Chapter 12 and Time Series

1. Consider the regression of (college) gpa by predicting variables of sat and high school grade point average (~~hsgpa~~ hsavg). The proposed model consists of a linear trend over sat and a quadratic trend over hsavg. A random set of n students are obtained and for each the gpa, sat, hsavg is recorded.
   a. Present this model for a generic student i of the sample.
   b. Over all i = 1, 2, …, n, show the above model in terms of vectors and matrices (e.g. one component would show the terms that would be in the "Z" matrix).
   c. How would you test the null hypothesis of a reduced model that excluded hsavg as a predicting variable? (You should clearly state the null and alternative hypothesis, present the test statistic, state its distributon under the null hypothesis, explain how each component of the test statistic would be computed, and explain how you would decide whether or not to reject the null hypothesis, at the $\alpha = 0.05$ level.

(ALSO SEE 1.d. ON NEXT PAGE)

2. Fertilizer treatment A is given to n_A=18 randomly selected plots from a large region; also, Fertilizer treatment B is given to n_B=15 independently selected random plots from the same region. Prior to treatment, the *organic matter content* (omc) of each sample plot was measured. It is desired to model corn growth (y) in the plots as being potentially assoiciated with which treatment is given and also a possible linear effect of a plot's organic matter content, with the assumption that there is no interaction between treatment and orgaic matter effects.
   a. Write this model in multiple regression form: with β's, variables and error, being sure to clearly describe the variables.
   b. Now, if interaction between treatment and organic matter is added, then how would the model (in regression form) change and what would be the practical result in terms of the plot of E(corn growth) vs omc, shown separately for each respective treatment.

3. Consider that the monthly unemployment rate (y) can be modeled as the sum of a linear long term trend, a linear influence of the previous month's gdp, a continuous (cos and sin) seasonal adjustment, and first-order autocorrelated errors. Data on these are recoreded for the last 48 months (t=1, 2, …, 48)
   a. Write the above model in terms of y_t, β's, variables and ϕ.
   b. How would you predict next month's unemployment rate?

1.(a) Let y = gpa, x_1 = sat, x_2 = hsavg.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 (x_{i2})^2 + \epsilon_i$$

(b) $$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & x_{12}^2 \\ 1 & x_{21} & x_{22} & x_{22}^2 \\ & & & \\ 1 & x_{n1} & x_{n2} & x_{n2}^2 \end{pmatrix}}_{Z_c} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{\beta_c} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

(c) Reduced model of excluding hsavg. is same as model where $\beta_2 \& \beta_3 = 0$

so test H_0: $\beta_2 = \beta_3 = 0$ (Reduced model is that $y_i = \beta_0 + \beta_1 x_{i1} + \tilde{\epsilon}_i$)

H_A: not H_0

Complete model is model of part (a), above.

Test statistic:

$$\frac{\frac{SSE_R - SSE_C}{2}}{\frac{SSE_C}{n-(3+1)}} \underset{H_0}{\sim} F_{\text{numdf}=2,\ \text{dendf}=n-(3+1)}$$

where $SSE_C = \sum_{i=1}^{n} (y_i - \hat{y}_{i_C})^2$

where $\hat{y}_{i_C}$ is estimate of y_i under the complete model, namely

$\hat{y}_{i_C} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i2}^2$,

where $\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = (Z_c^T Z_c)^{-1} Z_c^T \underline{y}$,

similarly, $SSE_R = \sum_{i=1}^{n} (y_i - \hat{y}_{i_R})^2$

where $\hat{y}_{i_R}$ is estimate of y_i under the reduced model, namely

$\hat{y}_{i_R} = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1}$, where $\begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \end{pmatrix} = (Z_R^T Z_R)^{-1} Z_R^T \underline{y}$, where $Z_R = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ & \vdots \\ 1 & x_{n1} \end{pmatrix}$

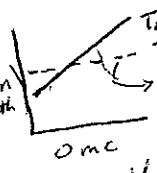
You reject H_0 iff the data's value for test statistic exceeds $F_{.05}$ (numdf = 2, dendf = n−(3+1))

(2)(a) Set up dummy r.v. $x_{i1} = \begin{cases} 1 & \text{if plot } i \text{ got Trt B} \\ 0 & \text{otherwise} \end{cases}$ so Trt A is base "level"

for $i = 1, 2, \ldots, 18+15$, $y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot omc_i + \epsilon_i$ (no interaction)

(b) if interaction: $y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot omc_i + \beta_3 \cdot (x_{i1} \cdot omc_i) + \epsilon_i$

for interaction



→ the slopes of the 2 lines will differ,

This means that the per unit change in E(corn growth) as omc changes, would differ for Trt A and Trt B. (Thus the estimated effect of Applying Trt A vs Trt B, depends on the omc of the plot.

(3)(a) long-term linear

$$y_t = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot gdp_{t-1} + \beta_3 \left(\cos \tfrac{2\pi t}{12}\right) + \beta_4 \left(\sin \tfrac{2\pi t}{12}\right) + R_t$$

and $R_t = \phi R_{t-1} + \epsilon_t$, for $t = 2, 3, \ldots, 48$.

(b) Now $R_{49} = \phi R_{48} + \epsilon$, so $\hat{R}_{49} = \hat{\phi} \hat{R}_{48}$ and where

$\hat{R}_{48} = y_{48} - \left(\hat{\beta}_0 + \hat{\beta}_1 \cdot 48 + \hat{\beta}_2 gdp_{47} + \hat{\beta}_3 \left(\cos \tfrac{2\pi(48)}{12}\right) + \hat{\beta}_4 \sin \left(\tfrac{2\pi(4}{12}\right.\right.$

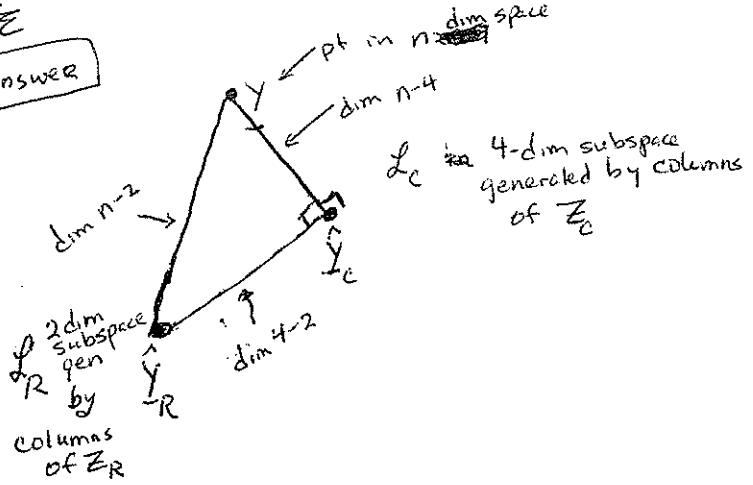
So, predict next month's unemployment rate by:

$\hat{y}_{49} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 49 + \hat{\beta}_2 \cdot gdp_{48} + \hat{\beta}_3 \left(\cos \tfrac{2\pi(49)}{12}\right) + \hat{\beta}_4 \left(\sin \tfrac{2\pi(49)}{12}\right) + \hat{R}_4$

see above for $\hat{R}_{49}$.

1.d. Present the n-dimensional geometric representation of this test.

Answer



$\mathcal{L}_C$ the 4-dim subspace generated by columns of Z_C

Test statistic $\dfrac{[(\hat{y}_C - \hat{y}_R)'(\hat{y}_C - \hat{y}_R)]/(4-2)}{[(y - \hat{y}_C)'(y - \hat{y}_C)]/(n-4)} \overset{H_0}{\sim} F_{\substack{\text{numdf}=4-2 \\ \text{dendf}=n-4}}$

FOR ADDITIONAL EXPLANATION, SEE 1-PAGE SUMMARY OF EFRON ARTICLE ON NEXT PAGE

Other suggested problems from Chapter 12.

In book, Chpt. 12 #1, 2, 5, 12, 16, 28, 37, 32, 39, 44, 54, 60, 68, 70, 75, 88, 90, 101, 109

The spaces $\mathscr{L}(1)$ and $\mathscr{L}(2)$ are indicated in Fig. 3, along with the corresponding least-squares points $\hat{\mu}(1)$ and $\hat{\mu}(2)$. Because $\mathscr{L}(1) \subset \mathscr{L}(2)$ we must have

$$(3.5) \qquad \sum_{i=1}^{n} (y_i - \hat{y}_{i(2)})^2 = SSE_C \leq \|\mathbf{y} - \hat{\mu}(2)\|^2 \leq \|\mathbf{y} - \hat{\mu}(1)\|^2 = SSE_R = \sum_{i=1}^{n} (y_i - \hat{y}_{i(1)})^2$$

In other words, increasing the explanatory space $\mathscr{L}$ decreases the residual vector $\mathbf{r} = \mathbf{y} - \hat{\mu}$. As we shall see this does *not* mean that big models are always better than small ones.

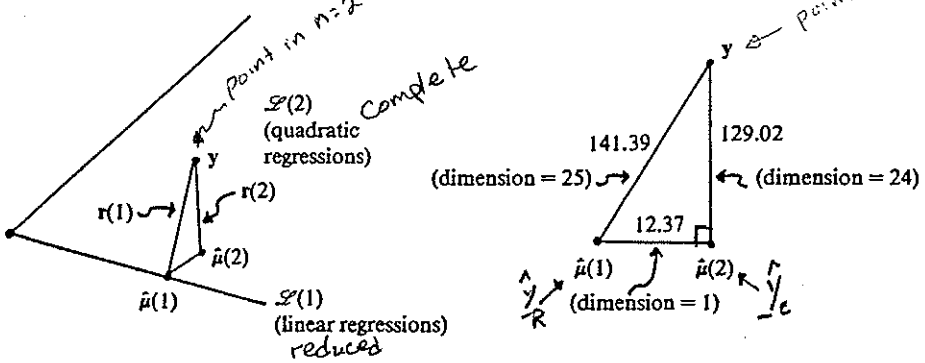


FIG. 3. *On the left is a schematic diagram comparing linear and quadratic models for the hormone data of* §1. *Because* $\mathscr{L}(1) \subset \mathscr{L}(2)$, *the residual vector must be larger for* $\mathscr{L}(1)$. *The squared lengths* $\|\mathbf{y} - \hat{\mu}(1)\|^2$ *and* $\|\mathbf{y} - \hat{\mu}(2)\|^2$ *are shown at right. Traditional theory indicates that there is no strong reason to believe that quadratic regression is superior to linear regression for the hormone data.*

Figure 3 indicates the squared residual lengths $\|\mathbf{y} - \hat{\mu}(1)\|^2 = 141.39$ and $\|\mathbf{y} - \hat{\mu}(2)\|^2 = 129.02$ for the hormone data, from which Pythagorus' theorem gives $\|\hat{\mu}(2) - \hat{\mu}(1)\|^2 = 12.37$. Does a decrease of 12.37 in squared residual length indicate a genuine advantage to the bigger model? Here is the traditional answer to that question:

(i) $\mathbf{y} - \hat{\mu}(2)$ takes its value in the 24-dimensional subspace of $\mathbb{R}^{27}$ orthogonal to the three-dimensional space $\mathscr{L}(2)$.

(ii) $\hat{\mu}(2) - \hat{\mu}(1)$ takes its value in a one-dimensional space, the portion of $\mathscr{L}(2)$ orthogonal to $\mathscr{L}(1)$.

(iii) If there is no true quadratic effect, that is, if the true mean vector μ lies in $\mathscr{L}(1)$, then we expect the ratio

$$(3.6) \qquad F = \frac{\|\hat{\mu}(2) - \hat{\mu}(1)\|^2/1}{\|\mathbf{y} - \hat{\mu}(2)\|^2/24}$$

to approximately equal 1 (since then there is nothing special about $\hat{\mu}(2) - \hat{\mu}(1)$ compared to the other 24 components of $\mathbf{y} - \hat{\mu}(1) = [\hat{\mu}(2) - \hat{\mu}(1)] + [\mathbf{y} - \hat{\mu}(2)]$).

(iv) The observed value of F in this case is $12.37/(129.02/24) = 2.30$, so we have to decide whether or not 2.30 "approximately equals 1."

(v) The theoretical distribution of F can be computed for model (2.6), (2.7); assuming that the true quadratic effect is zero,

$$(3.7) \qquad \text{Prob}\{F > 2.30\} = .14.$$

(vi) An "achieved significance level" of .14 is not considered significant evidence for the existence of a genuine quadratic component in μ. We expect bigger values of F, and smaller achieved significance levels, if μ is genuinely quadratic. The conventional borderline for significance is .05, in this case $F \geq 4.26$. To state things in