DS342 - Data Analytics

Lecture 8
Regression Analysis: Estimating Relationships

Edited Version

Dr. Marwa Sabry

Introduction
(slide 1 of 2)

- **Regression analysis** is the study of relationships between variables.
- There are two potential objectives of regression analysis: to understand how the world operates and to make predictions.
- Two basic types of data are analyzed:
  - ***Cross-sectional*** *data* are usually data gathered from approximately the same period of time from a population.
  - ***Time series*** *data* involve one or more variables that are observed at several, usually equally spaced, points in time.
    - Time series variables are usually related to their own past values—a property called *autocorrelation*—which adds complications to the analysis.

Introduction

(slide 2 of 2)

- In every regression study, there is a single variable that we are trying to explain or predict, called the **dependent** variable.
  - It is also called the **response** variable or the **target** variable.
- To help explain or predict the dependent variable, we use one or more **explanatory** variables.
  - They are also called **independent** or **predictor** variables.
- If there is a single explanatory variable, the analysis is called **simple regression**.
- If there are several explanatory variables, it is called **multiple regression**.
- Regression can be *linear* (*straight-line* relationships) or *nonlinear* (curved relationships).
  - Many nonlinear relationships can be *linearized* mathematically.

Scatterplots:
Graphing Relationships

- Drawing scatterplots is a good way to begin regression analysis.
- A scatterplot is a graphical plot of two variables, an *X* and a *Y*.
- If there is any relationship between the two variables, it is usually apparent from the scatterplot.

Example 10.1: Drugstore Sales.xlsx (slide 1 of 2)

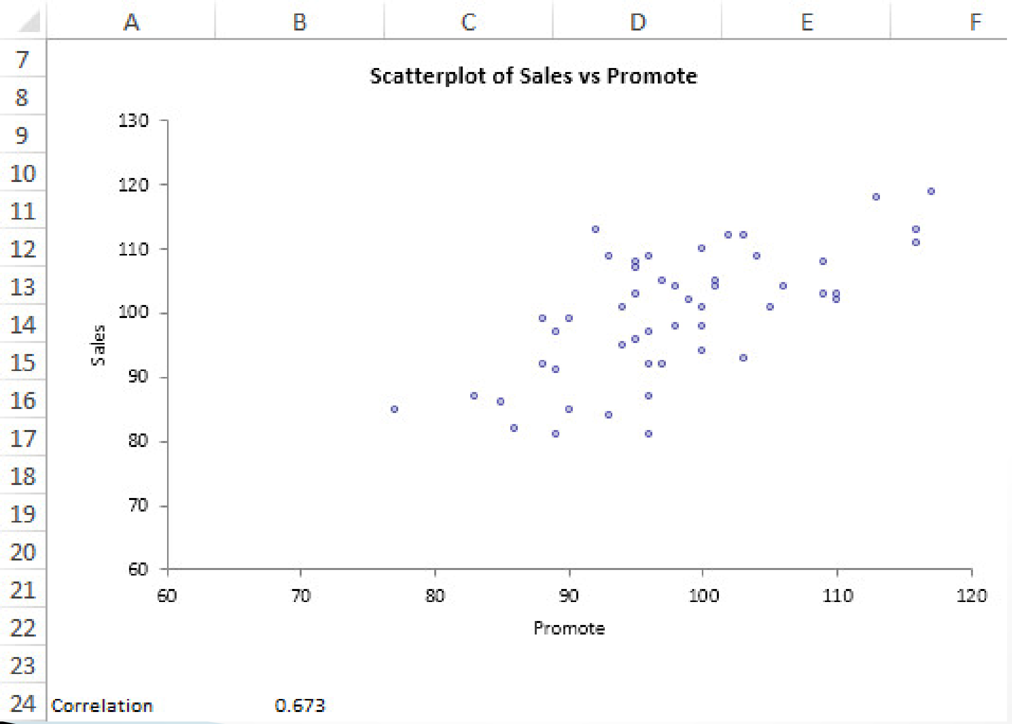
- **Objective**: To use a scatterplot to examine the relationship between promotional expenditures and sales at Pharmex.
- **Solution**: Pharmex has collected data from 50 randomly selected metropolitan regions.
- There are two variables: Pharmex's promotional expenditures as a percentage of those of the leading competitor ("Promote") and Pharmex's sales as a percentage of those of the leading competitor ("Sales").
- A partial listing of the data is shown below.

	A	B	C	D	E	F	G
1	Region	Promote	Sales				
2	1	77	85				
3	2	110	103				
4	3	110	102				
5	4	93	109				
6	5	90	85				
7	6	95	103				
50	49	95	108				
51	50	96	87				

Each value is a percentage of what the leading competitor did.

Example 10.1: Drugstore Sales.xlsx (slide 2 of 2)

- Use Excel's Chart Wizard Scatterplot to create a scatterplot.
  - Sales is on the vertical axis and Promote is on the horizontal axis because the store believes that large promotional expenditures tend to "cause" larger values of sales.



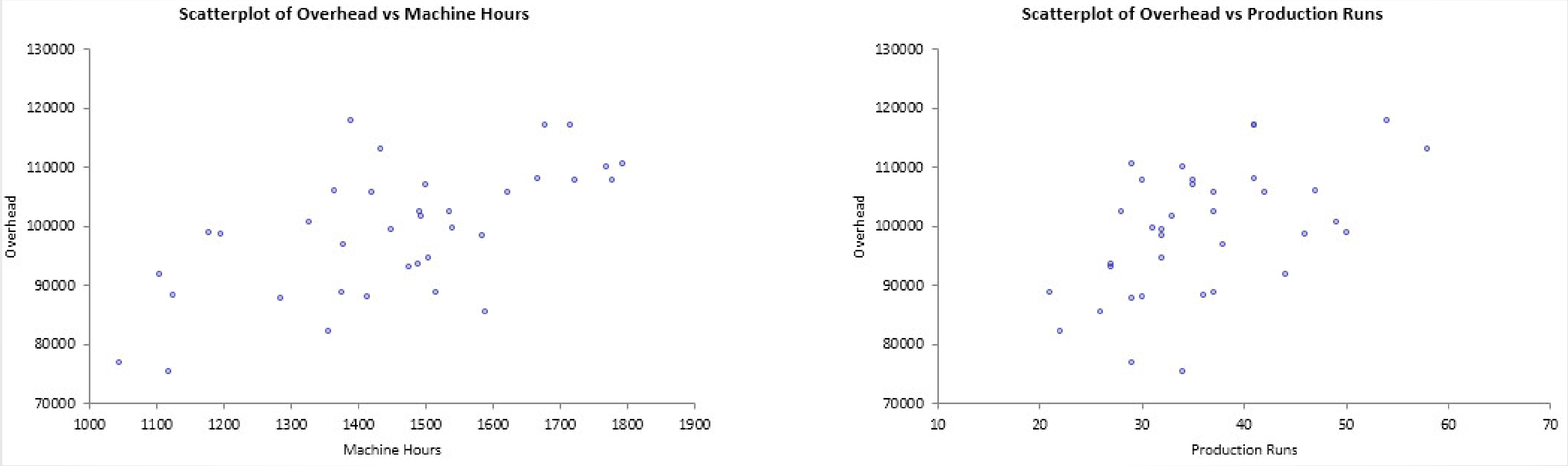
Example 10.2: Overhead Costs.xlsx (slide 1 of 3)

- **Objective**: To use scatterplots to examine the relationships among overhead, machine hours, and production runs at Bendrix.
- **Solution**: Data file contains observations of overhead costs, machine hours, and number of production runs at Bendrix.
- Each observation (row) corresponds to a single month.

	A	B	C	D
1	Month	Machine Hours	Production Runs	Overhead
2	1	1539	31	99798
3	2	1284	29	87804
4	3	1490	27	93681
5	4	1355	22	82262
6	5	1500	35	106968
34	33	1678	41	117183
35	34	1723	35	107828
36	35	1413	30	88032
37	36	1390	54	117943

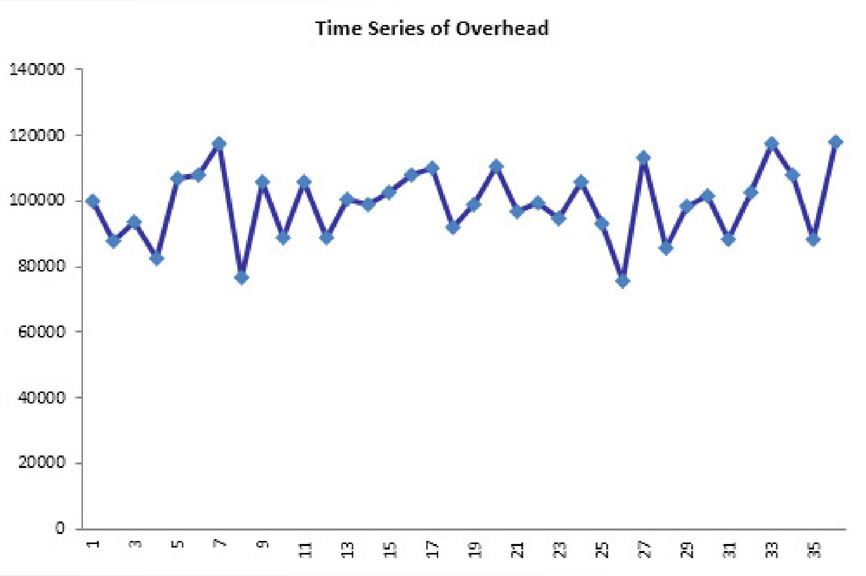
Example 10.2: Overhead Costs.xlsx (slide 2 of 3)

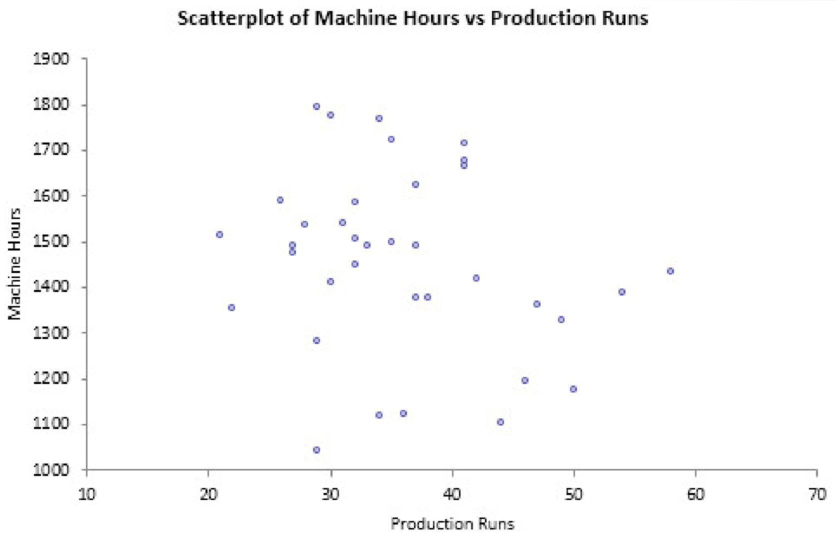
- Examine scatterplots between each explanatory variable (Machine Hours and Production Runs) and the dependent variable (Overhead).



Example 10.2: Overhead Costs.xlsx (slide 3 of 3)

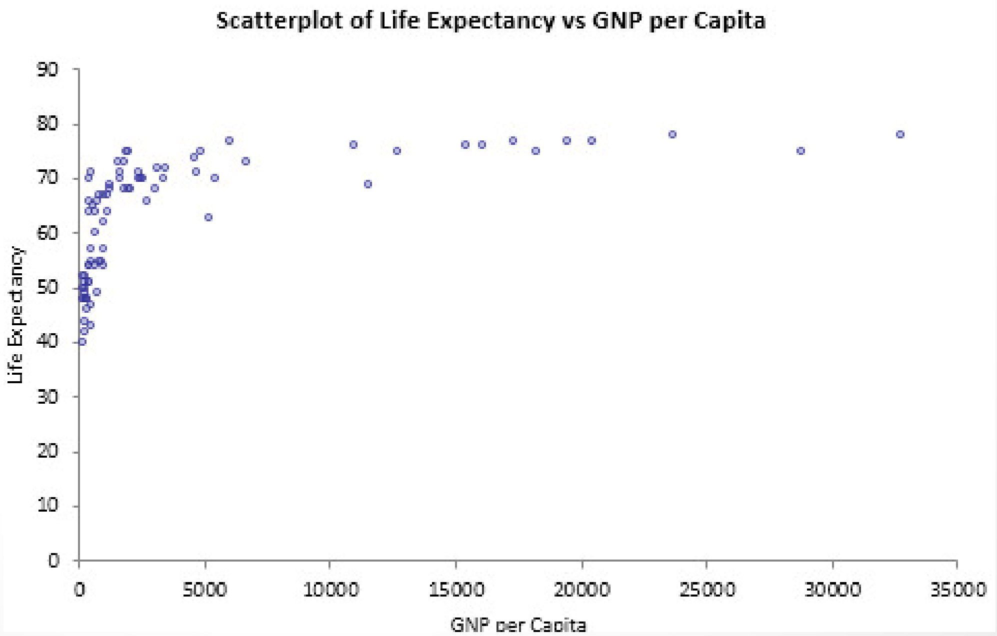
- Check for possible time series patterns, by creating a time series graph for any of the variables.
- Check for relationships among the multiple explanatory variables (Machine Hours versus Production Runs).





Linear versus Nonlinear Relationships

- Scatterplots are useful for detecting relationships that may not be obvious otherwise.
- The typical relationship you hope to see is a straight-line, or *linear*, relationship.
  - This doesn't mean that all points lie on a straight line, but that the points tend to cluster around a straight line.
- The scatterplot below illustrates a relationship that is clearly *nonlinear*.

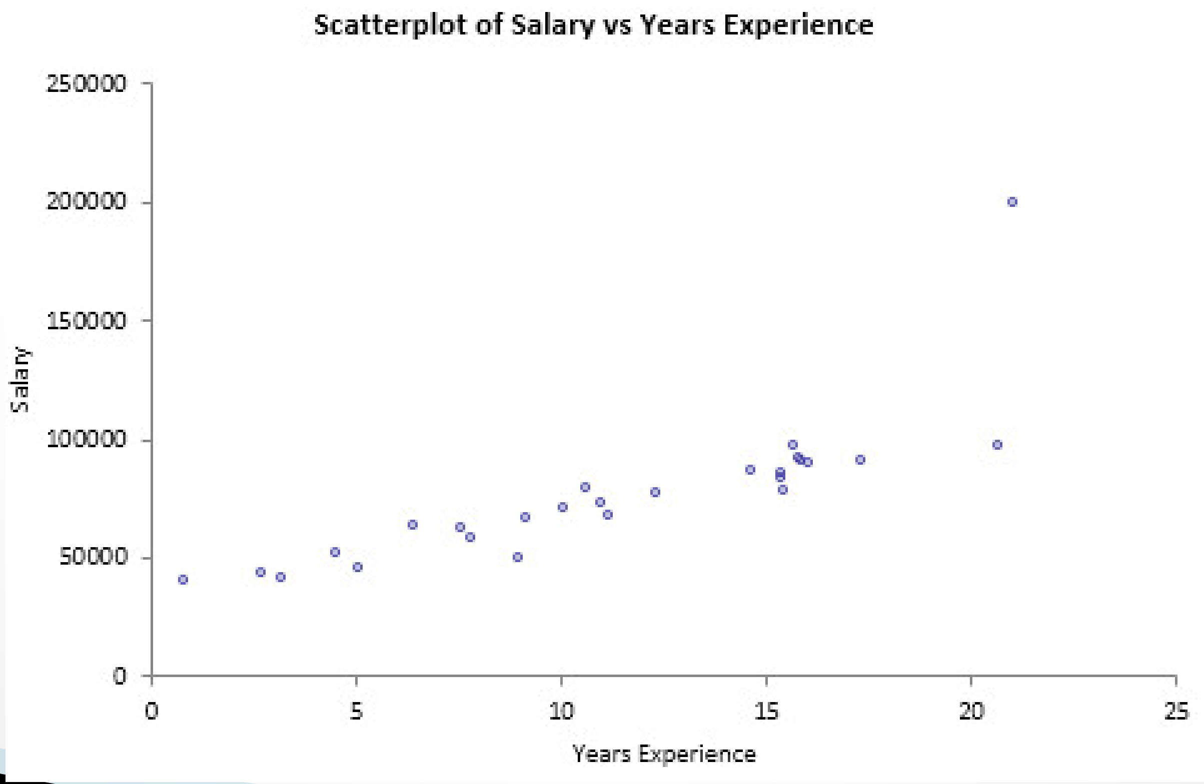


Outliers
(slide 1 of 2)

- Scatterplots are especially useful for identifying **outliers**—observations that fall outside of the general pattern of the rest of the observations.
  - If an outlier is clearly not a member of the population of interest, then it is probably best to delete it from the analysis.
  - If it isn't clear whether outliers are members of the relevant population, run the regression analysis with them and again without them.
    - If the results are practically the same in both cases, then it is probably best to report the results with the outliers included.
    - Otherwise, you can report both sets of results with a verbal explanation of the outliers.

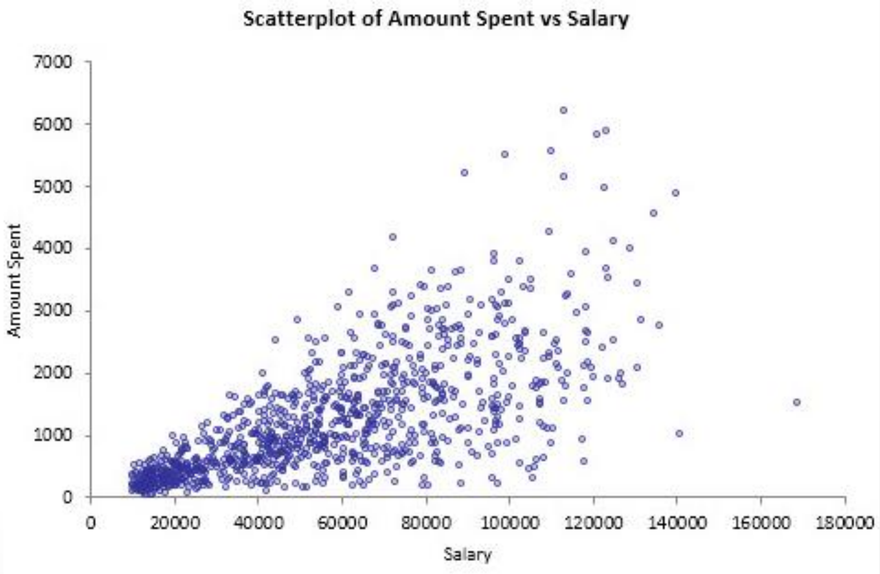
Outliers
(slide 2 of 2)

- In the figure below, the outlier (the point at the top right) is the company CEO, whose salary is well above that of all of the other employees.



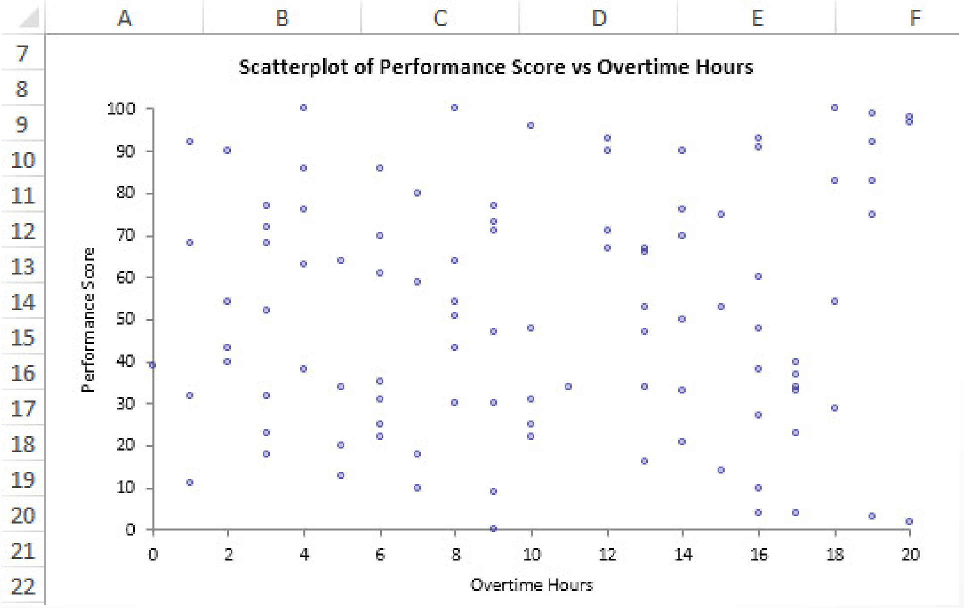
Unequal Variance

- Occasionally, the variance of the dependent variable depends on the value of the explanatory variable.
- The figure below illustrates an example of this.
  - There is a clear upward relationship, but the variability of amount spent increases as salary increases—which is evident from the *fan* shape.
- This unequal variance violates one of the assumptions of linear regression analysis, but there are ways to deal with it.



No Relationship

- A scatterplot can also indicate that there is *no* relationship between a pair of variables.
  - This is usually the case when the scatterplot appears as a shapeless swarm of points.



Correlations: Indicators of Linear Relationships (slide 1 of 2)

- **Correlations** are numerical summary measures that indicate the strength of linear relationships between pairs of variables.
  - A correlation between a pair of variables is a single number that summarizes the information in a scatterplot.
  - It measures the strength of *linear* relationships only.
  - The usual notation for a correlation between variables X and Y is r_{xy}.

Correlations: Indicators of Linear Relationships (slide 2 of 2)

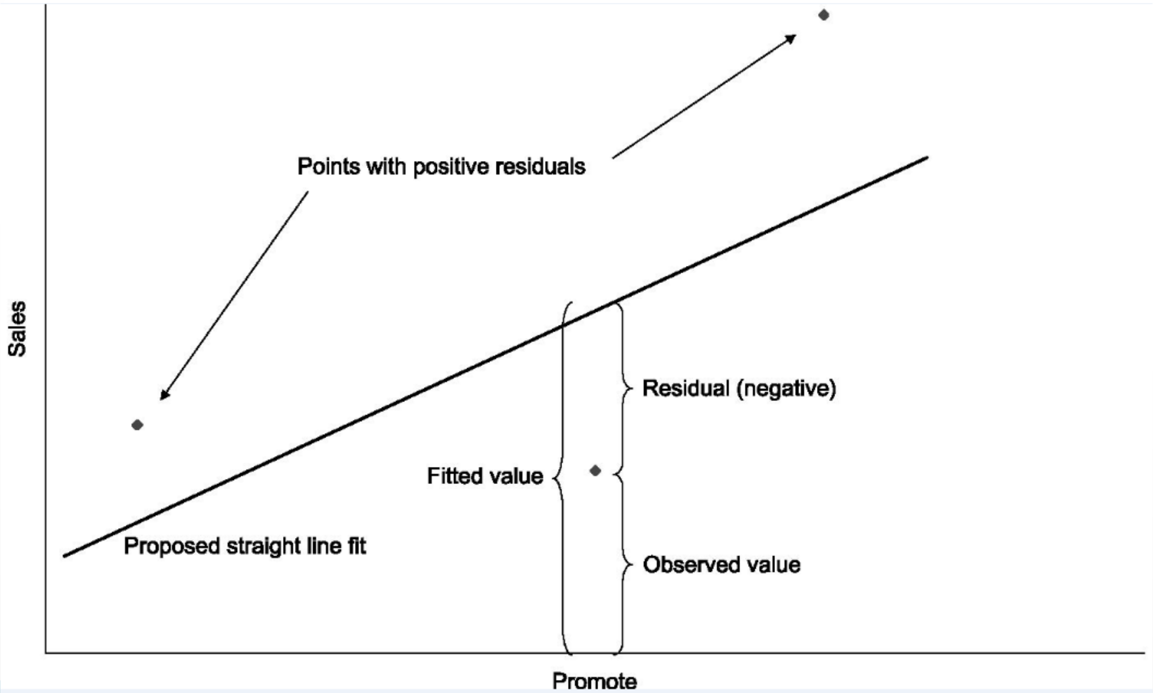
- By looking at the sign of the correlation—plus or minus—you can tell whether the two variables are positively or negatively related.
- Correlations are completely unaffected by the units of measurement.
  - A correlation equal to 0 or near 0 indicates practically **no linear relationship**.
  - A correlation with magnitude close to 1 indicates a **strong linear relationship**.
  - A correlation equal to -1 (negative correlation) or +1 (positive correlation) occurs only when the linear relationship between the two variables is perfect.
- Be careful when interpreting correlations—they are relevant descriptors only for ***linear*** relationships.

Simple Linear Regression

- Scatterplots and correlations indicate linear relationships and the strengths of these relationships, but they do not *quantify* them.
- Simple linear regression quantifies the relationship where there is a *single* explanatory variable.
- A straight line is fitted through the scatterplot of the dependent variable *Y* versus the explanatory variable *X.*

Least Squares Estimation

- When fitting a straight line through a scatterplot, choose the line that makes the vertical distance from the points to the line as small as possible.
- A **fitted value** is the predicted value of the dependent variable.
  - Graphically, it is the height of the line above a given explanatory value.



Least Squares Estimation

- True values for the slope and intercept are not known so they are estimated using sample data

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

where

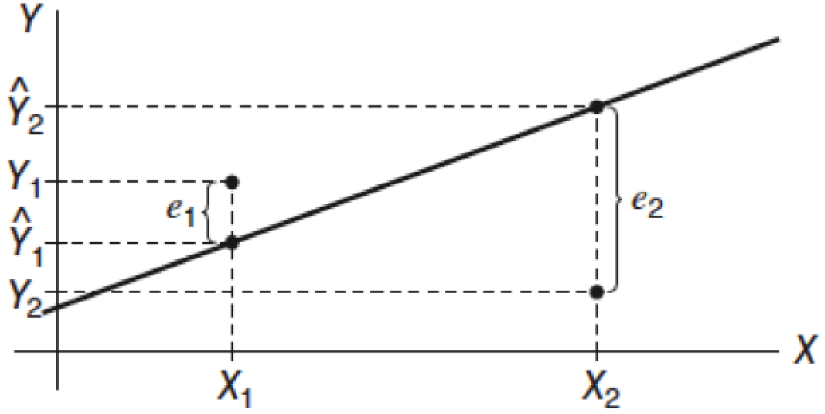
Y = dependent variable (response)
X = independent variable (predictor or explanatory)
b_0 = intercept (value of Y when X = 0)
b_1 = slope of the regression line

- The best-fitting line minimizes the sum of squares of the residuals.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - [b_0 + b_1 X_i])^2$$

$$b_1 = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$



Errors associated with individual observations

Least Squares Estimation

- The **residual** is the difference between the actual and fitted values of the dependent variable.
- Fundamental Equation for Regression:

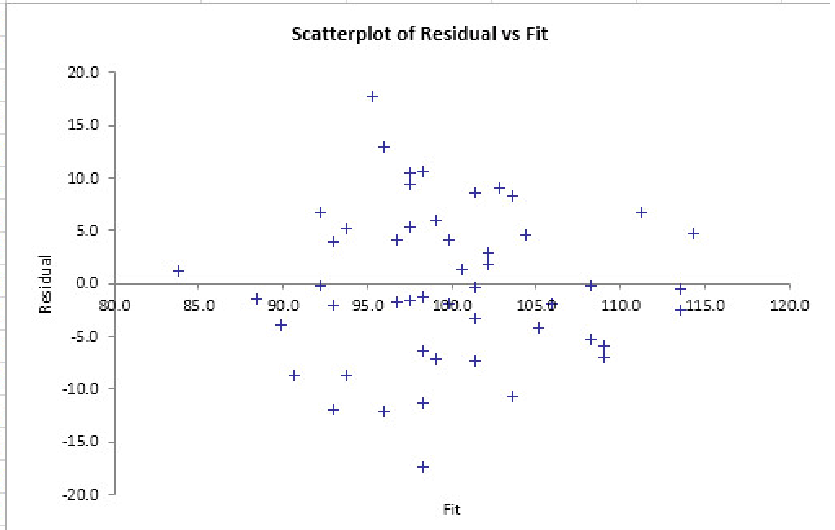
  Observed Value = Fitted Value + Residual
- The best-fitting line through the points of a scatterplot is the line with the *smallest sum of squared residuals*.
  - This is called the **least squares line**.
  - It is the line quoted in regression outputs.
- The least squares line is specified completely by its slope and intercept.
  - Equation for Slope in Simple Linear Regression:
  - Equation for Intercept in Simple Linear Regression:

Example 10.1 (continued): Drugstore Sales.xlsx (slide 1 of 2)

- **Objective**: To use Excel Regression Data Analysis procedure to find the least squares line for sales as a function of promotional expenses at Pharmex.
- **Solution**: Select Regression from the Data Analysis tools.
- Use Sales as the dependent variable and Promote as the explanatory variable.
- The regression output is shown below and on the next slide.

	A	B	C	D	E	F	G
7	*Multiple Regression for Sales*						
8		Multiple	R-Square	Adjusted	StErr of		
9	*Summary*	R		R-Square	Estimate		
10		0.6730	0.4529	0.4415	7.394732934		
11							
12		Degrees of	Sum of	Mean of	F-Ratio	p-Value	
13	*ANOVA Table*	Freedom	Squares	Squares			
14	Explained	1	2172.880392	2172.880392	39.7366	< 0.0001	
15	Unexplained	48	2624.739608	54.68207516			
16							
17		Coefficient	Standard	t-Value	p-Value	Confidence Interval 95%	
18	*Regression Table*		Error			Lower	Upper
19	Constant	25.12642006	11.8825852	2.1146	0.0397	1.234881256	49.01795886
20	Promote	0.762296485	0.120928454	6.3037	< 0.0001	0.519153532	1.005439438

Example 10.1 (continued): Drugstore Sales.xlsx (slide 2 of 2)



Graph Data	Sales	Fit	Residual
1	85	83.8232494	1.176750604
2	103	108.9790334	-5.979033397
3	102	108.9790334	-6.979033397

- The equation for the least squares line is:

  Predicted Sales = 25.1264 + 0.7623***Promote**

Regression Statistics

- **Multiple R:** | *r* |, where *r* is the sample correlation coefficient. The value of *r* varies from -1 to +1 (*r* is negative if slope is negative)
- **R Square:** coefficient of determination, R^2, which varies from 0 (no fit) to 1 (perfect fit)
- **Adjusted R Square:** adjusts R^2 for sample size and number of *X* variables
- **Standard Error:** variability between observed and predicted *Y* values. This is formally called the **standard error of the estimate**, S_{YX}.

Example 10.2 (continued): Overhead Costs.xlsx (slide 1 of 2)

- **Objective**: To use the Excel Regression Data Analysis procedure to regress overhead expenses at Bendrix against machine hours and then against production runs.
- **Solution**: The Bendrix manufacturing data set has two potential explanatory variables, Machine Hours and Production Runs.
- The regression output for Overhead with Machine Hours as the single explanatory variable is shown below.

	A	B	C	D	E	F	G
7	*Multiple Regression for Overhead*						
8		Multiple	R-Square	Adjusted	StErr of		
9	*Summary*	R		R-Square	Estimate		
10		0.6319	0.3993	0.3816	8584.739353		
11							
12		Degrees of	Sum of	Mean of	F-Ratio	p-Value	
13	*ANOVA Table*	Freedom	Squares	Squares			
14	Explained	1	1665463368	1665463368	22.5986	< 0.0001	
15	Unexplained	34	2505723492	73697749.75			
16							
17		Coefficient	Standard	t-Value	p-Value	Confidence Interval 95%	
18	*Regression Table*		Error			Lower	Upper
19	Constant	48621.35463	10725.3327	4.5333	< 0.0001	26824.85615	70417.85312
20	Machine Hours	34.70223642	7.299902097	4.7538	< 0.0001	19.86705047	49.53742238

Example 10.2 (continued): Overhead Costs.xlsx (slide 2 of 2)

- The output when Production Runs is the only explanatory variable is shown below.

	A	B	C	D	E	F	G
7	*Multiple Regression for Overhead*						
8		Multiple	R-Square	Adjusted	StErr of		
9	*Summary*	R		R-Square	Estimate		
10		0.5205	0.2710	0.2495	9457.239463		
11							
12		Degrees of	Sum of	Mean of	F-Ratio	p-Value	
13	*ANOVA Table*	Freedom	Squares	Squares			
14	Explained	1	1130247999	1130247999	12.6370	0.0011	
15	Unexplained	34	3040938861	89439378.26			
16							
17		Coefficient	Standard	t-Value	p-Value	Confidence Interval 95%	
18	*Regression Table*		Error			Lower	Upper
19	Constant	75605.51571	6808.610629	11.1044	< 0.0001	61768.75415	89442.27728
20	Production Runs	655.0706602	184.2746779	3.5549	0.0011	280.5794579	1029.561862

- The two least squares lines are therefore:

  Predicted Overhead = 48621 + 34.7*MachineHours

  Predicted Overhead = 75606 + 655.1*ProductionRuns

Standard Error of Estimate

- The magnitude of the residuals provide a good indication of how useful the regression line is for predicting *Y* values from *X* values.
- Because there are numerous residuals, it is useful to summarize them with a single numerical measure.
  - This measure is called the **standard error of estimate** and is denoted s_e.
  - It is essentially **the standard deviation of the residuals**.

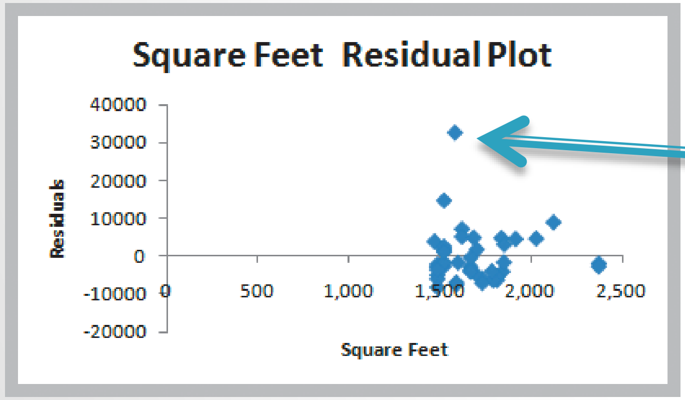
- The usual empirical rules for standard deviation can be applied to the standard error of estimate.
- In general, the standard error of estimate indicates the level of accuracy of predictions made from the regression equation.
  - The smaller it is, the more accurate predictions tend to be.

The Percentage of Variation Explained: R-Square

- R^2 is an important measure of the goodness of fit of the least squares line.
  - It is the percentage of variation of the dependent variable explained by the regression.
  - It always ranges between 0 and 1.
  - The better the linear fit is, the closer R^2 is to 1.
  - In simple linear regression, R^2 is the square of the correlation between the dependent variable and the explanatory variable.

Residual Analysis and Regression Assumptions

- **Residual** (error) = Actual Y value − Predicted Y value
- **Standardized residual** = residual / standard deviation
- **Rule of thumb:** Standardized residuals outside of ± 2 or ± 3 are potential outliers.



This point has a standard residual of 4.53

Multiple Regression

- To obtain improved fits in regression, several explanatory variables could be included in the regression equation. This is the realm of *multiple* regression*.*
  - Graphically, you are no longer fitting a *line* to a set of points. If there are two explanatory variables, you are fitting a *plane* to the data in three-dimensional space.
  - The regression equation is still estimated by the least squares method, but it is not practical to do this by hand.
  - There is a slope term for each explanatory variable in the equation, but the interpretation of these terms is different.
  - The standard error of estimate and R^2 summary measures are almost exactly as in simple regression.
  - Many *types* of explanatory variables can be included in the regression equation.

Interpretation of Regression Coefficients

- If Y is the dependent variable, and X_1 through X_k are the explanatory variables, then a typical multiple regression equation has the form shown below, where a is the Y-intercept, and b_1 through b_k are the slopes.
- General Multiple Regression Equation:

$$\text{Predicted } Y = a + b_1X_1 + b_2X_2 + \dots + b_kX_k$$

- Collectively, *a* the *b*s in the equation are called the **regression coefficients**.
- Each slope coefficient is the expected change in Y when this particular X increases by one unit *and the other Xs in the equation remain constant.*
  - This means that the estimates of the *b*s depend on which other *X*s are included in the regression equation.

Example 10.2 (continued): Overhead Costs.xlsx

- **Objective**: To use StatTools's Regression procedure to estimate the equation for overhead costs at Bendrix as a function of machine hours and production runs.
- **Solution**: Select Regression from the StatTools Regression and Classification dropdown list. Then choose the Multiple option and specify the single *D* variable and the two *I* variables.
- The coefficients in the output below indicate that the estimated regression equation is: Predicted Overhead = 3997 + 43.54Machine Hours + 883.62Production Runs.

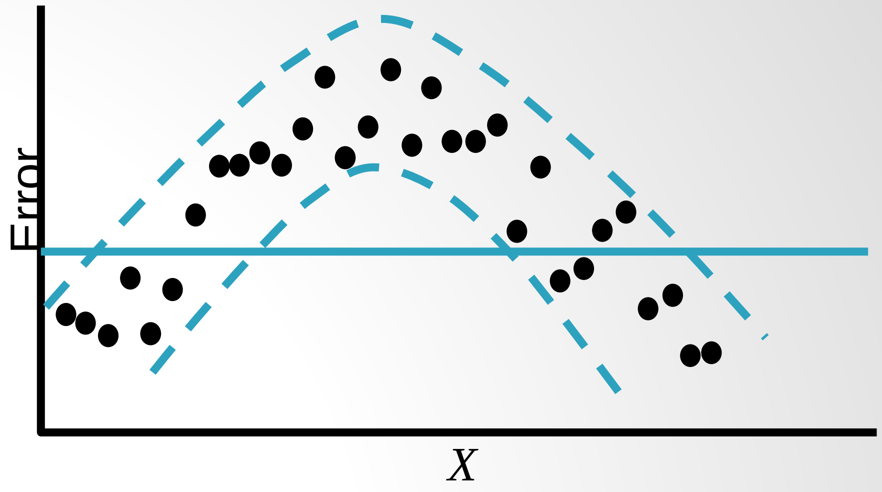
	A	B	C	D	E	F	G
7	*Multiple Regression for Overhead*						
8		Multiple	R-Square	Adjusted	StErr of		
9	*Summary*	R		R-Square	Estimate		
10		0.9308	0.8664	0.8583	4108.99309		
11							
12		Degrees of	Sum of	Mean of	F-Ratio	p-Value	
13	*ANOVA Table*	Freedom	Squares	Squares			
14	Explained	2	3614020661	1807010330	107.0261	< 0.0001	
15	Unexplained	33	557166199.1	16883824.22			
16							
17		Coefficient	Standard	t-Value	p-Value	Confidence Interval 95%	
18	*Regression Table*		Error			Lower	Upper
19	Constant	3996.678209	6603.650932	0.6052	0.5492	-9438.550632	17431.90705
20	Machine Hours	43.53639812	3.5894837	12.1289	< 0.0001	36.23353862	50.83925761
21	Production Runs	883.6179252	82.25140753	10.7429	< 0.0001	716.2761784	1050.959672

Checking Assumptions

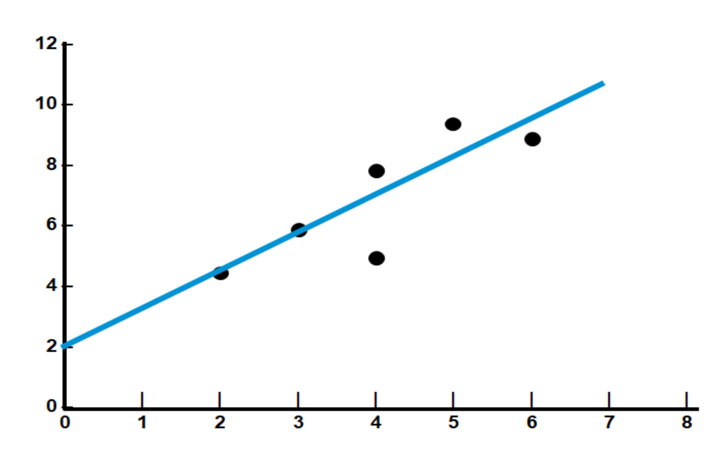
- ***Linearity***
- ***Normality of Errors***
- ***Homoscedasticity:*** variation about the regression line is constant
- ***Independence of Errors*****:** successive observations should not be related.

Checking Assumptions

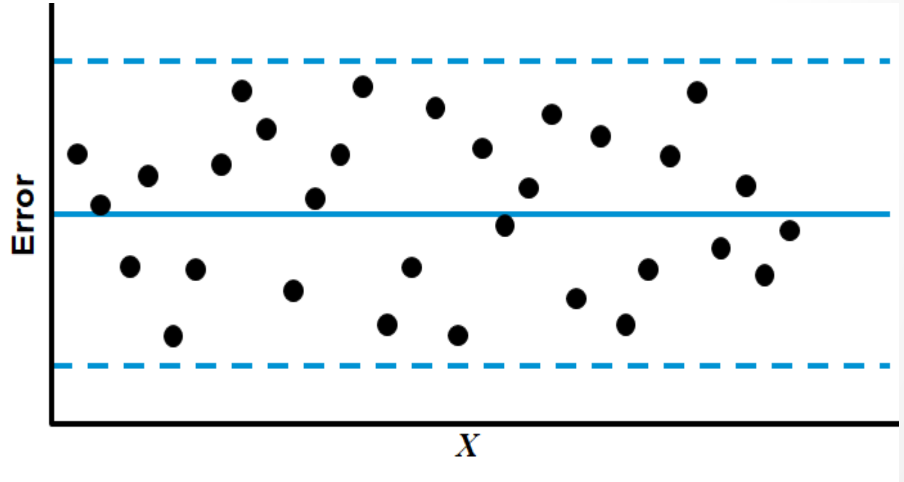
- ***Linearity:***
  - ✓ examine scatter diagram (should appear linear)
  - ✓ examine residual plot (should appear random)



- **Nonlinear relationship**



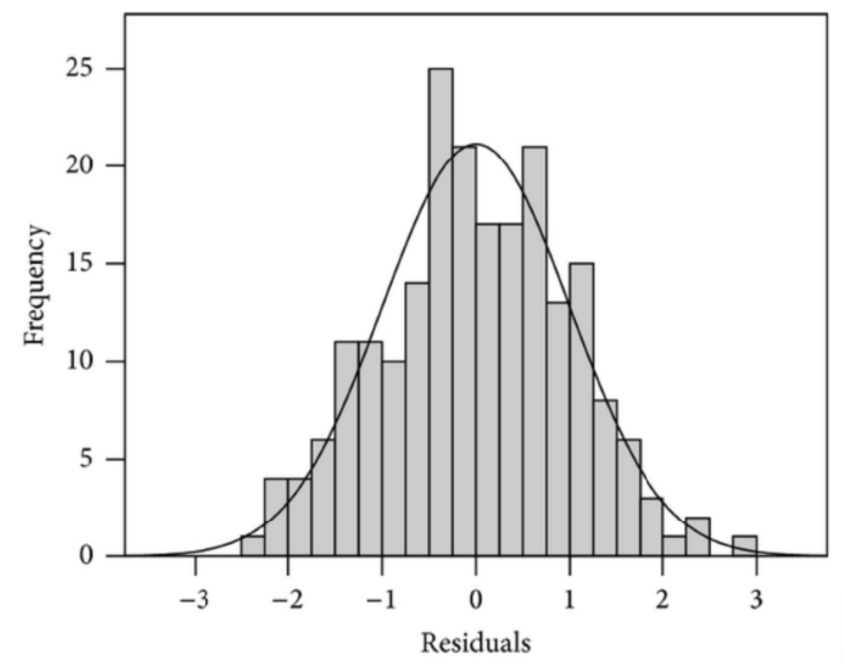
Linear relationship between *X* and *Y*



A random plot of residuals, errors seem random and no discernible pattern is present

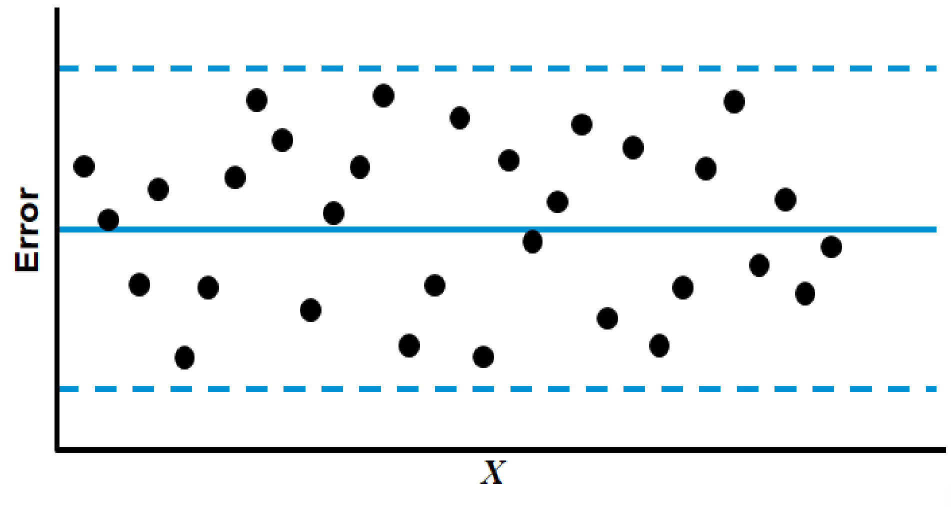
Checking Assumptions

- ***Normality of Errors:***
  - ✓ view a histogram of standardized residuals (normal errors with mean 0)
    - i.e., regression is robust to departures from normality



Checking Assumptions

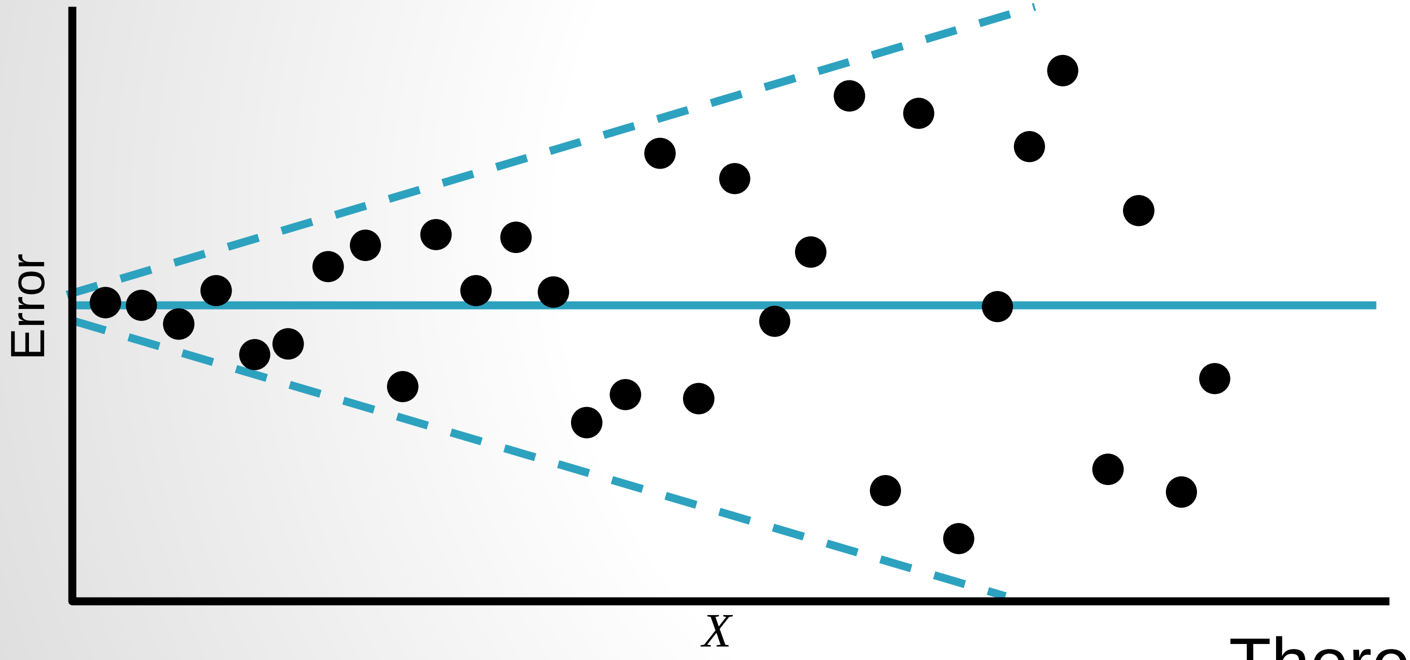
- ***Homoscedasticity:*** variation about the regression line is constant
  - ✓ examine the residual plot, residual plot shows no serious difference in the spread of the data for different *X* values.



A random plot of residuals, errors seem random and no discernible pattern is present

Residual Plots

- Nonconstant error variance, the errors increase as X increases!



Heteroscedasticity

There are two ways to deal with it:

- Use a different estimation method than least squares, called *weighted least squares*.
- Use a logarithmic transformation of the dependent variable.

Checking Assumptions

- ***Independence of Errors*: successive observations should not be related.**
  - ✓ This is important when the independent variable is time.
  - ✓ This assumption means that information on some of the errors provides no information on the values of the other errors.

1. For cross-sectional data, this assumption is usually taken for granted.
2. For time-series data, this assumption is often violated.
   - This is because of a property called ***autocorrelation***.
   - The **Durbin-Watson (DW) statistic** is one measure of autocorrelation. It always have a value ranging between 0 and 4.
   - ✓ A value of 2.0 indicates there is no autocorrelation.
   - ✓ Values from 0 to less than 2 point to positive autocorrelation.
   - ✓ values from 2 to 4 means negative autocorrelation.

Multicollinearity

- One other assumption is important for numerical calculations: No explanatory variable can be an *exact* linear combination of any other explanatory variables.
  - The violation occurs if one of the explanatory variables can be written as a weighted sum of several of the others.
    - This is called *exact* ***multicollinearity****.*
    - If it exists, there is *redundancy* in the data.
  - A more common and serious problem is *multicollinearity*, where explanatory variables are highly correlated.

Multicollinearity

- You will get two extra columns in the Regression table section of the regression output: **VIF** (variance inflation factor) and **R-Square**.
  - ✓ The R-Square for any X variable is the usual R-square value from a regression with that X as the dependent variable and the other X's as the explanatory variables. It indicates how related that X is to the other X's.
  - ✓ VIF is considered large when > 10. However a cutoff of 5 is commonly used. A value between 1 and 5 is moderate correlation. A value of 1 indicates no correlation.

Multicollinearity

Figure 11.7 Regression Output with Multicollinearity Diagnostics

	A	B	C	D	E	F	G	H	I
1	*Multiple Regression for Salary*	Multiple	R-Square	Adjusted	Std. Err. of	Rows	Outliers		
2	*Summary*	R		R-square	Estimate	Ignored			
3		0.9755	0.9516	0.9509	4964.209	0	1		
4									
5		Degrees of	Sum of	Mean of	F	p-Value			
6	*ANOVA Table*	Freedom	Squares	Squares					
7	Explained	4	1.42789E+11	35697211405	1448.552	< 0.0001			
8	Unexplained	295	7269795846	24643375.75					
9									
10		Coefficient	Standard	t-Value	p-Value	Confidence Interval 95%		Multicollinearity Checking	
11	*Regression Table*		Error			Lower	Upper	VIF	R-Square
12	Constant	49714.255	3815.818	13.028	<0.0001	42204.580	57223.930		
13	Gender (Female)	−2967.637	616.555	−4.813	<0.0001	−4181.041	−1754.232	1.001	0.001
14	Age	−309.744	148.123	−2.091	0.0374	−601.255	−18.233	31.504	0.968
15	Experience	388.641	190.562	2.039	0.0423	13.608	763.675	38.153	0.974
16	Seniority	2997.964	94.860	31.604	<0.0001	2811.276	3184.652	5.595	0.821
17									
18									
19	*Correlation Matrix*	Salary	Gender (Female)	Age	Experience	Seniority			
20	Salary	1.000	−0.093	0.857	0.882	0.973			
21	Gender (Female)	−0.093	1.000	−0.028	−0.026	−0.033			
22	Age	0.857	−0.028	1.000	0.984	0.884			
23	Experience	0.882	−0.026	0.984	1.000	0.905			
24	Seniority	0.973	−0.033	0.884	0.905	1.000			

Measuring the Fit of the Regression Model

- Regression models can be developed for any variables X and Y
- How do we know the model is actually helpful in predicting Y based on X?
- Three measures of variability are
  - SST – Total variability about the mean
  - SSE – Variability about the regression line
  - SSR – Total variability that is explained by the regression model

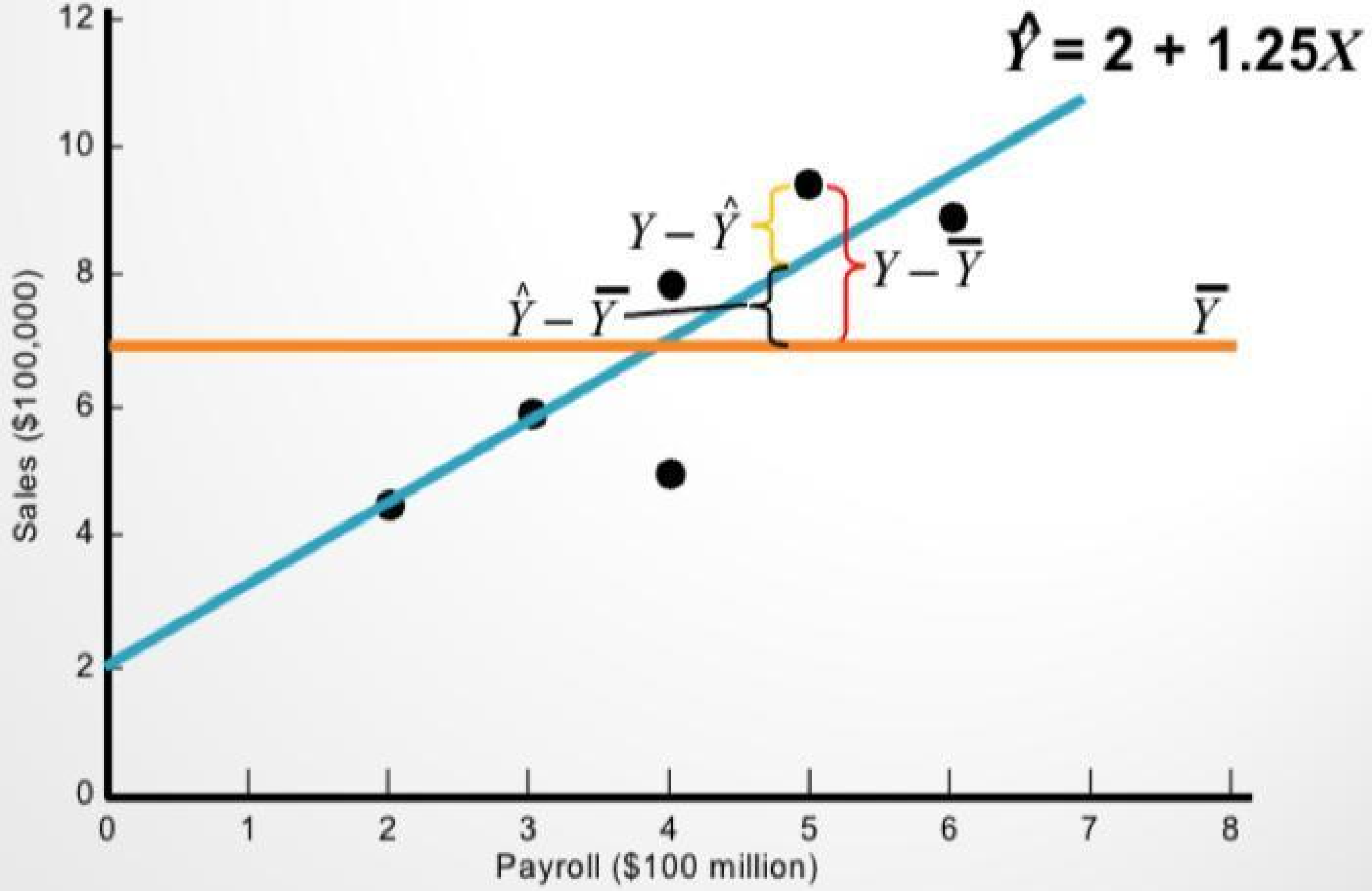
Measuring the Fit of the Regression Model

- Three measures of variability are

SST – Total variability about the mean

SSR – Total Variability that is explained by the regression line

SSE – variability about the regression model



Testing the Model for Significance

Testing the Significance

Testing the Significance of the Overall Model

Testing the Significance of the coefficients of the Model

Testing the Model for Significance

- When the sample size is too small, you can get good values for MSE and r^2 even if there is no relationship between the variables
- Testing the model for significance helps determine if the values are meaningful
- We do this by performing a statistical hypothesis test

Regression as Analysis of Variance

ANOVA conducts an *F*-test to determine whether variation in *Y* is due to varying levels of *X* (to test for *significance of regression).*

We start with the general linear model

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

H_0: population slope coefficient (β_1)= 0
H_1: population slope coefficient (β_1) $\neq$ 0

- **If β_1 = 0, the null hypothesis is that there is *no* relationship between *X* and *Y***
- **The alternate hypothesis is that there *is* a linear relationship ($\beta_1 \neq 0$)**
- **If the null hypothesis can be rejected, we have proven there is a relationship**

Analysis of Variance (ANOVA) Table

- When software is used to develop a regression model, an ANOVA table is typically created that shows the observed significance level (p-value) for the calculated F value
- This can be compared to the level of significance (α) to make a decision

	DF	SS	MS	F	SIGNIFICANCE
Regression	**k**	**SSR**	**MSR = SSR/k**	**MSR/MSE**	**P(F > MSR / MSE)**
Residual	**n - k - 1**	**SSE**	**MSE = SSE/(n - k - 1)**		
Total	**n - 1**	**SST**			

Testing the Model for Significance

- If there is very little error, the MSE would be small and the F-statistic would be large indicating the model is useful.

- If the F-statistic is large, the significance level (p-value) will be low, indicating it is unlikely this would have occurred by chance.

- So when the F-value is large, we can reject the null hypothesis and accept that there is a linear relationship between X and Y and the values of the MSE and r^2 are meaningful.

Testing the Model for Significance

Make a decision using one of the following methods

a) Reject the null hypothesis if the test statistic is greater than the F-value from the statistical tables. Otherwise, do not reject the null hypothesis:

$$\text{Reject if } \boxed{F_{calculated} > F_{\alpha, df_1, df_2}}$$
$$df_1 = k$$
$$df_2 = n - k - 1$$

b) Reject the null hypothesis if the observed significance level, or p-value, is less than the level of significance (α). Otherwise, do not reject the null hypothesis:

$$p\text{-value} = P(F > \text{calculated test statistic})$$
$$\text{Reject if } \boxed{p\text{-value} < \alpha}$$

Drugstore Sales Example

Given $\alpha = 0.05$

Calculate the value of the test statistic

$$F = \frac{MSR}{MSE} = \frac{2172.88}{54.68} = 39.74$$

	A	B	C	D	E	F	G
7	*Multiple Regression for Sales*						
8		Multiple	R-Square	Adjusted	StErr of		
9	*Summary*	R		R-Square	Estimate		
10		0.6730	0.4529	0.4415	7.394732934		
11							
12		Degrees of	Sum of	Mean of	F-Ratio	p-Value	
13	*ANOVA Table*	Freedom	Squares	Squares			
14	Explained	1	2172.880392	2172.880392	39.7366	< 0.0001	
15	Unexplained	48	2624.739608	54.68207516			
16							
17		Coefficient	Standard	t-Value	p-Value	Confidence Interval 95%	
18	*Regression Table*		Error			Lower	Upper
19	Constant	25.12642006	11.8825852	2.1146	0.0397	1.234881256	49.01795886
20	Promote	0.762296485	0.120928454	6.3037	< 0.0001	0.519153532	1.005439438

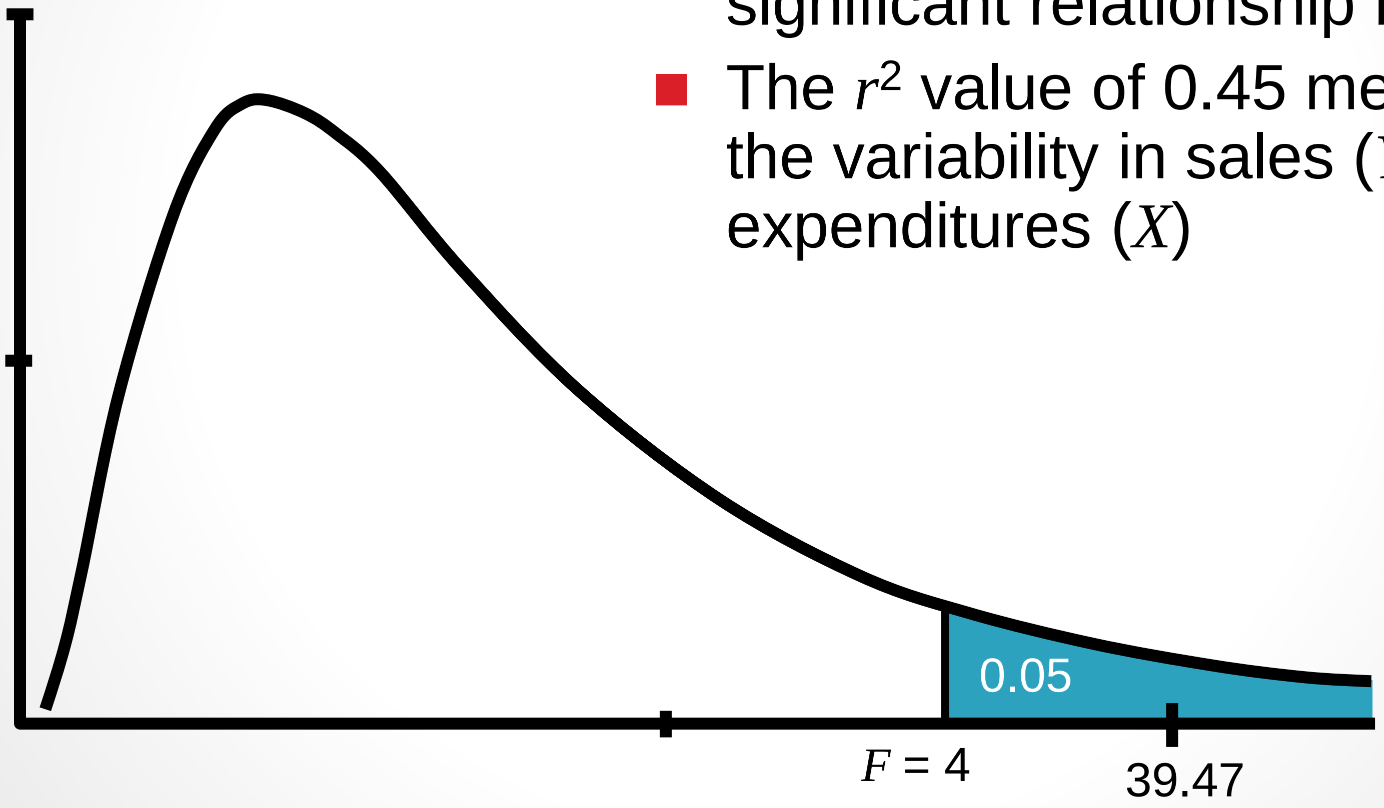
The value of F associated with a 5% level of significance and with degrees of freedom 1 and 48 from tables is

$F_{0.05,1,48} = 4$

$F_{calculated} = 39.74$

Reject H_0 because 39.74 > 4

Drugstore Sales Example



- We can conclude there is a statistically significant relationship between X and Y
- The r^2 value of 0.45 means about 45% of the variability in sales (Y) is explained by expenditures (X)

Evaluating Multiple Regression Models

- Evaluation is similar to simple linear regression models
  - The p-value for the F-test and r^2 are interpreted the same

- The hypothesis is different because there is more than one independent variable
  - The F-test is investigating whether all the coefficients are equal to 0

Evaluating Overhead Cost Example 10.2

- Both explanatory variables are significant since both have too low p-values (<0.0001)

	A	B	C	D	E	F	G
7	*Multiple Regression for Overhead*						
8		Multiple	R-Square	Adjusted	StErr of		
9	*Summary*	R		R-Square	Estimate		
10		0.9308	0.8664	0.8583	4108.99309		
11							
12		Degrees of	Sum of	Mean of	F-Ratio	p-Value	
13	*ANOVA Table*	Freedom	Squares	Squares			
14	Explained	2	3614020661	1807010330	107.0261	< 0.0001	
15	Unexplained	33	557166199.1	16883824.22			
16							
17		Coefficient	Standard	t-Value	p-Value	Confidence Interval 95%	
18	*Regression Table*		Error			Lower	Upper
19	Constant	3996.678209	6603.650932	0.6052	0.5492	-9438.550632	17431.90705
20	Machine Hours	43.53639812	3.5894837	12.1289	< 0.0001	36.23353862	50.83925761
21	Production Runs	883.6179252	82.25140753	10.7429	< 0.0001	716.2761784	1050.959672

Thank You ☺