LECTURE NOTES, PART 2

1.1 The Classical Linear Regression model

The variable in question (called the dependent variable, the response, the regressand) is related to several other variables (called the regressors or the explanatory variables or independent variables). Suppose we observe n values for these variables. Let y_i be the i^{th} observation of the dependent variable in question and let $(X_{i1}, X_{i2}, \dots, X_{ik})$ be the ith observation of the k regressors. The sample or data is a collection of these n observations.

Both the dependent and independent variables will be treated as random variables (which is unlike in many books and Section II. of Lecture Notes 1, where independent variables are considered as fixed). The classical regression model is a set of joint distributions satisfying the following FOUR assumptions:

Assumption 1. (LINEARITY) $\boxed{y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i \quad i=1,2,\dots,n}$

where β's are unknown parameters to be estimated, and ε_i is the unobserved error term with certain properties to be specified later.

β_i's are regression coefficients

$\sum_{j=1}^{k} \beta_j X_{ij}$ is called the regression function. $\quad \frac{\partial y_i}{\partial x_{ij}} = \beta_j$

Note: Linearity implies that the marginal effect of a regressor does not depend on the level of regressors. The error term represents the part of the dependent variable left unexplained by the regressors.

ex. $con_i = \beta_1 + \beta_2 YD_i + \varepsilon_i$ (simple regression model)

- con_i: consumption of i^{th} household
- YD_i: disposable income of i^{th} household

$\log(wage_i) = \beta_1 + \beta_2 S_i + \beta_3 tenure_i + \beta_4 expr_i + \varepsilon_i$

- $wage$
- S_i: education in yrs
- $tenure_i$: yrs in current position
- $expr_i$: experience in the labor force

Matrix notation

$$y_i = x_i'\beta + \varepsilon_i \quad i=1,2,\dots,n \quad \text{where} \quad x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{pmatrix}_{k\times 1}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}_{k\times 1}$$

If one lets: $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}$

$$X = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}_{n\times k}$$

, then the classical regression model in matrix notation is:

$$\boxed{y = X\beta + \varepsilon}$$

Assumption 2 (STRICT EXOGENEITY) $\boxed{E[\varepsilon_i | X] = 0 \quad i=1,2,\dots,n}$

Note: What exactly does this mean? For any given observation i, take the joint distribution of $nk+1$ random variables $f(\varepsilon_i, x_1, x_2, \dots, x_n)$ and consider the conditional distribution $f(\varepsilon_i | x_1, \dots, x_n)$. The conditional mean $E[\varepsilon_i | x_1, \dots, x_n]$, in general, is a nonlinear function of $(x_1, \dots, x_n)$. The strict exogeneity assumption says that this function is a constant of value zero!

Important! Assuming that this constant is zero is not restrictive if the regressors include a constant. Indeed, suppose that $E[\varepsilon_i | X] = \mu$ and $x_{i1} = 1$, $i=1,2,\dots,n$

Then rewrite $y_i = \beta_1 + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$ as

$$y_i = (\beta_1 + \mu) + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + (\varepsilon_i - \mu)$$ and

redefine β_1 as $\beta_1 + \mu$, and ε_i as $\varepsilon_i - \mu$. Then $E[\varepsilon_i | X] = 0$.

In virtually all applications, the regressors include a constant term!

Implications of Assumption 2:

a) By Law of total expectation $E[E[\varepsilon_i | X]] = E[\varepsilon_i]$, hence $\boxed{E[\varepsilon_i] = 0}$, i.e. the unconditional mean of the error term is 0.

b) Regressors are orthogonal to the error term for all observations, i.e.

$\boxed{E[x_{jl}\varepsilon_i] = 0}$ for all $j=1,2,\dots,n$; $i=1,2,\dots,n$; $l=1,2,\dots,k$.

proof of b): By the Law of iterated expectations:

$$E[\varepsilon_i | x_{je}] \overset{\text{Law of iterated expectations}}{=} E[\underbrace{E[\varepsilon_i | X]}_{=0 \text{ by strict exogeneity}} | x_{je}] = 0$$

Then $E[x_{je}\varepsilon_i] \overset{\text{Law of total expectations}}{=} E[E[x_{je}\varepsilon_i | x_{je}]] = E[x_{je} E[\varepsilon_i | x_{je}]] = 0.$

c) $Cov(\varepsilon_i, x_{je}) = E[x_{je}\varepsilon_i] - E[x_{je}]E[\varepsilon_i] \overset{a)}{=} E[x_{je}\varepsilon_i] \overset{b)}{=} 0.$

"regressors are uncorrelated with the error term".

If i is time, then Assumption 2 states that the regressors are orthogonal to the past, current and the future error terms, which is clearly not true in time series models.

ex. AR(1) $y_i = \beta y_{i-1} + \varepsilon_i \quad i = 1, 2, \ldots, n.$

Suppose, in the spirit of strict exogeneity assumption, that $E[y_{i-1}\varepsilon_i] = 0$. Then, $E[y_i\varepsilon_i] = E[(\beta y_{i-1} + \varepsilon_i)\varepsilon_i] = \beta E[y_{i-1}\varepsilon_i] + E[\varepsilon_i^2] = E[\varepsilon_i^2]$ which is not zero unless the error term is always zero.

Regressors in time series models are not orthogonal to the past error term

Assumption 3 (NO MULTICOLLINEARITY) $\boxed{\text{rank}(X) = k \text{ with probability } 1}$

Assumption 4 (SPHERICAL ERROR VARIANCE)

$E[\varepsilon_i^2 | X] = \sigma^2 > 0 \quad i = 1, 2, \ldots, n$ (homoskedasticity)

$E[\varepsilon_i \varepsilon_j | X] = 0 \quad i, j = 1, 2, \ldots, n; i \neq j$ (no correlation between observations)

Note: rank of a matrix is the # of linearly independent columns.

Assumption 3 implies that $n \geq k$, i.e. # of observations $\geq$ # of regressors

Assumption 4 says that the conditional second moment, which is in general a nonlinear function of X, is a constant σ^2.

Assumption 4 is equivalent to $Var(\varepsilon_i|X)=\sigma^2$ ($i=1,2,\ldots,n$) and $Cov(\varepsilon_i,\varepsilon_j|X)=0$ ($i,j=1,\ldots,n$; $i\neq j$)

(due to Assumption 2.)

Assumption 4 can be written as $\boxed{E[\varepsilon\varepsilon'|X]=\sigma^2 I_{n\times n}}$ ← identity $n\times n$ matrix

or as $Var(\varepsilon|X)=\sigma^2 I_{n\times n}$.

Note: If (y, X) is a random sample, then $\{y_i, x_i\}$ are i.i.d. across $i=1,2,\ldots,n$

Then Assumption 2. becomes $E[\varepsilon_i|x_i]=0$; while Assumption 4. becomes $E[\varepsilon_i^2|x_i]=\sigma^2>0$.

1.2 OLS

The residual for observation i is $y_i - x_i'\tilde{\beta}$, where $\tilde{\beta}$ is some hypothetical value of the regression coefficient vector β.

The sum of squared residuals SSR is: $\boxed{SSR(\tilde{\beta})=\sum_{i=1}^{n}(y_i-x_i'\tilde{\beta})^2}$, or, equivalently,

$\boxed{SSR(\tilde{\beta})=(y-X\tilde{\beta})'(y-X\tilde{\beta})}$

The OLS estimate b, of β, is the value $\tilde{\beta}$ that minimizes $SSR(\tilde{\beta})$.

Note that b will depend on the data y, X, and is in general, different from the true value β.

By having squared residuals in the objective function, the OLS method imposes heavy penalty on large residuals. So, it prevents large residuals for a few observations at the expense of tolerating relatively small residuals for many other observations

$$SSR(\tilde{\beta})=(y-X\tilde{\beta})'(y-X\tilde{\beta})=(y'-\tilde{\beta}'X')(y-X\tilde{\beta})=y'y-\tilde{\beta}'X'y-y'X\tilde{\beta}+\tilde{\beta}'X'X\tilde{\beta}$$

$$=y'y-2y'X\tilde{\beta}+\tilde{\beta}'X'X\tilde{\beta}=y'y-2a'\tilde{\beta}+\tilde{\beta}'A\tilde{\beta}, \text{ where}$$

(since $y'X\tilde{\beta}$ is just a scalar)

$$a=X'y \text{ and } A=X'X.$$

$\nabla SSR(\tilde{\beta}) = 0$ (set the gradient equal to zero!)

$\Downarrow$

$-2a + 2A\tilde{\beta} = 0$ (since $\frac{\partial(a'\tilde{\beta})}{\partial\tilde{\beta}} = a$ and $\frac{\partial(\tilde{\beta}'A\tilde{\beta})}{\partial\tilde{\beta}} = 2A\tilde{\beta}$ for symmetric matrix A — check this!)

$\Downarrow$

$a = A\tilde{\beta} \Rightarrow \boxed{\underbrace{X'X}_{k\times k} b = X'y}$ ($k\times 1$)

"first-order conditions" or so-called "normal equations"

The vector of residuals evaluated at $\tilde{\beta} = b$,

$\boxed{e_{n\times 1} := y - Xb}$ is called the vector of OLS residuals

Its i^{th} element is $e_i = y_i - x_i'b$.

Note: Rearranging the normal equations $X'Xb = X'y$ as $X'(y - Xb) = 0$ or as $X'e = 0$ we get that

$$\left.\begin{aligned} x_{11}e_1 + x_{21}e_2 + \ldots + x_{n1}e_n &= 0 \\ x_{12}e_1 + x_{22}e_2 + \ldots + x_{n2}e_2 &= 0 \\ &\vdots \\ x_{1k}e_1 + x_{2k}e_2 + \ldots + x_{nk}e_n &= 0 \end{aligned}\right\} + \Rightarrow$$

$$\Rightarrow \boxed{\sum_{i=1}^{n} x_i e_i = 0} \quad \text{or} \quad \frac{1}{n}\sum_{i=1}^{n} x_i e_i = 0$$

which shows that the normal equations can be interpreted as the sample analogue of the orthogonality conditions $E[x_i\varepsilon_i] = 0$.

Note: $\nabla SSR(\tilde{\beta}) = 0 \Leftrightarrow X'Xb = X'y$ are just a necessary condition for minimization. We have to check the second-order condition to make sure that b achieves the minimum, not the maximum. However, here it is clear it is a minimum, since $X'X$ is positive definite.

So, the normal equations $X'Xb = X'y$ is a system of k linear equations in k unknowns in $b = \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix}$. The solution can be written as:

$$\boxed{b = (X'X)^{-1}X'y}$$ the OLS estimator

Note: Since $(X'X)^{-1}X'y = (X'X/n)^{-1}X'y/n$, we can also rewrite this as:

$$b = S_{xx}^{-1} s_{xy}, \text{ where } S_{xx} = \frac{1}{n}X'X = \frac{1}{n}\sum_{i=1}^{n} x_i x_i'$$

and $$s_{xy} = \frac{1}{n}X'y = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

i.e. S_{xx} can be regarded as the sample average of $x_i x_i'$,

s_{xy} ——————— ‖ ——————— of $x_i y_i$.

More regression "lingo"

Def. The fitted value for observation i is defined as $\hat{y}_i := x_i'b$

The vector of fitted values is $\hat{y} = Xb$, so the vector of OLS residuals can be written as $e = y - \hat{y}$.

Def. The projection matrix P and the annihilator M are:

$$P_{n\times n} = X(X'X)^{-1}X', \quad M_{n\times n} = I_{n\times n} - P$$

properties: a) both P and M are symmetric and idempotent (i.e. $P^2 = P, M^2 = M$)

b) $PX = X$ (hence the "projection")

$MX = 0$ (hence the "annihilator")

c) $SSR(b) = (y-Xb)'(y-Xb) \overset{\text{by def. of } e}{=} e'e = \varepsilon' M \varepsilon$ (your homework)

One can think of the OLS method geometrically in terms of the projection $\hat{y}$ of y into the k-dimensional linear space $\mathcal{X}$ spanned by the columns of the data matrix X. Since $\hat{y}$ is orthogonal to $y - \hat{y}$, we have: $\|y\|^2 = \|\hat{y}\|^2 + \|y - \hat{y}\|^2$

Note that $\|y-\hat{y}\|^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = SSR$ and $\hat{y} = Xb = X(X'X)^{-1}X'y$, i.e. $\hat{y} = Hy$ where H is the projection matrix defined above.

Now, $M = I-H$ is clearly associated with projection onto the linear space $\mathcal{L}^{\perp}$ consisting of vectors $u \in \mathbb{R}^n$ that are orthogonal to the column vectors of X.

Computational remark: For a nonnegative definite matrix V, there exists an orthogonal matrix Q such that $V = QDQ'$ where D is the diagonal matrix whose elements are eigenvalues of V. This is known as the singular-value decomposition of V, which is then used to compute the inverse $V^{-1} = QD^{-1}Q'$. Note here that if $D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$, then $D^{-1} = \begin{pmatrix} 1/\lambda_1 & & 0 \\ & \ddots & \\ 0 & & 1/\lambda_n \end{pmatrix}$

So, for $V \equiv X'X$, one can use the singular-value decomposition to compute the OLS estimator $b = (X'X)^{-1}X'y$.

However, it is often more convenient to use the QR decomposition $X = QR$ where Q is an $n \times n$ orthogonal matrix and R is an $n \times k$ matrix with zero elements apart from the first k rows, which form a $k \times k$ upper-triangular matrix

$$R = \begin{pmatrix} R_1 \\ 0_{(n-k)\times k} \end{pmatrix} \qquad Q = (Q_1 \; Q_2), \text{ where}$$

R_1 is $k \times k$ upper triangular matrix and Q_1 consists of the first k columns of Q.

So, $X = QR = Q_1R_1$. Note that $X'X = R'Q'QR = R'R = R_1'R_1$ and

$$X'y = R_1'Q_1'y$$

So, the normal equations $(X'X)b = X'y$ can be written as:

$$R_1'R_1 b = R_1'Q_1'y, \text{ i.e. as } \boxed{R_1 b = Q_1'y}$$

which can now be solved by back-substitution (starting with the last row of R_1) since R_1 is upper triangular.

Def. The OLS estimate of σ^2 (the variance of the error term) is denoted by s^2, and is defined as: $\boxed{s^2 = \frac{SSR}{n-k} = \frac{e'e}{n-k}}$

Note: We divide by $n-k$ rather than by n (the sample size) in order to make s^2 an unbiased estimate for σ^2 (as is shown later).

Intuitive reason is that k parameters β have to be estimated before obtaining the residual vector e used to calculate s^2. More precisely, e has to satisfy the k normal equations $X'e = 0$ from page 5, which limits the variability of the residual.

Def. $s = \sqrt{s^2}$ is called the standard error of the regression (SER).

It's an estimate of the standard deviation of the error term.

Def. The sampling error is $b-\beta$. It will be useful later on to relate it to ε, as follows:

$$b-\beta = (X'X)^{-1}X'y - \beta = (X'X)^{-1}X'(X\beta+\varepsilon) - \beta = (X'X)^{-1}(X'X)\beta + (X'X)^{-1}X'\varepsilon - \beta$$

i.e. $\boxed{b-\beta = (X'X)^{-1}X'\varepsilon}$

Def Uncentered R^2

One measure of the variability of the dependent variable is the sum of squares $\sum_{i=1}^{n} y_i^2 = y'y$. We have the following decomposition of that "variability":

$$y'y = (\hat{y}+e)'(\hat{y}+e) = \hat{y}'y + 2\hat{y}'e + e'e = \hat{y}'\hat{y} + 2b'X'e + e'e. \text{ So.}$$

($y = \hat{y}+e$ since $e = y-\hat{y}$; $\hat{y} = Xb$; $X'e = 0$ (see page 5.))

$$\boxed{y'y = \hat{y}'\hat{y} + e'e}$$

The uncentered R^2 is $\boxed{R^2_{uc} = 1 - \frac{e'e}{y'y} = \frac{\hat{y}'\hat{y}}{y'y}}$

Since $\hat{y}'\hat{y}$ and $e'e$ are nonnegative, $0 \le R^2_{uc} \le 1$

R^2_{uc} is the fraction of the variability of the dependent variable that is attributable to the variation in the explanatory variables. The closer the fitted values track the dependent variable, the closer is the uncentered R^2 to one.

Def. Centered R^2, also known as "the coefficient of determination")

If the only regressor is a constant, i.e. $k=1$ and $X_{i1}=1$, $i=1,2,\dots,n$, then from $b=(X'X)^{-1}X'y$ it easily follows that $b=\bar{y}$, the sample mean of the dependent variable, which means that $\hat{y}_i=\bar{y}$, for all $i=1,2,\dots,n$; $\hat{y}'\hat{y}=n(\bar{y})^2$ and

$$e'e=\sum_{i=1}^{n}(y_i-\bar{y})^2$$

In general, if the regressors also include nonconstant variables, it can be shown (homework) that problem 10

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}e_i^2\,,\quad \circledast$$

$$\text{i.e. } (y-\bar{y})'(y-\bar{y})=(\hat{y}-\bar{y})'(\hat{y}-\bar{y})+e'e.$$

The Centered R^2 is

$$R^2:=1-\frac{e'e}{(y-\bar{y})'(y-\bar{y})}=1-\frac{\sum_{i=1}^{n}e_i^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

or equivalently

$$\boxed{R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{(\hat{y}-\bar{y})'(\hat{y}-\bar{y})}{(y-\bar{y})'(y-\bar{y})}}$$

Provided that the regressors include a constant so that $\circledast$ is valid, we have $0\le R^2\le 1$. R^2 is hence a measure of the explanatory power of the nonconstant regressors.

Caveat: If regressors do not include a constant (rare in practice), R^2 could turn out to be negative, since without the benefit of the intercept, the regression could do worse than the sample mean in terms of tracking the dependent variable.

Note: $\circledast$ is sometimes stated as: "total sum of squares" $\sum_{i=1}^{n}(y_i-\bar{y})^2$ = "regression sum of squares" $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ + "residual error sum of squares" $\sum_{i=1}^{n}e_i^2$ (see HW#2 Problem 10)

A practical advice

Since the method of least squares seeks to prevent a few large residuals at the expense of incurring many relatively small ones, only a few observations can be extremely influential in the sense that dropping them from the sample changes some elements of b substantially. There is a systematic way to find the influential observations.

Let $b_{(-i)}$ be the OLS estimator of β that would be obtained if OLS were used on a sample from which the i^{th} observation was omitted. Then it can be shown (not easy! See homework problem 11) that

$$\boxed{b_{(-i)} - b = -\frac{(X'X)^{-1}x_i}{1-h_{ii}} \cdot e_i}$$, where x_i is the i^{th} row of X as before, $e_i = y_i - x_i'b$ and h_{ii} is defined as $h_{ii} = x_i'(X'X)^{-1}x_i$, i.e. h_{ii} is the i^{th} diagonal element of the projection matrix P. It can be shown that $0 \le h_{ii} \le 1$ and $\sum_{i=1}^{n} h_{ii} = k$, so that h_{ii}, on average, is k/n.

Note: $tr(P) = k$ as shown later on page 13

Observations i with h_{ii} being well above the average k/n are influential, and probably should be dropped from the sample, as they represent "outliers".

1.3 Statistical properties of OLS estimators

b is linear in y since $b = (X'X)^{-1}X'y$

Theorem 1 Under our assumptions, we have:

A) unbiasedness: $E[b|X] = \beta$

B) $Var(b|X) = \sigma^2(X'X)^{-1}$

C) (Gauss-Markov) The OLS estimator is efficient in the class of all linear unbiased estimators; that is, for any other unbiased estimator $\hat{\beta}$ linear in y, we have $Var(\hat{\beta}|X) \ge Var(b|X)$

Note: This inequality is in the matrix sense, as both $Var(\hat{\beta}|X)$ and $Var(b|X)$ are clearly $k \times k$ matrices! What we mean by this is that the matrix $Var(\hat{\beta}|X) - Var(b|X)$ is positive semidefinite.

Note: Another way to state C) is to say that the OLS estimator is BLUE, i.e. the best linear unbiased estimator

D) $Cov(b, e|X) = 0$ (Homework: problem 7)

Notes: - Take a to be the k-dimensional vector whose all entries are 0, except for the ℓ^{th} entry which is 1, $1 \le \ell \le k$. Then, by C), we have

$$a'(Var(\hat{\beta}|X) - Var(b|X))a \ge 0 \quad , \text{ i.e.}$$

$$Var(\hat{\beta}_\ell|X) \ge Var(b_\ell|X) \qquad \ell = 1, 2, \ldots, k$$

So, for any regression coefficient, the variance of the OLS estimator is no larger than that of any other linear unbiased estimator.

- What is the true meaning of (A)?

$b = (X'X)^{-1}X'y$. Since y, X are random, then so is b. Imagine if you fix X at some given value, and then calculate b for all samples corresponding to all possible realizations of y, and then take the average of b. This average is $E[b|X]$. (A) states that this average equals the true value β

(A) also implies a weaker statement $E[b] = \beta$, since by Law of total expectations $E[E[b|X]] = E[b]$. In other words, if one calculates b for all possible different samples, differing not only in y but also in X, the average would be the true value β.

Similarly, one can show that (C) implies the unconditional statement $Var(\hat{\beta}) \ge Var(b)$ as well (where $\hat{\beta}$ is any other linear unbiased estimator)

proof of Theorem:

(A) We know from pages that the sampling error $b - \beta = A\varepsilon$ where $A = (X'X)^{-1}X'$

So, $E[b - \beta|X] = E[A\varepsilon|X] = AE[\varepsilon|X] = 0$.

(A as a function of X can be treated as nonrandom so linearity of expectation applies)

So, $E[b|X] = \beta$

Note: Anything short of strict exogeneity $E[\varepsilon|X] = 0$ assumption will not give unbiasedness!

(B) $\text{Var}(b|X) = \text{Var}(b-\beta|X) = \text{Var}(A\varepsilon|X) = A\,\text{Var}(\varepsilon|X)A' = A\,E[\varepsilon\varepsilon'|X]A' =$

(β not random; $A=(X'X)^{-1}X'$ is a function of X; $E[\varepsilon|X]=0$)

$= A(\sigma^2 I_{n\times n})A' = \sigma^2 AA' = \sigma^2 (X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}$

(Assumption 4)

(C) Let $\hat{\beta}$ be any unbiased estimator of β, that is linear in y. Let $\hat{\beta} = Cy$ for some matrix C, which possibly is a function of X. Let $D = C - A \Leftrightarrow C = D + A$, where as before $A = (X'X)^{-1}X'$. Then

$\hat{\beta} = (D+A)y = Dy + Ay = D(X\beta+\varepsilon) + b = DX\beta + D\varepsilon + b$ ⊛, so:

taking conditional expectations of both sides:

$E[\hat{\beta}|X] = DX\beta + E[D\varepsilon|X] + E[b|X]$ ⊕

Since both b and $\hat{\beta}$ are unbiased: $E[b|X] = E[\hat{\beta}|X] = 0$. Furthermore, $E[D\varepsilon|X] = DE[\varepsilon|X] = 0$, since D is a function of X and can be treated as nonrandom. So, from ⊕, we get: $DX\beta = 0$. Since this has to be true for any given β, it is necessary that $DX = 0$. So, from ⊛, we get $\hat{\beta} = D\varepsilon + b$, i.e.

$\hat{\beta} - \beta = D\varepsilon + (b-\beta) = D\varepsilon + A\varepsilon = (D+A)\varepsilon$.

So, $\text{Var}(\hat{\beta}|X) = \text{Var}(\hat{\beta}-\beta|X) = \text{Var}((D+A)\varepsilon|X) = (D+A)\,\text{Var}(\varepsilon|X)(D+A)' =$

(D & A both functions of X)

$= \sigma^2(D+A)(D+A)' =$

($\text{Var}(\varepsilon|X) = \sigma^2 I_{n\times n}$)

$= \sigma^2(DD' + AD' + DA' + AA') = \sigma^2(DD' + AA') =$

($DA' = DX(X'X)^{-1} = 0$ since $DX = 0$)

($AA' = (X'X)^{-1}$)

$= \sigma^2(DD' + (X'X)^{-1}) \geq \sigma^2(X'X)^{-1} = \text{Var}(b|X)$

(since DD' is positive semidefinite)

Theorem 2 $E[s^2|X] = \sigma^2$ (and hence $E[s^2] = \sigma^2$). In other words, the OLS estimator $s^2 = \frac{e'e}{n-k}$ of σ^2 is also unbiased.

(see page 8.)

proof: We need to show $E[e'e|X] = (n-k)\sigma^2$

You will show in homework problems that $e'e = \varepsilon' M \varepsilon$, where M is the annihilator matrix. Two steps in our proof: 1) $E[\varepsilon' M \varepsilon | X] = \sigma^2 tr(M)$; 2) $tr(M) = n-k$

(trace of a matrix)

Step 1 $\varepsilon' M \varepsilon = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}\varepsilon_i\varepsilon_j \Rightarrow$

$$\Rightarrow E[e'e|X] = E[\varepsilon' M \varepsilon | X] = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij} E[\varepsilon_i\varepsilon_j|X] =$$

(linearity; m_{ij}'s are functions of X)

(Assumption 4: $E[\varepsilon_i^2|X] = \sigma^2$, $E[\varepsilon_i\varepsilon_j|X] = 0$)

$$= \sum_{i=1}^{n} m_{ii}\sigma^2 = \sigma^2 \sum_{i=1}^{n} m_{ii} = \sigma^2 \cdot tr(M)$$

Step 2 $tr(M) = tr(I_{n\times n} - P) = tr(I_{n\times n}) - tr(P) = n - tr(P) = n-k$

$tr(P) = tr\left(X(X'X)^{-1}X'\right) = tr\left((X'X)^{-1}X'X\right) = tr(I_{k\times k}) = k$

($tr(CD) = tr(DC)$)

So, $E[s^2|X] = E\left[\frac{e'e}{n-k}|X\right] = \frac{1}{n-k}E[e'e|X] = \frac{1}{n-k}E[\varepsilon' M \varepsilon|X] =$

$= \frac{1}{n-k}\cdot\sigma^2 tr(M) = \frac{1}{n-k}\cdot\sigma^2\cdot(n-k) = \sigma^2$ □

Note A natural estimate of $Var(b|X) = \sigma^2(X'X)^{-1}$ is

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

this is one of the statistics included in any regression software package/procedure.

1.4 Hypothesis testing under normality

Very often, in practice, we want to test the "theory" that a regression coefficient is equal to some specific value. Although Theorem 1 from 1.3. guarantees that on average the OLS estimator of that coefficient should be equal to that specific value, the OLS estimator may not exactly be equal to that value for a particular sample at hand! Clearly, we should not conclude that our "theory" is false just because of that. In order for us to decide whether the sampling error is "too large" for the "theory" to be true, we need to construct from the sampling error some test statistic whose probability distribution is known given the truth of our "theory".

At first, it appears that doing so requires one to specify the joint distribution of (X, ε), because after all, the sampling error $b-\beta = (X'X)^{-1}X'\varepsilon$ is a function of (X, ε). As it turns out (surprising!!!), the distribution of an appropriate test statistic can be derived from the following intuitive assumption, that we now add to the list of regression model assumptions:

Assumption 5. The distribution of ε conditional on X is jointly normal.

Together with Assumptions 2 & 4, this implies $\boxed{\varepsilon | X \sim N(0, \sigma^2 I_{n\times n})}$

Here, clearly, 0 denotes the $n \times 1$ zero vector.

Several observations are in order. Normal distribution is specified by its mean and its variance. Note that the mean and the variance of ε CONDITIONAL ON X do not depend on X. Hence, the marginal (i.e. unconditional) distribution of ε is ALSO $N(0, \sigma^2 I_{n\times n})$. ε and X are independent according to this assumption as well. Theorem 1. from Section 1.3 now easily yields:

Claim Since $\varepsilon | X$ is normal, then $b-\beta | X$ is also normal (recall $b-\beta = A\varepsilon$, where $A = (X'X)^{-1}X'$). Moreover, $\boxed{(b-\beta) | X \sim N(0_{k\times 1}, \sigma^2 (X'X)^{-1})}$

1.4.1 Testing hypotheses about individual regression coefficients

$H_0 : \beta_\ell = \beta_\ell^0$ for some $\ell \in \{1, 2, \ldots, k\}$

where $\beta_\ell^{(0)}$ is some known value specified by the null hypothesis.
We want to test H_0 against the alternative hypothesis $H_1 : \beta_\ell \neq \beta_\ell^0$ at a significance level α. Looking at just the ℓ^{th} component of

$(b-\beta) | X \sim N(0_{k\times 1}, \sigma^2 (X'X)^{-1})$, we obtain that under the null hypothesis H_0: $(b_\ell - \beta_\ell^0) | X \sim N(0, \sigma^2 ((X'X)^{-1})_{\ell\ell})$ where

$((X'X)^{-1})_{\ell\ell}$ is the (ℓ,ℓ)-element of $k\times k$ matrix $(X'X)^{-1}$!

Consider the following statistic: $z_\ell := \dfrac{b_\ell - \beta_\ell^0}{\sqrt{\sigma^2 \cdot ((X'X)^{-1})_{\ell\ell}}}$

Then $z_\ell | X \sim N(0,1)$

For a second, suppose that σ^2 is known. Then z_ℓ has some nice properties as a test statistic:

1) it can be calculated from the sample
2) its distribution conditional on X does not depend on X
(which shall not be confused with the fact that the value of z_ℓ depends on X).
So, z_ℓ and X are independently distributed, and regardless of the value of X, the distribution of z_ℓ is the same as its unconditional distribution. This is convenient, because different samples differ not only in y, but also in X.
3) The distribution is known.

So, apart from σ^2 being the nuisance parameter, using this statistic, we can determine whether or not the sampling error $b_\ell - \beta_\ell^0$ is too large: it is too large if the test statistic takes on a value that is surprising for a realization from its distribution.
A natural idea is to replace σ^2 by its OLS estimator s^2.

Theorem 1 Suppose Assumptions 1–5 hold. Under the null hypothesis $H_0: \beta_\ell = \beta_\ell^0$, the t-ratio defined as:

$$t_\ell := \frac{b_\ell - \beta_\ell^0}{\sqrt{s^2 \cdot (X'X)^{-1}_{\ell\ell}}}$$ is distributed as t_{n-k} → t-distribution with $n-k$ degrees of freedom

Note: σ^2 in z_ℓ is replaced by s^2 in t_ℓ.

The denominator $\sqrt{s^2 \cdot (X'X)^{-1}_{\ell\ell}}$ of t_ℓ is called the standard error $SE(b_\ell)$ of the OLS estimate b_ℓ of β_ℓ.

proof:
$$t_\ell = \frac{b_\ell - \beta_\ell^0}{\sqrt{\sigma^2 \cdot (X'X)^{-1}_{\ell\ell}}} \cdot \sqrt{\frac{\sigma^2}{s^2}} = \frac{z_\ell}{\sqrt{s^2/\sigma^2}} = \frac{z_\ell}{\sqrt{\frac{e'e}{(n-k)\sigma^2}}} = \frac{z_\ell}{\sqrt{\frac{q}{n-k}}},$$

where $q := \frac{e'e}{\sigma^2}$. We have already shown that $z_\ell \sim N(0,1)$.

Next we show that

Ⓐ $q \mid X \sim \chi^2_{n-k}$

Ⓑ two random variables q and z_ℓ are independent conditional on X

Clearly, Ⓐ and Ⓑ yield $t_\ell \sim t_{n-k}$.

proof of Ⓐ: $q = \frac{e'e}{\sigma^2} = \frac{\varepsilon' M \varepsilon}{\sigma^2} = \frac{\varepsilon'}{\sigma} M \frac{\varepsilon}{\sigma}$.

↑ (homework problem 6)

FACT If $a \sim N(0_{m\times 1}, I_{m\times m})$ and A is an idempotent $m \times m$ matrix, then the quadratic form "$a'Aa$" has a χ^2-distribution with # of degrees of freedom $= \text{rank}(A)$

Since $\varepsilon \mid X \sim N(0_{n\times 1}, \sigma^2 I_{n\times n})$ by Assumption 5, then $\frac{\varepsilon}{\sigma} \mid X \sim N(0_{n\times 1}, I_{n\times n})$

Now, by FACT, $\frac{\varepsilon'}{\sigma} M \frac{\varepsilon}{\sigma} \mid X \sim \chi^2$ where # of degrees of freedom is $\text{rank}(M)$. But $\text{rank}(M) = \text{tr}(M)$, since M is idempotent. In Step 2 a page 13, we

proved that $tr(M) = n-k$. So, $q \mid X \sim \chi^2_{n-k}$.

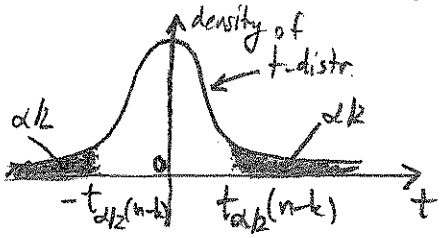
proof of (B): Since $b - \beta = (X'X)^{-1}X'\varepsilon$ (page 8) and $e = M\varepsilon$ (homework problem 6), we see that both b and e are linear functions of ε, so they are jointly normal conditional on X. Homework problem 7 ((D) of Theorem 1 in Section 1.3.) asks you to prove $Cov(b, e \mid X) = 0$. So, b and e are uncorrelated conditional on X. So, b and e are independently distributed conditional on X! (by FACT 7., Page 11. of Lecture Notes, part 1)
But z_ℓ is a function of b, while q is a function of e. So, z_ℓ and q are independently distributed conditional on X. □

t-test for $H_0: \beta_\ell = \beta_\ell^0$

Step1. Given the hypothesized value β_ℓ^0 of β_ℓ, calculate the t-ratio $t_\ell = \dfrac{b_\ell - \beta_\ell^0}{\sqrt{s^2((X'X)^{-1})_{\ell\ell}}}$ from the data.

"Too large" a deviation of t_ℓ from 0 is a sign of the failure of H_0.
Next step specifies how large is "too large".

Step2.



Find (the value) $t_{\alpha/2}(n-k)$ s.t.

$$P\left(-t_{\alpha/2}(n-k) < t < t_{\alpha/2}(n-k)\right) = 1-\alpha$$

where $t \sim t_{n-k}$.

Step3 Accept H_0 if $-t_{\alpha/2}(n-k) < t_\ell < t_{\alpha/2}(n-k)$, where t_ℓ is from Step 1.
Reject H_0 otherwise.
Since $t_\ell \sim t_{n-k}$ (under H_0) by Theorem 1, the probability of rejecting H_0 when H_0 is true is α. So, the "size" (or the "significance level") of our t-test is indeed α.

We can also use Theorem 1. to set up confidence intervals for b_ℓ.

Since $\frac{b_\ell - \beta_\ell^\circ}{SE(b_\ell)} \sim t_{n-k}$, then $P\left(-t_{\alpha/2}(n-k) < \frac{b_\ell - \beta_\ell^\circ}{SE(b_\ell)} < t_{\alpha/2}(n-k)\right) = 1-\alpha$

Rearranging the inequalities, we get that the $(1-\alpha)$-confidence interval for β_ℓ is

$$\boxed{\left[b_\ell - SE(b_\ell)\, t_{\alpha/2}(n-k),\ b_\ell + SE(b_\ell)\, t_{\alpha/2}(n-k)\right]}$$

Note: The interval is narrower the smaller the standard error $SE(b_\ell) = \sqrt{s^2 (X'X)^{-1}_{\ell\ell}}$.

p-value Our t-test can also be restated using the p-value:

Step 1 same as before

Step 2 Rather than finding the critical value $t_{\alpha/2}(n-k)$, calculate $p = 2P(t > |t_\ell|)$,
i.e. $p = 1 - P(-|t_\ell| < t < |t_\ell|)$.

Step 3 Accept H_0 if $p > \alpha$. Otherwise, reject.

1.4.2. Testing linear hypotheses

In practice, one often wants to test a hypothesis about several regression coefficients simultaneously, e.g. $H_0: R\beta = r$ where R is an $m \times k$ matrix (and r is an $m \times 1$ vector). To make sure that there are no redundancy among these m equations, we require that $\text{rank}(R) = m$

Example: In $\log(\text{wage}_i) = \beta_1 + \beta_2 S_i + \beta_3 \text{tenure}_i + \beta_4 \text{exp}_i + \varepsilon_i$ regression model from page 1 with $k = 4$, one could for example test the null hypothesis

$H_0: \beta_2 = \beta_3$ and $\beta_4 = 0$

which can be written as $H_0: R\beta = r$, where

$$R = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \text{ and } r = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Note: clearly, rows of R are independent so $\text{rank}(R) = 2$

To test such linear hypotheses, we look for a test statistic that has a known distribution under the null hypothesis.

Theorem 2 Suppose Assumptions 1–5 hold. Under the null hypothesis $H_0: R\beta = r$, where $R_{m \times k}$ with $\text{rank}(R) = m$, the F-ratio defined as:

$$F := \frac{\frac{(Rb-r)'(R(X'X)^{-1}R')^{-1}(Rb-r)}{m}}{s^2}$$

which can be also written as: $F = \frac{(Rb-r)'(R\widehat{Var(b|X)}R')^{-1}(Rb-r)}{m}$, according to the bottom of page 13

is $F_{m,n-k}$ distributed.

proof: As in Theorem 1, we prove that the distribution of F conditional on X is $F_{m,n-k}$. Because the F-distribution does not depend on X, it is also the unconditional distribution of the statistic!

Since $s^2 = \frac{e'e}{n-k}$, we can write $F = \frac{\frac{w}{m}}{\frac{q}{n-k}}$, where

$w := (Rb-r)'(\sigma^2 R(X'X)^{-1}R')^{-1}(Rb-r)$ and $q := \frac{e'e}{\sigma^2}$

We proceed in 3 steps:

Step 1 $w|X \sim \chi^2_m$.

Why? Let $v := Rb - r$. Under H_0, $Rb - r = Rb - R\beta = R(b-\beta)$. Since $(b-\beta)|X \sim N(0_{k\times 1}, \sigma^2(X'X)^{-1})$ (see page 14) we have that v, being a linear transformation of the k-variate normal $b-\beta$, is also normal (conditional on X) with mean $0_{m\times 1}$ and variance:

$Var(v|X) = Var(R(b-\beta)|X) = R\,Var(b-\beta|X)R' = \sigma^2 R(X'X)^{-1}R'$

So, $w = v'Var(v|X)v$. We'll now use the following easy fact:

FACT If $Z \sim N(\mu_{m\times 1}, \Sigma_{m\times m})$, with Σ non-singular, then the quadratic form $(Z-\mu)'\Sigma^{-1}(Z-\mu) \sim \chi^2_m$.

Since $rank(R) = m$ and $X'X$ is non-singular, then $\sigma^2 R(X'X)^{-1}R'$ ($= Var(v|X)$) is also non-singular.

Therefore, $w = v'Var(v|X)v$ and $v|X \sim N(0_{m\times 1}, Var(v|X))$ imply that

$w|X \sim \chi^2_m$

Step 2 See (A) in proof of Theorem 1. $\Rightarrow q|X \sim \chi^2_{n-k}$

Step 3 w is a function of b and q is a function of e. Then, as in part (B) of Theorem 1 proof, we conclude that w and q are independently distributed conditional on X

Now, $w|X \sim \chi^2_m$ and $q|X \sim \chi^2_{n-k}$, w, q independent cond. on X $\Rightarrow$

$$\Rightarrow F = \frac{\frac{w}{m}}{\frac{q}{n-k}} \sim F_{m,n-k} \qquad \square$$

We use the F-statistic to set up the F-test for $H_0: R\beta = r$.

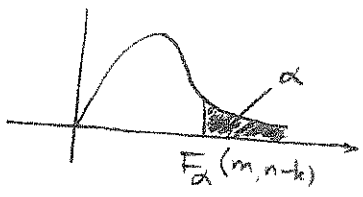
If H_0 is true, we expect $Rb - r$ to be small, so large values of F should be taken as evidence against H_0. We look only at the upper tail of the distribution of the F-statistic

F-test for $H_0: R\beta = r$

Step 1 Calculate $F = \dfrac{(Rb-r)'(R(X'X)^{-1}R')^{-1}(Rb-r)/m}{s^2}$ based on the data.

Step 2 Find the critical value $F_\alpha(m, n-k)$

Step 3 Accept H_0 if the F-ratio from Step 1 is less than $F_\alpha(m, n-k)$. Reject otherwise



One can also restate this test in terms of the p-value, where p in this case is the area of the upper tail of the F-distribution to the right of the F-ratio from Step 1. Accept H_0 if $p >$

Calculating the F-ratio from Theorem 2 requires matrix inversion and multiplication, which is computationally expensive. However, there is an alternative way to arrive at that F-ratio, which is more convenient. Theorem 2 is usually called the "Wald principle", while the method described below is branded as the "likelihood ratio principle".

Regression amounts to the optimization problem: $\min_{\tilde{\beta}} SSR(\tilde{\beta})$. The OLS estimator b of β is the value $\tilde{\beta}$ that achieves the minimum. In what follows, we will call this the unrestricted regression, b will be called the unrestricted OLS estimator of β, and most importantly, $SSR(b)$ will be denoted as $\boxed{SSR_u}$ (where "u" stands for "unrestricted"). If one wants to test $H_0: R\beta = r$, the one could consider the restricted optimization problem: $\boxed{\min_{\tilde{\beta}} SSR(\tilde{\beta}) \text{ so that } R\beta = r}$, which will be branded as the "restricted regression". This is a constrained optimization problem subject to the additional constraint $R\tilde{\beta} = r$ specified by the null hypothesis H_0. As such, it can be solved via Lagrangian

$$L = \frac{1}{2}(y - X\tilde{\beta})'(y - X\tilde{\beta}) + \lambda'(R\tilde{\beta} - r), \text{ where } \lambda \text{ is the } m \times 1 \text{ vector}$$

of Lagrange multipliers. In Homework 2, problem 9, you're asked to solve this optimization problem and find $\hat{\beta}$, the restricted OLS estimator of β, in terms of b.

Let $\boxed{SSR_R} = SSR(\hat{\beta})$, i.e. SSR_R is the minimized sum of squared residuals for the restricted regression. It can be shown (Problem 9c) that the F-ratio from Theorem 2

also equals $\boxed{F = \dfrac{\dfrac{SSR_R - SSR_u}{m}}{\dfrac{SSR_u}{n-k}}}$ ← "difference in the objective function deflated by the estimate of the error variance"

This formula for F-ratio is analogous to how the likelihood-ratio statistic is derived in the MLE as the difference in log likelihood with and without the imposition of the null hypothesis H_0.

Note: One important use of the F-test is for $H_0: \beta_{p+1} = \beta_{p+2} = \ldots = \beta_k = 0$ $(p<k)$ which means that the regression involves only p of the k input variables. In this case, rewrite H_0 as $H_0: R\beta = r$, where $m = k-p$,

$$R = \left(O_{(k-p)\times p} \,\middle|\, I_{(k-p)\times(k-p)} \right) \text{ and } r = O_{(k-p)\times 1}.$$ Then, proceed as above.

(See Homework 2, problem 10. for an application of this to proving the decomposition of $\sum_{i=1}^{n}(y_i - \bar{y})^2$ on page 9.)

t-tests vs F-tests

Because hypotheses about individual coefficients are linear hypotheses, the t-test of $H_0: \beta_\ell = \beta_\ell^0$ is just a special case of the F-test. Indeed, $H_0: \beta_\ell = \beta_\ell^0$ can be rewritten as $R\beta = r$ with $R_{1\times k} = (0\ 0\ldots 0\ \underset{\ell}{1}\ 0\ 0\ldots 0)$ and $r = \beta_\ell^0$ $(m=1)$

Then the F-ratio comes out to be $(b_\ell - \beta_\ell^0)\left(s^2\cdot((X'X)^{-1})_{\ell\ell}\right)^{-1}(b_\ell - \beta_\ell^0)$, which is exactly the square of the t-ratio t_ℓ on page 16. This is a perfect match, since $F_{1,n-k}$ is just the square of t_{n-k}.

Now, let's consider a null hypothesis that states that a set of individual regression coefficients equal certain values. For example, $H_0: \beta_1 = 1$ and $\beta_2 = 0$, and assume $k=2$. This can be written as a linear hypothesis $R\beta = r$ for $R = I_{2\times 2}$ and $r = (1,0)'$. So, one uses the F-test naturally. However, it's tempting to conduct the t-test separately for each individual coefficient of the hypothesis. We might accept H_0 if both $H_0^{(1)}: \beta_1 = 1$ and $H_0^{(2)}: \beta_2 = 0$. This amounts to testing whether $(1,0)$ belongs to the confidence region

$$\left\{(\beta_1,\beta_2) \,\middle|\, b_1 - SE(b_1)t_{\alpha/2}(n-2) < \beta_1 < b_1 + SE(b_1)t_{\alpha/2}(n-2),\ b_2 - SE(b_2)t_{\alpha/2}(n-2) < \beta_2 < b_2 + SE(b_2)t_{\alpha/2}(n-2)\right\}$$

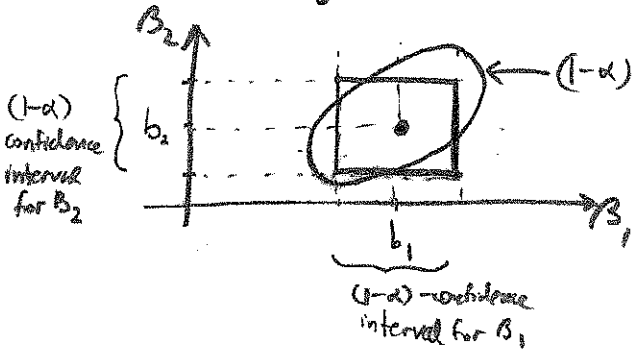
which is a rectangular region in the (β_1, β_2) plane.

On the other hand, the confidence region for the F-test is

$$\left\{(\beta_1,\beta_2)\,\middle|\,(b_1-\beta_1,\ b_2-\beta_2)\left(\widehat{Var(b|X)}\right)^{-1}\begin{pmatrix}b_1-\beta_1\\ b_2-\beta_2\end{pmatrix}<2F_\alpha(2,n-2)\right\}$$

which (since $\widehat{Var(b|X)}$ is positive definite, is an ellipse in the (β_1,β_2)-plane.

Typically, the two regions look like:



The F-test should be preferred to the test using two t-tests for two reasons:

1) If the size (significance level) in each of the two t-tests is α, then the overall size (the probability that $(1,0)$ is outside the rectangular region) is not α!
2) The F-test is a likelihood ratio test, and as such, has desirable properties. (This will be explained in the next Section.)

1.5. Relation to maximum likelihood

Having specified the distribution of the error vector ε, we can use the maximum likelihood principle to estimate the model parameters β, σ^2. We will show that b, the OLS estimator of β, is also the MLE, and the OLS estimator of σ^2 differs only slightly from the MLE.

Assumption 5. yields $\varepsilon|X \sim N(0_{n\times 1}, \sigma^2 I_{n\times n})$. Since $y = X\beta+\varepsilon$ (Assumption 1), we have $\boxed{y|X \sim N(X\beta, \sigma^2 I_n)}$, so the conditional density of y, given X, is

$$f(y|X) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)}$$

Replacing the true parameters (β,σ^2) by their hypothetical values $(\tilde{\beta},\tilde{\sigma}^2)$ and taking logs, we obtain the loglikelihood function (conditional on X):

$$\ell(\tilde{\beta},\tilde{\sigma}^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2}(y-X\tilde{\beta})'(y-X\tilde{\beta}) \qquad \circledast$$

The MLE for (β,σ^2) is the $(\tilde{\beta},\tilde{\sigma}^2)$ that maximizes this loglikelihood.

Much like in Example 2 of Section 2 in Lecture Notes, Part 1, we maximize $\ell(\tilde{\beta},\tilde{\sigma}^2)$ in two stages. First, we maximize over $\tilde{\beta}$ for any given $\tilde{\sigma}^2$. This amounts to minimizing the sum of squares $(y-X\tilde{\beta})'(y-X\tilde{\beta})$. The $\tilde{\beta}$ that does it is none other than the OLS estimator b, and the minimized sum of squares is $e'e$. Plug b in place of $\tilde{\beta}$ in $\circledast$ to get loglikelihood $= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2}e'e$.

This is a function of $\tilde{\sigma}^2 := \tilde{\sigma}^2$ alone. Taking the derivative with respect to $\tilde{\sigma}^2$ and setting it to zero (note that $e'e$ is not a function of $\tilde{\sigma}^2$), we get:

<u>Claim</u> Suppose Assumptions 1–5 hold. Then the MLE for β is the OLS estimator b, and the MLE for σ^2 is $\frac{1}{n}e'e = \frac{SSR}{n} = \frac{n-k}{n}s^2$

<u>Note</u>: We know that s^2 is unbiased ((A) in Theorem 1 on page 10.).
Since $\sigma^2 = \frac{n-k}{n}s^2$, we have that σ^2 is biased, although the bias becomes arbitrarily small as the sample size n increases for any given fixed k.

The maximized loglikelihood is $-\frac{n}{2}\log\left(\frac{2\pi}{n}\right) - \frac{n}{2} - \frac{n}{2}\log(SSR)$

and the maximized likelihood is $\left(\frac{2\pi}{n}\right)^{-\frac{n}{2}} \cdot e^{-\frac{n}{2}} \cdot (SSR)^{-n/2}$. $\qquad \oplus$

Let's calculate the Fisher information matrix for the $(k+1)$-dimensional parameter vector $\theta := (\beta', \sigma^2)'$.

According to <u>page 24 (Lecture Notes Part 1)</u>, we have that the information matrix is the negative of the expected value of the Hessian. (Our notation is a bit off, since now θ_0 on page 24. stands for θ here, and θ on page 24. stands for $\tilde{\theta} = (\tilde{\beta}', \tilde{\sigma}^2)'$ here and $f_\theta(x)$ on page 24. stands for $\ell(\tilde{\theta}) = \ell(\tilde{\beta}, \tilde{\sigma}^2)$ here).

So, $I(\theta) = -E_\theta\left[\left.\frac{\partial^2 \ell(\tilde{\theta})}{\partial\tilde{\theta}\partial\tilde{\theta}'}\right|_{\tilde{\theta}=\theta}\right]$. We need to calculate the $(k+1)\times(k+1)$ matrix

$$\left.\frac{\partial^2 \ell(\tilde{\theta})}{\partial\tilde{\theta}\partial\tilde{\theta}'}\right|_{\tilde{\theta}=\theta} = \begin{pmatrix} \left.\frac{\partial^2 \ell(\tilde{\theta})}{\partial\tilde{\beta}\partial\tilde{\beta}'}\right|_{\tilde{\theta}=\theta} & \left.\frac{\partial^2 \ell(\tilde{\theta})}{\partial\tilde{\beta}\partial\tilde{\gamma}}\right|_{\tilde{\theta}=\theta} \\ \left.\frac{\partial^2 \ell(\tilde{\theta})}{\partial\tilde{\gamma}\partial\tilde{\beta}'}\right|_{\tilde{\theta}=\theta} & \left.\frac{\partial^2 \ell(\tilde{\theta})}{\partial^2\tilde{\gamma}}\right|_{\tilde{\theta}=\theta} \end{pmatrix} \quad (\heartsuit), \text{ where}$$

(as before $\tilde{\gamma} \equiv \tilde{\sigma}^2$ and) $\frac{\partial^2 \ell(\tilde{\theta})}{\partial\tilde{\beta}\partial\tilde{\beta}'}$ is a $k\times k$ matrix, $\frac{\partial^2 \ell(\tilde{\theta})}{\partial\tilde{\gamma}\partial\tilde{\beta}'}$ is a $1\times k$ matrix, $\frac{\partial^2 \ell(\tilde{\theta})}{\partial\tilde{\beta}\partial\tilde{\gamma}}$ is a $k\times 1$ matrix and $\frac{\partial^2 \ell(\tilde{\theta})}{\partial^2\tilde{\gamma}}$ is a 1×1 matrix (i.e. just a number).

Now, it's easy to derive from ⊛ on page 24. $\left(\ell(\tilde{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2}(y-X\tilde{\beta})'(y-X\tilde{\beta})\right)$

$$\frac{\partial \ell(\tilde{\theta})}{\partial\tilde{\beta}} = \frac{\partial}{\partial\tilde{\beta}}\left(-\frac{1}{2\tilde{\gamma}}(y-X\tilde{\beta})'(y-X\tilde{\beta})\right) = -\frac{1}{2\tilde{\gamma}}\cdot\frac{\partial}{\partial\tilde{\beta}}\left(y'y - y'X\tilde{\beta} - \tilde{\beta}'X'y + \tilde{\beta}'X'X\tilde{\beta}\right)$$

recall from lin. alg. that $\frac{\partial(a'\tilde{\beta})}{\partial\tilde{\beta}} = a$ for a vector a and $\frac{\partial(\tilde{\beta}'A\tilde{\beta})}{\partial\tilde{\beta}} = 2A\tilde{\beta}$ for symmetric matrix A

$$= -\frac{1}{2\tilde{\gamma}}\cdot\left(-X'y - X'y + 2X'X\tilde{\beta}\right) = \frac{1}{\tilde{\gamma}}X'(y-X\tilde{\beta})$$

$$\frac{\partial \ell(\tilde{\theta})}{\partial\tilde{\gamma}} = -\frac{n}{2\tilde{\gamma}} + \frac{1}{2\tilde{\gamma}^2}(y-X\tilde{\beta})'(y-X\tilde{\beta})$$

So, the second derivatives of $\ell(\tilde{\theta})$ evaluated at the true values $\theta = (\beta, \gamma)$ of unknown parameters are:

$$\left.\frac{\partial^2 \ell(\tilde{\theta})}{\partial\tilde{\beta}\partial\tilde{\beta}'}\right|_{\tilde{\theta}=\theta} = -\frac{1}{\gamma}X'X \; ; \; \left.\frac{\partial^2 \ell(\tilde{\theta})}{\partial^2\tilde{\gamma}}\right|_{\tilde{\theta}=\theta} = \frac{n}{2\gamma^2} - \frac{1}{\gamma^3}(y-X\beta)'(y-X\beta)$$

and $$\left.\frac{\partial^2 \ell(\tilde{\theta})}{\partial\beta\partial\tilde{\gamma}}\right|_{\tilde{\theta}=\theta} = \left.\frac{\partial^2 \ell(\tilde{\theta})}{\partial\tilde{\gamma}\partial\tilde{\beta}'}\right|_{\tilde{\theta}=\theta} = -\frac{1}{\gamma^2}X'(y-X\beta)$$

Plugging these into the matrix $(\heartsuit)$ above, while recalling that $y - X\beta = \varepsilon$ and, since everything (including the log-likelihood function $\ell(\tilde{\theta})$) is conditional on X, also $E[\varepsilon|X] = 0$ (Assumption 2), $E[\varepsilon'\varepsilon|X] = n\sigma^2$ (implication of Assumption 4),

we get, after taking expectations (conditional on X), that

$$I(\theta) = \begin{bmatrix} \frac{1}{\sigma^2}X'X & 0_{k\times 1} \\ 0_{1\times k} & \frac{n}{2\sigma^4} \end{bmatrix}.$$

This block diagonal matrix can be inverted easily, so

$$(I(\theta))^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0_{k\times 1} \\ 0_{1\times k} & \frac{2\sigma^4}{n} \end{bmatrix}_{(k+1)\times(k+1)}$$

Note: The Asymptotic variance result from Section 2 of Lecture Notes Part 1 now states that

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N\left(0_{(k+1)\times 1}, (I(\theta))^{-1}\right) \text{ as } n \to \infty, \text{ where}$$

$\hat{\theta} = \left(b', \frac{e'e}{n}\right)'$ is the $(k+1)\times 1$ vector of the MLEs for the $(k+1)\times 1$ true parameter vector $\theta = (\beta', \sigma^2)'$.

Note: **Cramer-Rao inequality**: Let z be a vector of random variables (not necessarily independent) whose joint density is given by $f(\theta)$, where θ is an m-dimensional vector of unknown parameters in some parameter space Ω. Let $\ell(\tilde{\theta}) := f(\tilde{\theta})$ be the likelihood function, and let $\hat{\theta}$ be any unbiased estimator of θ with a finite variance-covariance matrix. Then under some regularity conditions (not stated here):

$$\text{Var}(\hat{\theta}) \geq (I(\theta))^{-1}$$

(These are $m\times m$ matrices, so the inequality is in the matrix sense, i.e. $\text{Var}(\hat{\theta}) - (I(\theta))^{-1}$ is positive semidefinite.)

Now, recall that $\text{Var}(b|X) = \sigma^2(X'X)^{-1}$ (see B) in Theorem 1, Section 1.3)

So, b, as an unbiased estimator, achieves the Cramer-Rao bound!

We have just proved an important result:

Claim Under assumptions 1–5, the OLS estimator b of β is BUE in that any other unbiased (but not necessarily linear) estimator has larger conditional variance in the matrix sense.

This should be distinguished from the Gauss-Markov theorem on page 10. that states that b is the minimum variance estimator among all estimators that are unbiased and linear in y. The above claim states that b has minimum variance in a much larger class of estimators! Note that this stronger result was obtained under additional Assumption 5 which was not used in the Gauss–Markov theorem.

Note: The MLE of σ^2 is biased, as mentioned on page 24; so Cramer-Rao bound does not apply here. But the OLS estimator s^2 of σ^2 is unbiased. Does it achieve the Cramer–Rao bound? In Homework 2 Problem 8 you're asked to show that $Var(s^2|X) = \frac{2\sigma^4}{n-k}$ under Assumption 1–5.

The Cramer-Rao bound is $\frac{2\sigma^4}{n}$; so almost achieved! Actually, it has been shown that an unbiased estimator of σ^2 with variance lower than $\frac{2\sigma^4}{n-k}$ does not exist (not shown here!)

Relation of the F-test with MLE method

The likelihood ratio test of the null hypothesis H_0 compares L_U, the maximized likelihood without the imposition of the restriction specified in H_0, with L_R, the likelihood maximized subject to the restriction specified in H_0.

The likelihood ratio is L_U/L_R and is usually denoted by λ. If it is too large, then this should be a sign that H_0 is false.

Recall the F-test from Section 1.4.2 for $H_0: R\beta = r$.

Using the expression for maximized likelihood ⊕ from page 24, we get that:

$$L_U \equiv \max_{\tilde{\beta}, \tilde{\sigma}^2} L(\tilde{\beta}, \tilde{\sigma}^2) = \left(\frac{2\pi}{n}\right)^{-n/2} e^{-\frac{n}{2}} \cdot (SSR_U)^{-n/2}$$

$$L_R \equiv \max_{\substack{\tilde{\beta}, \tilde{\sigma}^2 \\ \text{s.t. } H_0}} L(\tilde{\beta}, \tilde{\sigma}^2) = \left(\frac{2\pi}{n}\right)^{-n/2} \cdot e^{-\frac{n}{2}} (SSR_R)^{-n/2}, \text{ where}$$

SSR_U is the unrestricted sum of squared residuals

SSR_R is the restricted sum of squared residuals, as defined/explained on page -21-

Then the likelihood ratio is:

$$\lambda = \frac{L_U}{L_R} = \left(\frac{SSR_U}{SSR_R}\right)^{-n/2}$$

Compare this with the F-ratio on page 21. It is easy to see that $\boxed{F = \frac{n-k}{m}(\lambda^{2/n} - 1)}$

So, the F-ratio is just a monotone transformation of the likelihood ratio λ

So, the two tests are the same!

Caveat: All the results in Section 1.5. assume the Assumption 5 (Normality of error term).

1.6 Generalized least squares (GLS)

The Assumption 4 that the errors ε_i have the same variance may be too restrictive

For example, if y denotes a firm's profit and X is some measure of firm's size, then $Var(y)$ is likely to increase with X. Random disturbances (errors) with nonconstant variances are called heteroskedastic and often arise in cross-sectional studies in which one only has access to data that have been averaged within groups of different sizes.

Besides heteroskedasticity, the assumption of uncorrelated ε_i may also be untenable e.g. when the y_t's are computed via moving averages, as in the case where the rate of wage change is determined by $y_t = (w_t - w_{t-4})/w_{t-4}$ in quarter t from the wage indices w_t and w_{t-4}.

These considerations lead to the usual modification of Assumption 4, where $E[\varepsilon\varepsilon'|X] = \sigma^2 I_{n\times n}$ is relaxed and replaced by $\boxed{E[\varepsilon\varepsilon'|X] = \sigma^2 V(X)}$, where $V(X)$ is a nonsingular $n\times n$ matrix whose entries are in general nonlinear functions of X!

The classical regression model with this assumption and Assumptions 1–3 is called the <u>generalized regression model</u>.

After changing our Assumptions, the Gauss-Markov theorem no longer holds for the OLS estimator $b = (X'X)^{-1}X'y$. The BLUE is some other estimator!
The t-ratio is not distributed as the t-distribution; hence, the t-test is no longer valid. Same comments apply to the F-test.
However, the OLS estimator b is still unbiased, as proving part A of Theorem 1 in Section 1.3 did not require Assumption 4.

<u>Question</u>: If the matrix $V(X)$ is known, does there exist a BLUE for the generalized regression model?

Answer: Yes! The estimator is called the generalized least squares estimator (GLS). We find it next!

The main idea is to transform the generalized regression model (Assumptions 1, 2, 3 and $E[\varepsilon\varepsilon'|X] = \sigma^2 V(X)$, $V(X)$ nonsingular and known) into a model that satisfies Assumptions 1, 2, 3, 4 of the classical regression model.

Since $V \equiv V(X)$ is by construction symmetric and positive definite, so there exists a nonsingular $n \times n$ matrix C such that $V^{-1} = C'C$.

This decomposition is not unique with more than one choice for C, but, as it will be clear later, the choice of C does not matter!

Note: For example, C could be $D^{-1/2}Q'$ from Computational remarks on page 7.

Now consider creating a new regression model by transforming (y, X, ε) by C as:

$$\tilde{y} \equiv Cy, \quad \tilde{X} = CX, \quad \tilde{\varepsilon} \equiv C\varepsilon.$$

Then Assumption 1 $y = X\beta + \varepsilon$ implies $\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon}$. Also,

$$E[\tilde{\varepsilon}|\tilde{X}] = E[\tilde{\varepsilon}|X] = E[C\varepsilon|X] = CE[\varepsilon|X] = 0$$

- $E[\tilde{\varepsilon}|\tilde{X}] = E[\tilde{\varepsilon}|X]$: C is nonsingular so X and $\tilde{X}$ contain the same information.
- $E[\tilde{\varepsilon}|X] = E[C\varepsilon|X]$: linearity of conditional expectation
- $CE[\varepsilon|X] = 0$: Assumption 2 $E[\varepsilon|X] = 0$

It is also not difficult to prove that since V is positive definite, the no-multicollinearity assumption (Assumption 3) is also satisfied. Next, we consider

$$E[\tilde{\varepsilon}\tilde{\varepsilon}'|\tilde{X}] = E[\tilde{\varepsilon}\tilde{\varepsilon}'|X] = CE[\varepsilon\varepsilon'|X]C' = C\sigma^2 VC' = \sigma^2 CVC'$$

by new relaxed Assumption

Now, $V^{-1} = C'C$, so $(C')^{-1}V^{-1}C^{-1} = I_{n\times n}$

or $CVC' = I_{n\times n}$

Hence, $E[\tilde{\varepsilon}\tilde{\varepsilon}'|\tilde{X}] = \sigma^2 I_{n\times n}$

Finally, the distribution of $\tilde{\varepsilon}|\tilde{X}$ is the same as the distribution of $\tilde{\varepsilon}|X$. Moreover, since $\tilde{\varepsilon} = C\varepsilon$ is just a linear transformation of ε, if one assumes that $\varepsilon|X$ is normal, then $\tilde{\varepsilon}|\tilde{X}$ is normal as well.

Therefore, the transformed model satisfies the Assumptions 1–5.

The Gauss-Markov Theorem (Theorem 1 in Section 1.3) can now be applied to the transformed model, and it implies that the BLUE for β in the generalized regression model is:

$$\hat{\beta}_{GLS} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} = ((CX)'(CX))^{-1}(CX)'Cy = (X'C'CX)^{-1}(X'C'Cy)$$

i.e. $$\boxed{\hat{\beta}_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}y}$$

Its conditional variance is

$$Var(\hat{\beta}_{GLS}|X) = (X'V^{-1}X)^{-1}X'V^{-1}\cdot Var(y|X)\cdot V^{-1}X(X'V^{-1}X)^{-1} =$$

$$= (X'V^{-1}X)^{-1}X'V^{-1}\cdot\sigma^2 V\cdot V^{-1}X(X'V^{-1}X)^{-1} =$$

(since $Var(y|X) = Var(\varepsilon|X)$)

$$= \sigma^2(X'V^{-1}X)^{-1}$$

Note: Note that equivalently:

$$\hat{\beta}_{GLS} = (X'(Var(\varepsilon|X))^{-1}X)^{-1}X'(Var(\varepsilon|X))^{-1}y$$

Compare this to the OLS estimator $b = (X'X)^{-1}X'y$ which is also unbiased, as mentioned before. The GLS estimator is more efficient unbiased estimator, since its variance is smaller in the matrix sense. The gain in efficiency is achieved by exploiting the heteroskedasticity and correlation between observations in the error term, which operationally is to insert the inverse of $Var(\varepsilon|X)$ in the usual OLS estimator formula.

Summary: A) Under assumptions 1-3; $E[\hat{\beta}_{GLS}|X] = \beta$ (unbiasedness)

B) ——— ||- ——— and $E[\varepsilon\varepsilon'|X] = \sigma^2 V(X)$: $Var[\hat{\beta}_{GLS}|X] = \sigma^2(X'(V(X))^{-1}X)^{-1}$

C) $\hat{\beta}_{GLS}$ is BLUE for β.

Note: Important special case of generalized regression model is the so-called weighted regression or weighted least squares (WLS).

In weighted regression there is no correlation in the error term between observations, so that V is diagonal. Let $v_i(X)$ be the i^{th} diagonal element of $V(X)$.

So, $E[\varepsilon_i^2 | X] = Var(\varepsilon_i | X) = \sigma^2 \cdot v_i(X)$

and $E[\varepsilon_i \varepsilon_j | X] = 0 \quad i,j = 1, \ldots, n, \; i \neq j$.

If C is such that $C'C = V^{-1}$ (as above), then it is easy to see that C is also diagonal, with $\frac{1}{\sqrt{v_i(X)}}$ in the i^{th}-diagonal entry. Therefore,

$$\tilde{y} = \frac{y_i}{\sqrt{v_i(X)}} \quad \text{and} \quad \tilde{x}_i = \frac{x_i}{\sqrt{v_i(X)}} \qquad i = 1, 2, \ldots, n$$

Therefore, efficient estimation under a known form of heteroskedasticity is first to weigh each observation by the reciprocal of the standard deviation $\sqrt{v_i(X)}$ and then apply OLS. This is the weighted regression!

Caveat: Very rarely do we have a priori information specifying $V(X)$!
The specification of V may involve background knowledge of the data and the empirical study. It may also arise from examination of the residuals in situations where there are multiple observations (X_i, y_i) at distinct input values x_i. In this case, heteroskedasticity can be revealed from the residual plots and unbiased estimates of $Var(y_i)$ at each distinct input value can be obtained from the multiple observations there. The form of V may also arise as part of the model, as you'll see in time series modeling of asset returns and their volatilities, where ML will be used to estimate both the regression parameters and the entries of V. However, if $\hat{V}$ is estimated from the data, this affects the distribution and nice properties of $\hat{\beta}_{GLS}$.

1.7 Regression in practice

We'll borrow some examples from Ruppert ([3] on syllabus) and Lai & Xing ([4] on syllabus).

Example: The dependent variable is the change in the corporate AAA bond yield (aaa_dif in the code below)

The regressors are: change in the 10-year Treasury rate (cm10_dif in the code below)
change in the 30-year Treasury rate (cm30_dif in the code below)

Data is weekly. Code is in SAS.

Note: cm30 has missing values at the beginning of the data set, hence ⊛

```
options linesize = 72 ;
data WeeklyInterest ;
infile 'C:\book\SAS\WeeklyInterest.dat' ;
input month day year ff tb03 cm10 cm30 discount prime aaa ;
if lag(cm30) > 0 ; ⊛
aaa_dif = dif(aaa) ;
cm10_dif = dif(cm10) ;
cm30_dif = dif(cm30) ;
id = _N_ ; #of observations used, turns out to be 879.
run ;
title 'Weekly Interest Rates' ;
proc reg ;
model aaa_dif = cm10_dif cm30_dif / ss1 ss2 vif ;
output out=WeeklyInterest predicted=predicted rstudent=rstudent cookd=cookd h=leverage ;
run ;
proc gplot ;
plot rstudent*predicted ;
plot (rstudent cookd leverage cm10_dif cm30_dif)*id ;
plot cm10_dif*cm30_dif ;
run ;
```

Here is the output from this program.

The REG Procedure
Dependent Variable: aaa_dif

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	11.35366	5.67683	1357.95	<.0001
Error	876	3.66206	0.00418		
Corrected Total	878	15.01572			

"ANOVA" or "AOV" table 1.6.1

Root MSE	0.06466	R-Square	0.7561
Dependent Mean	-0.00130	Adj R-Sq	0.7556
Coeff Var	-4986.33904		

Parameter Estimates 1.6.2.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	-0.00010686	0.00218	-0.05	0.9609	0.00148
cm10_dif	1	0.36041	0.04456	8.09	<.0001	11.20585
cm30_dif	1	0.29655	0.04987	5.95	<.0001	0.14781

Parameter Estimates

Variable	DF	Type II SS	Variance Inflation
Intercept	1	0.00001004	0
cm10_dif	1	0.27353	14.03581
cm30_dif	1	0.14781	14.03581

Next, we explain the output for this model: $aqq\_dif_i = \beta_1 + \beta_2 \cdot cm10\_dif_i + \beta_3 \cdot cm30\_dif_i + \varepsilon_i$ $i = 1, 2, \ldots, 879$

1.7.1 **ANOVA table** $k = 3$ regressors (one is the constant) $n = 879$

"Analysis of variance table".

Sums of squares and R^2 (again!)

The total variation of y can be partitioned into two parts: the variation that can be predicted by X and the part that cannot be predicted.

Total variation of y is measured by the total sum of squares (total SS), which is (as already mentioned on page 9): $\boxed{\text{total SS} = \sum_{i=1}^{n} (y_i - \bar{y})^2}$

It is called "Corrected Total" in the ANOVA table, and is 15.01572

The variation that can be predicted by X is measured by the regression sum of squares which is $\boxed{\text{regression SS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}$ and it is called "model sum of squares" in the ANOVA table (equal to 11.35366)

Finally, part of variation that cannot be predicted by a linear function of X is measured by the residual error sum of squares (also denoted by SSR in Notes before) $\boxed{\text{residual error SS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2}$

and it is called the "error sum of squares" in the ANOVA table (equal to 3.66206)

As mentioned on page 9, you'll have to prove in Homework 2, Problem 10 that

$\boxed{\text{total SS} = \text{regression SS} + \text{residual error SS}}$

$15.01572 = 11.35366 + 3.66206$

R^2, denoted by R-square in the output is, as explained on page 9, equal to $R^2 = \dfrac{\text{regression SS}}{\text{total SS}} = 1 - \dfrac{\text{residual error SS}}{\text{total SS}}$

and measures the proportion of the total variation of y that can be linearly predicted by X. In our case, $R^2 = 0.7561 = 1 - \dfrac{3.66206}{15.01572}$ which indicates

(being close to 1) a strong support for our model.

Note: Recall that the correlation coefficient between two random variables X and Y is defined as $\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$ where σ_X, σ_Y are the std. deviations of X and Y respectively.

Since $|Cov(X,Y)| \leq \sqrt{Var(X)} \cdot \sqrt{Var(Y)}$ (Cauchy-Schwarz inequality), we have $|\rho| \leq 1$. The method-of moments estimate (which replaces population moments by their sample counterparts) of ρ based on the sample data $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ is the sample correlation coefficient

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Note that $r_{XY}^2 = \hat{\beta}_2^2 \cdot \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$

where $\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ is the OLS estimator of the regression coefficient β_2 of the y_i on the x_i in the model: $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, $i=1,\ldots,n$

(Check page 63. in the Lecture Notes, Part 1!)

If one looks at it from this new perspective, one sees that $R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$ could be thought of as the multiple correlation coefficient for the multiple regression model $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$

$i=1,2,\ldots,n$ and $x_{i1}=1$.

Degrees of freedom (DF in ANOVA table)

There are degrees of freedom associated with each of these "sources" of variation. The degrees for regression is $p = k-1$, i.e. the # of regressors aside from the constant one. The total df is $n-1$. The residual error df is $n-p-1 = n-k$. Here is a way to think of df. Initially, there are n df, one for each observation. Then one df is allocated to estimation of the constant regressor β_1, that is, the "intercept" of our model. This leaves a total of $n-1$ df left for modeling the effects of each of the $p = k-1$ nonconstant regressors. Each of these uses one df for estimation. This leaves $(n-1)-p = n-k$ df remaining for estimation of σ^2 using the residuals. The point we are making here is even more evident if one thinks of the OLS geometrically, as explained on page 6. of the notes.

Mean sums of squares

The mean sum of squares (MS) for any "source" is its sum of squares divided by its degrees of freedom. The total MS is $\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$ (which is the sample variance of y_i's). So, in that output, for example, $\frac{11.33366}{2} = 5.67683$.

The residual error MS (0.00418 in the output) is our estimator s^2 of σ^2. Indeed, $s^2 = \frac{e'e}{n-k} = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p-1}$

Standard error of regression

denoted by ROOT MSE in the output, is just $s = \sqrt{s^2}$ (see page 8. of Notes).

$$0.06466 = \sqrt{0.00418}$$

F value in the output is the value of the F-ratio on page 21 in the Notes for the null hypothesis $H_0: \beta_2 = 0, \beta_3 = 0$; that is, the hypothesis that the slopes β_2 and β_3 are both zero. In other words, it is the value of the F-ratio for

the hypothesis that there is no linear relationship between y (aaa_dif) and any of the regressors (cm10_dif, cm30_dif).

Homework2 Problem10 will convince you that this F-ratio value (1357.95) is actually $\frac{\text{regression MS}}{\text{residual error MS}} = \frac{5.67683}{0.00418} = 1357.95$

The entry in the column labeled <u>Pr>F</u> is the p-value of this test. Here, $p<0.0001$ is very small, which is a very strong evidence <u>against</u> H_0.

We conclude that there <u>IS</u> a relationship between changes in cm10 and/or cm30 on one hand and changes in aaa on the other

<u>Adj R-Sq</u> in the output (or Adjusted R^2)

Consider the formula for R^2. $\quad R^2 = 1 - \frac{\text{residual error SS}}{\text{total SS}}$

Let's rewrite it as $R^2 = 1 - \frac{\frac{1}{n}(\text{residual error SS})}{\frac{1}{n}(\text{total SS})}$

Now, the top of this expression is actually the MLE for σ^2 (see page 24 of Notes) and as such is a biased estimate of σ^2, the variance of ε_i.

Moreover, $\frac{1}{n}(\text{total SS})$ is a biased estimate of the variance of y_i's.

The fact that both estimators are biased is what biases R^2 towards larger models.

The biases can be removed by replacing $\frac{1}{n}$ by $\frac{1}{n-p-1}$ on top, and $\frac{1}{n}$ by $\frac{1}{n-1}$ in the bottom of that expression. We get:

$$\boxed{\text{adjusted } R^2 = 1 - \frac{\text{residual error MS}}{\text{total MS}}}$$

In our output

$$\text{adj } R^2 = 0.7556 = 1 - \frac{0.00418}{\frac{15.01572}{878}}$$

Sequential and partial sums of squares

When there are two or more regressors, the regression $SS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ can be divided into the portion due to the first non-constant regressor, portion due to the second non-constant regressor, etc.

With each regressor, there are two different sums of squares that SAS associates with it:

Sequential sum of squares (type I SS in SAS)

This is the increase in regression sum of squares when that regressor is added to the model that contains all other regressors that precede it in the program.

Clearly, some thought on the order in which to list the regressors is needed if you plan to use sequential sums of squares for model selection.

Sequential sums of squares is very useful when there is a natural order to regressors e.g. in polynomial regression, where the regressors are X, X^2, X^3 etc. (see homework2 problem 1).

partial sum of squares (type II SS in SAS)

This is the increase in regression sum of squares if that regressor is added to the model containing ALL of the other regressors.

Here the order of the regressors in the program is irrelevant.

By definition, Type I and Type II sums of squares should be equal for the regressor that is last in the list, cm30_dif in our example.

Any regressor that has an insignificant type II SS could be dropped from the model, provided that all the other regressors are retained.

In our example, regression SS = 11.35366 = 11.20585 + 0.14781

- regression SS with only cm_10 in the model ← type I SS for cm10_dif
- type I SS for cm30_dif → increase in regression SS when cm30_dif is added to the model already containing cm10_dif

The smallness of 0.14781 relative to 11.20585 should not be interpreted as cm30_dif being less important than cm10_dif. The variables that enter the model first generally have larger type I SS. The point here is that adding a second regressor to the model

(whichever that is, cm10_dif or cm30_dif) does not improve prediction very much, not necessarily that cm10_dif is "better" than cm30_dif. This point can be seen in the partial (type II) sums of squares which are small for both variables.

1.7.2 Parameter estimates

We got that: $b_1 = -0.00010686$, the estimator for β_1 (intercept)

$b_2 = 0.36041$, —"— β_2 (slope for cm10_dif)

$b_3 = 0.29655$, —"— β_3 (slope for cm30_dif)

Standard errors $\sqrt{s^2((X'X)^{-1})_{ll}}$ (see page 16 of Notes)

$SE(b_1) = 0.00218$, $SE(b_2) = 0.04456$, $SE(b_3) = 0.04987$

These can be used to set up confidence intervals for b_i, $i=1,2,3$ (see page 18 of Notes)

t value in output

for the l^{th} regressor is actually the value of the t-ratio t_l on page 16 in Notes.

When the null hypothesis is $H_0: \beta_l = 0$ (i.e. $\beta_l^0 = 0$).

In other words, the t-value stands for $t_l = \dfrac{b_l}{\sqrt{s^2 \cdot ((X'X)^{-1})_{ll}}} = \dfrac{b_l}{SE(b_l)}$

For example, in output, the t-value of β_1 (the intercept) is $-0.05 = \dfrac{-0.00010686}{0.00218}$

"Pr>|t|" column in output contains p-values of these t-tests.

We can see that cm10_dif and cm30_dif both have very small p-values (<0.0001) indicating that they are statistically significant, i.e. the null hypotheses $H_0: \beta_2 = 0$ and $H_3: \beta_3 = 0$ are both rejected.

However, $H_0: \beta_1 = 0$ is accepted (p-value is $0.9609 > 0.1$), and the intercept is not statistically significant (can be dropped, as it often happens with interest rate data!)

Note: The last conclusion that both cm10_dif and cm30_dif should be in the model is somewhat against the conclusion we made based on the sequential/partial SS. It could be that the 879 observations is too small of a sample, so a small effect of no practical significance could be statistically significant.

1.7.3 Model selection

Question: How to select the best subset of regressors out of a set of possible ones?

There are two principles to balance:

1) Models with larger # of regressors have less bias and they would give the best prediction if all coefficients could be estimated without error.
2) When unknown coefficients are replaced by estimates, then the prediction becomes less accurate, and this effect is worse when there are more coefficients to estimate. Larger models do have less bias, but also have more variability!

Models with too few regressors and sizeable bias are said to be underfit, while models with too many parameters are said to overfit.

t-test traditional approach

Select an order of regressor to be added to the model one by one.

Specifically, letting x_j denote the latest regressor entered into our model, perform the t-test of $H_0: \beta_j = 0$ (i.e. read its t-value or associated p-value in the output).

Sometimes, this t-value is actually called the "partial F-statistic associated with x_j" (in the presence of previously entered regressors).

Stepwise inclusion of regressors terminates as soon as the latest regressor entered is not significantly different from 0 (i.e. its p-value is too large).

There are many weaknesses of this approach. First, it depends heavily on the order in which potential regressors are added to the model (unless we have some reasonable

a priori information on the order of relevance of potential regressors.
Secondly, this stepwise procedure depends on the $\alpha \equiv$ the significance level (often 5%). Note, however, that since the procedure carries a sequence of t-tests, the overall significance level is probably much different from α. Thirdly, the procedure only considers the type I error, and the type II error of accepting H_0 when $\beta_j \neq 0$ is not used at all by the procedure.

R^2 and adjusted R^2 approach

R^2 is not a useful statistic for comparing models of different sizes. It is biased towards larger models and it always chooses the largest model. Adjusted R^2 statistic is adjusted to remove biases and CAN be used to select models.

Forward-backward regressor selection Suppose there are k potential regressors to choose from. To begin, we introduce the partial correlation coefficient that is used in the forward inclusion of variables. Given response variables u_i, v_i and regressors $(x_{i1}, x_{i2}, \ldots, x_{ik})$, $1 \leq i \leq n$, we can regress the u_i (respectively v_i) on $x_{i1}, x_{i2}, \ldots, x_{ik}$ and denote the residuals by e_i^u (respectively e_i^v). The correlation coefficient (see r_{XY} on page 35 of Notes, with $x_i \equiv e_i^u$ and $y_i \equiv e_i^v$) between e_i^u and e_i^v is called the partial correlation coefficient between u and v adjusted for $x_1, x_2, \ldots, x_k$ and is denoted by $r_{uv|x_1 \ldots x_k}$. The forward selection procedure is now defined as follows:

FORWARD ↑

To select the regressor that first enters the model, compute the USUAL correlation coefficients r_j of $\{(y_i, x_{ij}) \mid 1 \leq i \leq n\}$ for $j = 1, 2, \ldots, k$ and choose the regressor x_j with the largest r_j. After p regressors have been entered, we compute the partial correlation coefficients between y and the regressors not entered into

FORWARD

the model and then include the regressor with the largest partial correlation coefficient. This forward stepwise procedure terminates when the t-statistic associated with the latest regressor entered tells us that its regression coefficient is not significantly different from zero.

BACKWARD

Backward elimination procedure begins with the full model (with all k regressors) and computes the t-statistic for each regressor. The smallest one, call it T_L, is compared to pre-specified cutoff value T^* associated with the $\alpha/2$-quantile of the t-distribution, where α is usually chosen to be 0.05 or 0.01. see page 17. of Notes

If $T_L \geq T^*$, terminate the elimination procedure and choose the full model

If $T_L < T^*$, conclude that β_L is not significantly different from 0 and remove X_L from the regressor set. With the set of remaining regressors treated as the full regression model at every stage, carry out this backward elimination procedure.

Forward-backward procedure

This is a hybrid of the previous two procedures. It's computationally appealing to perform the forward selection procedure followed by the backward elimination.

Next, you can see an example of a backward elimination procedure.

which is called the *generalized least squares* (GLS) estimator. Moreover, $\widehat{\beta}_{\text{GLS}}$ is unbiased and

$$\text{Cov}(\widehat{\beta}_{\text{GLS}}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} \tag{1.54}$$

under assumption (1.52). Note that for the special case $\mathbf{V} = \sigma^2\mathbf{I}$, the right-hand side of (1.54) reduces to $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, which agrees with (1.9).

To prove (1.53) and (1.54), we use a result from Section 2.2: For a symmetric and positive definite matrix $\mathbf{V}$, there exists a symmetric and positive definite matrix $\mathbf{P}$ such that $\mathbf{PP} = \mathbf{V}$; i.e., $\mathbf{P} = \mathbf{V}^{1/2}$. Multiplying the regression model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ by $\mathbf{P}^{-1}$ yields

$$\mathbf{P}^{-1}\mathbf{Y} = \mathbf{P}^{-1}\mathbf{X}\beta + \mathbf{u}, \tag{1.55}$$

where $\mathbf{u} = \mathbf{P}^{-1}\epsilon$ has covariance matrix $\mathbf{P}^{-1}\text{Cov}(\epsilon)\mathbf{P}^{-1} = \mathbf{P}^{-1}\mathbf{PPP}^{-1} = \mathbf{I}$. Thus the model (1.55) has $\text{Cov}(\mathbf{u}) = \mathbf{I}$, for which the OLS estimate is of the form

$$[(\mathbf{P}^{-1}\mathbf{X})^T(\mathbf{P}^{-1}\mathbf{X})]^{-1}(\mathbf{P}^{-1}\mathbf{X})^T\mathbf{P}^{-1}\mathbf{Y} = (\mathbf{X}^T\mathbf{P}^{-1}\mathbf{P}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{P}^{-1}\mathbf{P}^{-1}\mathbf{Y},$$

which is the same as $\widehat{\beta}_{\text{GLS}}$ in (1.53) since $\mathbf{P}^{-1}\mathbf{P}^{-1} = (\mathbf{PP})^{-1} = \mathbf{V}^{-1}$. Therefore, using the transformation (1.55), GLS can be transformed to OLS and thus shares the same properties of OLS after we replace $\mathbf{X}$ by $\mathbf{P}^{-1}\mathbf{X}$. In particular, (1.54) follows from (1.9) after this transformation.

The specification of $\mathbf{V}$ may involve background knowledge of the data and the empirical study, as noted in the first paragraph of this section. It may also arise from examination of the residuals in situations where there are multiple observations $(\mathbf{x}_t, y_t)$ at distinct input values. In this case, heteroskedasticity can be revealed from the residual plots and unbiased estimates of $\text{Var}(y_t)$ at each distinct input value can be obtained from the multiple observations there. The form of $\mathbf{V}$ may also arise as part of the model, as in time series modeling of asset returns and their volatilities in Chapter 6, where Section 6.4 uses maximum likelihood to estimate both the regression parameters and the parameters of $\mathbf{V}$.

1.8 Implementation and illustration

To implement the methods described in this chapter, one can use the following functions in `R` or `Splus`:

```
lm(formula, data, weights, subset, na.action)
predict.lm(object, newdata, type)
bootstrap(data, statistic, B)
step(object, scope, scale, direction)
```

Example from Lai-Xing ([4] on syllabus)

For details, see Venables and Ripley (2002, pp. 139–163). Alternatively one can use the Matlab function `regress`. We illustrate the application of these methods in a case study that relates the daily log returns of the stock of Microsoft Corporation to those of several computer and software companies: see Section 3.1 for the definition and other details of asset returns.

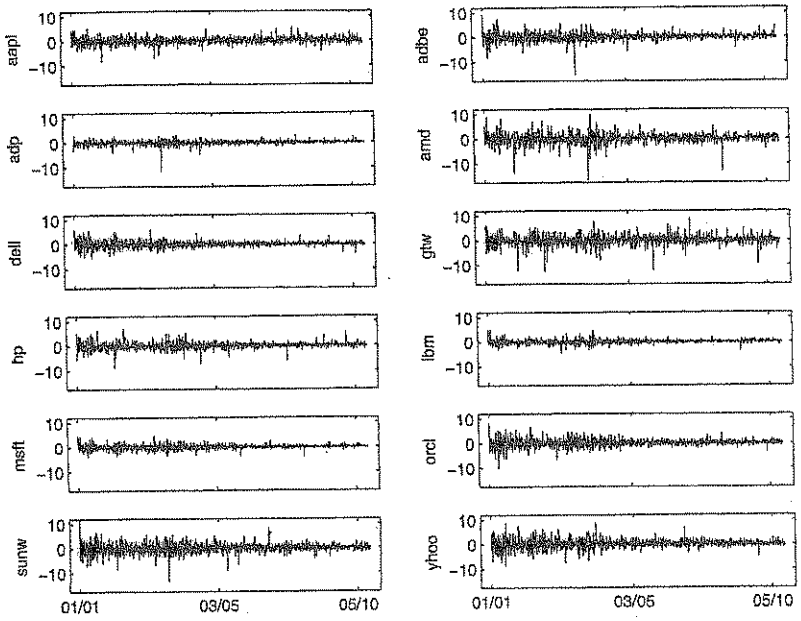


Fig. 1.4. The daily log returns of Apple Computer, Adobe Systems, Automatic Data Processing, Advanced Micro Devices, Dell, Gateway, Hewlett-Packard Company, International Business Machines Corp., Microsoft Corp., Orcale Corp., Sun Microsystems, and Yahoo! stocks from January 3, 2001 to December 30, 2005.

Microsoft Corp. (`msft`) is a leader in the application software industry. We collected a set of firms whose stock returns may be strongly correlated with `msft`. These firms include Apple Computer (`aapl`), Adobe Systems (`adbe`), Automatic Data Processing (`adp`), Dell (`dell`), Gateway (`gtw`), Hewlett-Packard Company (`hp`), International Business Machines Corp. (`ibm`), Orcale Corp. (`orcl`), Sun Microsystems (`sunw`), and Yahoo! (`yhoo`). Figure 1.4 plots the daily log returns of the stocks of these firms from January 3, 2001 to December 30, 2005. The sample size is $n = 1255$.

The correlation matrix of the daily log returns in Table 1.2 shows substantial correlations among these stocks. Figure 1.5 gives a matrix of scatterplots

showing all pairs of stocks. The scatterplots also reveal a strong relationship between msft and each predictor and a strong pairwise relationship among the predictors.

Table 1.2. Pairwise correlation coefficients of daily log returns.

	aapl	adbe	adp	amd	dell	gtw	hp	ibm	msft	orcl	sunw
adbe	.387										
adp	.285	.305									
amd	.448	.382	.310								
dell	.515	.486	.316	.470							
gtw	.355	.266	.195	.347	.368						
hp	.431	.425	.342	.429	.528	.380					
ibm	.430	.451	.387	.445	.532	.282	.477				
msft	.479	.526	.366	.457	.620	.375	.493	.603			
orcl	.407	.453	.319	.385	.510	.284	.446	.539	.587		
sunw	.425	.406	.273	.440	.511	.369	.472	.472	.455	.517	
yhoo	.422	.497	.322	.422	.499	.279	.422	.441	.493	.469	.414

Regression for full model

We use OLS to fit a linear regression model to the returns data. The regression coefficient estimates and their standard errors, t-statistics, and p-values are shown in Table 1.3. The p-values in Table 1.3 measure the effect of dropping that stock from the regression model; a p-value less than 0.01 shows the regression coefficient of the corresponding stock to be significantly nonzero by the t-test in (1.22). We can use Table 1.3 together with (1.11) to construct confidence intervals of the regression coefficients. For example, a 95% confidence interval of the regression coefficient for dell is $0.1771 \pm t_{1243,0.025} 0.0213$.

Variable selection

Starting with the full model, we find that the stocks hp and sunw, with relatively small partial F-statistics, are not significant at the 5% significance

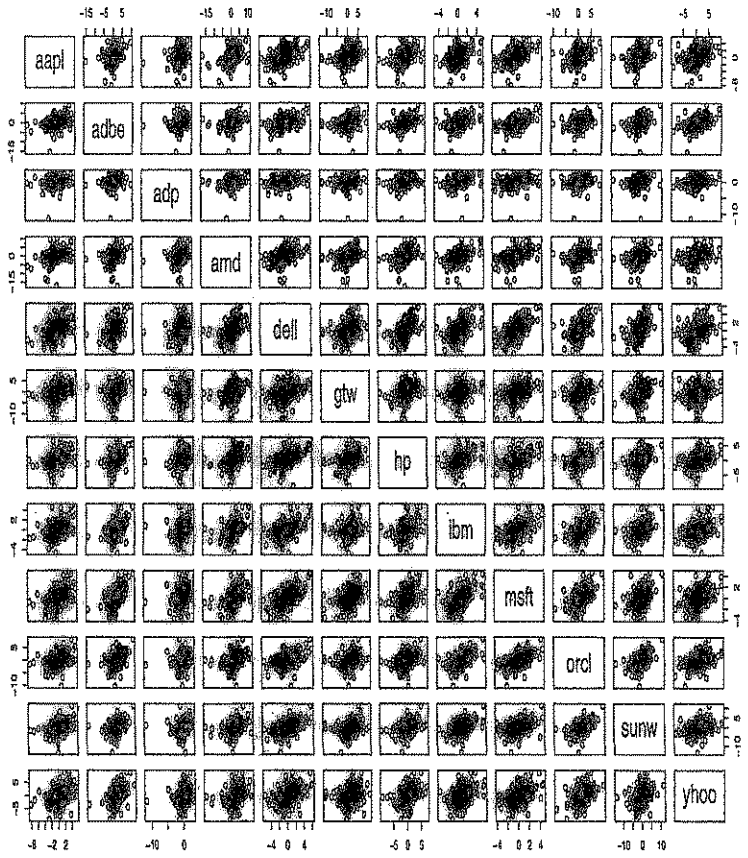


Fig. 1.5. Scatterplot matrix of daily log returns.

level. If we set the cutoff value at $F^* = 10$, which corresponds to a significance level< 0.01, then hp and sunw are removed from the set of predictors after the first step. We then refit the model with the remaining predictors and repeat the backward elimination procedure with the cutoff value $F^* = 10$ for the partial F-statistics. Proceeding stepwise in this way, the procedure terminates with six predictor variables: aapl, adbe, dell, gtw, ibm, orcl. The variables that are removed in this stepwise procedure are summarized in Table 1.4.

Table 1.5 gives the regression coefficients of the selected model

$$\texttt{msft} = \beta_0 + \beta_1 \texttt{aapl} + \beta_2 \texttt{adbe} + \beta_3 \texttt{dell} + \beta_4 \texttt{gtw} + \beta_5 \texttt{ibm} + \beta_6 \texttt{orcl} + \varepsilon. \quad (1.56)$$

Table 1.3. Regression coefficients of the full model

Term	Estimate	Std. Error	t-statistic	p-value
Intercept	0.0028	0.0163	0,170	0.8649
aapl	0.0418	0.0162	2.573	0.0102
adbe	0.0782	0.0139	5.639	0.0000
adp	0.0489	0.0227	2.158	0.0311
amd	0.0192	0.0107	1.796	0.0727
dell	0.1771	0.0213	8.325	0.0000
gtw	0.0396	0.0099	3.991	0.0000
hp	0.0270	0.0177	1.526	0.1273
ibm	0.2302	0.0275	8.360	0.0000
orcl	0.1264	0.0154	8.204	0.0000
sunw	−0.0206	0.0121	−1.696	0.0902
yhoo	0.0258	0.0124	2.085	0.0372

RSS = 411.45, d.f.= 1243, $s = 0.575$, adjusted $R^2 = 56.9\%$.

Table 1.4. Stepwise variable selection.

Step	1	2	3
Variables removed	sunw, hp	amd	adp, yhoo

The selected model shows that, in the collection of stocks we studied, the msft daily log return is strongly influenced by those of its competitors. Instead of employing backward elimination using partial F-statistics, we can proceed with stepwise forward selection using the AIC. Applying the R or Splus function stepAIC with 6 as the maximum number of variables to be included, stepAIC also chooses the regression model (1.56).

Regression diagnostics

For the selected model (1.56), Figure 1.6 shows diagnostic plots from R. The top panels give the plots of residuals versus fitted values and the Q-Q plot. The

C_p, AIC, Schwarz's (BIC) information criteria

Both AIC and C_p consist of a term measuring how well the model fits the data and another term that penalizes the number of regressors.

The C_p-statistic introduced by Mallows is defined as $\boxed{C_p = \frac{SSR_p}{s_k^2} + 2p - n}$

when a subset of p input regressors is chosen from the full set of potential regressors $\{x_1, x_2, \ldots, x_k\}$. Here SSR_p denotes the residual error SS for the model that uses only that subset of p regressors, while s_k^2 denotes the OLS estimate of σ^2 when ALL k regressors are used.

We want SSR_p to be as small as possible. The quantity $2p$ is the penalty for the # of regressors used.

Mallows in the 70's proposed choosing the subset that has the smallest C_p value. Why?

If our model with $p<k$ regressors has little bias, then $\frac{SSR_p}{s_k^2}$ should be close to $n-p$, since from page 13. we know that $E[e'e|X] = (n-p)\sigma^2$ for a model with p regressors, while s_k^2 is the OLS estimator of σ^2.

Hence, for a model with little bias, C_p should be equal to p, the # of regressors.

For models with bias $E[y_i] \neq E[x_i'b]$ and therefore $E[e_i] = E[y_i - x_i'b] \neq 0$

Since $E[e_i^2] = (E[e_i])^2 + Var(e_i)$, one gets:

$$E[SSR_p] = \sum_{i=1}^{n} E[e_i^2] = \sum_{i=1}^{n} (E[y_i - x_i'b])^2 + tr(Cov(e))$$

But $e = M\varepsilon$ (Homework 2 Problem 6), (M idempotent: page 6)

$$tr(Cov(e)) = tr(Cov(M\varepsilon)) = tr(M Cov(\varepsilon) M) = tr(M\sigma^2 I_{n\times n} M) =$$

$$= \sigma^2 tr(M) = (n-p)\sigma^2$$

(using page 13, step 2 with $k \equiv p$)

So, $E[SSR_p] = \sum_{i=1}^{n} (E[y_i - x_i'b])^2 + (n-p)\sigma^2$ ⊛

Therefore, if the model has bias (due to a poorly chosen subset of $p<k$ regressors), then $E[y_i - x_i'b] \neq 0$, so ⊛ implies that C_p will be significantly larger than p, the #of regressors chosen.

For models with many regressors, typically C_p is close to p but both of these quantities are large, in particular substantially larger than the minimum value of C_p. Such models should be considered overfits, that is, not biased but with too many parameters.

Akaike introduced an information criterion **AIC** based on the likelihood theory. It's defined as

$$\{-2\log(\text{maximized likelihood}) + 2(\#\text{of parameters})\}/n \quad \oplus$$

Including more input regressors in the regression model increases the model "information" as measured by $2(\log(\text{maximized likelihood}))$, but it also increases the penalty term $2\cdot(\#\text{of parameters})$, which is the same as in Mallows' C_p. The selection procedure is to <u>choose the model with the smallest AIC</u>!

<u>Note</u>: Since the MLE for β is the same as b, the OLS estimator, while the MLE $\hat{\sigma}_p^2$ of σ^2 is $\frac{RSS_p}{n}$, choosing the regression model with the smallest ⊕ is equivalent to <u>minimizing</u>

(see page 24 of Notes)

$$\boxed{AIC(p) = \log(\hat{\sigma}_p^2) + \frac{2p}{n}}$$

where p denotes the #of regressors used.

Indeed, by ⊕ on page 24 of the Notes, the maximized log likelihood is

$-\frac{n}{2}\log\left(\frac{2\pi}{n}\right) - \frac{n}{2} - \frac{n}{2}\log(SSR_p)$, so from ⊕ on previous page, maximizing

$$\frac{1}{n}\left\{-2\log(\text{max. likelihood}) + 2p\right\} = \frac{1}{n}\left\{+n\log\left(\frac{2\pi}{n}\right) + n + n\log(n\hat{\sigma}_p^2) + 2p\right\} =$$

$$= \log\left(\frac{2\pi}{n}\right) + 1 + \log(n\hat{\sigma}_p^2) + \frac{2p}{n} \quad \text{is}$$

equivalent to maximizing $AIC(p) = \log(\hat{\sigma}_p^2) + \frac{2p}{n}$

Schwarz, using Bayesian methods, derived the Bayesian information criterion

$$\boxed{BIC(p) = \log(\hat{\sigma}_p^2) + \frac{p\log n}{n}}$$

Note
BIC is sometimes denoted as SBC.

The selection procedure is to choose the model with the smallest BIC
Note that whereas the AIC uses 2 as the penalty factor, the BIC penalizes the number of regressors more heavily by using $\log n$ instead of 2!
After all the listed regressor selection methods, probably the best one is some sort of a forward-backward selection using AIC, BIC and C_p in each step. Using only AIC, BIC and/or C_p leads to a large combinatorial optimization problem of checking all possible subsets of k potential regressors, that can be very computationally expensive unless k is small (there are 2^k subsets!).
Next is the continuation of the example from page 33. Now, aside from cm10 and cm30, one has ff_dif = weekly change in the Federal funds rate
prime_dif = weekly change in the prime rate
as possible regressors that can be used to predict aaa_dif.

```
options linesize = 72 ;
data WeeklyInterest ;
infile 'C:\book\SAS\WeeklyInterest.dat' ;
input month day year ff tb03 cm10 cm30 discount prime aaa ;
if lag(cm30) > 0 ;
aaa_dif = dif(aaa) ;
cm10_dif = dif(cm10) ;
cm30_dif = dif(cm30) ;
ff_dif = dif(ff) ;
prime_dif = dif(prime) ;
run ;
title 'Weekly Interest Rates' ;
proc reg ;
model aaa_dif = cm10_dif cm30_dif ff_dif prime_dif/selection=rsquare adjrsq cp sbc aic ;
run ;
```

The REG Procedure
Model: MODEL1
Dependent Variable: aaa_dif

R-Square Selection Method

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	SBC
1	0.7463	0.7460	35.4718	-4778.8065	-4769.24795
1	0.7379	0.7376	65.5166	-4750.2675	-4740.70994
1	0.0625	0.0615	2489.033	-3630.0113	-3620.45378
1	0.0320	0.0309	2598.720	-3601.8083	-3592.25074
2	0.7561	0.7556	2.1482	-4811.5872	-4797.25086
2	0.7463	0.7458	37.2491	-4777.0205	-4762.68417
2	0.7463	0.7457	37.4036	-4776.8714	-4762.53501
2	0.7391	0.7385	63.2227	-4752.2898	-4737.95341
2	0.7379	0.7373	67.5166	-4748.2675	-4733.93116
2	0.0727	0.0706	2454.497	-3637.6104	-3623.27404
3	0.7563	0.7555	3.5415	-4810.1968	-4791.08170
3	0.7562	0.7553	3.9224	-4809.8141	-4790.69896
3	0.7464	0.7455	39.0751	-4775.1885	-4756.07337
3	0.7392	0.7383	64.8002	-4750.6866	-4731.57137
4	0.7564	0.7553	5.0000	-4808.7412	-4784.84732

Number in Model	R-Square	Variables in Model
1	0.7463	cm10_dif
1	0.7379	cm30_dif
1	0.0625	ff_dif
1	0.0320	prime_dif
2	0.7561	cm10_dif cm30_dif
2	0.7463	cm10_dif prime_dif
2	0.7463	cm10_dif ff_dif
2	0.7391	cm30_dif ff_dif
2	0.7379	cm30_dif prime_dif
2	0.0727	ff_dif prime_dif
3	0.7563	cm10_dif cm30_dif ff_dif
3	0.7562	cm10_dif cm30_dif prime_dif
3	0.7464	cm10_dif ff_dif prime_dif
3	0.7392	cm30_dif ff_dif prime_dif
4	0.7564	cm10_dif cm30_dif ff_dif prime_dif

The best model according to adjusted R^2, C_p, AIC and BIC (SBC) uses two regressors cm10_dif and cm30_dif

There are 3 models competing, the 2-regressor model (cm10_dif, cm30_dif) and two 3-regressor models (cm10_dif, cm30_dif, ff_dif) (cm10_dif, cm30_dif, prime_dif) that are clearly above the rest of the models, and interestingly enough all contain cm10_dif and cm30_dif

1.7.4 Collinearity and variance inflation

When two or more regressors are highly correlated with each other, then it is difficult to separate their effects on the response. For example, cm10_dif and cm30_dif have correlation of 0.96! If you regress aaa_dif on cm10_dif only then the adj R^2 is 0.746, on cm30_dif only 0.738, and on both cm10_dif and cm30_dif 0.7556. So they are both related to aaa_dif, but using both does not increase adj R^2 too much. Why? The problem here is that cm10_dif and cm30_dif provide redundant information, because of their high correlation. This is the multicollinearity problem (see Assumption 3.) Collinearity increases standard errors

The variance inflation factor (VIF) of a regressor X_ℓ tells us how much the variance of b_ℓ is increased by having the other regressors in the model.

More precisely, suppose you have regressors $X_1, \ldots, X_k$. Then the VIF of X_ℓ is found as follows. Regress X_ℓ on the $k-1$ other regressors $X_1, X_2, \ldots, X_{\ell-1}, X_{\ell+1}, \ldots, X_k$, and let R_ℓ^2 be the R^2 of this regression. So, R_ℓ^2 measures how well X_ℓ can be predicted from the other Xs. The VIF of X_ℓ is

$$\boxed{VIF_\ell = \frac{1}{1-R_\ell^2}}$$

A value of R_ℓ^2 close to 1 implies a large VIF

The more accurately that X_ℓ can be predicted from the other Xs, the more redundant it is and the higher its VIF. The minimum value of VIF_ℓ is 1, when $R_\ell^2 = 0$. VIF_ℓ becomes infinite as $R_\ell^2 \to 1$

In our example on page 33, both VIFs are 14.03581 (they have to be equal by the definition since there are only two regressors). This means that their standard errors are increased by a factor of $3.75 = \sqrt{14.03581}$, once one is added to the model already containing the other.

For the model that uses all four regressors mentioned on page 51, one gets the following VIFs:

VIFs for cm10_dif and cm30_dif are not changed too much by the inclusion of ff_dif and prime_dif, since cm10_dif and cm30_dif are not correlated much with ff_dif and prime_dif. In addition, prime_dif and ff_dif are not highly correlated, hence the low VIFs.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	-0.00010103	0.00218	-0.05	0.9631	0.00148
cm10_dif	1	0.35510	0.04517	7.86	<.0001	11.20586
cm30_dif	1	0.30093	0.05010	6.01	<.0001	0.14781
ff_dif	1	0.00531	0.00553	0.96	0.3371	0.00254
prime_dif	1	-0.00788	0.01071	-0.74	0.4620	0.00227

Parameter Estimates

Variable	DF	Type II SS	Variance Inflation
Intercept	1	0.00000897	0
cm10_dif	1	0.25860	14.41205
cm30_dif	1	0.15096	14.15236
ff_dif	1	0.00386	1.19941
prime_dif	1	0.00227	1.14743

Practical recipe to get rid of multicollinearity is centering the data, that is, if $X_{1i}, X_{2i}, \ldots, X_{ni}$ are the values of the i^{th} regressor, then use instead $X_{1i} - \overline{X_i}, X_{2i} - \overline{X_i}, \ldots, X_{ni} - \overline{X_i}$, where $\overline{X_i} = \frac{1}{n}\sum_{j=1}^{n} X_{ji}$ is the sample mean of the i^{th} regressor.

(In Homework 2, Problem 3, you'll see an example how centering reduces collinearity)

From Ruppert ([3] on syllabus) 1.7.5 Influence diagnostics and residuals

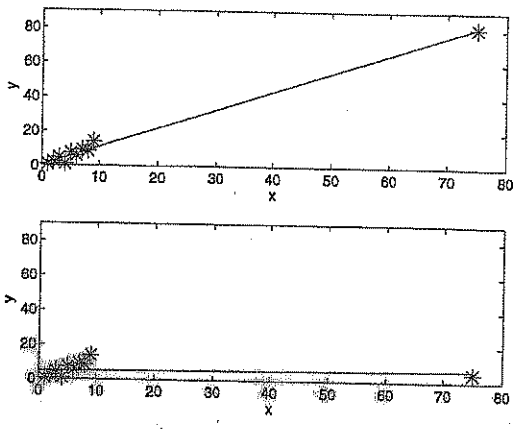


Fig. 6.4. *Linear regression example with a high leverage point.* **Top:** *no residual outliers.* **Bottom:** *residual outlier at high leverage point.*

Leverages

The *leverage* of the ith observations, denoted by H_{ii}, is a measure of how much influence Y_i has on its own fitted value $\widehat{Y}_i$. We do not go into the algebraic details. The end result is that there are weights H_{ij} depending on the values of the predictor variables but *not* on $Y_1, \ldots, Y_n$ such that

$$\widehat{Y}_i = \sum_{j=1}^{n} H_{ij} Y_j. \qquad \circledast$$

In other words, H_{ii} is the weight of Y_i in the determination of $\widehat{Y}_i$. It is bad if H_{ii} is large since that means that $\widehat{Y}_i$ is too much determined by Y_i itself with the result that the residual $\widehat{\epsilon}_i = Y_i - \widehat{Y}_i$ will be small and not a good estimate of ϵ_i. Also, the standard error of $\widehat{Y}_i$ is $\sigma\sqrt{H_{ii}}$, so a high value of H_{ii} means a fitted value with low accuracy. The leverage value H_{ii} is large when the predictor variables for the ith case are atypical of those values in the data, e.g., because one of the predictor variables for that case is extremely outlying. It can be shown by some elegant algebra that the average of $H_{11}, \ldots, H_{nn}$ is $(p+1)/n$, where $p+1$ is the number of parameters (one intercept and p slopes). A value of H_{ii} exceeding $2(p+1)/n$, that is, over twice the average value, is generally considered to be too large and therefore a cause for concern (Belsley, Kuh, and Welsch, 1980). In Figure 6.4, the high leverage point on the right has a leverage of 0.984 and $2(p+1)/n = (2)(2)/11 = 0.364$, so the leverage of this point is extreme. The H_{ii}s are sometimes called the *hat diagonals*.

Homework 2 Problem 6

Example 6.11. The output statement in the program of Example 6.3 caused the leverages to be included in the output data set and named `Leverage`. The

Note:

page 6. of Notes

$\hat{y} = Xb = X(X'X)^{-1}X'y$

i.e. $\hat{y} = Py$

$n \times n$ projection matrix

So,

$\hat{y}_i = H_{ii}y_i + \sum_{j \neq i} H_{ij}y_j$ $\circledast$

where $H_{ij} = x_i'(X'X)^{-1}x_j$

is the (i,j)th entry of P.

Note:

$\operatorname{tr}(P) = \sum_{i=1}^{n} H_{ii}$

$=$ # of regressors

(see Step 2. on page 13. of Notes)

PROC GPLOT step plotted the leverages against the case number (id). For convenience, these statements are printed again here.

```
proc reg ;
model aaa_dif = cm10_dif cm30_dif / ssi ss2 vif ;
output out=WeeklyInterest predicted=predicted rstudent=rstudent cookd=cookd h=leverage ;
run ;
proc gplot ;
plot rstudent*predicted ;
plot (rstudent cookd leverage cm10_dif cm30_dif)*id ;
plot cm10_dif*cm30_dif ;
run ;
```

In general, h=variable-name outputs the leverages and names them.

Residuals ↑ or e_i in our notation!

The *raw residual* is just $\widehat{\epsilon}_i = Y_i - \widehat{Y}_i$. The size of the raw residuals depends on σ_ϵ so we do not know how large a residual to consider unusually large. This problem is to some extent solved by using *standardized residuals*. The ith standardized residual is $\widehat{\epsilon}_i/s$, where s is the estimate of σ_ϵ. Under ideal circumstances such as a reasonably large sample and no outliers or high leverage points, the standardized residuals are approximately $N(0,1)$, so absolute values greater than 2 are outlying and greater than 3 are extremely outlying. However, circumstances are often not ideal. In the bottom plot of Figure 6.4, the standardized residual of the residual outlier/high leverage point is -0.36, not at all outlying. One problem with standardization is that the residuals do not have the same standard errors. They have all been standardized by the same value, s, but they should be standardized by their standard errors. The standard error of $\widehat{\epsilon}_i$ is $s\sqrt{1-H_{ii}}$. The *studentized residual*,[11] sometimes called the *internally studentized residual*, is $\widehat{\epsilon}_i$ divided by its standard error; that is, $\widehat{\epsilon}_i/(s\sqrt{1-H_{ii}})$. There is still one problem with studentized residuals. An extreme residual outlier can inflate s causing its studentized residual to appear too small. The solution is to redefine the ith studentized residual with an estimate of σ_ϵ that does not use the ith data point. Thus, the *externally studentized residual*, often denoted by *RSTUDENT*, is defined to be $\widehat{\epsilon}_i/\{s_{(-i)}\sqrt{1-H_{ii}}\}$ where $s_{(-i)}$ is the estimate of σ_ϵ computed by fitting the model to the data with the ith observation deleted.[12]

In the bottom plot of Figure 6.4, the high leverage point has a standardized residual of -0.36, an internally studentized residual of -2.79, and an RSTUDENT of -7.18. A rule of thumb is that a standardized or studentized residual is outlying if its absolute value exceeds 2 and extremely outlying if it exceeds 3. Thus, we see that the standardized residual completely failed to indicate a problem, the internally studentized residual did show a problem but did not indicate just how extreme the problem was, while the RSTUDENT

[11] *Studentization* means dividing a statistic by its standard error.
[12] The notation $(-i)$ signifies the deletion of the ith observation.

Note: We know (Homework 2, Problem 6) that $e = M\varepsilon$, so

$Cov(e|X) = Cov(M\varepsilon|X)$

$= M\,Cov(\varepsilon|X)M =$ ↑ M symmetric

$= M\sigma^2 I_{n\times n} M =$

$= \sigma^2 M^2 = \sigma^2 M$ ↑ M idempotent

where M is the annihilator matrix (see page 6 of Notes)

So, since $M = I - P$, using the notation from previous page (Note), we get:

$Var(e_i) = \sigma^2(1-H_{ii})$

$Cov(e_i, e_j) = -\sigma^2 H_{ij}, i \neq j$

↓ standard error of e_i is $s\sqrt{1-H_{ii}}$

Note: We denote these by $e_i' = \dfrac{e_i}{s\sqrt{1-H_{ii}}}$

Note: We denote RSTUDENT by $e_{(-i)} = \dfrac{e_i}{s_{(-i)}\sqrt{1-H_{ii}}}$

The cool relation between $e_{(-i)}$ and e_i' is: (Homework 2 Problem 11)

$$e_{(-i)} = \frac{e_i'}{\sqrt{\dfrac{n-k-(e_i')^2}{n-k-1}}}$$

, so no need to refit the model each time that an observation is omitted.

both revealed the problem and indicated its true magnitude. I recommend that RSTUDENT be used for all diagnostics.

Cook's D

A high leverage value or a large absolute value of RSTUDENT indicates a potential problem with a data point, but not how much influence a data point has on the estimates. *Cook's distance*, often called *Cook's D*, tells us how much the fitted values change if the ith observations is deleted. Let $\widehat{Y}_j(-i)$ be the jth fitted value using estimates of the $\widehat{\beta}$s obtained with the ith observation deleted. Then Cook's D is

$$D_i = \frac{\sum_{j=1}^{n}\{\widehat{Y}_j - \widehat{Y}_j(-i)\}^2}{(p+1)s^2}. \qquad (6.17)$$

The numerator in (6.17) is the sum of squared changes in the fitted values when the ith observation is deleted. A large value of this sum indicates a large influence of the ith case on the $\widehat{\beta}$s. The denominator standardizes this sum by dividing by the number of estimated parameters and an estimate of σ_ϵ^2.

The easiest way to use Cook's D is to plot the values of Cook's D against case number and look for unusually large values. See Example 6.11 for an illustration of plotting Cook's D.

Example 6.12. Figure 6.5 shows the values of Cook's D in the simulated data example in Figure 6.4. Clearly, Cook's D for the high leverage point is very large both when that point is a residual outlier and when it is not, but Cook's D is even more extreme in the latter case. In both cases, the estimated slope is almost entirely determined by this one observation, and that is worrisome because the X value of this data point is extreme. There is no way of knowing for sure that the linear regression model holds all the way out to $X = 80$ even if it is a good fit for the other data that are between $X = 0$ and $X = 9$.

Example 6.13. It was mentioned in Example 6.3 that there were missing values of cm30 at the beginning of the data set that were coded as zeros. In fact, there were 371 weeks of missing data for cm30. I started to analyze the data without realizing this problem. This created a huge outlying value of cm30_dif at observation number 372 when cm30 jumps from 0 to the first nonmissing value. Fortunately, plots of RSTUDENT, leverages, and Cook's D all reveal a serious problem somewhere between the 300th and 400th observation; see Figure 6.6. The nature of the problem is not evident from these plots, so I plotted each of the series aaa, cm10, and cm30. When I saw the initial zero values of the latter series, the problem was then obvious. Please remember this lesson: *ALWAYS look at the data.*

Note. Here $p+1 = k$ the total # of regressors (including the constant)

It can be shown that Cook's D for the i^{th} observation is also given by

$$D_i = \frac{(e_i')^2}{k}\,\frac{H_{ii}}{1-H_{ii}}$$

Note: Observations with $D_i \geq \frac{4}{n}$ are considered large!

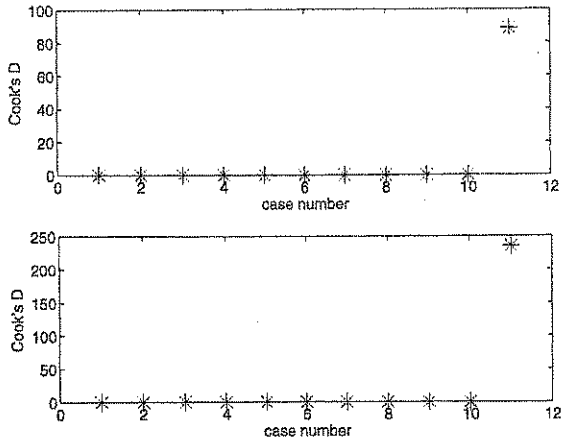


Fig. 6.5. *Cook's D in the linear regression example with a high leverage point.* ***Top:*** *no residual outliers.* ***Bottom:*** *residual outlier at high leverage point. Notice the very different vertical scales on the top and bottom. Cook's D of the leverage point is much larger in the bottom plot.*

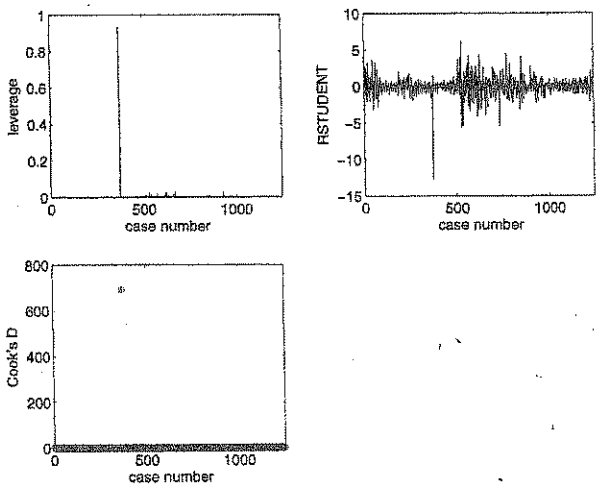


Fig. 6.6. *Weekly interest data. Full data set including the first* 371 *weeks of data where* cm30 *was missing and assigned a value of* 0*. This caused severe problems at case number* 372 *which are detected by Cook's D, the leverages, and RSTUDENT*

6.12.2 Residual analysis

Because the ith residual $\widehat{\epsilon}_i$ estimates the "noise" ϵ_i, the residuals can be used to check the assumptions behind regression. Residual analysis generally consists of various plots of the residuals, each plot being designed to check one or more of the regression assumptions. Regression software will output the several types of residuals discussed in Section 6.12.1. I recommend using RSTUDENT.

Problems to look for include:

- Nonnormality of the residuals;
- Nonconstant variance of the residuals;
- Correlation of the residuals; and
- Nonlinearity of the effects of the predictor variables on the response.

Nonnormality

Nonnormality can be detected by a normal probability plot, boxplot, and histogram of the residuals. Not all three are needed, but I strongly recommend looking at a normal plot. Moreover, inexperienced data analysts have trouble with the interpretation of normal plots. Looking at a normal plot and a histogram, side by side, is very helpful when learning to use normal probability plots.

The residuals often appear nonnormal because there is an excess of outliers relative to the normal distribution. We have defined a value of RSTUDENT to be outlying if the absolute value of the raw residual exceeds 2 and extremely outlying if it exceeds 3. Of course these cutoffs of 2 and 3 are arbitrary and only intended to give rough guidelines.

It is the presence of outliers, particularly extreme outliers, that is a concern when we have nonnormality. A deficiency of outliers relative to the normal distribution is not a problem. Sometimes outliers are due to errors, such as mistakes in the entry of the data or, as in Example 6.13, misinterpreting a zero as a true data value rather than the indicator of a missing value. If possible, outliers due to mistakes should be corrected, of course. However, in financial time series, outliers are often "good observations" due to excess volatility in the markets on certain days. An excess of only positive outlying residuals occurs when the response is right skewed and similarly left skewness is associated with an excess of negative outlying residuals. If the residuals appear to be roughly symmetric with an excess of both negative and positive outlying residuals, then the noise distribution, that is, the distribution of $\epsilon_1, \ldots, \epsilon_n$, has **heavy tails**.

Another possible reason for an excess of both positive and negative outlying residuals in a normal probability plot is nonconstant residual variance, a problem which is explained shortly. Normal probability plots assume that all observations come from the same distribution, in particular, that they have

the same variance. The purpose of that plot is to determine if the common distribution is normal or not. If there is not a common distribution, for example, because of nonconstant variance, then the normal plot is not readily interpretable. Therefore, one should check for a constant variance before attempting to interpret a normal plot.

Outliers can be a problem because they have an unduly large influence on the estimation results. A common solution to the problem of outliers is transformation of the response. It is always wise to check whether outliers are due to erroneous data, e.g., typing errors or other mistakes in data collection and entry. Of course, errors should be corrected if possible, but if this is not possible then erroneous data should be removed. Removal of outliers that are not known to be erroneous is dangerous and not recommended as routine statistical practice. However, reanalyzing the data with outliers removed is a sound practice. If the analysis changes drastically when the outliers are deleted, then one knows there is something about which to worry. On the other hand, if deletion of the outliers does not change the conclusions of the analysis then there is less reason to be concerned with whether the outliers were erroneous data.

Nonconstant variance

Nonconstant residual variance means that the conditional variance of the response given the predictor variables is not constant as assumed by standard regression models. Nonconstant variance is also called **heteroscedasticity**. Nonconstant variance can be detected by an absolute residual plot, that is, by plotting the absolute residuals against the predicted values ($\widehat{Y}_i$s) and, perhaps, also against the predictor variables. If the absolute residuals show a systematic trend, then this is an indication of nonconstant variance. Economic data often have the property that bigger responses are more variable. A more technical way of stating this is that the conditional variance of the response (given the predictor variables) is an increasing function of the conditional mean of the response. This type of behavior can be detected by plotting the absolute residuals versus the predicted values and looking for an increasing trend.

Often, trends are difficult to detect just by looking at the plotted points and adding a so-called *scatterplot smoother* is very helpful. A scatterplot smoother fits a smooth curve to a scatterplot. Nonparametric regression estimators such as loess, lowess,[13] and smoothing splines are commonly used scatterplot smoothers available in statistical software packages.

A potentially serious problem caused by nonconstant variance is inefficiency, that is, too-variable estimates. Here is a very simple example to illustrate the problem of inefficiency. We assume the simplest possible regression model with only an intercept and no predictor variables; that is, $Y_i = \beta_0 + \epsilon_i$. Not surprisingly, the estimate of the intercept is the sample

[13] Loess and lowess are different, but closely related, algorithms.

mean of the responses. Assume Y_1, Y_2, Y_3 are independent, all with mean β_0 and $\text{Var}(Y_1) = \text{Var}(Y_2) = 1$ and $\text{Var}(Y_3) = 10$. If we use all the data then

$$\text{Var}(\widehat{\beta}_0) = \text{Var}\{(Y_1 + Y_2 + Y_3)/3\} = (1 + 1 + 10)/9 = 12/9 = \frac{4}{3}.$$

If we delete Y_3 because it is less accurate than the other observations, then

$$\text{Var}(\widehat{\beta}_0) = \text{Var}\{(Y_1 + Y_2)/2\} = (1 + 1)/4 = \frac{1}{2} < \frac{4}{3}.$$

Our estimator is improved by deleting the inaccurate data point. However, deleting Y_3 is *not* the best thing we can do here. Although Y_3 contains less information about β_0 than either Y_1 or Y_2, it still contains some information. The most efficient (least variable) estimate is obtaining by *downweighting* Y_3 rather than deleting it altogether. The optimal weights are the reciprocal variances. Using these weights give us

$$\text{Var}(\widehat{\beta}_0) = \text{Var}\{(Y_1 + Y_2 + 0.1Y_3)/2.1\} = \{1 + 1 + (.1)^2 10\}/(2.1)^2 = \frac{1}{2.1} < \frac{1}{2}.$$

Another serious problem is that standard errors and confidence intervals assume a constant variance and can be seriously wrong if there is substantial nonconstant variance.

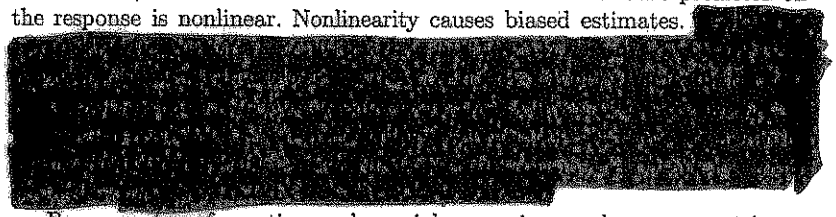
Transformation of the response and weighting are common solutions to the problem of nonconstant variance. Response transformations are presented in Section 6.13. Weighted least squares estimates $\boldsymbol{\beta}$ by minimizing

$$\sum_{i=1}^{n} w_i \{Y_i - f(\boldsymbol{X}_i; \widehat{\boldsymbol{\beta}})\}^2, \tag{6.18}$$

with w_i an estimate of the inverse (i.e., reciprocal) conditional variance of Y_i given $\boldsymbol{X}_i$. Estimation of the conditional variance function to determine the w_is is discussed in more advanced textbooks such as Carroll and Ruppert (1988).

Nonlinearity

If a plot of the residuals versus a predictor variable shows a systematic nonlinear trend, then this is an indication that the effect of that predictor on the response is nonlinear. Nonlinearity causes biased estimates.



Response transformation, polynomial regression, and nonparametric regression (splines, loess) are common solutions to the problem of nonlinearity when it exists.

Residual correlation

Note: You will learn about SACF and ARIMA next semester

If the data (X_i, Y_i) are a multivariate times series, then it is likely that the ϵ_i noises are correlated. This problem can be detected by looking at the SACF of the residuals. Similarly, in SAS, one can save the residuals in an output SAS data set, and then run, for example, PROC ARIMA on them. In PROC ARIMA, one can run the "identify" statement without an "estimate" statement if one wants only the SACF without the fitting of an ARMA model. Autocorrelation of the residuals is called *serial correlation* by some authors.

Residual correlation causes standard errors and confidence intervals to be incorrect. In particular, the coverage probability of confidence intervals can be much less than the nominal value. A solution to this problem is to model the ϵ_i as an AR process. This can be done using SAS's PROC AUTOREG.

Example 6.14. Data were simulated to get an example illustrating many of the ways to diagnose problems. In the example there are two predictor variables, X_1 and X_2. The assumed model is multiple linear regression.

Figure 6.7, which shows the responses plotted against each of the predictors, does not indicate any problems with this example, except the heteroscedasticity is suggested because the data are more scattered on the right sides of the plots. The point is that plots of the raw data often fail to reveal problems. Rather, it is plots of the residuals that can more reliably detect heteroscedasticity, nonnormality, and other difficulties.

Figure 6.8 is a normal plot and a histogram of the residuals. Notice the right skewness which suggests that a response transformation to remove right skewness, e.g., a square root or log transformation, should be investigated.

Figure 6.9 is a plot of the residuals versus X_1. The residuals appear to have a nonlinear trend. This is better revealed by adding a spline curve fit to the residuals. The curvature of the spline is evident. This pattern suggests that Y is not linear in X_1. A possible remedy is to add X_1^2 as a third predictor so that y is modeled as quadratic in X_1. Figure 6.10, a plot of the residuals against X_2, shows somewhat random scatter, indicating that Y appears to be linear in X_2. The spline fit does dip downward at the right side of Figure 6.10, but spline estimates are highly variable near the boundaries of the data and this feature is likely to be just random variation.

Figure 6.11 is a plot of the absolute residuals versus the predicted values. Note that the absolute residuals are largest where the fitted values are also largest, which is a clear sign of heteroscedasticity. A spline smooth has been added to make the heteroscedasticity clearer.

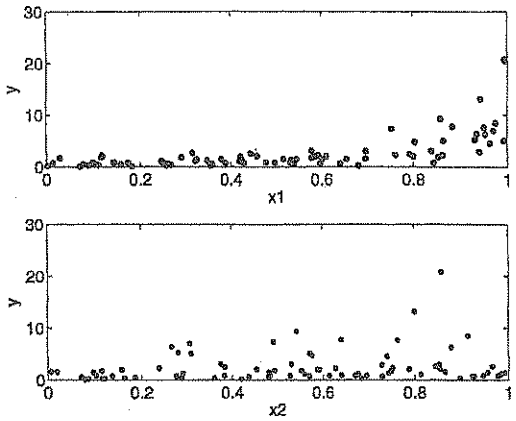


Fig. 6.7. *Simulated data. Responses plotted against the two predictor variables.*

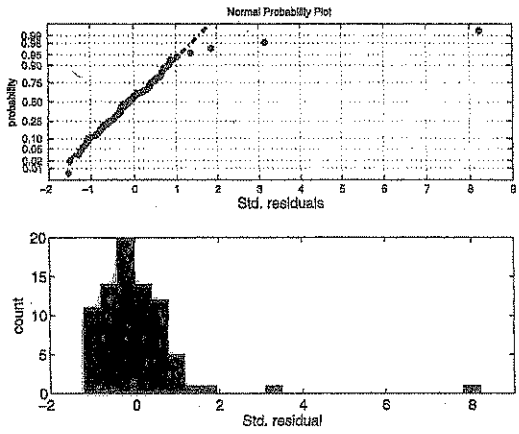


Fig. 6.8. *Simulated data. Normal plot and histogram of the studentized residuals. Right skewness is enough and perhaps a square root or log transformation of Y would be helpful.*

Example 6.15. (Estimating Default Probabilities) This example illustrates both nonlinear regression and the detection of heteroscedasticity by residual plotting.

Credit risk is the risk to a lender that a borrower will default on contractual obligations, in short, that a loan will not be repaid in full. A key parameter in the determination of credit risk is the probability of default. Bluhm, Overbeck,

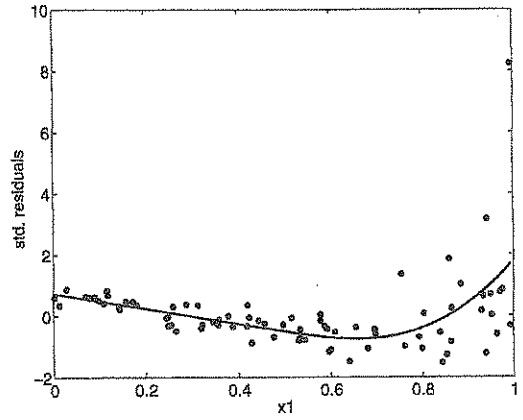


Fig. 6.9. *Simulated data. Plot of studentized residuals versus the X_1 with a spline smooth. This plot suggests that Y is not linearly related to X_1 and perhaps a model quadratic in X_1 is needed.*

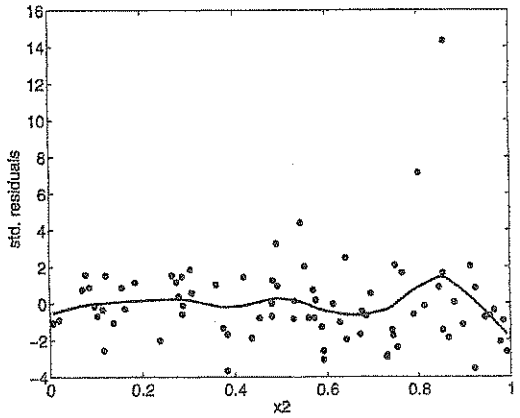


Fig. 6.10. *Simulated data. Plot of studentized residuals versus the X_2 with a spline smooth. This plot suggests that Y is linearly related to X_2 so that the component of the model relating Y to X_2 is satisfactory.*

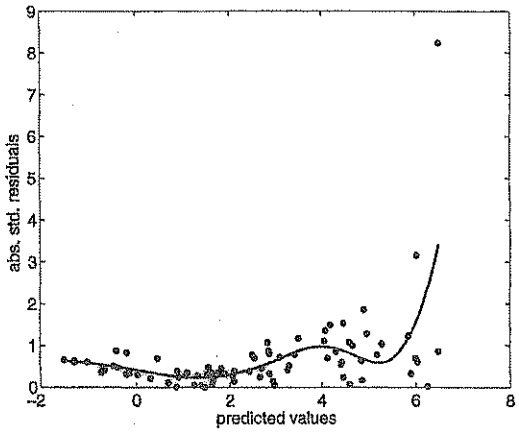


Fig. 6.11. *Simulated data. Plot of the absolute studentized residuals versus the predicted values with a spline smooth. This plot reveals heteroscedasticity.*

and Wagner (2003) illustrate how one can calibrate Moody's credit rating to estimate default probabilities. These authors use observed default frequencies for bonds in each of 16 Moody's ratings from Aaa (best credit rating) to B3 (worse rating). They convert the credit ratings to a 1 to 16 scale (Aaa $= 1, \ldots,$ B3 $= 16$). Figure 6.12 shows default frequencies plotted against the ratings. The data are from Bluhm, Overbeck, and Wagner (2003). The relationship is clearly nonlinear and the linear regression fit labelled "linear" in the figure has two obvious and serious problems. First, it does not follow the data closely, and, second, it gives negative estimated default probabilities at the better ratings. Unsurprisingly, Bluhm, Overbeck, and Wagner do not consider a linear fit to the default frequencies. Instead they assume that the default probability is an exponential function of the rating; that is,

$$Pr\{\text{default}|\text{rating}\} = \exp\{\beta_0 + \beta_1 \text{rating}\}. \tag{6.19}$$

To use this model they fit a linear function to the logarithms of the default frequencies. One difficulty with doing this is that many of the default frequencies are zero giving a log transformation of $-\infty$.

Bluhm, Overbeck, and Wagner address this issue by labelling default frequencies equal to zero as "unobserved" and not using them in the estimation process. The problem with their technique is that they have deleted the data with the lowest observed default frequencies. This biases their estimates of default probabilities in an upward direction. As we show, the bias is sizeable. Bluhm, Overbeck, and Wagner argue that an observed default frequency of zero does not imply that the true default probability is zero. This is certainly true. However, the default frequencies, even when they are zero, are unbiased estimates of the true default probabilities. I do not wish to seem critical of

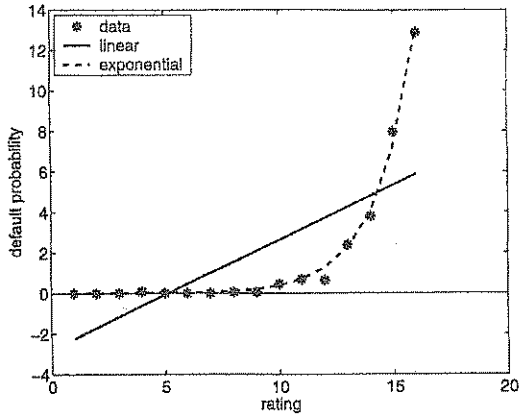


Fig. 6.12. *Default frequencies versus rating with a linear regression fit. The default frequency is expressed as a percentage. "Rating" is a conversion of the Moody's Rating to a* 1 *to* 16 *point scale as follows:* 1 = *Aaa*, 2 = *Aa*1, 3 = *Aa*3, 4 = *A*1, ..., 16 = *B*3. *Default probabilities are expressed as percentages.*

So, what can one do here? Where next?

nonlinear regression / nonparametric regression/splines (Chapter 13 in Ruppert, Chapter 7 in Lai-Xing)
Transform-Both-Sides regression (Chapter 6 in Ruppert)
Robust regression (not even in software yet!)

Residual diagnostics for Example on page 43.

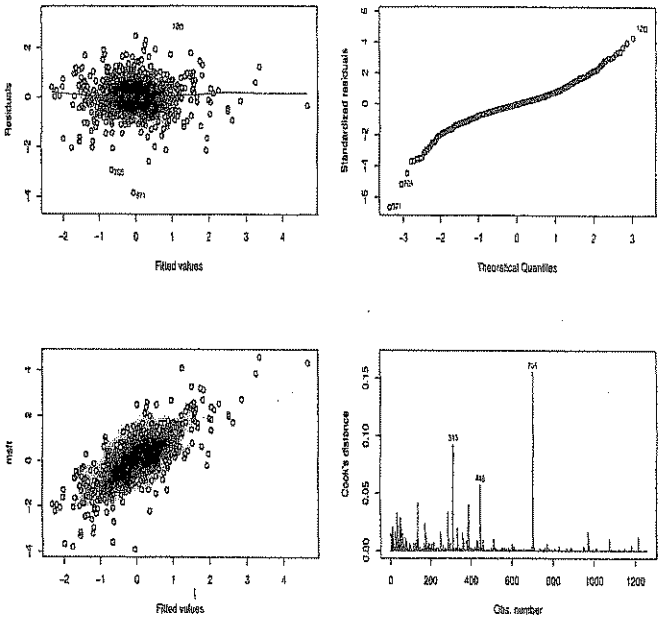


Fig. 1.6. Diagnostic plots of the fitted regression model (1.56).

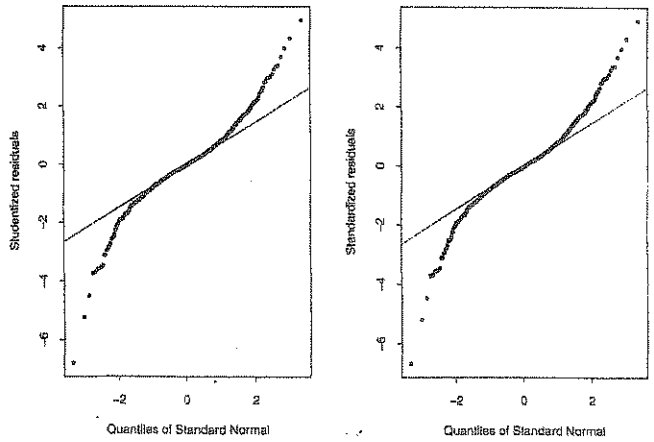


Fig. 1.7. Q-Q plots of studentized (left panel) and standardized (right panel) residuals.

Table 1.5. Regression coefficients of the selected regression model.

Term	Estimate	Std. Error	t-statistics	p-value
`Intercept`	0.0040	0.0164	0.2406	0.8099
`aapl`	0.0536	0.0159	3.3713	0.0008
`adbe`	0.0911	0.0134	6.7822	0.0000
`dell`	0.1920	0.0205	9.3789	0.0000
`gtw`	0.0437	0.0097	4.5218	0.0000
`ibm`	0.2546	0.0266	9.5698	0.0000
`orcl`	0.1315	0.0149	8.8461	0.0000

RSS = 418.20, d.f. = 1248, $s = 0.579$, adjusted $R^2 = 56.6\%$.

bottom panels plot (i) the `msft` log returns versus fitted values and (ii) Cook's distance for each observation, which identifies the 313th (April 8, 2002) and the 705th (October 23, 2003) observations as outliers that have substantial influence on the estimated regression coefficients.

Figure 1.7 replaces the residuals in the Q-Q plot (the left panel in the second row) of Figure 1.6 by studentized and standardized residuals, respectively. The Q-Q plots of the studentized and standardized residuals are quite similar. The tails in these Q-Q plots show substantial departures from the assumption of normal errors. The nonnormality of stock returns will be discussed further in Section 6.1.

Exercises

1.1. Given observations (x_i, y_i), $i = 1, \cdots, n$, fit a quadratic polynomial regression model

$$y_i = f(x_i) + \epsilon_i, \quad \text{where } f(x) = \alpha + \beta_1 x + \beta_2 x^2.$$

Construct a 95% confidence interval for $f(x)$ at a given x.

1.2. Consider the linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \quad 1 \le i \le n.$$

Using the F-test of a general linear hypothesis, show how to test the following hypotheses on the regression coefficients: