Lecture Notes for Multiple Linear Regression

Revised by Nicoleta Serban from Dr. Jeffrey Simonoff's original Regression and Multivariate Data Analysis class notes and from Dr. Kathryn Roeder and Dr. Larry Wasserman's original Regression Course notes

1 Multiple Linear Regression

1.1 Model

If Y depends on several variables, then we can extend our simple linear regression model to include more X's. For example we might predict the height of a child based on the height of father, height of mother, and sex of the child.

For each of n cases observed, values for the *response* (the y variable) and for each of the *predictors* (the x variables) are collected. The data will form an $n \times (k+1)$ array or *matrix*. The matrix decomposition for a linear regression model with n observations and k predictors is below

$$\left|\begin{array}{c|c|c|c|c|c|c} i= & Y & X_1 & X_2 & X_3 & \cdots & X_k \\ 1 & y_1 & x_{11} & x_{12} & \cdots & & x_{1k} \\ 2 & y_2 & x_{21} & x_{22} & \cdots & & x_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ n & y_n & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{array}\right|$$

The multiple linear regression model is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{1}$$

where $\beta = (\beta_0, \ldots, \beta_p)^T$ and $X = (1, X_1, \ldots, X_p)^T$. The value of the j^{th} covariate for the i^{th} subject is denoted by X_{ij}. Thus

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i. \tag{2}$$

At this point, it is convenient to use matrix notation. Let

$$\underset{n\times p}{\mathbb{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix}.$$

Each subject corresponds to one row. **The number of columns of $\mathbb{X}$ corresponds to the number of features plus 1 for the intercept** $q = p + 1$ Now define,

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}. \tag{3}$$

We can then rewrite (1) as

$$Y = \mathbb{X}\beta + \epsilon \tag{4}$$

Notational conventions.

Following some common notational conventions, we will denote a feature by the symbol X. If X is a vector, its components can be accessed by the subscripts X_j. An output, or response variable, is denoted by Y. We use uppercase letters such as X and Y when referring to the variables. Observed values are written as lower case, for example, the i'th observation of X is x_i. Matrices are represented using "mathbold font", for example, a set of n input p-vectors, $x_i, i = 1, \ldots, n$ would be represented by the $n \times q$ matrix $\mathbb{X}$. In general vectors will not be bold, except when they have n components; this convention distinguishes a q-vector of inputs x_i for the i'th observation from the n-vector $\mathbf{x}_j$ consisting of all the observations on the variable X_j. Since all vectors are assumed to be column vectors, the i'th row of $\mathbb{X}$ is x_i^T, the vector transpose of x_i.

Model flexibility.

When $k = 2$ the equation

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

gives the equation of a two dimensional plane or surface (plus some disturbances due to error).

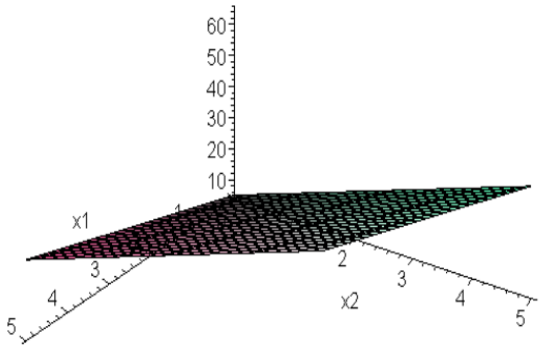


Figure 1: Graph of the regression plane $y = 5 + 5x_1 + 7x_2$: This is a natural extension of a simple linear regression model. When x_2 is constant, the surface in two dimensions is the regression line $y = 5 + 5x_1$; when x_1 is held constant, the surface in two dimensions is the line $y = 5 + 7x_2$

In general it is often desirable for some predictors to be mathematical functions of others in the sense that the resulting model may be much more successful in explaining variation than any model without such predictors.

The point here is that the linear model, when extended beyond many variables, is extremely flexible. The "linear" in linear regression refers to estimation of the response $\widehat{y}$ as a linear function of the observed data. For example, for the case of two independent variables x_1, x_2 four useful regression models are:

1. The first order model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
2. The second order, no interaction model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 {x_1}^2 + \beta_4 {x_2}^2 + \epsilon$.
3. The first-order interaction model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$
4. The complete second order model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 {x_1}^2 + \beta_4 {x_2}^2 + \beta_5 x_1 x_2 + \epsilon$

Model 1 is the most straightforward generalization of simple linear regression. It states that, for a fixed value of either variable, the expected value of Y is a linear function of the other variable and that the expected change in Y for a unit increase in x_1, x_2 is β_1, β_2 independent of the level of x_2, x_1, respectively. Thus if we graph a regression function as a function of only one variable, say x_1 for several different values of x_2 we obtain as *contours* of the regression function a collection of lines or curves.

In general, the expected change in Y for a one unit increase in a variable, say x_i, is the difference in predictions for $x_1 + 1$ and x_1 (can you guess why?).

$$E(Y_{x_1+1} - Y_{x_1}) = E(Y_{x_1+1}) - E(Y_{x_1})$$

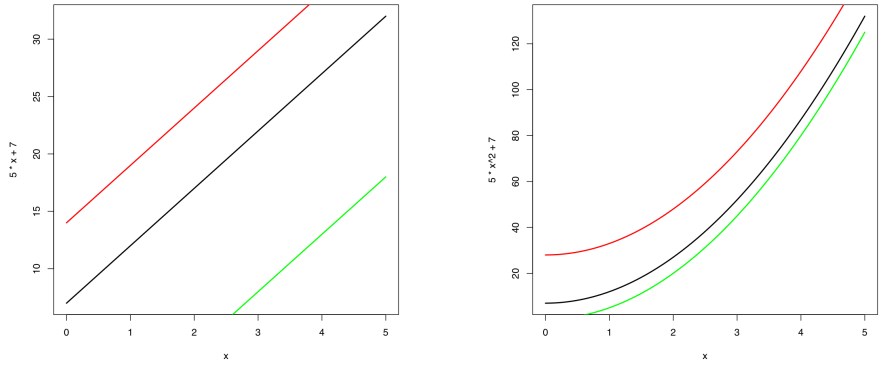
For example, for Model 2 (above, the second order no interaction model), when we fix x_2, the expected change in Y for a one unit increase in x_1 is: $\beta_1 + 2\beta_3 x_1$.

The contours of the regression function for model 3 are non parallel straight lines — for any *interaction* model they would be non parallel. For model 3, the expected change in Y when x_1 is increased by 1 is. $\beta_1 + \beta_3 x_2$. For any interaction model the expected change in Y will depend on the other variable(s). Figure out the expected change in Y for a unit increase in x_1 for model 4. Remember that these interaction models will fit the standard framework - for instance in model 3, let $x_1 x_2 = x_3$, say.

Remark: Why is this expected change talk important? Model interpretation, i.e. how do the predictors affect the response. In the bivariate (one x one y), simple linear regression case the expected change in Y given a unit change in x is always just the β_1 coefficient. In these multiple regression models, the interpretation is more difficult.

Qualitative variables.

Implicit in the discussion thus far is the assumption that predictor variables are quantitative. It is also possible to build models involving one or more qualitative variable. Suppose x_1 is a quantitative variable and the other variable of interest is qualitative with three different levels (say treatments 1,2 and 3). Then let



(a) (b)

Figure 2: (a): Contour graph of regression surface $y = 5 + 5x_1 + 7x_2$., when x_2 is held constant at $1, 2, -1$; (b): Contour graph of regression surface $y = 5 + 5{x_1}^2 + 7{x_2}^2$. when x_2 is held constant at $1, 2, 0$. Each two dimensional graph is the projection of the regression surface onto just the (y, x_1) plane. Each two dimensional graph illustrates the relationship between y and x_1 when all else is constant. These graphs are *non-interaction* models, hence the contour graphs are parallel.

$x_2 = \begin{cases} 1, & \text{if treatment 2;} \\ 0, & \text{otherwise.} \end{cases}$

and

$x_3 = \begin{cases} 1, & \text{if treatment 3;} \\ 0, & \text{otherwise.} \end{cases}$

Then the values $x_2 = x_3 = 0$ correspond to the second variable being at level 1. The predictors x_2 and x_3 can be included in the data matrix in the ordinary fashion and are called *indicator* variables or *dummy* variables.

Let D be a dummy variable. Consider

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X + \beta_2 D.$$

Then

coefficient	intercept	slope
$d = 0$	β_0	β_1
$d = 1$	$\beta_0 + \beta_2$	β_1

These are parallel lines. Now consider this model:

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 \, X \, D$$

Then:

coefficient	intercept	slope
$d = 0$	β_0	β_1
$d = 1$	$\beta_0 + \beta_2$	$\beta_1 + \beta_3$

These are nonparallel lines. To include a discrete variable with k levels, use $k - 1$ dummy variables. For example, if $z \in \{1, 2, 3\}$, do this:

z	d_1	d_2
1	1	0
2	0	1
3	0	0

In the model

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 X + \epsilon$$

we see

$$\begin{aligned} \mathbb{E}(Y|z=1) &= \beta_0 + \beta_1 + \beta_3 X \\ \mathbb{E}(Y|z=2) &= \beta_0 + \beta_2 + \beta_3 X \\ \mathbb{E}(Y|z=3) &= \beta_0 + \beta_3 X \end{aligned}$$

You should not create k dummy variables because they will not be linearly independent. Then $\mathbb{X}^T\mathbb{X}$ is not invertible.

1.2 Estimation

The extension of the least squares estimation approach to multiple linear regression is natural. Here the equation we minimize is:

$$f(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})]^2$$

This generates $(k+1)$ estimating or *normal equations* and thus for $(k+1)$ coefficients. Let $\widehat{\beta}$ be a $k \times 1$ vector. Then the least squares normal equations are:

$$\mathbb{X}^T\mathbb{X}\widehat{\beta} = \mathbb{X}^T\mathbf{Y}$$

which yields

$$\widehat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}.$$

We re-write the least squares estimator a

$$\widehat{\beta} = SY \text{ where } S = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \tag{5}$$

assuming that $(\mathbb{X}^T\mathbb{X})$ is invertible.

The vector of fitted values is

$$\begin{aligned} \widehat{Y} &= \mathbb{X}\widehat{\beta} \\ &= \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y \\ &= HY, \end{aligned}$$

The vector of residuals is $\widehat{\epsilon} = Y - \widehat{Y}$. The variance is then estimated by

$$\widehat{\sigma}^2 = \frac{\widehat{\epsilon}^T\widehat{\epsilon}}{n-p-1} = \frac{\text{RSS}}{n-q}. \tag{6}$$

Properties of Estimators

The estimators satisfy the following properties.

1. $\mathbb{E}(\widehat{\beta}) = \beta$.
2. $\mathbb{V}(\widehat{\beta}) = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1} \equiv \Sigma$.
3. $\widehat{\beta} \approx MN(\beta, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1})$.
4. *An approximate* $1-\alpha$ *confidence interval for* β_j *is*

$$\widehat{\beta}_j \pm z_{\alpha/2}\, \widehat{\mathsf{se}}(\widehat{\beta}_j) \tag{7}$$

where $\widehat{\mathsf{se}}(\widehat{\beta}_j)$ *is the square root of the appropriate diagonal element of the matrix* $\widehat{\sigma}^2(\mathbb{X}^T\mathbb{X})^{-1}$.

Let's prove the first two assertions. Note that

$$\mathbb{E}(\widehat{\beta}) \;=\; \mathbb{E}(SY) = S\mathbb{E}(Y) = S\mathbb{X}\beta = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\beta = \beta.$$

Also, by assumption $\mathbb{V}(Y) = \sigma^2 I$, where I is the identity matrix,

$$\begin{aligned}\mathbb{V}(\widehat{\beta}) &= \mathbb{V}(SY) = S\mathbb{V}(Y)S^T = \sigma^2 SS^T = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\Big((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\Big)^T \\ &= \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}.\end{aligned}$$

The ANOVA table is

Source	df	SS	MS	F
Regression	$q-1=p$	SS_{reg}	SS_{reg}/p	MS_{reg}/MSE
Residual	$n-q=n-p-1$	RSS	$RSS/(n-p-1)$	
Total	n-1	SS_{total}		

where $SS_{reg} = \sum_i(\widehat{Y}_i - \bar{Y})^2$ and $SS_{total} = \sum_i(Y_i - \bar{Y})^2$. We often loosely refer to "the degrees of freedom", but we should indicate whether we mean the df model (p) or the df error ($n-q = n-p-1$).

The F test $F = MS_{reg}/MSE$ is distributed $F_{p,n-p-1}$. This is testing the hypothesis

$$H_0 : \beta_1 = \cdots \beta_p = 0$$

Testing this hypothesis is of limited value. More frequently we test $H_0 : \beta_j = 0$. Based on an assumption of asymptotic normality one typically performs a t-test. The test statistic is of the form

$$T = \frac{\widehat{\beta}_j}{se_{\widehat{\beta}_j}}.$$

Reject H_0 if $|T|$ is large relative to a t-statistic with $(n-q)$ degrees of freedom. This tests if the j'th variable is associated with Y, after controlling for all the other variables. It is not the same as performed a test of association between X_j and Y "marginally". For instance, if X_1 and X_j are highly correlated, then β_j might not differ significantly from 0, even if X_j and Y are strongly associated.

Interpretation of the Estimated Coefficients

We must be very clear about the interpretation of a multiple regression coefficient. As usual, the constant term $\widehat{\beta}_0$ is an estimate of the expected value of the target variable when the predictors equal zero (only now there are several predictors). The estimate $\widehat{\beta}_j, j = 1, \ldots, p$ represents the estimated expected change in y associated with a one unit change in x_j holding all else in the model fixed. Consider the following example. Say we take a sample of college students and determine their College grade point average ($COLGPA$), High school GPA ($HSGPA$), and SAT score (SAT). We then build a model of $COLGPA$ as a function of $HSGPA$ and SAT:

$$COLGPA = 1.3 + .7HSGPA - .0003SAT$$

It is tempting to say (and many people do) that the coefficient for SAT has the "wrong sign," because it says that higher values of SAT are associated with lower values of College GPA. **This is absolutely incorrect!** What it says is that higher values of SAT are associated with lower values of College GPA, holding High school GPA fixed. High school GPA and SAT are no doubt correlated with each other, so changing SAT by one unit holding High school GPA fixed may not ever happen! The coefficients of a multiple regression must not be interpreted marginally! If you really are interested in the relationship between College GPA and just SAT, you should simply do a regression of College GPA on only SAT.

We can see whats going on here with some simple algebra. Consider the twopredictor regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i.$$

The least squares coefficients solve $(XX')^{-1}\beta = X'y$. In this case those equations are as follows:

$$\begin{aligned} n\beta_0 + \left(\sum x_{1i}\right)\beta_1 + \left(\sum x_{2i}\right)\beta_2 &= \sum y_i \\ \left(\sum x_{1i}\right)\beta_0 + \left(\sum x_{1i}^2\right)\beta_1 + \left(\sum x_{1i}x_{2i}\right)\beta_2 &= \left(\sum x_{1i}y_i\right) \\ \left(\sum x_{2i}\right)\beta_0 + \left(\sum x_{1i}x_{2i}\right)\beta_1 + \left(\sum x_{2i}^2\right)\beta_2 &= \left(\sum x_{2i}y_i\right) \end{aligned}$$

It is apparent that calculation of $\widehat{\beta}_1$ involves the variable x_2; similarly, the calculation of $\widehat{\beta}_2$ involves the variable x_1. That is, the form (and sign) of the regression coefficients depend on the presence or absence of whatever other variables are in the model. In some circumstances, this conditional statement is exactly what we want, and the coefficients can be interpreted directly, but in many situations, the "natural" coefficient refers to a marginal relationship, which the multiple regression coefficients do not address.

One of the most useful aspects of multiple regression is its ability to statistically represent a conditioning action that would otherwise be impossible. In experimental situations, it is common practice to change the setting of one holding others fixed, thereby isolating its effect, but this is not possible with observational data. Multiple regression provides a statistical version of this practice. This is the reasoning behind the use of "control variables" in multiple regression variables that are not necessarily of direct interest, but ones that the researcher wants to "correct for" in the analysis.

2 Multiple Linear Regression: Inference

2.1 The Hat Matrix

Recall that

$$\widehat{Y} = \mathbb{X}\widehat{\beta} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y = HY \tag{8}$$

where

$$H = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \tag{9}$$

is called the **hat** matrix. We can re-write the residuals as follows

$$\widehat{\epsilon} = Y - \widehat{Y} = Y - HY = (I - H)Y. \tag{10}$$

The hat matrix has the following properties.

1. $H\mathbb{X} = \mathbb{X}$.
2. *H is symetric and idempotent:* $H^2 = H$
3. *H projects Y onto the column space of $\mathbb{X}$.*
4. $\text{rank}(\mathbb{X}) = \text{tr}(H)$.

The residuals have the following properties.

1. *True residuals:* $\mathbb{E}(\epsilon) = 0$, $\mathbb{V}(\epsilon) = \sigma^2 I$.
2. *Estimated residuals:* $\mathbb{E}(\widehat{\epsilon}) = 0$, $\mathbb{V}(\widehat{\epsilon}) = \sigma^2(I - H)$.
3. $\sum_i \widehat{\epsilon}_i = 0$.
4. $\mathbb{V}(\widehat{\epsilon}_i) = \sigma^2(1 - h_{ii})$ *where h_{ii} is diagonal element of H.*

Below you will find the proofs of some of these properties for your reference although you will not need to know how to derive the properties,

$$\begin{aligned}
\mathbb{E}(\widehat{\epsilon}) &= (I - H)\mathbb{E}(Y) = (I - H)\mathbb{X}\beta \\
&= \mathbb{X}\beta - H\mathbb{X}\beta \\
&= \mathbb{X}\beta - \mathbb{X}\beta \quad \text{since } H\mathbb{X} = \mathbb{X} \\
&= 0.
\end{aligned}$$

Next,

$$\begin{aligned}
\mathbb{V}(\widehat{\epsilon}) &= (I-H)\mathbb{V}(Y)(I-H)^T \\
&= \sigma^2(I-H)(I-H)^T \\
&= \sigma^2(I-H)(I-H) \\
&= \sigma^2(I-H-H+H^2) \\
&= \sigma^2(I-H-H+H) \quad \text{since } H^2=H \\
&= \sigma^2(I-H).
\end{aligned}$$

2.2 Inference on Mean Response and Prediction

Call $\mathbf{X}_h = (1, X_{h,1}, ..., X_{h,k})$, a vector of observed predictors. Under the model $E(Y_h) = \mathbf{X}_h^T\beta$ and the fitted values $\widehat{Y_h} = \mathbf{X}_h^T\widehat{\beta}$. Of course, this fitted value the estimate of the expected value, or: $\widehat{Y_h} = \widehat{E(Y_h)}$.

We know this estimator is unbiased ($E(\widehat{Y_h}) = \mathbf{X_h}^T\beta$) and we can compute its variance ($Var(\widehat{Y_h}) = Var(\mathbf{X}_h\widehat{\beta}) = \sigma^2 \cdot \mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h = \mathbf{X}_h^T\Sigma_{\widehat{\beta}}\mathbf{X}_h$); its distribution is also the distribution of the residuals:

$$\widehat{Y_h} = N(\mathbf{X}_h^T\beta, \mathbf{X}_h^T\Sigma_{\widehat{\beta}}\mathbf{X}_h)$$

If we have to use an estimate of the variance, i.e. we replace σ^2 with $\widehat{\sigma}^2 = MSE$ then $\widehat{Var(\widehat{Y}_h)} = MSE \cdot \mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h = \mathbf{X}_h^T\widehat{\Sigma}_{\widehat{\beta}}\mathbf{X}_h$.

Of course this yields a $t-dist$ for inference: the $(1-\alpha)$ CI for $E(Y_h)$ — the CI for the mean response — is

$$\widehat{Y}_h \pm t_{\alpha/2,n-k} \cdot \sqrt{MSE \cdot \mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h}$$

For the CI for the *regression surface* — i.e. the mean response for all possible $\mathbf{X}_h$ — we replace the probability quantile t with F and k

$$\widehat{Y}_h \pm \sqrt{k \cdot F_{\alpha,k,n-k}} \cdot \sqrt{MSE \cdot \mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h}$$

This is the *Working-Hoteling* confidence band. This CI is also a conservative (wide) *simultaneous confidence interval for several mean responses*, in that all possible $\mathbf{X}_h$ vectors are covered in the interval/band. Alternately, the Bonferroni correction on the univariate intervals can be used:

$$\widehat{Y}_h \pm t_{\frac{\alpha}{2g},n-k} \cdot \sqrt{MSE \cdot \mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h}$$

where g is the number of confidence intervals you need.

The so-called *prediction interval* is just the CI for one particular value Y_h. The prediction interval for the mean of m new observations at predictors $\mathbf{X}_h$ is

$$\widehat{Y}_h \pm t_{\frac{\alpha}{2},n-k} \cdot \sqrt{MSE(\frac{1}{m} + \mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h)}$$

If $m = 1$ this is just an ordinary prediction interval for 1 new observation, each with the same observed predictors $\mathbf{X}_h$. When $m = m$ this is a *prediction interval* for the mean of m new observations. Wider bands for m new observations (not just the mean of m observations) are the *Scheffe* limits:

$$\widehat{Y}_h \pm \sqrt{m \cdot F_{\alpha;m,n-k}} \cdot \sqrt{MSE(\frac{1}{1} + \mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h)}$$

or the Bonferroni limits:

$$\widehat{Y}_h \pm t_{\frac{\alpha}{2m},n-k} \cdot \sqrt{MSE(\frac{1}{1} + \mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h)}$$

2.3 Testing Subsets of Coefficients

Let's revisit the ANOVA of a linear regression. The sum of total squares (SST) is decomposed into SSReg and SSE, the first is due to the regression and the second is due to error. However, we can further decompose SSReg as described below. Assuming that the order of predictors entering the model is X_1 enters first, X_2 enters second and so on, then we can decompose $SSReg = SS(X_1, X_2, \ldots, X_k)$ as follows

$$SSReg(X_1, X_2, \ldots, X_k) = SS(X_1) + SS(X_2|X_1) + SS(X_3|X_1, X_2) + \ldots + SS(X_k|X_1, \ldots, X_{k-1})$$

where
• $S(X_1)$ is the sum of squares explained by using only X_1 to predict Y
• $S(X_2|X_1)$ is the extra sum of squares explained by using only X_2 in addition to X_1 to predict Y
• $S(X_3|X_1, X_2)$ is the extra sum of squares explained by using only X_3 in addition to X_1 and X_2 to predict Y
and so on.

We can use this decomposition to answer the following questions: Does X_1 alone significantly aid in predicting Y? Does the addition of X_2 significantly contribute to the prediction of Y after we account (or control) for the contribution of X_2? Does the addition of X_3 significantly contribute to the prediction of Y after we account (or control) for the contribution of X_1 and X_2?

More generally, we may be interested in testing adding a subset of predictors significantly improves the prediction of Y. Thus, for a regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \beta_0 + \alpha_1 Z_1 + \cdots + \alpha_m Z_m + \epsilon$$

The hypothesis test is:

$$H_0 : \alpha_1 = \ldots = \alpha_m = 0 \text{ vs } H_A : \text{ at least one of the } \alpha_1, \ldots, \alpha_m \text{ not equal to zero}$$

For this, we use the partial F test where the F statistics is

$$F_{partial} = \frac{SS(Z_1, \ldots, Z_m | X_1, \ldots, X_k)}{MSE(X_1, \ldots, X_k, Z_1, \ldots, Z_m)}$$

We derive the extra sum of squares

$$SS(Z_1, \ldots, Z_m | X_1, \ldots, X_k) = SS(Z_1, \ldots, Z_m, X_1, \ldots, X_k) - SS(_1, \ldots, X_k) = SS_{full} - SS_{reduced}$$

or the difference between the sum of squares of the full model minus the sum of squares of the reduced model.

$MSE(X_1, \ldots, X_k, Z_1, \ldots, Z_m)$ is sum of squared errors of the full model.

The distribution of $F_{partial}$ under H_0 is a $F_{a,b}$ distribution, where df means degrees of freedom error, $a = df_{reduced} - df_{full}$ and $b = df_{full}$.

3 Diagnostics

Figure 3 shows a famous example. Four different data sets with the same fit. The moral: looking at the fit is not enough. We should also use some diagnostics. Generally, we diagnose problems by looking at the residuals. When we do this, we are looking for: (1) outliers, (2) influential points, (3) nonconstant variance, (4) nonlinearity, (5) nonnormality. The remedies are:

Problem	**Remedy**
1. Outliers	Non-influential: don't worry about it. Influential: remove or use robust regression.
2. Influential points	Fit regression with and without the point and report both analyses.
3. Nonconstant variance	Use transformation or nonparametric methods. Note: doesn't affect the fit too much; mainly an issue for confidence intervals.
4. Nonlinearity	Use transformation or nonparametric methods.
5. Nonnormality	Large samples: not a problem. Small samples: use transformations

Three types of residuals:

Name	Formula	R command (assume `lm` output is in `tmp`)
residual	$\widehat{\epsilon_i} = Y_i - \widehat{Y_i}$	`resid(tmp)`
standardized residual	$\frac{Y_i - \widehat{Y}_i}{\widehat{\sigma}\sqrt{1-h_{ii}}}$	`rstandard(tmp)`
studentized residual	$\frac{Y_i - \widehat{Y}_i}{\widehat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$	`rstudent(tmp)`

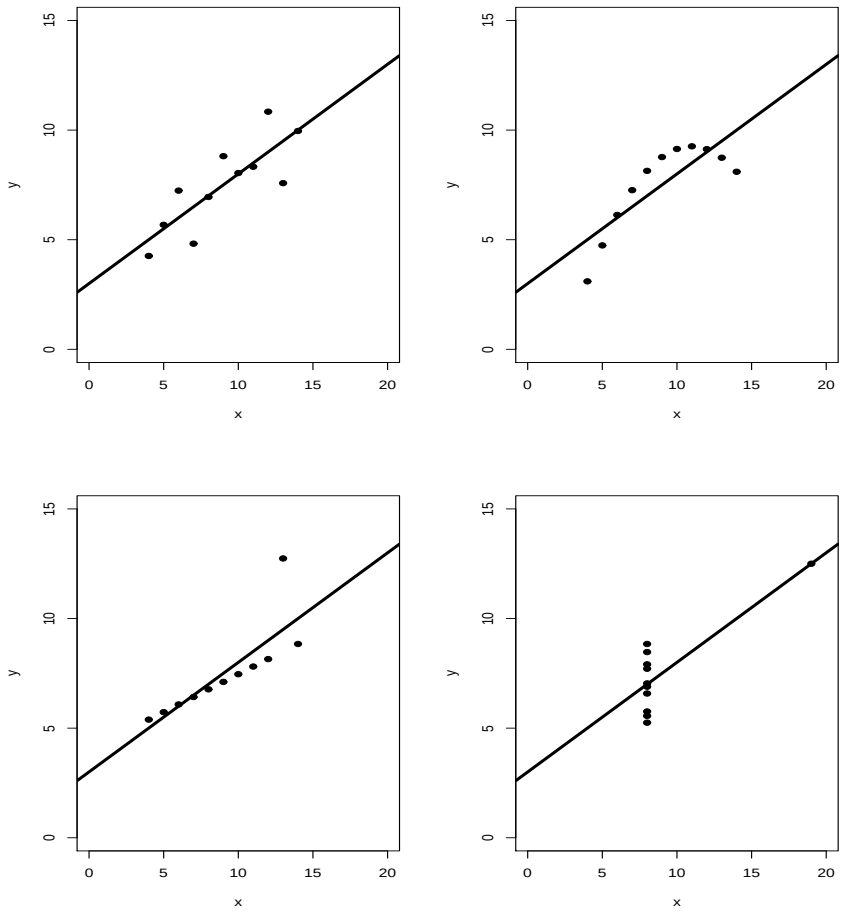


Figure 3: The Ansombe Example

3.1 Outliers

Can be found (i) graphically or (ii) by testing. Let us write

$$Y_j = \begin{cases} X_j^T\beta + \epsilon_j + \delta & j = i \\ X_j^T\beta + \epsilon_j & j \neq i. \end{cases}$$

Test

$$H_0 : \text{case } i \text{ is not an outlier} \quad \text{versus} \quad H_1 : \text{case } i \text{ is an outlier}$$

Do the following: (1) Delete case i. (2) Compute $\widehat{\beta}_{(i)}$ and $\widehat{\sigma}_{(i)}$. (3) Predict the deleted case: $\widetilde{Y}_i = X_i^T\widehat{\beta}_{(i)}$. (4) Compute

$$t_i = \frac{Y_i - \widetilde{Y}_i}{\widehat{\mathsf{se}}}.$$

(5) Reject H_0 if p-value is less than α/n.

Note that

$$\mathbb{V}(Y_i - \widetilde{Y}_i) = \mathbb{V}(Y_i) + \mathbb{V}(\widetilde{Y}_i) = \sigma^2 + \sigma^2 x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i.$$

So,

$$\widehat{\mathsf{se}}(Y_i - \widetilde{Y}_i) = \widehat{\sigma}\sqrt{1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i}.$$

How do the residuals come into this? Internally studentized residuals:

$$r_i = \frac{\widehat{\epsilon}_i}{\widehat{\sigma}\sqrt{1 - h_{ii}}}.$$

Externally studentized residuals:

$$r_{(i)} = \frac{\widehat{\epsilon}_i}{\widehat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}.$$

The relationship between the internally and externally studentized residuals is

$$t_i = r_i\sqrt{\frac{n - p - 2}{n - p - 1 - r_i^2}} = r_{(i)}.$$

3.2 Influence

Cook's distance

$$D_i \;=\; \frac{(\widehat{Y}_{(i)} - Y)^T(\widehat{Y}_{(i)} - Y)}{q\widehat{\sigma}^2} = \frac{1}{q} r_i^2 \left(\frac{h_{ii}}{1 - h_{ii}}\right)$$

where $\widehat{Y} = X\widehat{\beta}$ and $\widehat{Y}_{(i)} = X\widehat{\beta}_{(i)}$. Points with $D_i \geq 1$ might be influential. Points near the edge are typically the influential points.

3.3 Multicollinearity

An issue related to the interpretation of regression coefficients is that of multicollinearity. When the explanatory variables are highly correlated with each other, this can lead to instability in the regression coefficients, and the t statistics for the variables can be deflated. Mathematically speaking, if one of the predictor variables is a linear combination of the others, then we say that the variables are **collinear**. The result is that $\mathbb{X}^T\mathbb{X}$ is not invertible. Formally, this means that the standard error of $\widehat{\beta}$ is infinite and the standard error for predictions is infinite.

For example, suppose that $x_{1i} = 2$ and suppose we include an intercept. Then the $\mathbb{X}$ matrix is

$$\begin{pmatrix} 1 & 2 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 2 \end{pmatrix}$$

and so

$$\mathbb{X}^T\mathbb{X} = n \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$$

which is not invertible. The implied model in this example is

$$Y = \beta_0 + \beta_1 X_1 + \epsilon_i = \beta_0 + 2\beta_1 + \epsilon_i \equiv \widetilde{\beta}_0 + \epsilon_i$$

where $\widetilde{\beta}_0 = \beta_0 + 2\beta_1$. We can estimate $\widetilde{\beta}_0$ using $\overline{Y}$ but there is no way to separate this into estimates for β_0 and β_1.

Sometimes the variables are close to collinear. The result is that it may be difficult to invert $\mathbb{X}^T\mathbb{X}$. However, the bigger problem is that the standard errors will be *huge*.

From a practical point of view, multicollineary can lead to two problems:

1. If one value of one of the x variables is changed only slightly, the fitted regression coefficients can change dramatically.

2. It can happen that the overall F statistic is significant, yet each of the individual t statistics is not significant. Another indication of this problem is that the p value for the F test is considerably smaller than those of any of the individual coefficient t tests.

One problem that multicollinearity does not cause to any serious degree is inflation or deflation of overall measures of fit (R^2) since adding unneeded variables cannot reduce R^2 (it can only leave it roughly the same). Another problem with multicollinearity comes from attempting to use the regression model for prediction. In general, simple models tend to forecast better than more complex ones, since they make fewer assumptions about what the future must look like. That is, if a model exhibiting collinearity is used for prediction in the future, the implicit assumption is that the relationships among the predicting variables, as well as their relationship with the target variable, remain the same in the future. This is less likely to be true if the predicting variables are collinear.

How can we diagnose multicollinearity? We can get some guidance by looking again at a two predictor model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i.$$

It can be shown that in this case

$$V(\widehat{\beta}_1) = \sigma^2 \left[\sum x_{1i}^2 (1 - r_{12}^2)\right]^{-1}$$

and

$$V(\widehat{\beta}_2) = \sigma^2 \left[\sum x_{2i}^2 (1 - r_{12}^2)\right]^{-1}$$

where r_{12} is the correlation between x_1 and x_2. Note that as collinearity increases $|r_{12}| \to 1$ thus both variances tend to ∞. This effect can be quantified as follows:

r_{12}	Ratio of $V(\widehat{\beta}_1)$ to that if $r_{12} = 0$
0.00	1.00
0.50	1.33
0.70	1.96
0.80	2.78
0.90	5.26
0.95	10.26
0.97	16.92
0.99	50.26
0.995	100.00
0.999	500.00

This ratio describes by how much the variance of the estimated coefficient is inflated due to observed collinearity relative to when the predictors are uncorrelated.

A diagnostic to determine this in general is the *variance inflation factor* (VIF) for each predicting variable, which is defined as

$$VIF_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the R^2 of the regression of the variable x_j on the other predicting variables. The VIF gives the proportional increase in the variance of $\widehat{\beta}_j$ compared to what it would have been if the predicting variables had been completely uncorrelated. How big a VIF indicates a problem? A good guideline is that values satisfying

$$VIF < max\left(10, \frac{1}{1 - R_{model}^2}\right)$$

where R_{model}^2 is the usual R^2 for the regression fit, mean that either the predictors are more related to the target variable than they are to each other, or they are not related to each other very much. In these circumstances coefficient estimates are not very likely to be very unstable, so collinearity is not a problem.

What can we do about multicollinearity? Don't use all the variables; use variable selection (stay tuned...). Multicollinearity is just an extreme example of the *bias-variance tradeoff* we face whenever we do regression. If we include too many variables, we get poor predictions due to increased variance (more later).

3.4 Tweaking the Regression

If residual plots indicate some problem, we need to apply some remedies.

Possible remedies are:

- Transformation
- Robust regression
- nonparametric regression

Examples of transformations:

$$\sqrt{Y}, \log(Y), \log(Y+c), 1/Y$$

These can be applied to Y or x. We transform to make the assumptions valid, **not** to chase statistical significance.

3.4.1 *Transforming the Data

Often the relationship between x and y is not linear. There may be a theoretical basis, or not, for a non-linear relationship between the variables.

The necessity for an alternative model to the linear probabilistic model $Y = \beta_0 + \beta_1 x_1 + \epsilon$ may be suggested either by a theoretical argument or by examining plots from a linear regression analysis.

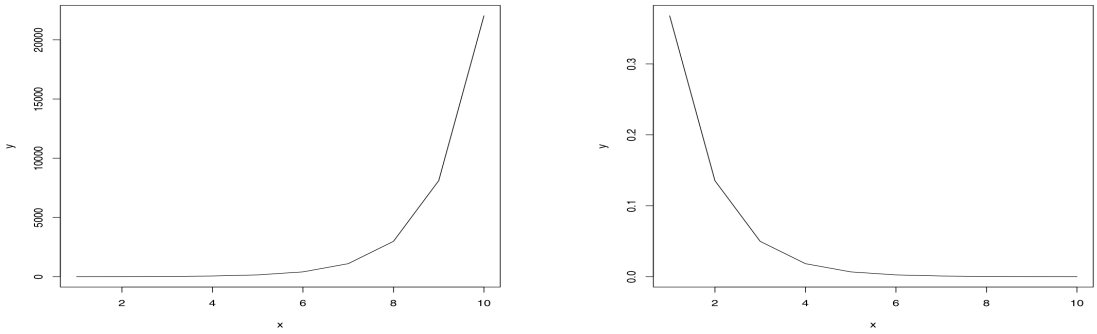
The goal is still to settle on a model whose parameters can easily be estimated.A probabilistic model relating Y to x is *intrinsically linear* if it can be reduced to a linear probabilistic model. Four of the most commonly used intrinsically linear probabilistic models are below:[1]

Try the below transformations when the data approximates the patterns of the graphs.

Exponential

The exponential transformation is useful in places where a theory behind the data supports the transformation.

[1]Lest you think all models can be intrinsically linear here are two that aren't: $y = \beta_0 + \gamma e^{\beta x}$ and $y = \alpha + \gamma x^{\beta}$



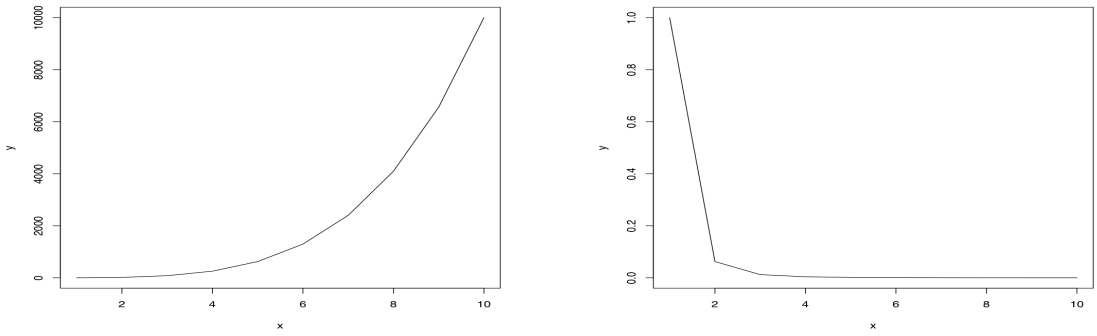
(a) (b)

Figure 4: Fig a:Graph function $y = e^x$ Fig b: Graph of function $y = e^{-x}$ Transform $y' = ln(y)$ to yield the linear form for ordinary linear regression.

The model is $y = \beta_0 e^{\beta x} \cdot \epsilon$. The linear form is $ln(y) = y' = \beta_0 + \beta_1 x' + \epsilon'$. The transformations are $x' = x, \beta_0 = ln(\beta_0), \beta_1 = \beta_1$ and $\epsilon' = ln(\epsilon)$.

Power

Power models are useful when we believe a nonlinear pattern exists in the data; often a theoretical reason.



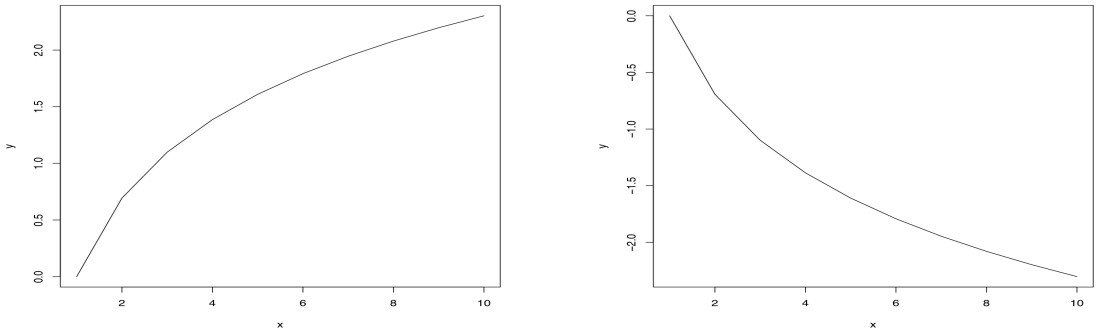
(a) (b)

Figure 5: Fig a:Graph of $y = \beta_0 x^{\beta} \cdot \epsilon$ Fig b: Graph of $y = \beta_0 x^{-\beta} \cdot \epsilon$ Here $log(y) = y' = \beta_0 + \beta_1 x' + \epsilon'$ with $x' = log(x), \beta_0 = log(\beta_0), \beta_1 = \beta, \epsilon' = log(\epsilon)$

The model is $y = \beta_0 x^{\beta} \cdot \epsilon$. The linear form is $log(y) = y' = \beta_0 + \beta_1 x' + \epsilon'$. The transformations are $x' = log(x), \beta_0 = log(\beta_0), \beta_1 = \beta$ and $\epsilon' = log(\epsilon)$.

Log x

The log transformation on x is useful as a *variance stabilizing transformation*



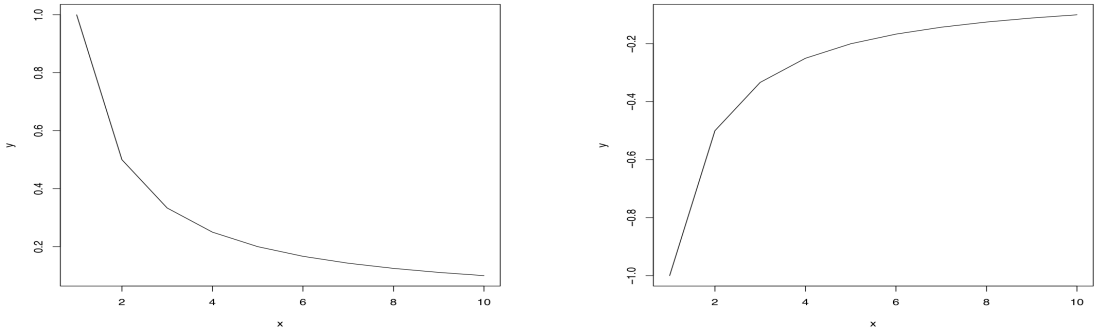
(a) (b)

Figure 6: Fig a:Graph of $y = \beta_0 + \beta \cdot log(x) + \epsilon$ Fig b: Graph of $y = \beta_0 + \beta \cdot log(-x) + \epsilon$. Here $x' = log(x)$ immediately linearizes the model.

The model is $y = \beta_0 + \beta log(x) + \epsilon$. The linear form is immediately generated by letting $x' = \log(x)$

Reciprocal

The reciprocal model is a specific case of the more general class of polynomial models.



(a) (b)

Figure 7: Fig a:Graph of $y = \beta_0 + \beta_1 \frac{1}{x} + \epsilon$ Fig b: Graph of $y = \beta_0 + \beta_1 \cdot \frac{1}{-x} + \epsilon$. Here $x' = \frac{1}{x}$ immediately linearizes the model.

Here the model is $y = \beta_0 + \beta \cdot \frac{1}{x} + \epsilon$ so that $x' = \frac{1}{x}$ yields a linear model.

Box-Cox

The normal error regression model with Y a member of the family of power transformations is

$$Y_i^\lambda = \beta_0 + \beta_1 X_i + \epsilon_i \sim N(0, \sigma^2)$$

The parameters to be estimated are now $\beta_0, \beta_1, \sigma^2$ and λ. Generally, these parameters are estimated via MLE.

By hand, you can estimate λ via numerical search in a range of potential values using

$$W_i = \begin{cases} K_1(Y_i^\lambda - 1) & \lambda \neq 0 \\ K_2(logY_i) & \lambda = 0 \end{cases}$$

with $K_2 = (\prod_{i=1}^n Y_i)^{1/n}$ and $K_1 = \frac{1}{\lambda K_2^{\lambda-1}}$.

The $W_i's$ are the standardized observations: for each value of λ you obtain the corresponding W_i's. Then use OLS regression treating the W_i's as predictors and select value of λ with the lowest SSE.

Interpretation or Words to the Wise

The major advantage of the intrinsically linear models is that the parameters (the β's) of the transformed model can be immediately estimated using the principle of least squares simply substituting x' and y' into the ordinary least squares estimating equations. Parameters of the original nonlinear model can then be estimated by transforming back the estimated (the $\widehat{\beta}$'s). For example, in the exponential and power models, when σ^2 is relatively small an approximate confidence interval for the fitted value of y given x^* results from taking the antilogs for the intervals generated by the nonlinear model.

4 Interaction effects for continuous predictors

4.1 Testing for interactions

The linear regression model is undoubtedly the most commonly-used statistical model, and has the advantage of wide applicability and ease of interpretation. The model has the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i$$

where y is the response variable, $\{x_1, \cdots, x_p\}$ are predictor variables, and ϵ is an error term. An implication of this model is that the partial relationship between y and any predictor x_j (given the other predictors are held fixed) is the same across all values of the predictors; specifically that holding all else fixed, a one unit change in x_j is associated with an expected β_j unit change in y, for any value of x_j and any values of the other predictors. When considering the constant relationship

between y and x_j for any value of another predictor, this is often referred to as the lack of an interaction effect of x_j on y given the value of a third variable. From a mathematical point of view, this is represented by the fact that the partial derivative

$$\frac{\partial y}{\partial x_j} = \beta_j$$

is a constant.

It is not uncommon for researchers and data analysts to consider the possibility that the effect of a predictor on the response could be different depending on the value of a third variable; that is, the presence of an interaction effect. The classic situation of this occurring is if the third variable is defining subgroups in the data, with the implication being that the slope of x_j differs depending on group membership. It is well-known that such a model can be fit by including in a regression model a set of indicator variables to define the groups, and all of the pairwise products of the indicator variables and the variable x_j (this can also be accomplished using effect codings; see Mayhew and Simonoff, 2015, for a full discussion of the use of effect codings to define subgroups in a data set). Consider the simplest situation of the presence of two subgroups A and B in the data and a single predictor x. Say an indicator variable I defines group membership, with $I = 0$ corresponding to membership in group A and $I = 1$ corresponding to membership group B. Fitting the regression model based on I, x, and their product Ix

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 I_i + \beta_3 I_i x_i + \epsilon_i$$

is equivalent to fitting the two separate lines

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

for members of group A($I = 0$) and

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 + \beta_3 x_i + \epsilon_i = \beta_0^* + \beta_1^* x_i + \epsilon_i$$

for members of group B ($I = 1$) . As can be seen, by including the product of I and x in the regression model, different slopes for the two groups are implied, representing the interaction effect of group membership and the numerical variable x. This generalizes for more than two subgroups to an analysis of covariance model (see Chatterjee and Simonoff, 2013, for extensive discussion of fitting such models).

This fact has had the unfortunate effect of resulting in researchers attempting to represent interactions between two numerical variables in the same way, by including their product as a predictor in a fitted regression,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i \tag{11}$$

This is problematic because using the t-test for whether the slope of the product variable equals 0 as an interaction test potentially results in errors of both types, Type I (mistakenly identifying a pattern that does not correspond to an interaction effect as an interaction) and Type II (mistakenly

deciding that no interaction effect is present when it actually is), no matter how large the sample is or how strong the underlying relationships are.

We will treat each of these issues in turn in the next two sections, illustrating them with simulated data. The data are a deliberately simplified version of the problem where the patterns are obvious, in order to illustrate the issues clearly; in a real data situation with multiple additional predictors the patterns could easily be less obvious to the eye, but just as serious. We will then discuss how to graphically uncover an interaction effect between two numerical variables, and how the use of additive models (a generalization of the linear model) can be an appropriate way to avoid mistakenly identifying a supposed interaction effect. We will then suggest a simple alternative approach for identifying interactions between numerical variables.

4.2 Problems with the product test for interactions

Mistakenly identifying nonlinearity as an interaction (Type I error)

The key idea is to recognize that (15) is not an interaction equation, but rather a nonlinear one. If nonlinearity is mistakenly identified as an interaction, a Type I error occurs. This can easily happen if the variables x_1 and x_2 are correlated with each other. Consider the following situation. Say the true underlying relationship is a quadratic one on variable x_1 alone; that is,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \epsilon_i.$$

The model on only x_1 and x_2 clearly cannot account for this quadratic relationship. If the product model (15) is fit instead, and if x_1 and x_2 are highly correlated,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \epsilon_i \approx \beta_0^* + \beta_1^* x_{1i} + \beta_2^* x_{2i} + \beta_3^* x_{1i} x_{2i} + \epsilon_i,$$

because up to constant terms or terms in x_1 or x_2 alone $x_1^2 \approx x_{1i} x_{2i}$. Thus, if a product term is included in the regression its t-statistic will be statistically significant, implying an interaction between x_1 and x_2 , when in fact what is present is a nonlinear relationship in x_1 alone.

Consider the following simulated example. The following regression output is based on fitting a regression with two predictors, x_1 and x_2 :

```
The regression equation is
y = 0.683 + 2.22 x1 - 1.80 x2

Predictor     Coef  SE Coef      T       P
Constant   0.6831   0.1449   4.71   0.000
x1          2.218    1.491   1.49   0.140
x2         -1.799    1.512  -1.19   0.237

S = 1.42172   R-Sq = 9.0%  R-Sq(adj)=7.2%

Analysis of Variance
```

```
Source          DF       SS     MS     F      P
Regression       2   19.496  9.748  4.82  0.010
Residual Error  97  196.066  2.021
Total           99  215.562
```

The overall regression is statistically significant, but neither predictor is; the reason for this is that the two predictors are highly correlated (the correlation between them is .994). The product test for an interaction now adds the product variable to the regression:

```
The regression equation is
y = - 0.0625 + 1.51 x1 - 0.592 x2 + 1.02 x1x2

Predictor      Coef  SE Coef      T      P
Constant   -0.06252  0.08655  -0.72  0.472
x1           1.5143   0.7624   1.99  0.050
x2          -0.5924   0.7756  -0.76  0.447
x1x2        1.01737  0.06125  16.61  0.000

S = 0.726073   R-Sq = 76.5%   R-Sq(adj) = 75.8%

Analysis of Variance

Source          DF       SS      MS       F      P
Regression       3  164.952  54.984  104.30  0.000
Residual Error  96   50.610   0.527
Total           99  215.562
```

The t-test is extremely highly statistically significant, apparently indicating an extremely strong interaction between the two predictors, but that is not in fact the case. The scatter plot in Figure 8 demonstrates what is actually going on: there is a quadratic relationship between y and x_1 , and the high correlation between x_1 and x_2 has resulted in the product of the two variables taking the place of the x_1^2 term. Thus, a nonlinear relationship in a single predictor has been misidentified as an interaction effect involving two predictors.

Mistakenly missing the presence an interaction (Type II error)

The product term in equation (15) can be viewed as an interaction effect on the response, as it does correspond to a differential effect of x_1 on y given the value of x_2 ; specifically,

$$\frac{\partial y}{\partial x_1} = \beta_1 + \beta_2 x_2.$$

The problem with the test is that this is a very specific form of an effect, and many interaction effects do not correspond to a relationship even close to this form. As a result, there are many

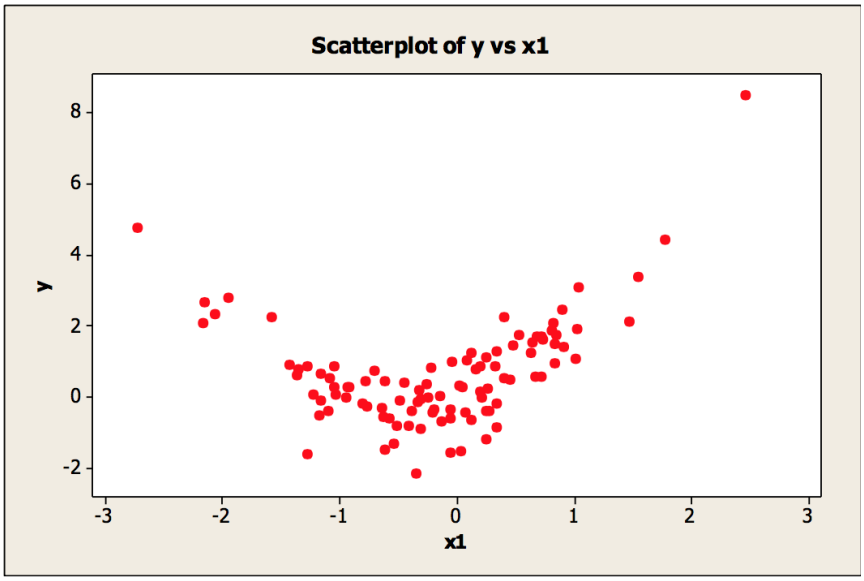


Figure 8: An illustration of a quadratic relationship.

situations where an actual interaction will be missed by the test of whether the slope of the product term equals 0.

Consider the following simulated example. The following regression output is based on fitting a regression with two predictors, x_1 and x_2 (note that y and x_2 are not the same as in the previous example):

```
The regression equation is
y = - 0.16 + 2.65 x1 - 0.0039 x2

Predictor       Coef  SE Coef      T      P
Constant     -0.159    1.956  -0.08  0.936
x1            2.652    1.148   2.31  0.023
x2         -0.00387  0.03347  -0.12  0.908
S = 10.2390   R-Sq = 5.2%   R-Sq(adj) = 3.3%

Analysis of Variance

Source           DF       SS     MS     F      P
Regression        2    562.4  281.2  2.68  0.073
Residual Error   97  10169.3  104.8
Total            99  10731.7
```

The overall regression is marginally statistically significant, as is the slope coefficient for x_1 . The product test for an interaction now adds the product variable to the regression:

```
The regression equation is
y = 0.03 + 3.62 x1 - 0.0066 x2 - 0.0219 x1x2

Predictor       Coef  SE Coef       T       P
Constant       0.031    1.999    0.02   0.988
x1             3.621    2.216    1.63   0.106
x2          -0.00658  0.03402   -0.19   0.847
x1x2        -0.02189  0.04276   -0.51   0.610

S = 10.2782   R-Sq = 5.5%   R-Sq(adj) = 2.5%

Analysis of Variance

Source           DF       SS      MS     F      P
Regression        3    590.1   196.7  1.86  0.141
Residual Error   96  10141.6   105.6
Total            99  10731.7
```

As is apparent, the product variable is not close to being statistically significant here, apparently implying that there is no interaction effect, but that is not in fact the case. There is in fact a very strong interaction effect: if $x_2 < 35$ or $x_2 > 70$ the slope $\beta_1 = 10$, and otherwise the slope $\beta_1 = -10$. This can be seen in the scatter plot in Figure 9, where the regions are labeled Low, Mid, and High:

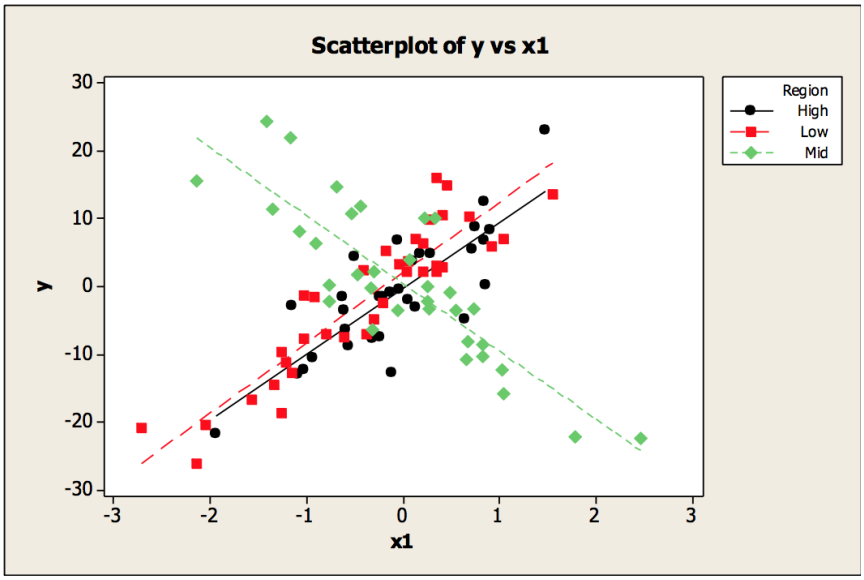


Figure 9: An illustration of an interaction.

Since this interaction does not "look like" a product term, the test has no power to identify it, even though doing so correctly would result in a strong fit (an R^2 more than 75% and a highly

statistically significant interaction effect corresponding to different slopes for the three "regions" of x_2).

4.3 Identifying interaction effects

Given the deficiencies in using the product of two numerical predictors to test for the presence of an interaction effect, a natural question to ask is whether there are better methods.. We will describe a graphical technique (termed a *trellis display*) that can help expose the presence of an interaction effect.

A trellis display is a version of a conditioning plot; it highlights patterns in the data conditioning on the value of a specific variable. Since this is precisely what an interaction effect in regression represents (the relationship between the response and a predictor changing based on the value of another variable), such a display is ideal for exploring graphically the possibility of an interaction effect. The display in Figure 10 gives a display for the second data set given above prepared using the lattice package of the R software package (Sarkar, 2008). Recall that in that data set the slope between y and x_1 changes depending on the value of x_2 . The plot is constructed by defining subregions based on the conditioning variable x_2 ; a simple default (used here) is to divide the data into regions with roughly equal numbers of observations. Each panel of the display is a scatter plot of y versus x_1 for the observations in that x_2 subregion. The subregions go from smallest values of x_2 in the lower left to largest values in the upper right, and are identified by the shading at the top of each plot in the display.

It is apparent in the display that for smaller values of x_2 there is a direct relationship between y and x_1 , for moderate values there is an inverse relationship, and for large values there is again a direct relationship. Thus, the plot easily summarizes the interaction effect in the data. As is true for any scatter plot in a multiple regression the display is in general only suggestive, since it cannot account for the effects of predictors other than x_1 and x_2 on the relationship between y and x_1 given x_2 , but it is certainly worth constructing if the possibility of an interaction effect is contemplated.

References

Chatterjee, S. and Simonoff, J.S. (2013), *Handbook of Regression Analysis*, Wiley: Hoboken, NJ.
Mayhew, M.J. and Simonoff, J.S. (2015), "Nonwhite, No More: Effect Coding as an Alternative to Dummy Coding with Implications for Researchers in Higher Education," *Journal of College Student Development*, 56, 170-175.
Sarkar, D. (2008), Lattice: *Multivariate Data Visualization with R*, Springer: New York.

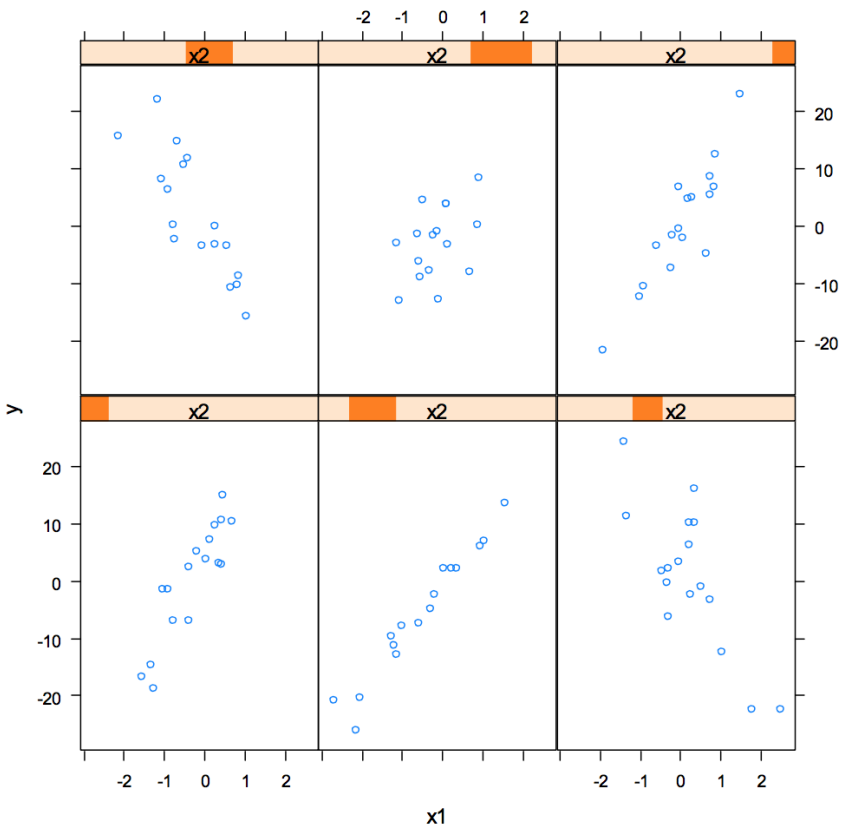


Figure 10: Another illustration of the presence of an interaction.

5 Optional Material: Ordinary least squared estimation and time series data

One of the assumptions underlying ordinary least squares (OLS) estimation is that the errors be uncorrelated. Of course, this assumption can easily be violated for time series data, since it is quite reasonable to think that a prediction that is (say) too high in June could also be too high in May and July. That kind of cyclical effect is indicative of positive autocorrelation, and it is quite common in time series data. But say we ignore this fact; why is it a problem to use OLS if the errors are autocorrelated?

The following two tables can help to answer that. Consider a simple regression problem, and let ρ be the first order autocorrelation of the errors (i.e., $\rho = corr(\epsilon_i, \epsilon_{i+1})$) and λ be the first order autocorrelation of the predicting variable x (it's likely that this, too, would exhibit autocorrelation; this is not a violation of any assumptions, but it can affect the properties of OLS estimators if there is also autocorrelation of errors). Further, assume the particular autocorrelation structure known as a *first order autoregressive model* (we'll talk more about this a little later).

The first problem with using OLS estimates in the context of autocorrelated errors is that they are *inefficient*; that is, they have higher variability (as estimates of the true parameters) than they

should. The following table gives efficiency of the OLS estimator of β_1 compared to the best possible estimator (the efficiency is simply the ratio variances):

	ρ								
λ	–0.9	–0.8	–0.5	–0.2	0	0.2	0.5	0.8	0.9
0.0	10.5	22.0	60.0	92.3	100.0	92.3	60.0	22.0	10.5
0.2	12.6	25.4	63.2	92.9	100.0	92.3	58.4	19.8	9.1
0.5	18.5	34.3	71.4	94.6	100.0	93.5	60.0	18.4	7.9
0.8	35.9	56.2	85.4	97.5	100.0	96.6	71.4	22.0	8.4
0.9	52.8	71.8	92.0	98.7	100.0	98.1	81.3	29.3	10.5

Figure 11

It is apparent that if errors are autocorrelated, the OLS estimator can be seriously inefficient. For example, if $\rho = \lambda = .9$, the variance of the OLS estimator is 10 times that of the best estimator. For positively autocorrelated errors, the inefficiency is fairly insensitive to the autocorrelation of the predictor, but for negatively autocorrelated errors, a positively autocorrelated predictor can actually help (it's fairly unlikely, however, that the signs of the autocorrelations of the predictor and of the errors would be different). Note, by the way, that results for negatively autocorrelated predictors mimic those above, except that the role of the sign of the autocorrelation of the errors is reversed (a negatively autocorrelated predictor is more trouble for negatively autocorrelated errors).

This is not good, but an even bigger problem also exists. The standard error of $\widehat{\beta}_1$ that is output by the computer is no longer correct; it is estimating the wrong thing. The following table gives the percentage bias in estimating $var(\widehat{\beta}_1)$ if the OLS computer output is used:

	ρ								
λ	–0.9	–0.8	–0.5	–0.2	0	0.2	0.5	0.8	0.9
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.2	43.9	38.1	22.2	8.3	0.0	–7.7	–18.2	–27.5	–30.5
0.5	163.6	133.3	40.0	22.2	0.0	–18.2	–40.0	–57.1	–62.1
0.8	514.3	355.6	133.3	38.1	0.0	–27.6	–57.1	–78.0	–83.7
0.9	852.6	514.3	163.6	43.9	0.0	–30.5	–62.1	–83.7	–89.5

Figure 12

It is apparent that using the usual measures of fit can lead to very misleading inferences. For example, if $\rho = \lambda = .9$, the estimated variance of $\widehat{\beta}_1$ is about 10% of its true value. This implies that the t–statistic for β_1 is about 3.1 times too large (a similar inflation of F and R^2 values also occurs). Thus, if left uncorrected, an insignificant relationship (say $t = 1.5$) can be mistakenly viewed as highly significant (apparent $t = 4.65$). It often happens that a regression on time series data with $R^2 = .8$ has the R^2 drop down to .3 or .4 when autocorrelation is addressed. Note also that if λ and ρ are of opposite sign, the apparent strength of the regression is too low, rather than too high.

Identifying autocorrelation

Since it is autocorrelation of the errors that is a violation of regression assumptions, it shouldn't be surprising that it is the (standardized) residuals that are used to identify possible autocorrelation. We've already talked about one way that this is done–whenever there is a time ordering to the data, a time series plot of the residuals should be constructed and examined for possible evidence of cyclical behavior. Note that the observations in all time series data must be ordered in correct chronological order (earliest to latest), rather than reverse order, or else tests and estimation methods are incorrect.

Consider a hypothetical data set $\{x_1, x_2, \cdots, x_p, y\}$, and a hypothesized linear regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i.$$

In addition to a time series plot of the residuals, there are several formal tests that can be used to identify the presence of autocorrelation in the residuals.

The *Durbin–Watson test* is a highly parametric test for autocorrelation. It is assumed that under the null, all of the usual assumptions for regression hold:

$$H_0 : \epsilon_i \sim N(0, \sigma^2) \text{ for all } i,\ corr(\epsilon_i, \epsilon_j) = 0 \text{ for } i \neq j.$$

Further, the form of possible autocorrelation is also specified as being from an autoregressive model of order 1 [AR(1)]:

$$H_a : \epsilon \sim AR(1), \rho \neq 0$$

The AR(1) model says that autocorrelation is due to the following structure for the errors:

$$\epsilon_i = \rho\epsilon_{i-1} + z_i, \quad i = 2, \cdots, n,$$

where $|\rho| < 1$ and the z_i are independent $N(0, \sigma^2)$ random variables. This implies that $\rho_s = \rho^s$, where ρ_s is the s^{th} order autocorrelation $[corr(\epsilon_i, \epsilon_{i+s})]$. So, for example, if $\rho = .7$, then $\rho_1 = .7$, $\rho_2 = .49$, $\rho_3 = .34$, and so on. That is, the autocorrelation in the errors goes down geometrically as the distance between them goes up.

The Durbin–Watson test is simply

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2},$$

where e_i is the i^{th} residual. Small values of DW indicate positive autocorrelation, while large values indicate negative autocorrelation.

The enclosed tables give critical values for the test. The Durbin–Watson statistic has the unfortunate property that it is not *pivotal*; that is, its distribution under the null hypothesis is a function of the actual data values. This means that there is no single critical value (given the number of observations and predictors) for which you can determine statistical significance, but rather only a range of critical values (see the attached tables). For large samples ($n \geq 100$, say), an approximate z–statistic for the Durbin–Watson test is

$$z = (\frac{DW}{2} - 1)\sqrt{n},$$

which can be compared to a Gaussian critical value. Note that a positive value of z indicates negative autocorrelation, while a negative value indicates positive autocorrelation.

Since the Durbin–Watson test has so many assumptions, it is important to check them if you're going to use it. That is, you must look at residual plots to check homoscedasticity and normality, and you can look at an autocorrelation function (ACF) plot to see if the observed autocorrelations appear to be consistent with the AR(1) model; this plot gives correlations of a variable with itself shifted by one time period, two time periods, etc. If the autocorrelations decay roughly at the geometric rate consistent with an AR(1) process, that supports the use of the Durbin–Watson test. Note that a complete lack of evidence of any autocorrelation in an ACF plot is consistent with an AR(1) process (one with $\rho = 0$). The ACF plot also can be used to identify other types of autocorrelation, such as seasonality and nonstationarity, which we will discuss later.

The individual residual autocorrelations that are used to construct the autocorrelation function also can be used to test for autocorrelation in the errors, since they should be close to zero if there is no autocorrelation in the errors. For each order autocorrelation, the hypotheses being tested are

$$H_0 : \rho_j = 0$$

versus

$$H_a : \rho_j \neq 0.$$

The ACF plot gives $\alpha = .05$ limits for each estimated autocorrelation, corresponding to a t–test for the significance of the estimated coefficient relative to the true coefficient equaling zero. This test requires the usual assumptions on the errors for small samples, but normality is not needed for large samples (as the Central Limit Theorem applies).

The runs test, being nonparametric in nature, requires virtually no assumptions about the data. The hypotheses being tested are very general:

$$H_0 : \text{ there is no autocorrelation in the errors}$$

versus

$$H_a : \text{ there is autocorrelation.}$$

Let n_+ be the number of positive residuals, and n_- be the number of negative ones in a sample of size n. A "run" is defined as a set of consecutive observations where the residuals have the same sign. In the presence of positive autocorrelation you would expect positive and negative residuals to tend to occur together, resulting in fewer than expected runs; in the presence of negative autocorrelation you would expect positive residuals to tend to be followed by negative ones (and vice versa), resulting in more runs than expected. It can be shown that under H_0 the expected number of runs is $\mu = \frac{2n_+n_-}{n} + 1 \approx \frac{n}{2} + 1$, while the variance of the number of runs is $\sigma^2 = \frac{2n_+n_-(2n_+n_- - n)}{n^2(n-1)} \approx \frac{n^2-2n}{4(n-1)}$. . The runs test is a z–test, comparing the observed number of runs u to the expected number:

$$z = \frac{|u - \mu| - \frac{1}{2}}{\sigma}$$

(the "$-\frac{1}{2}$" is a continuity correction). For large enough n, z is approximately normally distributed, and a tail probability can be obtained.

Note, by the way, that neither the runs test nor ACF plot can be produced in Minitab if there are missing values in the residuals (other than at the very beginning or very end of the series).

Say we've identified autocorrelation in the residuals from a regression. What should we do? There are many different possibilities, ranging from relatively simple approaches to quite complicated ones. Note, by the way, that addressing autocorrelation is not the full story – this is still a regression problem, and all of the usual checks (scatter plots, residual plots, diagnostics, etc.) are still essential.

Detrending and deseasonalizing

The structure in time series data is often greatly simplified if broad trends and seasonal effects are removed. If a time series plot of a variable shows steadily increasing (or decreasing) values over time, the variable can be detrended by running a regression on a time index variable (that is,

the case number), and then using the residuals as the detrended series. If the series has natural seasonal effects, these too can be handled using regression. For example, say the variable being examined is quarterly sales of ice cream. We wouldn't be surprised to see seasonal effects in such a variable. It can be deseasonalized by running a regression on three indicator variables that identify first, second, and third quarter observations, respectively, and then using the residuals as the deseasonalized series. Seasonal effects are often apparent in the ACF plot of the residuals as a large autocorrelation at the lag corresponding to the period of the seasonality (lag 4 for quarterly data, lag 12 for monthly data, etc.). So, for example, unemployment rates are usually reported in deseasonalized form, which roughly corresponds to taking the residuals from a regression on the quarter effects and adding the overall average unemployment rate back.

Usually, we are not interested in creating as a final product detrended or deseasonalized variables; rather, we would just like to include trend and/or seasonal effects as part of a time series regression model. In this situation, all that is required is to add a time index and/or seasonal indicator variables as additional predictors in the regression model (that is, you typically don't need to detrend or deseasonalize each variable in the model separately). Of course, this implies that you should look at a time series plot of the target variable (that is, a plot of the target variable versus the time index), and side–by–side boxplots of the target variable separated by season to see if detrending or deseasonalizing seem to be worth considering. Indeed, you should look at these plots routinely in any regression on time series data (the boxplots for a seasonal effect only if it is meaningful in your context, of course), since trend and seasonal effects are so common.

Sometimes trend effects are actually reflecting more fundamental properties of the variables that should be addressed. For example, variables that are related to population will increase over time simply because of increasing population. This is not what we're typically interested in, so such effects should be removed before analyzing the data by converting to per capita measures. Similarly, variables measured in current dollars will increase over time because of inflation; such variables should be converted to constant dollars using a price deflator like the consumer price index.

Lagging and differencing

Autocorrelation can sometimes be handled by using values of the target variable from the previous time period(s) as predictors in the model. So, for example, this quarter's sales might be regressed on last quarter's sales. Such variables are called lagged variables. A plot of the target versus the lagged target will often show a strong relationship between the two. Even more importantly, in a regression , autocorrelation has pretty much disappeared (an important point, however: the assumed distribution for the Durbin–Watson statistic is not valid if a lagged version of the target is used as a predictor, so the observed statistic should not be evaluated for statistical significance). Obviously, if you are considering using the lagged response as a predictor, you need to look at a scatter plot of the response versus the lagged response, since you need to look at scatter plots of the response versus **all** potential predictors.

In a multiple regression, if you decide to include a lagged target variable as a potential predictor, it is just that – a potential predictor – and should be treated as such. So, for example, performing a best subsets regression based on all of your available predictors is appropriate, since the presence

of the lagged target might change the usefulness of other variables in the model (that is, your previously chosen predictors might no longer be the appropriate choices).

A transformation related to lagging is differencing the data. This operation is appropriate when the **change** in a variable from one time period to the next is hypothesized to be uncorrelated with the changes at other time points. Data that follow a random walk, and more generally series that are *nonstationary*, satisfy this condition, and benefit from differencing, which helps explain why stock returns are much more interesting than stock prices (indeed, stock returns are almost **always** the correct thing to study, rather than stock prices). Differencing the target variable is often particularly useful when an ACF plot of the residuals indicates slowly decaying autocorrelations.

It is important to remember that once a target variable is differenced, the problem has changed. The goal is no longer to try to build a model that predicts the (original) target, but rather now the *change* in the target. The close connection between differencing a target variable and lagging one can be seen from the regression model form after differencing:

$$y_i - y_{i-1} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i$$

is equivalent to a regression including a lagged value of y as a predictor,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \beta_{p+1} y_{i-1} + \epsilon_i,$$

with the coefficient of the lagged $y(\beta_{p+1})$ equaling one.

If the target variable y follows a random walk, this relationship is particularly clear. A random walk is characterized by the change in y being random Gaussian noise; that is, $\beta_1 = \cdots = \beta_p = 0$ in the models above. A regression of y on lagged y will be highly significant (with coefficient for lagged y close to one), while one with differenced y as the target will be insignificant, yet **the two models are reflecting exactly the same relationship**. This can be seen if the standard errors of the estimate for the two models are compared; since both models have the same error structure, the estimates of σ will be very close, even though one model has a high R^2 and the other has a low one. Stock prices are often modeled as following a geometric random walk. What that implies is that logged stock price yesterday is a very good predictor of logged stock price today. What it also means is that the change in logged stock price (the return) is pretty much unrelated to anything else. The latter model, with its low R^2 and F statistics, is the correct representation of the relationship, even though it "doesn't look as good."

It is important to note, however, that differencing data can sometimes result in what is called *overdifferencing*, where the resultant series exhibits significant negative autocorrelation. In this circumstance, using a lagged version of the target as the predictor can be considerably more effective than differencing the data, which would be reflected in the estimated coefficient for the lagged target variable being significantly different from 1.

Using a lagged response variable as a predictor can also help address seasonality. For example, in quarterly data, using the response variable lagged by four periods (that is, the response from the period one year earlier) can account for quarterly seasonal effects in an effective way.

Predicting variables also can be included in a regression in lagged or differenced form. It is important to recognize, however, that this is only sensible if there is a good reason to believe that such variables have predictive power. So, for example, if you believed that higher interest rates

might cause higher unemployment, but only after a three month lag, it would be sensible to use interest rates from three months earlier as a predictor.

6 Optional Material: Weighted Least Squares

So far we have assumed that the ϵ_i's are independent and have the same variance. What happens if the variance is not constant? For example, Sheather's text gives a simple example about a cleaning company. The building maintenance company keeps track of how many crews it has working (X) and the number of rooms cleaned (Y). The number of crews varied from 2 to 16, and for each level of X, at least 4 observations of Y are available. A plot of X versus Y reveals that the relationship is linear, but that the variance grows as X increases. Because we have several observations for each level of X we can estimate σ^2 as a function of X. (Of course, we don't usually have multiple measures of Y for each level of X, so we will need more subtle ways of handling this problem.)

For another example, suppose D_i is the number of diseased individuals in a population of size m_i and $Y_i = \frac{D_i}{m_i}$. Under certain assumptions, it might be reasonable to assume that $D_i \sim \text{binomial}$, in which case $\mathbb{V}[Y_i]$ would be proportional to $1/m_i$. If the disease is contagious the binomial assumption would not be correct. Nevertheless, provided m_i is large for each i, it might be reasonable to assume that Y_i is approximately normal with mean $\beta_0 + \beta_1 x_i$ and variance $\frac{\sigma^2}{m_i}$. In this case the variance is a function of m_i, and we could model this variance as described below.

Suppose that

$$Y = \mathbb{X}\beta + \epsilon$$

where

$$\mathbb{V}(\epsilon) = \Sigma.$$

Suppose we use the usual least squares estimator $\widehat{\beta}$. Then,

$$\begin{aligned} \mathbb{E}(\widehat{\beta}) &= \mathbb{E}((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y) \\ &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{E}(Y) \\ &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\beta = \beta. \end{aligned}$$

So $\widehat{\beta}$ is still unbiased. Also, under weak conditions, it can be shown that $\widehat{\beta}$ is consistent (converges to β as we get more data). The usual estimator has reasonable properties. However, there are two problems.

First, with constant variance, the usual least squares estimator is not just unbiased, it is an optimal estimator in the sense that it is they are the *minimum variance, linear, unbiased estimator*. This is no longer true with non-constant variance. **Second, and more importantly**, the formula for the standard error of $\widehat{\beta}$ is wrong. To see this, recall that $\mathbb{V}(AY) = A\mathbb{V}(Y)A^T$. Hence,

$$\mathbb{V}(\widehat{\beta}) = \mathbb{V}((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y) = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{V}(Y)\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\Sigma\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}$$

which is different than the usual formula.

It can be shown that minimum variance, linear, unbiased estimator is obtained by minimizing

$$\text{RSS}(\beta) = (Y - \mathbb{X}\beta)^T\Sigma^{-1}(Y - \mathbb{X}\beta).$$

The solution is

$$\widehat{\beta} = SY \tag{12}$$

where

$$S = (\mathbb{X}^T \Sigma^{-1} \mathbb{X})^{-1} \mathbb{X}^T \Sigma^{-1}. \tag{13}$$

This is unbiased with variance

$$\mathbb{V}(\widehat{\beta}) = (\mathbb{X}^T \Sigma^{-1} \mathbb{X})^{-1}.$$

This is called **weighted least squares.**

Here is the derivation. Let B denote the square root of Σ. Thus, B is a symmetric matrix that satisfies

$$B^T B = B B^T = \Sigma.$$

It can be shown that B^{-1} is the square root of Σ^{-1}. Let $Z = B^{-1}Y$. Then we have

$$\begin{aligned} Z &= B^{-1}Y = B^{-1}(\mathbb{X}\beta + \epsilon) \\ &= B^{-1}\mathbb{X}\beta + B^{-1}\epsilon \\ &= M\beta + \delta \end{aligned}$$

where

$$M = B^{-1}\mathbb{X}, \quad \text{and, } \delta = B^{-1}\epsilon.$$

Moreover,

$$\mathbb{V}(\delta) = B^{-1}V(\epsilon)B^{-1} = B^{-1}\Sigma B^{-1} = B^{-1}BBB^{-1} = I.$$

Thus we can simply regress Z on M and do ordinary regression.

Let us look more closely at a special case. If the residuals are uncorrelated then

$$\Sigma = \begin{pmatrix} \frac{\sigma^2}{w_1} & 0 & 0 & \dots & 0 \\ 0 & \frac{\sigma^2}{w_2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & 0 & 0 & \frac{\sigma^2}{w_n} \end{pmatrix}.$$

In this case,

$$\text{RSS}(\beta) = (Y - \mathbb{X}\beta)^T \Sigma^{-1} (Y - \mathbb{X}\beta) \propto \sum_{i=1}^{n} w_i (Y_i - x_i^T \beta)^2.$$

Thus, in weighted least squares we are simply giving lower weight to the more variable (less precise) observations.

Now we have to address the following question: where do we get the weights? Or equivalently, how do we estimate $\sigma_i^2 = \mathbb{V}(\epsilon_i)$? There are four approaches.

(1) Do a transformation to make the variances approximately equal. Then we don't need to do a weighted regression.

(2) Use external information. There are some cases where other information (besides the current data) will allow you to know (or estimate) σ_i. These cases are rare but they do occur. For example σ_i^2 could be the measurement error of the instrument.

(3) Use replications. If there are several Y values corresponding to each x value, we can use the sample variance of those Y values to estimate σ_i^2. However, it is rare that you would have so many replications.

(4) Estimate $\sigma(x)$ as a function of x. Just as we can estimate the regression line, we can also estimate the variance, thinking of it as a function of x. We could assume a simple model like

$$\sigma(x_i) = \alpha_0 + \alpha_1 x_i$$

for example. Then we could try to find a way to estimate the parameters α_0 and α_1 from the data. In fact, we will do something more ambitious. We will estimate $\sigma(x)$ assuming only that it is a smooth function of x. We will do this later in the course when we discuss nonparametric regression.

In R we simply include weights in the lm command:
lm($Y \sim X$, weights= $1/\text{StdDev}^2$), where StdDev^2 is simply an estimate of $\mathbb{V}[Y|X]$.