Som i enkel lineær regresjon kan vi for hver hver fast, fiksert verdisett av forklaringsvariablene x_1, x_2,…, x_p tenke oss at finns en **subpopulasjon** av responser. Hver subpopulasjonen karakteriseres ved at fordelingene har en bestemt forventning. Forøvrig er fordelingene like. Spesielt har de samme standardavvik σ.

Eksempel: Suksess i college.

Respons y: GPA, kumulativ «grade point average» etter tre semestre

Forklaringsvariable: x_1 karakter i matematikk på videregående, HSM
x_2 karakter i naturfag på videregående, HSS
x_3 karakter i engelsk på videregående, HSE

som kan gi modeller av typen

$$\mu_{GPA} = \beta_0 + \beta_1\, HSM + \beta_2 HSS + \beta_3 HSE$$

Minitab-utskrift for GPA-data (editert)

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0,726103	22,77%	21,19%	18,01%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	0,069	0,454	0,15	0,879
HSM	0,1232	0,0549	2,25	0,026
HSS	0,1361	0,0700	1,95	0,054
HSE	0,0585	0,0654	0,89	0,373

Regression Equation

GPA = 0,069 + 0,1232 HSM + 0,1361 HSS + 0,0585 HSE

Den **statistiske modellen for multippel lineær regresjon** er

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

for i=1, …,n. Her er forventningen til responsverdien

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

en lineær funksjon av forklaringsvariablene.

Avvikene ϵ_i antas å være uavhengige normalfordelte variable med forventning 0 og standardavvik σ , dvs. N(0,σ) . Parametrene i modellen er β_0, β_1, … ,β_p og σ .

Antagelsen i multippel lineær regresjon, mer detaljert.

Vi kan mer spesifikt skrive opp modellen i 4 punkter:

- Linearitet: $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
- Uavhengighet: Gitt x-ene er $y_1, y_2, \cdots, y_n$ (samt $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$) uavhengige.
- Konstant varians: Gitt x-ene har $y_1, y_2, \cdots, y_n$ (samt $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$) samme standardavvik σ.
- Normalitet: Gitt x-ene er $y_1, y_2, \cdots, y_n$ (samt $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$) normalfordelte

Metoden for å finne estimatorer b_0, b_1,..., b_p for parametrene β_0, β_1,..., β_p er ved **minste kvadraters metode**. Estimert respons for den i'te enheten er

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$$

Det tilsvarende residualet er som tidligere definert som avviket mellom observert verdi og predikert verdi

$$\begin{aligned} e_i &= \textit{observert respons} - \textit{predikert respons} \\ &= y_i - \hat{y}_i \\ &= y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_p x_{ip}. \end{aligned}$$

Minste kvadraters estimatorene består nå i å velge de b-ene slik at summen av kvadratavvikene minimeres, altså $(b_0, b_1, b_2, \cdots, b_p)$ slik at

$$\sum (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_p x_{ip})^2$$

blir minst mulig.

Untatt i helt spesielle situasjoner finnes det ikke enkle formler for de enkelte b_j , men det er ikke noe problem å regne dem ut på en datamaskin.

For å estimere (residual) varians σ^2 bruker vi formelen

$$s^2 = \frac{1}{n-p-1} \sum (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_p x_{ip})^2$$

altså som et "gjennomsnitt" av kvadrerte residualer og estimat for σ blir som tidligere $s = \sqrt{s^2}$

Grunnen til at vi deler på n-p-1 i

$$s^2 = \frac{1}{n-p-1} \sum (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_p x_{ip})^2$$

er at vi har estimert p+1 regresjonsparametre $(b_0, b_1, b_2, \cdots, b_p)$.

Man kan vise at s^2 dermed vil være forventningsrett for σ^2.

Vi sier at vi har brukt p+1 frihetsgrader og at vi har n-p-1 frihetsgrader igjen.

Dette er helt tilsvarende enkel lineær regresjon der p=1 og vi brukte p+1 =2 til estimeringen og hadde n-2 frihetgrader igjen.

Også tilsvarende enkel lineær regresjon vil vi ved multippel regresjon ha at minste kvadraters estimatorene b_j alle vil være forventnings-rette

$$\mu_{b_j} = \beta_j$$

Dessuten kan man beregnes deres standarfeil SE_j (dvs. deres standardavvik). Formlene for disse er kompliserte, men det lett å beregne dem på en datamaskin. De er for øvrig proposjonale med standardavviket *s*.

Videre vil vi ha at b_j er normalfordelte når feilleddene $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ er normalfordelte. Hvis ikke denne egenskapen holder vil likevel estimatorene b_j være tilnærmet normalfordelte når n er stor.

Dermed får vi også t-statistikker $t_j = \dfrac{b_j}{SE_j}$

som blir t-fordelt med n-p-1 frihetsgrader hvis $\beta_j = 0$

Minitab-utskrift for GPA-data (igjen, men med fokus på estimater og standardfeil)

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0,726103	22,77%	21,19%	18,01%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	0,069	0,454	0,15	0,879
HSM	0,1232	0,0549	2,25	0,026
HSS	0,1361	0,0700	1,95	0,054
HSE	0,0585	0,0654	0,89	0,373

Regression Equation

GPA = 0,069 + 0,1232 HSM + 0,1361 HSS + 0,0585 HSE

Dette gir resultater som generaliserer det vi har sett:

Et **konfidensintervall** med konfidenskoeffisient lik C for β_j har grenser

$$b_j \quad \pm \quad t^* \, SE_{b_j}.$$

Her er t* den verdien som gjør at arealet under tetthetskurven til en t(n-p-1) fordeling mellom -t* og t* er C .

Minitab-utskrift for GPA-data (med konfidensintervall)

Her er n=150 og p=3, så vi har n-p-1 =146 frihetsgrader for t-fordelingen og 97.5 persentilen i t_{146} er $t^* = 1.976346$

```
Coefficients

Term        Coef   SE Coef          95% CI
Constant   0,069     0,454   ( -0,827;  0,966)
HSM       0,1232    0,0549   ( 0,0148; 0,2317)
HSS       0,1361    0,0700   (-0,0021; 0,2744)
HSE       0,0585    0,0654   (-0,0708; 0,1878)
```

Vi kan legge merke til at kun et av disse intervallene ikke inneholder verdien null, dermed vil bare parameteren svarende til HSM være signifikant forskjellig fra 0.

For å teste nullhypotesene H_0: $\beta_j = 0$ benyttes altså t-statistikken

$$t = \frac{b_j}{SE_{b_j}}$$

Når H_0: $\beta_j = 0$ holder vil denne t-statistikken være trukket fra

en T-fordeling med n-p-1 frihetsgrader.

La i det følgende T være en tilfeldig variabel fra denne fordelingen.

P-verdien for testen vil avhenge av hvilket alternativ vi bruker. Standard er å rapportere P-verdier fra tosidige tester.

P-verdien for $H_0: \beta_j = 0$ blir

med ensidig alternativ hypotese $H_a: \beta_j < 0$:

$$P(T < t)$$

med ensidig alternativ hypotese $H_a: \beta_j > 0$:

$$P(T > t)$$

med tosidig alternativ hypotese $H_a: \beta_j \neq 0$:

$$2\,P(T > |t|)$$

Minitab-utskrift for GPA-data: Testing

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	0,069	0,454	0,15	0,879
HSM	0,1232	0,0549	2,25	0,026
HSS	0,1361	0,0700	1,95	0,054
HSE	0,0585	0,0654	0,89	0,373

I samsvar med hva vi så fra konfidensintervallene var det altså bare HSM som hadde en signifikant sammenheng med GPA, men p-verdien for HSS (0.054) er nær det vanlige signifikanskravet.

Dette er p-verdier ved tosidige tester. Hvis man har bestemt seg for et ensidig alternativ kan man bare dele de oppgitte p-verdiene på 2. (så med ensidig ”>” alternativ er også HSS signifikant).

Estimering av forventet respons og predikasjon av ny verdi *y**.

Akkurat som ved enkel lineær regresjon vil vi være interessert i å estimere, gitt forklaringsvariable $x_1, x_2, \dots, x_p$,

- forventningen til y med disse forklaringsvariablene

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- verdien av ny *y** med de samme forklaringsvariablene

$$y^* = \mu_y + \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

Her er ε et feilledd som antas å ha forventning 0 og standardavvik σ, og kanskje er normalfordelt.

For begge størrelsene benytter vi samme punktestimat

$$\hat{\mu}_y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

men variasjonen må behandles på ulike måter.

Usikkerhet i forventet respons - og i predikert ny verdi *y**:

Standardfeilen $SE_{\hat{\mu}_y}$ til $\hat{\mu}_y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$ avhenger av usikkerheten (varianser og kovarianser) til minste kvadraters estimatorene $b_0, b_1, b_2, \cdots, b_p$ samt av verdiene av forklarings-variablene $x_1, x_2, \ldots, x_p$.

Et 95% konfidensintervall for $\hat{\mu}_y$ blir nå gitt ved $\hat{\mu}_y \pm t^* SE_{\hat{\mu}_y}$ der t^* er 97.5 persentilen i t-fordelingen med n-p-1 frihetsgrader.

Tilsvarende blir til et 95% prediksjonsintervall for ny *y** gitt ved

$$\hat{\mu}_y \pm t^* SE_{y^*}$$

der $SE^2_{y^*} = s^2 + SE^2_{\hat{\mu}}$ er estimert varians for ny *y**.

Usikkerhet i forventet respons og predikert ny GPA verdi *y** ved ulike verdier av HSM, HSS, HSE

Variable Setting: HSM = 5, HSS = 5, HSE = 5

Fit	SE Fit	95% CI	95% PI
1,659	0,202	(1,259; 2,058)	(0,169; 3,148)

Variable Setting: HSM = 2, HSS = 2, HSE = 2

Fit	SE Fit	95% CI	95% PI	
0,705	0,352	(0,009; 1,401)	(-0,890; 2,300)	XX

Variable Setting: HSM = 10, HSS = 10, HSE = 10

Fit	SE Fit	95% CI	95% PI
3,248	0,088	(3,074; 3,422)	(1,802; 4,693)

Merk: Ved HSM=HSS =HSE får vi negativ nedre grense, umulig! Tilsvarende øvre grense ved HSM=HSS=HSE=10 større enn 4 (også umulig!)

Forklart andel av varians: R^2

Ved enkel lineær regresjon hadde vi at forklart andel av variasjon

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

faktisk var lik korrelasjonskoeffisienten opphøyd i 2 ($R^2 = r^2$).

En så enkel sammenheng har vi ikke ved multippel regresjon. Men vi er fortsatt interessert i å se hvor mye av variasjonen i de opprinnelige dataene som kan forklares ved regresjonen og denne størrelsen er gitt ved den generelle definisjonen, altså

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

som lett kan regnes ut også i dette generaliserte tilfellet.

Kvadratroten $R = \sqrt{R^2}$ kalles den multiple korrelasjonskoeffisient

Minitab-utskrift for GPA-data: nå uthevd for R^2

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0,726103	22,77%	21,19%	18,01%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	0,069	0,454	0,15	0,879
HSM	0,1232	0,0549	2,25	0,026
HSS	0,1361	0,0700	1,95	0,054
HSE	0,0585	0,0654	0,89	0,373

Regression Equation

GPA = 0,069 + 0,1232 HSM + 0,1361 HSS + 0,0585 HSE

Noen egenskaper ved R^2

- $0 \leq R^2 \leq 1$
- R^2 er kvadratet av korrelasjonskoeffisienten mellom observasjonene y_i og prediksjonene $\hat{y}_i$.
- R^2 vil øke (kan ikke avta) når vi inkluderer en ny forklaringsvariabel
- R^2 er større enn kvadrert korrelasjon mellom alle forklaringsvariable x_j og respons y.

Siden R^2 vil øke med antall forklaringsvariable også når disse har helt marginal betydning vil den overestimere betydningen av alle forklaringsvariablene. Derfor oppgis også andre varianter av dette målet i statistikkpakker, bl.a.: Justert (adjusted) og predikert R^2

```
Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0,726103  22,77%     21,19%      18,01%
```

Antagelsen i multippel lineær regresjon, punktvis.

- Linearitet: $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
- Uavhengighet: Gitt x-ene er $y_1, y_2, \cdots, y_n$ (samt $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$) uavhengige.
- Konstant varians: Gitt x-ene har $y_1, y_2, \cdots, y_n$ (samt $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$) samme standardavvik σ.
- Normalitet: Gitt x-ene er $y_1, y_2, \cdots, y_n$ (samt $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$) normalfordelte

I det følgende skal først se på hvordan antagelsene sjekkes grafisk.

Deretter følger litt diskusjon av konsekvenser og mulige forbedringer.

Linearitet: $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

De to vanligste plottene for å sjekke om lineariteten holder er

○ Plott residualer mot predikerte verdier

○ Plott residualene mot hver av forklaringsvariablene

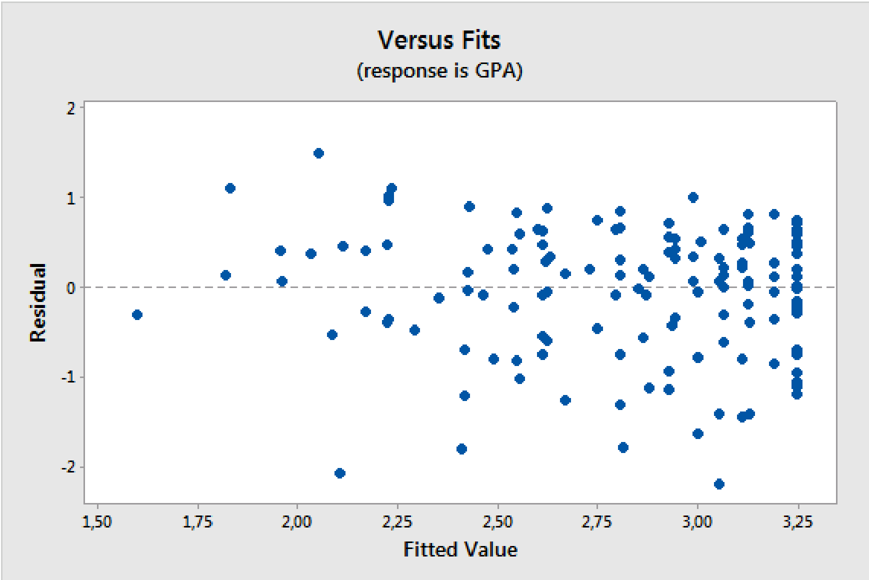
Hvis man ser et mønster (f.eks. først positive residualer, deretter negative og så til slutt positive residualer igjen) tyder dette på avvik fra linearitet.

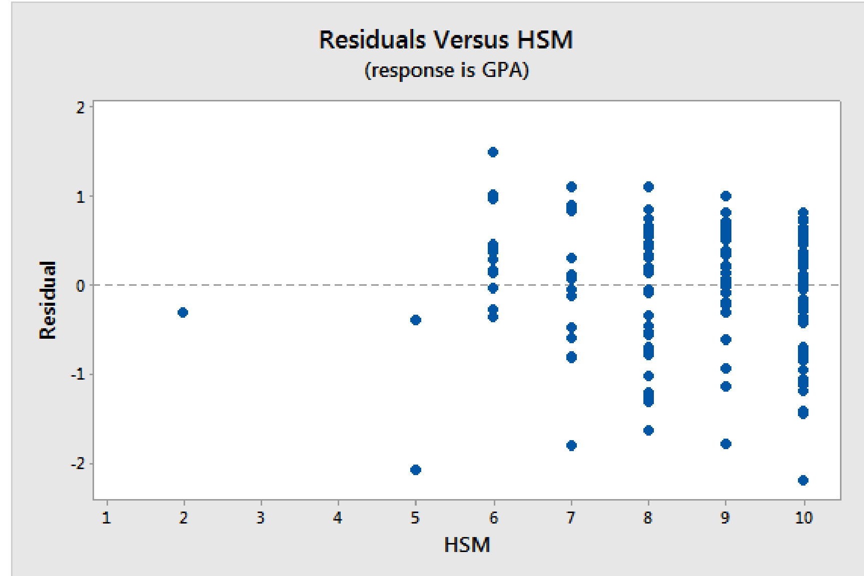
Merk: For enkel lineær regresjon er det samme form på disse plottene – siden predikert verdi er en lineærtransformasjon av forklaringsvariablen.

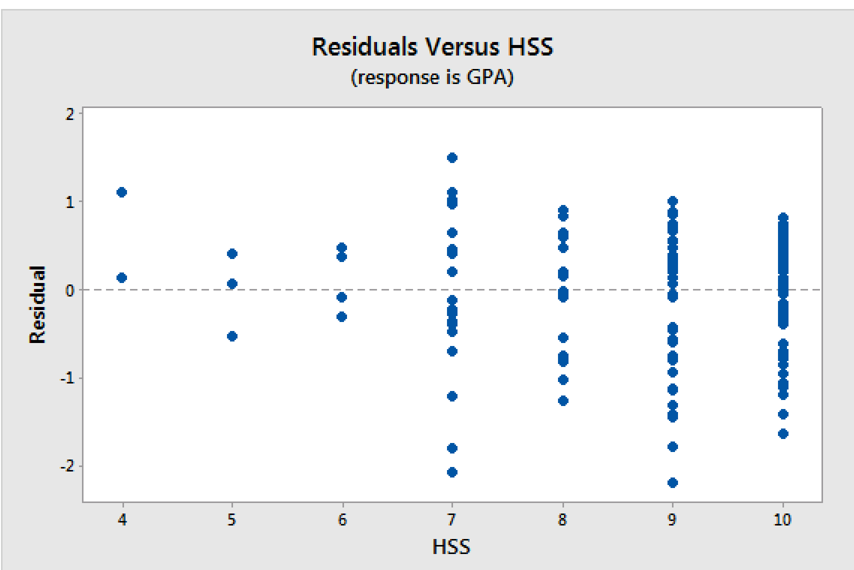
På neste side er det satt inn

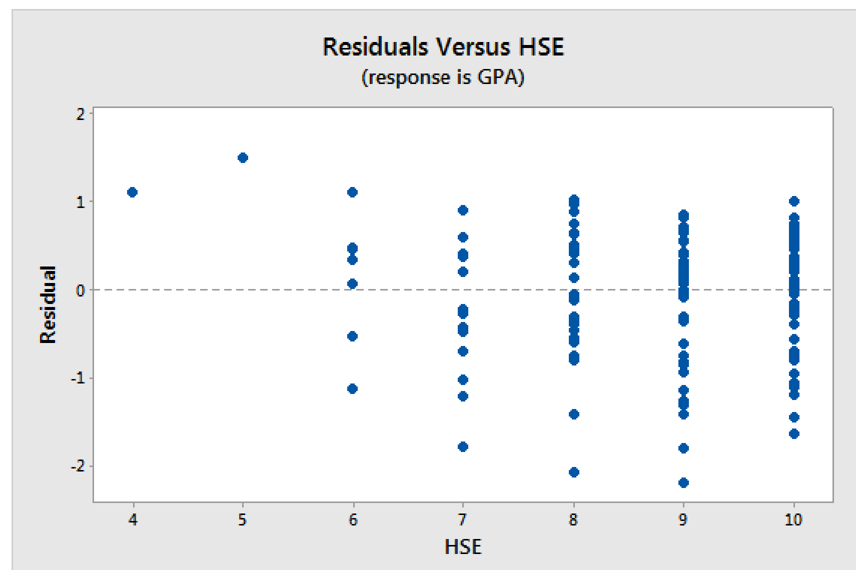
- Plott av residualer mot predikerte verdier (oppe til høyre)
- Plott av residualer mot HSM (oppe til venstre)
- Plott av residualer mot HSS (nede til høyre)
- Plott av residualer mot HSE (nede til venstre)

Her er det ikke klare avvik fra linearitet.


Versus Fits
(response is GPA)
Residual
Fitted Value
1,50
1,75
2,00
2,25
2,50
2,75
3,00
3,25
2
1
0
-1
-2


Residuals Versus HSM
(response is GPA)
Residual
HSM
1
2
3
4
5
6
7
8
9
10
2
1
0
-1
-2


Residuals Versus HSS
(response is GPA)
Residual
HSS
4
5
6
7
8
9
10
2
1
0
-1
-2


Residuals Versus HSE
(response is GPA)
Residual
HSE
4
5
6
7
8
9
10
2
1
0
-1
-2

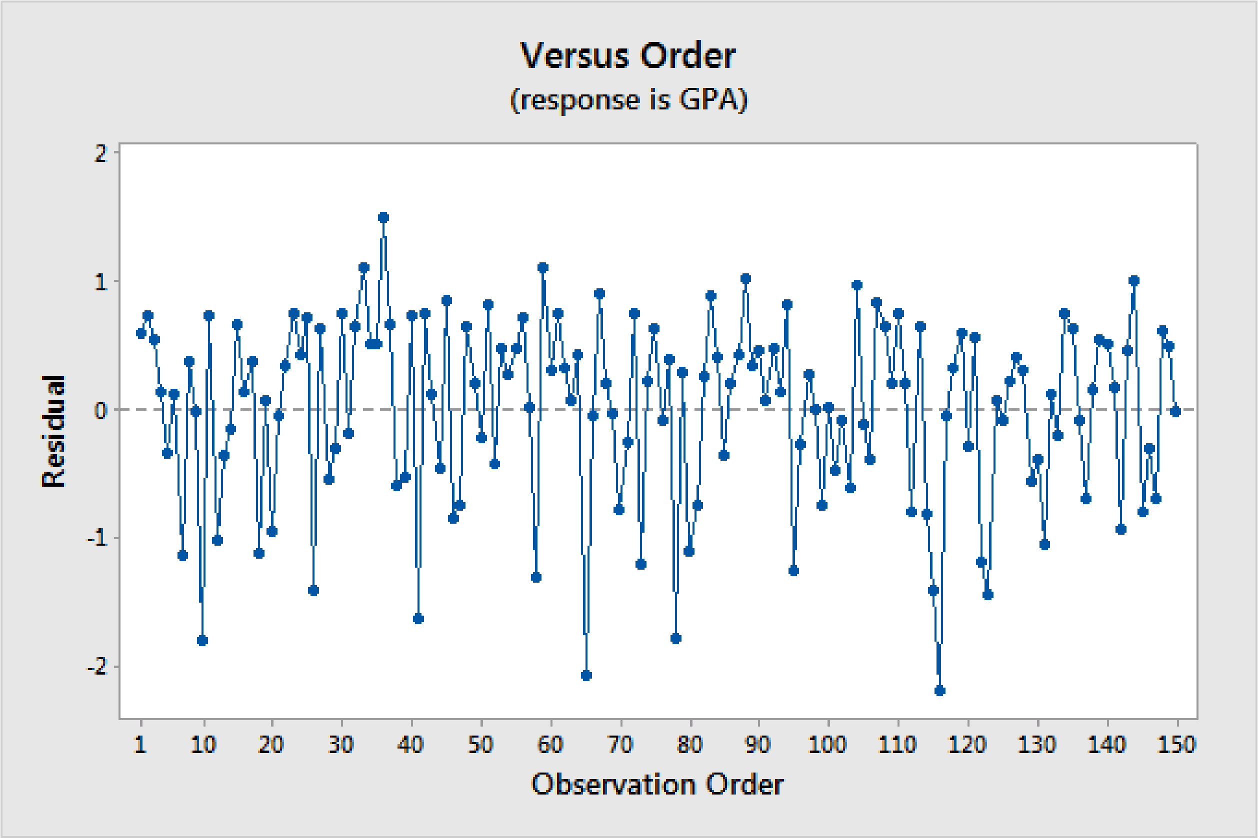
Uavhengiget:

Hvis dataene er hentet inn i en rekkefølge er det naturlig å plotte residualer mot observasjonsnummer.

Igjen, mønstre indikerer avhengighet.

Også andre typer avhengigheter som innen familier, skoleklasser e.l. kan oppdages: F.eks. hvis residualene innen en klasse systematisk nesten alle er positive.

Men: i mange situasjoner er det ikke mulig å sjekke dette, f.eks. ved GPA-dataene. For illustrasjonens skyld tar ser vi likevel på et plott av residualene mot ordning.



Versus Order
(response is GPA)
Residual
Observation Order
2
1
0
-1
-2
1
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150

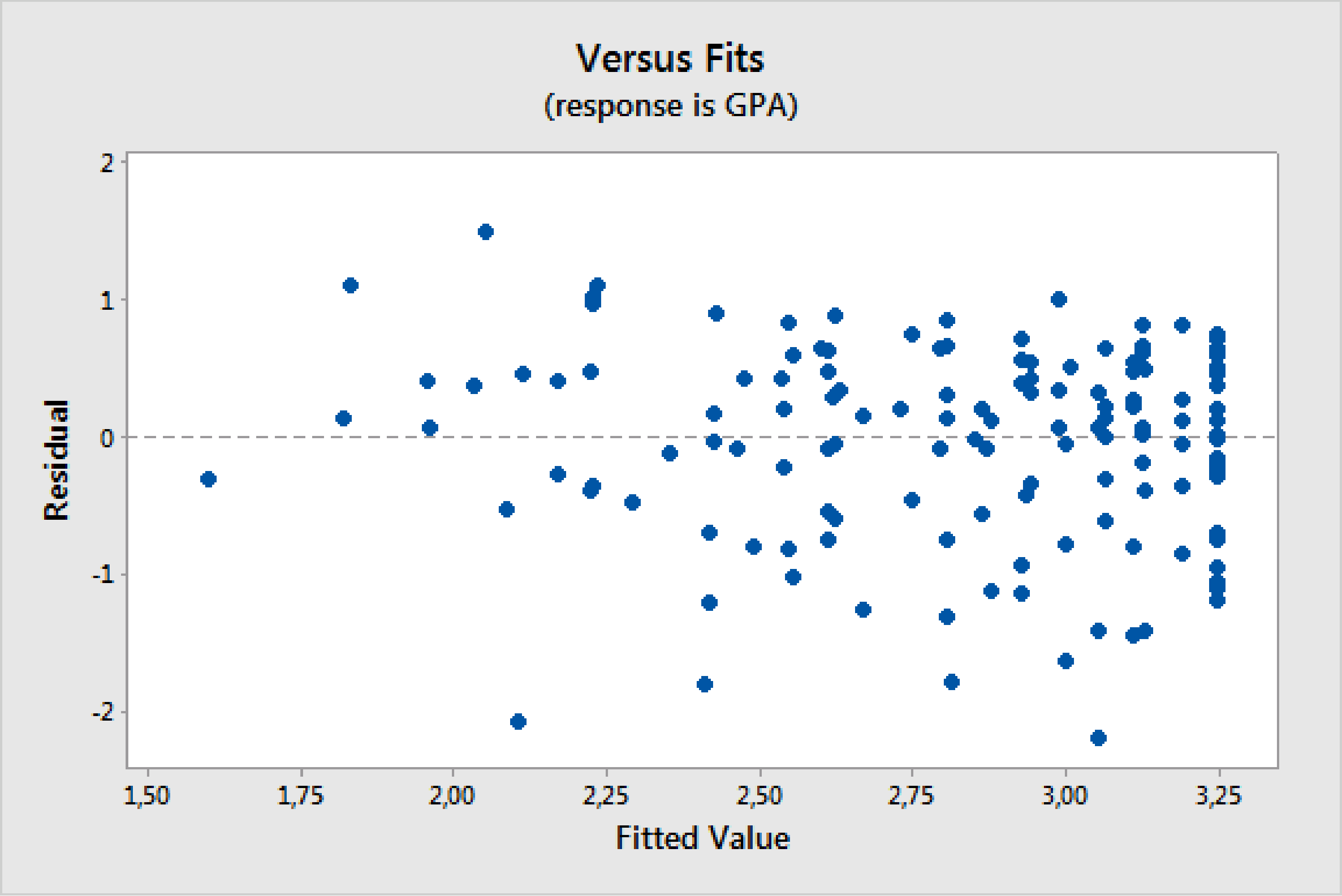
Konstant varians:

- Plott residualer mot predikerte verdier

- Plott absoluttverdi av residualene mot predikerte verdier

Hvis man ser et mønster, f.eks. en ”vifteform” i residualen, dvs. stadig større spredning er det grunn til å tro at antagelsen ikke holder.

På neste side gjentar vi plottet residualer mot predikerte verdier for GPA-dataene. Det er ikke opplagt at det er økende (eller avtagende) variasjon med de predikerte verdiene.


Versus Fits
(response is GPA)
Residual
Fitted Value
2
1
0
-1
-2
1,50
1,75
2,00
2,25
2,50
2,75
3,00
3,25

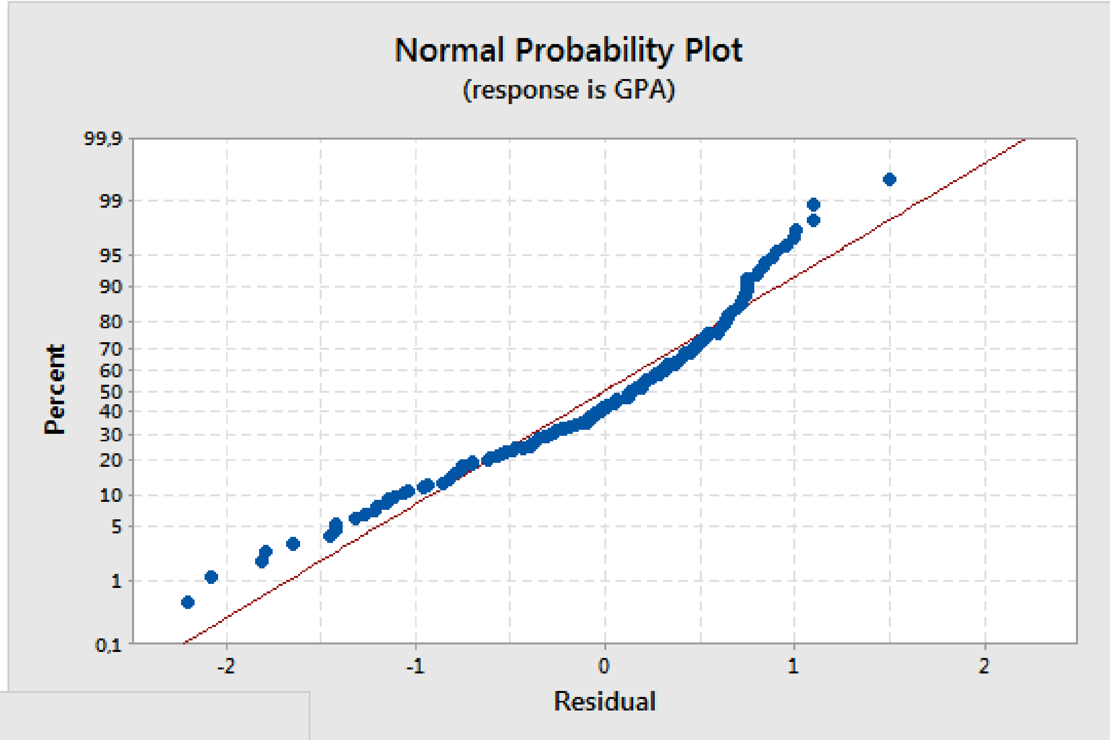
Normalitet:

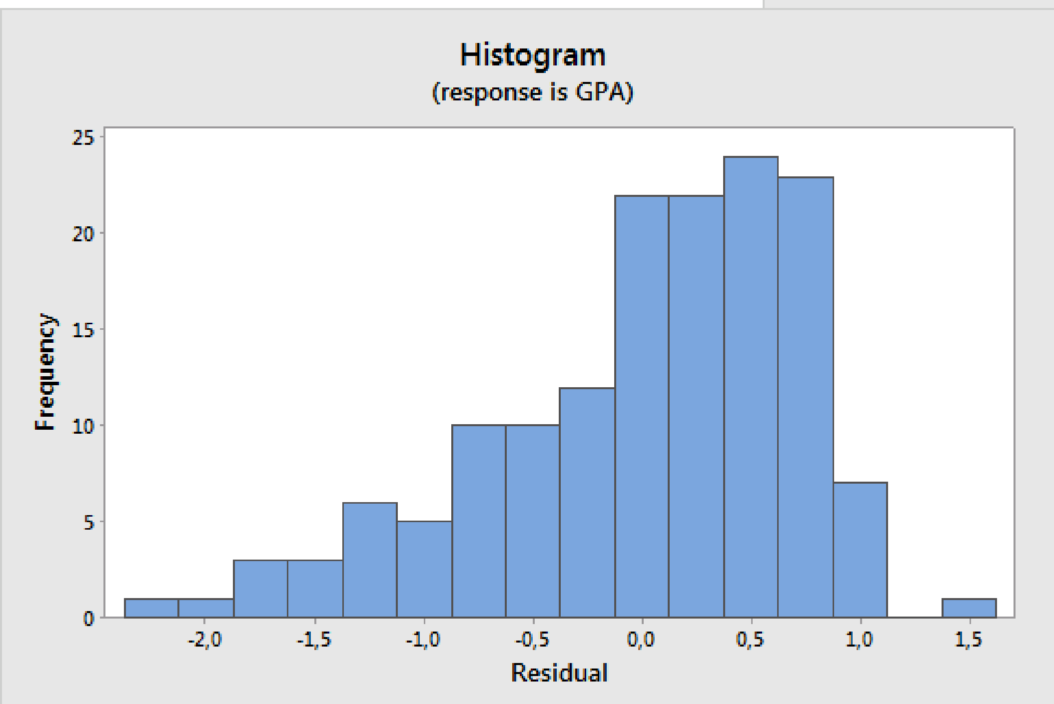
- Normalplott av residualene

- Histogram over residualene

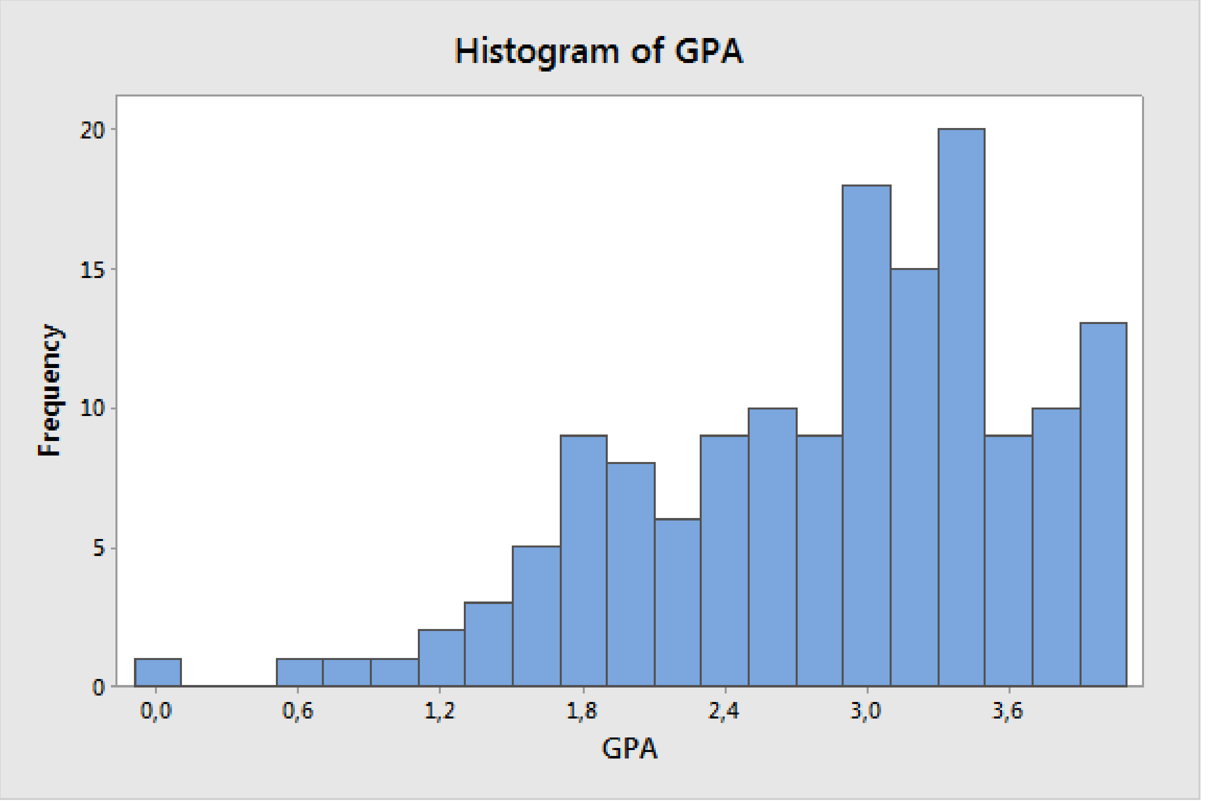
Disse plottene skal helst indikere normalfordeling.

I plottene på neste side for GPA-dataene ser vi klare avvik fra normalitetsantagelsen, men det er ingen opplagte outliere.


Normal Probability Plot
(response is GPA)
Percent
99,9
99
95
90
80
70
60
50
40
30
20
10
5
1
0,1
-2
-1
0
1
2
Residual


Histogram
(response is GPA)
Frequency
25
20
15
10
5
0
-2,0
-1,5
-1,0
-0,5
0,0
0,5
1,0
1,5
Residual

Til sammenligning histogrammet over GPA
(det ligner litt på histogrammet over residualene. Dette siden vi forklarer så lite)



Konsekvenser av avvik fra antagelsene i multippel lineær regresjon.

- Linearitet: Modellen er gal og hvis avvikene er markante må dette rettes opp!

- Uavhengighet: Standardfeil kan bli gale. Dette vil medføre at inferens (konf.int/tester) kan bli gale. Men estimatene er likevel OK.

- Konstant varians: Standardfeil kan bli gale. Dette vil medføre at inferens (konf.int/tester) blir gale. Men estimatene er likevel OK.

- Normalitet: Vi har ikke lenger at teststatistikker etc. er t-fordelt. Men dette er ikke kritisk hvis utvalget (n) er stort såsant det ikke er outliere.

Forbedring av lineær regresjonsmodelle, Linearitet.

Hvis linearitetsantagelsen $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ svikter kan man prøve bl.a.

- Transformer x-er (f.eks. log-trans.).
- Innfør et nytt ledd $\beta_{p+1} x_j^2$ i tillegg til $\beta_j x_j$ (polynomiell regresjon).
- Avansert: glattingsteknikker
- Transformer y-er

Antagelsen i multippel lineær regresjon, korreksjon.

- Uavhengighet: Hvis gruppene er av avhengige enheter er store kan man innføre variable som angir gruppen.

- Konstant varians: Transformer y-ene

- Normalitet: Sjekk outliere

Eksperimenter vs. observasjonelle studier

Planlagte studier – Eksperimenter

- Kan velge verdiene av forklaringsvariablene
- F.eks. slik at de er ukorrelerte
- Dette har en rekke fordeler i multippel regresjon

Observasjonellle studier

- Må ta de verdiene av forklaringsvariablene man observerer
- Typisk blir de korrelerte
- Må håndtere denne kolineariteten

Eks: GPA-dataene

Dette er en observasjonellle studie!

Ser at det er ganske store korrelasjoner mellom HSM, HSS og HSE:

Correlation: GPA; HSM; HSS; HSE

```
        GPA      HSM      HSS
HSM   0,420
      0,000

HSS   0,443    0,670
      0,000    0,000

HSE   0,359    0,485    0,695
      0,000    0,000    0,000

Cell Contents: Pearson correlation
               P-Value
```

For disse dataene er enkle lineære langt på vei like informative!

Regression Analysis: GPA versus HSM

S	R-sq	R-sq(adj)	R-sq(pred)
0,744866	17,62%	17,06%	15,31%

Regression Equation

GPA = 0,825 + 0,2349 HSM

Regression Analysis: GPA versus HSS

S	R-sq	R-sq(adj)	R-sq(pred)
0,735841	19,60%	19,06%	17,35%

Regression Equation

GPA = 0,558 + 0,2596 HSS

Regression Analysis: GPA versus HSE

S	R-sq	R-sq(adj)	R-sq(pred)
0,766027	12,87%	12,28%	10,14%

Regression Equation

GPA = 0,795 + 0,2318 HSE

Grunnen er at HSM, HSS og HSE her inneholder essensielt den samme informasjonen og det er ikke så mye ekstra informasjon å se på alle 3 samtidig!

Det er likevel litt å hente å bruke en modell bare med HSM og HSS:

S	R-sq	R-sq(adj)	R-sq(pred)
0,725607	22,35%	21,29%	19,23%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0,257	0,402	0,64	0,524	
HSM	0,1250	0,0548	2,28	0,024	1,81
HSS	0,1718	0,0574	2,99	0,003	1,81

Regression Equation

GPA = 0,257 + 0,1250 HSM + 0,1718 HSS

Et annet poeng: Minste kvadraters estimatene er veldig forskjellige i de ulike modellene

```
GPA = 0,825 + 0,2349 HSM
GPA = 0,558 + 0,2596 HSS
GPA = 0,795 + 0,2318 HSE

GPA = 0,257 + 0,1250 HSM + 0,1718 HSS

GPA = 0,069 + 0,1232 HSM + 0,1361 HSS + 0,0585 HSE
```

Spesielt gjelder dette de enkle lineære regresjonene.

Grunnen er det som er av effekt i f.eks. HSS og HSE forplantes over i HSM i henhold til korrelasjonen mellom forklaringsvariablene når HSS og HSE ikke er med i modellen.