Regression Analysis Using Weighted Least Squares

Robert Drucker, University of Washington (May 1996)

1. Theory

The j-th order least-squares regression problem can be stated in matrix notation as

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

where $\mathbf{X} = [\mathbf{1}\ \mathbf{x}\ \mathbf{x^2} \ldots \mathbf{x^j}]$, $\mathbf{x^n} = \left[x_i^n\right]$, is the independent variable matrix, $\mathbf{Y}$ is the dependent variable column vector $\left[y_i\right]$, and β and ε are respectively the linear model parameters and residual errors. The parameters β are estimated by $\mathbf{b} = \hat{\beta}$ so as to minimize the residuals

$$\hat{\varepsilon} = \mathbf{Y} - \mathbf{Xb}$$

in a least-squares sense, with solution

$$\mathbf{b} = (\mathbf{X}' \mathbf{X}) \setminus (\mathbf{X}' \mathbf{Y})$$

where the backslash operator \ denotes left-hand matrix division.

Conventional regression analysis assumes that the residuals $\hat{\varepsilon}$ are independent identically distributed (iid) random variables with distribution $N(0, \sigma^2)$. For our analysis, the results of the Durbin-Watson test for lag-1 serial correlation on our temperature series support the assumption of independence, while the close agreement of our results using a Monte Carlo simulation with those from a standard t-test analysis supports the assumption of normally distributed errors (normal distribution is also nearly guaranteed by the large sample size N>100 of each estimate.) However, it is not clear that the errors are identically distributed; in fact, the sample variance of the observations varies with both season and station, so that the standard error of the mean of each

sample period can be expected to vary proportionately. In addition, each sample may be derived from 1-4 ice stations deployed concurrently in the area. Thus we may not make the assumption of identically distributed errors and we must compute appropriate estimates of the mean and uncertainty of the temperatures for each period based on available data.

Draper and Smith[1] show that if the covariance matrix $\mathbf{V}(\varepsilon)$ of the residuals is not of the form $\mathbf{I}\sigma^2$ but is instead of the form $\mathbf{D}\sigma^2$, where $\mathbf{D}$ is a diagonal matrix with unequal nonzero diagonal elements, then the regression equation may still be solved using a transformation of variables, and has solution

$$\mathbf{b} = (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}) \setminus (\mathbf{X}'\mathbf{D}^{-1}\mathbf{Y}) \qquad [1.1]$$

Moreover, the covariance matrix of $\mathbf{b}$ is given by

$$\mathbf{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}\sigma^2 \qquad [1.2]$$

where the diagonal elements of $\mathbf{V}(\mathbf{b})$ give the variances of the fitted parameters $\mathbf{b}$. We may also estimate σ^2 by the residual mean square

$$s^2 = (\mathbf{Y}'\mathbf{D}^{-1}\mathbf{Y} - \mathbf{Y}'\mathbf{D}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{Y}) / \upsilon \qquad [1.3]$$

where υ= N-2 is the number of degrees of freedom.

2. Weighted Averages

For each station that satisfies the minimum of 95% operation in a sample period (ie, $N \geq 0.95M$, where M = 4 times the number of days in the sample period), we compute a mean temperature μ and a standard deviation of the mean

$$\sigma_\mu = \frac{\sigma_s}{\sqrt{N}}$$

where σ_s is the standard deviation of the sample. We then combine k = 1 to 4 concurrent stations into a single data point using the weighted averages

$$\bar{\mu} = \frac{\sum_{i=1}^{k} w_i \cdot \mu_i}{\sum_{i=1}^{k} w_i} \quad \text{and} \quad \bar{\sigma} = \left(\sum_{i=1}^{k} w_i \right)^{-1/2}$$

where the weights

$$w_i = 1/\sigma_{\mu_i}^2, \qquad i = 1..k$$

are the inverse squares of the individual uncertainties σ_μ computed above[2].

3. Weighted Regressions

Having computed the weighted means $\bar{\mu}$ and $\bar{\sigma}$ for each data point, we solve the first-order weighted regression problem by Equation 1.1 using $\mathbf{Y} = \left[\bar{\mu}\right]$ and the diagonal weighting matrix D composed of diagonal elements $\bar{\sigma}^2$. The trend (slope) b1 is taken as the second diagonal element of **b**. We compute s^2 using equation 1.3, then compute the covariance matrix

$$\hat{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1} s^2$$

where we substitute the estimator s^2 for σ^2 in equation 1.2 (we may also now estimate the variance matrix $\mathbf{D}\sigma^2$ by $\mathbf{D}s^2$ which provides estimates of the sample uncertainties.) The square root of the second diagonal element of $\hat{V}(\mathbf{b})$ gives the estimated standard deviation of the slope:

$$\hat{\sigma}_{b_1} = \sqrt{\hat{V}[2,2]}$$

4. Significance of the Regression

The significance of the regression may be calculated from the estimated standard deviation of the slope $\hat{\sigma}_{b_1}$, which was taken as the square root of the second diagonal element of the covariance matrix $\mathbf{V}(\mathbf{b})$. We define the *t*-statistic

$$t = \frac{b_1}{\hat{\sigma}_{b_i}}$$

and the *F*-statistic $F = t^2$. If $\beta_1 = 0$, F follows an *F*-distribution with 1 and υ degrees of freedom[1]. Thus we can calculate

$$P(F|\upsilon_1, \upsilon_2) = I_{\frac{\upsilon_2}{\upsilon_2 + \upsilon_1 F}}\left(\frac{\upsilon_2}{2}, \frac{\upsilon_2}{2}\right)$$

where $I_x(a, b)$ is the incomplete Beta function[3]. P represents the probability that F would be as large as it is given the null hypothesis $H_0 : \beta_1 = 0$; if P<.05 we can reject the null hypothesis with a confidence level of 95%.

Note that since the *t*-distribution approaches the normal distribution as $\upsilon \to \infty$, we could estimate P for values of υ greater than about 30 by integrating the two tails of a normal distribution with variance F:

$$P \approx \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-\frac{x^2}{2}} dx$$

$$= \frac{2}{\sqrt{\pi}} \int_{0}^{t/\sqrt{2}} e^{-s^2} ds$$

$$= erf\left(\frac{t}{\sqrt{2}} \right)$$

Where *erf()* denotes the error function.

We can also use the *t*-statistic to determine 95% confidence intervals for β_1:

$$\beta_1 = b_1 \pm t(\upsilon, 0.975) \cdot \hat{\sigma}_{b_1}$$

where the above statement is made with 95% confidence.

Table 1 gives the confidence intervals and P values for the monthly, seasonal and annual trends. Trends for which the 95% confidence interval does not include 0, or, equivalently, for which P<.05, are significant at the 95% confidence level.

References:

1. N. Draper and H.Smith, *Applied Regression Analysis, 2nd Edition*
2. John R. Taylor, *An Introduction to Error Analysis*
3. W. H. Press *et. al., Numerical Recipes in FORTRAN, 2nd Edition*

Table 1: Trends from Weighted Least-square Regression

		Temperatures		Anomalies	
Period	**n**	**Trend (^{o}C/decade)**	**P**	**Trend (^{o}C/decade)**	**P**
Jan	30	0.85 +/- 1.33	0.200	0.85 +/- 1.27	0.179
Feb	30	0.20 +/- 1.32	0.760	0.25 +/- 1.20	0.671
Mar	29	0.14 +/- 1.11	0.803	0.14 +/- 1.00	0.770
Apr	28	-0.11 +/- 0.89	0.800	-0.27 +/- 0.99	0.579
May	29	**0.89 +/- 0.71**	**0.016**	**0.84 +/- 0.77**	**0.035**
Jun	30	**0.43 +/- 0.36**	**0.020**	**0.41 +/- 0.33**	**0.016**
Jul	30	0.02 +/- 0.13	0.715	0.03 +/- 0.16	0.723
Aug	30	0.05 +/- 0.35	0.769	0.02 +/- 0.29	0.867
Sep	30	0.35 +/- 0.85	0.398	0.23 +/- 0.89	0.597
Oct	30	0.17 +/- 0.84	0.675	0.30 +/- 0.85	0.485
Nov	30	**-1.14 +/- 1.15**	**0.053**	-0.62 +/- 1.20	0.301
Dec	30	0.01 +/- 0.87	0.979	-0.06 +/- 1.00	0.897
Winter	28	0.35 +/- 0.65	0.273	0.35 +/- 0.61	0.246
Spring	29	0.11 +/- 0.67	0.731	0.21 +/- 0.66	0.521
Summer	30	0.14 +/- 0.18	0.119	**0.20 +/- 0.17**	**0.025**
Autumn	30	-0.38 +/- 0.68	0.258	-0.21 +/- 0.60	0.473
Annual	27	0.07 +/- 0.46	0.747	0.14 +/- 0.39	0.463