2017 Prelim

Will Gertsch

8/19/2020

1

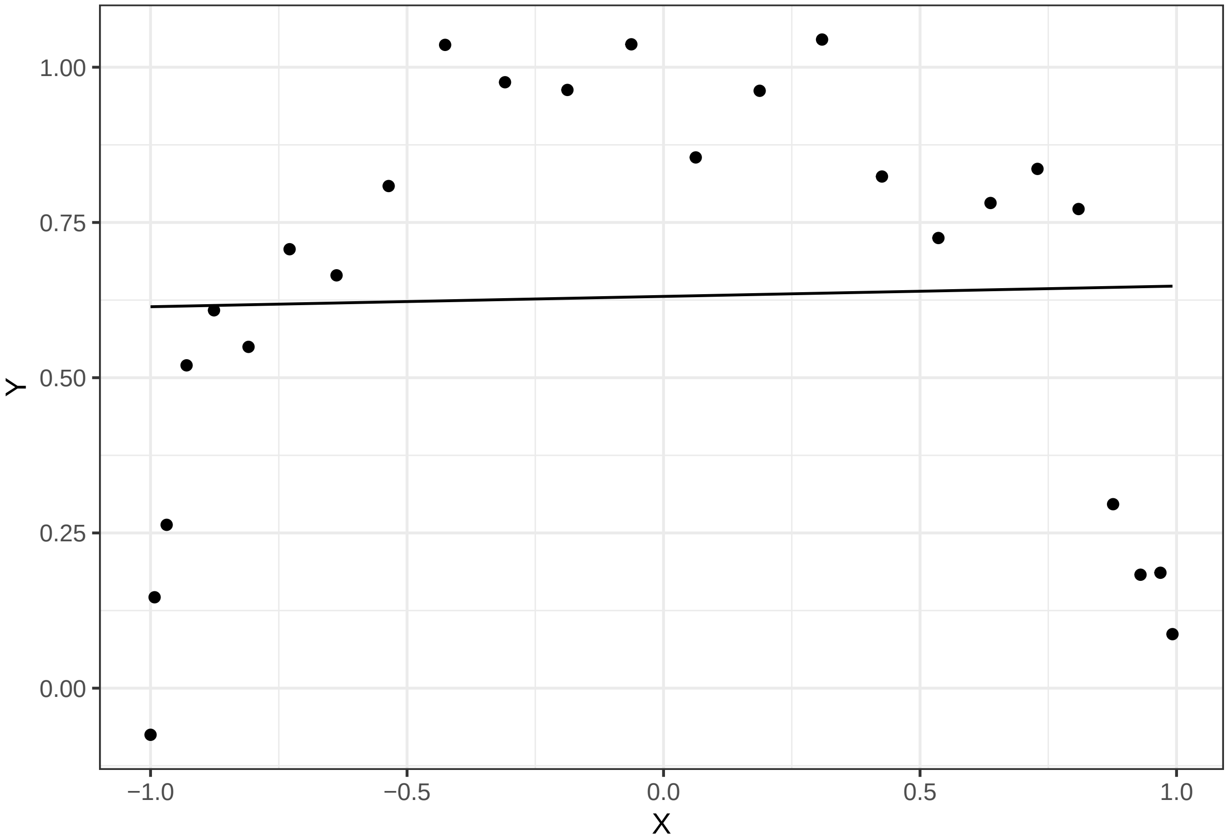
a

The line should be flat, similar to the residual plot since the model overestimates, underestimates, and then overestimates.

Let's fit the model and see if I am right.

```r
U = seq(0.04, 1.00, 0.04)
X = cos(pi*U)
Y = sin(pi*U) + rnorm(25, 0, sqrt(.01))
mod <- lm(Y ~ X)

library(ggplot2)
ggplot(data = NULL, aes(x = X)) +
  geom_point(aes(y = Y)) +
  geom_line(aes(y = predict(mod))) +
  theme_bw()
```



The line is mostly flat, but most of the time has a slight positive slope.

b

Correlation is a measure of linear relationship between two variables. Since the relationship is quadratic on this interval, the correlation will be close to 0.

The actual value is

```
cor(X,Y)
```

```
## [1] 0.03535826
```

c

The parabola will point downwards so in the corresponding quadratic $ax^2 + bx + c$, a will be negative. Therefore β_2 will be negative.

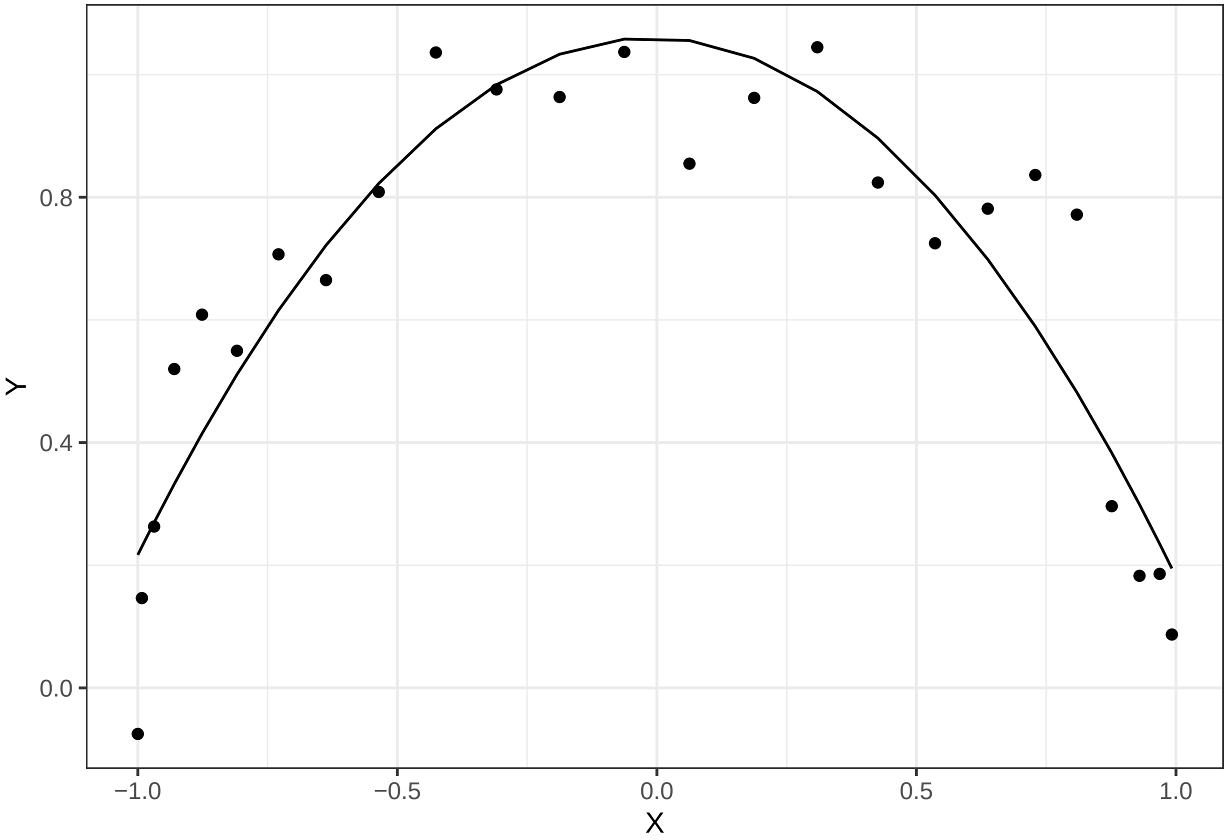
Let see if that is true.

```
mod2 <- lm(Y ~ X + I(X^2))
summary(mod2)
```

```
##
## Call:
```

```
## lm(formula = Y ~ X + I(X^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29196 -0.07830 -0.02096  0.08253  0.28966
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.06012    0.04941  21.458 3.05e-16 ***
## X           -0.01794    0.04051  -0.443    0.662
## I(X^2)      -0.86117    0.08090 -10.645 3.82e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1425 on 22 degrees of freedom
## Multiple R-squared:  0.8376, Adjusted R-squared:  0.8229
## F-statistic: 56.75 on 2 and 22 DF,  p-value: 2.068e-09
```

```r
ggplot(data = NULL, aes(x = X)) +
  geom_point(aes(y = Y)) +
  geom_line(aes(y = predict(mod2))) +
  theme_bw()
```



## d Standard deviation is (roughly) the average of the distances from the mean. It appears that the assumptions of linear regression are being met. Therefore the distribution of the residuals should be $N(0, \sigma^2)$. Since the residuals are mostly within ±0.2 of 0, the standard deviation should be close to 0.1.

Let's test that

```
sd(residuals(mod2))
```

```
## [1] 0.1364756
```

e

It depends on the range of your data. A quadratic will work fine when only looking at half the period of the trig function. Trying to model more than half a period with a quadratic will be a bad idea.

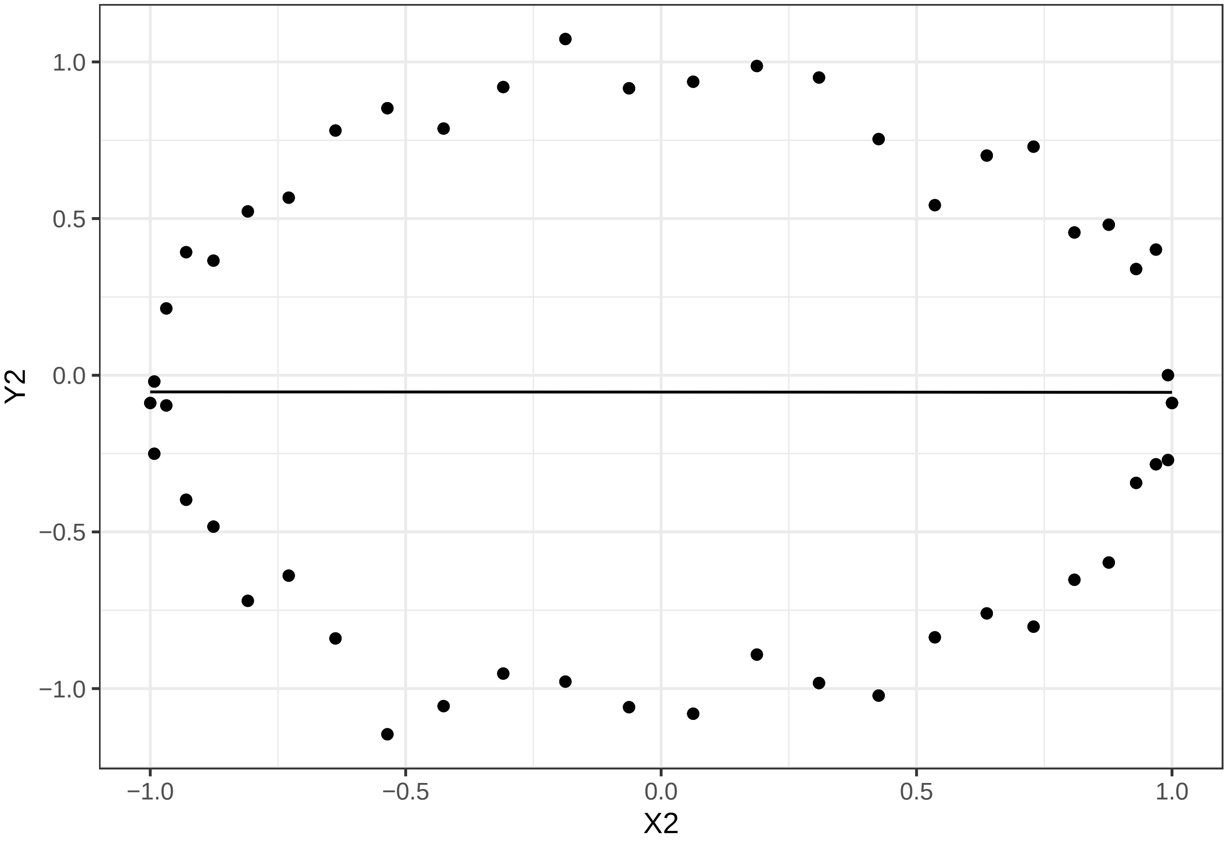
Let's see what happens

```
U2 = seq(-2, 2, 0.04)
X2 = cos(pi*U2)
Y2 = sin(pi*U2) + rnorm(25, 0, sqrt(.01))
```

```
## Warning in sin(pi * U2) + rnorm(25, 0, sqrt(0.01)): longer object length is not
## a multiple of shorter object length
```

```
mod3 <- lm(Y2 ~ X2 + I(X2^2))

library(ggplot2)
ggplot(data = NULL, aes(x = X2)) +
  geom_point(aes(y = Y2)) +
  geom_line(aes(y = predict(mod3))) +
  theme_bw()
```



2

a

$$R^2 = \frac{SSM}{SST} = \frac{2016.12237}{8704.81686} = 0.23161$$

About 23% of the variation in MFSI can be explained with a linear relationship with the predictors.

b

`age10` is the age in years divided by 10. Therefore, every 1 unit increase in `age10`, an increase of 10 years, is associated with a .68 decrease in MFSE holding all the other predictors constant. Divide the coefficient by 10 to get the yearly effect.

c

`log2daysout` is the log base 2 of number of days since date of diagnosis. For every doubling in the number of days since diagnosis, the MFSI increases by 1.040066, holding the other predictors constant.

d

The coefficient of marital status is 1.890183 while the coefficient of history of MDD is -.1695029. The interaction effect between the two is 3.48601. This suggests that being married increases the MFSI while MDD has a very small, statistically insignificant, decrease in MFSI. The interaction effect suggests that the effect of a history of MDD greatly increases when also married.

e

Interaction is not the same thing as correlation. Just because there is an interaction effect between 2 variables does not mean that they are related.

f

Not sure that these are techinically nested models, but let's assume that they are.

$$F^* = \frac{\frac{2026.35534-2016.12237}{11-10}}{\frac{6678.46151}{249}} = 0.3815264$$

Under H_0, $F^* \sim F(1, 249) \equiv t(249)$. Since the degrees of freedom are so large, we can make the approximation $t_{.975}(249) \approx Z_{.975} = 1.96$. Therefore, we fail to reject H_0 and conclude that the dummy variable model provides little additional explanation of the variation in MFSI compared to the original model.

3

a

Since we know σ is known, we can use a Z-test instead of t-test to derive the interval. $E(\bar{X} - \bar{Y}) = 0$ since $\mu_x - \mu_y = 0$. $\text{Var}(\bar{X} - \bar{Y}) = \frac{25^2}{100} + \frac{25^2}{300}$ since the samples are independent. Therefore, $\bar{X} - \bar{Y} \dot{\sim} N(0, \frac{25^2}{100} + \frac{25^2}{300})$.

Therefore we can use the usual normal CI.

$$\bar{X}-\bar{Y}-1.96*25\sqrt{\frac{1}{100}+\frac{1}{300}}<0<\bar{X}-\bar{Y}+1.96*25\sqrt{\frac{1}{100}+\frac{1}{300}}$$

Simplify and move $\bar{X}-\bar{Y}$ to the middle

$$-5.658033<\bar{X}-\bar{Y}<5.658033$$

Transform to standard normal.

$$\frac{-5.658033}{\sqrt{8.33}}<\frac{\bar{X}-\bar{Y}}{\sqrt{8.33}}<\frac{5.658033}{\sqrt{8.33}}$$

Therefore

$$P(\text{CI includes }0)=\Phi\left(\frac{5.658033}{\sqrt{8.33}}\right)-\Phi\left(\frac{-5.658033}{\sqrt{8.33}}\right)=0.95005$$

We could have gotten this result right away by noting that the definition of a 95% CI is that will include the true value with probability .95.

b

Using the usual normal confidence intervals, the CI for μ_x is

$$\bar{X}\pm1.96*\frac{25}{\sqrt{100}}$$

and the CI for μ_y is

$$\bar{Y}\pm1.96*\frac{25}{\sqrt{300}}$$

The CI's will overlap if

$$\bar{Y}-1.96*\frac{25}{\sqrt{300}}<\bar{X}+1.96*\frac{25}{\sqrt{100}}$$

and

$$\bar{X}-1.96*\frac{25}{\sqrt{100}}<\bar{Y}+1.96*\frac{25}{\sqrt{300}}$$

Therefore we have

$$-1.96\left(\frac{25}{\sqrt{100}}+\frac{25}{\sqrt{300}}\right)<\bar{X}-\bar{Y}<1.96\left(\frac{25}{\sqrt{100}}+\frac{25}{\sqrt{300}}\right)$$

which is

$$-7.729016<\bar{X}-\bar{Y}<7.729016$$

Given that $\mu_x=\mu_y$,

$$\bar{X}-\bar{Y}\dot{\sim}N\left(0,\frac{25^2}{100}+\frac{25^2}{300}\right)=N(0,8.33)$$

Transform to get to a standard normal.

$$\frac{\bar{X}-\bar{Y}-0}{\sqrt{8.33}}\sim N(0,1)$$

Apply the transformation to the bounds found earlier.

$$(-7.729016,7.729016)\rightarrow(-2.677945,2.677945)$$

Now we can compute the probability

$$P(\text{CI overlap})=\Phi(2.677945)-\Phi(-2.677945)=0.9925925$$

This means that checking to see if the two CI's overlap is more conservative than seeing if the CI for 2 means includes 0.

c

We repeat out work for part a), but this time

$$\bar{X} - \bar{Y} \dot{\sim} N\left(-6, \frac{25^2}{100} + \frac{25^2}{300}\right) = N(-6, 8.33)$$

We get the interval

$$\bar{X} - \bar{Y} \pm 1.96 * 25\sqrt{\frac{1}{100} + \frac{1}{300}}$$

Therefore, for interval to include 0, the following must be true

$$\bar{X} - \bar{Y} - 1.96 * 25\sqrt{\frac{1}{100} + \frac{1}{300}} < 0 < \bar{X} - \bar{Y} + 1.96 * 25\sqrt{\frac{1}{100} + \frac{1}{300}}$$

Simplify.

$$\bar{X} - \bar{Y} - 5.658033 < 0 < \bar{X} - \bar{Y} + 5.658033$$

Subtract $\bar{X} - \bar{Y}$ and multiply by -1 to get

$$-5.658033 < \bar{X} - \bar{Y} < 5.658033$$

Now use a z-transform to get to a standard normal.

$$\frac{-5.658033 + 6}{\sqrt{8.33}} < \frac{\bar{X} - \bar{Y} + 6}{\sqrt{8.33}} < \frac{5.658033 + 6}{\sqrt{8.33}}$$

Therefore

$$P(\text{CI includes } 0) = \Phi\left(\frac{5.658033 + 6}{\sqrt{8.33}}\right) - \Phi\left(\frac{-5.658033 + 6}{\sqrt{8.33}}\right) = 0.4528151$$

d

Once again we have the CIs

$$\bar{X} \pm 1.96 * \frac{25}{\sqrt{100}}$$

and

$$\bar{Y} \pm 1.96 * \frac{25}{\sqrt{300}}$$

We can repeat the work from b) to find the bounds on $\bar{X} - \bar{Y}$ needed for the CIs to intersect.

$$-7.729016 < \bar{X} - \bar{Y} < 7.729016$$

We then normalize using the new value of $\mu_x - \mu_y$.

$$\frac{-7.729016 + 6}{\sqrt{8.33}} < \frac{\bar{X} - \bar{Y} + 6}{\sqrt{8.33}} < \frac{7.729016 + 6}{\sqrt{8.33}}$$

Now compute the probability.

$$P(\text{CIs intersect}) = \Phi\left(\frac{7.729016 + 6}{\sqrt{8.33}}\right) - \Phi\left(\frac{-7.729016 + 6}{\sqrt{8.33}}\right) = 0.7254354$$

Simulation

```r
# change the parameters for a,b and c,d
S = 1e4 # number of simulations
nx = 100
ny = 300
mux = 0
muy = 0
sigma = 25
mudiff = mux-muy
se = sigma*sqrt(1/nx + 1/ny)


# difference in means CI
diff_cover = logical(S)
for (i in 1:S) {
  sx_i <- rnorm(nx, mux, 25)
  sy_i <- rnorm(ny, muy, 25)
  L = mean(sx_i - sy_i) - 1.96*se
  U = mean(sx_i - sy_i) + 1.96*se
  diff_cover[i] = L < 0 & U > 0
}

# 2 confidence intervals
intersect = logical(S)
for (i in 1:S) {
  sx_i <- rnorm(nx, mux, 25)
  sy_i <- rnorm(ny, muy, 25)
  L_x = mean(sx_i) - 1.96*(sigma/sqrt(nx))
  U_x = mean(sx_i) + 1.96*(sigma/sqrt(nx))
  L_y = mean(sy_i) - 1.96*(sigma/sqrt(ny))
  U_y = mean(sy_i) + 1.96*(sigma/sqrt(ny))
  intersect[i] = L_x < U_y & U_x > L_y
}

mean(diff_cover)
```

```
## [1] 0.9533
```

```r
mean(intersect)
```

```
## [1] 0.9929
```

Running the simulation gives answers of 0.9493 and 0.992 for a) and b) respectively. For parts c) and d), the simulation gives 0.4579 and 0.7258.

4

a

Integrate out X_1 to get the marginal pdf of X_2.

$$f_{X_2}(x_2) = 2e^{-x_2}\int_0^{x_2} e^{-x_1}dx_1 = 2e^{-x_2}\left[-e^{-x_1}\right]_0^{x_2} = 2e^{-x_2} - 2e^{-2x_2}$$

b

To find the expected value, we will need the pdf of $X_1|X_2$. Recall that

$$P(X_1|X_2) = \frac{P(X_1 \cap X_2)}{P(X_2)}$$

Therefore, we can find the conditional pdf by dividing the joint pdf by the marginal pdf of X_2.

$$f_{X_1|X_2}(x_1) = \frac{2e^{-(x_1+x_2)}}{2e^{-x_2} - 2e^{-2x_2}} = \frac{e^{-x_1}e^{-x_2}}{e^{-x_2}(1-e^{-x_2})} = \frac{e^{-x_1}}{(1-e^{-x_2})}$$

Therefore

$$E[X_1|X_2 = 2] = \int_0^2 x_1 \frac{1}{1-e^{-2}} e^{-x_1} dx_1 = \frac{1}{1-e^{-2}} \int_0^2 x_1 e^{-x_1} dx_1$$

Integrate by parts: $u = x_1$, $du = dx_1$, $dv = e^{-x_1}$, $v = -e^{-x_1}$.

$$= \frac{1}{1-e^{-2}} \left(-x_1 e^{-x_1} |_0^2 - \int_0^2 -e^{-x_1} dx_1 \right) = \frac{1}{1-e^{-2}} \left(-3e^{-2} + 1 \right) \approx 0.6869647$$

c

Use the cdf method.

$$P(Y \le y) = P(X_1 + X_2 \le y) = P(X_2 \le y - X_1)$$

5

a

$$E[S^2] = E\left[\frac{1}{n-1}\sum(x_i - \bar{x})^2\right] = \frac{1}{n-1} E\left[\sum((x_i-\mu) - (\bar{x}-\mu))^2\right]$$
$$= \frac{1}{n-1} E\left[\sum(x_i-\mu)^2 - 2(x_i-\mu)(\bar{x}-\mu) + (\bar{x}-\mu)^2\right]$$
$$= \frac{1}{n-1} E\left[\sum(x_i-\mu)^2 - n(\bar{x}-\mu)^2\right] = \frac{1}{n-1}\left(n\sigma^2 - n\,\mathrm{Var}(\bar{x})\right)$$
$$= \frac{1}{n-1}\left(n\sigma^2 - \sigma^2\right) = \sigma^2$$

b

Using some work from the previous part

$$S^2 = \frac{1}{n-1}\left(\sum(x_i-\mu)^2 - n(\bar{x}-\mu)^2\right)$$

Now let $Y_i = (x_i - \mu)^2$. Note that $EY_i = \sigma^2$. Therefore we have

$$S^2 = \frac{1}{n-1}\left(\sum Y_i - EY\right)$$

since $n(\bar{x}-\mu)^2 = n\,\mathrm{Var}(\bar{X}) = \sigma^2 = EY$. Now we can send $n \to \infty$

$$S^2 = \frac{1}{n-1}\left(n\bar{Y} - EY\right) = \frac{n}{n-1}\bar{Y} - \frac{EY}{n-1} \to^p 1 \cdot EY = \sigma^2$$

by the continuous mapping theorem and the weak law of large numbers.

c

The variance of $\bar{X}$ is θ/n. I bet it acheives the CR lower bound. Let's take a look at the log-likelihood.

$$\ell(\theta) = -n\theta + \sum x_i \log\theta - \log\prod x_i!$$

Now take derivatives.

$$\ell'(\theta) = -n + \frac{\sum x_i}{\theta}$$

$$\ell''(\theta) = -\frac{\sum x_i}{\theta^2}$$

Therefore the information is

$$I = -E\left[-\frac{\sum x_i}{\theta^2}\right] = \frac{n}{\theta}$$

Therefore, $\bar{X}$ achieves the Cramer-Rao lower bound, i.e. there is no estimator with less variance than $\bar{X}$. I haven't checked to see if S^2 hits the lower bound (my guess is that it does not, but I would have to do the calculation to know for sure), but I would say $\bar{X}$ is a better estimator because it hits the lower bound for variance and is simpler than S^2.