Applied Statistical Regression
AS 2012 – Week 01

Marcel Dettling

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

ETH Zürich, September 24, 2012

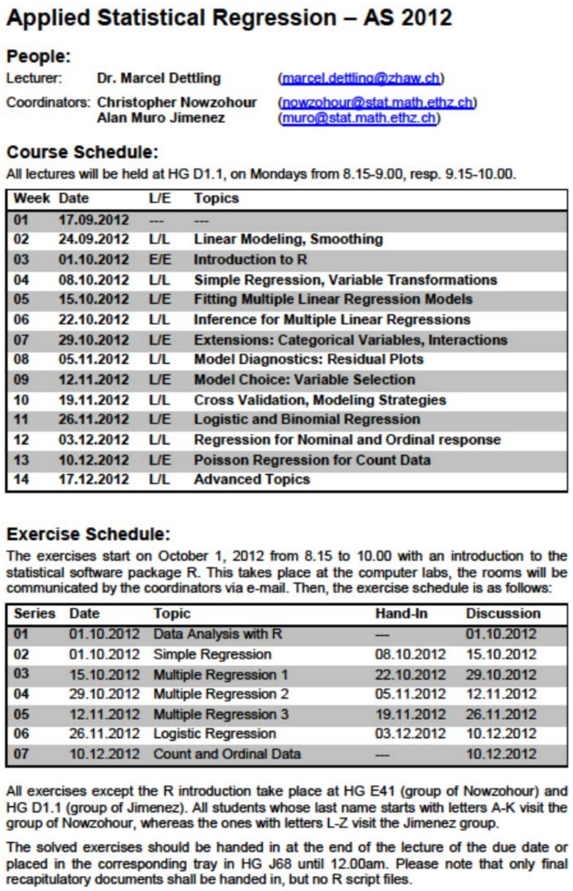
Applied Statistical Regression
AS 2012 – Week 01

Your Lecturer

Name:	Marcel Dettling
Age:	38 years
Civil Status:	Married, 2 children
Education:	Dr. Math. ETH
Position:	Lecturer at ETH Zürich and ZHAW Senior Researcher at IDP, a ZHAW institute
Hobbies:	Rock climbing, Skitouring, Paragliding, …

Applied Statistical Regression
AS 2012 – Week 01

Course Organization



Applied Statistical Regression – AS 2012

People:

Lecturer: **Dr. Marcel Dettling** (marcel.dettling@zhaw.ch)

Coordinators: **Christopher Nowzohour** (nowzohour@stat.math.ethz.ch)
Alan Muro Jimenez (muro@stat.math.ethz.ch)

Course Schedule:

All lectures will be held at HG D1.1, on Mondays from 8.15-9.00, resp. 9.15-10.00.

Week	Date	L/E	Topics
01	17.09.2012	---	---
02	24.09.2012	L/L	Linear Modeling, Smoothing
03	01.10.2012	E/E	Introduction to R
04	08.10.2012	L/L	Simple Regression, Variable Transformations
05	15.10.2012	L/E	Fitting Multiple Linear Regression Models
06	22.10.2012	L/L	Inference for Multiple Linear Regressions
07	29.10.2012	L/E	Extensions: Categorical Variables, Interactions
08	05.11.2012	L/L	Model Diagnostics: Residual Plots
09	12.11.2012	L/E	Model Choice: Variable Selection
10	19.11.2012	L/L	Cross Validation, Modeling Strategies
11	26.11.2012	L/E	Logistic and Binomial Regression
12	03.12.2012	L/L	Regression for Nominal and Ordinal response
13	10.12.2012	L/E	Poisson Regression for Count Data
14	17.12.2012	L/L	Advanced Topics

Exercise Schedule:

The exercises start on October 1, 2012 from 8.15 to 10.00 with an introduction to the statistical software package R. This takes place at the computer labs, the rooms will be communicated by the coordinators via e-mail. Then, the exercise schedule is as follows:

Series	Date	Topic	Hand-In	Discussion
01	01.10.2012	Data Analysis with R	—	01.10.2012
02	01.10.2012	Simple Regression	08.10.2012	15.10.2012
03	15.10.2012	Multiple Regression 1	22.10.2012	29.10.2012
04	29.10.2012	Multiple Regression 2	05.11.2012	12.11.2012
05	12.11.2012	Multiple Regression 3	19.11.2012	26.11.2012
06	26.11.2012	Logistic Regression	03.12.2012	10.12.2012
07	10.12.2012	Count and Ordinal Data	—	10.12.2012

All exercises except the R introduction take place at HG E41 (group of Nowzohour) and HG D1.1 (group of Jimenez). All students whose last name starts with letters A-K visit the group of Nowzohour, whereas the ones with letters L-Z visit the Jimenez group.

The solved exercises should be handed in at the end of the lecture of the due date or placed in the corresponding tray in HG J68 until 12.00am. Please note that only final recapitulatory documents shall be handed in, but no R script files.

What is Regression?

The answer to an everyday question:

How does a target variable of special interest depend on several other (explanatory) factors or causes.

Examples:

- growth of plants, depends on fertilizer, soil quality, …
- apartment rents, depends on size, location, furnishment, …
- car insurance premium, depends on age, sex, nationality, …

Regression:

- quantitatively describes relation between predictors and target
- high importance, most widely used statistical methodology

What is Regression?

Example: *Fresh Water Tank on* edelweiss air *Planes*

- **Earlier**: it was impossible to predict the amount of fresh water needed, the tank was always filled to 100% at Zurich airport.
- **Goal**: Minimizing the amount of fresh water that is carried. This lowers the weight, and thus fuel consumption and cost.
- **Task***:* Modelling the relation between fresh water consumption and # *of passengers*, *flight duration*, *daytime*, *destination*, … Furthermore, quantifying what is needed as a reserve.
- **Method:** *Multiple linear regression model*

Applied Statistical Regression
AS 2012 – Week 01

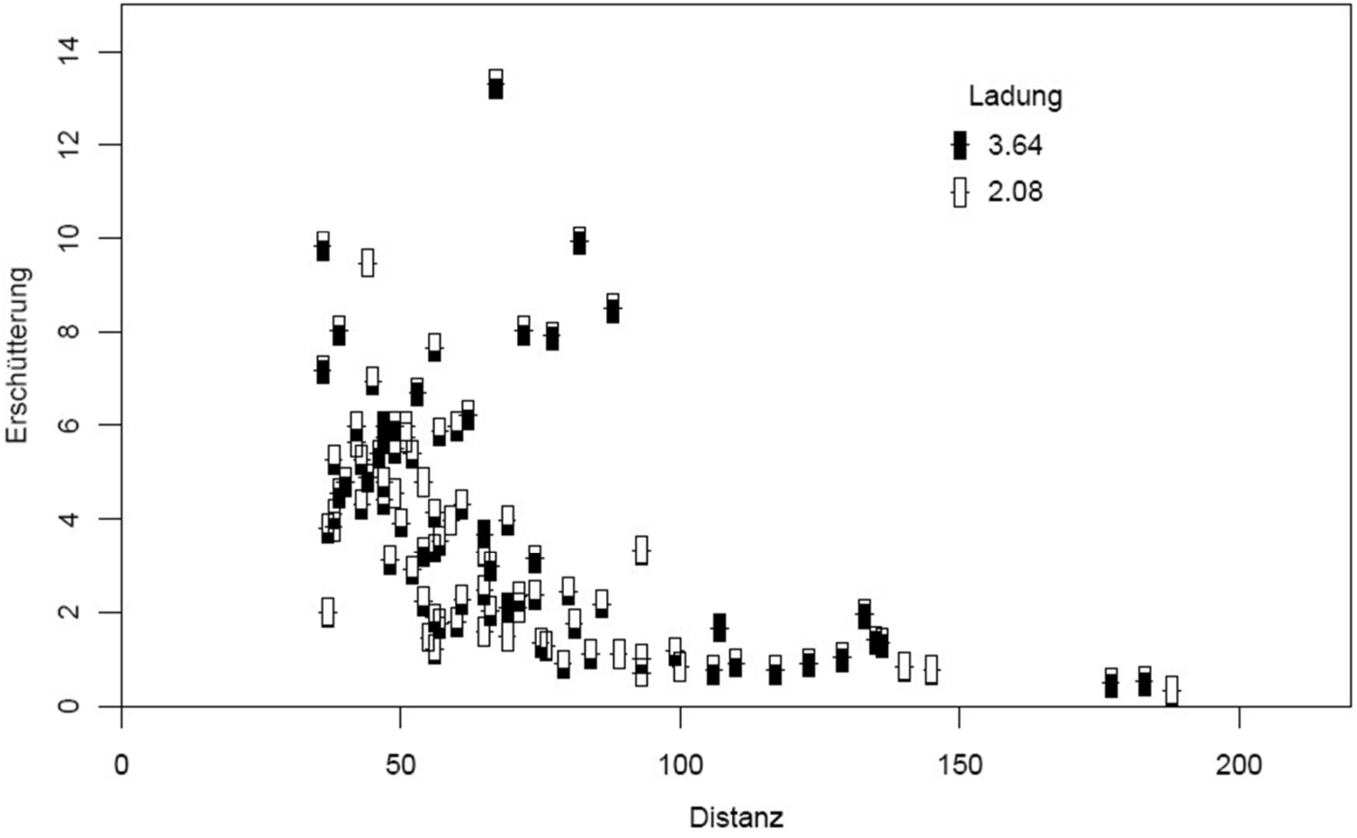
Regression Mathematics

→ See blackboard...

Applied Statistical Regression
AS 2012 – Week 01

Versatility of Linear Modeling

"Only" linear models: is that a problem? → **NO**



Goals with Linear Modeling

→To understand the causal relation, doing inference

- Does the fertilizer positively affect plant growth?
- Regression is a tool to give an answer on this
- However, showing causality is a different matter

→Target value prediction for new configurations

- What are the expected claims for auto insurance?
- Regression analysis formalizes “prior experience”
- It also provides an idea on the uncertainty of the prediction

Interpretation of Linear Models

- Linear models are mathematical formulae that formalize the relation between a target variable and a number of predictors.
- This formalization involves a deterministic / systematic part and an error term which stands for the random, non-systematic part.
- Usually, such linear models are a simplification of reality and are descriptive, but not causal.

Example: Automobile insurance…

Simple Regression

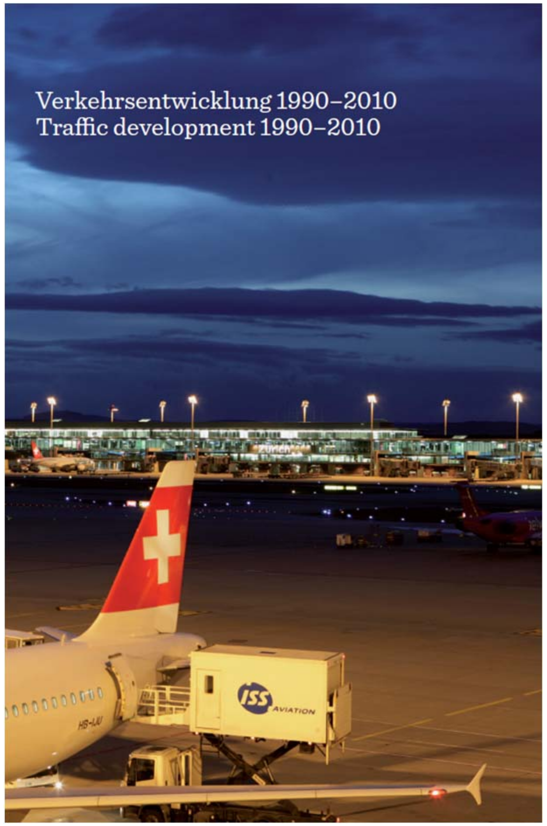
- Simple = There is only 1 predictor variable
- Advantage: easy visualization in a scatterplot
- Amounts to fitting a straight line or a curve
- Mathematically easier than multiple predictors

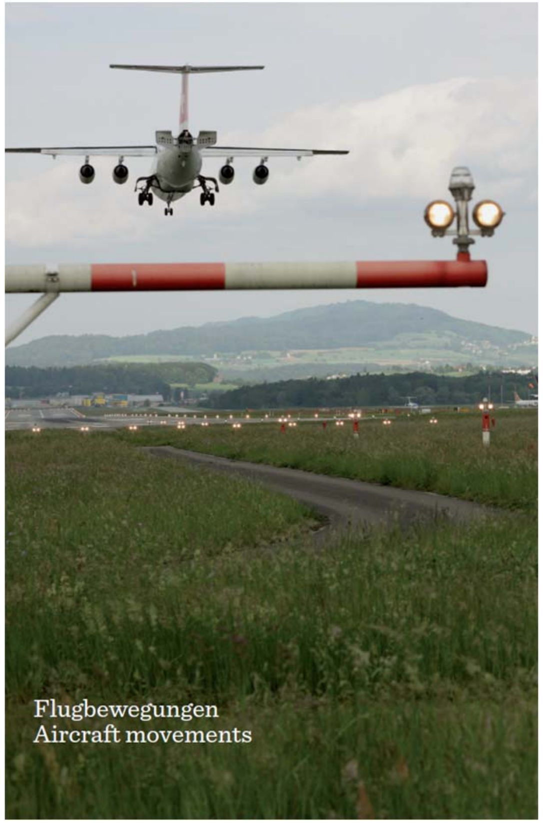
→ **an ideal start**

We will do non-parametric curve fitting first, then turn our attention to linear modelling. Later we do multiple regression, with the main focus on linear modelling.

Example: Airline Passengers

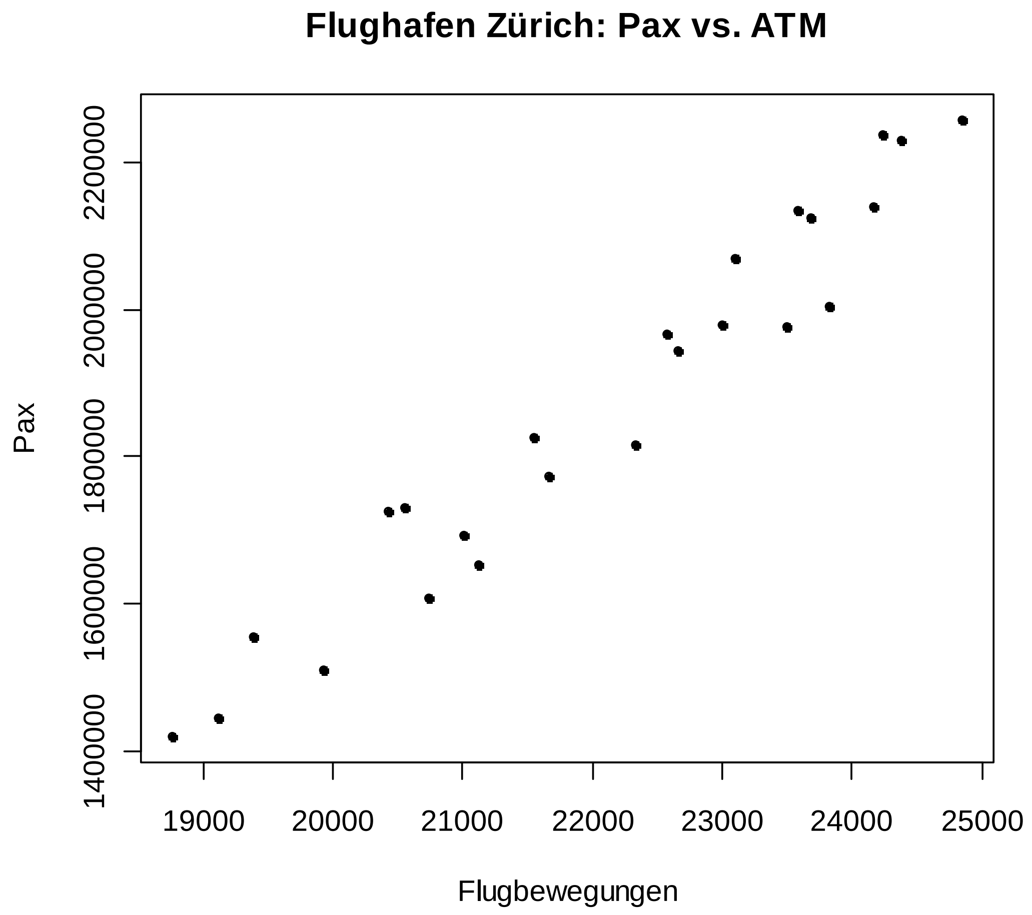
Each month, Zurich Airport publishes the number of air traffic movements and airline passengers. We study their relation.





Example: Airline Passengers

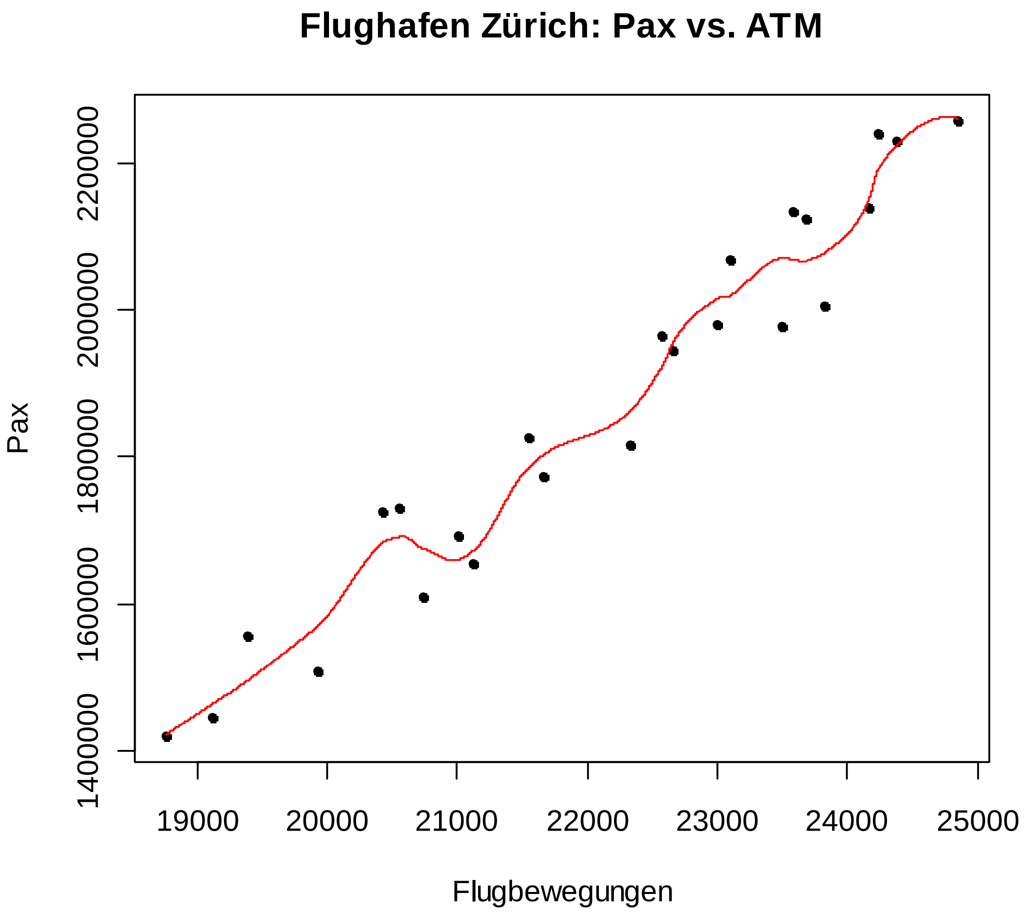
Month	Pax	ATM
2010-12	1‘730‘629	22‘666
2010-11	1‘772‘821	22‘579
2010-10	2‘238‘314	24‘234
2010-09	2‘139‘404	24‘172
2010-08	2‘230‘150	24‘377
...	...	...



Smoothing

We may use an arbitrary smooth function $f(\cdot)$ for capturing the relation between Pax and ATM.

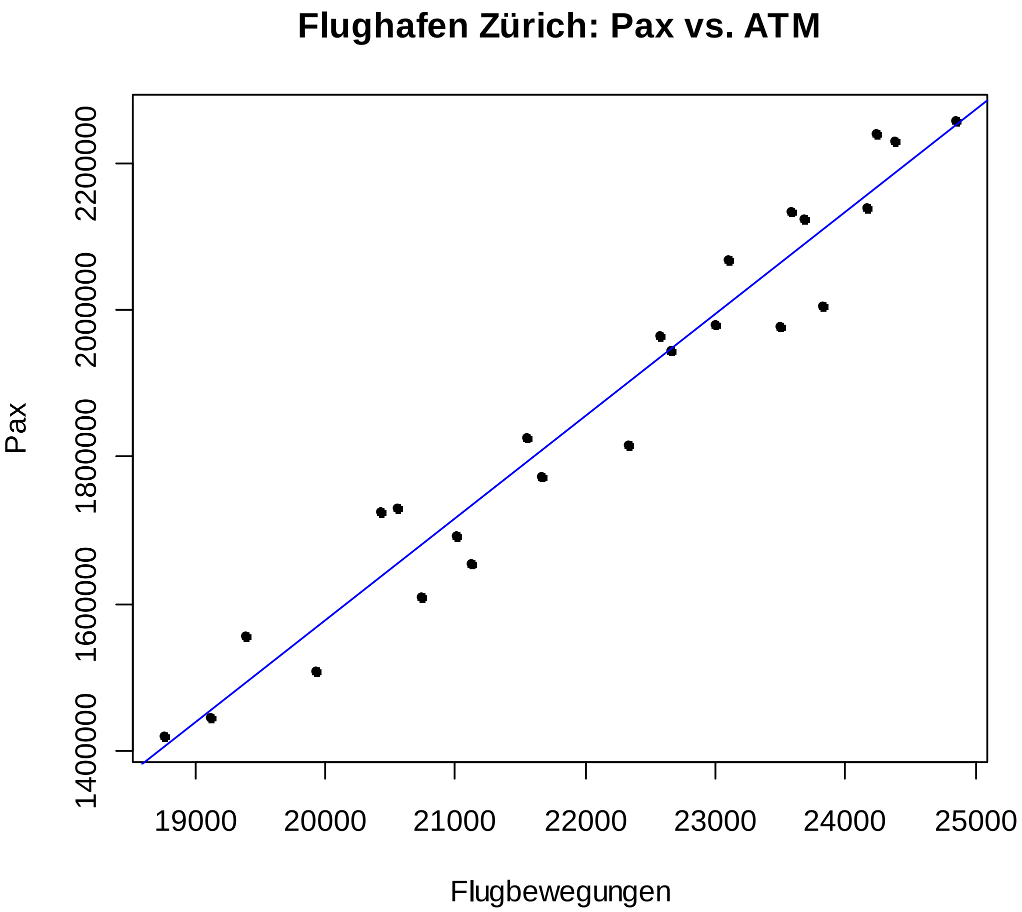
- It should fit well, but not follow the data too closely.
- The question is how the line/function are obtained.



Linear Modeling

A straight line represents the systematic relation between Pax and ATM.

- Only appropriate if the true relation is indeed a straight line
- The question is how the line/function are obtained.



Smoothing vs. Linear Modeling

Advantages and disadvantages of *smoothing*:

+ Flexibility
+ No assumptions are made
- Functional form remains unknown
- Danger of overfitting

Advantages and disadvantages of *linear modelling*:

+ Formal inference on the relation is possible
+ Better efficiency, i.e. less data required
- Only reasonable if the relation is linear
- Might falsely imply causality

Smoothing

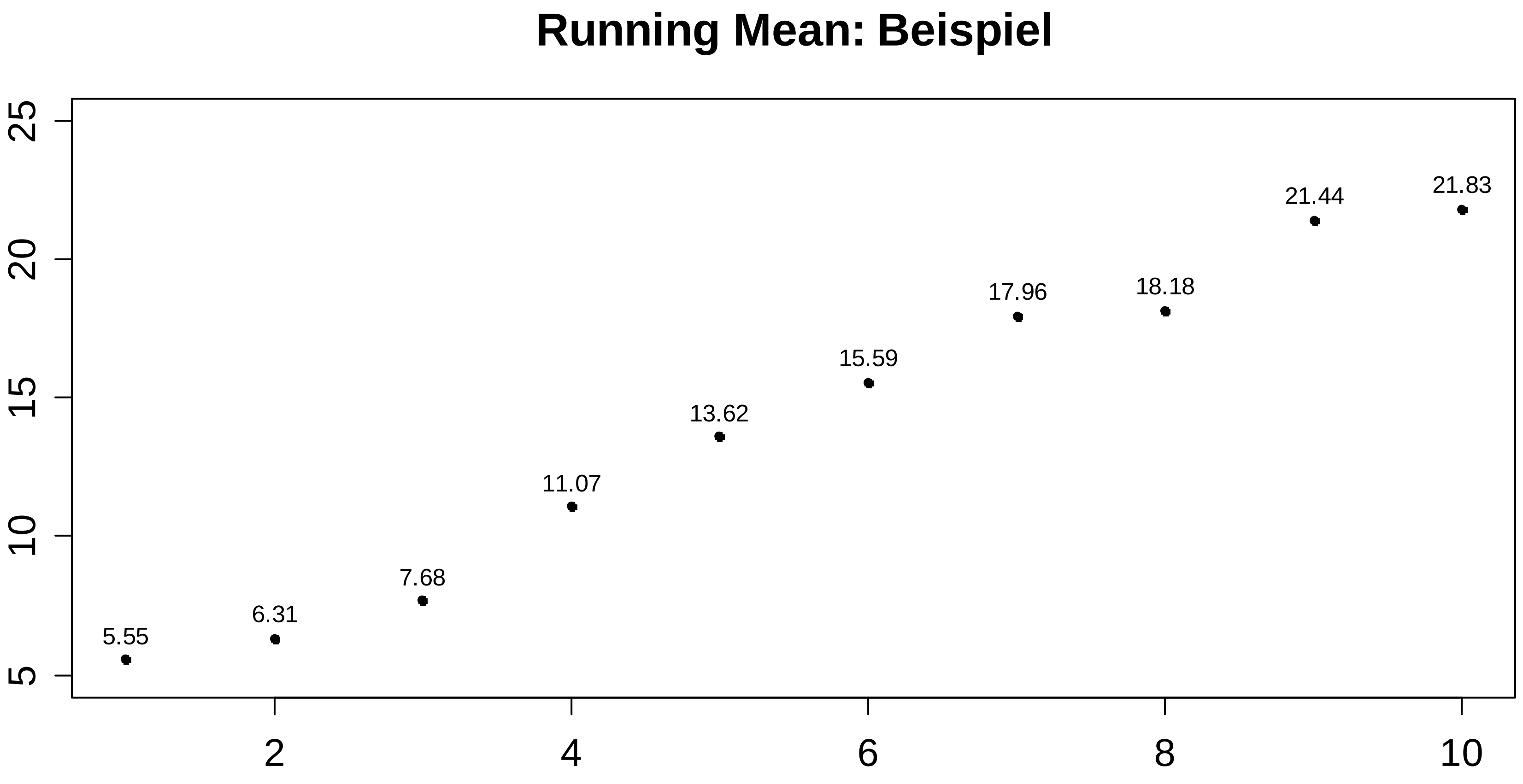
Our goal is *visualizing* the relation between the Y / response variable Pax and the x / predictor variable ATM.

→ we are not after a functional description of $f(\cdot)$

Since there is no parametric function that describes the response vs. predictor relation, smoothing is also termed **non-parametric regression analysis**.

Method/Idea: *"Running Mean"*
- take a window of x-values
- compute the mean of the y-values within the window
- this is an estimate for the function value at the window center

Running Mean: Example



Running Mean: Mathematics

RunningMean(x) = Mean of y-values over a window with width $\pm\lambda/2$ around x.

The *estimate* for $f(\cdot)$, denoted as $\hat{f}_\lambda(\cdot)$, is defined as follows:

$$\hat{f}_\lambda(x) = \frac{\sum_{j=1}^{n} w_j y_j}{\sum_{j=1}^{n} w_j} \quad ,$$

The *weights* are defined as $w_j = \begin{cases} 1 & falls\ |x - x_j| \leq \lambda/2 \\ 0 & sonst \end{cases}$,
and λ is the *window width*.

Running Mean: R-Implementation

- As an introductory exercise, it is instructive to code a function that computes and visualizes the running mean.

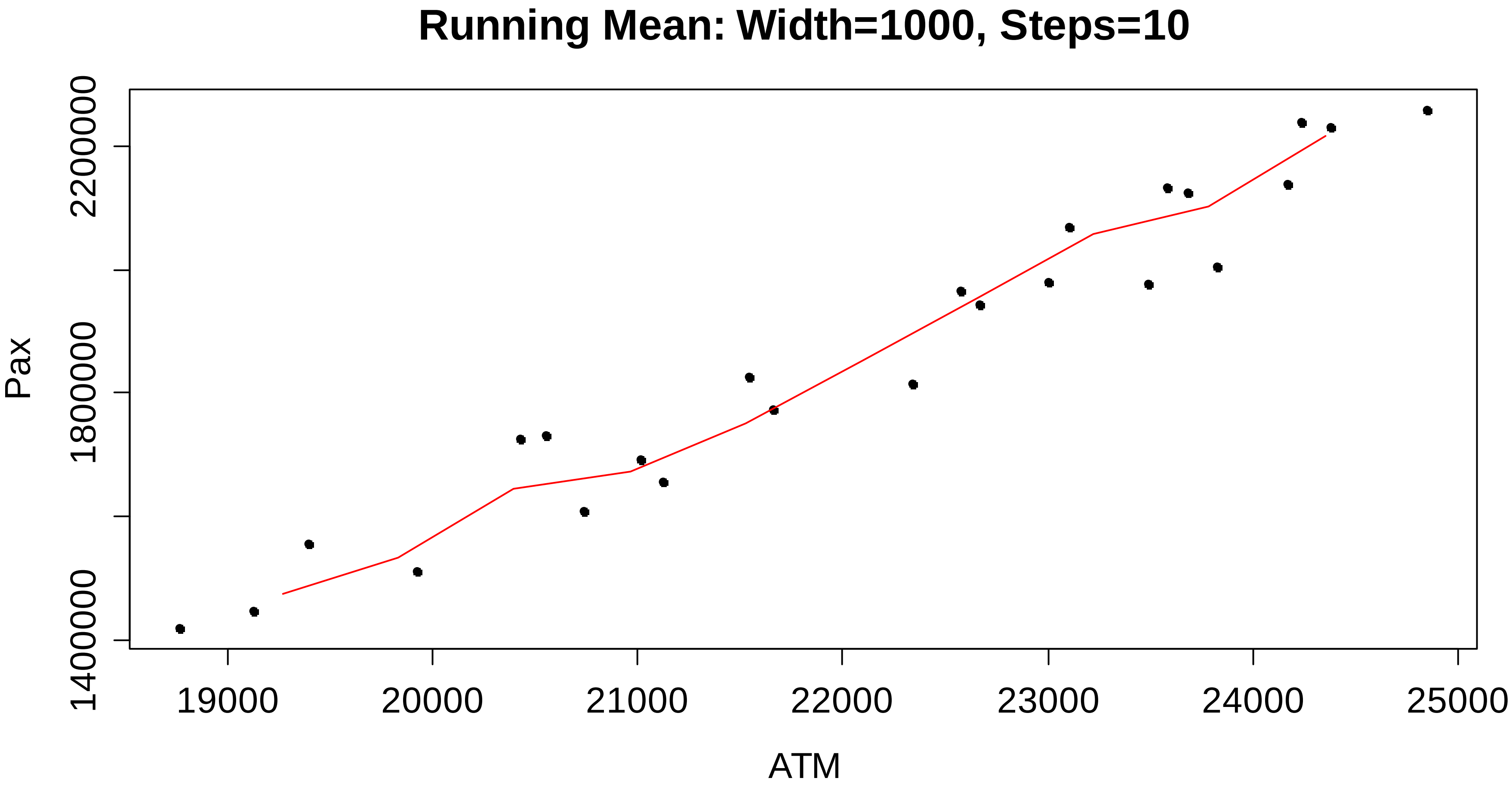
Arguments:	`xx=`	x values
	`yy=`	y values
	`width=`	window width
	`steps=`	# of points computed
	`plot=`	should the result be plotted?

- Alternatively, one can also use function `ksmooth()`with default settings. The window width can be adjusted by using argument `bandwidth=`.

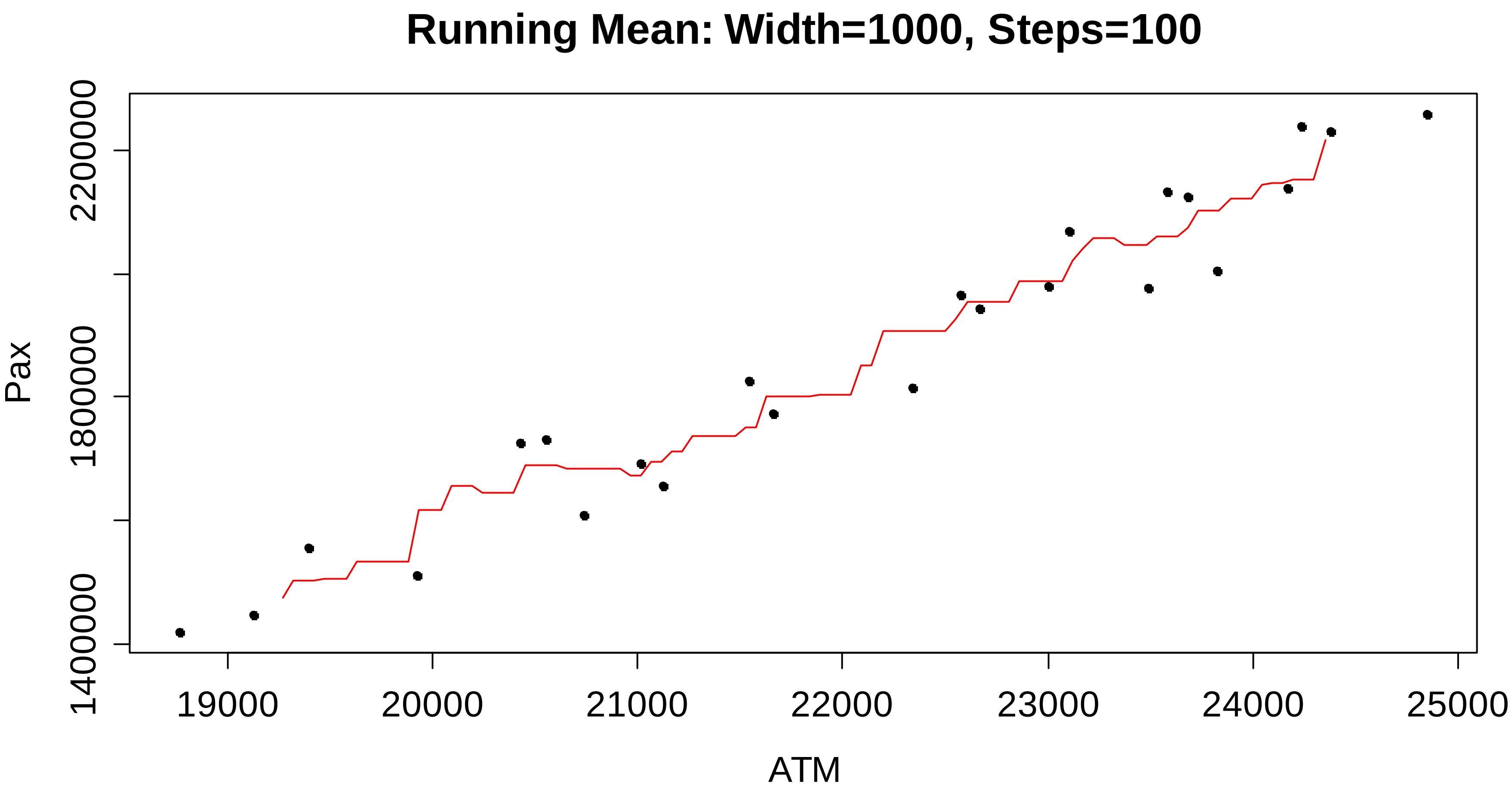
→ **We will now study the running mean fit...**

Applied Statistical Regression
AS 2012 – Week 01

Running Mean: Unique-Data



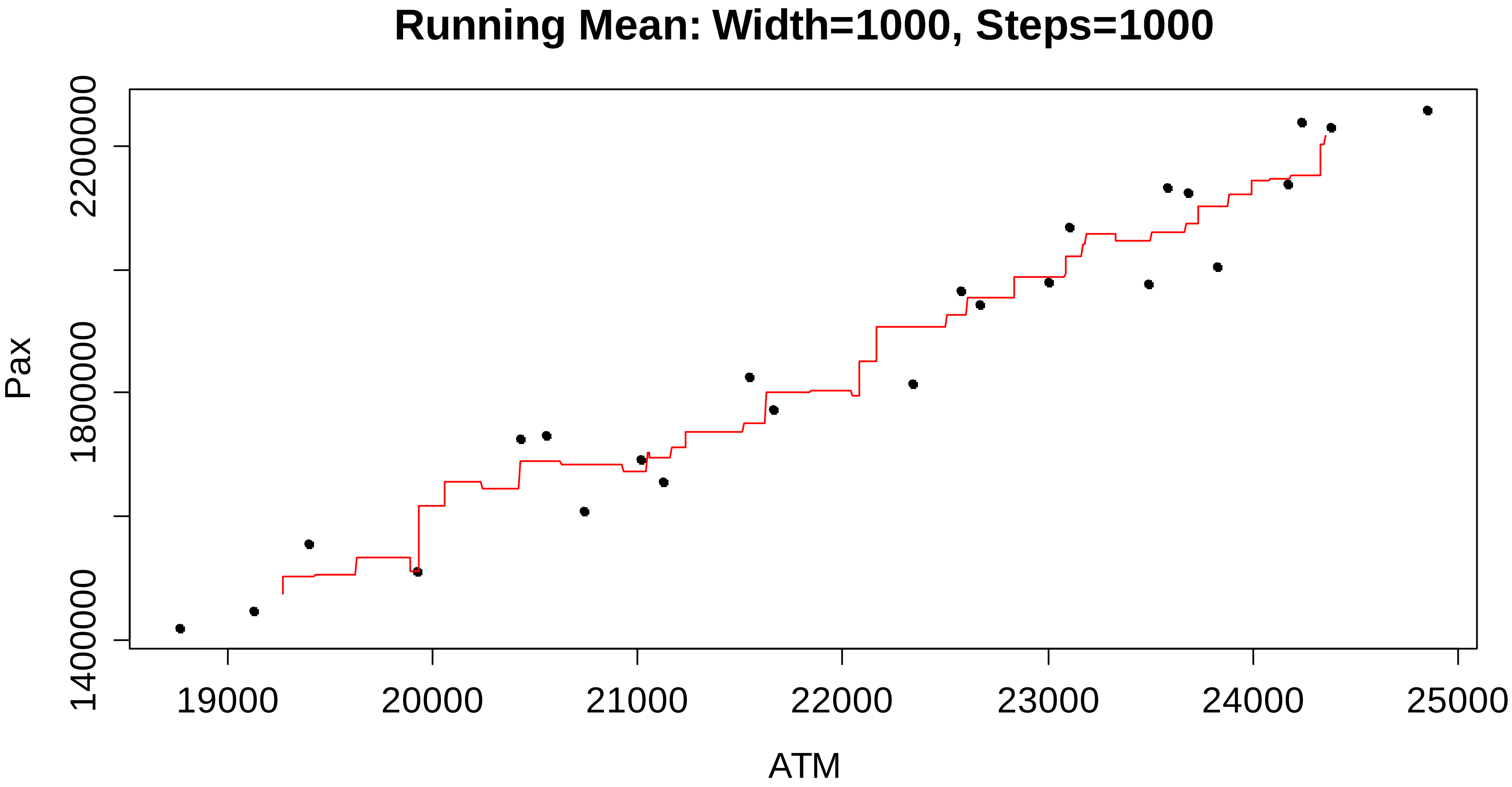
Running Mean: Unique-Data



Applied Statistical Regression

AS 2012 – Week 01

Running Mean: Unique-Data



Running Mean: Drawbacks

- The finer grained the evaluation points are, the less smooth the fitted function turns out to be. This is unwanted.
  Reason: *datapoints are "lost" abruptly.*

- For large window width, we loose a lot of information on the boundaries. For small windows however, we may have too few points withing the window, and thus instability.

→ ***There are much better smoothing algorithms!***

We will introduce:

a) a *Gaussian Kernel Smoother*, and

b) the robust *LOESS-Smoother*

Gaussian Kernel Smoother

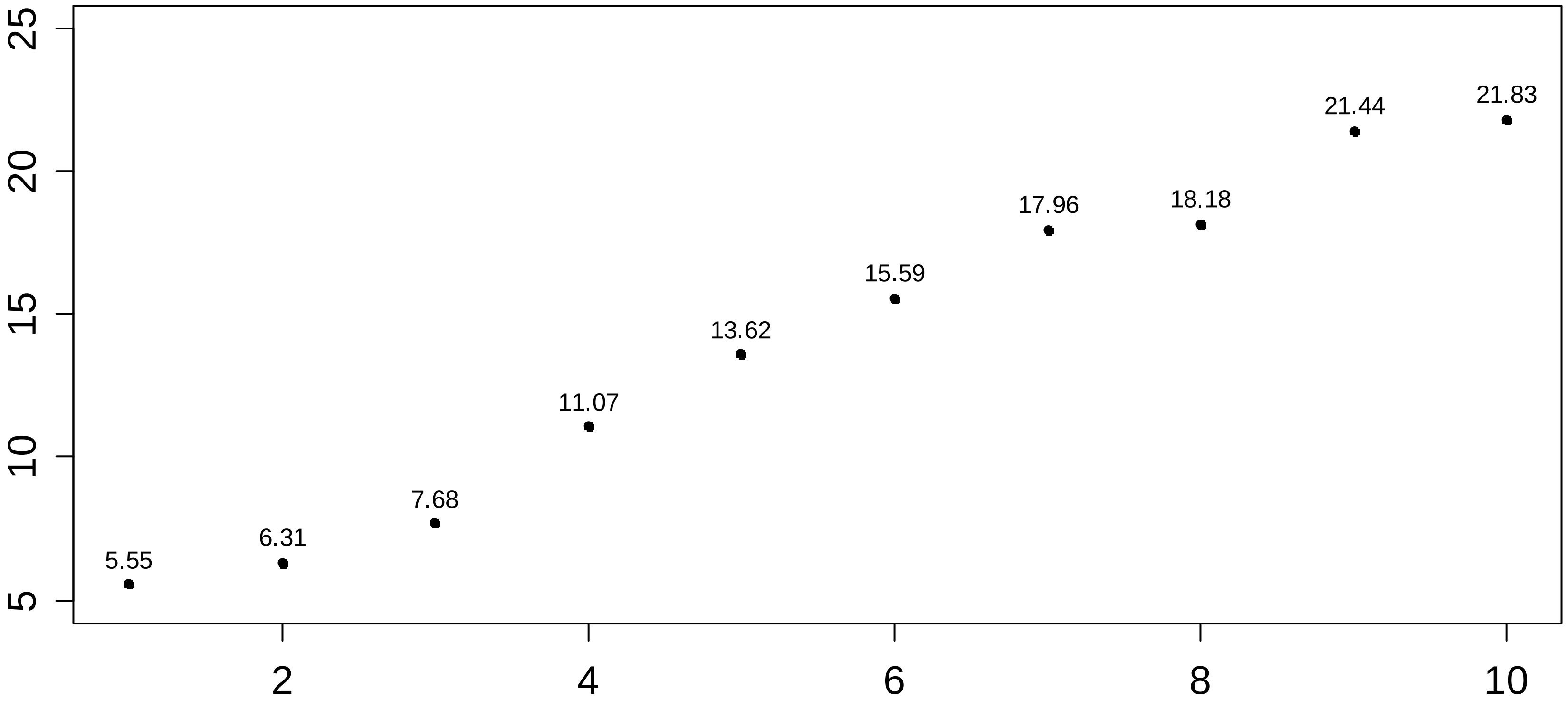
KernelSmoother(x) = Gaussian bell curve weighted average of y-values around x.

The estimate for $f(\cdot)$, denoted as $\hat{f}_\lambda(\cdot)$, is defined as:

$$\hat{f}_\lambda(x) = \frac{\sum_{j=1}^{n} w_j y_j}{\sum_{j=1}^{n} w_j} \quad ,$$

The weights are defined as: $w_j = \exp\left(\frac{(x - x_j)^2}{\lambda}\right)$, i.e.
the window is infinitely wide,
but distant observation obtain little weight.

Gaussian Kernel Smoother: Idea

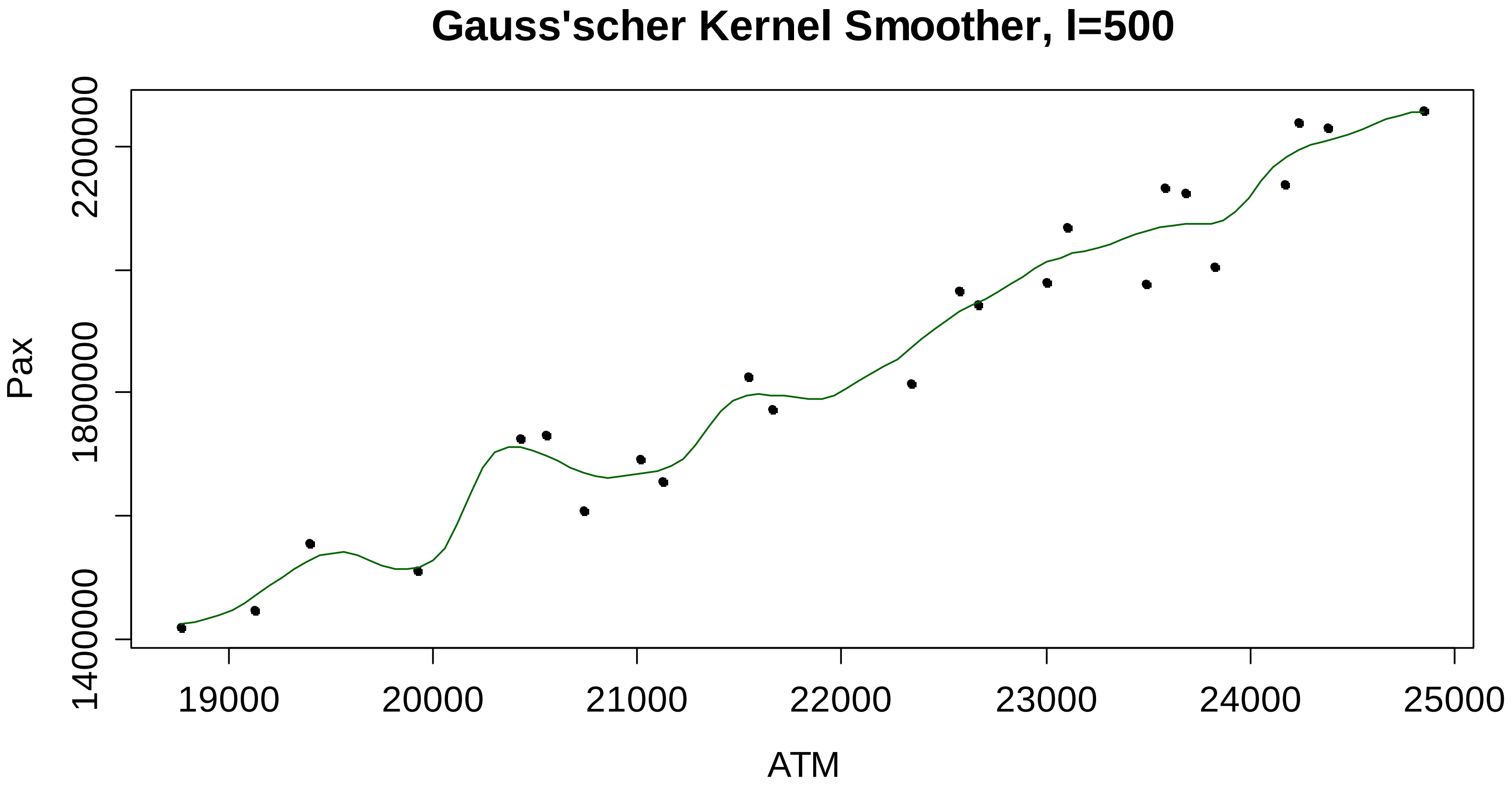


Applied Statistical Regression

AS 2012 – Week 01

Gaussian Kernel Smoother: Unique-Data

```
> ks.gauss <- ksmooth(ATM, Pax, kernel="normal", band=500)
> plot(ATM, Pax, xlab="ATM", ylab="Pax", pch=20)
> lines(ks.gauss$x, ks.gauss$y, col="darkgreen", lwd=1.5)
```



LOESS-Smoother

The LOESS-Smoother is better, more flexible and more robust than the Gaussian Kernel Smoother. It should be prefered!

It works as follows:

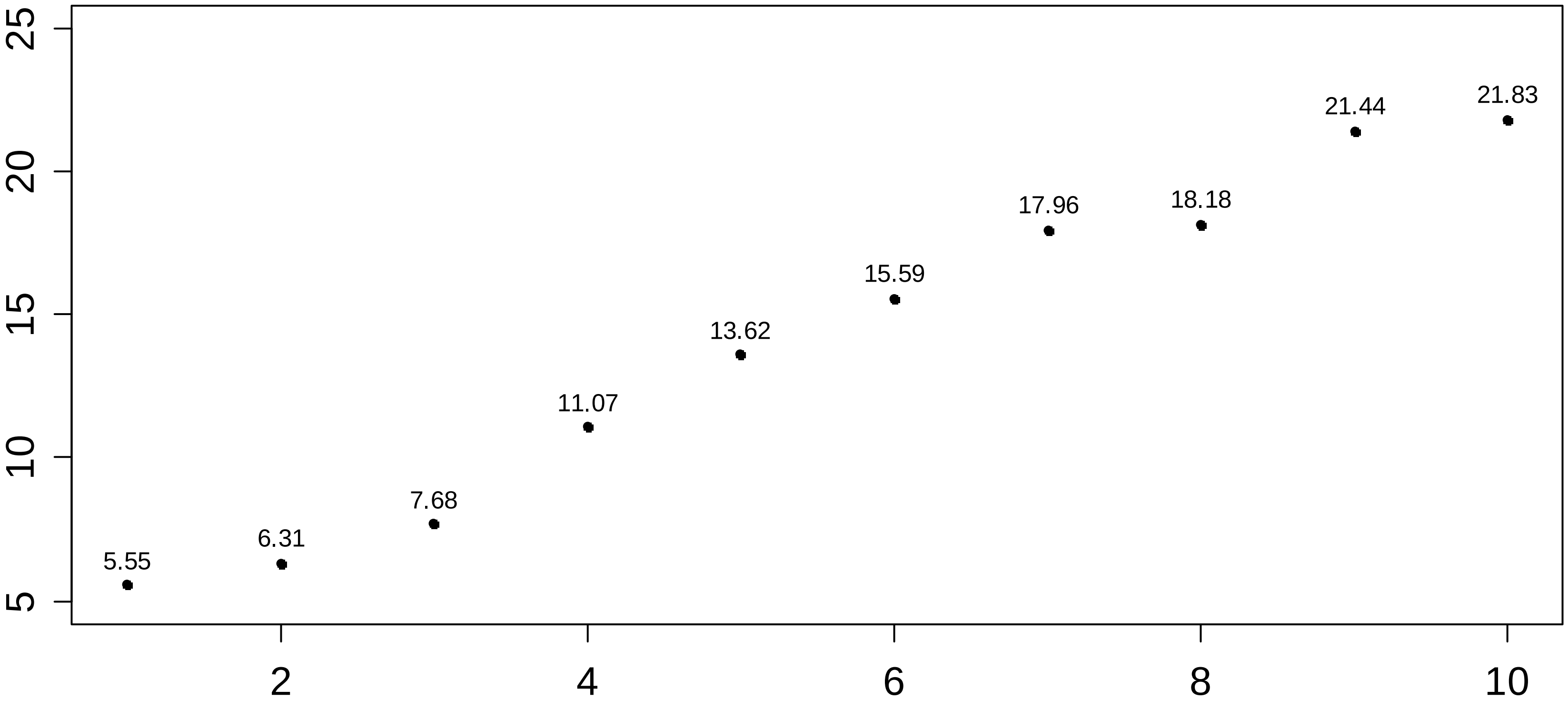
1) Choose a window of fixed width

2) For this window, a straight line (or a parabola) is fitted to the datapoints within, using a robust fitting method.

3) Predicted value at window center := fitted value

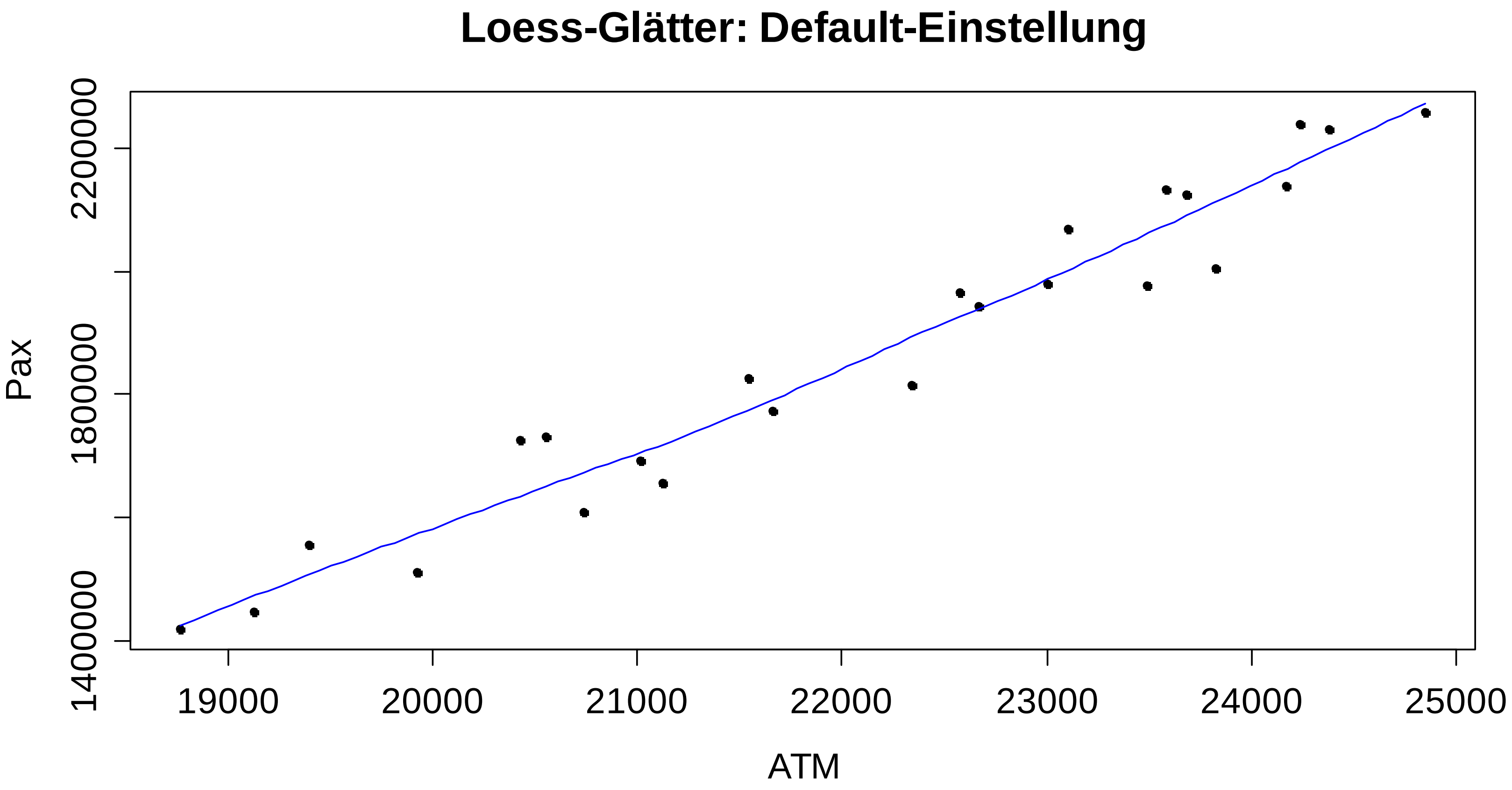
4) Slide the window over the entire x-range

LOESS-Smoother: Idea



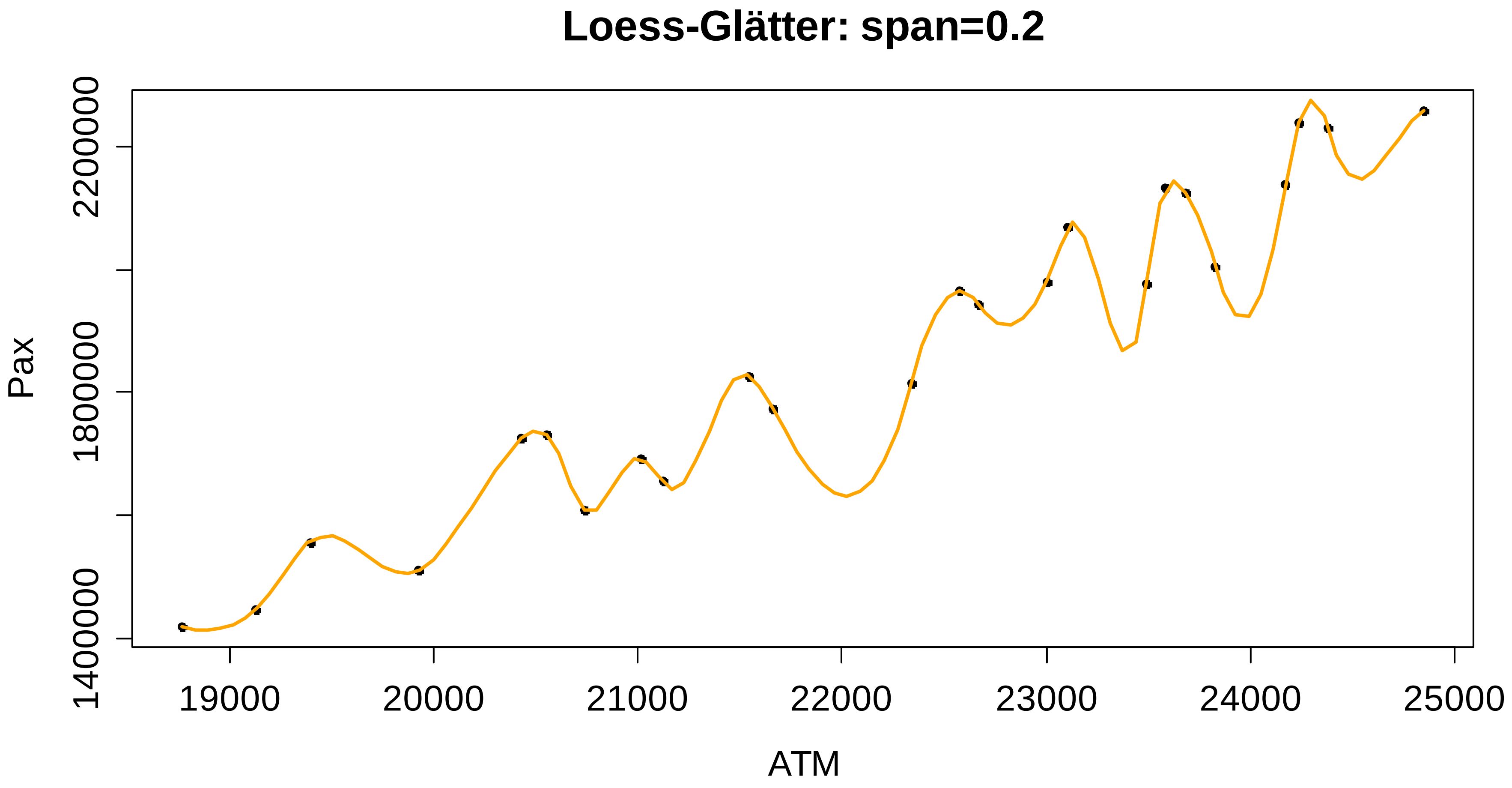
LOESS-Smoother: Unique-Data

```
> fit <- loess(Pax~ATM, data=unique2010)
> new.x <- seq(min(ATM), max(ATM), length=100)
> new.y <- predict(fit, newdata=data.frame(ATM=new.x))
```



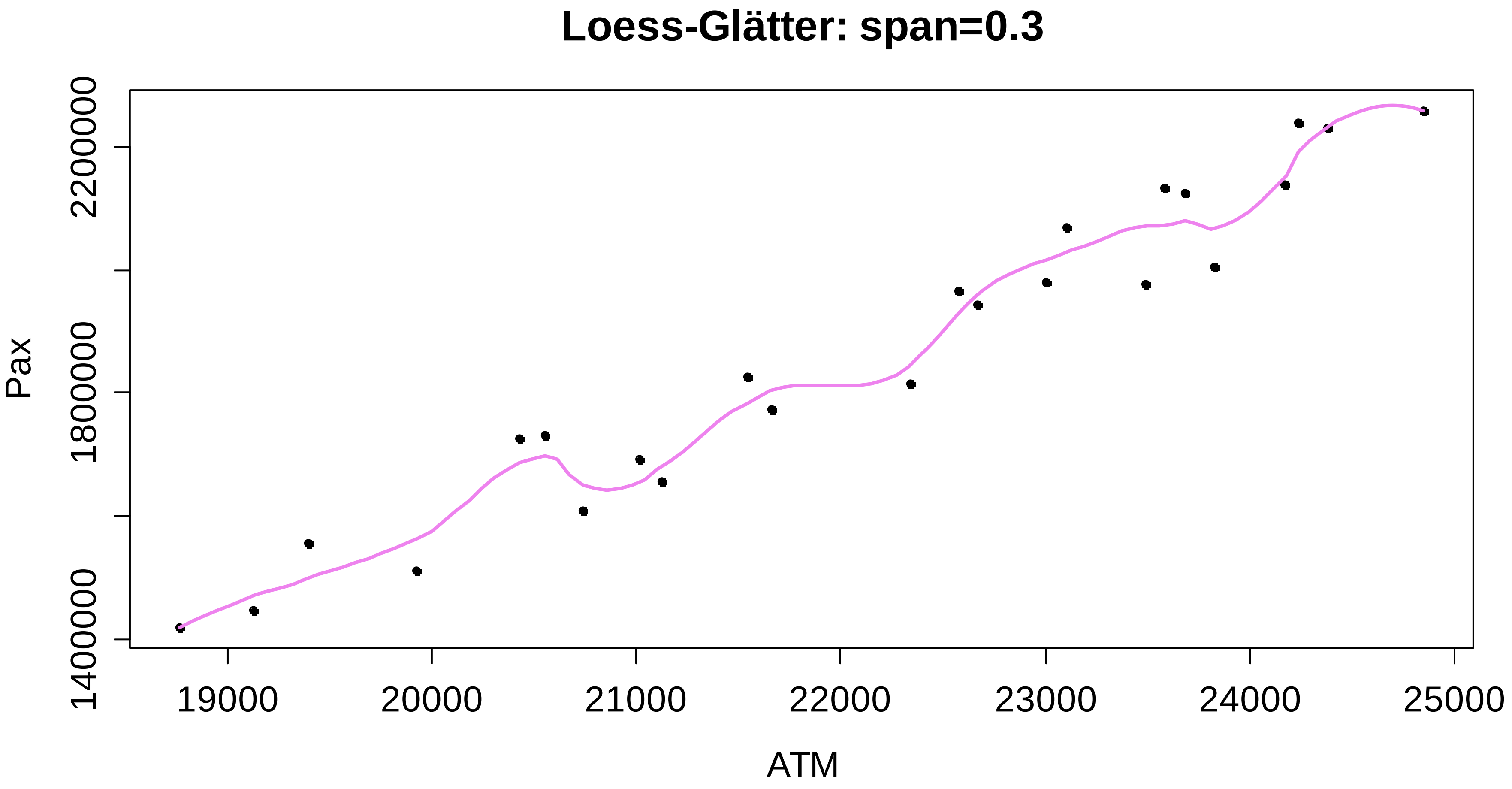
Choice of the Smoothing Parameter

→ Is usually done by try and error eyeballing.



Choice of the Smoothing Parameter

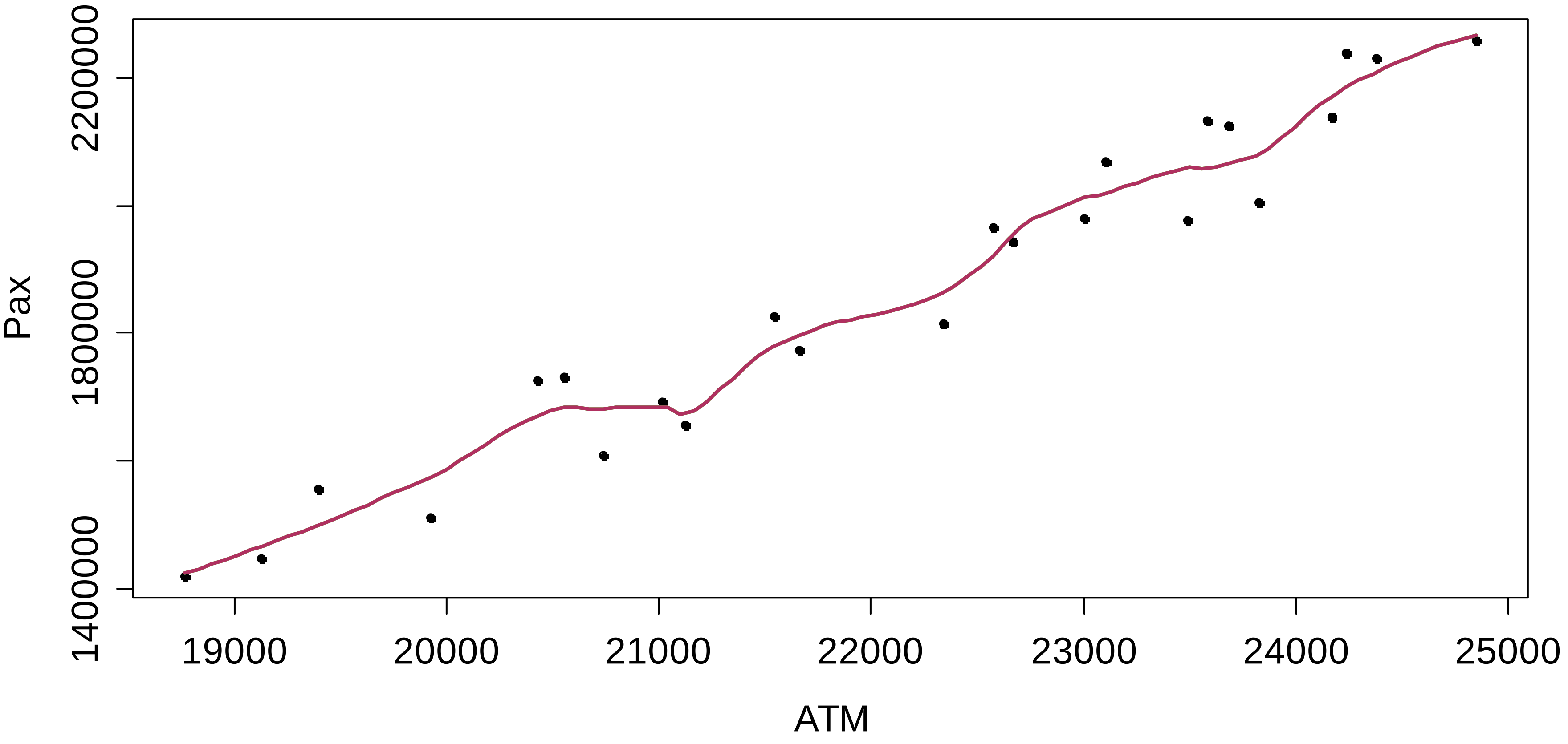
→ Is usually done by try and error eyeballing.



Choice of the Smoothing Parameter

→ Is usually done by try and error eyeballing.

Loess-Glätter: span=0.4



Simple Linear Regression

The more air traffic movements, the more passengers there are. The relation seems to be linear, which is of course also the mathematically most simple way of describing the relation.

$$f(x) = \beta_o + \beta_1 x, \text{ resp. } Pax = \beta_0 + \beta_1 \cdot ATM$$

Name/meaning of the two parameters in the equation: $\beta_0 =$ "Intercept" $\beta_1 =$ "Slope"

Fitting a straight line into a 2-dimensional scatter plot is known as **simple linear regression**. This is because:

- there is just one single predictor variable ("*simple*").
- the relation is linear in the parameters ("*linear*").

Model, Data & Random Errors

No we are bringing the data into play. The regression line will not run through all the data points. Thus, there are random errors:

$$y_i = \beta_0 + \beta_1 x_i + E_i, \text{ for all } i = 1, \ldots, n$$

Meaning of variables/parameters:

y_i is the response variable (Pax) of observation i .

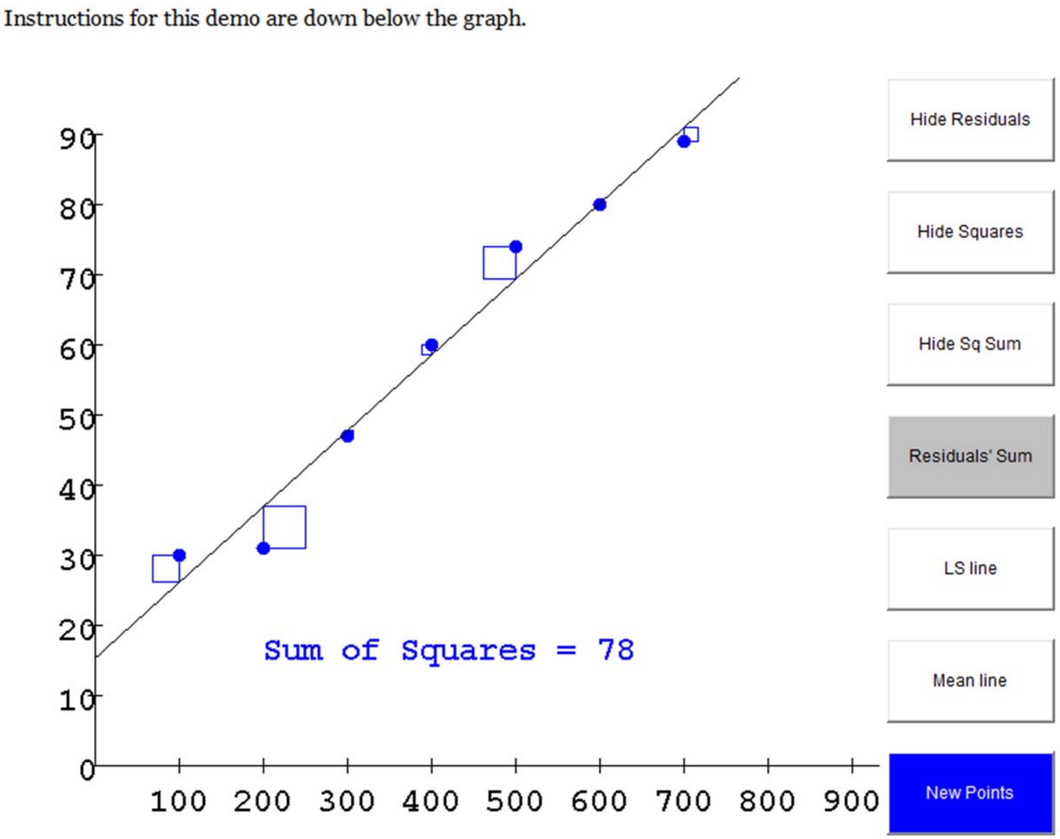
x_i is the predictor variable (ATM) of observation i .

β_0, β_1 are the regression coefficients. They are unknown previously, and need to be estimated from the data.

E_i is the residual or error, i.e. the random difference between observation and regression line.

Least Squares Fitting

→ http://hspm.sph.sc.edu/courses/J716/demos/LeastSquares/LeastSquaresDemo.html



Instructions for this demo are down below the graph.

We need to fit a straight line that fits the data well.

Many possible solutions exist, some are good, some are worse.

Our paradigm is to fit the line such that the squared errors are minimal.

Least Squares: Mathematics

The paradigm in verbatim...

Given a set of data points $(x_i, y_i)_{i=1,\dots,n}$, the goal is to fit the regression line such that the sum of squared differences between observed value y_i and regression line is minimal. The function

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta x_i))^2 = \min!$$

measures, how well the regression line, defined by β_0, β_1, fits the data. The goal is to minimize the function.

Solution: → **see next slide...**

Solution Idea: Partial Derivatives

- We are taking partial derivatives on the function $Q(\beta_0, \beta_1)$ with respect to both arguments β_0 and β_1. As we are after the minimum of the function, we set them to zero:

$$\frac{\partial Q}{\partial \beta_0} = 0 \text{ and } \frac{\partial Q}{\partial \beta_1} = 0$$

- This results in a linear equation system, which (here) has two unknowns β_0, β_1, but also two equations. These are also known under the name *normal equations*.

- The solution for β_0, β_1 can be written explicitly as a function of the data pairs $(x_i, y_i)_{i=1,\ldots,n}$, **see next slide...**

Least Squares: Solution

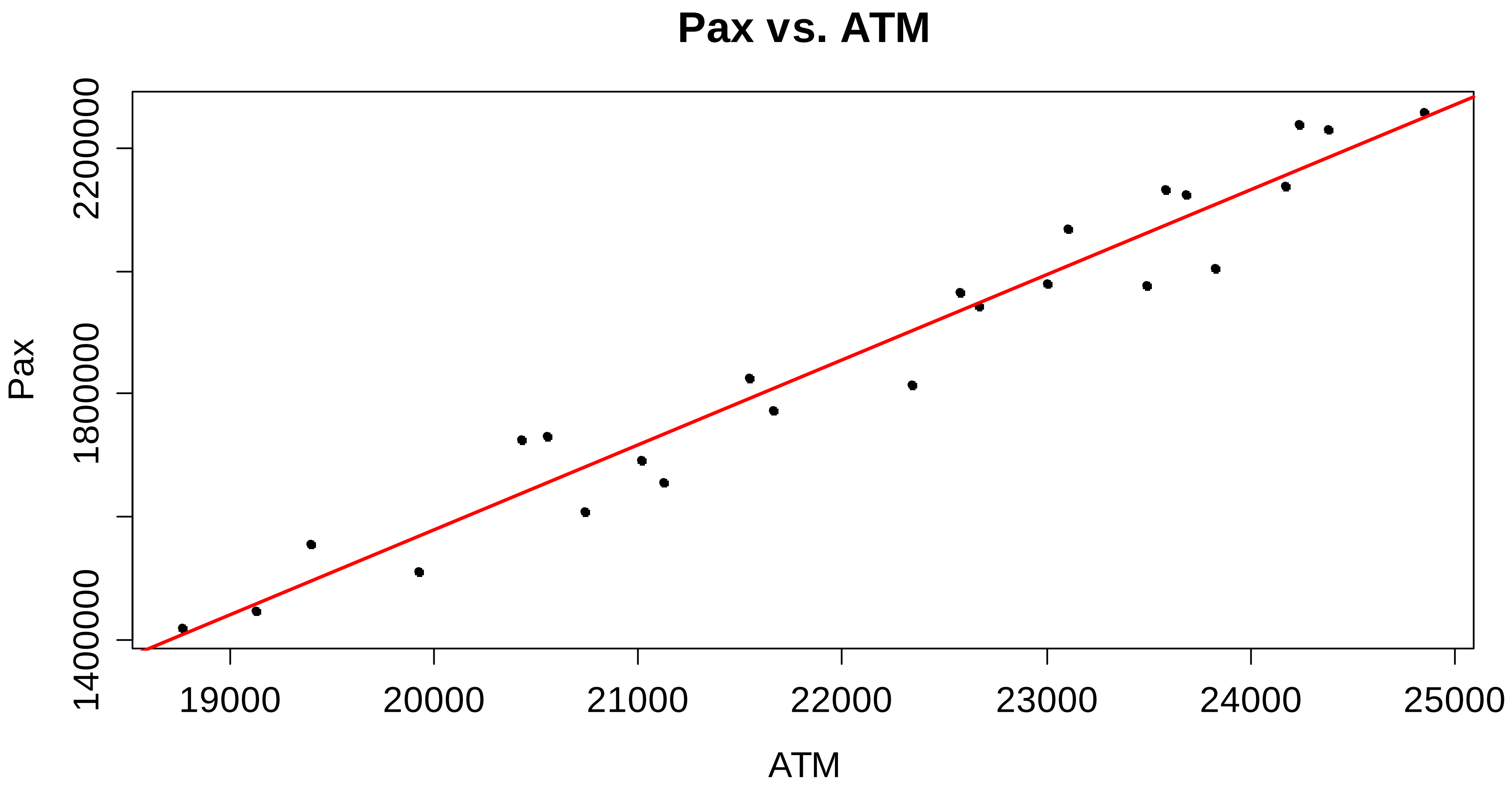
According to the least squares paradigm, the best fitting regression line is, i.e. the optimal coefficients are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad \text{und} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

- For a given set of data points $(x_i, y_i)_{i=1,\ldots,n}$, we can determine the solution with a pocket calculator (...or better, with R).
- **The solution for our example Pax vs. ATM:**
  $\hat{\beta}_1 = 138.8,\ \hat{\beta}_0 = -1'197'682$

→ `lm(Pax ~ ATM, data=airpax)`

Least Squares Regression Line



Is This a Good Model for Predicting the Pax Number from the ATM?

a) Beyond the range of observed data

Unknown, but most likely not...

b) Within the range of observed data

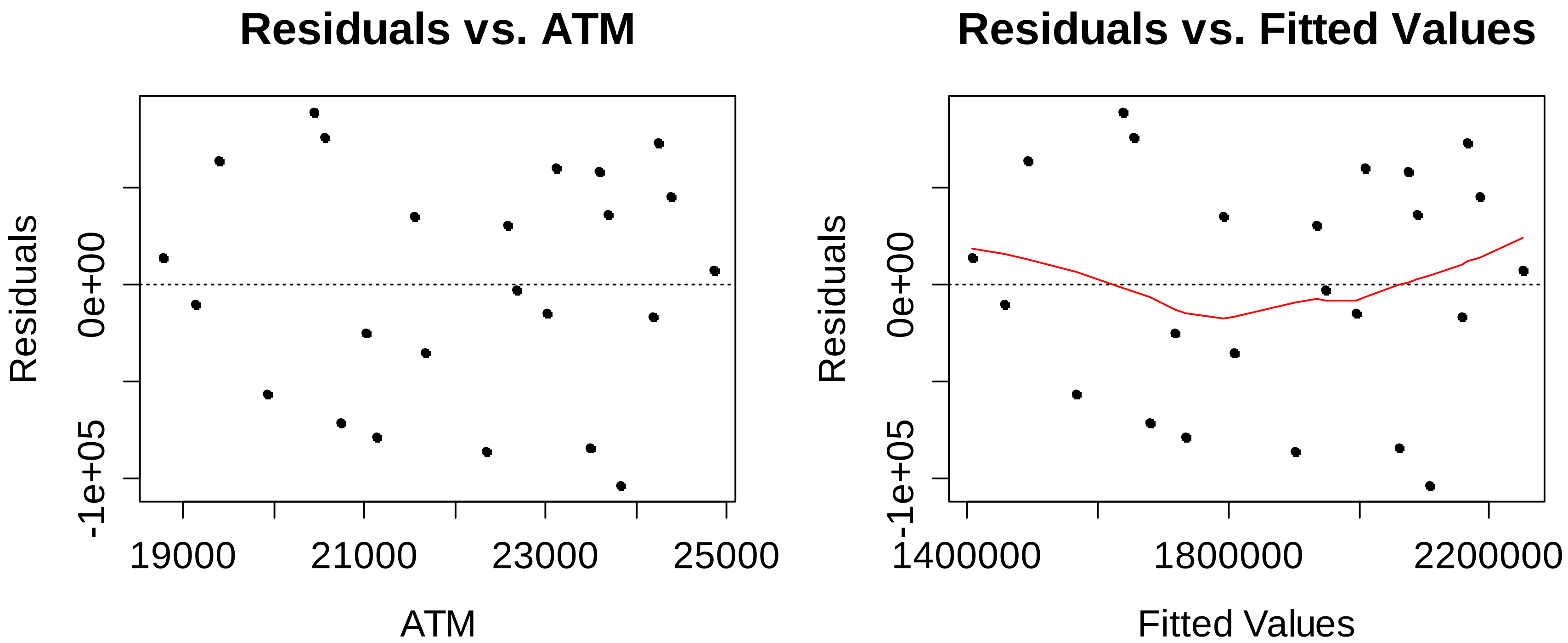
Yes, under the following conditions:

- the relation is in truth a straight line, i.e. $E[E_i]=0$
- the scatter of the errors is constant, i.e. $Var(E_i)=\sigma^2$
- the data are uncorrelated (from a representative sample)
- the errors are approximately normally distributed

→ **Fodder for thougt: 9/11, SARS, Eyjafjallajökull...?**

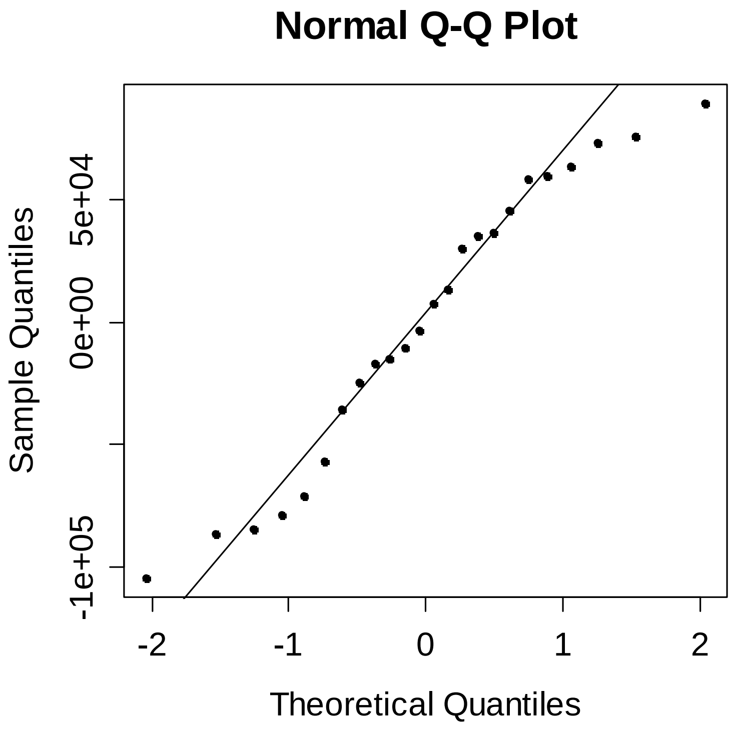
Model Diagnostics

For assessing the quality of the regression line, we need to (at least roughly) check whether the assumptions are met: $E[E_i]=0$ and $Var(E_i)=\sigma^2$ can be reviewed by:



Model Diagnostics

For assessing the quality of the regression line, we need to (at least roughly) check whether the assumptions are met: Gaussian distribution can be reviewed by:



We will revisit model diagnostics again later in this course, where it will be discussed more deeply.

"Residuals vs. Fitted" and the "Normal Plot" will always stay at the heart of model diagnostics.

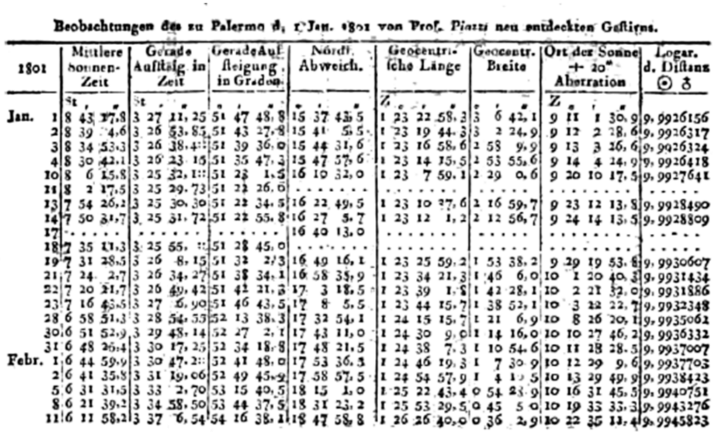
Applied Statistical Regression
AS 2012 – Week 01

Why Least Squares?

History...

Within a few years (1801, 1805), the method was developed independently by Gauss and Legendre. Both were after solving applied problems in astronomy...

Source: → http://de.wikipedia.org/wiki/Methode_der_kleinsten_Quadrate



Beobachtungen des zu Palermo d. 1. Jan. 1801 von Prof. Piazzi neu entdeckten Geſtirns.

1801	Mittlere Sonnen-Zeit	Gerade Aufſteig. in Zeit	Gerade Aufſteigung in Graden	Nördl. Abweich.	Geocentriſche Länge	Geocentr. Breite	Ort der Sonne + 20" Aberration	Logar. d. Diſtanz ⊙ ♂
	St ′ ″	St ′ ″	° ′ ″	° ′ ″	Z ° ′ ″	° ′ ″	Z ° ′ ″	
Jan. 1	8 43 27,8	3 27 11,25	51 47 48,8	15 37 43,5	1 23 22 58,3	3 6 42,1	9 11 1 30,9	9,9926156
2	8 39 4,6	3 26 53,85	51 43 27,8	15 41 5,5	1 23 19 44,3	3 2 24,9	9 12 2 28,6	9,9926317
3	8 34 53,3	3 26 38,4::	51 39 36,0	15 44 31,6	1 23 16 58,6	2 58 9,9	9 13 3 26,6	9,9926324
4	8 30 42,1	3 26 23,15	51 35 47,3	15 47 57,6	1 23 14 15,5	2 53 55,6	9 14 4 24,9	9,9926418
10	8 6 15,8	3 25 32,1::	51 23 1,5	16 10 32,0	1 23 7 59,1	2 29 0,6	9 20 10 17,5	9,9927641
11	8 2 17,5	3 25 29,73	51 22 26,0	....	....	......	....	.....
13	7 54 26,2	3 25 30,30	51 22 34,5	16 22 49,5	1 23 10 27,6	2 16 59,7	9 23 12 13,8	9,9928490
14	7 50 31,7	3 25 31,72	51 22 55,8	16 27 5,7	1 23 12 1,2	2 12 56,7	9 24 14 13,5	9,9928809
17	......	....	....	16 40 13,0	....	......	....	.....
18	7 35 11,3	3 25 55, ::	51 28 45,0	....	....	......	....	.....
19	7 31 28,5	3 26 8,15	51 32 2,3	16 49 16,1	1 23 25 59,2	1 53 38,2	9 29 19 53,8	9,9930607
21	7 24 2,7	3 26 34,27	51 38 34,1	16 58 35,9	1 23 34 21,3	1 46 6,0	10 1 20 40,3	9,9931434
22	7 20 21,7	3 26 49,42	51 42 21,3	17 3 18,5	1 23 39 1,8	1 42 28,1	10 2 21 32,0	9,9931886
23	7 16 43,5	3 27 6,90	51 46 43,5	17 8 5,5	1 23 44 15,7	1 38 52,1	10 3 22 22,7	9,9932348
28	6 58 51,3	3 28 54,55	52 13 38,3	17 32 54,1	1 24 15 15,7	1 21 6,9	10 8 26 20,1	9,9935062
30	6 51 52,9	3 29 48,14	52 27 2,1	17 43 11,0	1 24 30 9,0	1 14 16,0	10 10 27 46,2	9,9936332
31	6 48 26,4	3 30 17,25	52 34 18,8	17 48 21,5	1 24 38 7,3	1 10 54,6	10 11 28 28,5	9,9937007
Febr. 1	6 44 59,9	3 30 47,2::	52 41 48,0	17 53 36,3	1 24 46 19,3	1 7 30,9	10 12 29 9,6	9,9937703
2	6 41 35,8	3 31 19,06	52 49 45,9	17 58 57,5	1 24 54 57,9	1 4 19,5	10 13 29 49,9	9,9938423
5	6 31 31,5	3 33 2,70	53 15 40,5	18 15 1,0	1 25 22 43,4	0 54 23,9	10 16 31 45,5	9,9940751
8	6 21 39,2	3 34 58,50	53 44 37,5	18 31 23,2	1 25 53 29,5	0 45 5,0	10 19 33 33,3	9,9943276
11	6 11 58,2	3 37 6,54	54 16 38,1	18 47 58,8	1 26 26 40,0	0 36 2,9	10 22 35 11,4	9,9945823

Carl Friedrich Gauss

Adrien-Marie Legendre

Why Least Squares?

Mathematics...

- Least Squares is simple in the sense that the solution is known in closed form as a function of $(x_i, y_i)_{i=1,\ldots,n}$.
- The line runs through the center of gravity $(\overline{x}, \overline{y})$
- The sum of residuals adds up to zero: $\sum_{i=1}^{n} r_i = 0$
- Some deeper mathematical optimality can be shown when analyzing the large sample properties of the estimates $\hat{\beta}_0, \hat{\beta}_1$ This is especially true under the assumption of normally distributed errors E_i.

Gauss-Markov-Theorem

Mathematical optimality result for the Least Squares line.

It only holds if the following conditions are met:

- the relation is in truth a straight line, i.e. $E[E_i]=0$
- the scatter of the errors is constant, i.e. $Var(E_i)=\sigma^2$
- the errors are uncorrelated, i.e. $Cov(E_i,E_j)=0,\ if\ \ i\neq j$

Not yet required:

- ~~the errors are normally distributed: $E_i \sim N(0,\sigma_E^2)$~~

Gauss-Markov-Theorem:

- Least Squares yields the *best linear unbiased estimates*

Properties of the Least Square Estimates

Under the conditions above, the estimates are unbiased:

$$E[\hat{\beta}_0] = \beta_0 \text{ and } E[\hat{\beta}_1] = \beta_1$$

The variances of the estimates are as follows:

$$Var(\hat{\beta}_0) = \sigma_E^2 \cdot \left(\frac{1}{n} + \frac{\overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right) \text{ and } Var(\hat{\beta}_1) = \frac{\sigma_E^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Precise estimates are obtained with:

- a large number of observations n
- a good scatter in the predictor x_i
- an informative/useful predictor, making σ_E^2 small