TMA4267 Linear Statistical Models V2017 (L8)

Part 2: Linear regression: Modelling the effects of covariates [F:3.1.3] Parameter estimation: Estimator for β [F:3.2.1]

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: February 10, 2017

The classical linear model

The model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

is called a classical linear model if the following is true:

1. $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$.
2. $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2 \boldsymbol{I}$.
3. The design matrix has full rank $\mathrm{rank}(\boldsymbol{X}) = k + 1 = p$.

The classical *normal* linear regression model is obtained if additionally

4. $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I})$

holds. For random covariates these assumptions are to be understood conditionally on $\boldsymbol{X}$.

Model assumptions for the classical linear model [F:3.1.2]

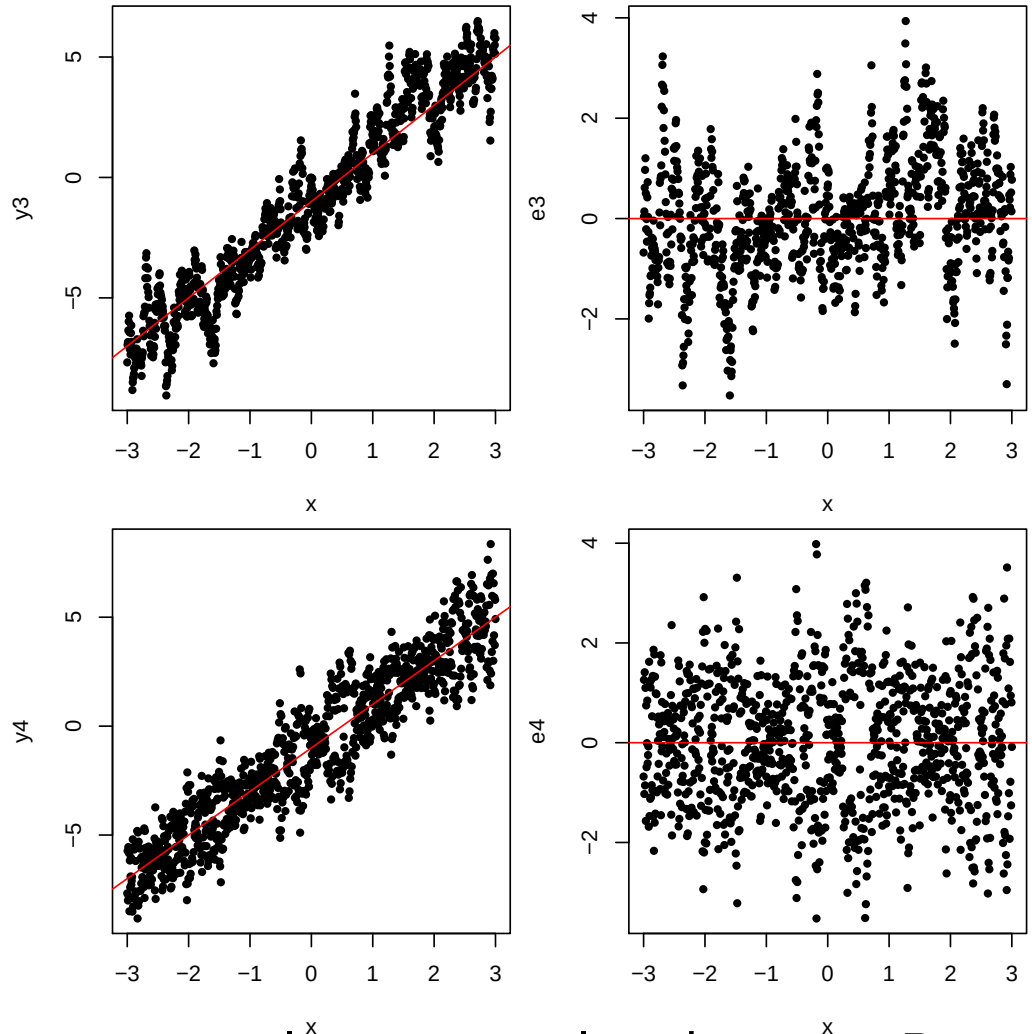
What are our model assumptions, how can we spot violations and what can we do to amend the violations.

1. Linearity of covariates: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
2. Homoscedastic error variance: $Var(\varepsilon_i) = \sigma^2$.
3. Uncorrelated errors: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$.
4. Additivity of errors: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$

We mainly use plots to assess this (more on model fit in F:3.4 Model choice and variable seletion)

- Covariate vs response (for each covariate)
- Covariate vs error (when we have simulated data and know the truth)
- Covariate vs residual (estimated error),
- Predicted response vs residual.

Uncorrelated errors?



Top: positively autocorrelated errors. Bottom: negatively correlated errors. Right: x vs y. Left: x vs error. Example from Fahrmeir et al (2013): Regression. Springer. (p.81). R code from TMA4267 lectures tab.

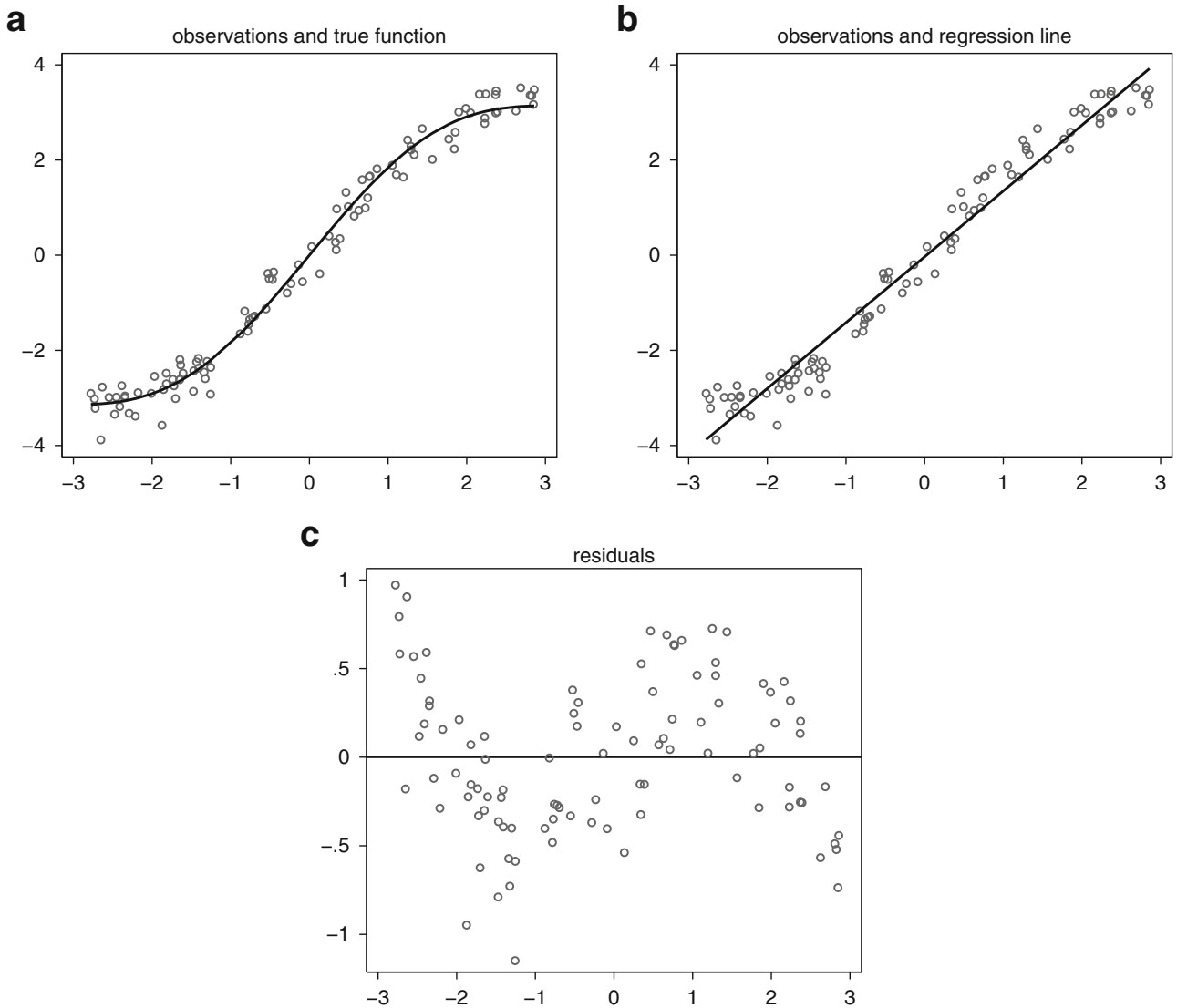


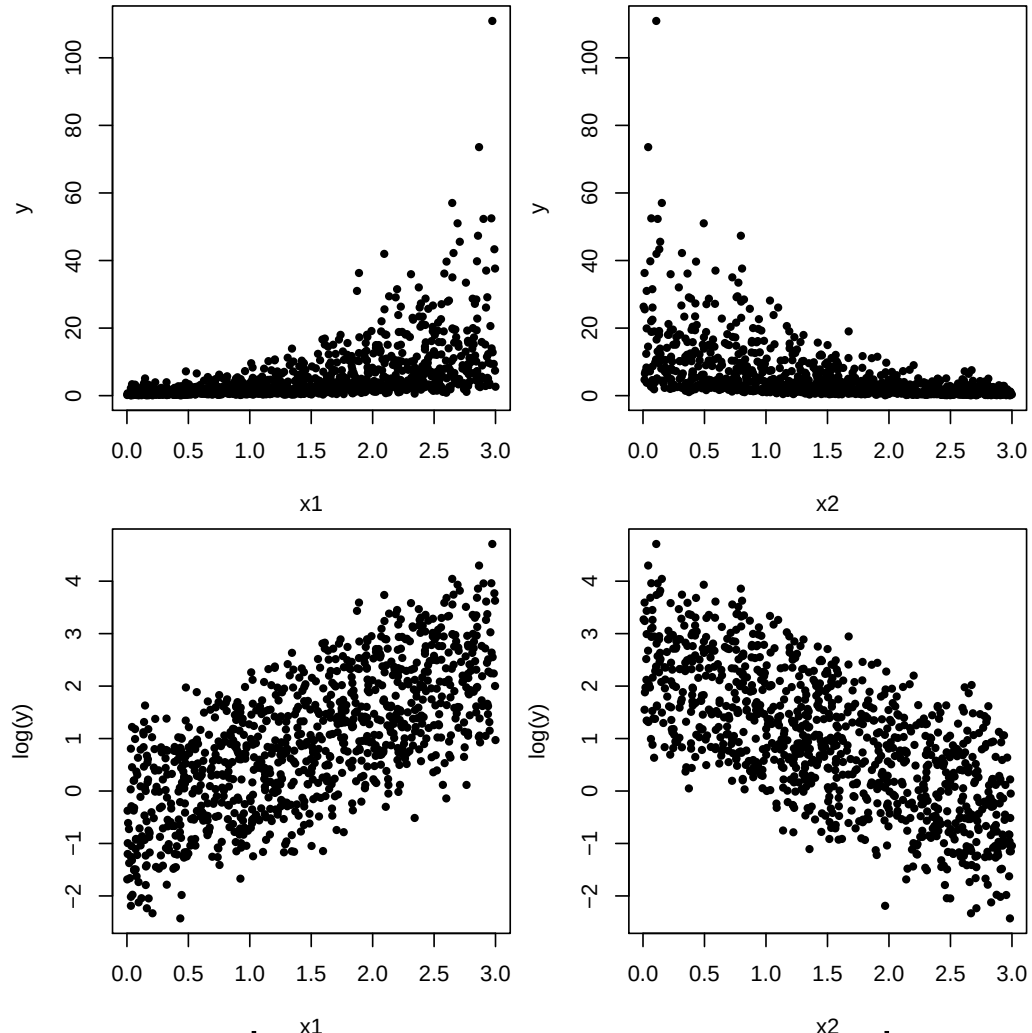
Fig. 3.4 Illustration for correlated residuals when the model is misspecified: Panel (**a**) displays (simulated) data based on the function $\mathrm{E}(y_i \mid x_i) = \sin(x_i) + x_i$ and $\varepsilon_i \sim \mathrm{N}(0, 0.3^2)$. Panel (**b**) shows the estimated regression line, i.e., the nonlinear relationship is ignored. The corresponding residuals can be found in panel (**c**)

Fahrmeir et al (2013): Regression. Springer. (p.82)

Multiplicative errors

```
x1=runif(n,0,3)
x2=runif(n,0,3)
e=rnorm(n,0,0.4)
y=exp(1+x1-x2+e)
plot(x1,y,pch=20)
plot(x2,y,pch=20)
plot(x1,log(y),pch=20)
plot(x2,log(y),pch=20)
```

Multiplicative errors



Top: x1 and x2 vs y. Bottom: x1 and x2 vs log(y). Example from Fahrmeir et al (2013): Regression. Springer. (p.85). R code from TMA4267 lectures tab.

Covariates - how to include in the linear regression?

1. Continuous covariates: as is, transformed or using polynomials.
2. Categorical covariates: dummy variable or effect coding.
3. Interactions between covariates.

Munich rent index data

```
> colnames(ds)
[1] "rent" "rentsqm" "area" "yearc" "location" "bath"
[7] "kitchen"  "cheating" "district"
> apply(ds[,1:4],2,summary)
           rent rentsqm   area yearc
Min.      40.51  0.4158  20.00  1918
1st Qu.  322.00  5.2610  51.00  1939
Median   427.00  6.9800  65.00  1959
Mean     459.40  7.1110  67.37  1956
3rd Qu.  559.40  8.8410  81.00  1972
Max.    1843.00 17.7200 160.00  1997
> unlist(apply(ds[,5:8],2,table))
location.1 location.2 location.3  bath.0  bath.1 kitchen.0
      1794       1210         78       2891   191   2951
 kitchen.1 cheating.0 cheating.1
       131        321       2761
```

How to code categorical covariates: rentsqm vs location with linear coding

- Location average=1, good=2 and top=3, and regression model

$$\text{rentsqm}_i = \beta_0 + \beta_1 \text{location}_i + \varepsilon_i$$

- Parameter estimate: $\hat{\beta}_1 = 0.39$. What does that mean?
  - Flat of average location: $\widehat{\text{rentsqm}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1$
  - Flat of good location: $\widehat{\text{rentsqm}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 2$
  - Flat of top location: $\widehat{\text{rentsqm}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 3$
- What is the difference in predicted rentsqm between top and good location, and between good and average location?
- So, the difference between a top and a good location is the same as the difference between good and average. Is this what we want?

Linear coding

```
> fit1=lm(rentsqm~as.numeric(location),data=ds)
> summary(fit1)
Call:
lm(formula = rentsqm ~ as.numeric(location), data = ds)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.54390    0.12368  52.911  < 2e-16 ***
as.numeric(location)  0.39312    0.08016   4.904 9.88e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.427 on 3080 degrees of freedom
Multiple R-squared:  0.007748,Adjusted R-squared:  0.007425
F-statistic: 24.05 on 1 and 3080 DF,  p-value: 9.878e-07
```

rentsqm vs location with dummy variable coding

$$\text{aloc}_i = \begin{cases} 0 & \text{location}_i \text{ is not average} \\ 1 & \text{location}_i \text{ is average} \end{cases}$$

$$\text{gloc}_i = \begin{cases} 0 & \text{location}_i \text{ is not good} \\ 1 & \text{location}_i \text{ is good} \end{cases}$$

$$\text{tloc}_i = \begin{cases} 0 & \text{location}_i \text{ is not top} \\ 1 & \text{location}_i \text{ is top} \end{cases}$$

$$\text{rentsqm}_i = \beta_0 + \beta_1 \text{aloc}_i + \beta_2 \text{gloc}_i + \beta_3 \text{tloc}_i + \varepsilon_i$$

- Write down the design matrix for this regression model, when we have 1794 flats with average location, 1210 with good and 78 with top location.
- What is the rank of this design matrix?
- Is there a problem, and a solution?

3.4 Dummy Coding for Categorical Covariates

For modeling the effect of a covariate $x \in \{1, \ldots, c\}$ with c categories using dummy coding, we define the $c - 1$ dummy variables

$$x_{i1} = \begin{cases} 1 & x_i = 1, \\ 0 & \text{otherwise}, \end{cases} \quad \ldots \quad x_{i,c-1} = \begin{cases} 1 & x_i = c - 1, \\ 0 & \text{otherwise}, \end{cases}$$

for $i = 1, \ldots, n$, and include them as explanatory variables in the regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{i,c-1} x_{i,c-1} + \ldots + \varepsilon_i .$$

For reasons of identifiability, we omit one of the dummy variables, in this case the dummy variable for category c. This category is called reference category. The estimated effects can be interpreted by direct comparison with the (omitted) reference category.

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.97)

Dummy coding via contr.treatment

```
> contrasts(ds$location)=contr.treatment(3)
> fit2=lm(rentsqm~location,data=ds)
> summary(fit2)
Call:
lm(formula = rentsqm ~ location, data = ds)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.95654    0.05728 121.456  < 2e-16 ***
location2    0.31570    0.09025   3.498 0.000475 ***
location3    1.21579    0.28060   4.333 1.52e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.426 on 3079 degrees of freedom
Multiple R-squared:  0.008867,Adjusted R-squared:  0.008223
F-statistic: 13.77 on 2 and 3079 DF,  p-value: 1.109e-06
```

Effect coding via contr.sum

```
> contrasts(ds$location)=contr.sum(3)
> fit3=lm(rentsqm~location,data=ds)
> summary(fit3)
Call:
lm(formula = rentsqm ~ location, data = ds)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.46704    0.09638  77.477  < 2e-16 ***
location1   -0.51050    0.10189  -5.010 5.75e-07 ***
location2   -0.19479    0.10445  -1.865   0.0623 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.426 on 3079 degrees of freedom
Multiple R-squared:  0.008867,Adjusted R-squared:  0.008223
F-statistic: 13.77 on 2 and 3079 DF,  p-value: 1.109e-06
```

Response: birth weight

Covariates: glucose level of mother and BMI of mother.

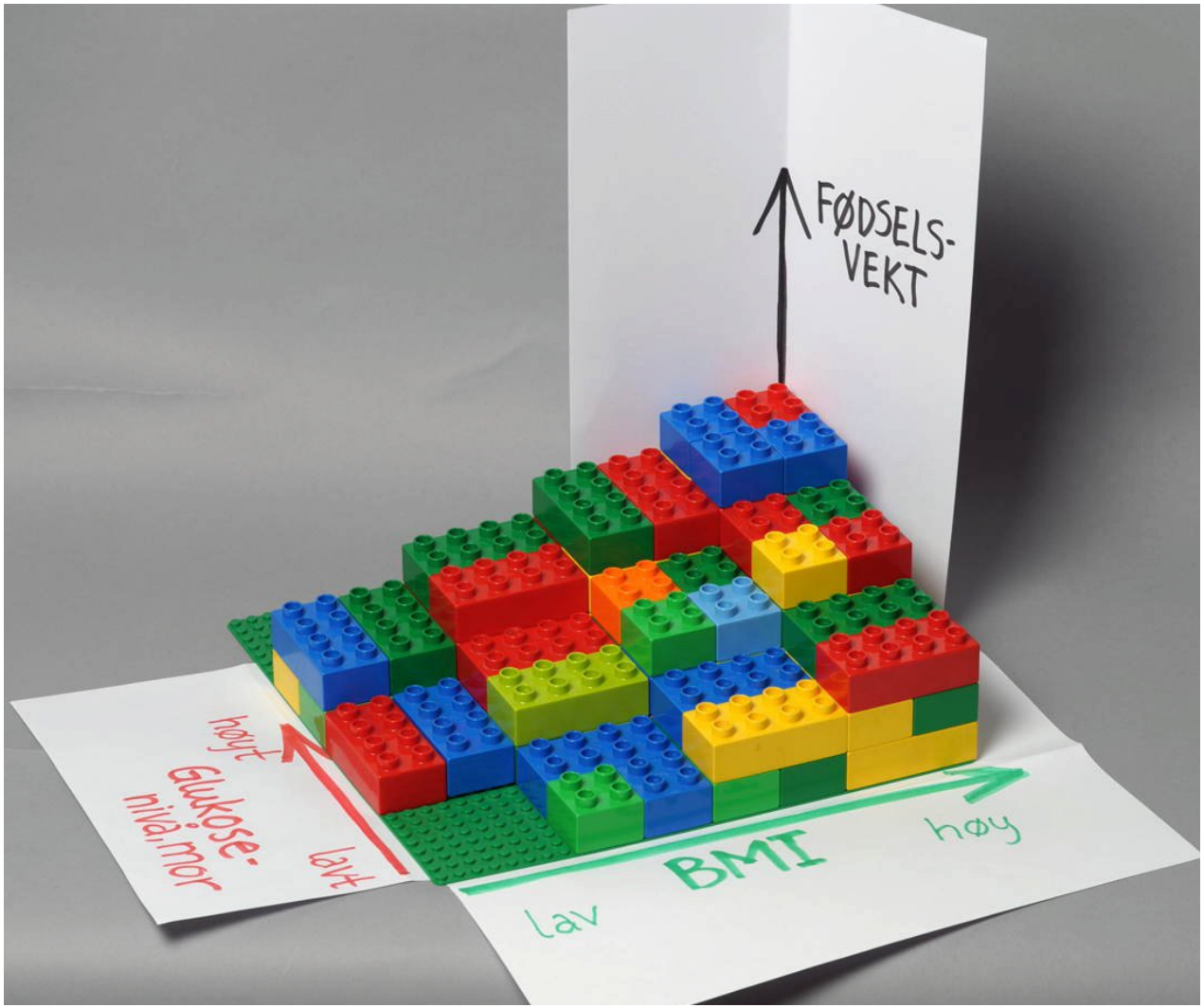


Figure from Kathrine Frey Frøslie.

Response: birth weight

Covariates: glucose level of mother and BMI of mother - with interaction.

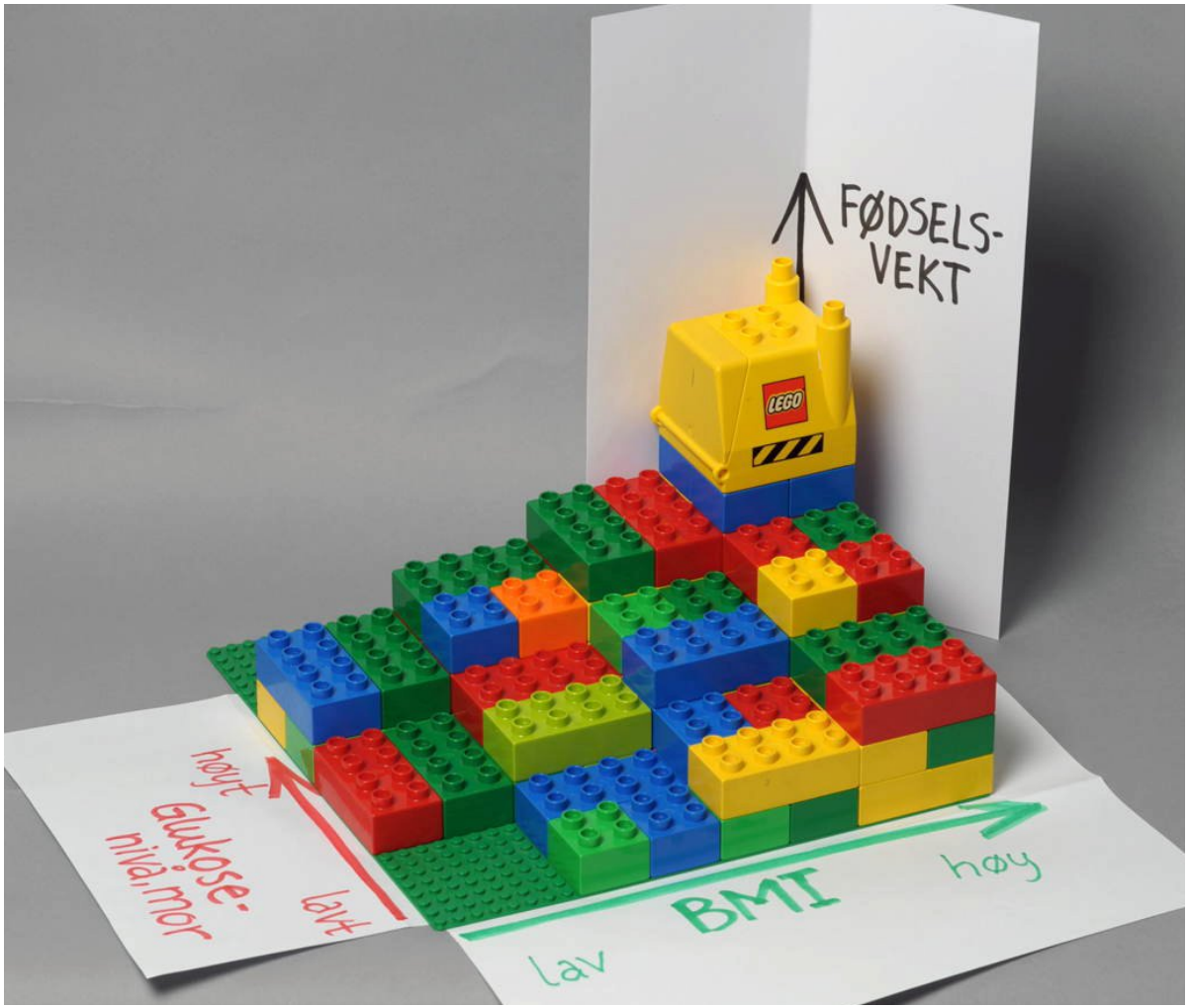


Figure from Kathrine Frey Frøslie.

The classical linear model

$$\underset{(n\times 1)}{\boldsymbol{Y}} = \underset{(n\times p)}{\boldsymbol{X}}\underset{(p\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

$$E(\boldsymbol{\varepsilon}) = \underset{(n\times 1)}{\mathbf{0}} \quad \text{and} \quad Cov(\boldsymbol{\varepsilon}) = \underset{(n\times n)}{\sigma^2\boldsymbol{I}}$$

where

- $\boldsymbol{\beta}$ and σ^2 are unknown parameters and
- the design matrix $\boldsymbol{X}$ has ith row $[x_{i1}x_{i2}\cdots x_{ip}]$.

Next: find the estimator $\hat{\boldsymbol{\beta}}$.

Today

- Model assessment: residual plots.
- Covariates: how to include in linear regression?
- Least squares and maximum likelihood estimator for $\boldsymbol{\beta}$.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$$