

Statistik och Dataanalys I

Föreläsning 20 - Inferens i linjär regression - konfidensintervall, hypotestest och prediktionsintervall

Mattias Villani



Statistiska institutionen
Stockholms universitet



mattiasvillani.com



@matvil



@matvil



mattiasvillani

- **Konfidensintervall** och **hypotesttest** i enkel regression
- **Prediktionsintervall**
- **Inferens i multipel linjär regression**

Standardfel för b_1

- Estimatoren för lutningskoefficienten

$$b_1 = \frac{s_{xy}}{s_x^2}$$

- Hur b_1 varierar mellan olika stickprov:

$$\sigma_{b_1} = SD(b_1) = \frac{\sigma_\varepsilon}{s_x \sqrt{n-1}}$$

- σ_{b_1} skattas med **standardfelet**

$$s_{b_1} = SE(b_1) = \frac{s_e}{s_x \sqrt{n-1}}$$

- Formel för $SE(b_0)$ slipper ni på SDA1. 😊
- lifespan data [`sd(spending) = 1.097516`]

$$s_{b_1} = \frac{1.678}{\sqrt{29-1} \cdot 1.097516} \approx 0.289$$

Standardfel för b_1 i R

```
> library(sda123)
> lifespan_no_usa = lifespan[1:29,] # remove the outlier USA
> model = lm(lifespan ~ spending, data = lifespan_no_usa)
> summary(model)
```

Call:

```
lm(formula = lifespan ~ spending, data = lifespan_no_usa)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-3.3108	-0.7016	-0.0507	1.1458	3.8860

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	74.1639	0.8782	84.45	< 2e-16 ***
spending	1.7629	0.2890	6.10	1.63e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.678 on 27 degrees of freedom

Multiple R-squared: 0.5795, Adjusted R-squared: 0.5639

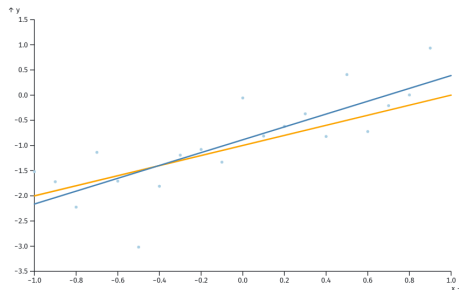
F-statistic: 37.21 on 1 and 27 DF, p-value: 1.626e-06

Samplingfördelning i regression - interaktivt



Skattat intercept: $b_0 = -0.8866$

Skattad lutning: $b_1 = 1.275$



Populationsmodell
Skattad regressionslinje

Konfidensintervall för b_1

- Estimatoren b_1 följer en **t-fördelning** med $n - 2$ **frihetsgrader**:

$$\frac{b_1 - \beta_1}{s_{b_1}} \sim t_{n-2}$$

- Varför $n - 2$? Skattar två parametrar, β_0 och β_1 . Förlorar två frihetsgrader.
- 95%-igt konfidensintervall för β_1**

$$b_1 \pm t_{0.025, n-2} \cdot s_{b_1}$$

- lifespan data: $n = 29$, och $t_{0.025, 27} = 2.052$ från tabell.
- 95%-igt konfidensintervall för β_1

$$1.763 \pm 2.052 \cdot 0.289 = (1.170, 2.356)$$

Konfidensintervall i R

■ R:

```
> model = lm(lifespan ~ spending, data = lifespan_no_usa) # utan USA  
> confint(model)
```

■ sda123-paketet:

```
> model = lm(lifespan ~ spending, data = lifespan_no_usa) # utan USA  
> reg_summary(model, conf_intervals = TRUE, anova = FALSE)
```

Measures of model fit

```
-----  
Root MSE      R2    R2-adj  
1.67836  0.57952  0.56394
```

Parameter estimates

```
-----  
                Estimate Std. Error t value  Pr(>|t|)  2.5 %  97.5 %  
(Intercept)  74.1639    0.87822  84.4482 2.9262e-34 72.362 75.9658  
spending      1.7629    0.28900   6.1002 1.6256e-06 1.170 2.3559
```

Hypotesttest för β

- Hypotesttest för lutningen i regressionen

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

- Teststatistiska

$$T = \frac{b_1 - 0}{s_{b_1}}$$

- Under H_0 har vi att $T \sim t_{n-2}$.
- Vi förkastar nollhypotesen på signifikansnivån $\alpha = 0.05$ om

$$|t_{obs}| > t_{crit}$$

där det kritiska värdet t_{crit} hämtas från tabell:

$$t_{crit} = t_{0.025, n-2}$$

- **P-värde** räknas som tidigare, men från t_{n-2} fördelning.

Hypotesttest för β - lifespan data

- $n = 29$, så $n - 2 = 27$, och $t_{\text{crit}} = t_{0.025}(27) = 2.052$.

$$t_{\text{obs}} = \frac{1.763 - 0}{0.289} = 6.100$$

- $|t_{\text{obs}}| > t_{\text{crit}}$ så vi **förkastar nollhypotesen** på 5% signifikansnivå.
- Vi förkastar nollhypotesen att spending inte är korrelerad med lifespan.
- spending är en **signifikant förklarande variabel** för livslängd på signifikansnivå 5%.
- Testets p -värde visar att vi tokförkastar H_0 !

$$p = 1.6256e - 06 = 0.0000016256$$

- 1.6256e-06. Flytta punkten/kommat sex steg till vänster.

Hypotestest i R

```
> library(sda123)
> lifespan_no_usa = lifespan[1:29,] # remove the outlier USA
> model = lm(lifespan ~ spending, data = lifespan_no_usa)
> summary(model)
```

Call:

```
lm(formula = lifespan ~ spending, data = lifespan_no_usa)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.3108	-0.7016	-0.0507	1.1458	3.8860

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	74.1639	0.8782	84.45	< 2e-16 ***
spending	1.7629	0.2890	6.10	1.63e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.678 on 27 degrees of freedom

Multiple R-squared: 0.5795, Adjusted R-squared: 0.5639

F-statistic: 37.21 on 1 and 27 DF, p-value: 1.626e-06

Prediktionsintervall

- Antag att vi gör en prognos vid ett nytt $x = x_*$

$$\hat{y}_* = b_0 + b_1 x_*$$

- **Prediktionsintervall** för \hat{y}_* - **två källor av osäkerhet**:

- ▶ De **okända parametrarna** β_0 och β_1 , dvs osäkerhet om regressionslinjen vid x_* .
- ▶ **Variationen i de enskilda y -värdena kring regressionlinjen.**
Alla observationer "träffas av ett ε " med standardavvikelse σ_ε .

- **Prediktionsvariansen**:

$$\sigma_{\text{prediktion}}^2 = \sigma_{\text{regressionslinjen vid } x_*}^2 + \sigma_\varepsilon^2$$

- **95%-igt prediktionsintervall** för enskild observation vid x_*

$$\hat{y}_* \pm t_{0.025, n-2} \cdot \sqrt{\frac{s_e^2}{n} + s_{b_1}^2 (x_* - \bar{x})^2 + s_e^2}$$

Prediktionsintervall

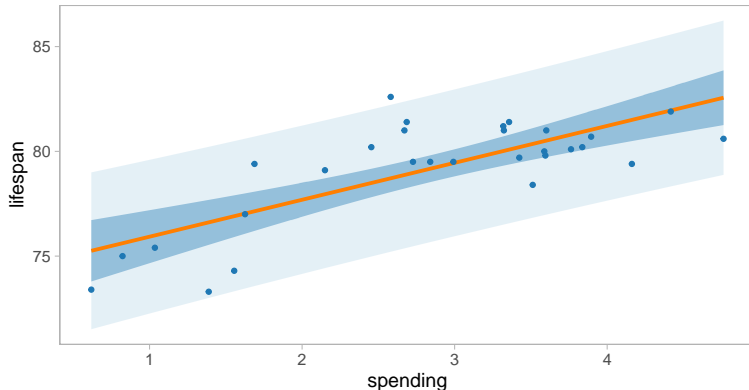
```
> model = lm(lifespan ~ spending, data = lifespan_no_usa)
> predict(model, newdata = data.frame(spending = 3.323))
      1
80.02209
> predict(model, newdata = data.frame(spending = 4.323))
      1
81.78502
> predict(model, newdata = data.frame(spending = 4.323), interval = "prediction")
      fit      lwr      upr
1 81.78502 78.17388 85.39616
```

Plot av prediktionsintervall



```
> library(sda123)
> reg_predict(lifespan ~ spending, data = lifespan_no_usa)
```

Konfidens- och prediktionsintervall



■ Ljusblå band är prediktionsintervall (för ett x i taget).

Multipel regression - modell och samplingfördelning

- **Populationsmodell** för **multipel regression** med k förklarande variabler

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon)$$

- Varje β_j skattas med b_j med minsta-kvadrat-metoden.
- Estimatorn b_j följer en **t -fördelning** med $n - k - 1$ **frihetsgrader**:

$$\frac{b_j - \beta_j}{s_{b_j}} \sim t_{n-k-1}$$

- Varför $n - k - 1$? Skattar k lutningskoefficienter $(\beta_1, \beta_2, \dots, \beta_k)$ och ett intercept (β_0) .
- Formlerna för minsta-kvadratskattningar b_j och standardfelen s_{b_j} är komplicerade. Datorn får göra jobbet. 😊

Multipel regression - konfidensintervall och test

■ Populationsmodell **multipel regression**

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon)$$

■ **95%-igt konfidensintervall för β_1**

$$b_j \pm t_{0.025, n-k-1} \cdot s_{b_j}$$

■ **Hypotestest för lutningen** i regressionen

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

■ **Teststatistiska**

$$T = \frac{b_j - 0}{s_{b_j}}$$

■ **Under H_0** har vi att $T \sim t_{n-k-1}$.

■ Om vi **förkastar H_0** så drar vi slutsatsen att $\beta_j \neq 0$ och säger att x_j **är en signifikant förklarande variabel**.

Multipel regression i R

```
> model = lm(lifespan ~ spending + gdp + doctorvisits, data = lifespan_no_usa)
> summary(model)
```

Call:

```
lm(formula = lifespan ~ spending + gdp + doctorvisits, data = lifespan_no_usa)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.4860	-0.8975	-0.0762	1.1654	3.7609

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	74.07091	1.34241	55.178	< 2e-16 ***
spending	2.10379	0.55123	3.817	0.000792 ***
gdp	-0.02993	0.04230	-0.708	0.485723
doctorvisits	0.02842	0.10867	0.262	0.795813

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.726 on 25 degrees of freedom

Multiple R-squared: 0.5884, Adjusted R-squared: 0.5391

F-statistic: 11.92 on 3 and 25 DF, p-value: 4.894e-05

Simulera data med sda123 paketet



```
> library(sda123)
> simdata <- reg_simulate(n = 500, betavect = c(1, -2, 1, 0), sigma_eps = 2)
> head(simdata)
```

	y	X1	X2	X3
1	1.9710435	-0.02922743	-0.2445304	1.00699482
2	-0.1156157	0.32641792	0.1161198	1.34818268
3	4.2054858	0.40102782	1.5955417	-1.85657317
4	7.3999811	-1.34540553	1.3106342	1.22095959
5	0.4633449	-1.31315970	-0.5062060	-0.08381122
6	2.1395357	-0.30667637	-0.7820189	1.51466922

```
> fit = lm(y ~ X1 + X2 + X3, data = simdata)
> summary(fit)
```

Call:

```
lm(formula = y ~ X1 + X2 + X3, data = simdata)
```

Residuals:

Min	1Q	Median	3Q	Max
-5.5043	-1.2584	-0.0609	1.2501	6.8445

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.21579	0.08731	13.925	<2e-16	***
X1	-1.98033	0.08854	-22.367	<2e-16	***
X2	0.88075	0.08493	10.370	<2e-16	***
X3	-0.02364	0.08639	-0.274	0.784	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.947 on 496 degrees of freedom

Multiple R-squared: 0.5343, Adjusted R-squared: 0.5315

F-statistic: 189.7 on 3 and 496 DF, p-value: < 2.2e-16