Simple Regression

Tom Bruning

2018-03-14

Simple Regression

In this chapter, you learn:

- How to use regression analysis to predict the value of a dependent variable based on a value of an independent variable.
- The meaning of the regression coefficients b0 and b1
- How to evaluate the assumptions of regression analysis and know what to do if the assumptions are violated.
- To make inferences about the slope and correlation coefficient.
- To estimate mean values and predict individual values.

Correlation vs. Regression

- A scatter plot can be used to show the relationship between two variables.
- Correlation analysis is used to measure the strength of the association (linear relationship) between two variables.
  - Correlation is only concerned with strength of the relationship.
  - No causal effect is implied with correlation.
  - Scatter plots were first presented in Ch. 2
  - Correlation was first presented in Ch. 3

  There are four types of relationships between two variables:

- Strong - As X goes up (or down) Y goes up (or down) and the association is closely linked.
- Weak - As X goes up (or down) Y goes up (or down) and the association is more varied.
- Linear - One variable has the predominant effect on another
- Curvilinear - More than one variable has the predominant effect on another

Introduction to Regression Analysis

Regression analysis is used to:

- Predict the value of a dependent variable based on the value of at least one independent variable.
- Explain the impact of changes in an independent variable on the dependent variable

Two Variable Types:

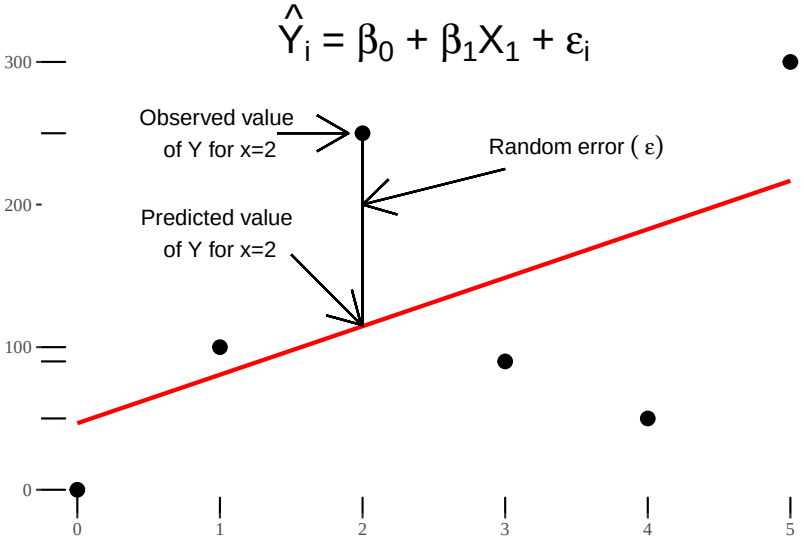
- **Dependent variable:** the variable we wish to predict or explain
- **Independent variable:** the variable used to predict or explain the dependent variable

Simple Linear Regression Model

- Only one independent variable, X
- Relationship between X and Y is described by a linear function
- Changes in Y are *assumed* to be related to changes in X
- The basic model;

$$Y_i = B_o + B_1 X_i + e_i \tag{1}$$

Where:
Y_i is the Dependent variable
B_o is the Population Y intercept
B_1 is the Population Slope coefficient
X_i is the Independent variable
e_i is the random error term
$B_o + B_1 X_i$ is the linear component e_i is the random error component



Why Graphing the Data is so important

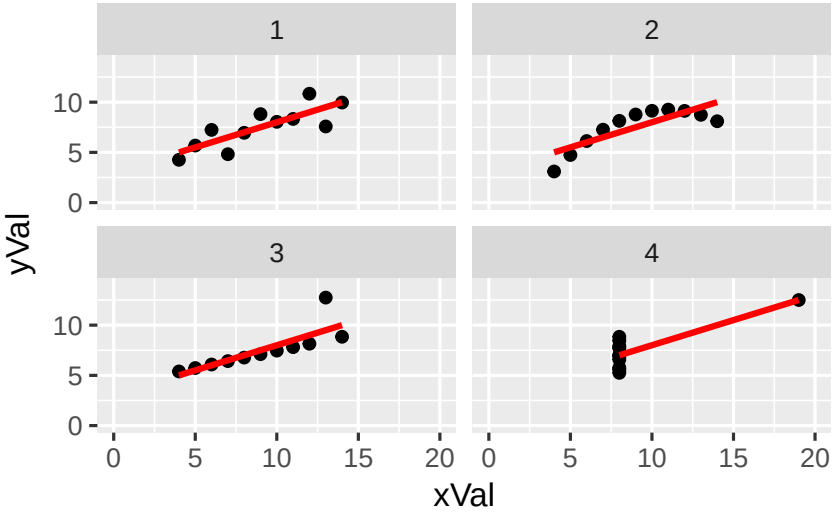
Suppose you have 4 data sets (DS1-DS4), with paired values (X and Y) like so:

DS1_x	DS1_y	DS2_x	DS2_y	DS3_x	DS3_y	DS4_x	DS4_y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

The following table, using the data from the dataset above, produces the summary data, displaying the variables that are used to compute the linear regression model for the four datasets. Because this data is identical, to 2 decimal places, the regression equation is identical for each dataset.

Dataset	Mean_x	Mean_y	Var_x	Var_y	Cor
1	9	7.5	11	4.13	0.82
2	9	7.5	11	4.13	0.82
3	9	7.5	11	4.12	0.82
4	9	7.5	11	4.12	0.82

The following scatterplots show the four datasets, and the regression line for each dataset. As you can see, the data doesn't look at all similar, even though the regression is identical. The formula that produces this regression line is $\hat{Y} = 3.00 + .50X$.



Simple Linear Regression Equation (Prediction Line)

The simple linear regression equation provides an estimate of the population regression line

$$\hat{Y} = b_0 + b_1 X_i \tag{2}$$

Where:
$\hat{Y}$ is the estimated (or predicted) Y value for observation i
b_0 is the estimate of the regression intercept
b_1 is the estimate of the regression slope
X_i is the value of X for observation i

The Least Squares Method

b_0 and b_1 are obtained by finding the values of that minimize the sum of the squared differences between Y and $\hat{Y}$:

$$min \sum (y_i - \hat{y})^2 = min \sum (y_i - (b_o + b_1 x_i))^2 \tag{3}$$

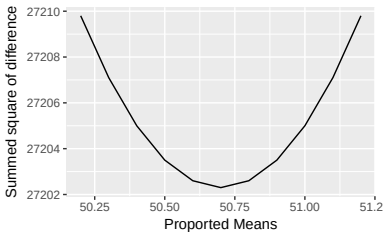
The least squares method is analogous to the mean of the residuals. For the mean of a data set, the difference between the sum of all elements in the dataset and any other value (OV) (not in the dataset) squared is minimized when the other value (OV) is the mean of the values in the dataset.

$$\mu = \sum_{i-1}^{N} = Min \sum_{1}^{n} (x_i - OV)^2 \tag{4}$$

Example

I have created a dataset with 30 values, randomly chosen from a range from 0-100.“

The mean of these 30 values is 50.7. I created a second dataset whose values ranged from 50.2 to 51.2 whose midpoint is the mean of the random data. I summed the squared difference of each value in the dataset by a value between 50.2 and 51.2 and plotted these sums. The minimum value of the Y-axis in the following plot is the where the x-value is the mean of the 30 item dataset.



Interpretation of the Slope and the Intercept

- b_0 is the estimated average value of Y when the value of X is zero
- b_1 is the estimated change in the average value of Y as a result of a one-unit increase in X

Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet).
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

Table 3: Housing Data

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Scatterplot of the Data

Simple Linear Regression Example: Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

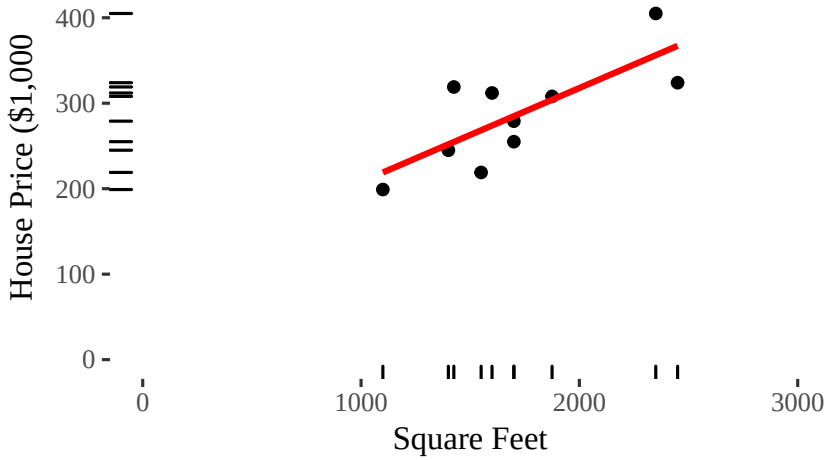
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

‘‘‘

House price model: Scatter Plot and Prediction Line



- Estimated House Price = 98.24833 + 0.10977(sq ft)

Simple Linear Regression Example: Interpretation of b_o

- b_0 is the estimated average value of Y when the value of X is zero (if X = 0 is in the range of observed X values).
- Because a house cannot have a square footage of 0, b_0 has no practical application.

Simple Linear Regression Example: Interpreting b_1

- b_1 estimates the change in the average value of Y as a result of a one-unit increase in X.
- Here, b_1 = 0.10977 tells us the mean value of a house increases by .10977($1000) $109.77, on average, for each additional square foot.

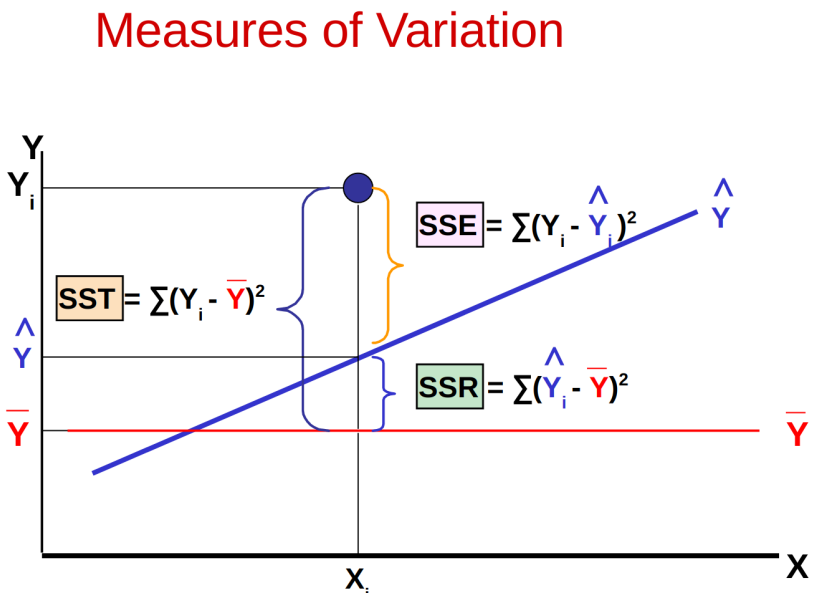
Simple Linear Regression Example: Making Predictions

- Predict the price for a house with 2000 square feet:
  Estimate price = 98.25 + 0.1098(sq.ft.)
  = 98.25 + 0.1098(2000)
  = 317.85
  The predicted price for a house with 2000 square feet is 317.85($1,000s)
  = $317,850

Beware of overrunning your headlights

- When using a regression model for prediction, only predict within the relevant range of data.
- Relevant range for interpolation - 1,000 to 2,500 square feet.
- Do not try to extrapolate beyond the range of observed X's

Measures of Variation



Total variation is made up of two parts:

$$SST = SSR + SSE \tag{5}$$

Total Sum of Squares = Regression Sum of Squares + Error Sum of Squares

$$SST = \sum (Y_i - \bar{Y})^2 \tag{6}$$

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2 \tag{7}$$

$$SSE = \sum (Y_i - \hat{Y})^2 \tag{8}$$

where: $\bar{Y}$ = Mean value of the dependent variable
Y_i = Observed value of the dependent variable
$\hat{Y}$ = Predicted value of Y for the given X_i value

- SST = total sum of squares (Total Variation)
  - Measures the variation of the Yi values around their mean Y
- SSR = regression sum of squares (Explained Variation)
  - Variation attributable to the relationship between X and Y
- SSE = error sum of squares (Unexplained Variation)
  - Variation in Y attributable to factors other than X

Coefficient of Determination, r^2

- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable.
- The coefficient of determination is also called r-squared and is denoted as r^2

$$r^2 = \frac{SSR}{SST} = \frac{Regression\ Sum\ of\ Squares}{Total\ Sum\ Squares} \tag{9}$$

- Note: $0 \leq r^2 \leq 1$

Assumptions of Regression L.I.N.E

- Linearity
  - The relationship between X and Y is linear
- Independence of Errors
  - Error values are statistically independent
- Normality of Error
  - Error values are normally distributed for any given value of X
- Equal Variance (also called homoscedasticity)
  - The probability distribution of the errors has constant variance

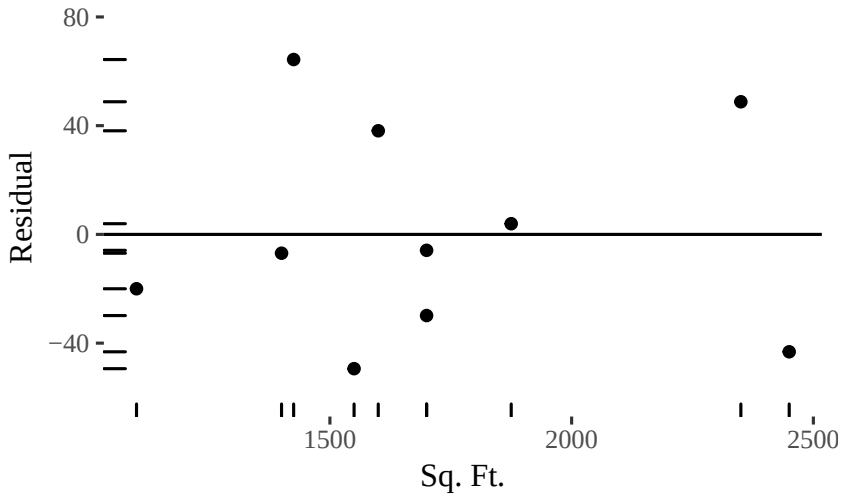
Residual Analysis

- The residual for observation i, e_i, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Evaluate independence assumption
  - Evaluate normal distribution assumption
  - Examine for constant variance for all levels of X (homoscedasticity)
- Graphical Analysis of Residuals Can plot residuals vs. X

Table 4: Residual Analysis

House Price in $1000s	Square Feet (X)	Predicted Price	Residual
245	1400	251.9263	-6.92633
312	1600	273.8803	38.11967

House Price in $1000s	Square Feet (X)	Predicted Price	Residual
279	1700	284.8573	-5.85733
308	1875	304.0671	3.93292
199	1100	218.9953	-19.99533
219	1550	268.3918	-49.39183
405	2350	356.2078	48.79217
324	2450	367.1848	-43.18483
319	1425	254.6706	64.32942
255	1700	284.8573	-29.85733



Checking for Normality

- Examine the Stem-and-Leaf Display of the Residuals
- Examine the Boxplot of the Residuals
- Examine the Histogram of the Residuals
- Construct a Normal Probability Plot of the Residuals

Confidence Interval Estimate for the Slope

Remember: When using the confidence interval method for hyptothesis testing, the null hypothesis is that (in this case) the slope is 0, and the alternative hypothesis is the slope is not equal to zero.

$H_0 : \beta_1 = 0$ and

$H_1 : \beta_1 \neq 0$

Which translate into: *If the interval does not include 0 reject* H_0

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858).

Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.74 and $185.80 per square foot of house size.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Figure 1: "Excel Output"

This 95% confidence interval does not include 0.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance

Estimating Mean Values and Predicting Individual Values

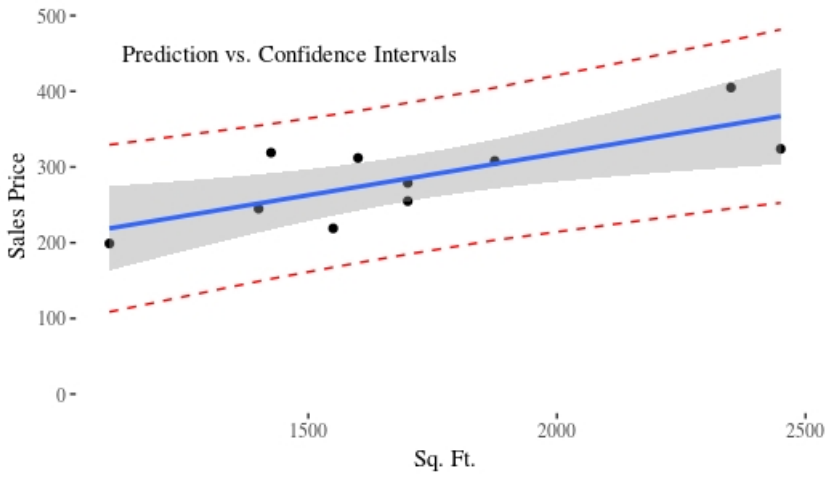
Goal: Form intervals around Y to express uncertainty about the value of Y for a given X_i

There are two intervals we are concerned with:

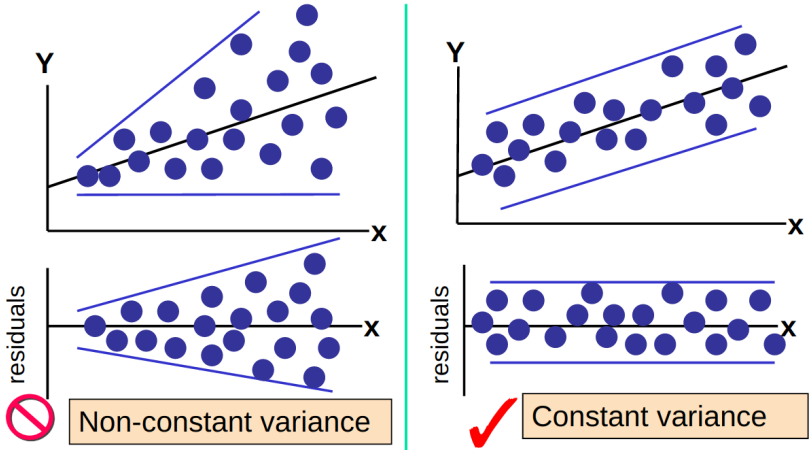
- Confidence Interval for the **mean** of Y, given X_i
- Prediction Interval for an **individual** Y, given X_i

The chart belows these concepts.

- The points represent the data pairs,
- the solid line represents the regression,
- the shaded area represents limits of the mean of Y given X_i,
- the dotted line represents the upper and lower limits for an individual Y, given X_i

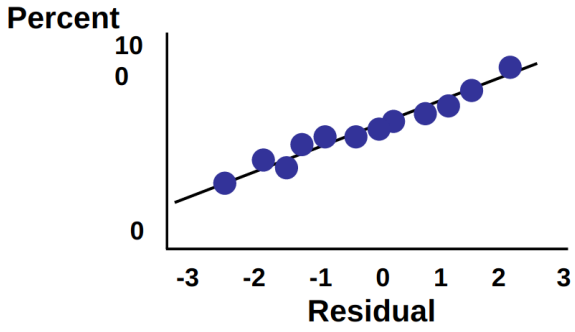


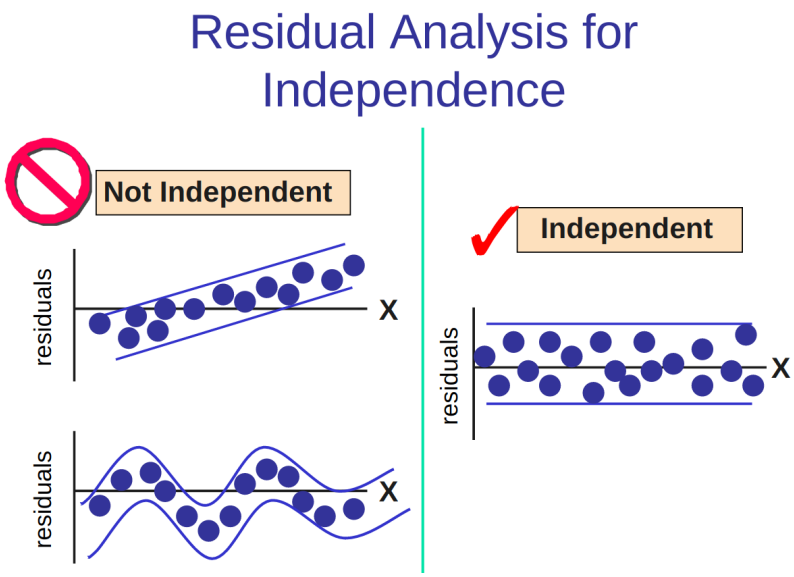
Residual Analysis for Equal Variance



Residual Analysis for Normality

When using a normal probability plot, normal errors will approximately display in a straight line




Residual Analysis for Independence
Not Independent
residuals
X
residuals
X
Independent
residuals
X