Last Class!!

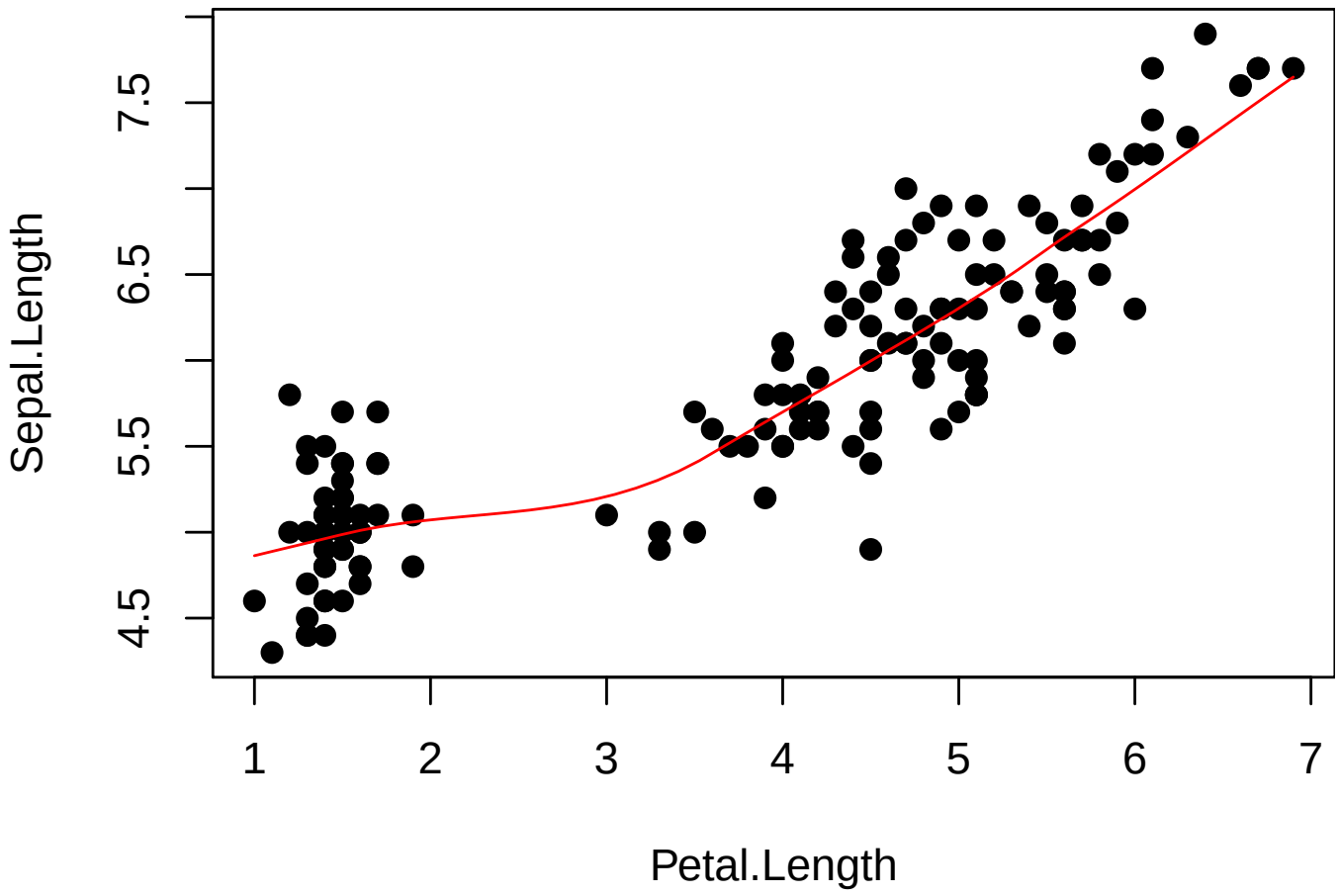
Applied Regression Analysis (STAT 757)

Paul J. Hurtado

5 Apr, 2016

Scatterplot Smoothing: LOESS

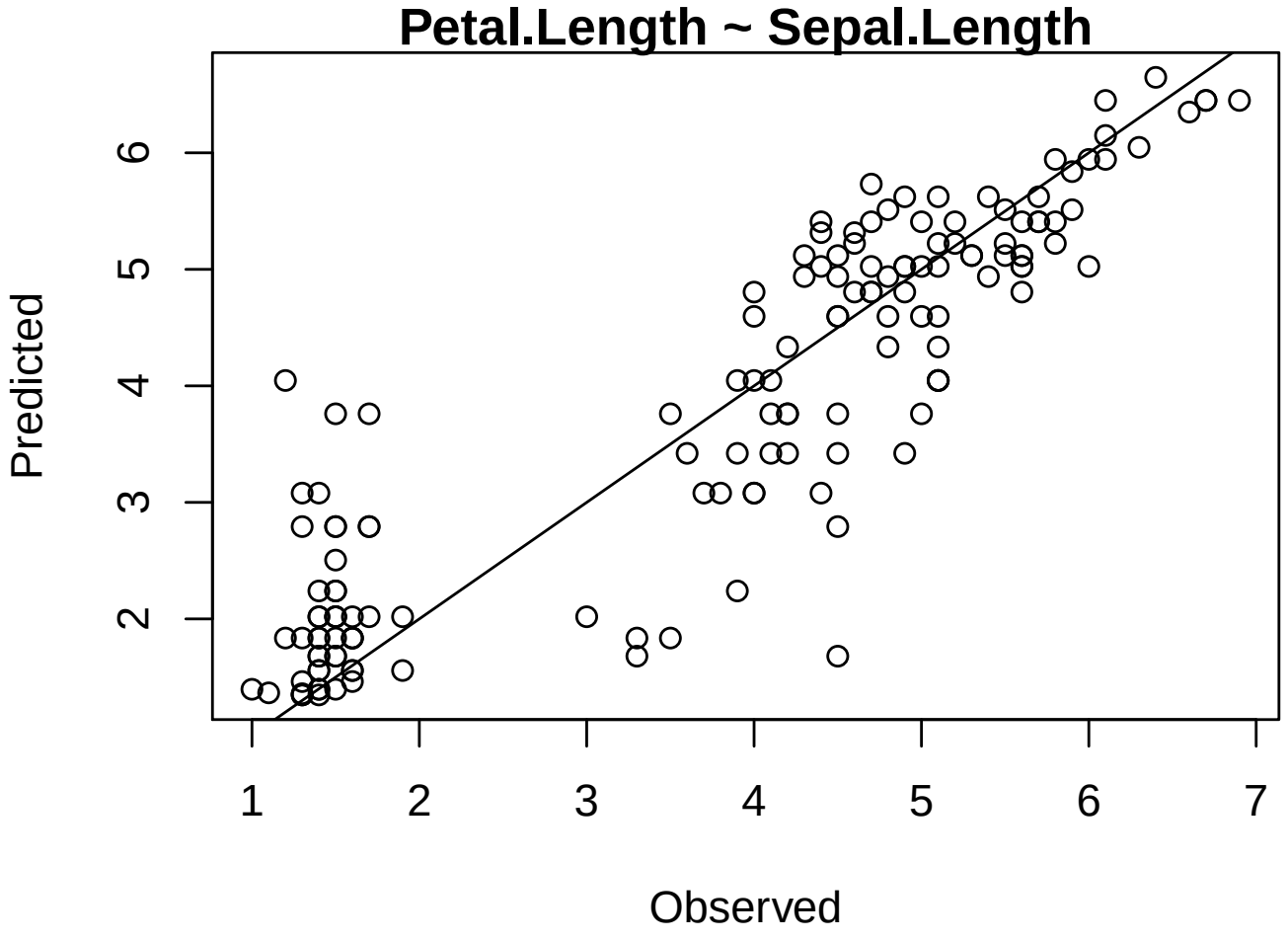
```
scatter.smooth(Petal.Length, Sepal.Length, pch=19, lpars=list(col="red"))
```



Scatterplot Smoothing: LOESS

Local Polynomial Regression via `loess()`

```
npfit = loess(Petal.Length ~ Sepal.Length, data=iris)
plot(iris$Petal.Length, predict(npfit, iris),
        xlab="Observed", ylab="Predicted", main="Petal.Length ~ Sepal.Length")
```



Nonparametric Multiple Regression

Polynomial approximation to f where instead of assuming (MLR)

$$E(Y|x) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

assume

$$E(Y_i|x) = f(x_{1i}, \cdots, x_{ki})$$

R packages:
`np`, `KernSmooth`, `gam::gam` and `mgcv::gam`, `sm`, `gss`...

Additive Regression Models

Additive Nonparametric Regression is similar to MLR, but with nonparametric transformation of predictor variables:

$$E(Y_i|x) = \beta_0 + f_1(x_{1i}) + \cdots + f_k(x_{ki})$$

Generalized Additive Models (GAMs) are similar to GLMs, but with nonparametric transformation of predictor variables:

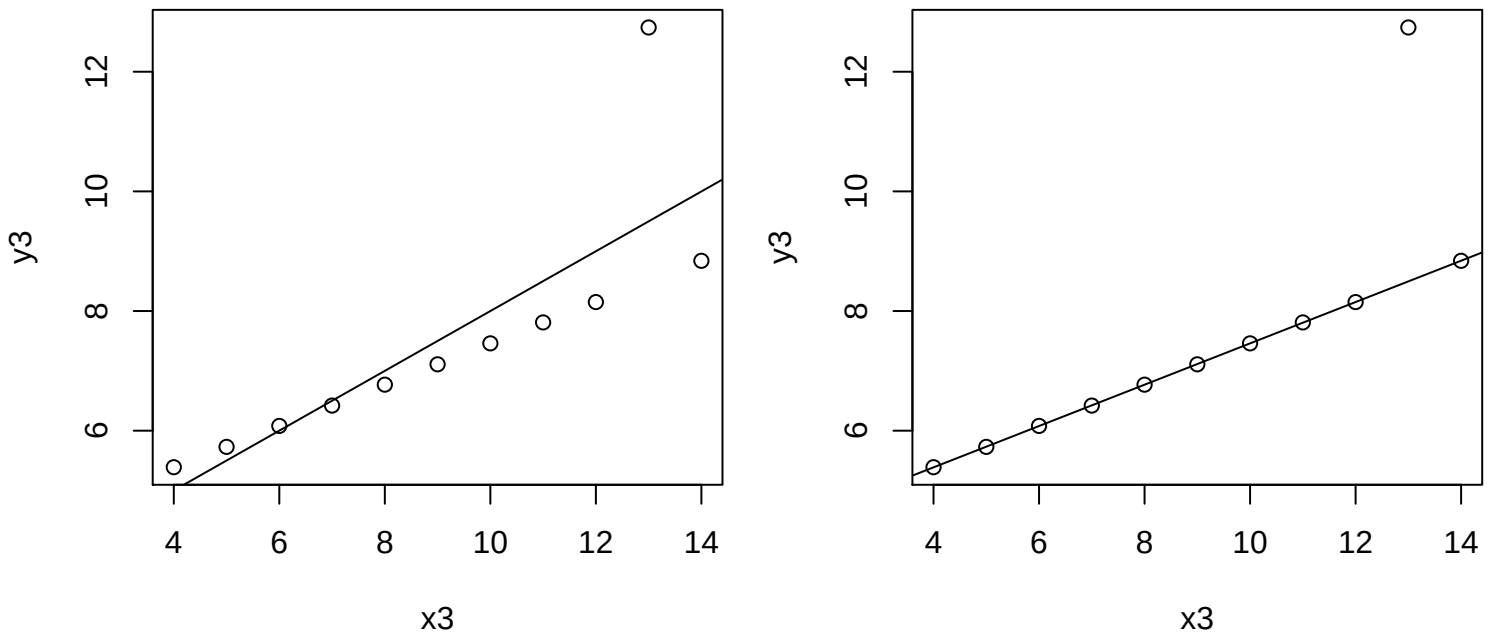
$$\eta_i = \beta_0 + f_1(x_{1i}) + \cdots + f_k(x_{ki})$$

For more, see the Nonparametric Regression appedix to Fox & Weisberg at: https://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Nonparametric-Regression.pdf

Robust Regression

Robust Estimation methods use estimators that are *efficient*[1] even if the distribution is heavy-tailed. MM-estimators are a popular approach.

```
library(robustbase)
fit=lm(y3~x3,data=anscombe)
rfit=lmrob(y3~x3,data=anscombe, maxit.scale=1000)
```



[1] Efficiency is a relative measure of the variance of an unbiased estimator, relative to the theoretical minimum given by the Cramer-Rao lower bound on estimator variances.

Time Series

Time Series Analysis is often done via ARIMA models (see `forecast` package), using time series objects created with the `ts()` function, and decomposed (e.g. to separate seasonal effects) using the `stl()` function.

More at:
`http://www.statmethods.net/advstats/timeseries.html`
CRAN Task View: Time Series Analysis
Time Series Analysis (STAT 758) in Fall 2016!

Bootstrapping

A kind of *resampling method* that replace distributional assumptions with data-driven computations.

1. Iteratively resample your data with replacement.
2. Calculate quantities of interest using the resampled data.

Population :: Sample as Sample :: Bootstrap Sample

More at:
http://www.statmethods.net/advstats/bootstrapping.html

Cross Validation

An alternative to the *training data vs test data* approach in Exam 2 is Cross Validation (CV).

Leave-One-Out CV:

1. Fit the model to your data N times, omitting the i^{th} data point from the i^{th} model fit.
2. Calculate

$$CV = \frac{\sum_{i=1}^{N}((\widehat{y_i} - y_i)^2}{N}$$

The model with the lowest average squared deviation from the leave-one-out predictions is considered the best model.

```
#
```

In *p-fold Crosss Validation* the data are divided into p subsets of approximately equal size, and each group is omitted and compared to predictions as above.

Machine Learning

Categories:
Supervised vs Unsupervised vs Reinforcement Learning vs ...

- K-Nearest Neighbors Classification Algorithm
- Decision Trees
- Association Rules
- Neural Networks
- Bayesian Networks
- ...

More at:
CRAN Task View: Machine Learning & Statistical Learning