Regression Analysis: Model Building

Introduction

Model building is the following process of developing an estimated regression equation that describes the relationship between the dependent and one or more independent variables:

- Establishing the framework for model building by introducing the concept of a general linear model.
- Determining when to add or delete independent variables.
- Variable selection procedures, including stepwise regression, the forward selection procedure, the backward elimination procedure, and best-subsets regression.

General Linear Model

As a general framework for developing an estimated regression equation that provides the best relationship between the dependent and p independent variables, we introduce the concept of a **general linear model**.

$$\boldsymbol{y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_p z_p + \epsilon}$$

Where

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the parameters of the model.

$z_1, z_2, \ldots, z_p$ are a function of the variables $x_1, x_2, \ldots, x_k$ for which the data are collected.

ϵ is the error term that accounts for the variability in y not explained by $z_1, z_2, \ldots, z_p$.

In multiple regression analysis, the word linear in the term “general linear model” refers only to the fact that $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ all have exponents of 1.

It does not imply that the relationship between y and the x_i’s is linear.

Modeling Curvilinear Relationships

The simplest case is the *simple first-order model with one predictor variable* in which we collect data for just one variable $z_1 = x_1$ and want to predict y by using a straight-line relationship.

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

However, in a first-order model, the relationship between an independent variable z_i and the dependent variable y remains constant at a value of β_i.

To model more complex associations between an independent variable and a dependent variable that changes with the values of the variables as a smooth and nonlinear curve, we need a **curvilinear relationship**.

One type of curvilinear relationship is the *second-order model with one predictor variable*, in which we set $z_1 = x_1$ and $z_2 = x_1^2$.

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon}$$

Let us consider an application of a curvilinear relationship.

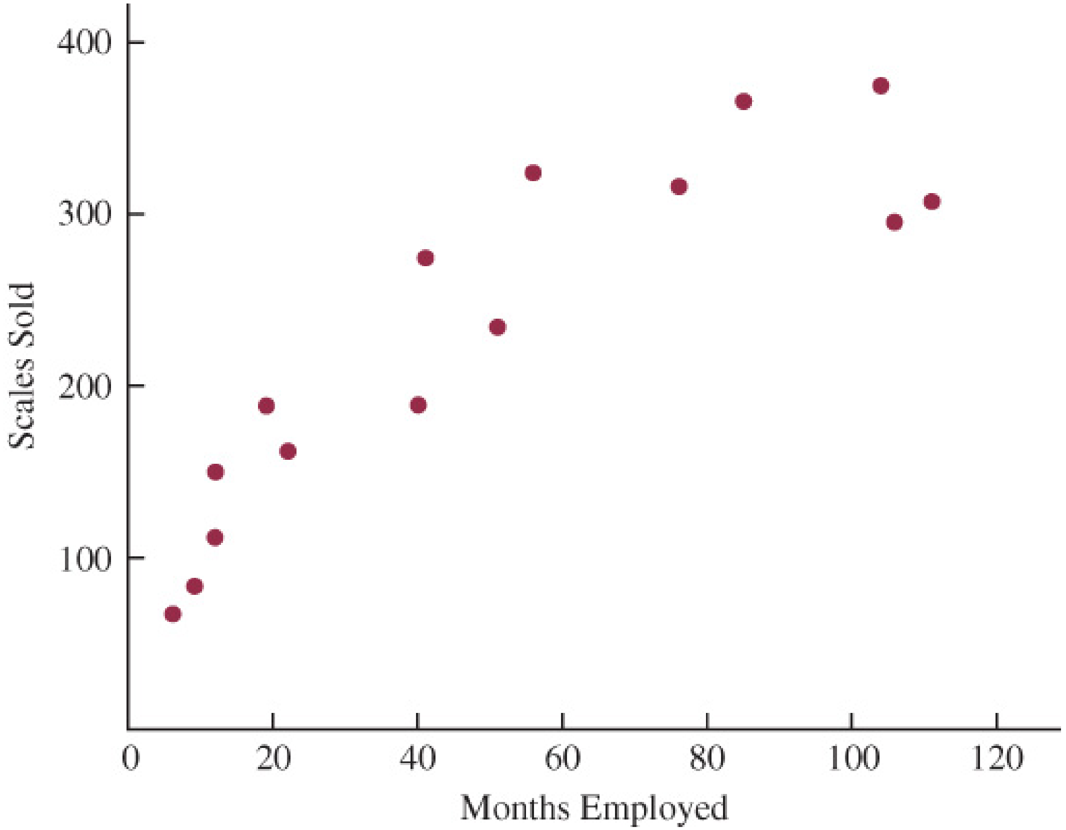
An Application of a Curvilinear Relationship

Managers at Reynolds, Inc., a manufacturer of industrial scales and laboratory equipment, want to investigate the relationship between length of employment of their salespeople and the number of electronic laboratory scales sold.

The managers collected data about the number of scales sold over the most recent sales period by 15 randomly selected salespersons, also recording the length of employment of each salesperson at the firm.

DATAfile: *Reynolds*

The scatter diagram indicates a possible curvilinear relationship between the length of time employed and the number of units sold.



A First-Order Model for the Reynolds Example

Before considering a curvilinear relationship for Reynolds, let us consider the estimated regression for a simple first-order model.

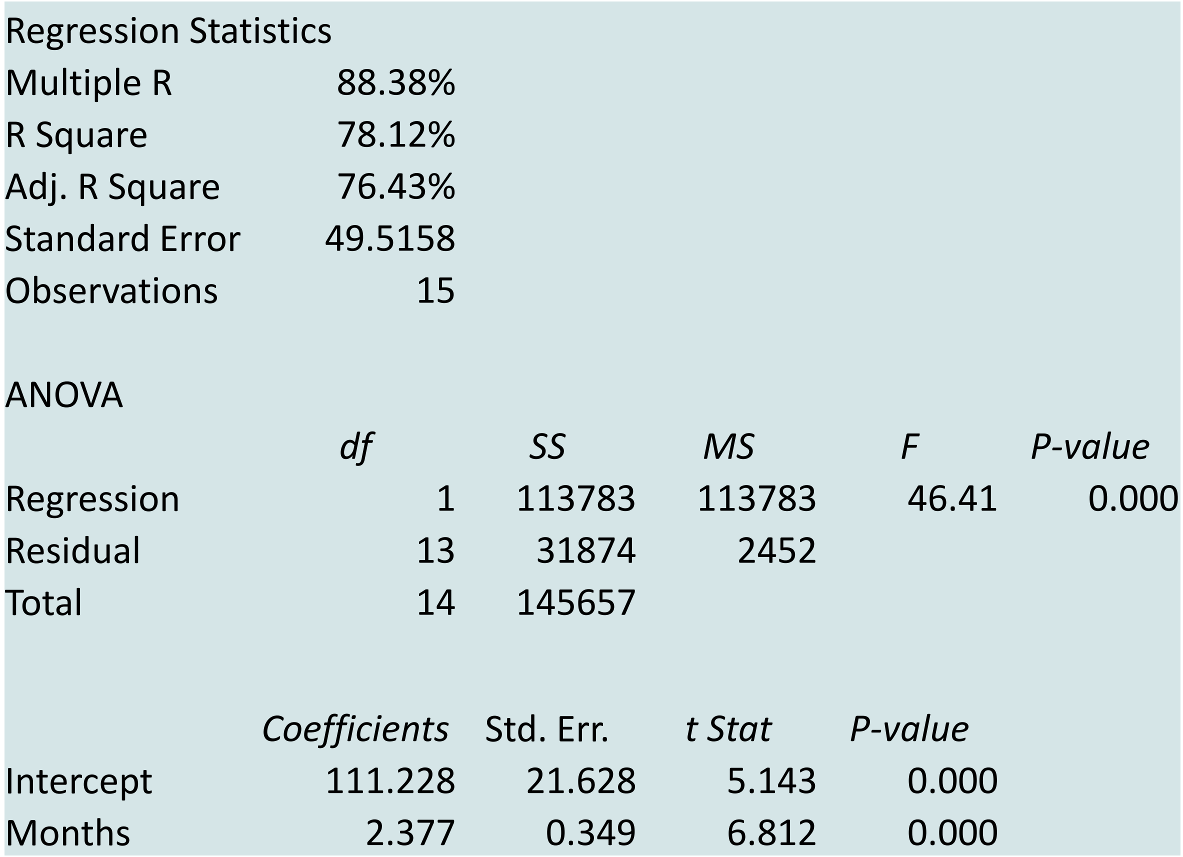
$$\text{Sales} = 111.2 + 2.377\ \text{Months}$$

Where

Sales = number of scales sold

Months = salesperson employment length.

The computer output shows how the relationship is significant (*p*-value = 0.000), and that a linear relationship with length of employment explains a high variability in sales (R-sq = 78.12%.)



Regression Statistics	
Multiple R	88.38%
R Square	78.12%
Adj. R Square	76.43%
Standard Error	49.5158
Observations	15

ANOVA

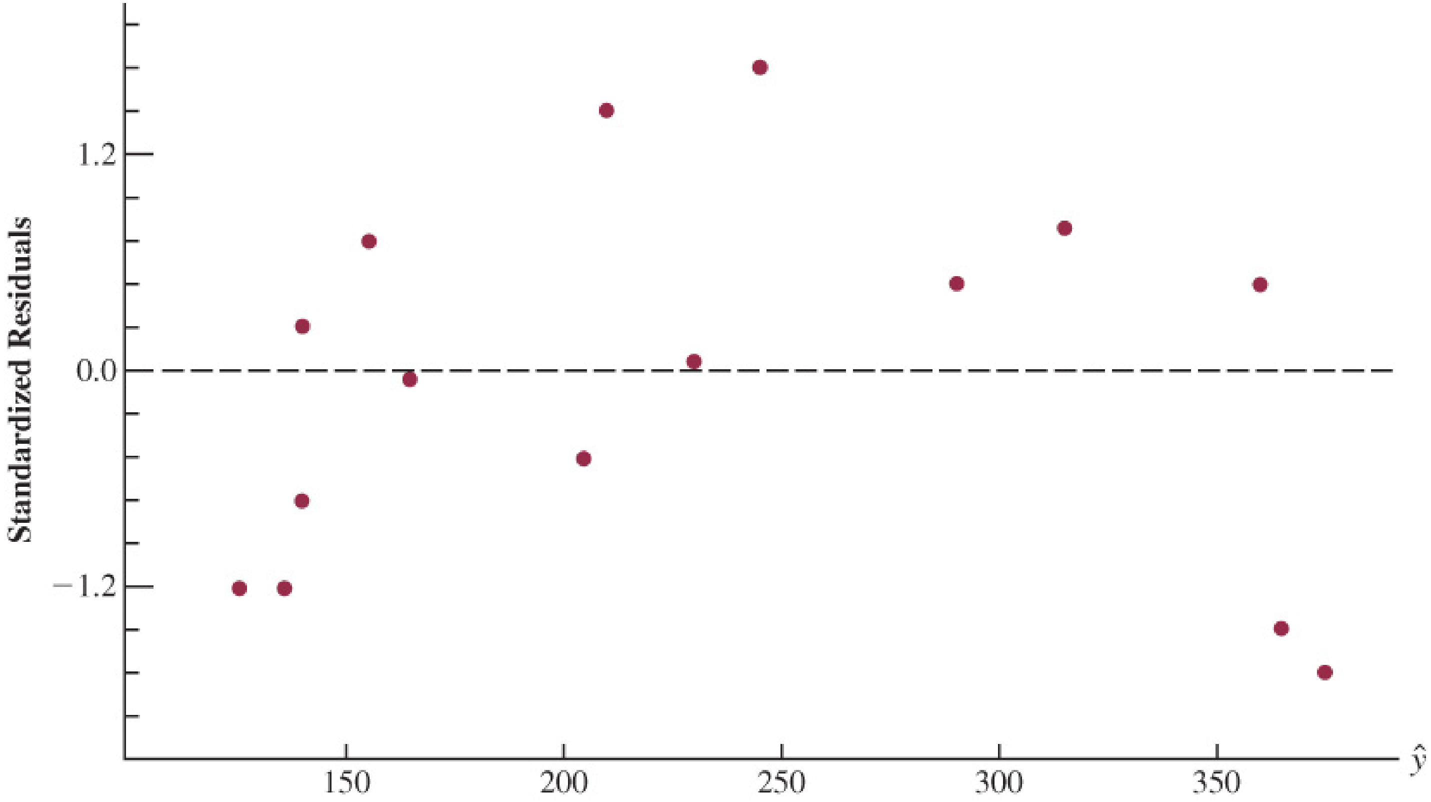
	df	*SS*	*MS*	*F*	*P-value*
Regression	1	113783	113783	46.41	0.000
Residual	13	31874	2452		
Total	14	145657			

	Coefficients	Std. Err.	*t Stat*	*P-value*
Intercept	111.228	21.628	5.143	0.000
Months	2.377	0.349	6.812	0.000

A First-Order Model Standardized Residual Plot

However, the standardized residual plot reveals a pattern in which the residuals are negative for low and high values of the predicted dependent variable, $\hat{y}$, and positive in the middle.

This suggests that the relationship between scales sold and months of employment may be curvilinear.



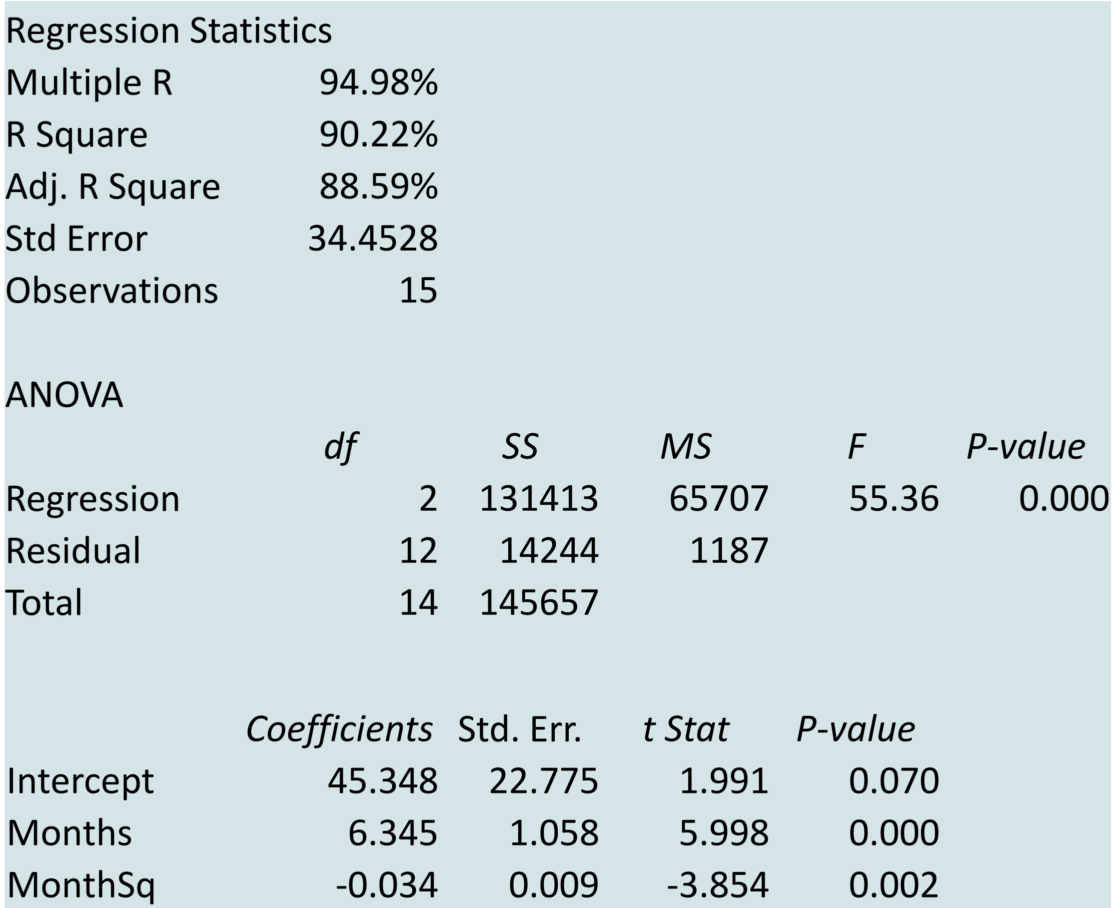
A Second-Order Model for the Reynolds Example

To account for the curvilinear relationship with a second-order model, we add a second independent variable, MonthSq, defined as the square of the number of months the salesperson has been with the firm.

We use statistical software to produce the estimated regression equation.

$$\text{Sales} = 45.3 + 6.34\ \text{Months} - 0.03449\ \text{MonthSq}$$

The computer output shows that the *F* test and the individual *t* tests are statistically significant, and a curvilinear relationship with the length of employment explains a very high variability in sales (R-sq = 90.22%.)



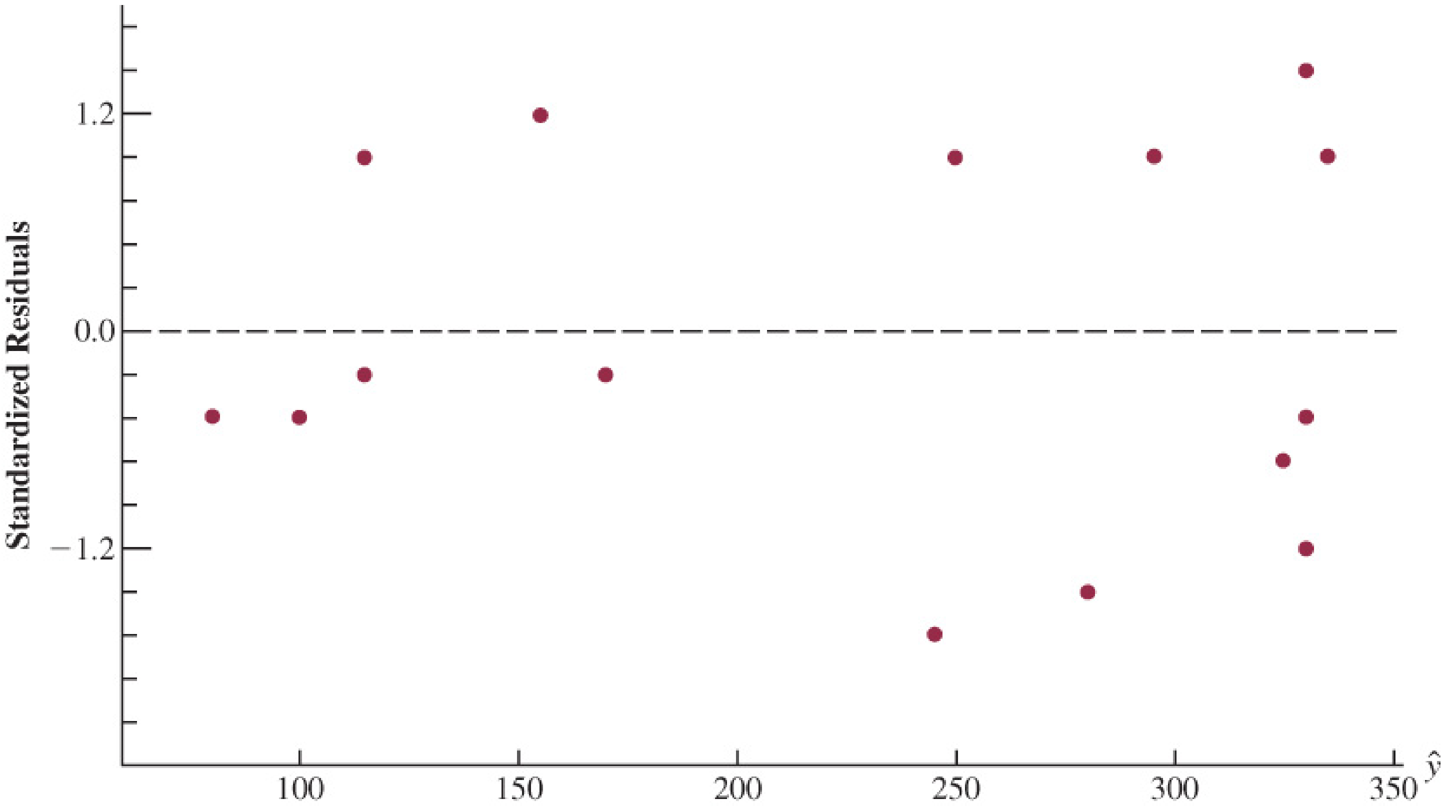
```
Regression Statistics
Multiple R          94.98%
R Square            90.22%
Adj. R Square       88.59%
Std Error          34.4528
Observations            15

ANOVA
                       df         SS        MS         F    P-value
Regression              2     131413     65707     55.36      0.000
Residual               12      14244      1187
Total                  14     145657


             Coefficients   Std. Err.    t Stat    P-value
Intercept          45.348      22.775     1.991      0.070
Months              6.345       1.058     5.998      0.000
MonthSq            -0.034       0.009    -3.854      0.002
```

A Second-Order Model Standardized Residual Plot

The standardized residual plot shows that the addition of the quadratic term removed the previous curvilinear pattern.

We can conclude that adding a quadratic term MonthsSq to the first-order model involving Months is statistically significant.

Furthermore, comparison of the R-sq(adj) reveals an increase from 76.43% to 88.59% with the addition of the quadratic term.



Interaction

If the original data set consists of observations for y and two independent variables, x_1 and x_2, we can use the general linear model to develop a *second-order model with two predictor variables* by setting $z_1 = x_1$, $z_2 = x_2$, $z_3 = x_1^2$, $z_4 = x_2^2$, and $z_5 = x_1 x_2$.

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon}$$

The variable $z_5 = x_1 x_2$ is added to account for the **interaction** effect between the variables.

If there is a belief that the dependent variable y has a linear relationship with the independent variables x_1 and x_2, a simplified version of the general linear model with the interaction term can be written as follows.

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon}$$

Where $z_1 = x_1$, $z_2 = x_2$, and $z_3 = x_1 x_2$.

Let us consider an application of the simplified second-order model with an interaction term.

Tyler Personal Care: an Application of Interaction

To provide an illustration of interaction and what it means, let us review the regression study conducted by Tyler Personal Care for one of its new shampoo products.

Two factors believed to have the most influence on sales are:

- unit selling price
- advertising expenditure

To investigate the effects of these two variables on sales, prices of $2.00, $2.50, and $3.00 were paid with advertising expenditures of $50,000 and $100,000 in 24 test markets.

DATAfile: *Tyler*

A summary study, shown to the right, reveals the diminishing effects that increased advertising ($50,000 vs. $100,000) has on mean sales, as the selling price increases from $2.00 to $3.00.

		Price	
Advertising Expenditure	**$2.00**	**$2.50**	**$3.00**
$50,000	461	364	332
$100,000	808	646	375

The Tyler Personal Care Example

We use statistical software to produce the estimated regression equation.

$$\text{Sales} = -276 + 175\ \text{Price} + 19.68\ \text{Advert}$$
$$-6.08\ \text{PriceAdvert}$$

Where the interaction term PriceAdvert was created by multiplying the values of Price and Advertising for each of the 24 observations.

The computer output shows that the *F* test and the individual *t* tests are statistically significant, including the *t* test for PriceAdvert. Thus, the interaction term is also statistically significant.

The regression model with the interaction term explains 97.81% of the variability in sales.

Regression Statistics	
Multiple R	94.98%
R Square	90.22%
Adj. R Square	88.59%
Std Error	34.4528
Observations	15

ANOVA

	df	SS	MS	F	P-value
Regression	2	131413	65707	55.36	0.000
Residual	12	14244	1187		
Total	14	145657			

	Coefficients	Std. Err.	t Stat	P-value
Intercept	-275.833	112.842	-2.444	0.024
Price	175.000	44.547	3.928	0.001
Advert	19.680	1.427	13.788	0.000
PriceAdvert	-6.080	0.563	-10.790	0.000

Transformations Involving the Dependent Variable

At times, it is worthwhile to consider transformations involving the dependent variable y rather than focusing attention on transformations involving one or more of the independent variables.

When the underlying assumptions for the tests of significance do not appear to be satisfied, such as in the presence of nonconstant variance, we are not justified in reaching any conclusions about the statistical significance of the resulting estimated regression equation.

Often, the problem of nonconstant variance can be corrected by transforming the dependent variable to a different scale, such as:

- The logarithmic transformation, using either the base-10 (common log) or the base *e* = 2.71828... (natural log).
- The reciprocal transformation, $1/y$ as the dependent variable instead of y.

When comparing the fit of a model with a transformed dependent variable with the one based on the original dependent variable, care must be taken to convert the estimated transformed model back into the same units of the original dependent variable.

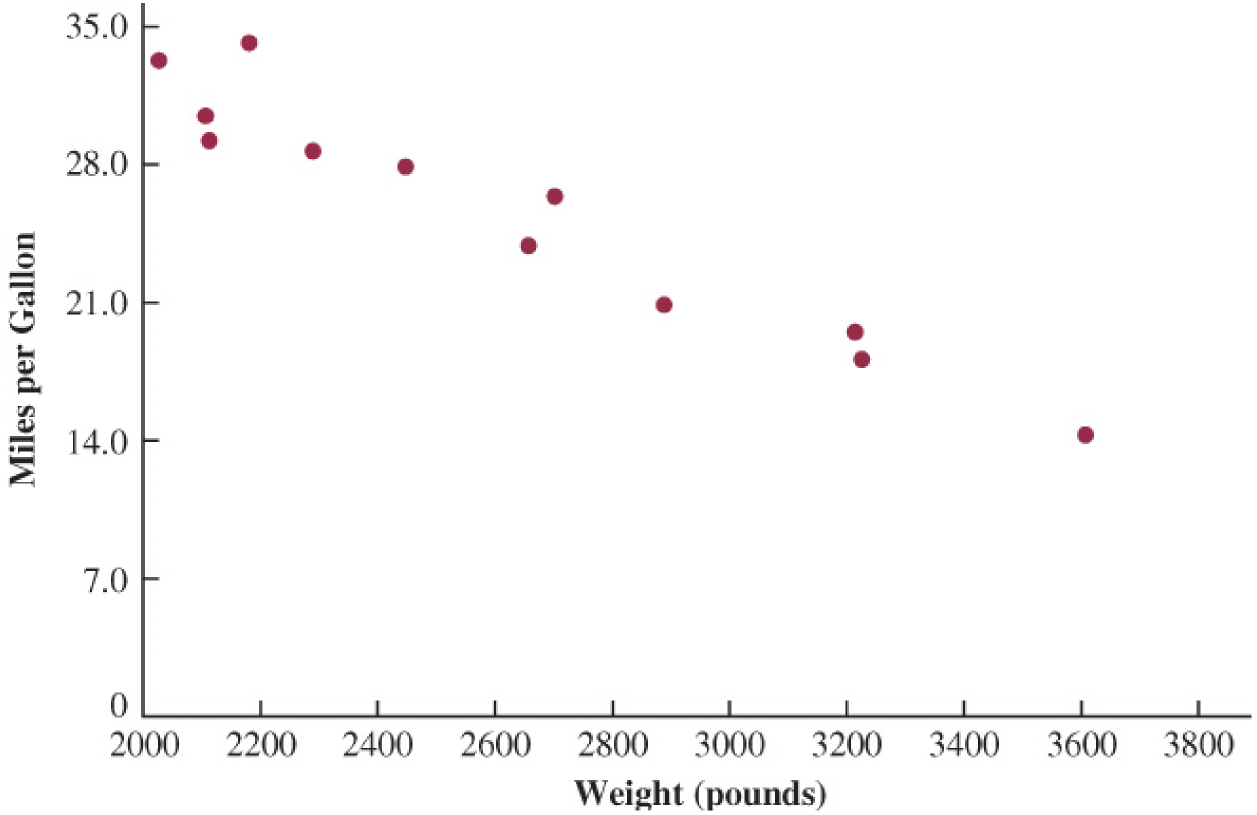
An Application Involving the Transformation of the Dependent Variable

As an illustration of when we might want to transform the dependent variable, consider the data about the miles-per-gallon ratings (y) and weight in pounds for 12 automobiles (x).

DATAfile: *MPG*

The scatter diagram indicates a negative linear relationship between these two variables.

Therefore, we use a simple first-order model to relate the two variables.



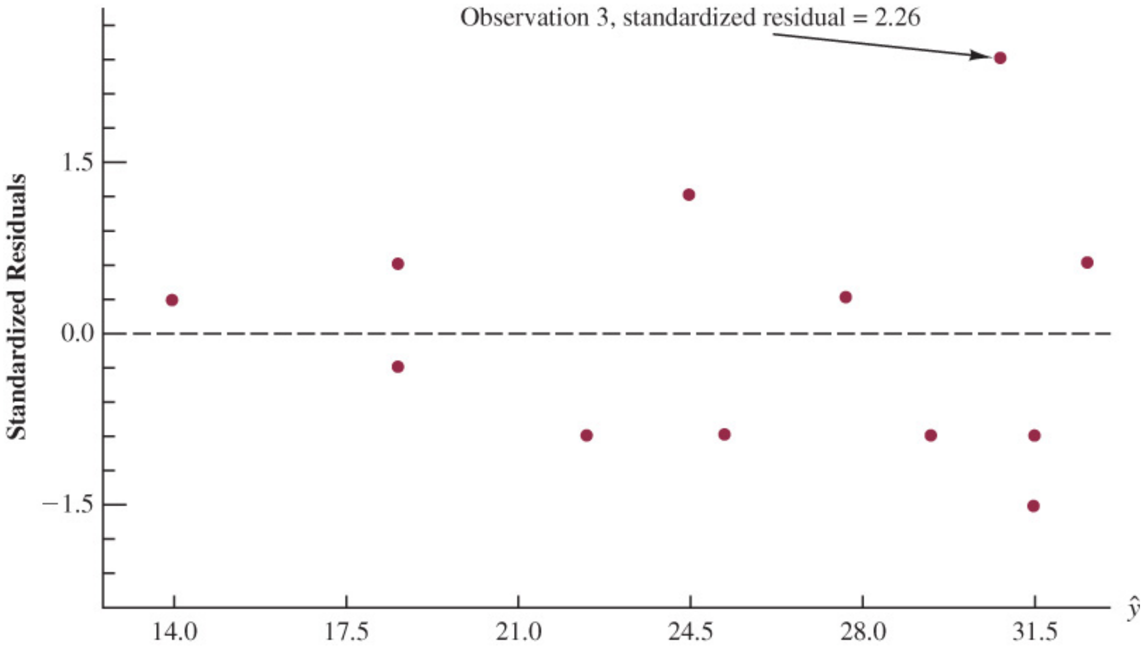
A First-Order Model of the MPG Example

The computer output produced the estimated regression equation

$$\mathrm{MPG} = 56.1 - 0.011644\ \mathrm{Weight}$$

The output shows that the model is significant with R-Sq equal to 93.94%, but the standardized residual plot shows a wedged shape pattern and an influential observation.

Regression Statistics					
Multiple R	96.72%		Std Error	1.6705	
R Square	93.54%		Observations	12	
Adj. R Square	92.89%				
ANOVA					
	df	SS	MS	F	P-value
Regression	1	403.98	403.98	144.76	0.000
Residual	10	27.91	2.79		
Total	11	431.88			
	Coefficients	Std. Err.	t Stat	P-value	
Intercept	56.096	2.582	21.725	0.000	
Weight	-0.011644	0.000968	-12.032	0.000	



Logarithmic Transformation of the MPG Variable

The computer output with the transformed LnMPG produced the following regression equation

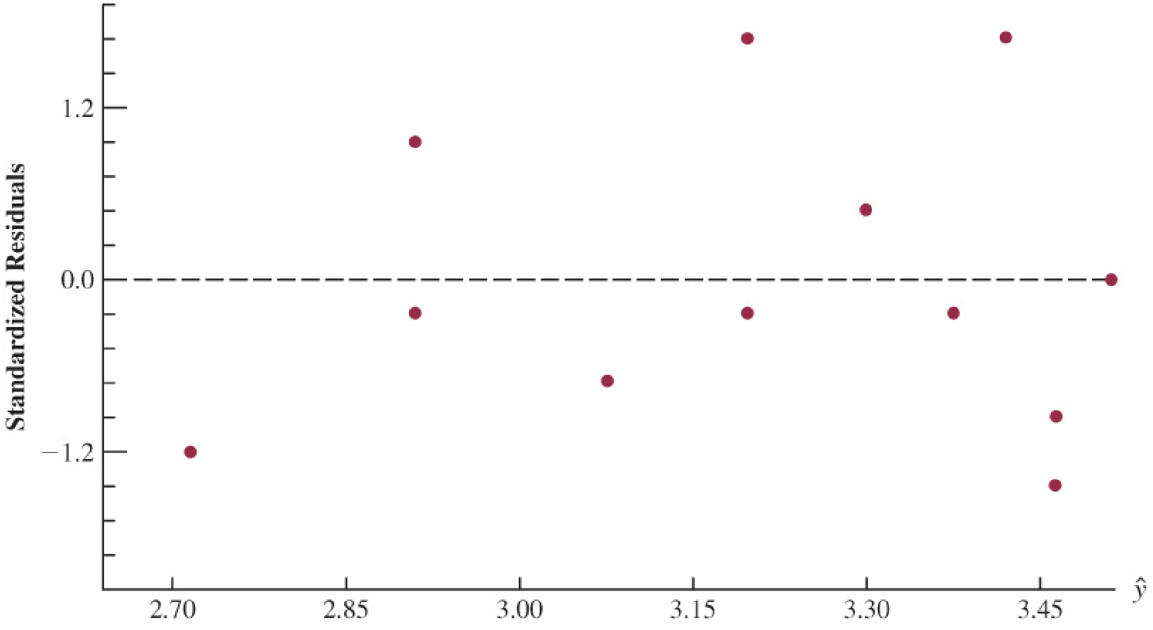
$$\ln \text{MPG} = 4.5242 - 0.000501\ \text{Weight}$$

The output shows that the model is significant with a high R-Sq equal to 94.77% and, more importantly, the nonconstant variance and unusual observation have now disappeared.

Regression Statistics			
Multiple R	97.35%	Std Error	0.0643
R Square	94.77%	Observations	12
Adj. R Square	94.25%		

ANOVA					
	df	SS	MS	F	P-value
Regression	1	0.7482	0.7482	181.22	0.000
Residual	10	0.0413	0.0041		
Total	11	0.7895			

	Coefficients	Std. Err.	t Stat	P-value
Intercept	4.5242	0.0993	45.5527	0.000
Weight	-0.000501	0.000037	-13.462	0.000



Comparing Models with a Transformed Dependent Variable

In the case of the MPG example, to compare the model with the logarithmic transformation to the untrasformed one, we need to first perform the following conversion.

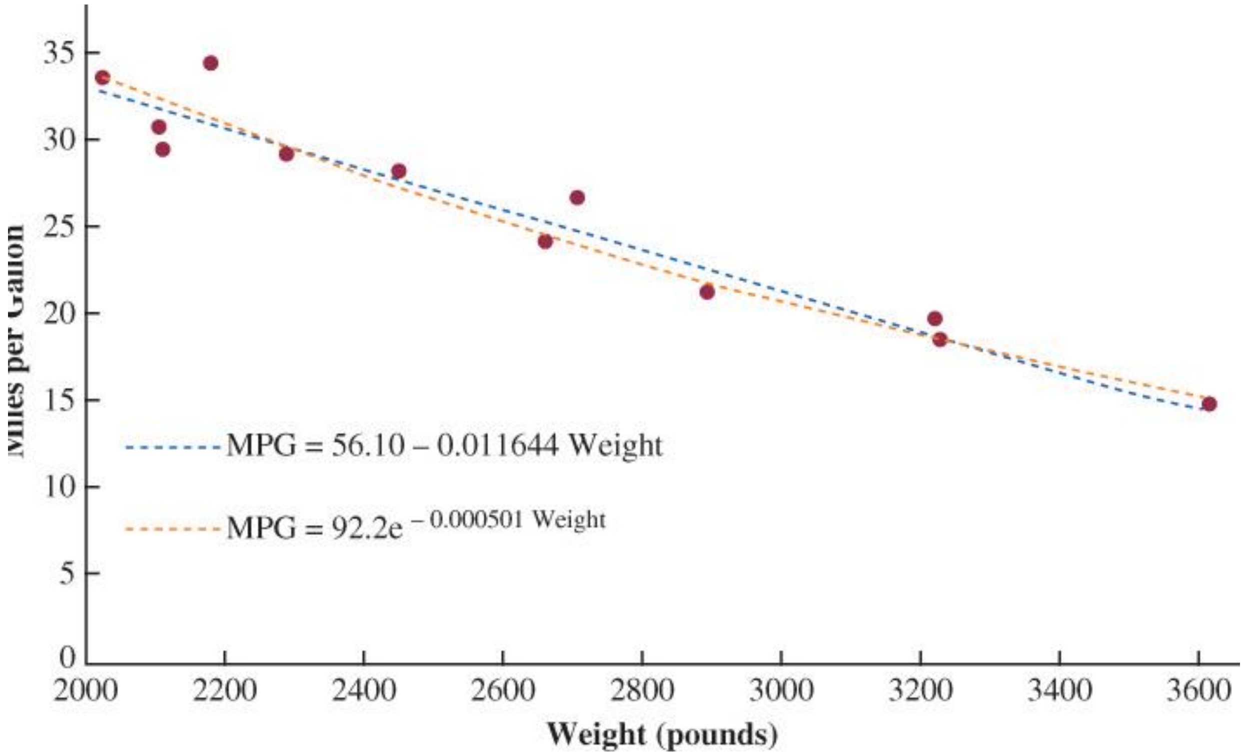
$$\ln \mathrm{MPG} = 56.1 - 0.011644\ \mathrm{Weight}$$

If we elevate both sides to the power of *e*, we obtain

$$\mathrm{MPG} = e^{56.10 - 0.011644\ \mathrm{Weight}}$$

$$MPG = 92.2e^{-0.011644\ \mathrm{Weight}}$$

The error resulting from the predicted values from both models can now be compared, as shown in the scatter diagram with the two fitted curves.



Nonlinear Models That Are Intrinsically Linear

Models in which the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ have exponents other than one are called nonlinear models.

In some cases, we can perform a transformation of variables that will enable us to use regression analysis with the general linear model.

One of these models is the exponential model, which involves the regression equation:

$$E(y) = \beta_0 \beta_1^x$$

We can transform this nonlinear regression equation to a linear regression equation by taking the natural logarithm of both sides of equation.

$$\ln E(y) = \ln \beta_0 + x \ln \beta_1$$

Now, if we let $y' = \ln E(y)$, $\beta_0' = \ln \beta_0$, and , $\beta_1' = \ln \beta_1$, we can rewrite the model as

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

Determining When to Add or Delete Variables

Consider the following multiple regression model involving q independent variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \epsilon$$

If we add the variables $x_{q+1}, x_{q+2}, \ldots x_p$, we obtain a model involving p independent variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \beta_{q+1} x_{q+1} + \beta_{q+2} x_{q+2} + \ldots + \beta_p x_p + \epsilon$$

To test whether the addition of $x_{q+1}, x_{q+2}, \ldots x_p$, is statistically significant, we test the following.

$$H_0: \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0$$

$$H_a: \text{One or more of the parameters is not equal to zero}$$

The test statistic follows an F distribution with $p - q$ degrees of freedom at the numerator and p degrees of freedom at the denominator.

$$F = \frac{\dfrac{SSE(\text{with } q \text{ variables}) - SSE(\text{with } p \text{ variables})}{p - q}}{\dfrac{SSE(\text{with } p \text{ variables})}{n - p - 1}}$$

Butler Trucking Company Application

Butler Trucking wants to improve the prediction model it uses to schedule deliveries.

Presently, the estimated regression equation predicts the total delivery time (y) using the miles traveled (x_1) as the only independent variable from a random sample of $n = 10$ deliveries.

$\hat{y} = 1.27 + 0.0678x_1$ with $q = 1$ and $SSE = 8.029$

When the number of deliveries is added as a second independent variable, x_2, the estimate regression equation becomes

$\hat{y} = -0.869 + 0.06113x_1 + 0.923x_2$ with $p = 2$ and $SSE = 2.2994$

The test statistic follows an F distribution with $p - q = 1$ degrees of freedom at the numerator and $n - p - 1 = 7$ degrees of freedom at the denominator.

$$F = \frac{\dfrac{SSE(\text{reduced}) - SSE(\text{full})}{p - q}}{\dfrac{SSE(\text{full})}{n - p - 1}} = \frac{\dfrac{8.029 - 2.2994}{2 - 1}}{\dfrac{8.029}{10 - 2 - 1}} = \frac{\dfrac{5.7296}{1}}{\dfrac{8.029}{7}} = \frac{5.7296}{0.3285} = 17.44$$

p-Value Approach to Testing for Variable Addition and Deletion

Using a significance level, $\alpha = 0.05$, Table 4 in Appendix B provides a value of $F_{0.05} = 5.59$ for a combination of 1 and 7 degrees of freedom for numerator and denominator.

Because $F = 17.44 > F_{0.05} = 5.59$, we can reject the null hypothesis that x_2 is not significant.

We can conclude that adding the number of deliveries to the model improves the precision of the Butler Trucking scheduling process.

It can be proven that the F statistic we just computed is the square of the *t* statistic used to test the significance of one parameter added to the model. Thus, we can conclude:

- Add an individual variable to the model if *t* test indicates its parameter to be significant.
- Conversely, drop an individual variable from the model if its parameter is not significant.
- However, do not drop more than one variable from the model if the individual *t* tests show that two or more parameters are not significant.

Application of Variable Selection Procedures

To provide an illustration of variable selection procedures, we introduce the Cravens data set (DATAfile: *Cravens)* for a company that sells products using a single sales representative located in each of 25 sales territories.

We use regression analysis to predict sales with 8 independent variables, defined as follows.

Sales	Total sales credited to the sales representative
Time	Length Of time employed in months
Poten	Market potential; total industry sales in units for the sales territory*
AdvExp	Advertising expenditure in the sales territory
Share	Market share; weighted average for the past four years
Change	Change in the market share over the previous four years
Accounts	Number of accounts assigned to the sales representative*
Work	A weighted index based on annual purchases and concentrations of accounts
Rating	Sales representative overall rating on a 1-7 scale

Correlation Matrix for the Cravens Data Set

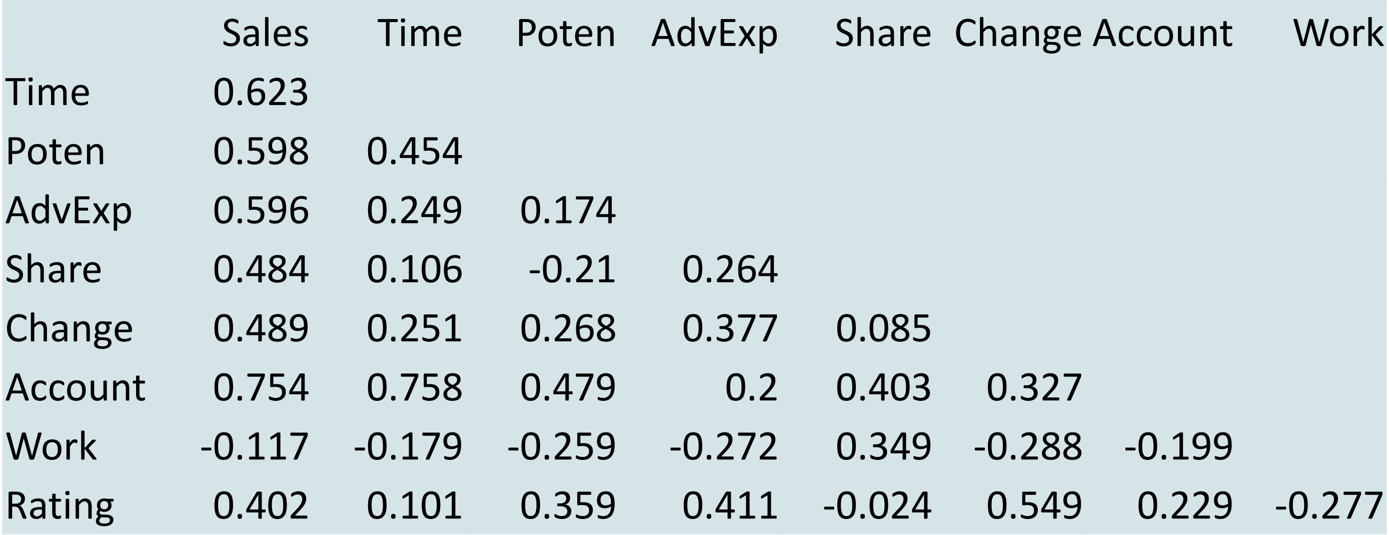
Let us consider the sample correlation coefficients between each pair of variables.

Multicollinearity can cause problems with model interpretability if the absolute value of the sample correlation coefficient exceeds 0.7 for any two of the independent variables.

Thus, we should avoid including Accounts and Time in the same regression model ($r = 0.758$), and we should also keep an eye on Rating and Change ($r = 0.549$.)

When considering sample correlation coefficients with the dependent variable Sales:

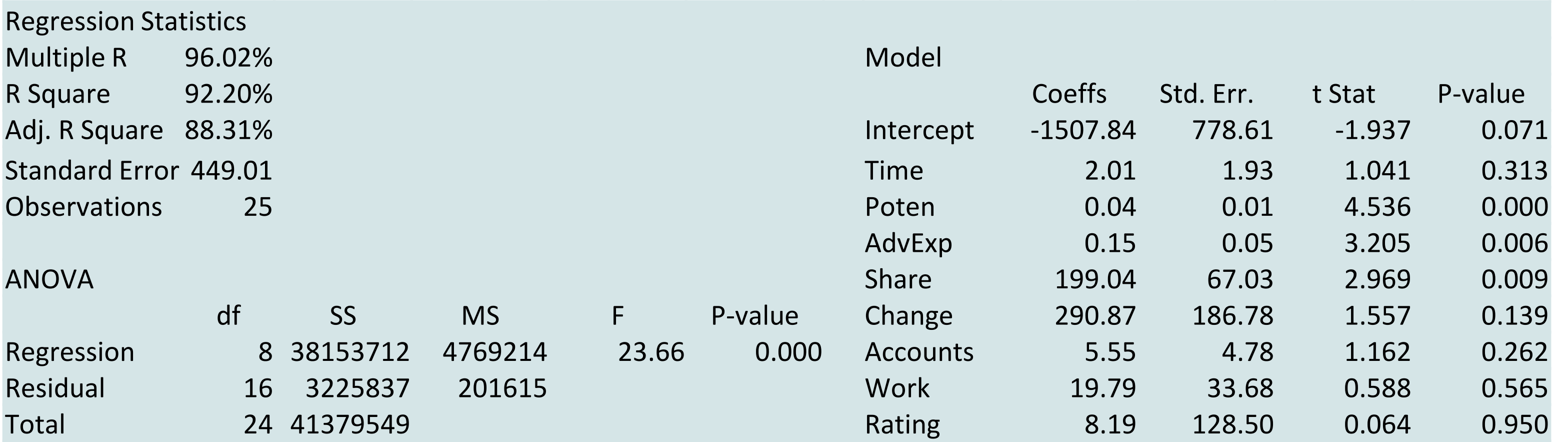
- Accounts has the highest $r = 0.754$.
- Next come Time, Poten, and AdvExp, all with *r* around 0.6.



	Sales	Time	Poten	AdvExp	Share	Change	Account	Work
Time	0.623							
Poten	0.598	0.454						
AdvExp	0.596	0.249	0.174					
Share	0.484	0.106	-0.21	0.264				
Change	0.489	0.251	0.268	0.377	0.085			
Account	0.754	0.758	0.479	0.2	0.403	0.327		
Work	-0.117	-0.179	-0.259	-0.272	0.349	-0.288	-0.199	
Rating	0.402	0.101	0.359	0.411	-0.024	0.549	0.229	-0.277

Initial Multiple Regression Models

The full model with all eight independent variables has $R_a^2 = 88.31\%$, but only three significant variables: Poten, AdvExp, & Share.



Regression Statistics	
Multiple R	96.02%
R Square	92.20%
Adj. R Square	88.31%
Standard Error	449.01
Observations	25

ANOVA

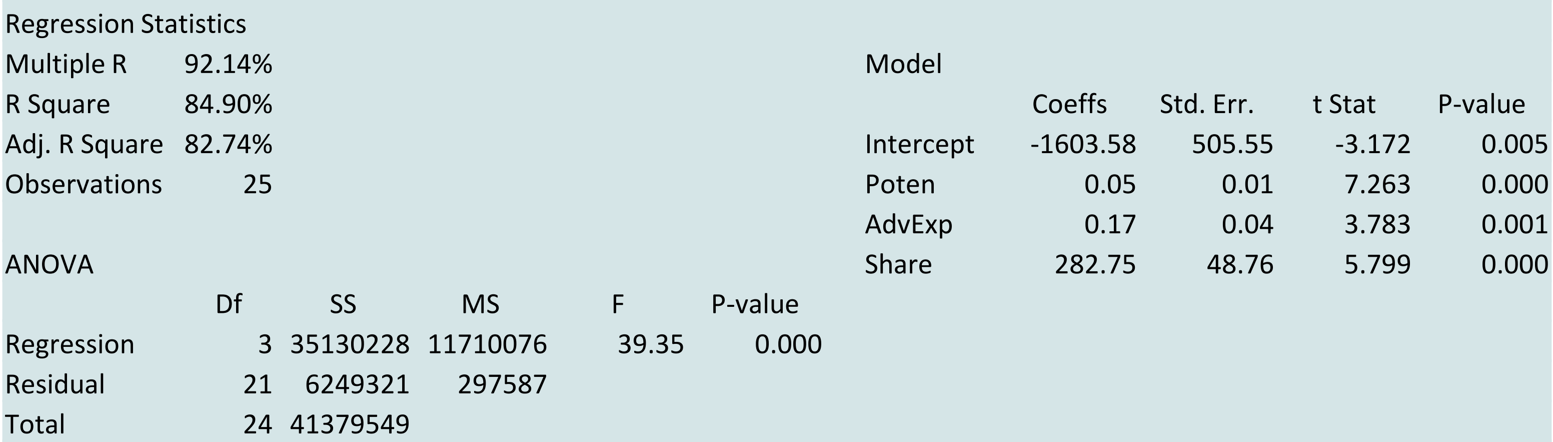
	df	SS	MS	F	P-value
Regression	8	38153712	4769214	23.66	0.000
Residual	16	3225837	201615		
Total	24	41379549			

Model

	Coeffs	Std. Err.	t Stat	P-value
Intercept	-1507.84	778.61	-1.937	0.071
Time	2.01	1.93	1.041	0.313
Poten	0.04	0.01	4.536	0.000
AdvExp	0.15	0.05	3.205	0.006
Share	199.04	67.03	2.969	0.009
Change	290.87	186.78	1.557	0.139
Accounts	5.55	4.78	1.162	0.262
Work	19.79	33.68	0.588	0.565
Rating	8.19	128.50	0.064	0.950

When we run a model with only those three variables, we obtain a significant model with $R_a^2 = 82.74\%$.

How do we find the best estimated regression equation?



Regression Statistics	
Multiple R	92.14%
R Square	84.90%
Adj. R Square	82.74%
Observations	25

ANOVA

	Df	SS	MS	F	P-value
Regression	3	35130228	11710076	39.35	0.000
Residual	21	6249321	297587		
Total	24	41379549			

Model

	Coeffs	Std. Err.	t Stat	P-value
Intercept	-1603.58	505.55	-3.172	0.005
Poten	0.05	0.01	7.263	0.000
AdvExp	0.17	0.04	3.783	0.001
Share	282.75	48.76	5.799	0.000

Variable Selection Procedures

There are four major variable selection procedures we can use to find the best estimated regression equation for a set of independent variables:

1. Stepwise Regression
2. Forward Selection
3. Backward Elimination
4. Best-Subsets Regression

The first three procedures are Iterative; one independent variable at a time is added or deleted based on the F statistic but offer no guarantee that the best model will be found.

In the fourth procedure, all possible subsets of the independent variables are evaluated.

Variable Selection: Stepwise Regression

In the **stepwise regression** procedure for variable selection, two levels of significance are set:

- *α-to-leave* determines if an independent variable should be removed from the model.
- *α-to-enter* determines if an independent variable should be added to the model.

The procedure starts with a one-variable model: the most significant variable.

At each iteration, the first consideration is to see whether the least significant variable *already in the model* can be removed because its F value has *p*-value > *α-to-leave*.

If no variable can be removed, the procedure checks to see whether the most significant variable not in the model can be added because its F value has *p*-value ≤ *α-to-enter*.

In the stepwise regression procedure, an independent variable can enter the model at one step, be removed at a subsequent step, and enter the model at a later step.

The procedure stops when no independent variable can be removed from or entered into the model.

Stepwise Regression Output for the Cravens Data

The stepwise regression procedure used on the Cravens data uses

α-to-leave = *0.05*

α-to-enter = 0.05.

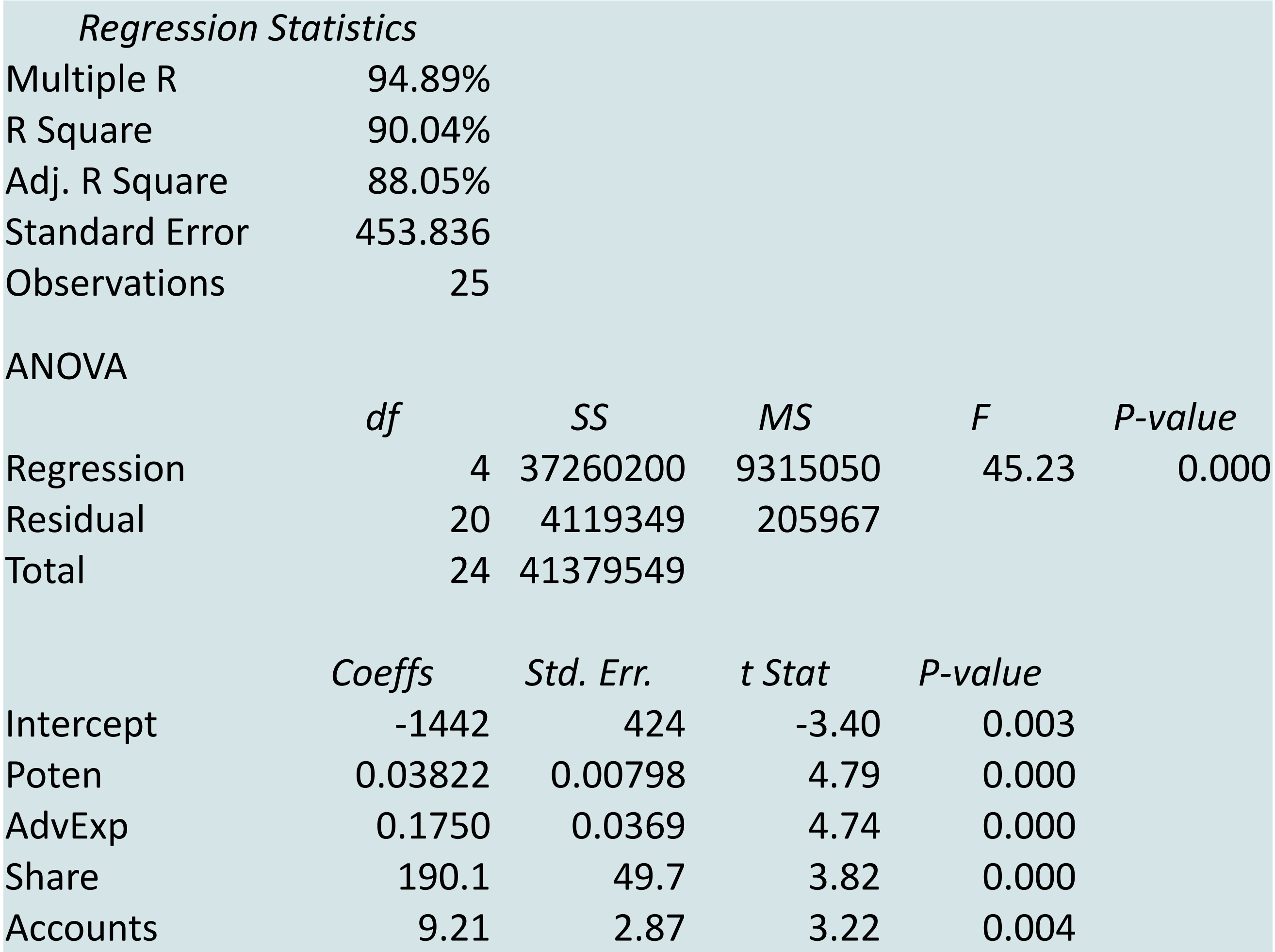
It started with a one-variable model (Accounts), and the final estimated regression equation contained 4 independent variables:

Poten, AdvExp, Share, and Accounts

After four steps, the procedure improved to

$s = 453.84$

$R^2 = 90.04\%$



Regression Statistics	
Multiple R	94.89%
R Square	90.04%
Adj. R Square	88.05%
Standard Error	453.836
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	4	37260200	9315050	45.23	0.000
Residual	20	4119349	205967		
Total	24	41379549			

	Coeffs	*Std. Err.*	*t Stat*	*P-value*
Intercept	-1442	424	-3.40	0.003
Poten	0.03822	0.00798	4.79	0.000
AdvExp	0.1750	0.0369	4.74	0.000
Share	190.1	49.7	3.82	0.000
Accounts	9.21	2.87	3.22	0.004

Variable Selection: Forward Regression

In the **forward regression** procedure for variable selection, the significance level *α-to-enter* determines whether an independent variable should enter the model.

The forward selection procedure is similar to the stepwise regression procedure, but it does not permit a variable to be removed from the model once it has been entered.

The procedure starts with no variables in the model.

At each iteration, the procedure checks whether the most significant variable not in the model can be entered into the model because its F statistic has *p*-value ≤ *α-to-enter*.

The procedure stops when no independent variables can be entered into the model.

The estimated regression equation obtained using the forward selection procedure for the Cravens data set used *α-to-enter* = 0.05. The procedure ended with the same final estimated regression equation obtained with the stepwise regression procedure.

$$\text{Sales} = -1442 + 0.03822\ \text{Poten} + 0.1750\ \text{AdvExp} + 190.1\ \text{Shares} + 9.21\ \text{Accounts}$$

Variable Selection: Backward Regression

In the **backward regression** procedure for variable selection, the significance level *α-to-leave* determines whether an independent variable should be removed from the model.

This procedure begins with a model that includes all the considered independent variables.

At each iteration, the procedure checks whether the least significant variable *already in the model* can be removed because its F value has *p*-value > *α-to-leave*.

The procedure stops when no other independent variables can be removed from the model.

The estimated regression equation obtained using the backward selection procedure for the Cravens data set started with all eight independent variables and *α-to-leave* = 0.05.

The procedure ended with a final estimated regression equation that includes the variable Time, rather than the variable Accounts, as it was for the previous procedures.

$$\text{Sales} = -1312 + 3.82\ \text{Time} + 0.03822\ \text{Poten} + 0.1750\ \text{AdvExp} + 190.1\ \text{Shares}$$

The procedure improved to $s = 463.93$ and $R^2 = 89.60\%$ after four steps.

Variable Selection: Best Subsets Regression

Stepwise regression, forward selection, and backward elimination are one-variable-at-a-time heuristic procedures that offer no guarantee the best model for a given number of variables will be found.

The **best-subsets regression** procedure enables the user to find, given a specified number of independent variables, the best regression model.

However, the computation of all the possible subsets for all independent variables may involve great computational power and require the review of a large volume of computer output.

For example, a best-subsets regression of the Cravens data set involves the analysis of 255 separate estimated regression equations.

In the next slide, we show a portion of the output obtained by using the best-subsets procedure for the Cravens data set, including the two best models for each number of independent variables from 1 to 7, and the full model with 8 independent variables.

Best Subsets Regression Output for the Cravens Data

For models with the same number of predictors, we use R^2 to select the better model. For the choice of the best regression model, consider the following five:

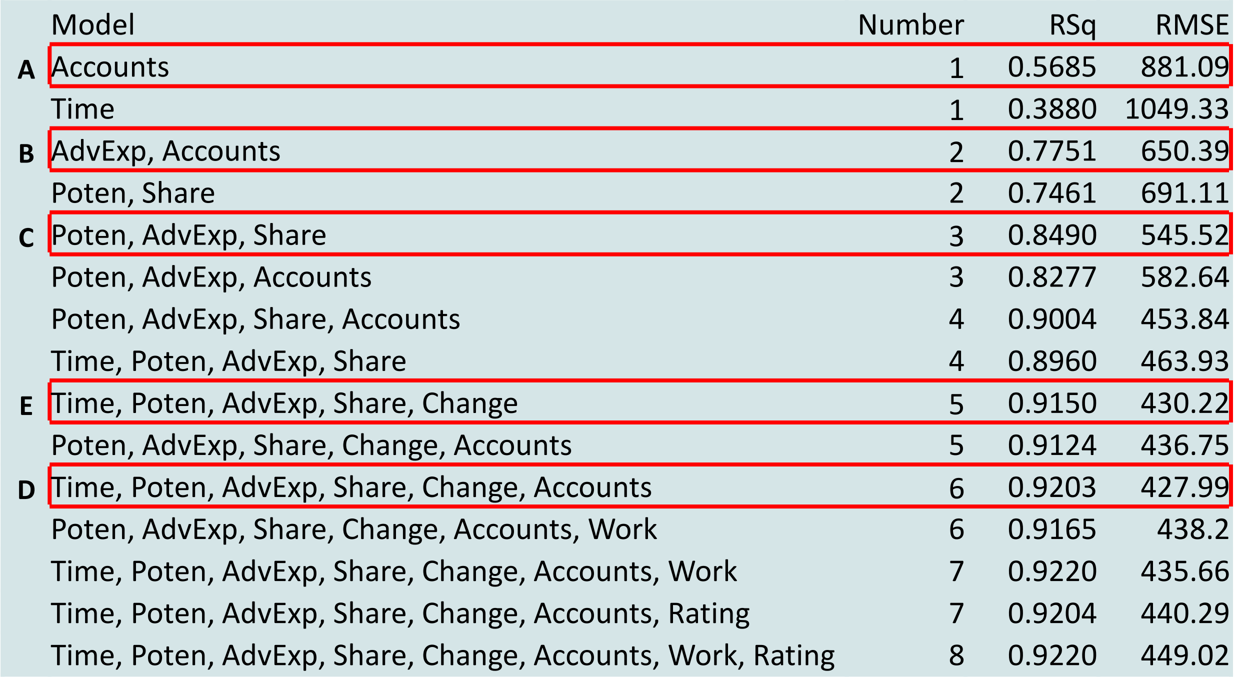
A. Accounts provides the best 1-variable model ($R^2 = 56.85\%$)

B. AdvExp and Accounts provide the best 2-variable model ($R^2 = 77.51\%$)

C. Poten, Adv.Exp. & Share provide the best 3-variable model ($R^2 = 84.90\%$)

D. This 6-variable model has the best $RMSE = 427.99$, but includes both Time and Accounts, known to be collinear.

E. This 5-variable model is simpler, has similar R^2 and $RMSE$ to the 6-variable one, but without any collinearity issues.



	Model	Number	RSq	RMSE
A	Accounts	1	0.5685	881.09
	Time	1	0.3880	1049.33
B	AdvExp, Accounts	2	0.7751	650.39
	Poten, Share	2	0.7461	691.11
C	Poten, AdvExp, Share	3	0.8490	545.52
	Poten, AdvExp, Accounts	3	0.8277	582.64
	Poten, AdvExp, Share, Accounts	4	0.9004	453.84
	Time, Poten, AdvExp, Share	4	0.8960	463.93
E	Time, Poten, AdvExp, Share, Change	5	0.9150	430.22
	Poten, AdvExp, Share, Change, Accounts	5	0.9124	436.75
D	Time, Poten, AdvExp, Share, Change, Accounts	6	0.9203	427.99
	Poten, AdvExp, Share, Change, Accounts, Work	6	0.9165	438.2
	Time, Poten, AdvExp, Share, Change, Accounts, Work	7	0.9220	435.66
	Time, Poten, AdvExp, Share, Change, Accounts, Rating	7	0.9204	440.29
	Time, Poten, AdvExp, Share, Change, Accounts, Work, Rating	8	0.9220	449.02

Making the Final Variable Selection

The analysis performed on the Cravens data to this point is good preparation for choosing a final model, but a careful analysis of the residuals should be conducted before the final choice.

If the analysis of the residuals does not highlight any problem, we can use the results of the best-subsets procedure to help us choose the final model.

The best-subsets procedure shows that:

- The model with just the AdvExp and Accounts variables is good, with $R^2 = 77.51\%$.

  It may be the favorite, if the variable market potential (Poten) is difficult to measure.

- The best four-variable model (Poten, AdvExp, Share, and Accounts) also happens to be the four-variable model identified with the stepwise and forward regression procedures.

  This may very well be the statisticians' favorite, barring any limitations on the measurement of variables.

- Models with more than four variables provide only marginal gains in R^2.

Multiple Regression Approach to Experimental Design

The use of dummy variables in a multiple regression equation can provide another approach to solving analysis of variance and experimental design problems.

We will apply the multiple regression approach to experimental design to the Chemitech, Inc. completely randomized design problem we introduced in Chapter 13 (DATAfile: *Chemitech.)*

The table to the top right shows the number of filtration systems assembled by the 15 workers using one of the three methods.

Method	A	B	C
	58	58	48
	64	69	57
	55	71	59
	66	64	47
	67	68	49
$\bar{x}$	62	66	52

The sample means are also summarized for each method. The experiment includes $k = 3$ treatments; thus, we will define $k = 3 - 1 = 2$ dummy variables as follows:

A = 1, B = 0: Observation is associated with assembly method A

A = 0, B = 1: Observation is associated with assembly method B

A = 0, B = 0: Observation is associated with assembly method C

Multiple Regression Model for the Chemitech Completely Randomized Design

We can use the dummy variables to relate the number of units assembled per week, y, to the method of assembly the employee uses.

$$E(y) = \text{Mean number of units assembled per week} = \beta_0 + \beta_1 A + \beta_2 B$$

For each of the three methods, the multiple regression equation reduces to:

Method A: $E(y) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$

Method B: $E(y) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$

Method C: $E(y) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$

Thus, $\beta_0 + \beta_1$, $\beta_0 + \beta_2$, and β_0 represent the mean number of units assembled with method A, B, and C, respectively.

We can now estimate the coefficients β_0, β_1, and β_2 and develop an estimate of the mean number of units assembled per week for each method running multiple regression on the shown 15 observations (DATAfile: *Chemitech2*.)

Multiple Regression Output for the Chemitech Completely Randomized Design

From the multiple regression output we see that the estimates of β_0, β_1, and β_2 are $b_0 = 52$, $b_1 = 10$, and $b_2 = 14$.

The best estimate of the mean number of units assembled per week for each assembly method is

Assembly Method	**Prediction of $E(y)$**
A	$b_0 + b_1 = 52 + 10 = 62$
B	$b_0 + b_2 = 52 + 14 = 66$
C	$b_0 = 52$

Note that the above estimate of the mean number of units produced with each of the three assembly methods is the same as the sample mean.

Regression Statistics	
Multiple R	77.76%
R Square	60.47%
Adj. R Square	53.88%
Standard Error	5.323
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	2	520	260.00	9.176	0.004
Residual	12	340	28.33		
Total	14	860			

	Coeffs	*Std. Err.*	*t Stat*	*P-value*
Intercept	52	2.38	21.84	0.000
A	10	3.37	2.97	0.012
B	14	3.37	4.16	0.001

ANOVA Test with Multiple Regression Analysis

If the means for methods A, B, and C do not differ, we have:

$$E(y) \text{ for method A} - E(y) \text{ for method C} = 0$$

$$E(y) \text{ for method } B - E(y) \text{ for method C} = 0$$

Because of what previously observed for the multiple regression equation, we have:

$$(\beta_0 + \beta_1) - \beta_0 = \beta_1 = 0$$

$$(\beta_0 + \beta_2) - \beta_0 = \beta_2 = 0$$

Thus, the null hypothesis for a test for difference of means can be stated as $H_0\colon \beta_1 = \beta_2 = 0$.

Using a significance level $\alpha = 0.05$, the computer output for the multiple regression shows $F = 9.18$ and a *p*-value of 0.004 for the F test. Hence, we reject H_0 and conclude that the means for the three assembly methods are not the same.

Similarly, we can also use the regression *t* tests on β_1 and β_2 to conclude that the difference between the means for methods A and C, and for methods B and C, are significantly different.

Autocorrelation

Often, the data used for regression studies in business and economics are collected over time.

It is not uncommon for the value of y at one time period to be related to the value of y at previous time periods.

When that happens, we say **autocorrelation** (or **serial correlation**) is present in the data.

- First-order autocorrelation is present when the value of y in time period *t* is related to its value in time period $t - 1$.
- Second-order autocorrelation is present when the value of y in time period *t* is related to its value in time period $t - 2$.
- And so on.

One of the assumptions of the regression model is that the error terms are independent. However, when autocorrelation is present, this assumption is violated.

Without data independence, serious errors can be made in performing tests of significance based upon the assumed regression model.

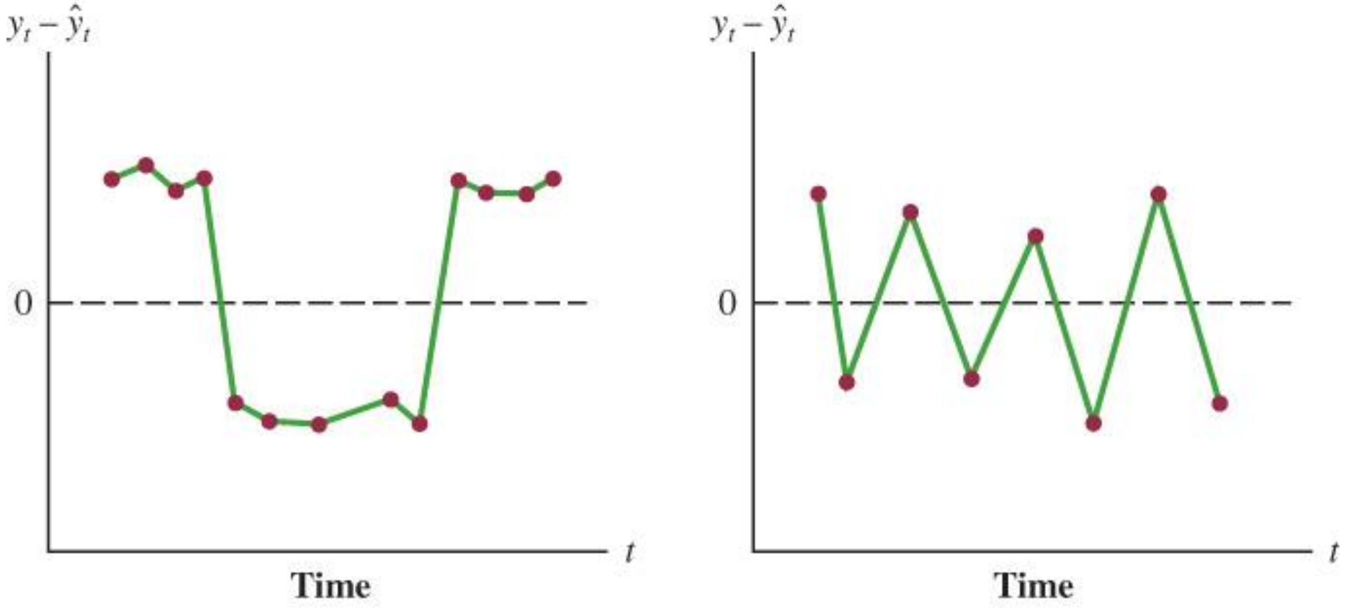
First-Order Autocorrelation

With positive autocorrelation, as shown in Panel A, we expect

- a positive residual in one period to be followed by a positive residual in the next period.
- a negative residual in one period to be followed by a negative residual in the next period.

With negative autocorrelation, as shown in Panel B, we expect a positive residual in one period to be followed by a negative residual in the next period, then a positive residual, and so on.

The Durbin-Watson statistic can be used to detect first-order autocorrelation.



Panel A. Positive Autocorrelation

Panel B. Negative Autocorrelation

The Durbin-Watson Test Statistic

The Durbin-Watson test statistic is defined as

$$\boldsymbol{d} = \frac{\sum_{t=2}^{n}(\boldsymbol{e}_t - \boldsymbol{e}_{t-1})^2}{\sum_{t=1}^{n}\boldsymbol{e}_t^2}$$

Where the residual for observation *t* is denoted

$$e_t = y_t - \hat{y}_t.$$

The Durbin-Watson test statistic has the following properties:

- The statistic ranges in value from zero to four.
- If positive autocorrelation is present, the statistic will be small (approaching zero.)
- If negative autocorrelation is present, the statistic will be large (approaching four.)
- A value near two indicates no autocorrelation.

The Durbin-Watson Test

Suppose that the values of ϵ (residuals) are not independent but are related as follows:

$$e_t = \rho e_{t-1} + z_t$$

Where ρ is a parameter with an absolute value less than one and z_t is a normally and independently distributed random variable with a mean of zero and variance of σ^2.

It follows that, if $\rho = 0$, the error terms are not related, as e_t also becomes a random variable.

The **Durbin-Watson test** uses the residuals to determine whether $\rho = 0$, by testing the following null hypothesis:

$$H_0\colon\ \rho = 0$$

There are three forms of alternative hypothesis, depending on whether we are testing for

positive autocorrelation:	$H_a\colon\ \rho > 0$
negative autocorrelation:	$H_a\colon\ \rho < 0$
positive or negative autocorrelation:	$H_a\colon\ \rho \neq 0$

Hypothesis Test for Autocorrelation Using the Durbin-Watson Test

Positive autocorrelation test (Panel A):

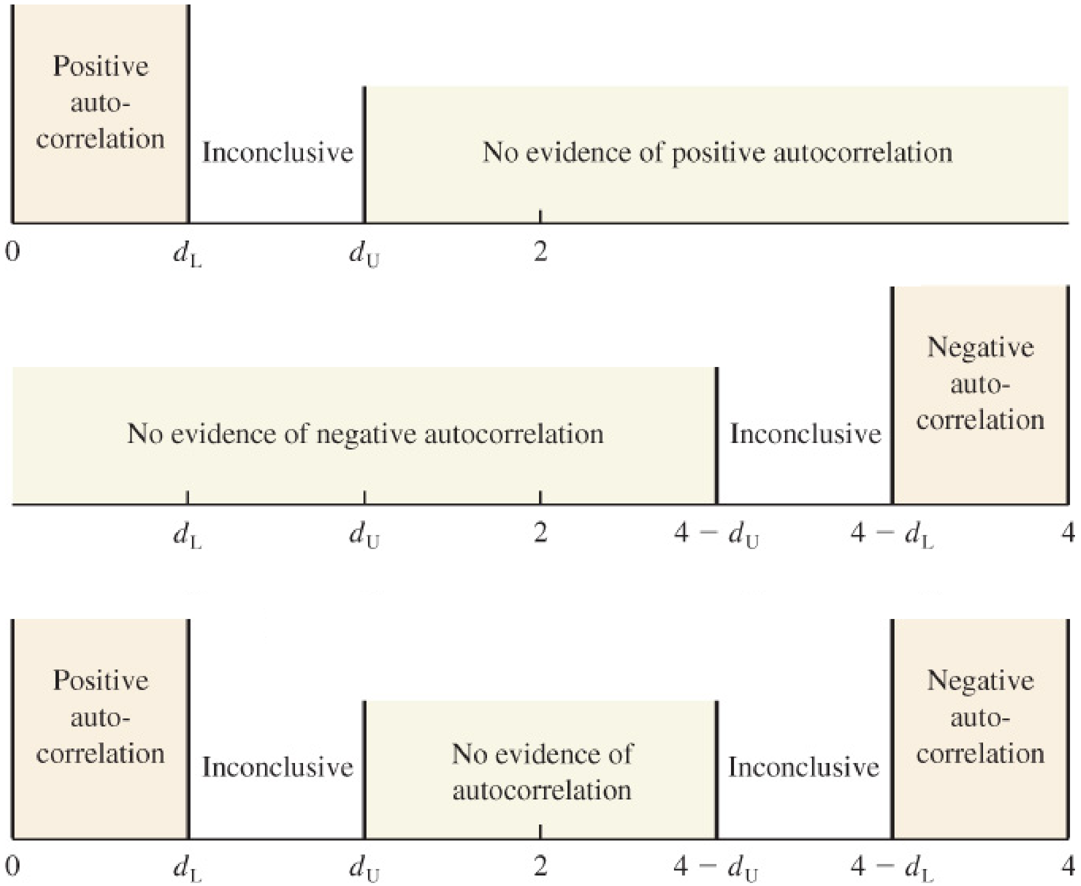
If $d < d_L$, positive autocorrelation is present. If $d_L \le d \le d_U$, test is inconclusive. If $d > d_U$, no evidence of positive autocorrelation.

Negative auto-correlation test (Panel B):

If $d > 4 - d_L$, negative autocorrelation is present. If $4 - d_U \le d \le 4 - d_L$, test is inconclusive. If $d < 4 - d_U$, no evidence of negative autocorrelation.

Two-tailed autocorrelation test (Panel C):

If $d < d_L$ or $d > 4 - d_L$ auto-correlation is present. If $d_L \le d \le d_U$ or $4 - d_U \le d \le 4 - d_L$, test is inconclusive. If $d_U < d < 4 - d_U$, no evidence of autocorrelation.


Positive auto-correlation
Inconclusive
No evidence of positive autocorrelation
0
d_L
d_U
2
No evidence of negative autocorrelation
Inconclusive
Negative auto-correlation
d_L
d_U
2
$4 - d_U$
$4 - d_L$
4
Positive auto-correlation
Inconclusive
No evidence of autocorrelation
Inconclusive
Negative auto-correlation
0
d_L
d_U
2
$4 - d_U$
$4 - d_L$
4

Critical Values for the Durbin-Watson Test for Autocorrelation

The one-tailed Durbin-Watson hypothesis test for autocorrelation uses the following lower and upper critical values (denoted d_L and d_U) at a significance level, $\alpha = 0.05$, and for a sample with *n* observations. For a two-tailed test, the level of significance is doubled.

	Number of Independent Variables									
	1		**2**		**3**		**4**		**5**	
n***	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Jefe de Cocina Example

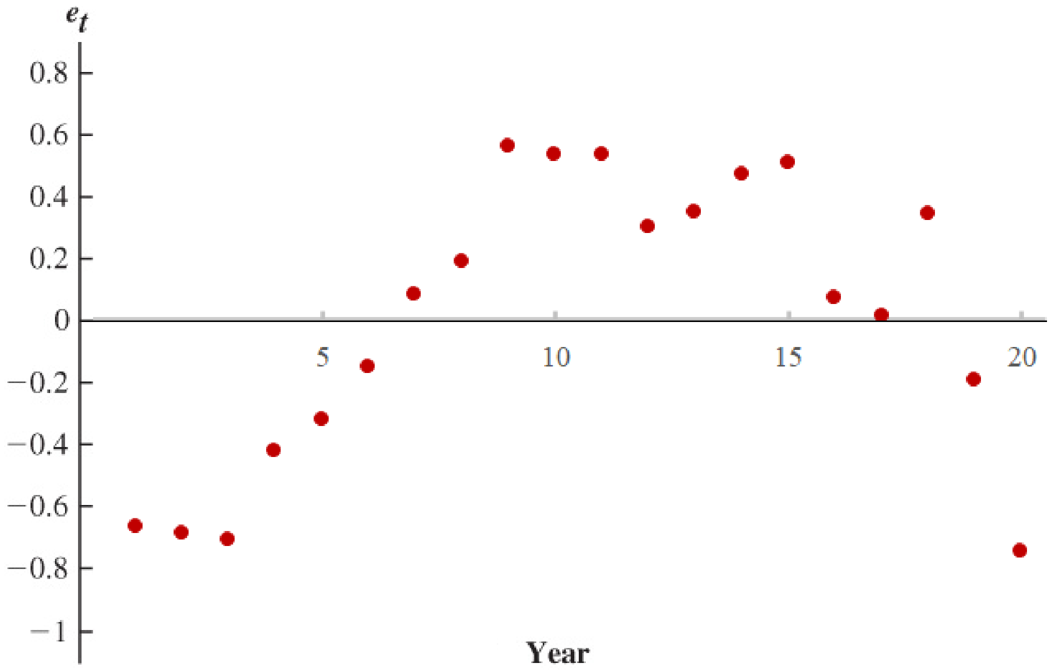
Jefe de Cocina provides locally sourced meals in waste-free packaging to regional households.

Using data collected on Jefe de Cocina's sales (in $ million) and the regional population's median income over the 20-year history of the company, an industry analyst has estimated the following regression model, with $R^2 = 97\%$.

$$\text{Sales} = -12.60671 + 0.00028\ \text{Income}$$

Based on the regression results, the industry analyst concludes that Jefe de Cocina's sales have been boosted by an increase in the disposable income of people living in its region.

However, because the data have been collected over time, the residual plot vs. Year indicates a presence of strong positive autocorrelation in the residuals.



Jefe de Cocina Durbin-Watson Test

The table to the right contains predicted sales for each of the 20 years as well as the residual calculations necessary to compute the Durbin-Watson test statistic

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = 1.7983/5.6354 = 0.319$$

From the table about the critical values, we see that, in case of one independent variable (in our case, Income), and $n = 20$, we have

$$d_L = 1.20 \text{ and } d_U = 1.41$$

Comparing $d = 0.319$ to these critical values, we observe that $d = 0.319 < 1.20 = d_L$.

We can reject the null hypothesis and conclude that the residuals are positively autocorrelated.

Year	Sales (S million)	Pred. Sales (S million)	e_t	$(e_t - e_{t-1})^2$	e_t^2
1	3.31	4.10	-0.79		0.6241
2	3.56	4.37	-0.81	0.0004	0.6561
3	3.61	4.45	-0.84	0.0009	0.7056
4	3.72	4.22	-0.50	0.1156	0.2500
5	4.04	4.42	-0.38	0.0144	0.1444
6	4.13	4.31	-0.18	0.0400	0.0324
7	4.27	4.17	0.10	0.0784	0.0100
8	4.58	4.36	0.22	0.0144	0.0484
9	5.09	4.43	0.66	0.1936	0.4356
10	5.72	5.09	0.63	0.0009	0.3969
11	6.36	5.73	0.63	0.0000	0.3969
12	6.77	6.42	0.35	0.0784	0.1225
13	7.29	6.88	0.41	0.0036	0.1681
14	8.18	7.63	0.55	0.0196	0.3025
15	8.84	8.25	0.59	0.0016	0.3481
16	9.25	9.17	0.08	0.2601	0.0064
17	10.01	10.00	0.01	0.0049	0.0001
18	11.22	10.82	0.40	0.1521	0.1600
19	12.55	12.78	-0.23	0.3969	0.0529
20	13.11			0.4225	0.7744
				1.7983	5.6354