# Inference for (Prediction Using ?) Functional Regression Models

**R. L. Eubank**[1]

**Department of Statistics**

**Texas A & M University**

[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:

$$\text{Jackpot} = .5\times(.64)\times(\text{annuity factor})\times(\text{cumulative sales for the } Run)$$

- *Cumulative sales at draw time are not known*

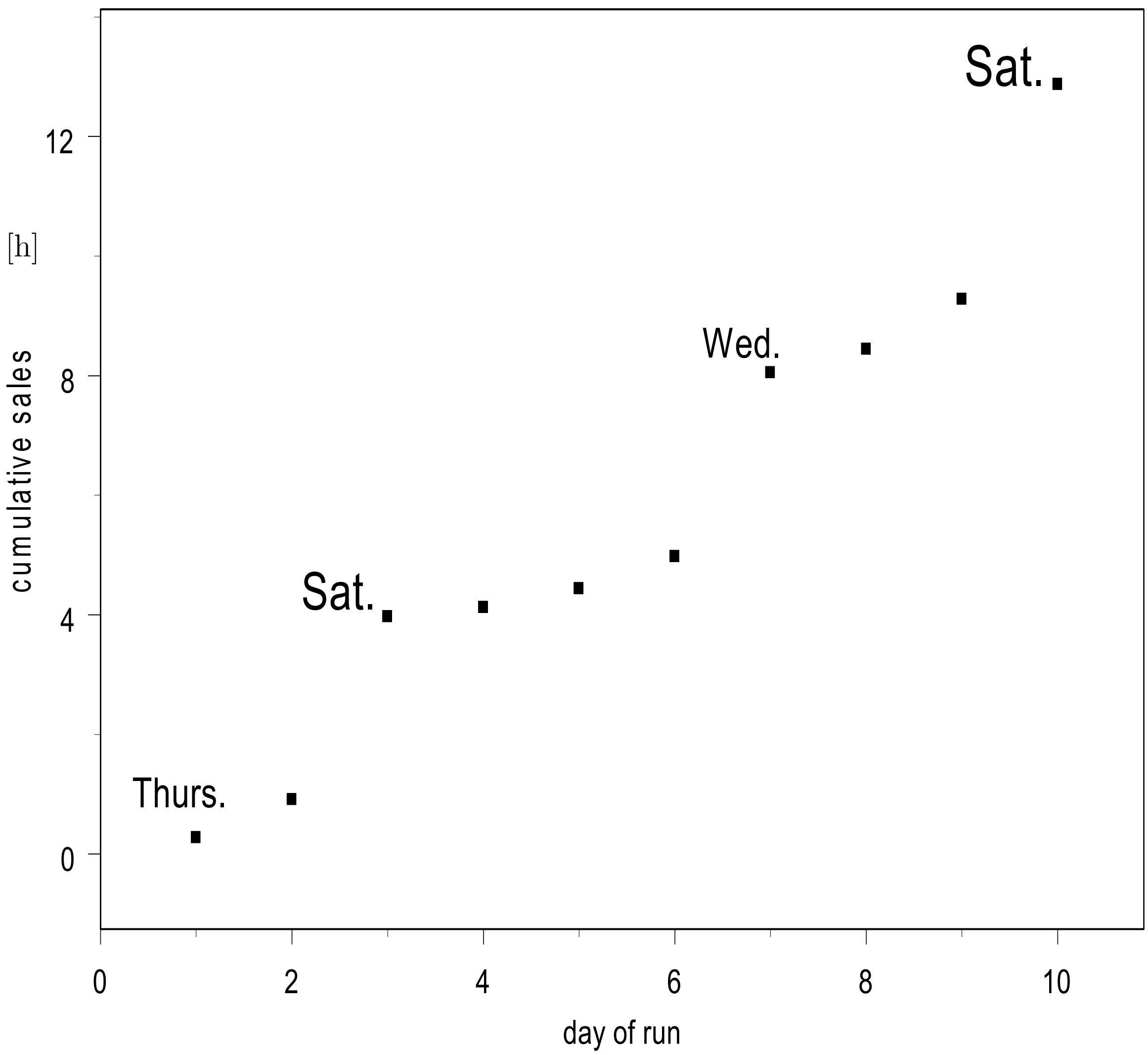
**Problems:** For a prospective Jackpot value, predict

- 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



## 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*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Time Index | 1 | 2 | | 3 | 4 | 4 2/3 | 5 1/3 | 6 |
| Thurs. Start | Thurs. | Fri. | | Sat. | Sun. | Mon. | Tues. | Wed. |
| Sun. Start | Sun. | Mon. | Tues. | Wed. | Thurs. | Fri. | | Sat. |
| Time Index | 1 | 1 2/3 | 2 1/3 | 3 | 4 | 5 | | 6 |

## 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)
  - Beenstock and Haitovsky (2001) *Journal of Economic Psychology 22*

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 $t$
- $(a, b)^T \sim \mathrm{N}\left( \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix} \right),$
- $\varepsilon_t \sim \mathrm{NID}(0, \sigma^2)$
- $\varepsilon_t,\ t = 1, \ldots, n,$ 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

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

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

- $(\hat{a}, \hat{b})^T = \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \left(X^T X + \Lambda^{-1}\right)^{-1} X^T [\mathbf{y} - \boldsymbol{\mu}],$
- with $\boldsymbol{\mu} = (\mu(t_1, z_{t_1}), \ldots, \mu(t_n, z_{t_n}))^T$,
- $X = [\mathbf{1}_n | \boldsymbol{\mu}]$ 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

$$\mathrm{Var}[\hat{y}(t^*)|\mathbf{y}] = \sigma^2(1, \mu(t^*, z_{t^*}))\left(X^T X + \Lambda^{-1}\right)^{-1} \begin{pmatrix} 1 \\ \mu(t^*, z_{t^*}) \end{pmatrix}.$$

## 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:

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

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

Method of moments estimators are then obtained from

- $\mathrm{E}\sum_{j=1}^{k}(\hat{b}_j - 1)^2 = k\sigma_b^2 + \sigma^2\sum_{j=1}^{k}\frac{1}{\mathrm{SS}_j}$,
- $\mathrm{E}\sum_{j=1}^{k}\hat{a}_j^2 = k\sigma_a^2 + \sigma^2\sum_{j=1}^{k}\left[\frac{1}{r_j} + \frac{1}{\mathrm{SS}_j}\right]$ and
- $\mathrm{E}\sum_{j=1}^{k}\hat{\sigma}_j^2 = k\sigma^2$.

## 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

  $$\hat{\mathbf{y}}_\lambda = A(\lambda)\mathbf{y}$$

  with

  $$A(\lambda) = I - (M^{-1} - M^{-1}V(V^T M^{-1} V)^{-1} V^T M^{-1})$$

  for

  $$V = [HT|\mathbf{1}]$$

  and

  $$M = HQH^T + I$$

  .

- The coefficient estimators are in $\boldsymbol{\beta}_\lambda = C(\lambda)\mathbf{y}$, where

  $$C(\lambda) = QH^TM^{-1} - (QH^TM^{-1}V - [T|\mathbf{0}])(V^TM^{-1}V)^{-1}V^TM^{-1}$$

  .

*Efficient Computation:*

- For any $n$-vector $\mathbf{u}$ it is possible to compute

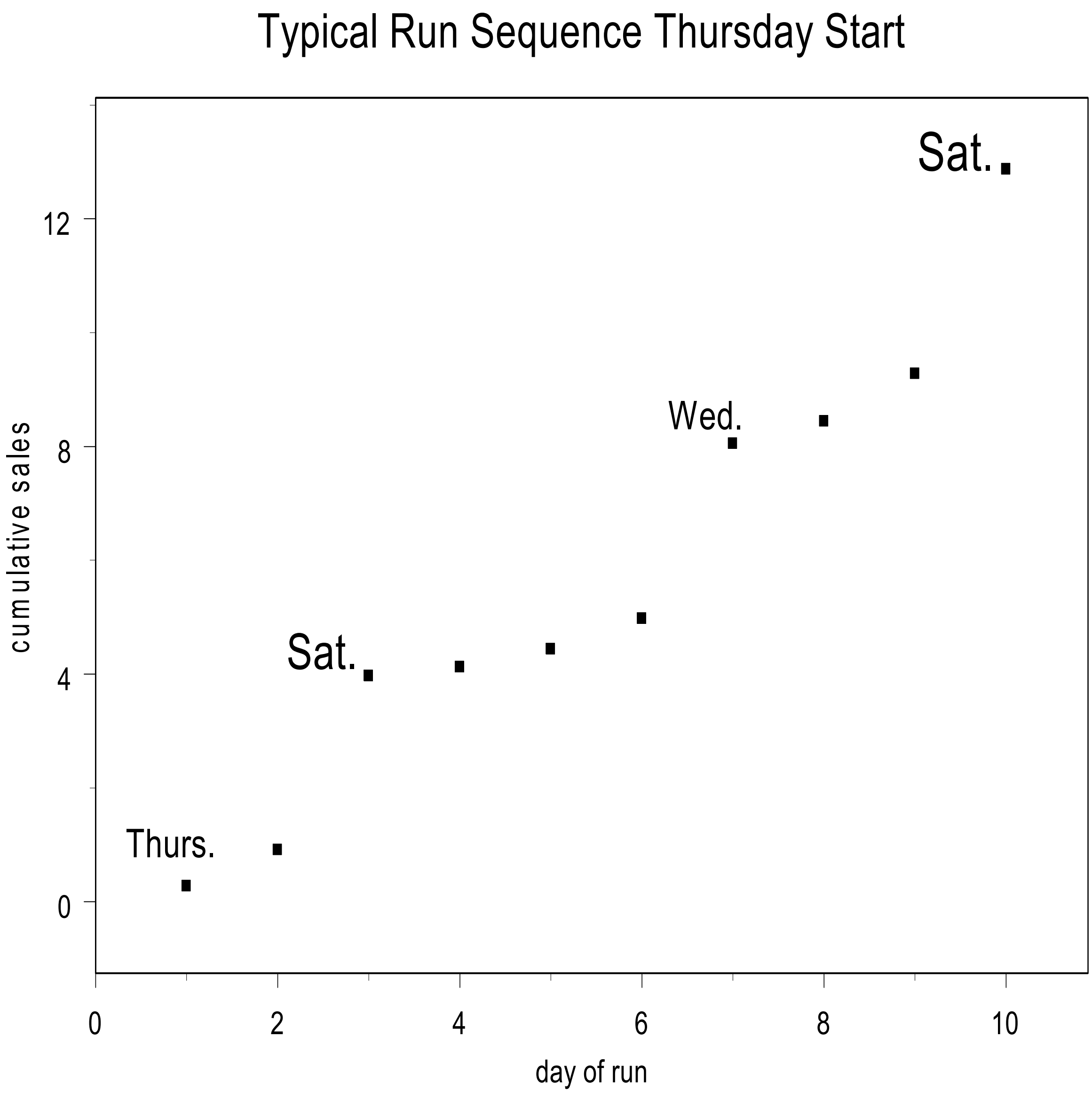
  $$M^{-1}\mathbf{u},$$

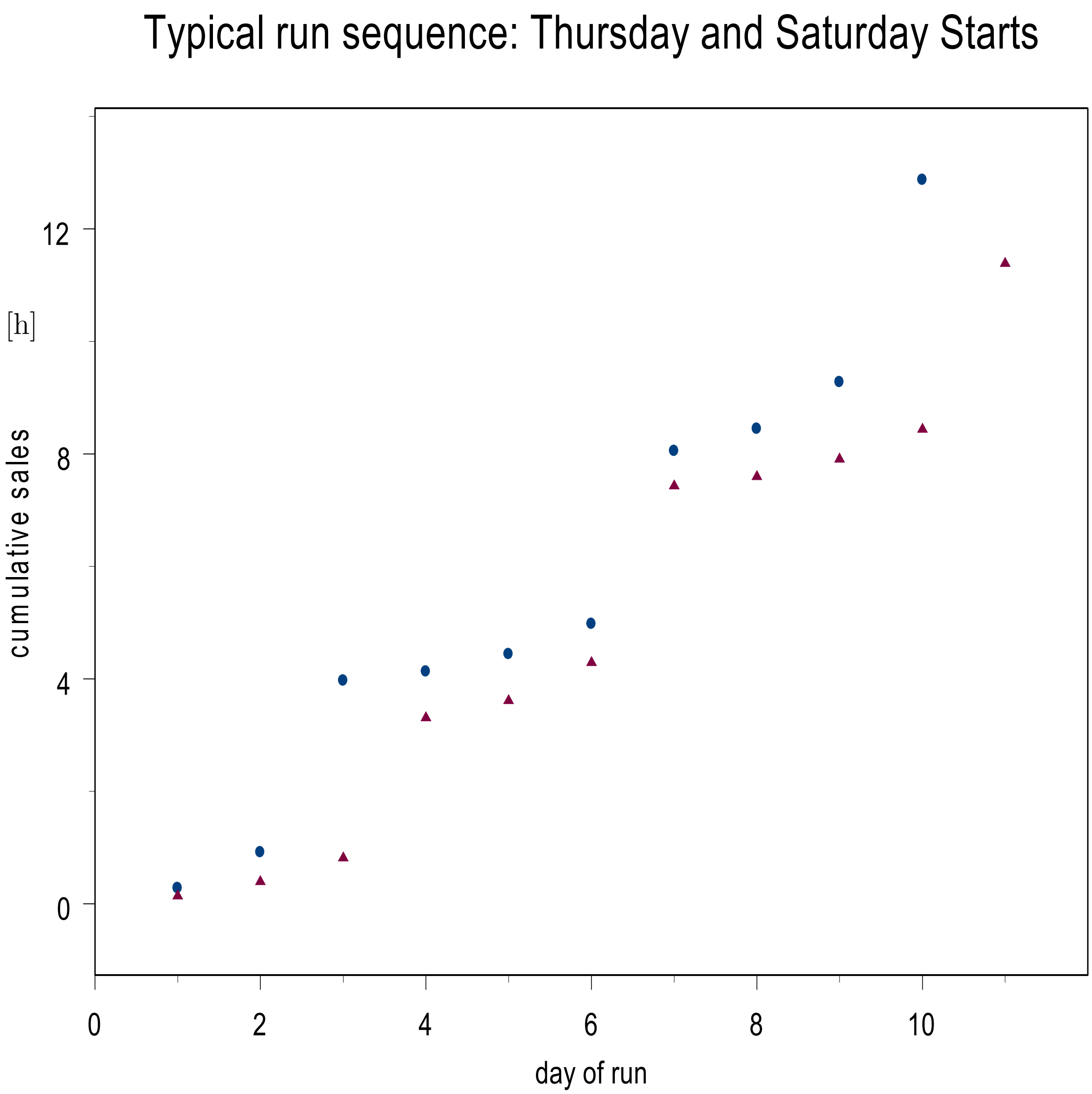
  $$QH^TM^{-1}\mathbf{u},$$

  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.

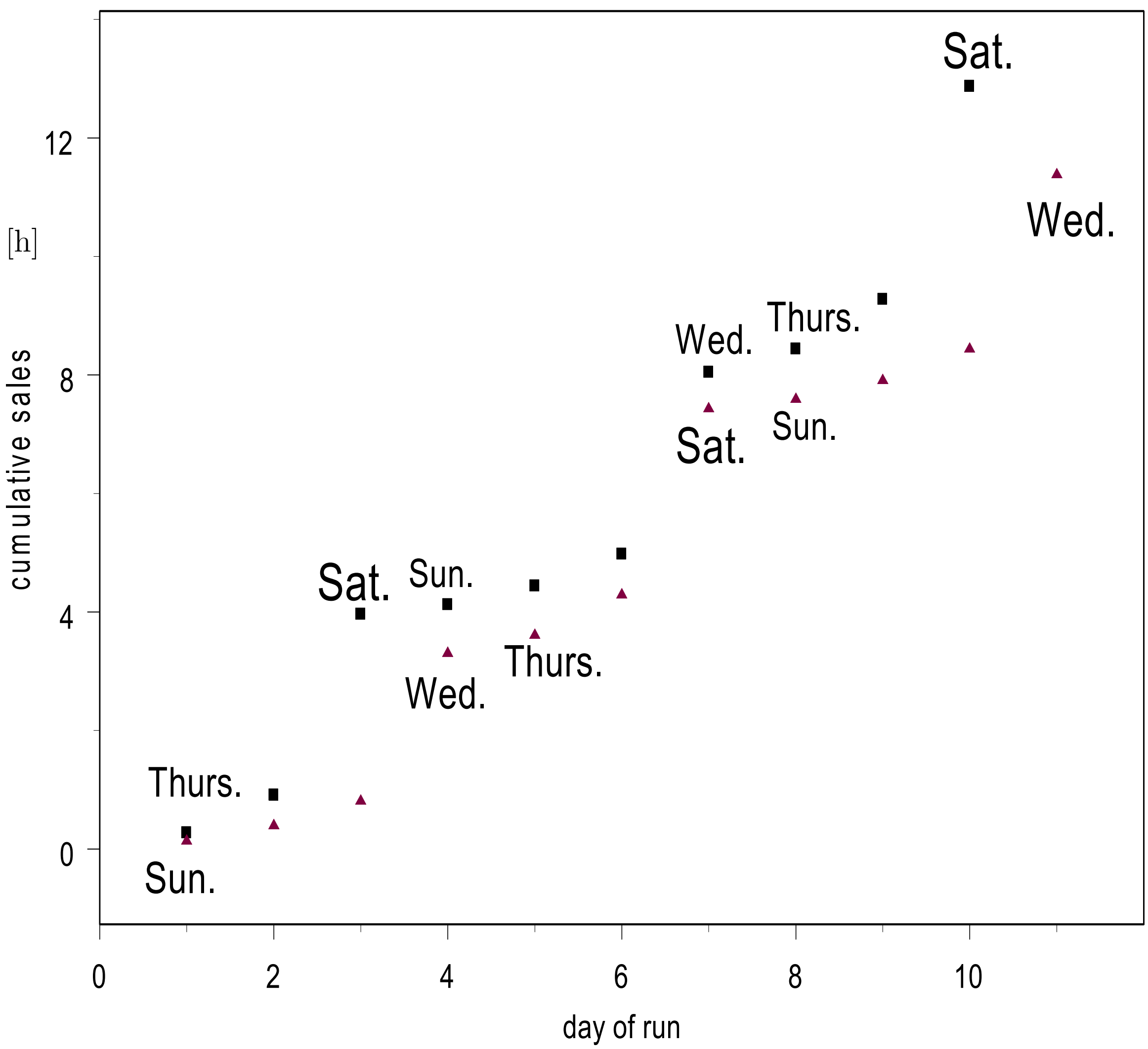
- C++ code available in March


Typical Run Sequence Thursday Start
cumulative sales
day of run
0
4
8
12
0
2
4
6
8
10
Thurs.
Sat.
Wed.
Sat.

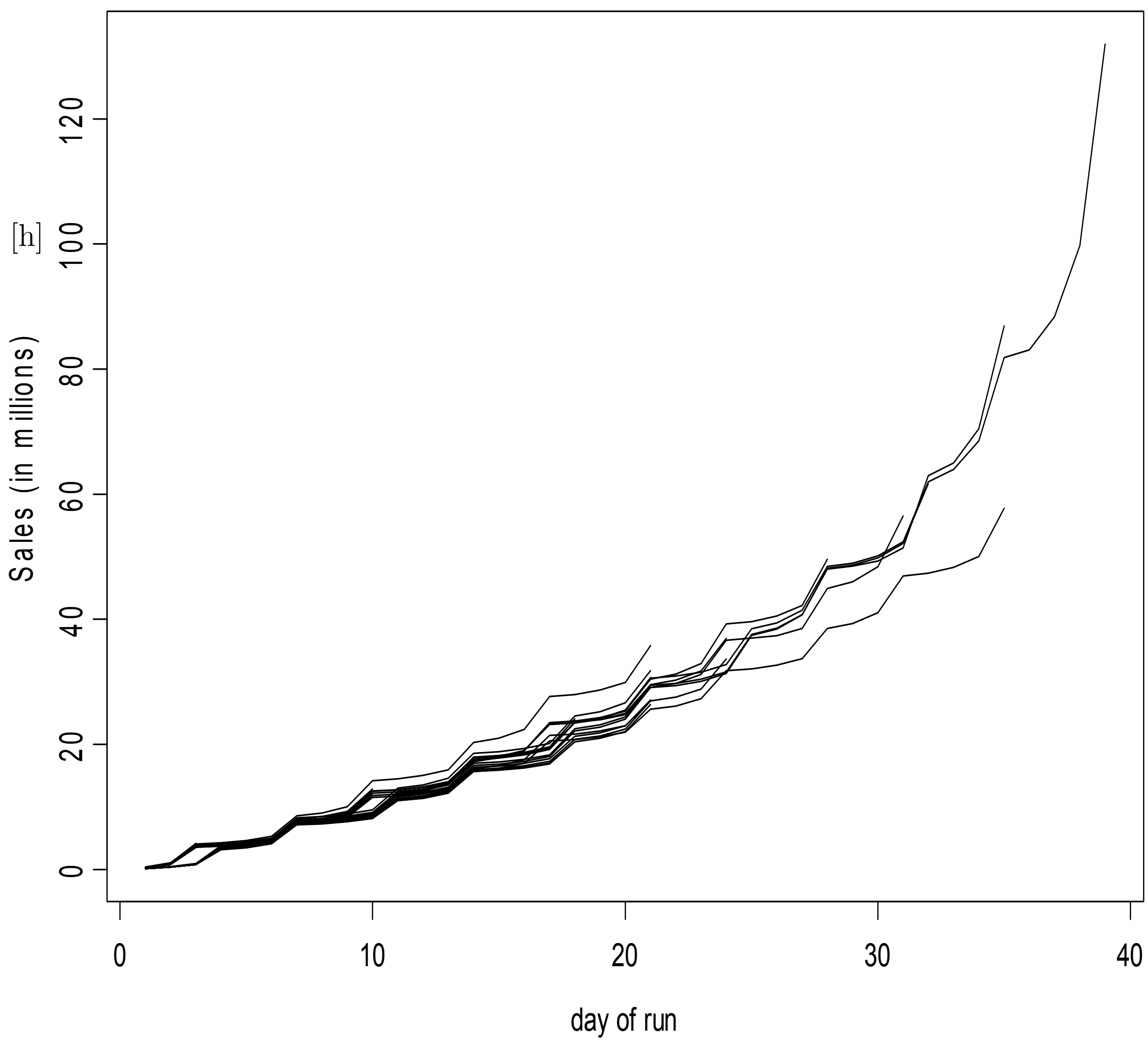


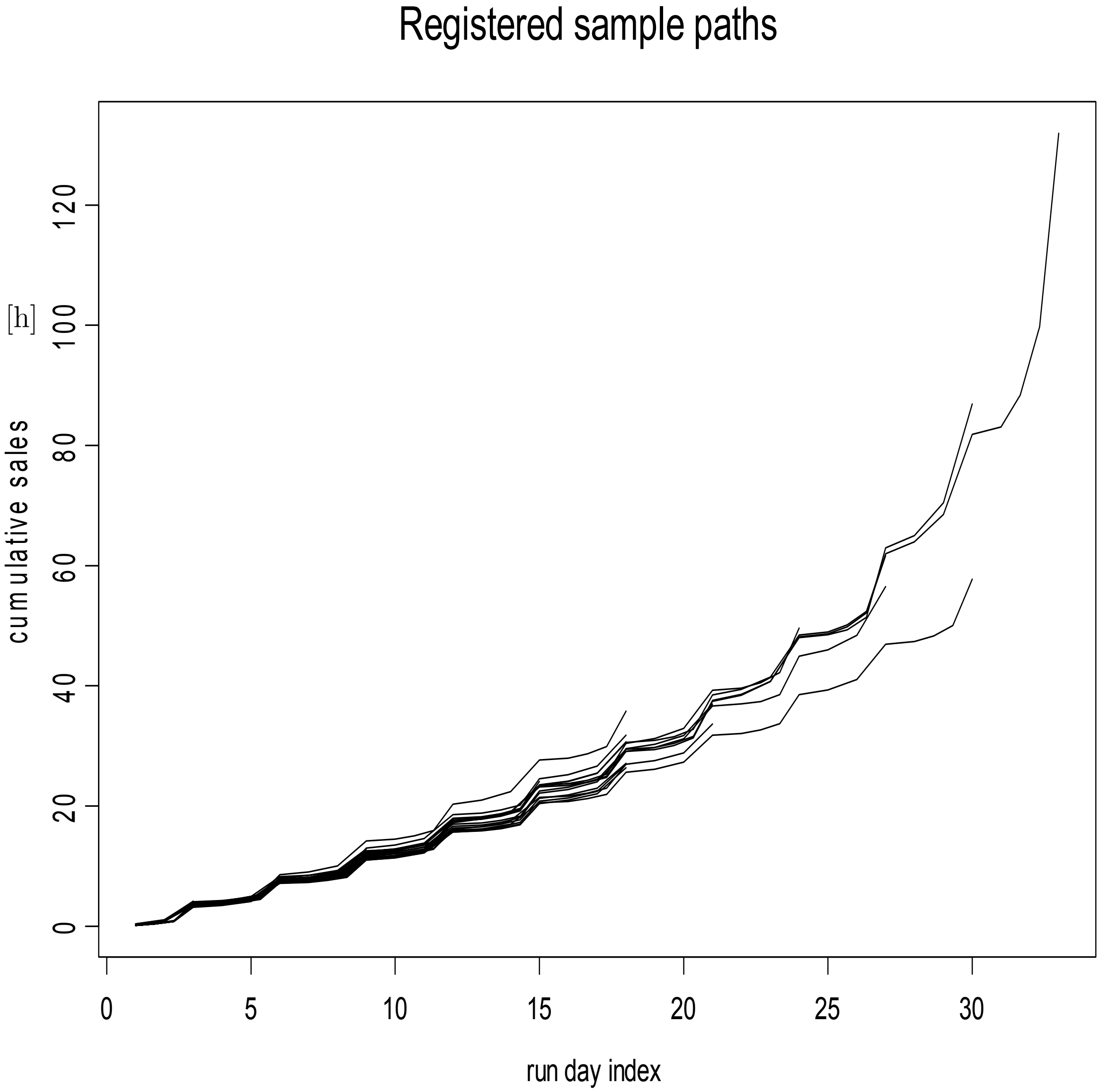
Typical run sequence: Thursday and Saturday Starts
[h]
cumulative sales
0
4
8
12
0
2
4
6
8
10
day of run

Typical Run Sequences: Thursday and Saturday Starts

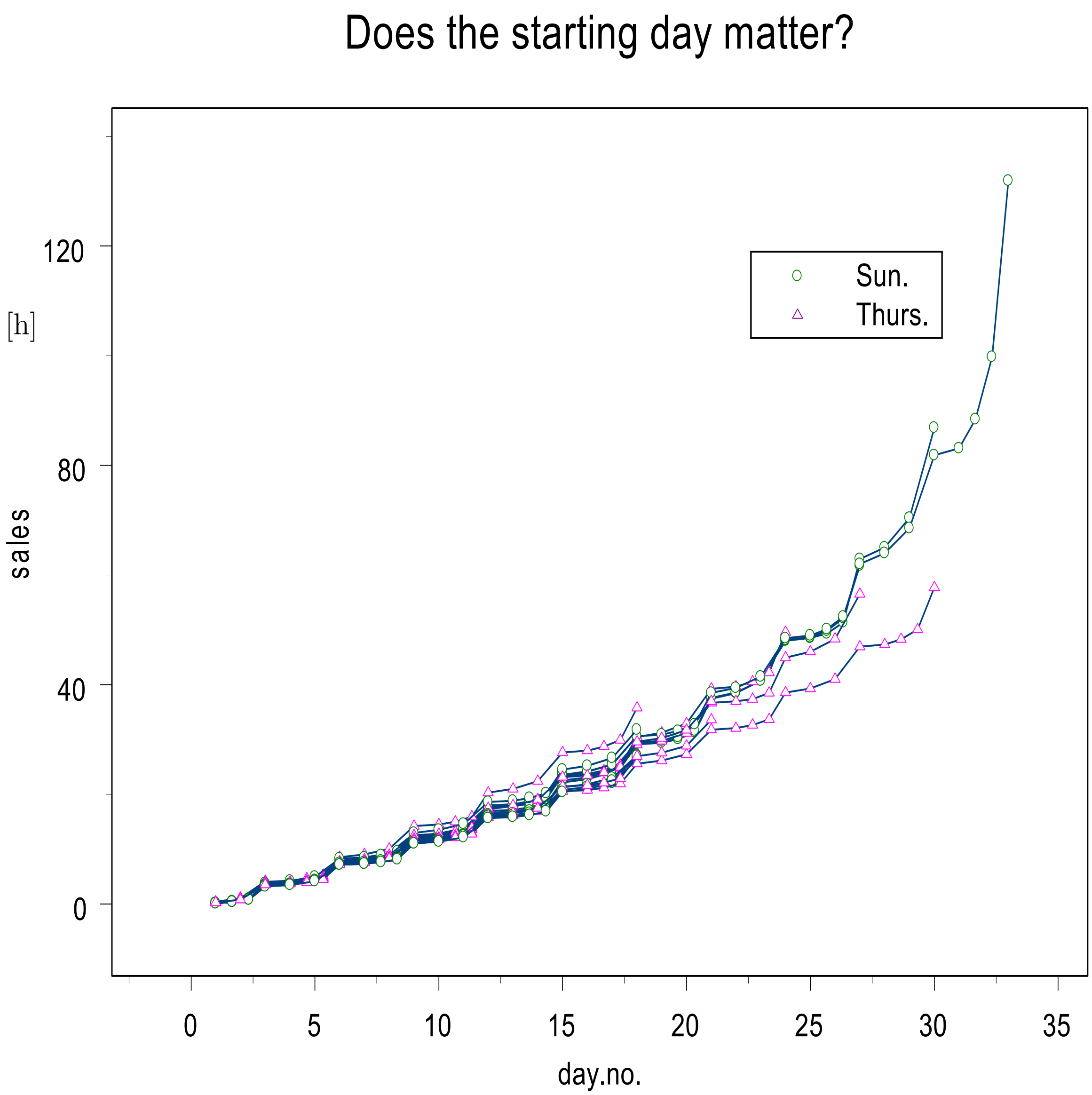


# Sample Paths for 24 Lotto Texas Runs



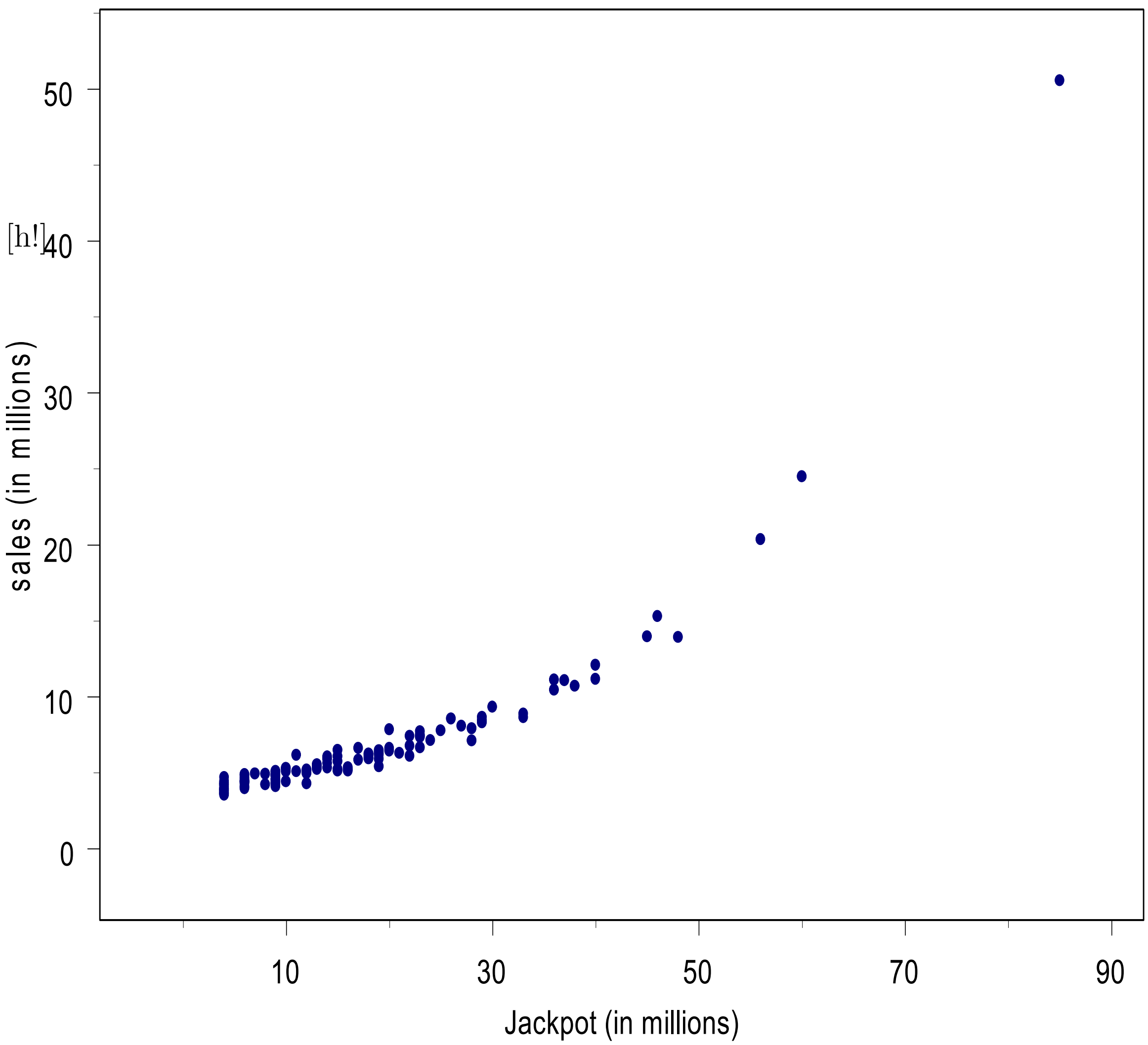


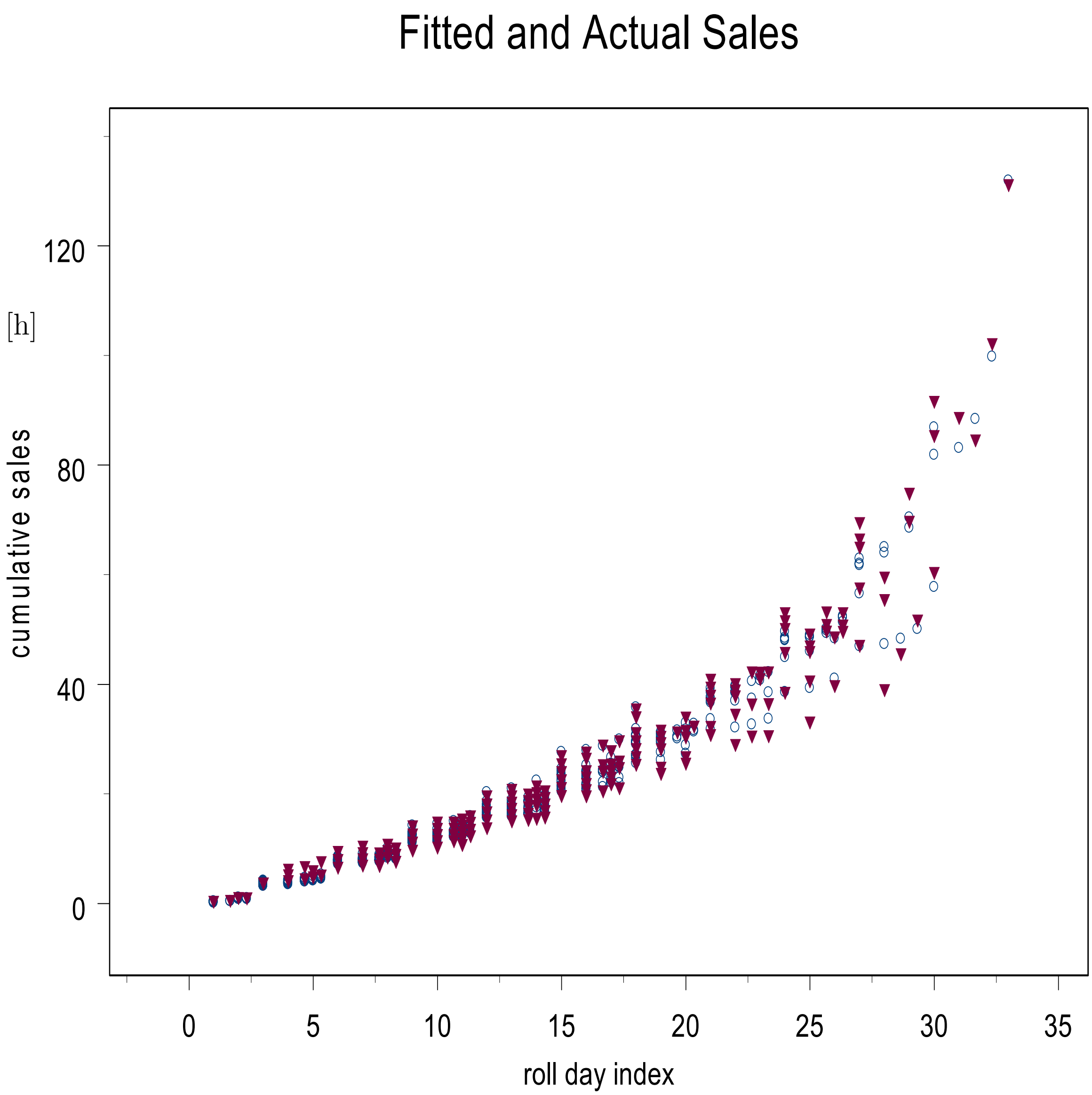
Registered sample paths
[h]
cumulative sales
0
20
40
60
80
100
120
0
5
10
15
20
25
30
run day index



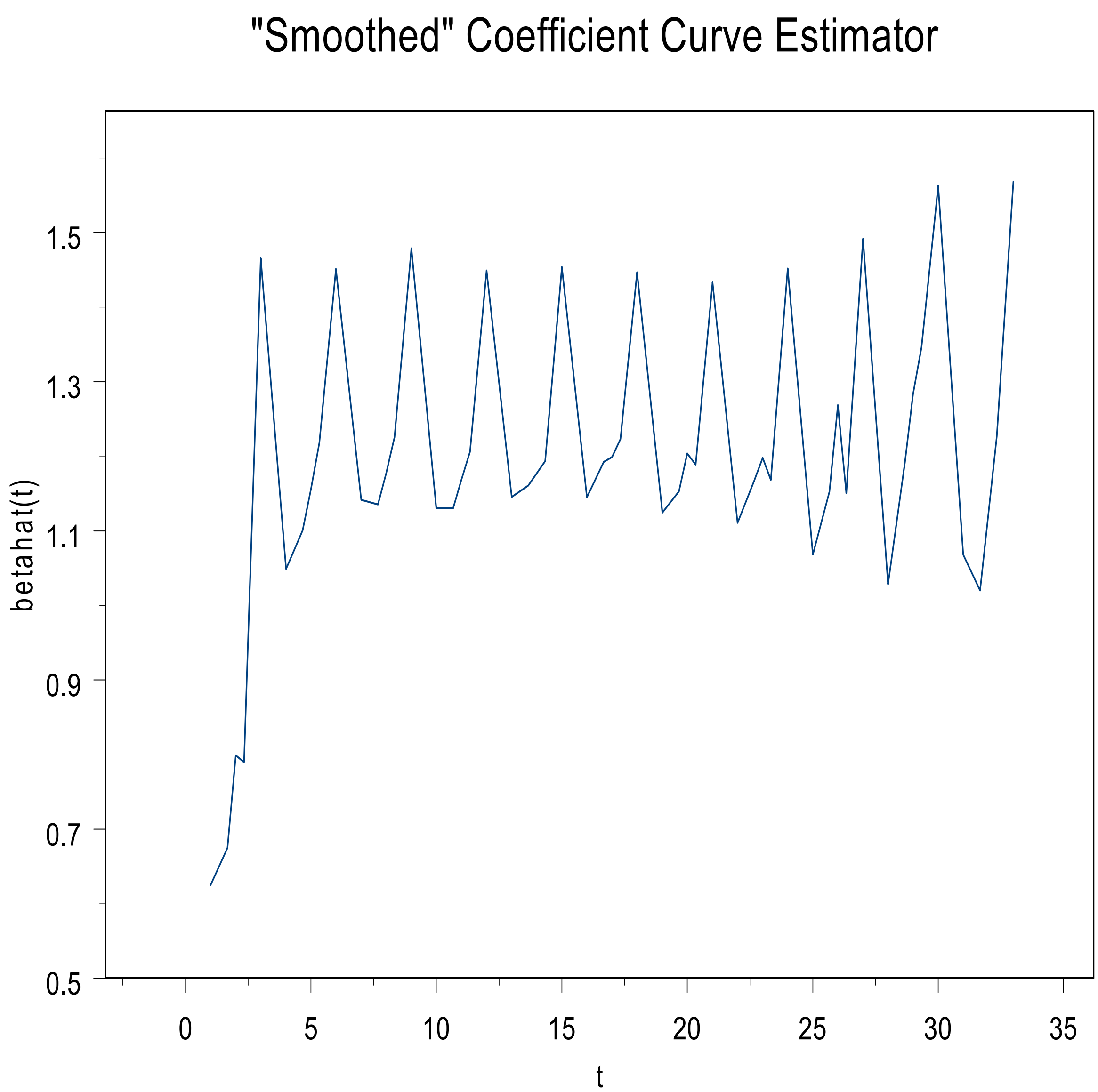
Does the starting day matter?
Sun.
Thurs.
[h]
sales
day.no.
0
40
80
120
0
5
10
15
20
25
30
35

# Sales versus Jackpot for Lotto Texas

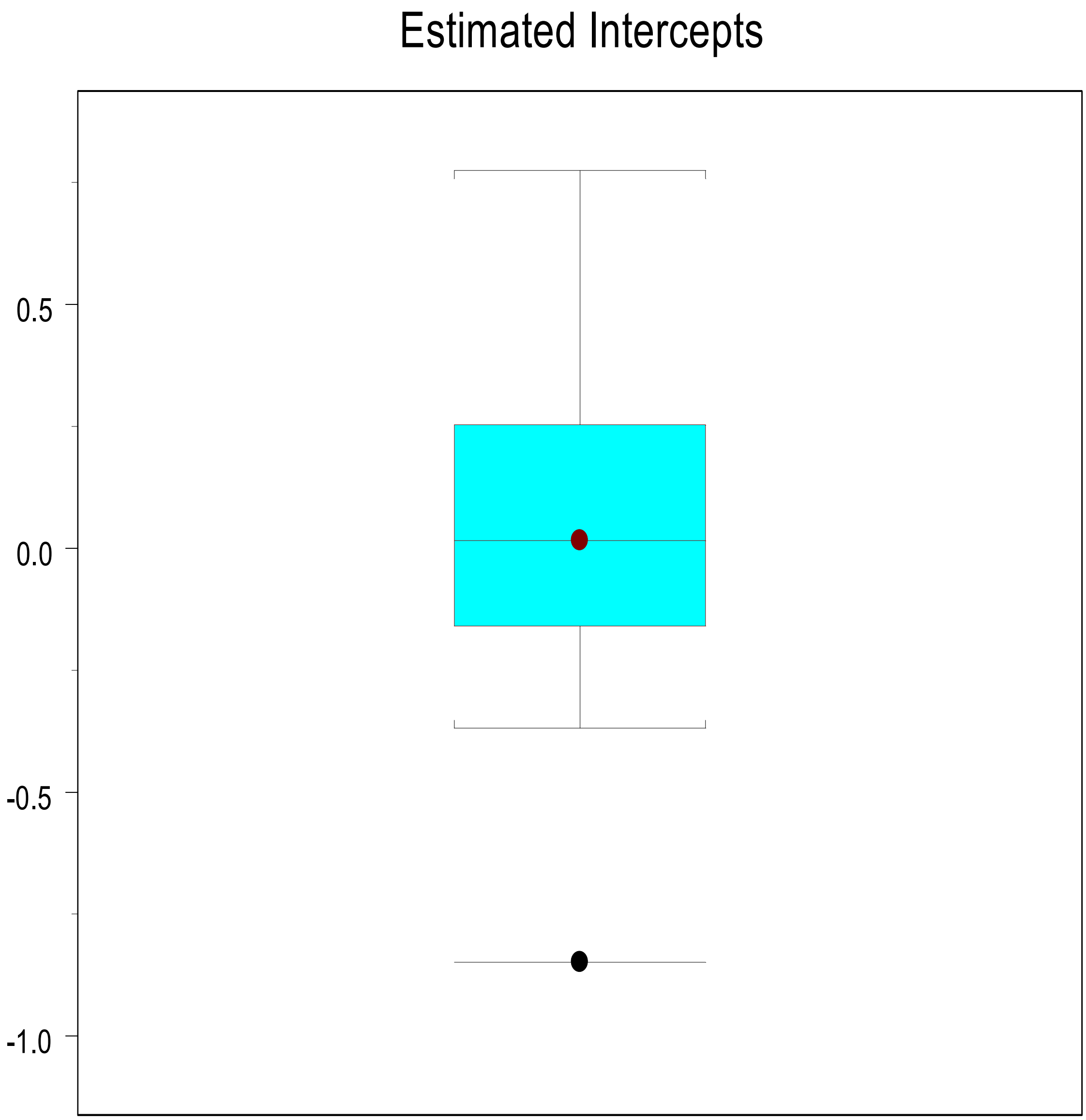


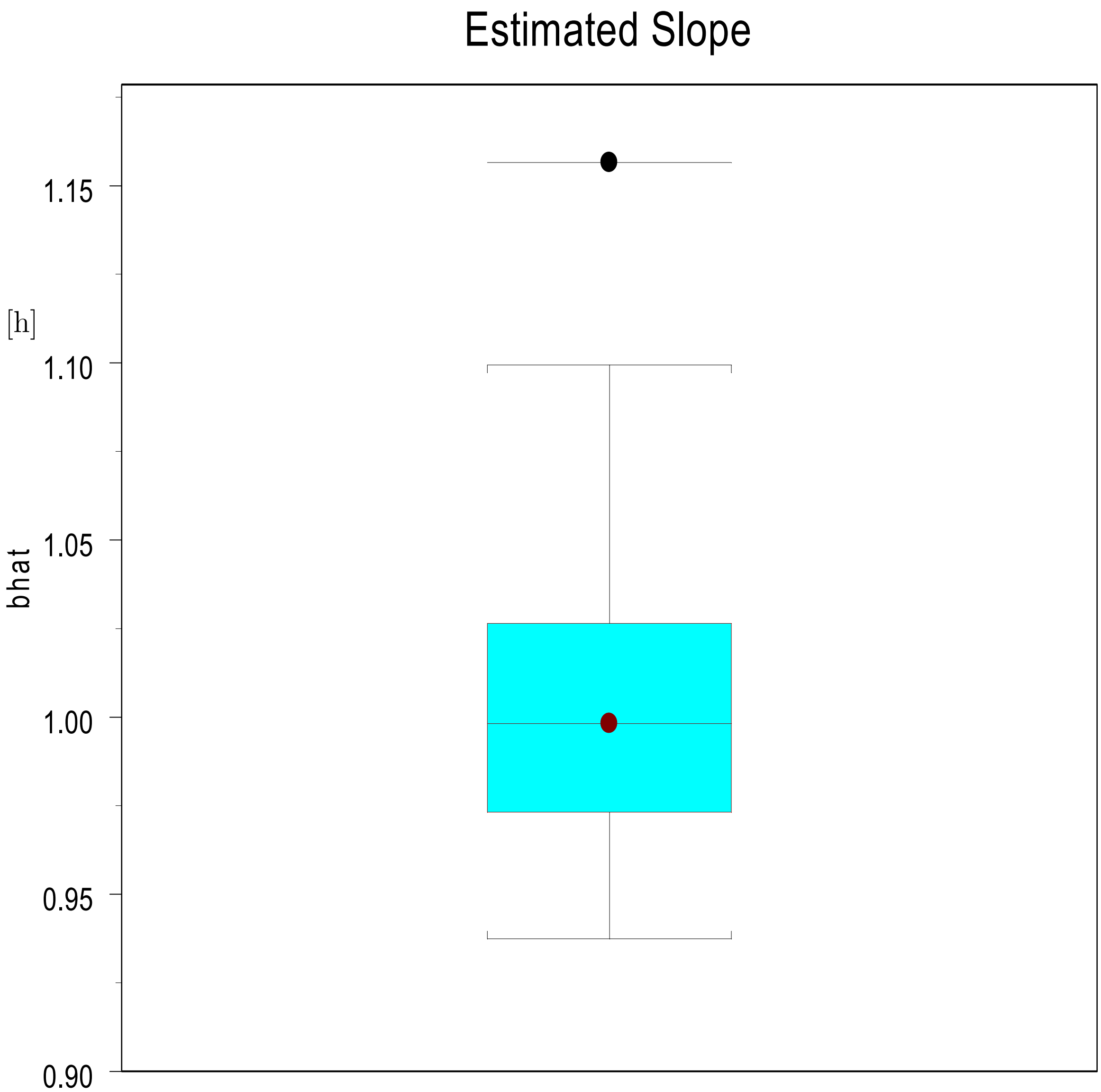


Fitted and Actual Sales
[h]
cumulative sales
0
40
80
120
0
5
10
15
20
25
30
35
roll day index

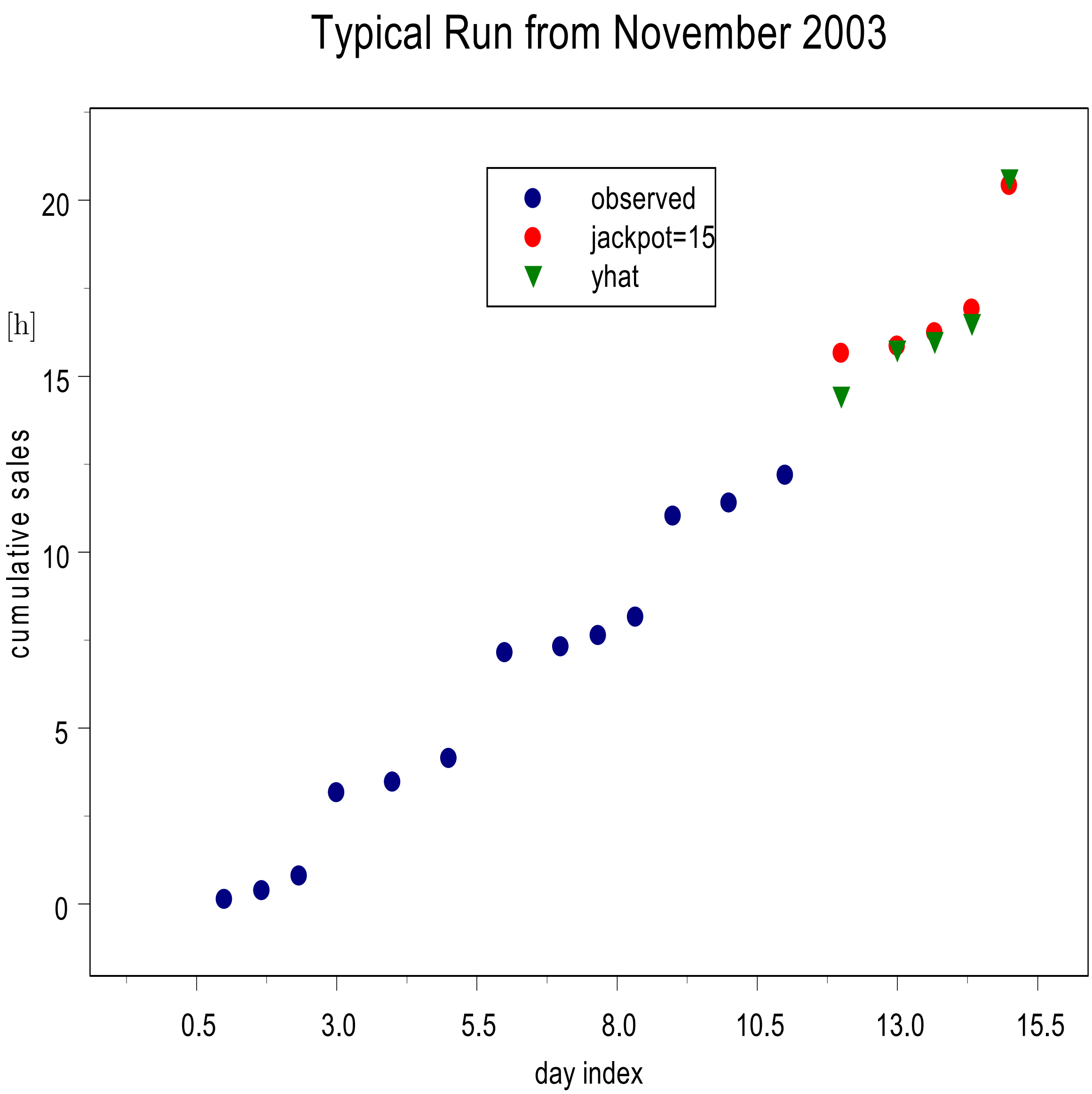

"Smoothed" Coefficient Curve Estimator
betahat(t)
1.5
1.3
1.1
0.9
0.7
0.5
0
5
10
15
20
25
30
35
t

[h]

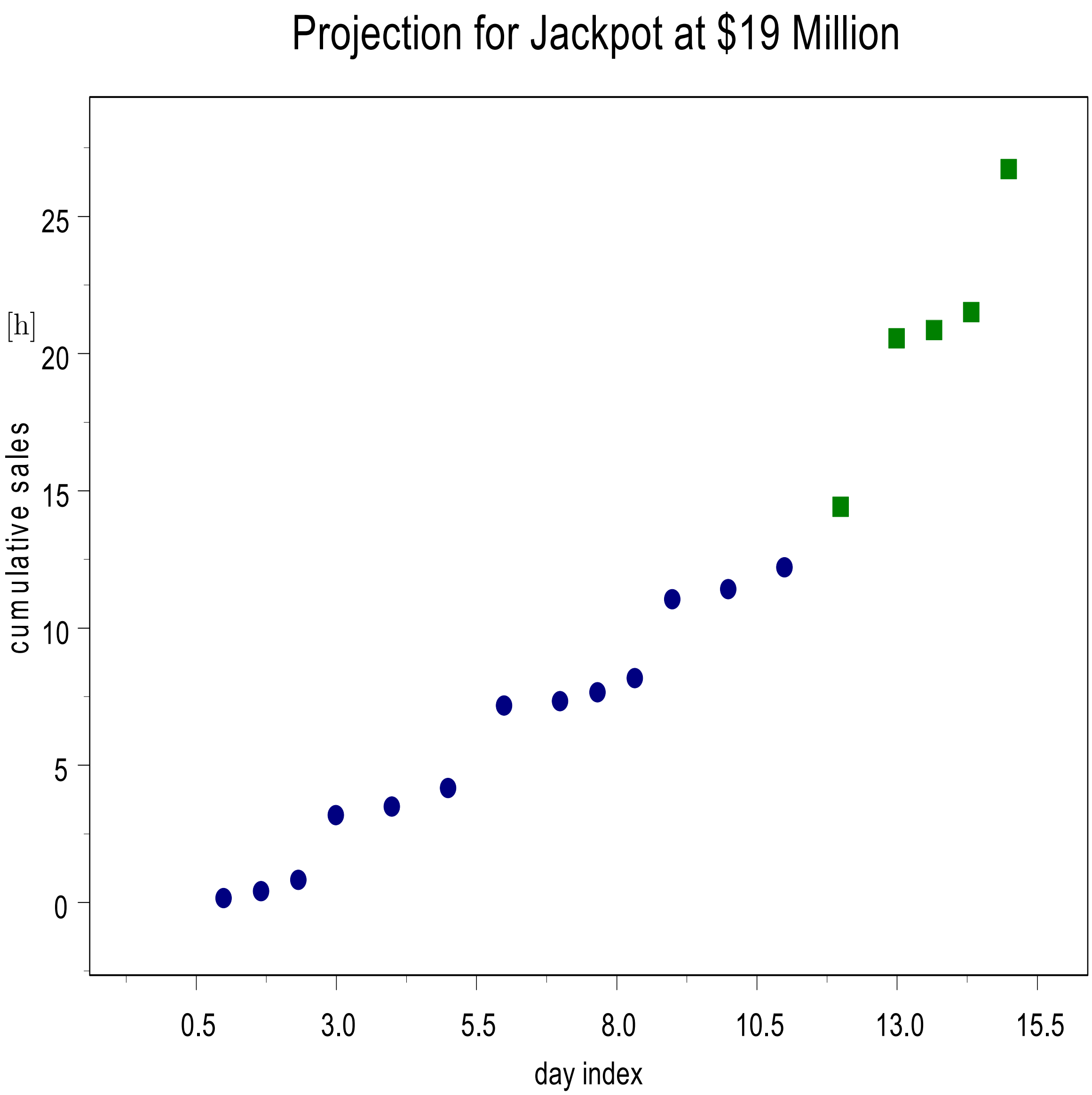




Estimated Slope
[h]
bhat
1.15
1.10
1.05
1.00
0.95
0.90



Typical Run from November 2003
observed
jackpot=15
yhat
[h]
cumulative sales
0
5
10
15
20
0.5
3.0
5.5
8.0
10.5
13.0
15.5
day index


Projection for Jackpot at $19 Million
[h]
cumulative sales
0
5
10
15
20
25
0.5
3.0
5.5
8.0
10.5
13.0
15.5
day index