STA314, Lecture 3

Sept 20, 2019
Stanislav Volgushev

Linear models

'All models are wrong but some are useful' - George Box

Linear regression models - introduction/recap

Recall: we have assumed regression model

$$y_i = f(x_i) + \varepsilon_i.$$

K-nn makes smoothness assumptions on f and takes local averages

Linear regression assumes a 'linear' form for f.

Simple linear regression for one-dimensional x_i assumes

$$f(x_i) = b_1 + b_2 x_i$$

where b_1, b_2 unknown parameters that need to be learned from data.

Multiple linear regression: if $x_i = (1, x_{i,1}, ..., x_{i,d})^\top$ are (d+1)-dimensional vectors assume

$$f(x_i) = x_i^\top b = b_1 + b_2 x_{i,1} + ... + b_{d+1} x_{i,d}$$

with parameter vector $b = (b_1, ..., b_{d+1})^\top$.

- notation $x_i = (1, x_{i,1}, ..., x_{i,d})^\top$ including 1 as the first entry is for notational convenience. This is not uniform across different sources, be careful!

Estimating multiple linear regression

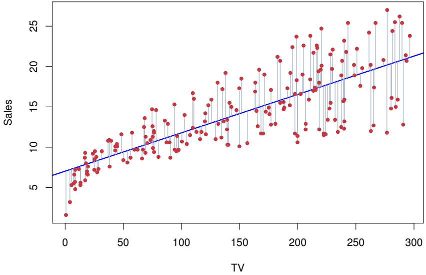
$$f(x_i) = x_i^\top b = b_1 + b_2 x_{i,1} + \ldots + b_{d+1} x_{i,d}$$

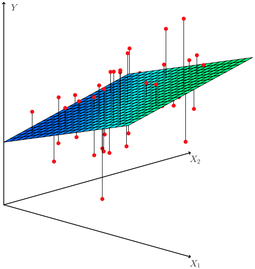
Estimation: minimize *residual sum of squares* (how does this relate to training error?)

$$RSS(b) := \sum_{i=1}^{n} \left(y_i - x_i^\top b \right)^2,$$

i.e. define

$$\hat{b} = \text{argmin}_{b \in \mathbb{R}^{d+1}} RSS(b).$$





Left: linear regression with $d = 1$ blue line: $\widehat{f}$, right: $d = 2$, plane: $\widehat{f}$.

Multiple linear regression: some calculations

Let $x_i = (1, x_{i,1}, ..., x_{i,d})^\top$. Define

$$\mathbf{X} := \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,d} \\ \vdots & \vdots & \dots & \vdots & \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,d} \end{pmatrix} \in \mathbb{R}^{n \times (d+1)}, \quad Y := \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

RSS has representation (see calculations in lectures)

$$RSS(b) = \sum_{i=1}^{n} \left(y_i - x_i^\top b \right)^2 = \|Y - \mathbf{X}b\|_2^2.$$

$RSS(b)$ has unique minimizer if and only if *Gram matrix* $\mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{(d+1)\times(d+1)}$ is invertible. In that case solution is given by (see calculations in lectures)

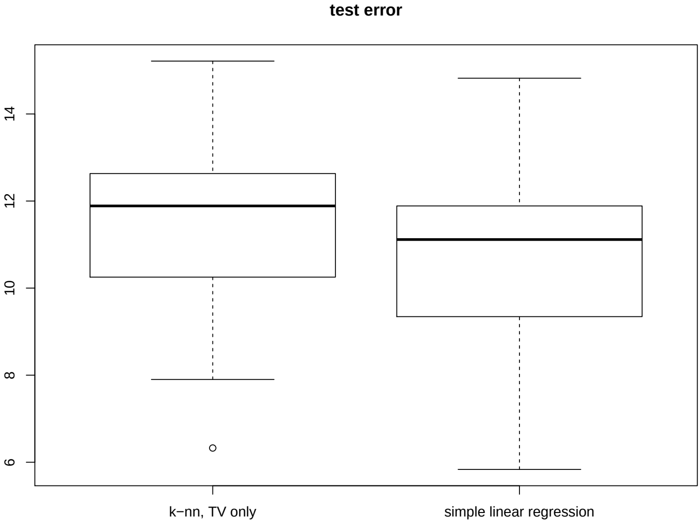
$$\hat{b} = \operatorname{argmin}_{b \in \mathbb{R}^{d+1}} \|Y - \mathbf{X}b\|_2^2 = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y.$$

- Explicit solution, no numerical optimization required!
- Easy to compute even for large data sets.
- Question: what can you say about $\hat{b}$ if the Gram matrix is not invertible?

Multiple linear regression in R

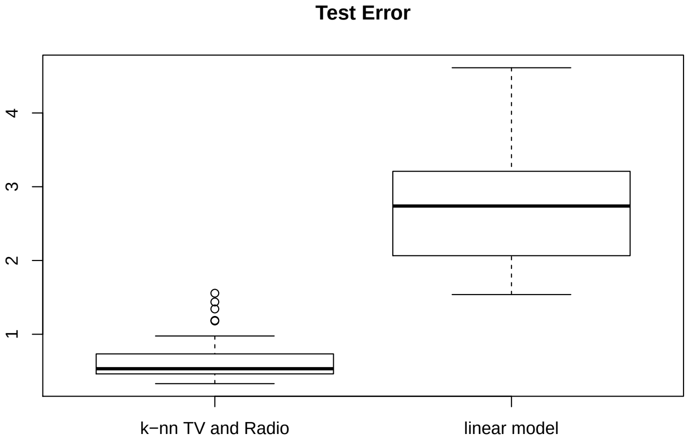
See R code.

Comparison with k-nn: one predictor



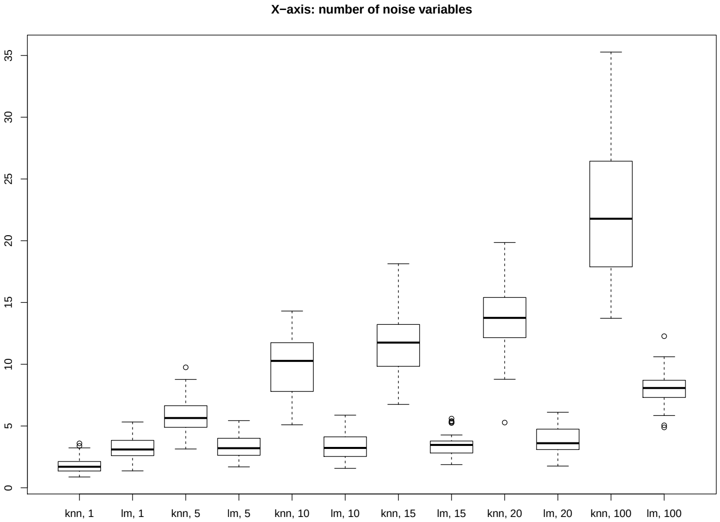
Comparison between linear model with just TV and k-nn with just TV (k selected via 10-fold cross-validation).

Comparison with k-nn: TV, Radio and Newspaper included



Comparison between linear model and k-nn with 2 predictors, TV and Radio.

Comparison with k-nn: adding many noise variables



Comparison between linear model and k-nn with different number of noise variables.

The story up to here

- Linear models are easy to fit (explicit formula).
- Performance can be similar to k-nn if regression function f is approximately linear.
- Typically do better than k-nn when there are many predictors.
- STA302: linear models are very interpretable and it is easy to do inference.
- Overall, 'basic' linear models as discussed in lectures do not have a very competitive prediction performance. However, they can provide good building blocks for more complex models that give better predictions - see coming lectures!

Multiple linear regression: some important statistics

The *total sum of squares* (poor man's training error)

$$TSS := \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad \bar{y} := \frac{1}{n} \sum_{i=1}^{n} y_i.$$

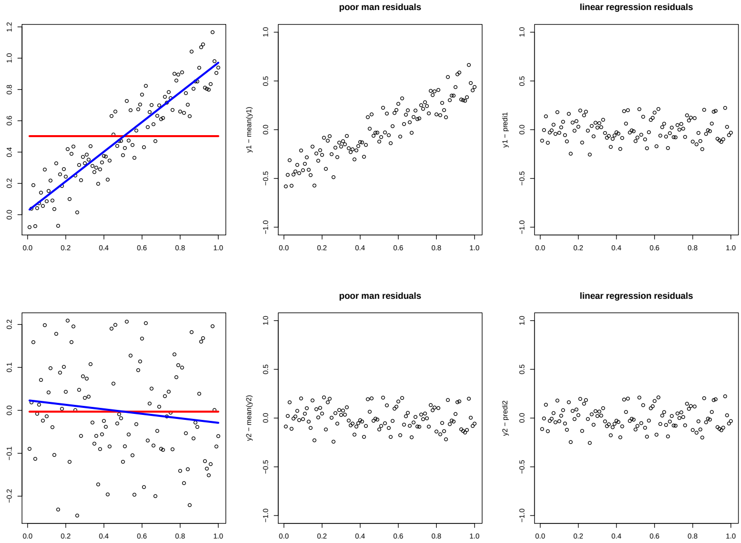
The R^2 statistic (*aka proportion of variance explained*)

$$R^2 := 1 - \frac{RSS(\widehat{b})}{TSS}, \quad RSS(b) = \sum_{i=1}^{n} (y_i - x_i^\top b)^2.$$

Interpretation:

- $0 \leq R^2 \leq 1$. How do you prove that?
- $RSS(\hat{b})$ measures *training error* of prediction $\hat{f}(x) = ...$ (see lectures). TSS measures training error of prediction $\hat{f}(x) = ...$ (see lectures).
- If R^2 close to 1 that means $RSS(\widehat{b})$ 'a lot' smaller compared to TSS, i.e. linear model has much better prediction accuracy compared to 'poor man's prediction' $\bar{y}$ which ignores all predictors.
- Usually R^2 is taken as a measure of how much variation in y_i the linear model can explain. Some caution is needed when interpreting it...

Examples



Red line: $\hat{f}(x) = \bar{y}$. Blue line: estimated linear regression function. Which of the two cases has bigger R^2?

Some things that R^2 does not mean

1. A small R^2 **does not mean** that the relationship between x and y is not linear.
2. A small R^2 **does not mean** that there is no relationship between predictor and outcome.
3. A large R^2 **does not mean** that the relationship is linear.

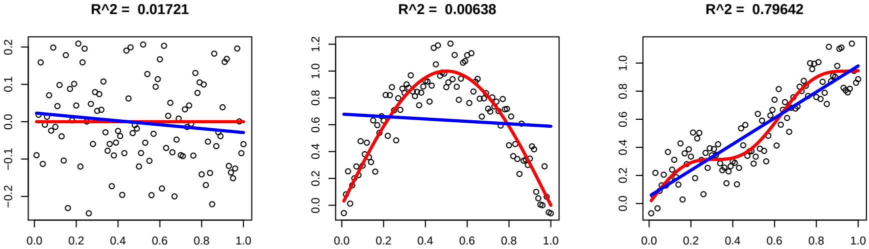


Figure: Red curve: true regression function. Blue line: estimated linear regression function.

Multiple linear regression in R: reading output

Running

```
lm(Sales~TV)
```

gives the output (assuming `Sales` and `TV` in workspace)

```
Call:
lm(formula = Sales ~ TV)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Multiple linear regression in R: reading output

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- Coefficients: information about $\hat{b}$ in linear regression model.
- (Intercept): information about $\hat{b}_1$, the intercept
- TV: information the coefficient of predictor TV
- Estimate: value of $\hat{b}_1$, $\hat{b}_2$ etc

t value: value of t-test statistic. $\Pr(>|t|)$: p-value of test for t-test. This is a test for $H_0 : b_j = 0$, i.e. predictor x_j *not informative in the presence of other predictors*.

Multiple linear regression in R: reading output

```
Residuals:
    Min      1Q  Median      3Q     Max 
-8.3860 -1.9545 -0.1913  2.0671  7.2124 

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099 
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

- Residuals: summary statistics about $\hat{\varepsilon}_i := y_i - x_i^\top \hat{b}$.
- Residual standard error (scaled square root of training error): $\sqrt{\frac{1}{n-d-1}\sum_{i=1}^{n} \hat{\varepsilon}_i^2}$
- Multiple R-squared: R^2 discussed earlier.

F-statistic and corresponding p value: for testing $H_0 : b_2 = ... = b_{d+1} = 0$, i.e. none of the predictors have linear effect on outcome (why this is called F test and what df means: see STA302).

Multiple linear regression: residual plots

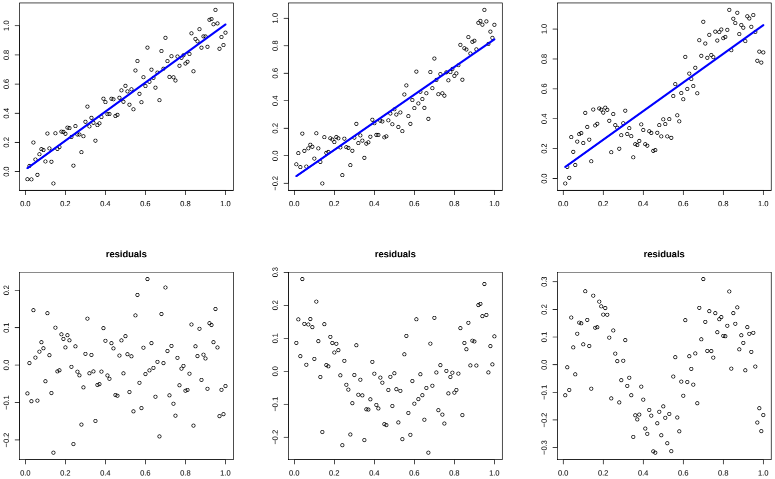
To get an overall impression of how well a regression function fits the data and if the linear model is correct, we often look at *residual plots*.

- If the predictor is one-dimensional, residuals $\widehat{\varepsilon}_i = Y_i - \widehat{f}(x_i)$ are plotted against x_i.
- If the predictor has more than one dimension, this is not feasible. In that case, residuals $\widehat{\varepsilon}_i$ are often plotted against predicted values $\widehat{f}(x_i)$.
- This will not always reveal deviations of the true model from linearity, so it is helpful to also plot against each predictor individually. This can also help with improving models.

Things to look for in a residual plot:

- *A pattern the mean of residuals* indicates that the regression function does not completely capture true relationship between predictor and outcome.
- *A pattern in the spread of residuals* indicates that the assumption of i.i.d. errors is not valid, this is sometimes called *heteroscedasticity*. Problematic for inference.

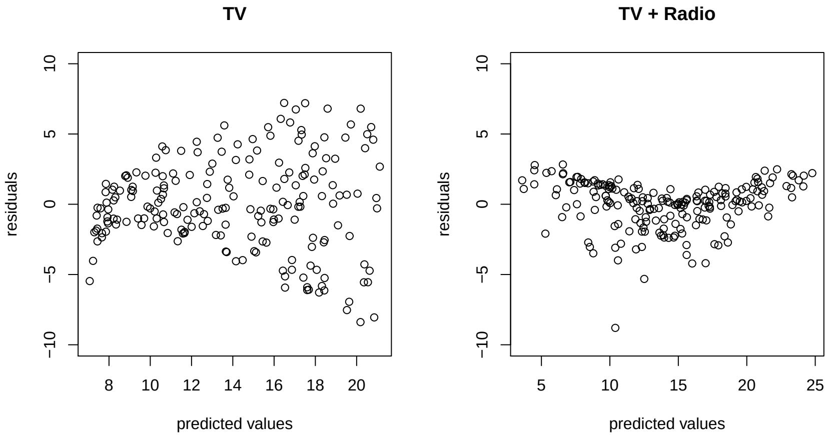
Examples of residual plots: pattern in mean



- Upper row: data (dots) and estimated linear regression functions in blue.
- Lower row: residuals plotted against predictor values.
- Second and third residual plot show pattern in mean of residuals.

Challenge: come up with an example where the true regression function is not linear but no pattern in the mean of residuals can be seen from plotting residuals against predicted values.

Multiple linear regression: residual plots for Advertising data



- Both plots show clear patterns in mean of residuals. So we can think about further improving the model.
- Left plot shows clear pattern in spread of residuals, right plot shows some pattern in spread of residuals.

Multiple linear regression: interactions

So far we have only allowed each predictor to influence the response independently of the values of other predictors. This is not always realistic.

- Example: in the model $\mathit{Sales} = b_1 + b_2 \times \mathit{TV} + b_3 \times \mathit{Radio} + \varepsilon$, the effect of spending additional money on TV does not depend on how much we spent on Radio.

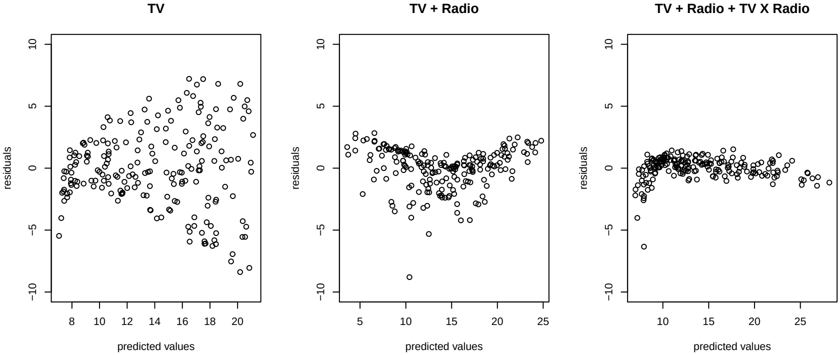
A simple way to model such effects while keeping the simple linear model structure is to allow for *interactions*. Including an interaction term between predictors $x_{i,1}, x_{i,2}$ means creating an additional predictor $x_{i,1}x_{i,2}$, i.e. a model of the form

$$f(x_i) = b_1 + b_2 x_{i,1} + b_3 x_{i,2} + b_4 x_{i,1} x_{i,2}$$

Note: this changes the interpretation of b_2, b_3. Now increasing $x_{i,1}$ by one unit will increase $f(x_i)$ by $b_2 + b_4 x_{i,2}$ units!

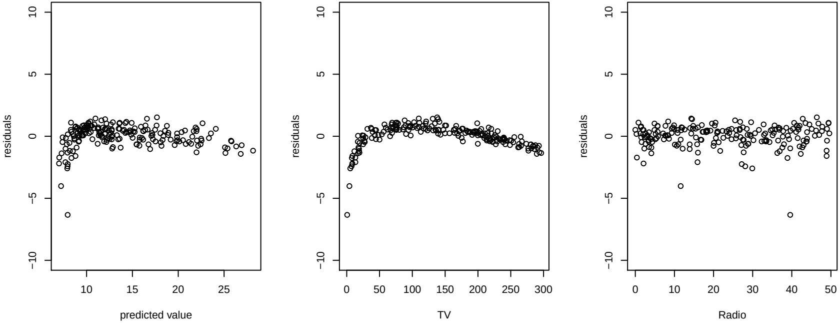
- Example: for the Advertising data we could consider a regression function of the form
$$\mathit{Sales} = b_1 + b_2 \times \mathit{TV} + b_3 \times \mathit{Radio} + b_4 \times (\mathit{TV} \times \mathit{Radio}) + \varepsilon.$$

Example: Advertising data set with interaction terms



- Mean of squared residuals for 3 models: 10.51, 2.78, 0.87. Including interactions leads to improvements in training error (this will always be the case! why?). If this translates to improved test error will need to be accessed via the approach discussed earlier.
- Model with interaction still shows clear pattern in mean of residuals, so there is still something this model misses. Pattern in spread of residuals almost gone.
- Let's look at some R code.

Advertising data: a closer look at residuals



- No pattern in residuals against `Radio`.
- Strong pattern in residuals against `TV`.
- Idea: allow for more flexible influence of `TV`.

Multiple linear regression: non-linear transformations of predictors

A popular way to allow non-linear effects of predictors on response is *polynomial regression of order k* where $f(x)$ is a polynomial of the form

$$f(x) = b_1 + b_2 x + b_3 x^2 + ... + b_{k+1} x^k.$$

- this is still *linear in the coefficients* $b_1, ..., b_{k+1}$.
- if we treat $x, x^2, ..., x^k$ as new predictors this behaves like a multiple linear regression model with $k + 1$ predictors.

In general, any model of the form (recall: $x_i = (1, x_{i,1}, ..., x_{i,d})^\top$)

$$f(x_i) = b_1 + b_2 g_1(x_i) + b_3 g_2(x_i) + ... + b_L g_{L+1}(x_i)$$

for **given** functions $g_1, ..., g_L$ and unknown coefficients $b_1, ..., b_{L+1}$ inherits many properties of a linear model (can you write the linear model we considered so far in this form?). A popular class of models is given by *basis expansions*, this will be discussed later in this course.

Examples of non-linear transformations of predictors

Which of the functions on this slide can be written as

$$f(x_i) = b_1 + b_2 g_1(x_i) + b_3 g_2(x_i) + ... + b_{d+1} g_d(x_i)$$

Example 1: $x_i = (1, x_{i,1}, x_{i,2})^\top$,

$$f(x_i) = b_1 + b_2 x_{i,1} + b_3 x_{i,2} + b_4 x_{i,1} x_{i,2}$$

Example 2: $x_i = (1, x_{i,1}, x_{i,2})^\top$

$$f(x_i) = b_1 + b_2 \sin(x_{i,1}) + b_3 \cos(1 + x_{i,2}) + b_4 x_{i,2}^3$$

Example 3: $x_i = (1, x_{i,1}, x_{i,2})^\top$

$$f(x_i) = b_1 + \sin(b_2 + x_{i,1}) + b_3 x_{i,2}$$

R syntax for linear models

```
X:Z
```

is an interaction term between `X,Z`.

```
X*Z
```

means include `X,Z` and interaction `XZ`.

```
I(X^k)
```

means include $\mathtt{X}^k$ as predictor. Some other functions such as sin, cos, exp work directly but $I(\cdot)$ always works. Type `?formula` for details. See also function `poly` for polynomial regression.

If `D` is a data frame that contains a column with name `Y`

```
lm(Y ~ ., data = D)
```

will run a linear regression with Y as response and all other columns of `D` as predictors.

If `D` is a data frame that contains columns with names `Y`, `Z`

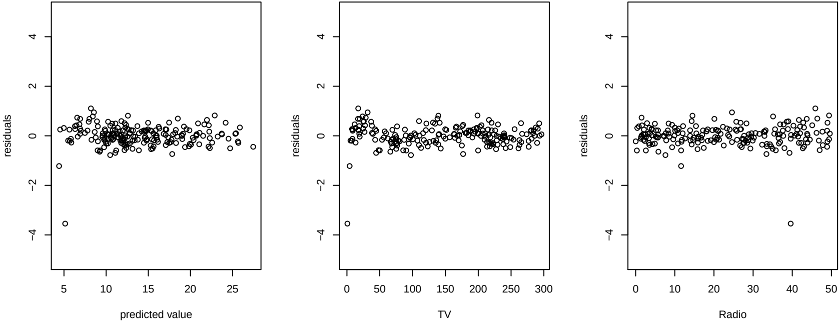
```
lm(Y ~ . - Z, data = D)
```

will run a linear regression with Y as response and all other columns of `D` **except Z** as predictors.

Non-linear transformation of predictors: Advertising data revisited

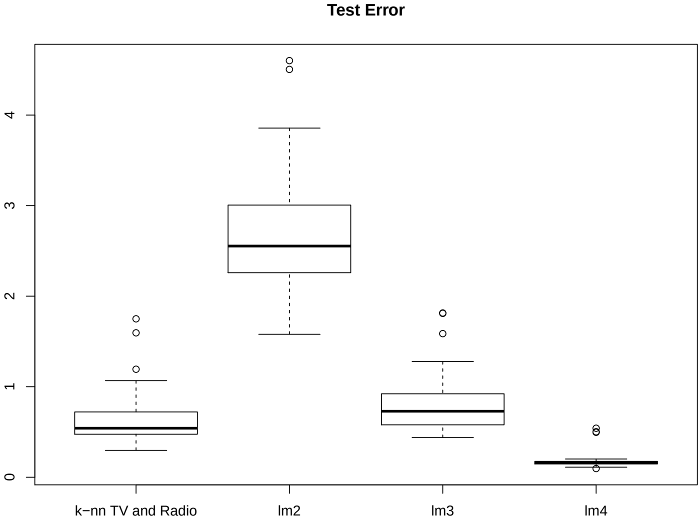
The residual plot of `Sales` against `TV` suggested that the effect of `TV` on `Sales` can not be modelled as a simple linear function. Try model

$$Sales = b_1 + b_2 \times Radio + b_3 \times (TV \times Radio) + \sum_{k=1}^{5} b_{3+k} \times TV^k + \varepsilon.$$



- Pattern in residuals almost disappeared.
- Mean of squared residuals decreased from 0.87 to 0.19, to be verified on test set.

Multiple linear regression vs K-nn: Advertising example



- lm2: additive form, lm3: included interaction, lm4: included interaction and non-linear transformation of TV. lm4 beats k-nn with TV and Radio by considerable margin

STA314, Lecture 3

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.