Data Analysis Project

Tintin P.

May 15, 2020

Description

This data set includes 103 data points. The data were collected from a Stanford University Heart Transplant Study, where they tried to determine the correlation between heart transplantation and length of survival. Participants who were admitted to the program were those considered to be in need of new heart and has potential to benefit from the transplant.

Relevant variables:

- Age: age at the time of admittance to the program
- Survived: status of life before the end of the program (dead, alive)
- Survtime: number of days the participants are alive until the end of the program
- Transplant: those who received a new heart are labeled as "treatment", otherwise, they are labeled "control".

Analysis #1, Binomial Proportion

We will examine the proportion of participants who were are alive until the end of the program.

```
library (Bolstad)
```

```
##
## Attaching package: 'Bolstad'
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, sd, var
```

```
load ("./heart_transplant.rda")
heart_transplant = heart_transplant[-c(1:2,6,8)]
```

Analysis of survival status of participants at the end of the program:

```
summary(heart_transplant$survived)
```

```
## alive  dead
##    28    75
```

```r
surv.alive = (heart_transplant$survived == "alive")
(n.surv.alive = sum (surv.alive))
```

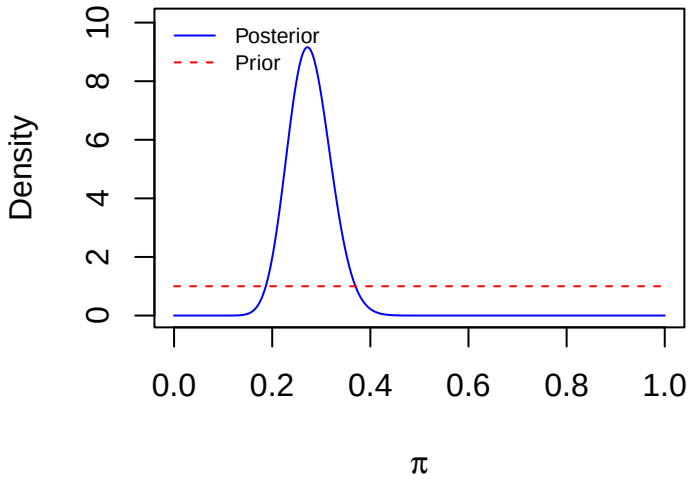
```
## [1] 28
```

```r
(sample.size = length (heart_transplant$survived))
```

```
## [1] 103
```

(a) Using a uniform prior distribution

```r
binobp (n.surv.alive, sample.size, a = 1, b = 1)
```

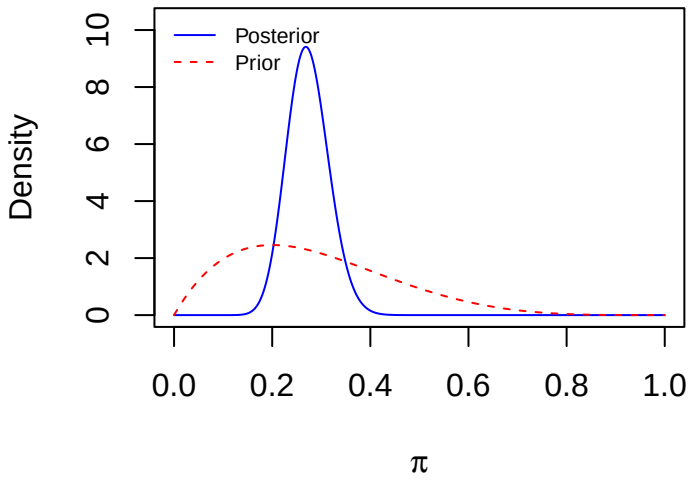



```
## Posterior Mean           :  0.2761905
## Posterior Variance       :  0.0018859
## Posterior Std. Deviation :  0.0434274
##
## Prob.    Quantile
## ------   ---------
## 0.005    0.1732322
## 0.010    0.1820187
## 0.025    0.1953552
## 0.050    0.2072276
## 0.500    0.2747649
## 0.950    0.3500227
## 0.975    0.3650994
## 0.990    0.3828302
## 0.995    0.3950103
```

Using a uniform prior, the estimated proportion of participants who were alive until the end of the program is 0.276. The probability is 95% that the true proportion of participants who were alive until the end of the program is between 0.195 and 0.365.

(b) Using a beta (a, b) prior that is not uniform - choose your own a and b.

```
binobp (n.surv.alive, sample.size, a = 2, b = 5)
```




```
## Posterior Mean           :  0.2727273
## Posterior Variance       :  0.0017869
## Posterior Std. Deviation :  0.0422719
##
## Prob.    Quantile
## ------   ---------
## 0.005    0.1724367
## 0.010    0.1810111
## 0.025    0.1940180
## 0.050    0.2055902
## 0.500    0.2713457
## 0.950    0.3445832
## 0.975    0.3592624
## 0.990    0.3765316
## 0.995    0.3883989
```

Using a beta (2,5) prior, the estimated proportion of participants who were alive until the end of the program is 0.273. The probability is 95% that the true proportion of participants who were alive until the end of the program is between 0.194 and 0.359.

(c) Using a frequentist confidence interval

```
# Agresti-Coull confidence interval
y = n.surv.alive
n = sample.size
(p.est = (y+2)/(n+4))
```

```
## [1] 0.2803738
```

```
zquant = qnorm (0.975)
se.p.est = sqrt (p.est*(1-p.est)/(n+4))
# calculate the confidence interval
p.est + c(-1, 1)*zquant * se.p.est
```

```
## [1] 0.1952643 0.3654834
```

The frequentist estimate for the proportion of participants who were alive until the end of the program is 0.280. With 95% confidence, the true proportion is between 0.195 and 0.365.

```
# Frequentist confidence interval as defined in the book
(pihat = y/n)
```

```
## [1] 0.2718447
```

```
pihat + c(-1, 1) * zquant * sqrt (pihat * (1 - pihat) / sample.size)
```

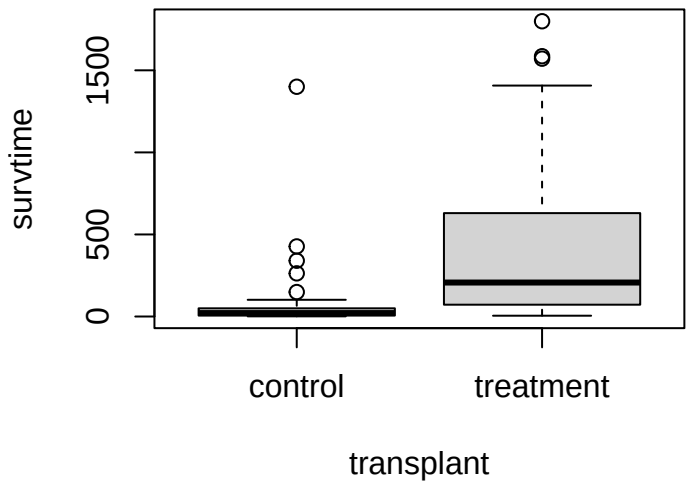
```
## [1] 0.1859232 0.3577662
```

Alternatively, as defined in the textbook, the frequentist estimate for the proportion of participants who were alive until the end of the program is 0.272. With 95% confidence, the true proportion is between 0.186 and 0.358.

Analysis #2, Compare Two Means

We will compare the length of survival between participants who received a heart transplant (treatment) and those who didn't receive a heart transplant (control).

```
with (heart_transplant, boxplot (survtime ~ transplant))
```



The variance of life span is different between the two groups, so we will use the unequal variances analysis for comparing two means. First, we determine the posterior distribution for the mean of each group.

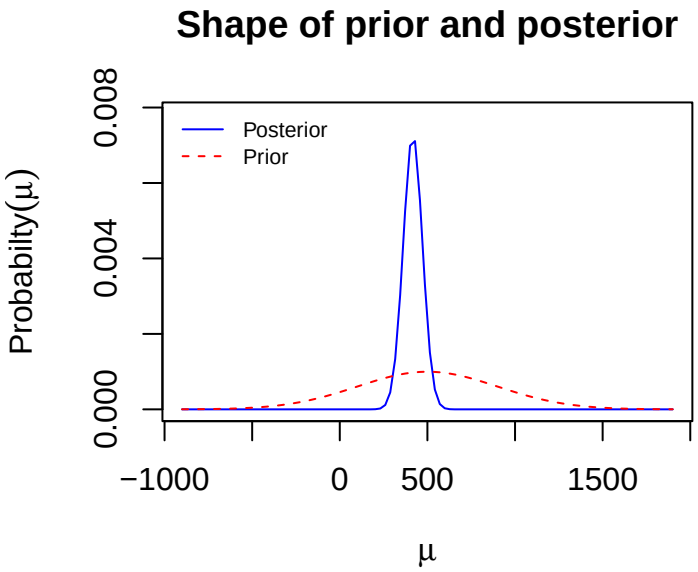
(a) Using a normal prior

```
transplant.treatment = (heart_transplant$transplant == "treatment")
transplant.control = !(heart_transplant$transplant == "treatment")

treatment.group = heart_transplant$survtime [transplant.treatment]
control.group = heart_transplant$survtime [transplant.control]

postT = normnp (treatment.group, 500, 400)
```

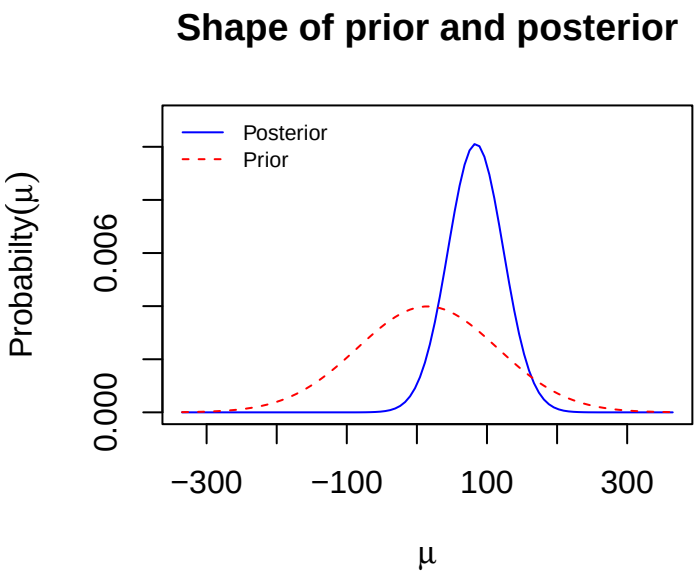
```
## Standard deviation of the residuals :458.7
## Posterior mean           : 416.9876782
## Posterior std. deviation : 54.6986496
```




```
##
## Prob.    Quantile
## ------   ----------
## 0.005    276.0932937
## 0.010    289.7395910
## 0.025    309.7802949
## 0.050    327.0164060
## 0.500    416.9876782
## 0.950    506.9589504
## 0.975    524.1950614
## 0.990    544.2357654
## 0.995    557.8820627
```

```
postF = normnp (control.group, 15, 100)
```

```
## Standard deviation of the residuals :250.3
## Posterior mean           : 83.9161644
## Posterior std. deviation : 39.4489252
```




```
##
## Prob.    Quantile
## ------   ----------
## 0.005    -17.6975331
## 0.010    -7.8557588
## 0.025     6.5976918
## 0.050    19.0284567
## 0.500    83.9161644
## 0.950    148.8038720
## 0.975    161.2346370
## 0.990    175.6880876
## 0.995    185.5298619
```

The posterior distributions are normal. The posterior mean and SD in the treatment group (receive transplant) are 417.00 and 54.70 days. The posterior mean and SD in the control group (didn't receive transplant) are 83.92 and 39.45 days.

Summary Table of Life Span

Participant Group	Mean Length of Survival (days)	SD Length of Survival (days)
Treatment (receive transplant)	417.00	54.70
Control (didn't receive transplant)	83.92	39.45

(b) Bayesian credible interval for the difference of two means

```
m1p = mean (postT)
s1p = sd (postT)
m2p = mean (postF)
s2p = sd (postF)
sighat1 = sd (heart_transplant$survtime [transplant.treatment])
sighat2 = sd (heart_transplant$survtime [transplant.control])
n1 = length (heart_transplant$survtime [transplant.treatment])
n2 = length (heart_transplant$survtime [transplant.control])
sw.df = (sighat1^2/n1 + sighat2^2/n2)^2 /
        ((sighat1^2/n1)^2 / (n1-1) + (sighat2^2/n2)^2 / (n2-1))
# The interval
(m1p - m2p) + c(-1, 1) * qt (0.975, sw.df) * sqrt (s1p^2 + s2p^2)
```

```
## [1] 199.2702 466.8729
```

The probability is 95% that the true difference in population mean length of survival is between 199.27 and 466.87 days, showing that participants who received a transplant tend to have a longer average length of survival.

Note that this conclusion is only an observed association, not a cause-and-effect statement.

(c) Frequentist hypothesis test & confidence interval for the difference between two means

Two-sided Frequentist hypothesis test: $H_0 : \mu_T - \mu_C = 0$ vs. $H_1 : \mu_T - \mu_C \neq 0$, at 5% level of significance.

```
with (heart_transplant, t.test (treatment.group, control.group))
```

```
##
##  Welch Two Sample t-test
##
## data:  treatment.group and control.group
## t = 4.5578, df = 99.864, p-value = 1.467e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  180.0214 457.5549
## sample estimates:
## mean of x mean of y
## 415.40580  96.61765
```

The p-value is $0 < 0.05 = \alpha$. Thus, we reject the null hypothesis H_0, and conclude that there is a significant difference between the length of survival of the participants that received a heart transplant (treatment) and those who didn't receive a heart transplant (control).

Also, the frequentist point estimate for the difference of mean length of survival is the observed difference of sample means, which is 415.4 - 96.62 = 318.79 days. With 95% confidence, the difference in mean length of survival is between 180.02 and 457.55 days. Thus, we can conclude that participants who received a transplant tend to have a longer average length of survival.

Since these data are from an observational study, we cannot conclude any cause-and-effect relationship between length of survival and whether or not a participant received a heart transplant.

Analysis #3, Simple Linear Regression

The purpose of this analysis is to study the relationship between length of survival (survtime) from age, given that the participants received the transplant (treatment) and were alive until the end of the program

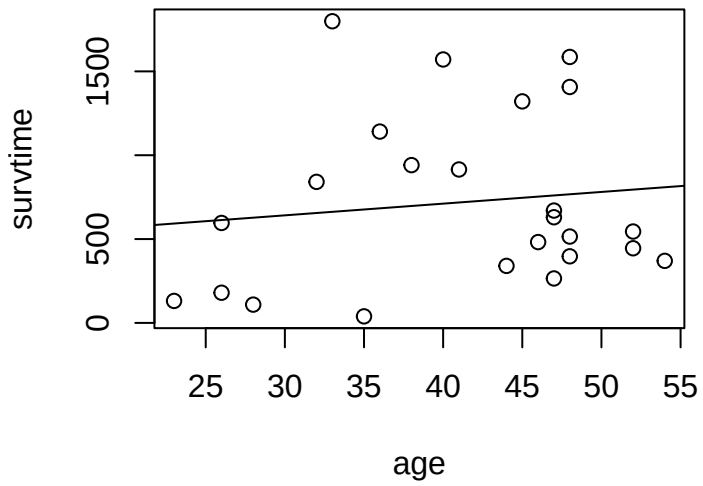
Linear Regression

```
# participants who have received the transplant and were alive until the end of the program
narrowed_heart_transplant = subset(heart_transplant, subset = (transplant == "treatment" & survived == "

# Plot with fitted line
plot (survtime ~ age, data=narrowed_heart_transplant)
fit1 = lm (survtime ~ age, data=narrowed_heart_transplant)
(sfit1 = summary (fit1))
```

```
##
## Call:
## lm(formula = survtime ~ age, data = narrowed_heart_transplant)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -637.3 -407.7 -190.0  297.3 1136.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   431.98     497.19   0.869    0.394
## age             6.98      11.85   0.589    0.562
##
## Residual standard error: 519.7 on 22 degrees of freedom
## Multiple R-squared:  0.01553,    Adjusted R-squared:  -0.02921
## F-statistic: 0.3471 on 1 and 22 DF,  p-value: 0.5617
```

```
abline (fit1)
```



The Frequentist estimated Y-intercept is 431.98. The estimated slope is 6.98 days of survival per increment in year of age. The estimated variance about the regression line is $520^2 = 270400$ (residual standard error). The Standard Deviation Error is 11.85.

Visually, we note from the plot above that the relationship does not appear to be linear, but it does seem to be increasing from left to right.

(a) Using Frequestist confidence interval for the slope

```
# 95% confidence interval for the slope:
confint (fit1)
```

```
##                  2.5 %     97.5 %
## (Intercept) -599.13871 1463.09159
## age          -17.58985   31.55035
```

We are 95% confident that true slope is between -17.60 and 31.55 days of survival per year of age. In otherwords, we are 95% confident that among this group of participants, for every increment in year of their age, they'll have a longer length of survival by between -17.59 and 31.55 days.

(b) Using normal prior distributions to obtain a Bayesian credible interval for the slope

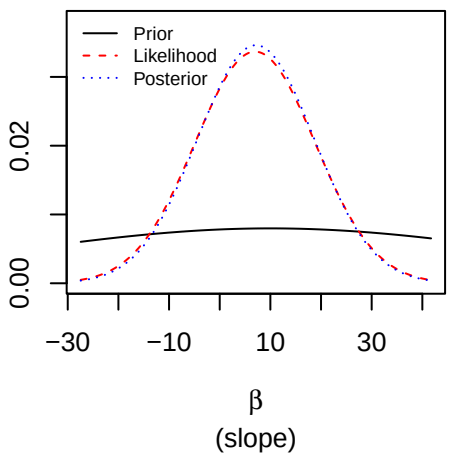
We assume a Normal $(10, 50^2)$ prior distribution for the slope and a flat prior for the intercept.

```
# inference for the slope
blr1 = bayes.lin.reg (narrowed_heart_transplant$survtime,
                      narrowed_heart_transplant$age, slope.prior = "normal",
                      intcpt.prior = "flat", mb0=10, sb0=50, plot.data=TRUE)
```

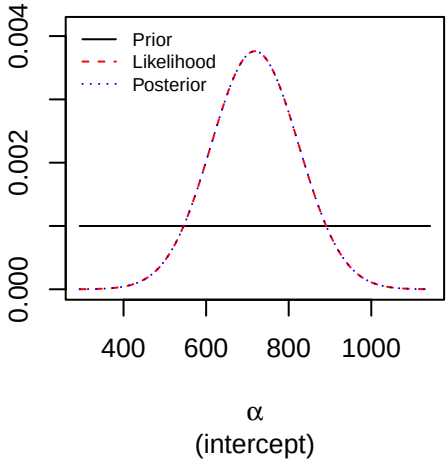
```
## Standard deviation of residuals:  520
```

```
##             Posterior Mean Posterior Std. Deviation
##             -------------- ------------------------
## Intercept:  718.2          106.08
## Slope:      7.141          11.528
```

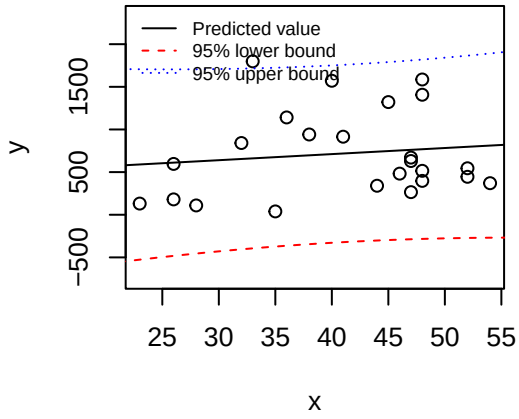
Prior, likelihood and posterior for β



Prior, likelihood and posterior for $\alpha_{\bar{x}}$



Predicitions with 95% bounds




```
# 95% credible interval for the slope:
quantile (blr1$slope, c(0.025, 0.975))
```

```
## [1] -15.22842  29.97373
```

The posterior probability is 95% that the true slope is between -15.23 and 29.97 days of survival per increment in year of age.

(c) Two-Sided Bayesian Hypothesis Test for Slope

Since the best fit least square line's slope is approximately 7, that implies the older the participant, the longer they are to be alive. My initial hypothesis was that the younger the participant, the longer they are to live; so this result sounds doubtful, so I'll run a two-sided bayesian hypothesis test for slope.

Perform a Bayesian test of $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$, at 5% level of significance.

Since zero is in the credible interval (-15.23, 29.97), we fail to reject the null hypothesis and conclude that there is no statistically significant relationship between age and length of survival of this group of participants. Therefore, we cannot use regression model to predict participants length of survival from their age (given that they received the transplant and were alive until the end of the program).