TMA4267 Linear Statistical Models V2017 [L7]

Part 2: Linear regression [F p73-86]
Model definition [F3.1], Parameters and residuals [F3.1.1], Model check [F3.1.2]

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: February 7, 2017

Part 2: Linear regression

Part 2: Linear regression

- Fahrmeir et al (2013): Regression. Chapter 3.1, 3.2, 3.4 and required parts of 3.5 and Appendix B.

Part 3: Hypothesis testing and analysis of variance

- Fahrmeir et al (2013): Regression. Chapter 3.3 and required parts of 3.5 and Appendix B.
- Härdle et al (2015): Applied Multivariate Statistical Analysis. Chapter 8.1.1. (ANOVA).
- A short note on multiple testing (to be written).

File TMA4267Part2and3.pdf available from course www-page.

Age-predicted maximal heart rate in healthy subjects: The HUNT Fitness Study

B. M. Nes, I. Janszky, U. Wisløff, A. Støylen, T. Karlsen (2012) in Scandinavian Journal of Medicine and Science in Sports.

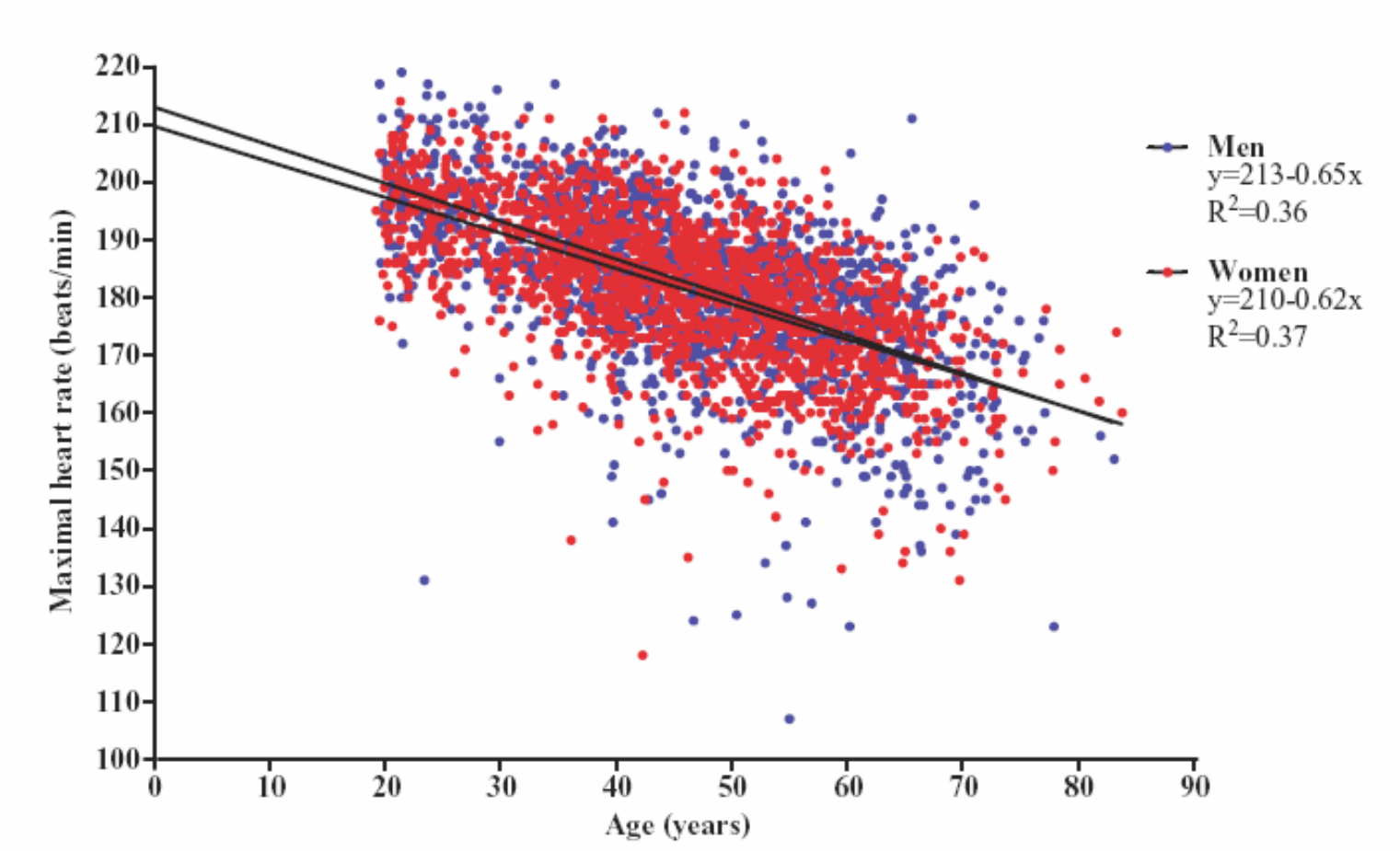
- *HRmax describes the highest heart rate achieved by a subject exercising to exhaustion and is verified by a plateau of heart rate despite increasing workload. In the literature, HRmax commonly refers to the peak heart rate at termination of a graded maximal exercise test.*
- *However, in clinical settings, a maximal exercise test is not always feasible and there is a need to predict HRmax from age prior to testing to be able to adequately assess heart rate response and relative intensity of effort at submaximal levels.*

Age-predicted maximal heart rate in healthy subjects: The HUNT Fitness Study

- *HRmax at a given age is frequently estimated by the "220 - age" formula.*
- *The aim of the present study was to develop a new prediction formula for HRmax through analysis of HRmax measured at VO2peak in a diverse population of 4635 healthy subjects and compare this formula with three commonly used prediction formulas. Furthermore, we wanted to investigate the relationship between HRmax and gender, physical activity status, BMI, and objectively measured aerobic fitness.*

Age-predicted maximal heart rate in healthy subjects: The HUNT Fitness Study - Statistical procedures

- Only subjects that fulfilled the criteria of a maximal test, with registered maximal heart rate (HRmax), were included in the analysis (n = 3320).
- General linear modeling was used to determine the effect of age on HRmax. HRmax was entered as the dependent variable and age as the independent variable. Nonlinearity of the relationship between age and HRmax was investigated by including polynomial terms to the regression model.
- In a subsequent analysis, the effects of gender, BMI, physical activity status, and maximal oxygen uptake were examined by entering these factors as independent variables in addition to age. In further subsequent models, interaction terms were included as well to assess effect modification.
- The continuous variables were checked for normality, homogeneity of variances, and heteroscedasticity of the residuals.



Nes et al (2012): Age-predicted maximal heart rate in healthy subjects: The HUNT Fitness Study. $n = 3320$ individuals.

Munich Rent Index data set

described in Fahrmeir et al (2013) on pages 19-20.

```
> library("gamlss.data")
> ds=rent99
> dim(ds)
[1] 3082    9
> colnames(ds)
[1] "rent"  "rentsqm"  "area"  "yearc"    "location" "bath"
[7] "kitchen"  "cheating" "district"
```

```
> summary(ds)
      rent           rentsqm            area           yearc
 Min.   :  40.51   Min.   : 0.4158   Min.   : 20.00   Min.   :1918
 1st Qu.: 322.03   1st Qu.: 5.2610   1st Qu.: 51.00   1st Qu.:1939
 Median : 426.97   Median : 6.9802   Median : 65.00   Median :1959
 Mean   : 459.44   Mean   : 7.1113   Mean   : 67.37   Mean   :1956
 3rd Qu.: 559.36   3rd Qu.: 8.8408   3rd Qu.: 81.00   3rd Qu.:1972
 Max.   :1843.38   Max.   :17.7216   Max.   :160.00   Max.   :1997
 location bath     kitchen  cheating    district
 1:1794   0:2891   0:2951   0: 321   Min.   : 113
 2:1210   1: 191   1: 131   1:2761   1st Qu.: 561
 3:  78                               Median :1025
                                      Mean   :1170
                                      3rd Qu.:1714
                                      Max.   :2529
```

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 $\boldsymbol{X}$.

Conditional mean and covariance

If we believe that the vector with elements Y and $\boldsymbol{X}$ are multivariate normal $N_{k+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we may look at the partition

$$\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \sim N_{k+1}\left(\begin{pmatrix} \mu_Y \\ \boldsymbol{\mu_X} \end{pmatrix}, \begin{pmatrix} \Sigma_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{pmatrix} \right)$$

The conditional distributions of the components are (multivariate) normal, with conditional mean and variance of $Y \mid \boldsymbol{X} = \boldsymbol{x}$ are

$$\begin{aligned} \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) &= \mu_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_X) \\ \mathrm{Var}(Y \mid \boldsymbol{X} = \boldsymbol{x}) &= \Sigma_Y - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY} \end{aligned}$$

Observe: mean is linear in $\boldsymbol{x}$ and variance independent of $\boldsymbol{x}$.

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.

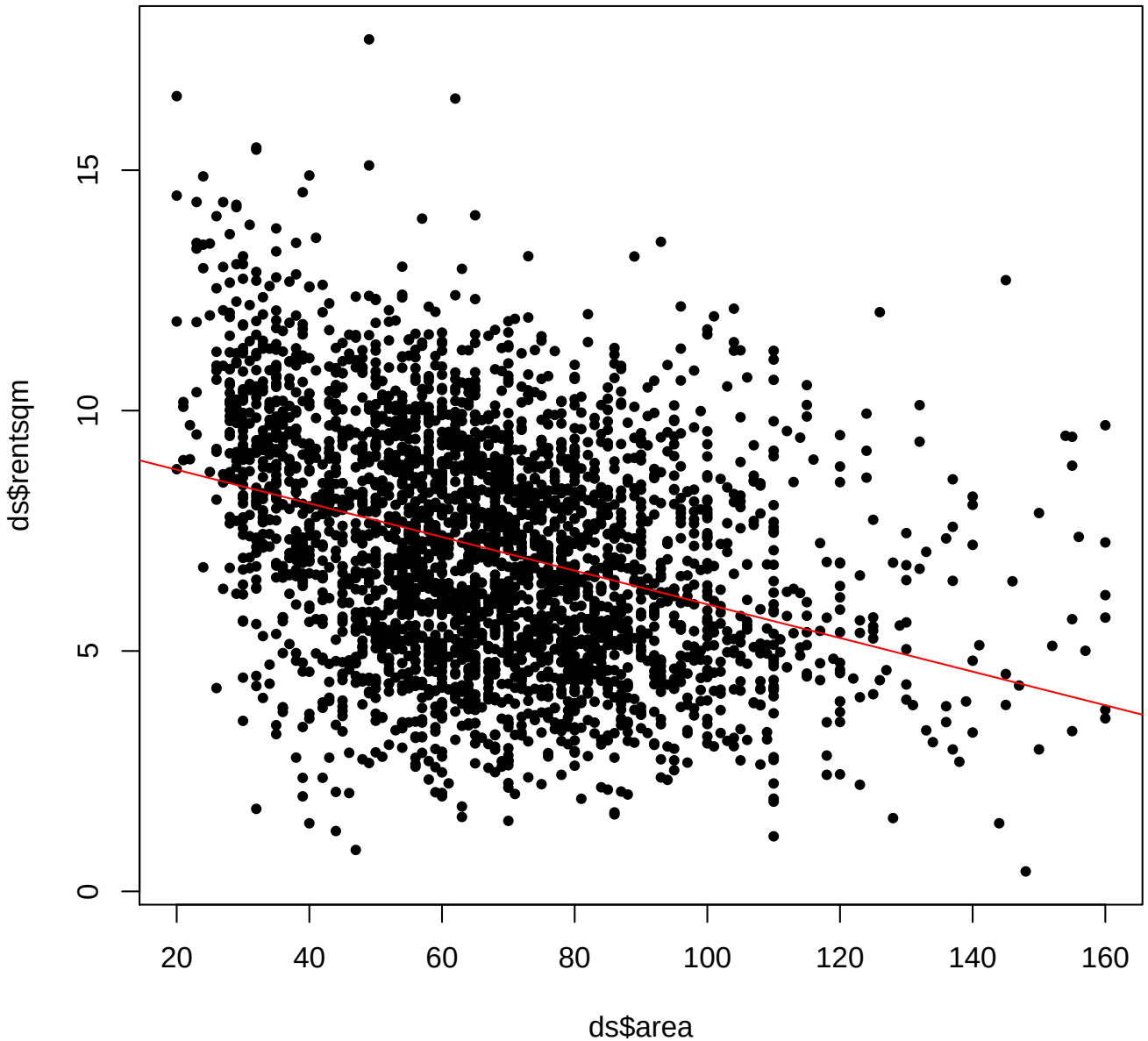
1. Linearity of covariates: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
2. Homoscedastic error variance: $Cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$.
3. Uncorrelated errors: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$.
4. Additivity of errors: $\boldsymbol{Y}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

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 (to be popular later).

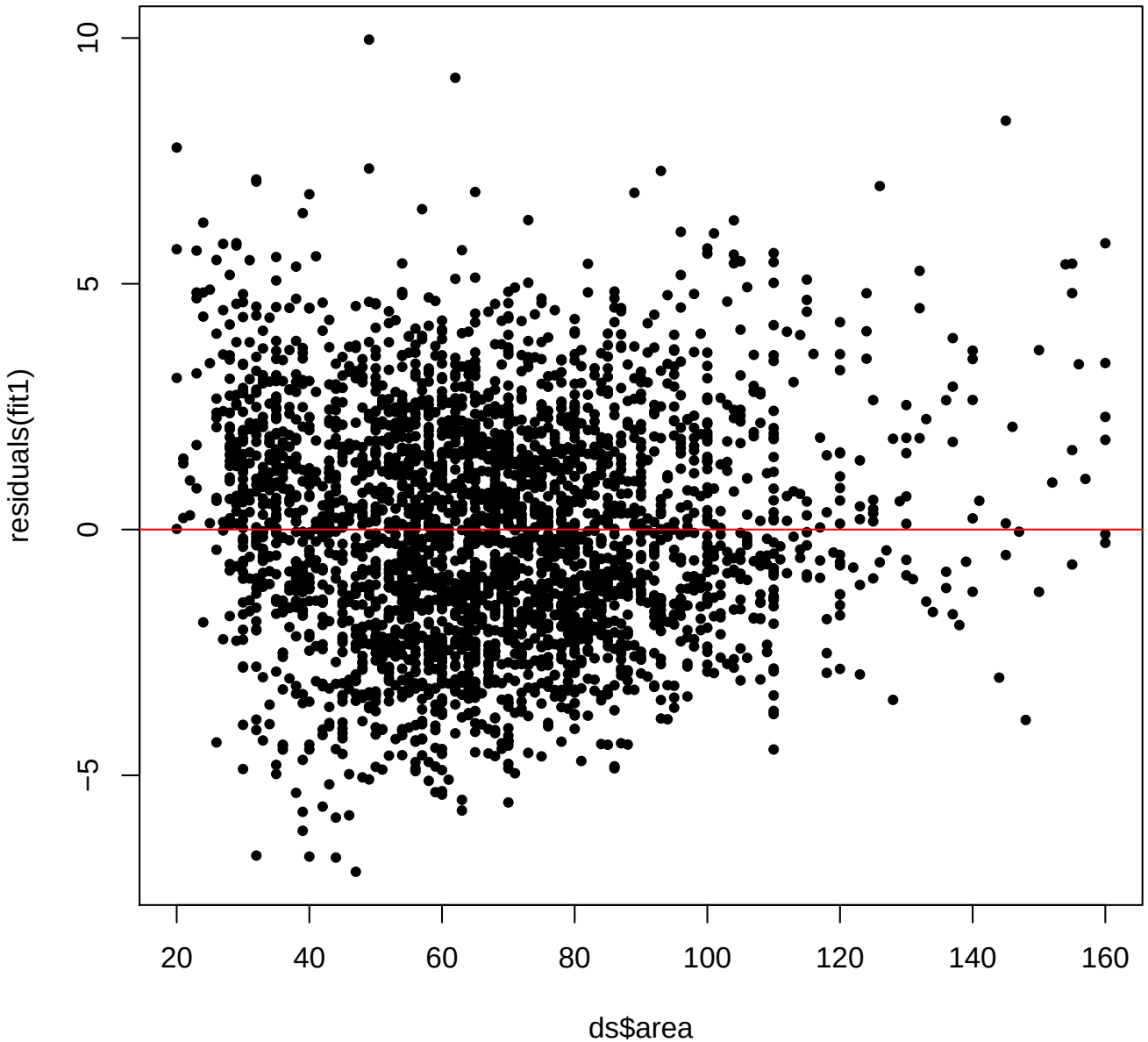
Linearity of covariates: Covariate vs. response

Munich Rent Index: area vs rentsqm



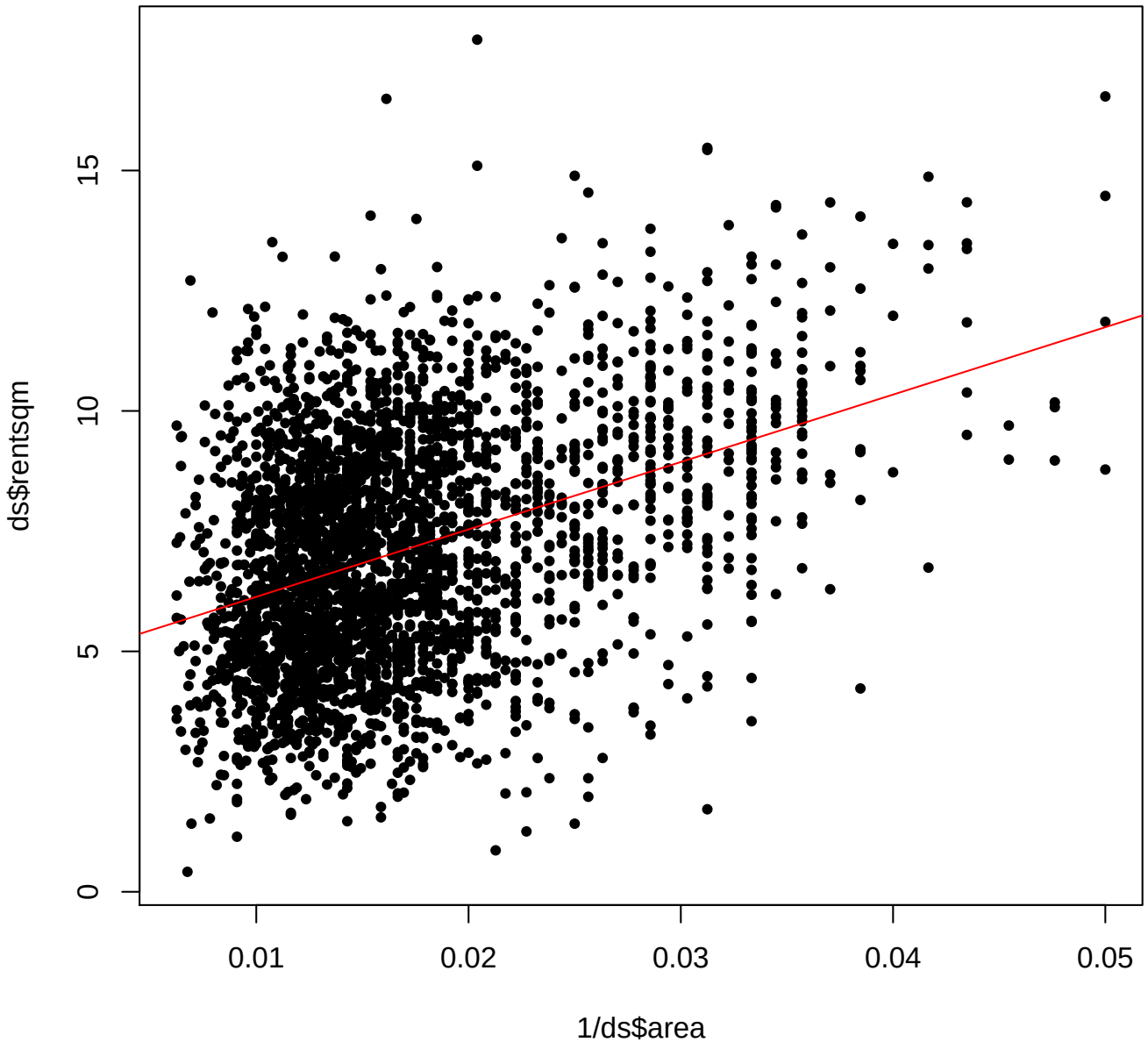
Linearity of covariates: Covariate vs. residual (residual plot)

Munich Rent Index: area vs residual



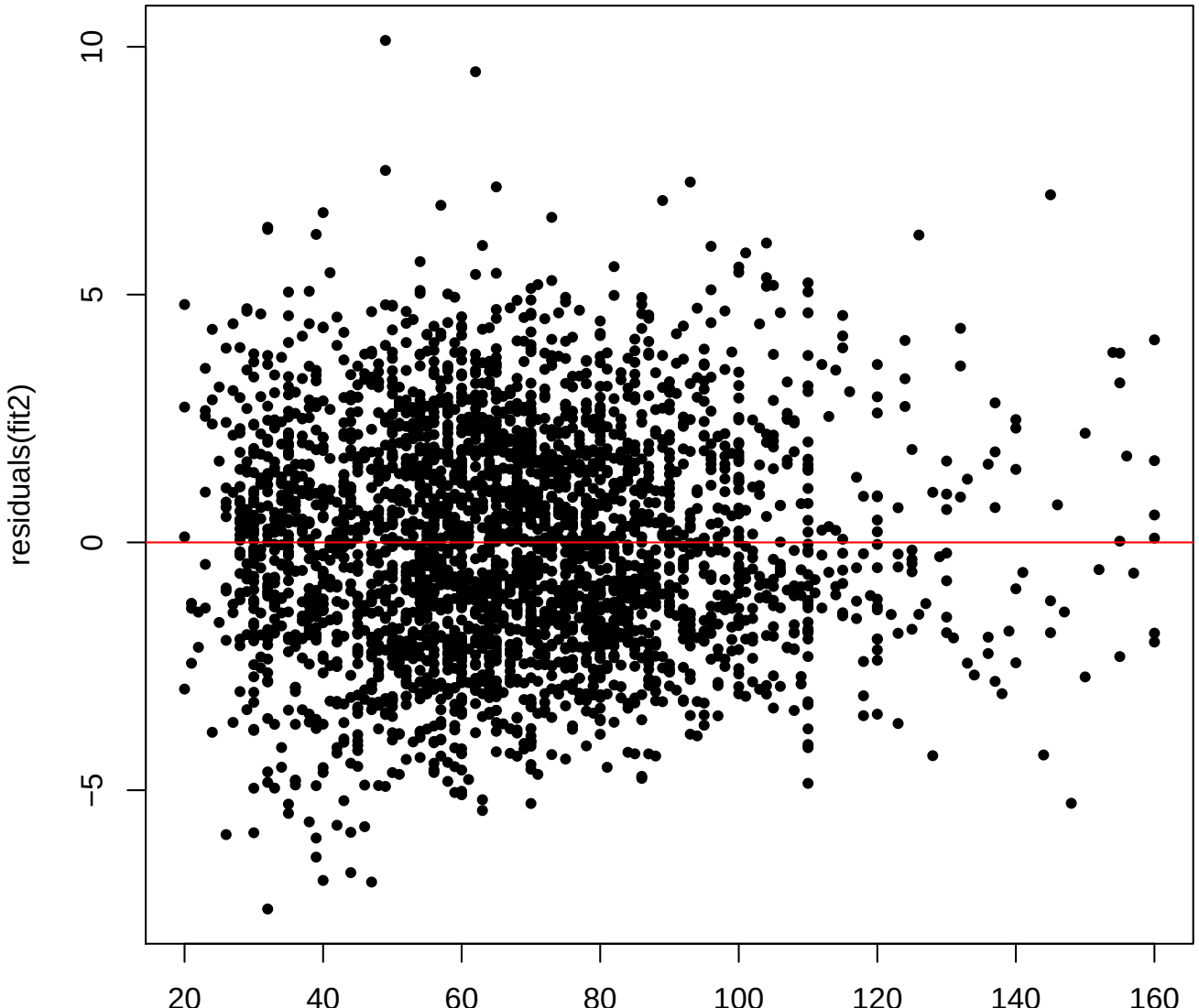
Linearity of covariates: Transformed covariate vs. response

Munich Rent Index: 1/area vs rentsqm



Linearity of covariates: Transformed covariate vs. residual (residual plot)

Munich Rent Index: 1/area vs residual



3.2 Modeling Nonlinear Covariate Effects Through Variable Transformation

If the continuous covariate z has an approximately nonlinear effect $\beta_1 f(z)$ with known transformation f, then the model

$$y_i = \beta_0 + \beta_1 f(z_i) + \ldots + \varepsilon_i$$

can be transformed into the linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \ldots + \varepsilon_i,$$

where $x_i = f(z_i) - \bar{f}$. By subtracting

$$\bar{f} = \frac{1}{n} \sum_{i=1}^{n} f(z_i),$$

the estimated effect $\hat{\beta}_1 x$ is automatically centered around zero. The estimated curve is best interpreted by plotting $\hat{\beta}_1 x$ against z (instead of x).

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.94)

3.3 Modeling Nonlinear Covariate Effects Through Polynomials

If the continuous covariate z has an approximately polynomial effect $\beta_1 z + \beta_2 z^2 + \ldots + \beta_l z^l$ of degree l, then the model

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \ldots + \beta_l z_i^l + \ldots + \varepsilon_i$$

can be transformed into the linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \ldots + \beta_l x_{il} + \ldots + \varepsilon_i,$$

where $x_{i1} = z_i$, $x_{i2} = z_i^2$, ..., $x_{il} = z_i^l$.
The centering (and possibly orthogonalization) of the vectors $\boldsymbol{x}^j = (x_{1j}, \ldots, x_{nj})'$, $j = 1, \ldots, l$, to $\boldsymbol{x}^1 - \bar{\boldsymbol{x}}_1, \ldots, \boldsymbol{x}^l - \bar{\boldsymbol{x}}_l$ with the mean vector $\bar{\boldsymbol{x}}_j = (\bar{x}_j, \ldots, \bar{x}_j)'$ facilitates interpretation of the estimated effects. A graphical illustration of the estimated polynomial is a useful way to interpret the estimated effect of z.

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.95)

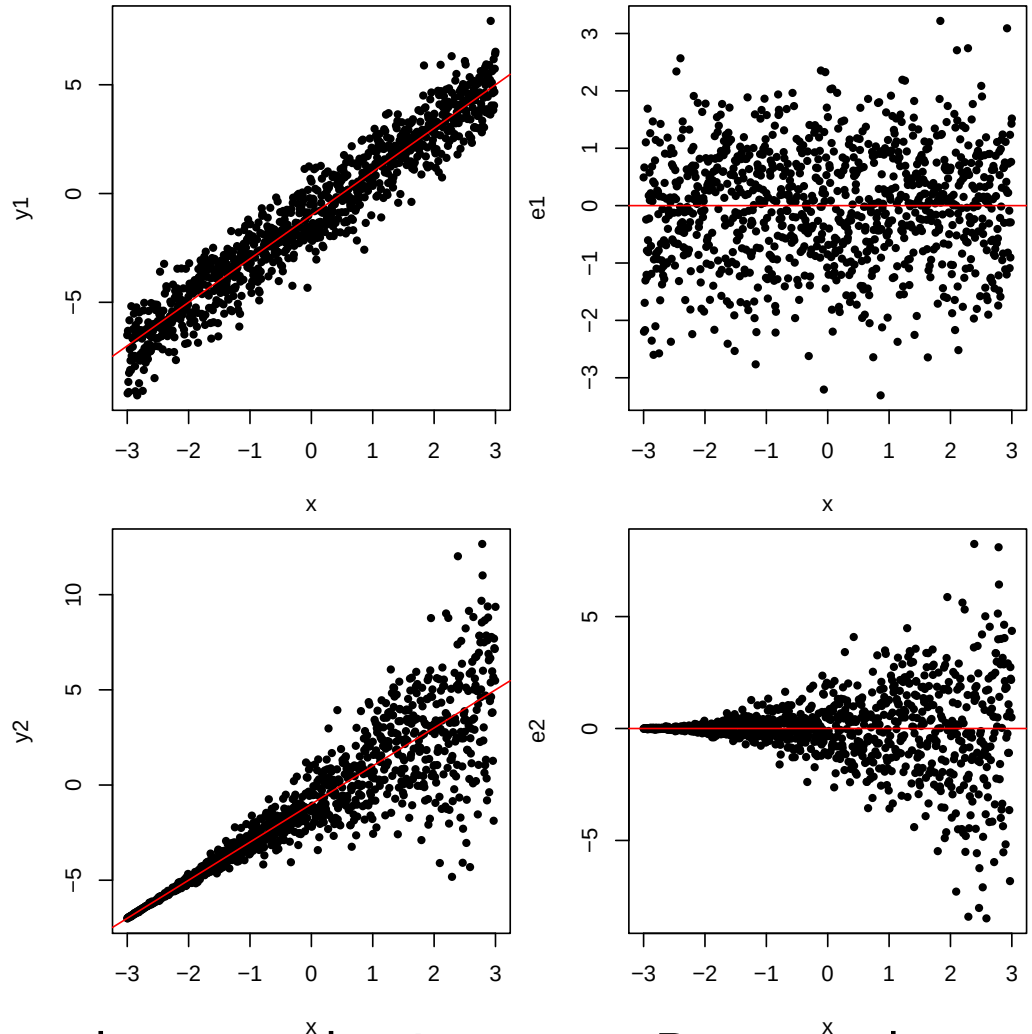
Homoscedastic errors

```
n=1000
x=seq(-3,3,length=n)
beta0=-1
beta1=2
xbeta=beta0+beta1*x
sigma=1
e1=rnorm(n,mean=0,sd=sigma)
y1=xbeta+e1
ehat1=residuals(lm(y1~x))
plot(x,y1,pch=20)
abline(beta0,beta1,col=1)
plot(x,e1,pch=20)
abline(h=0,col=2)
```

Heteroscedastic errors

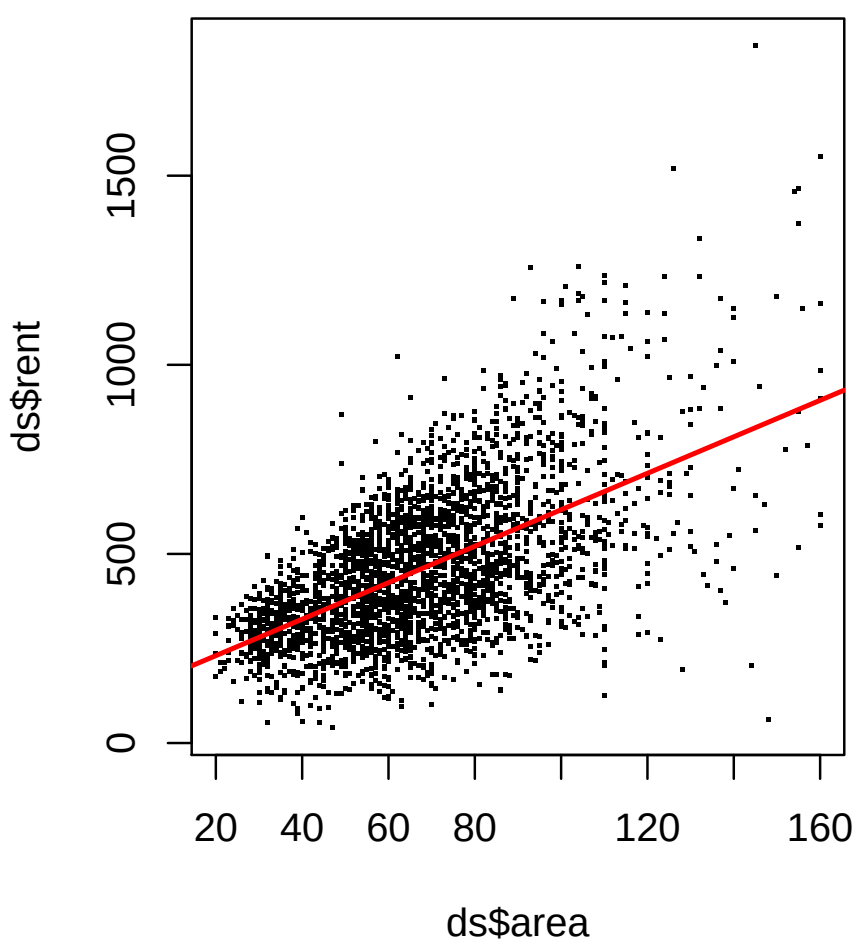
```
sigma=(0.1+0.3*(x+3))^2
e2=rnorm(n,0,sd=sigma)
y2=xbeta+e2
ehat2=residuals(lm(y2~x))
plot(x,y2,pch=20)
abline(beta0,beta1,col=2)
plot(x,e2,pch=20)
abline(h=0,col=2)
```

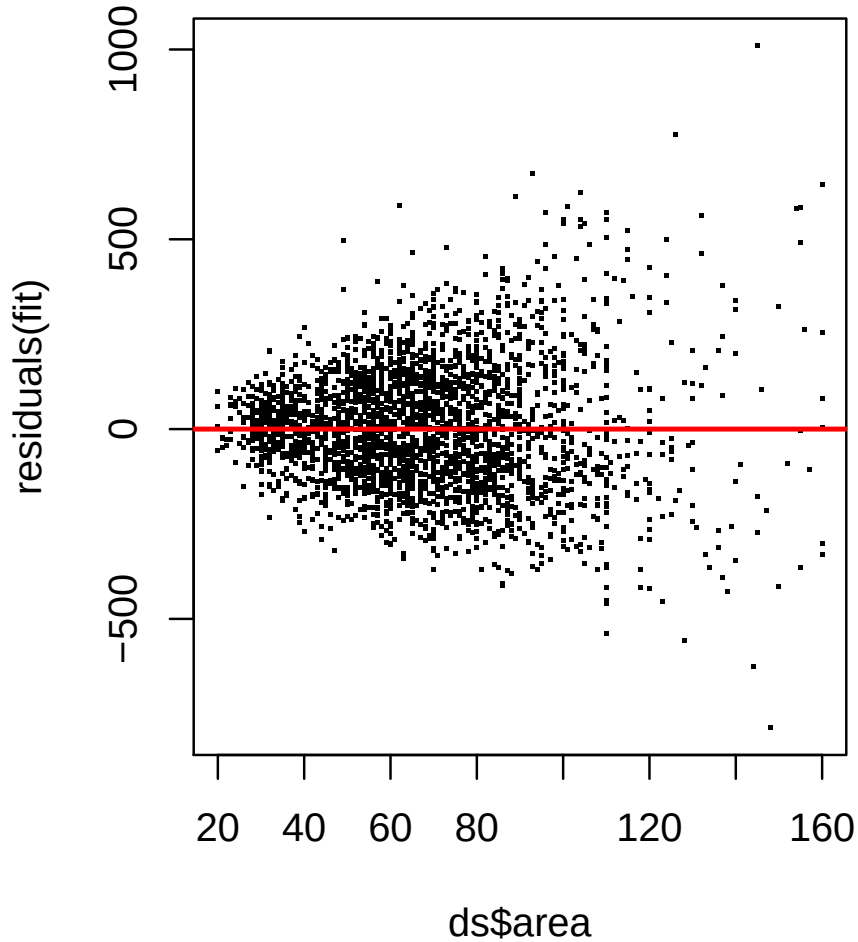
Homo- and heteroscedastic errors



Top: homoscedastic errors. Bottom: heteroscedastic errors. Right: x vs y. Left: x vs error. Example from Fahrmeir et al (2013): Regression. Springer. (p.79). R code from TMA4267 lectures tab.

Homoscedastic errors?





Left: area vs rent, right: area vs residuals. Fahrmeir et al (2013): Regression. Springer. (p.80). R code from TMA4267 lectures tab.

Today

- Normal linear model: implication for Y.
- Model parameters $\boldsymbol{\beta}, \sigma^2$, parameter estimators $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$, residuals $\hat{\boldsymbol{\varepsilon}} = Y - \boldsymbol{X}\hat{\boldsymbol{\beta}}$.
- Model assumptions.
- Next: covariates- how to include in linear regression, and then parameter estimation.