

Linear Regression - Part 2

Anders Stockmarr

Course developers: Anders Stockmarr, Elisabeth Wreford Andersen

DTU Department of Applied Mathematics and Computer Science
Section for Statistics and Data Analysis
Technical University of Denmark
`anst@dtu.dk`

January 7th, 2025

Plan for this week

Monday Statistical inference, and the t-test

Tuesday Simple and Multiple regression

Wednesday ANOVA, ANCOVA, and linear models

Thursday Categorical data, statistical report writing,
logistic Regression

Friday Introduction to repeated measures , Principal
Component Analysis

Learning objectives

After this session you should be able to:

- 1 Understand what a *multiple linear regression* (MLR) models is and be able to fit it to data

Learning objectives

After this session you should be able to:

- 1 Understand what a *multiple linear regression* (MLR) models is and be able to fit it to data
- 2 Interpret the result from a *multiple linear regression*

Learning objectives

After this session you should be able to:

- 1 Understand what a *multiple linear regression* (MLR) models is and be able to fit it to data
- 2 Interpret the result from a *multiple linear regression*
- 3 Understand and use interactions.

Learning objectives

After this session you should be able to:

- 1 Understand what a *multiple linear regression* (MLR) models is and be able to fit it to data
- 2 Interpret the result from a *multiple linear regression*
- 3 Understand and use interactions.
- 4 Do model reduction as backwards selection.

Overview

- 1 Multiple Linear Regression
- 2 Estimation
 - First MLR in Example
- 3 Building a MLR
 - GAM
 - Interaction
 - Trees
- 4 New model for the ozone data
- 5 Model Check
- 6 Testing
- 7 Exercises

Multiple Linear Regression

- The association between several continuous variables.
- Y response / outcome / dependent variable
- X_1, \dots, X_p explanatory / covariates / independent variables.

Data

Observations of sets $(x_{1i}, \dots, x_{pi}, y_i)$ for all $i = 1, \dots, n$ individuals or units.

Unit	x_1	x_2	\cdots	x_p	y
1	x_{11}	x_{12}	\cdots	x_{1p}	y_1
2	x_{21}	x_{22}	\cdots	x_{2p}	y_2
2	x_{31}	x_{32}	\cdots	x_{3p}	y_3
\vdots	\vdots	\vdots	\cdots	\vdots	\vdots
n	x_{n1}	x_{n2}	\cdots	x_{np}	y_n

The Multiple Linear Regression (MLR) model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Aim: Identify one (or several) reasonable model(s) that are:

- As simple as possible
- Captures the relevant structures in the data

General rule: **Keep variables that contribute — drop variables that don't**

Important Issues to Consider

- Which explanatory variables to include
- Curvature in the response to the explanatory variables
- Interactions between explanatory variables (will return to this)
- Correlation between explanatory variables

Use and Abuse of Multiple Linear Regression?

- Multiple Linear Regression may correspond to the scientific question of interest.
- With multiple explanatory variables, the predictions become more precise (more of the variability is explained).
- Assume that one of the explanatory variables is of greatest interest. There may be another variable connected to both the variable of interest and the outcome:

Use and Abuse of Multiple Linear Regression?

- Multiple Linear Regression may correspond to the scientific question of interest.
- With multiple explanatory variables, the predictions become more precise (more of the variability is explained).
- Assume that one of the explanatory variables is of greatest interest. There may be another variable connected to both the variable of interest and the outcome:
- **Example: Drowning and ice cream sales**
 - ❶ It seems that the higher the sales of ice cream the more drowning accidents.

Use and Abuse of Multiple Linear Regression?

- Multiple Linear Regression may correspond to the scientific question of interest.
- With multiple explanatory variables, the predictions become more precise (more of the variability is explained).
- Assume that one of the explanatory variables is of greatest interest. There may be another variable connected to both the variable of interest and the outcome:
- **Example: Drowning and ice cream sales**
 - ① It seems that the higher the sales of ice cream the more drowning accidents.
 - ② Is this because people eat ice cream at the beach, and then cannot swim?

Use and Abuse of Multiple Linear Regression?

- Multiple Linear Regression may correspond to the scientific question of interest.
- With multiple explanatory variables, the predictions become more precise (more of the variability is explained).
- Assume that one of the explanatory variables is of greatest interest. There may be another variable connected to both the variable of interest and the outcome:
- **Example: Drowning and ice cream sales**
 - ① It seems that the higher the sales of ice cream the more drowning accidents.
 - ② Is this because people eat ice cream at the beach, and then cannot swim?
 - ③ Or is there a 3rd variable (season) influencing both sale of ice cream and drowning accidents?

Example: Air pollution studies

How is ozone concentration related to wind speed, air temperature and solar radiation?

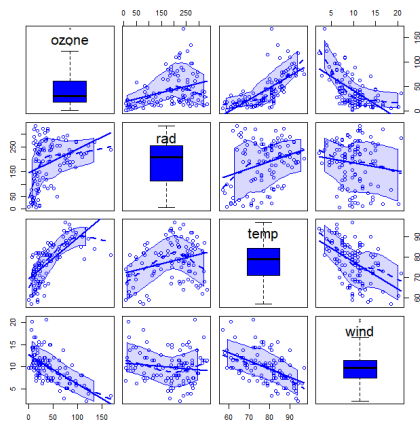
- We have 111 observations of ozone, wind speed, temperature and radiation.

The outcome (response) Y is the ozone concentration and the explanatory variables are X_1 radiation (rad), X_2 temperature (temp) and X_3 wind speed.

Always start by plotting the data!

Scatter plot

```
library(car)
scatterplotMatrix(~ ozone + rad + temp + wind,
                  diagonal = list(method="boxplot"), data = oz)
```



Overview

- 1 Multiple Linear Regression
- 2 Estimation**
 - First MLR in Example
- 3 Building a MLR
 - GAM
 - Interaction
 - Trees
- 4 New model for the ozone data
- 5 Model Check
- 6 Testing
- 7 Exercises

The Regression Model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$$

Traditional assumptions:

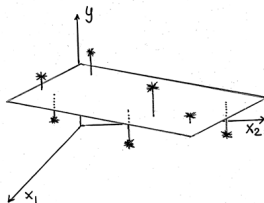
$\varepsilon_i \sim N(0, \sigma^2)$, independent.

Least squares:

Find the $\beta_0, \beta_1, \dots, \beta_p$ to minimize the sum of the squared distances:

$$SS(\beta_0, \beta_1, \dots, \beta_p) =$$

$$\sum (y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}))^2$$



Matrix Notation

If $n = 6$ and $p = 3$ then we can write the model using matrix notation:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ 1 & x_{31} & x_{32} & x_{33} \\ 1 & x_{41} & x_{42} & x_{43} \\ 1 & x_{51} & x_{52} & x_{53} \\ 1 & x_{61} & x_{62} & x_{63} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

Using compact notation we have:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Estimation

Using Least Squares method for estimation we get:

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

The estimated uncertainty on the estimate (variance):

$$\text{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

When we have estimates for $\beta_0, \beta_1 \dots, \beta_p$ then we can calculate the expected values for the outcome:

$$\hat{y} = X\hat{\beta}$$

The value \hat{y}_i is called the fitted value, or expected value. This corresponds to the value on the regression line.

Estimation continued

As for simple linear regression we also have the **residuals** (what is left):

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Using the matrix notation:

$$\hat{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}}$$

The model variance σ^2 is estimated:

$$\hat{\sigma}^2 = s^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n - (p + 1)} = MSE$$

Multiple Linear Regression

```
reg1 <- lm(ozone ~ rad + temp + wind, data = oz)
summary(reg1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	-64.2321	23.0420	-2.79	0.0063 **
## rad	0.0598	0.0232	2.58	0.0112 *
## temp	1.6512	0.2534	6.52	2.4e-09 ***
## wind	-3.3376	0.6538	-5.10	1.4e-06 ***

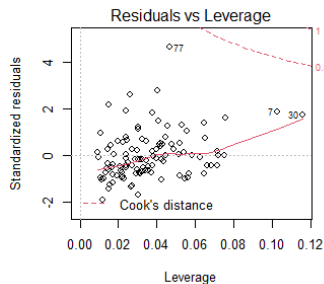
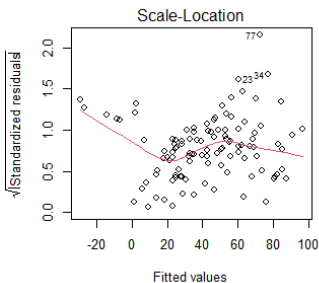
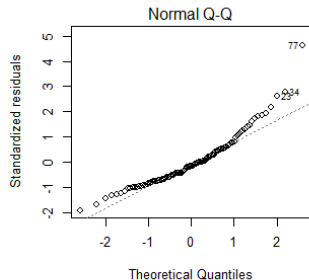
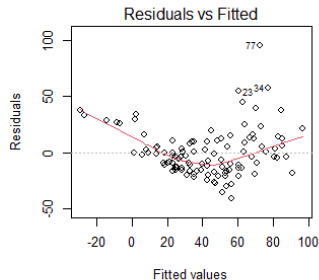
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

Residual standard error: 21.2 on 107 degrees of freedom

Multiple R-squared: 0.606, Adjusted R-squared: 0.595

F-statistic: 54.9 on 3 and 107 DF, p-value: <2e-16



Interpreting the Estimates

- It is **always important to interpret the model parameters**
 - give an explanation in words of the results.

Interpreting the Estimates

- It is always important to interpret the model parameters
 - give an explanation in words of the results.
- $\hat{\beta}_i$ is the effect of variable X_i on Y , when all other variables are fixed.

Interpreting the Estimates

- It is always important to interpret the model parameters
 - give an explanation in words of the results.
- $\hat{\beta}_i$ is the effect of variable X_i on Y , when all other variables are fixed.
- $\hat{\beta}_i$ is a slope. The expected change in Y when X_i changes one unit and the remaining variables are unchanged.

Interpreting the Estimates

- It is **always important to interpret the model parameters**
 - give an explanation in words of the results.
- $\hat{\beta}_i$ is the effect of variable X_i on Y , **when all other variables are fixed**.
- $\hat{\beta}_i$ is a slope. The expected change in Y when X_i changes one unit and the remaining variables are unchanged.
- The effect is corrected for the effect of the other explanatory variables.

The Estimates from the MLR

##	Estimate	Std..Error	Lower	Upper	p.value
## (Intercept)	-64.23	23.04	-109.91	-18.55	0.00628
## rad	0.06	0.02	0.01	0.11	0.01124
## temp	1.65	0.25	1.15	2.15	< 0.001
## wind	-3.34	0.65	-4.63	-2.04	< 0.001

And the variance is estimated by:

$$\hat{\sigma}^2 = 21.2^2 = 449.44$$

The **Intercept** = $\hat{\beta}_0$ is the expected ozone when wind=0, rad=0 and temp=0, not so interesting.

The Estimates from the MLR

##	Estimate	Std..Error	Lower	Upper	p.value
## (Intercept)	-64.23	23.04	-109.91	-18.55	0.00628
## rad	0.06	0.02	0.01	0.11	0.01124
## temp	1.65	0.25	1.15	2.15	< 0.001
## wind	-3.34	0.65	-4.63	-2.04	< 0.001

And the variance is estimated by:

$$\hat{\sigma}^2 = 21.2^2 = 449.44$$

The **Intercept** = $\hat{\beta}_0$ is the expected ozone when wind=0, rad=0 and temp=0, not so interesting.

temp = $\hat{\beta}_2$ is a slope. The ozone level increases by 1.65 when the temperature increases by 1 **for fixed wind and radiation**.

Overview

- 1 Multiple Linear Regression
- 2 Estimation
 - First MLR in Example
- 3 Building a MLR**
 - **GAM**
 - **Interaction**
 - **Trees**
- 4 New model for the ozone data
- 5 Model Check
- 6 Testing
- 7 Exercises

Building a MLR

- Perhaps this model was too simple.
- We want to include radiation, temperature and wind but we don't know whether it is reasonable with linear effects.
- Perhaps we need a curve.

GAM

- This morning we talked about adding squared terms of continuous explanatory variables. After looking at residual plots and seeing a non-random shape.

GAM

- This morning we talked about adding squared terms of continuous explanatory variables. After looking at residual plots and seeing a non-random shape.
- In many cases, however, we have one or more continuous explanatory variables, but no a priori reason to choose one particular parametric form over another for describing the shape of the relationship between the response variable and the explanatory variable(s).

GAM

- This morning we talked about adding squared terms of continuous explanatory variables. **After looking at residual plots and seeing a non-random shape.**
- In many cases, however, we have one or more continuous explanatory variables, but no a priori reason to choose one particular parametric form over another for describing the shape of the relationship between the response variable and the explanatory variable(s).
- Generalized additive models (GAMs) are useful in such cases because they allow us to capture the shape of a relationship between y and x **without having to choose a particular parametric form beforehand.**

GAM

- We are replacing the linear form in the regression model

$$\sum_j \beta_j X_j$$

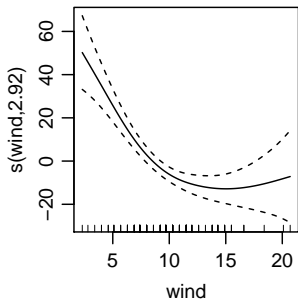
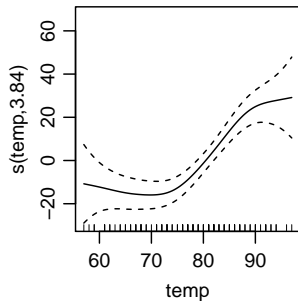
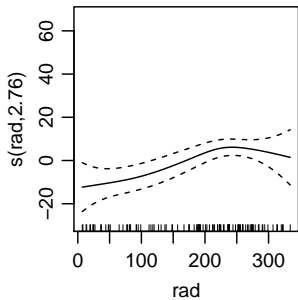
- By the sum of smooth functions

$$\sum_j s_j(X_j)$$

- The functions s_j are unspecified smooth functions estimated using a non-parametric smoother

GAM for the ozone example

```
library(mgcv)
par(mfrow = c(2,2), mgp = c(2,0.7,0), mar =
c(3,3,1,1))
model <- gam(ozone ~ s(rad) + s(temp) + s(wind), data = oz)
plot(model)
par(mfrow = c(1,1))
```



Ideas from GAM

- The confidence intervals are sufficiently narrow to suggest that the curvature in the relationship between ozone and temperature is real
- The curvature of the relationship with wind is questionable
- A linear model may well be all that is required for solar radiation

What if the effect of temperature depends on wind speed?

We had the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

Where y_i is the observed ozone concentration i , x_{1i} the radiation, x_{2i} the temperature and x_{3i} the wind speed.

We are assuming that temperature and wind have an **additive effect** on the ozone concentration.

The effect of temperature is assumed the same for all wind speeds.

What if the effect of temperature depends on wind speed?

We had the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

Where y_i is the observed ozone concentration i , x_{1i} the radiation, x_{2i} the temperature and x_{3i} the wind speed.

We are assuming that temperature and wind have an **additive effect** on the ozone concentration.

The effect of temperature is assumed the same for all wind speeds.

Perhaps the additive model is too simple, we can include a **multiplicative term** (also called an interaction).

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 (x_{2i} \cdot x_{3i}) + \varepsilon_i$$

where β_4 accounts for the **interaction** between temperature and wind.

Taylor expansion of real functions

Why is a multiplicative term a relevant idea?

- Power series expansion of a smooth one-dimensional function:

$$f(x) = f(x_0) + \sum_{n=1}^{\infty} a_n (x - x_0)^n, \quad a_n = \frac{f^{(n)}(x_0)}{n!}$$

- 1st order Taylor expansion:

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0) \cdot (x - x_0) + r_1(x - x_0) \\ &= \underbrace{f(x_0) - f'(x_0) \cdot x_0}_{\alpha} + \underbrace{f'(x_0)}_{\beta} \cdot x + r_1(x - x_0) \\ &= \alpha + \beta x + r_1(x - x_0) \end{aligned}$$

Taylor expansion of real functions

- If the relation between Y and x is really f :

$$Y = \alpha + \beta x + \varepsilon$$

$\varepsilon = r_1(x - x_0) + \epsilon$ covers both model aberrations and stochasticity.

- If the model aberration is too big to be handled by the general uncertainty ε , we may resort to a finer model description, Taylor expansion to the 2^{nd} order (here in arbitrary dimensions):

$$\begin{aligned} f(x) = & f(x_0) + \langle f'(x_0), x - x_0 \rangle \\ & + \frac{1}{2}(x - x_0)^T \cdot f''(x_0) \cdot (x - x_0) + r_2(x - x_0) \end{aligned}$$

- The matrix in the second term contains the coefficients to the 2^{nd} order multiplicative model terms.

Trees

- We need to get some ideas about which interactions to include.

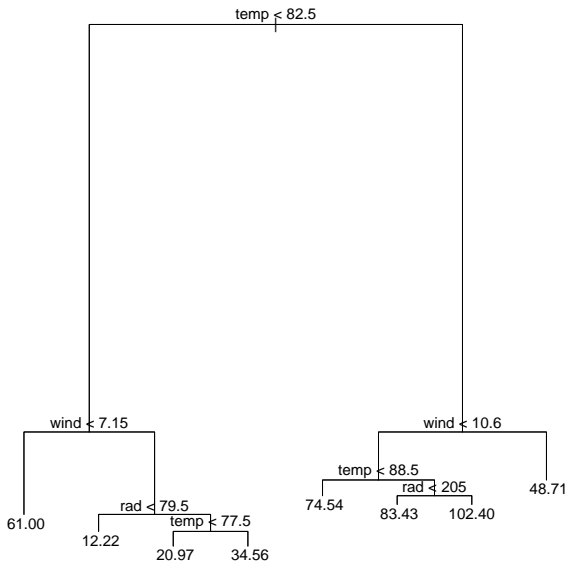
Trees

- We need to get some ideas about which interactions to include.
- Trees can help identify interactions
- Good for initial data inspection
- The splits that gives the largest reduction in variance for each part.
- At the leaves we have the mean value in that subset of the data

Tree in example

```
library(tree)
model<-tree(ozone~., data = oz)
plot(model)
text(model)
```

Tree in example



Trees

- If the number of covariates is large, the single tree method will no longer work. One will then need to resort to sparse methodology.
- One such bases itself on the Random Forest methodology, selecting subsets of covariates at random and construct corresponding trees.
- Selection of interacting terms, after correcting for correlation/dependence, can be made following *Behr et al 2022*, <https://doi.org/10.1073/pnas.2118636119>.
- The procedure by *Behr et al 2022* is, however, outside the scope of this course.

Ideas from the tree

- Temperature is by far the most important
- Wind speed important at both high and low values. Low wind is associated to higher mean ozone levels.
- Possible interaction between wind and temperature and wind and radiation.

Overview

- 1 Multiple Linear Regression
- 2 Estimation
 - First MLR in Example
- 3 Building a MLR
 - GAM
 - Interaction
 - Trees
- 4 New model for the ozone data**
- 5 Model Check
- 6 Testing
- 7 Exercises

New model for the ozone data

- We are now ready with a more complex model for the ozone data.
- We will include curvature for wind and temperature and interactions between wind and temperature and wind and radiation.

$$\begin{aligned}y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{2i}^2 + \beta_5 x_{3i}^2 \\ &+ \beta_6 (x_{2i} \cdot x_{3i}) + \beta_7 (x_{1i} \cdot x_{3i}) + \varepsilon_i\end{aligned}$$

- y_i =ozone, x_{1i} =radiation, x_{2i} =temp, x_{3i} =wind and $\varepsilon_i \sim N(0, \sigma^2)$.

New model for the ozone data

- We are now ready with a more complex model for the ozone data.
- We will include curvature for wind and temperature and interactions between wind and temperature and wind and radiation.

$$\begin{aligned}y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{2i}^2 + \beta_5 x_{3i}^2 \\ &+ \beta_6 (x_{2i} \cdot x_{3i}) + \beta_7 (x_{1i} \cdot x_{3i}) + \varepsilon_i\end{aligned}$$

- y_i =ozone, x_{1i} =radiation, x_{2i} =temp, x_{3i} =wind and $\varepsilon_i \sim N(0, \sigma^2)$.
- The next step is to simplify the model

New model for the ozone data

- We are now ready with a more complex model for the ozone data.
- We will include curvature for wind and temperature and interactions between wind and temperature and wind and radiation.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{2i}^2 + \beta_5 x_{3i}^2 \\ + \beta_6 (x_{2i} \cdot x_{3i}) + \beta_7 (x_{1i} \cdot x_{3i}) + \varepsilon_i$$

- y_i =ozone, x_{1i} =radiation, x_{2i} =temp, x_{3i} =wind and $\varepsilon_i \sim N(0, \sigma^2)$.
- The next step is to simplify the model
- We must not forget to check the underlying assumptions!

The new model

```
reg2 <- lm(ozone ~ rad + temp + wind + I(temp^2) +
           I(wind^2) + temp:wind + rad:wind, data = oz)
summary(reg2)
```

Coefficients:

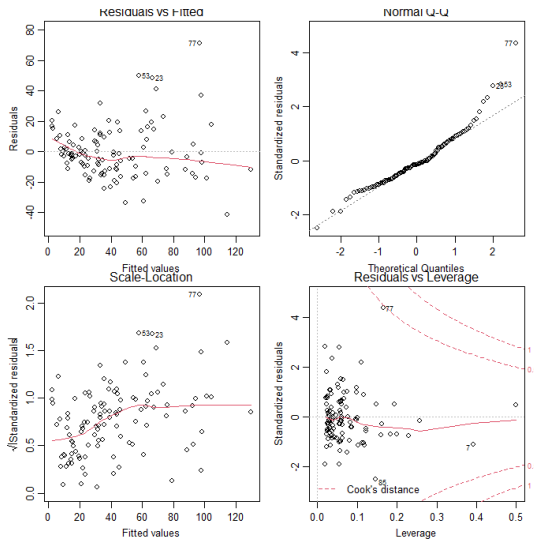
	Estimate	Std. Error	t value	Pr(> t)	
## (Intercept)	514.40147	193.78358	2.65	0.0092	**
## rad	0.21295	0.06928	3.07	0.0027	**
## temp	-10.65404	4.09489	-2.60	0.0106	*
## wind	-27.39197	9.61700	-2.85	0.0053	**
## I(temp^2)	0.06780	0.02241	3.03	0.0031	**
## I(wind^2)	0.61940	0.14577	4.25	4.7e-05	***
## temp:wind	0.16967	0.09446	1.80	0.0754	.
## rad:wind	-0.01356	0.00609	-2.23	0.0281	*

##

Residual standard error: 17.9 on 103 degrees of freedom

Multiple R-squared: 0.729, Adjusted R-squared: 0.711

F-statistic: 39.6 on 7 and 103 DF, p-value: <2e-16



Work with the person next to you

- Use the model with interaction.
- What is the expected ozone concentration:
 - If the level of radiation is 100, temperature is 60 and wind speed is 20.
 - If the level of radiation is 185, temperature is 80 and wind speed is 10.
 - If the level of radiation is 185, temperature is 80 and wind speed is 5.

Overview

- 1 Multiple Linear Regression
- 2 Estimation
 - First MLR in Example
- 3 Building a MLR
 - GAM
 - Interaction
 - Trees
- 4 New model for the ozone data
- 5 Model Check**
- 6 Testing
- 7 Exercises

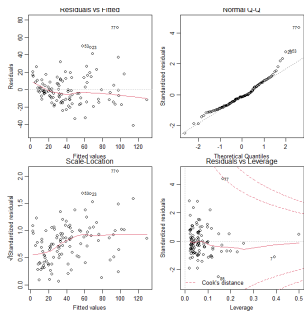
Model Check in detail

As in the simple linear regression we must check our model assumptions before we interpret our model too much. We have to check:

- Normal residuals (observed - fitted), using qq-plots.
- Variance homogeneity (one σ^2), residual plots against fitted values.
- Linear effect of X_1, \dots, X_p , residual plots against each covariate.

Model Check in example

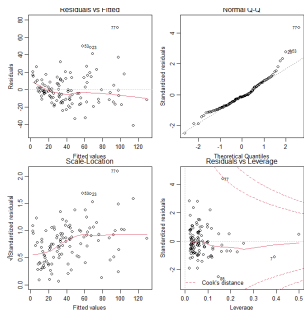
```
par(mfrow=c(2, 2))
plot(reg2, which=1:4)
```



- The model check still did not look good!
- Problem with variance homogeneity.

Model Check in example

```
par(mfrow=c(2, 2))
plot(reg2, which=1:4)
```

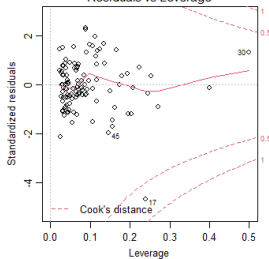
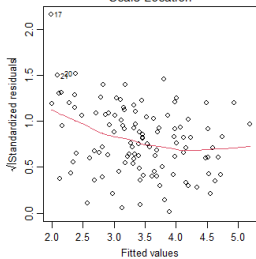
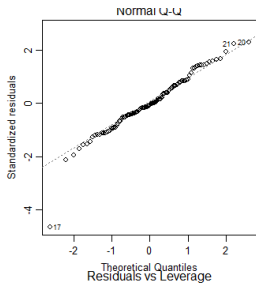
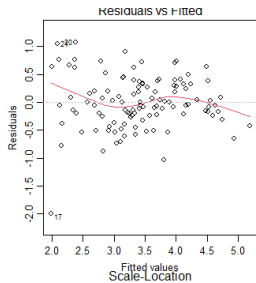


- The model check still did not look good!
- Problem with variance homogeneity.
- What should we do?

New model for $\log(\text{ozone})$

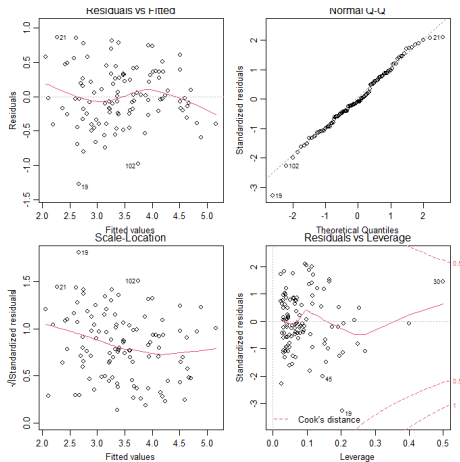
- We need to start from scratch with all the original explanatory variables included.
- We would expect the curvature to have changed.
- We can run a new GAM and do a new Tree;
- With just a few explanatory variables, we can also choose brute force and include all 2nd order effects.
- The new starting model should be:

```
reg3 <- lm(log(ozone) ~ rad + temp + wind + I(temp^2) +  
            I(wind^2) + I(rad^2) +  
            rad:temp + rad:wind +  
            temp:wind, data = oz)
```



Removing Outlier

```
reg3 <- update(reg3, data=oz[-17,])
```



Model Estimates

```
summary(reg3)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.378e+01	5.064e+00	2.721	0.00768	**
rad	-1.670e-05	5.470e-03	-0.003	0.99757	
temp	-2.208e-01	1.076e-01	-2.051	0.04287	*
wind	-6.291e-01	2.425e-01	-2.594	0.01092	*
I(temp^2)	1.270e-03	6.014e-04	2.111	0.03725	*
I(wind^2)	1.038e-02	3.608e-03	2.876	0.00493	**
I(rad^2)	-1.353e-05	6.260e-06	-2.161	0.03308	*
rad:temp	9.794e-05	6.208e-05	1.578	0.11780	
rad:wind	-3.984e-05	1.552e-04	-0.257	0.79794	
temp:wind	4.429e-03	2.385e-03	1.857	0.06628	.

Overview

- 1 Multiple Linear Regression
- 2 Estimation
 - First MLR in Example
- 3 Building a MLR
 - GAM
 - Interaction
 - Trees
- 4 New model for the ozone data
- 5 Model Check
- 6 Testing**
- 7 Exercises

Testing

- When we finally have a satisfactory starting model then we often want to simplify.
- Sometimes we have **specific questions**, i.e.
 - Does the crime rate depend on the level of education?
 - Did the intervention make the children eat more healthily?
- Other times we have a lot of variables and are mainly looking for **structures** in the data.

Model selection

- It is **not trivial** to choose a good starting model. Often it is a combination of theoretical knowledge and/or practical experience, and then the hypotheses we want to test.

Model selection

- It is **not trivial** to choose a good starting model. Often it is a combination of theoretical knowledge and/or practical experience, and then the hypotheses we want to test.
- Sometimes we have a large amount of data with many variables and not much knowledge. Here we will often take a more exploratory approach with some sensible strategies, and an automated approach when looking for a model.

Model selection

- It is **not trivial** to choose a good starting model. Often it is a combination of theoretical knowledge and/or practical experience, and then the hypotheses we want to test.
- Sometimes we have a large amount of data with many variables and not much knowledge. Here we will often take a more exploratory approach with some sensible strategies, and an automated approach when looking for a model.
- **Be cautious when using automated approaches.** If the number of variables is not too large, then it is better to think it through and keep an eye on what is happening.

Model selection

- It is **not trivial** to choose a good starting model. Often it is a combination of theoretical knowledge and/or practical experience, and then the hypotheses we want to test.
- Sometimes we have a large amount of data with many variables and not much knowledge. Here we will often take a more exploratory approach with some sensible strategies, and an automated approach when looking for a model.
- **Be cautious when using automated approaches.** If the number of variables is not too large, then it is better to think it through and keep an eye on what is happening.
- A rule of thumb: The number of parameters in the model should be less than $(\text{number of observations})/5$.

Backwards and Forwards selection

Backwards

- Start with the largest model, the most complex, and remove variables which are not significant, **one at a time**. Continue until all variables are significant.

Backwards and Forwards selection

Backwards

- Start with the largest model, the most complex, and remove variables which are not significant, **one at a time**. Continue until all variables are significant.

Forwards

- Start with the model that only includes an intercept. Add a variable one at a time starting with the most significant. Continue until none of the remaining variables are significant.

Tests of main effects and interactions

Never remove a main effect if it is part of an interaction

Tests of main effects and interactions

Never remove a main effect if it is part of an interaction

In R the function `drop1()` will help you not to make mistakes.

Tests of main effects and interactions

Never remove a main effect if it is part of an interaction

In R the function `drop1()` will help you not to make mistakes.

Never remove a lower order term if it is part of a higher order term

Tests of main effects and interactions

Never remove a main effect if it is part of an interaction

In R the function `drop1()` will help you not to make mistakes.

Never remove a lower order term if it is part of a higher order term

You don't know if there is (evidence of) an interaction unless you look for it

Model reduction in example

```
drop1(reg3,test="F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(temp^2) + I(wind^2) + I(rad^2) +
            rad:temp + rad:wind + temp:wind
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			19.020	-173.05		
I(temp^2)	1	0.84769	19.867	-170.25	4.4570	0.037253 *
I(wind^2)	1	1.57299	20.593	-166.31	8.2704	0.004925 **
I(rad^2)	1	0.88826	19.908	-170.03	4.6703	0.033077 *
rad:temp	1	0.47342	19.493	-172.35	2.4891	0.117796
rad:wind	1	0.01253	19.032	-174.98	0.0659	0.797937
temp:wind	1	0.65574	19.675	-171.32	3.4477	0.066284 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model reduction in example

```
drop1(reg3, test="F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(temp^2) + I(wind^2) + I(rad^2) +
            rad:temp + rad:wind + temp:wind
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			19.020	-173.05		
I(temp^2)	1	0.84769	19.867	-170.25	4.4570	0.037253 *
I(wind^2)	1	1.57299	20.593	-166.31	8.2704	0.004925 **
I(rad^2)	1	0.88826	19.908	-170.03	4.6703	0.033077 *
rad:temp	1	0.47342	19.493	-172.35	2.4891	0.117796
rad:wind	1	0.01253	19.032	-174.98	0.0659	0.797937
temp:wind	1	0.65574	19.675	-171.32	3.4477	0.066284 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Notice the use of `test="F"`. We will remove `rad:wind`.

Model reduction in example

```
reg4 <- update(reg3, ~. -rad:wind)
drop1(reg4, test = "F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(temp^2) + I(wind^2) + I(rad^2) +
  rad:temp + temp:wind
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)		
<none>			19.032	-174.98				
I(temp^2)	1	0.83641	19.869	-172.25	4.4387	0.037611	*	
I(wind^2)	1	1.56240	20.595	-168.30	8.2914	0.004864	**	
I(rad^2)	1	0.88847	19.921	-171.96	4.7150	0.032243	*	
rad:temp	1	0.56397	19.596	-173.77	2.9929	0.086687	.	
temp:wind	1	0.66581	19.698	-173.20	3.5333	0.063029	.	

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'
	0.1	' '		' '			1	

Model reduction in example

```
reg4 <- update(reg3, ~. -rad:wind)
drop1(reg4, test = "F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(temp^2) + I(wind^2) + I(rad^2) +
  rad:temp + temp:wind
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)		
<none>			19.032	-174.98				
I(temp^2)	1	0.83641	19.869	-172.25	4.4387	0.037611	*	
I(wind^2)	1	1.56240	20.595	-168.30	8.2914	0.004864	**	
I(rad^2)	1	0.88847	19.921	-171.96	4.7150	0.032243	*	
rad:temp	1	0.56397	19.596	-173.77	2.9929	0.086687	.	
temp:wind	1	0.66581	19.698	-173.20	3.5333	0.063029	.	

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'
	0.1	' '						1

We will remove rad:temp. **We will continue like this until all variables**

Model reduction in example

```
reg4 <- update(reg4, ~. -rad:temp)
drop1(reg4, test = "F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(temp^2) + I(wind^2) + I(rad^2) +
temp:wind
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)	
<none>			19.596	-173.77			
rad	1	1.58795	21.184	-167.20	8.2655	0.004919	**
I(temp^2)	1	1.27863	20.875	-168.81	6.6554	0.011310	*
I(wind^2)	1	1.60681	21.203	-167.10	8.3636	0.004679	**
I(rad^2)	1	0.69249	20.289	-171.95	3.6045	0.060450	.
temp:wind	1	0.57900	20.175	-172.56	3.0138	0.085580	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model reduction in example

```
reg4 <- update(reg4, ~. -rad:temp)
drop1(reg4, test = "F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(temp^2) + I(wind^2) + I(rad^2) +
temp:wind
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)					
<none>			19.596	-173.77							
rad	1	1.58795	21.184	-167.20	8.2655	0.004919	**				
I(temp^2)	1	1.27863	20.875	-168.81	6.6554	0.011310	*				
I(wind^2)	1	1.60681	21.203	-167.10	8.3636	0.004679	**				
I(rad^2)	1	0.69249	20.289	-171.95	3.6045	0.060450	.				
temp:wind	1	0.57900	20.175	-172.56	3.0138	0.085580	.				

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

We will remove temp:wind.

Model reduction in example

```
reg4 <- update(reg4, ~. -temp:wind)
drop1(reg4, test = "F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(temp^2) + I(wind^2) + I(rad^2)
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)	
<none>			20.175	-172.56			
rad	1	1.44844	21.623	-166.94	7.3948	0.0076784	**
temp	1	0.33447	20.509	-172.75	1.7076	0.1942126	
wind	1	2.32037	22.495	-162.59	11.8462	0.0008361	***
I(temp^2)	1	0.69977	20.875	-170.81	3.5725	0.0615534	.
I(wind^2)	1	1.09518	21.270	-168.75	5.5913	0.0199242	*
I(rad^2)	1	0.57877	20.754	-171.45	2.9548	0.0886278	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model reduction in example

```
reg4 <- update(reg4, ~. -temp:wind)
drop1(reg4, test = "F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(temp^2) + I(wind^2) + I(rad^2)
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)	
<none>			20.175	-172.56			
rad	1	1.44844	21.623	-166.94	7.3948	0.0076784	**
temp	1	0.33447	20.509	-172.75	1.7076	0.1942126	
wind	1	2.32037	22.495	-162.59	11.8462	0.0008361	***
I(temp^2)	1	0.69977	20.875	-170.81	3.5725	0.0615534	.
I(wind^2)	1	1.09518	21.270	-168.75	5.5913	0.0199242	*
I(rad^2)	1	0.57877	20.754	-171.45	2.9548	0.0886278	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

We will remove the squared effect of rad.

Model reduction in example

```
reg4 <- update(reg4, ~. -I(rad^2))
drop1(reg4, test = "F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(temp^2) + I(wind^2)
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>		20.754	-171.45			
rad	1	4.1707	24.924	-153.31	20.8996	1.339e-05 ***
temp	1	0.2728	21.027	-172.02	1.3669	0.2450244
wind	1	2.3402	23.094	-161.70	11.7269	0.0008827 ***
I(temp^2)	1	0.6390	21.393	-170.12	3.2022	0.0764518 .
I(wind^2)	1	1.0639	21.818	-167.95	5.3312	0.0229249 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model reduction in example

```
reg4 <- update(reg4, ~. -I(rad^2))
drop1(reg4, test = "F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(temp^2) + I(wind^2)
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)	
<none>		20.754	-171.45				
rad	1	4.1707	24.924	-153.31	20.8996	1.339e-05	***
temp	1	0.2728	21.027	-172.02	1.3669	0.2450244	
wind	1	2.3402	23.094	-161.70	11.7269	0.0008827	***
I(temp^2)	1	0.6390	21.393	-170.12	3.2022	0.0764518	.
I(wind^2)	1	1.0639	21.818	-167.95	5.3312	0.0229249	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

We will remove the squared effect of temp.

Model reduction in example

```
reg4 <- update(reg4, ~. -I(temp^2))
drop1(reg4, test = "F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(wind^2)
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)	
<none>			21.393	-170.12			
rad	1	3.9973	25.390	-153.27	19.6192	2.325e-05	***
temp	1	11.5647	32.958	-124.58	56.7617	1.807e-11	***
wind	1	3.3253	24.718	-156.22	16.3212	0.000102	***
I(wind^2)	1	1.6759	23.069	-163.82	8.2258	0.004993	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model reduction in example

```
reg4 <- update(reg4, ~. -I(temp^2))
drop1(reg4, test = "F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(wind^2)
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			21.393	-170.12		
rad	1	3.9973	25.390	-153.27	19.6192	2.325e-05 ***
temp	1	11.5647	32.958	-124.58	56.7617	1.807e-11 ***
wind	1	3.3253	24.718	-156.22	16.3212	0.000102 ***
I(wind^2)	1	1.6759	23.069	-163.82	8.2258	0.004993 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

We will stop the model reduction as all variables are statistically significant.

Words of caution

- Do not blindly use automatic stepwise variable selection procedures
- Don't confuse and combine model search and selection with *confirmatory hypothesis testing*; we need a fitting and sensible model for the latter
- The model space is often very large — there may be more than one model that may explain the data equally well
- Always consider the use of interactions, polynomials and transformations; even though you may decide against them in the end.

Words of caution

- Do not blindly use automatic stepwise variable selection procedures
- Don't confuse and combine model search and selection with *confirmatory hypothesis testing*; we need a fitting and sensible model for the latter
- The model space is often very large — there may be more than one model that may explain the data equally well
- Always consider the use of interactions, polynomials and transformations; even though you may decide against them in the end.

Model selection is an art — it takes practice to master

Final Model

We have subsequently removed $rad : wind$, $rad : temp$, $temp : wind$, rad^2 , and $temp^2$ through backwards selection. Coefficients in final model:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.1932358	0.5990022	1.992	0.048963	*
rad	0.0022097	0.0004989	4.429	2.33e-05	***
temp	0.0419157	0.0055635	7.534	1.81e-11	***
wind	-0.2208189	0.0546589	-4.040	0.000102	***
I(wind^2)	0.0068982	0.0024052	2.868	0.004993	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Common issues arising in MLR

- **Differences in the measurement scales** of the explanatory variables, leading to large variation in the sums of squares and hence to an ill-conditioned matrix.
 - Can consider **standardizing**, i.e. subtracting the mean and dividing by the standard deviation.
- **Multicollinearity**, in which there is a near-linear relation between two of the explanatory variables (nearly the same information), leading to unstable parameter estimates.
 - Perhaps choose only one of several colinear variables, or use PCA (Lecture 10)
- **Parameter proliferation** where quadratic and interaction terms soak up more degrees of freedom than our data can afford.
 - Careful selection of interaction and quadratic terms, for example through the methods discussed today, trees and GAM.

Learning objectives

After this session you should be able to:

- 1 Understand what a *multiple linear regression* (MLR) models is and be able to fit it to data
- 2 Interpret the result from a *multiple linear regression*
- 3 Understand and use interactions.
- 4 Do backwards selection.

Overview

- 1 Multiple Linear Regression
- 2 Estimation
 - First MLR in Example
- 3 Building a MLR
 - GAM
 - Interaction
 - Trees
- 4 New model for the ozone data
- 5 Model Check
- 6 Testing
- 7 Exercises**

2 Exercises for Multiple linear regression

- 1 Process: Understand process loss as a function of other continuous variables
- 2 Cheese: Describe the taste of matured cheese as a function of chemical descriptors.