Nonlinear Models and Transformations

Sometimes relationships between Y and one or more of the X_i's is nonlinear. Remember that powers of the X_i's still keep the model linear in terms of the slope coefficients β_i. When we talk about nonlinear models, we mean models which are not linear in the regression coefficients β_i.

Luckily many nonlinear models can be made linear by appropriate transformations. Such models are called intrinsically linear.

Consider for example the multiplicative model:

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \cdots X_k^{\beta_k} \epsilon,$$

which may be linearized by the logarithmic transformation into the form:

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

In the special case of only one independent variable one obtaines the power model:

$$Y = \beta_0 X^{\beta_1} \epsilon,$$

which linearizes as

$$Y^* = \beta_0^* + \beta_1^* X^* + \epsilon^*$$

where $Y^* = \log Y$, $\beta_0^* = \log \beta_0$, $\beta_1^* = \beta_1$, $X^* = \log X$ and $\epsilon^* = \log \epsilon$. The estimates for the original parameters are recovered by the inverse transformations

$$\hat{\beta}_1 = \hat{\beta}_1^* \quad \text{and} \quad \hat{\beta}_0 = e^{\hat{\beta}_0^*}.$$

Example: (continued.)
The estimated parameters of the power model for the sales and advertising data are obtained from regressing the log of sales upon the log of advertising as

$\hat{\beta}_1 = 0.553$ and $\hat{\beta}_0 = e^{1.701} = 5.479$, that is,

$$\hat{Y} = 5.479 \cdot X^{0.553}.$$

Linear Regression Results

Model: Linear_Regression_Model
Dependent Variable: LogSales LogSales

Number of Observations Read	21
Number of Observations Used	21

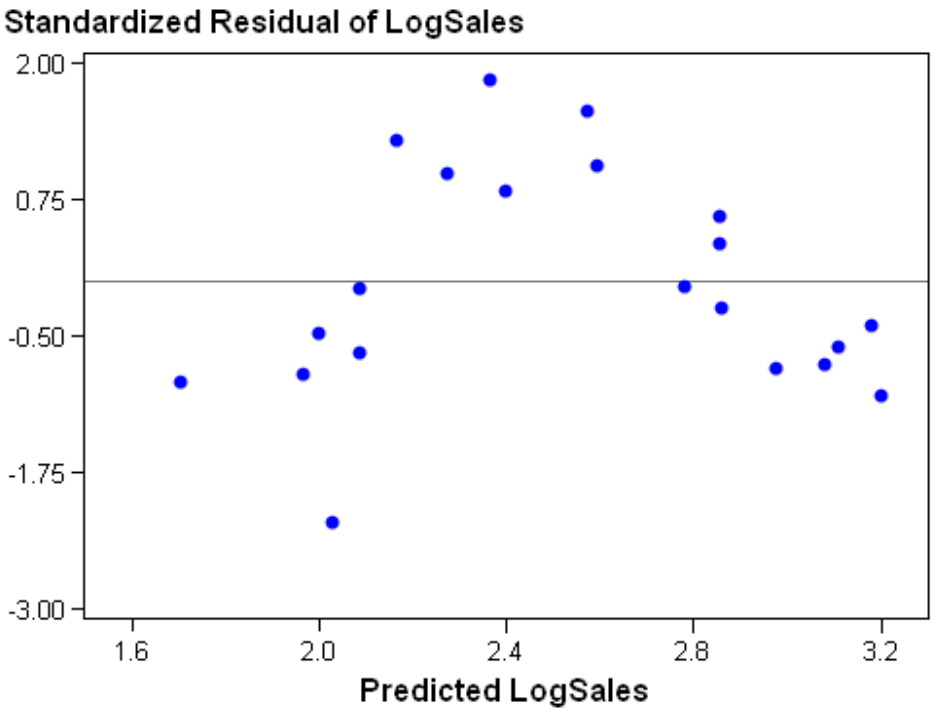
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.27217	4.27217	337.56	<.0001
Error	19	0.24047	0.01266		
Corrected Total	20	4.51263			

Root MSE	0.11250	**R-Square**	0.9467
Dependent Mean	2.52692	**Adj R-Sq**	0.9439
Coeff Var	4.45205		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.70082	0.05123	33.20	<.0001
Log_Advt	Log_Advt	1	0.55314	0.03011	18.37	<.0001



The exponential model is:

$$Y = \beta_0 e^{\beta_1 X_1 + \cdots + \beta_k X_k} \epsilon.$$

It is linearized by:

$$\log Y = \log \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \log \epsilon.$$

So we regress the logarithm of the dependent variable upon the untransformed independent variables.

The reciprocal model is:

$$Y = \frac{1}{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon}.$$

It is linearized by:

$$\frac{1}{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon.$$

So we regress the reciprocal of the dependent variable upon the untransformed independent variables.

The logarithmic model is:

$$Y = \beta_0 + \beta_1 \log X + \epsilon.$$

This model is already linear in the β's, so there is no need to transform any regression output. We only have to remember to regress Y upon $\log X$ rather than X itself.

Example: (continued.)

The estimated parameters of the logarithmic model for the sales and advertising data are obtained from regressing sales upon the log of advertising as

$$\widehat{\beta}_0 = 3.668 \quad \text{and} \quad \widehat{\beta}_1 = 6.784,$$

that is,

$$\widehat{Y} = 3.668 + 6.784 \log X.$$

This is the best model for the sales and advertising data that we have tried.

Linear Regression Results

Model: Linear_Regression_Model
Dependent Variable: Sales Sales

Number of Observations Read	21
Number of Observations Used	21

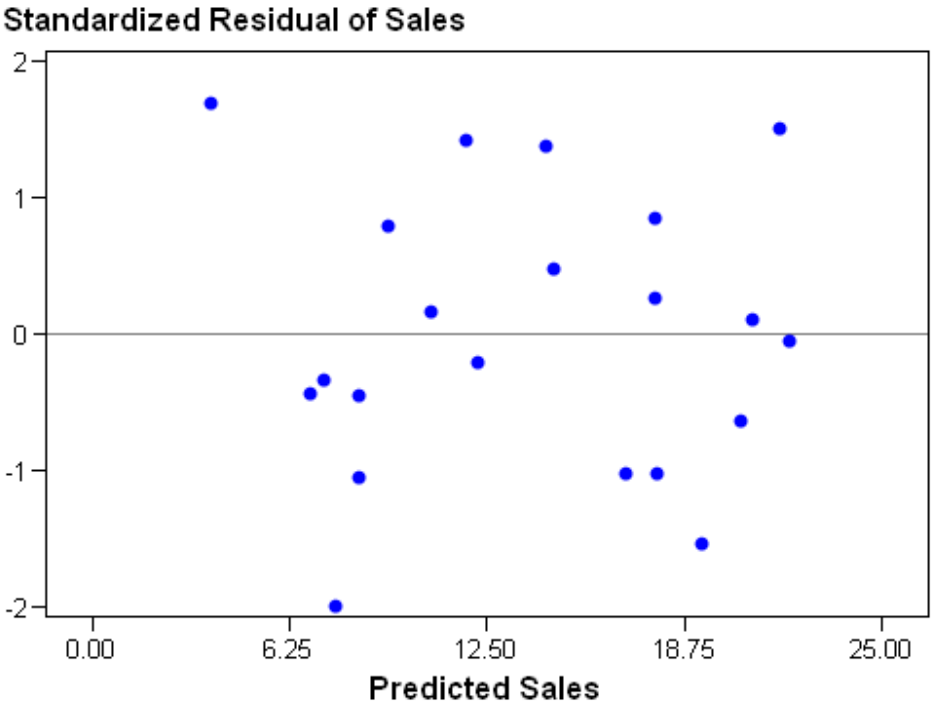
Analysis of Variance

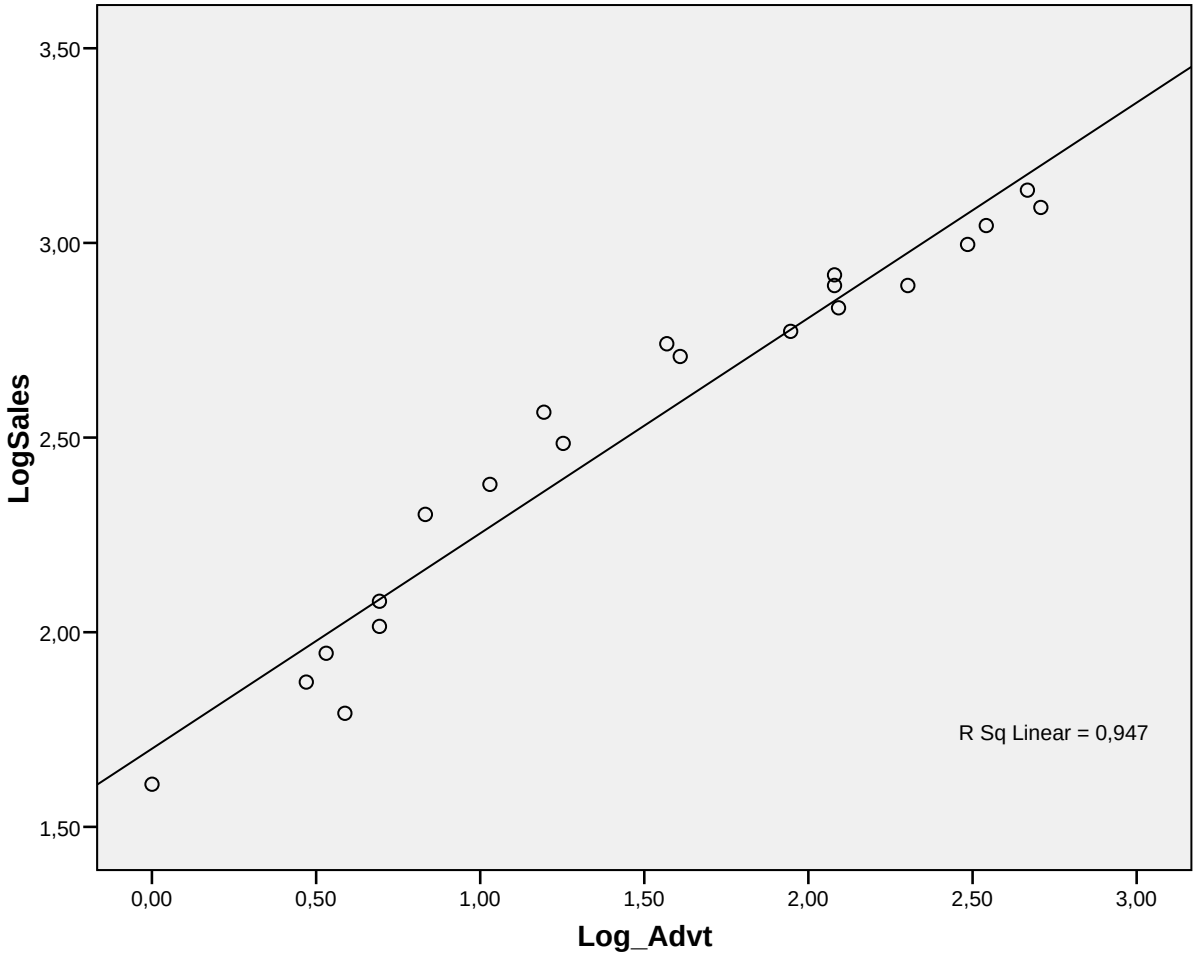
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	642.62235	642.62235	826.24	<.0001
Error	19	14.77765	0.77777		
Corrected Total	20	657.40000			

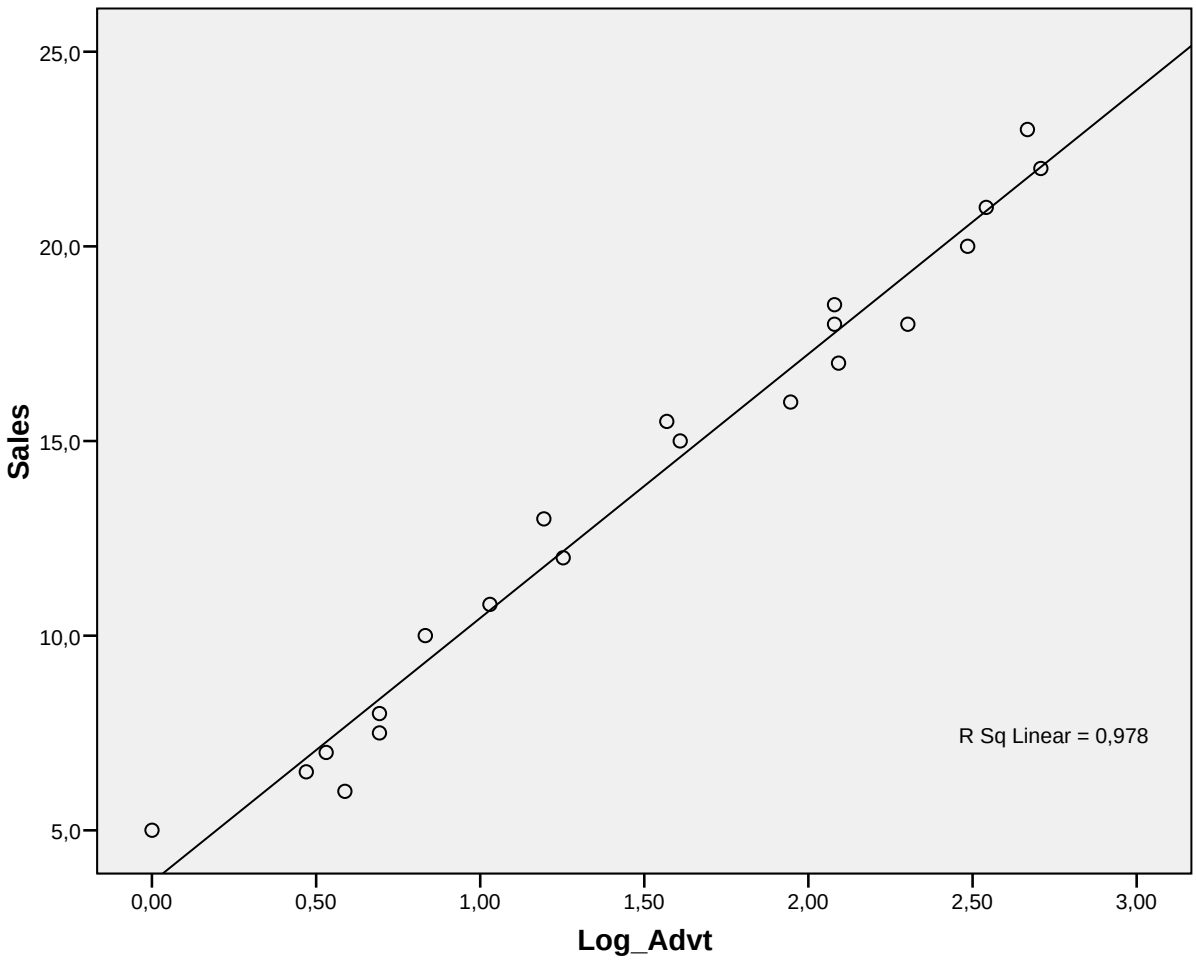
Root MSE	0.88191	**R-Square**	0.9775
Dependent Mean	13.80000	**Adj R-Sq**	0.9763
Coeff Var	6.39068		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	3.66825	0.40159	9.13	<.0001
Log_Advt	Log_Advt	1	6.78400	0.23601	28.74	<.0001




3,50
3,00
2,50
2,00
1,50
LogSales
0,00
0,50
1,00
1,50
2,00
2,50
3,00
Log_Advt
R Sq Linear = 0,947


25,0
20,0
15,0
10,0
5,0
Sales
0,00
0,50
1,00
1,50
2,00
2,50
3,00
Log_Advt
R Sq Linear = 0,978

Logistic Regression

With logistic regression we are interested to model some probability p as a function of an explanatory variable x. The naive approach to set $p = \beta_0 + \beta_1 x + \epsilon$ doesn't work because given normally distributed residuals ϵ it is inconsistent with the fact that $0 \leq p \leq 1$.

Instead one divides the probability p by $(1-p)$ in order to obtain the so called odds

$$ODDS = \frac{p}{1-p}$$

and postulates that the log odds or logit

$$z = \log(ODDS) = \log\left(\frac{p}{1-p}\right)$$

is a linear function of x:

$$z = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x.$$

This is the (binary) logistic regression model.

Solving $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$ for p yields the so called logistic function:

$$p = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}},$$

such that:

$$\begin{array}{lll} p \to 1 & \text{for} & \beta_1 x \to \infty, \\ p \to 0 & \text{for} & \beta_1 x \to -\infty. \end{array}$$

The parameter $b_1 = \widehat{\beta}_1$ estimates the change in the logit caused by a unit change x. From

$$ODDS = e^z = e^{\beta_0+\beta_1 x} = e^{\beta_0} e^{\beta_1 x}$$

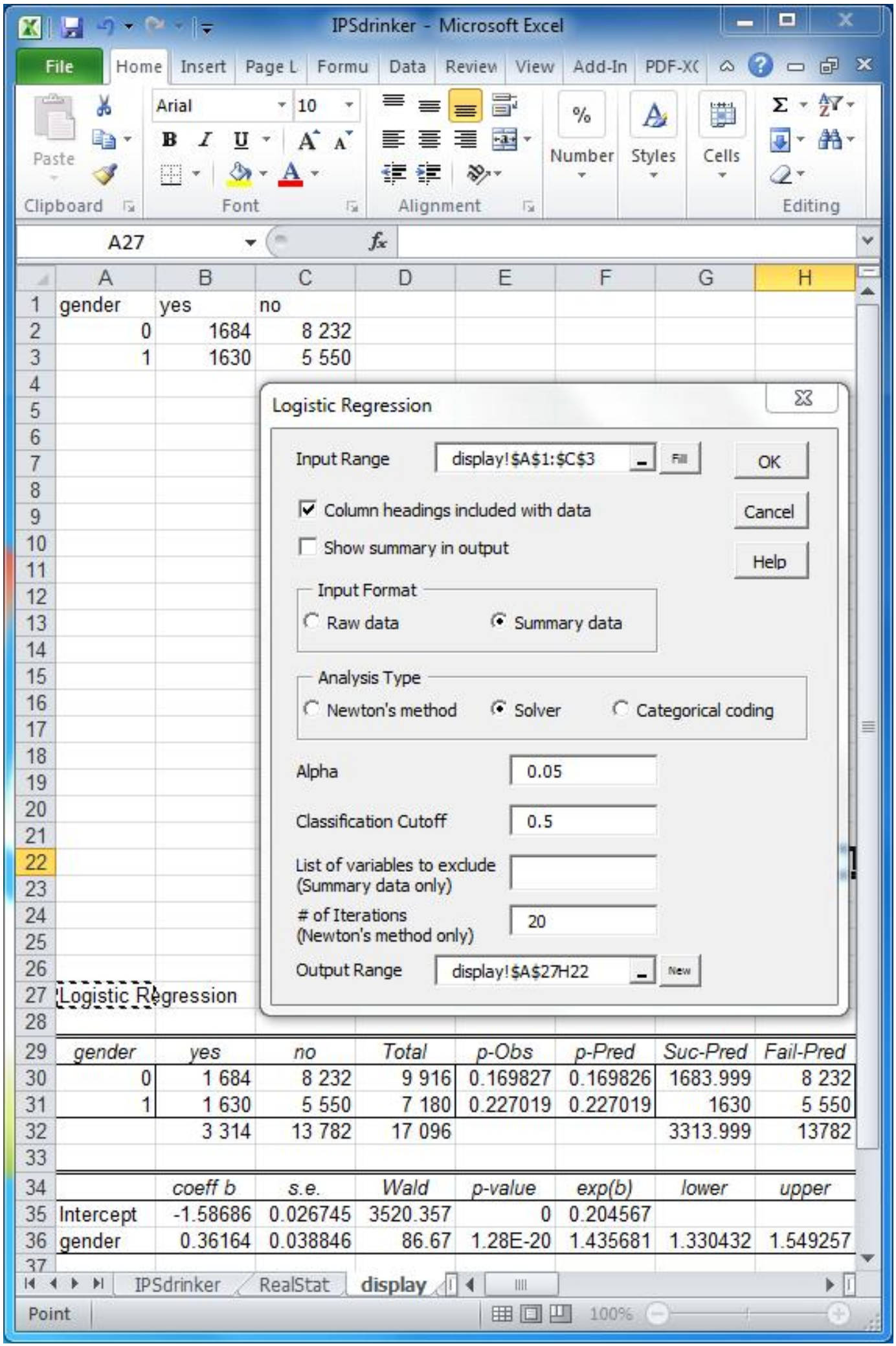
we see that the odds ratio (vetosuhde)

$$\frac{ODDS_{x+1}}{ODDS_x} = \frac{e^{\beta_0} e^{\beta_1 (x+1)}}{e^{\beta_0} e^{\beta_1 x}} = e^{\beta_1}.$$

is the factor by which the odds of the event change for a one-unit change in the explanatory variable.

In the special case that x is a dummy variable, the logistic regression model with only one regressor may be fitted exactly, such that

$$ODDS_{\{x=0\}} = e^{\beta_0} \quad \text{and} \quad ODDS_{\{x=1\}} = e^{\beta_0+\beta_1}.$$



	A	B	C	D	E	F	G	H
1	gender	yes	no					
2	0	1684	8 232					
3	1	1630	5 550					

29	*gender*	*yes*	*no*	*Total*	*p-Obs*	*p-Pred*	*Suc-Pred*	*Fail-Pred*
30	0	1 684	8 232	9 916	0.169827	0.169826	1683.999	8 232
31	1	1 630	5 550	7 180	0.227019	0.227019	1630	5 550
32		3 314	13 782	17 096			3313.999	13782

34		*coeff b*	*s.e.*	*Wald*	*p-value*	*exp(b)*	*lower*	*upper*
35	Intercept	-1.58686	0.026745	3520.357	0	0.204567		
36	gender	0.36164	0.038846	86.67	1.28E-20	1.435681	1.330432	1.549257

Example: (Moore/McCabe Example 16.4)

In the preceding contingency table, the probabilities for being a frequent binge drinker are

$$p_{\text{men}} = \frac{1630}{7180} = 0.227, \qquad p_{\text{women}} = \frac{1684}{9916} = 0.170.$$

The corresponding odds are:

$$ODDS_{\text{men}} = \frac{0.227}{0.773} = 0.294, \; ODDS_{\text{women}} = \frac{0.170}{0.830} = 0.205$$

with associated log odds or logits:

$$z_{\text{men}} = \log(0.294) = -1.23, \;\; z_{\text{women}} = \log(0.205) = -1.59.$$

Coding $x = 1$ for men and $x = 0$ for women, we obtain by inserting into $z = b_0 + b_1 x$:

$$z_{\text{men}} = b_0 + b_1 = -1.23, \; z_{\text{women}} = b_0 = -1.59.$$

Solving for b_1 yields:

$$b_1 = -1.23 - (-1.59) = 0.36,$$

such that the odds ratio becomes

$$e^{0.36} = 1.43,$$

which indeed coincides with

$$\frac{ODDS_{\text{men}}}{ODDS_{\text{women}}} = \frac{0.294}{0.205} = 1.43.$$

Inference for Logistic Regression

A level α confidence interval for β_1 (slope) is

$$b_1 \pm z_{\alpha/2} SE_{b_1}.$$

A level α confidence interval for e^{β_1} (odds ratio) is obtained by transforming the confidence interval for the slope β_1 into

$$\left[e^{b_1 - z_{\alpha/2} SE_{b_1}}, e^{b_1 + z_{\alpha/2} SE_{b_1}}\right].$$

To test $H_0: \beta_1 = 0$, compute the test statistic:

$$z = \frac{b_1}{SE_{b_1}} \sim N(0,1) \quad \text{under } H_0.$$

Computer output usually reports the square of this statistic, called the Wald statistic:

$$z^2 = \left(\frac{b_1}{SE_{b_1}}\right)^2 \sim \chi^2(1) \quad \text{under } H_0.$$

Note that $\beta_1 = 0$ corresponds to $e^{\beta_1} = 1$, that is identical odds for the event occuring regardless of the value of x.

Example: (continued.)

By looking up in a table, or by calling NORMSINV in excel, we find $z_{0.05/2} = 1.96$. A 95% confidence interval for β_1 is:

$$\begin{aligned} b_1 \pm z_{\alpha/2} SE_{b_1} &= 0.3616 \pm 1.96 \cdot 0.0388 \\ &= [0.2855,\ 0.4376]. \end{aligned}$$

The 95% confidence interval for e^{β_1} is:

$$\begin{aligned} \left[e^{b_1 - z_{\alpha/2} SE_{b_1}}, e^{b_1 + z_{\alpha/2} SE_{b_1}}\right] &= \left[e^{0.2855},\ e^{0.4376}\right] \\ &= [1.33,\ 1.55]. \end{aligned}$$

The Wald statistic for β_1 is:

$$\left(\frac{b_1}{SE_{b_1}}\right)^2 = \left(\frac{0.3616}{0.0388}\right)^2 = 86.67$$

with associated p-value:

$$\text{CHIDIST}(86.67;1) = 1.3 \cdot 10^{-20},$$

leaving no doubt that being a man raises the odds of being a frequent binge drinker.

The Real Statistics toolpack offers logistic regression under 'Multinomial logistic regression' as one of the choices within the 'Regression' tool. Set the number of variables to 1 in order to get binary logistic regression.

Multicollinearity

The purpose of this section is to convince ourselves that for a "reliable" regression it is important to have explanatory variables that are as "unrelated to each other" as possible.

What does "reliability" mean?

With reliability I mean in this context both efficient estimation in the sense that standard errors and confindence bands for individual regression parameters are small, but also that the parameter estimates are robust with respect to changing numbers of regressors, or adding or deleting a few data points.

What does "unrelatedness" mean?

With unrelatedness I mean that it should not be possible to predict any of the regressors as a linear function of the other regressors. The mathematical term for this is multicollinearity. Intuitively, multicollinearity means that some regressors are so similiar, that the regression has a hard time to decide who is responsible for changes in the dependent variable Y.

Consider first the case of only two explanatory variables. X_1 and X_2 are said to be perfectly collinear if we can find two constants a and b such that:

$$X_1 = a + bX_2.$$

In such a case both variables fall on a straight line, and one of them perfectly determines the other. Therefore, no information about Y is gained by adding X_2 to a regression that contains already X_1 (or vice versa).

An obvious measure of collinearity between only two variables is their *correlation* r_{12}. A pair of variables with $|r_{12}| = 1$ is perfectly collinear, with decreasing degree of collinearity for decreasing $|r_{12}|$.

Unfortunately, in the case of more than two regressors things become more complicated. High correlations remain a warning signal of multicollinearity. But we may still have multicollinarity, even though taken individually, the correlation coefficients look moderate.

Definition of Multicollinearity

Exact or perfect multicollinearity between the regressors $X_1, \ldots, X_k$ is said to exist if the regressors are *linearly dependent*, that is, we may express one or more of the regressors as a linear combination of the other regressors, e.g.:

$$X_1 = c_2 X_2 + c_3 X_3 + \cdots + c_k X_k.$$

We talk about near multicollinearity if the equation above holds only approximately:

$$X_1 \approx c_2 X_2 + c_3 X_3 + \cdots + c_k X_k.$$

There is obiously no objective rule to decide whether a set of regressors is near muliti-collinear or not, since multicollinearity is a characteristic of degree. We shall later give some rules of thumb when multicollinearity starts being a problem.

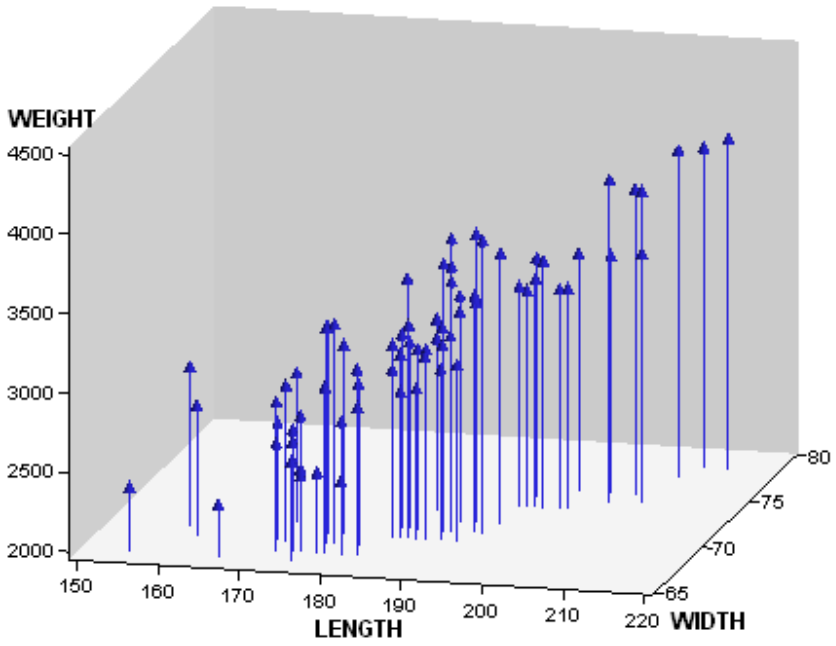
Geometric Interpretation of Multicollinearity

Geometrically, if there is multicollinearity among the regressors $X_1, \ldots, X_k$, then taking the observations as points in the k-dimensional space with the ith coordinate given by the the value for the ith regressor fall roughly on a so called *hyperplane* or lower-dimensional plane with less than k dimensions.

For example, when there are two predictor variables ($k = 2$), multicollinearity means that a two-dimensional scatterplot of the values of X_1 and X_2 falls roughly on a straight line. For $k = 3$ predictor variables multicollinearity means that a three-dimensional scatterplot of the values of X_1, X_2 and X_3 falls roughly on either a plane or a straight line.

Example. (Weiss: Example B.22)

Scatter Plot



Correlation Analysis

3 Variables: WEIGHT WIDTH LENGTH

Pearson Correlation Coefficients, N = 82			
	WEIGHT	**WIDTH**	**LENGTH**
WEIGHT WEIGHT	1.00000	0.84110	0.84001
WIDTH WIDTH	0.84110	1.00000	0.86464
LENGTH LENGTH	0.84001	0.86464	1.00000

Causes of Multicollinearity

• Sometimes multicollinearity arises because the regressors are naturally related to each other. In the previous example, weight, length and width measured essentially the same concept, namely the size of the car.

• We can also introduce multicollinearity by creating regressors from another regressor, as is done in polynomial regression through the taking of powers. We might also combine two or more regressors to obtain another regressor. For example, we might sum the exam scores X_1 and X_2 in order to obtain the total exam score X_3. Then there will be multicollinearity among X_1, X_2, and X_3.

• An incorrect use of dummy variables may introduce multicollinearity, e.g. using r instead of $r-1$ indicator variables to describe a qualitative variable with r levels.

• A data collection method may produce multicollinearity, when gathering data with related values on several variables. For example, if we run a regression of home size Y versus family income X_1 and family size X_2, then we will get multicollinearity if we happen to sample mainly small families with low income and large families with high income, rather than also collecting data from large families with low income and small families with high income.

• Sometimes constraints on the data may force us to introduce multicollinearity. For example, if we run a regression of chemical yield Y on the concentration of two elements X_1 and X_2 with their sum being constant, then as one chemical increases in concentration, the other one must decline. So X_1 and X_2 are (negatively) correlated, and multicollinearity is present.

Consequences of Exact Multicollinearity

Recall the least square estimate for the unknown parameter vector $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

which we obtained by premultiplying the normal equations

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y} \qquad \text{by} \qquad (\mathbf{X}'\mathbf{X})^{-1}.$$

Now if one or more regressors may be expressed as linear combinations of other regressors, the columns of X will be linearly dependent, which implies that the rank of $\mathbf{X}'\mathbf{X}$ is less than $k+1$ and the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist! That is, we simply cannot calculate any least square estimate for β.

The normal equations $(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$ remain a valid description of the minimization problem faced in least square estimation, but an unambigious solution can no longer be found. This is only logical, since several combinations of the regressors serve equally well in explaining the dependent variable.

Consequences of Near Multicollinearity

1. Lack of Robustness with respect to inclusion/ exclusion of regressors

The first important point to understand when assessing the impact of multicollinearity, is that in the general case we may expect the slope coefficient of any regressor to change when other regressors are added or eliminated from the regression equation. To see this, write the regression equation as

$$Y_{(n\times 1)} = \beta_{0\,(n\times 1)} + \mathbf{X}_{(n\times k)}\beta_{(n\times k)} + \epsilon^*_{(n\times 1)}.$$

Taking the arithmetic mean yields:

$$\overline{Y}_{(n\times 1)} = \beta_{0\,(n\times 1)} + \overline{\mathbf{X}}_{(n\times k)}\beta_{(n\times k)} + \overline{\epsilon^*}_{(n\times 1)}.$$

Subtracting both equations from each other yields the regression model in *deviation form*:

$$\underbrace{Y - \overline{Y}}_{\mathbf{y}} = \underbrace{X - \overline{X}}_{\mathbf{x}}\beta + \underbrace{\epsilon^* - \overline{\epsilon^*}}_{\epsilon}.$$

Note that $\mathbf{X}$ and $\mathbf{x}$ are of size $(n \times k)$ and have no leading columns of 1's.

In order to assess the impact of adding/ deleting a group of regressors from the regression equation, we divide the regressors into 2 groups: group 1 say, consisting of the first m regressors $X_1, \ldots, X_m$, and group 2 consisting of the remaining $r = k - m$ regressors $X_{m+1}, \ldots X_{m+r}$.

That is, we split the parameter vector β into $\beta_1 = (\beta_1, \ldots, \beta_m)'$ and $\beta_2 = (\beta_{m+1}, \ldots, \beta_{m+r})'$ and the design matrix into $\mathbf{x}_1$ and $\mathbf{x}_2$ behind the m'th column:

$$\mathbf{x}_{(n\times k)} = \underbrace{\begin{pmatrix} x_{1,1} & \cdots & x_{m,1} \\ \vdots & & \vdots \\ x_{1,n} & \cdots & x_{m,n} \end{pmatrix}}_{\mathbf{x}_{1\,(n\times m)}} \underbrace{\begin{pmatrix} x_{m+1,1} & \cdots & x_{m+r,1} \\ \vdots & & \vdots \\ x_{m+1,n} & \cdots & x_{m+r,n} \end{pmatrix}}_{\mathbf{x}_{2\,(n\times r)}},$$

such that the regression equation becomes

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \epsilon,$$

and the normal equations $(\mathbf{x}'\mathbf{x})\mathbf{b} = \mathbf{x}'\mathbf{y}$ are:

$$\begin{pmatrix} \mathbf{x}_1'\mathbf{x}_1 & \mathbf{x}_1'\mathbf{x}_2 \\ \mathbf{x}_2'\mathbf{x}_1 & \mathbf{x}_2'\mathbf{x}_2 \end{pmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1'\mathbf{y} \\ \mathbf{x}_2'\mathbf{y} \end{pmatrix}.$$

We wish to find the least square estimate $\mathbf{b}_1$ for the slope coefficients $\beta_1, \ldots \beta_m$ of the first m regressors $X_1, \ldots, X_m$ under presence of r additional regressors $X_{m+1}, \ldots, X_{m+r}$. For that purpose we premultiply the first row of the normal equations

$$\mathbf{x}_1'\mathbf{x}_1\mathbf{b}_1 + \mathbf{x}_1'\mathbf{x}_2\mathbf{b}_2 = \mathbf{x}_1'\mathbf{y}$$

by $(\mathbf{x}_1'\mathbf{x}_1)^{-1}$ in order to obtain:

$$\begin{aligned} & \mathbf{b}_1 + (\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1'\mathbf{x}_2\mathbf{b}_2 = (\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1'\mathbf{y} \\ \Leftrightarrow\ & \mathbf{b}_1 = (\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1' \underbrace{(\mathbf{y} - \mathbf{x}_2\mathbf{b}_2)}_{\mathbf{u}_2}, \end{aligned}$$

This is in general not the same as $\mathbf{b}_1 = (\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1'\mathbf{y}$ which we would have got if we had regressed y on $x_1, \ldots, x_m$ alone.

Is there special situation in which adding an additional set of r regressors does not change the slope estimates $b_1, \ldots, b_m$, that is,

$$\mathbf{b}_1 = (\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1'\mathbf{u}_2 = (\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1'\mathbf{y}$$

except for the trivial case that $\mathbf{u}_2 = \mathbf{y}$ when the additional regressors have no explanatory power, that is $\mathbf{x}_2\mathbf{b}_2 = 0$? Yes, there is!

Consider the case that $\mathbf{x}_1'\mathbf{x}_2 = \mathbf{0}_{(m\times r)}$, then:

$$\begin{aligned}\mathbf{b}_1 &= (\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1'\mathbf{y} - (\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1'\mathbf{x}_2\mathbf{b}_2 \\ &= (\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1'\mathbf{y}.\end{aligned}$$

In order to understand the meaning of the requirement $\mathbf{x}_1'\mathbf{x}_2 = \mathbf{0}$ we calculate $\mathbf{x}_1'\mathbf{x}_2$:

$$\begin{aligned}\mathbf{x}_1'\mathbf{x}_2 &= \begin{pmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & & \vdots \\ x_{m,1} & \cdots & x_{m,n}\end{pmatrix}\begin{pmatrix} x_{m+1,1} & \cdots & x_{m+r,1} \\ \vdots & & \vdots \\ x_{m+1,n} & \cdots & x_{m+r,n}\end{pmatrix} \\ &= \begin{pmatrix} \sum x_1x_{m+1} & \cdots & \sum x_1x_{m+r} \\ \vdots & & \vdots \\ \sum x_mx_{m+1} & \cdots & \sum x_mx_{m+r}\end{pmatrix} \stackrel{!}{=} \mathbf{0}_{(m\times r)}.\end{aligned}$$

Now recall that

$$\sum x_ix_j = \sum(X_i - \bar{X}_i)(X_j - \bar{X}_j) \propto r_{ij},$$

because

$$r_{ij} = \frac{\sum(X_i - \bar{X}_i)(X_j - \bar{X}_j)}{\sqrt{\sum(X_i - \bar{X}_i)^2\sum(X_j - \bar{X}_j)^2}}.$$

That is, *adding/deleting a group of useful regressors to/from the model will leave the other coefficients unchanged only if all regressors from group 1 are uncorrelated with all regressors from group 2!*

Linear Regression Results

Model: Linear_Regression_Model
Dependent Variable: MPG MPG

Number of Observations Read	82
Number of Observations Used	82

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	582.30758	194.10253	105.01	<.0001
Error	78	144.18022	1.84846		
Corrected Total	81	726.48780			

Root MSE	1.35958	**R-Square**	0.8015
Dependent Mean	23.51220	**Adj R-Sq**	0.7939
Coeff Var	5.78246		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	48.61827	4.98552	9.75	<.0001	0
WEIGHT	WEIGHT	1	-0.00586	0.00072062	-8.13	<.0001	4.12927
LENGTH	LENGTH	1	0.01391	0.02680	0.52	0.6051	4.78615
WIDTH	WIDTH	1	-0.13021	0.11170	-1.17	0.2473	4.81613

Example. (Weiss: Example B.23)

Consider the preceding regression of a cars fuel efficiency (mpg) upon its weight, length and width. The estimated regression is:

$$\widehat{mpg} = 48.618 - 0.0059\text{weight} + 0.014\text{length} - 0.13\text{width}.$$

The regression estimates for the slope coefficients for all combinations of regressors are:

Variables included in regression equation	Slope estimates for weight	length	width
weight	-0.0063	–	–
length	–	-0.183	–
width	–	–	-0.787
weight, length	-0.0062	-0.003	–
weight, width	-0.0057	–	-0.099
length, width	–	-0.076	-0.512
weight, length, width	-0.0059	0.014	-0.130

While the slope coefficient of weight is fairly stable with respect to the presence/ absence of other regressors, there are large swings in both the slope coefficients for length (-0.183 − 0.014) and weight (-0.099 − -0.787). In the regression with all regressors present, the slope coefficient for length has even the wrong sign!

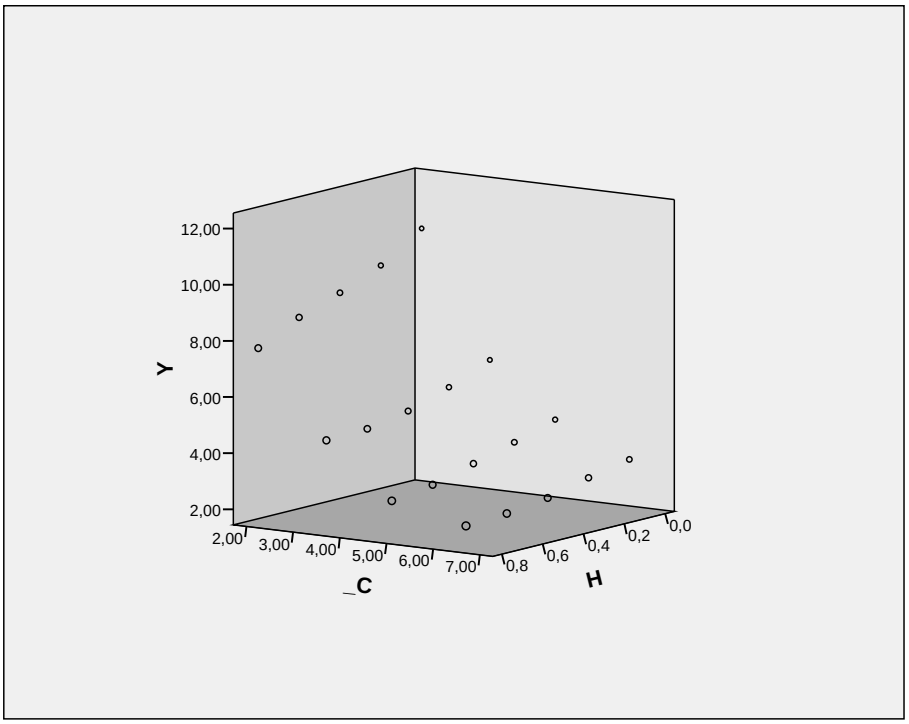
A Counter-Example: Uncorrelated Predictors (Weiss: Example B.27)

The adjoining table shows the outcome of an experiment, where the whiteness of paper (Y) has been measured as a function of two bleaching agents (C and H) used in its production.

Chlorine dioxide C	Hydrogen peroxide H	Whiteness measure Y
2.06	0.0	10.55
2.06	0.2	9.59
2.06	0.4	8.98
2.06	0.6	8.46
2.06	0.8	7.73
3.52	0.0	6.16
3.52	0.2	5.55
3.52	0.4	5.06
3.52	0.6	4.79
3.52	0.8	4.74
4.92	0.0	4.32
4.92	0.2	3.87
4.92	0.4	3.47
4.92	0.6	3.08
4.92	0.8	2.87
6.51	0.0	3.22
6.51	0.2	2.93
6.51	0.4	2.57
6.51	0.6	2.38
6.51	0.8	2.30

The scatterplot of H versus C shows a rectangular mesh of points with no trend. The correlation coefficient between C and H is 0. The regression output for regressing Y on either C or H alone, or both C and H shows that the sample regression coefficients remain the same as we include different sets of predictors in the regression equation.

3D Scatterplot of Dependent vs. Explanatory Variables

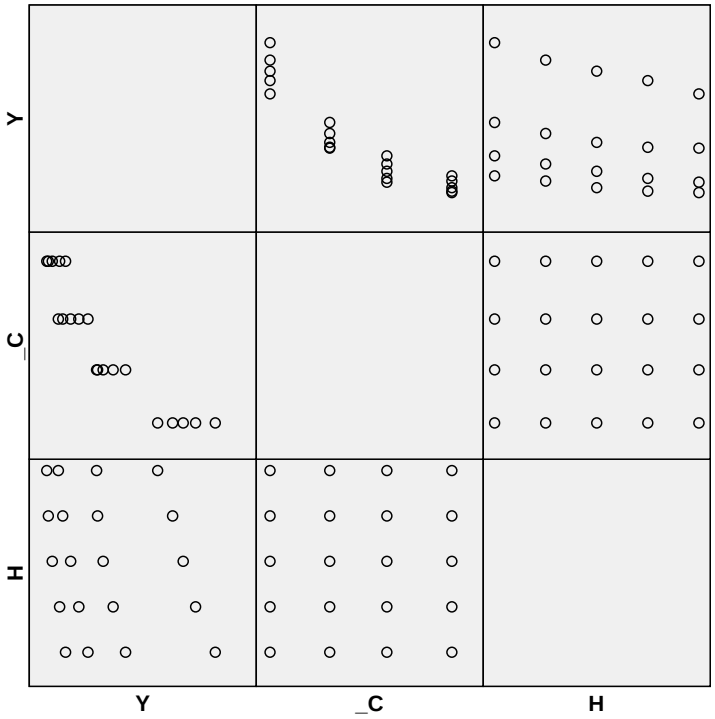


Correlations of Explanatory Variables

Correlations

		_C	H
_C	Pearson Correlation	1	.000
	Sig. (2-tailed)		1.000
	N	20	20
H	Pearson Correlation	.000	1
	Sig. (2-tailed)	1.000	
	N	20	20

Matrix Scatter Plot



Predictor variables are C and H

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.945[a]	.893	.880	.90087

a. Predictors: (Constant), H, _C

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	11.939	.626		19.083	.000		
	_C	-1.407	.122	-.916	-11.526	.000	1.000	1.000
	H	-2.056	.712	-.230	-2.887	.010	1.000	1.000

a. Dependent Variable: Y

Predictor variable is C

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.916[a]	.840	.831	1.06879

a. Predictors: (Constant), _C

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	11.116	.661		16.822	.000		
	_C	-1.407	.145	-.916	-9.715	.000	1.000	1.000

a. Dependent Variable: Y

Predictor variable is H

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.230[a]	.053	.000	2.59922

a. Predictors: (Constant), H

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	5.954	1.007		5.914	.000		
	H	-2.056	2.055	-.230	-1.001	.330	1.000	1.000

a. Dependent Variable: Y

2. *Lack of Robustness with respect to small changes in the data*

Rather than proving the general case we shall here only illustrate lack of robustness for the case of two strongly correlated regressors. For only two regressors, we may write the normal equations $(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$ in deviation form as

$$\begin{pmatrix} \sum x_1^2 & \sum x_1 x_2 \\ \sum x_1 x_2 & \sum x_2^2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} \sum x_1 y \\ \sum x_2 y \end{pmatrix}.$$

In our illustration we shall further restrict ourselves to the case that $\sum x_1^2 = \sum x_2^2 = 1$, as this allows us to interpret the off-diagonal terms $\sum x_1 x_2$ as the correlation coefficient r_{12} between X_1 and X_2, since then

$$\begin{aligned} r_{12} &= \frac{\sum (X_1 - \bar{X}_1)(X_2 - \bar{X}_2)}{\sqrt{\sum (X_1 - \bar{X}_1)^2 \sum (X_2 - \bar{X}_2)^2}} \\ &= \frac{\sum x_1 x_2}{\sqrt{\sum x_1^2 \sum x_2^2}} = \sum x_1 x_2. \end{aligned}$$

As an example of the normal equations

$$\begin{pmatrix} \sum x_1^2 & \sum x_1 x_2 \\ \sum x_1 x_2 & \sum x_2^2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} \sum x_1 y \\ \sum x_2 y \end{pmatrix}$$

for two highly correlated regressors consider

$$\begin{pmatrix} 1 & 0.999 \\ 0.999 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 2.998 \\ 2.999 \end{pmatrix}$$

with solutions $b_1 = 1$ and $b_2 = 2$.

Now consider adding/removing a few data points, such that the values on the right side of the normal equations change by about 1%:

$$\begin{pmatrix} 1 & 0.999 \\ 0.999 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 3.0285 \\ 2.9685 \end{pmatrix}$$

The new solutions are $b_1 = 31.5$ and $b_2 = -28.5$, completely unrelated to the original coefficients under almost the same data!

Compare this with uncorrelated regressors:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 2.998 \\ 2.999 \end{pmatrix},$$

where a 1% change on the right hand side yields but a 1% change in the coefficients.

3. *Inflated Standard Errors of the Sample Regression Coefficients*

Recall that the variance covariance matrix for the slope estimates $\mathbf{b}$ is $\text{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, such that in the case with only two regressors and design matrix in deviation form:

$$\begin{aligned}\text{Var}(\mathbf{b}) &= \begin{pmatrix} V(b_1) & \text{Cov}(b_1,b_2) \\ \text{Cov}(b_1,b_2) & V(b_2) \end{pmatrix} = \sigma^2(\mathbf{x}'\mathbf{x})^{-1} \\ &= \sigma^2 \begin{pmatrix} \mathbf{x}_1'\mathbf{x}_1 & \mathbf{x}_1'\mathbf{x}_2 \\ \mathbf{x}_2'\mathbf{x}_1 & \mathbf{x}_2'\mathbf{x}_2 \end{pmatrix}^{-1} = \sigma^2 \begin{pmatrix} \sum x_1^2 & \sum x_1x_2 \\ \sum x_1x_2 & \sum x_2^2 \end{pmatrix}^{-1} \\ &= \frac{\sigma^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2} \begin{pmatrix} \sum x_2^2 & -\sum x_1x_2 \\ -\sum x_1x_2 & \sum x_1^2 \end{pmatrix},\end{aligned}$$

by using the matrix inversion formula

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

This implies e.g. for the variance of b_1:

$$\begin{aligned}V(b_1) &= \frac{\sigma^2 \sum x_2^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2} = \frac{\sigma^2}{\sum x_1^2 - \frac{(\sum x_1x_2)^2}{\sum x_2^2}} \\ &= \frac{\sigma^2}{\sum x_1^2 \left(1 - \frac{(\sum x_1x_2)^2}{\sum x_1^2 \sum x_2^2}\right)} = \frac{\sigma^2}{\sum x_1^2(1-r_{12}^2)}.\end{aligned}$$

That is, the variance of the slope estimate may approach infinity as the correlation between the regressors approaches one.

Example. (Weiss: Example B.24)

Standard errors of the slope coefficients for all combinations of regressors for mpg:

Variables included in regression equation	Standard errors for weight	length	width
weight	0.0003532	–	–
length	–	0.01783	–
width	–	–	0.0706
weight, length	0.0006551	0.02263	–
weight, width	0.0006526	–	0.0937
length, width	–	0.03292	0.1368
weight, length, width	0.0007206	0.02680	0.1117

The standard errors of the slope coefficients vary by a factor of about 2. For each regressor the smallest standard error occurs for the regression which includes only that regressor alone. Adding predictors increases the standard error, even though MSE as an estimator of the residual variance is a decreasing function of the number of regressors, such that also $SE_{b_{i-1}} = \sqrt{\text{MSE}(\mathbf{X}'\mathbf{X})^{-1}_{ii}}$ should decrease. This increase in standard errors is a clear indication of multicollinearity.

4. Unreliable t-statistics

Since the t-statistic for testing the utility of individual regression parameters is calculated as $t = b_i/SE_{b_i}$, both of which we found to be severely affected by the presence of multicollinearity, it may come at no surprise that multicollinearity undermines the usefulness of this test. In severe cases of multicollinearity it is even possible that the F-test for the overall utility of the regression is significant, but that all t-tests turn out non-significant.

Example. (Weiss: Example B.25)

Variables included in regression model	t-statistic for utility of weight	length	width
weight	-17.78^{***}	–	–
length	–	-10.26^{***}	–
width	–	–	-11.15^{***}
weight, length	-9.48^{***}	-0.13	–
weight, width	-8.73^{***}	–	-1.06
length, width	–	-2.32^{*}	-3.75^{***}
weight, length, width	-8.13^{***}	0.52	-1.17

*/*** denotes significance at 5%/ 0.1%.

Length and width are judged significant only in regresssions not including weight.

No Effect of Multicollinearity on Prediction

Although multicollinearity affects the values and our understanding of the sample regression coefficients, it does not adversely affect the predicted value and the prediction interval for the response variable.

Example. (Weiss: Example B.26)

Consider the predicted value and prediction interval for mpg for a car with weight=3000, length=190, and width=70:

Variables included in regression model	R^2	s	Predicted mpg	Prediction interval for mpg
weight	0.798	1.354	24.582	21.868–27.296
length	0.568	1.980	22.181	18.852–26.784
width	0.609	1.886	23.618	19.843–27.393
weight, length	0.798	1.363	24.559	21.804–27.313
weight, width	0.801	1.353	24.496	21.779–27.214
length, width	0.634	1.836	23.291	19.604–26.978
weight, length, width	0.802	1.360	24.580	21.830–27.329

Based on R^2 and $s = \sqrt{\text{MSE}}$, regressions including weight are superior to those that don't ($R^2 \approx 0.8$, $s \approx 1.35$ vs. $R^2 \approx 0.6$, $s \approx 1.9$). The predicted values for mpg and the prediction intervals are similiar.

Detecting Multicollinearity:
Variance Inflation Factors (VIF's)

Recall the variance of the slope estimates b_i, $i = 1, 2$ in the case of two regressors:

$$V(b_i) = \frac{\sigma^2}{\sum x_i^2(1-r^2)},$$

where r was the correlation coefficient between the observations for the two regressors. The ratio of $V(b_i)$ to what it would be if the regressors were uncorrelated is:

$$\text{VIF} = \frac{\sigma^2}{\sum x_i^2(1-r^2)} \Big/ \frac{\sigma^2}{\sum x_i^2} = \frac{1}{1-r^2},$$

called the variance inflation factor of b_i.
Note that $r^2 = R^2$ in a regression of X_i on X_j. The variance inflation factor in the case of arbitrarily many regressors is:

$$\text{VIF}(b_i) = \frac{1}{1-R_i^2},$$

where R_i^2 denotes the coefficient of determination in a regression of X_i on all other regressors. Regressors with VIF$>$10 ($R^2 > 0.9$) should be eliminated from the regression.

Solutions to the Multicollinearity Problem

1. Make sure you haven't made any errors such as using too many dummies for too few cases or calculating some regressor as a (near) linear combination of others.

2. Get more and better data. Sometimes multicollinearity is just due to small sample size or bad sampling schemes.

3. Combine or apply some nonlinear transformation to (some of) your collinear regressors. (But be aware of the consequences for the normality assumption of the error term!)

4. Use out-of-sample information. Suppose prior research has shown that $\beta_1 = 2\beta_2$. Then, create a new variable $X_3 = X_1 - X_2$ and regress Y upon X_3 instead of both on X_1 and X_2. $\Rightarrow$ $b_3 = \widehat{\beta}_2$, $2b_3 = \widehat{\beta}_1$.

5. Use a partial F-test instead of t-tests for individual coefficients. That is, if X_1, X_2 and X_3 are highly correlated, test $\beta_1 = \beta_2 = \beta_3 = 0$ instead of testing $\beta_i = 0$, $i = 1, 2, 3$, individually.

6. The most simple and most often used solution is to drop collinear variables from the regression equation based on criteria such as the variance inflation factor. If that is done properly, we manage to reduce multicollinearity with only minor reduction in goodness of fit.

   However, erroneously dropping a relevant variable will result in a misspecified model, that is, the coefficient estimates will be biased with standard errors which underestimate the true variance of the estimators. So never drop a variable based upon software output alone, but reflect on whether the variable makes sense or not!