Simple Linear Regression

Simple Linear Regression

- Our objective is to study the relationship between two variables X and Y.
- One way to study the relationship between two variables is by means of **regression.**
- Regression analysis is the process of estimating a functional relationship between X and Y. A regression equation is often used to predict a value of Y for a given value of X.
- Another way to study relationship between two variables is **correlation.** It involves measuring the direction and the strength of the **linear** relationship.

First-Order Linear Model = Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where

y = dependent variable

x = independent variable

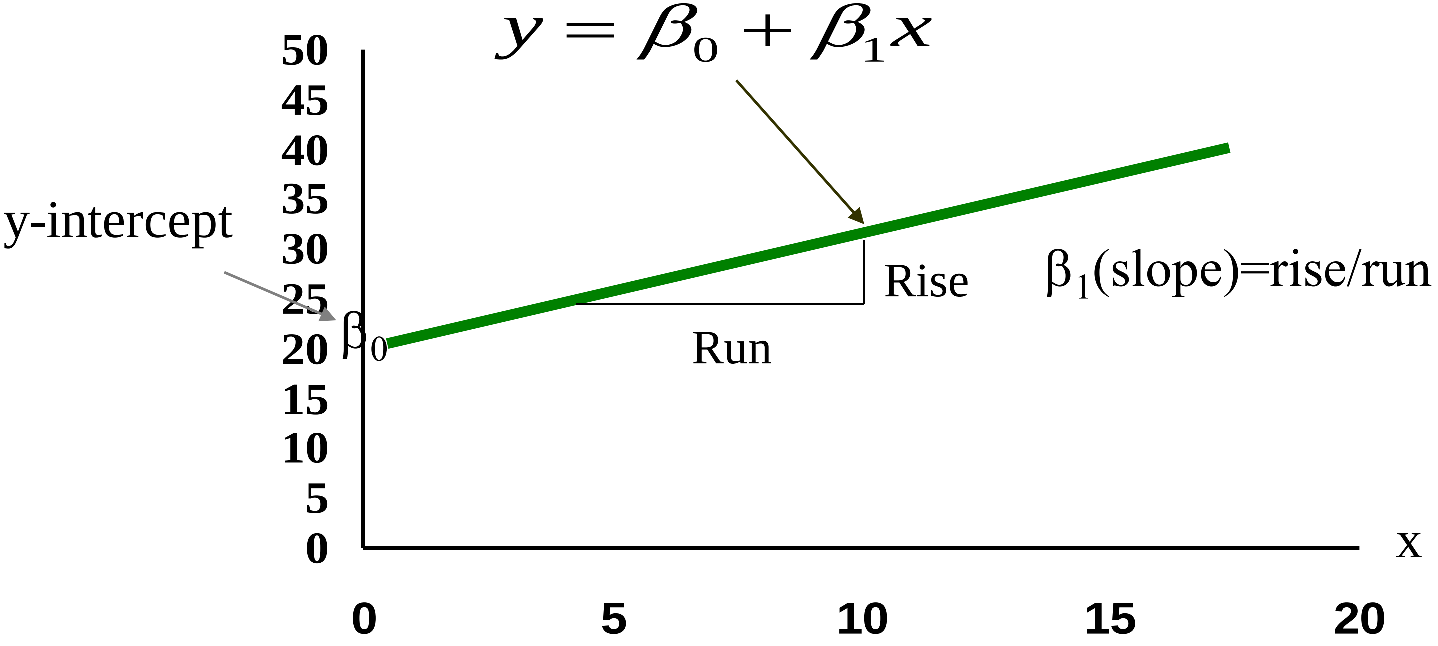
β_0= y-intercept

β_1= slope of the line

ε = error variable

Deterministic Component of Model

$$y = \beta_0 + \beta_1 x$$



LEAST SQUARES MODEL (Best Possible Fit)

- To estimate the parameters β$_0$ and β$_1$, we use least square method.

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$SS_x = \sum (x_i - \bar{x})^2$$

$$= \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

Example

Data: Determine the straight line that fits this data:

x	**2**	**4**	**8**	**10**	**13**	**16**
y	**2**	**7**	**25**	**26**	**38**	**50**

$$\sum x_i = 53$$

$$\sum y_i = 148$$

$$\sum x_i^2 = 609$$

$$\sum x_i y_i = 1{,}786$$

$$\bar{x} = \frac{53}{6} = 8.8333$$

$$\bar{y} = \frac{148}{6} = 24.6666$$

Example (Contd.)

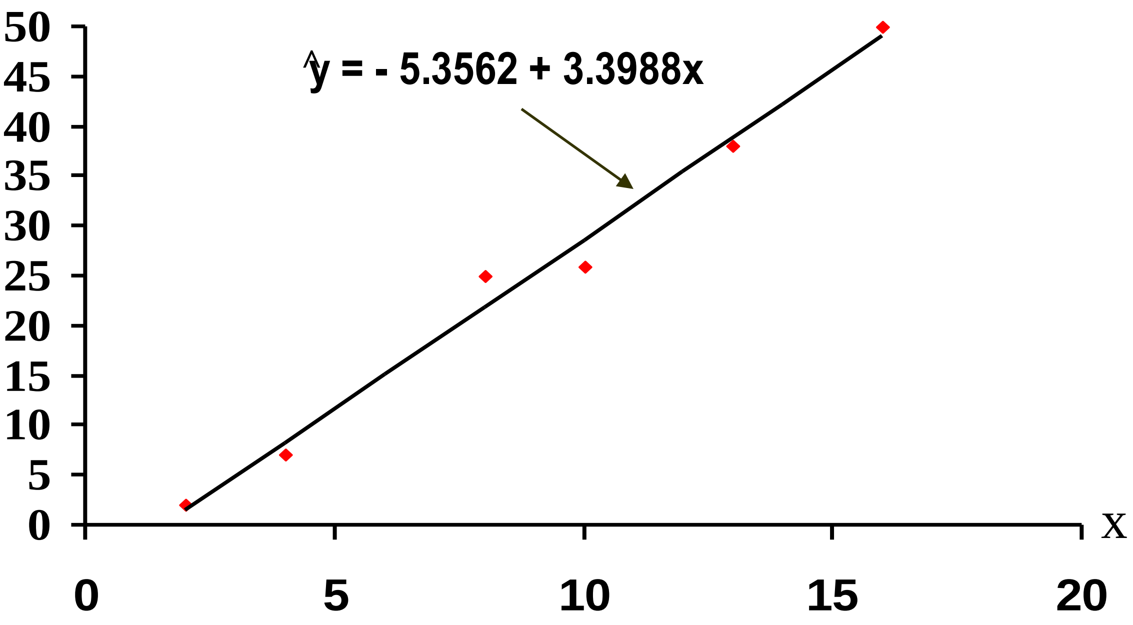
$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{1,786 - \frac{(53)(148)}{6}}{609 - \frac{(53)^2}{6}}$$

$$= \frac{478.667}{140.833} = 3.399$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \left(\frac{148}{6} \right) - 3.399 \left(\frac{53}{6} \right)$$

$$= -5.356$$

Scatter Plot and Fitted line



ERROR

- **The scatterplot shows that the points are not on a line, and so, in addition to the the relationship we also describe error:**

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

- The Y's are the responses (or dependent) variable. The x's are the predictors or independent variable, and the epsilon's are the errors. We assume they are normal, mutually independent, and have variance σ^2.

- LEAST SQUARES: Minimize

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

- The minimizing y-intercept and slope are given by $\hat{\beta}_0$ and $\hat{\beta}_1$. We use the notation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- The quantities $R_i = y_i - \hat{y}_i$ are called the residuals. If we assume a normal error, these should look normal.

WHAT FORM DOES THE ERROR TAKE?

- Each observation may be decomposed into two parts:

$$y = \hat{y} + (y - \hat{y})$$

- The first part is used to determine the fit, and the second to estimate the error.
- We estimate the variance of the error by using the sum of squares error:

$$SSE = \sum (Y - \hat{Y})^2 = S_{yy} - \left(\frac{S_{xy}^2}{S_{xx}} \right)$$

ESTIMATE OF σ^2

- We estimate σ^2 by

$$s^2 = \frac{SSE}{n-2} = MSE$$

Simple Linear Regression Model

- The Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- The Least Squares Regression Line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x}$$

$$SS_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$SS_x = \sum (x_i - \overline{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

EXAMPLE

- An educational economist wants to establish the relationship between an individual’s income and education. He takes a random sample of 10 individuals and asks for their income (in $1000s) and education (in years). The results are shown below. Find the least squares regression line.

Education	11	12	11	15	8	10	11	12	17	11
Income	25	33	22	41	18	28	32	24	53	26

DEPENDENT AND INDEPENDENT VARIABLES

- The dependent variable is the one that we want to forecast or analyze.
- The independent variable is hypothesized to affect the dependent variable.
- In this example, we wish to analyze income and we choose the variable individual’s education that most affects income. Hence, y is income and x in individual’s education

FIRST STEP

$$\sum x_i = 118$$

$$\sum x_i^2 = 1450$$

$$\sum y_i = 302$$

$$\sum y_i^2 = 10072$$

$$\sum x_i y_i = 3779$$

SUM OF SQUARES

$$SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 3779 - \frac{(118)(302)}{10} = 215.4$$

$$SS_x = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 1450 - \frac{(118)^2}{10} = 57.6$$

Therefore,

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{215.4}{57.6} = 3.74$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{302}{10} - 3.74\frac{118}{10} = -13.93$$

The Least Squares Regression Line

- The least squares regression line is

$$\hat{y} = -13.93 + 3.74x$$

- Interpretation of coefficients:

*The sample slope $\hat{\beta}_1 = 3.74$ tells us that on average for each additional year of education, an individual's income rises by $3.74 thousand.

* The y-intercept is $\hat{\beta}_0 = -13.93$. This value has no meaning.

ERROR VARIABLE

- ε is normally distributed.
- $E(\varepsilon) = 0$
- The variance of ε is σ^2.
- The errors are independent of each other.
- The estimator of σ^2 is

$$s_\varepsilon^2 = \frac{SSE}{n-2} = MSE$$

where

$$SSE = \sum (Y - \hat{Y})^2 = S_{yy} - \left(\frac{S_{xy}^2}{S_{xx}} \right) \quad \textbf{and} \quad SS_y = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

EXAMPLE (contd.)

- For the previous example

$$SS_y = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 10072 - \frac{(302)^2}{10} = 951.6$$

Hence, SSE is

$$SSE = S_{yy} - \left(\frac{S_{xy}^2}{S_{xx}}\right) = 951.6 - \frac{(215.4)^2}{57.6} = 146.09$$

Therefore,

$$s_\varepsilon^2 = \frac{SSE}{n-2} = \frac{146.09}{10-2} = 18.26 \text{ and } s_\varepsilon = \sqrt{s_\varepsilon^2} = 4.27$$

INTERPRETATION OF s_{ε}^{2}

- The value of s can be compared with the mean value of y to provide a rough guide as to whether s is small or large.
- Since $\bar{y} = 30.2$ and s_{ε}=4.27, we would conclude that s is relatively small, which indicates that the regression line fits the data quite well.

EXAMPLE

- Car dealers across North America use the red book to determine a cars selling price on the basis of important features. One of these is the car's current odometer reading.
- To examine this issue 100 three year old cars in mint condition were randomly selected; their selling price and odometer reading were observed.

Portion of the data file

Odometer	Price
37388	5318
44758	5061
45833	5008
30862	5795
…..	…
34212	5283
33190	5259
39196	5356
36392	5133

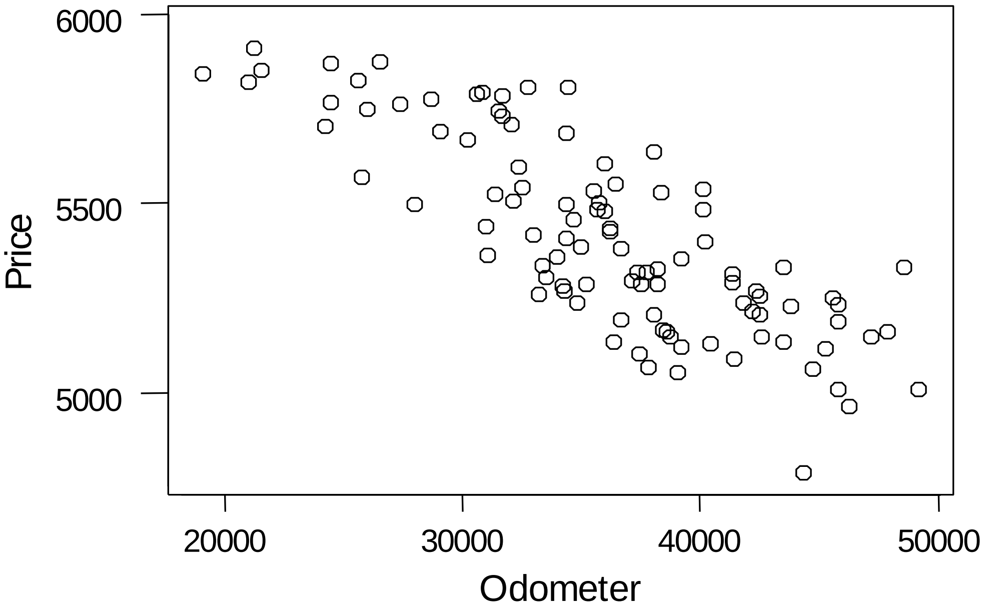
Example
(Minitab Output)

```
Regression Analysis

The regression equation is
Price = 6533 - 0.0312 Odometer

Predictor      Coef       StDev          T     P
Constant     6533.38      84.51      77.31 0.000(SIGNIFICANT)
Odometer   -0.031158    0.002309    -13.49 0.000(SIGNIFICANT)

S = 151.6       R-Sq = 65.0%     R-Sq(adj) = 64.7%

Analysis of Variance

Source       DF         SS         MS         F        P
Regression    1    4183528    4183528    182.11    0.000
Error        98    2251362      22973
Total        99    6434890
```

Example

- The least squares regression line is

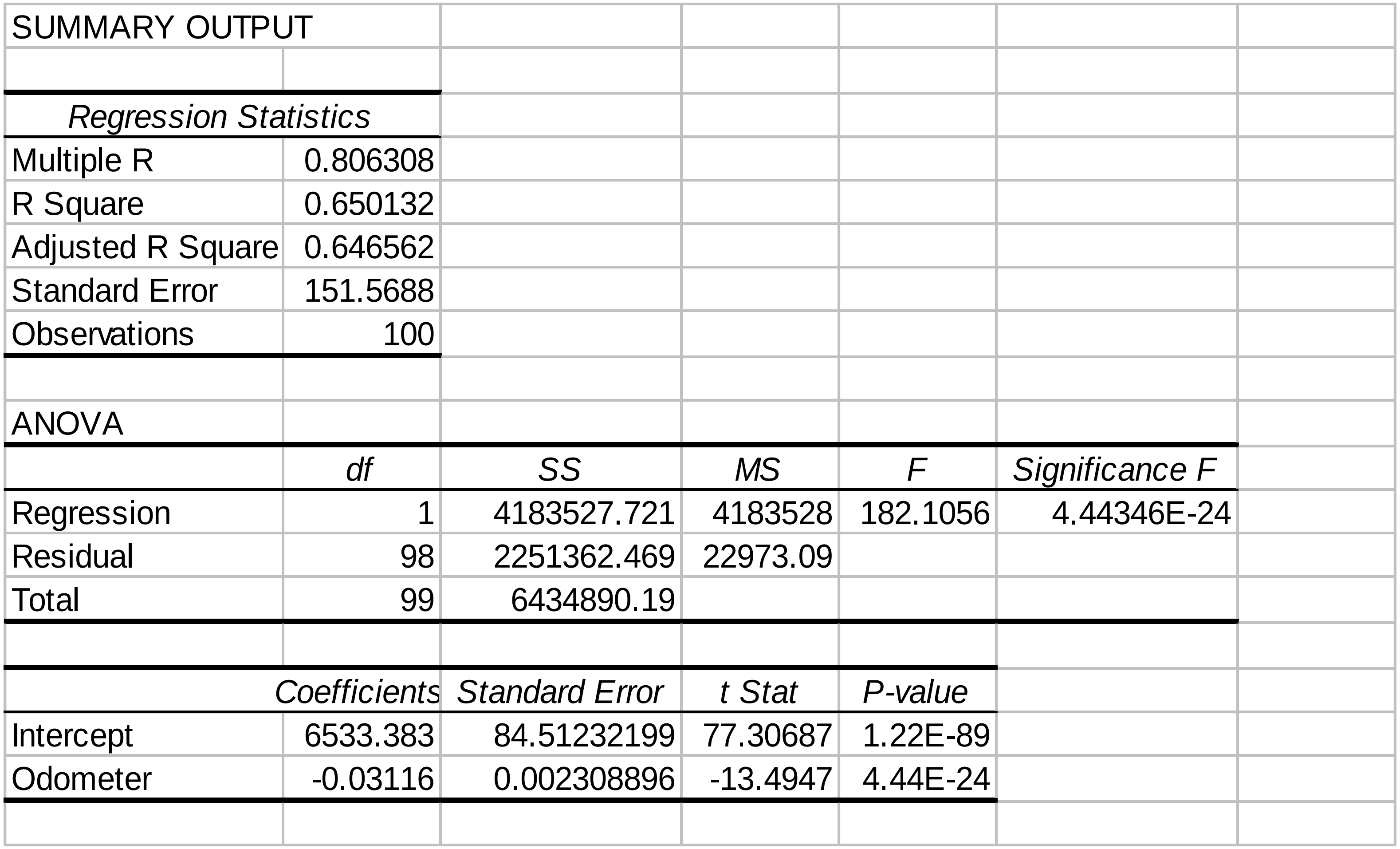
$$\hat{y} = 6533.38 - 0.031158x$$



Interpretation of the coefficients

- $\hat{\beta}_1 = -0.031158$ means that for each additional mile on the odometer, the price decreases by an average of 3.1158 cents.
- $\hat{\beta}_0 = 6533.38$ means that when x = 0 (new car), the selling price is $6533.38 but x = 0 is not in the range of x. So, we cannot interpret the value of y when x=0 for this problem.
- R^2=65.0% means that 65% of the variation of y can be explained by x. The higher the value of R^2, the better the model fits the data.

Excel Example



SUMMARY OUTPUT

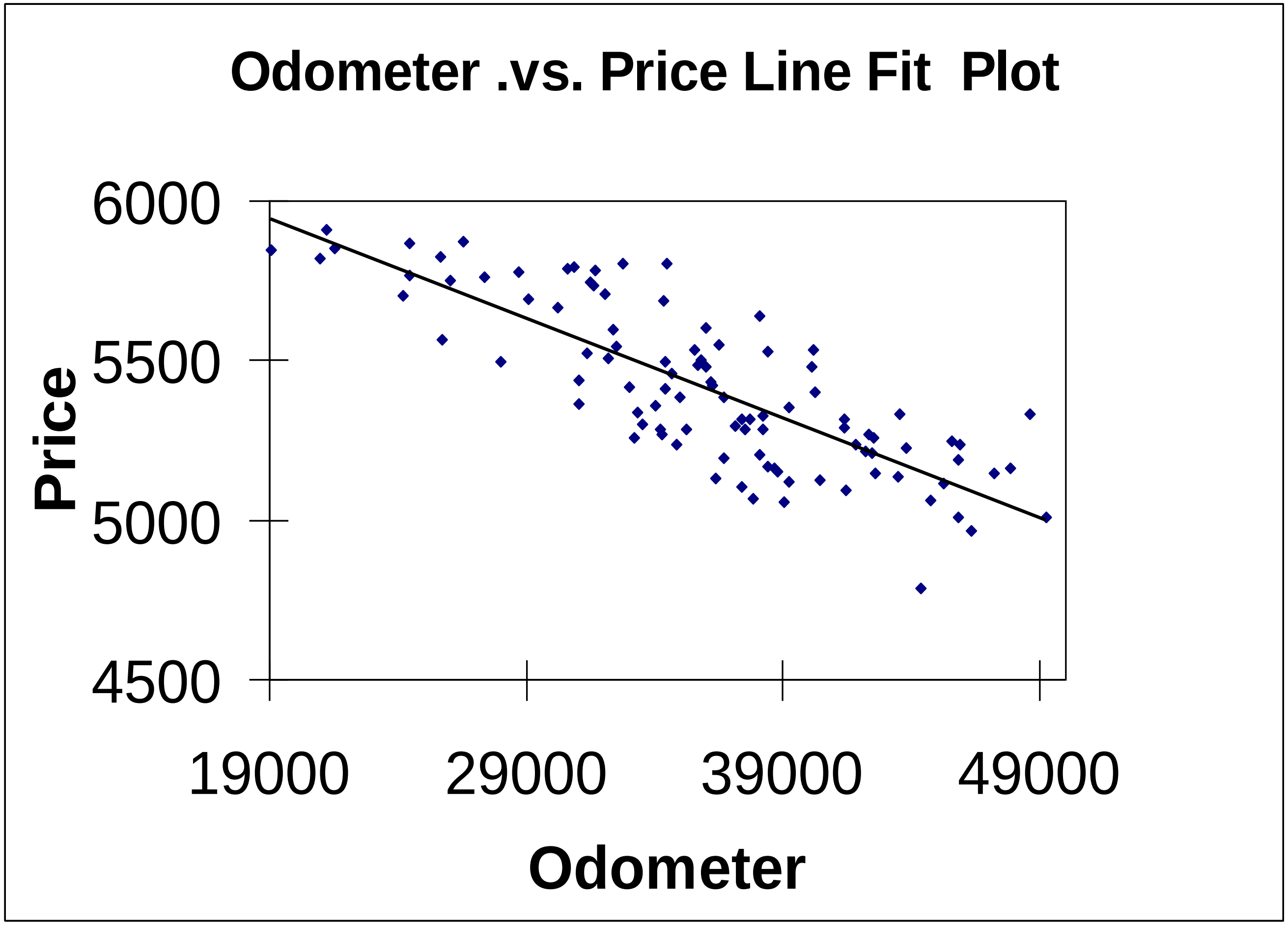
Regression Statistics	
Multiple R	0.806308
R Square	0.650132
Adjusted R Square	0.646562
Standard Error	151.5688
Observations	100

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4183527.721	4183528	182.1056	4.44346E-24
Residual	98	2251362.469	22973.09		
Total	99	6434890.19			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	6533.383	84.51232199	77.30687	1.22E-89
Odometer	-0.03116	0.002308896	-13.4947	4.44E-24

Example
(Excel Scatter Plot)



TESTING THE SLOPE

- **Are X and Y linearly related?**

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

•Test Statistic:

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \quad \text{where} \quad s_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{SS_x}}$$

TESTING THE SLOPE

- The Rejection Region: Reject H_0 if

  $t < -t_{\alpha/2,n-2}$ or $t > t_{\alpha/2,n-2}$.

- If we are testing that high x values lead to high y values, H_A: $\beta_1 > 0$. Then, the rejection region is $t > t_{\alpha,n-2}$.

- If we are testing that high x values lead to low y values or low x values lead to high y values, H_A: $\beta_1 < 0$. Then, the rejection region is $t < - t_{\alpha,n-2}$.

Assessing the model Example

- Excel output

```
                 Coefficients    Standard Error   t Stat    P-value
Intercept        6533.4          84.512322        77.307    1E-89
Odometer         -0.031          0.0023089        -13.49    4E-24
```

- Minitab output

```
Predictor      Coef       StDev         T       P
Constant     6533.38      84.51     77.31   0.000
Odometer   -0.031158   0.002309    -13.49    0.000
```

Coefficient of Determination

$$R^2 = \frac{SS_{xy}^2}{SS_x - SS_y} = 1 - \frac{SSE}{SS_y}$$

For the data in the example we obtain:

$$R^2 = 1 - \frac{SSE}{SS_y} = 1 - \frac{2,251,363}{6,434,890}$$

$$= 1 - .3499 = .6501$$

Using the Regression Equation

- From the fitted line for the example:
- Suppose we would like to predict the selling price for a car with 40,000 miles on the odometer

$$\begin{aligned}\hat{y} &= 6{,}533 - 0.0312x \\ &= 6{,}533 - 0.0312(40{,}000) \\ &= \$5{,}285\end{aligned}$$

Prediction and Confidence Intervals

- **Prediction Interval of y for x=x$_g$:** The confidence interval for ***predicting the particular value of y*** for a given x

$$\hat{y} \pm t_{\alpha/2,n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{SS_x}}$$

- **Confidence Interval of E(y|x=x$_g$):** The confidence interval for ***estimating the expected value of y*** for a given x

$$\hat{y} \pm t_{\alpha/2,n-2} s_e \sqrt{\frac{1}{n} + \frac{(x_g - \bar{x})^2}{SS_x}}$$

Solving by Hand (Prediction Interval)

- From previous calculations we have the following (example):

$$\hat{y} = 5285, s_{\varepsilon} = 151.6, SS_x = 4309340160, \bar{x} = 36,009$$

- Thus a 95% **prediction interval** for x=40,000 is:

$$5,285 \pm 1.984(151.6)\sqrt{1 + \frac{1}{100} + \frac{(40,000 - 36,009)^2}{4,309,340,160}}$$

$$5,285 \pm 303$$

•The prediction is that the selling price of the car will fall between $4982 and $5588.

Solving by Hand (Confidence Interval)

- Thus a 95% **confidence interval** of

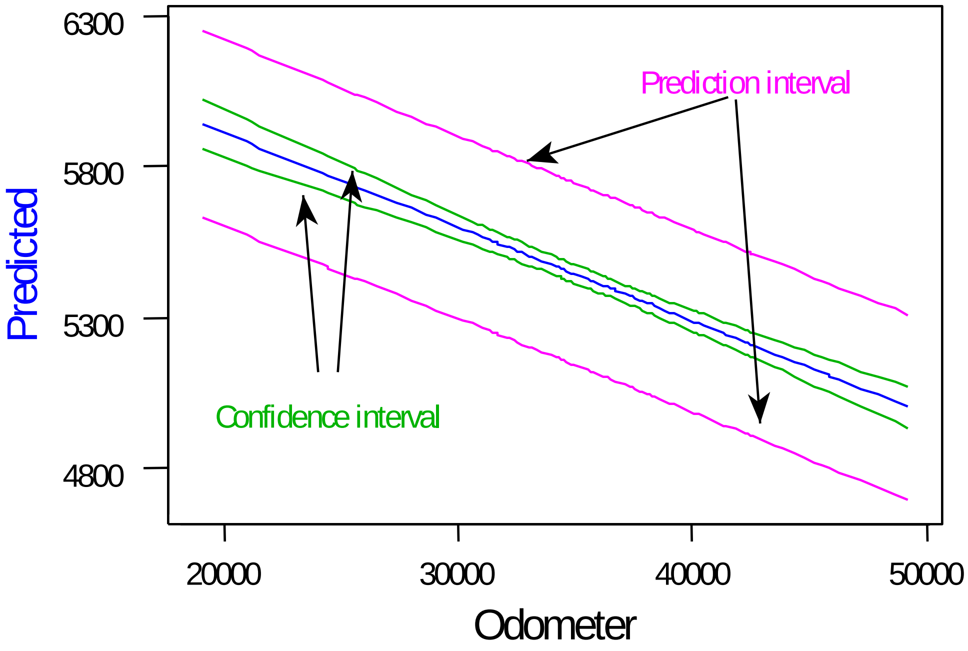
  E(y| x=40,000) is:

$$5{,}285 \pm 1.984(151.6)\sqrt{\frac{1}{100}+\frac{(40{,}000-36{,}009)^2}{4{,}309{,}340{,}160}}$$

$$5{,}285 \pm 35$$

•The mean selling price of the car will fall between $5250 and $5320.

Prediction and Confidence Intervals Graph



Regression Diagnostics

- How to diagnose violations and how to deal with observations that are unusually large or small.
- **Residual Analysis**:

⇒**Non-normality**

⇒**Heteroscedasticity**

⇒**Non-independence of the errors**

⇒**Outlier**

⇒**Influential observations**

STANDARDIZED RESIDUALS

- The standardized residuals are calculated as

$$\text{Standardized residual} = \frac{r_i}{s_\varepsilon}$$

where $r_i = y_i - \hat{y}_i$. The standard deviation of the i-th residual is

$$s_{r_i} = s_\varepsilon \sqrt{1 - h_i} \text{ where } h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{SS_x}$$

NON-NORMALITY

- The errors are normally distributed. To check the normality of errors,we use histogram of the residuals or normal probability plot of residuals.

HETEROSCEDASTICITY

- The error variance σ_{ε}^{2} should be constant. When this requirement is violated, the condition is called heteroscedasticity.
- To diagnose hetersocedastisticity or homoscedasticity, one method is to plot the residuals against the predicted value of y. If the points are distributed evenly around the expected value of errors which is 0, this means that the error variance is constant.

NON-INDEPENDENCE OF ERROR VARIABLE

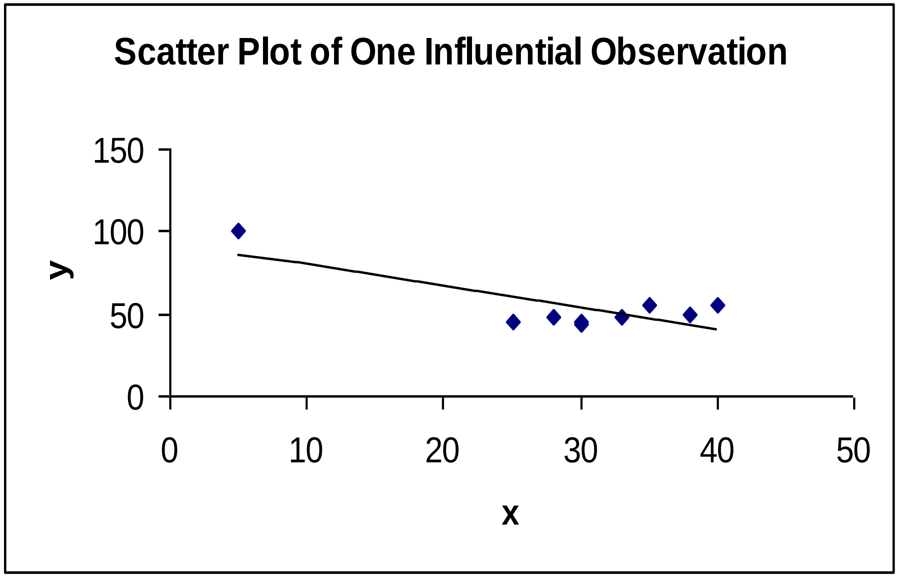
- The values of error should be independent. When the data are time series, the errors often are correlated (i.e., autocorrelated or serially correlated). To detect autocorrelation we plot the residuals against the time periods. If there is no pattern, this means that errors are independent.

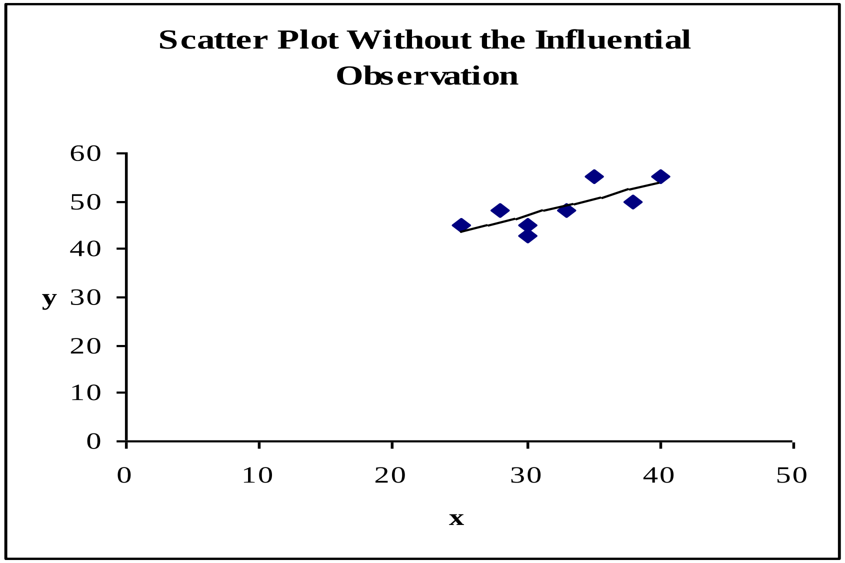
OUTLIER

- An outlier is an observation that is unusually small or large. Several possibilities to have an outlier are
- Error in recording the data.⇒ Detect the error and correct it
- The point should not have been included in the data (belongs to another population) ⇒ Discard the point from the sample
- The observation is unusually small or large although it belong to the sample and there is no recording error. ⇒ There is nothing to do.

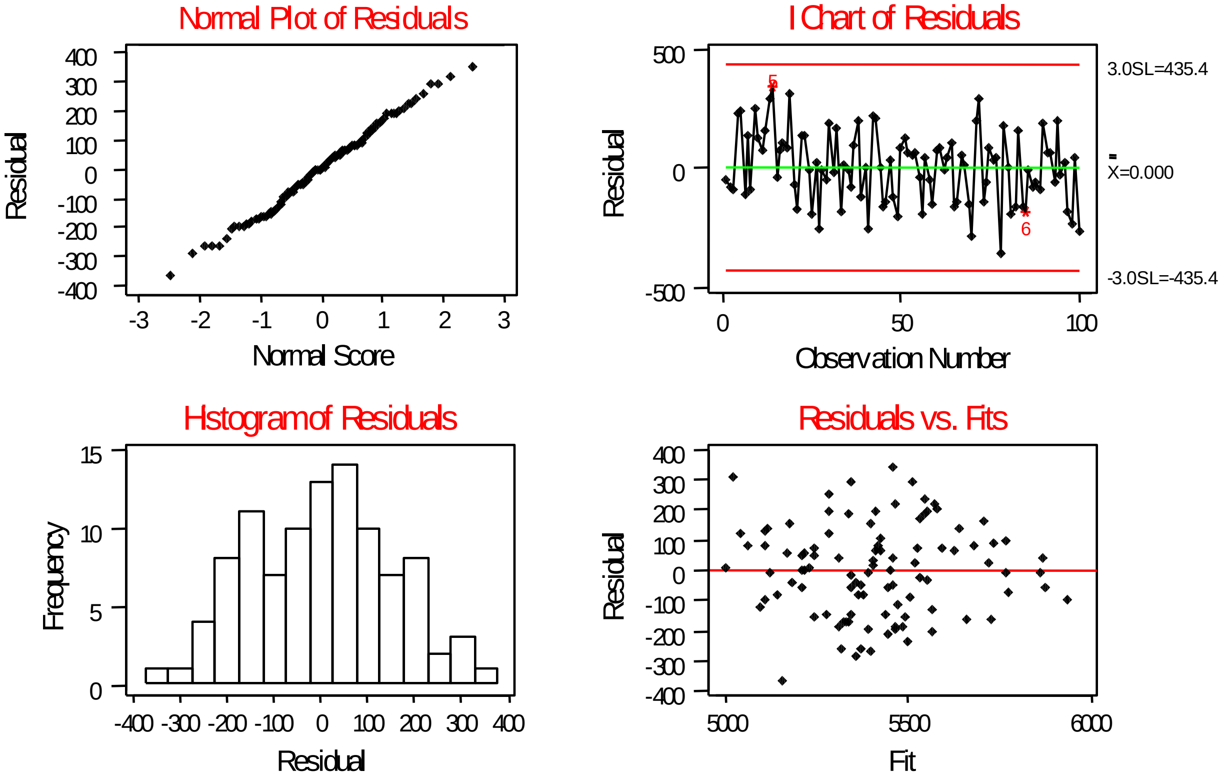
INFLUENTIAL OBSERVATIONS

- One or more observations have a large influence on the statistics.





Regression Diagnostics



EXERCISE

- It is doubtful that any sports collects more statistics than baseball. The fans are always interested in determining which factors lead to successful teams. The table below lists the team batting average and the team winning percentage for the 14 American League teams at the end of a recent season.

Team-B-A	Winning%
0.254	0.414
0.269	0.519
0.255	0.500
0.262	0.537
0.254	0.352
0.247	0.519
0.264	0.506
0.271	0.512
0.280	0.586
0.256	0.438
0.248	0.519
0.255	0.512
0.270	0.525
0.257	0.562

y = winning % and x = team batting average

a) LS Regression Line

$$\sum x_i = 3.642, \sum x_i^2 = 0.949$$

$$\sum y_i = 7.001, \sum y_i^2 = 3.549$$

$$\sum x_i y_i = 1.824562$$

$$SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 1.824562 - \frac{(3.642)(7.001)}{14} = 0.0033$$

$$SS_x = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 0.948622 - \frac{(3.642)^2}{14} = 0.00118$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{0.003302}{0.001182} = 0.7941$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 0.5 - (0.7941)0.26 = 0.2935$$

- The least squares regression line is

$$\hat{y} = 0.2935 + 0.7941x$$

- The meaning $\hat{\beta}_1 = 0.7941$ is for each additional batting average of the team , the winning percentage increases by an average of 79.41%.

b) STANDARD ERROR OF ESTIMATE

$$SSE = S_{yy} - \left(\frac{S_{xy}^2}{S_{xx}}\right) = \left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right) - \left(\frac{S_{xy}^2}{S_{xx}}\right)$$

$$= (3.548785 - \frac{7.001^2}{14}) - \frac{0.003302^2}{0.00182} = 0.03856$$

So,

$$s_\varepsilon^2 = \frac{SSE}{n-2} = \frac{0.03856}{14-2} = 0.00321 \text{ and } s_\varepsilon = \sqrt{s_\varepsilon^2} = 0.0567$$

- Since s_ε=0.0567 is small, we would conclude that s is relatively small, which indicates that the regression line fits the data quite well.

c) Do the data provide sufficient evidence at the 5% significance level to conclude that higher team batting average lead to higher winning percentage?

$$H_0: \quad \beta_1 = 0$$

$$H_A: \quad \beta_1 > 0$$

Test statistic: $t = \dfrac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = 1.69$ (p-value=.058)

Conclusion: Do not reject H_0 at $\alpha = 0.05$. The higher team batting average do not lead to higher winning percentage

d) Coefficient of correlation

$$R^2 = \frac{SS_{xy}^2}{SS_x - SS_y} = 1 - \frac{SSE}{SS_y} = 1 - \frac{0.03856}{0.04778} = 0.1925$$

The 19.25% of the variation in the winning percentage can be explained by the batting average.

e) Predict with 90% confidence the winning percentage of a team whose batting average is 0.275.

$$\hat{y} = 0.2935 + 0.7941(0.275) = 0.5119$$

$$\hat{y} \pm t_{\alpha/2,n-2} s_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_g - \overline{x})^2}{SS_x}} =$$

$$0.5119 \pm (1.782)(0.0567)\sqrt{1 + \frac{1}{14} + \frac{(0.275 - 0.2601)^2}{0.001182}}$$

$$0.5119 \pm 0.1134$$

90% PI for y: $(0.3985, 0.6253)$

- The prediction is that the winning percentage of the team will fall between 39.85% and 62.53%.