Multiple Linear Regression

Peter Büchel

HSLU W

SA: W 08

Linear Regression

- Generalisation of Anova (DoE)
- Now including hypothesis tests
- Linear regression is a steping stone into Machine Learning

Introduction, Example

- Job for statistician of a company: Analysis, to work out strategy how to increase sales of a certain product
- Company provides data on advertising budget and sales
- Data set `Advertising` consists of:
  - `sales` of this product in 200 different markets
  - Advertising budget of this product in these markets for three different media: `TV`, `radio` and `newspaper`
- Code:

```
adv <- read.csv("../Data/Advertising.csv")[, -1]
head(adv, 3)
##      TV radio newspaper sales
## 1 230.1  37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
```

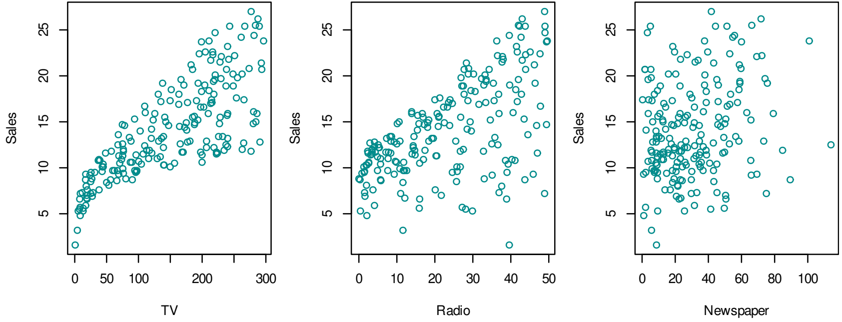
- Data shown in scatter plots:

```r
TV <- adv[, 1]
Radio <- adv[, 2]
Newspaper <- adv[, 3]
Sales <- adv[, 4]

plot(Sales ~ TV, col = "darkcyan", xlab = "TV", ylab = "Sales")

plot(Sales ~ Radio, col = "darkcyan", xlab = "Radio", ylab = "Sales")

plot(Sales ~ Newspaper, col = "darkcyan", xlab = "Newspaper",
    ylab = "Sales")
```



- For company not possible to directly increase sales of the product
- But it can control advertising spending in the three media
- Aim: Establish a link between advertising and sales so that companies can adjust their advertising budgets to indirectly increase sales
- Aim: Develop a *model* as accurately as possible, so that on the basis of the three media budgets the sale of the product can be *predicted*
- `TV`: Clear relationship between advertising and sales of product
- The more money invested in advertising, the greater the sales figures
- Question: What *form* does this relationship take?
- `newspaper`: No relationship at all: No need for newspaper advertising

- Mathematical view: Look for function f which determines the sale Y depending on the advertising budgets X_1 (`TV`), X_2 (`radio`) and X_3 (`newspaper`):

$$Y \approx f(X_1, X_2, X_3)$$

- Relationship above: No equal sign, since scatter plots do *not* represent graphs of a function

- Function f can only display the relationship between X_1, X_2, X_3 and Y *approximately*

- Notation:
    - Variable Y: *Response variable*
    - X_1, X_2 and X_3: *Predictors*, *explanatory variable*

- Generally: Quantitative response variable Y and p different predictors $X_1, X_2, \ldots, X_p$

- Assume: There is *somehow* a relationship between Y and $X_1, X_2, \ldots, X_p$

- General form:
$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

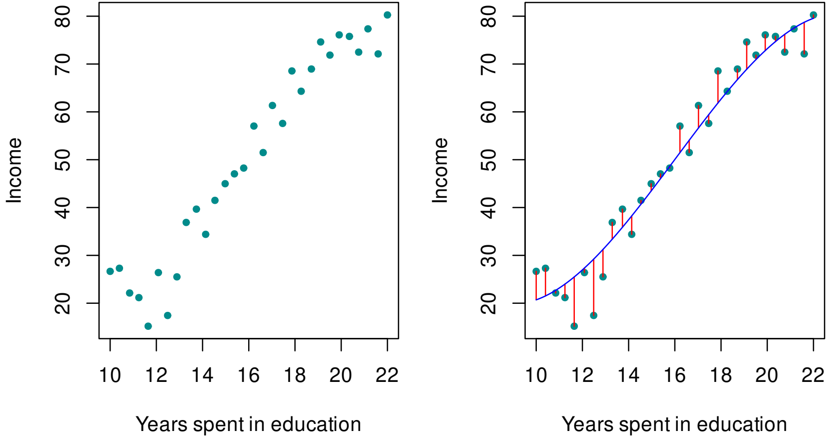
- f fixed but *unknown* function of $X_1, X_2, \ldots, X_p$

- Quantity ε: *Random error term* independent of $X_1, X_2, \ldots, X_p$ with mean

- Meaning of error term ε: Following example

Example: Income

- Figure left: `income` of 30 individuals as a function of `education` (in years)
- Graph indicates: `income` can be calculated from `education`



- But: Function f which links predictors and response variables, usually unknown
- In this situation: *Estimate* f from the data
- Data set simulated: Function f known (blue curve) in right figure
- Some observations are above, others below the blue curve
- Red vertical lines: Represent the error term ε
- Overall, errors have an empirical mean close to 0
- Aim of the regression: *Estimate* function f

- Estimation in stochastics: Calculation of values
- Estimation is an approximation of true quantity
- Estimated quantity is marked with hat $\widehat{\ }$
- $\widehat{Y}$: Estimate of unknown quantity Y
- $\widehat{f}$: Estimate of unknown function f

Example

- Predictors $X_1, X_2, \ldots, X_p$: Values of various characteristics of a blood sample that the patient's family doctor can determine in his laboratory
- Response variable Y: Measure of the risk that the patient will suffer severe side effects when using a particular drug
- Physician: Wants to predict Y based on $X_1, X_2, \ldots, X_p$ when prescribing a drug Y so that he does not prescribe a drug to patients who are at high risk for side effects of this drug - i.e. where Y is large

Questions for example of the `advertising`

- Which media contribute to the sale of the product?
- Which media have the greatest influence on sales ?
- What increase in sales does a particular increase in TV advertising result in?

Estimate of f

- Several procedures to estimate f

- Here only *parametric method*

- Procedure:
  - *Assume* functional form of f
  - Simplest assumption: f linear in $X_1, X_2, \ldots, X_p$:

$$f(X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

  - Choice of model: Procedure that fits the data into the model *best*
  - Linear Model: Estimate parameter $\beta_0, \beta_1, \ldots \beta_p$
  - Parameter so that

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

  - Most common method for determining $\beta_0, \beta_1, \ldots \beta_p$: *Least squares method*

Examples

- Example `advertising`: Linear model:

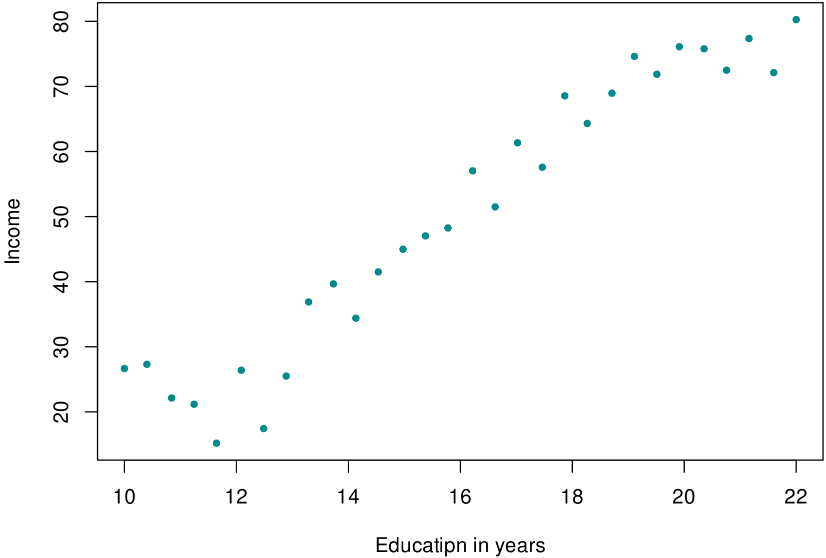
$$\texttt{Sales} \approx \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{radio} + \beta_3 \cdot \texttt{newspaper}$$

- Example `Income`: Linear model:

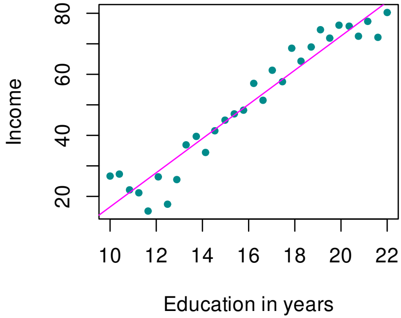
$$\texttt{Income} \approx \beta_0 + \beta_1 \cdot \texttt{Education}$$

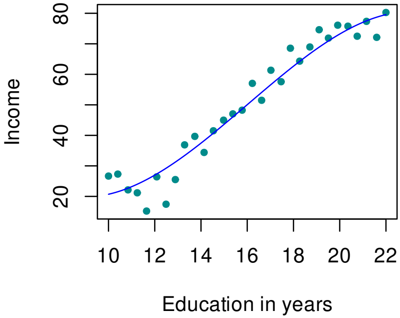
Example

- Data set `income`:



- Question: Which *model* to choose, or which shape should f have





- From data: Linear model (top left):

$$f(X) = \beta_0 + \beta_1 X$$

- Also cubic model (polynomial 3rd degree) possible (top right):

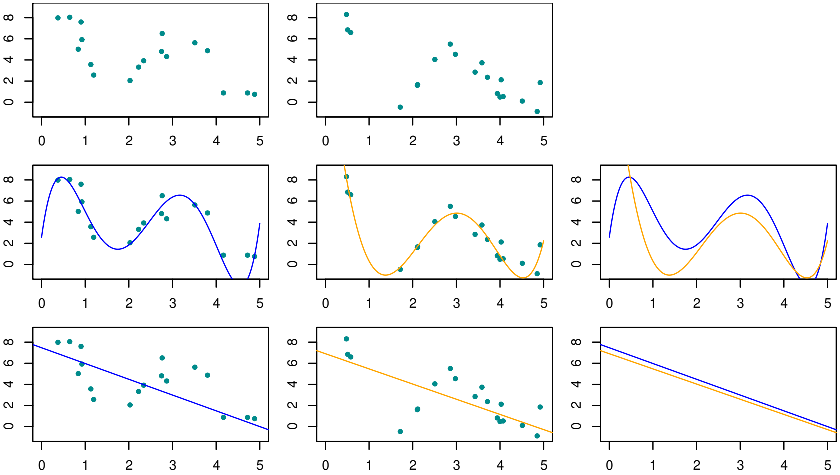
$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

- Many other models conceivable
- But which is the "correct" one?
- Wrong question
- Function f unknown: It is up to us to choose the "best" model
- Statistics: Assisting in decision-making
- Which model is the "better" one in our example?
- Cubic model seems to fit better, but is more complicated
- Simpler linear model (slightly less accurate) has an advantage: The parameters β_0 and β_1 can be interpreted geometrically:
    - β_0 is the y intercept
    - β_1 the slope of line
- Parameters in cubic model are *not* interpretable (except β_0)

Remarks

- More complicated models do *not* have to be the better models

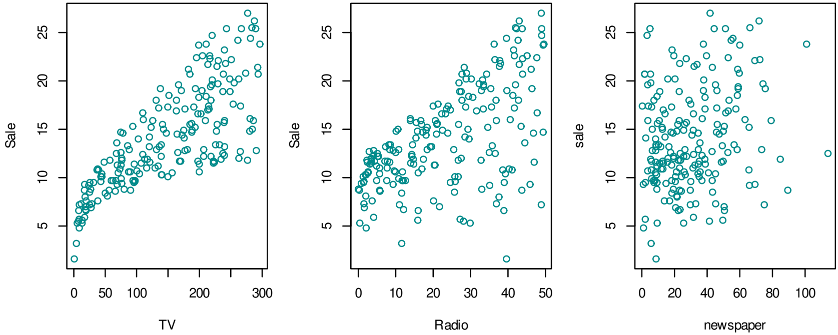
- Phenomenon: *Overfitting*



- Errors or outliers are taken too much into account
- In a lot of cases: Linear model sufficient
- Keep it simple often works best

Linear regression

- Data set `advertising`:



- `Sales` for a given product (in units of one thousand products sold) as a function of advertising budgets (in units of one thousand CHF) for `TV`, `radio` and `newspaper`

- Based on this data: Statisticians draw up a marketing plan that should lead to higher sales next year
- What information is useful for drawing up such recommendations?

Simple regression model

- *Simple linear regression*: Very simple procedure to obtain a quantitative output Y on the basis of a single predictors X
- Assumption: Approximately linear relationship between X and Y
- Mathematically: Linear relationship:

$$Y \approx \beta_0 + \beta_1 X$$

- „$\approx$" stands for „is approximately modelled by"

Example

- Example `Advertising`: Quantity X `TV` and quantity Y `sales`

- According to the linear regression model, it follows

$$\texttt{sales} \approx \beta_0 + \beta_1 \cdot \texttt{TV}$$

- Variables β_0 and β_1 are unknown constants representing the intercept and slope of the linear model

- β_0 and β_1: *Coefficients* or *parameters* of model

- Coefficients are estimated from the given data
- Estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for the model coefficients
- If these coefficients are known, future sales can be predicted on the basis of a specific advertising budget for TV
- Calculation by means of:

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

where $\widehat{y}$ denotes the prediction of Y based on the input $X = x$

Example

- Example `Advertising`: $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and determine the regression line:

```
lm(Sales ~ TV)
##
## Call:
## lm(formula = Sales ~ TV)
##
## Coefficients:
## (Intercept)           TV
##     7.03259      0.04754
```

- Value under `Intercept`: $\widehat{\beta}_0$: y intercept

- Value under `TV`: $\widehat{\beta}_1$ of regression line

- Linear Model:

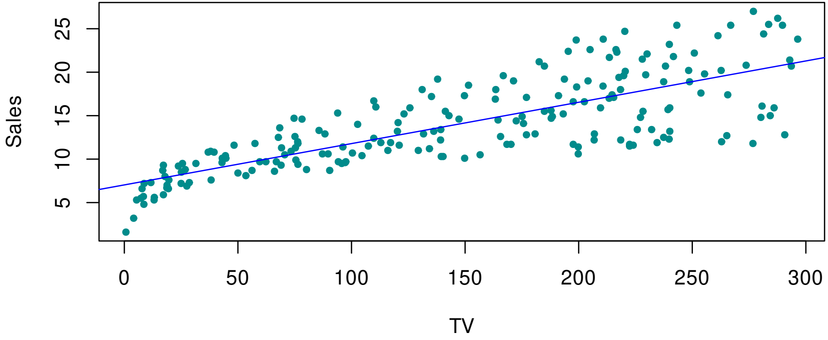
$$Y \approx 7.03 + 0.0475X$$

- According to approximation: For additional CHF 1000 advertising expenses 47.5 additional units of the product are sold

- Scatter plot with regression line

```r
plot(TV, Sales, col = "darkcyan", xlab = "TV", ylab = "Sales",
    pch = 20)

abline(lm(Sales ~ TV), col = "blue")
```



Hypothesis test: Statistical significance of β_1

- Most common hypothesis test: Testing the *null hypothesis*

$$H_0 : \quad \text{There is } \textit{no} \text{ relationship between } X \text{ and } Y$$

- *Alternative hypothesis*

$$H_A : \quad \text{There is } \textit{a} \text{ relationship between } X \text{ and } Y$$

- Mathematically:

$$H_0 : \quad \beta_1 = 0$$

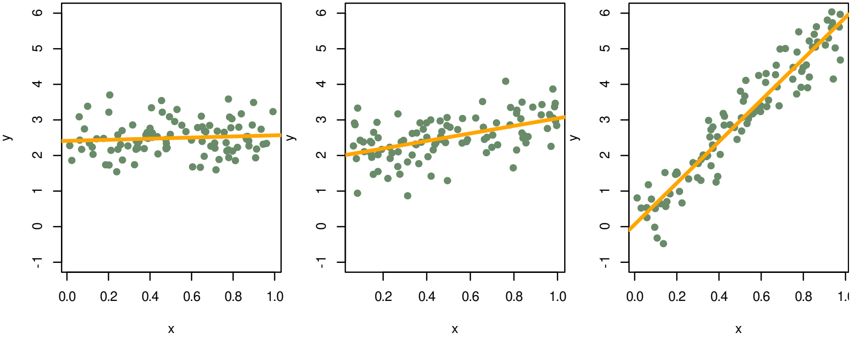
- Alternative:

$$H_A : \quad \beta_1 \neq 0$$

- $\beta_1 = 0$, then:

$$Y = \beta_0 + \varepsilon$$

- Sketch:



- Y does *not* depend on X

- Testing null hypothesis: $\widehat{\beta}_1$ sufficiently far from 0 so that β_1 is not 0

- With t test

Example

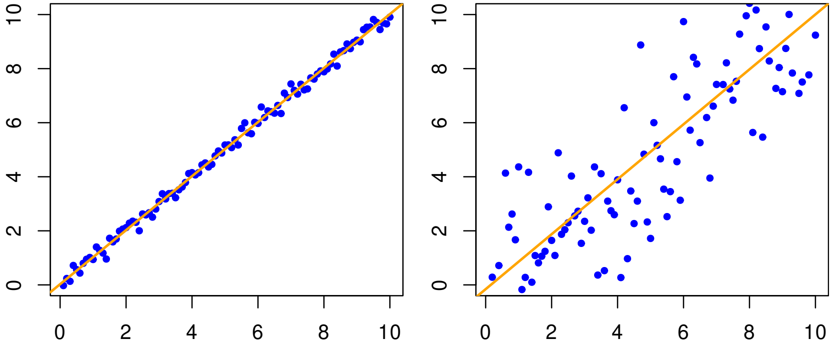
- p value of β_1 in the example `advertising` calculate:

```
summary(lm(Sales ~ TV))
##
## 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
```

- Entry `Coefficients` under `Pr(>|t|)`: p value $2 \cdot 10^{-16}$
- By far less than 0.05
- Reject null hypotheses $\beta_1 = 0$: $\beta_1 \neq 0$
- Clear indications of the link between `TV` and `sales`

Evaluation of the accuracy of the model: R^2

- Null hypothesis rejected: *To what extent does the model fit the data?*
- Figure:



- Left: Ascending line fits points very well with
- Right: Ascending line fits does *not* points well

- Accuracy of linear regression estimated by the *residual standard error* (RSE) and the R^2 statistics
- R^2 more important
- R^2-statistics: Value between 0 and 1
- It indicates to what proportion of the variability in Y is explained by X using the model
- Value close to 1: A large proportion of the variability is explained by the regression. The model therefore describes the data very well.
- Value close to 0: Regression does not explain variability of response variable
- Again: Graphical "derivation" (lecture notes)

Remarks:

- Empirical correlation only indicates the accuracy of *linear* regressions
- R^2 can be used for *any* regression
- Standard interpretation of R^2: "Proportion of the variability which is explained by the model"
- However: Pretty useless
- See `https://data.library.virginia.edu/is-r-squared-useless/?s=03`

Example

- In example of TV advertising the R^2 value is 0.61

```r
summary(lm(Sales ~ TV))$r.squared
## [1] 0.6118751
```

- Thus almost two thirds of the variability in `Sales` is explained by `TV` with linear regression

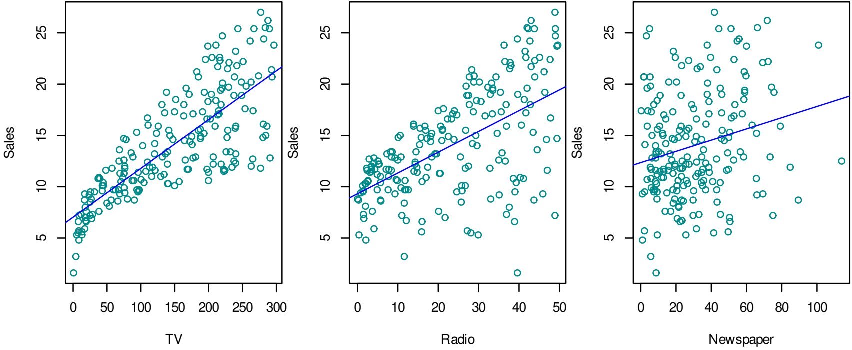
Multiple Linear Regression

- Simple linear regression: Useful procedure to predict output based on *one* single predictor
- In practice: Output often depends on more than one predictor

Example

- Dataset `Advertising`: Have seen relationship between `TV` advertising and `Sales`
- Data on advertising for `Radio` and `Newspaper` also available
- Question: Do one or both of these advertising expenses affect sales?
- Extend analysis of sales figures: Consider both additional inputs

- Possible: Perform a simple regression for each separate advertising budget
- Figure:



- Parameters and other important data in tables below

- Simple regression from `TV` to `Sales`:

	Coefficient	Std.error	t statistics	p value
`Intercept`	7.033	0.458	15.36	< 0.0001
`TV`	0.048	0.003	17.67	< 0.0001

- Simple regression from `Radio` to `Sales`:

	Coefficient	Std.error	t statistics	p value
`Intercept`	9.312	0.563	16.54	< 0.0001
`Radio`	0.203	0.020	9.92	< 0.0001

- Simple regression from `Newspaper` to `Sales`:

	Coefficient	Std.error	t statistics	p value
`Intercept`	12.351	0.621	19.88	< 0.0001
`Newspaper`	0.055	0.017	3.30	< 0.0001

- Separate simple linear regressions: Not satisfactory
- First: Not clear how to make a prediction for sales for given values of three predictors:
  - Each input linked to sales by *different regression equation*
- Second: Each of three regression equations ignores other two predictors for determining coefficients
- May lead to very misleading estimates of effect on sales of advertising expenses for each medium if three predictors are correlated

- Better: All predictors directly taken into account
- Each predictor is assigned *own* slope coefficient in *one* equation
- General: p different predictors
- *Multiple linear regression model*:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

- X_j: j-th predictor
- β_j: Relationship between *this* predictor and response variable Y
- β_j: Average change of response variable when changing X_j by one unit, *if all other predictors are kept constant*
- In other words: Slope in direction of X_j

Example

- Multiple linear regression model for dataset `Advertising`:

$$\texttt{Sales} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{Radio} + \beta_3 \cdot \texttt{Newspaper} + \varepsilon$$

- So:

$$\texttt{Sales} \approx \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{Radio} + \beta_3 \cdot \texttt{Newspaper}$$

- Multiple linear model generalises simple linear model
- Calculations and interpretations for multiple model similar, although usually more complicated than linear model
- Graphical methods: Virtually no use for multiple linear regression
- Data points for previous example: Not possible, as three axes are already needed for predictors

Example: Income

- Graphical representation possible for two predictors
- Dataset Income

```
In <- read.csv("../Data/Income2.csv")[, -1]
head(In)
```

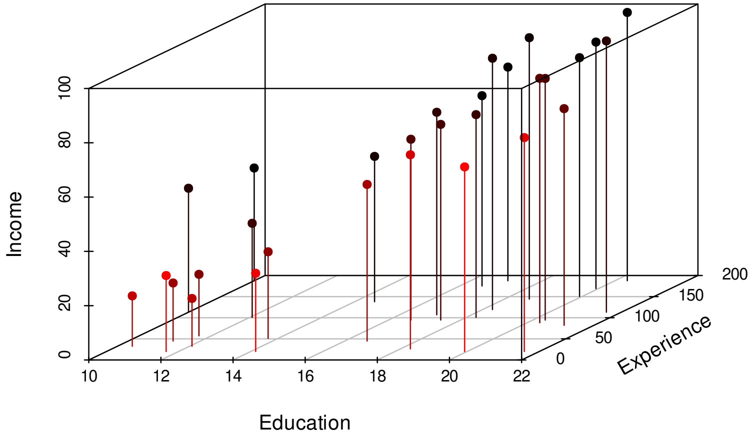
```
##   education experience   income
## 1  21.58621   113.1034 99.91717
## 2  18.27586   119.3103 92.57913
## 3  12.06897   100.6897 34.67873
## 4  17.03448   187.5862 78.70281
## 5  19.93103    20.0000 68.00992
## 6  18.27586    26.2069 71.50449
```

- So far: Education single predictor
- Income also depends on Experience (number of professional months)

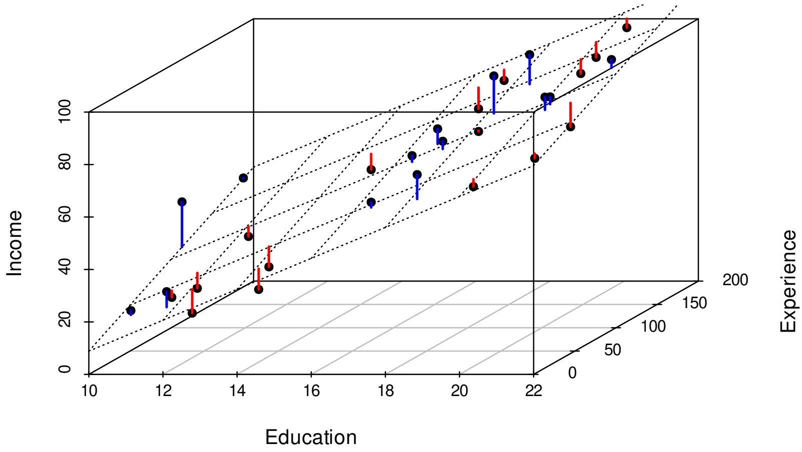
- Multiples linear model:

$$\texttt{Income} = \beta_0 + \beta_1 \cdot \texttt{Education} + \beta_2 \cdot \texttt{Experience} + \varepsilon$$

- Data points in 3d space:



- Analogous simple linear regression model: Look for *plane* that fits data points best



- Procedure analogous to simple linear regression
- Determine plane such that sum of squares of distances of data points from plane becomes minimal
- Lines:
  - Blue: Points above plane
  - Red: Points below plane
- Differences from point to plane: *Residuals*
- Use again: *Least squares method*

- Estimate of β_0, β_1 and β_2 with R:

$$\widehat{\beta}_0 = -50.086, \qquad \widehat{\beta}_1 = 5.896, \qquad \widehat{\beta}_2 = 0.173$$

```
coef(lm(Income ~ Education + Experience))
## (Intercept)   Education  Experience
## -50.0856387   5.8955560   0.1728555
```

- Multiple linear model:

$$\texttt{Income} \approx -50.086 + 5.896 \cdot \texttt{Education} + 0.173 \cdot \texttt{Experience}$$

Interpretation of Coefficients

- $\widehat{\beta}_0 = -50.086$:
  - If person has no education and no experience, earns CHF $-50\,086$
  - Interpretation makes no sense of course
- $\widehat{\beta}_1 = 5.896$:
  - With constant experience, you earn CHF 5896 more for each year of additional education
- $\widehat{\beta}_2 = 0.173$:
  - With a constant education, you earn CHF 173 more per additional month of work experience

General: Estimation of Regression Coefficients

- Like simple linear regression: Regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ generally unknown

- Estimation from data:

$$\widehat{\beta}_0, \quad \widehat{\beta}_1, \quad \ldots, \quad \widehat{\beta}_p$$

- Based on estimates, one can make predictions:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \ldots + \widehat{\beta}_p x_p$$

- Estimate parameters: Use again least squares method

- dialling $\beta_0, \beta_1 \ldots, \beta_p$ so that the sum of the residual squares RSS

$$
\begin{aligned}
\text{RSS} &= \sum_{i=1}^{n} r_i^2 \\
&= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\
&= \sum_{i=1}^{n} (y_i - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \ldots - \widehat{\beta}_p x_{ip})^2
\end{aligned}
$$

  is minimised

- Where x_{ij} is i-th observation of j-th predictor

- principle same as for simple linear regression

Example

- R: Multiple linear regression model for `Advertising`:

```
coef(lm(Sales ~ TV + Radio + Newspaper))
## (Intercept)          TV       Radio   Newspaper
## 2.938889369 0.045764645 0.188530017 -0.001037493
```

- It follows:

$$\texttt{Sales} \approx 2.94 + 0.046 \cdot \texttt{TV} + 0.189 \cdot \texttt{Radio} - 0.001 \cdot \texttt{Newspaper}$$

- Coefficients:
  - For given advertising expenses for radio and newspapers, an additional CHF 1000 of advertising expenses for TV will result in sale of about 46 more units
  - For given TV and newspaper advertising expenses, an additional CHF 1000 of advertising expenses for radio will result in sale of approximately 189 more units
  - Interesting: For newspaper you would sell *less* products if you invested *more*

- Table: Other important values:

	coefficient	Std.error	t statistics	p value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`Radio`	0.189	0.0086	21.89	< 0.0001
`Newspaper`	-0.001	0.0059	-0.18	0.8599

- Code: Replace `coef` by `summary`

```
fit <- lm(Sales ~ TV + Radio + Newspaper)

summary(fit)
##
## Call:
## lm(formula = Sales ~ TV + Radio + Newspaper)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## Newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

- Coefficient of separate simple linear regressions in slide 38
- Slopes of multiple linear regression for `TV` and `Radio` very similar:
  - `TV`: 0.46 (multiple), 0.48 (single)
  - `Radio`: 0.189 (multiple), 0.203 (single)
- Estimated regression coefficient $\widehat{\beta}_3$ for `Newspaper` shows different behaviour:
  - Simple: 0.055 (not equal to 0)
  - Multiple: -0.001 (almost equal to 0)
- Corresponding p values:
  - Simple: < 0.0001 (highly significant)
  - Multiple: 0.86 (far from being significant)

- Simple and multiple regression coefficients can be very different
- Simple regression: Slope indicates change in response `Sales` when spending CHF 1000 more on newspaper advertising, while other two predictors `TV` and `Radio` are *ignored*
- Multiple linear regression: Slope for `Newspaper` describes change in response `Sales` when spending CHF 1000 more on newspaper advertising, while other two predictors `TV` and `radio` are hold *constant*
- Does it make sense that multiple regression does not suggest a relationship between `Sales` and `Newspaper`, but simple regression implies opposite?

- It does make sense indeed

- Table with correlation coefficients:

	TV	Radio	Newspaper	Sales
TV	1.0000	0.0548	0.0567	0.7822
Radio		1.0000	0.3541	0.5762
Newspaper			1.0000	0.2283
Sales				1.0000

- Code:

```
cor(data.frame(TV, Radio, Newspaper, Sales))
##                   TV      Radio  Newspaper     Sales
## TV        1.00000000 0.05480866 0.05664787 0.7822244
## Radio     0.05480866 1.00000000 0.35410375 0.5762226
## Newspaper 0.05664787 0.35410375 1.00000000 0.2282990
## Sales     0.78222442 0.57622257 0.22829903 1.0000000
```

- Correlation coefficient `Radio` and `Newspaper`: 0.35
- What does this mean?
- Shows a tendency to invest more in advertising for `Newspaper` when advertising expenses for `Radio` is increased
- Assume: Multiple regression model *correct*
- Expenses on `Newspaper`: No direct influence on `Sales`
- Advertising expenses for `Radio`: Higher sales
- In markets where more is invested in radio advertising, expenses on `Newspaper` is also higher, as correlation coefficients of 0.35

- Simple linear regression: Only correlation between `Newspaper` and `Sales`, whereby for higher values of `Newspaper` also higher values of `Sales` are observed
- Simple linear regression only “sees” increase in `Sales`
- But: Newspaper advertising does *not* influences sales
- Higher values for `Newspaper` due to correlation also result in higher values for `Radio`: *This* quantity influences `Sales`
- `Newspaper` “takes credit” for success of `Radio` on `Sales`
- This result conflicts with intuition
- Occurs frequently in real situations

Absurd example

- Simple regression: Relationship between shark attacks and ice cream sales on a given beach
- The greater the ice cream sales, the more frequent shark attacks
- Absurd idea: Ban ice cream sales on this beach so that there are no more shark attacks
- But where is the connection?
- Reality: In hot weather more people come to beach $\rightarrow$ more ice cream sales $\rightarrow$ more shark attacks
- Confounder: Temperature
- Multiple regression model of shark attacks with ice cream sales *and* temperature: Ice cream sales no longer influence shark attacks, but air temperature does

Is there a relationship between predictors and response variable?

- Multiple linear regression with p predictors: *All* regression coefficients except β_0 are zero (no variable has influence):

$$\beta_1 = \beta_2 = \ldots = \beta_p = 0$$

- Null hypothesis:

$$H_0 : \quad \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

- Alternative hypothesis:

$$H_A : \quad \text{At least one } \beta_i \text{ is not equal to } 0$$

- Calculation of *F statistics* with p-value

Example

- *p*-value for multiple linear model for dataset **Advertising**:

```
summary(lm(Sales ~ TV + radio + newspaper, data = adv))

##
## Call:
## lm(formula = Sales ~ TV + radio + newspaper, data = adv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

- `R`-output `p-value` in line for F-statistic: p-value for multiple linear model practically zero
- Very convincing hint: At least one predictor is responsible for an increase in `Sales` with increased advertising expenses

Example

- Why don't we just look at individual p-values?
- If one is below significance level, then we know that at least this variable has an influence
- But: Because of principle of hypothesis testing, statistically significant p-value is randomly erroneous
- Following example: No variable is significant
- All β_1-values near 0
- But: Gives random deviations where corresponding p-values becoming significant
- Therefore: If there are many variables, one is almost always significant, although in reality there are not

- Code:

```r
set.seed(4)
v <- 20
d <- 500

df <- matrix(rnorm(v * d), nrow = d)
# head(df)
df <- data.frame(df)

Y <- rnorm(d)
# Y

df$Y <- Y

fit <- lm(Y ~ ., , data = df)
summary(fit)
```

- Output:

```
##
## Call:
## lm(formula = Y ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62976 -0.66857  0.00927  0.64462  2.81840
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.029669   0.047272  -0.628   0.5305
## X1          -0.010970   0.048886  -0.224   0.8225
## X2          -0.036943   0.049150  -0.752   0.4526
## X3          -0.005961   0.047734  -0.125   0.9007
## X4          -0.018073   0.047726  -0.379   0.7051
## X5           0.005827   0.048524   0.120   0.9045
## X6          -0.127798   0.049554  -2.579   0.0102 *
## X7          -0.052386   0.049816  -1.052   0.2935
## X8           0.020574   0.048557   0.424   0.6720
## X9          -0.015178   0.047941  -0.317   0.7517
## X10         -0.015107   0.046988  -0.322   0.7480
## X11          0.005580   0.046517   0.120   0.9046
## X12         -0.004676   0.046583  -0.100   0.9201
## X13         -0.021652   0.049114  -0.441   0.6595
## X14         -0.093800   0.046075  -2.036   0.0423 *
## X15          0.019740   0.047451   0.416   0.6776
## X16          0.042796   0.045267   0.945   0.3449
## X17         -0.074511   0.049061  -1.519   0.1295
## X18          0.041733   0.047568   0.877   0.3808
## X19         -0.078238   0.047492  -1.647   0.1001
## X20         -0.057475   0.048156  -1.194   0.2333
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.042 on 479 degrees of freedom
```

Determination of important predictors

- First: Do predictors have any influence on response variable?
- Decision: With help of F statistics and corresponding p value
- If at least one variable influence response variable (null hypothesis rejected): *Which* predictors are these?
- Can view individual p-values as in table

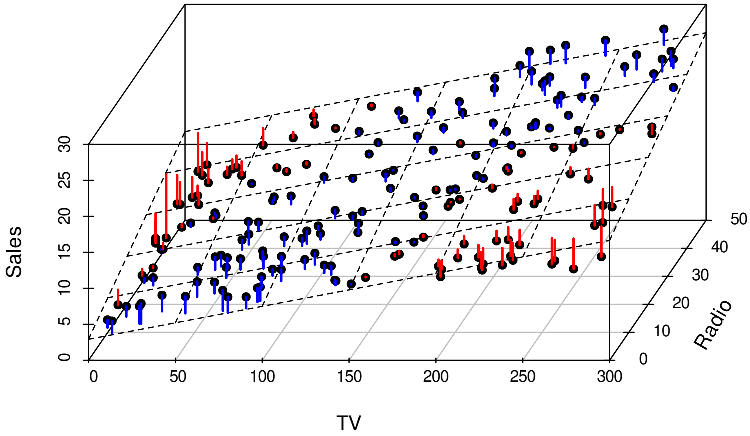
- Possible: All predictors influence response variable, but usually only a few
- Goal: Determine variables and then set up a model containing only these variables
- Interested in simplest possible model that fits data: Easier to interpret
- Which variables are important?
- Procedure: *Variable selection* (see lecture notes)

How well does model fits the data?

- Measure of determination R^2
- Dataset `Advertising` is the R^2 value 0.8972
- R^2 increases the more predictors are considered

No linear regression

- Graphical overview: Show problems with model that are invisible to numerical values:



- Three-dimensional scatter plot: Only `TV` and `Radio` taken into account
- Dotted lines: Regression plane
- Observation: Values of plane too large if advertising expenses was spent exclusively on either `TV` or `Radio`
- Back left: Advertising only for `Radio`
- Front right: Only for `TV`
- Values of plane are too low if advertising expenses is distributed equally between `TV` and `Radio`
- Nonlinear pattern: Cannot be accurately described by a linear regression
- Plot indicates *interaction* or *synergy effect*: Larger sales if advertising expenses is divided

Cancellation of assumption regarding additivity

- Interaction effects
- Example advertising:

```
fit <- lm(Sales ~ TV + Radio + TV * Radio)

summary(fit)

##
## Call:
## lm(formula = Sales ~ TV + Radio + TV * Radio)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3366 -0.4028  0.1831  0.5948  1.5246
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
## TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
## Radio       2.886e-02  8.905e-03   3.241   0.0014 **
## TV:Radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9435 on 196 degrees of freedom
## Multiple R-squared:  0.9678,Adjusted R-squared:  0.9673
## F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

- p-values for TV, Radio and interaction term TV $\cdot$ Radio: Statistically significant
- Seems clear: All these variables should be included in model
- Possible: p value for interaction term is very small, but p values of main effects (here TV and Radio) are not