Surveymetodik
Föreläsning 9

Måns Magnusson

Avd. Statistik, LiU

Översikt

Section 1

Regressionsestimation

Introduktion till regressionsestimation

- Kvotestimation använder modellen

$$y_i = \hat{B}x_i$$

- Om vi inte kan anta att $y = 0$ då $x = 0$ så kan inte kvotestimation användas

Introduktion till regressionsestimation

- Kvotestimation använder modellen

$$y_i = \hat{B} x_i$$

- Om vi inte kan anta att $y = 0$ då $x = 0$ så kan inte kvotestimation användas

- Kan istället använda **regressionsestimation** med modellen

$$y = \mathbf{x}^T \mathbf{B} = B_0 + B_1 x_1 + ... + B_p x_p$$

- **Precis som tidigare, två situationer**: Antingen känner vi till populationstotalerna för $\mathbf{x}$ ($\mathbf{t_x}$), eller inte.

Introduktion till regressionsestimation II

- Exempel på användning om vi **inte känner** till $\mathbf{t_x}$
  - Vi kan vara intresserad av **populationsregressionskoefficienterna**

$$\mathbf{B} = B_0, B_1, ..., B_p$$

Introduktion till regressionsestimation II

- Exempel på användning om vi **inte känner** till $\mathbf{t_x}$
  - Vi kan vara intresserad av **populationsregressionskoefficienterna**

$$\mathbf{B} = B_0, B_1, ..., B_p$$

- Exempel på användning om vi **känner** till $\mathbf{t_x}$
  - Vi kan använda $\mathbf{t_x}$ för att **förbättra precisionen** i $\hat{\bar{y}}_{\mathcal{U}}$ eller $\hat{t}_y$

Introduktion till regressionsestimation II

- Exempel på användning om vi **inte känner** till $\mathbf{t_x}$
  - Vi kan vara intresserad av **populationsregressionskoefficienterna**

$$\mathbf{B} = B_0, B_1, ..., B_p$$

- Exempel på användning om vi **känner** till $\mathbf{t_x}$
  - Vi kan använda $\mathbf{t_x}$ för att **förbättra precisionen** i $\hat{\bar{y}}_{\mathcal{U}}$ eller $\hat{t}_y$
  - Vi kan använda $\mathbf{t_x}$ för att **kalibrera** $\hat{\bar{y}}_{\mathcal{U}}$ eller $\hat{t}_y$ till kända $\mathbf{t_x}$.
  - Detta är den vanligaste metoden för att hantera bortfallsfel och ramfel.

Regressionsestimation: Estimatorer

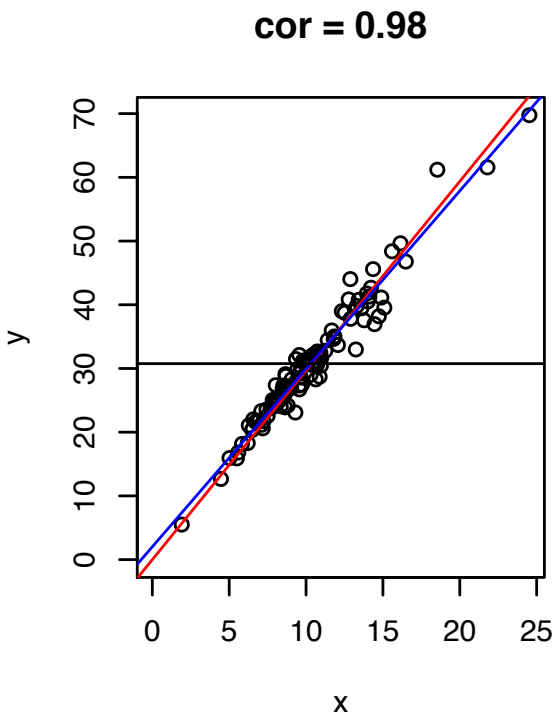
- Skattningen av regressionskoefficienterna **vid OSU** görs på exakt samma sätt som vid vanlig regression:

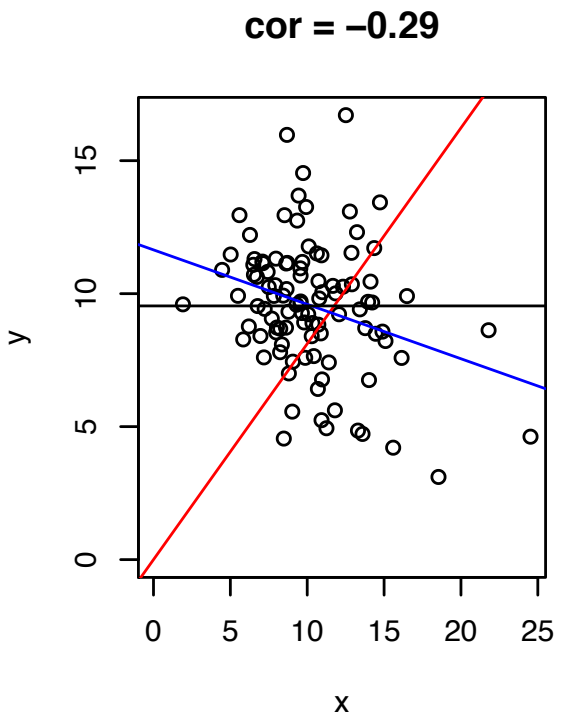
$$\hat{B}_1 = \frac{\sum_{i \in \mathcal{S}} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in \mathcal{S}} (x_i - \bar{x})^2} \text{ och } B_0 = \bar{y}_{\mathcal{S}} - \hat{B}_1 \bar{x}_{\mathcal{S}}$$

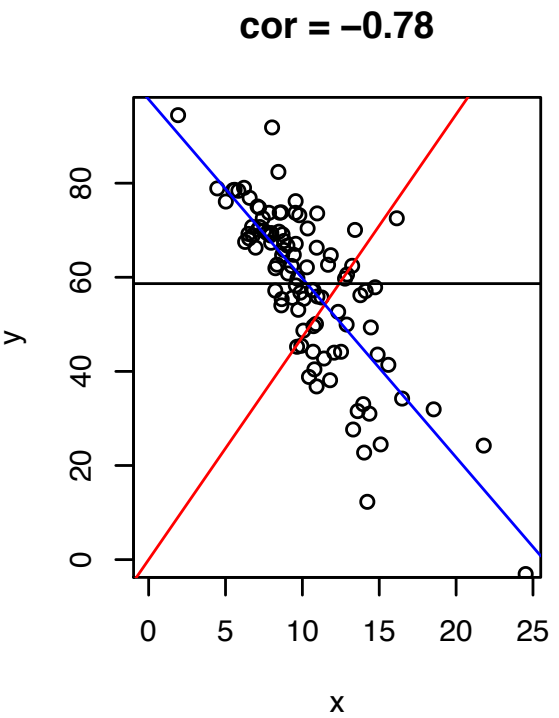
eller $(*)$

$$\hat{\mathbf{B}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Exempel: Exempel på olika modeller







Regressionsestimation: Estimatorer II

- Om vi känner till $\bar{\mathbf{x}}$ kan vi använda $\mathbf{x}$ som **hjälpvariabel(er)** för att skatta $\bar{y}_{\mathcal{U}}$ med **bättre precision** på följande sätt $(*)$:

$$\hat{\bar{y}}_{reg} = \hat{B}_0 + \hat{B}_1 \bar{x}_{1,\mathcal{U}} = \text{regressionsskattning av } \bar{y}_{\mathcal{U}}$$

eller

$$\hat{\bar{y}}_{reg} = \bar{y}_{\mathcal{S}} + (\bar{x}_U - \bar{x}_{\mathcal{S}})\hat{B}_1$$
$$\hat{t}_{yreg} = \hat{t}_y + (t_x - \hat{t}_x)\hat{B}_1$$

eller mer generellt

$$\hat{\bar{y}}_{reg} = \bar{y}_{\mathcal{S}} + (\bar{\mathbf{x}}_{\mathcal{S}} - \bar{\mathbf{x}}_U)^T \hat{\mathbf{B}}$$
$$\hat{t}_{yreg} = \hat{t}_y + (\mathbf{t_x} - \hat{\mathbf{t}}_\mathbf{x})^T \hat{\mathbf{B}}$$

- Precis som kvotskattningen är regressionsskattningar **inte** väntevärdesriktiga. Varför?

Regressionsestimation: Estimatorer II

- Om vi känner till $\bar{\mathbf{x}}$ kan vi använda $\mathbf{x}$ som **hjälpvariabel(er)** för att skatta $\bar{y}_{\mathcal{U}}$ med **bättre precision** på följande sätt $(*)$:

$$\hat{\bar{y}}_{reg} = \hat{B}_0 + \hat{B}_1 \bar{x}_{1,\mathcal{U}} = \text{regressionsskattning av } \bar{y}_{\mathcal{U}}$$

eller

$$\hat{\bar{y}}_{reg} = \bar{y}_{\mathcal{S}} + (\bar{x}_U - \bar{x}_{\mathcal{S}})\hat{B}_1$$
$$\hat{t}_{yreg} = \hat{t}_y + (t_x - \hat{t}_x)\hat{B}_1$$

eller mer generellt

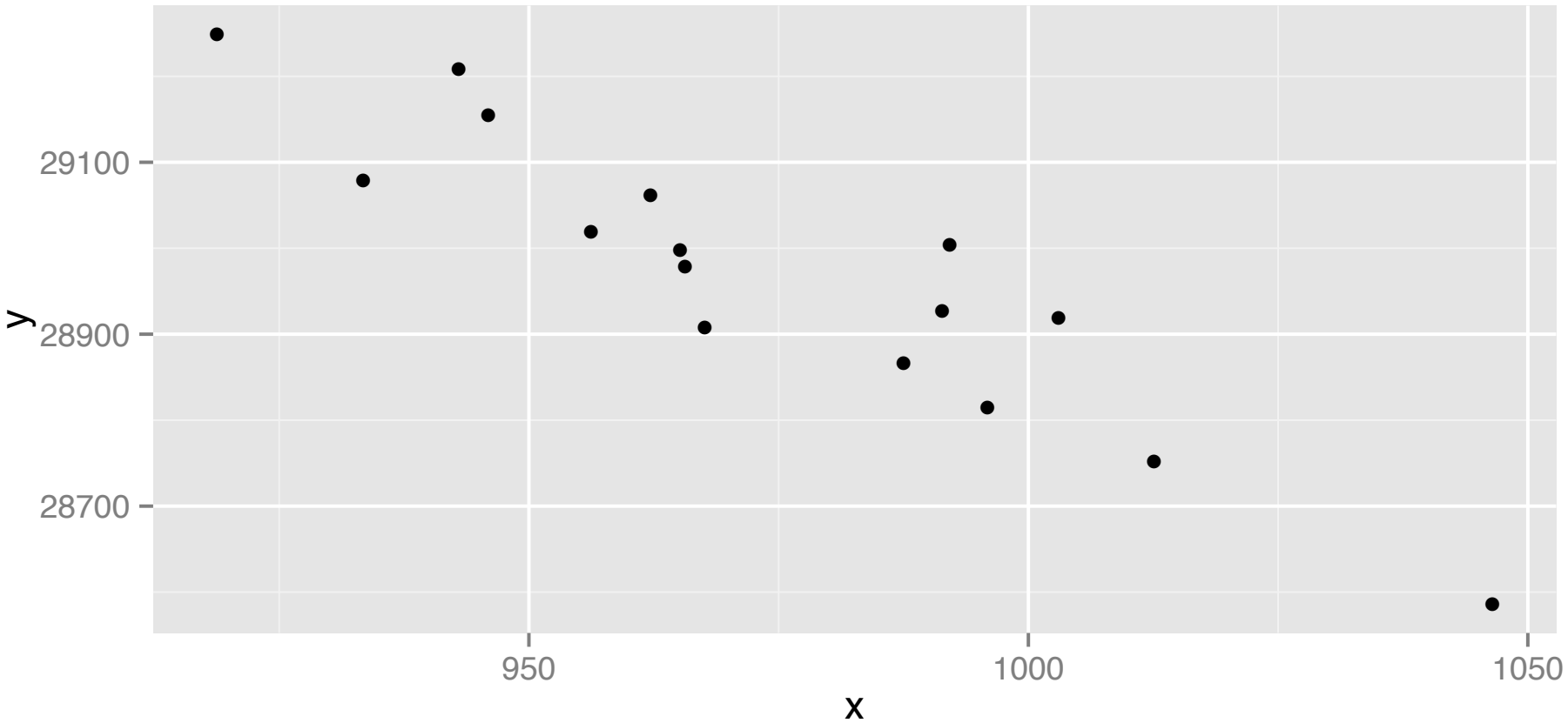
$$\hat{\bar{y}}_{reg} = \bar{y}_{\mathcal{S}} + (\bar{\mathbf{x}}_{\mathcal{S}} - \bar{\mathbf{x}}_U)^T \hat{\mathbf{B}}$$
$$\hat{t}_{yreg} = \hat{t}_y + (\mathbf{t_x} - \hat{\mathbf{t}}_\mathbf{x})^T \hat{\mathbf{B}}$$

- Precis som kvotskattningen är regressionsskattningar **inte** väntevärdesriktiga. Varför?

- Men vad händer om vi använder en modell som är fel?

Exempel: Genomsnittsinkomster

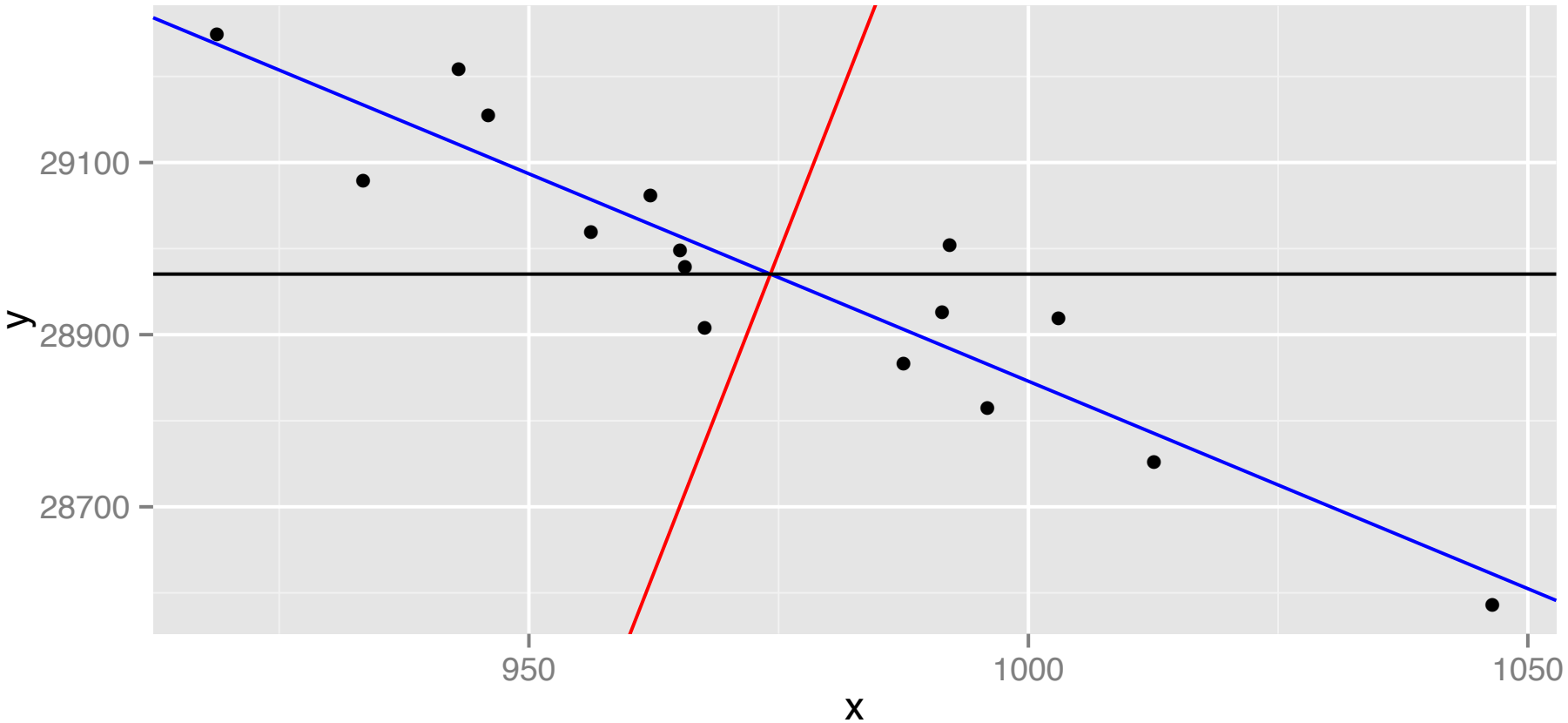
Vi vill uppskatta genomsnittsinkomsten i 16 bostadsområden (y). Vi vet det genomsnittliga utgifter för försörjningsstöd (x) för alla områden och drar ett urval på $n = 5$.



I populationen (alla områden) är $R = -0.927$ och $B_1 = -4.821$.

Exempel: Genomsnittsinkomster II

Skillnaden mellan modellen för kvotestimatorn (röd), regressionsestimatorn (blå) och den 'vanliga' skattningen $\bar{y}_S$ (svart).



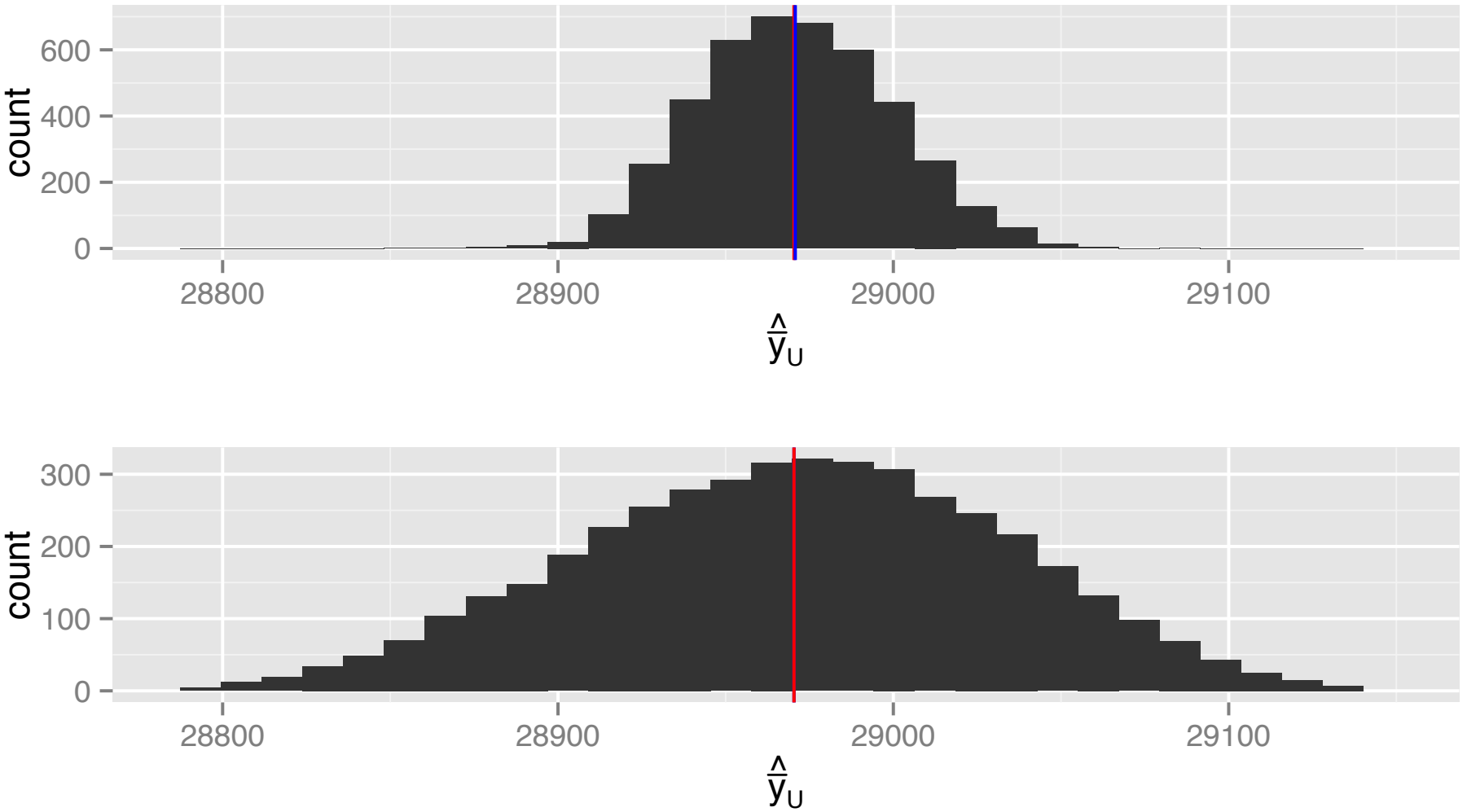
Exempel: Genomsnittsinkomster III

Teoretiska fördelningen med K =4368 stycken teoretiska urval.

	obs.1	obs.5	P_S	mean_hat_x	mean_hat_y	B0	B1	mean_hat_yreg
1289	28867	29020	0.000229	950	29100	34787	-5.99	28957
2905	29079	29004	0.000229	971	28961	31733	-2.86	28951
3918	28586	29020	0.000229	981	28925	34412	-5.59	28965
1631	28979	29004	0.000229	963	29016	30304	-1.34	29000
1615	28979	29004	0.000229	961	29025	31442	-2.51	28992
381	28867	29249	0.000229	978	28940	33425	-4.59	28957
919	28867	29249	0.000229	967	29013	34489	-5.66	28974
4088	28586	29004	0.000229	996	28880	34375	-5.52	28999
1456	28979	29155	0.000229	956	29070	33031	-4.14	28994
3165	28919	29249	0.000229	965	29042	33349	-4.46	29000
1262	28867	28752	0.000229	974	28977	35302	-6.49	28977
4147	28815	29155	0.000229	973	28985	35281	-6.47	28975
2982	29079	28752	0.000229	955	29057	33957	-5.13	28957
3538	28919	29020	0.000229	977	28951	33298	-4.45	28964
96	28867	28752	0.000229	1003	28821	33677	-4.84	28960

Exempel: Genomsnittsinkomst IV

Samlingfördelningen för $\hat{\bar{y}}_{reg}$ och $\hat{\bar{y}}_{\mathcal{U}}$ då $\bar{y}_{\mathcal{U}}$ =28970.373.



Effekten av kvotestimation

- Skillnaden mellan regressionsestimatorn och den "vanliga" estimatorn

$$
\begin{array}{rclrcl}
E(\hat{\bar{y}}_{reg}) & = & 28970.754 & E(\hat{\bar{y}}_{\mathcal{U}}) & = & 28970.373 \\
Var(\hat{\bar{y}}_{reg}) & = & 807.838 & Var(\hat{\bar{y}}_{\mathcal{U}}) & = & 3957.185 \\
Bias(\hat{\bar{y}}_{reg}) & = & 0.382 & Bias(\hat{\bar{y}}_{\mathcal{U}}) & = & 0 \\
MSE(\hat{\bar{y}}_{reg}) & = & 807.984 & MSE(\hat{\bar{y}}_{\mathcal{U}}) & = & 3957.185
\end{array}
$$

Regressionsestimation: Variansberäkningar

- Vi är (som vanligt) intresserade av $Var(\hat{\bar{y}}_{reg})$ för att kunna skapa ett konfidensintervall för $\hat{\bar{y}}_{\mathcal{U}}$
- Detta görs (precis som för kvotestimatorn) genom att beräkna residualerna

$$e_i = y_i - \hat{B}_0 - \hat{B}_1 x_i$$

eller mer generellt

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\mathbf{B}}$$

Regressionsestimation: Variansberäkningar

- Vi är (som vanligt) intresserade av $Var(\hat{\bar{y}}_{reg})$ för att kunna skapa ett konfidensintervall för $\hat{\bar{y}}_{\mathcal{U}}$
- Detta görs (precis som för kvotestimatorn) genom att beräkna residualerna

$$e_i = y_i - \hat{B}_0 - \hat{B}_1 x_i$$

eller mer generellt

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\mathbf{B}}$$

- Det är sedan residualerna som används för att beräkna $Var(\hat{\bar{y}}_{reg})$ på följande sätt

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

där

$$s_e^2 = \frac{\sum_{i \in \mathcal{S}} e_i^2}{n-1} \text{ eller } s_e^2 = \frac{\sum_{i \in \mathcal{S}} e_i^2}{n-p}$$

där p är antalet parametrar i modellen.

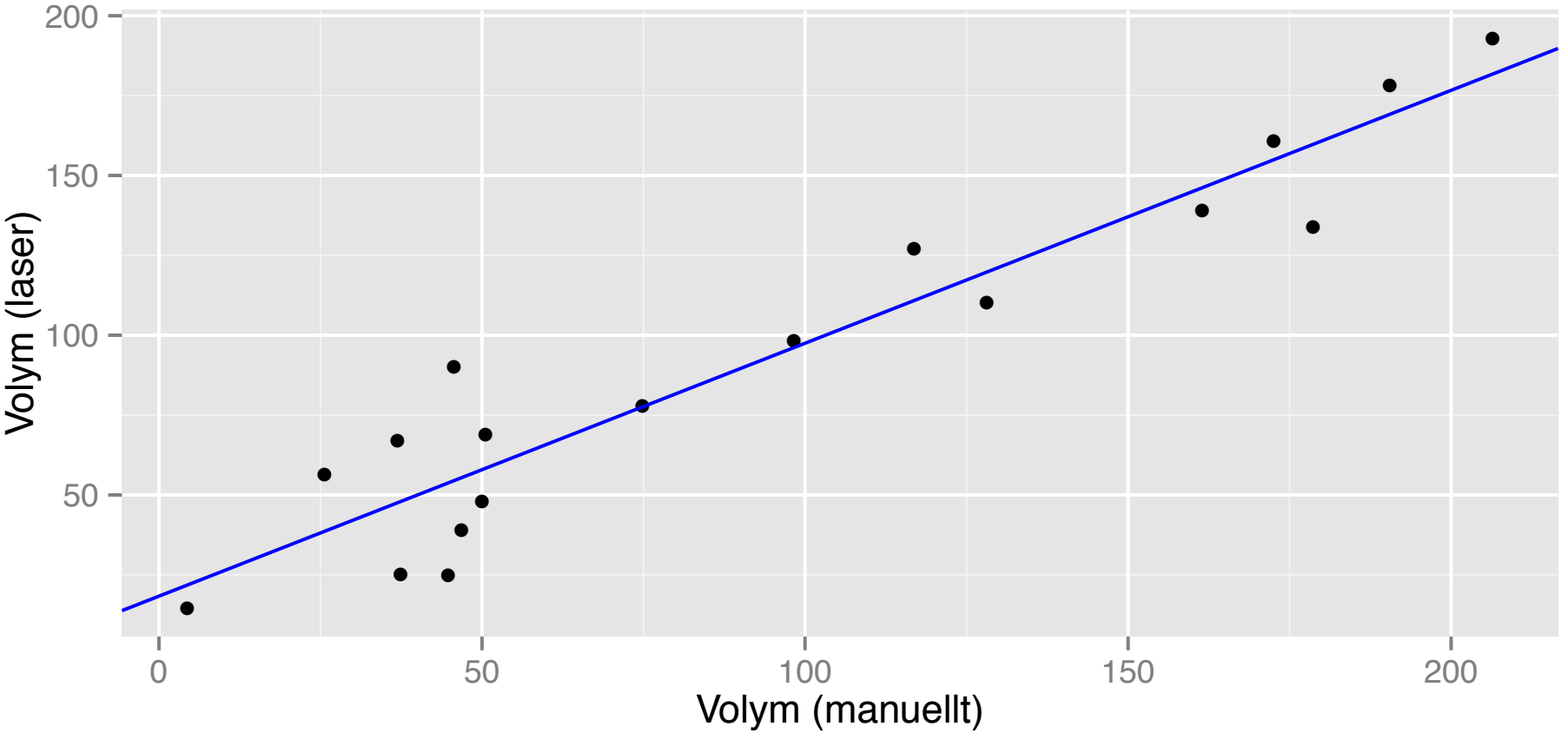
Exempel: Skogsvolym

- Inom skogsindustrin är det av intresse att uppskatta trävolym per hektar
- Skogsföretag A vill uppskatta den totala volymen träd för ett område på 17010 ha.
- Som hjälpinformation finns laserscanning av hela området.
- Skogsföretag samlar slumpmässigt in volymen från 18 slumpmässiga ytor (1 ha).
- Laserscanning har gjorts för hela området med en uppskattning av volymen till 1807641 m^3 träd

Exempel: Skogsvolym II

```
(Intercept)           x
     18.317       0.792
       y      x y_hat       e
1   98.1  98.28  96.1   2.015
2  192.8 206.44 181.7  11.048
3   56.5  25.64  38.6  17.859
4   25.1  37.39  47.9 -22.829
5   66.9  36.91  47.5  19.322
6   68.9  50.52  58.3  10.627
7   90.1  45.65  54.5  35.671
8   14.6   4.38  21.8  -7.189
9  127.1 116.88 110.8  16.235
10 160.7 172.54 154.9   5.747
11  77.8  74.81  77.5   0.212
12 178.2 190.49 169.1   9.079
13 110.3 128.12 119.7  -9.477
14 133.8 178.61 159.7 -25.927
15  47.9  50.01  57.9  -9.994
16  39.1  46.80  55.4 -16.277
17 138.9 161.47 146.1  -7.209
18  24.8  44.75  53.7 -28.913
```

Exempel: Skogsvolym III



Subsection 1

Regressionsestimation som kalibrering

Regressionsestimation som kalibrering

- Regressionsestimatorn kan användas för kalibrering (precis som kvotestimatorn).
- Denna estimator kallas ibland **generalized regression** estimatior (GREG) och utrycks som
(se Lohr (2009, s. 458 f.))

$$\hat{t}_{yGREG} = \hat{t}_y + (\mathbf{t_x} - \hat{\mathbf{t}}_\mathbf{x})\hat{\mathbf{B}}$$

Regressionsestimation som kalibrering

- Regressionsestimatorn kan användas för kalibrering (precis som kvotestimatorn).
- Denna estimator kallas ibland **generalized regression** estimatior (GREG) och utrycks som
(se Lohr (2009, s. 458 f.))

$$\hat{t}_{yGREG} = \hat{t}_y + (\mathbf{t_x} - \hat{\mathbf{t}}_\mathbf{x})\hat{\mathbf{B}}$$

$$\hat{t}_{yGREG} = \sum_{i \in \mathcal{S}} w_i g_i y_i$$

där

$$g_i = 1 + (\mathbf{t_x} - \hat{\mathbf{t}}_\mathbf{x})^T \left(\sum_{j \in \mathcal{S}} w_j \frac{1}{\sigma_j^2} \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \frac{1}{\sigma_i^2} \mathbf{x}_i$$

Regressionsestimation som kalibrering

- Regressionsestimatorn kan användas för kalibrering (precis som kvotestimatorn).
- Denna estimator kallas ibland **generalized regression** estimatior (GREG) och utrycks som
(se Lohr (2009, s. 458 f.))

$$\hat{t}_{yGREG} = \hat{t}_y + (\mathbf{t_x} - \hat{\mathbf{t}}_\mathbf{x})\hat{\mathbf{B}}$$

$$\hat{t}_{yGREG} = \sum_{i \in \mathcal{S}} w_i g_i y_i$$

där

$$g_i = 1 + (\mathbf{t_x} - \hat{\mathbf{t}}_\mathbf{x})^T \left(\sum_{j \in \mathcal{S}} w_j \frac{1}{\sigma_j^2} \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \frac{1}{\sigma_i^2} \mathbf{x}_i$$

- På detta sätt kalibreras skattningarna till de kända totalerna $\mathbf{t_x}$

$$\hat{t}_{xGREG} = \sum_{i \in \mathcal{S}} w_i g_i x_i = t_x$$

Referenser

Lohr, S., 2009. Sampling: design and analysis, 2nd Edition. Thomson.