Kapittel 2

Utforske og beskrive data

Sammenhenger mellom variable

Forrige uke: Kap. 2.1 om assosiasjon og kryssplott

Denne uken: Kap. 2.2, 2.3, 2.4 om korrelasjon og minste kvadraters regresjon

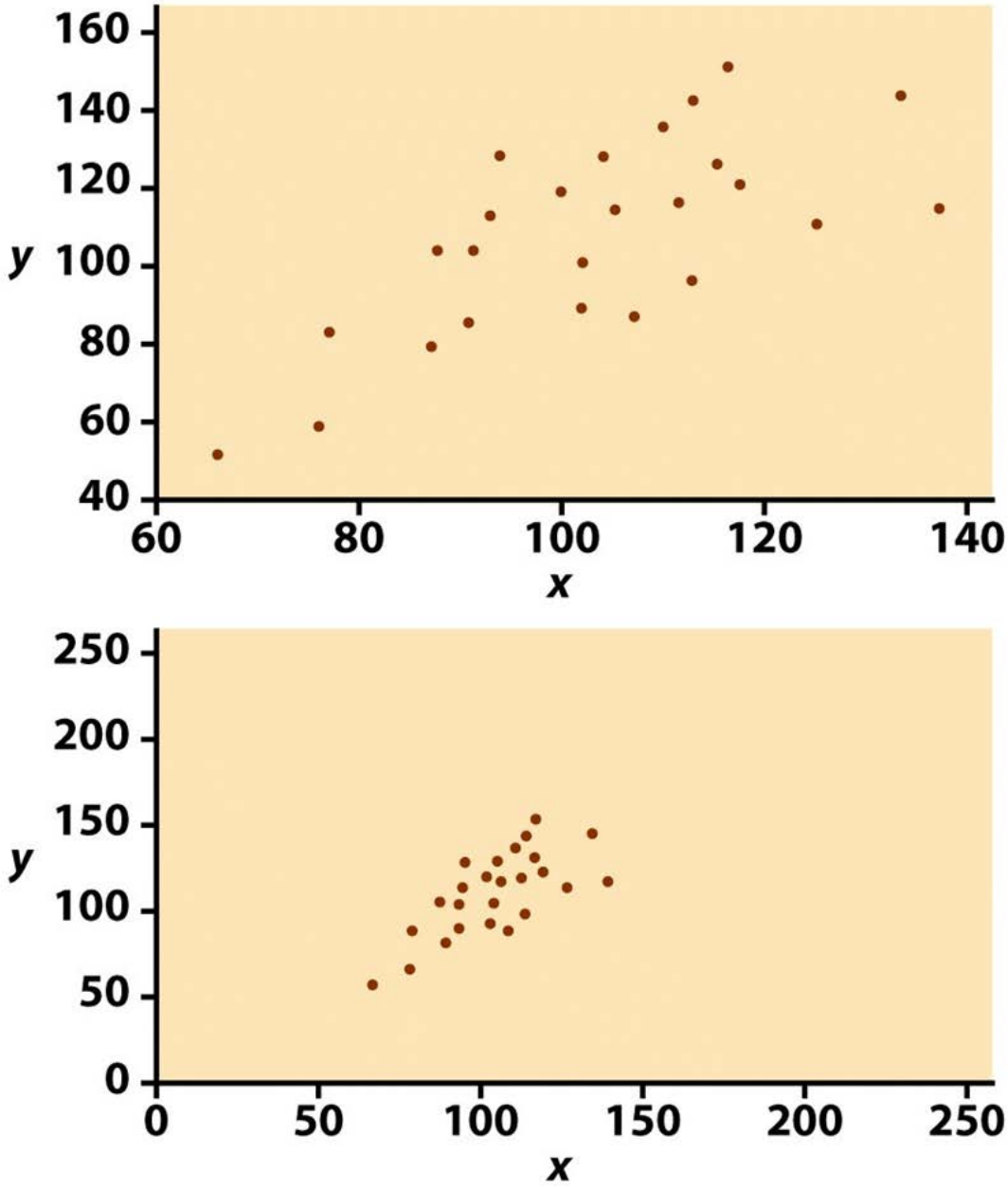


Figure 2-9
Introduction to the Practice of Statistics, Fifth Edition

- To kryssplott av samme datasett, men med forskjellig skala på aksene
- Det nederste plottet ser ut til å vise en sterkere sammenheng?!

2.2 Korrelasjon

- Kryssplott viser oss formen, retningen og en idé om styrken på sammenhengen mellom to kvantitative variable
- Lineære sammenhenger er sterke dersom punktene ligger nærme en rett linje, og svake dersom de er veldig mye spredt rundt linjen
- Men: Våre øyne ikke gode til å bedømme hvor sterk en sammenheng er
- Trenger kvantitativt mål på sammenheng
- Korrelasjon et slikt mål

Korrelasjon

- Korrelasjon: Kvantitativt mål på lineær sammenheng mellom to kvantitative variable
- Observasjoner av variablene *x* og *y* for *n* individer: Individ *i* har datapunktene (x_i,y_i)
- Korrelasjonen *r* mellom *x* og *y* er

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

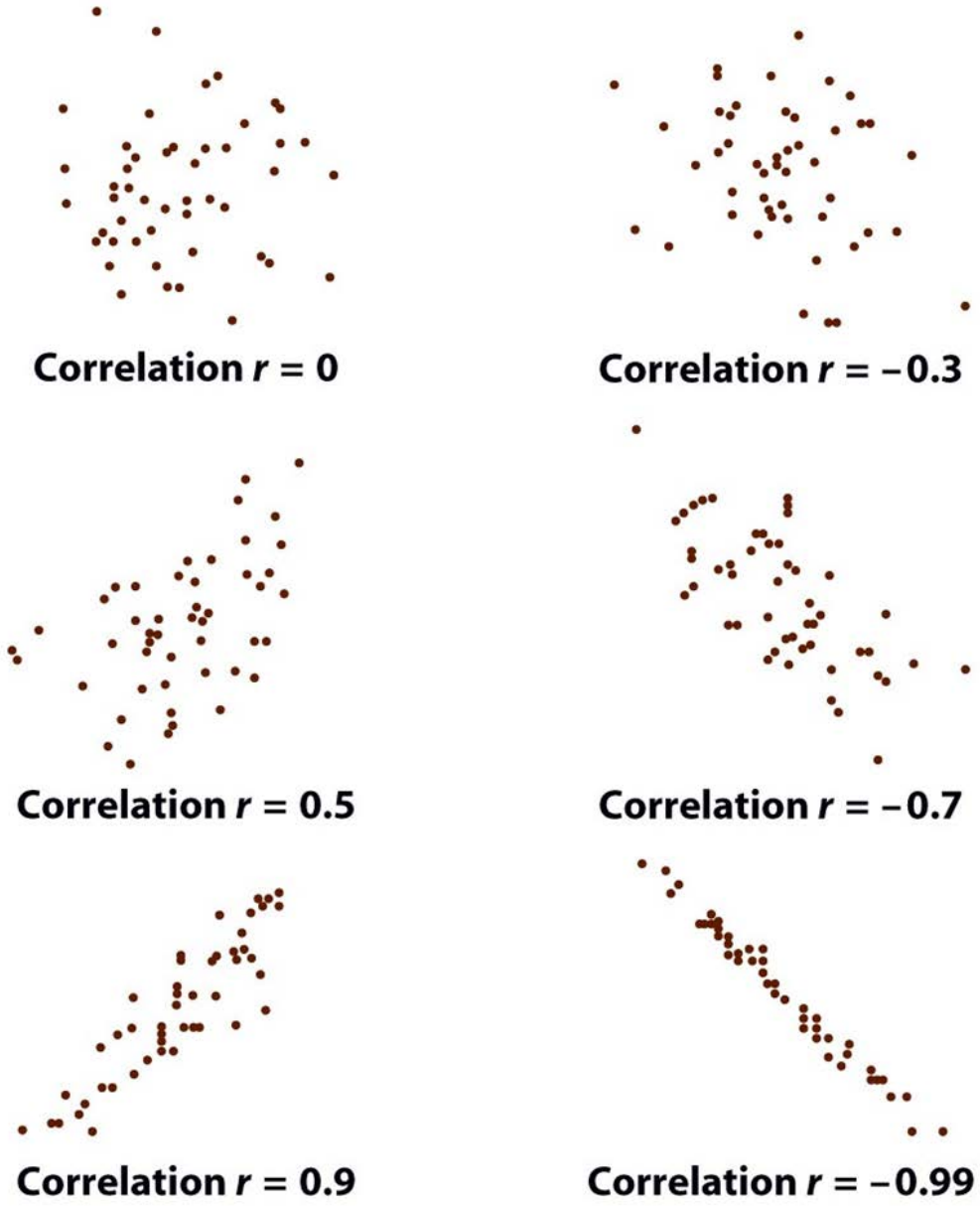
(Σ betyr "sum over alle individene")

- *r* er et gjennomsnitt av produktet av de standardiserte observasjonene for hvert individ

Egenskaper korrelasjon

- Positiv *r* svarer til positiv sammenheng og vice versa
- *r* ligger mellom -1 og 1
- Perfekt korrelasjon, r=1 eller -1, svarer til alle punkt på en rett linje
- Basert på standardiserte verdier, uavhengig av senterpunkt, skala
- Måler styrke av lineær sammenheng
- Lite robust for ekstreme verdier
- Skiller ikke mellom forklarings- og respons-variable
- Krever at begge variable er kvantitative

Korrelasjon: Måler styrke og retning av lineær sammenheng



Korrelasjon: Måler styrke og retning av lineær sammenheng

r = -0.877

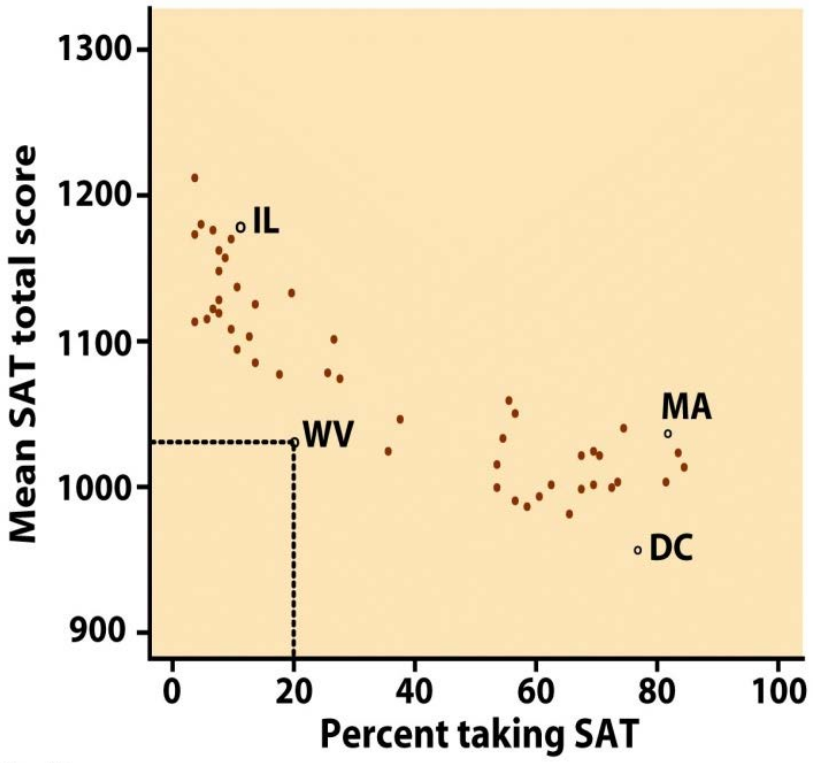


Figure 2-1
Introduction to the Practice of Statistics, Fifth Edition

r = -0.113

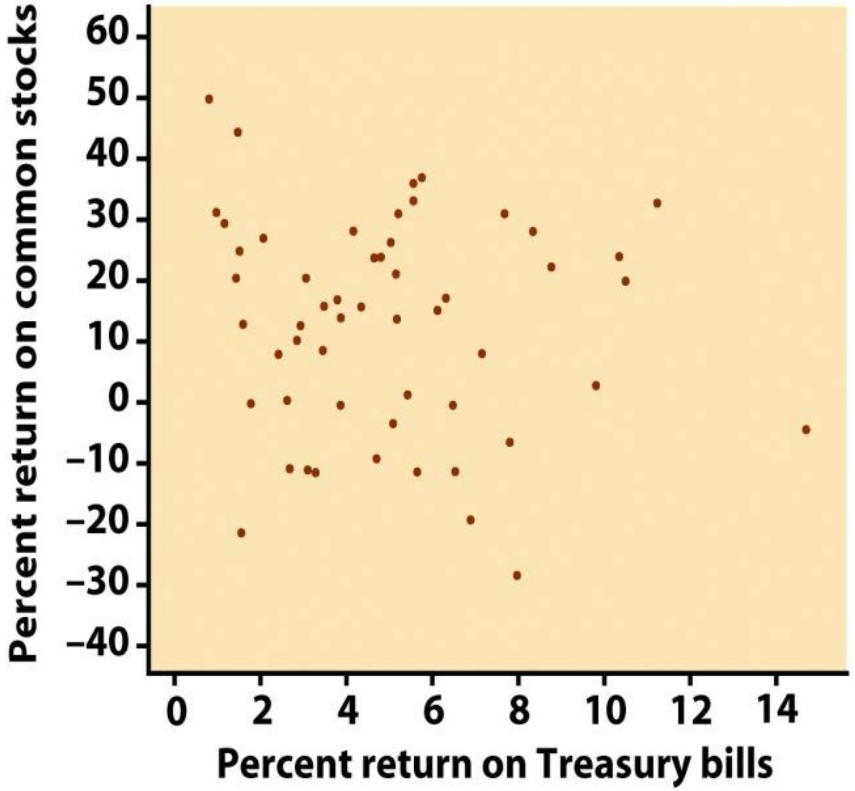
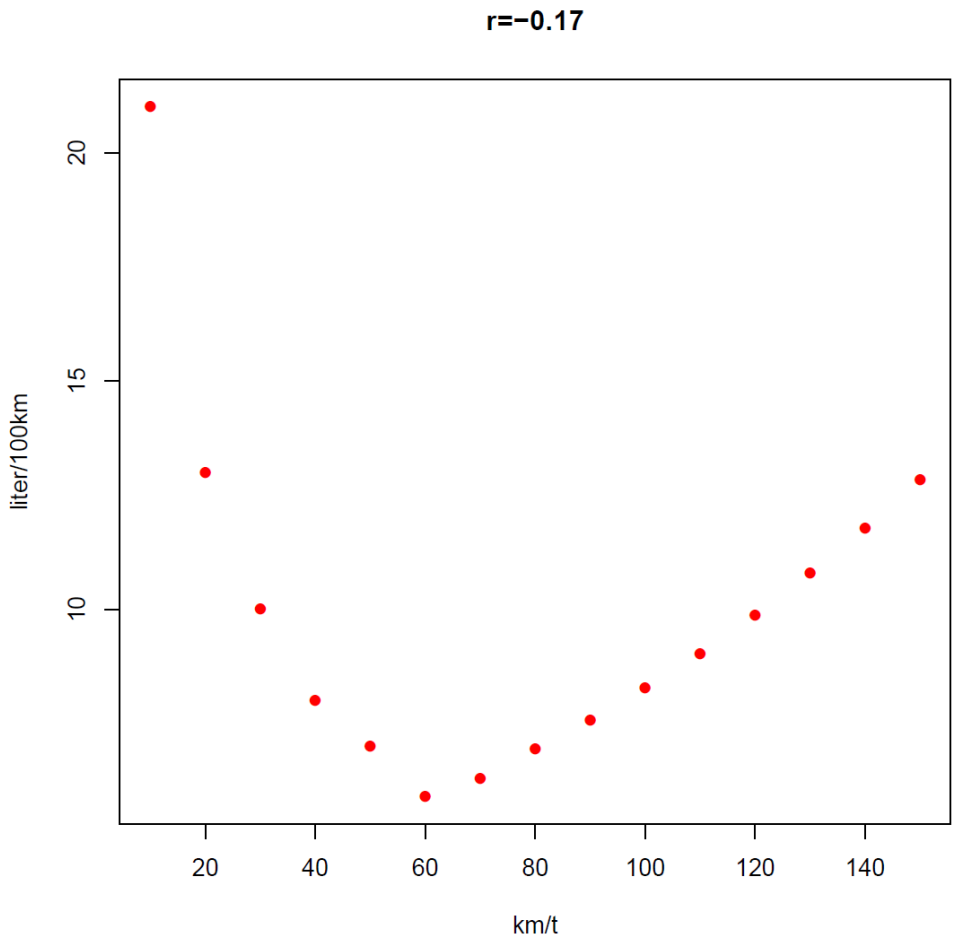


Figure 2-7
Introduction to the Practice of Statistics, Fifth Edition

Korrelasjon: Måler styrke og retning av lineær sammenheng

- Spredningsplott av forskjellige hastigheter for en bil mot målinger av hvor mye drivstoff bilen bruker ved de ulike hastighetene
- Ser en klar sammenheng, men…



2.3 Regresjon

En regresjonslinje er en rett linje som beskriver hvordan responsvariabelen *y* endrer seg når forklaringsvariabelen *x* skifter verdier.

Vi sier ofte at regresjonslinjen predikerer verdien av *y* for en gitt verdi av *x*.

Krever en responsvariabel og en forklarings-variabel

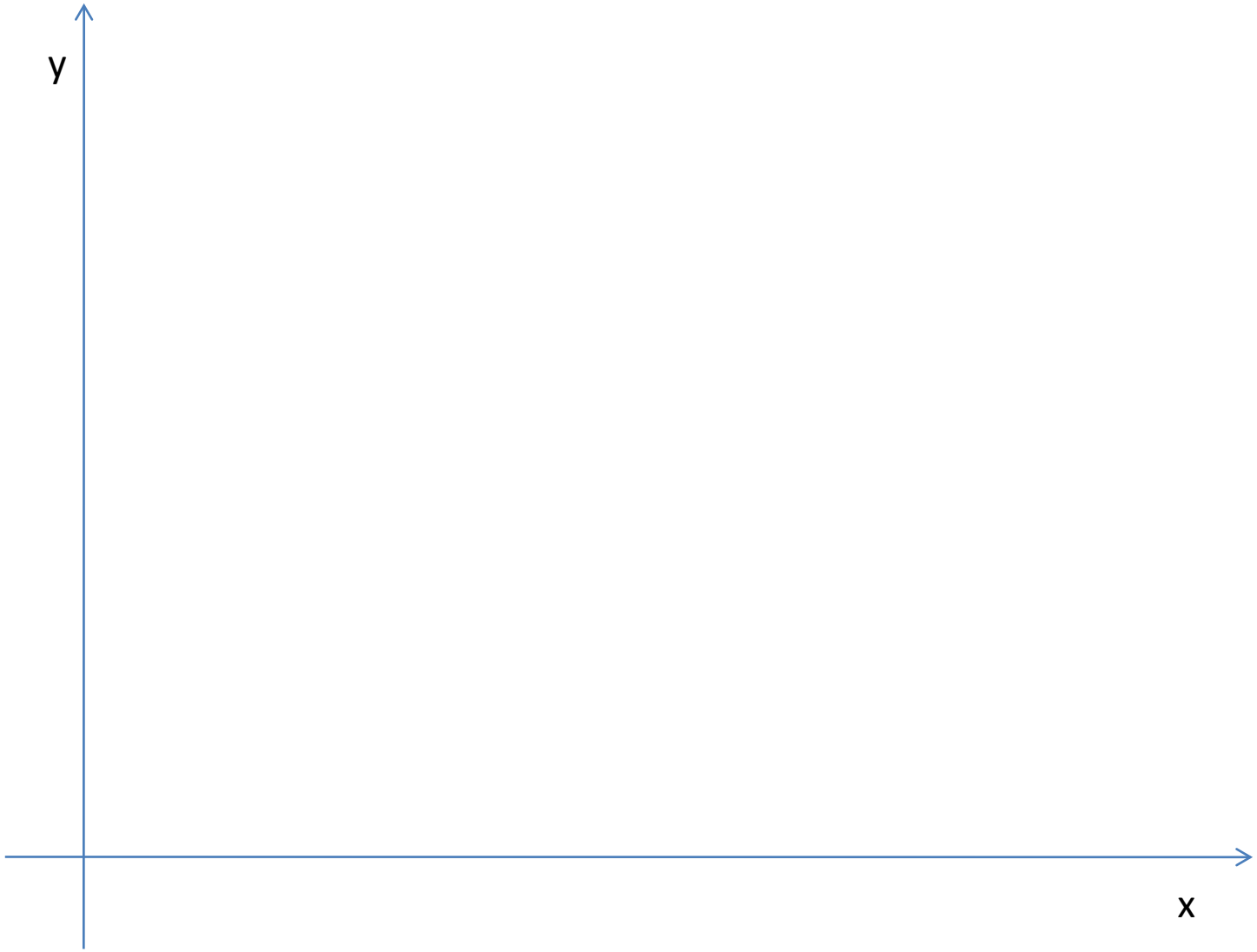
Rette (lineære) linjer

En rett linje som relaterer y til x har en likning på formen

$$y=a+bx$$

b kalles stigningstallet, mengden y endrer seg når x endrer seg med en enhet.

a kalles skjæringspunktet, verdien y tar for $x=0$


y
x

Eks 2.18 Rastløshet og vekt

- Holder rastløshet deg slank?

- 16 friske forsøkspersoner overspiser i 8 uker.
- Noen legger mye på seg, andre ikke.
- Kan dette forklares av økt 'rastløshet'?

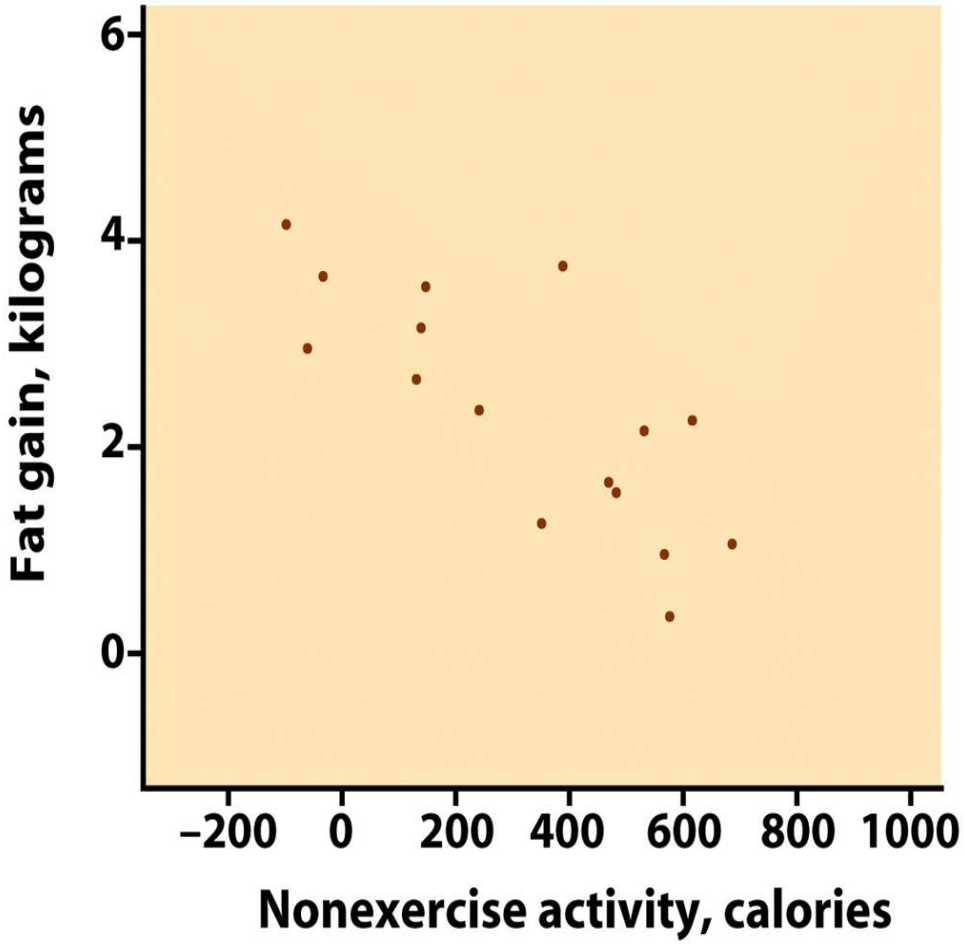


Figure 2-11
Introduction to the Practice of Statistics, Fifth Edition

• y = fettøkning, x = økt energiforbruk (utenom-trenings-aktivitet)

MODELL for sammenhengen

$$y = a + bx = 3.505 - 0.00344\,x$$

• a=3.505 er skjæringspunktet og beskriver vektøkning hvis utenom-trenings-aktiviteter ikke øker

•b= -0.00344 er stigningstallet og beskriver at vektøkningen går ned med 0.0034 kg for hver ekstra kalori utenom-trenings-aktiviteter

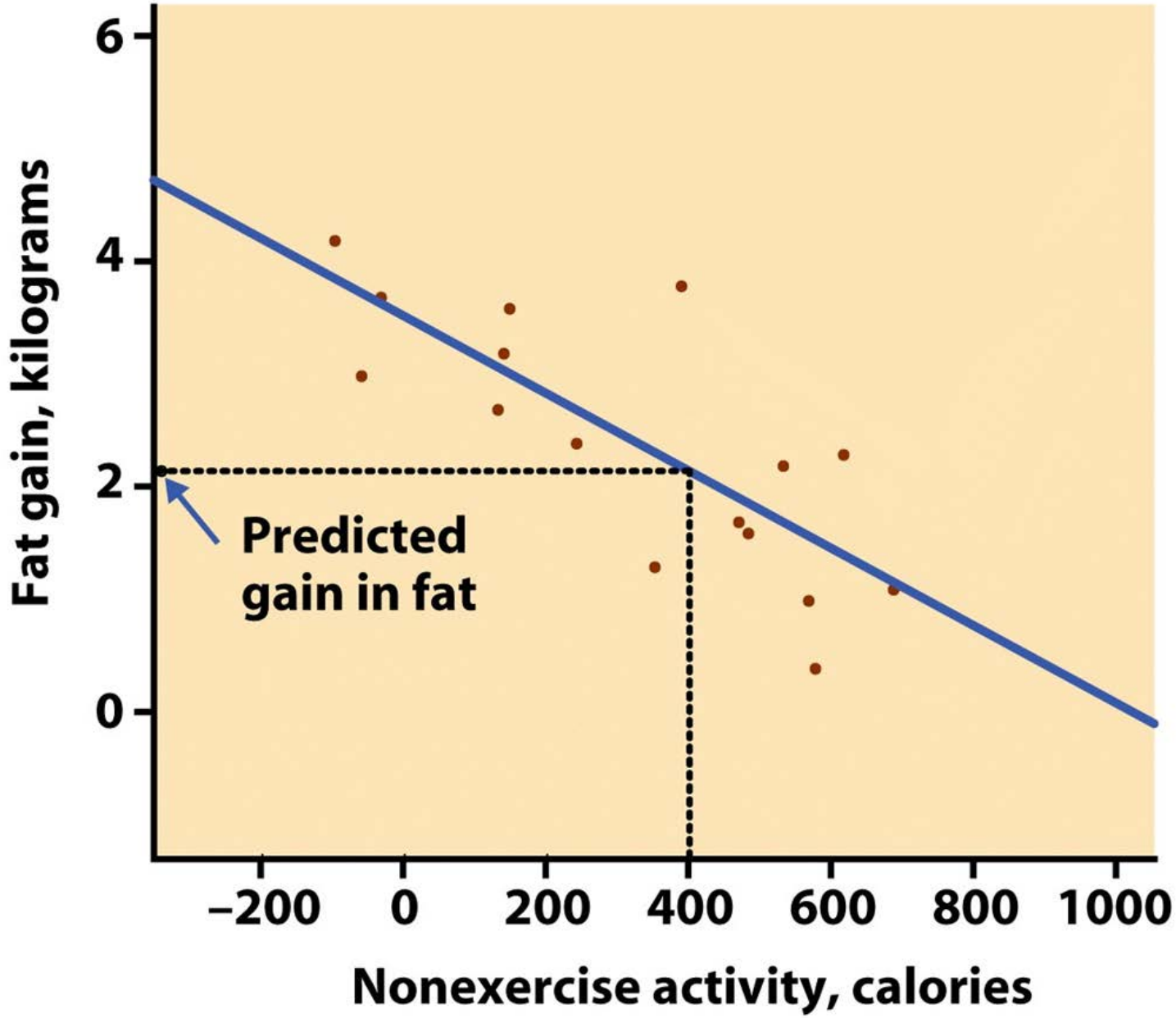


Figure 2-12
Introduction to the Practice of Statistics, Fifth Edition

Prediksjon

- Regresjonslinje kan brukes til å predikere respons *y* for en gitt verdi av forklaringsvariabel *x*
- $x = 400$ kalorier
- $y = 3.505 - 0.00344 * 400 = 2.13$ kg

Ekstrapolering

- Ekstrapolering er å bruke regresjonslinjen langt utenfor området av verdier på *x* i datasettet
- x=*1500* gir *y = 3.505-0.00344*1500=-1.66* kg
- x=1500 var utenfor de målte x-verdiene
- Slike prediksjoner ofte ikke særlig presise!!!

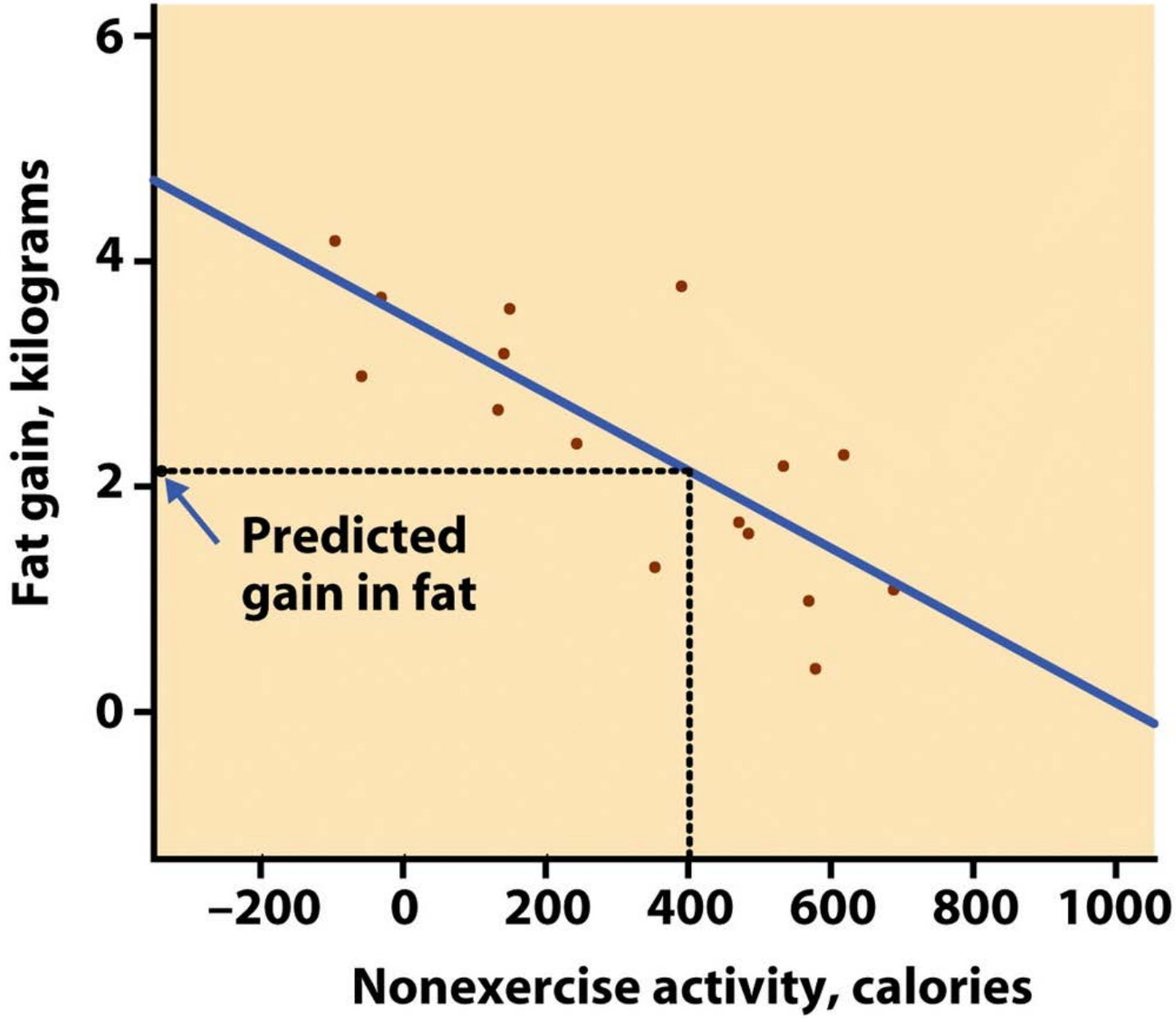


Figure 2-12
Introduction to the Practice of Statistics, Fifth Edition

Minste kvadraters regresjon

- Hvordan finne «beste» *a* og *b* fra data?
- Ingen linje vil gi perfekt tilpasning
- Ønsker *vertikal* avstand mellom linje og observert *y* verdi minst mulig –mist mulig prediksjonsfeil

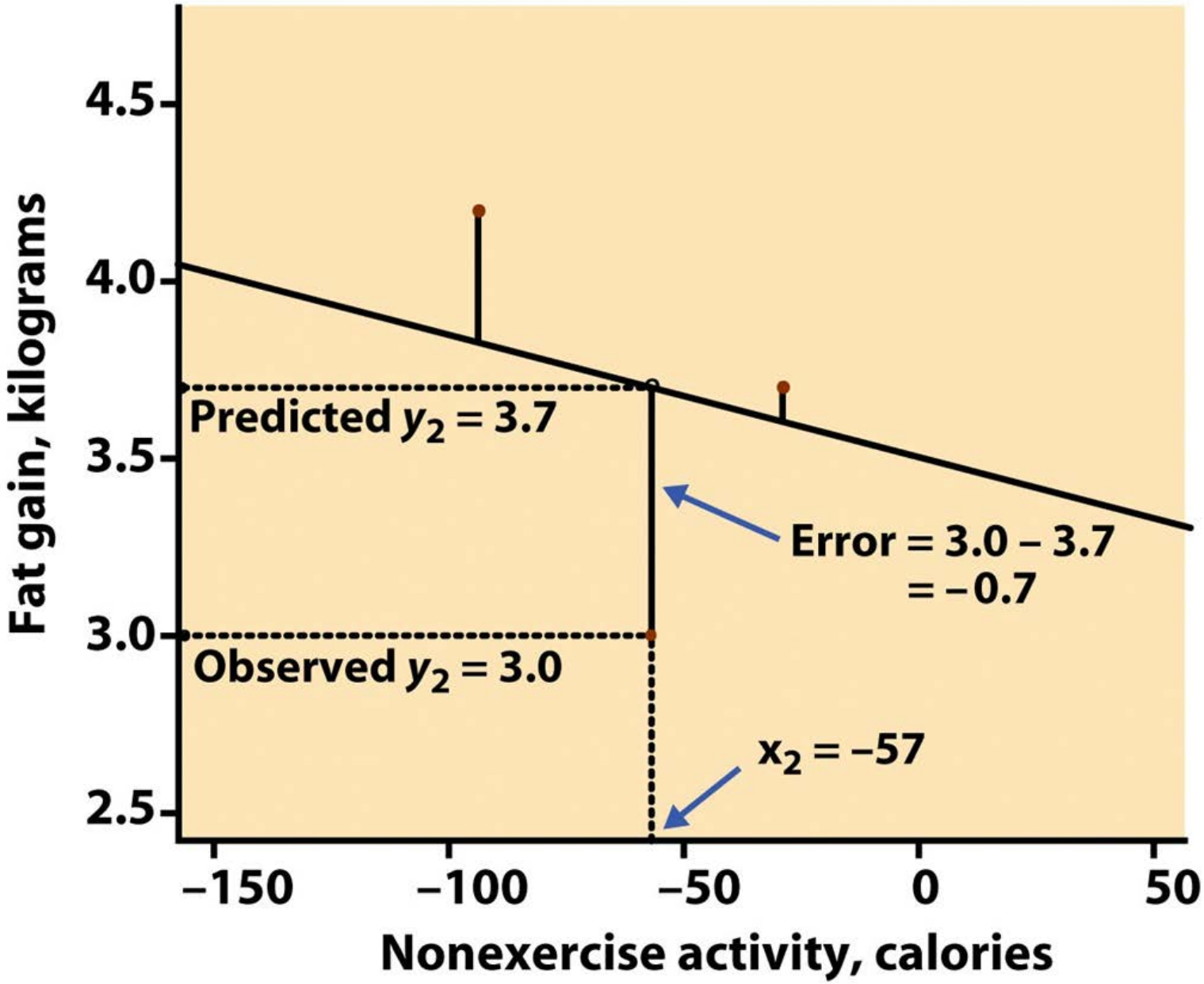


Figure 2-13
Introduction to the Practice of Statistics, Fifth Edition

Minste kvadraters regresjonslinje: Linjen som gjør kvadratsummen av vertikale avstander minst mulig

Observasjoner $(x_1,y_1),\ldots,(x_n,y_n)$

Minimerer

$$\sum (error)^2 = \sum (y_i - a - bx_i)^2$$

Skal minimere

$$\sum (error)^2 = \sum (y_i - a - bx_i)^2$$

med hensyn på a og b

Partiell-deriverer mhp a og b, setter lik 0.
Gir to ligninger med to ukjente, løses mhp a og b.

Prøv selv!! (ikke pensum)

Likninger for minste kvadraters regresjonslinje

Regresjonslinje $\hat{y} = a + bx$

der stigningstallet er $b = r\, s_y / s_x$

og skjæringspunktet er $a = \bar{y} - b\,\bar{x}$

Her er $\bar{x}$ og $\bar{y}$ gjennomsnittet av x- og y-verdiene

og s_x og s_y standardavvik for x og y

Vekt og rastløshet: Utregning av minste kvadraters regresjonslinje

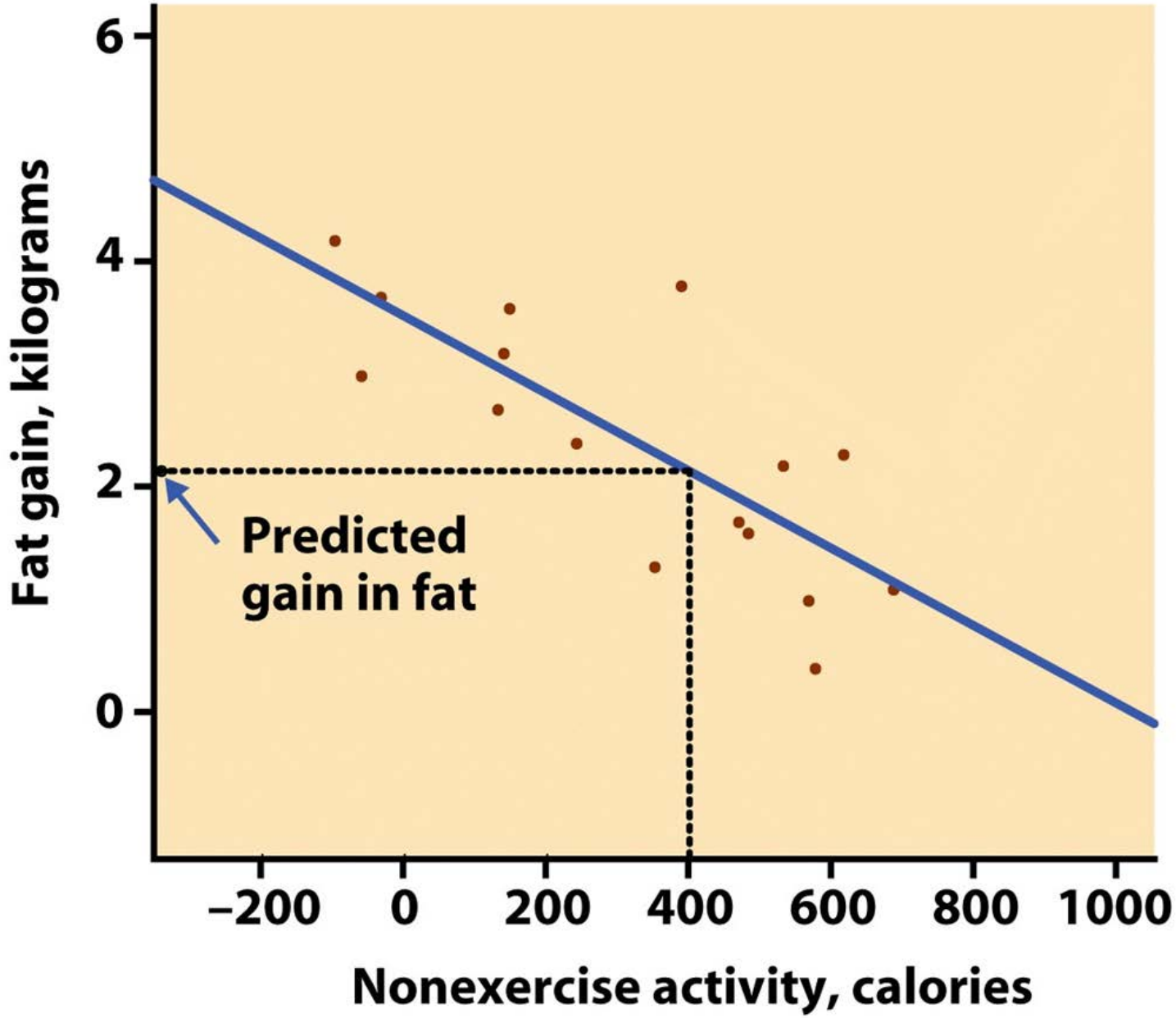


Figure 2-12
Introduction to the Practice of Statistics, Fifth Edition

Egenskaper regresjonslinje

- En endring på ett std.avvik i x gir en endring på r std.avvik i y
- Hvis $r = -1$ eller 1 gir ett std.avvik endring i x (– eller +) ett std.avvik endring i y. Ellers mindre endring i y, dvs. y responderer mindre på endring i x
- Linjen går alltid gjennom $(\bar{x}, \bar{y})$

Regresjon og korrelasjon

- Regresjonslinje basert på *y* respons og *x* forklaringsvariabel
- Korrelasjonen *r* er symmetrisk i *x* og *y*, irrelevant hvilken som er responsvariabel og hvilken som er forklaringsvariabel.

- En korrelasjon (r=0.7842), men to ulike regresjonslinjer avhengig av hvilke variable som blir valgt som responsvariabel og forklaringsvariabel
- Heltrukken linje: Fart er respons og avstand er forklaringsvariabel
- Stiplet linje: Omvendt

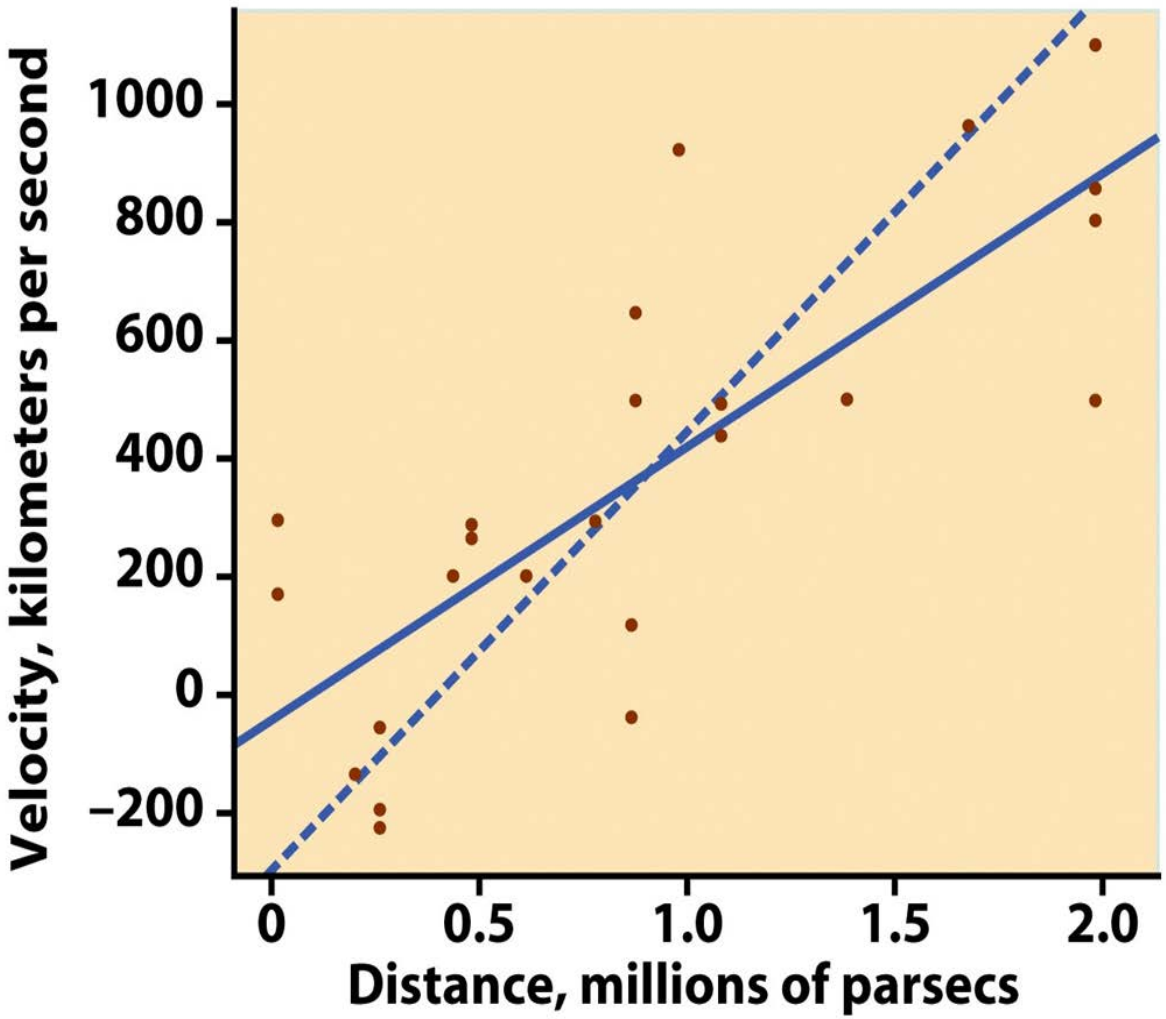


Figure 2-15
Introduction to the Practice of Statistics, Fifth Edition

Regresjon og korrelasjon

- Men det er en sammenheng mellom korrelasjon og regresjon
- r^2 forklarer andelen av variasjon i *y* som kan forklares av *x*
- To kilder til variasjon, variasjon langs linjen (forklart av *x*) og variasjon rundt linjen (ikke forklart av *x*)
- I eksempel,

$$r = -0.7786, r^2 = 0.606$$

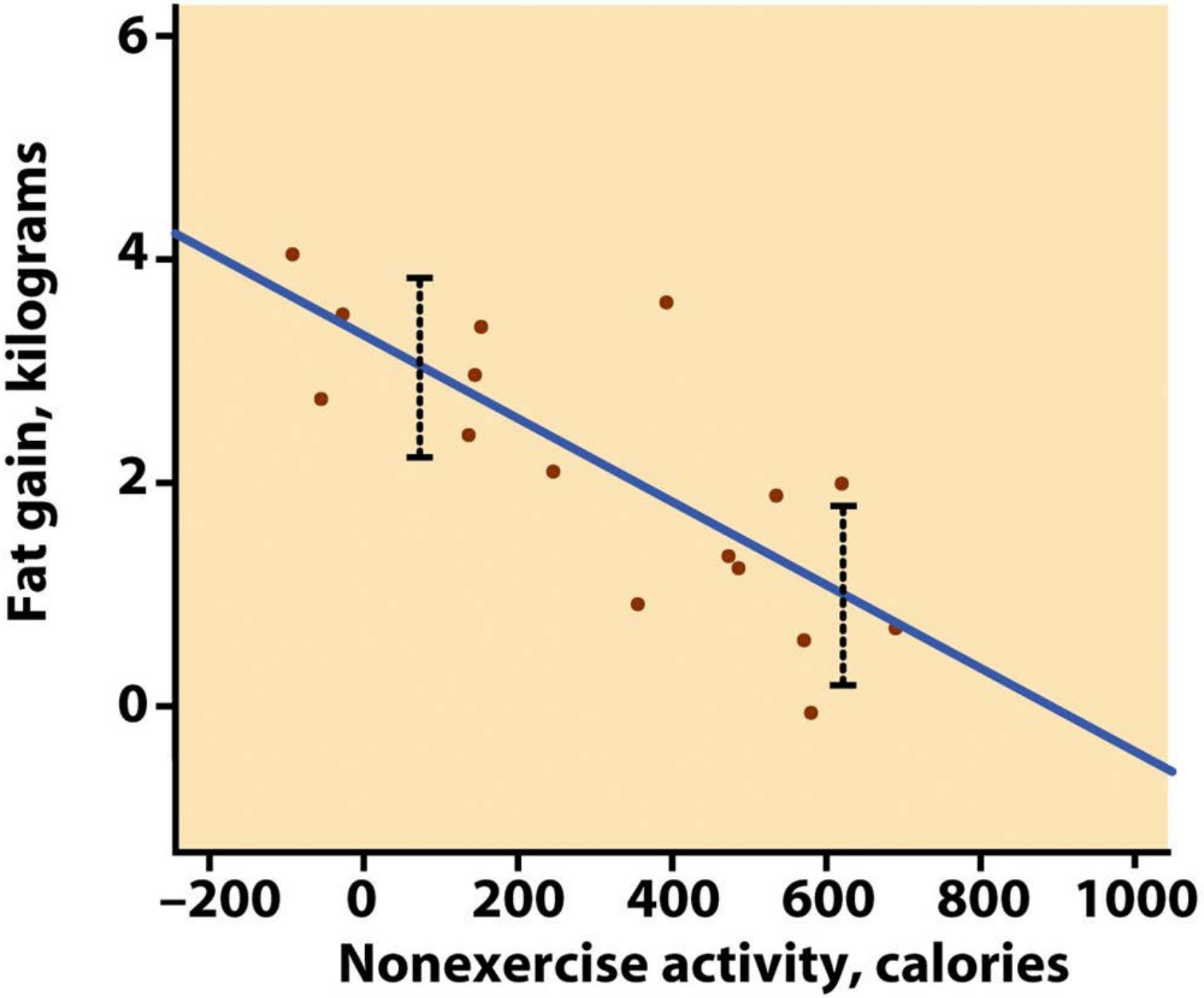


Figure 2-16
Introduction to the Practice of Statistics, Fifth Edition

Regresjon i Minitab

Vekt og rastløshet

Kapittel 2.4

Forsiktighetsregler ved bruk av regresjon og korrelasjon

Oppsummering

- Regresjonslinje: $\hat{y} = a + bx$ - en rett linje som beskriver hvordan responsvariabelen *y* endrer seg når forklaringsvariabelen *x* skifter verdier
- Minste kvadraters regresjonslinje: Linjen som gjør kvadratsummen av vertikale avstander minst mulig
- Minste kvadraters regresjonslinjen er gitt av
  - $stigningstallet\ b = r\, s_y / s_x$
  - $skjæringspunktet\ a = \bar{y} - b\,\bar{x}$
- Regresjonslinje kan brukes til å predikere respons *y* for en gitt verdi av forklaringsvariabel *x*

Regresjon: Eksempel

- Igjen: Rastløshet (NEA) og vektøkning
- Minste kvadraters regresjonslinje

- Prediksjon: Ett individ hadde en NEA-økning på 135 kalorier. Predikert vektøkning:

$$\hat{y} = 3.505 - 0.00344 * 135 = 3.04$$

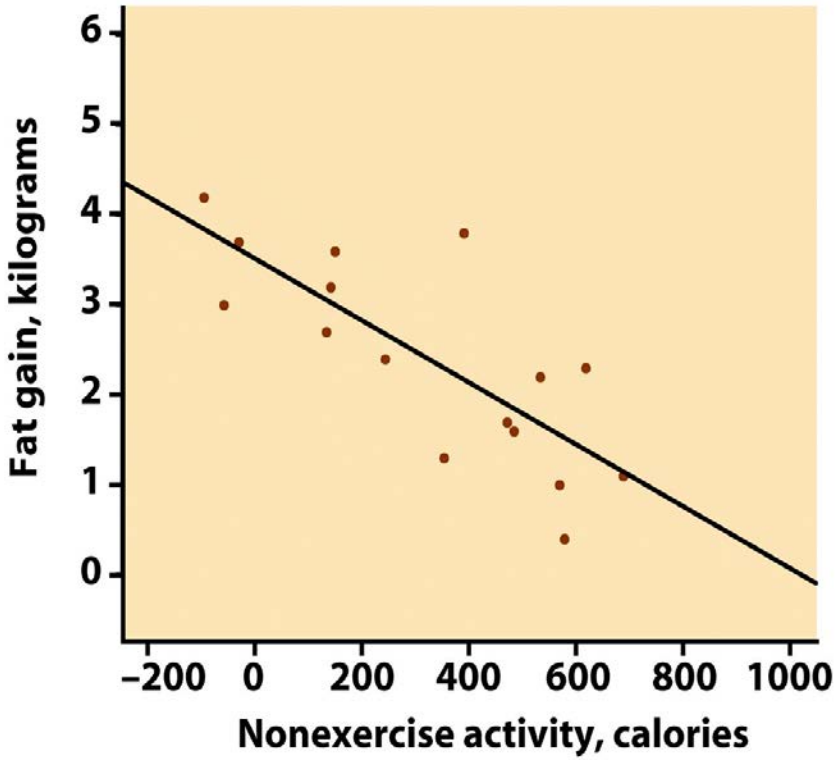


Figure 2-20a
Introduction to the Practice of Statistics, Fifth Edition

Residualer

- Residualer er differensen mellom observert verdi og predikert verdi:

$$\text{residual} = \text{observert } y \text{ - predikert } y =$$
$$y - \hat{y} = y - (a + bx)$$

- "Resten", det vi ikke har forklart ved forklaringsvariabelen gjennom regresjonslinjen
- Residual for hver observasjon: $e_i = y_i - (a + bx_i)$
- Residualer summerer seg til 0
- Nyttige for modell-sjekk

Residualer: Eksempel

- Fra forrige gang: Rastløshet (NEA) og vektøkning
- Regresjonslinje

$$\hat{y} = 3.505 - 0.00344x$$

- Ett individ hadde en NEA-økning på 135 kalorier. Vi predikerte vektøkning $\hat{y} = 3.04$. Observert vektøkning var $y = 2.7$

residual=observert *y* - predikert *y*=

$$y - \hat{y} = 2.7 - 3.04 = -0.34$$

- Tilsvarende kan vi finne de 15 andre residualene

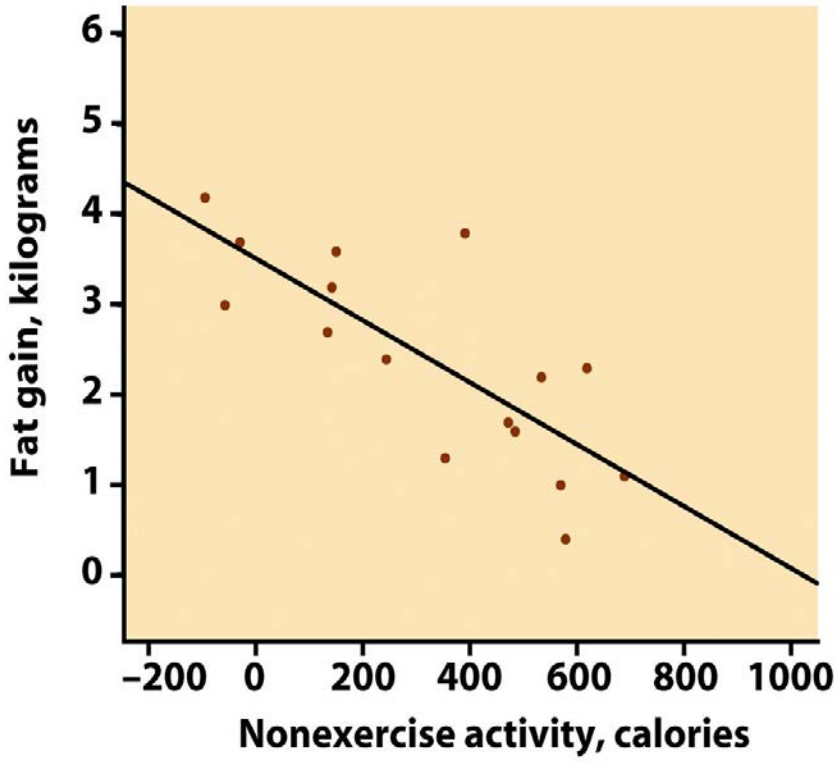


Figure 2-20a
Introduction to the Practice of Statistics, Fifth Edition

Residualplott

- Residualene beskriver hvor langt (vertikalt) dataene ligger fra regresjonslinjen
- Derfor nyttig å undersøke dem for å se hvor godt linjen beskriver dataene
- Ser på kryssplott av residualer mot forklaringsvariabel
- Hjelper til å vurdere tilpasningen av en regresjonslinje
- Legger gjerne på en horisontal linje gjennom 0 (som da tilsvarer regresjonslinjen)
- Bør *ikke* være noe mønster i residualene, dvs punktene skal være tilfeldig plassert rundt den horisontale linjen residual=0

Residualplott: Eksempel

- Utenom-trenings-aktivitet (NEA) og vektøkning
- Mønster?
- Nei, punktene ser ut til å være tilfeldig plassert rundt linjen residual=0

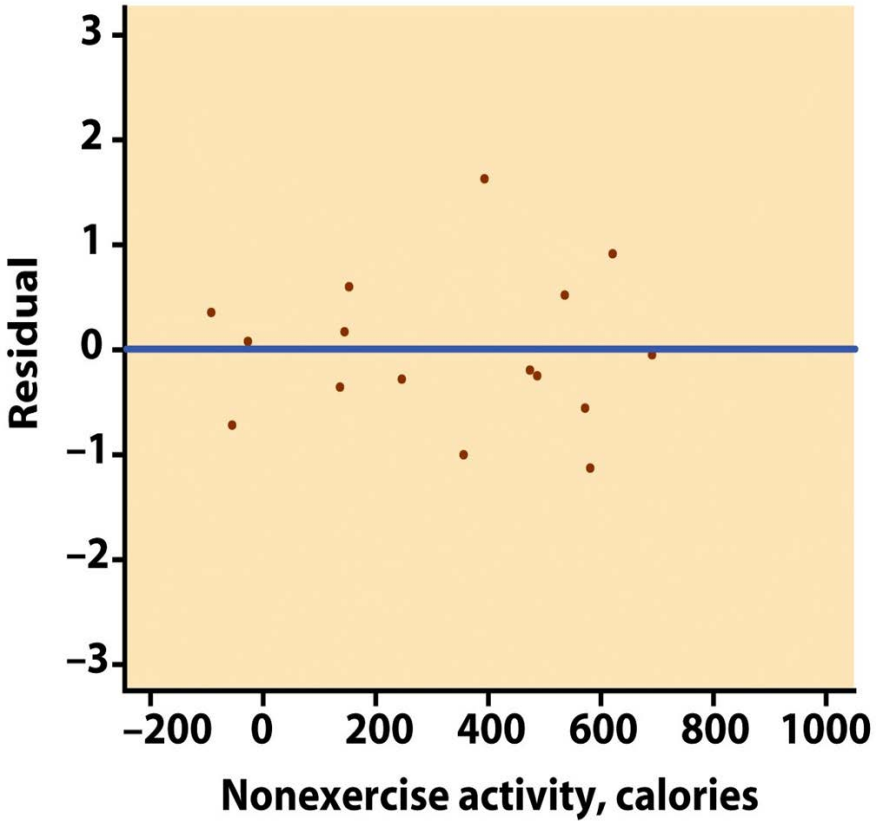


Figure 2-20b
Introduction to the Practice of Statistics, Fifth Edition

Residualplott: Eksempel

- Måling av defekter i oljerør, måling i felten og måling i laboratorie
- Den blå linjen y=x viser at feltmålingene tenderer til å være lavere enn laboratoremålingene for store defekter
- Regresjonslinjen (ulik y=x) går gjennom sentrum av punktene
- Fanger opp det at feltmålingene tenderer til å være lavere enn laboratoremålingene for store defekter

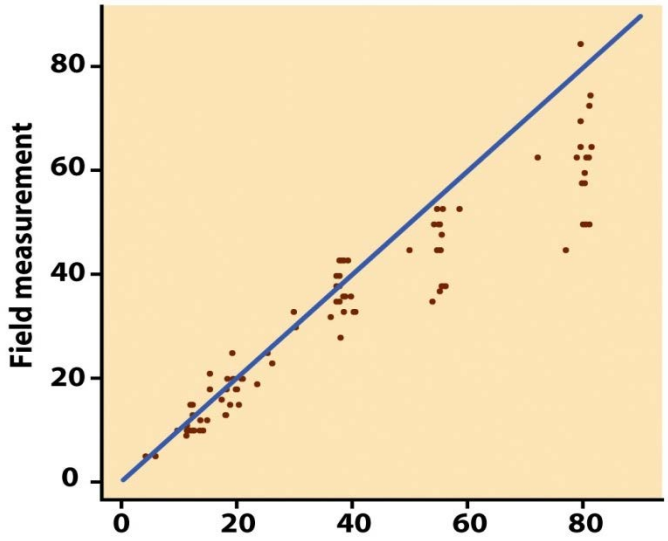
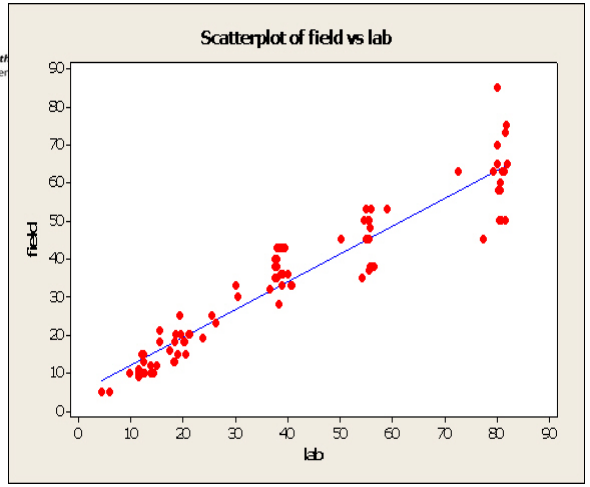


Figure 2-3
Introduction to th
© 2005 W. H. Freer



Residualplott: Eksempel

- Mønster?

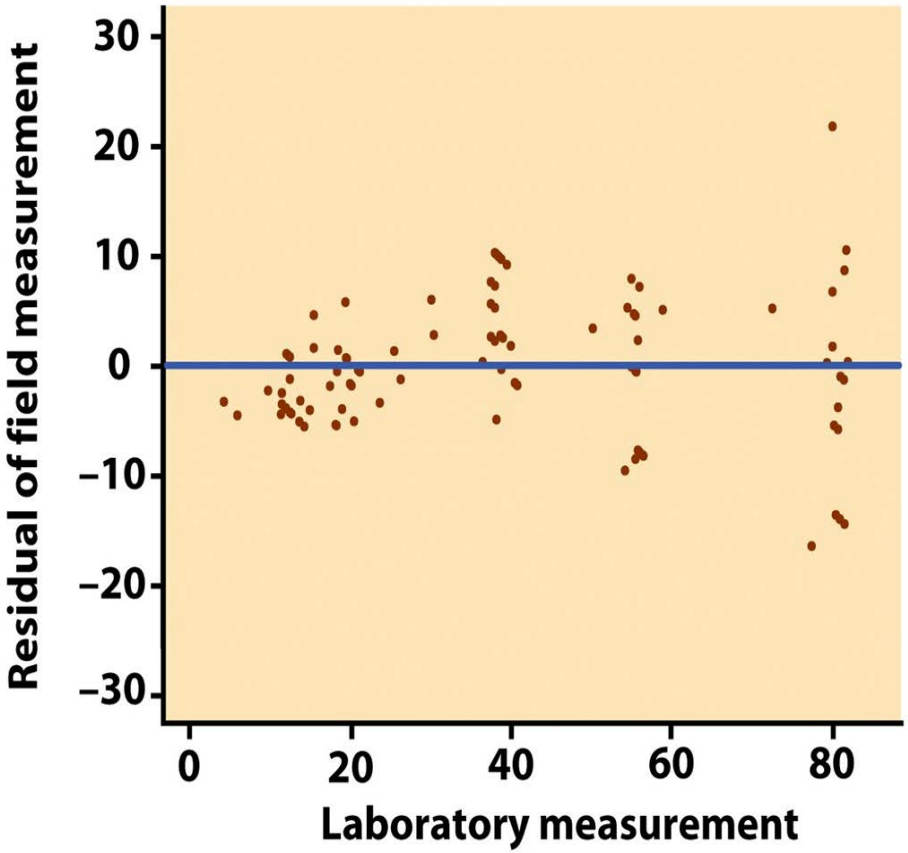


Figure 2-21
Introduction to the Practice of Statistics, Fifth Edition

Residualplott: Eksempel

- Mønster?
- Ser at stemmer ganske godt overens med lineær sammenheng
- Men, ser at residualene tenderer ti å ha større spredning for store defekter
- Fremheves av residualplottet

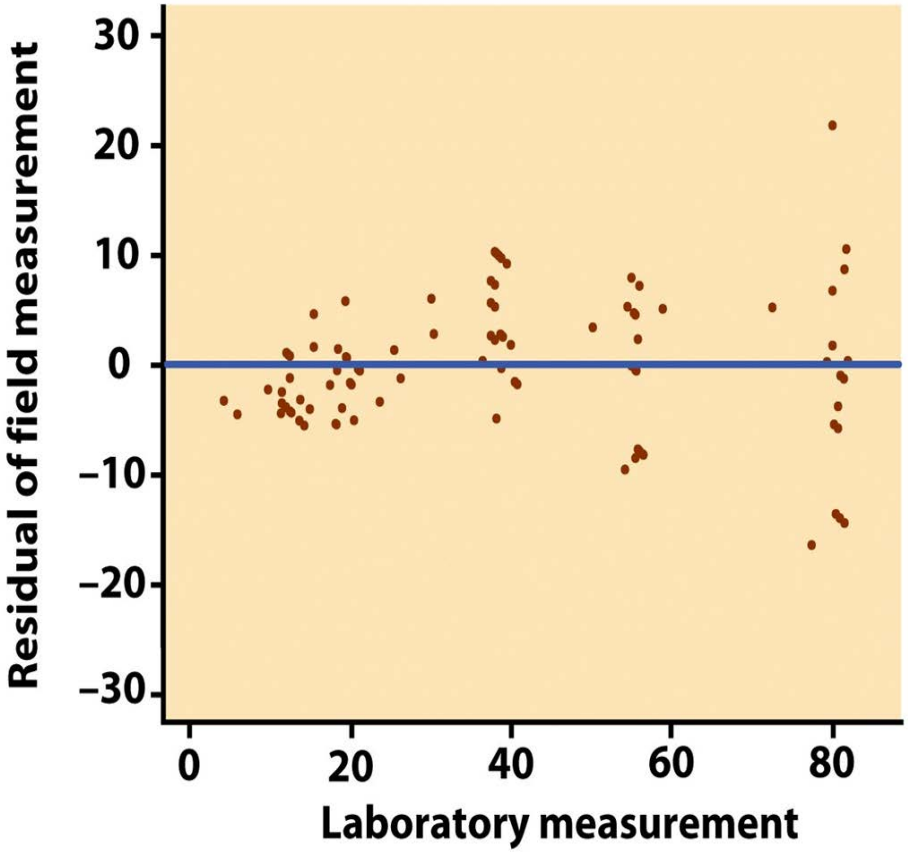
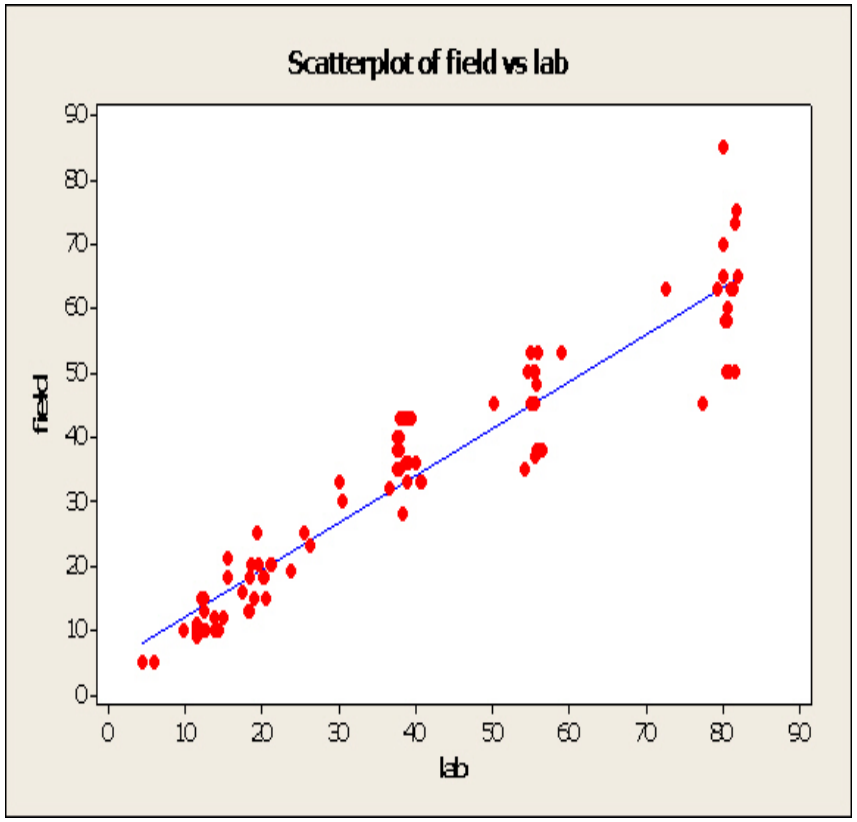


Figure 2-21
Introduction to the Practice of Statistics, Fifth Edition

Ser egentlig det samme i regresjonsplott og residualplott, men residualplott fremhever begrensninger ved regresjonslinjen. Lettere å se hvor store residualene er og om det er mønster



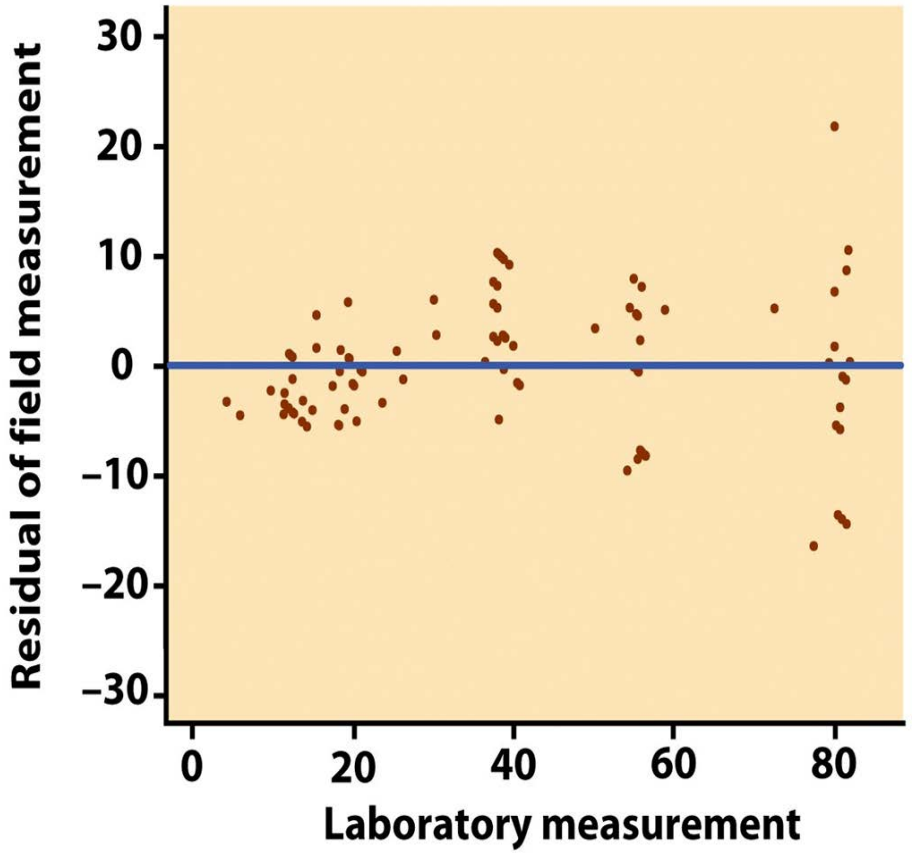
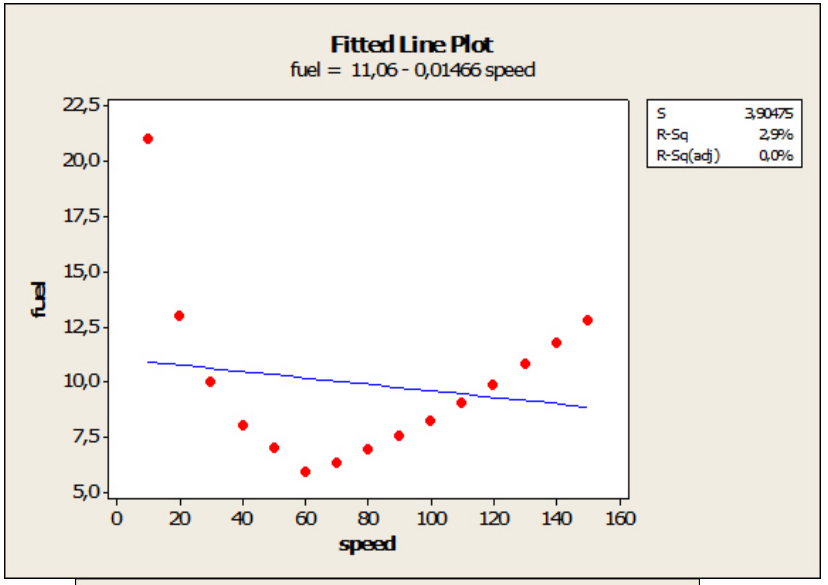
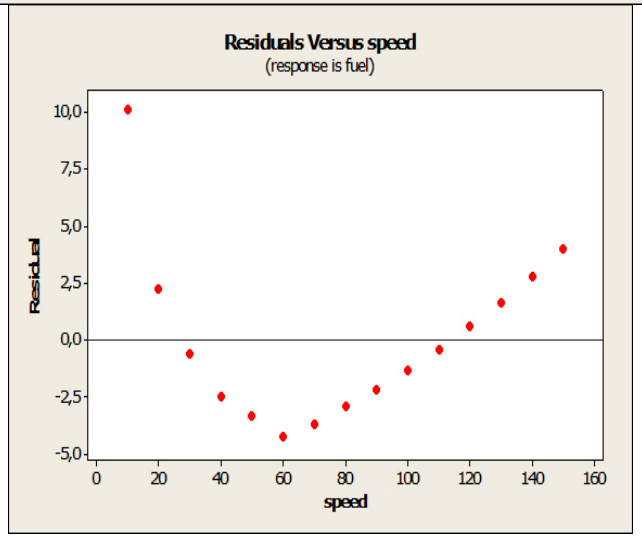


Figure 2-21
Introduction to the Practice of Statistics, Fifth Edition

Residualplott: Eksempel

- Drivstoff-forbruk og hastighet
- Ser fra kryssplott og korrelasjon at det er en veldig svak lineær sammenheng
- Tegner regresjonslinjen og residualplott i Minitab
- Klart mønster i residualplottet
- Det ikke-lineære kurve-mønsteret gjentar seg i residualplottet, for det fanges ikke opp av lineær regresjon





Uteliggere og innflytelsesrike observasjoner

- I tillegg til overordnede mønster, bør kryssplott og residualplott inspiseres for å identifisere uvanlige enkeltpunkter
  - Uteliggere
  - Innflytelsesrike observasjoner

Uvanlige punkter: Eksempel
(HbA måles hos legen, FPG måles hjemme)

TABLE 2.5

Two measures of glucose level in diabetics

Subject	HbA (%)	FPG (mg/ml)	Subject	HbA (%)	FPG (mg/ml)	Subject	HbA (%)	FPG (mg/ml)
1	6.1	141	7	7.5	96	13	10.6	103
2	6.3	158	8	7.7	78	14	10.7	172
3	6.4	112	9	7.9	148	15	10.7	359
4	6.8	153	10	8.7	172	16	11.2	145
5	7.0	134	11	9.4	200	17	13.7	147
6	7.1	95	12	10.4	271	18	19.3	255

Table 2-5
Introduction to the Practice of Statistics, Fifth Edition

Se etter uvanlige enkeltpunkter

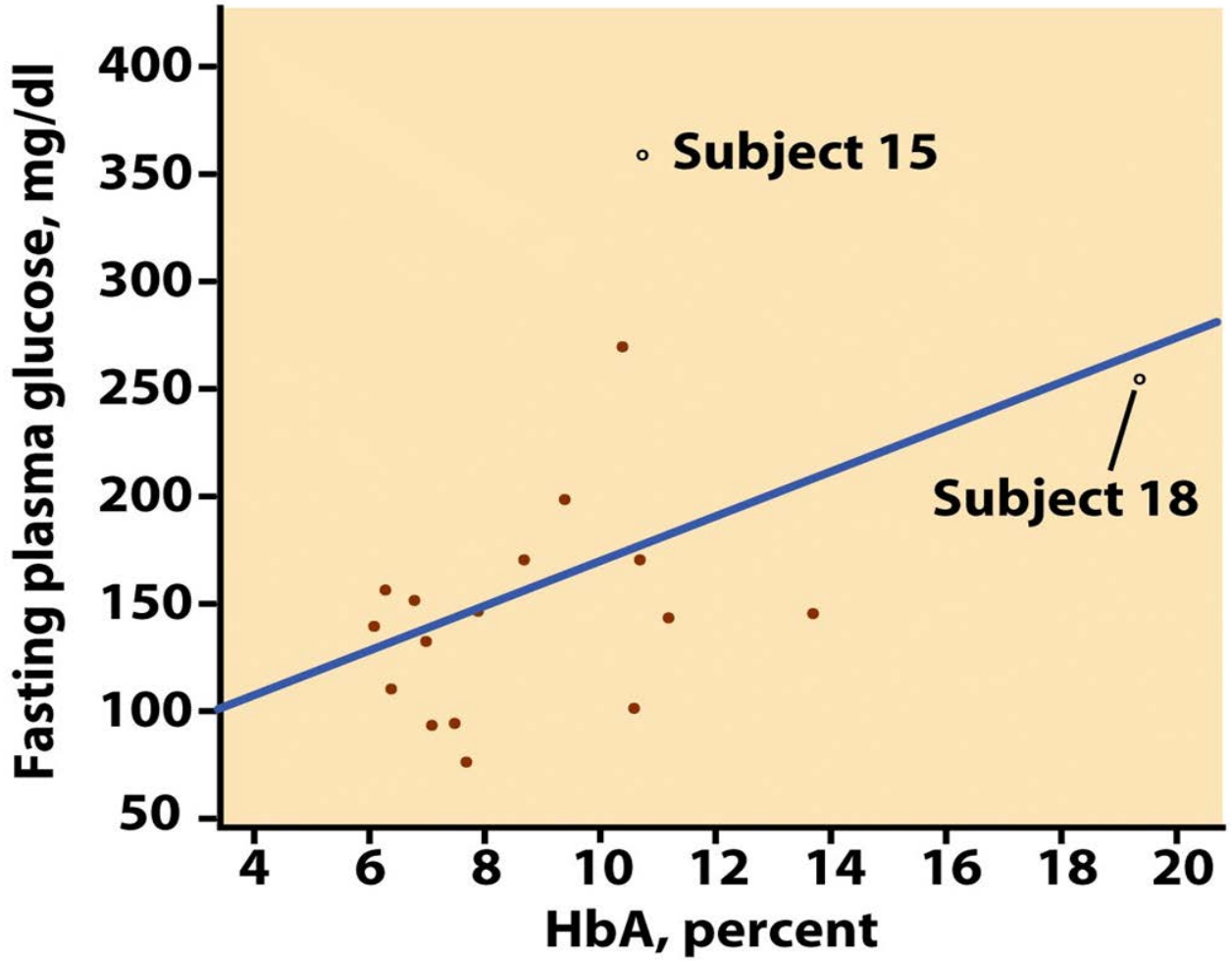


Figure 2-22
Introduction to the Practice of Statistics, Fifth Edition

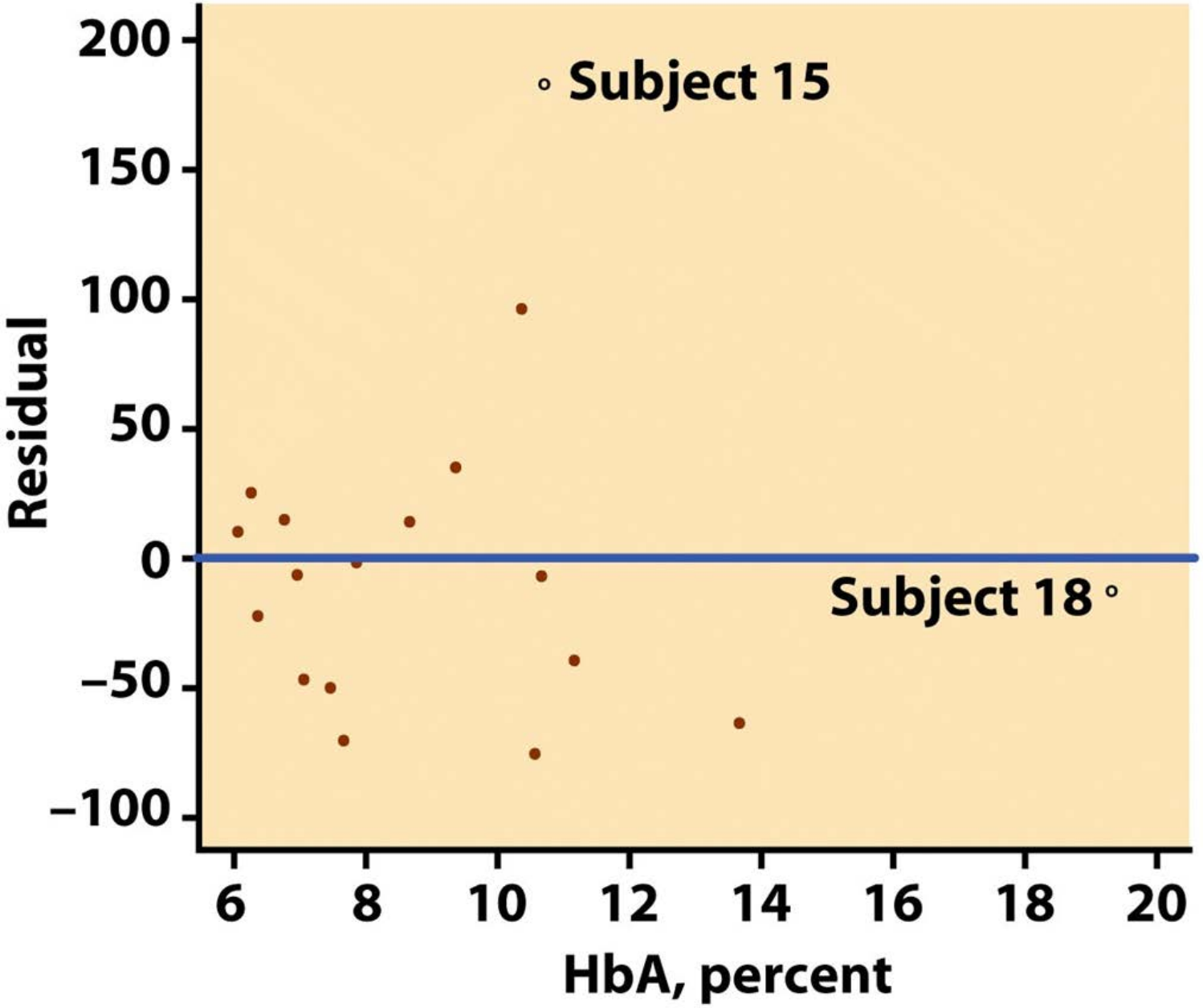


Figure 2-23
Introduction to the Practice of Statistics, Fifth Edition

Uteliggere og innflytelsesrike punkter

- En uteligger er et punkt som ligger utenfor det overordnede mønster av observasjoner
- Punkter som er uteliggere i *y* retningen (individ 15) har store residualer, mens uteliggere i *x* retningen (individ 18) behøver ikke å ha store residualer
- En observasjon er innflytelsesrik hvis fjerning av den resulterer i en klar endring av resultatene
- Punkter som er uteliggere i *x* retningen er ofte innflytelsesrike

Svak lineær sammenheng: r=0.4819

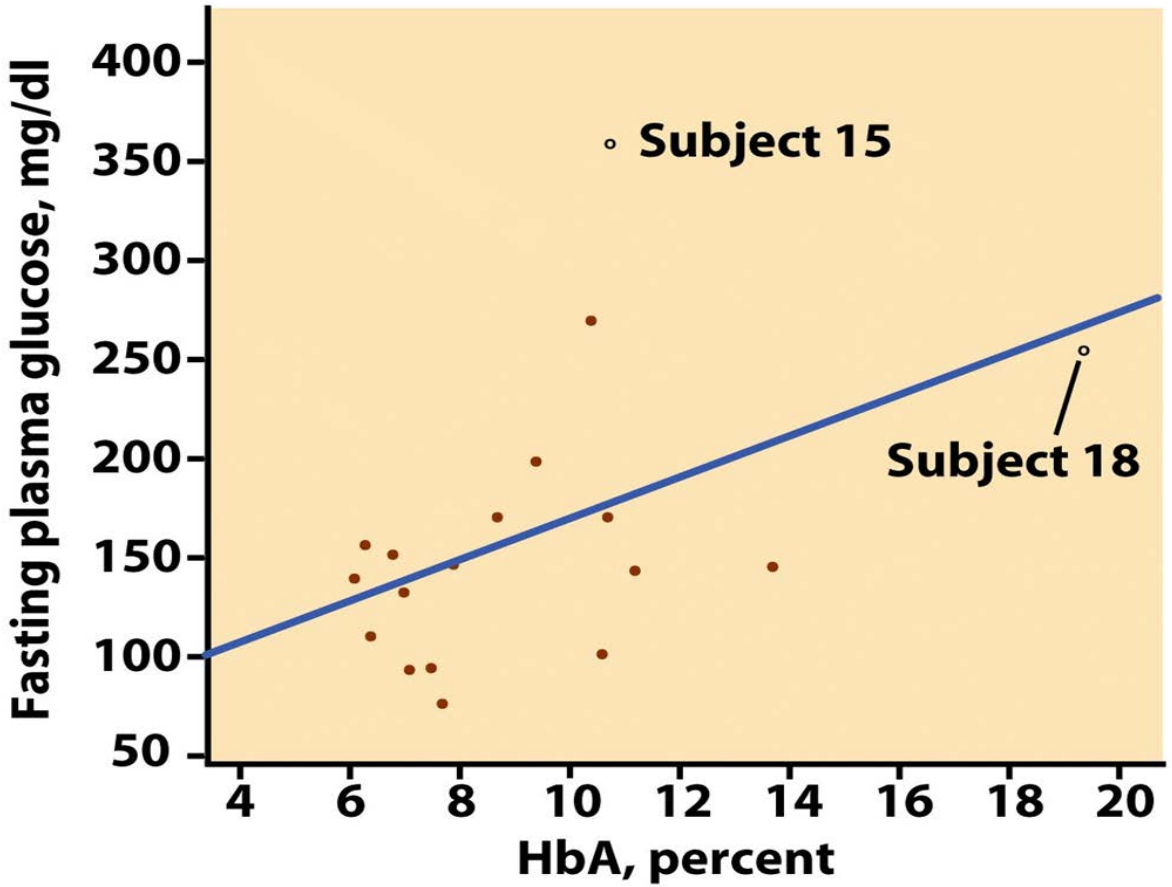


Figure 2-22
Introduction to the Practice of Statistics, Fifth Edition

Innflytelsesrike observasjoner

- Vanskelig å vite hvor innflytelsesrik en observasjon er uten å gjøre regresjonen både med og uten observasjonen
- Et punkt som er en uteligger i x-retning er ofte innflytelsesrikt, men dersom det ligger nær regresjonslinjen beregnet når dette punktet er utelatt, vil det ha liten innflytelse om det tas med i regresjonen
- Et punkt som er uteligger i y-retning kan være innflytelserikt dersom det ikke er mange punkter med lignende x-verdier som “holder linjen på plass”

- Observasjon 15 er uteligger i y-retning. Gjør den lineære sammenhengen svakere (korrelasjonen blir større uten)
- Observasjon 18 er uteligger i x-retning. Gjør den lineære sammenhengen sterkere (korrelasjonen blir mindre uten)
- Observasjon 15 er litt mer innflytelsesrik enn observasjon 18 (se regresjonslinjene)

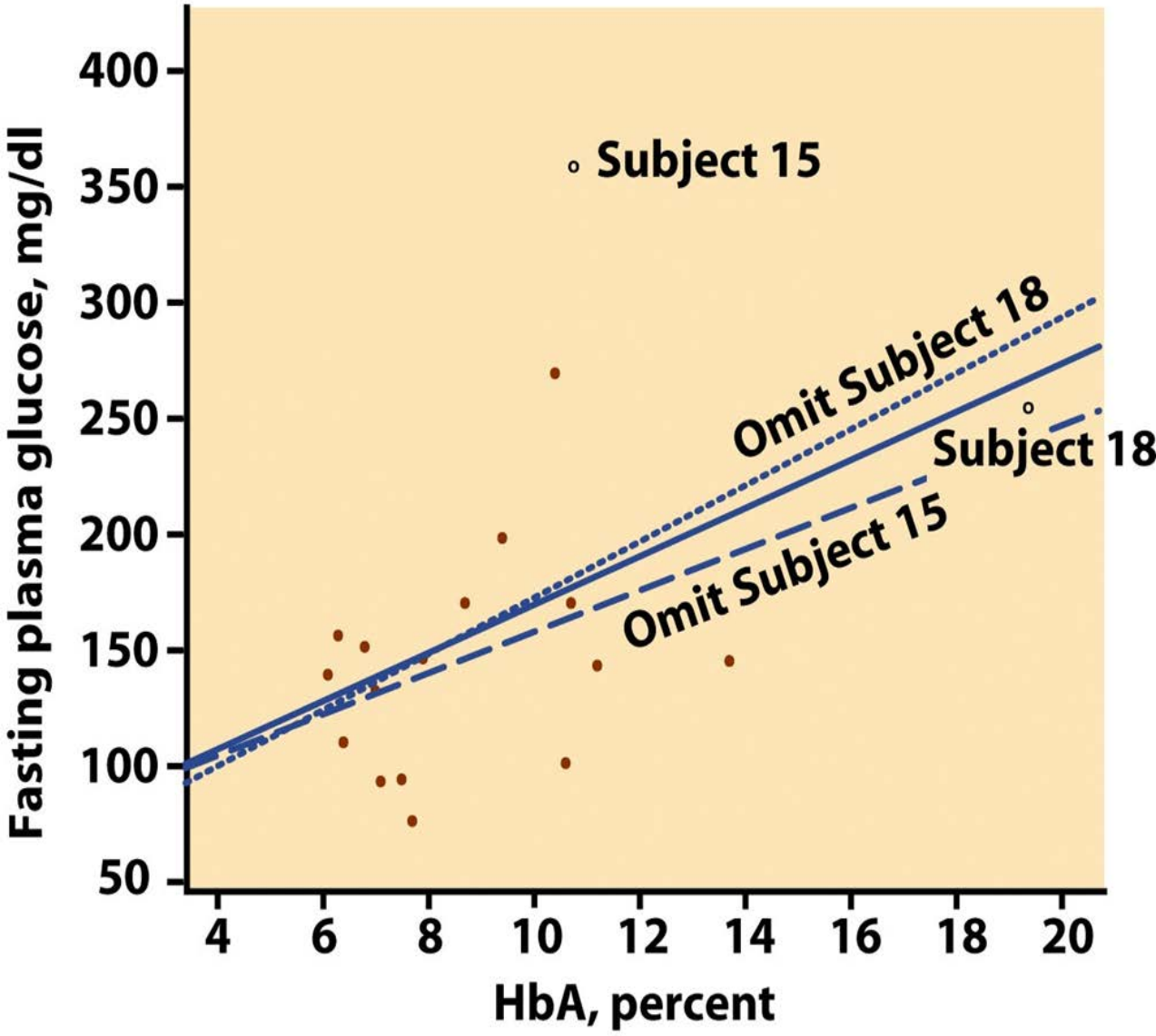


Figure 2-24
Introduction to the Practice of Statistics, Fifth Edition

$$r_{-15} = 0.5684\,,\; r_{-18} = 0.3837$$

Vær varsom!

- Korrelasjon måler kun lineær sammenheng
- Ekstrapolering (bruke modellen utenfor området en har data) gir ofte upålitelige prediksjoner
- Korrelasjon og minste kvadraters linjer er lite robuste
  - plott alltid dataene og se etter potensielt innflytelsesrike punkter

Underliggende variabel

- En underliggende variabel er en variabel som ikke er blant forklarings- eller respons-variablene, men som kan ha innflytelse på tolkningen av sammenhengen mellom disse variablene (*boken: lurking variable*)
  - Kan gjøre en korrelasjon eller en regresjon misvisende

Underliggende variabel: Eksempel

- Stor korrelasjon mellom hvor mye matte studenter tar på high school og suksess senere
- Betyr det at matte er nøkkelen til suksess?
- Mulig underliggende variabel: Familieforhold
- Velutdannede foreldre
  - Legger vekt på utdanning
  - Kan betale utdanning
  - Påvirker hvor mye matte man tar på high school

Underliggende variabel: Eksempel

- Kryssplott av verdien av varer importert inn til USA hvert år mellom 1990 og 2001 (forklaringsvariabel x) mot privat pengebruk på helse i de samme årene (responsvariabel y)
- Sterk lineær sammenheng
- $r=0.9749$, $r^2=0.9504$, dvs regresjonen forklarer 95% av variasjonen i y
- Ingen økonomisk sammenheng mellom x og y
- Sammenheng bare pga at begge variablene øker år for år, årstall er underliggende variabel
- Kryssplott og korrelasjon er korrekte, men hjelper oss ikke med å forstå virkeligheten

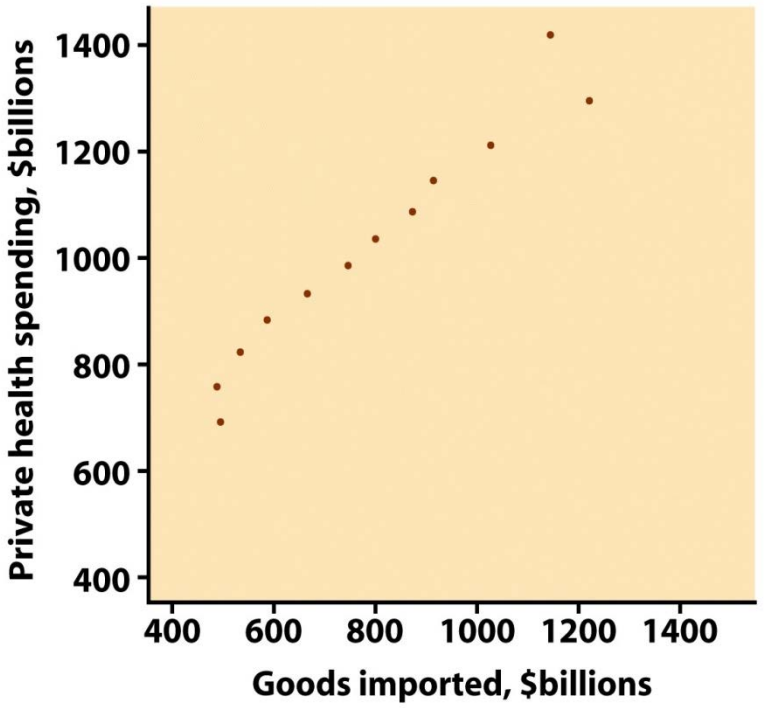


Figure 2-25
Introduction to the Practice of Statistics, Fifth Edition

Sammenheng og kausalitet

- Korrelasjonen i forrige eksempel er reell, men har ingen nyttig fortolkning:
  - variablene har ingen direkte sammenheng
  - verdien av den ene påvirker *ikke* verdien av den ande
- Sammenheng impliserer altså *ikke* kausalitet
- Andre underliggende variable kan forklare sammenhengen
- Kan også skjule faktisk sammenheng

Underliggende variabel skjuler sammenheng

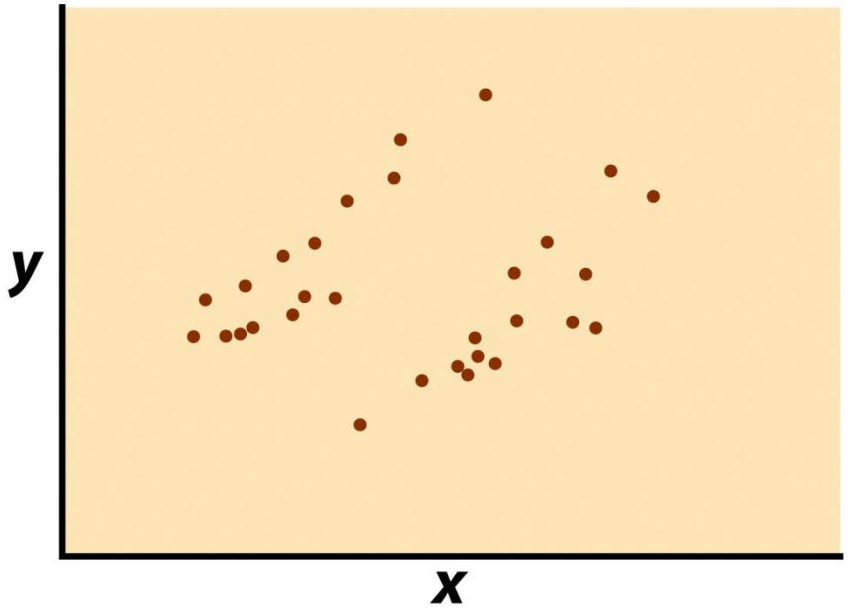


Figure 2-26
Introduction to the Practice of Statistics, Fifth Edition

Viktig å plotte data!!!