Ratio Mathematica
Vol. 32, 2017, pp. 63-75

ISSN: 1592-7415
eISSN: 2282-8214

Teaching Least Squares in Matrix Notation

Guglielmo Monaco[1], Aniello Fedullo[2*]

[1]Department of Chemistry and Biology "A. Zambelli", University of Salerno, Italy
gmonaco@unisa.it

[2]Department of Physics "E. R. Caianiello", University of Salerno, Italy
afedullo@unisa.it

Received on: 30-05-2017. **Accepted** on: 27-06-2017. **Published** on: 30-06-2017

doi:10.23755/rm.v32i0.335

Abstract

Material for teaching least squares at the undergraduate level in matrix notation is reported. The weighted least squares equations are first derived in matrix form; equivalence with the standard results obtained by standard algebra are then given for the weighted average and the simplest linear regression. Indicators of goodness of fit are introduced and interpreted. Eventually a basic equation for resampling is derived.

Keywords: coefficient of determination, weighted sample mean, resampling, undergraduate education.
2010 AMS subject classifications: 62J05.

1 Introduction

Statistics is a never missing topic in first degree courses of scientific programs. Very soon, often at the second year undergraduate, the basic knowledge of random variables and distributions, is complemented by the simple linear regression, as a necessary tool for the interpretation of experimental data gathered in the laboratories. Indeed, the critical practice of linear regressions often forms students' basic awareness of data analysis. The advent of powerful and handy softwares on the one hand has reduced the effort required to the students for accomplishing the needed calculations, on the other hand has given them the possibility to easily perform more advanced statistical analyses [1, 2], which they cannot really understand on the grounds of the course. One of simplest of such more advanced analyses is the consideration of more regressors, the starting point of multivariate data analysis [3]. Although a specific course at the last undergraduate or first graduate year can be much profitable, we experienced that, provided the students have a basic knowledge in linear algebra, the generalized least squares can be thought at the second year undergraduate with reasonable appreciation from the class.
Reference textbooks on the matter, seemingly more diffused in the community of econometrics [5] than in that of experimental sciences [6], are not missing. However, we needed to compact some fundamental concepts and equations, and still convince the students that the more general matrix form of the least squares allows to easily retrieve the results obtainable with standard algebra. Thus, we prepared the following material, and we presented it effectively in a 12 hours module together with numerical exercises. Although our lessons obviously have a significant overlap with reference textbooks, the revised simple linear regression and the introduction of the (adjusted) weighted coefficient of determination are not easily retrieved from any of the textbooks known to us.

2 Matrix Form of the Weighted Least Squares

We consider n measures $\{y_1, y_2, ..., y_n\}$ and for each of them, say the i-th one, the regressors $\{x_{i1}, x_{i2}, ..., x_{ip}\}$, here assumed constant, which are generally coming from different associated measures. We will assume that for each measure the first regressor equals one, $x_{i1} = 1$, in order to take into account the so called intercept. The linear regression model connects the above quantities by

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \qquad i = 1, 2, ..., n \tag{1}$$

where $\beta_1, \beta_2, ..., \beta_p$ are the parameters to be estimated and $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are random errors, assumed independent and possibly normally distributed, with mean 0 and standard deviations $\sigma_1, \sigma_2, ..., \sigma_n$. Ordinary least squares (OLS) and weighted least squares (WLS), also called homoskedastic and heteroskedastic regressions, are the names used to distinguish the special case of equal values for all standard deviations from the case of different values. The equations for WLS of course also apply to the special OLS case.

Dividing eq. 1 by σ_i, i.e. given $z_i := \frac{y_i}{\sigma_i}$, $q_{ij} := \frac{x_{ij}}{\sigma_i}$, $\varsigma_i := \frac{\varepsilon_i}{\sigma_i}$, and using the matrix notation, the model is written as

$$z \;=\; Q\beta + \varsigma, \tag{2}$$

or, equivalently,

$$W^{\frac{1}{2}} y \;=\; W^{\frac{1}{2}} X\beta + W^{\frac{1}{2}}\epsilon,$$

where W is a diagonal matrix whose elements $W_{ii} := w_i = \sigma_i^{-2}$ are known as statistical weights, z and β are column matrices of n and p elements, respectively, Q is a matrix of dimension $n \times p$. It should be noticed that $Q\beta$ is the expectation value of z, i.e. $Q\beta =< z >$.

Under these hypotheses the least squares method gives an estimate of the model parameters by the minimization with respect to β of the functional

$$SS := \varsigma^T\varsigma = (z - Q\beta)^T(z - Q\beta) \tag{3}$$

$$= \;(z - Q\beta)^T(z - Q\beta) = z^Tz - 2\beta^TQ^Tz + \beta^TQ^TQ\beta, \tag{4}$$

where it has been considered that $\beta^TQ^Tz = z^TQ\beta$.

The estimates of the parameters by the least squares method are the solutions of the equations $\frac{\partial SS}{\partial \beta_i} = 0$, for $i = 1, 2, ..., p$, one for each model parameter. The computation of the derivative with respect to the vector of the parameters gives:

$$-Q^Tz + Q^TQ\beta = 0, \tag{5}$$

whose solution

$$\hat{\beta} = VQ^Tz \tag{6}$$

is, by definition, the least squares estimator of β, where $V := C^{-1}$, and $C := Q^TQ$, which we will assume always invertible.

We note that $\hat{\beta}$ is an unbiased estimator of β, indeed form eqs. 2 and 6 we have

$$< \hat{\beta} >= VQ^T < z >= VQ^TQ\beta = VC\beta = \beta. \tag{7}$$

An unbiased behavior also characterizes the weighted sample mean. Indeed, eq. 5 for $\beta = \hat{\beta}$ gives $Q^Tz = Q^T\hat{z}$ which, rewritten in the original variables, is $X^TWy = X^TW\hat{y}$. From this and from the initial hypothesis $x_{i1} = 1$, for any i, one gets $\sum_i w_i y_i = \sum_i w_i \hat{y}_i$, which divided by $\sum_i w_i$ shows that the weighted sample mean of the fitted values equals the weighted sample mean of the measures:

$$\bar{y}_w = \overline{\hat{y}}_w. \tag{8}$$

Given $\delta := \hat{\beta} - \beta$ from eqs. 6 and 7 one gets

$$\delta \;\; = \;\; VQ^T\varsigma, \tag{9}$$

which allows to easily compute the covariance matrix of the parameters, showing that it coincides with V

$$< \delta\delta^T >= VQ^T < \varsigma\varsigma^T > QV = VQ^TIQV = V,$$

where I denotes the identity matrix.

The standard deviations of the estimators of the parameters are given by the square roots of the diagonal elements of V.

Using the fitted values, one can write

$$z = \hat{z} + (z - \hat{z}) = Q\hat{\beta} + e,$$

where e is known as the vector of residuals, whose analysis is object of much concern in literature.

The fitted values are often written as

$$\hat{z} = Q\hat{\beta} = QVQ^Tz =: Hz, \tag{10}$$

where we have introduced the symmetric matrix H, which is known as hat matrix as it 'puts the hat on z'. This matrix is readily verified to be idempotent, $H^2 = QVQ^TQVQ^T = H$, a feature which readily allows to demonstrate the useful property of orthogonality of residuals and fitted values:

$$(z - \hat{z})^T\hat{z} = z^T(I - H)Hz = 0.$$

Given

$$SSE := min_\beta SS = e^T e,$$

the expansion 4, with ς in place of z and δ in place of β, can be rewritten as:

$$SSE = (\varsigma - Q\delta)^T(\varsigma - Q\delta) = \varsigma^T\varsigma - 2\delta^T Q^T\varsigma + \delta^T C\delta = \varsigma^T\varsigma - \delta^T C\delta,$$

where we have considered that $Q^T\varsigma = C\delta$ thanks to eq. 9.

Given $SSR := \delta^T C\delta$, which as SS and SSE is non-negative, the preceding equation becomes

$$SSE = SS - SSR$$

whose interpretation is that the error in the estimation of the parameters, yielding a nonzero SSR, reduces the sum of squares SS which could have been computed with the expectation value $\langle z\rangle = Q\beta$.

The average of SSE can be easily computed considering that $\delta^T C\delta = Tr\left[\delta\delta^T C\right]$, and then

$$\langle SSE\rangle \;=\; \left\langle\varsigma^T\varsigma\right\rangle - Tr\left[\left\langle\delta\delta^T\right\rangle C\right] = n - Tr\,(VC) = n - p,$$

known as the number of degrees of freedom, denoted by ν.

Notation. In the following $s_{xx,w}$, $S_{xx,w}$ e $s_{xy,w}$ indicate respectively the sample variance, sum of squares and weighted covariance, defined from the weighted sample mean $\bar{y}_w := \frac{\sum_i w_i y_i}{\sum_i w_i}$ in analogous manner to the corresponding unweighted means. We recall that their expressions are $s_{xx,w} = \overline{x^2_w} - \bar{x}^2_w$, $S_{xx,w} = s_{xx,w}\sum_i w_i$ e $s_{xy,w} = \overline{xy}_w - \bar{x}_w\bar{y}_w$, where $xy := (x_1y_1, ..., x_ny_n)$.

3 Indicators for the Goodness of Fit

Besides reporting the best-fit parameters and the resulting fitted values, it is customary to give compact indicators of the goodness of fit.

A method which is widely used in the analysis of experimental data consists in the chi-squared test: the hypothesis that the model is correct is not rejected, at the appropriate level of significance, if SSE assumes values close to $\langle SSE\rangle$, i.e., for any number of parameters, if $\chi^2_r = \frac{SSE}{\nu}$ is close to 1. Values of χ^2_r larger or smaller than 1 are then considered as indicators of a poor fit or, respectively, overfitting.

A different approach considers weighted sample means. Defining the *weighted coefficient of determination* R^2_w as the square of the weighted sample correlation

coefficient $\frac{s_{y\hat{y},w}}{\sqrt{s_{yy,w}s_{\hat{y}\hat{y},w}}}$ between data y and fitted values $\hat{y} = X\hat{\beta}$ and thus limited by $0 \leq R_w^2 \leq 1$, one has that

$$1 - R_w^2 = \frac{SSE}{S_{yy,w}} = \frac{s_{ee}}{s_{yy,w}}, \tag{11}$$

showing that $R_w^2 = 1$ iff $SSE = 0$, i.e. iff all residuals are zero. Therefore the greater the value of R_w^2 the better the agreement. Eq. 11 can be proven thanks to the orthogonality relation discussed above. The vector $\overline{y_w}w^{\frac{1}{2}}$, where $w^{\frac{1}{2}}$ is a column vector of elements $w_i^{\frac{1}{2}}$, is orthogonal to the vector of residuals $z - \hat{z}$, by virtue of eq. 8. Therefore the orthogonality of residuals and fitted values, eq. 2, still holds if the fitted values are translated by $\overline{y_w}w^{\frac{1}{2}}$. The vector relationship

$$z - \overline{y_w}w^{\frac{1}{2}} = (\hat{z} - \overline{y_w}w^{\frac{1}{2}}) + (z - \hat{z}), \tag{12}$$

graphically sketched in Figure 1, allows to assess that

$$S_{yy,w} = S_{\hat{y}\hat{y},w} + SSE, \tag{13}$$

whose interpretation is that $S_{\hat{y}\hat{y},w}/S_{yy,w}$ is the fraction of variability of the data explained by the knowledge of Q, i.e. by the regression, and $SSE/S_{yy,w}$ is the unexplained one, i.e. that coming from errors.

Still from eq. 12 one gets

$$S_{\hat{y}y,w} = (\hat{z} - \overline{\hat{y}_w}w^{\frac{1}{2}})^T(z - \overline{y_w}w^{\frac{1}{2}}) = (\hat{z} - \overline{\hat{y}_w}w^{\frac{1}{2}})^T(\hat{z} - \overline{\hat{y}_w}w^{\frac{1}{2}}) = S_{\hat{y}\hat{y},w} \tag{14}$$

and then

$$R_w^2 = \frac{S_{\hat{y}y,w}^2}{S_{\hat{y}\hat{y},w}S_{yy,w}} = \frac{S_{\hat{y}\hat{y},w}}{S_{yy,w}}. \tag{15}$$

Insertion of eq. 15 in eq. 13 readily gives eq. 11.

In order to discourage the introduction of models too complicated for the data examined, it has been introduced the adjusted determination coefficient

$$R_a^2 = 1 - (1 - R_w^2)\frac{n-1}{n-p},$$

obtained substituting the unbiased variances in the rhs of eq. 11.

It often happens that standard deviations of experimental data are only approximately known. A common assumption is that the standard deviations σ_i are known but for a factor k: $\sigma_i = k\tilde{\sigma}_i$, with the $\tilde{\sigma}_i$ known *a priori*. If the adjustment

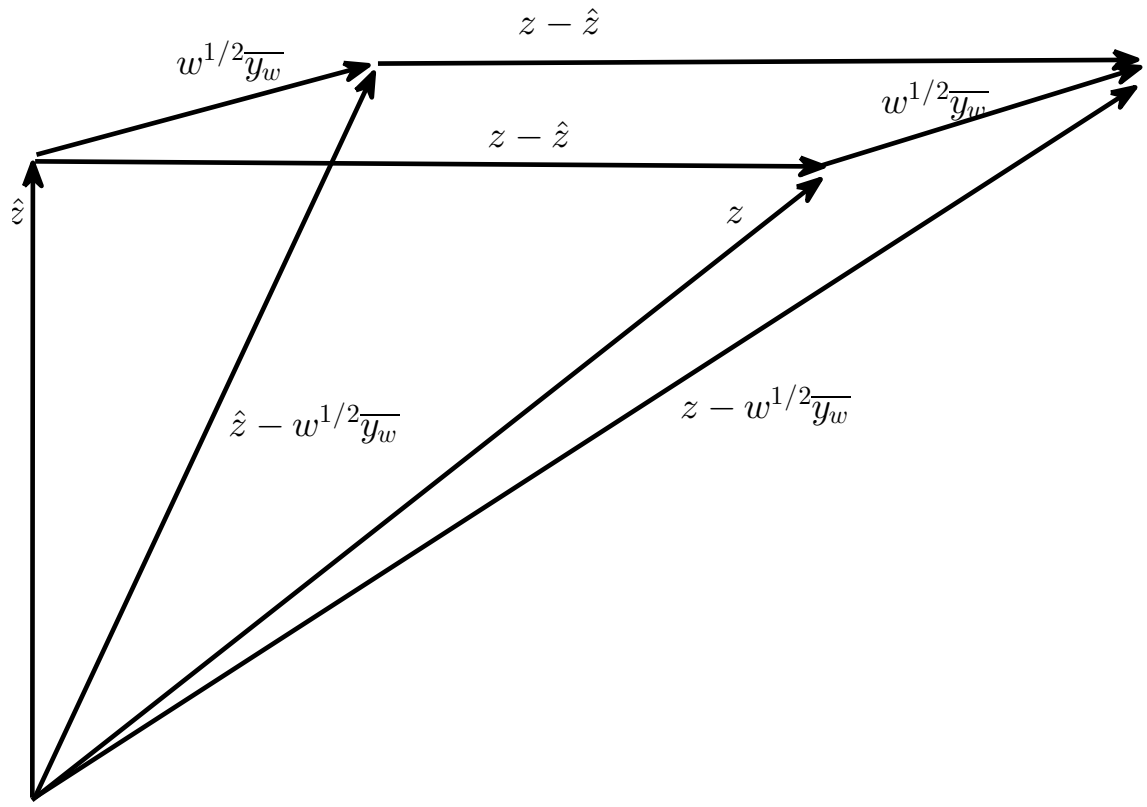

Figure 1: The residuals $z - \hat{z}$ are orthogonal to both the estimates $\hat{z}$ and the vector $\overline{y_w} w^{\frac{1}{2}}$.

of k leads to a good fitting for the model, χ^2_r should be close to ν. Using this value, one gets

$$\nu = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{k^2 \tilde{\sigma}_i^2},$$

and a trial value for k is obtained as

$$k = \sqrt{\frac{1}{\nu} \frac{\sum (y_i - \hat{y}_i)^2}{\tilde{\sigma}_i^2}}.$$

4 Basic Applications

4.1 (Weighted) mean

The model $y = \beta \mathbf{1} + \varepsilon$ has an $n \times 1$ matrix of relative regressors, whose i-th element is

$$q_{i1} = w_i^{\frac{1}{2}}$$

Application of eq. 7 soon gives as the best fit parameter the weighted mean

$$\hat{\beta} = VQz = \frac{\sum_i w_i y_i}{\sum_i w_i} = \bar{y}_w$$

and its variance is the sum of the weights: $\sigma^2_\beta = V_{11} = \sum_i w_i$.

4.2 WLS for a straight line

The standard linear regression considers the model $y = a + bx$. In the above notation $a = \beta_1$ and $b = \beta_2$ and the regressor matrix is

$$X = \begin{bmatrix} 1 & 1 & ... & 1 \\ x_1 & x_2 & ... & x_n \end{bmatrix}^T$$

The matrix of relative regressors will be then

$$Q = \begin{bmatrix} \sqrt{w_1} & \sqrt{w_2} & ... & \sqrt{w_n} \\ \sqrt{w_1}x_1 & \sqrt{w_2}x_2 & ... & \sqrt{w_n}x_n \end{bmatrix}^T,$$

the vector of relative data

$$z = \left[\begin{array}{cccc} \sqrt{w_1}y_1 & \sqrt{w_2}y_2 & \ldots & \sqrt{w_n}y_n \end{array} \right]^T,$$

and

$$C = \sum_i w_i \left[\begin{array}{cc} 1 & \bar{x}_w \\ \bar{x}_w & \overline{x_w^2} \end{array} \right],$$

whose inverse gives the covariance matrix of the parameters

$$V = \frac{1}{S_{xx,w}} \left[\begin{array}{cc} \overline{x_w^2} & -\bar{x}_w \\ -\bar{x}_w & 1 \end{array} \right].$$

The standard deviations of the estimators of the parameters will be then

$$\left[\begin{array}{c} \sigma_{\hat{a}} \\ \sigma_{\hat{b}} \end{array} \right] = \left[\begin{array}{c} \sqrt{V_{11}} \\ \sqrt{V_{22}} \end{array} \right] = \frac{1}{\sqrt{S_{xx,w}}} \left[\begin{array}{c} \sqrt{\overline{x_w^2}} \\ 1 \end{array} \right],$$

and the estimated parameters will be

$$\left[\begin{array}{c} \hat{a} \\ \hat{b} \end{array} \right] = VQ^T z = \frac{1}{s_{xx,w}} \left[\begin{array}{cc} \overline{x_w^2} & -\bar{x}_w \\ -\bar{x}_w & 1 \end{array} \right] \left[\begin{array}{c} y_w \\ \overline{xy}_w \end{array} \right] = \left[\begin{array}{c} \bar{y}_w - \frac{s_{xy,w}}{s_{xx,w}}\bar{x}_w \\ \frac{s_{xy,w}}{s_{xx,w}} \end{array} \right],$$

which in case of all equal weights (*homoskedastic regression*) have the simpler expression $\left[\begin{array}{cc} \bar{y} - \frac{s_{xy}}{s_{xx}}\bar{x} & \frac{s_{xy}}{s_{xx}} \end{array} \right]^T$.

4.3 Revised simple linear regression

We now give a simplified approach for the bivariate weighted linear regression: given $\mathbf{1} := [1\ 1\ \ldots\ 1]^T$, we subtract $\overline{\mathrm{y}}_w\mathbf{1}$ from the data and from the fitted data and, considering that $\overline{\mathrm{y}}_w = \overline{\hat{y}}_w = a + b\bar{x}_w$, we obtain

$$y - \overline{\mathrm{y}}_w\mathbf{1} = b(x - \bar{x}_w\mathbf{1}) + \varepsilon, \tag{16}$$

which, with $z := W^{\frac{1}{2}}(y - \overline{y}_w\mathbf{1})$, $q := W^{\frac{1}{2}}(x - \bar{x}_w\mathbf{1})$ e $\varsigma := W^{\frac{1}{2}}\varepsilon$, can be written as in eq. 2,

$$z = bq + \varsigma,$$

but here there is the single parameter b to be determined, as in the example of the weighted mean.

This means that matrix C is the scalar $S_{xx,w}$ readily invertible, and then $V = C^{-1} = \frac{1}{S_{xx,w}}$. On the other hand, as

$$q^T z \;=\; (x-\bar{x}_w\mathbf{1})^T W(y-\overline{\mathrm{y}}_w\mathbf{1}) = S_{xy,w} + \overline{\mathrm{y}}_w \sum_i w_i(x_i-\bar{x}_w) = S_{xy,w},$$

from eq. 6, one gets again $\hat{b} = \frac{s_{xy,w}}{s_{xx,w}}$. Writing now the model as $y - bx = a\mathbf{1} + \varepsilon$, the example in 4.1 gives for the intercept $\bar{y}_w - b\bar{x}_w$, from where, replacing b with its estimator[1], one finally gets $\hat{a} = \bar{y}_w - \hat{b}\bar{x}_w$, as in 4.2.

It is to be considered, however, that this simplification leads to loose information on the covariance of the a and b parameters, which should then be recover *ex post* (Appendix).

4.4 Resampling and the Best-fit Parameters

A remarkable representation of the p best-fit parameters can be obtained if one tries to determine them from the $\binom{n}{p}$ p-elements subsets of the original set of n measures [4]. Let $S_{(s)}$ be a $p \times n$ matrix obtained from the $n \times n$ identity matrix, upon selecting the p rows whose indices form subset s, with $s = 1, \ldots \binom{n}{p}$. Let also $M^{[k|v]}$ be the matrix obtained from matrix M upon replacing its k-th column with vector v.

For any p-elements subset s, the data needed for the WLS are stored in vector $z_{(s)} = S_{(s)}z$ and the square matrix $Q_{(s)} = S_{(s)}Q$; the best-fit parameters are

$$\hat{\beta}_{(s)} = Q_{(s)}^{-1} z_{(s)} = X_{(s)}^{-1} W_{(s)}^{-1/2} W_{(s)}^{1/2} y_{(s)} = X_{(s)}^{-1} y_{(s)}, \tag{17}$$

which shows that, for p measures, WLS and OLS give the same results.

Use of Cramer's rule on eqs. 5 and 17 gives

$$\hat{\beta}_k = \frac{\det Q^T Q^{[k|z]}}{\det Q^T Q}, \tag{18}$$

and

$$\hat{\beta}_{(s)k} = \frac{\det Q_{(s)}^{[k|z]}}{\det Q_{(s)}} = \frac{\det X_{(s)}^{[k|y]}}{\det X_{(s)}}, \tag{19}$$

Use of the Cauchy-Binet theorem to expand the determinants of the equation 18 leads to

$$\hat{\beta}_k = \frac{\sum_s \det Q_{(s)} \det Q_{(s)}^{[k|z]}}{\sum_s \det Q_{(s)} \det Q_{(s)}} = \frac{\sum_s w_s \hat{\beta}_{(s)k}}{\sum_s w_s}, \tag{20}$$

which is the equation for a weighted average of the OLS results $\hat{\beta}_{(s)k}$ with weights

$$w_s = (\det Q_{(s)})^2. \tag{21}$$

[1] Implicit use is made of the functional invariance of the estimator $\hat{b}$.

The above representation of the best-fit parameters is the starting point for robust modifications of WLS, where the basic idea is to exclude from the mean the more extreme values of $\beta_{(s)k}$ [7].

5 Conclusions

The least squares method, a fundamental piece of knowledge for students of all scientific tracks, is often introduced considering the simple linear regression with only two parameters to be determined. However, the availability of ever more large data sets prompts even undergraduate students to a sounder and wider knowledge of linear regression. Here, we have used the linear algebra formalism to compact the main results of the least squares method, encompassing ordinary and weighted least squares, goodness of fit indicators, and eventually a basic equation of re-sampling, which could be used to stimulate interested students in an even broader knowledge of data analysis. The compactness of the equations reported above allow their introduction at the undergraduate level, provided that basic linear algebra has been previously introduced.

Acknowledgements

Financial support from the MIUR (FARB2015) is gratefully acknowledged.

References

[1] R.J. Carroll and D. Ruppert. *Transformation and weighting in regression*. Chapmand and Hall, 1988.

[2] G. Casella and R.L. Berger. *Statistical inference*. Cengage Learning, 2001.

[3] A. J. Dobson and A. G. Barnett. *An Introduction to Generalised Linear Models*. Chapman and Hall, 2008.

[4] R. W. Farebrother. Relations among subset estimators: A bibliographical note. *Technometrics*, 27(1):85–86, 1985.

[5] W.H. Greene. *Econometric Analysis*. Prentice Hall, 2012.

[6] J. R. Taylor. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books, 1997.

[7] C. F. J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295, 1986.

Appendix

Moments of $\hat{a}$ e $\hat{b}$

Averages.

$< \hat{b} >= \frac{<s_{xy,w}>}{s_{xx,w}} = \frac{(x-\bar{x}_w\mathbf{1})^T W<y-\bar{y}_w\mathbf{1}>}{s_{xx,w}} = \frac{(x-\bar{x}_w\mathbf{1})^T W<y>}{s_{xx,w}} = \frac{(x-\bar{x}_w\mathbf{1})^T W<a\mathbf{1}+bx>}{s_{xx,w}} = b\frac{(x-\bar{x}_w\mathbf{1})^T W(x-\bar{x}_w\mathbf{1})}{s_{xx,w}} = b$;

$< \hat{a} >=< \bar{y}_w > -\bar{x}_w < \hat{b} >= a + b\bar{x}_w - \bar{x}_w b = a$

The estimators are then unbiased.

Variances.

We shall use the following auxiliary results:

i) $Cov(y_i, y_j) = \delta_{ij} w_i^{-1}$

ii) $Var(\bar{y}_w) = Tr(W)^{-1}$

iii) $Cov(\bar{y}_w, \hat{b}) = 0$

Given $d := W(x - \bar{x}_w\mathbf{1})$, we have that $d^T\mathbf{1} = \sum_i d_i = 0$ and then $S_{xy,w} = d^T(y - \bar{y}_w\mathbf{1}) = d^T y$; Then $Var(S_{xy,w}) = \sum_{ij} d_i d_j Cov(y_i, y_j) = \sum_i d_i^2 w_i^{-1}$

$= \sum_i w_i (x_i - \bar{x}_w)^2 = S_{xx,w}$ from which $Var(\hat{b}) = \frac{Var(S_{xy,w})}{S_{xx,w}^2} = \frac{1}{S_{xx,w}}$.

$Var(\hat{a}) = Var(\bar{y}_w) - \bar{x}_w Cov(\bar{y}_w, \hat{b}) + \bar{x}_w^2 Var(\hat{b}) = Tr(W)^{-1} + \frac{\bar{x}_w^2}{S_{xx,w}} = \frac{\overline{x^2}_w}{S_{xx,w}}$.

.

$Cov(\hat{a}, \hat{b}) = Cov(\bar{y}_w - \bar{x}_w\hat{b}, \hat{b}) = Cov(\bar{y}_w, \hat{b}) - \bar{x}_w Var(\hat{b}) = -\frac{\bar{x}_w}{S_{xx,w}}$.

Proof of the auxiliary results

i) $Cov(y_i, y_j) = Cov(a + bx_i + \varepsilon_i, a + bx_j + \varepsilon_j) = Cov(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma_i^2 = \delta_{ij} w_i^{-1}$.

ii) $Var(\bar{y}_w) = Tr(W)^{-2} \sum_{ij} w_i w_j Cov(y_i, y_j) = Tr(W)^{-2} \sum_i w_i = Tr(W)^{-1}$

iii) $Tr(W)Cov(\bar{y}_w, S_{xy,w}) = \sum_{ij} w_i d_j Cov(y_i, y_j) = \sum_i d_i = 0$ and then $Cov(\bar{y}_w, \hat{b}) = 0$.

□