STAT 4234/5234: Calculating the regression estimator for a population total

Consider the population $\{(x_i, y_i) : i = 1, \ldots, N\}$ and suppose we wish to estimate the population mean $\bar{y}_U$ based on a simple random sample of size n. We further suppose the value of $\bar{x}_U$, the population mean for the auxiliary variable, is known. In regression estimation we estimate $\bar{y}_U$ by

$$\hat{\bar{y}}_{\text{reg}} = \bar{y} + \hat{B}_1(\bar{x}_U - \bar{x}) = \bar{y} + r\frac{s_y}{s_x}(\bar{x}_U - \bar{x})$$

where s_x and s_y are the sample standard deviations of x and y, respectively, and r denotes the sample correlation coefficient.

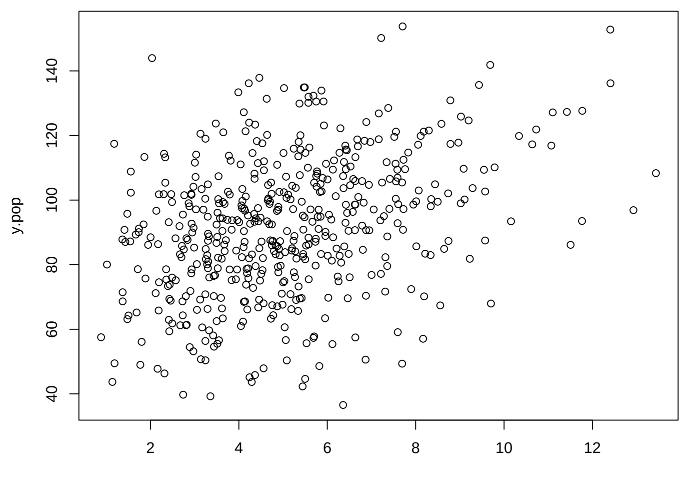
We have further seen that

$$\text{SE}\left(\hat{\bar{y}}_{\text{reg}}\right) = \sqrt{\frac{s_y^2(1-r^2)}{n}\left(1 - \frac{n}{N}\right)}$$

gives an expression for the standard error of the regression estimator.

To illustrate the computing for regression estimation, we create a fictional population of (x_i, y_i) as follows.

```
> set.seed(5234)
> x.pop <- rgamma(500, shape=5, rate=1)
> y.pop <- rnorm(500, mean=75+3*x.pop, sd=20)
> plot(x.pop, y.pop)
> cor(x.pop, y.pop)
[1] 0.3611513
> xbar.U <- mean(x.pop); xbar.U;
[1] 5.095567
```



The population correlation is $R = 0.36$, and the auxiliary variable population mean is $\bar{x}_U = 5.10$.

Now we take a simple random sample of size $n = 25$.

```
> N <- 500; n <- 25;
> samp <- sample(N, n)
> x.samp <- x.pop[samp]; y.samp <- y.pop[samp];
```

Calculate the regression estimator $\hat{\bar{y}}_{\text{reg}}$.

```
> xbar <- mean(x.samp); ybar <- mean(y.samp);
> fit <- lsfit(x.samp, y.samp)
> names(fit)
[1] "coefficients" "residuals"    "intercept"    "qr"
> fit$coefficients
Intercept         X
 84.77118   2.12263
> B1.hat <- as.numeric(fit$coefficients)[2]; B1.hat;
[1] 2.12263
> ybar.hat.reg <- ybar + B1.hat * (xbar.U - xbar)
> ybar.hat.reg
[1] 95.58719
```

And the standard error of our estimate.

```
> e <- fit$residuals
> V.hat <- var(e)/n * (1 - n/N)
> SE <- sqrt(V.hat); SE;
[1] 4.472836
```

Now a 95% confidence interval for $\bar{y}_U$ is

```
> ybar.hat.reg + c(-1,1) * 1.96 * SE
[1]  86.82043 104.35395
```

And a 95% CI for the population total t_y is

```
> N * ( ybar.hat.reg + c(-1,1) * 1.96 * SE )
[1] 43410.21 52176.97
```

Here is an R function that takes the sample data as inputs, along with N and $\bar{x}_U$, and returns the regression estimator of $\bar{y}_U$ along with its standard error.

```
regression.estimator.mean <- function(x.samp, y.samp, N, xbar.U)
{
 n <- length(y.samp)
 xbar <- mean(x.samp); ybar <- mean(y.samp);
 fit <- lsfit(x.samp, y.samp)
 B1.hat <- as.numeric(fit$coefficients)[2]
 ybar.hat.reg <- ybar + B1.hat * (xbar.U - xbar)
 e <- fit$residuals
 V.hat <- var(e)/n * (1 - n/N)
 SE <- sqrt(V.hat)
 answer <- c(point.est=ybar.hat.reg, std.error=SE)
 return(answer)
}
```

You can use this function for your homework if you wish.

```
> result <- regression.estimator.mean(x.samp=x.samp,
+   y.samp=y.samp, N=N, xbar.U=xbar.U)
> result
point.est std.error
95.587187  4.472836
```

A 95% confidence interval for $\bar{y}_U$ is

```
> result[1] + c(-1,1) * 1.96 * result[2]
[1]  86.82043 104.35395
```

and a 95% CI for the population total t_y is

```
> N * ( result[1] + c(-1,1) * 1.96 * result[2] )
[1] 43410.21 52176.97
```