STOCKHOLMS UNIVERSITET
Statistiska institutionen
Michael Carlson (examinator)

HT2022
2023-01-13 / rev 2023-02-02

Tentamen i Undersökningsmetodik (4,5 hp)

Kurs: Regressionsanalys och undersökningsmetodik

2023-01-13

Lösningförslag

Uppgift 1. (30p)

U = population med $N = 200$ företag, S = OSU u.å. med stickprovsstorlek $n = 30$. Låt y_k = omsättning för företag nummer k, och $\mu_y = \frac{1}{N}\sum_{k\in U} y_k$ = genomsnittlig omsättning samt $\tau_y = \sum_{k\in U} y_k$ = total omsättning för samtliga företag i hela populationen. Följande är givet:

$$\sum_{k\in S} y_k = 158 \qquad \sum_{k\in S} y_k^2 = 996$$

a) Punktskattning och ett 95% konfidensintervall för μ_y.

Skattning av μ_y: $\bar{y} = \frac{\sum_{k\in S} y_k}{n} = \frac{158}{30} = \boxed{\mathbf{5.2667}}$

Stickprovsvarians y_k: $s_y^2 = \frac{\sum_{k\in S} y_k^2 - n\bar{y}^2}{n-1} = \frac{996 - 30\cdot 5.2667^2}{29} = \mathbf{5.6506}$

Skattning av $V(\bar{y})$: $\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s_y^2}{n} = \left(1 - \frac{30}{200}\right)\frac{5.6506}{30} = \mathbf{0.16010}$

Standardfel: $SE(\bar{y}) = \sqrt{\hat{V}(\bar{y})} = \sqrt{0.16010} = \mathbf{0.40012}$

Felmarginal 95%: $z_{0.025}\cdot SE(\bar{y}) = 1.96 \cdot 0.40012 = \mathbf{0.78423}$

95% KI för μ_y: $= \mathbf{5.2667 \pm 0.78423}$ eller $(\mathbf{4.4824\,;\,6.0509})$

eller avrundat $\boxed{(\mathbf{4.48\,;\,6.05})}$

b) Punktskattning och ett 95% konfidensintervall för τ_y.

Kom ihåg: $\hat{V}(\hat{\tau}_y) = N^2\hat{V}(\bar{y}) \Rightarrow SE(\hat{\tau}_y) = N\cdot SE(\bar{y}) \Rightarrow \text{felmarg}(\hat{\tau}_y) = N\cdot\text{felmarg}(\bar{y})$

Skattning av τ_y: $\hat{\tau}_y = N\bar{y} = 200\cdot 5.2667 = \boxed{\mathbf{1\,053.3}}$

95% KI för τ_y: $(N\cdot\text{nedre gräns}\,;\,N\cdot\text{övre gräns}) =$

$= (200\cdot 4.48\,;\,200\cdot 6.05) = \boxed{(\mathbf{896\,;\,1\,210})}$

c) Felmarginalen ska vara högst hälften av det man fick i b)-uppgiften, dvs. när vi skattar τ_y.

Ett sätt:

Från b)-uppgiften beräknas att felmarginalen för $\hat{\tau}_y$ blev $(1210.2 - 896.85)/2 = 156.85$ (avrundat 157 är ok). Felmarginalen i den nya undersökningen ska vara högst hälften av detta:

max felmarginal: $$B_\tau = 1.96 \cdot D_\tau \leq \frac{\text{felmarginal i b)}}{2} = \frac{156.85}{2} = 78.425$$

max standardfel: $$D_\tau = \frac{B}{1.96} \leq \frac{78.425}{1.96} = 40.013$$

max varians: $$D_\tau^2 \leq 40.013^2 = 1601.0$$

Vi vet att $\tau_y = N\mu_y$ och att $V(\hat{\tau}_y) = N^2 V(\bar{y})$. I formeln för minsta stickprovsstorlek ersätts alltså σ_y^2 med $(N^2\sigma_y^2)$. Eftersom vi inte vet värdet på σ_y^2 använder vi s_y^2 från a)-uppgiften vilket ger

$$n \geq \frac{N(N^2\sigma_y^2)}{D_\tau^2(N-1) + \sigma_y^2} \approx \frac{N^3 s_y^2}{D_\tau^2(N-1) + (N^2 s_y^2)} = \frac{200^3 \cdot 5.6506}{1601.0 \cdot 199 + 200^2 \cdot 5.6506} = \mathbf{83.002}$$

dvs. $\boldsymbol{n \geq 84}$ (OBS! Avrunda alltid uppåt!)

NOT: beroende på hur du har avrundat kan du ha fått $n \geq 83$, vilket är ok.

Ett annat sätt (ska motiveras):

Vi vet att $\tau_y = N\mu_y$ och $\mu_y = \tau_y/N$ vilket betyder att den enda skillnaden mellan resultaten i a) och b) är att man multiplicerar $\bar{y}$ med $N = 200$ för att få $\hat{\tau}_y$. Alltså kan uppgiften lika gärna lösas genom att istället utgå ifrån a) och kräva att felmarginalen för $\bar{y}$ högst får vara lika med $B_\mu = 0.78423/2 = 0.39212$ (hälften av felmarginalen i a).

max felmarginal: $$B_\mu \leq 0.39212$$ alt. $$B_\mu = \frac{B_\tau}{N} \leq \frac{78.425}{200} = 0.39212$$

max standardfel: $$D_\mu = \frac{B}{z_{0.025}} \leq \frac{0.39212}{1.96} = 0.20006$$

max varians: $$D_\mu^2 = 0.20006^2 = 0.040025$$

Sedan används formeln i formelsamlingen med s_y^2 som din bästa gissning (skattning) för det okända σ_y^2:

$$n \geq \frac{N\sigma_y^2}{D_\mu^2(N-1) + \sigma_y^2} \approx \frac{N s_y^2}{D_\mu^2(N-1) + s_y^2} = \frac{200 \cdot 5.6506}{0.040025 \cdot 199 + 5.6506} = \mathbf{83.002}$$

dvs. $\boldsymbol{n \geq 84}$ (OBS! Avrunda alltid uppåt!)

d) Låt P = andelen företag i U som har en omsättning ≥ 8 mkr. Låt $x_k = 1$ om företag nummer k är sådan att $y_k \geq 8$ och $x_k = 0$ annars. I stickprovet S kan vi lätt konstatera att antalet företag med omsättning ≥ 8 är $\sum_{k\in S} x_k = 6$.

Punktskattning och ett 95% konfidensintervall för P:

Andelsskattning: $\hat{p} = \bar{x} = \frac{\sum_{k\in S} x_k}{n} = \frac{6}{30} = \mathbf{0.20}$

Skattning av $V(\hat{p})$: $\hat{V}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1} = \left(1 - \frac{30}{200}\right)\frac{0.16}{29} = \mathbf{0.0046897}$

Standardfel: $se(\hat{p}) = \sqrt{\hat{V}(\hat{p})} = \sqrt{0.0046897} = \mathbf{0.068481}$

Felmarginal: $z_{0.025} \cdot se(\hat{p}) = 1.96 \cdot 0.068481 = 0.13422$

95% KI för P: $\mathbf{0.20 \pm 0.13422}$ eller avrundat $(\mathbf{0.0658}\,;\mathbf{0.334})$

eller i %-enheter $\mathbf{20.0\% \pm 13.4\%}$ eller $(\mathbf{6.6\%}\,;\ \mathbf{33.4\%})$

e) Man vill redovisa genomsnittlig omsättning för kategorierna K = ”företag med kvinnlig vd” och M = ”företag med manlig vd” var för sig. Detta ska göras efter att urvalet har dragits med OSU u.å.

Vi vet inte på förhand vilka företag i U som tillhör respektive kategori och vi vet inte alltid hur stora delpopulationerna är, dvs. hur stora N_K och N_M är. Kategorierna brukar kallas ***domäner*** och skattningarna inom respektivedomän är ***domänskattningar***, t.ex. $\bar{y}_K$ och $\bar{y}_M$.

Stickprovsstorlekarna för respektive kategori, n_K respektive n_M, är slumpvariabler, eftersom vi inte i förväg har styrt upp hur många som ska dras från respektive kategori såsom man gör med ett stratifierat urval. Detta medför att n_K och n_M är stokastiska och att felmarginalerna blir något större.

KOMMENTAR: Om du har angett poststratifiering så får du delvis rätt. Poststratifiering är en skattningsmetod där domänskattningar för samtliga domäner vägs ihop till en skattning av en populationsparameter för hela populationen såsom vid ett stratifierat urval. Men då ska man helst känna till populationsstorlekarna N_K och N_M vilket inte alls är säkert. Och så måste man observera minst 2 värden i varje domän (post-strata) vilket inte är säkert särskilt om man har små stickprov och många domäner.

Alltså, planerat i förväg med fix stickprovsstorlek inom varje kategori så kallas det *stratum*, ej planerat i förväg med stokastisk stickprovsstorlek inom varje kategori så kallas det *domän* eller möjligen *post-strata*.

Uppgift 2. (20p)

$U =$ population bestående av $N = 1000$ hushåll, $S =$ stratifierat OSU u.å. med total stickprovsstorlek $n = 100$ fördelat på $L = 2$ strata med proportionell allokering. Stratumens populationsstorlekar N_h, stickprovsmedelvärden $\bar{y}_h$ och standardavvikelser s_h är givna i uppgiften, vissa andra storheter saknas och måste beräknas.

a) Punktskattning och ett 95% konfidensintervall för μ_y, det genomsnittliga sparandet för hela populationen.

Börja med att beräkna stratumvikterna $W_h = N_h/N$ och summera de viktade medelvärdena (grå celler i tabellen var ej givna, de måste beräknas):

$$\bar{y}_{\text{str}} = \sum_{h=1}^{L} W_h \bar{y}_h = W_1\bar{y}_1 + W_2\bar{y}_2 = 0.6 \cdot 90 + 0.4 \cdot 135 = \mathbf{108}$$

h	N_h	W_h	$\bar{y}_h$	$W_h\bar{y}_h$
1	600	0.6	90	54
2	400	0.4	135	54
Summa	1000	1.0	-	**108**

Beräkna sedan standardfelet och felmarginalen. Standardavvikelserna s_h var givna i uppgiften (kom ihåg att kvadrera dem!). Stickprovsstorlekarna fastställdes med proportionell allokering vilket innebär att $n_h = nW_h = 100W_h$ (grå celler i tabellen beräknas):

$$\hat{V}(\bar{y}_{\text{str}}) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h} = W_1^2\left(1 - \frac{n_1}{N_1}\right)\frac{s_1^2}{n_1} + W_2^2\left(1 - \frac{n_2}{N_2}\right)\frac{s_2^2}{n_2}$$

h	n_h	s_h	W_h^2	$1 - \frac{n_h}{N_h}$	s_h^2/n_h	$W_h^2(1 - \frac{n_h}{N_h})\frac{s_h^2}{n_h}$
1	60	50	0.36	0.9	41.667	13.50
2	40	70	0.16	0.9	122.50	17.64
Summa	100	-	-	-	-	**31.14**

Standardfel: $SE(\bar{y}_{\text{str}}) = \sqrt{\hat{V}(\bar{y}_{\text{str}})} = \sqrt{13.14} = \mathbf{5.5803}$

Felmarginal 95%: $z_{0.025} \cdot SE(\bar{y}) = 1.96 \cdot 1.8601 = \mathbf{10.937}$

95% KI för μ_y: avrundat $\mathbf{108 \pm 10.94}$ eller i heltal kronor $(\mathbf{97}\,;\mathbf{119})$

b) Den tillgängliga informationen är stratumstorlek N_h och observerade varianserna s_h^2. Vi använder s_h^2 då vi inte vet de sanna värdena σ_h^2. Använd formeln för formeln för optimal (Neyman) allokering:

$$n_h = n \cdot \frac{N_h \sigma_h}{\sum_{j=1}^{L} N_j \sigma_j} \approx n \cdot \frac{N_h s_h}{\sum_{j=1}^{L} N_j s_j} = 100 \cdot \frac{N_h s_h}{58\,000}$$

h	N_h	s_h	$N_h s_h$	$\frac{N_h s_h}{\sum_{j=1}^{L} N_j s_j}$	n_h	n_h av-rundat
1	600	50	30 000	0.51724	51.724	**52**
2	400	70	28 000	0.48276	48.276	**48**
Summa	1000	-	58 000	1	100	**100**

KOMMENTAR: Precis som väntat kommer man med den optimala allokeringen att dra lite fler från stratum 2 och något färre i stratum 1 jämfört med proportionell allokering som endast tar hänsyn till stratumens populationsstorlekar. Justeringen kommer förstås av att standardavvikelsen är större i stratum 2 än i stratum 1. Man kan för skojs skull plugga in de nya optimala stickprovsstorlekarna 52 resp. 48 i beräkningarna i a-uppgiften ovan och låta allt annat vara lika och jämföra. Då blir standardfelet $se(\bar{y}_{\text{str}}) = 5.49$ vilket endast är en liten förbättring jämfört med den ursprungliga allokeringens 5.58, en minskning om endast ca 1.6 %. Skillnaden i standardavvikelserna 50 resp. 70 är inte tillräckligt stor för att vinsten med optimal allokering ska bli särskilt märkbar.

c) y-värdena inom ett stratum ska likna varandra (homogenitet inom strata) dvs. ligga nära $\mu_h =$ stratumets medelvärde. Då blir $\sigma_h^2 =$ varianserna inom strata små. Idealt är det om alla σ_h^2 är mindre än $\sigma^2 =$ variansen för hela sammantagna populationen. Vidare bör stratumens olika medelvärden helst vara olika varandra (heterogenitet mellan strata); då fångar man med större sannolikhet upp hela bredden av y-värden som finns i populationen. Om detta är uppfyllt så är stratifierat OSU att föredra.

Om man i förväg planerar att redovisa statistik för delgrupper är det bättre att stratifiera populationen i förväg och dra urval från respektive delgrupp = stratum med i förväg fastställda stickprovsstorlekar från varje stratum. Alternativet är att man definierar dessa delgrupper = domäner i efterhand och att stickprovsstorlekarna blir slumpmässigt bestämda (jämför med uppgift 1c) ovan).

När man jämför urvalsdesigner och skattningsmetoder utgår man typiskt från skattningens varians eller standardfel samt eventuell bias. Dessa kan vägas ihop till det s.k. medelkvadratfelet $MSE = Varians + Bias^2$. Designer och metoder som leder till lägre MSE är att föredra. Man kan ibland acceptera lite bias om det blir en kraftig reducering i variansen (osäkerhet).

Uppgift 3. (20p)

U = population med $N = 90$ företag, S = OSU u.å. med stickprovsstorlek $n = 9$.

Låt y_k = personalkostnaderna och x_k = omsättningen för butik nummer k.

Givet: $\sum_{k\in S} y_k = 62.4$ $\quad \sum_{k\in S} y_k^2 = 552.1$

$\sum_{k\in S} x_k = 547.0$ $\quad \sum_{k\in S} x_k^2 = 45\,767.0$

$\sum_{k\in S} x_k y_k = 4\,964.4$ $\quad \sum_{k\in U} x_k = \tau_x = 4\,500$ $\quad \mu_x = \tau_x/N = 50$

Från detta kan följande beräknas:

$\bar{y} = 6.9333$ $\quad s_y^2 = 14.93$ $\quad \bar{x} = 60.778$ $\quad s_x^2 = 1565.2$

$s_{xy} = 146.48$ $\quad \mu_x = \tau_x/N = 50$

a) Skatta τ_y = den totala personalkostnaden med en kvotskattning $\hat{\tau}_{\text{kvot}}$ där omsättning används som hjälpvariabel och beräkna standardfelet för skattningen.

Skattning av kvoten: $\hat{R} = \dfrac{\sum_{k\in S} y_k}{\sum_{k\in S} x_k} = \dfrac{62.4}{547.0} = \dfrac{\bar{y}}{\bar{x}} = \dfrac{6.9}{60.8} = \mathbf{0.11408}$

Punktskattning: $\hat{\tau}_{\text{kvot}} = \hat{R} \cdot \tau_x = 0.11408 \cdot 4\,500 = \boxed{\mathbf{513.3455 \approx 513.3}}$

Beräkna summan: $\sum_{k\in S}(y_k - \hat{R}x_k)^2 = \sum_{k\in S} y_k^2 - 2\hat{R}\sum_{k\in S} x_k y_k + \hat{R}^2 \sum_{k\in S} x_k^2$

$= 552.1 - (2 \cdot 0.11408 \cdot 4\,964.4) + (0.11408^2 \cdot 45\,767) = \mathbf{15.024}$

Variansskattning: $\hat{V}(\hat{\tau}_{\text{kvot}}) = N^2\left(1 - \dfrac{n}{N}\right) \cdot \dfrac{1}{n} \cdot \left(\dfrac{\sum_{k\in S}(y_k - \hat{R}x_k)^2}{n-1}\right)$

$= 90^2\left(1 - \dfrac{9}{90}\right) \cdot \dfrac{1}{9} \cdot \left(\dfrac{15.024}{8}\right) = \mathbf{1521.166}$

Standardfel: $se(\hat{\tau}_{\text{kvot}}) = \sqrt{\hat{V}(\bar{y})} = \sqrt{151.166} = \boxed{\mathbf{39.002 \approx 39.0}}$

b) Skatta τ_y = den totala personalkostnaden med en regressionsskattning $\hat{\tau}_{\text{reg}}$ där omsättning används som hjälpvariabel och beräkna standardfelet för skattningen.

Lutningskoefficient: $b = \dfrac{n\sum x_k y_k - (\sum x_k)(\sum y_k)}{n\sum x_k^2 - (\sum x_k)^2} = \dfrac{9 \cdot 4\,964.4 - 547.0 \cdot 62.4}{9 \cdot 45\,767 - 547.0^2} =$

$= \dfrac{s_{xy}}{s_x^2} = \dfrac{146.48}{1565.2} = \mathbf{0.093588}$

forts. nästa sida

Punktskattning: $\hat{\tau}_{\text{reg}} = N\hat{\mu}_{\text{reg}} = N\left(\bar{y} + b(\mu_x - \bar{x})\right)$

$= 90\left(6.9 + 0.093588 \cdot (50 - 60.778)\right) = \boxed{\mathbf{533.2197 \approx 533.2}}$

Beräkna summan: $\sum_{k\in S}(y_k - \bar{y})^2 - b^2 \sum_{k\in S}(x_k - \bar{x})^2 = (n-1)s_y^2 - b^2(n-1)s_x^2$

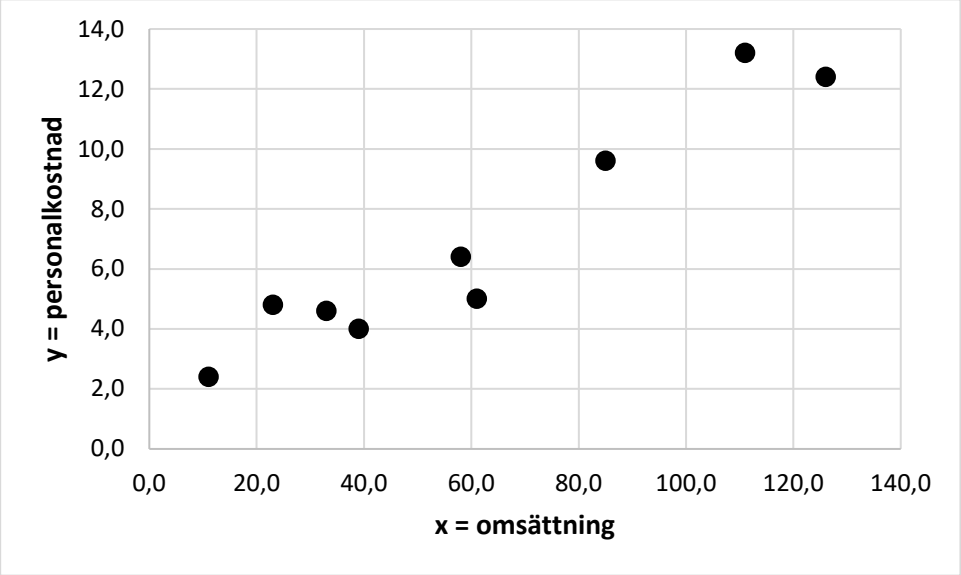
$= 8 \cdot 14.93 - 0.093588^2 \cdot 8 \cdot 1565.2 = \mathbf{9.7674}$

Variansskattning: $\hat{V}(\hat{\tau}_{reg}) = N^2\left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \left(\frac{\sum_{k\in S}(y_k - \bar{y})^2 - b^2 \sum_{k\in S}(x_k - \bar{x})^2}{n-2}\right)$

$= 90^2 \cdot \left(1 - \frac{9}{90}\right) \cdot \frac{1}{9} \cdot \left(\frac{9.7674}{7}\right) = \mathbf{1130.228}$

Standardfel: $se(\hat{\tau}_{\text{reg}}) = \sqrt{\hat{V}(\bar{y})} = \sqrt{1130.228} = \boxed{\mathbf{33.6189 \approx 33.6}}$

c) Spridningsdiagram

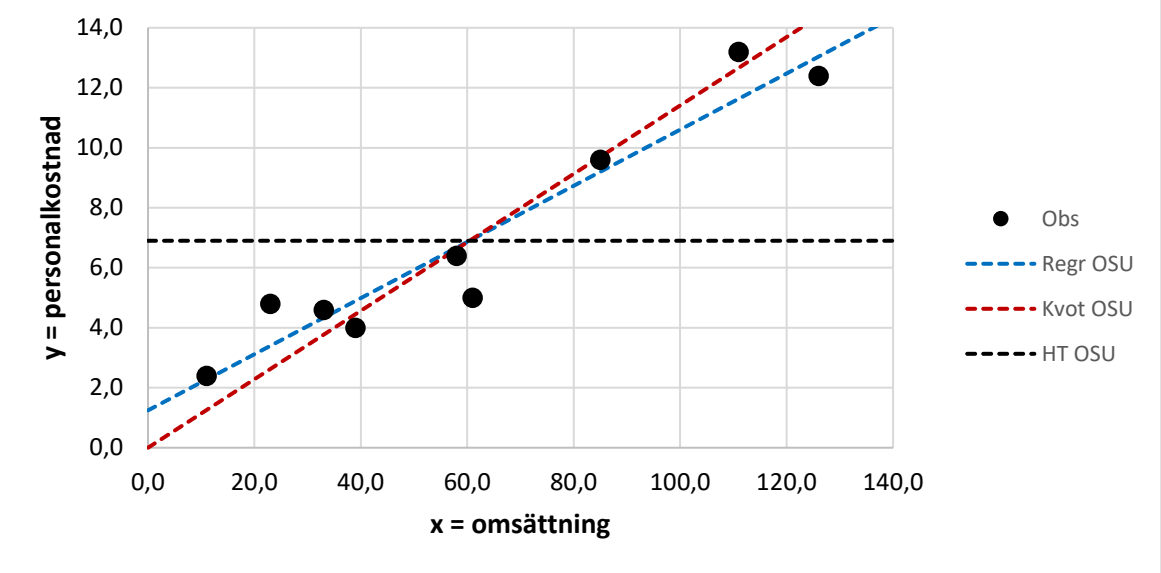


För regressionsskattning ska det finnas ett **linjärt samband**, när x ökar så ökar även y och det ser ut som om en rak linje skulle kunna anpassas till observationerna. Tveksamt om linjen går genom **origo**, linjen tycks skära y-axeln vid en punkt nära noll men kanske närmare $y = 1$. Det ser även ut att vara ganska **homoskedastiskt**, spridningen runt den tänkta linjen ser ut att vara lika stor oberoende av värdet på x.

För regressionsskattning ska det finnas ett **linjärt samband**, men linjens intercept ska vara noll med tolkningen att butiker utan omsättning ($x = 0$) i snitt har noll personalkostnader ($y = 0$) vilket i och för sig kan vara rimligt att anta. Om dessutom variansen för y_k ökar när x_k ökar så passar det också med kvotmodellen; det ska se ut som en ”**tratt**” i spridnings-diagrammet (heteroskedastiskt) vilket det inte ser ut att göra.

Slutsats: Sammantaget verkar det bli bättre med en regressionsskattning än en kvotskattning.

KOMMENTAR: Det efterfrågades inte i uppgiften men vi kan rita in de skattade linjerna för respektive modell i diagrammet och inkludera HT-skattningen av μ_y, dvs. $\bar{y} = 6.9333$. Regressions- och kvotlinjerna är $y = 1.2453 + 0.093588x$ respektive $y = 0.11401x$. Vi ser att den blå linjen verkar vara lite bättre anpassad till punkterna jämfört med den röda; antalet punkter som ligger närmare den blå linjen är något fler än antalet punkter som ligger närmare den röda.



Ingick inte heller i uppgiften men vi kan jämföra punktskattningarna och standardfelen med varandra och med HT-skattningen under OSU u.å.:

$$\hat{\tau}_{\text{HT}} = N\bar{y} = 90 \cdot 6.9333 = 624$$

$$se(\hat{\tau}_{\text{HT}}) = \sqrt{N^2\hat{V}(\bar{y})} = N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_y^2}{n}} = 200 \cdot \sqrt{0.9 \cdot \frac{14.93}{8}} = 109.97 \approx 110.0$$

Skattningsmetod	Punktskattning	Standardfel
HT	624.0	110.0
Kvot	513.2	39.0
Regression	533.2	33.6

Både kvot- och regressionsskattningarna justerar ner HT-skattningen. Detta är väntat eftersom vi såg att det observerade $\bar{x} = 60.778$ ligger en bra bit över det kända $\mu_x = 50$; man har fått lite för många företag med stor omsättning och därmed även lite större personalkostnader.

Båda kvot och regressionsskattning fungerar utmärkt i detta fall, både ger en kraftig reduktion i standardfelet. Regressionsskattningen är den som presterar bäst.

Vi kan inte utifrån det som gavs i uppgiften jämföra bias i kvot- och regressionsskattningarna som vi vet finns men den är förmodligen inte överväldigande.

Uppgift 4. (20p)

KOMMENTAR: Följande är förslag på svar, mer än vad som anges här kan nästan säkert sägas. Och kom ihåg att man inte behöver ge så här långa svar, kort och koncist räcker!

a) För ett *systematiskt urval*, utgår man från en ram där samtliga N urvalsenheter har listats i någon ordning. Stickprovet ska vara av storlek n, man beräknar kvoten $N/n = r$ och väljer slumpmässigt ett tal från $1, \dots, r$, säg att det blev $t =$ startpunkt. Slutligen väljs objekt nr t och sedan var r:te objekt till stickprovet $S = \{t,\ (t+r),\ (t+2r),\ (t+3r), \dots, (t+(n-1)r)\}$. Se F7 sid 9-12.

Fördelar: Det är enkelt, kan nästan göras för hand. Om ordningen i ramen följer y-värdenas storleksordning (dvs. om ordningsnummer och y-värde är starkt korrelerade) så kommer stickprovet innehålla en bra spridning av y-värden och standardfelet kommer att vara mindre än om det vore ett rent OSU eftersom de olika möjliga stickproven är ganska lika varandra. Detta kan dock vara svårt att veta något om i praktiken.

Nackdelar: Om ramen avspeglar en periodicitet i y-värdena (t.ex. dagliga mätningar, månads-data osv.) och r följer samma periodicitet riskerar man att startpunkter kraftigt påverkar punktskattningen och att man underskattar varians och standardfel.

b) *Mixed mode*; innebär att man inom samma undersökning använder en mix av flera datainsamlingsmetoder, t.ex. både webb och intervju.

Fördelar: då man anpassar insamlingen till respondenten kan man få bättre kvalitet t.ex. mindre mätfel och minskat bortfall.

Nackdelar: dyrare då det inte bara är en insamling och mätinstrument som ska designas. Det kan vara svårt att skapa likvärdiga mätningar - att alla tolkar frågan och svarar på ”samma” sätt oavsett mode – något vi kan kalla mode-bias.

c) Med *registerdata* menas de uppgifter man kan få från ett befintligt dataregister/databas som är i bruk för något syfte. Som ett alternativ eller komplement till primärdatainsamling:

Fördelar: billigare än primärdatainsamling - data finns ju redan; möjlighet att följa specifika objekt över tid; minskad uppgiftslämnarbörda då färre frågor behövs – data finns ju redan.

Nackdelar: risk för inaktuella uppgifter - sena inrapporteringar; täckningsproblem kan uppstå både över- och undertäckning; registrets ursprungliga syfte ej detsamma som undersökningen med oönskade variabeldefinitioner och andra avgränsningar/definitioner av populationen och objekten.

Hjälpmedel: kan användas som urvalsram och som källa för hjälpvariabler, både i urvalsfasen (t.ex. stratifiering) och skattningsfasen (t.ex. kvot- och regressionsskattningar).

d) Upplägget/designen/frågeformuleringen (exempel):

Målpopulationen är inte väldefinierad. Är det de som brukar lyssna (vilka är de i så fall?) eller de som råkade lyssna just den här gången (och vilka är det i så fall?).

Det är ett inte ett sannolikhetsurval utan ett självrekryterande urval och vi har ingen aning om vad inklusionssannolikheterna är. Är det då ett representativt urval? För vilken population i så fall?

Det faktum att man ska skicka in ett mejl med sin åsikt kan verka avskräckande hur sekretess- och integritetssynpunkt , att många avstår då de tycker att det är för känsligt.

Frågan är väldigt hypotetisk och kan vara svår att ta ställning till, ”om du kunde, skulle du …” osv. ”Om du kunde” kanske implicerar att man inte kan få barn, att man inte har råd mm. Är det typiska lyssnare? Dessutom har frågan en inbyggd filterfråga – ”skulle du välja att skaffa barn *igen*” antyder att frågan är riktad till de som redan har barn. Tydlig risk för olika tolkningar av vad som efterfrågas och det är oklart vad det är för fråga enskilda respondenter faktiskt svarar på.

Endast två svarsalternativ är olyckligt, man bör alltid överväga neutrala svarsalternativ som t.ex. ”vet ej”, ”ingen åsikt, ”kanske” osv. När man endast kan svara Ja eller Nej så kan många strunta i att delta eftersom de tycker att det blir för onyanserat och att alternativen inte passar just dem.

Resultatet (exempel):

Fler kvinnor än män svarade – vad kan det bero på? Är det fler kvinnliga lyssnare än män? Eller är kvinnor mer motiverade än män att svara på just den här frågan? Över- och underrepresentation kan man alltid kompensera för (t.ex. med poststratifiering) men populationen är som sagt odefinierad så det går inte.

Det vore intressant att kunna jämföra kvinnor och män vilket inte går i detta fall; andelen kvinnliga nej-svar ligger mellan 62.5% och 87.5% men vi vet inte eftersom det inte redovisas och då blir poststratifiering ännu mer omöjligt!

Var kritisk! Hur kan man veta att det var 80% kvinnor som svarade? Frågade om detta eller gick man på namnet? Hur gjorde man i så fall med könsneutrala e-postadresser? Ifrågasätt allt!

En stor andel svarade nej vilket kan vara en indikation på att det är problem med det självrekryterande urvalet, att de som känner starkt för att svara nej i större utsträckning har brytt sig om att skicka in ett mejl.

Hur definieras bortfall i denna undersökning? Alla som lyssnade minus alla som svarade? Som sagt vi har en odefinierad population och ett oklart urvalsförfarande så bortfallsanalys är bara att glömma.

Slutsats: Undersökningen har uppenbara vetenskapliga brister och resultaten går inte att generalisera till en större population. Jag skulle gömma det hela långt ner i en källare.