Analyzing and Optimizing Revenue Using Bayesian Hierarchical Models

Bryan Whiting

December 9, 2014

1 Introduction

Golf courses have difficulty in being able to assess what their customers are willing to pay for a given tee time. A local start-up has recently ventured to maximize profits by setting the right price for the right tee time. Traditionally, fixing prices has been a "gut" feeling, but the courses would like to used analytics to fix better prices, understand their customers, and optimize revenue.

Since $Revenue = Price * Quantity$, to maximize revenue, golf courses must establish the optimal price. Profitability is also established by the formula $Profit = Revenue$ - $Cost$, and should be predicted. With maximizing overall revenue of the golf course as the main purpose, the following questions are the focus of this analysis:

- What revenue can we expect at certain times?
- What tee times are "good" and which aren't?
- How does changing price affect Quantity and Revenue?

The goal of this analysis is to answer these questions using the posterior distributions $\pi(\theta|\underline{y})$ and $f(\hat{y}|y)$.

A tee time is the minute of the day that a golfer begins to play a round. A North American golf course tracks their tee-time bookings using "tee sheets" and data such as what price was charged and how many players played during that time. In total, the data in this analysis represent eight years of data over 2761 days and 48,765 tee times. There are 86 tee times that range between 6:20am and 8:30pm, usually in 10 minute increments. Furthermore, there can be 0-4 players for each tee time.

Critical to this data is that not all tee times are played every day. Times are sometimes not filled, and depending on sunlight and weather, some times aren't available for play. The following is a sample of tee times, average revenue at that time, and the number of observations.

Hour.Min	6.20	6.30	7.50	8.00	12.50	13.00	13.10	17.50	18.00	19.50	20.00
Avg.Rev	100.14	93.76	119.96	152.51	139.42	146.56	143.77	43.05	45.81	42.22	57.80
Num.Obs	15.00	124.00	443.00	646.00	808.00	945.00	839.00	444.00	513.00	45.00	16.00

2 Model

Choice of Model: Gleaning insight from an expert in pricing tee times, we gained intuition behind how our model should be framed. First, each tee time has its own revenue because demand varies at different times in the day. Secondly, the tee times have a different variance within each tee time as opposed to between each tee time. Lastly, there is an inherent correlation between tee times, as in all time-series data.

To represent this intuition, we will model θ_i to be the true mean revenue of each tee time, σ^2 to be the between-tee-time variance (variance of the population of tee times), and τ^2 to be the within-tee-time variance (variance around the true revenue per time). Revenue for each ith tee time and j observation is represented by Y_{ij}, and follows a normal distribution. The mean revenue for each ith time is represented by θ_i. Since the times are correlated with each time, we center each θ_i time at each θ_{i-1} time. Furthermore,

there are $i = 1, ..., k = 86$ tee times and $j = n_i$ observations for each time. Lastly, since σ^2, and τ^2 must be greater than zero, the inverse gamma distribution works well because of its conjugate properties.

Using Bayesian Hierarchical models, we assume the following likelihood and priors for our data.

$$Y_{ij} \sim N(\theta_i, \sigma^2) \quad -\infty < \theta_i < \infty, \sigma^2 > 0 \qquad \sigma^2 \sim IG(a_\sigma, b_\sigma) \quad a_\sigma, b_\sigma > 0 \tag{1}$$

$$\theta_i \sim N(\theta_{i-1}, \tau^2) \quad \tau^2 > 0 \qquad \tau^2 \sim IG(a_\tau, b_\tau) \quad a_\tau, b_\tau > 0 \tag{2}$$

$$\theta_0 \sim N(m, s^2) \quad -\infty < m < \infty, s^2 > 0 \tag{3}$$

Values for $a_\sigma, b_\sigma, a_\tau, b_\tau, m, s^2$ and intuition behind θ_0 are discussed below.

Under this model, the likelihood of the data and posterior distributions are derived under the following formulas.

$$f(\underline{y}|\underline{\theta}, \sigma^2) = \prod_{i=1}^{k} \prod_{n=1}^{n_i} (2\pi\sigma^2)^{-\frac{1}{2}} exp \frac{(y_{ij} - \theta_i)^2}{-2\sigma^2} \tag{4}$$

$$\pi(\underline{\theta}, \sigma^2, \tau^2|\underline{y}) \propto f(\underline{y}|\underline{\theta}, \sigma^2)\pi(\theta_i|\theta_{i-1}, \tau^2)\pi(\tau^2)\pi(\sigma^2)\pi(\theta_0) \tag{5}$$

Deriving Complete Conditionals: In order to get posterior draws of our parameters using MCMC, we need to derive the complete conditionals and use Gibb's Sampling. Under Bayes rule, the complete conditionals for any parameter γ follows the formula $[\gamma] \propto f(\underline{y}|\gamma)\pi(\gamma)$. Therefore, we derive the complete conditionals for θ_i, σ^2,and τ^2. Starting with θ_i, we see that because it as a normal likelihood and a normal prior, the posterior has a conjugate, normal distribution.

$$[\theta_i] \propto f(\underline{y}|\underline{\theta}, \sigma^2)\pi(\theta_i|\theta_{i-1}, \tau^2) \tag{6}$$

$$\propto \prod_{i=1}^{k} \prod_{n=1}^{n_i} N_{y_{ij}}(\theta_i, \sigma^2) \prod_{i=1}^{k} N_{\theta_i}(\theta_{i-1}, \tau^2) \tag{7}$$

$$\propto \prod_{j=1}^{n_i} exp \frac{(y_{ij} - \theta_i)^2}{-2\sigma^2} \exp \frac{(\theta_i - \theta_{i-1})^2}{-2\tau^2} \exp \frac{(\theta_{i+1} - \theta_i)^2}{-2\tau^2} \dots \exp \frac{(\theta_k - \theta_{k-1})^2}{-2\tau^2} \tag{8}$$

$$[\theta_i] \propto N(\frac{\sigma^2(\theta_{i-1} + \theta_{i+1}) + \bar{y} n_i \tau^2}{2\sigma^2 + \tau^2 n_i}, \frac{\tau^2\sigma^2}{2\sigma^2 + \tau^2 n_i}) \tag{9}$$

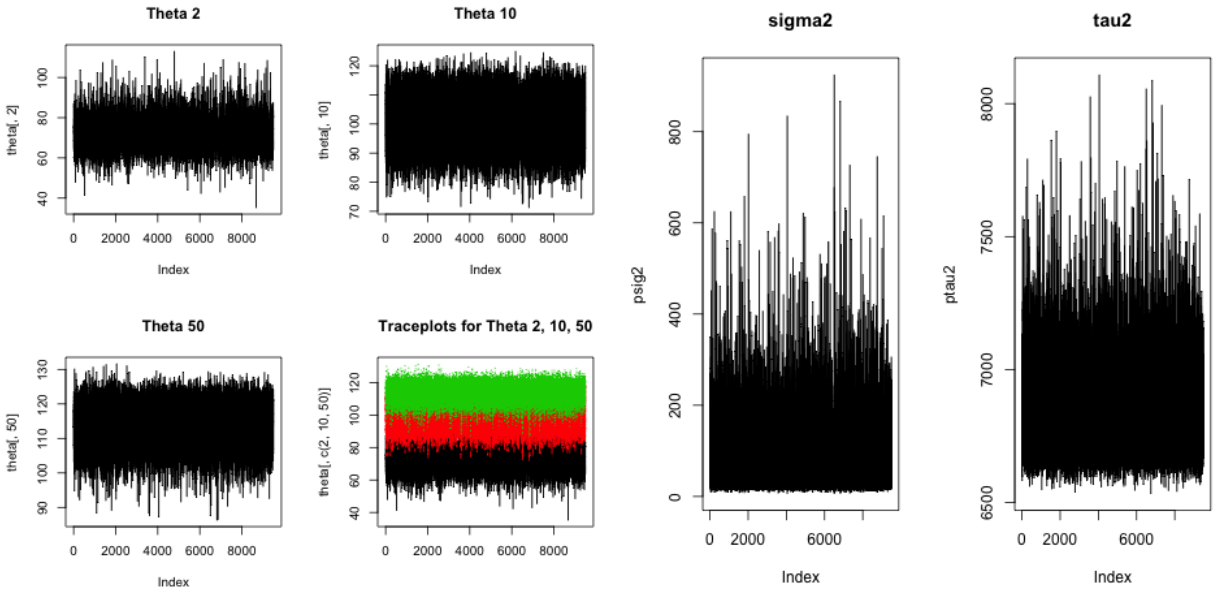
For the variance terms, complete conditional of σ^2, τ^2 are conjugate to their priors. That is,

$$[\sigma^2] \propto IG\left(a_\sigma + \frac{1}{2}\sum_{i=1}^{k} n_i, \left[\frac{1}{b_\sigma} + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \theta_i)^2\right]^{-1}\right) \tag{10}$$

$$[\tau^2] \propto IG\left(a_\tau + \frac{k}{2}, \left[\frac{1}{b_\tau} + \frac{1}{2}\sum_{i=1}^{k}(\theta_i - \theta_{i-1})^2\right]^{-1}\right) \tag{11}$$

*Assumptions & Interpretation:*By the construction of the model, we are assuming several things: (1) because $y_{ij} \sim N(\theta_i, \sigma^2)$, that implies tee times are similar. That is, no other effects such as day of week, month, weather, etc. affect the revenue; (2) τ^2 and σ^2 are independent, and τ^2 is uncorrelated across tee times. We believe this to be true because knowing the variance within tee times should yield no information about the variance between tee times; (3) we assume θ_i is correlated with θ_{i-1} and θ_{i+1}. By construction, each θ_i comes from a normal with mean as a function $f(\theta_{i-1}, \theta_{i+1})$. Therefore, by imposing this constraint, we account for the correlation in revenue between tee times; and (4) because we have indexed-based draws for θ_i, special caution with the model needs to be considered for the first and last times θ_1 and θ_k. Therefore, θ_0 is a randomly-generated starting point for the day, and could be interpreted as "the revenue for 6:10am." We used human judgement to model m and s^2 below. m would be interpreted as the average revenue of the tee time just before this course opens and s^2 would be the variance behind that mean. We note that $[\theta_k]$ lacks a θ_{k+1} term, and therefore has a different complete conditional.

Prior Values: For the hyper-parameters specified above, we modeled the prior distribution of σ^2 and τ^2 to identify good starting values. We believed that the variance between tee times would be greater than the



variance within tee times. Likewise, we believe that the variance within tee times would be rather small, considering the golf courses might generally charge the same fees over the years. Therefore, $\sigma^2 \sim IG(a_\sigma = 50, b_\sigma = 5)$ and $\tau^2 \sim IG(a_\tau = 15, b_\tau = 10)$. Lastly, choosing $m = 60$ with a low variance of 5.0 would imply a standard deviation of around 2.2. This would mean that we expect the starting revenue to vary between 53.3 and 67. This embeds enough uncertainty into our model.

3 Model Diagnostics and Convergence

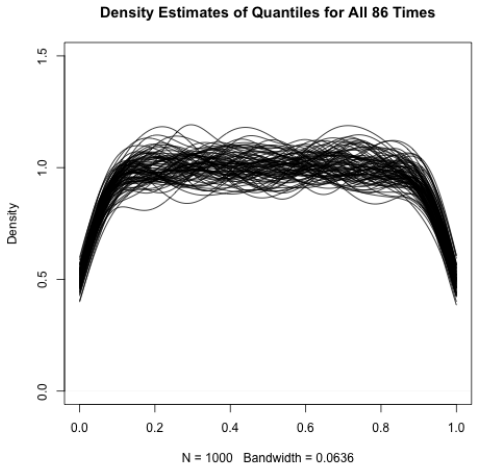
For our MCMC, we used took 10,000 draws and used a burn-in of 500, although a burn-in of 10 could have been sufficient. We can tell whether the model is a good one based off of the convergence diagnostics and the model of fit. We will review several convergence diagnostics and then review the posterior-predictive-checking goodness-of-fit test. *Trace Plots:* Trace plots that have stabilized are a good indication of whether a distribution has converged. Although not all can be represented here, this subset of trace plots of θ_i for i = 2, 10, and 50 demonstrates that the θ_i have converged. We can see further that the variance at θ_2 is lower than the variance at θ_{50}, which we expect considering we have less data at θ_2 and we do at the later times.

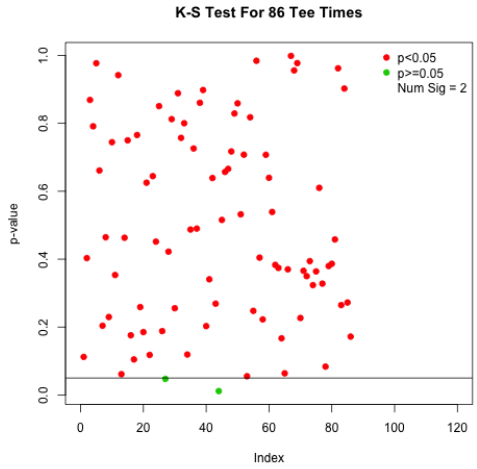
Raftery-Lewis Dependence Factors: A Raftery-Lewis evaluates the accuracy of the estimated or desired percentiles and reports the number of samples to reach the desired accuracy of the percentiles and reports a dependence factor. Dependence factors measure the deviation from posterior sample independence, and a value of 1 would indicate that the draws have reached independence. Since all of our draws are random draws, we would expect them to all be low. In our analysis, all of the dependence factors were close to 1 and much less than 5 (the conventional limit for dependence factors).

Geweke Diagnostic: Under the Geweke Diagnostic, we took samples of the first 10% of the posterior draws *post burn-in* and compared the mean to the sample mean of the last 50% of the draws. Under H_0: Mean of first 10% = mean of last 50%, we obtained favorable p-values that would indicate only a few distributions didn't converge under this hypothesis.

Autocorrelation: Another convergence diagnostic is to assess the autocorrelation within the data do see how correlated the posterior draws are. Considering we are not necessarily concerned with correlation in the posterior draws, we are still pleased to see that many of the posterior draws showed low autocorrelation.

Goodness of Fit via Posterior Predictive Checking: We used posterior predictive checking to assess whether the model fits the data as specified. We compared the predictive distribution $f(\hat{y}|\underline{y}, \underline{\theta})$ with the data $\underline{y}_i$. We assess what the quantiles of the data are in the posterior predictive distribution. For each y_i, we assess whether the quantiles are uniformly distributed and use a K-S test of significance to see if the quantiles are uniformly distributed. We report that only two distributions lacked uniform quantiles. Therefore, we can conclude that our model is quite a good fit for the data.

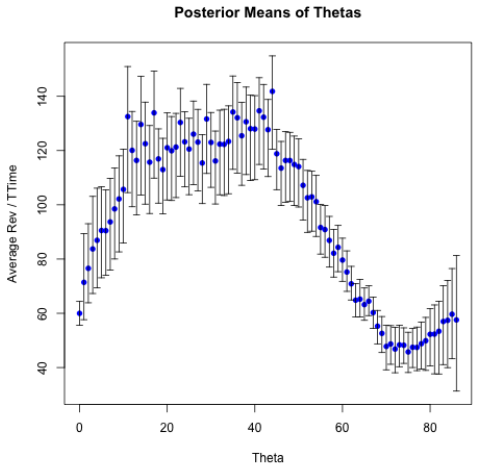


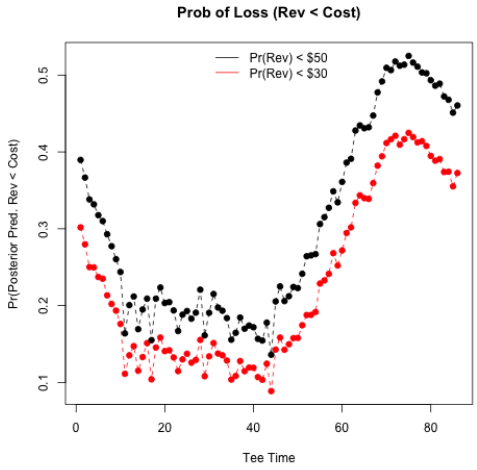


4 Results

Correlation in θs: The covariance and correlation plots demonstrated appropriate results and showed that there was a strong, positive correlation with the first terms and a strong, negative correlation with the later terms in the model. Considering that we would expect there to be a positive correlation in rising revenues earlier in the day and a negative correlation in falling revenues later in the day, we are pleased with the results.

Posterior Means and Predictive Revenue: The table for the posterior predictive means is given in the appendix, but a graphical representation is presented here along with HPD intervals. Furthermore, using the posterior predictive distribution of revenue, we can estimate the probability of a tee time given certain 'fixed' costs per round. Furthermore, and although outside the scope of this assignment, we were able to successfully identify revenue-maximizing prices at each tee time.





5 Conclusion

In conclusion, we were able to demonstrate that a Bayesian Hierarchical model was able to adequately model the revenues for individual tee times as well as provide a strong framework for being able to answer important questions such as how to maximize revenue and minimize loss. Generally, the model demonstrated strong convergence in trace plots as well as in other traditional diagnostics. Because of the strong nature of the model, we would feel confident in using this model to accurately predict future revenue for this golf course.

6 Appendix: Tables & Code

6.1 Sources

- MCMC Diagnostics resource: `http://www.people.fas.harvard.edu/~plam/teaching/methods/convergence/convergence_print.pdf`

6.2 Posterior Summary Statistics & Revenue Maximizing Prices

Table 1: Posterior Summary Statistics and Revenue-Maximizing Prices at Each Tee Time

	$E(\theta\|Y)$	.025%	.975%	$V(\theta\|Y)$	$\sqrt{Var(\theta\|Y)}$	$E(Rev\|P)$	Price	Time
θ_0	59.97	55.60	64.36	5.02	2.24	.	.	.
θ_1	71.41	57.65	89.37	61.79	7.86	62.22	43.21	6.2
θ_2	76.58	63.87	93.02	60.93	7.81	64.86	41.66	6.3
θ_3	83.72	67.30	103.10	102.54	10.13	70.73	42.61	6.4
θ_4	86.91	69.37	106.12	109.73	10.48	71.38	43.87	6.5
θ_5	90.52	73.09	106.62	90.28	9.50	73.77	45.48	7
θ_6	90.46	74.05	105.54	77.85	8.82	74.57	45.34	7.1
θ_7	93.69	75.98	109.73	91.70	9.58	77.76	45.59	7.2
θ_8	98.49	80.02	113.59	91.13	9.55	79.86	47.45	7.3
θ_9	102.15	82.64	118.04	103.10	10.15	83.4	46.7	7.4
θ_{10}	105.68	85.89	120.52	97.17	9.86	86.15	46.01	7.5
θ_{11}	132.52	104.49	150.97	186.48	13.66	106.63	52.52	8
θ_{12}	120.07	99.27	134.42	97.11	9.85	96.81	50.56	8.1
θ_{13}	116.37	96.27	130.78	94.22	9.71	94.09	48.16	8.2
θ_{14}	129.61	103.62	147.32	160.71	12.68	104.54	53.07	8.3
θ_{15}	122.51	100.24	137.85	115.69	10.76	98.09	46.68	8.4
θ_{16}	115.71	96.79	129.21	82.25	9.07	93.85	49.94	8.5
θ_{17}	133.91	109.83	149.26	129.94	11.40	106.57	53.05	9
θ_{18}	116.92	100.61	127.99	55.81	7.47	93.37	49.45	9.1
θ_{19}	112.97	96.71	124.51	59.49	7.71	90.34	47.4	9.2
θ_{20}	121.04	101.78	133.87	80.72	8.98	97	49.71	9.3
θ_{21}	119.92	101.58	132.54	75.81	8.71	95.6	50.35	9.4
θ_{22}	121.28	102.66	133.70	76.30	8.74	97.46	52.15	9.5
θ_{23}	130.36	110.47	142.91	85.16	9.23	103.94	53.18	10
θ_{24}	123.21	106.60	134.26	59.45	7.71	98.88	51.22	10.1
θ_{25}	120.54	103.73	131.90	61.35	7.83	96.88	49.95	10.2
θ_{26}	126.04	107.45	138.25	75.35	8.68	100.32	51.97	10.3
θ_{27}	123.10	105.15	135.19	71.35	8.45	98.76	50.37	10.4
θ_{28}	115.41	100.44	125.86	50.37	7.10	91.83	49.37	10.5
θ_{29}	131.64	111.55	144.40	89.28	9.45	106.07	54.78	11
θ_{30}	123.00	106.44	134.03	58.22	7.63	97.26	58.31	11.1
θ_{31}	116.17	100.29	127.19	56.37	7.51	93.31	49.01	11.2
θ_{32}	122.36	103.70	134.88	78.07	8.84	97.39	51.64	11.3
θ_{33}	122.27	103.43	135.22	79.60	8.92	98.34	49.69	11.4
θ_{34}	123.35	104.08	136.47	84.27	9.18	99.1	50.87	11.5
θ_{35}	134.20	113.07	147.49	95.07	9.75	106.64	54.52	12
θ_{36}	132.09	111.70	145.07	88.66	9.42	105.11	53.1	12.1
θ_{37}	125.45	107.19	137.72	72.18	8.50	100.61	51.2	12.2
θ_{38}	130.61	111.14	143.41	85.04	9.22	103.12	58.17	12.3
θ_{39}	128.05	108.96	140.44	78.40	8.85	103.53	53.55	12.4
θ_{40}	127.92	109.25	140.24	77.19	8.79	102.14	53.05	12.5
θ_{41}	134.65	114.87	146.86	83.68	9.15	106.73	55.7	13

θ_{42}	132.30	113.17	144.38	78.18	8.84	106.09	54.46	13.1
θ_{43}	127.68	110.40	138.82	63.50	7.97	102.14	50.62	13.2
θ_{44}	141.81	120.43	154.89	97.47	9.87	113.65	54.74	13.3
θ_{45}	118.81	105.58	127.76	35.37	5.95	95.29	49.98	13.4
θ_{46}	113.52	99.70	123.36	41.91	6.47	90.65	49.36	13.5
θ_{47}	116.35	100.96	126.95	52.74	7.26	93.67	48.44	14
θ_{48}	116.34	101.22	126.72	50.23	7.09	93.73	50.11	14.1
θ_{49}	114.86	99.71	125.33	50.77	7.13	91.04	50.08	14.2
θ_{50}	114.08	99.18	124.30	48.76	6.98	92.03	46.38	14.3
θ_{51}	107.18	94.34	116.75	37.43	6.12	86.89	49.17	14.4
θ_{52}	102.58	89.84	112.66	38.16	6.18	82.72	47.7	14.5
θ_{53}	102.93	90.27	112.36	35.87	5.99	82.93	46.79	15
θ_{54}	101.15	88.33	110.82	37.09	6.09	82.87	47.75	15.1
θ_{55}	91.63	81.83	100.15	23.51	4.85	75.2	44.8	15.2
θ_{56}	90.87	80.66	99.70	25.77	5.08	73.62	44.03	15.3
θ_{57}	86.87	77.03	95.72	24.19	4.92	71.51	42.83	15.4
θ_{58}	82.14	73.35	90.93	21.25	4.61	68.9	44.32	15.5
θ_{59}	84.32	75.34	92.39	20.46	4.52	70.36	44.52	16
θ_{60}	79.62	71.79	87.69	16.91	4.11	66.63	44.46	16.1
θ_{61}	75.22	68.08	83.01	14.79	3.85	63.84	41.97	16.2
θ_{62}	70.87	64.91	77.30	9.99	3.16	61.84	39.95	16.3
θ_{63}	64.83	58.68	70.98	9.72	3.12	56.85	40.38	16.4
θ_{64}	65.21	58.63	72.47	12.30	3.51	55.97	39.24	16.5
θ_{65}	63.21	57.42	69.34	9.07	3.01	56.42	37.66	17
θ_{66}	64.46	59.14	70.12	7.75	2.78	56.49	39.44	17.1
θ_{67}	60.26	54.41	65.96	8.46	2.91	54.58	38.34	17.2
θ_{68}	55.25	48.73	61.06	9.95	3.15	50.82	39.43	17.3
θ_{69}	52.60	45.53	58.81	11.81	3.44	48.9	37.65	17.4
θ_{70}	47.79	39.17	55.55	18.55	4.31	47.16	38.46	17.5
θ_{71}	48.68	41.20	55.46	13.62	3.69	47.41	37.65	18
θ_{72}	46.83	38.08	54.77	18.67	4.32	45.55	38.27	18.1
θ_{73}	48.36	40.30	55.58	15.01	3.87	47.24	36.13	18.2
θ_{74}	48.22	41.53	54.63	11.39	3.37	46.63	36.17	18.3
θ_{75}	45.76	38.16	53.01	15.11	3.89	45.12	37.85	18.4
θ_{76}	47.47	40.00	54.42	14.07	3.75	46.31	36.28	18.5
θ_{77}	47.43	38.87	54.92	17.71	4.21	46.15	35.09	19
θ_{78}	48.79	39.43	56.76	19.93	4.46	47.44	35.88	19.1
θ_{79}	49.88	38.95	58.49	25.63	5.06	47.97	39.01	19.2
θ_{80}	52.27	40.56	61.70	29.12	5.40	49.37	37.43	19.3
θ_{81}	52.32	37.68	63.08	42.06	6.49	49.58	37.48	19.4
θ_{82}	53.38	37.63	64.41	46.53	6.82	49.86	36.15	19.5
θ_{83}	56.98	41.08	70.22	49.65	7.05	52.18	38.78	20
θ_{84}	57.41	39.97	72.14	59.40	7.71	51.89	37.1	20.1
θ_{85}	59.66	43.24	76.50	62.45	7.90	53.95	40.47	20.2
θ_{86}	57.56	31.32	81.26	140.75	11.86	52.72	38.66	20.3
τ^2	100.37	14.10	325.11	8444.68	91.89	.	.	.
σ^2	6818.45	6607.19	7314.53	38250.55	195.58	.	.	.

6.3 R Code

```
##############################################
# Author: Bryan Whiting
# Created: Nov 4, 2014
# Class: Stat 651, Bayes
# Assgnmt: Final Project
# Description: Estimate revenue using variables of interest.
##############################################

##############################################
# Packages and directories
##############################################
setwd("~/stat651-bayes/final-project")
library(MASS) #kde2d function
library(xtable)
library(sqldf)
library(ggplot2)
library(Hmisc) #errbar function
library(corrplot) # for graphing the covariance matrix
library(coda) #mcmc diagnostics, rf, autocorr

##############################################
# Resources
##############################################
# MCMC diagnostics
system("open http://www.people.fas.harvard.edu/~plam/teaching/methods/
    convergence/convergence_print.pdf")
# Coda package - MCMC diagnostics, etc
system("open http://cran.r-project.org/web/packages/coda/coda.pdf")
# MCMC diagnostics interpreted:
system("open http://support.sas.com/documentation/cdl/en/statug/63033/HTML/
    default/viewer.htm#statug_introbayes_sect008.htm")


##############################################
# Read in and Manipulate data
##############################################
#fd <- read.table("final.txt")
#
setwd("~/stat651-bayes/final-project/fromnile")
fd <- read.csv("fd GlenEagles.csv")
dat <- as.data.frame(fd)

setwd("~/stat651-bayes/final-project")
# Descriptions
# ttfee - average fee paid by players in round
#   rack rate = posted price
#     if nobody, it's rack rate'.
#   ttfee - average price paid by those who paid.

# Dependent Variables
# SPD
#   if empty, call it 0.
```

```
# temp
#   Max & Min of temp for that day
#   temp is hourly, min and max are daily
# pcp06
#   pcp24 is usable
# Dewp - humidity measure
#   humidity is good. If dewpoint < temp, you have fog.
# Pull in Macro

### Data Cleansing - only care about revenue
### Truncate Quantity so it maxes at 4
fd$Qtrunc <- fd$Quantity
fd$Qtrunc[which(fd$Qtrunc > 4)] <- 4

### Care only about Revenue
fd$Revenue <- with(fd,TTFee*Qtrunc)
dat <- as.data.frame(fd$Revenue)
dat$TTimes <- fd$TTimes
dat$Date <- fd$TDPosix
colnames(dat) <- c("Rev","TTime","Date")

# get n_i
dat$na <- as.numeric(!is.na(dat$Rev))
n_i <- sqldf("select sum(na) as n_i, TTime from dat group by TTime")

# Since 6.2 only appears on certain days, we'll have to add n_as so that they
    all add up to the total.
days.recorded <- sqldf("select TTime, count(TTime) as nidays from dat group by
    TTime")
# Sanity check
# x <- dat[which(dat$TTime == 6.3),] # dim(x) should equal the nidays value

# Get the total number of days on record
num.days <- length(unique(dat$Date))
num.days
uniquedays <- unique(dat$Date)
# Days per year on Record
table(substr(uniquedays,1,4))

# True n_i should be
n_i
ni <- n_i[,1]

# Assign 1:89 for each time slot
n_i$ind <- 1:89
dat <- sqldf("select dat.*, n_i.ind
      from dat
      left join n_i
      on dat.TTime=n_i.TTime")


##############################################
# Exploratory Analysis
##############################################
```

```
# MLEs for Revenue
avg <- sqldf("select avg(Rev) as avgrev, variance(Rev) as varrev, TTime from
    dat group by TTime")
plot(x = avg$TTime, y = avg$avgrev,
     main = "MLEs of Revenue at Tee Time", xlab = "Hour",
     ylab = "Revenue"
     )
abline(v = avg$TTime, col = 'grey') # REMEMBER- ttimes aren't real "times"

plot(x = avg$TTime, y = avg$varrev, col = "red", main = "VARIANCE of Rev @
    TTime")
abline(v = avg$TTime) # REMEMBER- ttimes aren't real "times"

# Plotting on the same doesn't help
# ggplot(avg, aes(TTime)) +
#   geom_line(aes(y = avgrev, colour = 1)) +
#   geom_line(aes(y = varrev, colour = 2))

qplot(x = TTime, y = avgrev,data = avg,
      main = "Average Revenue by Tee Time")
# + geom_vline(xintercept = avg$TTime)
#ts.plot(x = avg$TTime, y = avg$avgrev)

nrow(avg) #89 tee times

#### Create a table of times
tims <- cbind(avg$TTime, avg$avgrev,n_i$n_i)
obj <- round(t(tims),2)
obj <- obj[,c(1:2,10:11,40:42,70,71,82:83)]
rownames(obj) <- c("Hour.Min","Avg.Rev","Num.Obs")
obj <- xtable(obj)
print(obj,file = "data.txt",table.placement = "H",caption.placement = "top",
    include.rownames = T,backslash = T, include.colnames = F,
       sanitize.colnames.function = identity,
       sanitize.rownames.function = identity)




##############################################
# MCMC
##############################################
# Model:
# Y_ij ~ N(\tt_i,\sigma^2)
# \tt_i ~ N(\tt_{i-1}, v^2)
# \tau^2 ~ IG(a_{\tau},b_{\tau})
# \sigma^2 ~ IG(a_{\sigma},b_{\sigma})

# Interpretation:
# Y_ij: Revenue at tee time "i", number "j"
# \tt_i: Average revenue at tee time "i"
# \sigma^2: within tee time variance
# \tt_{i-1}: Average revenue at previous tee time
# \tau^2: variance of average in revenue "i", between-tee-time variance
```

```
### Choosing good Prior Values
png("prior-sig.png")
plot(density(1/rgamma(100000,a.s<<- 50, shape = (b.s<<-5))),main = "Prior Sig2
    ")
dev.off()
png("prior-tau.png")
plot(density(1/rgamma(100000,a.t<<- 15, shape = (b.t<<-10))),main = "Prior
    Tau2")
dev.off()

### Prior predictive distribution
npred = 10000
pr.pred <- list(NA)
for(i in 1:86){
  y.pred[[i]] <- rnorm(npred,theta[,(i+1)],sqrt(theta[,89])) #i+1 avoids 0th
      term
  # make the min 0 (nobody's gonna lose money on a round.)
  y.pred[[i]][which((y.pred[[i]] < 0))] <- 0
}


# Elements of the data.
n <- 86 # Only 86 ttimes have data.
k <- 86
# Set up the revenue list, remove the nas
rev <- list(NA)
for(i in 1:n) {
  r <- dat$Rev[which(dat$ind==i)]
  rev[[i]] <- r[!is.na(r)]
}

# what do the sums look like?
lapply(rev, mean)
lapply(rev, length)

# hyperpriors
m = 60
s2 = 5
revbar0 = 100
n0 = 10
a.sig <- a.s
b.sig <- b.s
a.tau <- a.t
b.tau <- b.t


golfmcmc<-function(niter=200){
  # Starting values
  theta <- matrix(60,ncol = (k-1), nrow = niter)
  theta0 <- rep(60,niter)
  thetak <- rep(60,niter)
  tau2 <- rep(1,niter)
  sig2 <- rep(1,niter)
```

```
  for(i in 2:niter)
  {
    ### GENERATE THETAS
    # Set-up
    prev <- i-1
    nxt  <- i+1
    tau2prev <- tau2[prev]
    sig2prev <- sig2[prev]

    # Get theta0 value
#     temp <- 1.0/(tau2prev*s2*n0 + sig2prev*(tau2prev+s2))
#     mustar <- (tau2prev*sig2prev*revbar0*n0 + theta[prev,1]*sig2prev*s2 + m*
    sig2prev*tau2)* temp
#     sigstar <- (sig2prev*tau2prev*s2) * temp
#     theta0[i] <- rnorm(1,mustar,sqrt(sigstar))

    # Try other way
    theta0[i] <- rnorm(1,m,sqrt(s2))


    # Get theta_1 value
      mu <- theta0[i] + theta[prev,2]
      temp <- 1.0 / (ni[1] * tau2prev + 2*sig2prev)
      mustar <- (mean(rev[[1]])*ni[1] * tau2prev + mu * sig2prev) * temp
      sigstar <- tau2prev * sig2prev * temp
      theta[i,1] <- rnorm(1, mustar, sqrt(sigstar))

    # Get theta_2 through theta_{k-1}
    for(j in 2:(k-1))
    {
      if(j == 85){ mu <- theta[i,j-1] + thetak[i] }
        else{
          mu <- theta[i,j-1] + theta[i,j+1]
        }
      temp <- 1.0 / (ni[j] * tau2prev + 2*sig2prev)
      mustar <- (mean(rev[[j]])*ni[j] * tau2prev + mu * sig2prev) * temp
      sigstar <- tau2prev * sig2prev * temp
      theta[i,j] <- rnorm(1, mustar, sqrt(sigstar))
    }

    # Get theta_k value
    mu <-  theta[i,k-1]
    temp <- 1.0 / (ni[k] * tau2prev + sig2prev)
    mustar <- (rev[[k]]*ni[k] * tau2prev + mu * sig2prev) * temp
    sigstar <- tau2prev * sig2prev * temp
    thetak[i] <- rnorm(1, mustar, sqrt(sigstar))

    ### GENERATE SIG2, TAU2
    # Calculate ssq for sigma2
    ssq <- 0
    ssq <- ssq + sum((revbar0*n0-theta0[i])^2) #theta0
    for(j in 1:(k-1)){
      ssq <- ssq + sum((rev[[j]]-theta[i,j])^2) #theta_1:theta_{k-1}
```

```
    }
    ssq <- ssq + sum((rev[[k]]-thetak[i])^2) #theta_k

    #generate sig2
    astar <- a.sig+sum(ni)*0.5 #Instead of dividing by 2
    bstar <- (1/b.sig+ssq/2)^(-1)
    sig2[i] <- 1/rgamma(1,astar,scale=bstar)

    # Calculate sstau for tau2
    sstau <- 0
    sstau <- sstau + sum((theta0[i]-theta0[i-1])^2)
    sstau <- sstau + sum((theta[i,]-theta[i-1,])^2)
    sstau <- sstau + sum((thetak[i]-theta[i,(k-1)])^2)

    #generate tau2
    astar <- a.tau + (k*0.5) #Instead of dividing by 2
    bstar <- (1/b.tau+.5*sstau)^(-1)
    tau2[i] <- 1/rgamma(1,astar,scale=bstar)
    cat("iter =",i,"\n")
  }
  out <- cbind(theta0,theta,thetak,tau2,sig2)
  colnames(out) <- c(paste("theta",0:k,sep = ""),"tau2","sig2")
  return(out)
}
niter <- 10000
theta.full <- golfmcmc(niter)

##############################################
# MCMC Diagnostics & Convergence:
##############################################
### Burn-in
burn <- 500
theta <- theta.full[-(1:burn),]

### Marginal Distributions ###
psig2 <- theta[,88]
ptau2 <- theta[,89]

### Trace Plots ###
plot(theta[,1],type = "l")
plot(theta[,2],type = "l")
matplot(theta[,c(2,10,50)],type = "l")
matplot(theta[,6:10],type = "l")

# a sampel of trace plots
png("trace-thetas.png")
par(mfrow = c(2,2))
plot(theta[,2],type = "l", main = "Theta 2")
plot(theta[,10],type = "l", main = "Theta 10")
plot(theta[,50],type = "l", main = "Theta 50")
matplot(theta[,c(2,10,50)],type = "l")
title("Traceplots for Theta 2, 10, 50")
dev.off()
```

```
# all trace plots
traceplots <- function(){
  for(i in 1:87){
    plot(theta[,i],type = "l", main = paste("Theta",i))
    Sys.sleep(.15)
  }
}
#saveGIF(traceplots) # need ImageMagick to work

# Trace plots of sigma2, tau2
png("trace-sig-tau.png")
par(mfrow = c(1,2))
plot(psig2,type = "l",main = "sigma2")
plot(ptau2,type = "l",main = "tau2")
dev.off()

### Auto-correlation
autocorr.plot(theta[,1],auto.layout = F)
autocorrplots <- function(){
  par(mfrow = c(1,1))
  for(i in 1:87){
    autocorr.plot(theta[,i], main = paste("Theta",i),auto.layout = F)
    Sys.sleep(.15)
  }
  dev.off()
}
autocorrplots()

png("diags-autoc-theta.png")
par(mfrow = c(2,2))
autocorr.plot(theta[,20], main = paste("Theta",20),auto.layout = F)
autocorr.plot(theta[,40], main = paste("Theta",40),auto.layout = F)
autocorr.plot(theta[,67], main = paste("Theta",67),auto.layout = F)
autocorr.plot(theta[,75], main = paste("Theta",75),auto.layout = F)
dev.off()

png("diags-autoc-vars.png")
par(mfrow = c(1,2))
autocorr.plot(theta[,88], main = paste("Tau2"),auto.layout = F)
autocorr.plot(theta[,89], main = paste("Sig2"),auto.layout = F)
dev.off()


### Raftery-Lewis
thetamcmc <- mcmc(theta)
raftery.diag(thetamcmc)
rfd <- raftery.diag(theta)
depfacts <- rfd$resmatrix[,4]
png("diags-rfdeps.png")
plot(depfacts,main = "Raftery-Lewis Dependence Factors\n DF < 5 Desired",
     ylab = "Depdendence Factor",
     pch = 19)
text(x = c(88,89),y = (depfacts[c(88,89)]-.05),labels = c("tau2","sig2"))
dev.off()
```

```
### Geweke Diagnostic
# H0: Difference of means between first 10% and last 50% = 0.
gvals <- geweke.diag(thetamcmc)
pvals <- 1-pnorm(abs(gvals$z),0,1)

# Create plot of pvalues
notsig <- sig05 <- which(pvals > 0.05)
sig05 <- which(pvals <= 0.05 & pvals > 0.01)
sig01 <- which(pvals <= 0.01)
cols <- 1:89
cols[notsig]<-1
cols[sig05]<- 2
cols[sig01] <- 4
# plot the p-values
png("diags-geweke-pvals.png")
plot(pvals,xlim = c(0,130),col = cols,pch = 19,
     main = "Geweke Diagnostic P-values: \n Ho: 1st 10% = Last 50%",
     ylab = "p-value")
abline(h = 0.05)
abline(h = 0.01)
legend("topright",col = c(1,2,4,0,0),pch = c(19,19,19,0,0),
       legend = c("p>0.05","p<=0.05","p<=0.01",paste("Num <0.05 =",length(
           sig05)),paste("Num <0.01 =",length(sig01))),
       bty ="n")
dev.off()

### Effective Sample Size
effectiveSize(theta)

### HPD
HPDinterval(thetamcmc)

##############################################
# Posterior Statistics Results:
##############################################
### Posterior Densities
# Sigma, tau
plot(density(psig2))
plot(density(ptau2))

### Posterior Means, Variances, & 95% CIs
newk <- k+1 # to include theta0
# Means, Quantiles (95\% PI), Variance, Stdevs
means <- colMeans(theta)
#post.thetas <- cbind(means[2:newk],avg$avgrev[1:k],1:k)
quants <- apply(theta,2,quantile,c(.025,.975))
vars <- apply(theta,2,var)
sds <- apply(theta,2,sd)
post.sum <- cbind(means,t(quants),vars,sds)
# Prettify the answers into a table.
colnames(post.sum) <- c("$E(\\theta|Y)$",".025\\%",".975\\%","$V(\\theta|Y)$
    ","$\\sqrt{Var(\\theta|Y)}$")
```

```
rownames(post.sum) <- c(paste("$\\theta_{",0:86,"}$",sep = "") ,"$\\tau^2$","$
    \\sigma^2$")
obj <- xtable(post.sum,caption ="Posterior Summary Statistics")
print(obj,file = "postsum.txt",table.placement = "H",caption.placement = "top
    ",include.rownames = T,backslash = T,
       sanitize.colnames.function = identity,
       sanitize.rownames.function = identity)

### Plotting Posterior Means
qplot(y = colMeans(theta)[1:newk],x = 1:newk)
#+ geom_point(y = avg$avgrev, x = 1:newk, col = 2)

# Plot of posterior means with error bars
png("results-postmeanstheta.png")
errbar(x = 0:86,y = means[1:87],yplus = quants[1,1:87],yminus= quants[2,1:87],
       ylab = "Average Rev / TTime",xlab = "Theta",
       col = "blue")
title("Posterior Means of Thetas")
dev.off()
boxplot(theta[,1:87],outline = F, "Boxplots")

# Posterior distribution of tau2 and sigma2
# not a good plot:
errbar(x = c(1,2),y = means[88:89],yplus = quants[1,88:89],yminus= quants
    [2,88:89],
       main = "Posterior Means",
       col = "blue")
boxplot(theta[,88],outline = F)
boxplot(theta[,89],outline = F)

### Posterior Correlation matrix
covs <- cov(theta[,1:87])
png("covplot.png")
corrplot(covs,is.corr = F,main = "Covariance Plot", tl.pos = "n"
         )
dev.off()

cors <- cor(theta[,1:87])
png("corrplot.png")
corrplot(cors,method = "color",tl.pos = "n",
         #outline = T)
         addgrid.col = "light grey")
dev.off()

### Posterior Predictive Distribution
y.pred <- list(NA)
for(i in 1:86){
  y.pred[[i]] <- rnorm(niter,theta[,(i+1)],sqrt(theta[,89])) #i+1 avoids 0th
      term
  # make the min 0 (nobody's gonna lose money on a round.)
  y.pred[[i]][which((y.pred[[i]] < 0))] <- 0
}

# Combine the predictive revenue means with ttime, and compare with MLEs
```

```
y.predmn <- lapply(y.pred,mean)
y.predmn <- unlist(y.predmn)
mles <- unlist(lapply(rev, mean))
prevbytime <- data.frame(TTime = n_i$TTime[1:86])
prevbytime$y.predmn <- y.predmn
prevbytime$mles <- mles

png("postpred-means.png")
plot(x=prevbytime$TTime, y = prevbytime$y.predmn,
     xlab = "Hour of Tee Time",
     main = "Bayes v. ML Estimates on Revenue",
     ylab = "Predicted Average Revenue",
     col = "blue", ylim = c(40,160),pch = 19, type = "o",lty = 2)
lines(x=prevbytime$TTime,y = prevbytime$mles, type = "o", pch = 19,lty = 2)
legend("topright",col = c("blue",1),lty = 1, bty = "n", legend = c("Bayes","
   MLE"))
abline(v = prevbytime$TTime,col = "light grey",lty = 2)
dev.off()

# Other way to plot it
# xx <- as.matrix(prevbytime)
# matplot(x = xx[,1],y = xx[,2:3],type = "o",pch = 19, lty = c(1,2))

# Plot posterior distribution of highest revenue
max.ind <- which(y.predmn == max(y.predmn))
xx <- y.pred[[max.ind]]
# Get hpd
source("~/stat651-bayes/bayes-functions.R")
hpd <- get.hpd(xx,.95)

png("postpred-130.png")
plot(density(xx), main = "Posterior Predictive for 1:30pm")
abline(v = mean(xx),col = "blue")
text(y = .002,x = mean(xx+10),labels = paste("mean = $",round(mean(y.pred
    [[44]])),sep = ""))
abline(v = quantile(xx,c(.025,.975)), col = "blue",lty = 2)
abline(v = hpd, col = "red",lty = 2, cex =3)
legend("topright",legend = c("95% PI","95% HPD"),col = c("blue","red"),lty =
    2,bty = "n")
dev.off()



##############################################
# Questions We Can Answer
##############################################
### Questions of Interest
# What hours are best? (define best)
# What hours are worst? (define worst, and find pr(rev <20)
# What is the expected revenue at each time slot each month?
# Is there an hour-to-hour effect? (t-test on different hours)
# are the times different within hours?

### What's the probability of making less than $30?
```

```
cost <- 30
y.less <- NA
  for(i in 1:length(y.pred)){
    y.less[i] <- mean(y.pred[[i]]<cost)
  }
y.less30 <- y.less
cost <- 50
y.less <- NA
for(i in 1:length(y.pred)){
  y.less[i] <- mean(y.pred[[i]]<cost)
}
y.less50 <- y.less

prob.cost <- cbind(y.less30,y.less50)
png("results-probloss.png")
matplot(prob.cost,type = "o",lty=2,pch = 19,col = c(2,1),
        ylab = "Pr(Posterior Pred. Rev < Cost)",xlab = "Tee Time",
        main = "Prob of Loss (Rev < Cost)")
legend("top",legend = c("Pr(Rev) < $50","Pr(Rev) < $30"),bty = "n",col = c
    (1,2),lty = 1)
dev.off()
### What



##############################################
# Model Selection
##############################################
# If I have time, I'll get to this

##############################################
# Goodness of Fit
##############################################
### Posterior Predictive Checking
# With Density plots
pval <- NA
png("quantiles.png")
for(i in 1:86){
  xx <- rnorm(niter,theta[,(1+i)],sqrt(psig2))
  q <- pnorm(xx,theta[,(1+i)],sqrt(psig2))
  if (i == 1) {
    plot(density(q,from = 0, to = 1),ylim = c(0,1.5),
    main = "Density Estimates of Quantiles for All 86 Times")
  }
  else{lines(density(q,from = 0, to = 1))}
  pval[i] <- ks.test(q,punif)$p.value
}
dev.off()
#Make it a moustach
# text(x = .3, y = 1.3, label = ".",cex = 5)
# text(x = .3, y = 1.3, label = "O",cex = 5)
# text(x = .7, y = 1.3, label = ".",cex = 5)
# text(x = .7, y = 1.3, label = "O",cex = 5)
# text(x = .5, y = .6, label = "\\___/",cex = 5)
```

```
# K-S test
# prepare the p-values
png("quantiles-pvals.png")
signif <- (as.numeric(pval<.05)+2)
num.sig <- sum(pval<0.05)
# plot the p-values
plot(pval,xlim = c(0,120),col = signif,pch = 19,
     main = "K-S Test For 86 Tee Times",
     ylab = "p-value")
abline(h = 0.05)
legend("topright",col = c(2,3,0),pch = c(19,19,0),legend = c("p<0.05","p
    >=0.05",paste("Num Sig =",num.sig)),bty ="n")
dev.off()




### Bayesian Chi2 test




##############################################
# Demand Curves
##############################################
## Backing Out Demand Curves given MCMC data and golf revenue structure.
price.optim <- function(tim=1,n=1,P=floor(seq(10,370, length = 40))){
  # n is the number of total days over all time. probably not the right
      amplifier. n should be the number of days in a month.
  REV <- y.pred[[tim]]
  #plot(density(REV))
  PQmat <- matrix(NA, ncol = 5, nrow = length(P))
  rownames(PQmat) <- as.character(P)
  colnames(PQmat) <- as.character(c(0,1,2,3,4))

  for(i in 1:length(P)){
    Q = floor(REV/P[i])
    Q[which(Q>=5)] <-4
    #which(Q==5)
    q0 = sum(Q==0)
    q1 = sum(Q==1)
    q2 = sum(Q==2)
    q3 = sum(Q==3)
    q4 = sum(Q==4)
    tab <- cbind(q0,q1,q2,q3,q4)
    PQmat[i,] <- prop.table(tab)
  }
  PQmat
  Weighted.PQmat <- sweep(PQmat,2,c(0,1,2,3,4),"*")
  # Expected revenue for 1 day:
  ExpRev <- sweep(Weighted.PQmat,1,P,'*')[,2:5]
  ExpTotRev <- rowSums(ExpRev)
```

```
  Demand.mat <- ceiling(PQmat*n)

  # Probability matrix
  #matplot(PQmat, type = c("o"), main = "Probability")
  # Demand Curves
  #matplot(cumDem[,2:5],type = "o", main = paste("Demand",tim))
  # Revenue Curves for Each Q #QQQ Why does this exist?
  maxrev <- round(max(ExpTotRev),2)
  pmax.ind <- max(ExpTotRev)==ExpTotRev
  pmax <- P[which(pmax.ind==1)]
  pmax <- round(pmax,2)

  plot(x = P, y =ExpTotRev,type = "o", main = paste("Daily Expected Revenue
      for theta",tim,sep = ""))
  legend("topright",legend = c(paste("Max Rev = $",maxrev,sep = ""),paste("@
      Price = $",pmax,sep = "")))
  list(rev= maxrev,price = pmax)
}
png("results-optimrev.png")
price.optim(40,n=1,P = seq(10,350,length = 1000))
dev.off()

prices1 = revs1<-NA
for(i in 1:86){
  obj <- price.optim(i,n=1,P = seq(10,350,length = 200))
  revs1[i] <- obj$rev
  prices1[i] <- obj$price
}

### Simplified Pricing function, focuses just on optimizing daily revenue.
price.to.rev <- function(P,tim){
  REV <- y.pred[[tim]]
  #plot(density(REV))
  Q = floor(REV/P)
  Q[which(Q>=5)] <-4
  q0 = sum(Q==0)
  q1 = sum(Q==1)
  q2 = sum(Q==2)
  q3 = sum(Q==3)
  q4 = sum(Q==4)
  Qprobs <- prop.table(cbind(q0,q1,q2,q3,q4))
  Weighted.Qprobs <- c(0,1,2,3,4)*Qprobs
  Exp.Rev <- sum(P*Weighted.Qprobs)
  -Exp.Rev #to get max value
}
price.to.rev(50.16,tim = 40)

# Return a list of the profit-maximizing prices for each tee-time
price = max.rev = NA
for(i in 1:86){
  # Get best
  best.p <- function(P){price.to.rev(P,tim = i)}
  # Optimize the Function
```

```
  obj <- optim(par = 100,best.p,method = "Brent", lower = 0, upper = 400)
  max.rev[i]<- -obj$value
  price[i] <- obj$par
}
revbyprice <- data.frame(max.rev)
revbyprice$price <- price
revbyprice$time <- n_i$TTime[1:86]

# Compare this method with the above, "by hand" methods
cbind(revs1,prices1,revbyprice) #seems to check out!

# Output maximum revenues in a table
obj <- revbyprice
colnames(obj) <- c("$E(Rev|P)$","Price","Time")
rownames(obj) <- c(paste("$\\theta_{",1:86,"}$",sep = ""))
obj <- xtable(obj,caption ="Revenue-Maximizing Prices at Each Tee Time")
print(obj,file = "revbyprice.txt",caption.placement = "top",include.rownames =
    T,backslash = T, tabular.environment='longtable', floating = F,
      sanitize.colnames.function = identity,
      sanitize.rownames.function = identity)

##############################################
# Results Tables
##############################################
mega <- cbind(post.sum,rbind(".",round(revbyprice,2),".","."))
colnames(mega) <- c("$E(\\theta|Y)$",".025\\%",".975\\%","$V(\\theta|Y)$","$\\
   sqrt{Var(\\theta|Y)}$","$E(Rev|P)$","Price","Time")
rownames(mega) <- c(paste("$\\theta_{",0:86,"}$",sep = ""),"$\\tau^2$","$\\
   sigma^2$")
obj <- xtable(mega,caption ="Posterior Summary Statistics and Revenue-
   Maximizing Prices at Each Tee Time")
print(obj,file = "mega.txt",caption.placement = "top",
      include.rownames = T,backslash = T,
      tabular.environment='longtable', floating = F,
      sanitize.colnames.function = identity,
      sanitize.rownames.function = identity)




##############################################
# To-Do:
##############################################
# Figure out why theta0 blows up.
# Get Posterior Means
# Get Posterior Variances
# Get Other important findings of relevance. What is the problem we're trying
   to solve?
# Go though class notes and see what he wants us to do.
# Posterior predictive
```