Yinan Kang

CSCI E-106: Data Modeling

Assignment 3

Due: February, 19 2019 at 7:19 pm EST

Instructions: Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit either scanned hand-written solution or typed solutions and two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document generated using knitr for .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

All questions are coming from Kutner, M. *et al*: Applied Linear Statistical Models, Fifth Edition.

1. (2.56)
2. (2.59)
3. (2.63)
4. (2.64)
5. (2.66)

2.56 $X = 1, 4, 10, 11, 14$

$\sigma = 0.6, \ \beta_0 = 5, \ \beta_1 = 3$

a) via lecture notes

$E(MSE) = \sigma^2$

$E(MSR) = \sigma^2 + \beta_1^2 \Sigma(X_i - \bar{X})^2$

$\boxed{E(MSE) = (0.6)^2 = 0.36}$

$\boxed{E(MSR) = 1026.36}$ calculation in .rmd file

b) Note: I am interpreting "whether or not a regression relation exists" to be same as "whether linear relationship" exists, which is why I bring up the 2-sided t-test: $H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$.

$\boxed{\text{Current set: } X's = [1, 4, 10, 11, 14]}$ would lead to greater $\Sigma(X_i - \bar{X})^2$, thus reducing variance, leading to higher t^*, meaning more likely Ha can be concluded, as

$|t^*| > t(\frac{1-\alpha}{2}, n-2)$ is decision criteria for concluding Ha.

The same logic applies when estimating a mean response. The wider spread in current set of X's reduces $s\{\hat{Y}_h\}$ (or $s\{pred\}$), leading to tighter confidence (or prediction) interval

$\therefore$ current set $X = [1, 4, 10, 11, 14]$ is more desirable in both situations.

Scanned by CamScanner

(2.59) Note: Replacing $f(Y_1, Y_2)$ with $f(X_i, Y_i)$ for clarity.

Density fn.:

$$f(X_i, Y_i) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}} \exp\left(-\frac{1}{2(1-\rho_{xy}^2)}\left[\left(\frac{X_i-\mu_x}{\sigma_x}\right)^2\right.\right.$$

a)

$$\left.\left. \cdots -2\rho_{xy}\left(\frac{X_i-\mu_x}{\sigma_x}\right)\left(\frac{Y_i-\mu_y}{\sigma_y}\right)+\left(\frac{Y_i-\mu_y}{\sigma_y}\right)^2\right]\right)$$

Likelihood fn:

$$\prod_{i=1}^{n} f(X_i, Y_i)$$

$$L(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho_{xy})$$

$$= \frac{1}{2\pi\sigma_x\sigma_y(1-\rho_{xy}^2)^{n/2}} \exp\left(-\frac{1}{2(1-\rho_{xy}^2)}\sum_{i=1}^{n}\left[\left(\frac{X_i-\mu_x}{\sigma_x}\right)^2\right.\right.$$

$$\left.\left. \cdots -2\rho_{xy}\left(\frac{X_i-\mu_x}{}\right)\left(\frac{Y_i-\mu_y}{}\right)+\left(Y_i-\mu_y\right)^2\right]\right)$$

$$\log_e L = -\log_e 2\pi - \log_e\sigma_x - \log_e\sigma_y - \frac{n}{2}\log_e(1-\rho_{xy}^2)$$

$$\cdots -\frac{1}{2(1-\rho_{xy}^2)}\sum_{i=1}^{n}\left[\left(\frac{X_i-\mu_x}{\sigma_x}\right)^2 - 2\rho_{xy}\left(\frac{X_i-\mu_x}{\sigma_x}\right)\left(\frac{Y_i-\mu_y}{\sigma_y}\right)\right.$$

$$\left. \cdots + \left(\frac{Y_i-\mu_y}{\sigma_y}\right)^2\right]$$

Log$_e$ Likelihood Fn:

rewriting for simplicity

$$\log_e L = -\log_e 2\pi - \log_e\sigma_x - \log_e\sigma_y - \frac{n}{2}\log_e(1-\rho_{xy}^2)$$

$$\cdots -\frac{1}{2}(1-\rho_{xy})^{-1}\sum_{i=1}^{n}\left[(X_i-\mu_x)^2\sigma_x^{-2} - 2\rho_{xy}(X_i-\mu_x)(Y_i-\mu_y)\sigma_x^{-1}\sigma_y^{-1}\right.$$

$$\left. \cdots + (Y_i-\mu_y)^2\sigma_y^{-2}\right]$$

$\hat{\sigma}_x$:
$$\frac{\partial \log_e L}{\partial \sigma_x} = -\frac{1}{\sigma_x} - \frac{1}{2}(1-\rho_{xy})^{-1}\sum_{i=1}^{n}\left[-2(X_i-\mu_x)^2\sigma_x^{-3} + 2\rho_{xy} \cdots (X_i-\mu_x)(Y_i-\mu_y)\sigma_x^{-2}\sigma_y^{-1}\right]$$

$\downarrow$ set $=0$, use estimators

$$\frac{1}{\hat{\sigma}_x} = -\frac{1}{2}(1-\hat{\rho}_{xy})^{-1}\sum_{i=1}^{n}\left[-2(\hat{X}_i-\hat{\mu}_x)^2\hat{\sigma}_x^{-3} + 2\hat{\rho}_{xy}(\hat{X}_i-\hat{\mu}_x)(\hat{Y}_i-\hat{\mu}_y) \cdots \hat{\sigma}_x^{-2}\hat{\sigma}_y^{-1}\right]$$

$(\times\hat{\sigma}_x^2)$ $\quad\rightarrow(\times\hat{\sigma}_x^2)$

$$\rightarrow \hat{\sigma}_x = -\frac{1}{2}(1-\hat{\rho}_{xy})^{-1}\sum_{i=1}^{n}\left[-2(\hat{X}_i-\hat{\mu}_x)^2\hat{\sigma}_x^{-1} + 2\hat{\rho}_{xy}(\hat{X}_i-\hat{\mu}_x)(\hat{Y}_i-\hat{\mu}_y)\hat{\sigma}_y^{-1}\right]$$

$\hat{\sigma}_y$:
$$\frac{\partial \log_e L}{\partial \sigma_y} = -\frac{1}{\sigma_y} - \frac{1}{2}(1-\rho_{xy})^{-1}\sum_{i=1}^{n}\left[2\rho_{xy}(X_i-\mu_x)(Y_i-\mu_y)\sigma_x^{-1}\sigma_y^{-2} - 2(Y_i-\mu_y)^2\sigma_y^{-3}\right]$$

$\downarrow$ set $=0$, use estimators

$$\frac{1}{\hat{\sigma}_y} = \ldots \text{ " } \text{ " }$$

$(\times\hat{\sigma}_y^2)$ $\quad (\times\hat{\sigma}_y^2)$

$$\rightarrow \hat{\sigma}_y = -\frac{1}{2}(1-\hat{\rho}_{xy})^{-1}\sum_{i=1}^{n}\left[2\hat{\rho}_{xy}(\hat{X}_i-\hat{\mu}_x)(\hat{Y}_i-\hat{\mu}_y)\hat{\sigma}_x^{-1} - 2(\hat{Y}_i-\hat{\mu}_y)^2\hat{\sigma}_y^{-1}\right]$$

$\hat{\rho}_{xy}$:
$$\frac{\partial \log_e L}{\partial \rho_{xy}} = -\frac{n}{2(1-\rho_{xy})}\left(\sum_{i=1}^{n}\left(-2(X_i-\mu_x)(Y_i-\mu_y)\sigma_x^{-1}\sigma_y^{-1}\right)\right)$$

$\downarrow$ set $=0$, use estimators

$$\left[-\frac{1}{2(1-\rho_{xy})^2}\cdot\left(\Sigma\cdots\right)\right] \leftarrow \text{forgot this term}$$

$$\cancel{\frac{n}{2(1-\hat{\rho}_{xy})}\cdot 2(1-\hat{\rho}_{xy})^2} = -\sum_{i=1}^{n}\left(-2(\hat{X}_i-\hat{\mu}_x)(\hat{Y}_i-\hat{\mu}_y)\hat{\sigma}_x^{-1}\hat{\sigma}_y^{-1}\right)$$

$$\rightarrow \hat{\rho}_{xy} = 1 - \left(\frac{-\sum_{i=1}^{n}\left(-2(\hat{X}_i-\hat{\mu}_x)(\hat{Y}_i-\hat{\mu}_y)\hat{\sigma}_x^{-1}\hat{\sigma}_y^{-1}\right)}{n}\right)$$

$$\frac{\ldots (Y_i - \mu_y)\sigma_x \sigma_y^{-1})}{n}$$

$\hat{\mu}_x$:

$$\frac{\partial \log_e L}{\partial \mu_x} = -\frac{1}{2}(1-\rho_{xy})^{-1}\sum_{i=1}^{n}\left[-2(X_i-\mu_x)\sigma_x^{-2} + 2\rho_{xy}(Y_i-\mu_y)\sigma_x^{-1}\sigma_y^{-1}\right]$$

↓ set = 0, use estimators

$$\rightarrow 0 = -\frac{1}{2}(1-\hat{\rho}_{xy})^{-1}\sum_{i=1}^{n}\left[-2(\hat{X}_i-\hat{\mu}_x)\sigma_x^{-2} + 2\hat{\rho}_{xy}(\hat{Y}_i-\hat{\mu}_y)\hat{\sigma}_x^{-1}\hat{\sigma}_y^{-1}\right]$$

unsure how to simplify further

$\hat{\mu}_y$:

$$\frac{\partial \log_e L}{\partial \mu_y} = -\frac{1}{2}(1-\rho_{xy})^{-1}\sum_{i=1}^{n}\left[2\hat{\rho}_{xy}(X_i-\mu_x)\sigma_x^{-1}\sigma_y^{-1} - 2(Y_i-\mu_y)\sigma_y^{-2}\right]$$

↓ set = 0, use estimators

$$\rightarrow 0 = -\frac{1}{2}(1-\hat{\rho}_{xy})^{-1}\sum_{i=1}^{n}\left[2\hat{\rho}_{xy}(\hat{X}_i-\hat{\mu}_x)\hat{\sigma}_x^{-1}\hat{\sigma}_y^{-1} - 2(\hat{Y}_i-\hat{\mu}_y)\hat{\sigma}_y^{-2}\right]$$

unsure how to simplify further

Scanned by CamScanner

(b)

$$\alpha_{x|y} = \mu_x - \mu_y \rho_{xy} \frac{\sigma_x}{\sigma_y}$$ 2.80(a)

$$\beta_{xy} = \rho_{xy} \frac{\sigma_x}{\sigma_y}$$ 2.80(b)

$$\sigma^2_{x|y} = \sigma_x^2 (1 - \rho_{xy}^2)$$ 2.80(c)

MLE :

$$\hat{\alpha}_{x|y} = \hat{\mu}_x - \hat{\mu}_y \hat{\rho}_{xy} \frac{\hat{\sigma}_x}{\hat{\sigma}_y}$$

$$\hat{\beta}_{xy} = \hat{\rho}_{xy} \frac{\hat{\sigma}_x}{\hat{\sigma}_y}$$

$$\hat{\sigma}^2_{x|y} = \hat{\sigma}_x^2 (1 - \hat{\rho}_{xy}^2)$$

Leaving expressions abbreviated, as the MLE estimators I derived in (a) were each very long (and on messier side).

(c) Regression LSE :

$$b_1 = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})^2}$$ 1.10a

$$b_0 = \bar{Y} - b_1 \bar{X}$$ 1.10b

While I cannot show that the MLE estimators I derived in (a) equate (1.10a) or (1.10b) when plugged into "$\alpha_{x|y}$" and "β_{xy}", I recognize that theoretically "$\hat{\beta}_{xy}$" serves as "b_1" for a bivariate normal distribution, and similarly "$\hat{\alpha}_{x|y}$" serves as "b_0".

Assignment3

Yinan Kang

2/16/2019

R Markdown

2.56 Calculation

```
data.X <- c(1,4,10,11,14)
var <- 0.6^2
sum.X <- 0
for (i in 1:length(data.X)) {
  sum.temp <- (data.X[i] - mean(data.X))^2
  sum.X <- sum.X + sum.temp
}

exp.msr <- var + (3^2)*sum.X
print(exp.msr)
```

```
## [1] 1026.36
```

Problem continued by hand.

Problem 2.63

```
cdi.df <- read.csv("/cloud/project/CDI.csv")
for (i in unique(cdi.df$Geographic.Region)) {
  data.temp <- dplyr::filter(cdi.df, cdi.df$Geographic.Region == i)
  lm.temp <- lm(data.temp$Per.Capita.Income ~ data.temp$X.Bachelor.s.degrees)
  assign(paste0("reg.lm.",i),lm.temp)
}

confint(reg.lm.1, level = 0.9)
```

```
##                                      5 %     95 %
## (Intercept)                    7809.8077 10637.82
## data.temp$X.Bachelor.s.degrees  460.5177   583.80
```

```
confint(reg.lm.2, level = 0.9)
```

```
##                                       5 %      95 %
## (Intercept)                    12627.0363 14535.774
## data.temp$X.Bachelor.s.degrees   193.4858   283.853
```

```
confint(reg.lm.3, level = 0.9)
```

```
##                                      5 %       95 %
## (Intercept)                    9516.0773 11543.4929
## data.temp$X.Bachelor.s.degrees  285.7076   375.5158
```

```r
confint(reg.lm.4, level = 0.9)
```

```
##                                     5 %       95 %
## (Intercept)                    6862.6967 10367.4086
## data.temp$X.Bachelor.s.degrees  364.7585   515.8729
```

The regression line slopes in different regions appear to have noticeable differences, with Region 2 having smaller slopes in the 90% confidence interval, Region 1 having higher slopes, and Regions 3 and 4's confidence intervals lie roughly in between.

So no, the regression lines in the different regions do not appear to have similar slopes.

Problem 2.64

```r
senic <- read.table("/cloud/project/APC1.DAT", quote="\"", comment.char="")
names(senic) <- c("ID", "length.of.stay", "age" , "infection.risk", "routine.culturing.ratio", "routine
head(senic,3)
```

```
##   ID length.of.stay  age infection.risk routine.culturing.ratio
## 1  1           7.13 55.7            4.1                     9.0
## 2  2           8.82 58.2            1.6                     3.8
## 3  3           8.34 56.9            2.7                     8.1
##   routine.chest.xray.ratio number.of.beds medical.school.affiliation
## 1                     39.6            279                          2
## 2                     51.7             80                          2
## 3                     74.0            107                          2
##   region average.daily.census number.of.nurses
## 1      4                  207              241
## 2      2                   51               52
## 3      3                   82               54
##   available.facilities.services
## 1                            60
## 2                            40
## 3                            20
```

```r
# Problem 2.64 refers to Problem 1.43, which calls for Y = "Average length of stay in a hospital",
# and three separate predictors ("Infection Risk", "Available Facilities
# and Services", and "Routine Chest X-ray Ratio")

lm.infection <- lm(senic$length.of.stay ~ senic$infection.risk)
lm.facility <- lm(senic$length.of.stay ~ senic$available.facilities.services)
lm.chest <- lm(senic$length.of.stay ~ senic$routine.chest.xray.ratio)

summary(lm.infection)
```

```
##
## Call:
## lm(formula = senic$length.of.stay ~ senic$infection.risk)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0587 -0.7776 -0.1487  0.7159  8.2805
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             6.3368     0.5213  12.156  < 2e-16 ***
## senic$infection.risk    0.7604     0.1144   6.645 1.18e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.624 on 111 degrees of freedom
## Multiple R-squared:  0.2846, Adjusted R-squared:  0.2781
## F-statistic: 44.15 on 1 and 111 DF,  p-value: 1.177e-09
```

```r
summary(lm.facility)
```

```
##
## Call:
## lm(formula = senic$length.of.stay ~ senic$available.facilities.services)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2712 -1.0716 -0.2816  0.7584  9.5433
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          7.71877    0.51020  15.129  < 2e-16
## senic$available.facilities.services  0.04471    0.01116   4.008 0.000111
##
## (Intercept)                         ***
## senic$available.facilities.services ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.795 on 111 degrees of freedom
## Multiple R-squared:  0.1264, Adjusted R-squared:  0.1185
## F-statistic: 16.06 on 1 and 111 DF,  p-value: 0.0001113
```

```r
summary(lm.chest)
```

```
##
## Call:
## lm(formula = senic$length.of.stay ~ senic$routine.chest.xray.ratio)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9226 -1.0810 -0.2708  0.8200  8.7008
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    6.566373   0.726094   9.043 5.67e-15 ***
## senic$routine.chest.xray.ratio 0.037756   0.008657   4.361 2.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.774 on 111 degrees of freedom
## Multiple R-squared:  0.1463, Adjusted R-squared:  0.1386
## F-statistic: 19.02 on 1 and 111 DF,  p-value: 2.906e-05
```

Using R.sq as the criterion, “X = Infection Risk” accounts for the largest reduction in the variability of the average length of stay, as its "adjusted R.sq" = 0.2781 is highest among the three.

Problem 2.66

```
rm(list=ls())

predictors <- c(4,8,12,16,20)

set.seed(39)
errors <- rnorm(5,mean = 0, sd = 5) #set sd=5 as variance = 25

outcomes <- predictors*4 +20 + errors

cumul.df <- data.frame(predictors = predictors, errors = errors, outcomes = outcomes)

attach(cumul.df)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     errors, outcomes, predictors
```

```
xbar <- mean(predictors)
ybar <- mean(outcomes)

b1 <- sum((predictors-xbar)*(outcomes-ybar))/(sum((predictors-xbar)^2))
b0 <- ybar - b1*xbar
y.hat <- b0 + b1*10

model.b <- lm(outcomes~predictors)

#'outcome.10' is the confidence interval for when Xh = 10 at 95% confidence (default)
outcome.10 <- predict(model.b,newdata=data.frame(predictors=10),interval="confidence")
```

b1 = 4.2423799
b0 = 15.1207714
Confidence Interval at 95% = 52.873717, 62.215424

2.66 (b)

```
# 'b1.vect' will store all calculated b1 values for part (c)
# 'outcome.df' will store all confidence intervals for part (d)

b1.vect <- c(rep(0,200))
outcome.df <- data.frame(lower_bound = 0, upper_bound = 0)
seeds <- sample(1:1000,200)

predictors <- c(4,8,12,16,20)

# Looping through 200 times:

for (j in 1:length(seeds)) {
```

```r
  set.seed(seeds[j])
  errors <- rnorm(5,mean = 0, sd = 5) #set sd=5 as variance = 25

  outcomes <- predictors*4 +20 + errors

  cumul.df <- data.frame(predictors = predictors, errors = errors, outcomes = outcomes)


  xbar <- mean(cumul.df$predictors)
  ybar <- mean(cumul.df$outcomes)

  b1 <- sum((cumul.df$predictors-xbar)*(cumul.df$outcomes-ybar))/(sum((cumul.df$predictors-xbar)^2))
  b0 <- ybar - b1*xbar
  y.hat <- b0 + b1*10

  model.b <- lm(cumul.df$outcomes~cumul.df$predictors)
  outcome.10 <- predict(model.b,newdata=data.frame("cumul.df$predictors"=10),interval="confidence")

  outcome.df[j,1] <- outcome.10[2]
  outcome.df[j,2] <- outcome.10[3]
  b1.vect[j] <- b1


}
# Showing snippets of 'outcome.df' and 'b1.vect'
head(outcome.df,3)
```

```
##   lower_bound upper_bound
## 1    51.37671    68.40979
## 2    50.15110    68.86338
## 3    51.94072    68.45251
```

```r
dim(outcome.df)
```

```
## [1] 200   2
```

```r
head(b1.vect,3)
```

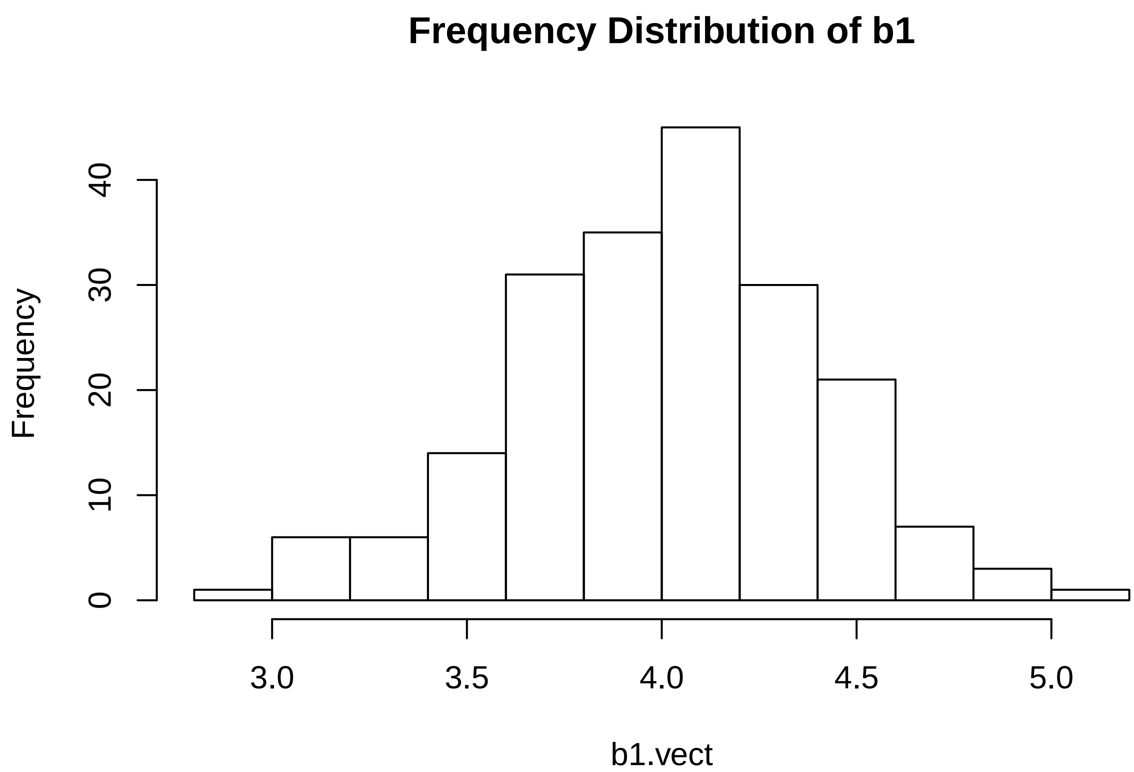
```
## [1] 4.258269 4.678071 4.127949
```

```r
length(b1.vect)
```

```
## [1] 200
```

2.66 (c)

```r
hist(b1.vect, main = "Frequency Distribution of b1")
```




```r
b1.sd <- sd(b1.vect)
b1.mean <- mean(b1.vect)
```

Standard Deviation of b1 = 0.3902433
Mean of b1 = 4.011482
These results are consistent with theoretical expectations, as 'b1.mean' approximately equals 'b1' calculated in (a), as well as the 'beta1' value when using lm(). Similarly for 'b1.sd', the standard deviation of b1 approximates the theoretical standard deviation.

2.66 (d)

```r
bad = 0
for (k in 1:nrow(outcome.df)) {
  if (outcome.df[k,2] > 60 && outcome.df[k,1] > 60) {
    bad = bad+1
  }
  if (outcome.df[k,2] < 60 && outcome.df[k,1] <60) {
    bad = bad+1
  }
}
print(bad)
```

```
## [1] 0
```

There are cases where confidence interval of E(Yh) when Xh = 10 does not include E(Yh). The proportion where E(Yh) IS included is 200%

While given an infinitely large # of trials, the proportion should be 95%, the result after 200 trials is within reason.