Applied Statistical Regression
AS 2013 – Week 02 & Week 04

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 23, 2013

Applied Statistical Regression
AS 2013 – Week 02

Your Lecturer

Name:	Marcel Dettling
Age:	39 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 2013 – Week 02

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.

Applied Statistical Regression
AS 2013 – Week 02

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

Applied Statistical Regression
AS 2013 – Week 02

Regression Mathematics

→ **See blackboard...**

Applied Statistical Regression
AS 2013 – Week 02

What is Regression?

Example: *Fresh Water Tank on*  *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 2013 – Week 02

Goals with Linear Modeling

→*To understand the relation between*

- 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

Regression: Goals

1) Understanding the relation between y and $x_1, ..., x_p$

The aim is to pin down which of the predictors have influence on the response variable, as well as to quantify the strength of this relation. There is a battery of statistics and tests that address these questions.

2) Prediction

The regression equation can be used for predicting the expected response value $\hat{y}$ for an arbitrary predictor configuration $x_1, ..., x_p$. We will not only generate point predictions, but can also attribute a prediction interval that quantifies the involved uncertainty.

Applied Statistical Regression
AS 2013 – Week 02

Simple Regression

In this course, we first discuss *simple regression*, where there is only one single predictor variable. Later, we will extend this to *multiple regression*, where many predictors can be present.

Advantages of discussing simple regression:

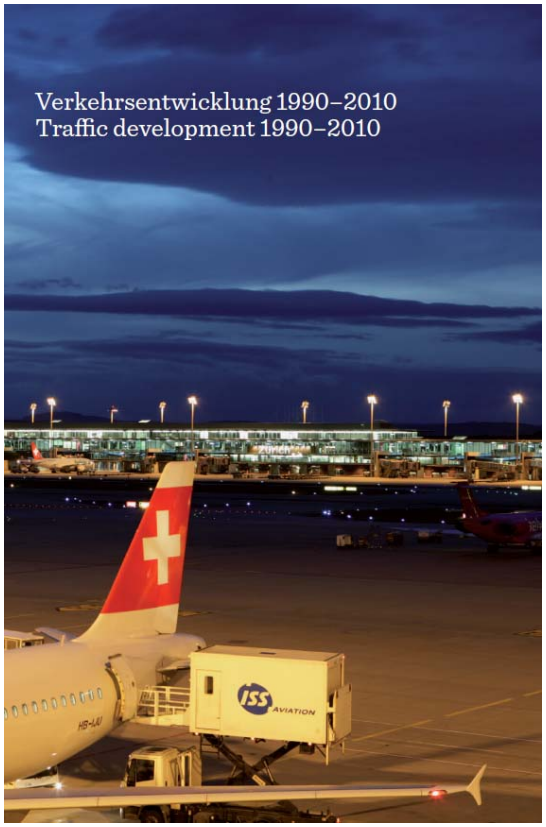
- **Visualization of data and fit is possible**
- **Corresponds to estimating a straight line or curve**
- **Is also mathematically simpler and more intuitive**

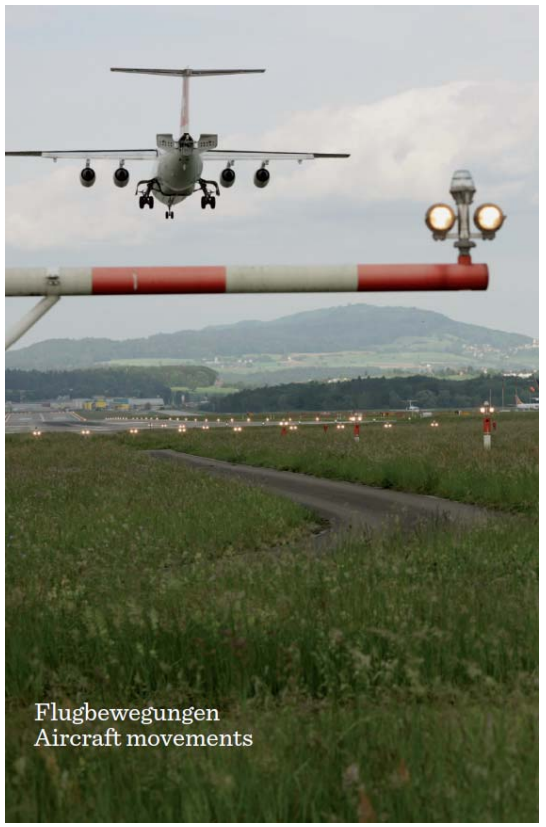
We start out with *smoothing*, i.e. fitting non-parametric curves. Then, we will proceed with discussing linear models, i.e. the classical parametric regression approach.

Applied Statistical Regression
AS 2013 – Week 02

Example: Airline Passengers

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

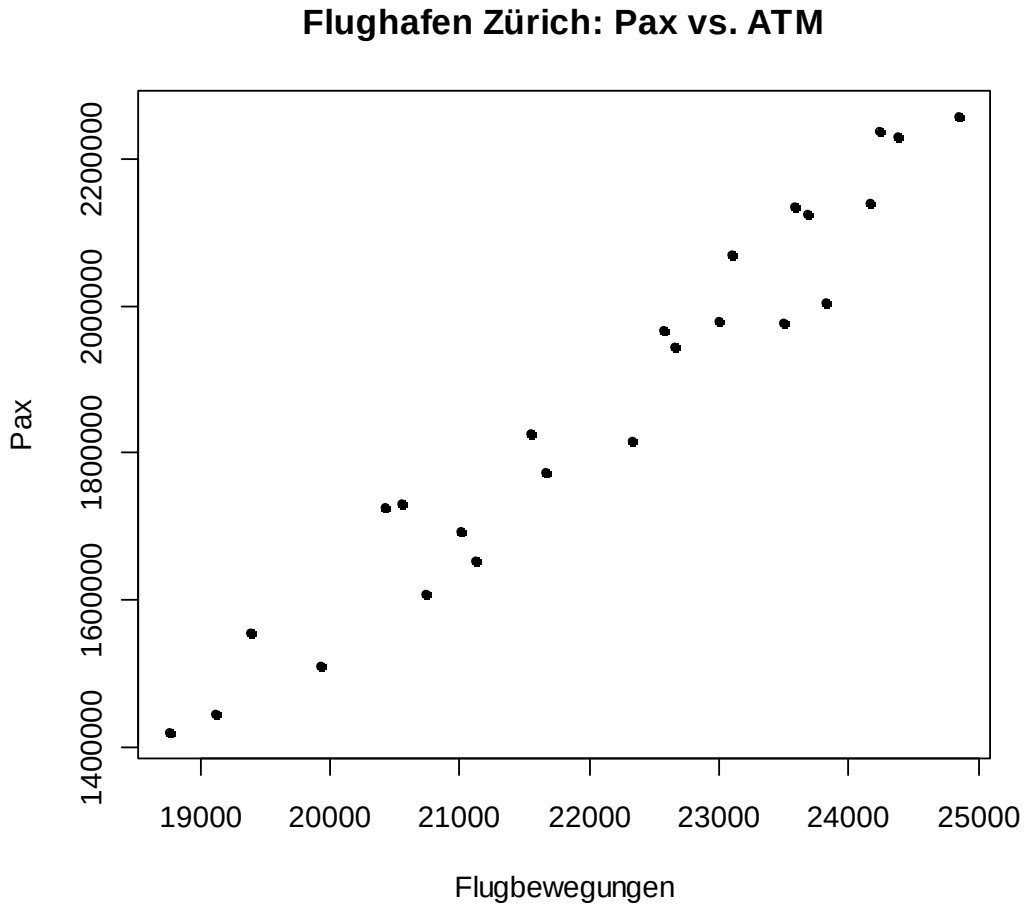




Applied Statistical Regression
AS 2013 – Week 02

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
…	…	…

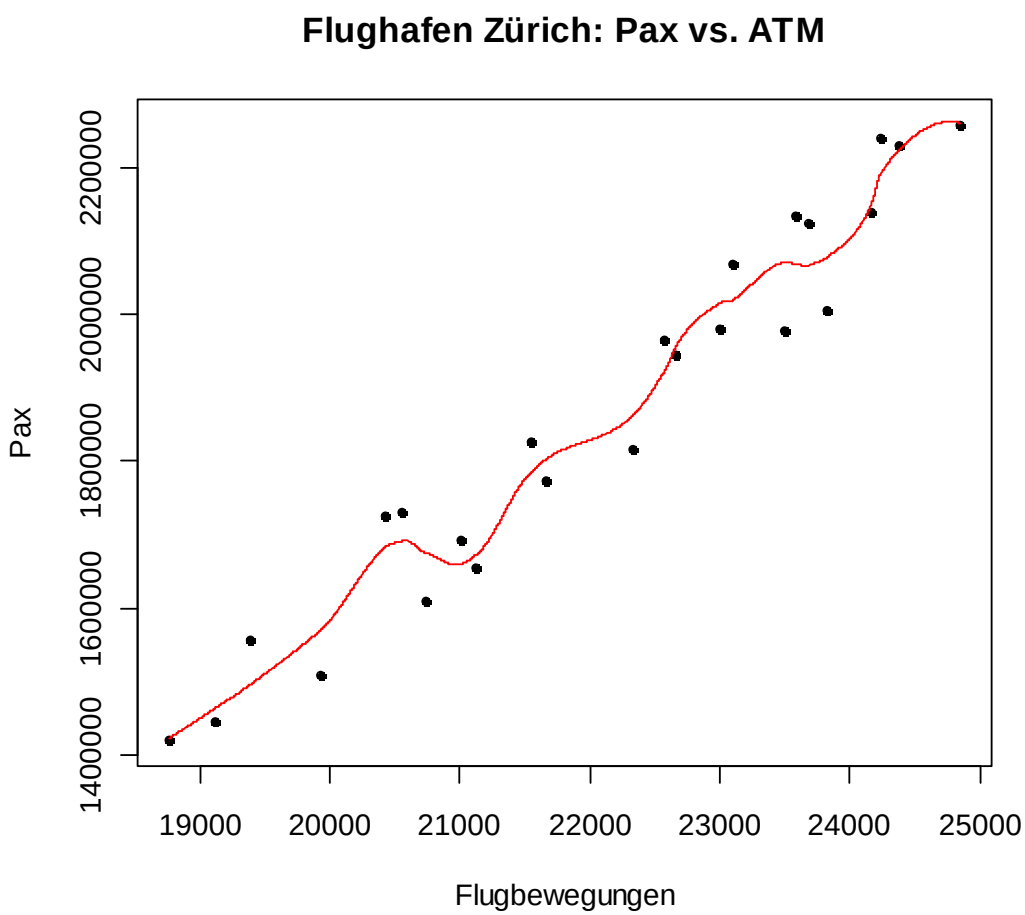


Applied Statistical Regression
AS 2013 – Week 02

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.

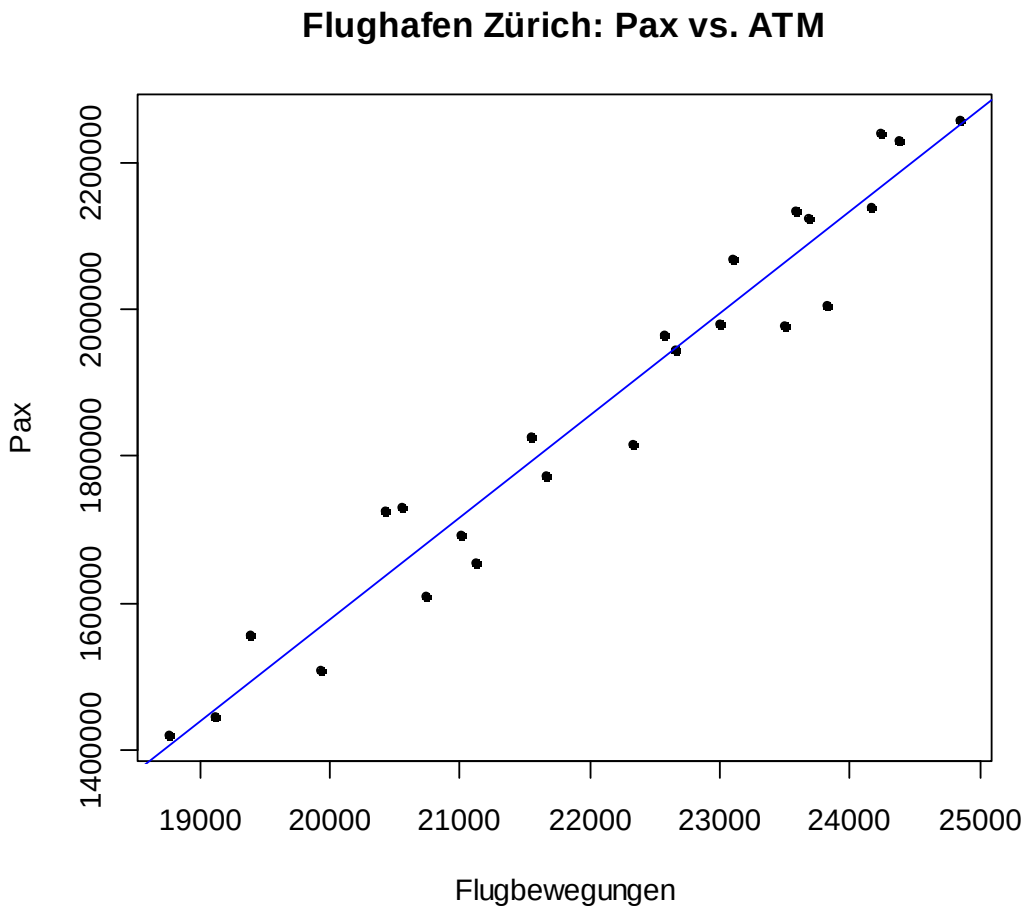


Applied Statistical Regression
AS 2013 – Week 02

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

Applied Statistical Regression
AS 2013 – Week 02

Smoothing

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

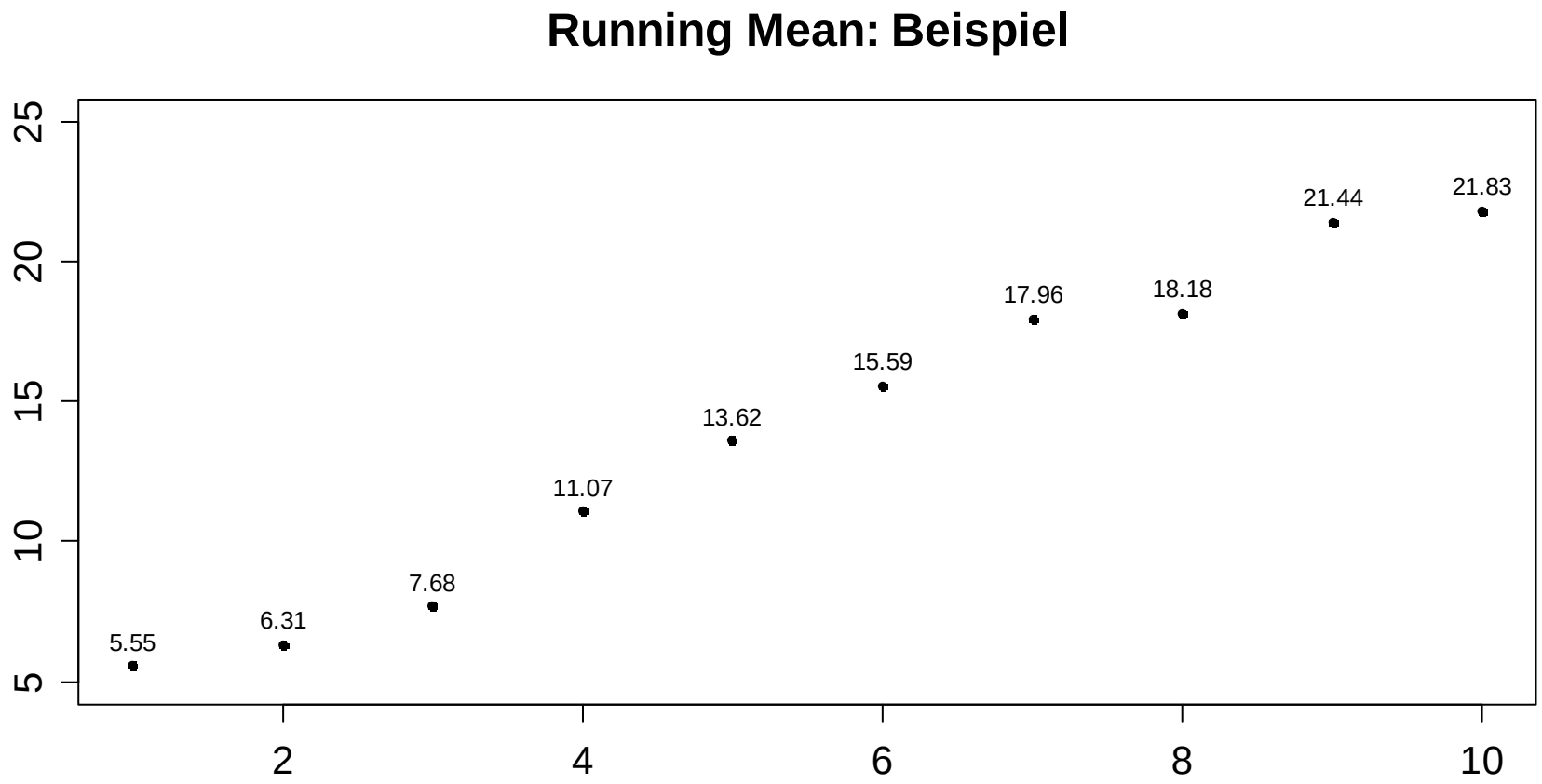
→ we are not after a functional description of $f(\cdot)$

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

Applied Statistical Regression
AS 2013 – Week 02

Running Mean: Example



Applied Statistical Regression
AS 2013 – Week 02

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:

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

The *weights* are defined as and λ is the *window width*.

$$w_i = \begin{cases} 1 & if \ |x - x_j| \leq \lambda/2 \\ 0 & else \end{cases} \quad ,$$

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

- Alternatively, one can simply use function `ksmooth()`. The window size can be adjusted by argument `bandwidth=`. Some other settings can be made, especially with respect to evaluation.

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

Applied Statistical Regression
AS 2013 – Week 02

Running Mean: R-Implementation

Kernel Regression Smoother

Description

The Nadaraya–Watson kernel regression estimate.

Usage

```
ksmooth(x, y, kernel = c("box", "normal"), bandwidth = 0.5,
        range.x = range(x),
        n.points = max(100, length(x)), x.points)
```

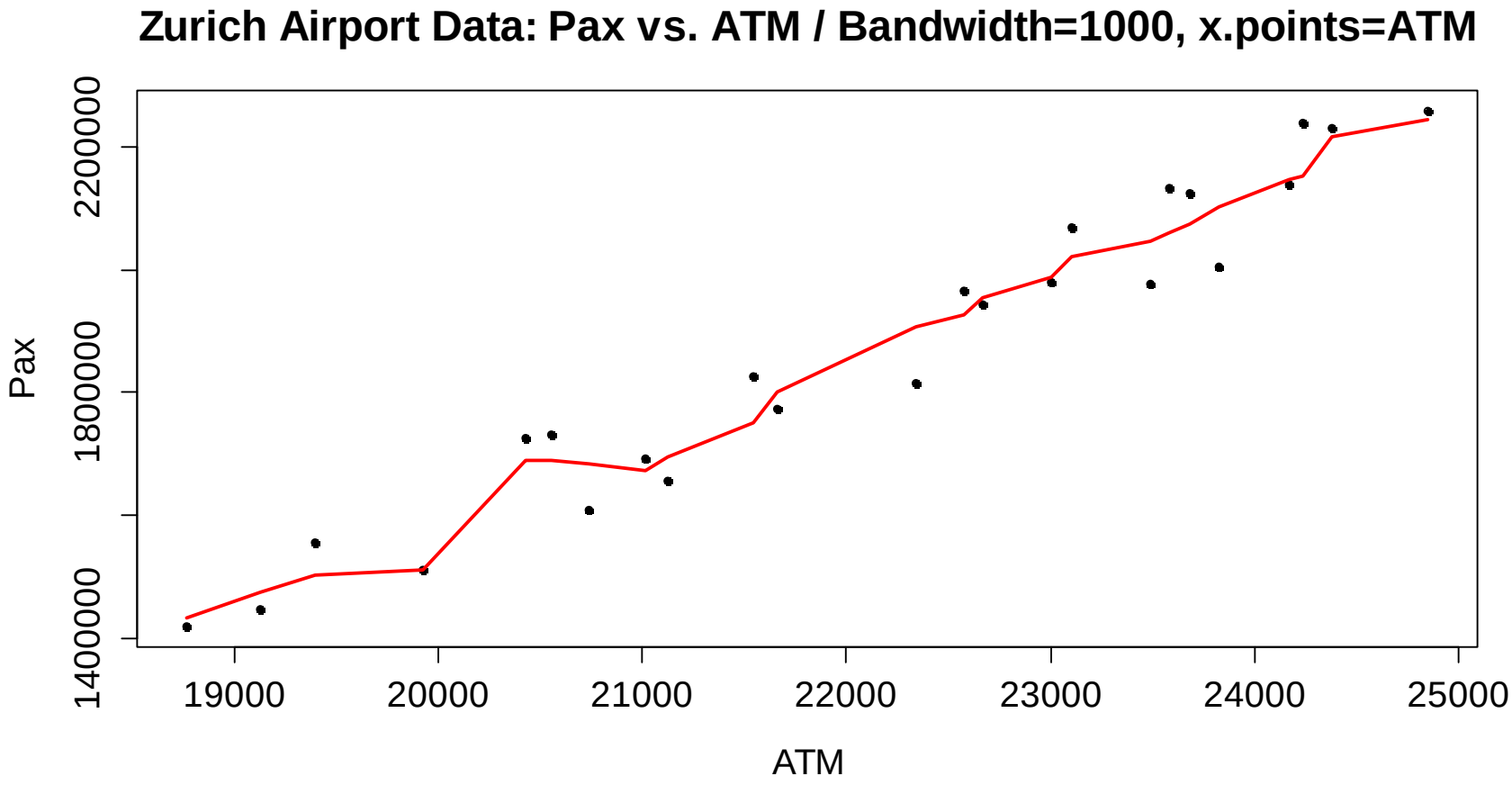
Arguments

`x`	input x values
`y`	input y values
`kernel`	the kernel to be used.
`bandwidth`	the bandwidth. The kernels are scaled so that their quartiles (viewed as probability densities) are at +/- `0.25*bandwidth`.
`range.x`	the range of points to be covered in the output.
`n.points`	the number of points at which to evaluate the fit.
`x.points`	points at which to evaluate the smoothed fit. If missing, `n.points` are chosen uniformly to cover `range.x`.

Applied Statistical Regression
AS 2013 – Week 02

Running Mean: Unique Data

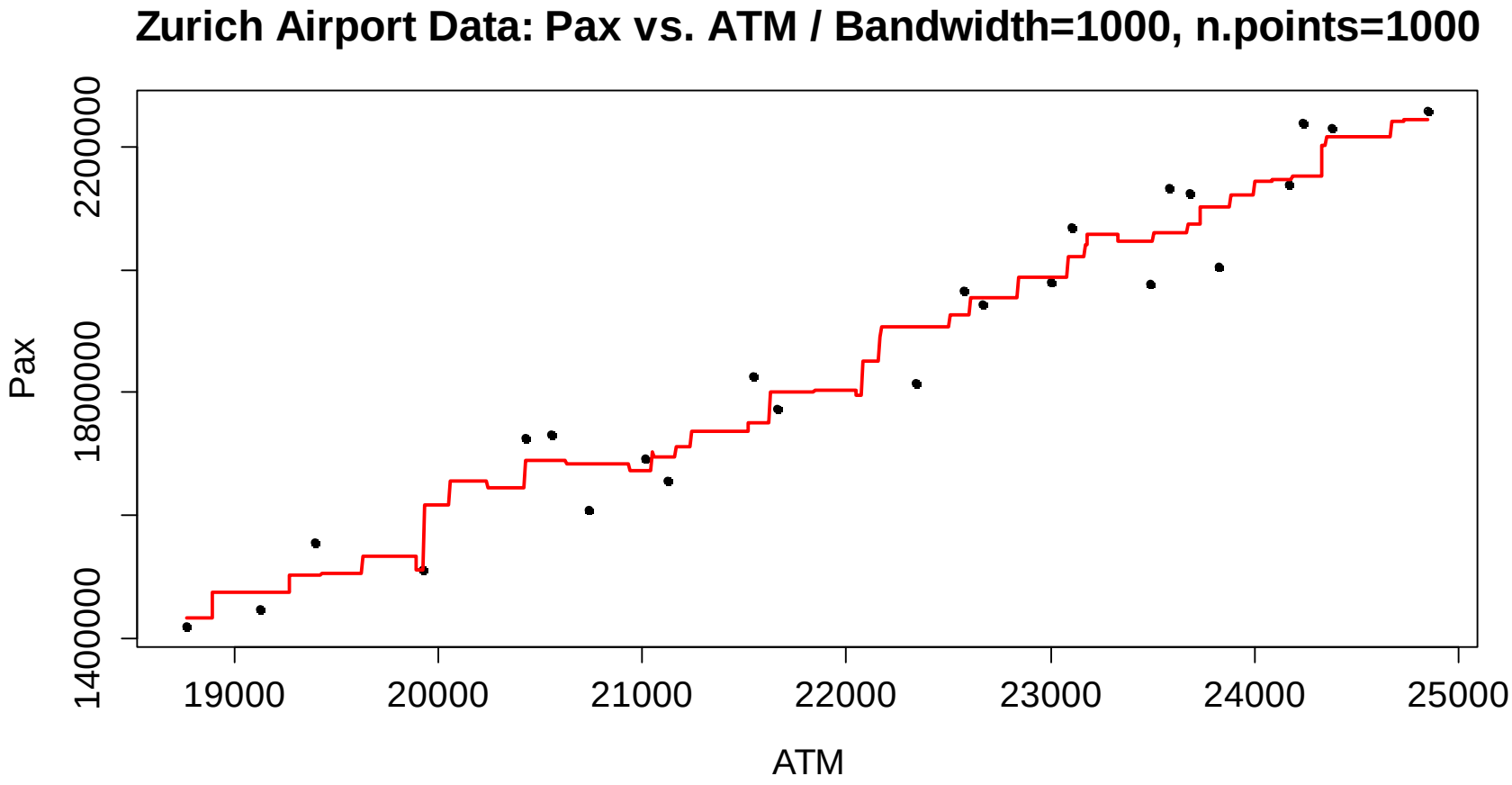
```
> plot(Pax ~ ATM, data=unique2010, main="...")
> lines(fit, col="red", lwd=2)
```



Applied Statistical Regression
AS 2013 – Week 02

Running Mean: Unique Data

```
> plot(Pax ~ ATM, data=unique2010, main="...")
> lines(fit, col="red", lwd=2)
```



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*

Applied Statistical Regression
AS 2013 – Week 02

Gaussian Kernel Smoother

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

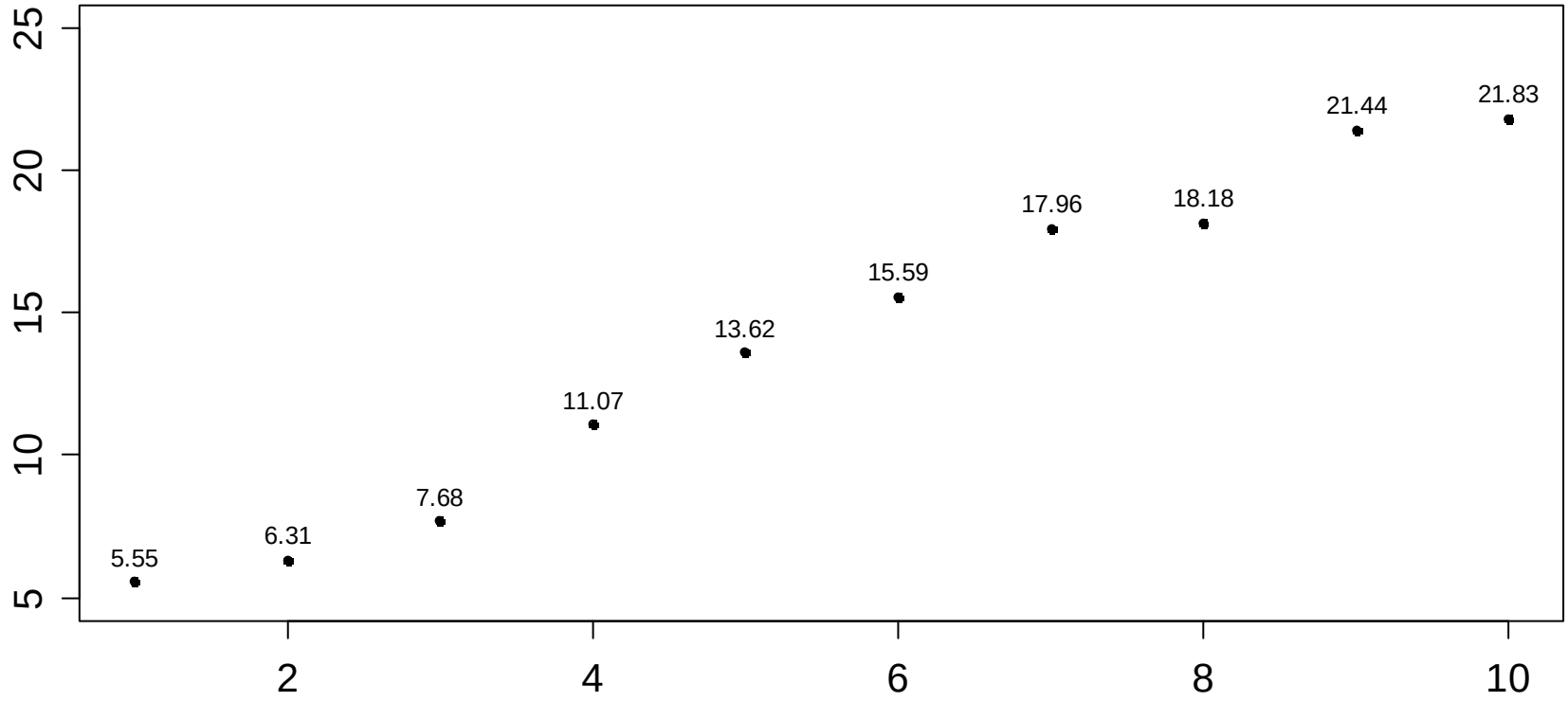
The estimate for $f(\cdot)$, denoted as $\hat{f}_\lambda(\cdot)$, is defined as:

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

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.

Applied Statistical Regression
AS 2013 – Week 02

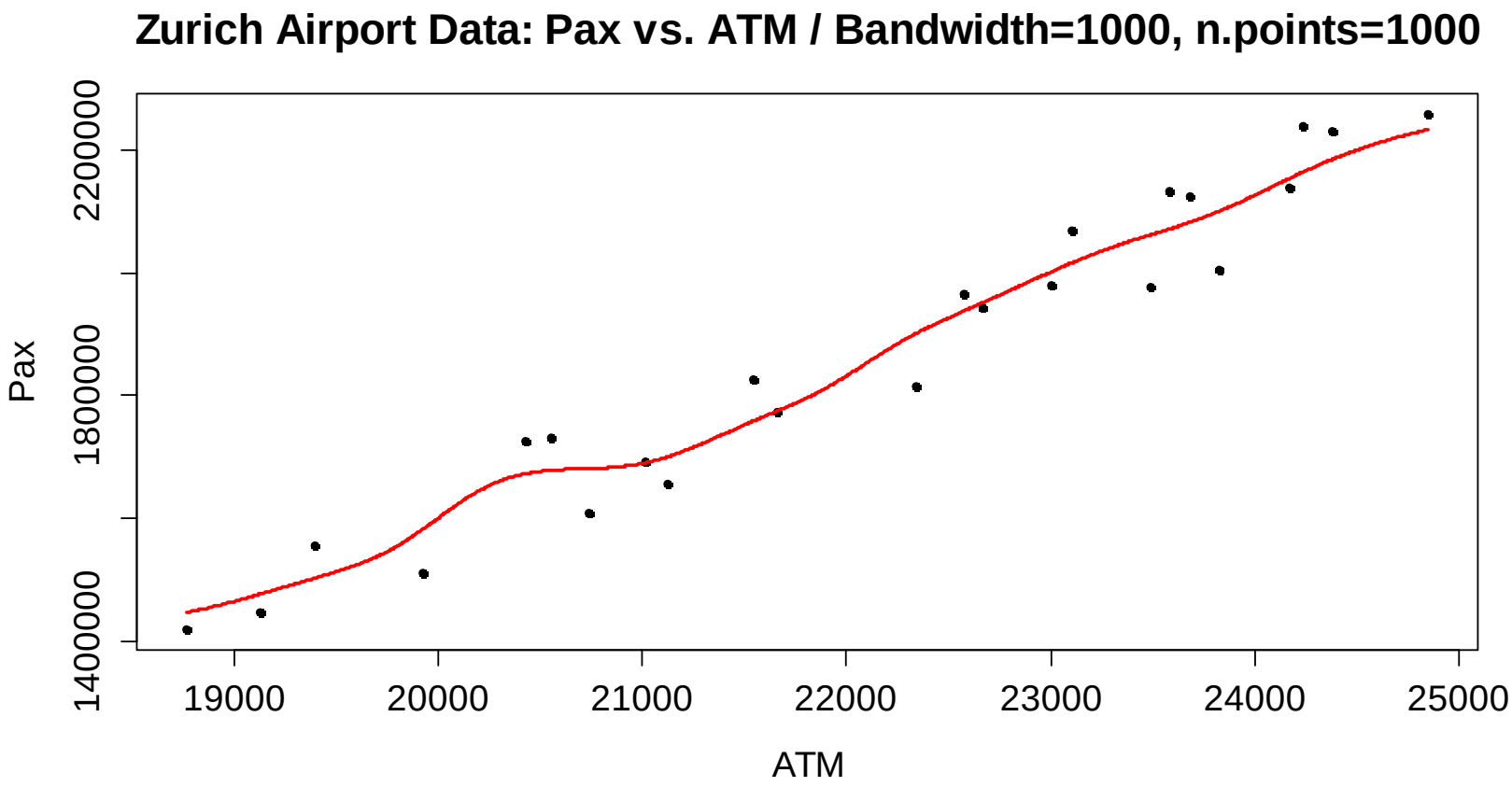
Gaussian Kernel Smoother: Idea



Applied Statistical Regression
AS 2013 – Week 02

Gaussian Kernel Smoother: Unique Data

```
> ks.gauss <- ksmooth(ATM, Pax, kernel="normal", band=1000)
> plot(ATM, Pax, xlab="ATM", ylab="Pax", pch=20)
> lines(ks.gauss, col="darkgreen", lwd=1.5)
```



LOESS-Smoother

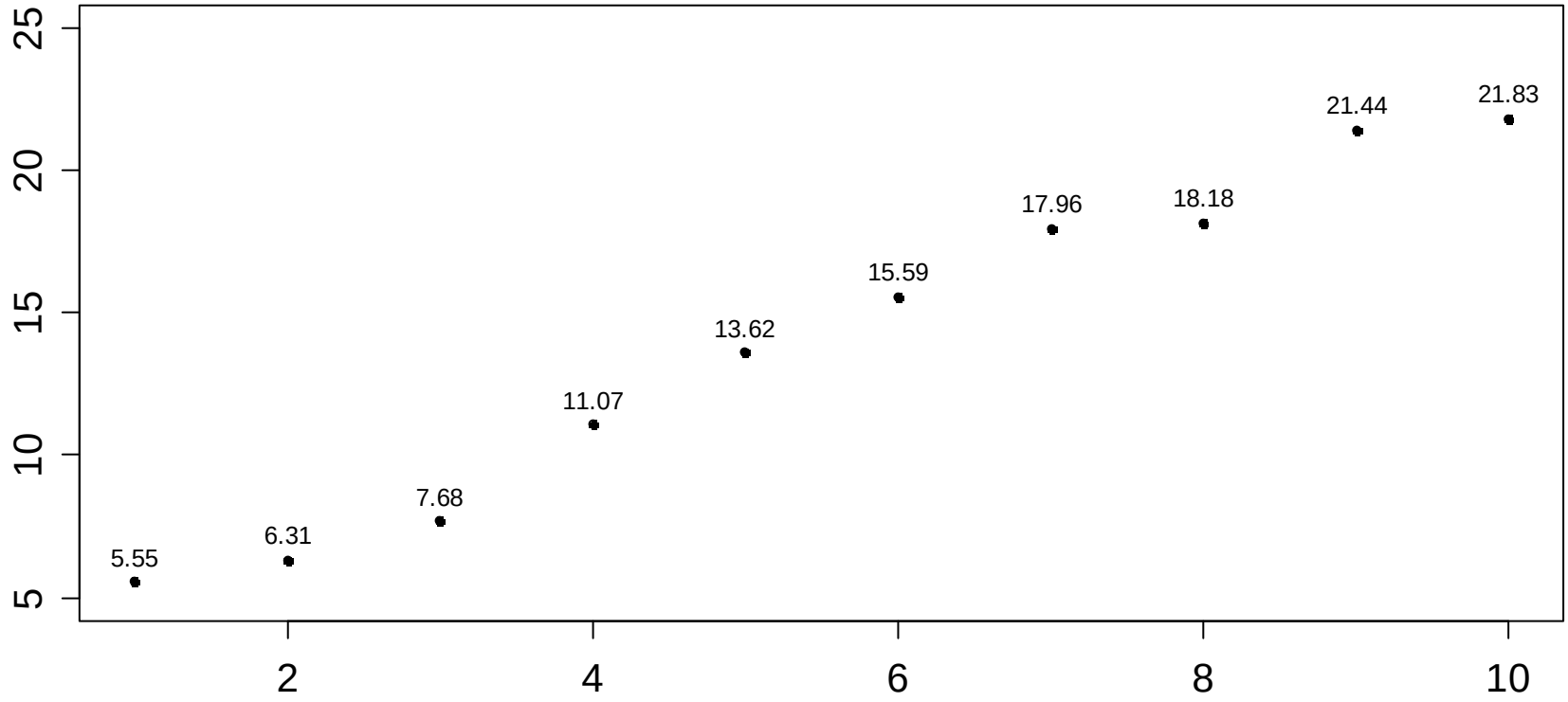
The LOESS-Smoother is better, more flexible and more robust than the Gaussian Kernel Smoother. It should be prefered!

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*

Applied Statistical Regression
AS 2013 – Week 02

LOESS-Smoother: Idea



Applied Statistical Regression
AS 2013 – Week 02

LOESS-Smoother: R-Implementation

Scatter Plot with Smooth Curve Fitted by Loess

Description

Plot and add a smooth curve computed by `loess` to a scatter plot.

```
loess.smooth(x, y, span = 2/3, degree = 1,
    family = c("symmetric", "gaussian"), evaluation = 50, ...)
```

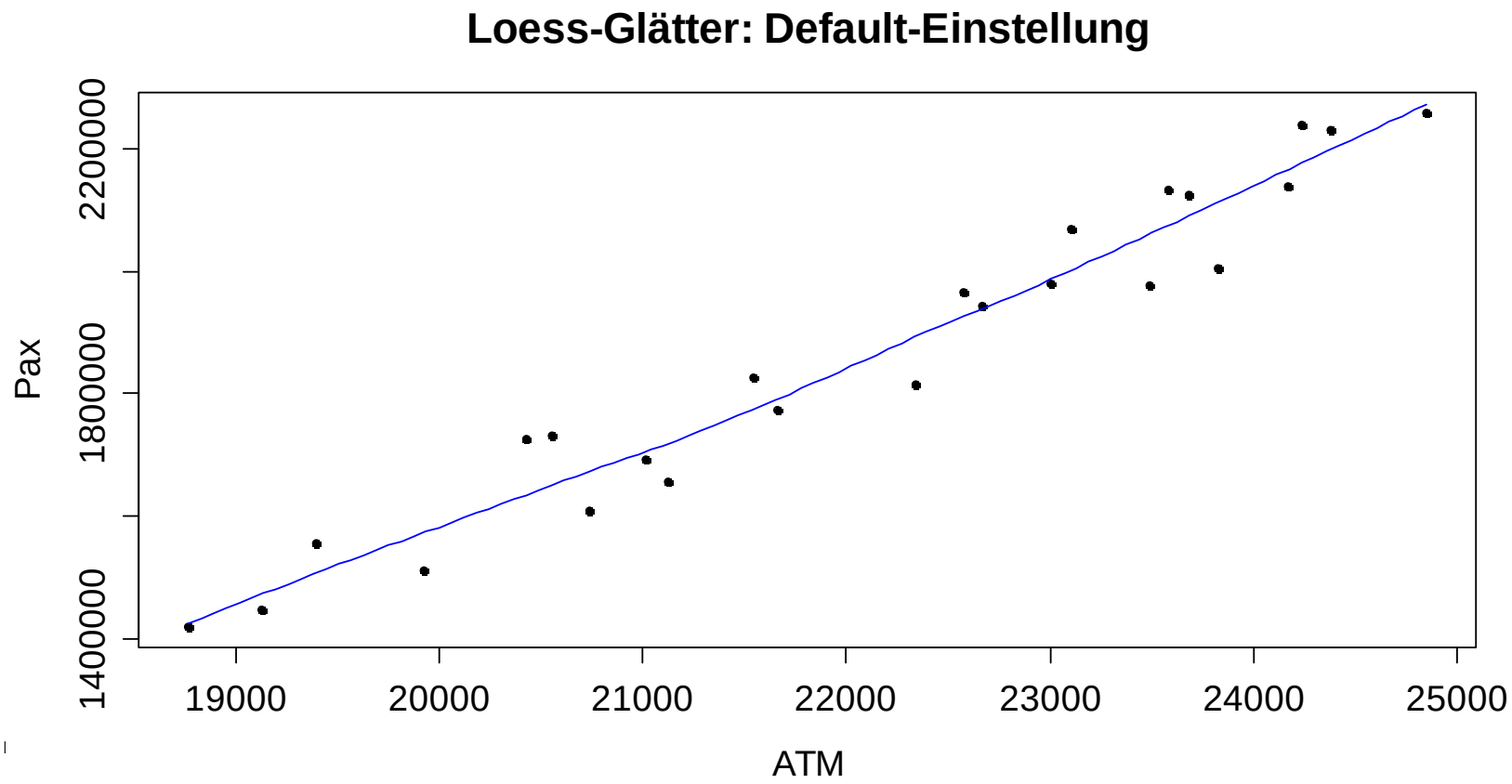
Arguments

`x,y`	the `x` and `y` arguments provide the x and y coordinates for the plot. Any reasonable way of defining the coordinates is acceptable. See the function `xy.coords` for details.
`span`	smoothness parameter for `loess`.
`degree`	degree of local polynomial used.
`family`	if `"gaussian"` fitting is by least-squares, and if `family="symmetric"` a re-descending M estimator is used.
`xlab`	label for x axis.
`ylab`	label for y axis.
`ylim`	the y limits of the plot.
`evaluation`	number of points at which to evaluate the smooth curve.

Applied Statistical Regression
AS 2013 – Week 02

LOESS-Smoother: Unique Data

```
> smoo  <- loess.smooth(unique2010$ATM, unique2010$Pax)
> plot(Pax ~ ATM, data=unique2010, main=...)
> lines(smoo, col="blue")
```

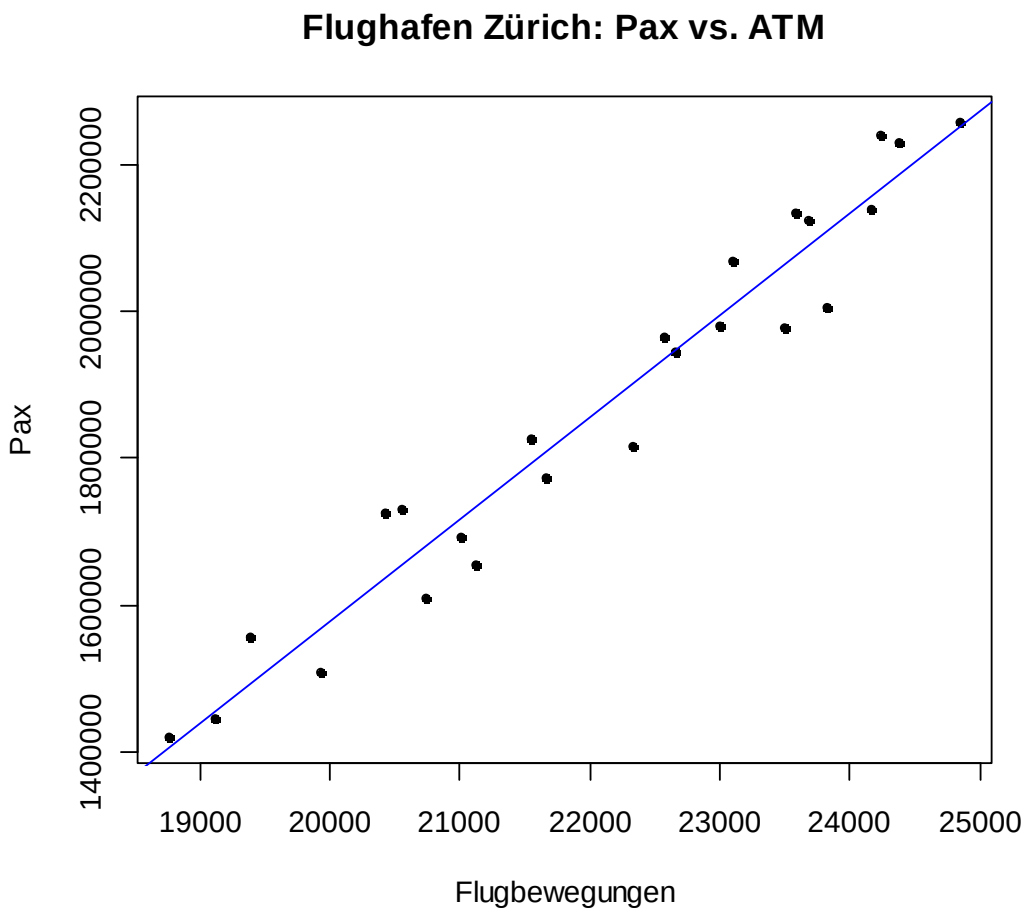


Applied Statistical Regression
AS 2013 – Week 04

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.



Applied Statistical Regression
AS 2013 – Week 04

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:

β_0 = "Intercept"

β_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.

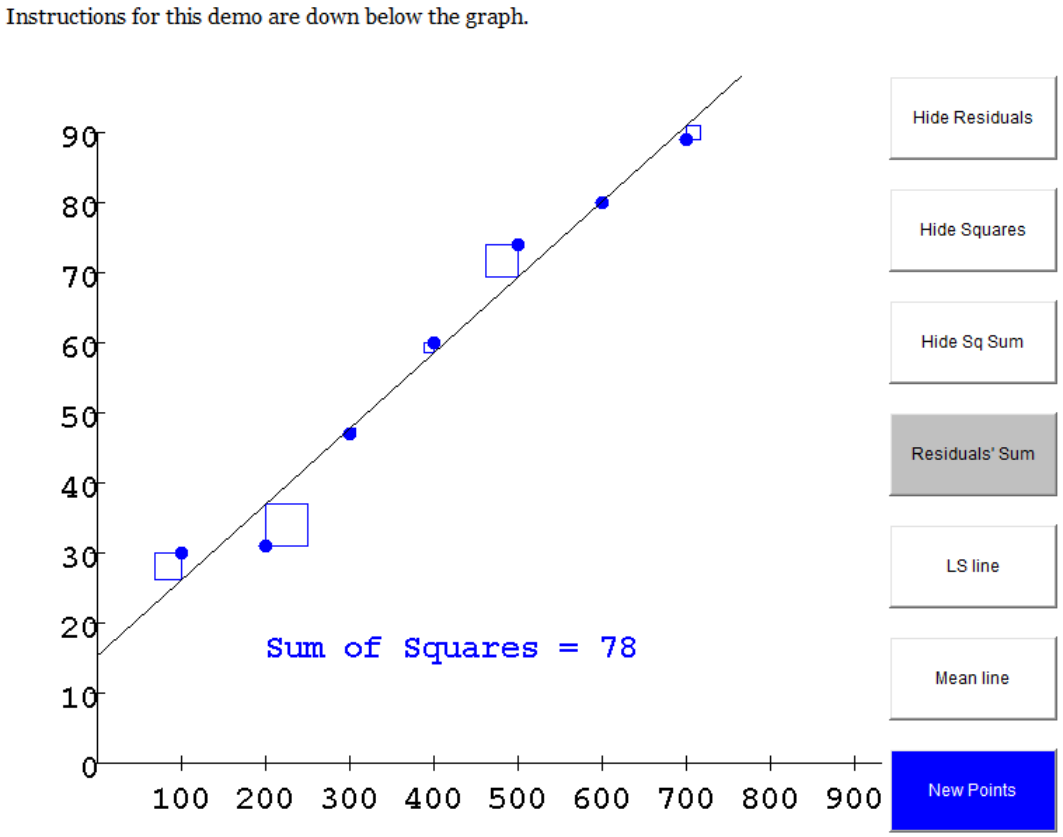
β_0, β_1 are the regression coefficients. They are unknown previously, and need to be estimated from the data.

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

Applied Statistical Regression
AS 2013 – Week 04

Least Squares Fitting

→ http://sambaker.com/courses/J716/demos/LeastSquares/LeastSquaresDemo.html



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 $(x_i, y_i)_{i=1,...,n}$, 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 this "quality 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:

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

- 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:

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

- 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:**
  $\hat{\beta}_1 = 138.8,\ \hat{\beta}_0 = -1'197'682$ obtained from

```
> lm(Pax ~ ATM, data=uni...)
```

Fitted Values

The estimated parameters $\hat{\beta}_0, \hat{\beta}_1$ can be used for determining the fitted values $\hat{y}$. Please note that mathematically, this is a conditional expected value::

$$\hat{y} = E[y \mid x] = \hat{\beta}_0 + \hat{\beta}_1 x$$

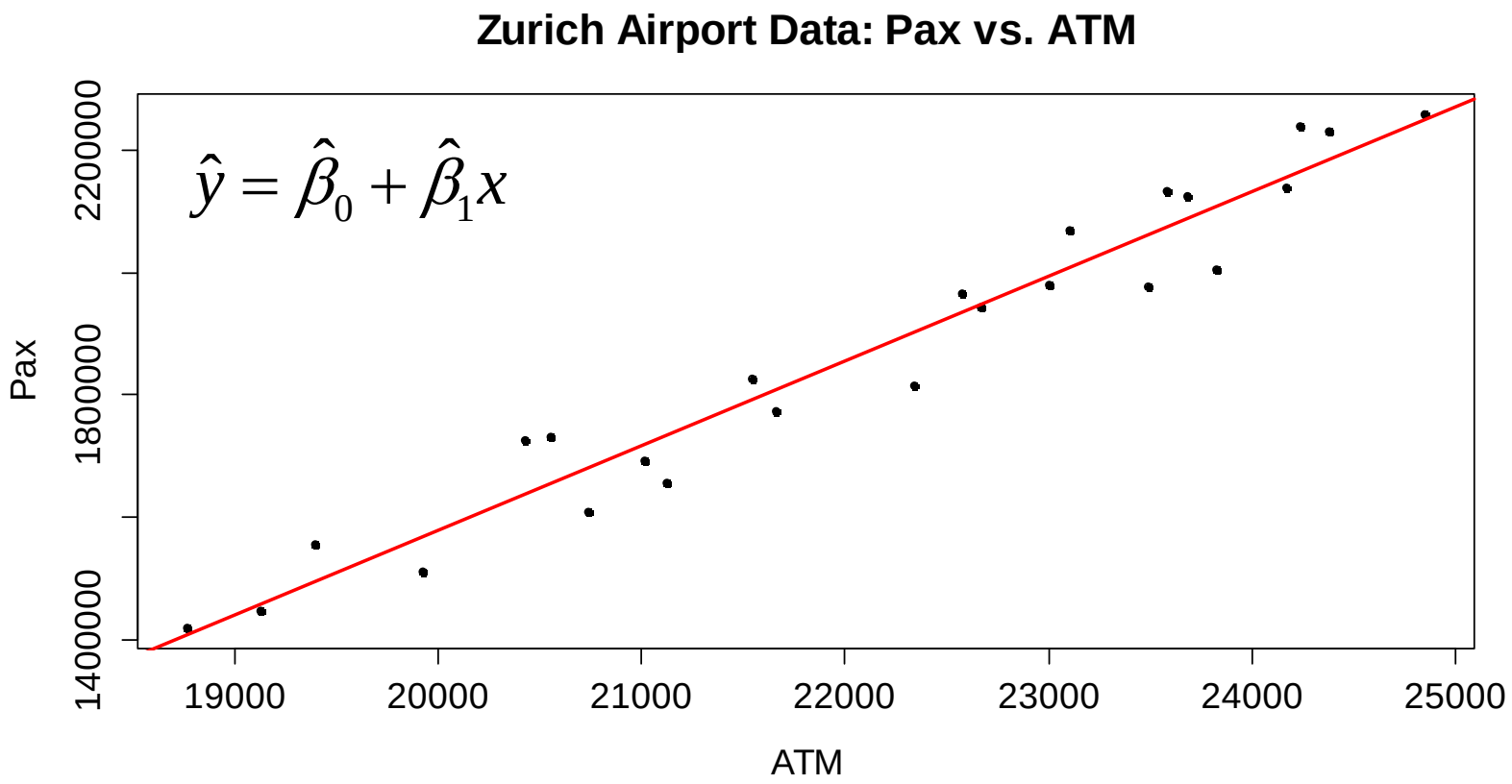
In R, the fitted values are obtained as follows:

```
> fit <- lm(Pax ~ ATM, data=unique2010)
> fitted(fit)
      1       2       3       4       5
1654841 1808312 2165068 2156465 2184911
      6       7       8       9     ...
2250545 2108731 2062107 1493184     ...
```

Applied Statistical Regression
AS 2013 – Week 04

Drawing the Regression Line

```
> plot(Pax ~ ATM, data=unique2010, pch=20)
> title("Zurich Airport Data: Pax vs. ATM")
> abline(fit, col="red", lwd=2)
```



Applied Statistical Regression
AS 2013 – Week 04

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:

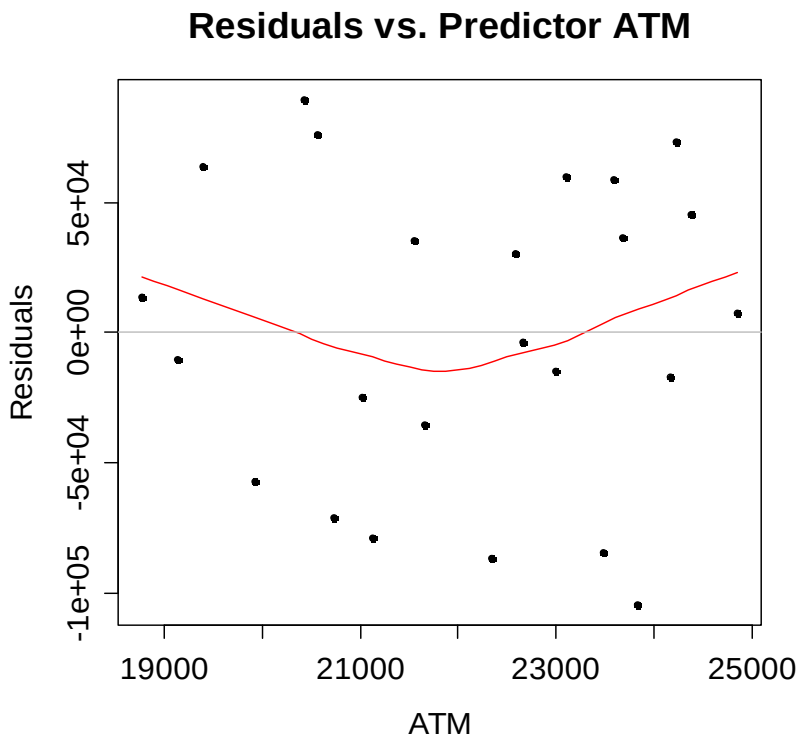
- the relation is in truth a straight line, i.e. $E[E_i] = 0$
- the scatter of the errors is constant, i.e. $Var(E_i) = \sigma^2$
- the data are uncorrelated (from a representative sample)
- the errors are approximately normally distributed

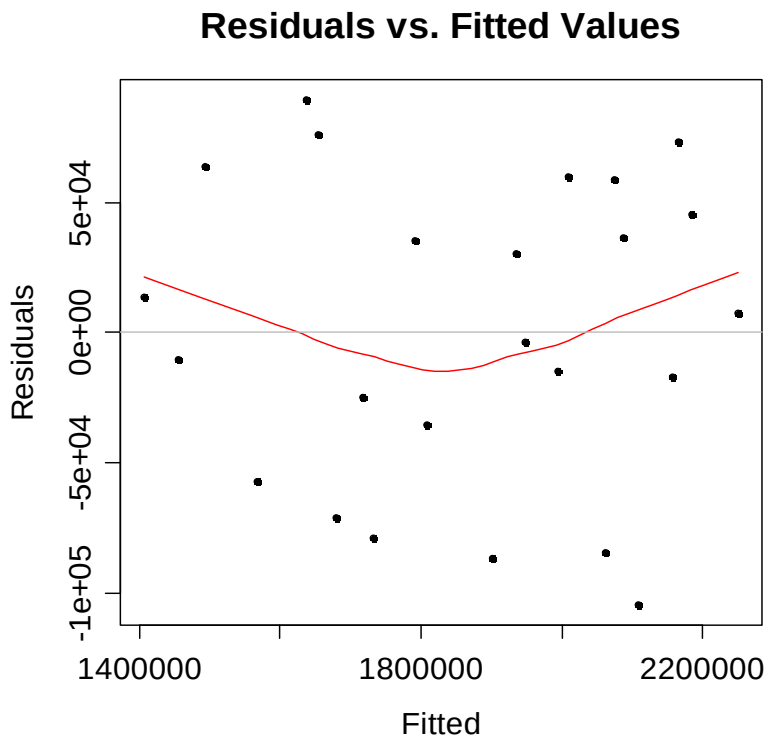
→ **Fodder for thougt: 9/11, SARS, Eyjafjallajökull...?**

Applied Statistical Regression
AS 2013 – Week 04

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:

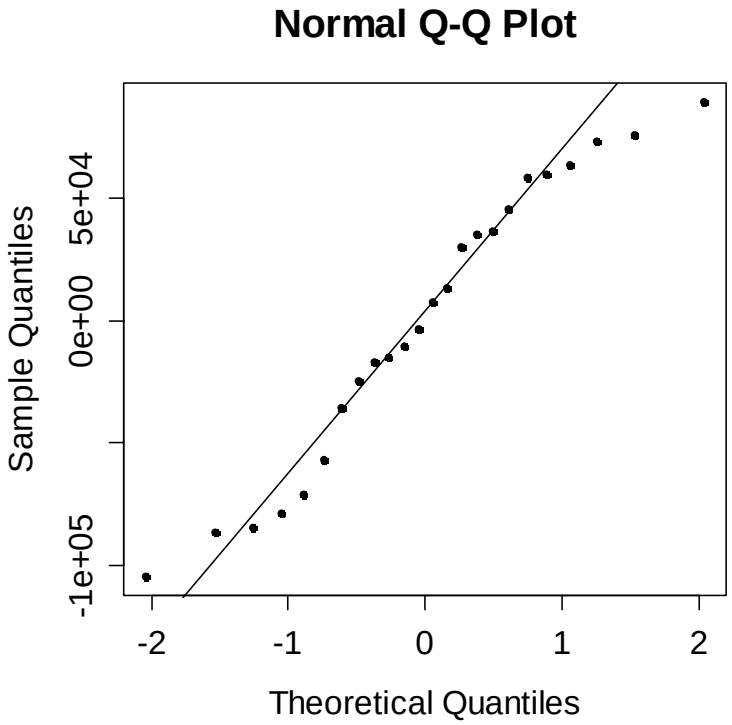




Applied Statistical Regression
AS 2013 – Week 04

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.

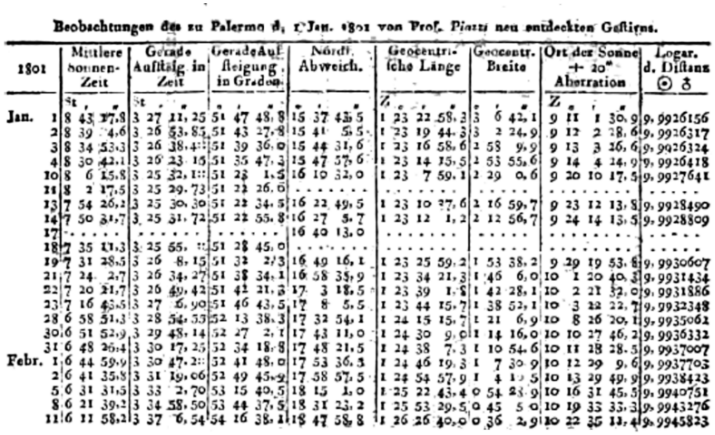
Applied Statistical Regression
AS 2013 – Week 04

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 ' "	° ' "	° ' "	♈ ° ' "	° ' "	♈ ° ' "	
Jan.	1	8 43 37,8	3 27 21, 25	51 47 48, 8	15 37 42, 5	1 23 22 58, 2	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, 2	3 2 24, 9	9 12 2 28, 6	9, 9926317
	3	8 34 32,3	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 0,1	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 13,3	3 25 55, 1	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 45 6, 0	10 1 20 40, 3	9, 9931434
	22	7 20 21,7	3 26 49, 42	51 41 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 45,6	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, 53	52 13 38, 3	17 32 54, 1	1 24 15 15, 7	1 21 6, 9	10 8 26 20, 1	9, 9935061
	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 36, 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, 2	1 24 46 19, 3	1 7 30, 9	10 12 29 9, 6	9, 9937703
	2	6 41 35,8	3 31 19, 66	52 49 45, 9	17 58 57, 5	1 24 54 57, 9	1 4 1, 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 12, 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 $(\bar{x}, \bar{y})$
- 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.

Applied Statistical Regression
AS 2013 – Week 04

Gauss-Markov-Theorem

A 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:

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

The variances of the estimates are as follows:

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

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
- (an error distribution which is approximately Gaussian)

Estimating the Error Variance

Besides the regression coefficients, we also need to estimate the *error variance*. It is a necessary ingredient for all tests and confidence intervals that will be discussed shortly.

The estimate is based on the *residual sum of squares* (RSS):

$$\hat{\sigma}_E^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

In R, the regression summary provides the estimate for the error's standard deviation as `Residual standard error`:

```
> summary(fit)
...
Residual standard error: 59700 on 22 degrees of freedom
```

Benefits of Linear Regression

- **Inference on the relation between y and x**

The goal is to understand if and how strongly the response variable depends on the predictor. There are performance indicators as well as statistical tests adressing the issue.

- **Prediction of (future) observations**

The regression line/equation can be employed to predict the PAX number for any given ATM value.

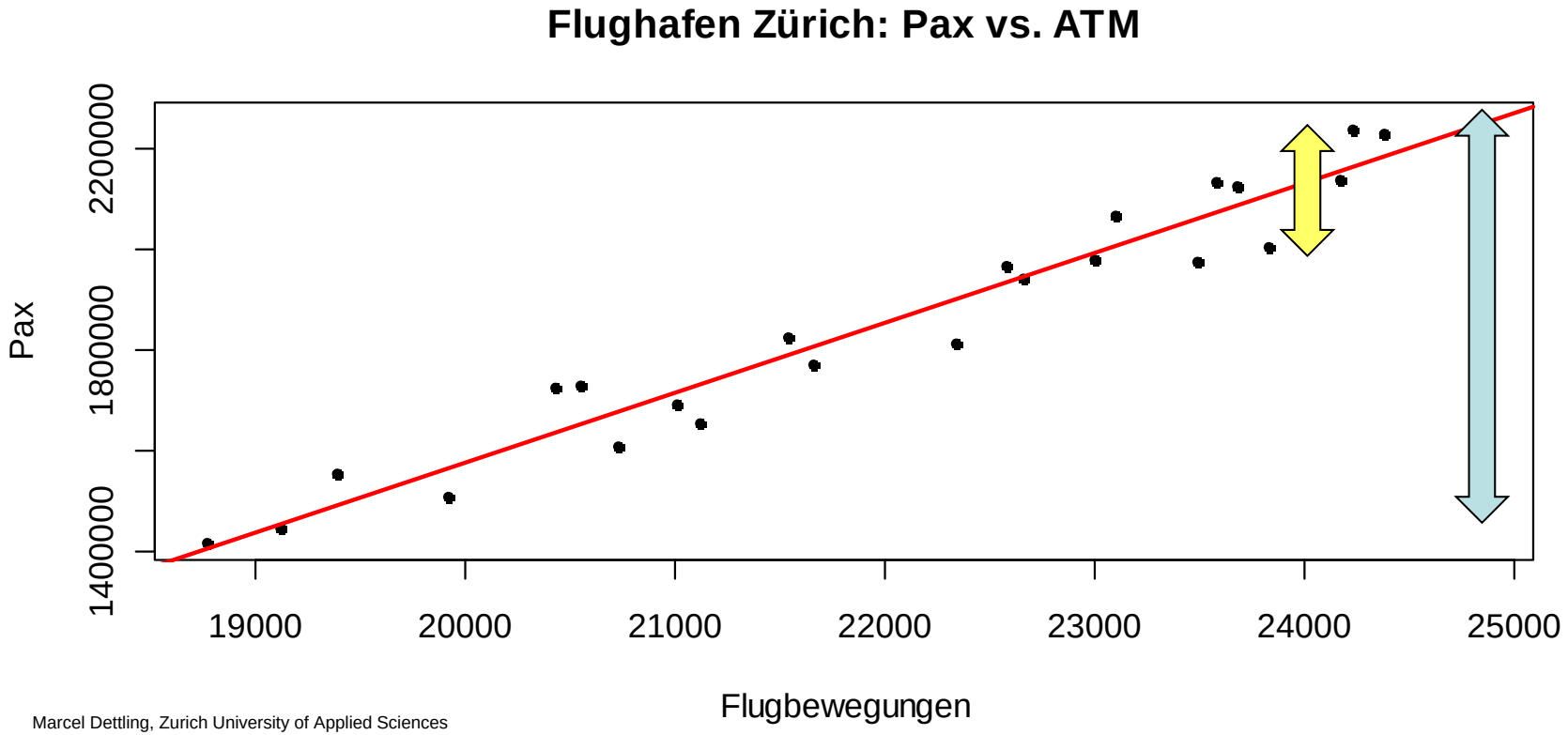
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

However, this mostly will not work well for extrapolation!

Applied Statistical Regression
AS 2013 – Week 04

R^2: *The Coefficient of Determination*

The coefficient of determination R^2 is also known as *multiple R-squared*. It tells which portion of the total variation is accounted for by the regression line.



Applied Statistical Regression
AS 2013 – Week 04

Computation of R^2

R^2 is the portion of the total variation that is explained through regression. It is determined as one minus the quotient of the yellow arrow divided by the blue arrow.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \in [0,1]$$

The closer to 1 the value is, the tighter the datapoints are packed around the regression line. However, there are no formal criteria which R^2 value needs to be met such that the regression can be said to be useful/valid.

Confidence Interval for the Slope β_1

A 95%-CI for the slope β_1 tells which values (besides the point estimate $\hat{\beta}_1$) are plausible, too. The uncertainty is due to estimation/sampling effects.

95%-CI for β_1 : $\hat{\beta}_1 \pm qt_{0.975;n-2} \cdot \hat{\sigma}_{\hat{\beta}_1}$, resp.

$$\hat{\beta}_1 \pm qt_{0.975;n-2} \cdot \sqrt{\hat{\sigma}_E^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

In R:

```
> fit <- lm(Pax ~ ATM, data=unique2010)
> confint(fit, "ATM")
        2.5 %  97.5 %
ATM 124.4983 153.025
```

Testing the Slope β_1

There is a statistical hypothesis test which can be used to check whether the slope is significantly different from zero, or any other arbitrary value b . The null hypothesis is:

$$H_0 : \beta_1 = 0, \text{ resp. } H_0 : \beta_1 = b$$

One usually tests two-sided on the 95%-level. The alternative is:

$$H_A : \beta_1 \neq 0, \text{ resp. } H_A : \beta_1 \neq b$$

As a test statistic, we use:

$$T_{H_0:\beta_1=0} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}, \text{ resp. } T_{H_0:\beta_1=b} = \frac{\hat{\beta}_1 - b}{\hat{\sigma}_{\hat{\beta}_1}}, \text{ both have a } t_{n-2} \text{ distribution.}$$

Applied Statistical Regression
AS 2013 – Week 04

Reading R-Output

```
> summary(lm(Pax ~ ATM, data=dat))

Call: lm(formula = Pax ~ ATM, data = dat)

Residuals:    Min      1Q  Median      3Q     Max
          -104188  -40885    2099   48588   89154

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.198e+06  1.524e+05  -7.858 7.94e-08 ***
ATM          1.388e+02  6.878e+00  20.176 1.11e-15 ***

---

Residual standard error: 59700 on 22 degrees of freedom
Multiple R-squared: 0.9487, Adjusted R-squared: 0.9464
F-statistic: 407.1 on 1 and 22 DF,  p-value: 1.110e-15
```

→ **Will be explained in detail on the blackboard!**

Testing the Slope β_1

Practical Example:

Use the Pax vs. ATM data and perform a statistical test for the null hypothesis $H_0 : \beta_1 = 150$. The information from the summary on slide 25 can be used as a basis. Then, also answer:

a) *Explain in colloquial language what was just tested. What is the benefit of this test? What claims could motivate the test?*

b) *How does the testing result relate with the 95%-CI that we computed on slide 23? Would we be able to tell the test results from the CI alone?*

→ **See blackboard for the answers**

Testing the Intercept β_0

An analogous test can be done for the intercept.

- No matter what the test result will be, the intercept should generally not be omitted from the regression model.
- The presence of the intercept protects against possible non-linearities and calibration errors of measurement devices. If it is kicked out of the model, the results are generally worse.
- If theory dictates that there should not be an intercept but it is still significant, take this as evidence that the linear relation does not hold when extrapolating to $x = 0$.

Applied Statistical Regression
AS 2013 – Week 04

Prediction

Using the regression line, we can predict the y-value for any desired x-value. The result is the expectation for y given x.

$$E[y \mid x] = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{a.k.a. } \textbf{\textit{"fitted value"}}$$

Example: With 24’000 air traffic movements, we expect

$$-1'197'682 + 24'000 \cdot 138.8 = 2'133'518 \text{ Passengers}$$

Be careful:

At best, interpolation within the range of observed x-values is trustworthy. Extrapolation with ATM values such as 50’000, 5'000 or even 0 usually produces completely useless results.

Applied Statistical Regression
AS 2013 – Week 04

Prediction with R

We can use the regression fit object for prediction. The syntax for obtaining the fitted value(s) is as follows:

```
> fit <- lm(Pax ~ ATM, data=unique2010)
> dat <- data.frame(ATM=c(24000))
> predict(fit, newdata=dat)
1 2132598
```

The x-values need to be provided in a data frame, where the variable/column name is identical to the predictor name.

Then, the `predict()` procedure is invoked with the regression fit and the new x-values as arguments.

Applied Statistical Regression
AS 2013 – Week 04

Confidence Interval for $E[y \mid x]$

We just computed the fitted value $\hat{\beta}_0 + \hat{\beta}_1 x$, i.e. the expected number of passengers for 24'000 ATMs. This is not a deterministic value, but an estimate that is subject to variability.

A 95%-CI for the fitted value at position x is given by:

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm qt_{0.975;n-2} \cdot \hat{\sigma}_E \cdot \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

In R:
```
> predict(fit, newdata=dat, interval="confidence")
        fit     lwr     upr
1 2132598 2095450 2169746
```

Applied Statistical Regression
AS 2013 – Week 04

Prediction Interval for y

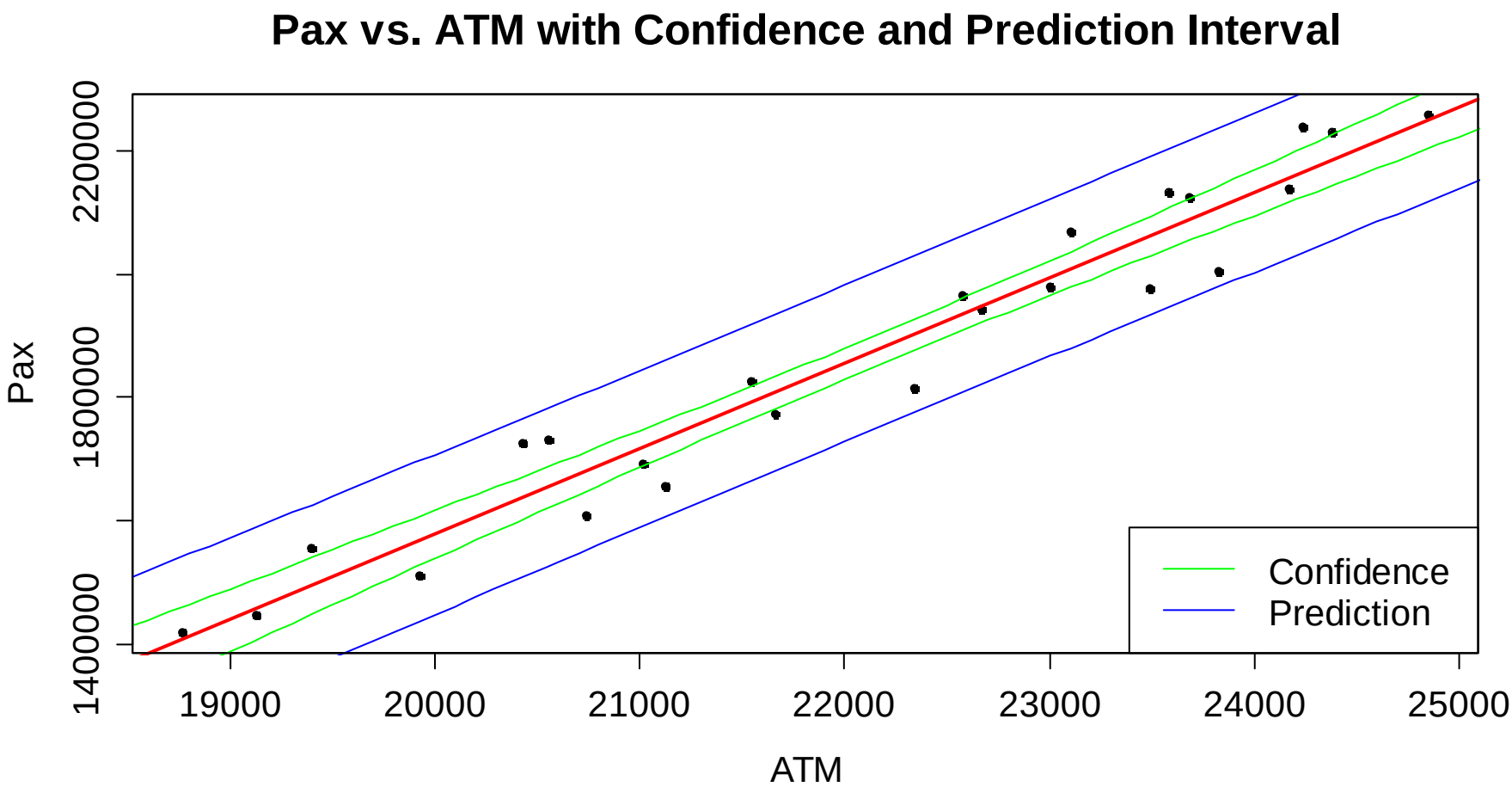
The confidence interval for $E[y \mid x]$ tells about the variabilty of the fitted value. It does not account for the scatter of the data points around the regression line and thus does not define a region where we have to expect the observed value. A 95% prediction interval at position x is given by:

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm qt_{0.975;n-2} \cdot \hat{\sigma}_E \cdot \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

In R:
```
> predict(fit,newdata=dat,interval="prediction")
        fit     lwr     upr
1 2132598 2003343 2261853
```

Applied Statistical Regression
AS 2013 – Week 04

Confidence and Prediction Interval



Applied Statistical Regression
AS 2013 – Week 04

Confidence and Prediction Interval

Note:

Visualizing the confidence and prediction intervals in R is not straightforward, but requires some tedious handwork.

R-Hints:

```
dat  <- data.frame(ATM=seq(...,..., length=200))
pred <- predict(fit, newdata=dat, interval=...)
plot(..., ..., main="...")
lines(dat$ATM, pred[,2], col=...)
lines(dat$ATM, pred[,3], col=...)
```