Statistical Weather Forecasting (Wilks, Chapter 6)

A. Without NWP

Linear regression

y_i : predictand at time t_i $i = 1,\ldots n$

$x_{i1},\ldots,x_{ik},\ldots,x_{iK}$: K predictors $k = 1,\ldots K$ at time t_i

Linear regression forecast

$$\hat{y}_i = b_0 + b_1 x_{i1} + \ldots + b_K x_{iK} \qquad y_i = \hat{y}_i + \varepsilon_i$$

at time t_i observed = forecast + error

Simple linear regression: one predictor (K=1)

$$\hat{y}_i = b_0 + b_1 x_i \quad \varepsilon_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$$

Choose b_0, b_1, to minimize $\sum_{i=1}^{n} \varepsilon_i^2 = f(b_0, b_1)$: least-squares regression

$$\frac{\partial \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i \right)^2}{\partial b_0} = -2 \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i \right) = 0$$

Now, $\sum_{i=1}^{n} y_i = n\bar{y}$ so that $n\bar{y} - nb_0 - nb_1\bar{x} = 0$

or

$$b_0 = \bar{y} - b_1 \bar{x}$$

i.e., the regression line goes through the means

Now, replacing b_0

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i \right)^2 = \sum_{i=1}^{n} \left(y_i - \bar{y} - b_1 (x_i - \bar{x}) \right)^2 = \sum_{i=1}^{n} \left(y_i' - b_1 x_i' \right)^2$$

So that for the second coefficient b_1

$$\frac{\partial \sum_{i=1}^{n} \left(y_i' - b_1 x_i' \right)^2}{\partial b_1} = -2 \sum_{i=1}^{n} \left(y_i' - b_1 x_i' \right) x_i' = 0$$

From here

$$\sum_{i=1}^{n} x_i' y_i' = b_1 \sum_{i=1}^{n} x_i' x_i' \text{ or } \qquad n\overline{x'y'} = b_1 n\overline{x'^2} \text{, i.e., } \qquad \underline{b_1 = \frac{\overline{x'y'}}{\overline{x'^2}}}$$

Exercise: From $y_i = b_0 + b_1 x_i + \varepsilon_i$ and the formulas for b_0, b_1 show that $\overline{x'\varepsilon'} = 0$, i.e., that the forecast error must be uncorrelated to the predictors.

Note that b_1 can be written as

$$b_1 = \frac{\overline{x'y'}}{\sqrt{\overline{x'^2}\,\overline{y'^2}}} \frac{\sqrt{\overline{y'^2}}}{\sqrt{\overline{x'^2}}} = \rho \frac{s_y}{s_x}$$

where ρ is the sample x-y correlation, and $s_y^2 = \sum_{i=1}^{n} \frac{y_i'^2}{n-1}$ is the sample estimate of the variance.

It is convenient to consider sums of squares: We can divide the "sum of squares" = "n*variance" of y in the following way:

SST	=	SSR	+	SSE
Total y-variance		Regression variance		Forecast error (residual) variance

$$SST = \sum_{i=1}^{n} y_i'^2 = n\overline{y'^2}$$

$$SSE = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i\right)^2 = \sum_{i=1}^{n} \left(y_i' - b_1 x_i'\right)^2 = \sum_{i=1}^{n} \left(y_i' - \frac{\overline{x'y'}}{\overline{x'^2}} x_i' \right)^2 =$$

$$= \sum_{i=1}^{n} \left(y_i'^2 - 2\frac{\overline{x'y'}}{\overline{x'^2}} x_i' y_i' + \frac{\overline{x'y'}^2}{\overline{x'^2}^2} x_i'^2 \right) = n\overline{y'^2} - 2\frac{\overline{x'y'}}{\overline{x'^2}} n\overline{x'y'} + \frac{\overline{x'y'}^2}{\overline{x'^2}^2} n\overline{x'^2} =$$

$$= n\overline{y'^2} - n\frac{\overline{x'y'}^2}{\overline{x'^2}} = n\overline{y'^2}\left(1 - \frac{\overline{x'y'}^2}{\overline{x'^2}\,\overline{y'^2}}\right) = n\overline{y'^2}\left(1-\rho^2\right)$$

So, since $SST = n\overline{y'^2}$ is the total variance of y (with *n-1* degrees of freedom)

$SSE = n\overline{y'^2}\left(1-\rho^2\right)$ is the residual (error) variance of y

Therefore , since ***SST=SSR+SSE***

$SSR = n\overline{y'^2}\rho^2$ is the "explained variance" of the regression forecast: the square of the correlation gives the percentage of "explained variance" in the dependent sample used to derive the regression coefficients.

In the case of multiple regression, this is also true, allowing the definition of a "coefficient of determination" R^2 (a generalized squared correlation):

$SST = n\overline{y'^2}$ has *n-1* degrees of freedom (one was used for $\overline{y}$). Hence $\frac{SST}{n-1}$ is the variance of y

$SSE = n\overline{y'^2}\left(1-R^2\right)$ is the residual error variance, which defines R^2

$SSR = n\overline{y'^2}R^2$ has K degrees of freedom, K is the number of predictors

SSE has *n-1-K* degrees of freedom. Therefore the <u>dependent</u> sample estimate of the forecast error variance is $s_\varepsilon^2 = \frac{1}{n-1-K}\sum_{i=1}^{n}\varepsilon_i^{\,2} = \frac{1}{n-1-K}SSE$.

Now, after we finished the "training" of the simple linear regression (finding b_0 , b_1) for the dependent data set, we have a new independent predictor $x_{0.}$ When we apply the formula we derived for the dependent sample to this independent new predictor, then the forecast error variance estimate is larger because the training data has sampling errors:

$$s_{\varepsilon_0}^2 = s_\varepsilon^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$$

Here the term $\frac{1}{n}$ is due to sampling errors of the mean, and $\frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ due to sampling errors of the slope.

If instead of simple regression we have multiple regression with K predictors, the error for a new independent predictor also increases compared to the dependent sample estimate:

$$s_{\varepsilon_0}^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n - K - 1} \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$$

So, if the new predictor is far from the mean, the expected error of the forecast is large! The prediction error with independent data is increased compared with the dependent (training) sample for two reasons: 1) the number of degrees of freedom is reduced by using K predictors, and 2) the dependent sampling errors are built into the prediction equation (and don't apply to an independent sample).