35
19147	5.13	3.50	19151	1.40	1.40
19107	4.98	3.01	19136	1.37	1.32
19123	4.81	3.55	19137	1.36	1.17
19130	3.77	3.23	19149	1.29	1.33
19146	3.42	3.39	19153	1.28	1.20
19125	3.11	2.38	19144	1.25	1.37
19127	2.96	2.63	19138	1.10	0.90
19119	2.72	2.65	19135	1.09	1.10
19115	2.68	2.20	19131	1.09	0.82
19116	2.67	2.15	19141	1.01	0.95
19128	2.64	2.22	19133	1.00	0.41
19122	2.53	1.80	19143	0.92	1.16
19129	2.46	2.43	19120	0.91	0.83
19148	2.34	1.86	19124	0.89	0.79
19104	2.23	1.60	19134	0.84	0.95
19145	2.22	1.68	19139	0.73	0.87
19114	2.09	1.85	19142	0.69	0.60
19154	2.01	1.86	19140	0.61	0.38
19152	1.96	1.80	19132	0.51	0.48

Figure 1: Housing Data

Here is the scatter plot of the data. Though not all the points are exactly on a straight line but the trend is an upward straight line. As a result we model the relation ship as $y = \beta_0 + \beta_1 x$ where β_0 is the intercept and β_1 is the slope.

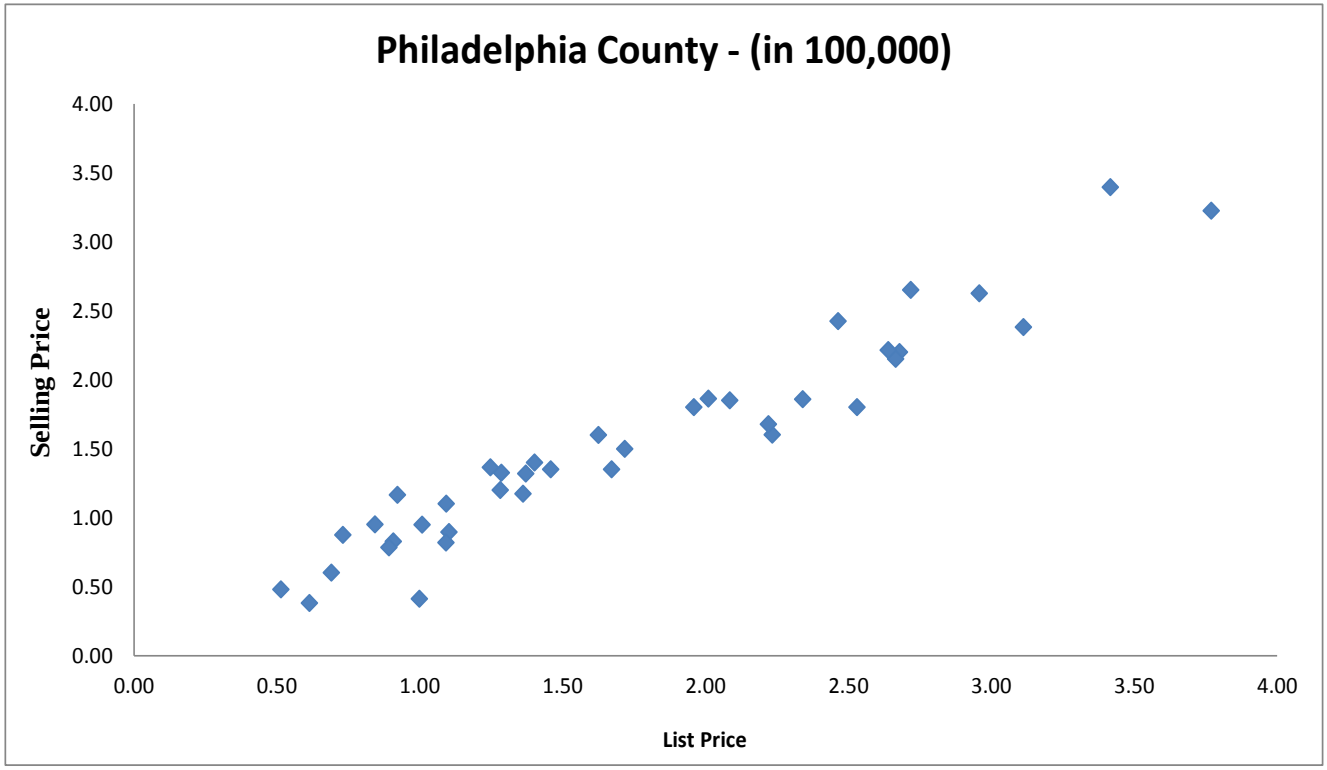

Figure 2: Housing Scatter Plot

If the true relationship between X and Y is the straight line then why they are not all on line. The answer is there is an error term involved in the relationship which we don't know. For example if 5 houses which are exactly same according to the listing price, will those be sold in the the same? The answer is, most likely not. It is very much possible that Y (=Selling Price) is impacted by some other variables which we don't have data for. Let the difference between the observed value of Y and the hypothetical straight line $(\beta_0 + \beta_1 x)$ be an error e. It is convenient to think of e as a statistical error (denoted by vertical bars in the graph). that is, it is a random variable that accounts for the failure of the model to fit the data exactly. The error may be made up of the effects of other variables on selling price, measurement errors, and so forth. Thus, a more plausible model for the Y = Selling Price is

$$Y = \beta_0 + \beta_1 X + e$$

The above equation is called Simple Linear Regression Model (SLR Model). Customarily X is called the independent variable and Y is called the dependent variable. But for many other reasons, X is also referred as the predictor or regressor variable and Y as the response variable. Because the above model involves only one regressor variable, it is called a **simple** linear regression model as oppose to **Multiple** Linear regression Model (MLR model) which involves

more than one regressors (will be considered later). The word "Linear" refers to additive nature between parameters β_0 & β_1.

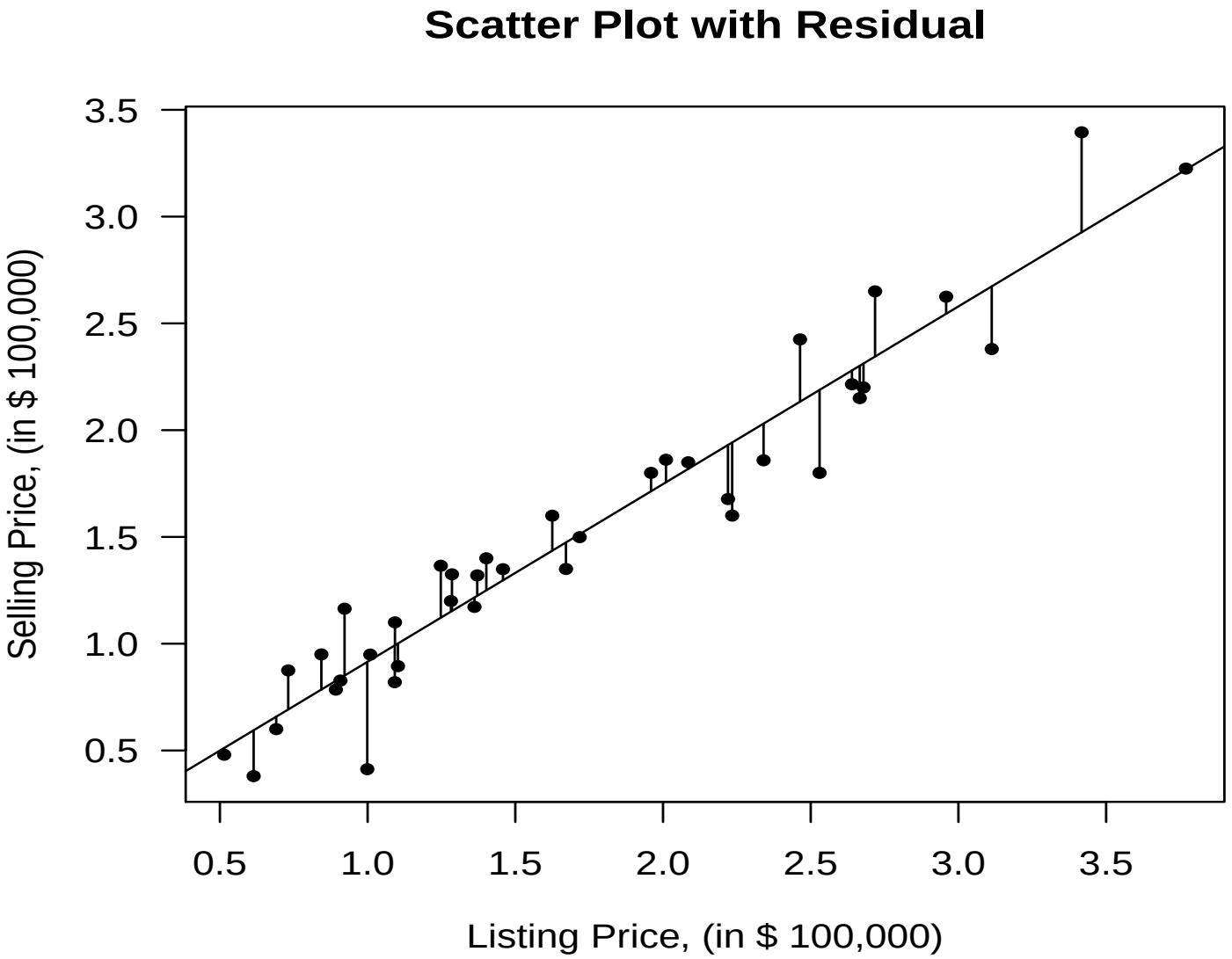

Figure 3: Housing Scatter Plot

LEAST-SQUARES ESTIMATION OF THE PARAMETERS:

Even though we assumed SLR model, but the parameters β_0 & β_1 are unknown and must be estimated using sample data. Suppose that we have n pairs of data, say $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, . Though there are many different methods are available to estimate the parameters, here we will consider the "method of least squares". The basic idea is to minimize the total error. Because the errors can be positive and negative, it does not make sense to minimize the sum of the errors as it can become as much negative as you wish. Hence we minimize the sum of "Squared" error as it is always positive. That is, we estimate β_0 & β_1 so that the sum of the squares of the differences between the observations y_i and the straight line is the minimum. Using SLR model, we may write

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad \text{for i=1,2,...,n.}$$

Note that, SLR model may be viewed as a population regression model while the above is a sample regression model, written in terms of the n pairs of data.

Now form above sample model we can write $e_i = y_i - \beta_0 - \beta_1 x_i$ which implies the sum of squared error (SSE) is

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

.

Now to minimize SSE with respect to β_0 & β_1, we simply take the derivative w.r.t both of them and set them equal to zero and then solve. Following are the two equations after we take the derivative,

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \tag{1}$$

and

$$\sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \tag{2}$$

Now solving the above two equations we get the solutions as

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i.y_i - n\,.\bar{x}\,.\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,.\bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

and

$$\hat{\beta}_0 = b_0 = \bar{y} - b_1 \times \bar{x}$$

S_{xy} & S_{xx} are called cross product and sum of squared terms respectively.

Hence, b_0 & b_1 are called the least square estimates of the population parameter β_0 & β_1. Now as we know the estimated values of the parameters, it allows us to predict any future value of Y given the value of X as

$$\hat{Y} = b_0 + b_1\,.X$$

It is called estimated regression line or least square line.

Now super imposing the the line on the scatter plot the graph looks as

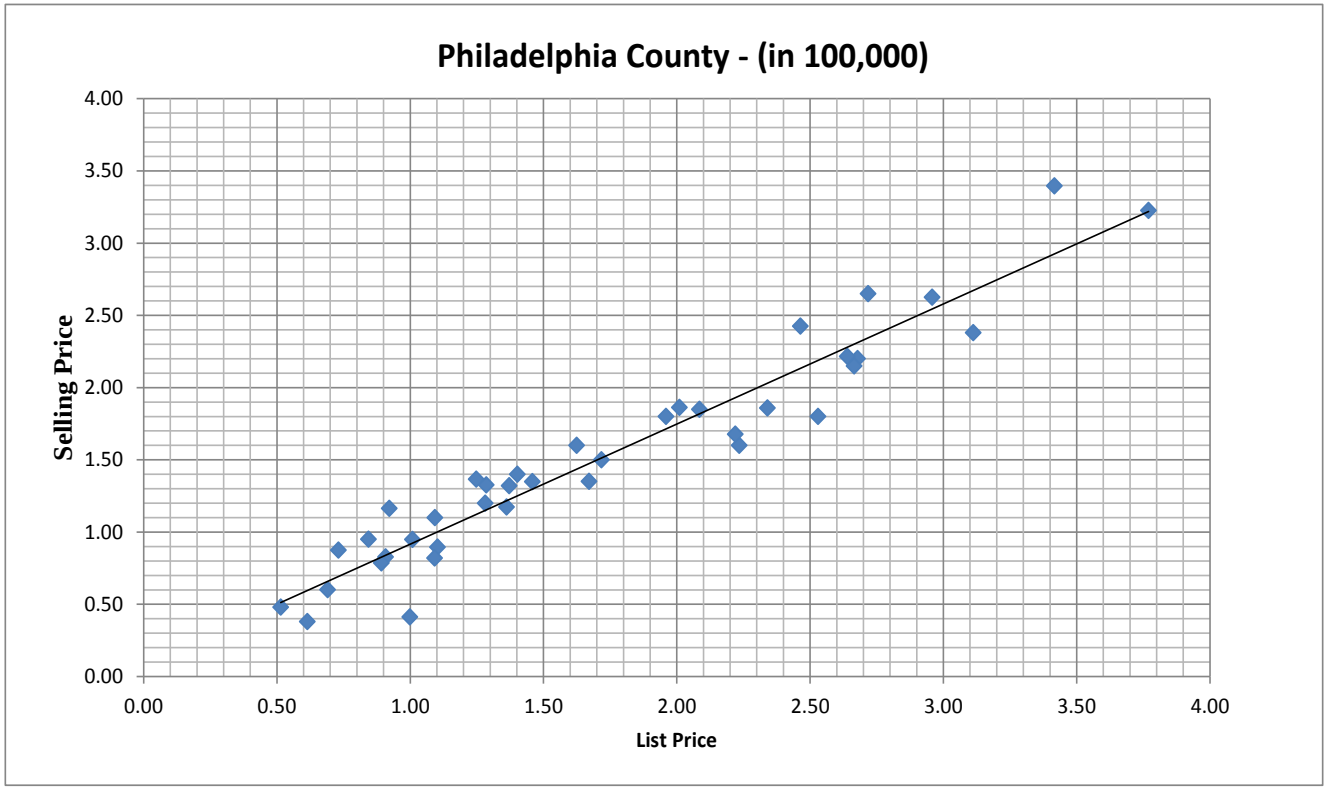

Figure 4: Housing Scatter Plot

Note that the vertical distance between the observed point and the line is the error. Hence, when point is above the line the error for that point is positive and if point is below the line then the error is negative. If we compute the error for each point and find the average, it will be zero.

The previous predicted value $\hat{Y}$ is a point prediction for Y but it does not give you any idea about how far it can be from the true value. For that we need to study the distribution of Y for a fixed value of X and it depends on the distribution of errors ($e_i's$). It is assumed that errors are independent and identically distributed as normal with mean 0 and variance σ^2. Hence

Assumption: $e_i \sim N(0, \sigma^2)$ for n = 1,2 ,...n and they are all independent

It implies

$Y_i \sim N(\beta_0 + \beta_1 .X_i, \sigma^2)$ for n = 1,2 ,...n and they are all independent

So to get any idea about the variation of Y we need to estimate σ^2 and it is done using the estimated sum of square. The estimator for σ^2 is

$$S^2_{y.x} = \hat{\sigma^2} = \frac{SSE}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum Y_i^2 - b_1 \sum X_i Y_i - b_o \sum Y_i}{n-2}$$

As a result the standard deviation is estimated by

$$\hat{\sigma} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum Y_i^2 - b_1 \sum X_i Y_i - b_o \sum Y_i}{n-2}}$$

The estimated standard deviation is also called standard error of estimate. Using the above estimate we can obtain a 100(1-α)% Prediction Interval for "One" future Y for given value of $X = x_o$ as

$$(b_o + b_1 x_o) \quad \pm \quad t_{\alpha/2,n-2}.S_{y.x}.\sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{X})^2}{(\sum X_i^2 - n\bar{X}^2)}}$$

Note that one future value of Y for any given value of X is a random variable but the true mean of all those Y's for fixed $X = X_0$ is a parameter namely $\mu_{Y|X} = \beta_0 + \beta_1 .X_0$. It is also of interest many times. The point estimate of $\mu_{Y|X}$ is $\hat{\mu}_{Y|X} = b_0 + b_1 .X_0$. and 100(1-$\alpha$)% Confidence Interval for Mean of Y for $X = x_o (i.e \mu_{Y|X})$

$$(b_o + b_1 x_o) \quad \pm \quad t_{\alpha/2,n-2}.S_{y.x}.\sqrt{\frac{1}{n} + \frac{(x_o - \bar{X})^2}{(\sum X_i^2 - n\bar{X}^2)}}$$

Computation:

To start the computation for all the estimates we need to know five quantities which are called sufficient statistic. Those are $\sum x,\ \sum y,\ \sum x^2,\ \sum y^2,\ \sum xy$.

For the previous example, those are as follows:

$\sum x = 67.64,\ \sum y = 59.53,\ \sum x^2 = 144.58,\ \sum y^2 = 111.34,$

$\sum xy = 125.94\ \&\ n = 39$

So, $\bar{x} = 1.73 \quad \text{and} \quad \bar{y} = 1.53$

Now following the formulas for the least square estimates of β's, we get

$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{SS_{xy}}{SS_{xx}} = \frac{125.94 - 39 \times 1.73 \times 1.53}{144.58 - 39 \times 1.73^2} = \frac{22.69}{27.28} = 0.8319 \quad and$

$b_o = \bar{Y} - b_1\bar{X} = 1.53 - 0.8319 \times 1.73 = 0.0836$

Hence the least-Square Line is $\hat{Y} = 0.0836 + 0.8319\ .X$

The estimated value of $\beta_1 = 0.8319$ is interpreted as the estimated change in Y for unit increment in X. In this example, it means that if a house is listed \$100,000 more then its sailing price is estimated to go up by \$83,190.

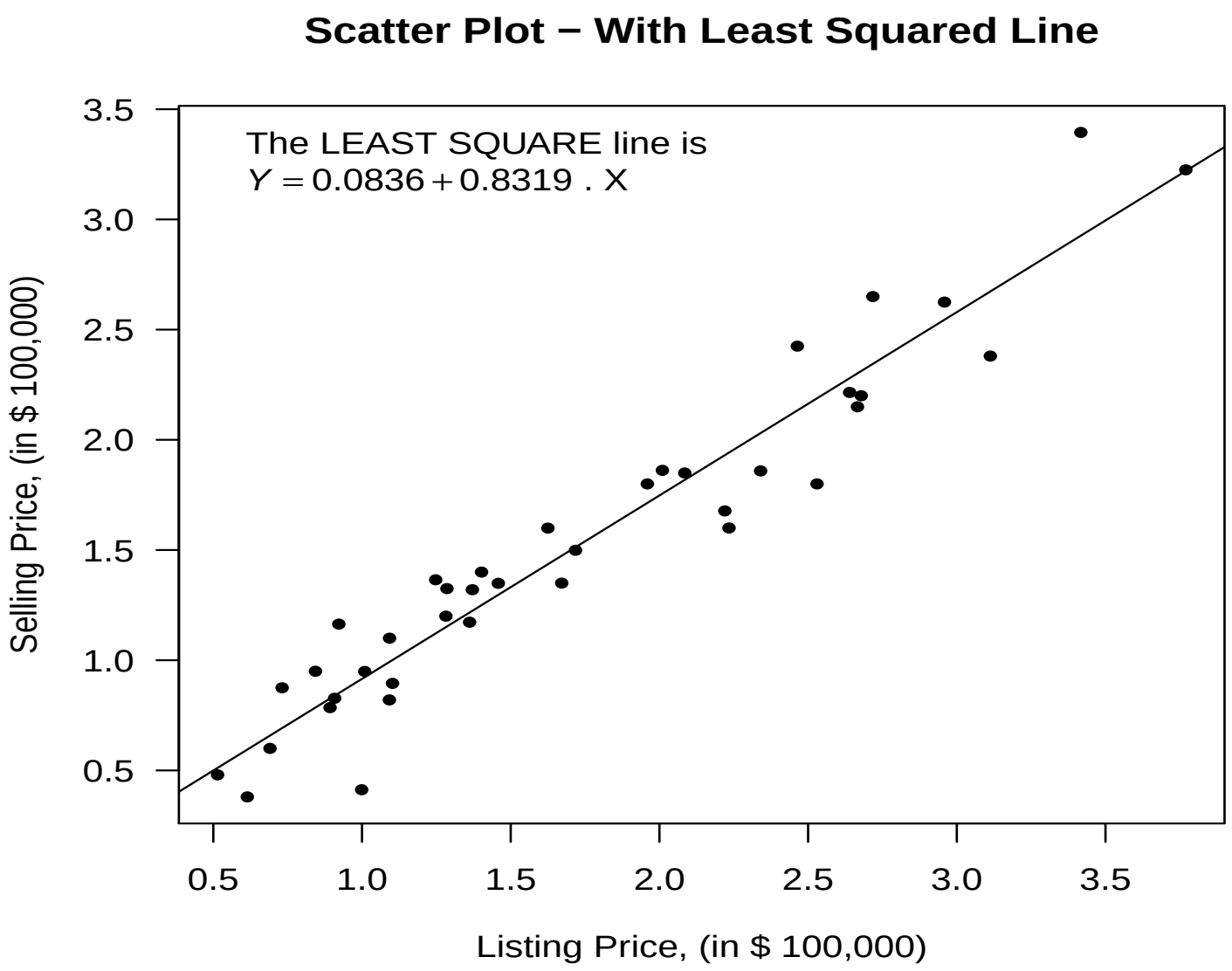

Figure 5: Housing Data - Least Square Line

On the other hand the value of $\hat{\beta}_0 = 0.0836$ should be interpreted as the expected value of Y when given value of X is 0. That implies, if a house is listed for \$0 then it is expected to sell for \$8,360 which is useless. The problem here is that value of X=0 is outside our data values of X, as result this regression line as no information how Y behaves when X is around 0. So it is to understand that predicting for a value of Y when X is outside the data range may be very dangerous.

Now to estimate the standard deviation of Y's (i.e σ) we use the formula based on SSE as

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum Y_i^2 - b_1 \sum X_i Y_i - b_o \sum Y_i}{n-2}}$$
$$= \sqrt{\frac{111.34 - 0.0836 \times 59.53 - 0.8319 \times 125.94}{39-2}} = 0.2078.$$

Suppose we want to find a 95% prediction interval for the selling price of a house which is listed for \$210,000. Because the unit is \$100,000, it means $x_0 = 2.1$. Hence the 95% confidence interval is

$$\begin{aligned}
(b_o + b_1 x_o) \quad &\pm \quad t_{\alpha/2,n-2}.S_{y.x}.\sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{X})^2}{(\sum X_i^2 - n\bar{X}^2)}} \\
= (0.0836 + 0.8319 \times 2.1) \quad &\pm \quad 2.026 \times 0.2078.\sqrt{1 + \frac{1}{39} + \frac{(2.1 - 1.73)^2}{27.28}} \\
= 1.8306 \quad &\pm \quad 2.026 \times 0.2078. \times 1.0152 \\
= 1.8306 \quad &\pm \quad 0.4273 = (1.4033, 2.2580)
\end{aligned}$$

On the other had if 95% confidence interval is needed for the same listing price then we get

$$(b_o + b_1 x_o) \quad \pm \quad t_{\alpha/2,n-2}.S_{y.x}.\sqrt{\frac{1}{n} + \frac{(x_o - \bar{X})^2}{(\sum X_i^2 - n\bar{X}^2)}}$$

$$= (0.0836 + 0.8319 \times 2.1) \quad \pm \quad 2.026 \times 0.2078.\sqrt{\frac{1}{39} + \frac{(2.1 - 1.73)^2}{27.28}}$$

$$= 1.8306 \quad \pm \quad 2.026 \times 0.2078. \times 0.1748$$

$$= 1.8306 \quad \pm \quad 0.0736 = (1.7571, 1.9042)$$

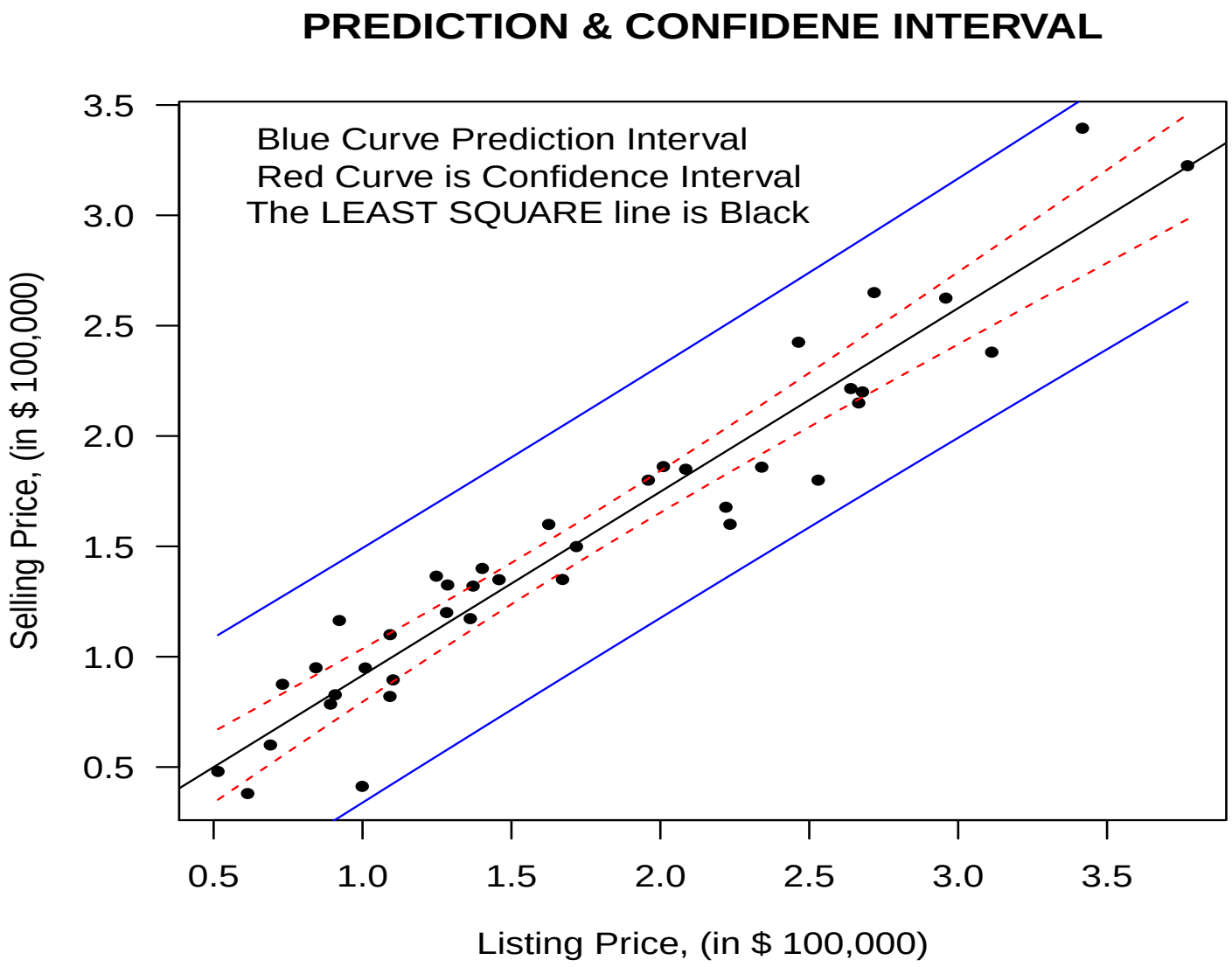

Figure 6: Housing Scatter Plot

Properties of Least-Square Estimators

As we noted before that estimators of intercept and slope (i.e $\hat{\beta}_0$ & $\hat{\beta}_0$) are linear combination (weighted) of y_i's.

For example,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i \quad \text{where } c_i = \frac{(x_i - \bar{x})}{S_{xx}}.$$

Property - 1: Assuming the model is correct, above estimators are unbiased. (i.e $\mathrm{E}(\hat{\beta}_i) = \beta_i,\ i = 0, 1.$).

$$E(\hat{\beta}_1) = E(\sum_{i=1}^{n} c_i y_i) = \sum_{i=1}^{n} c_i E(y_i) = \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) = \beta_0 . \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

Since $E(e_i) = 0$ by assumption. Also note that $\sum_{i=1}^{n} c_i = 0$ and $\sum_{i=1}^{n} c_i\, x_i = 1$ (Prove yourself).

Using that we an see $E(\hat{\beta}_1) = \beta_1$.

Similarly it can shown that $E(\hat{\beta}_0) = \beta_0$.

Property - 2 $\quad Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \quad$ and $\quad Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$

$$Var(\hat{\beta}_1) = Var(\sum_{i=1}^{n} c_i y_i) = \sum_{i=1}^{n} c_i^2 Var(y_i) = \sum_{i=1}^{n} c_i^2 \, \sigma^2$$

Note: Above line is possible because y_i's are uncorrelated and variance of y_i's are all σ^2. Else covariance terms needs to be included.

Now note that

$$\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{S_{xx}^2} = \frac{1}{S_{xx}} \quad \Rightarrow \quad Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

Note: Try to prove it for $\hat{\beta}_0$. (or Look at the book, page-19).

Property - 3: Sum of the estimated residuals is 0. i.e $\sum_1^n \hat{e}_i = 0$

Observe that

$$\sum_{1}^{n} \hat{e}_i = \sum_{1}^{n} (y_i - \hat{y}_i) = \sum_{1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 \, x_i) = n\bar{y} - n\hat{\beta}_0 - \hat{\beta}_1 \, n\bar{x} = 0$$

Note that the following facts are used in the previous line,

$$\sum_1^n y_i = n\,\bar{y},\ \sum_1^n x_i = n\,\bar{x}, \text{and}\ \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Property 4: $\sum_1^n \hat{y}_i = \sum_1^n y_i$

Observe that

$$\sum_1^n \hat{e}_i = 0 \quad \Rightarrow \quad \sum_1^n (y_i - \hat{y}_i) = 0 \quad \Rightarrow \quad \sum_1^n \hat{y}_i = \sum_1^n y_i$$

Property 5: Least square line passes through the $(\bar{x}, \bar{y})$ points.

It follows from the fact that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \quad \Rightarrow \quad \bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}$$

Property 6: The weighted sum of the residuals equals zero when weights are the corresponding value of the regressor variable. i.e. $\sum_1^n x_i\, e_i = 0$

Property 7: The weighted sum of the residuals equals zero when weights are the fitted y value. i.e. $\sum_1^n \hat{y}_i\, e_i = 0$