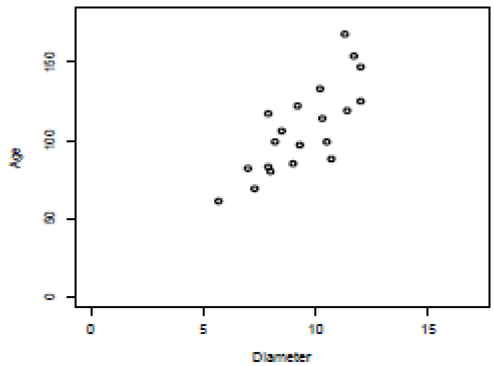
Referred Solution to Assignment 04

Part I : Problems from Chapter 4

4.3. (a) The scatterplot of age (y) vs. diameter (x):



(b) The diameter of a tree (x) is the auxiliary variable and $\bar{x}_{\mathrm{U}} = 10.3$.

$\hat{B} = \dfrac{\bar{y}}{\bar{x}} = 11.41946$ and $s_e^2 = \dfrac{1}{n-1}\sum_{i\in S}(y_i - \hat{B}x_i)^2 = 321.9330$.

(i) $\hat{\bar{y}}_r = \hat{B}\bar{x}_{\mathrm{U}} = 117.6204$.

(ii) $S.E(\hat{\bar{y}}_r) = \sqrt{\hat{V}(\hat{\bar{y}}_r)} = \sqrt{\left(1 - \dfrac{n}{N}\right)\left(\dfrac{\bar{x}_{\mathrm{U}}}{\bar{x}}\right)^2 \dfrac{s_e^2}{n}} = 4.3549$.

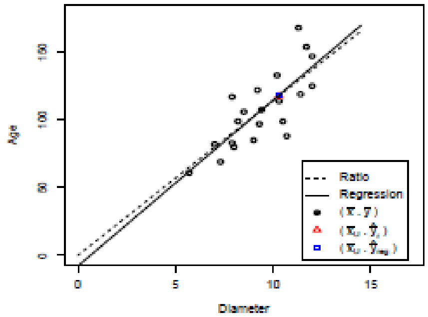
(c) $\hat{B}_1 = \dfrac{r s_y}{s_x} = 12.24966$ and $\hat{B}_0 = \bar{y} - \hat{B}_1\bar{x} = -7.808087$.

(i) $\hat{\bar{y}}_{\text{reg}} = \bar{y} + \hat{B}_1(\bar{x}_{\mathrm{U}} - \bar{x}) = 118.3634$.

(ii) $S.E(\hat{\bar{y}}_{\text{reg}}) = \sqrt{\left(1 - \dfrac{n}{N}\right)\dfrac{s_e^2}{n}} = 4.07077$

when using $s_e^2 = \dfrac{1}{n-2}\sum_{i\in S}(y_i - \hat{B}_0 - \hat{B}_1 x_i)^2 = 337.3848$.

(d) The plot along with the fitted line and labeled estimates:



From the plot and results in (b) and (c), we find that estimates based on ratio estimation and regression estimation are very close.

(e) (i) $\dfrac{|Bias(\hat{\bar{y}}_r)|}{\sqrt{V(\hat{\bar{y}}_r)}} \le \dfrac{\sqrt{V(\bar{x})}}{\bar{x}_U} = CV(\bar{x}) \Rightarrow \widehat{CV}(\bar{x}) = \dfrac{s_x/\sqrt{n}}{\bar{x}_U} = 3.97\%.$

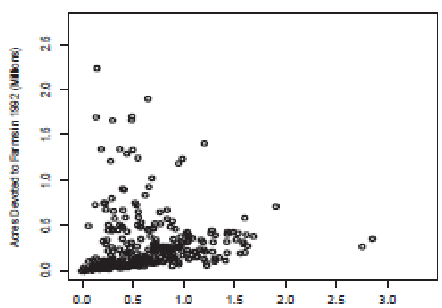
So, the bias of the ratio estimator is acceptable.

(ii) $|Bias(\hat{\bar{y}}_{\text{reg}})| = |-Cov(\hat{B}_1, \bar{x})| < SD(\hat{B}_1)SD(\bar{x})$

$\dfrac{|Bias(\hat{\bar{y}}_{\text{reg}})|}{\sqrt{V(\hat{\bar{y}}_{\text{reg}})}} \le \dfrac{SD(\hat{B}_1)SD(\bar{x})}{\sqrt{V(\hat{\bar{y}}_{\text{reg}})}} = Q \Rightarrow \hat{Q} = \dfrac{(2.304)(0.40894)}{4.07077} = 23.15\%.$

So, we are not sure whether the bias of the regression estimator is definitely problematic.

4.8. (a) Plot of acres devoted to farms in 1992 (y) vs. number of farms in 1987 (x):



(b) $\hat{B} = \dfrac{t_y}{\hat{t}_x} = \dfrac{\bar{y}}{\bar{x}} = 459.8975$ and $t_x = 2087759 \Rightarrow \hat{t}_{yr} = \hat{B}t_x = 960155061.$

(c) $\hat{B}_1 = \dfrac{rs_y}{s_x} = 47.65325,\ \bar{x}_U = 678.2843 \Rightarrow \hat{\bar{y}}_{\text{reg}} = \bar{y} + \hat{B}_1(\bar{x}_U - \bar{x}) = 299352.3.$

So, $\hat{t}_{y\text{reg}} = N\hat{\bar{y}}_{\text{reg}} = 921406265.$

(d)

Auxiliary Variable	$acres87$	$farms87$	$farms87$
Estimation Method	Ratio	Ratio	Regression
Standard Error	$S.E(\hat{t}_{yr,acres87})$	$S.E(\hat{t}_{yr,farms87})$	$S.E(\hat{t}_{yreg,farms87})$
	5546162	68446406	58163158

So, the ratio estimation with auxiliary variable $acres87$ is the most precise.

4.9. Herein, we continue exercise 4.8 to estimate the total number of acres devoted to farming in 1992 under subdomains by using the number of farms in 1987 as the auxiliary variable.

(a) Counties with fewer than 600 farms:
Define

$$x_i = \begin{cases} 1, & \text{if county } i \text{ with fewer than 600 farms in 1987} \\ 0, & \text{otherwise} \end{cases}$$

and define $u_i = y_i x_i$. Then, we have

(i) $\hat{t}_{yd} = \hat{t}_u = N\bar{u} = 473559072$,

(ii) $S.E(\hat{t}_{yd}) = N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_u^2}{n}} = 55528141$.

(b) Counties with 600 or more farms:
Similarly, define

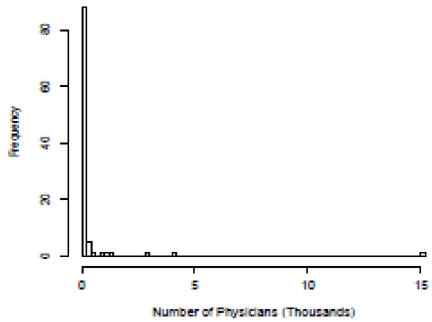
$$x_i = \begin{cases} 1, & \text{if county } i \text{ with 600 or more farms in 1987} \\ 0, & \text{otherwise} \end{cases}$$

and define $u_i = y_i x_i$. Then, we have

(i) $\hat{t}_{yd} = 443368037$,

(ii) $S.E(\hat{t}_{yd}) = 39595965$.

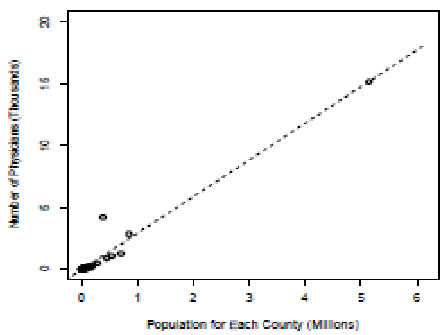
4.11. (a) Histogram of the number of physicians for the 100 counties:



(b) (i) $\hat{t}_{\text{SRS}} = N\bar{y} = 933411.$

(ii) $S.E(\hat{t}_{\text{SRS}}) = N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}} = 491982.8.$

(c) Plot of the number of physicians vs. population for each county:



Since the regression line does not go through the origin, we prefer the regression estimation. Regression result below ($\hat{B}_0 = -54.231$) also shows that the line does not go through the origin.

(d) $\hat{B}_1 = 0.002965,\ \hat{B}_0 = -54.23128,\ \bar{x}_{\text{U}} = 81209.02 \Rightarrow \hat{\bar{y}}_{\text{reg}} = 186.5236.$

(i) $\hat{t}_{y\text{reg}} = N\hat{\bar{y}}_{\text{reg}} = 585870.6.$

(ii) $S.E(\hat{t}_{y\text{reg}}) = N \cdot S.E(\hat{\bar{y}}_{\text{reg}}) = 105177.4.$

(e) From (a) and (d), we find that the regression estimation is closer to the true value.

4.20. (a) In large samples, we expect $\bar{x} \approx \bar{x}_{\text{U}}$. It is equivalent to show $s_e^2 = s_y^2 - 2\hat{B} r s_x s_y + \hat{B}^2 s_x^2$.

$$\begin{aligned}
s_e^2 &= \frac{1}{n-1} \sum_i (y_i - \hat{B} x_i)^2 \\
&= \frac{1}{n-1} \sum_i \left[y_i - \bar{y} - \hat{B}(x_i - \bar{x}) \right]^2 \quad \text{since } \bar{y} = \hat{B}\bar{x} \\
&= \frac{1}{n-1} \sum_i \left[(y_i - \bar{y})^2 - 2\hat{B}(y_i - \bar{y})(x_i - \bar{x}) + \hat{B}^2 (x_i - \bar{x})^2 \right] \\
&= \frac{\sum_i (y_i - \bar{y})^2}{n-1} - 2\hat{B} \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{n-1} + \hat{B}^2 \frac{\sum_i (x_i - \bar{x})^2}{n-1} \\
&= s_y^2 - 2\hat{B} r s_x s_y + \hat{B}^2 s_x^2.
\end{aligned}$$

(b) In example 4.2, we have $s_x^2 = 1.18716 \times 10^{11}$, $s_y^2 = 118907450529$, and $r = 0.995806$. So, when we do not truncate some of the significant digits on the calculation, we have the following results and conclude that it is exactly the same as the value computed by (4.10).

$\hat{V}(\hat{B})$	Exercise 4.20 (a)	Formula (4.10)
Use $\bar{x}_{\text{U}}$	3.070769×10^{-5}	3.070769×10^{-5}
Use $\bar{x}$	3.306794×10^{-5}	3.306794×10^{-5}

Part II : Extra Problems

6. $\hat{B} = \dfrac{\bar{y}}{\bar{x}} = \dfrac{16}{36}$, $t_x = 228000$, $\hat{t}_{yr} = \hat{B} t_x \approx 1011333$, and $S.E(\hat{t}_{yr}) = 10223.76$.

(i) 95% confidence interval for t_y:

$$\hat{t}_{yr} \pm t_{0.975,9} S.E(\hat{t}_{yr}) \Rightarrow (78205.57, 124461.10).$$

(ii) 95% Fieller confidence interval for B can be obtained form formula in the class note and replace $z_{\alpha/2}$ by $t_{0.975,9}$, namely, (L, U). So,

95% Fieller confidence interval for t_y: $(t_x L, t_x U) \Rightarrow (79068.35, 127466.62)$.

Since the sample size is very small, the Fieller confidence interval is more reliable.

7. If we ignore the correlation, we have $V(\hat{B}-\tilde{B}) = V(\hat{B})+V(\tilde{B})$. Hence, $\hat{V}(\hat{B}-\tilde{B}) = \hat{V}(\hat{B}) + \hat{V}(\tilde{B}) = 0.00824$. Then,

$$T = \frac{\hat{B} - \tilde{B}}{\sqrt{\hat{V}(\hat{B} - \tilde{B})}} = 1.9477.$$

Based on t distribution with $d.f = 33$, we obtain P-value $= 0.06$. At significant level α=0.05, we fail to reject H_0. It has no strong evidence that there is difference between these two treatments.

So, if we ignore the correlation, we may not statistically detect the difference between two treatments when the difference really exists.