Inference for (Prediction Using ?) Functional Regression Models

R. L. Eubank\textsuperscript{1}

Department of Statistics
Texas A & M University

\textsuperscript{1}Research supported by NSF
Texas Lottery: Background

• Texas Lottery Commission established in 1992

• Four on-line games and scratch-offs

• Ticket terminals and sales monitoring by GTECH

• Capstone game is Lotto Texas
  
  – Drawings on Wednesday and Saturday night (10 P.M.)
  
  – Select 6 from 54 ball game

  – Chance of winning jackpot prize is about 1 in 26 Million

  – Overall chance of winning are 1 in 71
Lotto Texas Jackpots

- State retains 50% of sale:  *Prize Pool* is remaining 50%
- Jackpot tier gets 64% of the *Prize Pool*
- Jackpot tier funds accumulate until there is a winner
  - **Terminology:**
    - A *Hit* occurs when there is one or more winner.
    - A *Run* is a sequence of consecutive draws without a *Hit*
- Jackpots start at $4 Million
- Advertised Jackpots during a *Run* are (ideally) the annuitized (over 25 years) value of the accumulated money in the Jackpot tier. The idealize formula is:

  \[
  \text{Jackpot} = 0.5 \times 0.64 \times (\text{annuity factor}) \times (\text{cumulative sales for the Run})
  \]

- *Cumulative sales at draw time are not known*

**Problems:** For a prospective Jackpot value, predict
- cumulative sales up to Saturday using information only up to Wednesday afternoon
- cumulative sales up to Wednesday using information only up to Friday afternoon
Methodology

- **Current Approach:**
  - “nearest” neighbor
    - Advantages:
      - Very Simple
      - Works very well
  - Disadvantages
    - Ad Hoc
      - No prediction error assessment

- **Possible Statistical Alternatives**
  - Simple linear regression of sales on Jackpot
  - Nonparametric estimation of the mean function
  - FDA analysis of runs as sample paths
Typical Run Sequence Thursday Start

Cumulative sales vs. day of run graph showing:
- Thurs.: cumulative sales increase with day of run.
- Wed.: cumulative sales remain stable.
- Sat.: cumulative sales show a significant increase with day of run.

Cumulative sales values range from 0 to 12.
Registration Issues

- Sales peek on the day of a Lotto draw

- Runs that start on Thursday and Sunday are on different time scales

- *Time Rescaling: Aligning Landmarks*

<table>
<thead>
<tr>
<th>Time Index</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>4 2/3</th>
<th>5 1/3</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Index</td>
<td>1</td>
<td>1 2/3</td>
<td>2 1/3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>
Modeling the Sample Path “Mean” Function

- Lotto Texas sales are driven by Jackpots
  - Shapira and Venezia (1992) *Organizational Behavior and Human Decision Processes* 50

- There is a Lottomania effect (i.e., rollover has more effect than would be expected from just the Jackpot increase)

Possible Statistical Implications:

- Relevant predictor variables are length of *Run* and Jackpot size

- A candidate mean function model might be the functional regression/time-varying coefficient model

  \[ \mu(t, z_t) = \beta_0 + \beta_1(t)z_t \]

  with \( t \) the day scale position in the *Run* and \( z_t \) the associated Jackpot
**Relationship of Sample Paths to Mean Function**

- Sales for a given *Run* have a tendency to lie above or below the “average” trend

A model that could describe this is

$$y(t) = a + b \mu(t, z_t) + \varepsilon_t, \quad t = 1, \ldots, n,$$

where

- $y(t)$ = cumulative sales at day index $t$
- $(a, b)^T \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix}\right)$
- $\varepsilon_t \sim \text{NID}(0, \sigma^2)$
- $\varepsilon_t, \ t = 1, \ldots, n$, and $(a, b)^T$ are independent
Prediction: Step 1

Suppose that $\mu(\cdot, \cdot), \sigma_a^2, \sigma_b^2, \sigma^2$ were known. Then, the best estimator of

$$E[y(t^*)|(a, b)] = a + b\mu(t^*, z_{t^*}),$$

given $y = (y(t_1), \ldots, y(t_n))^T$ is $\hat{a} + \hat{b}\mu(t^*, z_{t^*})$ where

- $(\hat{a}, \hat{b})^T = \begin{bmatrix} 0 \\ 1 \end{bmatrix} + (X^TX + \Lambda^{-1})^{-1}X^T[y - \mu],$
- with $\mu = (\mu(t_1, z_{t_1}), \ldots, \mu(t_n, z_{t_n}))^T,$
- $X = [1_n | \mu]$ and
- $\Lambda = \begin{bmatrix} \sigma_a^2/\sigma^2 & 0 \\ 0 & \sigma_b^2/\sigma^2 \end{bmatrix}$

This gives

$$\hat{y}(t^*) = \hat{a} + \hat{b}\mu(t^*, z_{t^*}),$$

with

$$\text{Var}[\hat{y}(t^*)|y] = \sigma^2(1, \mu(t^*, z_{t^*})) \left( X^TX + \Lambda^{-1} \right)^{-1} \begin{pmatrix} 1 \\ \mu(t^*, z_{t^*}) \end{pmatrix}.$$
Prediction: Step 2. Estimation of the “Variance Components”

If $\mu(\cdot, \cdot)$ is known then the past Run data

$$y_j(t_{ij}) = a_j + b_j\mu(t_{ij}, z_{t_{ij}}) + \varepsilon_{ij}, \quad i = 1, \ldots, r_j, \quad j = 1, \ldots, k,$$

gives us predictors of the $a_j, b_j$ and estimators of $\sigma^2$ as follows:

- $\hat{b}_j = \sum_{i=1}^{r_j} y_j(t_{ij})(\mu(t_{ij}, z_{t_{ij}}) - \bar{\mu}_j)/SS_j$,

- $\hat{a}_j = \bar{y}_j - \hat{b}_j\bar{\mu}_j$,

- $\hat{\sigma}^2_j = (r_j - 1)^{-1}\sum_{i=1}^{r_j}(y_j(t_{ij}) - \hat{a}_j - \hat{b}_j\mu(t_{ij}, z_{t_{ij}}))^2$ with

- $\bar{y}_j = r_j^{-1}\sum_{i=1}^{r_j} y_j(t_{ij})$, $\bar{\mu}_j = r_j^{-1}\sum_{i=1}^{r_j} \mu(t_{ij}, z_{t_{ij}})$, and

$$SS_j = \sum_{i=1}^{r_j}(\mu(t_{ij}, z_{t_{ij}}) - \bar{\mu}_j)^2.$$

Method of moments estimators are then obtained from

- $E\sum_{j=1}^{k}(\hat{b}_j - 1)^2 = k\sigma^2_b + \sigma^2\sum_{j=1}^{k} \frac{1}{SS_j}$,

- $E\sum_{j=1}^{k}\hat{a}^2_j = k\sigma^2_a + \sigma^2\sum_{j=1}^{k} \left[\frac{1}{r_j} + \frac{1}{SS_j}\right]$ and

- $E\sum_{j=1}^{k}\hat{\sigma}^2_j = k\sigma^2$. 
Prediction: Step 3. Estimation of the mean function

A partially linear, varying coefficient, smoothing spline estimator, \( \mu_\lambda(\cdot, \cdot) \), for \( \mu(\cdot, \cdot) \) is obtained by minimization of

\[
\sum_{j=1}^{k} \sum_{i=1}^{r_j} (y_j(t_{ij}) - b - g(t_{ij}) z_{t_{ij}})^2 + \lambda \int \left( g^{(m)}(t) \right)^2 dt.
\]

Form of the Estimator:

- the fitted values are
  \[
  \hat{y}_\lambda = A(\lambda) y
  \]
  with
  \[
  A(\lambda) = I - (M^{-1} - M^{-1} V (V^T M^{-1} V)^{-1} V^T M^{-1})
  \]
  for
  \[
  V = [HT | 1]
  \]
  and
  \[
  M = HQH^T + I
  \].
The coefficient estimators are in $\beta_\lambda = C(\lambda)y$, where
\[
C(\lambda) = QH^TM^{-1} - (QH^TM^{-1}V-[T|0])(V^TM^{-1}V)^{-1}V^TM^{-1}
\].

**Efficient Computation:**

- For any $n$-vector $u$ it is possible to compute
  \[
  M^{-1}u,
  \]
  \[
  QH^TM^{-1}u,
  \]
  and the diagonal elements of $M^{-1}$ and $QH^TM^{-1}HQ$ all in $O(n)$ operations using the ordinary (i.e., non-diffuse) Kalman filter.

- $\lambda$ can be selected using GML, etc.

- C++ code available in March
Typical Run Sequence Thursday Start

cumulative sales

day of run

Thurs.
Sat.
Wed.
Typical run sequence: Thursday and Saturday Starts
Typical Run Sequences: Thursday and Saturday Starts
Registered sample paths

![Graph showing cumulative sales over run day index. The x-axis represents the run day index, and the y-axis represents cumulative sales. The graph includes multiple lines, each representing a different sample path.](image-url)
Does the starting day matter?
Sales versus Jackpot for Lotto Texas
"Smoothed" Coefficient Curve Estimator
Estimated Intercepts
Typical Run from November 2003

Observed

Jackpot=15

Yhat
Projection for Jackpot at $19 Million

 cumulative sales

 day index