Inferens for regresjon

10.1 Enkel lineær regresjon
11.1-11.2 Multippel regresjon

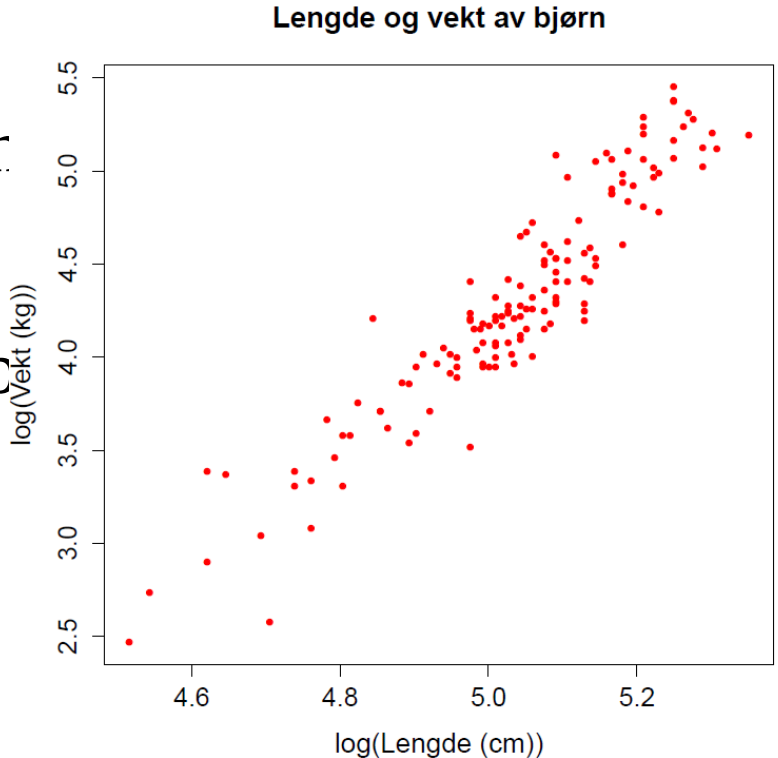
Denne uken: Enkel lineær regresjon

- Litt repetisjon fra kapittel 2
- Statistisk modell for enkel lineær regresjon
- Estimering av modell-parametre
- Konfidensintervall og hypotestetesting for modell-parametre
- Prediksjonsintervall

Respons- og forklaringsvariable

- Ofte har variable ulike roller i en studie
- Responsvariabel måler utfall av en studie
- En forklaringsvariabel brukes til å forklare endringer i en responsvariabel
- Forklaringsvariabelen kar

slik brukes til å forklare variasjonen i responsvariab



Regresjonslinje

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

$$y=b_0+b_1x$$

b_1 kalles stigningstallet, mengden *y* endrer seg når *x* endrer seg med en enhet. b_0 kalles skjæringspunktet, verdien *y* tar for $x=0$

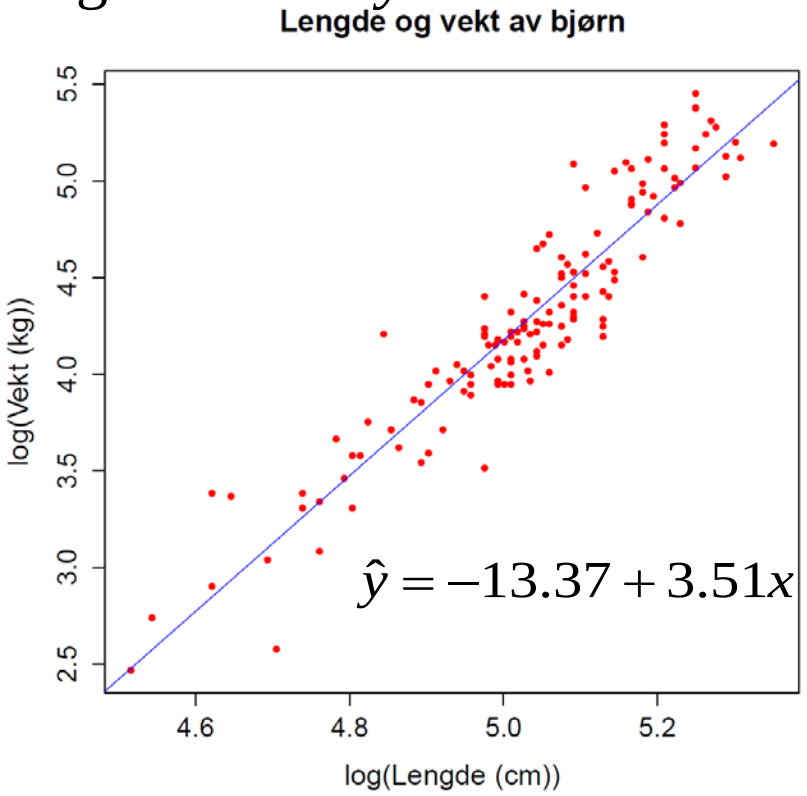
Minste kvadraters regresjon

- Hvordan finne «beste» b_0 og b_1 fra data?
- Ingen linje vil gi perfekt tilpasning
- Ønsker *vertikal* avstand mellom linje og observert y verdi minst mulig
- Minste kvadraters regresjonslinje:

Linjen som gjør kvadratsummen av vertikale avstander minst mulig

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

$$\sum (error)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$



Likninger for minste kvadraters regresjonslinje

Regresjonslinje $\hat{y} = b_0 + b_1 x$

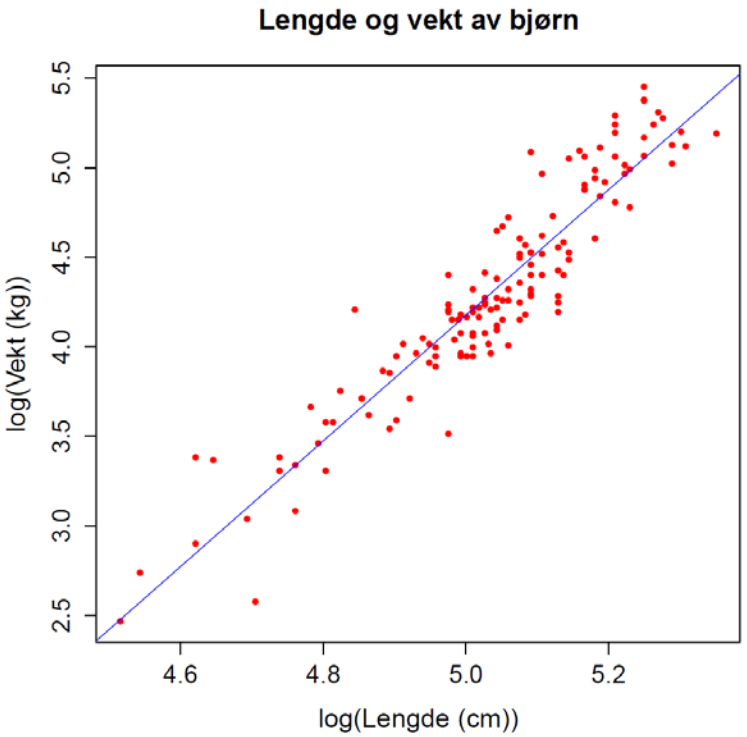
med stigningstall $b_1 = r\, s_y / s_x$

og skjæringspunkt $b_0 = \bar{y} - b_1 \bar{x}$

Skriver $\hat{y}$ ("y – hat")...

Regresjon og korrelasjon

- 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*)



Residualer

- Residualer er differensen mellom observert verdi og predikert verdi:

residual=observert y - predikert y= $y-\hat{y}=y-(b_0+b_1 x)$

- "Resten", det vi ikke har forklart ved forklaringsvariabelen gjennom regresjonslinjen
- Residual for hver observasjon: $e_i=y_i-(b_0+b_1 x_i)$

10.1 Inferens for enkel lineær regresjon

Data i et scatterplot er **tilfeldig utvalg** fra en populasjon med en linær sammenheng mellom *x* og *y*. Et annet utvalg ➔ litt annet scatterplot

Det er egentlig **populasjonsforventningen** μ_y vi modellerer som en lineær funksjon av *x:* $\mu_y = \beta_0 + \beta_1 x$.

Vi skal nå finne ut om den **observerte sammenhengen** er **statistisk signifikant** (og ikke et resultat av tilfeldigheter).

Sammenligning av respons på to behandlinger: Forventet respons varierer med type behandling

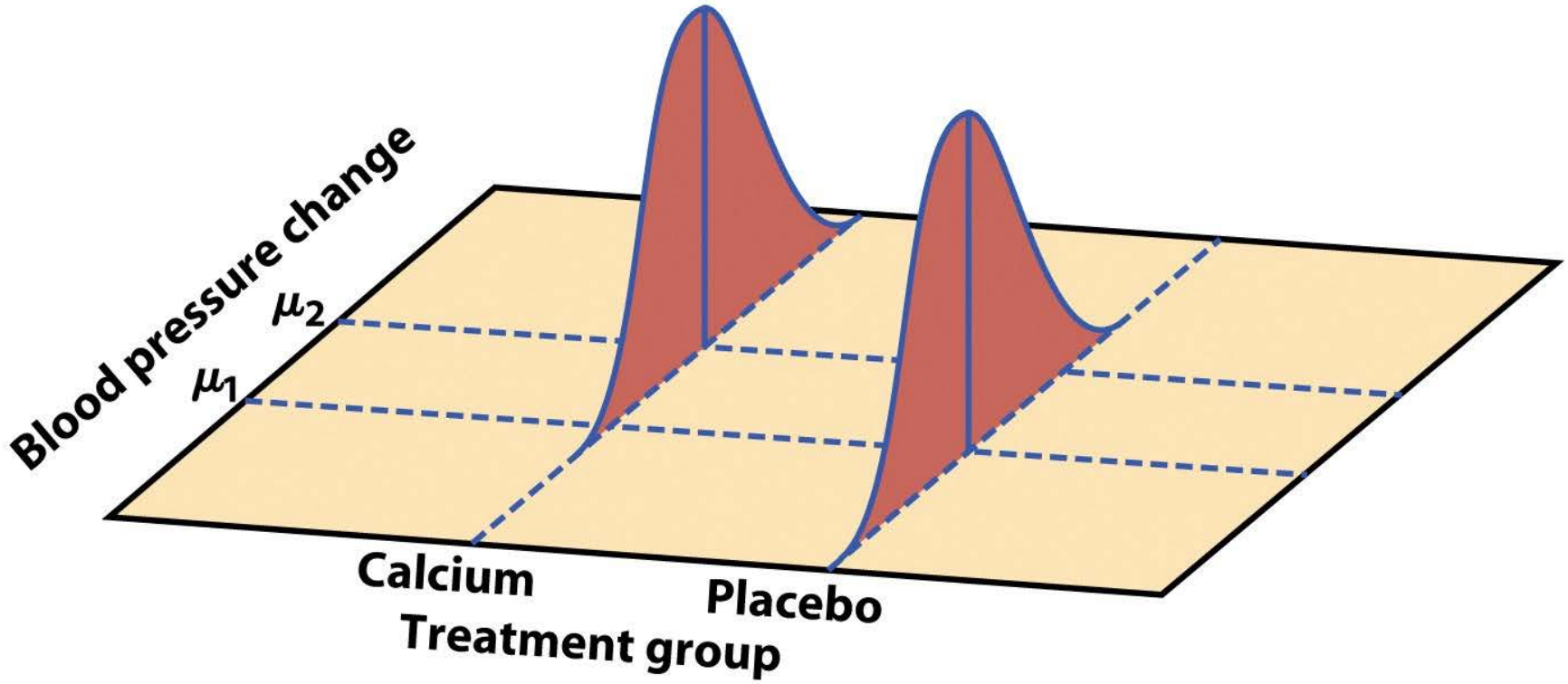
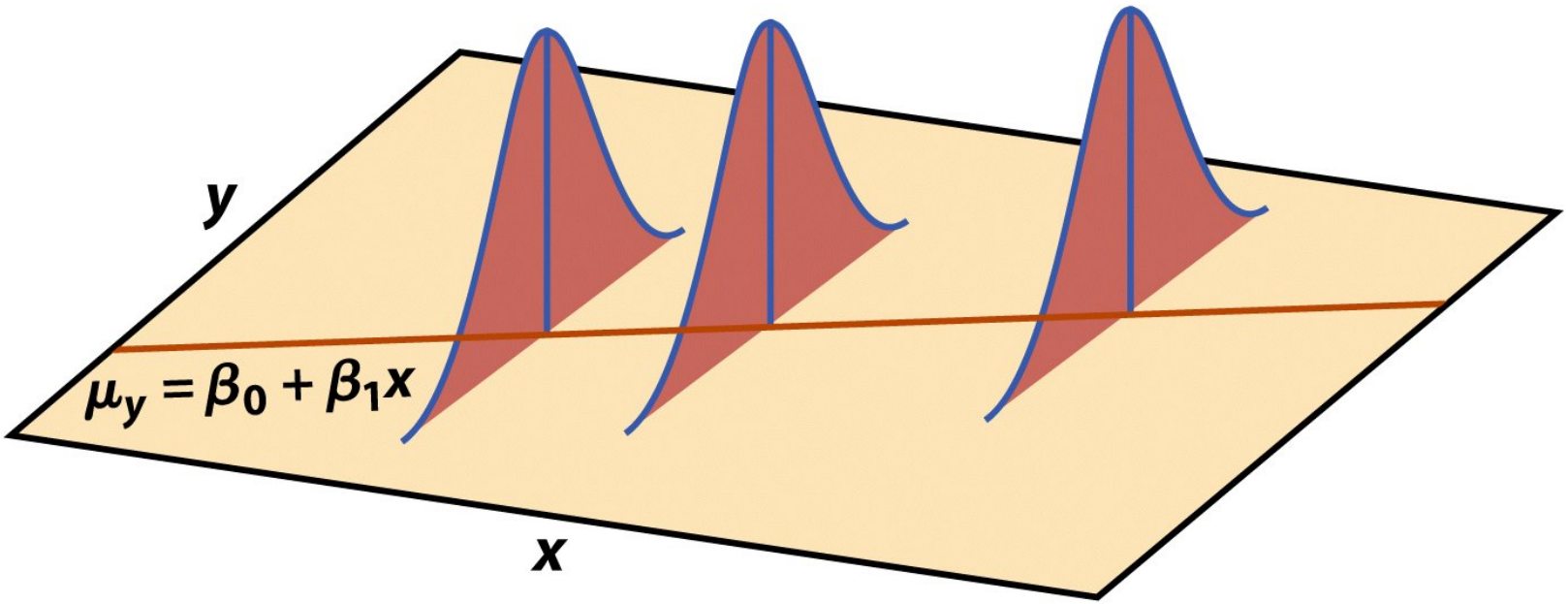


Figure 10-1
Introduction to the Practice of Statistics, Fifth Edition

- Regresjon: Sub-populasjoner, en for hver verdi av forklaringsvariabelen *x*
- Forventningen er en rettlinjet funksjon av *x*
- Observerte *y*-er for gitt verdi av *x* vil variere rundt denne forventningen gitt av *x*:

  Data = Forventning gitt av linjen + Residual
- Modellen antar at denne variasjonen rundt linjen, målt ved standardavviket σ, er den samme for alle verdier av *x*



•Statistisk modell for lineær regresjon

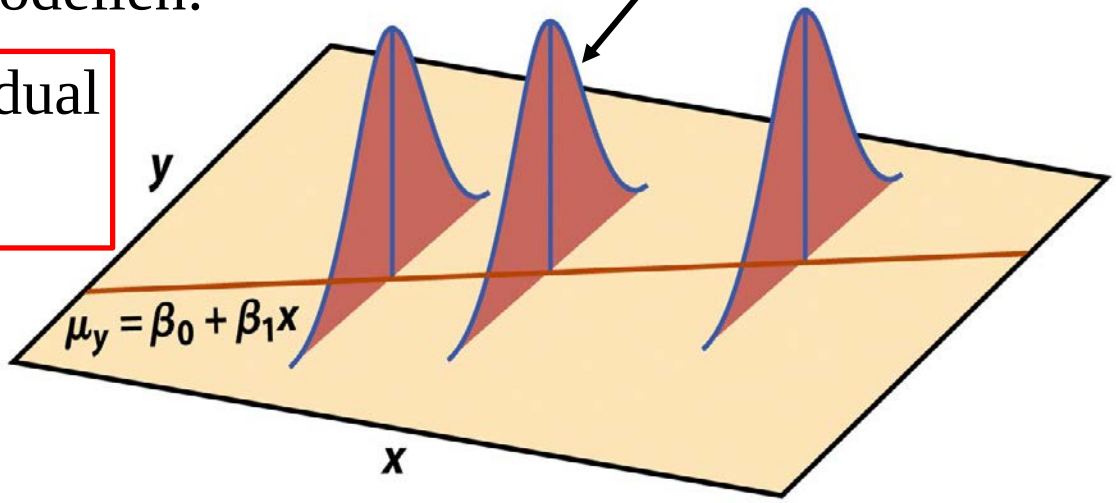
I <u>populasjonen</u> er den lineære regresjonsligningen

$\mu_y = \beta_0 + \beta_1 x.$

For enhver gitt verdi x, følger responsen y en normalfordeling med standardavvik σ

<u>Data</u> er observajoner fra modellen:

Data	=	fit	+	residual
y_i	=	$(\beta_0 + \beta_1 x_i)$	+	ε_i



hvor ε_i er **uavhengige**

og **Normalfordelte** $N(0,\sigma)$.

Enkel lineær regresjon antar **lik varians for alle y (homoskedasitet =** σ er den samme for alle verdier av x).

Modell for enkel lineær regresjon

- Modell: $\mu_y=\beta_0+\beta_1 x$
- Vil forvente variasjoner rundt μ_y
- Har *n* parvise observasjoner av forklarings- og responsvariablene $(x_1,y_1),...,(x_n,y_n)$
- Modell for y_i: $y_i= \beta_0+\beta_1 x_i + \varepsilon_i$
- ε_i antas uavhengige og $N(0,\sigma)$-fordelte
- 3 ukjente parametre: β_0, β_1, σ
- *Statistisk modell for enkel lineær regresjon*

Sammenheng med minste kvadraters metode

- Observasjoner $(x_1,y_1),...,(x_n,y_n)$

- Regresjonslinje $y = b_0 + b_1x$
  - Minimere $\sum(\text{error})^2=\sum(y_i-b_0-b_1x_i)^2$
  - $b_1=rs_y/s_x$, $b_0=\bar{y}-b_1\bar{x}$

- Modell for forventningen til *y* gitt *x:* $\mu_y=\beta_0+\beta_1x$

- b_0 er estimat for β_0 , b_1 er estimat for β_1

Inferens

- Inferens om ukjente størrelser

  + estimering av σ

  – Stigningstall β_1
  – Skjæringspukt β_0
  – Forventet respons μ_y for gitt verdi x
  – Individuell fremtidig respons y for gitt verdi x
- Antakelser:
  – observasjonene er uavhengige
  – Sammenhengen er lineær
  – y er normalfordelt rundt sin forventning
  – Variansen til y er konstant

Estimering

- Modell: $\mu_y=\beta_0+\beta_1 x$
- Minste-kvadraters estimater: $b_1=rs_y/s_x$, $b_0=\bar{y}-b_1\bar{x}$
- b_0 og b_1 er *forventningsrette* estimater for β_0 og β_1
- b_0 og b_1 er normalfordelte hvis ε_i er N(0,σ)
  - Tilnærmet normalfordelte generelt

- *Forventningsrett* estimat for μ_y for $x = x^*$: $\hat{\mu}_y = b_0 + b_1 x^*$

- *Prediksjon* av respons y for $x = x^*$: $\hat{y} = b_0 + b_1 x^*$

Residualer

e_i = observert respons − predikert respons

$= y_i - \hat{y}_i$

$= y_i - b_0 - b_1 x_i$

- e_i er et anslag på ϵ_i
- $\mu_{\varepsilon_i} = 0$ Sum av e_i er 0

- Brukes til estimering av σ og sjekk av modell

Estimering av σ

- Estimat for σ^2: $s^2 = \frac{\sum e_i^2}{n-2}$
- Deler på $n-2$ for å gjøre s^2 forventningsrett
- $n-2$ kalles antall frihetsgrader for s^2

- Estimererer σ ved $s = \sqrt{s^2}$

Eksempel: Sammenhengen mellom drivstoff-forbruk målt ved «miles per gallon» (MPG) og hastighet målt av «miles per hour» (MPH), 60 observasjoner

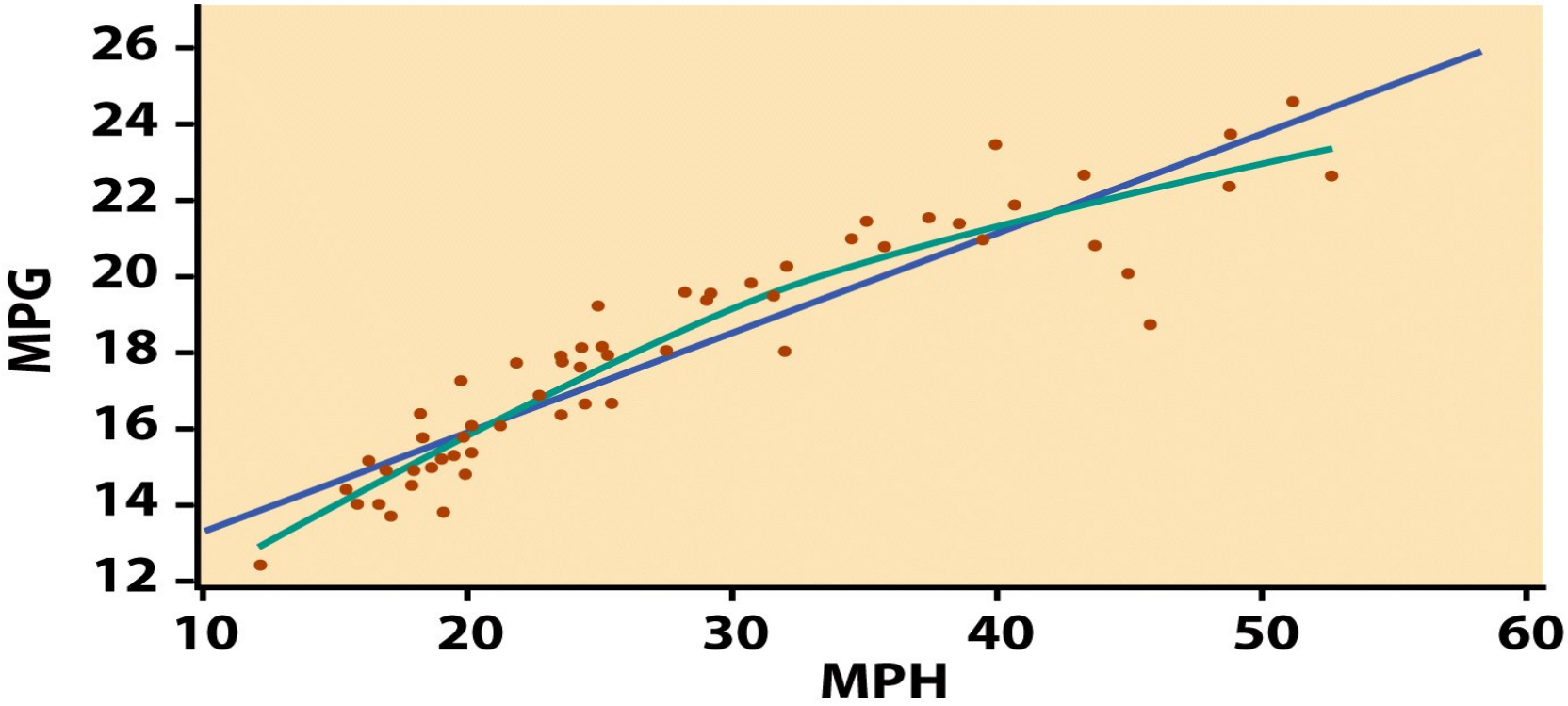
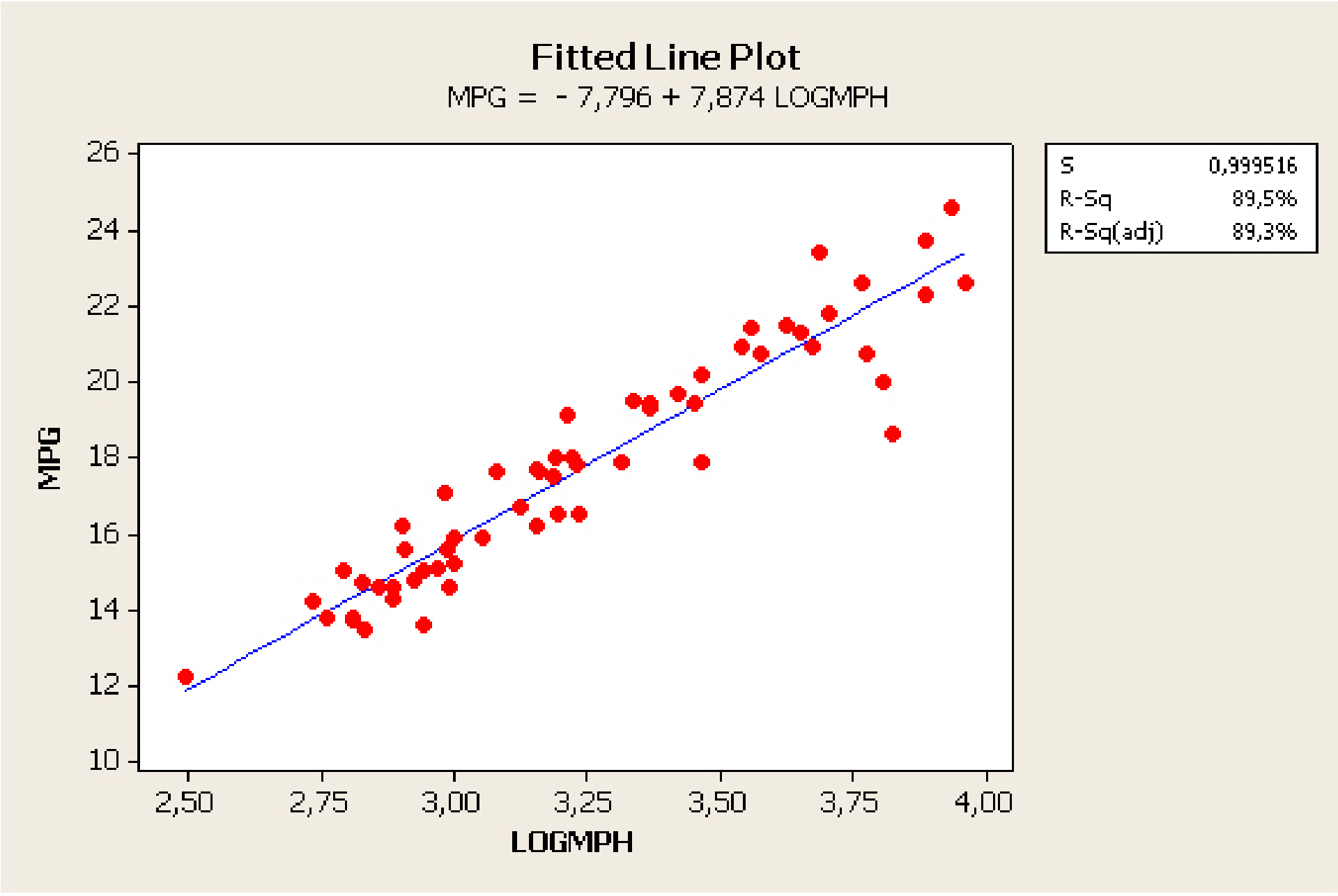


Figure 10-3
Introduction to the Practice of Statistics, Fifth Edition

MPG versus MPH

- Start alltid med *grafisk fremstilling* av data
- Eksempel indikerer *ikke-lineær* sammenheng
- Kan få lineær sammenheng ved *transformasjoner*
- I eksempel: log (ln) transformasjon av MPH


Fitted Line Plot
MPG = - 7,796 + 7,874 LOGMPH
S 0,999516
R-Sq 89,5%
R-Sq(adj) 89,3%
MPG
LOGMPH
26
24
22
20
18
16
14
12
10
2,50
2,75
3,00
3,25
3,50
3,75
4,00

Eksempelet analysert i Minitab

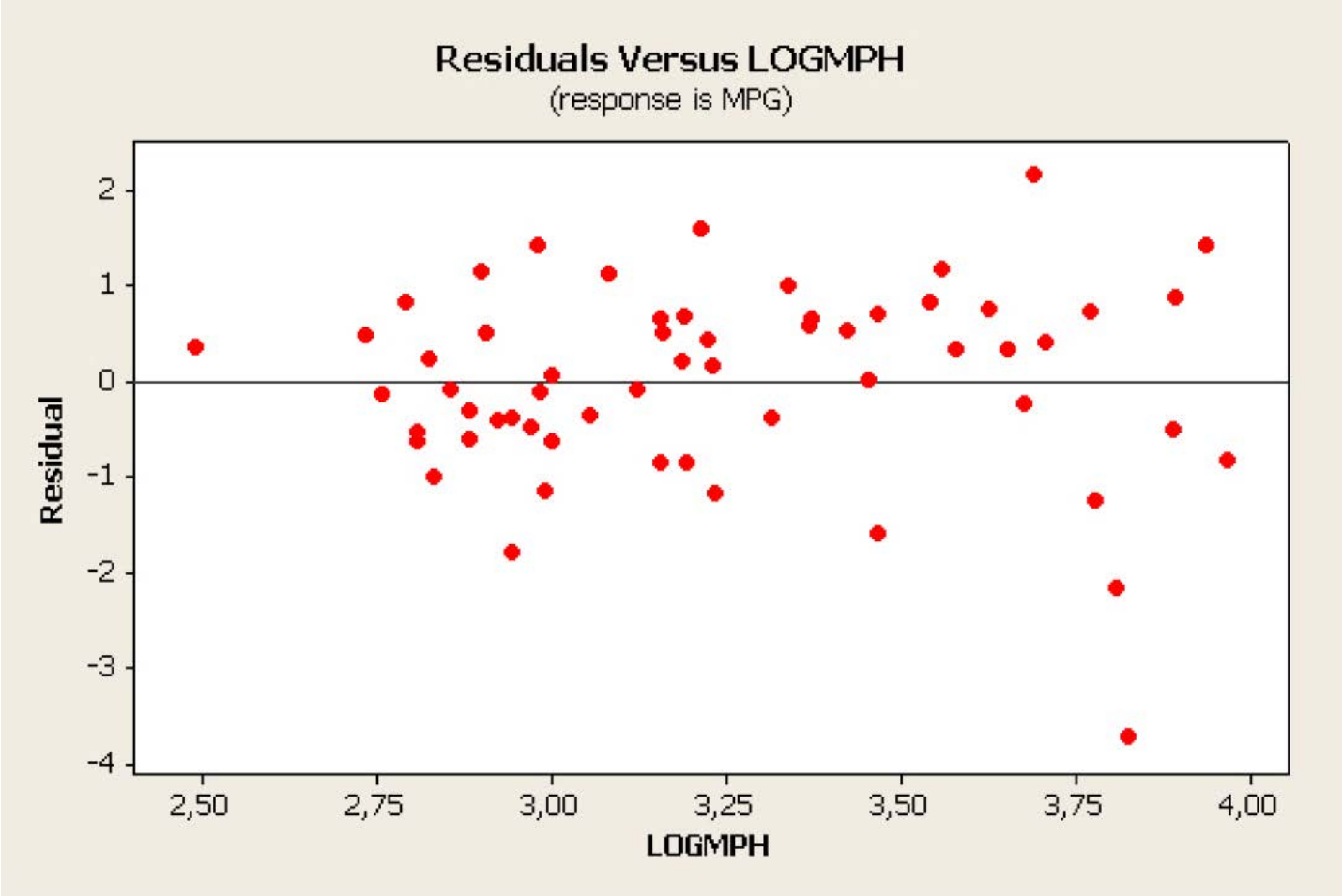
```
The regression equation is
MPG = - 7.80 + 7.87 logmph

Predictor          Coef        StDev            T          P
Constant         -7.796        1.155        -6.75      0.000
logmph           7.8742       0.3541        22.24      0.000

S = 0.9995       R-Sq = 89.5%      R-Sq(adj) = 89.3%
```

Figure 10-5b
Introduction to the Practice of Statistics, Fifth Edition

Residualplott- mot forklaringsvariabelen



Residualplott - kvantilplott

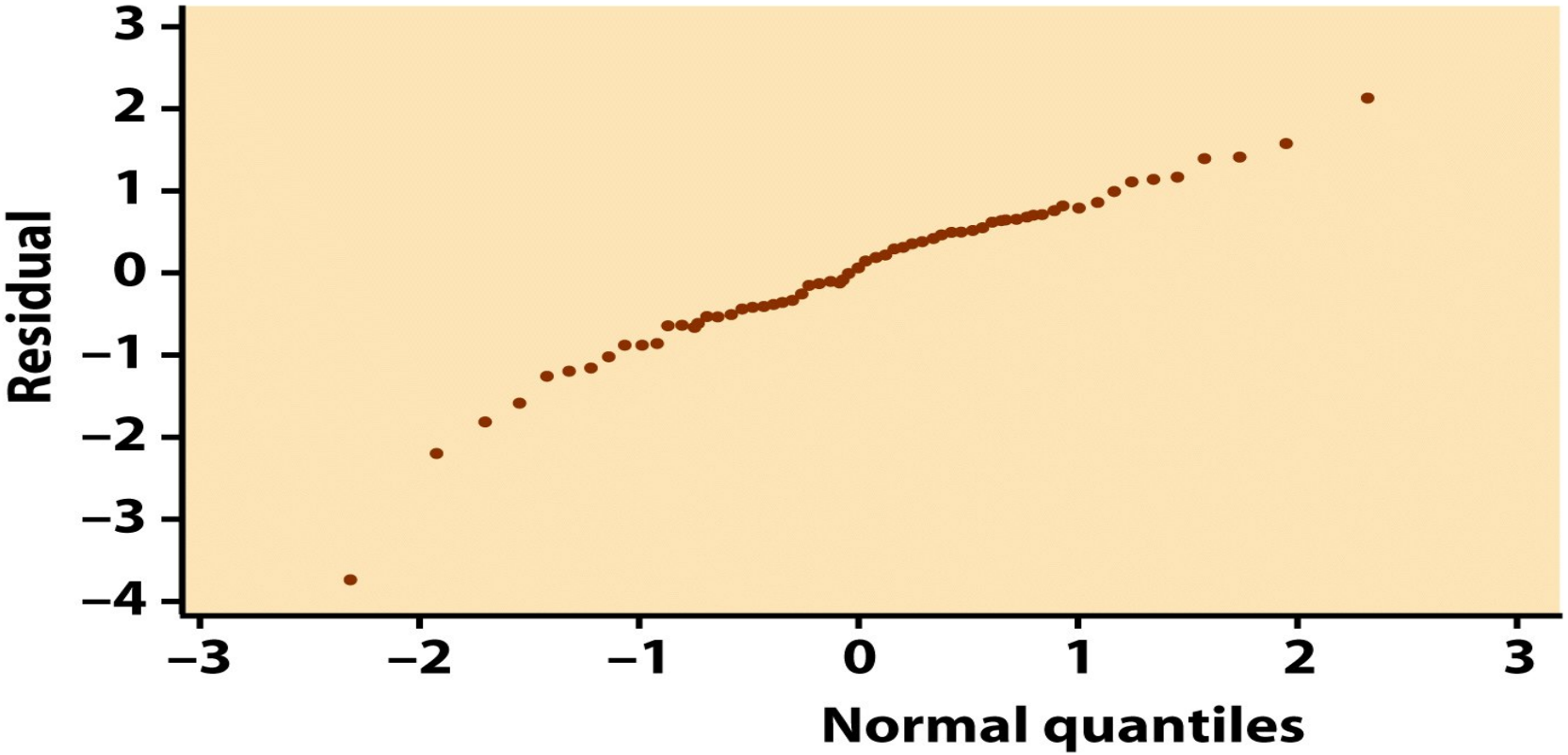


Figure 10-8
Introduction to the Practice of Statistics, Fifth Edition

Konfidensintervall

Estimatorene $\boldsymbol{b_0}$ og $\boldsymbol{b_1}$ for regresjonskoeffisientene er normalfordelte med forventning β_0, β_1. Vi må estimere variansene deres

➔Vi bruker t-fordeling med ***n* – 2 frihetsgrader.**

b_1 har estimert standardavvik SE_{b1}

b_0 har estimert standardavvik SE_{b0}

$$\text{Estimat} \mp t^* \, SE_{\text{estimat}}$$

t er kritisk verdi i t(n – 2)-fordelingen med areal C mellom –t* og +t*.*

CONFIDENCE INTERVALS AND SIGNIFICANCE TESTS FOR REGRESSION SLOPE AND INTERCEPT

A **level *C* confidence interval for the intercept** β_0 is

$$b_0 \pm t^* \mathrm{SE}_{b_0}$$

A **level *C* confidence interval for the slope** β_1 is

$$b_1 \pm t^* \mathrm{SE}_{b_1}$$

In these expressions t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

- SE_{b0} og SE_{b1} avhenger blant annet av *s*
- Dere finner dem ved bruk av dataprogram

Konfidensintervall – eksempel

n=60

```
The regression equation is
MPG = - 7.80 + 7.87 logmph

Predictor         Coef        StDev            T          P
Constant        -7.796        1.155        -6.75      0.000
logmph          7.8742       0.3541        22.24      0.000

S = 0.9995       R-Sq = 89.5%      R-Sq(adj) = 89.3%
```

Figure 10-5b
Introduction to the Practice of Statistics, Fifth Edition

Konfidensintervall for stignigstall β_1 for hånd- eksempel

Signifikanstester

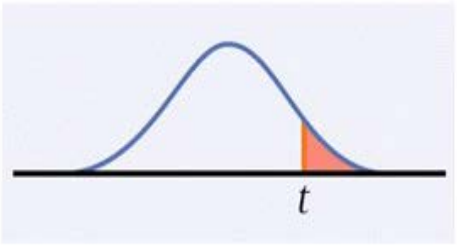
- Ofte: Ønsker å teste $H_0:\beta_1=0$
  - Svarer til at det ikke er noen sammenheng mellom x og y
  - $\mu_y=\beta_0$
- Testobservator $$t=\frac{b_1}{\mathrm{SE}_{b_1}}$$
- Software tester også: $H_0:\beta_0=0$
  - Sjeldent av interesse

To test the hypothesis $H_0: \beta_1 = 0$, compute the **test statistic**

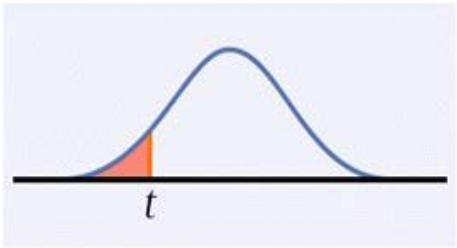
$$t = \frac{b_1}{\text{SE}_{b_1}}$$

The **degrees of freedom** are $n - 2$. In terms of a random variable T having the $t(n-2)$ distribution, the P-value for a test of H_0 against

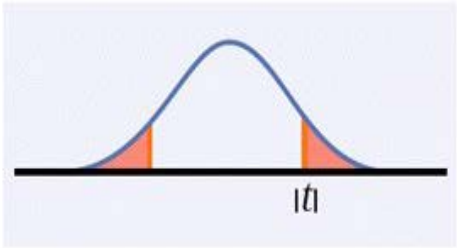
$H_a: \beta_1 > 0$ is $P(T \geq t)$



$H_a: \beta_1 < 0$ is $P(T \leq t)$



$H_a: \beta_1 \neq 0$ is $2P(T \geq |t|)$



Definition, pg 644

Introduction to the Practice of Statistics, Fifth Edition

Signifikanstest for β_1 i Minitab-eksempel

```
The regression equation is
MPG = - 7.80 + 7.87 logmph

Predictor         Coef       StDev          T          P
Constant       -7.796       1.155      -6.75      0.000
logmph         7.8742      0.3541      22.24      0.000

S = 0.9995      R-Sq = 89.5%      R-Sq(adj) = 89.3%
```

Figure 10-5b
Introduction to the Practice of Statistics, Fifth Edition

Signifikanstest for β_1 for hånd - eksempel

Konfidensintervall for forventet respons

For $x = x^*$ er forventet respons: $\mu_y = \beta_0 + \beta_1 x^*$

Estimert forventet respons: $\hat{\mu}_y = b_0 + b_1 x^*$

Konfidensintervall for μ_y: $\hat{\mu}_y \pm t^* \, SE_{\hat{\mu}_y}$

$SE_{\hat{\mu}_y}$ kan fås fra software

t^* (som før) fra t-fordeling med $n-2$ frihetsgrader

Konfidensintervall for forventet respons i Minitab- eksempel

Regression Analysis: MPG versus LOGMPH

The regression equation is
MPG = - 7,80 + 7,87 LOGMPH

Predictor	Coef	SE Coef	T	P
Constant	-7,796	1,155	-6,75	0,000
LOGMPH	7,8742	0,3541	22,24	0,000

S = 0,999516 R-Sq = 89,5% R-Sq(adj) = 89,3%

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	18,976	0,141	(18,694; 19,258)	(16,956; 20,997)

Values of Predictors for New Observations

New Obs	LOGMPH
1	3,40

Konfidensintervall for forventet respons for hånd- eksempel

Konfidensgrenser for forventet respons (stiplede linjer). Viser konfidensintervall for gitt x-verdi

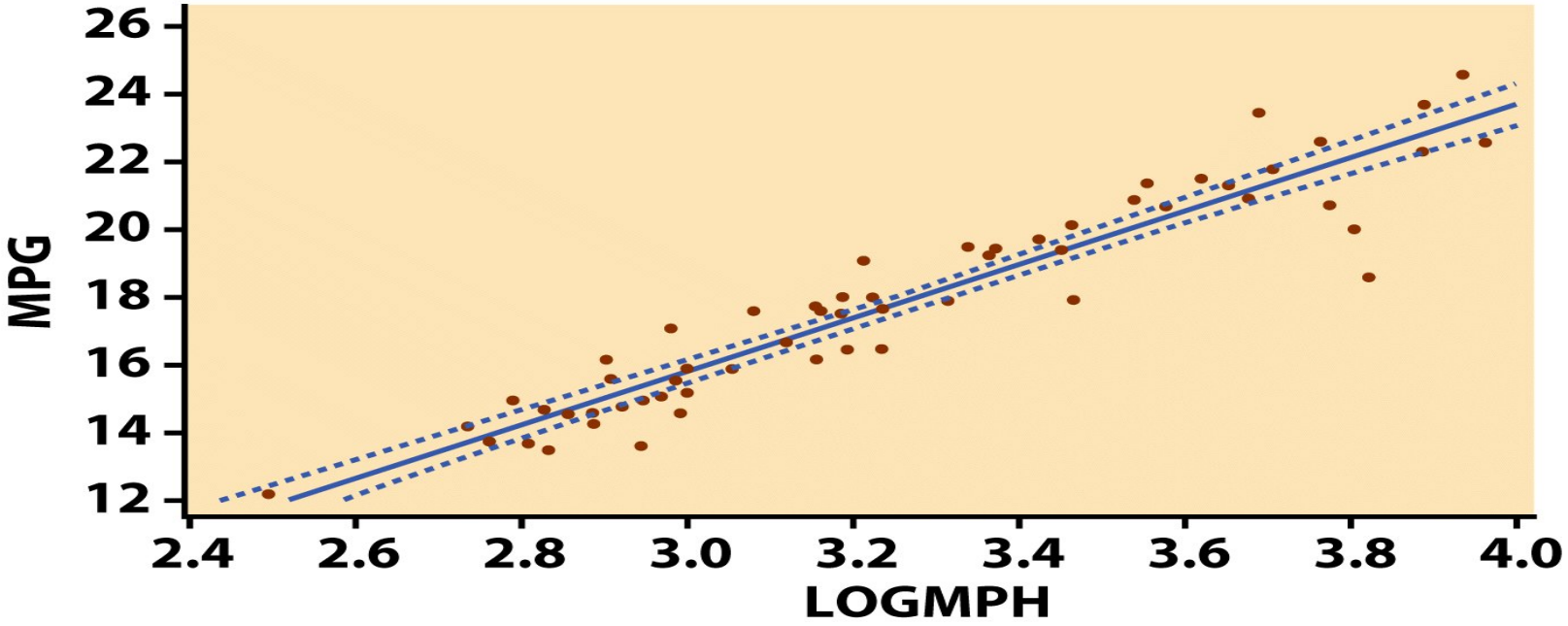


Figure 10-9
Introduction to the Practice of Statistics, Fifth Edition

Hvorfor er konfidensintervallene for x* i midten smalere enn for x* langt fra midten?

Prediksjonsintervall

Konfidensintervall for μ_y gir usikkerhet om forventet respons

Ofte av interesse å si noe om usikkerhet for individuell respons

Prediksjon av respons: $\hat{y} = b_0 + b_1 x^*$

Prediksjonsintervall: $\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}$ (husk CI for μ_y: $\hat{\mu}_y \pm t^* \mathrm{SE}_{\hat{\mu}_y}$)

$$\mu_y = \beta_0 + \beta_1 x^*, \quad y = \beta_0 + \beta_1 x^* + \varepsilon$$

$$\mathrm{SE}_{\hat{y}}^2 = \mathrm{SE}_{\hat{\mu}_y}^2 + s^2, \quad \mathrm{SE}_{\hat{y}} = \sqrt{\mathrm{SE}_{\hat{\mu}_y}^2 + s^2}$$

Formler for SE i kap. 10.2

Større usikkerhet i prediksjon av y enn i estimering av μ_y fra regresjonslinjen for gitt x^* pga ε

$SE^2_{\hat{\mu}_y}$ sier noe om usikkerheten i estimeringen av regresjonslinjen

s^2 tar i tillegg hensyn til variasjonen rundt regresjonslinjen/forventningen

Prediksjonssintervall for forventet respons for hånd- eksempel

Prediksjonsgrenser (stiplede linjer). Viser prediksjonsintervall for gitt x-verdi

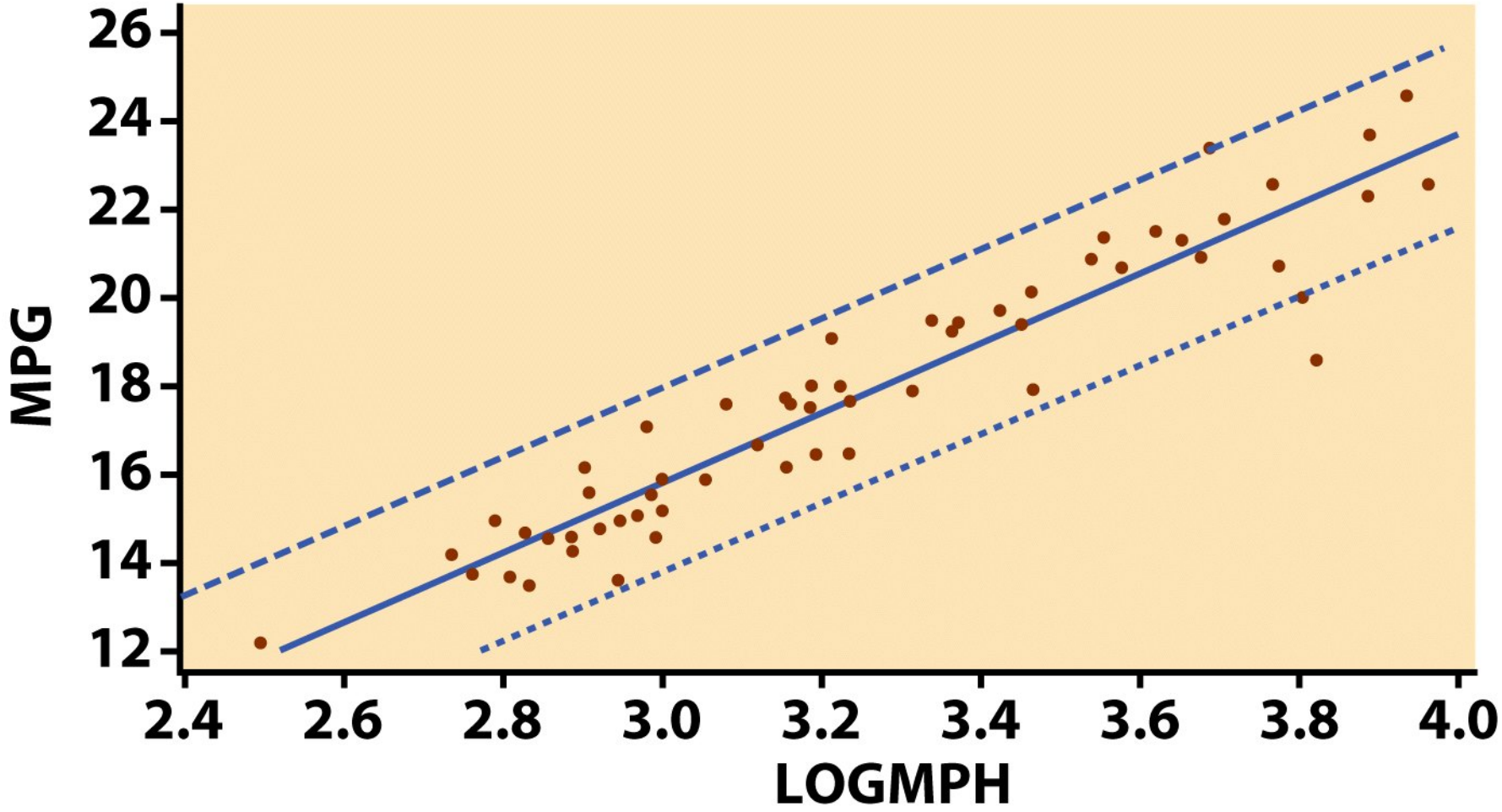
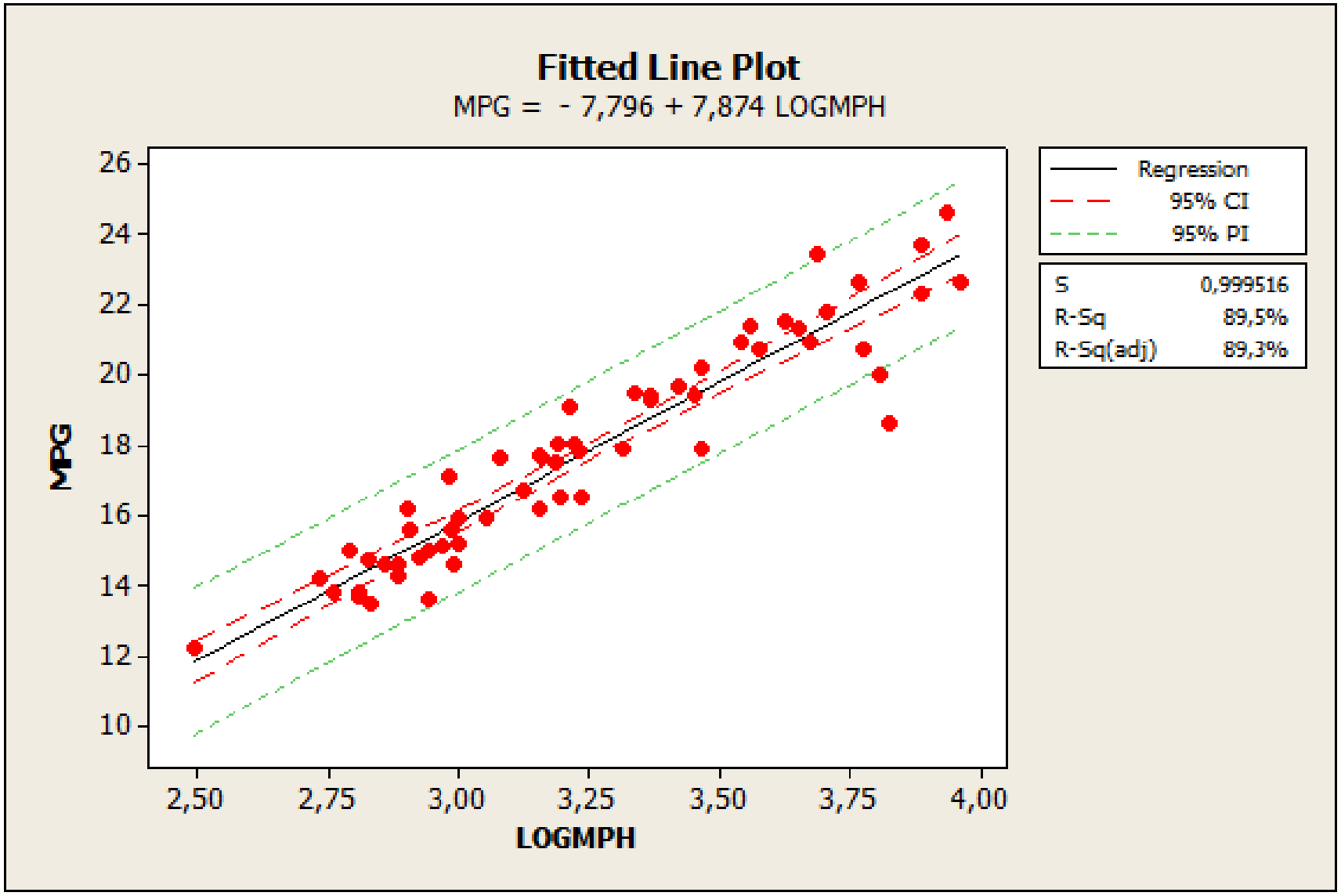


Figure 10-10
Introduction to the Practice of Statistics, Fifth Edition


Fitted Line Plot
MPG = - 7,796 + 7,874 LOGMPH
Regression
95% CI
95% PI
S 0,999516
R-Sq 89,5%
R-Sq(adj) 89,3%
MPG
LOGMPH
26
24
22
20
18
16
14
12
10
2,50
2,75
3,00
3,25
3,50
3,75
4,00