Module 2: MULTIPLE LINEAR REGRESSION

TMA4315 Generalized linear models H2018

Mette Langaas, Department of Mathematical Sciences, NTNU –
with contributions from Øyvind Bakke and Ingeborg Hem

30.08 and 06.09 [PL], 31.08 and 07.09 [IL]

(Latest changes: 29.08.2018. Theory for w1 is up to date, ILw1 is QCd.)

Overview

Learning material

- Textbook: Chapter 2.2, 3 and B.4. (Chapter 3 was on the reading list for TMA4267 Linear statistical 2016-2018, so much of this module is know from before - but not from a GLM point of view!)
- Classnotes 30.08.2018
- Classnotes 06.09.2018

Topics

First week

- Aim of multiple linear regression.
- Define and understand the multiple linear regression model - traditional and GLM way
- parameter estimation with maximum likelihood (and least squares)
- likelihood, score vector and Hessian (observed Fisher information matrix)
- properties of parameter estimators
- assessing model fit (diagnostic), residuals, QQ-plots
- design matric: how to code categorical covariates (dummy or effect coding), and how to handle interactions

Jump to IL for first week

Second week

- What did we do last week?
- big data implementation (if time)
- Statistical inference for parameter estimates
  - confidence intervals,
  - prediction intervals,
  - hypothesis test,
  - linear hypotheses
- SSE and deviance
- analysis of variance decompositions and R^2, sequential ANOVA table
- model selection with AIC and variants

Jump to IL for second week

FIRST WEEK

Aim of multiple linear regression

1. Construct a model to help understand the relationship between a response and one or several explanatory variables. [Correlation, or cause and effect?]
2. Construct a model to predict the reponse from a set of (one or several) explanatory variables. [More or less "black box"]

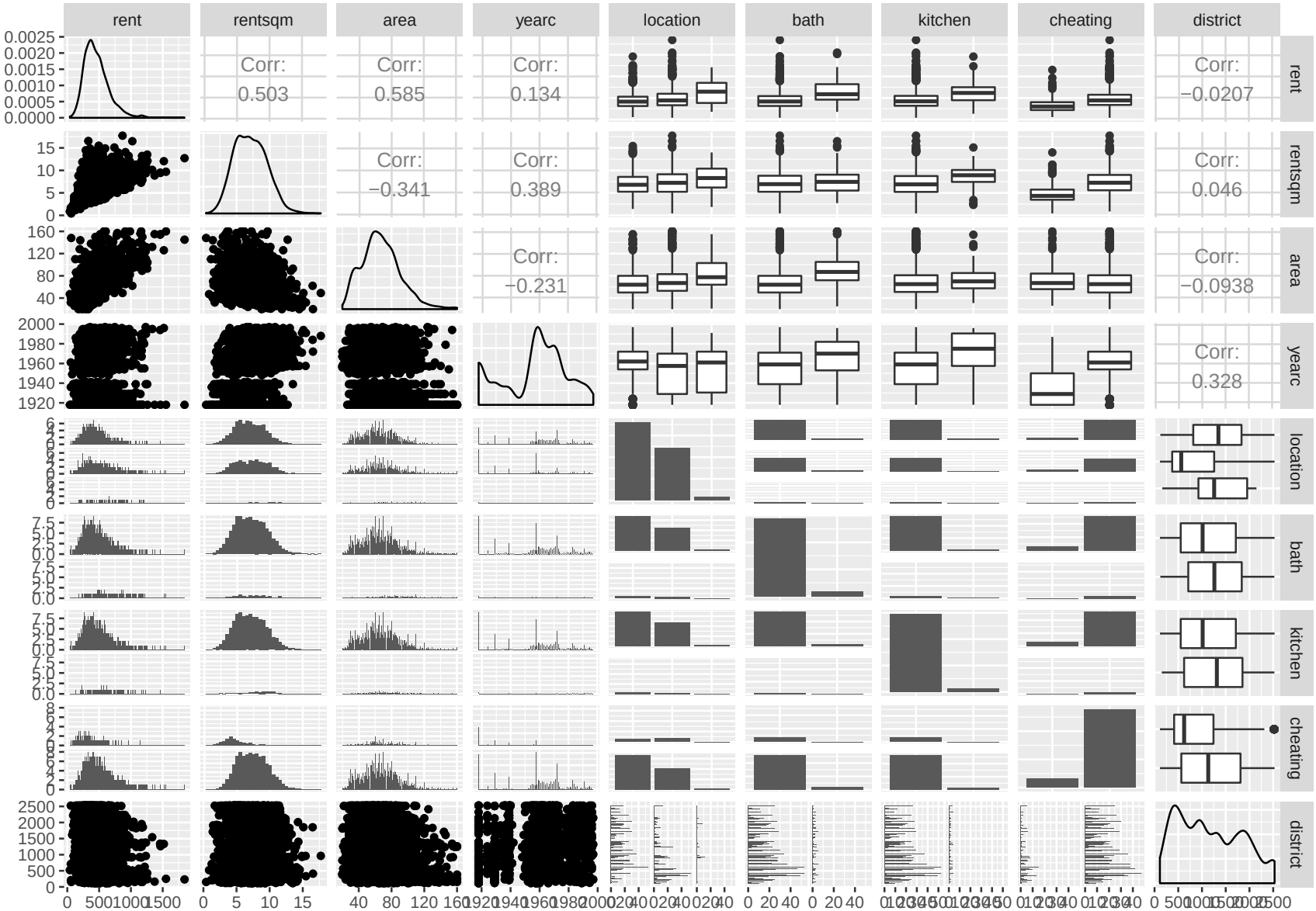
Munich rent index

Munich, 1999: 3082 observations on 9 variables.

- `rent`: the net rent per month (in Euro).
- `rentsqm`: the net rent per month per square meter (in Euro).
- `area`: living area in square meters.
- `yearc`: year of construction.
- `location`: quality of location: a factor indicating whether the location is average location, 1, good location, 2, and top location, 3.
- `bath`: quality of bathroom: a a factor indicating whether the bath facilities are standard, 0, or premium, 1.
- `kitchen`: Quality of kitchen: 0 standard 1 premium.
- `cheating`: central heating: a factor 0 without central heating, 1 with central heating.
- `district`: District in Munich.

More information in Fahrmeir et. al., (2013) page 5.

```
library("gamlss.data")
library(GGally)
ggpairs(rent99, lower = list(combo = wrap(ggally_facethist, binwidth =
```



Interesting questions

1. Is there a relationship between `rent` and `area`?
2. How strong is this relationship?
3. Is the relationship linear?
4. Are also other variables associated with `rent`?
5. How well can we predict the rent of an appartment?
6. Is the effect of `area` the same on `rent` for appartments at average, good and top `location`? (interaction)

Notation

$\mathbf{Y}$: $(n \times 1)$ vector of responses (random variable) [e.g. one of the following: rent, rent pr sqm, weight of baby, ph of lake, volume of tree]

$\mathbf{X}$: $(n \times p)$ design matrix [e.g. location of flat, gestation age of baby, chemical measurement of the lake, height of tree]

β : $(p \times 1)$ vector of regression parameters (intercept included, so $p = k + 1$)

ε : $(n \times 1)$ vector of random errors. Used in "traditional way".

We assume that pairs $(\mathbf{x}_i^T, y_i)$ $(i = 1, ..., n)$ are measured from sampling units. That is, the observation pair $(\mathbf{x}_1^T, y_1)$ is independent from $(\mathbf{x}_2^T, y_2)$, and so on.

Hands on: Munich rent index — response and covariates

Study the print-out and discuss the following questions:

1. What can be response, and what covariates? (using what you know about rents)
2. What type of response(s) do we have? (continuous, categorical, nominal, ordinal, discrete, factors, ...).
3. What types of covariates? (continuous, categorical, nominal, ordinal, discrete, factors, ...)
4. Explain what the elements of `model.matrix` are. (Hint: coding of `location`)

```r
library("gamlss.data")
ds = rent99
colnames(ds)
summary(ds)
dim(ds)
head(ds)
str(ds$location)
contrasts(ds$location)

X = model.matrix(rentsqm ~ area + yearc + location + bath +
    cheating + district, data = ds)
head(X)
```

```
## [1] "rent"     "rentsqm"  "area"     "yearc"    "locatio
## [7] "kitchen"  "cheating" "district"
##       rent             rentsqm            area
##  Min.   :  40.51   Min.   : 0.4158   Min.   : 20.00   Mi
##  1st Qu.: 322.03   1st Qu.: 5.2610   1st Qu.: 51.00   1s
##  Median : 426.97   Median : 6.9802   Median : 65.00   Me
##  Mean   : 459.44   Mean   : 7.1113   Mean   : 67.37   Me
```

Model

The traditional way

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

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

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

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

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

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

The GLM way

Independent pairs $(Y_i, \mathbf{x}_i)$ for $i = 1, \ldots, n$.

1. Random component: $Y_i \sim N$ with $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{Var}(Y_i) = \sigma^2$.
2. Systematic component: $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$.
3. Link function: linking the random and systematic component (linear predictor): Identity link and response function. $\mu_i = \eta_i$.

Questions

- Compare the traditional and GLM way. Have we made the same assumptions for both?
- What is the connection between each $\mathbf{x}_i$ and the design matrix?
- What is "full rank"? Why is this needed? Example of rank less than p?
- Why do you think we move from traditional to GLM way? Could we not just let ε be from binomial, Poisson, etc. distribution?

Parameter estimation

In multiple linear regression there are two popular methods for estimating the regression parameters in β: maximum likelihood and least squares. These two methods give the same estimator when we assume the normal linear regression model. We will in this module focus on maximum likelihood estimation, since that can be used also when we have non-normal responses (modules 3-6: binomial, Poisson, gamma, multinomial).

Likelihood theory (from B.4)

Likelihood $L(\beta)$

We assume that pairs of covariates and response are measured independently of each other: $(\mathbf{x}_i, Y_i)$, and Y_i follows the distribution specified above, and $\mathbf{x}_i$ is fixed.

$$L(\beta) = \prod_{i=1}^{n} L_i(\beta) = \prod_{i=1}^{n} f(y_i; \beta)$$

Q: fill in with the normal density for f and the multiple linear regression model.

Loglikelihood $l(\beta)$

The log-likelihood is just the natural log of the likelihood, and we work with the log-likelihood because this makes the mathematics simpler - since we work with exponential families. The main aim with the likelihood is to maximize it to find the maximum likelihood estimate, and since the log is a monotone function the maximum of the log-likelihood will be in the same place as the maximum of the likelihood.

$$l(\beta) = \ln L(\beta) = \sum_{i=1}^{n} \ln L_i(\beta) = \sum_{i=1}^{n} l_i(\beta)$$

Observe that the log-likelihood is a sum of invidual contributions for each observation pair i.

Q: fill in with the normal density for f and the multiple linear regression model.

Repetition: rules for derivatives with respect to vector

Hardle and Simes (2015), page 65.

- Let β be a p-dimensional column vector of interest,
- and let $\frac{\partial}{\partial \beta}$ denote the p-dimensional vector with partial derivatives wrt the p elements of β.
- Let $\mathbf{d}$ be a p-dimensional column vector of constants and
- $\mathbf{D}$ be a $p \times p$ symmetric matrix of constants.

Rule 1:

$$\frac{\partial}{\partial \beta}(\mathbf{d}^T \beta) = \frac{\partial}{\partial \beta}(\sum_{j=1}^{p} d_j \beta_j) = \mathbf{d}$$

Rule 2:

$$\frac{\partial}{\partial \beta}(\beta^T \mathbf{D} \beta) = \frac{\partial}{\partial \beta}(\sum_{j=1}^{p} \sum_{k=1}^{p} \beta_j d_{jk} \beta_k) = 2\mathbf{D}\beta$$

Rule 3: The Hessian of the quadratic form $\beta^T \mathbf{D} \beta$ is

$$\frac{\partial^2 \beta^T \mathbf{D} \beta}{\partial \beta \partial \beta^T} = 2\mathbf{D}$$

Score function $s(\beta)$

The score function is a $p \times 1$ vector, $s(\beta)$, with the partial derivatives of the log-likelihood with respect to the p elements of the β vector.

$$s(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial l_i(\beta)}{\partial \beta} = \sum_{i=1}^{n} s_i(\beta)$$

Again, observe that the score function is a sum of individual contributions for each observation pair i.

Q: fill in for the multiple linear regression model.

To find the maximum likelihood estimate $\hat{\beta}$ we solve the set of p equations:

$$s(\hat{\beta}) = 0$$

Q: fill in for the multiple linear regression model. Specify what the *normal equations* are.

For the normal linear regression model, these equations $s(\hat{\beta}) = 0$ have a solution to be written on closed form.

Least squares and maximum likelihood (ML) estimator for β:

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

Q: Least squares is found by minimizing $\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2$. How can you see that least squares and ML gives the same estimator?

Looking ahead: Hessian and Fisher information

But, for other distribution than the normal we get a set of non-linear equations when we look at $s(\hat{\beta}) = 0$, and then we will use the Newton-Raphson or Fisher Scoring iterative methods.

Observed Fisher information matrix $H(\beta)$

$$H(\beta) = -\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = -\frac{\partial s(\beta)}{\partial \beta^T}$$

so this is minus the Hessian of the loglikelihood.

- $H(\beta)$ may be considered as a *local measure of information* that the likelihood contains.
- The higher the curvature of the log-likelihood near its maximum the more information is provide by the likelihood about the unknown parameter.

Q: Calculate this for the multiple linear regression model. What is the dimension of $H(\beta)$?

In addition we also use the *expected Fisher information matrix* $F(\beta)$ which we may find in two ways, one is by taking the mean of the observed Fisher information matrix:

$$F(\beta) = E\left(-\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T}\right).$$

Q: Calculate this for the multiple linear regression model. What is the dimension of $F(\beta)$?

In Module 3 we need the Fisher information matrix in the Newton-Raphson method, and also to find the (asympotic) covariance matrix of our estimated coefficents $\hat{\beta}$ - so much more about this then.

Hands on: Munich rent index parameter estimates

Explain what the values under `Estimate` mean in practice.

```
fit = lm(rentsqm ~ area + yearc + location + bath + kitchen
    data = ds)
summary(fit)
```

```
##
## Call:
## lm(formula = rentsqm ~ area + yearc + location + bath +
##     cheating, data = ds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4303 -1.4131 -0.1073  1.3244  8.6452
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -45.475484   3.603775 -12.619  < 2e-16 ***
```

Projection matrices: idempotent, symmetric/orthogonal

(Optional - known from TMA4267)

First, we define predictions as $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$, and inserted the ML (and LS) estimate we get $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

We define the projection matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

called the *hat matrix*. This simplifies the notation for the predictions,

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

so the hat matrix is putting the hat on the response $\mathbf{Y}$.

In addition we define residuals as

$$\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$$
$$\hat{\varepsilon} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

so we have a second projection matrix

$$\mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

Geometry of Least Squares — involving our two projection matrices

(Optional - known from TMA4267)

- Mean response vector: $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\beta$
- As β varies, $\mathbf{X}\beta$ spans the model plane of all linear combinations. I.e. the space spanned by the columns of $\mathbf{X}$: the column-space of $\mathbf{X}$.
- Due to random error (and unobserved covariates), $\mathbf{Y}$ is not exactly a linear combination of the columns of $\mathbf{X}$.
- LS-estimation chooses $\hat{\beta}$ such that $\mathbf{X}\hat{\beta}$ is the point in the column-space of $\mathbf{X}$ that is closes to $\mathbf{Y}$.
- The residual vector $\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$ is perpendicular to the column-space of $\mathbf{X}$.
- Multiplication by $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ projects a vector onto the column-space of $\mathbf{X}$.
- Multiplication by $\mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ projects a vector onto the space perpendicular to the column-space of $\mathbf{X}$.

Restricted maximum likelihood estimator for σ^2

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{Y} - \mathbf{X}\hat{\beta})^T(\mathbf{Y} - \mathbf{X}\widehat{\beta}) = \frac{\text{SSE}}{n-p}$$

In the generalized linear models setting (remember exponential family from Module 1) we will look at the parameter σ^2 as a nuisance parameter $=$ parameter that is not of interest to us. Our our focus will be on the parameters of interest - which will be related to the mean of the response, which is modelled using our covariate - so the regression parameters β are therefore our prime focus.

However, to perform inference we need an estimator for σ^2.

The maximum likelihood estimator for σ^2 is $\frac{\text{SSE}}{n}$, which is found from maximizing the likelihood inserted our estimate of $\hat{\beta}$

$$L(\hat{\beta}, \sigma^2) = (\frac{1}{2\pi})^{n/2}(\frac{1}{\sigma^2})^{n/2} \exp(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}))$$

$$\begin{aligned} I(\hat{\beta}, \sigma^2) &= \ln(L(\hat{\beta}, \sigma^2)) \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) \end{aligned}$$

The score vector with respect to σ^2 is

$$\frac{\partial I}{\partial \sigma^2} = 0 - \frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$$

Solving $\frac{\partial I}{\partial \sigma^2} = 0$ gives us the estimator

$$\frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = \frac{\text{SSE}}{n}$$

But, this estimator is biased.

To prove this you may use the trace-formula, that is $\mathrm{E}(\mathbf{Y}^T\mathbf{A}\mathbf{Y}) = \mathrm{tr}(\mathbf{A}\mathrm{Cov}(\mathbf{Y})) + \mathrm{E}(\mathbf{Y})^T\mathbf{A}\mathrm{E}(\mathbf{Y})$, and we use that $\mathrm{SSE} = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}$. This was done in class notes from TMA4267 - lecture 10

But, the estimator is *asympotically* unbiased (unbiased when the sample size n increases to infinity).

When an unbiased version is preferred, it is found using *restricted maximum likelihood* (REML). We will look into REML-estimation in Module 7. In our case the (unbiased) REML estimate is

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = \frac{\text{SSE}}{n-p}$$

The restricted maximum likelihood estimate is used in `lm`.

Q: What does it mean that the REML estimate is unbiased? Where is the estimate $\hat{\sigma}$ in the regression output? (See output from `lm` for the rent index example.)

Properties for the normal linear model

To be able to do inference (=make confidence intervals, prediction intervals, test hypotheses) we need to know about the properties of our parameter estimators in the (normal) linear model.

- Least squares and maximum likelihood estimator for β:

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

  with $\hat{\beta} \sim N_p(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.
- Restricted maximum likelihood estimator for σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{Y} - \mathbf{X}\hat{\beta})^T(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \frac{\text{SSE}}{n-p}$$

  with $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$.
- Statistic for inference about β_j, c_{jj} is diagonal element j of $(\mathbf{X}^T\mathbf{X})^{-1}$.

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t_{n-p}$$

  This requires that $\hat{\beta}_j$ and $\hat{\sigma}$ are independent (see below).

However, when we work with *large samples* then $n - p$ becomes large and the t distribution goes to a normal distribution, so we may use the standard normal in place of the t_{n-p}.

Asymptotically we have:

$$\hat{\beta} \sim N_p(\beta, \tilde{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1})$$

and

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\tilde{\sigma}} \sim N(0,1)$$

where $\tilde{\sigma}^2 = \frac{\text{SSE}}{n}$ (the ML estimator).

Q: Pointing forwards: do you see any connection between the covariance matrix of $\hat{\beta}$ and the Fisher information?

Are $\hat{\beta}$ and SSE are independent? (optional)

Independence: Let $\mathbf{X}_{(p\times 1)}$ be a random vector from $N_p(\mu, \mathbf{\Sigma})$. Then $\mathbf{AX}$ and $\mathbf{BX}$ are independent iff $\mathbf{A\Sigma B}^T = \mathbf{0}$.

- $\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2\mathbf{I})$
- $\mathbf{AY} = \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, and
- $\mathbf{BY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$.
- Now $\mathbf{A}\sigma^2\mathbf{I}\mathbf{B}^T = \sigma^2\mathbf{A}\mathbf{B}^T = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{H}) = \mathbf{0}$
- since $\mathbf{X}(\mathbf{I} - \mathbf{H}) = \mathbf{X} - \mathbf{HX} = \mathbf{X} - \mathbf{X} = \mathbf{0}$.
- We conclude that $\hat{\beta}$ is independent of $(\mathbf{I} - \mathbf{H})\mathbf{Y}$,
- and, since SSE=function of $(\mathbf{I} - \mathbf{H})\mathbf{Y}$: SSE=$\mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}$,
- then $\hat{\beta}$ and SSE are independent, and the result with T_j being t-distributed with $n - p$ degrees of freedom is correct.

Remark: a similar result will exist for GLMs, using the concept of *orthogonal parameters*.

Checking model assumptions

In the normal linear model we have made the following assumptions.

1. Linearity of covariates: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$. Problem: non-linear relationship?
2. Homoscedastic error variance: $\text{Cov}(\varepsilon) = \sigma^2\mathbf{I}$. Problem: Non-constant variance of error terms
3. Uncorrelated errors: $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$.
4. Additivity of errors: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$
5. Assumption of normality: $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$

The same assumptions are made when we do things the GLM way for the normal linear model.

In addtion the following might cause problems:

- Outliers
- High leverage points
- Collinearity

General theory on QQ-plots

Read this for yourself. You do not need to understand this in detail, but is useful to have a basic idea why we look for a straight line in a QQ-plot. There is one question about this in the ILw1.

Go to separate R Markdown or html document: QQ–plot as Rmd or QQ–plot as html

Residuals

If we assume the normal linear model then we know that the residuals ($n \times 1$ vector)

$$\hat{\varepsilon} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

has a normal (singular) distribution with mean $\mathrm{E}(\hat{\varepsilon}) = \mathbf{0}$ and covariance matrix $\mathrm{Cov}(\hat{\varepsilon}) = \sigma^2(\mathbf{I} - \mathbf{H})$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

This means that the residuals (possibly) have different variance, and may also be correlated.

Our best guess for the error ε is the residual vector $\hat{\varepsilon}$, and we may think of the residuals as *predictions of the errors*. Be aware: don't mix errors (the unobserved) with the residuals ("observed").

But, we see that the residuals are not independent and may have different variance, therefore we will soon define variants of the residuals that we may use to assess model assumptions after a data set is fitted.

Q: How can we say that the residuals can have different variance and may be correlated? Why is that a problem?

We would like to check the model assumptions - we see that they are all connected to the error terms. But, but we have not observed the error terms ε so they can not be used for this. However, we have made "predictions" of the errors - our residuals. And, we want to use our residuals to check the model assumptions.

That is, we want to check that our errors are independent, homoscedastic (same variance for each observation), and not dependent on our covariates - and we want to use the residuals (observed) in place of the errors (unobserved). Then it would have been great if the residuals have these properties when the underlying errors have. To amend our problem we need to try to fix the residual so that they at least have equal variances. We do that by working with *standardized* or *studentized residuals*.

Standardized residuals:

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

where $h_i i$ is the ith diagonal element of the hat matrix $\mathbf{H}$.

In R you can get the standardized residuals from an `lm`-object (named `fit`) by `rstandard(fit)`.

Studentized residuals:

$$r_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}$$

where $\hat{\sigma}_{(i)}$ is the estimated error variance in a model with observation number i omitted. This seems like a lot of work, but it can be shown that it is possible to calculated the studentized residuals directly from the standardized residuals:

$$r_i^* = r_i\left(\frac{n - p - 1}{n - p - r_i^2}\right)^{1/2}$$

Plotting residuals - and what to do when assumptions are violated?

Some important plots

1. Plot the residuals, r_i^* against the predicted values, $\hat{y}_i$.

   - Dependence of the residuals on the predicted value: wrong regression model?
   - Nonconstant variance: transformation or weighted least squares is needed?

2. Plot the residuals, r_i^*, against predictor variable or functions of predictor variables. Trend suggest that transformation of the predictors or more terms are needed in the regression.

3. Assessing normality of errors: QQ-plots and histograms of residuals. As an additional aid a test for normality can be used, but must be interpreted with caution since for small sample sizes the test is not very powerful and for large sample sizes even very small deviances from normality will be labelled as significant.
4. Plot the residuals, r_i^*, versus time or collection order (if possible). Look for dependence or autocorrelation.

Residuals can be used to check model assumptions, and also to *discover outliers*.

Diagnostic plots in R

More information on the plots here:
http://data.library.virginia.edu/diagnostic-plots/ and
http://ggplot2.tidyverse.org/reference/fortify.lm.html

You can use the function `fortify.lm` in `ggplot2` to create a dataframe from an `lm`-object, which `ggplot` uses automatically when given a `lm`-object. This can be used to plot diagnostic plots.

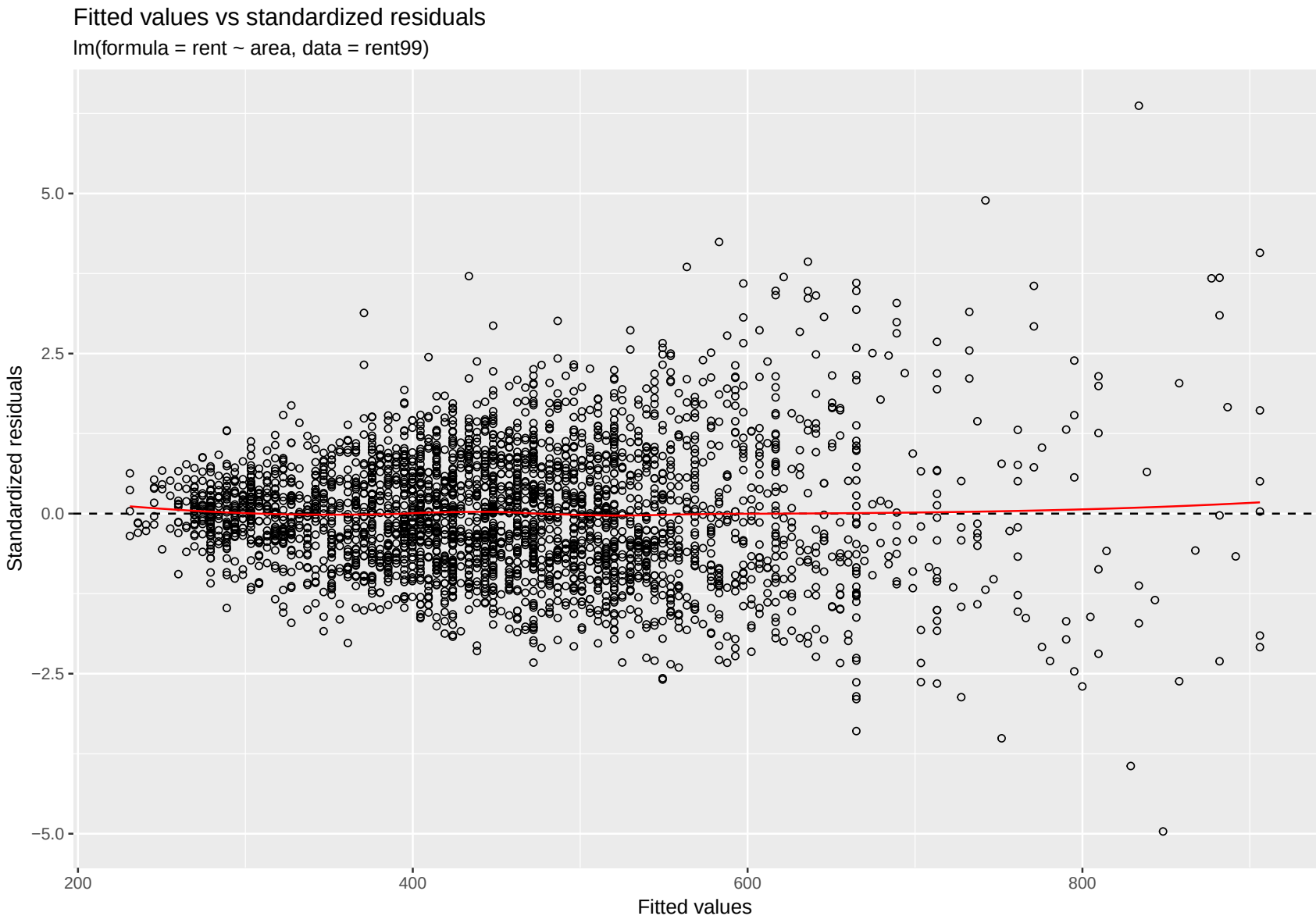
For simplicity we use the Munch rent index with `rent` as response and only `area` as the only covariate. (You may change the model to a more complex one, and rerun the code chunks.)

```
##    rent area     .hat .sigma   .cooksd .fitted  .resid .stdresid
## 1 109.9   26 0.001312  158.8 5.870e-04   260.0 -150.00   -0.9454
## 2 243.3   28 0.001219  158.8 1.678e-05   269.6  -26.31   -0.1658
## 3 261.6   30 0.001130  158.8 6.956e-06   279.2  -17.60   -0.1109
## 4 106.4   30 0.001130  158.8 6.711e-04   279.2 -172.83   -1.0891
## 5 133.4   30 0.001130  158.8 4.779e-04   279.2 -145.85   -0.9191
## 6 339.0   30 0.001130  158.8 8.032e-05   279.2   59.79    0.3768
```

Residuals vs fitted values

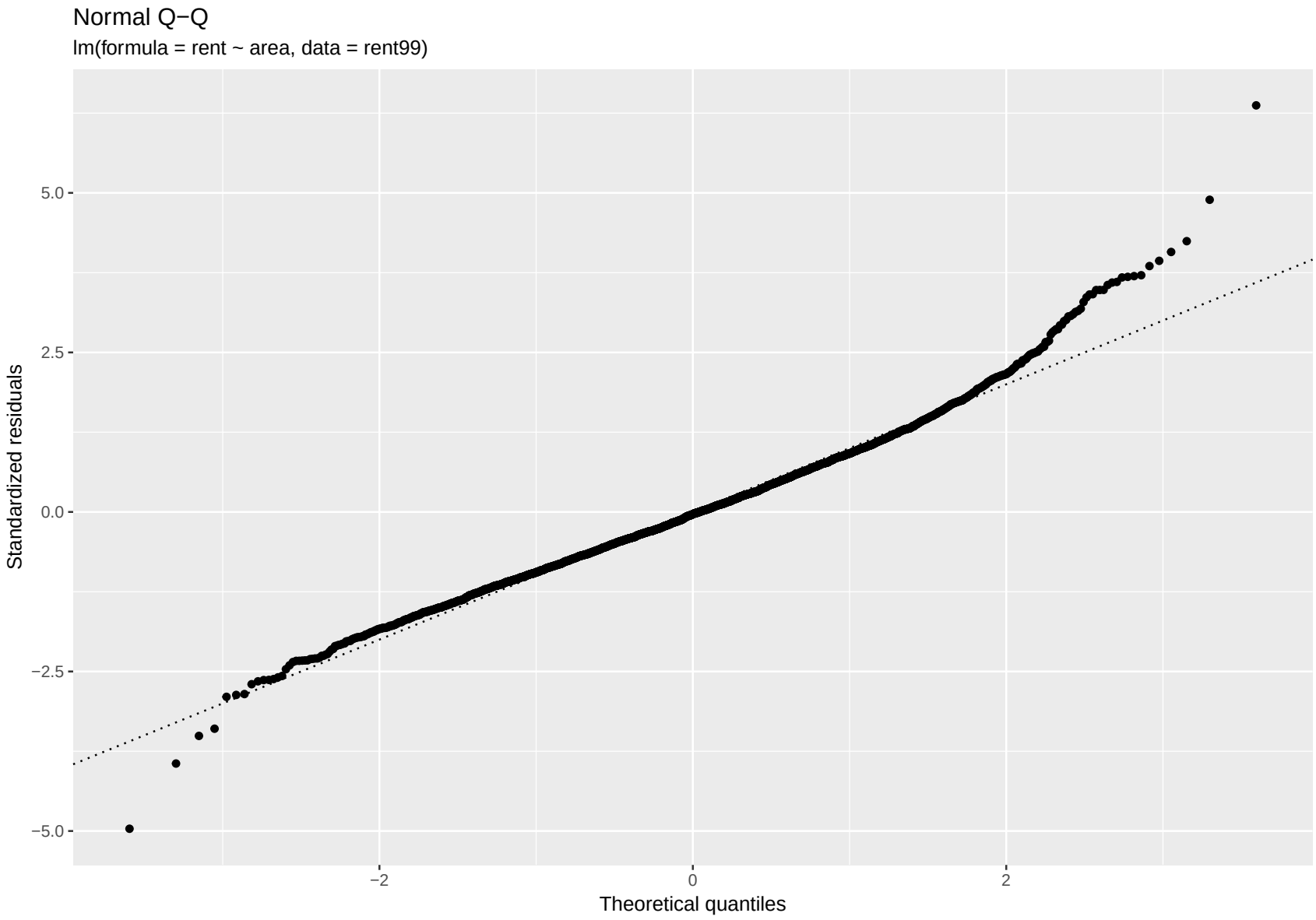
A plot with the fitted values of the model on the x-axis and the residuals on the y-axis shows if the residuals have non-linear patterns. The plot can be used to test the assumption of a linear relationship between the response and the covariates. If the residuals are spread around a horizontal line with no distinct patterns, it is a good indication on no non-linear relationships, and a good model.

Does this look like a good plot for this data set?


Fitted values vs standardized residuals
lm(formula = rent ~ area, data = rent99)
Standardized residuals
5.0
2.5
0.0
−2.5
−5.0
200
400
600
800
Fitted values

Normal Q-Q

This plot shows if the residuals are Gaussian (normally) distributed. If they follow a straigt line it is an indication that they are, and else they are probably not.


Normal Q−Q
lm(formula = rent ~ area, data = rent99)
Standardized residuals
5.0
2.5
0.0
−2.5
−5.0
−2
0
2
Theoretical quantiles

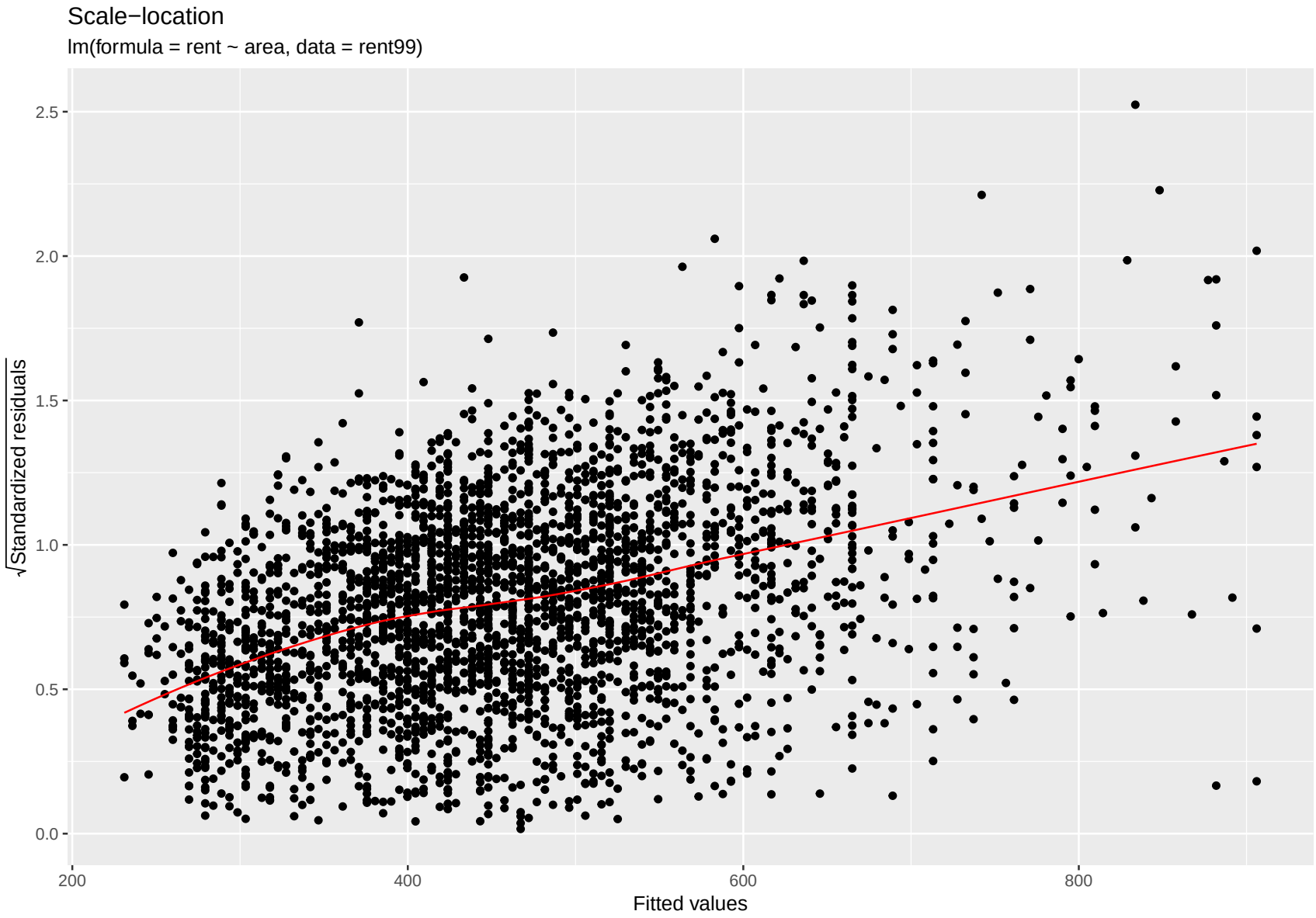
```r
library(nortest)
ad.test(rstudent(fit))
```

```
##
##  Anderson-Darling normality test
##
## data:  rstudent(fit)
## A = 6.4123, p-value = 9.809e-16
```

Scale-location

This is also called spread-location plot. It shows if the residuals are spread equally along the ranges of predictors. Can be used to check the assumption of equal variance (homoscedasticity). A good plot is one with a horizontal line with randomly spread points.

Is this plot good for your data?

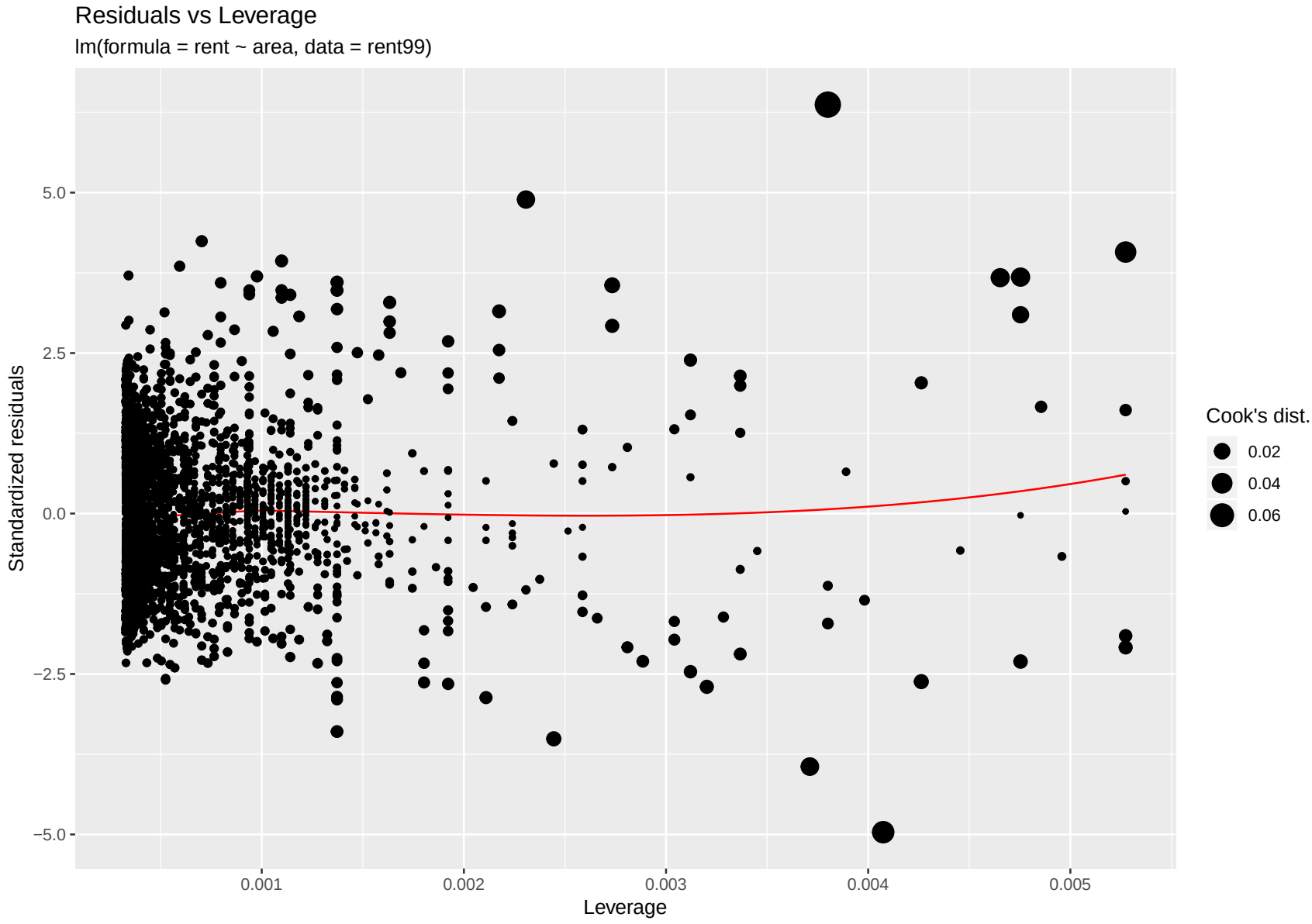

Scale−location
lm(formula = rent ~ area, data = rent99)
√Standardized residuals
2.5
2.0
1.5
1.0
0.5
0.0
200
400
600
800
Fitted values

Residual vs Leverage

This plot can reveal influential outliers. Not all outliers are influential in linear regression; even though data have extreme values, they might not be influential to determine the regression line (the results don't differ much if they are removed from the data set). These influential outliers can be seen as observations that does not get along with the trend in the majority of the observations. In `plot.lm`, dashed lines are used to indicate the Cook's distance, instead of using the size of the dots as is done here.

Cook's distance is the Euclidean distance between the $\widehat{\mathbf{y}}$ (the fitted values) and $\widehat{\mathbf{y}}_{(i)}$ (the fitted values calculated when the i-th observation is omitted from the regression). This is then a measure on how much the model is influences by observation i. The distance is scaled, and a rule of thumb is to examine observations with Cook's distance larger than 1, and give some attention to those with Cook's distance above 0.5.

Leverage is defined as the diagonal elements of the hat matrix, i.e., the leverage of the i-th data point is h_{ii} on the diagonal of $\mathbf{H} = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}$. A large leverage indicated that the observation (i) has a large influence on the estimation results, and that the covariate values $(\mathbf{x}_i)$ are unusual.



(Some observations does not fit our model, but if we fit a more complex model this may change.)

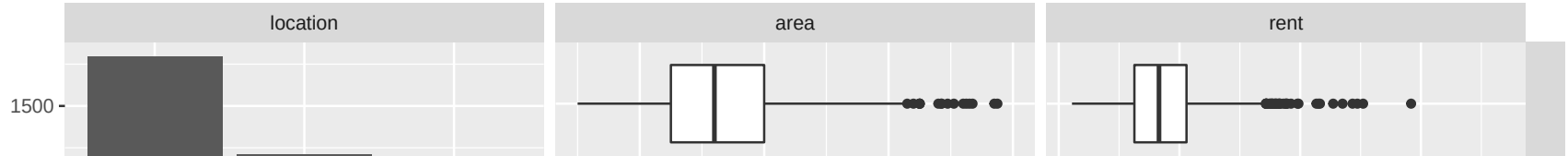
Categorical covariates - dummy and effect coding

(read for yourself - topic of ILw1)

Example: consider our `rent` dataset with `rent` as reponse, and continuous covariate `area` and categorical covariate `location`. Let the `location` be a factor with levels `average, good, excellent`.

```
library(gamlss.data)
library(tidyverse)
library(GGally)
```

```
ds = rent99 %>% select(location, area, rent)
levels(ds$location)
# change to meaningful names
levels(ds$location) = c("average", "good", "excellent")
ggpairs(ds)
```



Effect coding aka sum-zero-contrast:

This is an equally useful and popular coding - and this is the coding that is preferred when working with analysis of variance in general. The effect coding assumes that the sum of the effects for the levels of the factor sums to zero, and this is done with the following coding scheme (Model 3 with the original location and 4 with the releveled version.)

```
X3 = model.matrix(~area + location, data = ds, contrasts =
X3[c(1, 3, 69), ]
X4 = model.matrix(~area + locationRELEVEL, data = ds, cont
X4[c(1, 3, 69), ]
```

```
##    (Intercept) area location1 location2
## 1            1   26         0         1
## 3            1   30         1         0
## 69           1   55        -1        -1
##    (Intercept) area locationRELEVEL1 locationRELEVEL2
## 1            1   26                1                0
```

Interactions

(read for yourself)

To illustrate how interactions between covariates can be included we use the `ozone` data set from the `ElemStatLearn` library. This data set is measurements from 1973 in New York and contains 111 observations of the following variables:

- `ozone` : ozone concentration (ppm)
- `radiation` : solar radiation (langleys)
- `temperature` : daily maximum temperature (F)
- `wind` : wind speed (mph)

We start by fitting a multiple linear regression model to the data, with ozone as our response variable and `temperature` and `wind` as covariates.

ozone	radiation	temperature	wind
41	190	67	7.4
36	118	72	8.0
12	149	74	12.6
18	313	62	11.5
23	299	65	8.6
19	99	59	13.8

```
##
## Call:
## lm(formula = ozone ~ temperature + wind, data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.160 -13.209  -3.089  10.588  98.470
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

Interactive lectures- problem set first week

Theoretical questions

Problem 1

1. Write down the GLM way for the multiple linear regression model. Explain.
2. Write down the likelihood and loglikelihood. Then define the score vector. What is the set of equations we solve to find parameter estimates? What if we could not find a closed form solution to our set of equations - what could we do then?
3. Define the observed and the expected Fisher information matrix. What dimension does these matrices have? What can these matrices tell us?

4. A core finding is $\hat{\beta}$.

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

with $\hat{\beta} \sim N_p(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.

Show that $\hat{\beta}$ has this distribution with the given mean and covariance matrix. What does this imply for the distribution of the jth element of $\hat{\beta}$? In particular, how can we calculate the variance of $\hat{\beta}_j$?

5. Explain the difference between *error* and *residual*. What are the properties of the raw residuals? Why don't we want to use the raw residuals for model check? What is our solution to this?
6. That is the theoretical intercept and slope of a QQ–plot based on a normal sample? Hint: QQ–plot as html

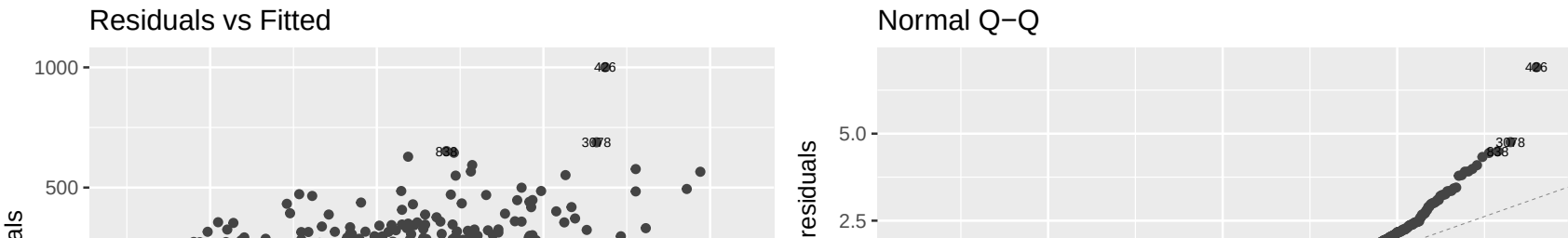
Interpretation and understanding

Problem 2: Munich Rent Index data

Fit the regression model with first `rent` and then `rentsqm` as reponse and following covariates: `area`, `location` (dummy variable coding using location2 and location3), `bath`, `kitchen` and `cheating` (central heating).

```
library(gamlss.data)
library(ggfortify)
`?`(rent99)

mod1 <- lm(rent ~ area + location + bath + kitchen + cheati
mod2 <- lm(rentsqm ~ area + location + bath + kitchen + che
autoplot(mod1, label.size = 2)
```



Problem 3: Simple vs. multiple regression

We look at a regression problem where both the response and the covariates are centered - that is, the mean of the response and the mean of each covariate is zero. We do this to avoid the intercept term, which makes things a bit more complicated.

1. In a design matrix (without an intercept column) orthogonal columns gives diagonal $\mathbf{X}^T\mathbf{X}$. What does that mean? How can we get orthogonal columns?
2. If we have orthogonal columns, will then simple (only one covariate) and multiple estimated regression coefficients be different? Explain.
3. What is multicollinearity? Is that a problem? Why (not)?

Problem 4: Dummy vs. effect coding in MLR

Background material for this task: [Categorical covariates - dummy and effect coding)(#categorical)

We will study a dataset where we want to model `income` as response and two unordered categorical covariates `gender`and `place` (location).

```
income <- c(300, 350, 370, 360, 400, 370, 420, 390, 400, 43
    300, 320, 310, 305, 350, 370, 340, 355, 370, 380, 360,
gender <- c(rep("Male", 12), rep("Female", 12))
place <- rep(c(rep("A", 4), rep("B", 4), rep("C", 4)), 2)
data <- data.frame(income, gender = factor(gender, levels =
    "Male")), place = factor(place, levels = c("A", "B", "C
```

1. First, describe the data set.

```
library(GGally)
GGally::ggpairs(data)
```

Problem 5: Interactions

This part of the module was marked "self-study". Go through this together in the group, and make sure that you understand.

Problem 6: Simulations in R (optional)

(a version this problem was also given as recommended exercise in TMA4268 Statistical learning)

1. For simple linear regression, simulate at data set with homoscedastic errore and with heteroscedastic errors. Here is a suggestion of one solution. Why this? To see how things looks when the model is correct and wrong. Look at the code and discuss what is done, and relate this to the plots of errors (which are usually unobserved) and plots of residuals.

```
# Homoscedastic errors
n = 1000
x = seq(-3, 3, length = n)
beta0 = -1
```

SECOND WEEK

UNDER CONSTRUCTION :-) A first version is ready before Monday September 3.

R packages

```
install.packages(c("formatR", "gamlss.data", "tidyverse", "
    "nortest"))
```

References and further reading

- Slightly different presentation (more focus on multivariate normal theory): Slides and written material from TMA4267 Linear Statistical Models in 2017, Part 2: Regression (by Mette Langaas).
- And, same source, but now [Slides and written material from TMA4267 Linear Statistical Models in 2017, Part 3: Hypothesis testing and ANOVA] (http://www.math.ntnu.no/emner/TMA4267/2017v/TMA4267V2017Part3.pdf)