Övning 3

Lite repetition och hopklipp från föreläsning 8-10 gällande korrelation, regression etc., se föreläsningsfilerna för detaljer.

Korrelation

n – antalet observationer

i – indexerar observationerna

Antag att du har två variabler, x och y

Standardavvikelse(x) = $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Standardavvikelse(y) = $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$

r = Korrelationskoefficient för sambandet mellan x och y = $\frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{s_x}\frac{(y_i-\bar{y})}{s_y}$

(från F8)

Enkel linjär regression - populationen

- Vi vill förstå hur ett samband ser ut i populationen
- Sambandet i populationen modelleras som
  $Y = \beta_0 + \beta_1 X + \varepsilon$

Enkel linjär regression - vårt urval/stickprov

- Vi **skattar** förhållandet i populationen genom att, **för vårt urval**, ta fram b_0 och b_1 i följande relation, med minsta kvadratmetoden,
  $Y = b_0 + b_1 X + e$

Enkel linjär regression – R2

- **R2** (eller R^2**, R-kvadrat)** (R-squared, coefficient of determination), används som ett mått på hur mycket en regressionsmodell förklarar av variationen i responsvariabeln.
- För enkel linjär regression gäller att $R^2 = r^2$, där r är korrelationskoefficienten mellan X och Y. (F8, s. 12).

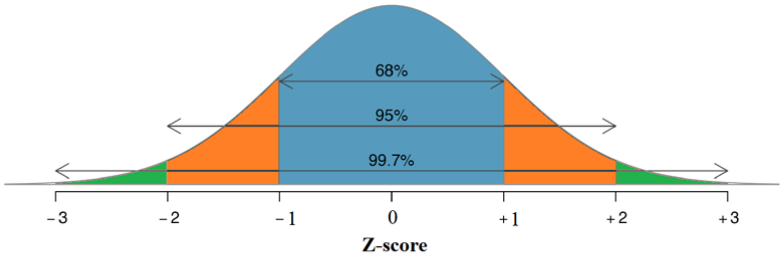
Inferens - hypotestest för lutningskoefficient

- Vi är typiskt intresserade av om det finns ett samband eller inte, mellan två variabler (dvs. lutningskoefficienten):
  - $H_0 : \beta_1 = 0$ (Det finns inget samband)
  - $H_A : \beta_1 \neq 0$ (Det finns ett samband)

Ta fram T-värde

- Vår skattning (estimate) av lutningen (b_1) och skattningens standardfel (Std. Error, SE) fås båda från regressionstabellen
- Nollhypotesen ("det skeptiska perspektivet") är (typiskt) **inget samband mellan variablerna (X och Y i populationen)**, dvs. "null value" $= 0$
- Testvariabeln blir:

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

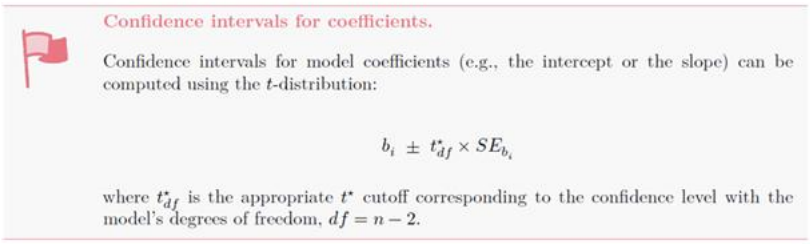


- Hypotestestet följer samma logik som vi sett tidigare
- Om vi hade haft ett t-värde mellan de kritiska värdena -2 och 2 hade vi inte kunnat förkasta nollhypotesen, med 95% säkerhet
- Men vi har $t_{obs} > 10$ - vi kan med hög säkerhet förkasta nollhypotesen
- p-värdet i output (som är lågt) - sannolikheten att observera den lutningskoefficient vi gör, om nollhypotesen hade varit sann

(Värdet för t_{obs} refererar till t-värdet i Uppsalaexemplet på F8, F9)

Konfidensintervall för regressionskoefficient

- Liknar vad vi gjorde på föreläsning 6, se också boken 24.5:



Confidence intervals for coefficients.

Confidence intervals for model coefficients (e.g., the intercept or the slope) can be computed using the t-distribution:

$$b_i \pm t^*_{df} \times SE_{b_i}$$

where t^*_{df} is the appropriate t^* cutoff corresponding to the confidence level with the model's degrees of freedom, $df = n - 2$.

- De värden vi behöver är punktskattningen (b_1) och dess standardfel, båda finns i regressionstabellen
- För ett 95% konfidensintervall sätter vi t-värdet=2 i formeln (egentligen 1.96, men 2 är OK)

Multipel linjär regression - populationen (3 variabler, osv..)

- Vi går från en till två till flera X-variabler
- Som tidigare: populationssamband, skattas med data från urval
- Vad förändras?
  - Tolkning av regressionskoefficienter (begreppet **"allt annat lika"**)
  - **Multikollinearitet**
  - **R2-måttet ökar för varje ny variabel** - åtgärd behövs
- Prediktion, Inferens (hypotestest och konfidensintervall gällande en lutningskoefficient) - liknar enkel linjär regression

Gällande inferens - håll reda på (från kap 25.1)

In Chapter 24, you learned that the hypothesis test for a linear model with **one predictor**[1] can be written as:

if only one predictor, $H_0 : \beta_1 = 0$.

That is, if the true population slope is zero, the p-value measures how likely it would be to select data which produced the observed slope (b_1) value.

With **multiple predictors**, the hypothesis is similar, however, it is now conditioned on each of the other variables remaining in the model.

if multiple predictors, $H_0 : \beta_i = 0$ given other variables in the model

Uppgift 1

Normalt tar vi inte fram korrelationskoefficienten manuellt, men det kan vara bra att gå igenom någon sådan beräkning.

Antag att du har följande data, du observerar x och y för tre (n=3) olika objekt (dina observationer) och får följande värden.

$\{x_1, x_2, x_3\} = \{2, 3, 4\}$

$\{y_1, y_2, y_3\} = \{2, 2, 5\}$

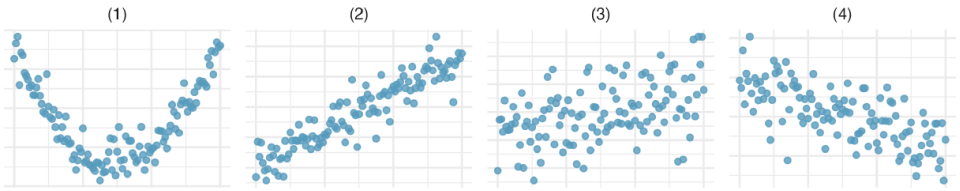
Båda sekvenserna har medelvärde 3, dvs $\bar{x} = 3, \bar{y} = 3$.

A) Rita ett spridningsdiagram med de tre datapunkterna
B) I övning 1 tog du fram standardavvikelsen för sekvensen {2, 3, 4}, vi fick s_x=1. Ta fram standardavvikelsen för sekvensen {2, 2, 5}, du ska få svaret s_y=$\sqrt{3}$
C) Ta fram korrelationskoefficienten mellan de två variablerna, dvs:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{s_x}\frac{(y_i-\bar{y})}{s_y}$$

Uppgift 2. Boken 7.7

7. **Match the correlation, I.** Match each correlation to the corresponding scatterplot.[11]



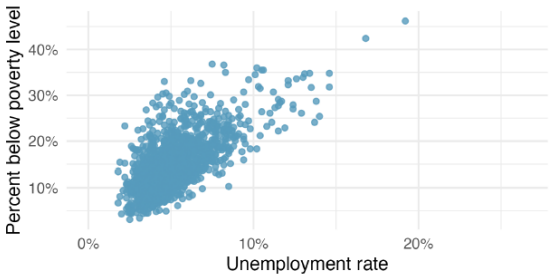
a. $r = -0.7$
b. $r = 0.45$
c. $r = 0.06$
d. $r = 0.92$

Uppgift 3. Boken 7.23

23. **Poverty and unemployment.** The following scatterplot shows the relationship between percent of population below the poverty level (`poverty`) from unemployment rate among those ages 20-64 (`unemployment_rate`) in counties in the US, as provided by data from the 2019 American Community Survey. The regression output for the model for predicting `poverty` from `unemployment_rate` is also provided.[20]

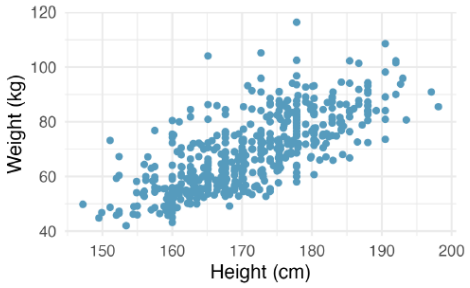
term	estimate	std.error	statistic	p.value
`(Intercept)`	4.60	0.349	13.2	<0.0001
`unemployment_rate`	2.05	0.062	33.1	<0.0001

a. Write out the linear model.
b. Interpret the intercept.
c. Interpret the slope.
d. The R^2 of this model is 46%. Interpret this value.
e. Calculate the correlation coefficient.



Uppgift 4. Boken 24.3

3. **Body measurements, mathematical test.** The scatterplot and least squares summary below show the relationship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals. (Heinz et al. 2003)



term	estimate	std.error	statistic	p.value
(Intercept)	-105.01	7.54	-13.9	<0.0001
hgt	1.02	0.04	23.1	<0.0001

a. Describe the relationship between height and weight.

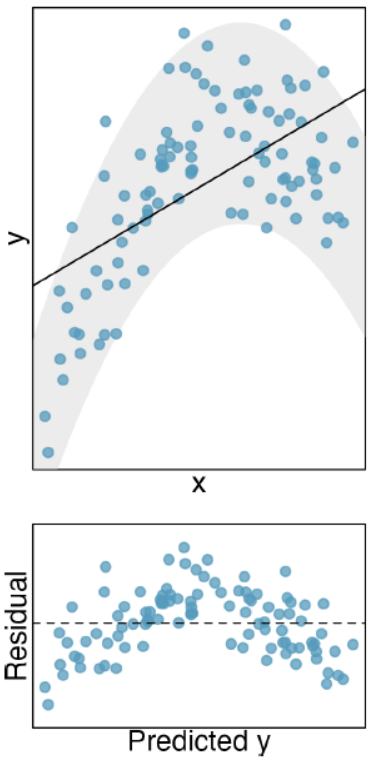
b. Write the equation of the regression line. Interpret the slope and intercept in context.

c. Do the data provide convincing evidence that the true slope parameter is different than 0? State the null and alternative hypotheses, report the p-value (using a mathematical model), and state your conclusion.

d. The correlation coefficient for height and weight is 0.72. Calculate R^2 and interpret it in context.

Uppgift 5. Boken, kap 24.6

Vi ser ett spridningsdiagram med två variabler, en rät linje anpassad med minsta kvadratmetoden och en residualplott. Kommentera gällande regressionsmodellens lämplighet.



Uppgift 6. Boken 25.7, fråga B

7. **Baby's weight, mathematical test.** US Department of Health and Human Services, Centers for Disease Control and Prevention collect information on births recorded in the country. The data used here are a random sample of 1,000 births from 2014. Here, we study the relationship between smoking and weight of the baby. The variable `smoke` is coded 1 if the mother is a smoker, and 0 if not. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in pounds, based on the smoking status of the mother.[6] (ICPSR 2014)

term	estimate	std.error	statistic	p.value
(Intercept)	-3.82	0.57	-6.73	<0.0001
weeks	0.26	0.01	18.93	<0.0001
mage	0.02	0.01	2.53	0.0115
sexmale	0.37	0.07	5.30	<0.0001
visits	0.02	0.01	2.09	0.0373
habitsmoker	-0.43	0.13	-3.41	7e-04

b. Using the regression output, evaluate whether the true slope of `habit` (i.e., whether the mother is a smoker) is different than 0, given the other variables in the model. State the null and alternative hypotheses, report the p-value (using a mathematical model), and state your conclusion.

Uppgift 7, Tolkning regressionstabell (röd ruta), Föreläsning 1, s. 12 (exempel på en tabell från en artikel, studien i sig ingår inte i kursen)