Uppgift 1 (lösning på uppgift C)

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 $\bar{x} = 3, \bar{y} = 3$.

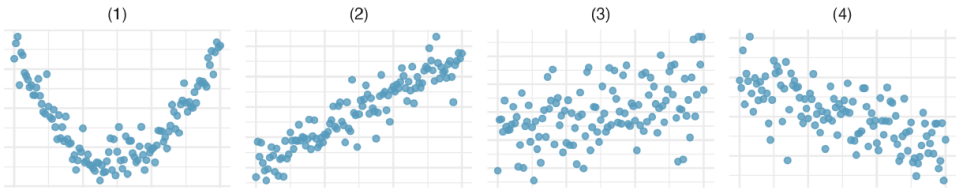
A) Rita ett spridningsdiagram med de tre datapunkterna
B) I övning 1 tog du fram standardavvikelsen för sekvensen {2, 3, 4}, vi fick s_x=1. Ta fram standardavvikelsen för sekvensen {2, 2, 5}, du ska få svaret s_y=$\sqrt{3}$
C) Ta fram korrelationskoefficienten mellan de två variablerna, dvs:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{s_x}\frac{(y_i-\bar{y})}{s_y} = \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{s_x}\frac{(y_i-\bar{y})}{s_y} = \frac{1}{2s_xs_y}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) =$$
$$\frac{1}{2\times1\times\sqrt{3}}((2-3)(2-3)+(3-3)(2-3)+(4-3)(5-3)) =$$
$$\frac{1}{2\times\sqrt{3}}((-1)(-1)+(0)(-1)+(1)(2)) =$$
$$\frac{1}{2\times\sqrt{3}}(1+0+2) = \frac{1}{2\times\sqrt{3}}(3) = \frac{3}{2\times\sqrt{3}} \approx 0.866$$

Vi får en hög korrelation, nära 1, gå tillbaka till spridningsdiagrammet i A och gå igenom om värdet känsligt rimligt.

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. $r = 0.92$

Svar finns i boken, s. 483.

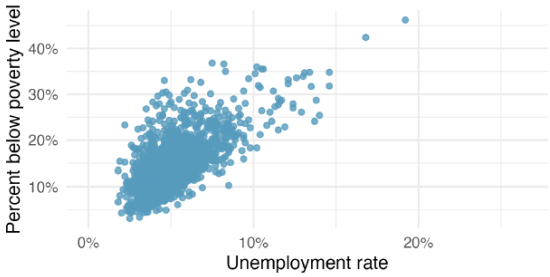
7. (a) $r = -0.7 \rightarrow (4)$. (b) $r = 0.45 \rightarrow (3)$. (c) $r = 0.06 \rightarrow (1)$. (d) $r = 0.92 \rightarrow (2)$.

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.



I bokens facit, s. 484, finns följande:

23. (a) $\widehat{\texttt{poverty}} = 4.60 + 2.05 \times \texttt{unemployment\_rate}$. (b) The model predicts a poverty rate of 4.60% for counties with 0% unemployment, on average. This is not a meaningful value as no counties have such low unexmployment, it just serves to adjust the height of the regression line. (c) For each additional percentage increase in unemployment rate, poverty rate is predicted to be higher, on average, by 2.05%. (d) Unemployment rate explains 46% of the variability in poverty levels in US counties. (e) $\sqrt{0.46} = 0.678$.

I tillägg:

a) Om frågan explicit gäller att skriva upp **populationsmodellen** kan vi skriva:

Definiera variabler:
Y – poverty (procentandel under fattigdomsgräns),
X – unemployment (procentandel arbetslösa),
ε – felterm

Populationsmodell:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

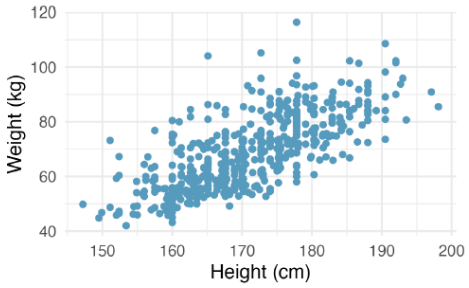
I ord: Procentandelen befolkning under fattigdomsgränsen är lika med en konstant (β_0 - beta0) (benämns typiskt **intercept**) plus en **lutningskoefficient** (β_1 - beta1) gånger procentandelen arbetslösa plus en **felterm** (ε - epsilon)

I facit har vi fått skattningen av modellen (regressionslinjen), med insatta värden för intercept och lutningskoefficient, dvs., sätt in skattade värden (och variabelnamn) i följande:

$$\hat{Y} = b_0 + b_1 X$$

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.

I bokens facit, s. 495, finns följande:

3. (a) The relationship is positive, moderate-to-strong, and linear. There are a few outliers but no points that appear to be influential. (b) $\widehat{\texttt{wgt}} = -105.0113 + 1.0176 \times \texttt{hgt}$. Slope: For each additional centimeter in height, the model predicts the average weight to be 1.0176 additional kilograms (about 2.2 pounds). Intercept: People who are 0 centimeters tall are expected to weigh -105.0113 kilograms. This is obviously not possible. Here, the y- intercept serves only to adjust the height of the line and is meaningless by itself. (c) H_0: The true slope coefficient of height is zero ($\beta_1 = 0$). H_A: The true slope coefficient of height is different than zero ($\beta_1 \neq 0$). The p-value for the two-sided alternative hypothesis ($\beta_1 \neq 0$) is incredibly small, so we reject H_0. The data provide convincing evidence that height and weight are positively correlated. The true slope parameter is indeed greater than 0. (d) $R^2 = 0.72^2 = 0.52$. Approximately 52% of the variability in weight can be explained by the height of individuals.

I tillägg:

Med ”least squares summary” avses skattningen av regressionsmodellen med minsta kvadrat-metoden.

a) Om frågan explicit gäller att skriva upp **populationsmodellen** kan vi skriva:

Definiera variabler:
Y – vikt (wgt) (enhet: kg),
X – längd (hgt) (enhet: cm),
ε – felterm

Populationsmodell:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

I facit har vi fått skattningen av modellen (regressionslinjen), med insatta värden för intercept och lutningskoefficient, dvs., sätt in skattade värden (och variabelnamn) i följande:

$$\hat{Y} = b_0 + b_1 X$$

b) Det är orealistiskt att ha personer med längd 0 cm och modellen har skattats med data från individer i längdintervallet (ca) 145-200 cm. Interceptet har ändå den teoretiska tolkning som ges i facit.

Om vi skulle använda modellen för att prediktera skulle vi inte prediktera för individer som har en längd långt utanför 145-200 cm.

c) Noll och alternativhypotes ges i facit, repeterade här:

$H_0 : \beta_1 = 0$ (Det finns inget samband mellan längd och vikt)
$H_A : \beta_1 \neq 0$ (Det finns ett samband mellan längd och vikt)

Vi har ett tvåsidigt test. Om vi exv. väljer signifikansnivån α=0.05 får vi ett kritiskt t-värde (z-värde) ± 1.96 (från standardnormalfördelningstabellen).

Från regressionsoutput ser vi att t-värdet (kallat ”statistic” i boken) är 23.1, dvs långt mycket högre än vad som krävs för att förkasta nollhypotesen på 5%-nivån. I linje med det höga t-värdet har vi ett mycket lågt p-värde.

Från F7 har vi definitionen av p-värde:

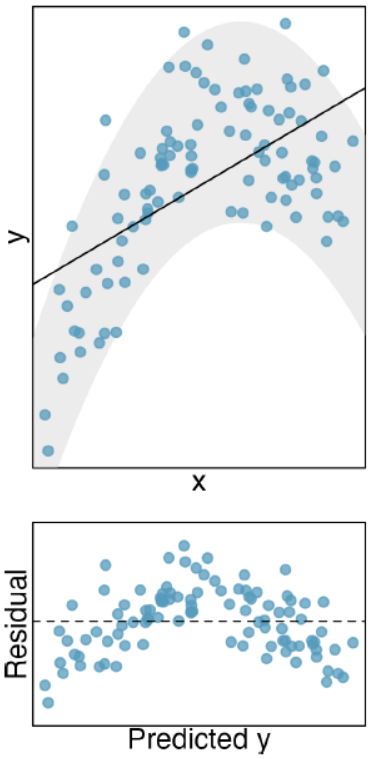
p-värde: Sannolikheten att observera ett värde som är minst lika extremt som det observerade, givet att nollhypotesen är sann

Dvs, om det inte finns ett samband mellan längd och vikt, vad är sannolikheten att observera det t-värde vi observerar? Vi ser i tabellen att sannolikheten är mindre än 0.0001, i detta fall är sannolikheten så låg att mjukvaran konstaterar att den är försumbar och skriver ut just < 0.001.

Vi förkastar nollhypotesen och finner istället stöd för alternativhypotesen, dvs att det finns ett samband mellan längd och vikt.

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.



För varje observation är residualen skillnaden mellan det faktiska och det skattade y-värdet (som återfinns på regressionslinjen), se boken s. 109.

För denna fråga kan vi använda samma svar som i boken s. 111, för ”Dataset 2”.

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.

I bokens facit, s. 495, finns följande, där fel/oklarheter i texten har korrigerats med röd text:

7. (b) H_0: The true slope coefficient of habit is zero ($\beta_5 = 0$). H_A: The true slope coefficient of habit is different than zero ($\beta_5 \neq 0$). The p-value for the two-sided alternative hypothesis ($\beta_5 \neq 0$)is incredibly low (0.0007)(smaller than 0.05), so we reject H_0. The data provide convincing evidence that habit and weight are positively correlated, given the other variables in the model. The true slope parameter is indeed greater than 0.

Stegen vi gått igenom för att komma fram till denna slutsats är:

Definiera variabler och ställ upp populationsmodellen:

$$weight = \beta_0 + \beta_1 weeks + \beta_2 mage + \beta_3 sexmale + \beta_4 visits + \beta_5 habitsmoker + \varepsilon$$

Skatta modellen:

$$\hat{weight} = b_0 + b_1 weeks + b_2 mage + b_3 sexmale + b_4 visits + b_5 habitsmoker$$

där värdet på koefficientskattningarna b_1 - b_5 finns i regressionstabellen. Notera hattsymbolen på utfallsvariabeln – skattat värde.

Ställ upp noll- och alternativhypotes gällande populationskoefficienten för habitsmoker (”habit”):

$H_0 : \beta_5 = 0$ (Det finns inget samband mellan rökning ("habit") och födelsevikt ("weight"))
$H_A : \beta_5 \neq 0$ (Det finns ett samband mellan rökning ("habit") och födelsevikt ("weight"))

Vi har ett tvåsidigt test. Om vi exv. väljer signifikansnivån α=0.05 får vi ett kritiskt t-värde (z-värde) ± 1.96 (från standardnormalfördelningstabellen).

Från regressionsoutput ser vi att t-värdet (t_{obs}) är -3.41, dvs ett mer negativt värde (ett mer extremt värde, ett värde längre från noll) än vad som krävs för att förkasta nollhypotesen på 5%-nivån (-1.96).

Vi förkastar nollhypotesen och finner istället stöd för alternativhypotesen, dvs. att det finns ett samband mellan mammans rökning och barnets vikt. Vi kan inte säga att vi har bevisat alternativhypotesen men vi har funnit mycket starkt stöd för alternativhypotesen.

(t-värdet är negativt eftersom sambandet mellan rökning och vikt är negativt (röker – lägre vikt))

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

Vi har en regressionstabell från en studie, i detta fall Cheungs m.fl. artikel från Ekonomisk Debatt, 2024:

Tabell 1 Skattningar av effekter

	Utflöde från arbetslöshet	Dagar i arbetslöshet första kvartalet	Dagar i arbetslöshet första året
Direkt effekt behandlingsgrupp	0,034*** (0,006)	−1,456*** (0,353)	−5,991*** (1,359)
Undanträngning kontrollgrupp	−0,015** (0,006)	0,695** (0,318)	4,160** (1,727)
Nettoeffekt behandlingsgrupp	0,018*** (0,005)	−0,761** (0,326)	−1,831 (1,484)
Medelvärde	0,390	73,78	187,0
Observationer	552 816	552 816	552 816

Anm: Tabellen redovisar regressioner av respektive utfallsvariabel på en indikator för deltagare (”Direkt effekt behandlingsgrupp”) och en indikator för att vara inskriven i ett aktivt arbetsförmedlingskontor (”Undanträngning kontrollgrupp”). Medelvärde avser kontrollkontoren. Standardfel klustrade på kontorsnivå inom parentes. *** p<0,01, ** p<0,05, *p<0,1.
Källa: Egna beräkningar.

Vi fokuserar enbart på siffrorna i den röda rutan. Författarna har skattat hur stor andel av individer i en viss grupp inskrivna på Arbetsförmedlingen som under en tremånadersperiod ”flödar ut” från arbetslöshet (i princip: får ett jobb) som resultat av deltagande i en viss åtgärd. Effekten skattas till 0.034 - i detta fall innebär effekten 3.4 procentenheter ökad sannolikhet att lämna arbetslöshet jämfört med om man inte deltagit i åtgärden.

Under 0.034 står 0.006 inom parentes och vi får reda på i tabellförklaringen att denna siffra är standardfelet. (Vi kan för vårt sammanhang bortse från kommentaren om klustring.)

Om vi tänker på ”Direkt effekt behandlingsgrupp” som vår variabel 1 (det är en dummyvariabel för om en individ har fått möjlighet att delta i fler möten med arbetsförmedlare) i en multipel linjär regression har vi följande hypoteser:

$H_0 : \beta_1 = 0$ (Det finns inget samband mellan deltagande och utflöde från arbetslöshet)
$H_A : \beta_1 \neq 0$ (Det finns ett samband mellan deltagande och utflöde från arbetslöshet)

I tabellen har vi den skattade koefficienten och dess standardfel. Vi har däremot inte fått t-värdet, men kan ta fram det, vi har att testvariabeln (F9, s. 24)

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

i vårt fall är lika med (vårt observerade t-värde)

$t_{obs} = \frac{0.034}{0.006} \approx 5.7$

(Vårt värde är ungefärligt eftersom författarna troligen rapporterar avrundade värden i tabellen).

Vi har ett tvåsidigt test. Om vi exv. väljer signifikansnivån $\alpha=0.05$ får vi ett kritiskt t-värde (z-värde) ± 1.96 (från standardnormalfördelningstabellen).

Vi förkastar nollhypotesen och finner istället stöd för alternativhypotesen, dvs att det finns ett samband mellan att vara med i gruppen som erbjuds fler arbetsförmedlarmöten och utflöde från arbetslöshet.

I vissa tidskrifter och regressionstabeller används stjärnor för att markera statistiskt signifikanta resultat.

Den andra raden i tabellen rapporterar skattad lutningskoefficient för förklaringsvariabel nummer 2.

De två andra kolumnerna i tabellen representerar skattningar, med samma förklaringsvariabler, men med andra utfallsvariabler.