## 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 $x_1, x_2, \ldots, x_p$ ,

- 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 $\varepsilon$ et feilledd som antas å ha forventning 0 og standardavvik $\sigma$, 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%
```