CHAPTER 10

REGRESSION AND CORRELATION

INTRODUCTION

- In this chapter we employ Regression Analysis to examine the relationship among quantitative variables.
- The technique is used to predict the value of one variable (the dependent variable - y)based on the value of other variables (independent variables x_1, x_2,…x_k.)

The Model

- The first order linear model or a simple regression model,

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

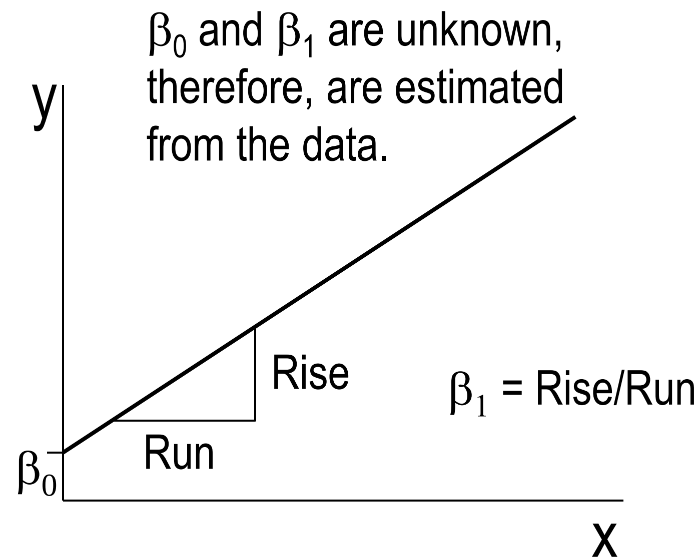
y = dependent variable

x = independent variable

β_0 = y-intercept

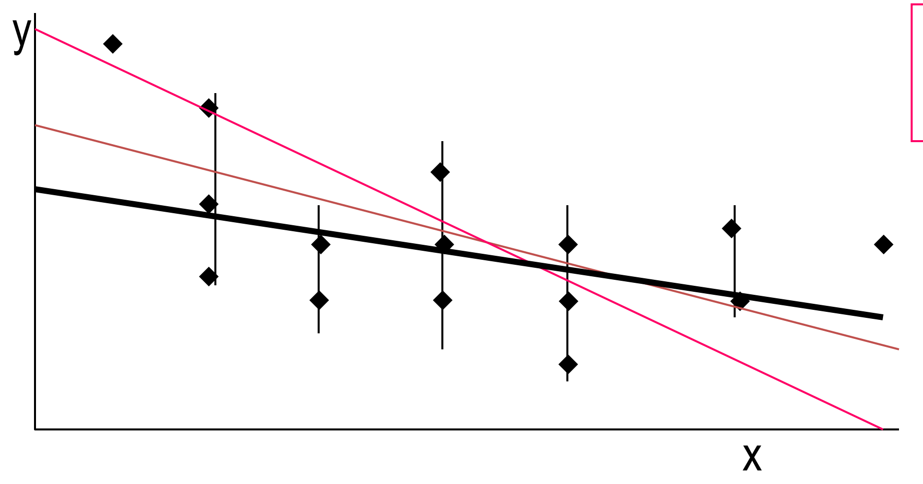
β_1 = slope of the line

ε = error variable



Estimating the Coefficients

- The estimates are determined by
  – drawing a sample from the population of interest,
  – calculating sample statistics.
  – producing a straight line that cuts into the data.



The question is:
Which straight line fits best?

Least Squares Method

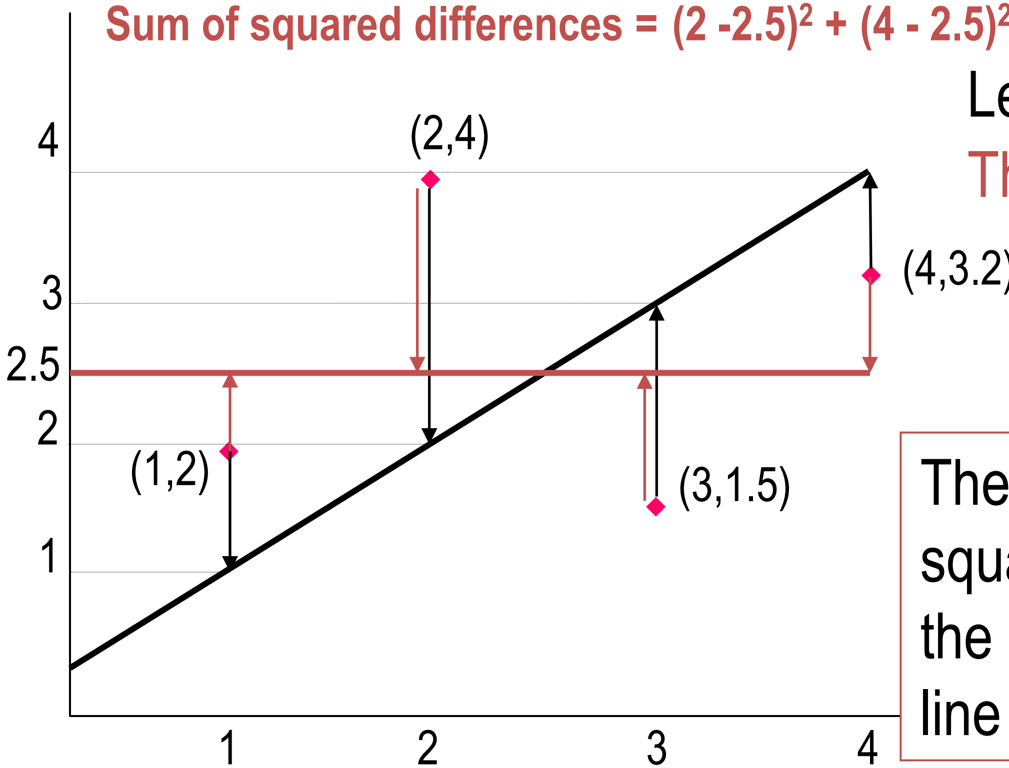
The best line is the one that minimizes the sum of squared vertical differences between the points and the line.

Sum of squared differences = $(2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89$

Sum of squared differences = $(2 - 2.5)^2 + (4 - 2.5)^2 + (1.5 - 2.5)^2 + (3.2 - 2.5)^2 = 3.99$

Let us compare two lines

The second line is horizontal



The smaller the sum of squared differences the better the fit of the line to the data.

To calculate the estimates of the coefficients that minimize the differences between the data points and the line, use the formulas:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{(\sum x_i \sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \text{ or } \frac{\sum x_i y_i - n\,\bar{x}\,\bar{y}}{\sum x_i^2 - n\,\bar{x}^2}$$

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

Now we define

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

$$SS_y = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = \sum y_i^2 - n\,\bar{y}^2$$

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

Then

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

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

The estimated simple linear regression equation that estimates the equation of the first order linear model is:

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

Example: Relationship between odometer reading and a used car's selling price.

- A car dealer wants to find the relationship between the odometer reading and the selling price of used cars.
- A random sample of 100 cars is selected, and the data
  recorded in file XM18-02.
- Find the regression line.

Car	Odometer	Price
1	37388	5318
2	44758	5061
3	45833	5008
4	30862	5795
5	31705	5784
6	34010	5359
.	.	.
.	.	.
.	.	.

Independent variable x

Dependent variable y

Solution

– Solving by hand

- To calculate $\hat{\beta}_0$ and $\hat{\beta}_1$, we need to calculate several statistics first;

$$\bar{x} = 36\,009.45; \qquad SS_x = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 4\,309\,340\,160$$

$$\bar{y} = 5\,411.41; \qquad SS_{xy} = \sum x_i y_i - \frac{(\sum x_i \sum y_i)}{n} = \text{-}134\,269\,296$$

where n = 100.

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{\text{-}134\,269\,296}{4\,309\,340\,160} = \text{-}.0312$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 5411.41 - (-.0312)(36\,009.45) = 6533.38$$

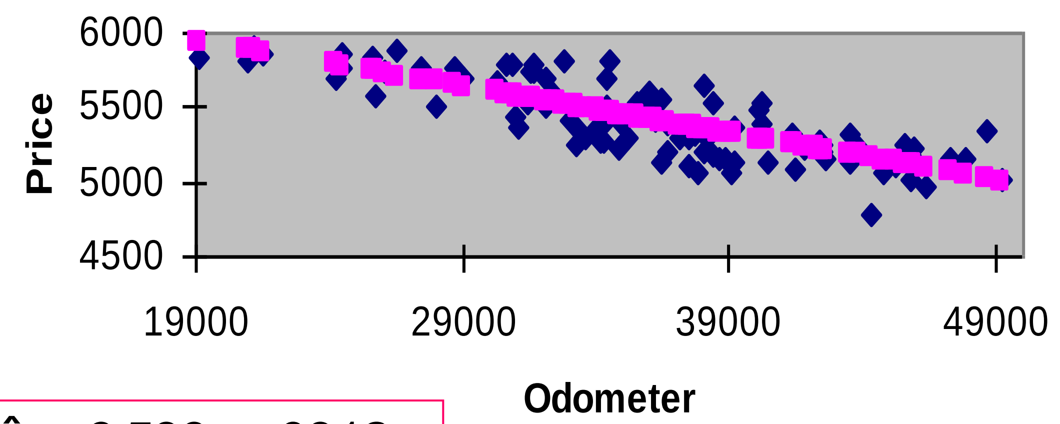
$$\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 6\,533 - .0312x}$$

– Using the computer (see file Xm18-02.xls)

Tools > Data analysis > Regression > [Shade the y range and the x range] > OK

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.806308
R Square	0.650132
Adjusted F	0.646562
Standard E	151.5688
Observatio	100

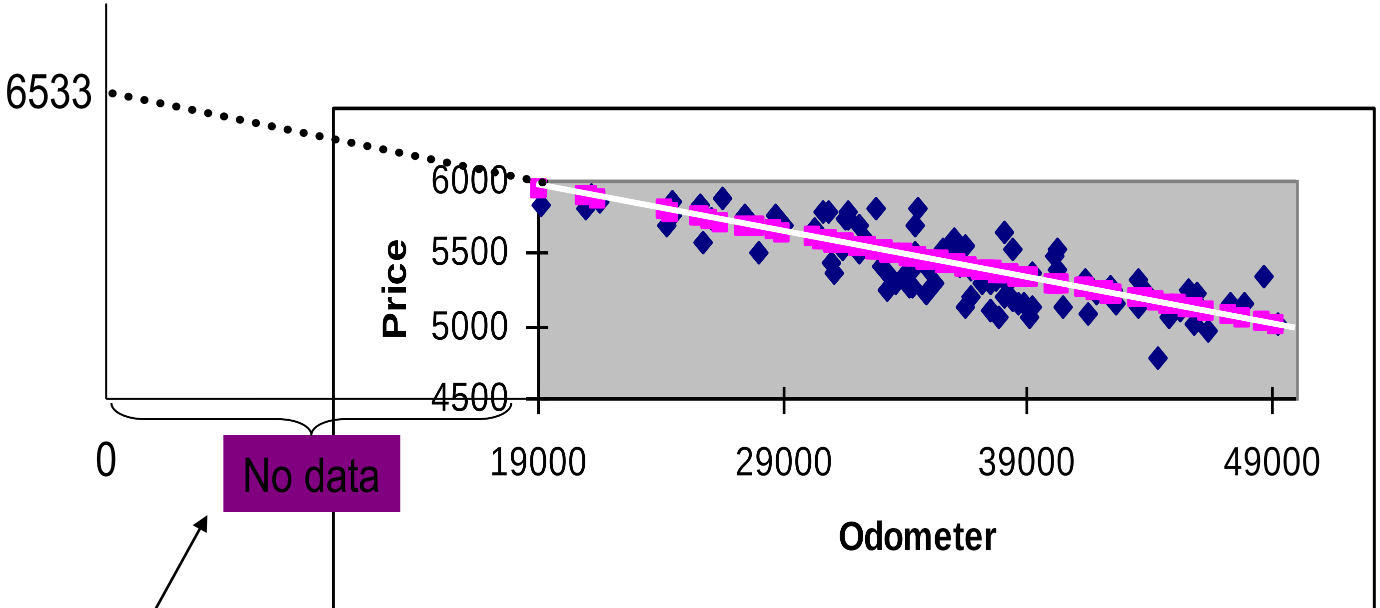


$$\hat{y} = 6\,533 - .0312x$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	1	4183528	4183528	182.1056	4.4435E-24
Residual	98	2251362	22973.09		
Total	99	6434890			

	Coefficients	*Standard Erro*	*t Stat*	*P-value*
Intercept	6533.383	84.51232	77.30687	1.22E-89
Odometer	-0.03116	0.002309	-13.4947	4.44E-24



$$\hat{y} = 6\,533 - .0312x$$

The intercept is b_0 = 6533.

This is the slope of the line.
For each additional mile on the odometer,
the price decreases by an average of \$0.0312

Do not interpret the intercept as the
"Price of cars that have not been driven"

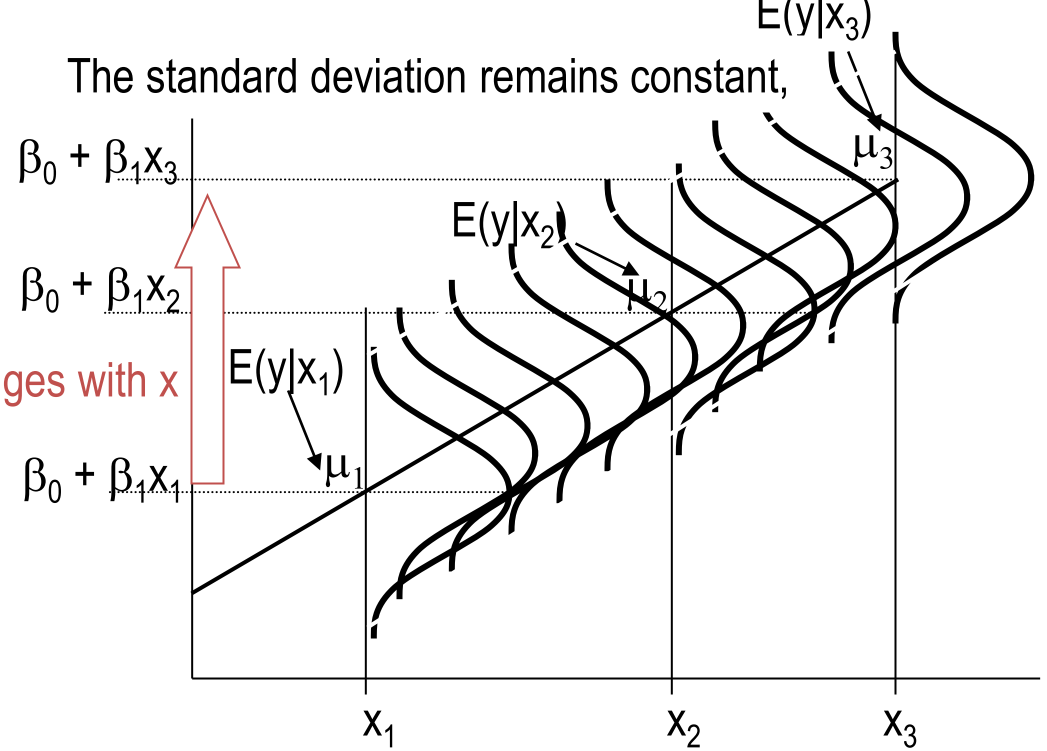
Error Variable: Required Conditions

- The error ε is a critical part of the regression model.
- Five requirements involving the distribution of ε must be satisfied.
  - The mean of ε is zero: $E(\varepsilon) = 0$.
  - The standard deviation of ε is a constant (σ_ε) for all values of x.
  - The errors are independent.
  - The errors are independent of the independent variable x.
  - The probability distribution of ε is normal.

From the first three assumptions we have: y is normally distributed with mean $E(y) = \beta_0 + \beta_1 x$, and a constant standard deviation σ_ε

The standard deviation remains constant,

but the mean value changes with x



Assessing the Model

- The least squares method will produce a regression line whether or not there is a linear relationship between x and y.
- Consequently, it is important to assess how well the linear model fits the data.
- Several methods are used to assess the model:
  - Testing and/or estimating the coefficients.
  - Using descriptive measurements.

- Sum of squares for errors
  - This is the sum of differences between the points and the regression line.
  - It can serve as a measure of how well the line fits the data. $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$

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

  - This statistic plays a role in every statistical technique we employ to assess the model.

- Standard error of estimate
  - The mean error is equal to zero.
  - If σ_ε is small the errors tend to be close to zero (close to the mean error). Then, the model fits the data well.
  - Therefore, we can, use σ_ε as a measure of the suitability of using a linear model.
  - An unbiased estimator of σ_ε^2 is given by s_ε^2

Standard Error of Estimate

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

- Example
  - Calculate the standard error of estimate for example 18.1, and describe what does it tell you about the model fit?
- Solution

Calculated before

$$SS_y = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 6\,434\,890$$

$$SSE = SS_y - \frac{SS_{xy}^2}{SS_x} = 6\,434\,890 - \frac{(-134\,269\,296)^2}{4\,309\,340\,160} = 2\,251\,363$$

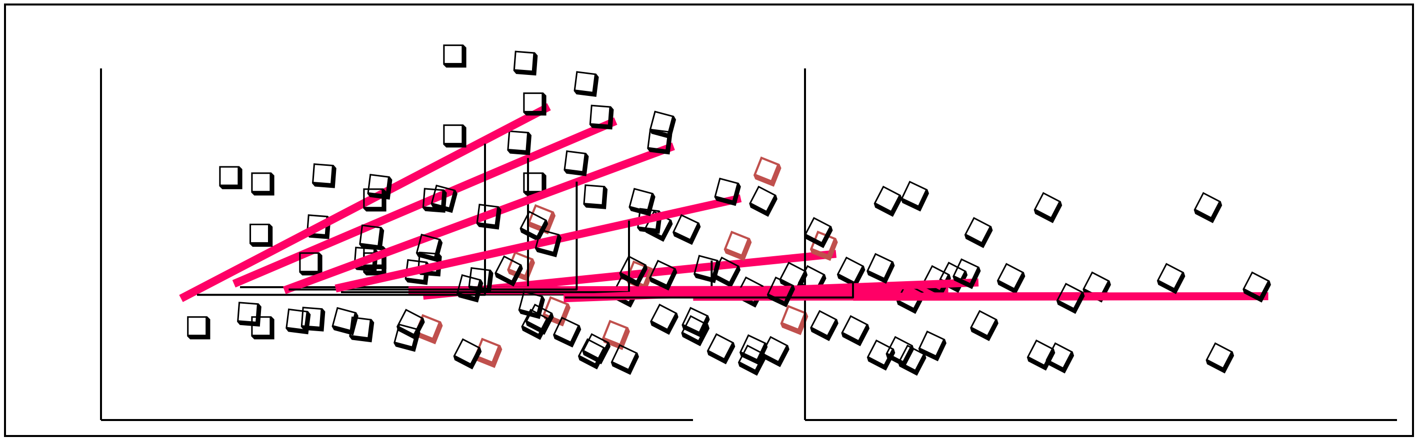
Thus,

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{2\,251\,363}{100-2}} = 151.6$$

It is hard to assess the model based on s_ε even when compared with the mean value of y.

$$s_\varepsilon = 151.6,\ \bar{y} = 5{,}411.4$$

- Testing the slope
  - When no linear relationship exists between two variables, the regression line should be horizontal.



Linear relationship.
Different inputs (x) yield
different outputs (y).

No linear relationship.
Different inputs (x) yield
the same output (y).

The slope is not equal to zero

The slope is equal to zero

- We can draw inference about β_1 from $\hat{\beta}_1$ by testing

  H_0: $\beta_1 = 0$

  H_A: $\beta_1 \neq 0$ (or < 0,or > 0)

  – The test statistic is

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$

where

$$s_{\hat{\beta}_1} = \frac{s_{\varepsilon}}{\sqrt{SS_x}}$$

The standard error of $\hat{\beta}_1$.

  – If the error variable is normally distributed, the statistic is Student t distribution with d.f. = n-2.

- Testing the slope - Example continued
  - Solving by hand

    To compute "t" we need the values of $\hat{\beta}_1$ and $s_{\hat{\beta}_1}$

- $$\hat{\beta}_1 = -.312$$

$$s_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{SS_x}} = \frac{151.6}{\sqrt{4\,309\,340\,160}} = -.00231$$

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = \frac{-.312 - 0}{.00231} = -13.49$$

There is overwhelming evidence to infer that the odometer reading affects the auction selling price.

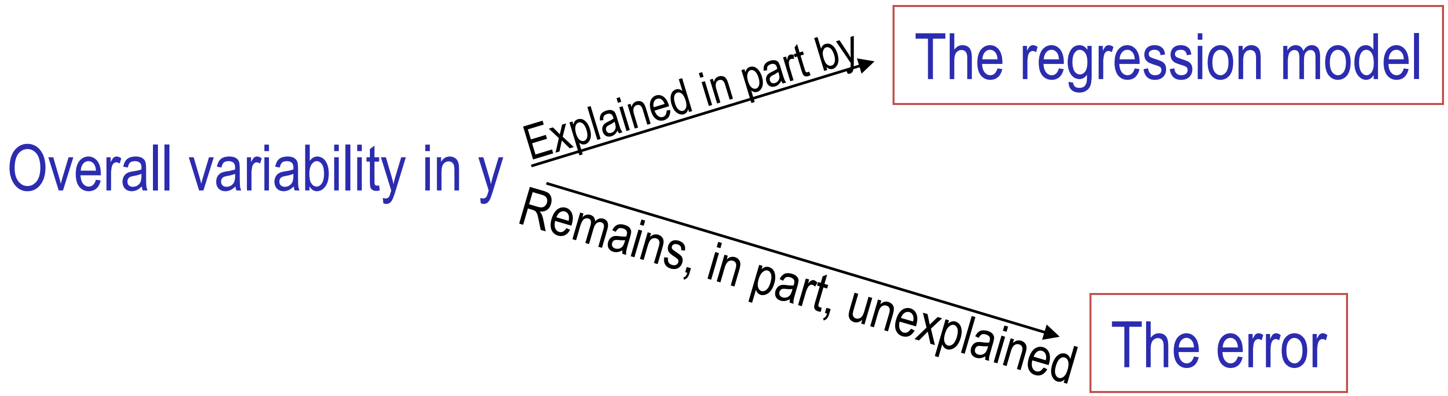
  - Using the computer

	Coefficients	Standard Error	t Stat	P-value
Intercept	6533.383035	84.51232199	77.30687	1.22E-89
Odometer	-0.031157739	0.002308896	-13.4947	4.44E-24

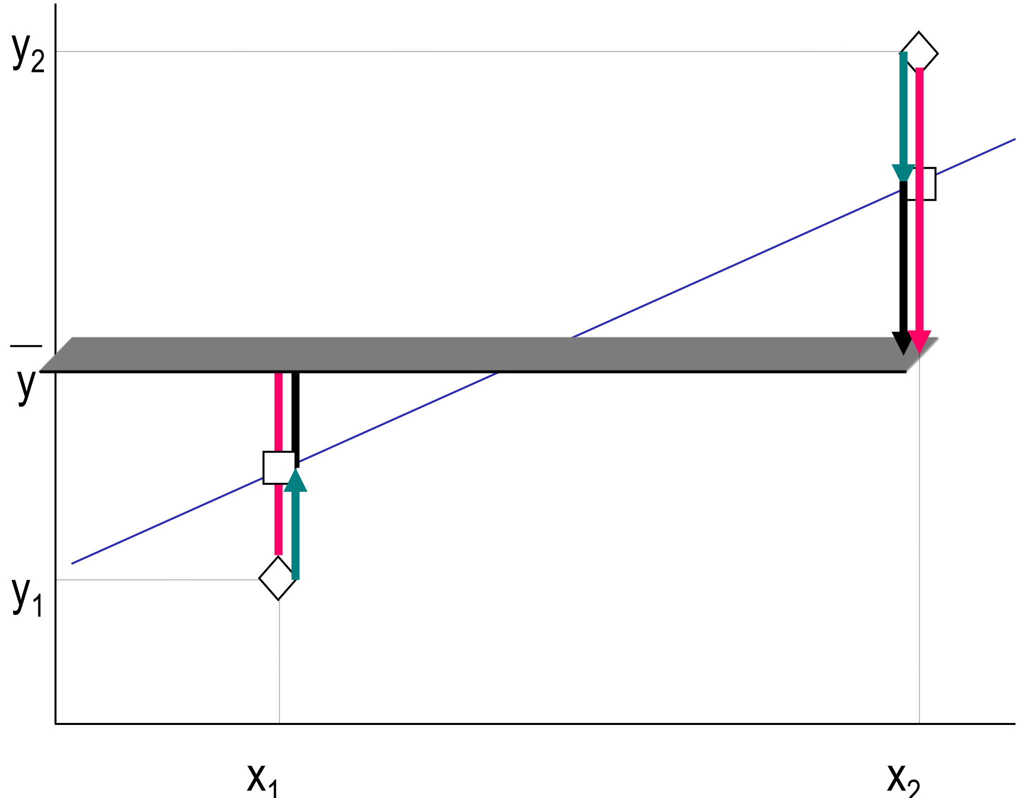
- Coefficient of determination
  - When we want to measure the strength of the linear relationship, we use the coefficient of determination.

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

– To understand the significance of this coefficient note:



Two data points (x_1,y_1) and (x_2,y_2) of a certain sample are shown.



Total variation in y = Variation explained by the regression line) + Unexplained variation (error)

$$(y_1-\bar{y})^2+(y_2-\bar{y})^2 = (\hat{y}_1-\bar{y})^2+(\hat{y}_2-\bar{y})^2 + (y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2$$

$$\boxed{SST = \text{Variation in } y = SSR + SSE}$$

- R^2 measures the proportion of the variation in y that is explained by the variation in x.

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{SSE}{SS_y}$$

- R^2 takes on any value between zero and one.

  $R^2 = 1$: Perfect match between the line and the data points.

  $R^2 = 0$: There are no linear relationship between x and y.

- Example
  - Find the coefficient of determination for example 18.1; what does this statistic tell you about the model?
- Solution
  - Solving by hand;

$$R^2 = 1 - \frac{SSE}{SS_y} = 1 - \frac{2\,251\,363}{6\,434\,890} = .6501$$

  - Using the computer
    - out we have

Regression Statistics	
Multiple R	0.8063
R Square	0.6501
Adjusted R Square	0.6466
Standard Error	151.57
Observations	100

65% of the variation in the auction selling price is explained by the variation in odometer reading. The rest (35%) remains unexplained by this model.

Using the Regression Equation

- Before using the regression model, we need to assess how well it fits the data.
- If we are satisfied with how well the model fits the data, we can use it to make predictions for y.
- Illustration
  - Predict the selling price of a three-year-old Laser with 40,000 km on the odometer (Example 18.1).

$$\hat{y} = 6533 - .0312x = 6533 - .0312(40{,}000) = 5{,}285$$

- Prediction interval and confidence interval
  - Two intervals can be used to discover how closely the predicted value will match the true value of y.
    - Prediction interval - for a particular value of y,

– The prediction interval

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

– The confidence interval

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

The prediction interval is wider than the confidence interval

- Example: Interval estimates for the car auction price

  – Provide an interval estimate for the bidding price on a Ford Laser with 40,000 km on the odometer.

  – Solution

    - The dealer would like to predict the price of a single car

    - The prediction interval(95%) =

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

$$t_{.025,98}$$

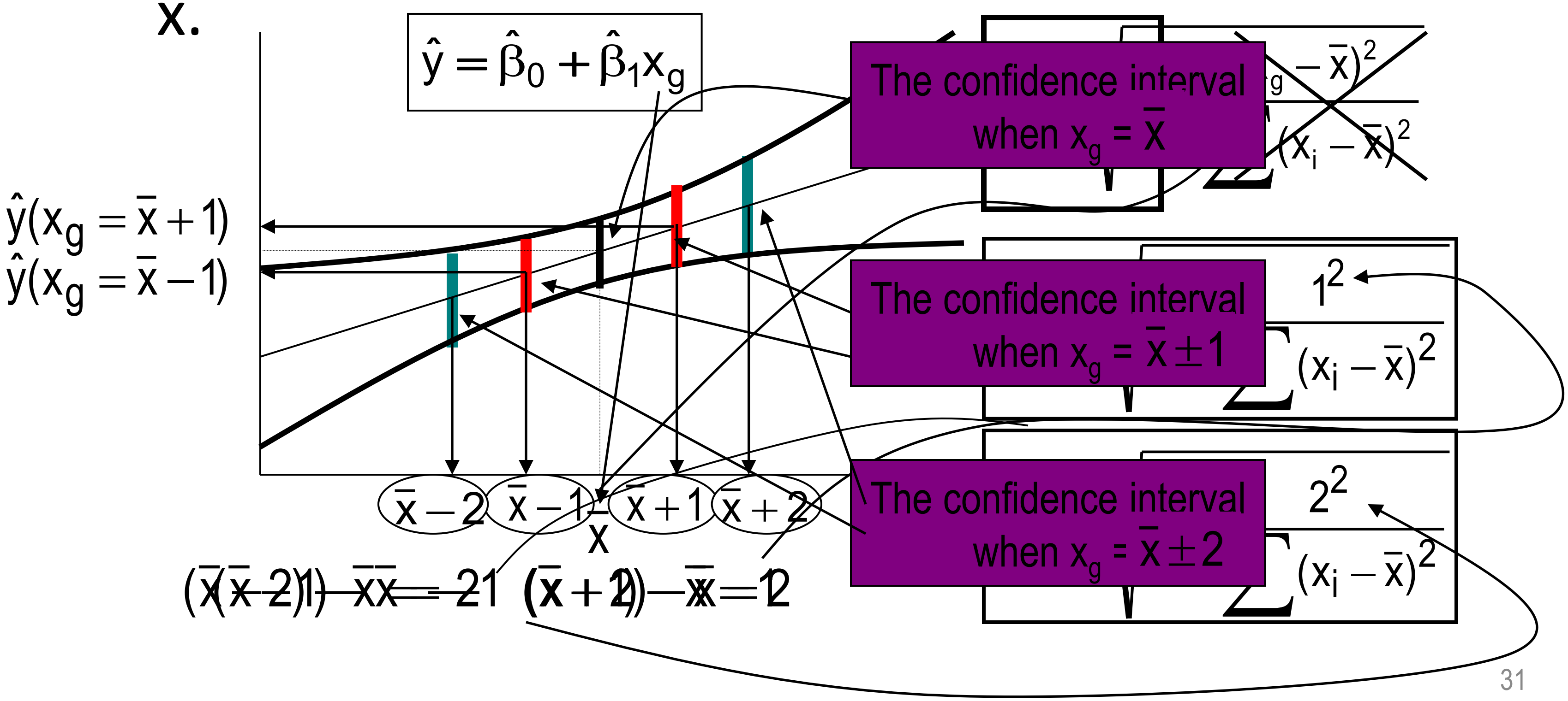
$$\overbrace{[6533 - .0312(40000)]} \pm 1.984(151.6) \sqrt{1 + \frac{1}{100} + \frac{(40{,}000 - 36{,}009)^2}{\sum 4{,}309{,}340{,}160}} = 5{,}285 \pm 303$$

– The car dealer wants to bid on a lot of 250 Ford Lasers, where each car has been driven for about 40,000 km.
– Solution
  - The dealer needs to estimate the mean price per car.
  - The confidence interval (95%) =

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

$$[6533 - .0312(40000)] \pm 1.984(151.6) \sqrt{\frac{1}{100} + \frac{(40{,}000 - 36{,}009)^2}{\sum 4{,}309{,}340{,}160}} = 5{,}285 \pm 35$$

- The effect of the given value of x on the interval
  - As x_g moves away from $\bar{x}$ the interval becomes longer. That is, the shortest interval is found at $\bar{x}$.



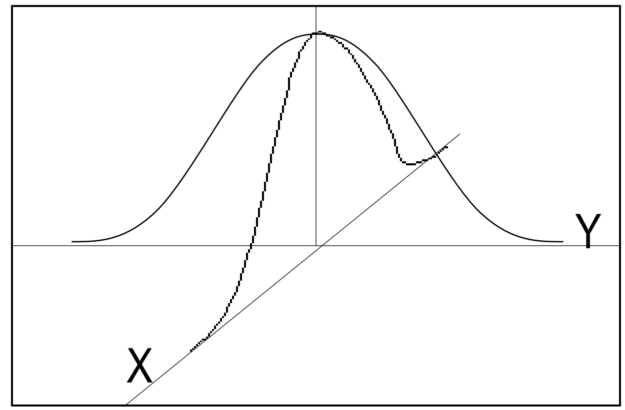
Coefficient of correlation

- The coefficient of correlation is used to measure the strength of a linear association between two variables.
- The coefficient values range between -1 and 1.
  - If ρ = -1 (perfect negative linear association) or ρ = +1 (perfect positive linear association) every point falls on the regression line.
  - If ρ = 0 there is no linear association.
- The coefficient can be used to test for linear relationship between two variables.

- Testing the coefficient of correlation
  - When there are no linear relationship between two variables, $\rho = 0$.
  - The hypotheses are:

    $H_0: \rho = 0$

    $H_A: \rho \neq 0$
  - The test statistic is:



$$\boxed{t = r\sqrt{\frac{n-2}{1-r^2}}}$$

The statistic is Student t distributed with d.f. = n - 2, provided the variables are bivariate normally distributed.

where r is the sample coefficient of correlation

calculated by $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}}$

- Example Testing for linear relationship
  - Test the coefficient of correlation to determine if linear relationship exists in the data of example 18.1.
- Solution
  - We test H_0: $\rho = 0$
    H_A: $\rho \neq 0$.
  - Solving by hand:
    - The rejection region is $|t| > t_{\alpha/2,n-2} = t_{.025,98} = 1.984$ or so.
    - The sample coefficient of correlation r= $SS_{xy} / \sqrt{SS_x SS_y}$ =-.806

The value of the t statistic is

$$t = r\sqrt{\frac{n-2}{1-r^2}} = -13.49$$

Conclusion: There is sufficient evidence at α = 5% to infer that there are linear relationship between the two variables.

Regression Diagnostics - I

- The three important conditions required for the validity of the regression analysis are:
  - the error variable is normally distributed.
  - the error variance is constant for all values of x.
  - The errors are independent of each other.
- How can we diagnose violations of these conditions?

- Residual Analysis
  - Examining the residuals (or standardized residuals), we can identify violations of the required conditions
  - Example 18.1 - continued
    - Nonnormality.
      - Use Excel to obtain the standardized residual histogram.
      - Examine the histogram and look for a bell shaped diagram with mean close to zero.

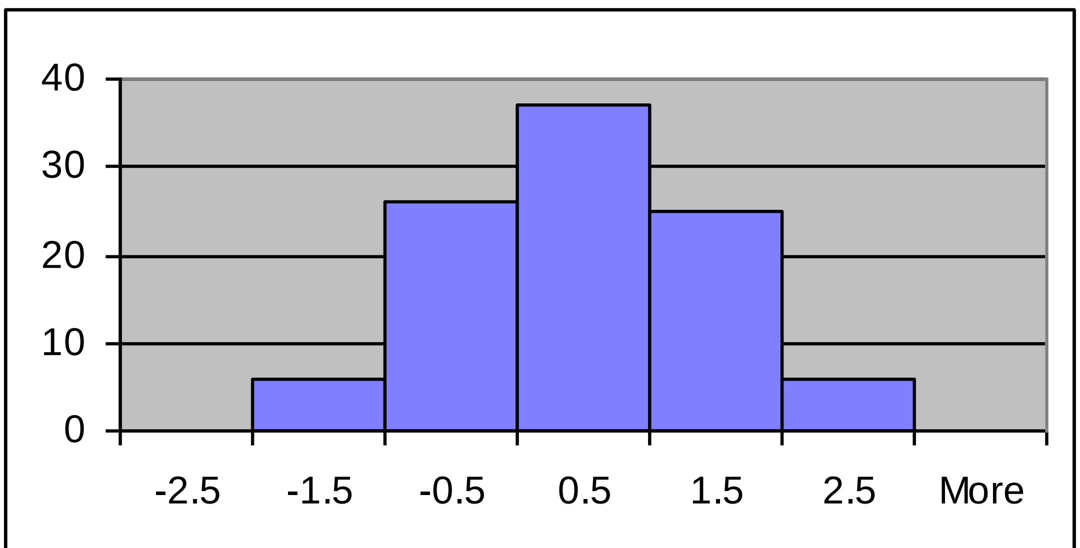
RESIDUAL OUTPUT		A Partial list of Standard residuals
Observation	*Residuals*	*Standard Residuals*
1	-50.45749927	-0.334595895
2	-77.82496482	-0.516076186
3	-97.33039568	-0.645421421
4	223.2070978	1.480140312
5	238.4730715	1.58137268

For each residual we calculate the standard deviation as follows:

$$s_{r_i} = s_\varepsilon \sqrt{1 - h_i} \quad \text{where}$$

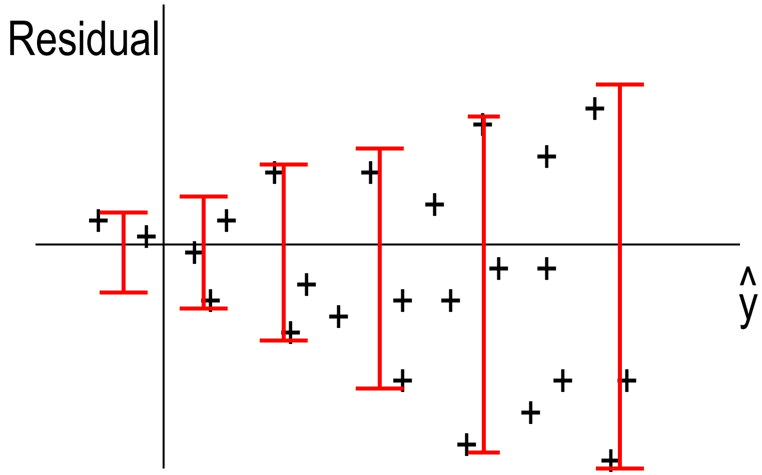
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}$$

Standardized residual i =
Residual i / Standard deviation

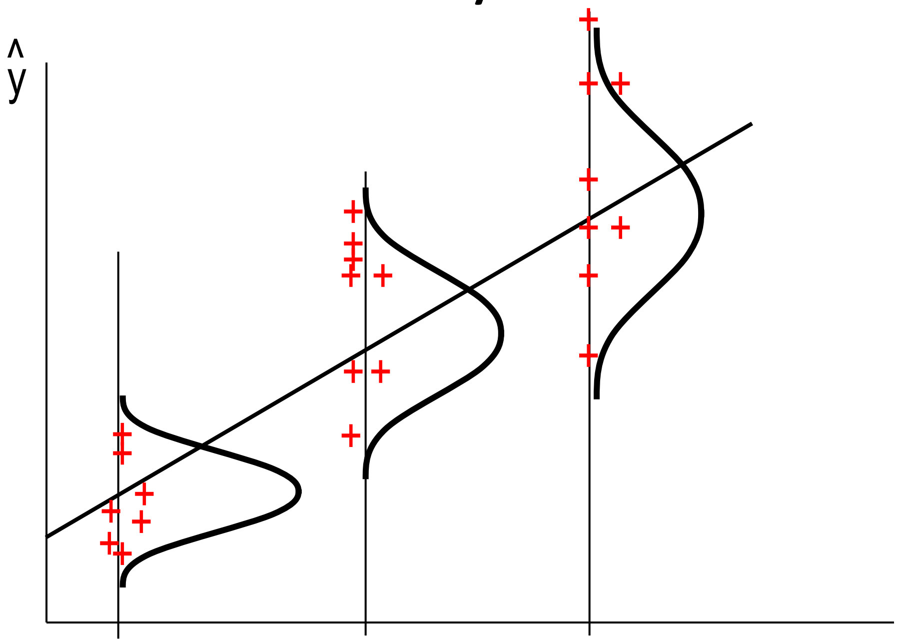


We can also apply the Lilliefors test or the χ^2 test of normality.

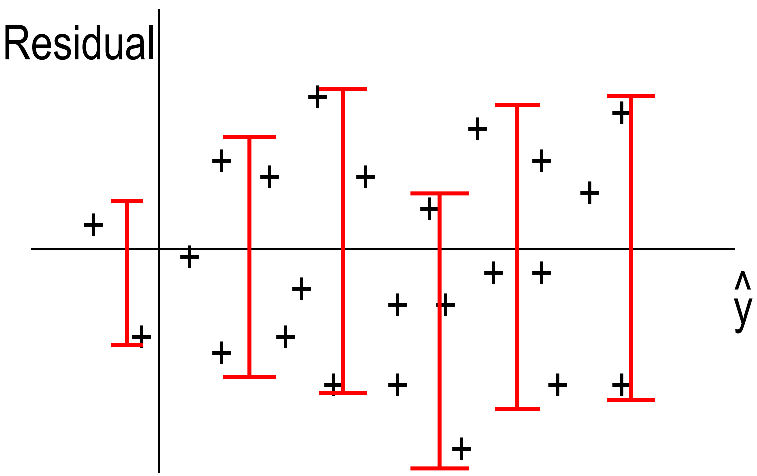
- Heteroscedasticity
  - When the requirement of a constant variance is violated we have heteroscedasticity.



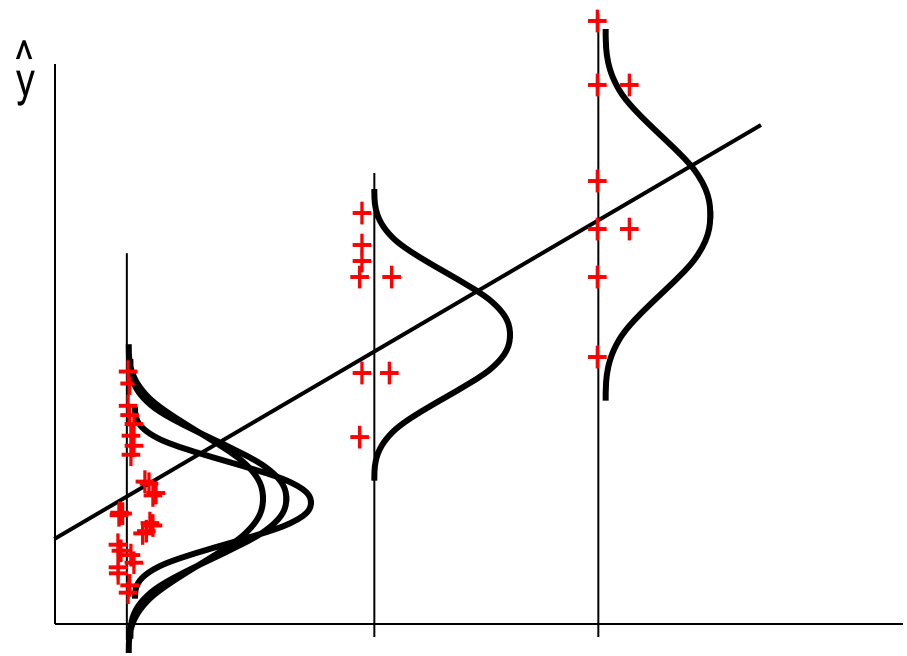
The spread increases with $\hat{y}$



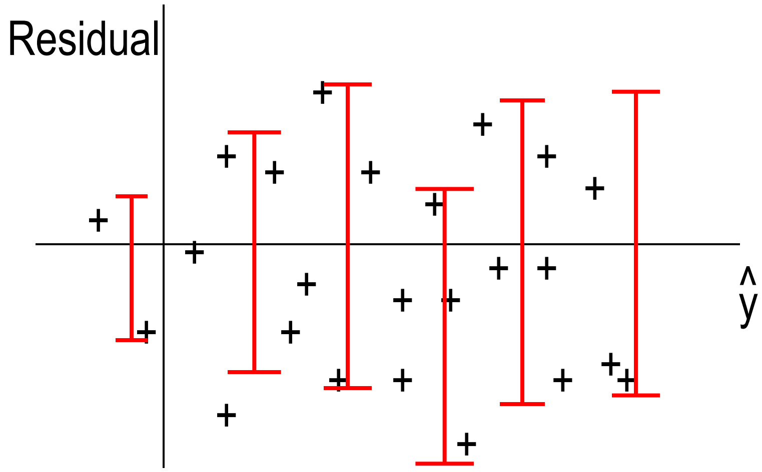
- When the requirement of a constant variance is not violated we have homoscedasticity.



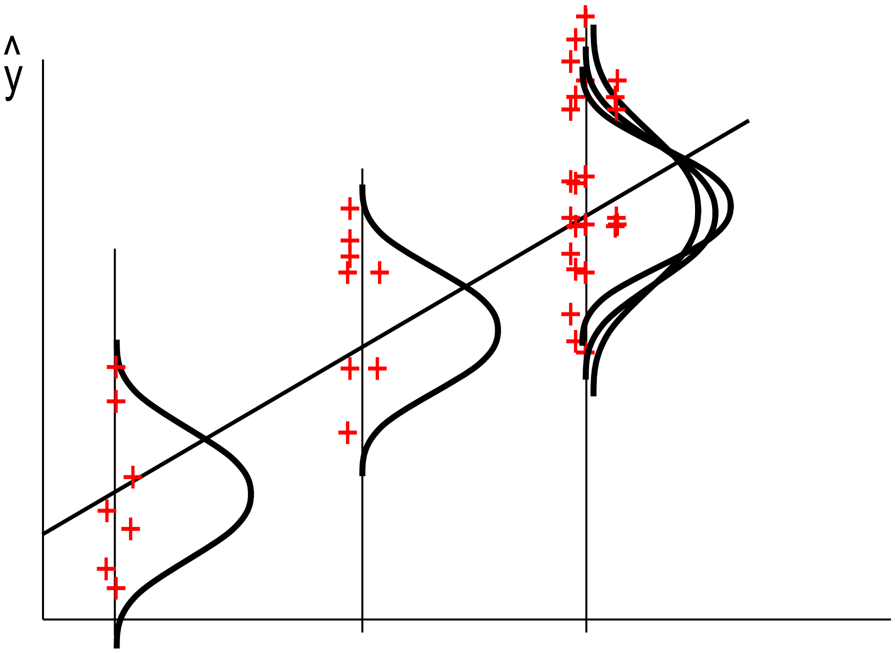
The spread of the data points does not change much.



- When the requirement of a constant variance is not violated we have homoscedasticity.

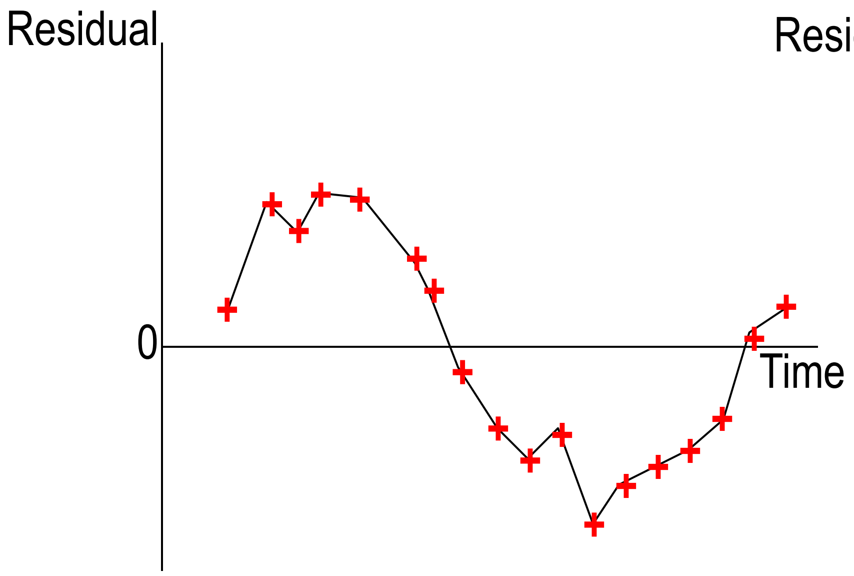


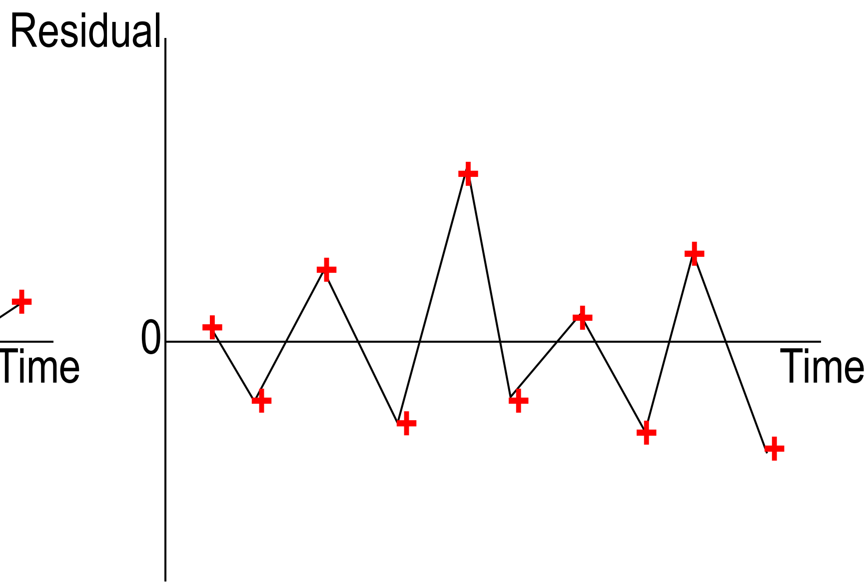
As far as the even spread, this is a much better situation



- Nonindependence of error variables
  - **A time series** is constituted if data were collected over time.
  - Examining the residuals over time, no pattern should be observed if the errors are independent.
  - When a pattern is detected, the errors are said to be autocorrelated.
  - Autocorrelation can be detected by graphing the residuals against time.

Patterns in the appearance of the residuals over time indicates that autocorrelation exists.

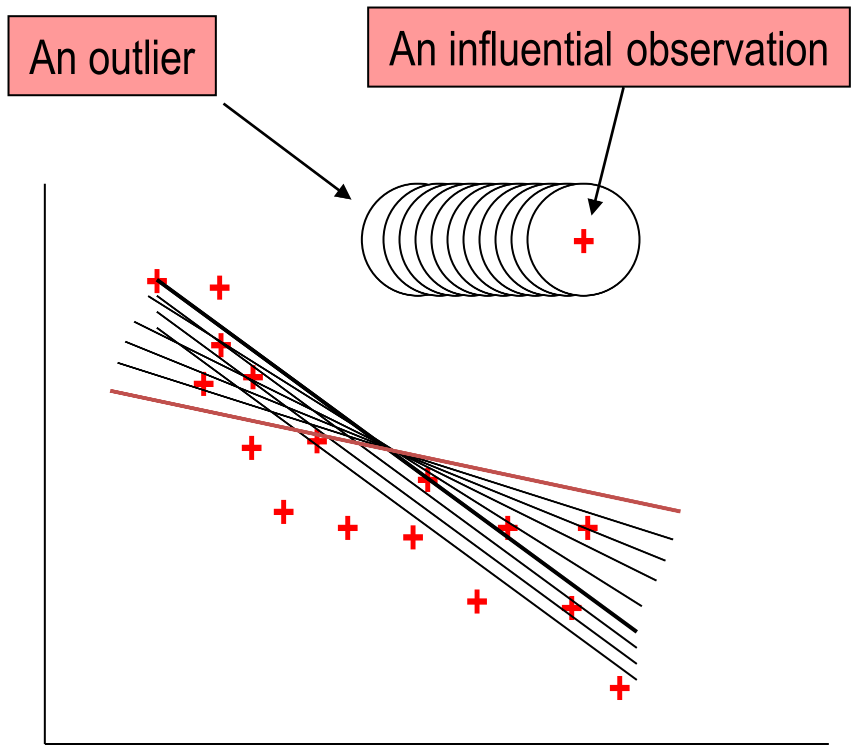




Note the runs of positive residuals, replaced by runs of negative residuals

Note the oscillating behavior of the residuals around zero.

- Outliers
  - An outlier is an observation that is unusually small or large.
  - Several possibilities need to be investigated when an outlier is observed:
    - There was an error in recording the value.
    - The point does not belong in the sample.
    - The observation is valid.
  - Identify outliers from the scatter diagram.
  - It is customary to suspect an observation is an outlier if its |standard residual| > 2



… but, some outliers may be very influential

The outlier causes a shift in the regression line

- Procedure for regression diagnostics
  - Develop a model that has a theoretical basis.
  - Gather data for the two variables in the model.
  - Draw the scatter diagram to determine whether a linear model appears to be appropriate.
  - Check the required conditions for the errors.
  - Assess the model fit.
  - If the model fits the data, use the regression equation.