Module 2: MULTIPLE LINEAR REGRESSION
Week 1

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]

(Lastest changes: 26.08.2018)

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.)
- 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)
- big data implementation (if time)
- properties of parameter estimators
- assessing model fit (diagnostic), residuals, QQ-plots
- design matrix: how to code categorical covariates (dummy or effect coding), and how to handle interactions

Second week

- What did we do last week?
- Statistical inference for parameter estimates
  - confidence intervals,
  - prediction intervals,
  - hypothesis test,
  - linear hypotheses
- analysis of variance decompositions and R^2, sequential ANOVA table
- DEVIANCE???
- model selection with AIC and variants

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?]

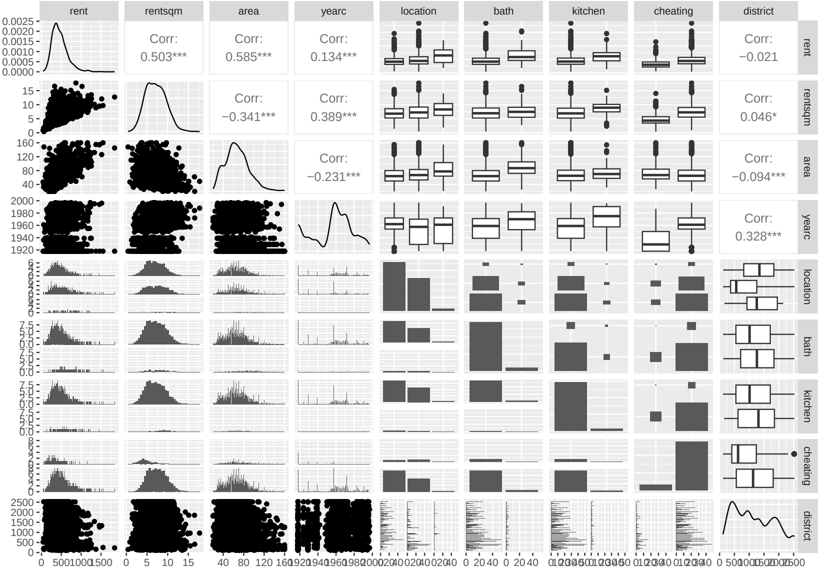
2. Construct a model to predict the response from a set of (one or several) explanatory variables. [More or less "black box"]

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.


rent
rentsqm
area
yearc
location
bath
kitchen
cheating
district
Corr:
0.503***
Corr:
0.585***
Corr:
0.134***
Corr:
−0.021
Corr:
−0.341***
Corr:
0.389***
Corr:
0.046*
Corr:
−0.231***
Corr:
−0.094***
Corr:
0.328***
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
15
10
5
0
160
120
80
40
2000
1980
1960
1940
1920
2500
2000
1500
1000
500
0
0 5001000 1500
0 5 10 15
40 80 120 160
1920 1940 1960 1980 2000
0 20 40
0 20 40
0 2500

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 apartment?
6. Is the effect of `area` the same on `rent` for apartments 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]

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

$\varepsilon : (n \times 1)$ 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

From the list of variable and the statement of the questions, answer these questions:

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

Model

The traditional way

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

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

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

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

4. $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$ 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 $\mathsf{E}(Y_i) = \mu_i$ and $\mathsf{Var}(Y_i) = \sigma^2$.
2. Systematic component: $\eta_i = \mathbf{x}_i^T \beta$.
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 $l(\beta)$

We work with the log-likelihood because this makes the mathematics simpler

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 individual 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 of vector

H¨{a}rdle 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: 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 providd by the likelihood about the unknown parameter. Since we look at minus the Hessian, we have a positive $H(\beta)$ near the maximum.

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.

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

```
##                Estimate  Std. Error    t value     Pr(>
## (Intercept) -45.47548356 3.603775035 -12.618846 1.251586
## area         -0.03233033 0.001647971 -19.618257 7.789203
## yearc         0.02695857 0.001845686  14.606265 9.119495
## location2     0.77713297 0.076870269  10.109669 1.168079
## location3     1.72506792 0.236062188   7.307684 3.447543
## bath1         0.76280784 0.157559037   4.841410 1.352865
## kitchen1      1.13690814 0.183087707   6.209637 6.024370
## cheating1     1.76526110 0.129067991  13.676986 2.212288
```

Restricted maximum likelihood estimator for σ^2

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

The regression parameters β are therefore our prime focus. We will look at the parameter σ^2 as a nuisance parameter $=$ parameter that is not of interest to us.

To perform inference we need an estimator for σ^2.

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

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

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

The score vector with respect to σ^2 is

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

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

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

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

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 $\tilde{\sigma}^2 = \frac{\mathsf{SSE}}{n}$ (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)

Can be proven using knowledge from TMA4267 on properties of the multivariate normal distribution.

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

We have:

- $\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2\mathbf{I})$
- $\mathbf{AY} = \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, 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: $\text{Cov}(\varepsilon) = \sigma^2\mathbf{I}$.
3. Uncorrelated errors: $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$.
4. Additivity: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$
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

Residuals

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

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

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

However, we have made "predictions" of the errors - our residuals.
And, we want to use our residuals to check the model assumptions.

We want to check that our errors are

- independent,
- homoscedastic (same variance for each observation),
- not dependent on our covariates

We want to use the residuals (observed) in place of the errors (unobserved).

It would have been great if the residuals have these properties when the underlying errors have.

Enter *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:

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

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

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.

Tests for normality can be used, but they can be useless: for small sample sizes the test is not 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

For simplicity we use the Munich rent index with `rent` as response and only `area` as the only covariate.

```r
fit <- lm(rent ~ area, data = rent99)  # Run a regression analysis
format(head(fortify(fit)), digits = 4L)
```

```
##    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
```

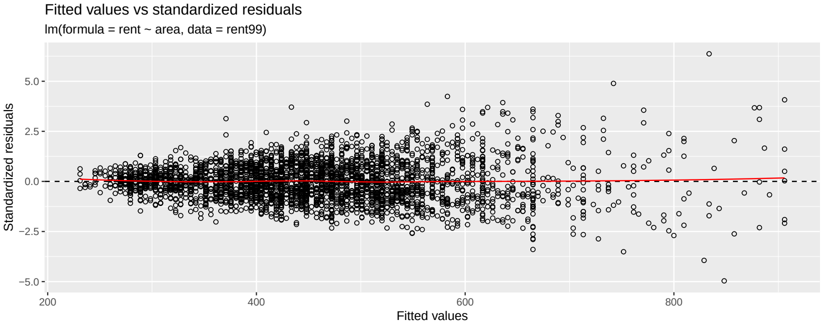
(`ggplot2::fortify.lm` creates a dataframe from an `lm`-object, which can be used to plot diagnostic plots. `ggplot` will do this automatically when asjked to plot)

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?

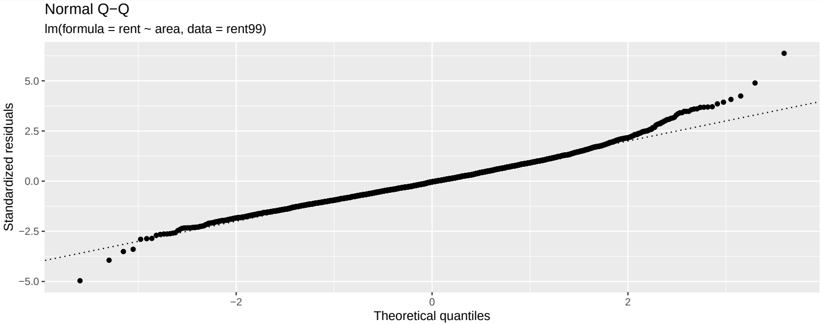
```r
ggplot(fit, aes(.fitted, .stdresid)) + geom_point(pch = 21)
    linetype = "dashed") + geom_smooth(se = FALSE, col = "r
    method = "loess") + labs(x = "Fitted values", y = "Star
    title = "Fitted values vs standardized residuals", subt
```



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.

```
ggplot(fit, aes(sample = .stdresid)) + stat_qq(pch = 19) +
    slope = 1, linetype = "dotted") + labs(x = "Theoretical
    y = "Standardized residuals", title = "Normal Q-Q", sub
```



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

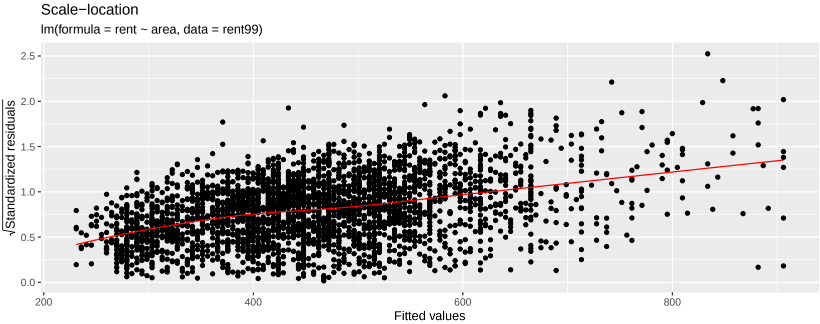
```
##
##  Anderson-Darling normality test
##
## data:  rstudent(fit)
## A = 6.4123, p-value = 9.809e-16
```

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?

```
ggplot(fit, aes(.fitted, sqrt(abs(.stdresid)))) + geom_poi
    col = "red", size = 0.5, method = "loess") + labs(x = "
    y = expression(sqrt("Standardized residuals")), title =
    subtitle = deparse(fit$call))
```



Residual vs Leverage

This plot can reveal influential outliers.
Not all outliers are influential in linear regression; the results might not change if they are removed from the data set
Influential outliers can be seen as observations that does not get along with the trend in the majority of the observations.

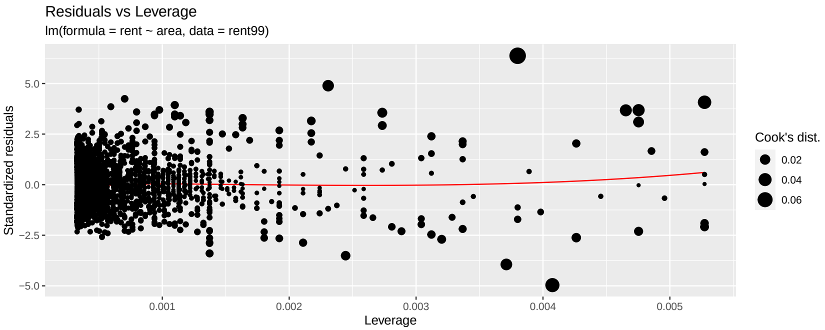
Cook's distance is the Euclidean distance between the $\hat{\mathbf{y}}$ (the fitted values) and $\hat{\mathbf{y}}_{(i)}$ (the fitted values calculated when the i-th observation is omitted from the regression).

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

```
ggplot(fit, aes(.hat, .stdresid)) + geom_smooth(se = FALSE,
    size = 0.5, method = "loess") + geom_point(aes(size =
    scale_size_continuous("Cook's dist.") + labs(x = "Lever
    title = "Residuals vs Leverage", subtitle = deparse(fit
```



(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)

Interactions

(if we have time)

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)

Ozone

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|)
## (Intercept) -67.2008    23.6083  -2.846  0.00529 **
## temperature   1.8265     0.2504   7.293 5.32e-11 ***
## wind         -3.2993     0.6706  -4.920 3.12e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 21.72 on 108 degrees of freedom
## Multiple R-squared:  0.5817, Adjusted R-squared:  0.574
## F-statistic:  75.1 on 2 and 108 DF,  p-value: < 2.2e-16
```

The model can be written as:

$$Y = \beta_0 + \beta_1 x_t + \beta_2 x_w + \varepsilon$$

In this model we have assumed that increasing the value of one covariate is independent of the other covariates. For example: by increasing the `temperature` by one-unit always increases the response value by $\beta_2 \approx 1.651$, regardless of the value of `wind`.

However, one might think that the covariate `wind` (wind speed) might act differently upon `ozone` for different values of `temperature` and vice verse.

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_t + \beta_2 x_w + \beta_3 \cdot (x_t \cdot x_w) + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 x_w) \cdot x_t + \beta_2 x_w + \varepsilon \qquad . \\ &= \beta_0 + \beta_1 x_t + (\beta_2 + \beta_3 x_t) \cdot x_w + \varepsilon \end{aligned}$$

We fit this model in `R`. An interaction term can be included in the model using the `*` symbol.

Q: Look at the `summary` below. Is this a better model than without the interaction term? It the term significant?

```
ozone.int = lm(ozone ~ temperature + wind + temperature * wind, data =
summary(ozone.int)
```

```
##
## Call:
## lm(formula = ozone ~ temperature + wind + temperature * wind,
##     data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.929 -11.190  -3.037   8.209  97.440
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -239.94146   48.59004  -4.938 2.92e-06 ***
## temperature        4.00151    0.59311   6.747 8.02e-10 ***
## wind              13.60882    4.28070   3.179  0.00193 **
## temperature:wind  -0.21747    0.05446  -3.993  0.00012 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.36 on 107 degrees of freedom
## Multiple R-squared:  0.636,  Adjusted R-squared:  0.6258
## F-statistic: 62.31 on 3 and 107 DF,  p-value: < 2.2e-16
```

Below we see that the interaction term is highly significant. The p-value is very small, so that there is strong evidence that $\beta_3 \neq 0$. Furthermore, R^2_{adj} has increased, indicating that more of the variability in the data has been explained by the model (than without the interaction).

Interpretation of the interaction term:

- If we now increase the `temperature` by 10° F, the increase in `wind` speed will be

  $$(\hat{\beta}_1 + \hat{\beta}_3 \cdot x_w) \cdot 10 = (4.0 - 0.22 \cdot x_w) \cdot 10 = 40 - 2.2x_w \text{ units.}$$

- If we increase the `wind` speed by 10 mph, the increase in `temperature` will be

  $$(\hat{\beta}_2 + \hat{\beta}_3 \cdot x_t) \cdot 10 = (14 - 0.22 \cdot x_t) \cdot 10 = 140 - 2.2x_t \text{ units.}$$

The hierarchical principle

It is possible that the interaction term is highly significant, but the main effects are not.

In our `ozone.int` model above: the main effects are `temperature` and `wind`. The hierarchical principle states that if we include an interaction term in our model, the main effects are also to be included, even if they are not significant. This means that if the coefficients $\hat{\beta}_1$ or $\hat{\beta}_2$ would be insignificant, while the coefficient $\hat{\beta}_3$ is significant, $\hat{\beta}_1$ and $\hat{\beta}_2$ should still be included in the model.

The hierarchical principle: why?

A model with interaction terms, but without the main effects is hard to interpret.

Removing a main effect is the same as setting $\beta_1 = 0$

$$Y = \beta_0 + \beta_1 x_t + \beta_2 x_w + \beta_3 \cdot (x_t \cdot x_w) + \varepsilon$$

i.e. saying the slope of the x_t effect is 0 when $x_w = 0$.

Interactions between qualitative (discrete) and quantitative (continuous) covariates

We create a new variable `temp.cat` which is a `temperature` as a qualitative covariate with two levels and fit the model:

$$y = \beta_0 + \beta_1 x_w + \begin{cases} \beta_2 + \beta_3 x_w & \text{if temperature="low"} \\ 0 & \text{if temperature = "high"} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot x_w & \text{if temperature="low"} \\ \beta_0 + \beta_1 x_w & \text{if temperature="high""} \end{cases}$$

Ozone: make temperature categorical

```
temp.cat = ifelse(ozone$temperature < mean(ozone$temperatur
    "high")
ozone2 = cbind(ozone, temp.cat)
kable(head(ozone2))
```

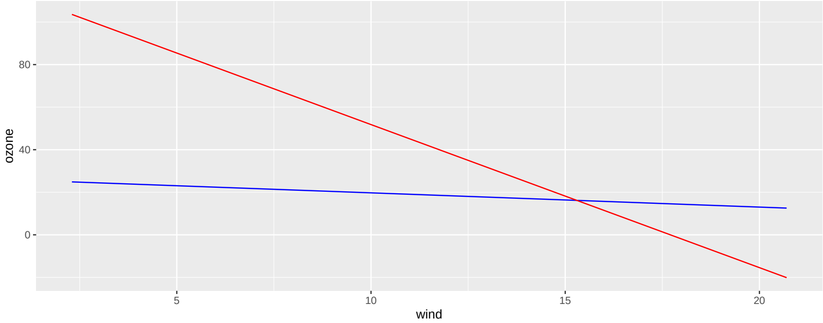
ozone	radiation	temperature	wind	temp.cat
41	190	67	7.4	low
36	118	72	8.0	low
12	149	74	12.6	low
18	313	62	11.5	low
23	299	65	8.6	low
19	99	59	13.8	low

Model with interaction

```r
ozone.int2 = lm(ozone ~ wind + temp.cat + temp.cat * wind,
summary(ozone.int2)$coefficients
```

```
##                     Estimate Std. Error   t value     Pr
## (Intercept)       119.045026  7.5004384 15.871742 6.9436
## wind               -6.723457  0.8195494 -8.203846 5.6099
## temp.catlow       -92.631612 12.9465805 -7.154910 1.0934
## wind:temp.catlow    6.054367  1.1999086  5.045690 1.8605
```

```r
interceptlow = coef(ozone.int2)[1] + coef(ozone.int2)[3]
slopelow = coef(ozone.int2)[2] + coef(ozone.int2)[4]
intercepthigh = coef(ozone.int2)[1]
slopehigh = coef(ozone.int2)[2]
ggplot(ozone) + geom_line(aes(y = interceptlow + slopelow
    col = "blue") + geom_line(aes(y = intercepthigh + slope
    x = wind), col = "red") + ylab("ozone")
```


ozone
wind