5. **Multiple Regression** (Regressioanalyysi)
(Azcel Ch. 11, Milton/Arnold Ch. 12)

The k-Variable Multiple Regression Model

The population regression model of a dependent variable Y on a set of k independent variables $X_1, X_2, \ldots, X_k$ is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon,$$

where β_0 is the intercept and $\beta_i,\ i = 1, \ldots, k$ are the slopes of the regression surface (also called response surface) with respect to X_i.

Model assumptions:

1. $\epsilon_j \sim NID(0, \sigma^2)$ for all observations $j = 1, \ldots, n$;

2. The variables X_i are considered fixed quantities (not random variables), that is, the only randomness in Y comes from the error term ϵ.

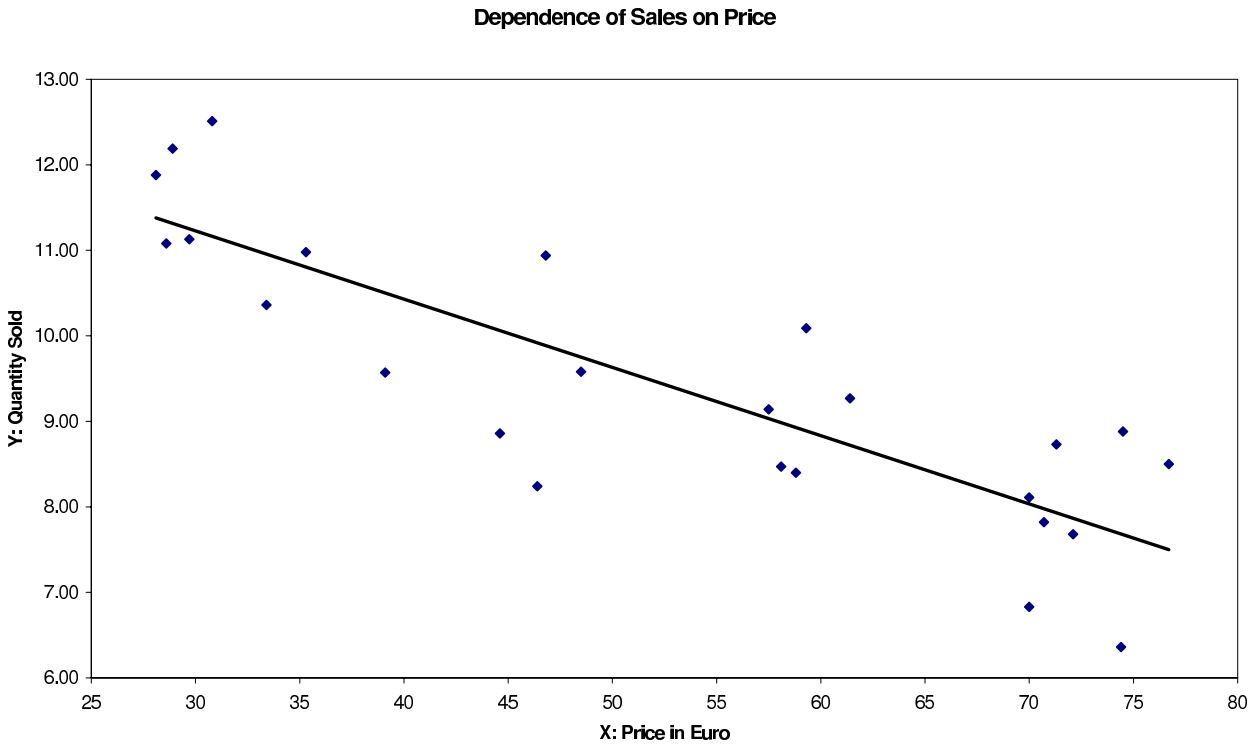
The parameters $\beta_0, \beta_1, \ldots \beta_k$ are estimated by the method of least squares (pns-menetelmä), as in the case with only 1 regressor.

Method of Least Squares with k Regressors
(Pienimmän neliösumman menetelmä)

The following table contains the quantity of a product sold (Y) as a function of the products price (X).

Observation Number	X:Price in Euro	Y: Quantity Sold
1	35.30	10.98
2	29.70	11.13
3	30.80	12.51
4	58.80	8.40
5	61.40	9.27
6	71.30	8.73
7	74.40	6.36
8	76.70	8.50
9	70.70	7.82
10	57.50	9.14
11	46.40	8.24
12	28.90	12.19
13	28.10	11.88
14	39.10	9.57
15	46.80	10.94
16	48.50	9.58
17	59.30	10.09
18	70.00	8.11
19	70.00	6.83
20	74.50	8.88
21	72.10	7.68
22	58.10	8.47
23	44.60	8.86
24	33.40	10.36
25	28.60	11.08

Such data may be conveniently illustrated in a so called scatterplot (hajontakuvio).



It appears, that the number of units sold Y is roughly a linear function of the price X, that is,

$$y_j \approx b_0 + b_1 x_j, \qquad \text{or}$$

$$y_j = b_0 + b_1 x_j + u_j, \quad j = 1 \ldots 25, \qquad (1)$$

where the residuals (jäännökset) u_j are small in some sense compared to the linear term $b_0 + b_1 x_j$. We shall in the following consider a technique of identifying a linear relationship in approximately linearly distributed data, known as Method of Least Squares or Regression Analysis, (pienimmän neliösumman menetelmä, regressioanalyysi).

Before doing that, let us rewrite the regression equation (1) in matrix form as

$$\mathbf{y}_{(25\times1)} = \mathbf{X}_{(25\times2)}\mathbf{b}_{(2\times1)} + \mathbf{u}_{(25\times1)}, \text{ where: } (2)$$

$$\mathbf{y} = \begin{pmatrix} 10.98 \\ 11.13 \\ 12.51 \\ \vdots \\ 10.36 \\ 11.08 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & 35.3 \\ 1 & 29.7 \\ 1 & 30.8 \\ \vdots & \\ 1 & 33.4 \\ 1 & 28.6 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \quad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{24} \\ u_{25} \end{pmatrix}.$$

Doing so has the advantage that we may easily generalize our method to the case where the dependent variable Y, called regressand (selitettävä muuttuja), depends linearly on more than one independent variable X, called regressor (selittävä muuttuja). For example, the number of units sold might not only depend on a constant ($X_1 = \mathbf{1}$) and price (X_2), but also on advertisement (X_3), bonuses to sales officers (X_4), and so on, such that

$$y_j = b_0 \cdot 1 + b_1 x_{1,j} + b_2 x_{2,j} + \ldots + b_k x_{k,j} + u_j \quad (3)$$

for observation $j \in (1, \ldots, n)$ of n observations and $k + 1$ regressors (including the 1).

The matrix formulation in this general case remains exactly the same as before except, that we need to add additional elements to $\mathbf{X}$ and $\mathbf{b}$ in order to incorporate the additional regressors, that is,

$$\mathbf{y}_{(n\times 1)} = \mathbf{X}_{(n\times(k+1))}\mathbf{b}_{((k+1)\times 1)} + \mathbf{u}_{(n\times 1)}, \quad \text{with} \tag{4}$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{k,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{k,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & x_{2,n} & \cdots & x_{k,n} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix},$$

$$\text{and as before: } \mathbf{y}' = (y_1, y_2, \ldots, y_n),$$
$$\mathbf{u}' = (u_1, u_2, \ldots, u_n).$$

Note that the so called design matrix $\mathbf{X}$ has one column more than the number of of regressors, because one column is needed for the constant term b_0 in the linear specification $y_j = b_0 \cdot 1 + \sum_{i=1}^{k} b_i x_{i,j} + u_j$.

We are now in a position to discuss the method of least squares in order to obtain estimates for the unknown parameter vector $\mathbf{b}$ from the observed vector $\mathbf{y}$ and the design matrix $\mathbf{X}$ for the general case of k regressors.

The method of least squares determines the unknown parameters $\beta_0, \ldots, \beta_k$ by minimizing the sum of all squared vertical distances between the observations y_j and their predicted values $\widehat{y_j} := b_0 + \sum_{i=1}^{k} b_i x_{i,j}$ on the regression surface.

That is, we seek to minimize the sum of squared residuals

$$f(\mathbf{b}) := \sum_{j=1}^{n} u_j^2 = \mathbf{u}'\mathbf{u} = (u_1, u_2, \ldots, u_n) \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \tag{5}$$

which yields, recalling $\mathbf{u} = \mathbf{y} - \mathbf{X}\mathbf{b}$:

$$\begin{aligned} f(\mathbf{b}) &= (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}. \end{aligned} \tag{6}$$

Now a necessary condition for a minimum of the function $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{b}$ is that $\frac{\partial f}{\partial \mathbf{b}} = 0$.

In order to check this condition we need to make use of the matrix differentiation rules

$$\frac{\partial(\mathbf{b}'\mathbf{a})}{\partial \mathbf{b}} = \mathbf{a} \qquad \text{and} \qquad (7a)$$

$$\frac{\partial(\mathbf{b}'\mathbf{A}\mathbf{b})}{\partial \mathbf{b}} = 2\mathbf{A}\mathbf{b} \quad \text{for } \mathbf{A} = \mathbf{A}'. \qquad (7b)$$

Applying these to

$$f(\mathbf{b}) = \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

yields

$$\frac{\partial f}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} \stackrel{!}{=} \mathbf{0}_{((k+1)\times 1)}, \text{ or} \quad (8)$$

$$\boxed{(\mathbf{X}'\mathbf{X})_{((k+1)\times(k+1))}\mathbf{b}_{((k+1)\times 1)} = \mathbf{X}'_{((k+1)\times n)}\mathbf{y}_{(n\times 1)},} \quad (9)$$

which are called the normal equations (normaaliyhtälöt). Solving this matrix equation (or set of scalar equations) yields the sought parameter vector $\mathbf{b}' = (b_0, b_1, \ldots, b_k)$.

Normal Equations in Scalar Form

We shall here only consider the case of 2 independent variables, that is:

$$(\mathbf{X}'\mathbf{X})_{(3\times3)}\mathbf{b}_{(3\times1)} = \mathbf{X}'_{(3\times n)}\mathbf{y}_{(n\times1)}, \quad \text{where}$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{1,n} & x_{2,n} \end{pmatrix}_{(n\times3)}, \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Now

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 1 & \cdots & 1 \\ x_{1,1} & \cdots & x_{1,n} \\ x_{2,1} & \cdots & x_{2,n} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum y_j \\ \sum x_{1,j}y_j \\ \sum x_{2,j}y_j \end{pmatrix}$$

and

$$\begin{aligned} \mathbf{X}'\mathbf{X} &= \begin{pmatrix} 1 & \cdots & 1 \\ x_{1,1} & \cdots & x_{1,n} \\ x_{2,1} & \cdots & x_{2,n} \end{pmatrix} \begin{pmatrix} 1 & x_{1,1} & x_{2,1} \\ \vdots & \vdots & \vdots \\ 1 & x_{1,n} & x_{2,n} \end{pmatrix} \\ &= \begin{pmatrix} n & \sum x_{1,j} & \sum x_{2,j} \\ \sum x_{1,j} & \sum x_{1,j}^2 & \sum x_{1,j}x_{2,j} \\ \sum x_{2,j} & \sum x_{1,j}x_{2,j} & \sum x_{2,j}^2 \end{pmatrix}, \end{aligned}$$

such that

$$\mathbf{X'Xb} = \begin{pmatrix} n & \sum x_{1,j} & \sum x_{2,j} \\ \sum x_{1,j} & \sum x_{1,j}^2 & \sum x_{1,j}x_{2,j} \\ \sum x_{2,j} & \sum x_{1,j}x_{2,j} & \sum x_{2,j}^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}$$

$$= \begin{pmatrix} nb_0 & + & b_1 \sum x_{1,j} & + & b_2 \sum x_{2,j} \\ b_0 \sum x_{1,j} & + & b_1 \sum x_{1,j}^2 & + & b_2 \sum x_{1,j}x_{2,j} \\ b_0 \sum x_{2,j} & + & b_1 \sum x_{1,j}x_{2,j} & + & b_2 \sum x_{2,j}^2 \end{pmatrix}$$

and the scalar form of the normal equations $\mathbf{X'y} = \mathbf{X'Xb}$ becomes:

$$\begin{aligned} \sum y_j &= nb_0 + b_1 \sum x_{1,j} + b_2 \sum x_{2,j}, \\ \sum x_{1,j}y_j &= b_0 \sum x_{1,j} + b_1 \sum x_{1,j}^2 + b_2 \sum x_{1,j}x_{2,j}, \\ \sum x_{2,j}y_j &= b_0 \sum x_{2,j} + b_1 \sum x_{1,j}x_{2,j} + b_2 \sum x_{2,j}^2; \end{aligned}$$

where all summations extend from the 1st to the nth observation.

Solving these equations yields the estimates b_0, b_1 and b_2 for the parameters β_0, β_1 and β_2 of the 2-variable regression model.

Solving the Normal Equations

To find the least-square estimates for $\beta_0, \ldots, \beta_k$ we need to solve the normal equations

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}.$$

We know that if the columns of X are linearly independent, that is, no column can be expressed as a linear combination of the others, then $(\mathbf{X}'\mathbf{X})$ has an inverse, which we shall denote by $(\mathbf{X}'\mathbf{X})^{-1}$. To solve the normal equations for $\mathbf{b}$, we premultiply both sides of the normal equations above by $(\mathbf{X}'\mathbf{X})^{-1}$ to obtain

$$\widehat{\beta}_{((k+1)\times 1)} = \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

the components of which are the sought least-square estimates $b_0, \ldots, b_k$ for $\beta_0, \ldots, \beta_k$.

As it is no easy task to find the inverse of a matrix by hand except in the simplest cases, the calculation of the expression above is in practice left to the statistical software we use.

Example: Milton/Arnold Examples 12.2.1–3
An equation is to be developed from which we can predict the gasoline mileage of an automobile as a linear function of its weight and temperature at the time of operation.

Car No.	1	2	3	4	5	6	7	8	9	10
mpg(y)	17.9	16.5	16.4	16.8	18.8	15.5	17.5	16.4	15.9	18.3
(x_1/tons)	1.35	1.90	1.70	1.80	1.30	2.05	1.60	1.80	1.85	1.40
($x_2/^o$F)	90	30	80	40	35	45	50	60	65	30

The model specification matrix $\mathbf{X}$, vector of parameter estimates $\mathbf{b}$, and vector of responses $\mathbf{y}$ are:

$$\mathbf{X} = \begin{pmatrix} 1 & 1.35 & 90 \\ 1 & 1.90 & 30 \\ \vdots & \vdots & \vdots \\ 1 & 1.40 & 30 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 17.9 \\ 16.5 \\ \vdots \\ 18.3 \end{pmatrix}.$$

We wish to solve the normal equations $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}$ by calculating $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. To this end, note that

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 1 & \cdots & 1 \\ 1.35 & \cdots & 1.40 \\ 90 & \cdots & 30 \end{pmatrix} \begin{pmatrix} 17.9 \\ 16.5 \\ \vdots \\ 18.3 \end{pmatrix} = \begin{pmatrix} 170 \\ 282.405 \\ 8887 \end{pmatrix}.$$

Furthermore,

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & \cdots & 1 \\ 1.35 & \cdots & 1.40 \\ 90 & \cdots & 30 \end{pmatrix} \begin{pmatrix} 1 & 1.35 & 90 \\ \vdots & \vdots & \vdots \\ 1 & 1.40 & 30 \end{pmatrix}$$
$$= \begin{pmatrix} 10 & 16.75 & 525 \\ 16.75 & 28.6375 & 874.5 \\ 525 & 874.5 & 31\,475 \end{pmatrix}$$

with inverse matrix

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 6.070769 & -3.02588 & -0.0171888 \\ -3.02588 & 1.738599 & 0.0021663 \\ -0.017189 & 0.002166 & 0.0002583 \end{pmatrix}.$$

The vector of parameter estimates is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$
$$= \begin{pmatrix} 6.070769 & -3.02588 & -0.0171888 \\ -3.02588 & 1.738599 & 0.0021663 \\ -0.017189 & 0.002166 & 0.0002583 \end{pmatrix} \begin{pmatrix} 170 \\ 282.405 \\ 8887 \end{pmatrix}$$
$$= \begin{pmatrix} 24.75 \\ -4.16 \\ -0.014897 \end{pmatrix}.$$

The estimated model is

$$Y = 24.75 - 4.16X_1 - 0.014897X_2 + \epsilon.$$

Based on this equation, we estimate the mileage of a car weighing 1.5 tons on a 70^{o}F day to be

$$\widehat{y} = 24.75 - 4.16 \cdot 1.5 - 0.014897 \cdot 70 = 17.47\text{mpg}.$$

The ANOVA F-test for Multiple Regression

In simple linear regression the F test from the ANOVA table is equivalent to the two-sided test of the hypothesis that the slope of the regression line is 0. For multiple regression there is a corresponding ANOVA F test, but it tests the hypothesis that *all* regression coefficients (except the intercept β_0) are 0.

The ANOVA table for multiple regression is

Source	Sum of Squares	DF	Mean Square	F Ratio
Regression	$\sum(\hat{y}_i - \bar{y})^2$	k	SSR/DFR	MSR/MSE
Error	$\sum(y_i - \hat{y}_i)^2$	$n-(k+1)$	SSE/DFE	
Total	$\sum(y_i - \bar{y})^2$	$n-1$	SST/DFT	

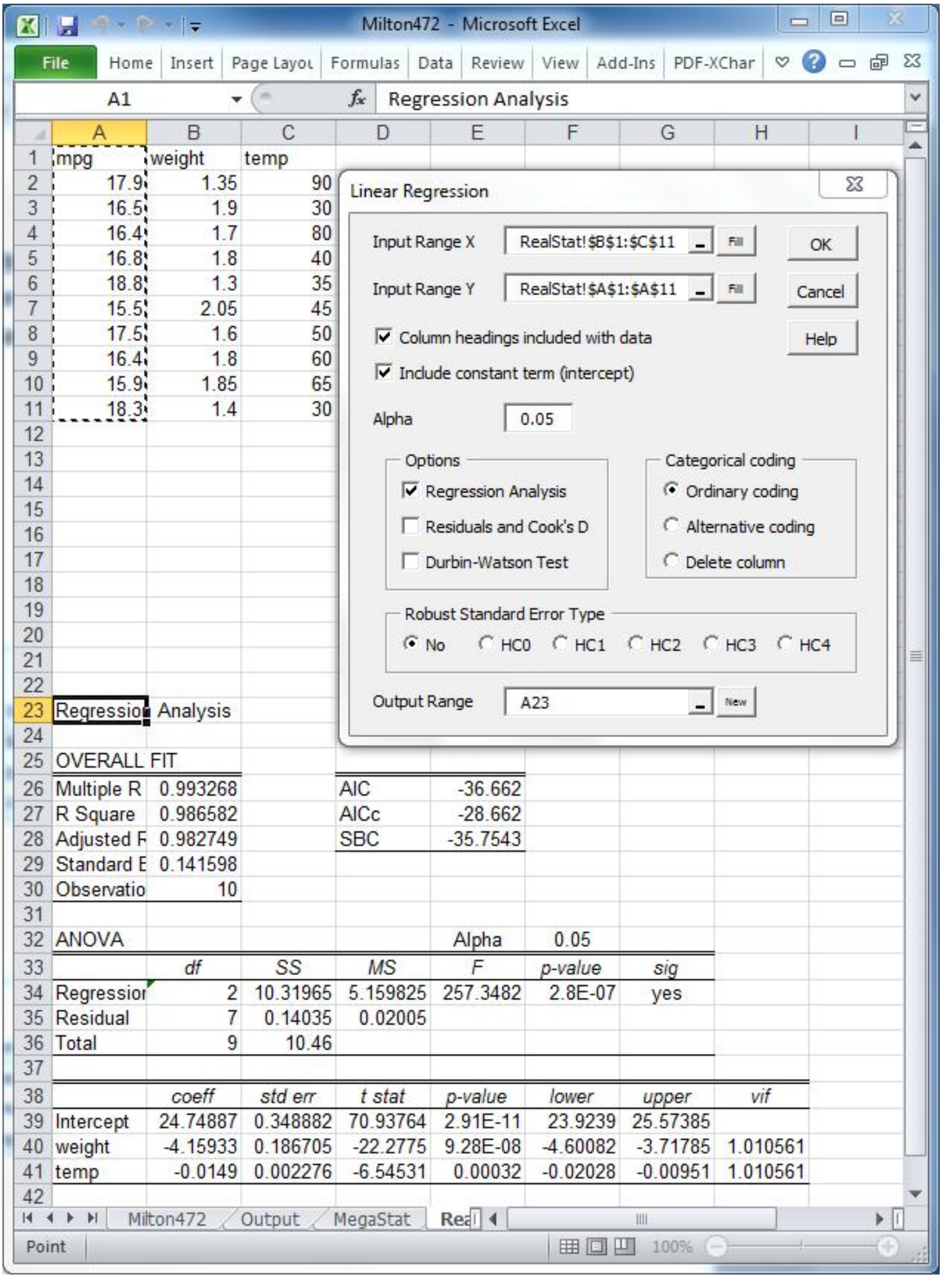
The ratio $\frac{\text{MSR}}{\text{MSE}}$ is an F statistic for testing

$$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

against

$$H_1\colon\ \text{Not all } \beta_i,\ i = 1, \ldots, k \text{ are zero.}$$

Under H_0: $\quad F = \dfrac{\text{MSR}}{\text{MSE}} \sim F(k, n-k-1).$



	A	B	C
1	mpg	weight	temp
2	17.9	1.35	90
3	16.5	1.9	30
4	16.4	1.7	80
5	16.8	1.8	40
6	18.8	1.3	35
7	15.5	2.05	45
8	17.5	1.6	50
9	16.4	1.8	60
10	15.9	1.85	65
11	18.3	1.4	30

Regression Analysis

OVERALL FIT

Multiple R	0.993268	AIC	-36.662
R Square	0.986582	AICc	-28.662
Adjusted F	0.982749	SBC	-35.7543
Standard E	0.141598		
Observatio	10		

ANOVA — Alpha 0.05

	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regressior	2	10.31965	5.159825	257.3482	2.8E-07	yes
Residual	7	0.14035	0.02005			
Total	9	10.46				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*	*vif*
Intercept	24.74887	0.348882	70.93764	2.91E-11	23.9239	25.57385	
weight	-4.15933	0.186705	-22.2775	9.28E-08	-4.60082	-3.71785	1.010561
temp	-0.0149	0.002276	-6.54531	0.00032	-0.02028	-0.00951	1.010561

How good is the regression?

The mean square error (jäännösvarianssi)

$$\text{MSE} = \frac{\text{SSE}}{n-(k+1)} = \frac{\sum(y_j - \widehat{y}_j)^2}{n-(k+1)}$$

displayed in the ANOVA table is an unbiased estimator of the variance σ^2 of the population errors ϵ. The square root of MSE is an estimator of the standard deviation σ, usually denoted by s and referred to as the standard error (keskivirhe) of estimate

$$s = \sqrt{\text{MSE}}.$$

The mean square error and its square root are measures of the size of the errors in regression but give no indication about the explained component of the regression fit.

As in the case with only one regressor, we measure the regression fit by

$$R^2 = \frac{s^2_{\widehat{y}}}{s^2_y} = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}},$$

the (multiple) coefficient of determination. (Note that again: SST = SSR + SSE.)

The multiple coefficient of determination R^2 measures the quality of the regression fit as the proportion of the variation in the dependent variable that is explained by the linear combination of the independent variables. Adding new variables to the model can never decrease the amount of variance explained, therefore R^2 will always increase when we add new variables (it will only stay constant if the variables we added are completely useless).

For comparing models with varying numbers of regressors it is useful to have a measure of regression fit which decreases under addition of variables of low explanatory power. Such is given by the adjusted multiple coefficient of determination (tarkistettu selitysaste)

$$\overline{R^2} = 1 - \frac{\text{MSE}}{\text{MST}} = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)},$$

which is related to the ordinary R^2 by

$$\overline{R^2} = 1 - (1 - R^2)\frac{n-1}{n-(k+1)}.$$

Note that the F test for $\beta_1 = \ldots = \beta_k = 0$ may just as well be regarded as a test of $R^2 = 0$, since

$$F = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)} = \frac{R^2}{1-R^2} \cdot \frac{n-(k+1)}{k}.$$

Example: (continued.)

$$\text{MSE} = \frac{\text{SSE}}{n-(k+1)} = \frac{0.14}{10-3} = 0.02$$

$$s = \sqrt{\text{MSE}} = 0.141\ldots$$

$$R^2 = \frac{10.32}{10.46} = 1 - \frac{0.14}{10.46} = 0.9866$$

$$\overline{R^2} = 1 - \frac{\text{MSE}}{\text{MST}} = 1 - \frac{0.02}{10.46/9} = 0.9828$$

Alternatively:

$$\overline{R^2} = 1-(1-R^2)\frac{n-1}{n-(k+1)} = 1-0.0134\cdot\frac{9}{7} = 0.9828$$

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-(k+1)}{k}$$
$$= \frac{0.9866}{1-0.9866} \cdot \frac{10-3}{2} = 257.7$$

The difference to $F = 257.3$ in the ANOVA table is due to rounding error.

Inference in Multiple Regression

Properties of the Least-Squares Estimators

Now that we learnt how to find the regression parameters (by the method of least squares) and how to assess the usefulness of the regression as a whole (by the ANOVA F test) we would also like to be able to tell for individual regression parameters whether they are statistically significant or not. This requires us to learn something about the sampling distribution of the least square estimator $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ for the unknown parameter vector β.

For that purpose we define the expected value $E(\mathbf{Y})$ of a vector of random variables $\mathbf{Y}$, that is, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$, as

$$E(\mathbf{Y}) = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{pmatrix}.$$

The calculation rules for expectations of random vectors resemble those of scalar expectations:

1. $E(C) = C$,
2. $E(C\mathbf{Y}) = CE(\mathbf{Y})$ ($\Rightarrow\ E(\mathbf{Y}'C') = E(\mathbf{Y}')C'$),
3. $E(\mathbf{Y} + \mathbf{Z}) = E(\mathbf{Y}) + E(\mathbf{Z})$;

where $\mathbf{Y}$ and $\mathbf{Z}$ denote $(n \times 1)$ random vectors and C denotes an $(m \times n)$ matrix of constants.

These rules may be used to show that the least squares estimator $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is an unbiased estimator of $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$ in the multiple regression model $Y = \mathbf{X}\beta + \epsilon$:

$$\begin{aligned} E(\mathbf{b}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon)] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta] + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon)] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\epsilon) \\ &= \beta. \end{aligned}$$

Before discussing the variance of $\mathbf{b}$, let us first refresh the definition and some calculation rules for the variance of scalar random variables:

1. $V(X) := E[(X-E(X))^2] = E(X^2)-E(X)^2$,
2. $V(c) = 0$ for c constant,
3. $V(aX + b) = a^2V(X)$ for a, b constants,
4. $V(X + Y) = V(X) + V(Y)$
   for X and Y independent.

If X and Y are not independent, then:

$$
\begin{aligned}
& V(X+Y) \\
=& E\left\{[(X+Y)-E(X+Y)]^2\right\} \\
=& E\left\{[(X-E(X))+(Y-E(Y))]^2\right\} \\
=& E[(X-E(X))^2]+E[(Y-E(Y))^2] \\
& +2E[(X-E(X))(Y-E(Y))] \\
=& V(X)+V(Y)+2E[(X-E(X))(Y-E(Y))].
\end{aligned}
$$

Thus, unlike the mean, the variance of a sum of two random variables is, in general, not the sum of the variances. The quantity

$$\text{Cov}(X, Y) := E[(X - E(X))(Y - E(Y))]$$

is called the covariance of X and Y.

Thus, we obtain the variance of a sum as:

$$V(X + Y) = V(X) + V(Y) + 2\text{Cov}(X, Y).$$

From the definition of covariance we obtain the two immediate consequences

$$\text{Cov}(X, X) = V(X), \ \text{Cov}(X, Y) = \text{Cov}(Y, X).$$

Furthermore,

$$(X-E(X))(Y-E(Y)) = XY - XE(Y) - YE(X) + E(X)E(Y),$$

and hence by taking expectations we see that

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

This implies that $\text{Cov}(X, Y) = 0$ whenever X and Y are independent, since then

$$E(XY) = E(X)E(Y).$$

For random vectors it is convenient to collect all covariances between the components of the vector in a single matrix, the so called variance-covariance matrix defined by $\text{Var}(\mathbf{Y})$

$$:= \begin{pmatrix} V(Y_1) & \text{Cov}(Y_1, Y_2) & \dots & \dots & \text{Cov}(Y_1, Y_n) \\ \text{Cov}(Y_1, Y_2) & V(Y_1) & & & \text{Cov}(Y_2, Y_n) \\ \text{Cov}(Y_1, Y_3) & \text{Cov}(Y_2, Y_3) & V(Y_3) & & \text{Cov}(Y_3, Y_n) \\ \vdots & \vdots & & \ddots & \vdots \\ \text{Cov}(Y_1, Y_n) & \text{Cov}(Y_2, Y_n) & \dots & \dots & V(Y_n) \end{pmatrix},$$

where $\mathbf{Y}$ denotes again a vector of random variables $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)'$.

Similiar to the calculation rules for variances of scalar random variables we have the following important matrix rule for variance:

$$\text{Var}(C\mathbf{Y} + \mathbf{d}) = C\text{Var}(\mathbf{Y})C'$$

where $\mathbf{Y}$ is again an $(n \times 1)$ random vector, C is an $(m \times n)$ constant matrix, and $\mathbf{d}$ an $(n \times 1)$ constant vector.

The assumption of the multiple regression model that $Y_1, Y_2, \ldots, Y_n$ are independent with common variance σ^2 may now be written as

$$\text{Var}\,(\mathbf{Y}) = \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma^2 \end{pmatrix} = \sigma^2 I,$$

where I is the $(n \times n)$ identity matrix, a matrix of 1's on the main diagonal and 0 elsewhere.

Recall that our least squares estimator $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is of the form $C\mathbf{Y}$ with $C = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. We may therefore use the rule $\text{Var}\,(C\mathbf{Y}) = C\text{Var}\,(\mathbf{Y})C'$, such that

$$\begin{aligned} \text{Var}\,(\mathbf{b}) &= \text{Var}\,[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}\,(\mathbf{Y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']', \end{aligned}$$

where, using $(AB)' = B'A'$ and $(A^{-1})' = (A')^{-1}$:
$[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' = X[(\mathbf{X}'\mathbf{X})^{-1}]' = X[(\mathbf{X}'\mathbf{X})']^{-1} = X(\mathbf{X}'\mathbf{X})^{-1}$.

$$\begin{aligned}\text{Var}\,(\mathbf{b}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}\,(\mathbf{Y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}\,(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2 I\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

Since σ^2 is unknown, we replace it by our usual estimator, the mean square error

$$s^2 = \text{MSE} = \frac{\text{SSE}}{n-(k+1)}$$

in order to obtain

$$\widehat{\text{Var}}(\mathbf{b}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

as our estimator for the variance covariance matrix of the least square parameter estimates $\mathbf{b} = (b_0, b_1, \ldots, b_k)'$ for the unknown parameter vector $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$.

The diagonal elemements $s^2(\mathbf{X}'\mathbf{X})^{-1}_{ii}$ of $\widehat{\text{Var}}(\mathbf{b})$ are the estimated variances of $b_0, b_1, \ldots, b_k$. Their standard errors are therefore given by

$$SE_{b_{i-1}} = s\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{ii}} = \sqrt{\text{MSE}(\mathbf{X}'\mathbf{X})^{-1}_{ii}}.$$

Example: (continued.)

We found earlier that MSE=0.02 and

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 6.070769 & -3.02588 & -0.0171888 \\ -3.02588 & 1.738599 & 0.0021663 \\ -0.017189 & 0.002166 & 0.0002583 \end{pmatrix}.$$

Our variance estimates for the least squares coefficients b_0, b_1 and b_2 are therefore:

$$V(\widehat{b}_0) = 0.02 \cdot 6.070769 = 0.1217,$$
$$V(\widehat{b}_1) = 0.02 \cdot 1.738599 = 0.03486,$$
$$V(\widehat{b}_2) = 0.02 \cdot 0.0002583 = 0.00000518;$$

with corresponding standard errors

$$SE_{b_0} = \sqrt{0.1217} = 0.349,$$
$$SE_{b_1} = \sqrt{0.03486} = 0.187,$$
$$SE_{b_2} = \sqrt{0.00000518} = 0.002.$$

Inference on Single Regression Parameters

Recall that X is normally distributed with parameters μ and σ^2 if its density is of the form

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

This is denoted by $X \sim N(\mu, \sigma^2)$.

We say that a random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ follows a multinormal distribution $N(\mu, \Sigma)$ if its density is of the form

$$f(\mathbf{y}) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{y}-\mu)'\Sigma^{-1}(\mathbf{y}-\mu)\right]$$

with

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}$$

denoting $\mu_i = E(Y_i)$, $\sigma_{ij} = \text{Cov}(Y_i, Y_j)$, and $\sigma_i^2 = \sigma_{ii} = V(Y_i)$ for $i, j = 1, \ldots, n$.

Multinormally distributed random vectors have the following two important properties:

1. Each single component of $\mathbf{Y}$ is normally distributed with mean μ_i and variance σ_i^2:

$$\mathbf{Y} = (Y_1, \ldots, Y_n)' \sim N(\mu, \Sigma)$$

$$\Rightarrow \; Y_i \sim N(\mu_i, \sigma_i^2).$$

2. Any arbitrary linear combination of the components of $\mathbf{Y}$ is also normally distributed. In matrix form this is written:

$$\mathbf{Z} = A\mathbf{Y} + c \sim N(\underbrace{A\mu + c}_{E(\mathbf{Z})}, \underbrace{A\Sigma A'}_{\mathrm{Var}(\mathbf{Z})}),$$

where A is any $(r \times n)$ matrix of constants and c is any $(r \times 1)$ vector of constants.

We are now in a position to restate the k-variable regression model in matrix form as:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I);$$

which implies by property 2 from above that

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2 I).$$

Our interest is in the sampling distribution of $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. We know already that

$$E(\mathbf{b}) = \beta \quad \text{and} \quad \text{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Furthermore, since $\mathbf{b}$ is just a linear combination of the components of $\mathbf{Y}$, we have again by property 2:

$$\mathbf{b} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}),$$

which implies by property 1:

$$b_i \sim N(\beta_i, \sigma^2_{b_i}), \quad i = 0, 1, \ldots, k;$$

where $\sigma^2_{b_{j-1}} = V(b_{j-1}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{jj}$ with $(\mathbf{X}'\mathbf{X})^{-1}_{jj}$ the j'th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$, and $j = i + 1$, so the first diagonal element is for b_0, the second diagonal element is for b_1, and so on.

Standardizing yields

$$z = \frac{b_i - \beta_i}{\sigma_{b_i}} \sim N(0,1).$$

Because the σ_{b_i} are unknown, we substitute $\sigma_{b_{i-1}}$ by their estimates

$$SE_{b_{i-1}} = s\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{ii}} = \sqrt{\mathsf{MSE}(\mathbf{X}'\mathbf{X})^{-1}_{ii}}.$$

and use instead the t-statistic:

$$t = \frac{b_i - \beta_i^*}{SE_{b_i}} \sim t(n-k-1) \text{ under } H_0\colon \beta_i = \beta_i^*.$$

Alternatively we may calculate $(1-\alpha)$ confidence intervals for β_i as

$$[b_i \pm t_{\frac{\alpha}{2}}(n-k-1) \cdot SE_{b_i}].$$

In particular, we may test for statistical significance of individual regression parameters by calculating the t-statistics

$$t = \frac{b_i}{SE_{b_i}} \sim t(n-k-1) \quad \text{under } H_0\colon \beta_i = 0$$

or by checking that

$$|b_i| > t_{\frac{\alpha}{2}}(n-k-1) \cdot SE_{b_i}.$$

Example: (continued.)

The t-statistics for the regression parameters β_0 (intercept), β_1 (weight), and β_2 (heat) are:

$$t_{\beta_0} = \frac{b_0}{SE_{b_0}} = \frac{24.75}{0.349} = 70.9,$$
$$t_{\beta_1} = \frac{b_1}{SE_{b_1}} = \frac{-4.159}{0.1867} = -22.3,$$
$$t_{\beta_2} = \frac{b_2}{SE_{b_2}} = \frac{-0.0149}{0.0023} = -6.5.$$

The degrees of freedom are:

$$df = n - (k + 1) = 10 - (2 + 1) = 7.$$

By calling T.INV.2T(0.001;7) in Excel or by looking up in a table we find that the 0.1% critical value of a 2-sided t-test with 7 degrees of freedom is 5.408, which is smaller than the absolute value of any of the t-statistics above. All regression parameters are therefore significant at 0.1%.

Using Multiple Regression for Prediction

Confidence Interval on Estimated Mean

We shall now find a confidence interval for the mean value of the response variable Y for a specific set of values $x_1, \ldots, x_k$ of the predictor variables, which have not necessarily been used in developing the regression equation. Let

$$\begin{aligned}\mu_{Y|\mathbf{x}} &:= E(Y|X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \\ &= \mathbf{x}'_{1\times(k+1)}\beta_{(k+1)\times 1}, \quad \text{where}\end{aligned}$$

$$\mathbf{x}' = (1, x_1, x_2, \ldots, x_k), \quad \beta = (\beta_0, \beta_1, \ldots, \beta_k)'.$$

An unbiased estimator for $\mu_{Y|\mathbf{x}}$ is

$$\widehat{\mu}_{Y|\mathbf{x}} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k = \mathbf{x}'\mathbf{b}.$$

The variance of $\widehat{\mu}_{Y|\mathbf{x}}$ is

$$\begin{aligned}\text{Var}\,(\widehat{\mu}_{Y|\mathbf{x}}) &= \text{Var}\,(\mathbf{x}'\mathbf{b}) = \mathbf{x}'\text{Var}\,(\mathbf{b})\mathbf{x} \\ &= \mathbf{x}'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x} = \sigma^2\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x},\end{aligned}$$

such that

$$\widehat{\mu}_{Y|\mathbf{x}} \sim N(\mathbf{x}'\beta, \sigma^2\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}).$$

Standardizing and replacing the unknown σ^2 by its estimator $s^2 = \text{MSE} = \frac{\text{SSE}}{n-(k+1)}$ yields

$$t = \frac{\widehat{\mu}_{Y|\mathbf{x}} - \mu_{Y|\mathbf{x}}}{s\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}} \sim t(n-k-1).$$

A $(1-\alpha)$ confidence interval on $\mu_{Y|\mathbf{x}}$ is thus

$$\left[\widehat{\mu}_{Y|\mathbf{x}} \pm t_{\frac{\alpha}{2}}(n-k-1)s\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}\right].$$

Example: (continued.) We estimated earlier the average gasoline mileage for a car weighing 1.5 tons operated on a 70^oF day as

$$\widehat{\mu}_{Y|\mathbf{x}} = 24.75 - 4.16 \cdot 1.5 - 0.0149 \cdot 70 = 17.47 \text{mpg}.$$

The standard error of the estimate was

$$s = \sqrt{\text{MSE}} = 0.1416.$$

We wish to find a 95% confidence interval on $\mu_{Y|\mathbf{x}}$ at

$$\mathbf{x}' = (1, 1.5, 70).$$

We know from previous work that

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 6.070769 & -3.02588 & -0.0171888 \\ -3.02588 & 1.738599 & 0.0021663 \\ -0.017189 & 0.002166 & 0.0002583 \end{pmatrix}.$$

A calculation by hand or calling MMULT in Excel yields for $\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$:

$$(1, 1.5, 70) \begin{pmatrix} 6.070769 & -3.02588 & -0.0171888 \\ -3.02588 & 1.738599 & 0.0021663 \\ -0.017189 & 0.002166 & 0.0002583 \end{pmatrix} \begin{pmatrix} 1 \\ 1.5 \\ 70 \end{pmatrix} = 0.22.$$

The $t_{\frac{0.05}{2}}$ critical value with 7 degrees of freedom is 2.365, such that a 95% confidence interval in the average gasoline milage for $x_1 = 1.5$ and $x_2 = 70$ is

$$\begin{aligned} & \widehat{\mu}_{Y|\mathbf{x}} \ \pm \ t_{\frac{\alpha}{2}} \cdot s\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}} \\ =& 17.47 \pm 2.365 \cdot 0.1416\sqrt{0.22} \\ =& 17.47 \pm 0.16 \end{aligned}$$

We can thus be 95% confident that the average gasoline milage of cars weighing 1.5 tons operated on a 70^{o}F day lies between 17.31 and 17.63 miles per gallon.

Prediction Interval on Single Response

Consider next predicting a single response $Y|\mathbf{x} = \mu_{Y|\mathbf{x}} + \epsilon$. The scalar product $\mathbf{x}'\mathbf{b}$ is also an unbiased estimator of $Y|\mathbf{x}$ since

$$E(Y|\mathbf{x}) = E(\mu_{Y|\mathbf{x}}) + E(\epsilon) = \mathbf{x}'\beta.$$

But the variance of $\widehat{Y}|\mathbf{x}$ is larger than the variance of $\widehat{\mu}_{Y|\mathbf{x}}$ due to the additional variation in ϵ. More specifically:

$$\begin{aligned}\text{Var}\,(\widehat{Y}|\mathbf{x}) &= \text{Var}\,(\widehat{\mu}_{Y|\mathbf{x}}) + \text{Var}\,(\epsilon) \\ &= \sigma^2\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x} + \sigma^2 \\ &= \sigma^2(1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}).\end{aligned}$$

That is,

$$\widehat{Y}|\mathbf{x} \sim N(\mathbf{x}'\beta, \sigma^2(1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x})).$$

A similiar argument as for the confidence interval on the estimated mean yields as a $(1-\alpha)$ prediction interval for an individual response:

$$\left[\widehat{Y}|\mathbf{x} \pm t_{\frac{\alpha}{2}}(n-k-1)s\sqrt{1+\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}\right].$$

Example: (continued.)
A 95% prediction interval for a car weighing 1.5 tons operating on a 70^oF day is

$$\begin{aligned}&[17.47 \pm 2.365 \cdot 0.1416\sqrt{1+0.22}]\\ =&[17.47 \pm 0.38]\\ =&[17.09, 17.85].\end{aligned}$$

(The corresponding confidence interval for the mean response was [17.31,17.63].)

Getting confidence intervals for the mean and individual predictions in excel requires use of the array functions MMULT for matrix multiplication and MINVERSE for calculating the matrix inverse. Before entering an array function you must mark an area exactly as large as the output matrix and finish your command with Ctrl+Shift+Enter.

Partial F Tests

Testing a Subset of Predictor Variables

In this section we shall present a test based on the F distribution (and in simple cases the t distribution) in order to test whether a subset of the original predictor variables is sufficient for prediction. Consider the regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon,$$

We refer to this model as the full model. Assume that we propose to reduce the number of predictor variables by deleting r of them, such that we obtain the reduced model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-r} X_{k-r} + \epsilon.$$

We wish to test

H_0: reduced model is appropriate
against H_1: full model is needed.

This may be done using a partial F test.

The method used to test H_0 is rather intuitive. We first find the residual sum of squares for both the full model (SSE_F) and for the reduced model (SSE_R) from the corresponding ANOVA tables. We know that for a given model the residual sum of squares reflects the variation in the response variable *not* explained by the model. If the predictor variables $X_{k-r+1}, X_{k-r+2}, \ldots, X_k$ are important, then deleting them from our model should result in a significant increase in the unexplained variation in Y. That is, SSE_R should become considerably larger than SSE_F. The partial F test makes use of this idea. It is given by:

$$F = \frac{(SSE_R - SSE_F)/r}{MSE_F} \sim F(r, n-(k+1))$$

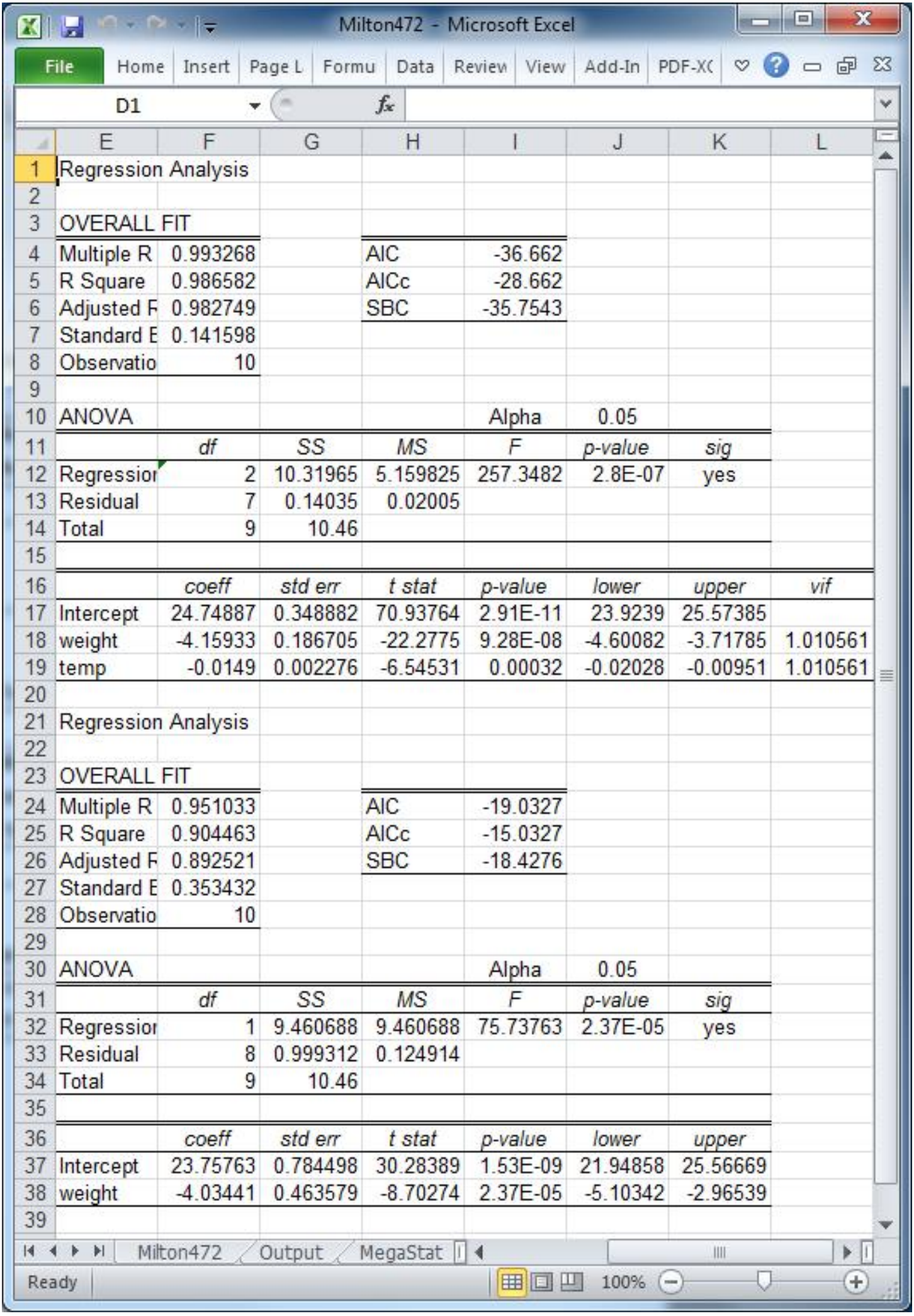
if the null hypothesis, that the reduced model is appropriate, holds true. We reject on the right tail of the distribution, that is, large values of the partial F statistics are taken as evidence that the full model is needed.

Example: (continued.)
Suppose we want to test whether the weight of the car alone is sufficient to predict the gasoline milage of an automobile.

From the ANOVA output for the full model we know: $SSE_F = 0.14$ and $MSE_F = 0.02$. A glance at the ANOVA table for the reduced model reveals that $SSE_R = 0.999$ and we consider deleting $r = 1$ variable from the model. The partial F statistic is therefore:

$$F = \frac{(SSE_R - SSE_F)/r}{MSE_F}$$
$$= \frac{(0.999 - 0.14)/1}{0.02} = 42.95.$$

Calling F.INV.RT(0.01;1;7) or looking up in a table reveals that $F_{0.01}(1, 7) = 12.2$. The p-value is F.DIST.RT(42.95;1;7)=0.032%. So we reject the null hypothesis that the weight of the car alone would suffice in predicting the gasoline milage of an automobile.



Regression Analysis

OVERALL FIT

Multiple R	0.993268	AIC	-36.662
R Square	0.986582	AICc	-28.662
Adjusted F	0.982749	SBC	-35.7543
Standard E	0.141598		
Observatio	10		

ANOVA — Alpha 0.05

	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regressior	2	10.31965	5.159825	257.3482	2.8E-07	yes
Residual	7	0.14035	0.02005			
Total	9	10.46				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*	*vif*
Intercept	24.74887	0.348882	70.93764	2.91E-11	23.9239	25.57385	
weight	-4.15933	0.186705	-22.2775	9.28E-08	-4.60082	-3.71785	1.010561
temp	-0.0149	0.002276	-6.54531	0.00032	-0.02028	-0.00951	1.010561

Regression Analysis

OVERALL FIT

Multiple R	0.951033	AIC	-19.0327
R Square	0.904463	AICc	-15.0327
Adjusted F	0.892521	SBC	-18.4276
Standard E	0.353432		
Observatio	10		

ANOVA — Alpha 0.05

	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regressior	1	9.460688	9.460688	75.73763	2.37E-05	yes
Residual	8	0.999312	0.124914			
Total	9	10.46				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	23.75763	0.784498	30.28389	1.53E-09	21.94858	25.56669
weight	-4.03441	0.463579	-8.70274	2.37E-05	-5.10342	-2.96539

In the special case that we consider to delete only one variable from the full model (as in the preceding example), the p-value for the partial F test coincides with the p-value for the single coefficient t test for the coefficient we are considering to delete from the full model. Indeed, in the previous example the p-value for the temp coefficient is T.DIST.2T(6.545;7)=0.032%, the same as in the partial F test.

So the t tests for significance of individual regression parameters may alternatively be interpreted as partial F tests for reducing the full model by the corresponding regression parameter alone.

That is so because the absolute value of the t statistic for each single regression parameter is just the square root of the partial F test for deleting the same parameter ($\sqrt{42.95} \approx 6.5$ in the preceding example, differences to the ANOVA output for the full model are due to rounding), and we know already that the square of a t-distributed random variable with ν degrees of freedom is $F(1, \nu)$-distributed.

Qualitative Independent Variables

Recall that the purpose of both ANOVA and regression is to forecast the value of a quantitative variable based on some other variable(s). The diffence between ANOVA and Regression is whether the explanatory variable is qualitative (ANOVA) or quantitative (Regression).

We shall next consider incorporating both quantitative and qualitative explanatory variables into a linear model for a quantitative dependent variable. As a start-off point we choose qualitative variabels with only two levels, such as available versus not available. Such a variable is called a dummy variable or indicator variable $\mathbb{I}_A$, because it indicates if some condition A holds. It has the value 1 when the condition A holds and the value 0 when the condition does not hold.

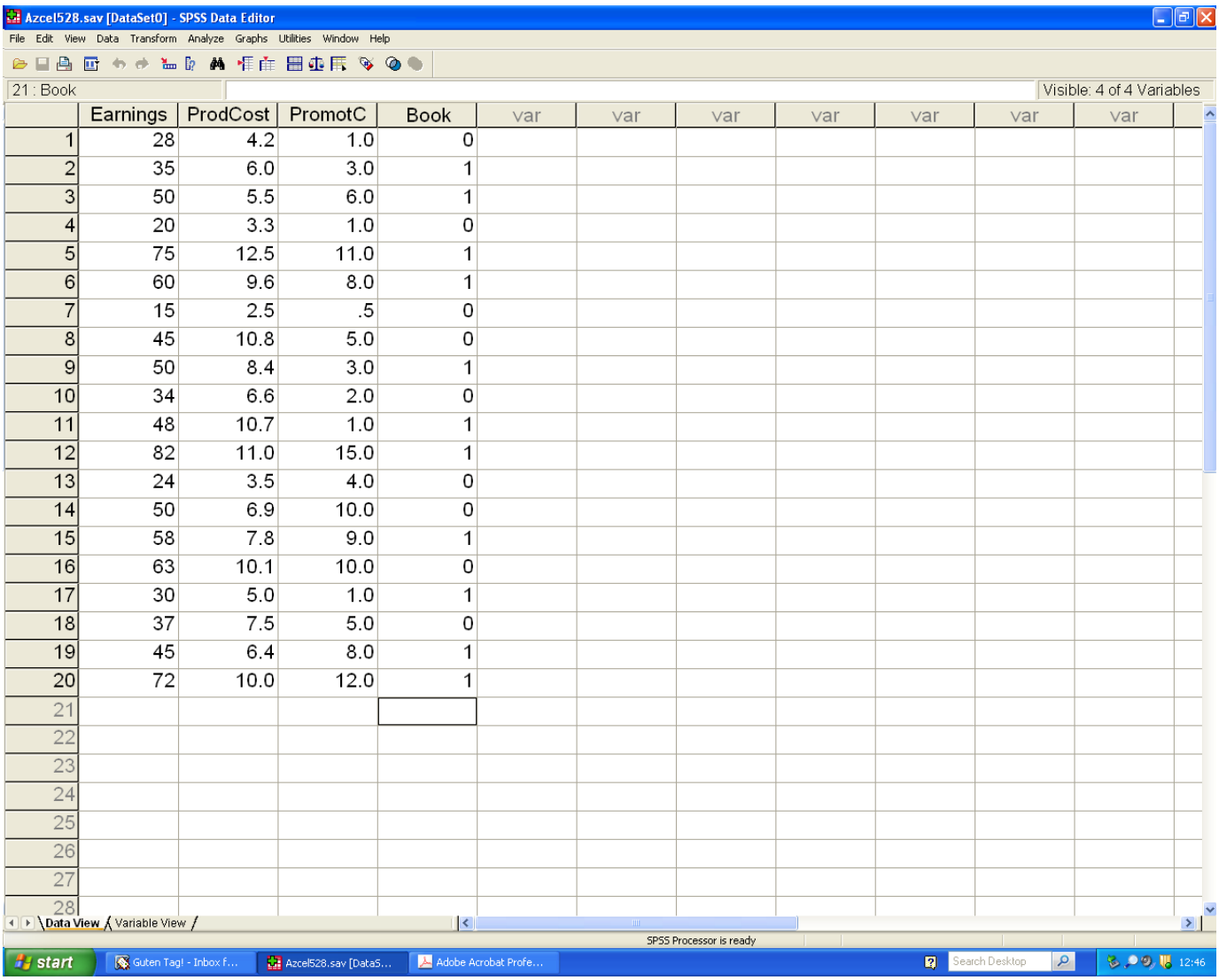
$$\mathbb{I}_A = \begin{cases} 1 & \text{if condition } A \text{ holds,} \\ 0 & \text{if condition } A \text{ does not hold.} \end{cases}$$

The use of indicator variables in regression analysis does not require any additional computational routines. We just include the indicator variable as an additional explanatory variable, coded as 1 if the quality of interest is obtained and 0 if it is not obtained.

Example: (Azcel, Example 11-3.)

A motion picture industry analyst wants to estimate the gross earnings generated by a movie (Y/mio \$) as a linear function of production costs (X_1/mio \$) and promotion costs (X_2/mio \$). As a third variable she wants to consider whether the film is based on a book ($X_3 = 1$) or not ($X_3 = 0$).

The estimated coefficient of 7.166 for X_3 means that having the movie based on a book increases the movie's gross earnings by an average of \$7.166 million.



Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.983[a]	.967	.960	3.690

a. Predictors: (Constant), Book, PromotC, ProdCost

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6325.151	3	2108.384	154.887	.000[a]
	Residual	217.799	16	13.612		
	Total	6542.950	19			

a. Predictors: (Constant), Book, PromotC, ProdCost

b. Dependent Variable: Earnings

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.836	2.333		3.358	.004
	ProdCost	2.848	.392	.447	7.258	.000
	PromotC	2.278	.253	.535	8.989	.000
	Book	7.166	1.818	.197	3.942	.001

a. Dependent Variable: Earnings

The p-value of 0.001 for "Book" means that we can reject $H_0 : \beta_3 = 0$ against $H_1 : \beta_3 \neq 0$ at a significance level of $\alpha = 0.1\%$.

This again implies that we have in fact two different regression models depending upon whether the film is based on a book ($X_3 = 1$) or not ($X_3 = 0$). The regression model for films based on a book is

$$\begin{aligned} Y &= 7.836 + 2.848X_1 + 2.278X_2 + 7.166 \cdot 1 + \epsilon \\ &= 15.002 + 2.848X_1 + 2.278X_2 + \epsilon \end{aligned}$$

whereas for films not based on a book it is

$$Y = 7.836 + 2.848X_1 + 2.278X_2 + \epsilon.$$

We see that the fact that $H_0 : \beta_3 = 0$ was rejected implies that different subsamples of the films are described by different regression models with different intercepts, but identical slope coefficients.

In general, if the regression model contains k quantitative regressors $X_1, \ldots, X_k$ and one dummy variable X_{k+1}, then rejection of $H_0 : \beta_{k+1} = 0$ against $H_1 : \beta_{k+1} \neq 0$ implies that there are two parallel regression models for the 2 subsamples corresponding to the value of the indicator variable X_{k+1}:

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_k X_k + \epsilon \quad \text{for } X_{k+1} = 0, \quad \text{and}$$

$$Y = (\beta_0 + \beta_{k+1}) + \sum_{i=1}^{k} \beta_k X_k + \epsilon \text{ for } X_{k+1} = 1.$$

Even though we could in principle go and fit own regression models for each subsample separately, we have good reasons to use the dummy variable approach instead:

1) It allows us to test statistically whether there are indeed two separate models needed.
2) By pooling the data from both groups, we improve the efficiency of our estimators for the common regression parameters $\beta_1, \ldots, \beta_k$ and σ^2.

Qualitative Variables with more than 2 levels

Rather than introducing a pseudo-indicator with more than 2 levels, we account for a qualitative variable with r levels by using $r-1$ indicator (0/1) variables as follows:
All indicator variables $= 0$ indicates the first level, all other $r-1$ levels are indicated by setting the corresponding indicator variable to 1 and the remaining dummies to 0.

Example: (continued.)
Suppose the analyst is interested not in whether the movie is based on a book, but rather in the category to which the movie belongs: adventure, drama, or romance. Assume furthermore for simplicity, that the only quantitative regressor is production costs (such that we get a regression line rather than a surface). We may then model the $r = 3$ subgroups adventure, drama, and romance by $r - 1 = 2$ dummy variables X_2 and X_3 as follows.

Category:	X_2	X_3
Adventure	0	0
Drama	1	0
Romance	0	1

Estimating the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

yields 3 estimated regression lines:

$$\begin{aligned}
\widehat{Y} &= b_0 + b_1 X_1 && \text{for adventure,} \\
\widehat{Y} &= (b_0 + b_2) + b_1 X_1 && \text{for drama,} \\
\widehat{Y} &= (b_0 + b_3) + b_1 X_1 && \text{for romance.}
\end{aligned}$$

Nonrejection of $H_0 \colon \beta_2 = 0$ would imply that adventure and drama films produce the same earnings, while nonrejection of $H_0 : \beta_3 = 0$ would imply that adventure and romance films produce the same earnings (given identical production costs X_1). A partial F test could be used to test $H_0 \colon \beta_2 = \beta_3 = 0$, that is that all three categories produce the same earnings (given identical production costs X_1).

Interactions between Qualitative and Quantitative Variables

So far we have assumed that the quantitative variable X_1 effects all levels of the qualitative variables in the same way, that is, all regression lines or surfaces are parallel (identical slope coefficients, only intercepts may differ). This assumption may be tested. In the case of one indicator variable X_2, with 2 levels, an appropriate model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

Estimation of this model yields

$$\begin{aligned} \widehat{Y} &= b_0 + b_1 X_1 && \text{for } X_2 = 0, \\ \widehat{Y} &= b_0 + b_1 X_1 + b_2 + b_3 X_1 \\ &= (b_0 + b_2) + (b_1 + b_3) X_1 && \text{for } X_2 = 1. \end{aligned}$$

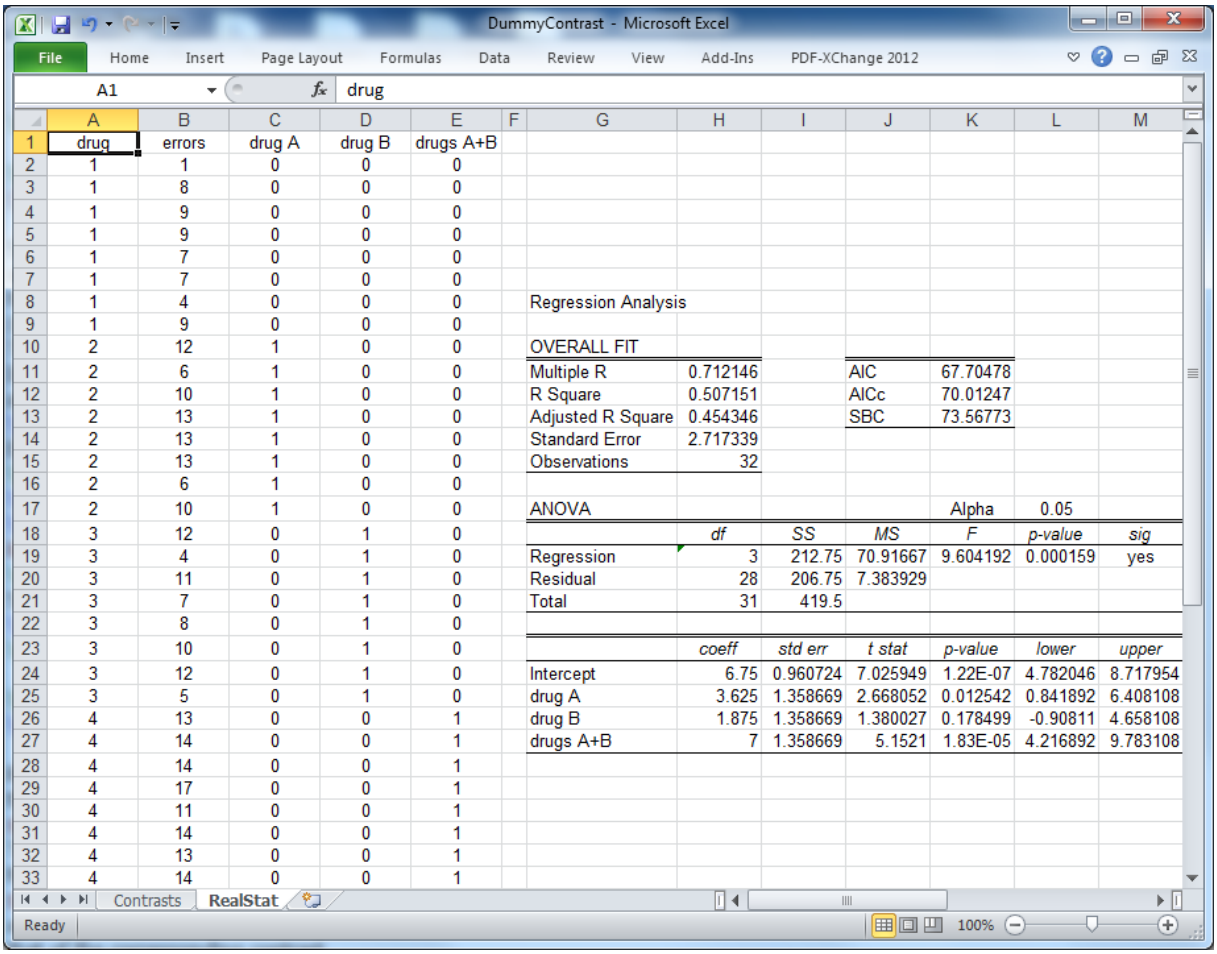
So we may test equality of slopes by testing

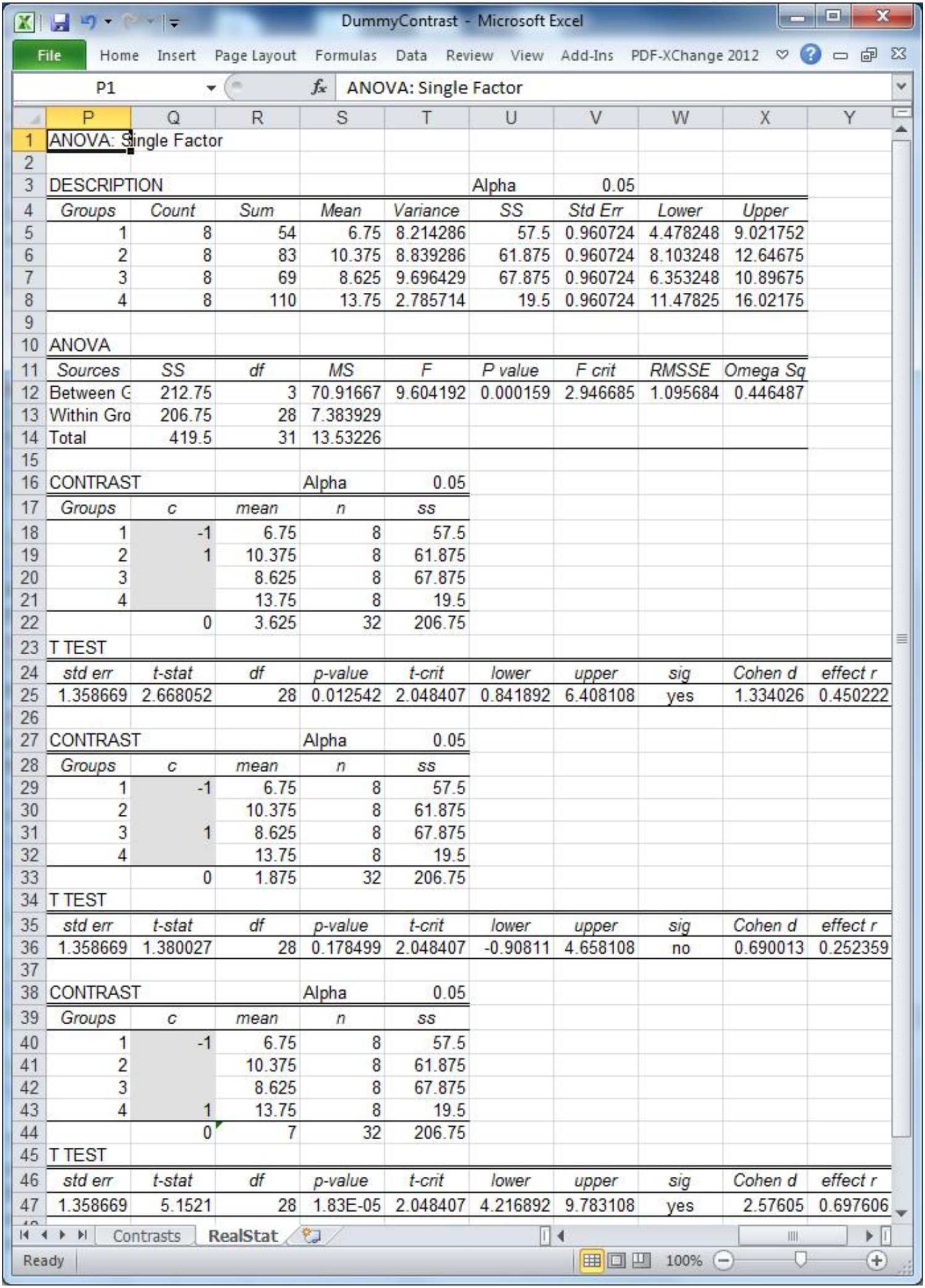
$$H_0 : \beta_3 = 0 \quad \text{against} \quad H_1 : \beta_3 \neq 0.$$

Dummy Regressions as Simple Contrasts

Regressing on dummy variables alone replicates the results of contrasts of the form $H_0: \mu_{\text{Dummy}=1} = \mu_{\text{Dummy}=0}$. The slope coefficient of a dummy variable equals then the difference in means $\bar{x}_{\text{Dummy}=1} - \bar{x}_{\text{Dummy}=0}$ and the p-value of the t-test is the same as that of the corresponding contrast.

Example. Consider again the errors made under influence of drug A, drug B, or both drugs.





DummyContrast - Microsoft Excel

File Home Insert Page Layout Formulas Data Review View Add-Ins PDF-XChange 2012

P1 ANOVA: Single Factor

	P	Q	R	S	T	U	V	W	X	Y
1	ANOVA: Single Factor									
2										
3	DESCRIPTION					Alpha	0.05			
4	*Groups*	*Count*	*Sum*	*Mean*	*Variance*	*SS*	*Std Err*	*Lower*	*Upper*	
5	1	8	54	6.75	8.214286	57.5	0.960724	4.478248	9.021752	
6	2	8	83	10.375	8.839286	61.875	0.960724	8.103248	12.64675	
7	3	8	69	8.625	9.696429	67.875	0.960724	6.353248	10.89675	
8	4	8	110	13.75	2.785714	19.5	0.960724	11.47825	16.02175	
9										
10	ANOVA									
11	*Sources*	*SS*	*df*	*MS*	*F*	*P value*	*F crit*	*RMSSE*	*Omega Sq*	
12	Between G	212.75	3	70.91667	9.604192	0.000159	2.946685	1.095684	0.446487	
13	Within Gro	206.75	28	7.383929						
14	Total	419.5	31	13.53226						
15										
16	CONTRAST			Alpha	0.05					
17	*Groups*	*c*	*mean*	*n*	*ss*					
18	1	-1	6.75	8	57.5					
19	2	1	10.375	8	61.875					
20	3		8.625	8	67.875					
21	4		13.75	8	19.5					
22		0	3.625	32	206.75					
23	T TEST									
24	*std err*	*t-stat*	*df*	*p-value*	*t-crit*	*lower*	*upper*	*sig*	*Cohen d*	*effect r*
25	1.358669	2.668052	28	0.012542	2.048407	0.841892	6.408108	yes	1.334026	0.450222
26										
27	CONTRAST			Alpha	0.05					
28	*Groups*	*c*	*mean*	*n*	*ss*					
29	1	-1	6.75	8	57.5					
30	2		10.375	8	61.875					
31	3	1	8.625	8	67.875					
32	4		13.75	8	19.5					
33		0	1.875	32	206.75					
34	T TEST									
35	*std err*	*t-stat*	*df*	*p-value*	*t-crit*	*lower*	*upper*	*sig*	*Cohen d*	*effect r*
36	1.358669	1.380027	28	0.178499	2.048407	-0.90811	4.658108	no	0.690013	0.252359
37										
38	CONTRAST			Alpha	0.05					
39	*Groups*	*c*	*mean*	*n*	*ss*					
40	1	-1	6.75	8	57.5					
41	2		10.375	8	61.875					
42	3		8.625	8	67.875					
43	4	1	13.75	8	19.5					
44		0	7	32	206.75					
45	T TEST									
46	*std err*	*t-stat*	*df*	*p-value*	*t-crit*	*lower*	*upper*	*sig*	*Cohen d*	*effect r*
47	1.358669	5.1521	28	1.83E-05	2.048407	4.216892	9.783108	yes	2.57605	0.697606

Contrasts RealStat

Ready 100%

Polynomial Regression

The mathematical framework of multiple regression may be used to model relationships between a response variable Y and a single predictor variable X, where the relationship between X and Y is *curved* rather than linear.

A one-variable polynomial regression model is

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_m X^m + \epsilon,$$

where m is the *degree* of the polynomial (the highest power of X in the equation), which is also called the *order* of the model.

Polynomial models of order higher than 2 are very rarely used in practice, due to the danger of overfitting, and because the dependence between different powers of X may result in difficulties to find the right regression parameters (so called multicollinearity, to be discussed later).

*Azcel540.sav [DataSet0] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

22 : Advert

Visible: 5 of 5 Variables

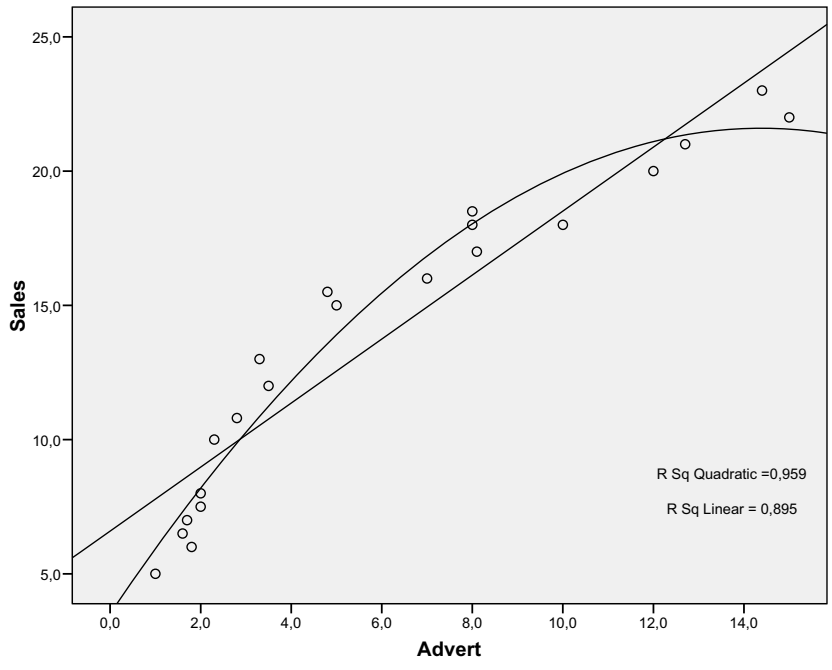
	Sales	Advert	AdvSQR	Log_Advt	LogSales	var	var	var	var	var	va
1	5.0	1.0	1.00	.00	1.61						
2	6.0	1.8	3.24	.59	1.79						
3	6.5	1.6	2.56	.47	1.87						
4	7.0	1.7	2.89	.53	1.95						
5	7.5	2.0	4.00	.69	2.01						
6	8.0	2.0	4.00	.69	2.08						
7	10.0	2.3	5.29	.83	2.30						
8	10.8	2.8	7.84	1.03	2.38						
9	12.0	3.5	12.25	1.25	2.48						
10	13.0	3.3	10.89	1.19	2.56						
11	15.5	4.8	23.04	1.57	2.74						
12	15.0	5.0	25.00	1.61	2.71						
13	16.0	7.0	49.00	1.95	2.77						
14	17.0	8.1	65.61	2.09	2.83						
15	18.0	8.0	64.00	2.08	2.89						
16	18.0	10.0	100.00	2.30	2.89						
17	18.5	8.0	64.00	2.08	2.92						
18	21.0	12.7	161.29	2.54	3.04						
19	20.0	12.0	144.00	2.48	3.00						
20	22.0	15.0	225.00	2.71	3.09						
21	23.0	14.4	207.36	2.67	3.14						
22											
23											
24											
25											
26											
27											
28											

Data View | Variable View

SPSS Processor is ready

start | Data analyzing - Inbo... | *Azcel540.sav [Data... | Riippu | Search Desktop | 13:10

Graph



Example: (Azcel Example 11–5.)
Sales response to advertising usually follows a curve reflecting the diminishing returns to advertising expenditure. As a firm increases its advertsing expenditure, sales increase, but the rate of increase drops continually after a certain point.

The preceding slide contains data on sales revenues as a function of advertising expenditure. As is evident from the scatterplot, sales as a function of advertising is better approximated by a polynomial of 2nd order than by a straight line. So we attempt to fit:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

and obtain (see next slide):

$$\hat{Y} = 3.515 + 2.515X - 0.0875X^2.$$

(Note that the regression model is not fully satisfactory as it is evident from the residual plot that there is left some autocorrelation in the residuals.)

Linear Regression Results

Model: Linear_Regression_Model
Dependent Variable: Sales Sales

Number of Observations Read	21
Number of Observations Used	21

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	630.25801	315.12901	208.99	<.0001
Error	18	27.14199	1.50789		
Corrected Total	20	657.40000			

Root MSE	1.22796	**R-Square**	0.9587
Dependent Mean	13.80000	**Adj R-Sq**	0.9541
Coeff Var	8.89827		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	3.51505	0.73847	4.76	0.0002
Advert	Advert	1	2.51478	0.25796	9.75	<.0001
AdvSQR	AdvSQR	1	-0.08745	0.01658	-5.28	<.0001