QUESTION 3:

a) Find and interpret 90% confidence interval for $\tau(y)$ total cost of bridge repair in Southern Ontario. You may not use additional explanatory variate information for this part of the question.

```
bridges <- read.table('Bridges.txt')
cost <- bridges[2:51, 2]
cost <- as.numeric(as.character(cost))
time <- bridges[2:51, 1]
time <- as.numeric(as.character(time))
urgency <- bridges[2:51, 3]
urgency <- as.numeric(as.character(urgency))

# a) 90% CI for total cost of bridge repairs

# estimates
a.n = length(cost)
a.mu.hat = mean(cost)
a.std.hat = sqrt(var(cost))

a.ci.lower = a.mu.hat - (qnorm(0.95))*(sqrt(1 - a.n/645) * (a.std.hat / sqrt(a.n)))
a.ci.upper = a.mu.hat + (qnorm(0.95))*(sqrt(1 - a.n/645) * (a.std.hat / sqrt(a.n)))

# 90% CI for avg cost
cat("90% CI for avg cost: (", a.ci.lower, ",", a.ci.upper , ")")

# 90% CI for total cost
cat("90% CI for total cost: (", a.ci.lower * 645, ",", a.ci.upper * 645 , ")")
```

```
## 90% CI for avg cost: ( 51.02935 , 60.44185 )
## 90% CI for total cost: ( 32913.93 , 38984.99 )
```

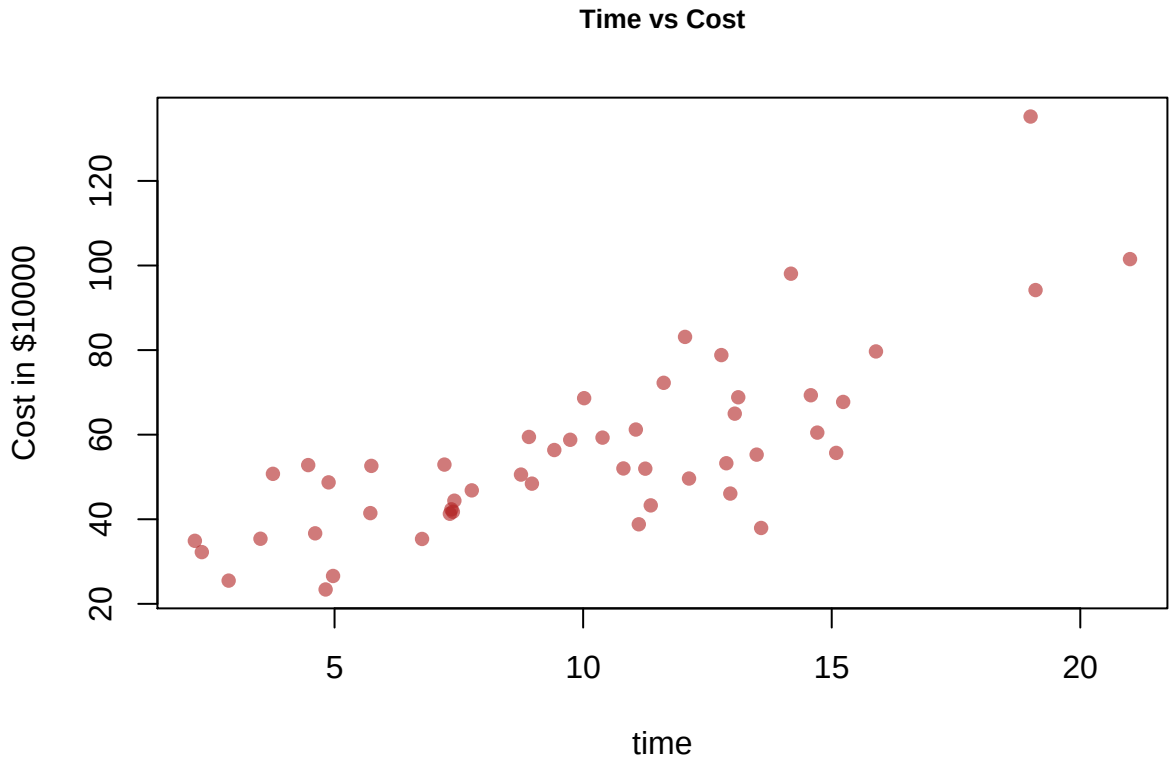
The 90% confidence interval for the average cost of bridge repairs (in $10,000) is (51.02935, 60.44185)

Given that there are 645 bridges, we estimate the 90% confidence interval for total cost (in $10,000) of bridge repairs to be (32913.93, 38984.99)

Hence we are 90% confident that the true value of the total cost (in $ 10,000) lies between $32913.93 and $38984.99

b) Plot the cost data vs. time and comment on the appropriateness of ratio and regression estimates of the total $\tau(y)$ in this example.

```
plot(x = time, y = cost, main = "Time vs Cost", pch = 16,
     col = adjustcolor("firebrick", 0.6), cex.main = 0.8,
     ylab = "Cost in $10000")
```



There looks to be a strong linear relationship between time and cost of repair, hence we can use ratio and regression estimates to produce a better estimate for the total cost compared to using the sample average.

c) For financial records, the population average number of years since last major repair is $\mu(x) = 9.1$. Using the ratio estimate of the mean, construct a 90% confidence interval for $\tau(y)$, the total cost of bridge repair, and interpret your finding.

```
mu.hat.x = mean(time)
mu.hat.y = mean(cost)
theta.hat = mu.hat.y / mu.hat.x
mu.x.true = 9.1
n = 50
N = 645

mu.hat.y.ratio <- theta.hat*mu.x.true
var.r <- var(cost - theta.hat*time)
var.mu.ratio <- (1/n) * (1 - n/N) * var.r
se.mu.ratio <- sqrt(var.mu.ratio)

# 90% CI for avg cost
mu.y.ratio.lower <- mu.hat.y.ratio - (qnorm(0.95))*se.mu.ratio
mu.y.ratio.upper <- mu.hat.y.ratio + (qnorm(0.95))*se.mu.ratio

# 90% CI for total cost
cat("90% CI for total cost: (", mu.y.ratio.lower * 645, ",", mu.y.ratio.upper * 645 , ")")
```

```
## 90% CI for total cost: ( 30442.52 , 35077.23 )
```

We see that the 90% confidence interval for the total cost (in \$10,000) is (30442.52, 35077.23)

So we are 90% confident that the true value of the total cost (in \$10,000) is between \$30442.52, \$35077.23

d) Repeat your analysis in part (c), but this time use the regression estimator of the mean.

```
mod <- lm(cost ~ time)
beta.hat <- summary(mod)$coef[2,1]
mu.hat.y.reg <- mu.hat.y + beta.hat*(mu.x.true - mu.hat.x)
y = cost
x = time

var.mu.reg <- (1/n) * (1 - n/N) * sum((y - mu.hat.y - beta.hat*(x - mu.hat.x))^2)/(n-1)
se.mu.reg <- sqrt(var.mu.reg)
mu.y.reg.lower <- mu.hat.y.reg - (qnorm(0.95))*se.mu.reg
mu.y.reg.upper <- mu.hat.y.reg + (qnorm(0.95))*se.mu.reg

# 90% CI for total cost
cat("90% CI for total cost: (", mu.y.reg.lower * 645, ",", mu.y.reg.upper * 645 , ")")
```

```
## 90% CI for total cost: ( 31965.69 , 35816.39 )
```

We see that the 90% confidence interval for the total cost (in $10,000) is (31965.69, 35816.39)

So we are 90% confident that the true value of the total cost (in $10,000) is between $31965.69, $35816.39

e) This survey was a pilot to test the engineering methodology and sampling protocol. Using all the results from the pilot survey, how large a sample is required so that we are 95% confident that π, the proportion of bridges needing urgent repair, is estimated within 0.025 of its true value?

We use formula from notes:

$$n = (\frac{1}{N} + \frac{l^2}{c^2\hat{\sigma}^2})^{-1}$$

we will use

$l = 0.025, c = 1.96, N = 645$

and estimated standard deviation

$$\hat{\sigma} \approx \sqrt{\hat{\pi}(1-\hat{\pi})}$$

```
# number of urgent bridges needing repair
urgent = urgency[urgency == 1]
pi.hat = length(urgent)
pi.hat
```

```
## [1] 14
```

so $\hat{\pi} = \frac{14}{50} = 0.28$

inputting the components together gives:

$$n \approx (\frac{1}{645} + \frac{(0.025)^2}{(1.96)^2(0.28 * (1-0.28))^2})^{-1} = 180.0699$$

So we need about 181 samples to satisfy the requirements

f) Based on the pilot survey results, it was decided to stratify the population using the time since the last major repair (x). One proposal was to use the following three strata for estimating the total cost of repair.

- Stratum 1 : at most 5 years sine the last major repair (160 bridges in this group)
- Stratum 2 : more than 5 years, but at most 10 years since the last major repair (230 bridges in this group)
- Stratum 3 : more than 10 years since the last major repair (255 bridges in this group)

For the purpose of this part of the question, assume that the pilot data was collected according to the stratified sampling protocol explained above. Find a 90% confidence interval for $\tau(y)$ the total cost of bridge repair, AND compare your finding to your answer in parts (a), (c), and (d). Use simple sample average to estimate the mean in each stratum. What do you observe?

```
# stratum weights
f.W1 = 160 / 645
f.W2 = 230 / 645
f.W3 = 255 / 645

# stratas
bridges.df = data.frame(time, cost, urgency)
f.strat1 = bridges.df[time <= 5, ] # n1 = 10
f.strat2 = bridges.df[time > 5 & time <= 10, ] # n2 = 14
f.strat3 = bridges.df[time > 10, ] # n3 = 26

# fpc for each stratum
f.f1 = 10 / 160
f.f2 = 14 / 230
f.f3 = 26 / 255

# mean
f.mu.hat = (f.W1 * mean(f.strat1$cost)) + (f.W2 * mean(f.strat2$cost)) + (f.W3 * mean(f.strat3$cost))

# std dev
f.var.mu = (f.W1)^2 * (1-f.f1) * (var(f.strat1$cost)/10) +
           (f.W2)^2 * (1-f.f2) * (var(f.strat2$cost)/14) +
           (f.W3)^2 * (1-f.f3) * (var(f.strat3$cost)/26)
f.sd.mu = sqrt(f.var.mu)

# 90 CI for total cost of bridge repair
mu.y.reg.lower <- 645 * (f.mu.hat - (qnorm(0.95))*f.sd.mu)
mu.y.reg.upper <- 645 * (f.mu.hat + (qnorm(0.95))*f.sd.mu)
cat("90% CI for total cost: (", mu.y.reg.lower, ",", mu.y.reg.upper, ")")
```

```
## 90% CI for total cost: ( 32004.62 , 36109.56 )
```

90% confidence interval for $\tau(y)$ in \$10,000:
- part (a): (32913.93, 38984.99)
- part (c): (30442.52, 35077.23)
- part (d): (31965.69, 35816.39)
- part (f): (32004.62, 36109.56)

The 90% confidence interval we found in this part has smaller length than the confidence intervals from part

(a), (c), but part (d) has the smallest length among all.

By observation, when we exploit the explanatory variate of time, it helps shorten the length of the confidence interval as the confidence intervals from part (c), (d), and (f) are much more shorter than part (a).

g) Repeat the analysis in part (f), but this time, use $\bar{y}$ in stratum 1, $\hat{u}_{ratio}(y)$ in stratum 2, and $\hat{u}_{reg}(y)$ in stratum 3. Is your confidence interval different from that of part (f)? Is this consistent with what you expected? Why or why not?

From financial records, the average number of years since the last major repair for stratum 1 is 3.5 years and for stratum 2 is 7 years.

1. First we find the population means of number of years since last major repair for stratums 2 and 3

```
# population avg for # of years since last major repair
g.mu.x.true.s1 = 3.5    # strata 1
g.mu.x.true.s2 = 7      # strata 2

# we know u(x) = 9.1 from part (c) and we also know
# stratum sizes:
# N1 = 160
# N2 = 230
# N3 = 255
g.mu.x.true = 9.1
g.mu.x.true.s3 = ((g.mu.x.true * 645) - (g.mu.x.true.s1 * 160) - (g.mu.x.true.s2 * 230)) / 255
cat("Population mean for x in strata 3: ", g.mu.x.true.s3)
```

```
## Population mean for x in strata 3:  14.50784
```

2. Now we set up the estimates for ratio and regression then construct the confidence intervals

```
# stratas
bridges.df = data.frame(time, cost, urgency)
f.strat1 = bridges.df[time <= 5, ]             # n1 = 10
f.strat2 = bridges.df[time > 5 & time <= 10, ] # n2 = 14
f.strat3 = bridges.df[time > 10, ]             # n3 = 26

# estimates for ratio
g.x2 = f.strat2$time
g.y2 = f.strat2$cost
g.mu.hat.x2 = mean(g.x2)
g.mu.hat.y2 = mean(g.y2)
g.theta.hat = g.mu.hat.y2 / g.mu.hat.x2

# estimates for regression
g.x3 = f.strat3$time
g.y3 = f.strat3$cost
g.mu.hat.x3 = mean(g.x3)
g.mu.hat.y3 = mean(g.y3)
g.mod <- lm(g.y3 ~ g.x3)
g.beta.hat <- summary(g.mod)$coef[2,1]
```

```r
# 1) point estimates
g.mu.hat.y.ratio <- g.theta.hat*g.mu.x.true.s2
g.mu.hat.y.reg <- g.mu.hat.y3 + g.beta.hat*(g.mu.x.true.s3 - g.mu.hat.x3)

# 2) variances
g.var.r <- var(g.y2 - g.theta.hat*g.x2)
g.var.mu.ratio <- (1/14) * (1 - 14/230) * g.var.r

g.var.mu.reg <- (1/26) * (1 - 26/255) * sum((g.y3 - g.mu.hat.y3 - g.beta.hat*(g.x3 -g.mu.hat.x3))^2)/(2


# 3) construct the 90% CI

# stratified estimate of pop avg
g.mu.hat = (f.W1 * mean(f.strat1$cost)) + (f.W2 * g.mu.hat.y.ratio) + (f.W3 * g.mu.hat.y.reg)

# stratified estimate of std dev
g.var.mu = (f.W1)^2 * (1-f.f1) * (var(f.strat1$cost)/10) +
  (f.W2)^2 * (g.var.mu.ratio/14) +
  (f.W3)^2 * (g.var.mu.reg/26)
g.sd.mu = sqrt(g.var.mu)

# 90 CI for total cost of bridge repair
g.mu.y.lower <- 645 * (g.mu.hat - (qnorm(0.95))*g.sd.mu)
g.mu.y.upper <- 645 * (g.mu.hat + (qnorm(0.95))*g.sd.mu)
cat("90% CI for total cost: (", g.mu.y.lower, ",", g.mu.y.upper, ")")
```

```
## 90% CI for total cost: ( 33398.12 , 35225.32 )
```

So the 90% confidence interval for $\tau(y)$ in \$10,000 from each part are:
- part (a): (32913.93, 38984.99)
- part (c): (30442.52, 35077.23)
- part (d): (31965.69, 35816.39)
- part (f): (32004.62, 36109.56)
- part (g): (33398.12, 35225.32)

We discovered in part (b) that there is a strong linear relationship between the time since last major repair and cost. Also we were able to use stratification including ratio and regression estimates for constructing the confidence interval in part (g), so it was expected that the confidence interval would be the shortest among the others.

From part (f), the confidence interval is (32004.62, 36109.56). Since we were able to exploit information on the explanatory variate of years since latest repair, the result of the shortened confidence interval was expected as ratio and regression estimates were used.

QUESTION 4:

a) Suppose the budget for the project allows for sampling an additional 300 bridges to the 50 samples in the pilot study. Using the optimal allocation, based on the results from the pilot study, how would you distribute these 300 samples across the three strata?

```
# std devs of stratums 1, 2, 3
sd.s1 = sd(f.strat1$cost)
sd.s2 = sd(f.strat2$cost)
sd.s3 = sd(f.strat3$cost)

# optimal allocation
op.denom = (f.W1 * sd.s1) + (f.W2 * sd.s2) + (f.W3 * sd.s3)
op.n1 = ((f.W1 * sd.s1) / op.denom) * 350
op.n2 = ((f.W2 * sd.s2)  / op.denom) * 350
op.n3 = ((f.W3 * sd.s3) / op.denom) * 350
```

Explanation:

1) Using the results from Question 3, we gather estimates:

$\hat{\sigma}_1 = 10.70222$, $\hat{\sigma}_2 = 7.356244$, $\hat{\sigma}_3 = 22.03112$
$W_1 = \frac{160}{645}, W_2 = \frac{230}{645}, W_3 = \frac{255}{645}$

2) We then use formula to get optimal allocation sample sizes

$n_h = \frac{(W_h \hat{\sigma}_h) n}{W_1 \hat{\sigma}_1 + W_2 \hat{\sigma}_2 + W_3 \hat{\sigma}_3}$

$n_1 = 66.42754 \approx 66$
$n_2 = 65.63542 \approx 66$
$n_3 = 217.937 \approx 218$

So when distributing the 300 samples:

Strata 1: we put in 66 - 10 = 56 samples
Strata 2: we put in 66 - 14 = 52 samples
Strata 3: we put in 218 - 26 = 192 samples

b) We would like to estimate the total cost for bridge repairs in Ontario based on the stratified random sampling protocol explained in part (f) of Question 4, such that we are 90% confident that the total cost is estimated within $10,000,000 of its true value. How many bridges must be samples if we follow an optimal allocation method? You may assume that the final population correction factor in each stratum is approximately 1.

1) Under optimal allocation, we will use formula from lecture

$$n \geq \left(\frac{c(W_1\hat{\sigma}_1+W_2\hat{\sigma}_2+W_3\hat{\sigma}_3)}{l}\right)^2$$

From the pilot study, we obtained

$\hat{\sigma}_1 = 10.70222$, $\hat{\sigma}_2 = 7.356244$, $\hat{\sigma}_3 = 22.03112$
$W_1 = \frac{160}{645}, W_2 = \frac{230}{645}, W_3 = \frac{255}{645}$

We will also need

$P(Z \leq c) = 0.95 \quad \Rightarrow \quad c = 1.644854$
$l = 1000$

2) Hence we get

$$n \geq \left(\frac{(1.644854)(\frac{160}{645}(10.70222)+\frac{230}{645}(7.356244)+\frac{255}{645}(22.03112))}{1000}\right)^2 = 0.002117497$$

Note there are 645 bridges in the population. To get number of samples so total cost is estimated within $10,000,000 of its true value with 90% confidence:

$$n \geq (645)^2 * 0.002117497 = 880.9317$$

so we need at least 881 samples

c) Answer the question in part (b) but this time, use proportional-to-size allocation. Compare you answer to that of part (b) and comment on your finding.

1) Under proportional-to-size allocation, we will use formula

$$n \geq \left(\frac{c\sqrt{W_1\hat{\sigma}_1^2+W_2\hat{\sigma}_2^2+W_3\hat{\sigma}_3^2}}{l}\right)^2$$

From the pilot study, we obtained

$\hat{\sigma}_1 = 10.70222$, $\hat{\sigma}_2 = 7.356244$, $\hat{\sigma}_3 = 22.03112$
$W_1 = \frac{160}{645}, W_2 = \frac{230}{645}, W_3 = \frac{255}{645}$

We will also need

$P(Z \leq c) = 0.95 \quad \Rightarrow \quad c = 1.644854$
$l = 1000$

2) Hence we get

$$n \geq \left(\frac{(1.644854)\sqrt{\frac{160}{645}(10.70222)^2+\frac{230}{645}(7.356244)^2+\frac{255}{645}(22.03112)^2}}{1000}\right)^2 = 0.0006482473$$

Note there are 645 bridges in the population. To get number of samples so total cost is estimated within \$10,000,000 of its true value with 90% confidence:

$$n \geq (645)^2 * 0.0006482473 = 269.6871$$

so we need at least 270 samples

In part (b), we need at least 881 samples which is significantly more than what we need under proportional-to-size allocation. It looks that it will be more expensive if we were to use optimal allocation to satisfy the requirement