JACOBS
UNIVERSITY

JTME-990203 –Statistical Modeling with R

Dr. Adalbert Wilhelm

Research IV, Room 111,

phone - 3402,

e-mail: a.wilhelm@jacobs-university.de

Office hours: Tuesday, 15:30-17:00, plus open door / by appointment

Statistics and Data Science

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

SIGN IN TO RECOMMEND
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

Enlarge This Image

Thor Swift for The New York Times
Carrie Grimes, senior staff engineer at Google, uses statistical analysis of data to help improve the company's search engine.

"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for dronish number nerds. They are finding themselves increasingly in demand — and even cool.

"I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. "And I'm not kidding."

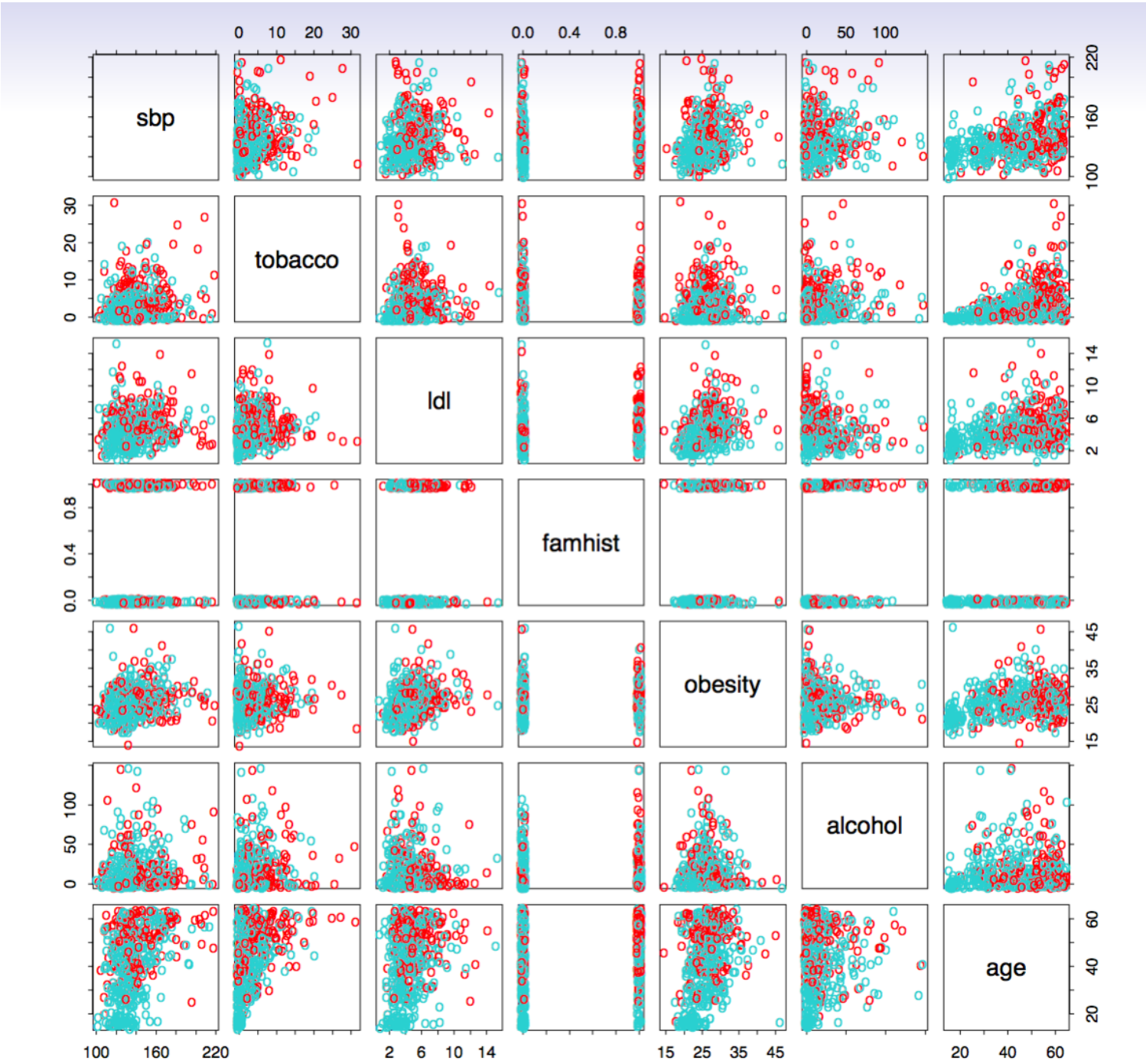
Multimedia

Quote of the Day, New York Times, August 5, 2009

"I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding." — HAL VARIAN, chief economist at Google.

Statistical Modeling = Learning from data

- Predict whether a patient who has been hospitalized due to a heart attack, will have a second heart attack. Using demographic, diet, and clinical measurements for that patient
- Predict the price of a stock in six months from now, on the basis of company performance measures, economic data and historic development
- Customize an email spam detection system
- Identify handwritten numbers and letters using a digitized image
- Establish the relationship between salary and demographic variables in population survey data.
- Segment customers to do targeted marketing based on demographics, past purchases, preferences, interests, …
- Group grocery items together in order to determine the store layout, special offerings, etc. using market basket data



Example: Heart attacks

Example: Spam Detection

- data from 4601 emails sent to an individual (named George, at HP labs, before 2000). Each is labeled as spam or email.
- goal: build a customized spam filter.
- input features: relative frequencies of 57 of the most commonly occurring words and punctuation marks in these email messages.

	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between spam *and* email.


Example: Wages
Wage
50
100
200
300
20
40
60
80
Age
Wage
50
100
200
300
2003
2006
2009
Year
Wage
50
100
200
300
1
2
3
4
5
Education Level

Statistical modeling and The Supervised Learning Problem

Starting point:

- Outcome measurement Y (also called dependent variable, response, target).
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables).
- In the regression problem, Y is quantitative (e.g price, blood pressure).
- In the classification problem, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
- We have training data $(x_1,y_1),\ldots,(x_N,y_N)$. These are observations (examples, instances) of these measurements.

The Supervised Learning Problem

Objectives:

- On the basis of the training data we would like to:
  - Accurately predict unseen test cases.
  - Understand which inputs affect the outcome, and how.
  - Assess the quality of our predictions and inferences.

Philosophy

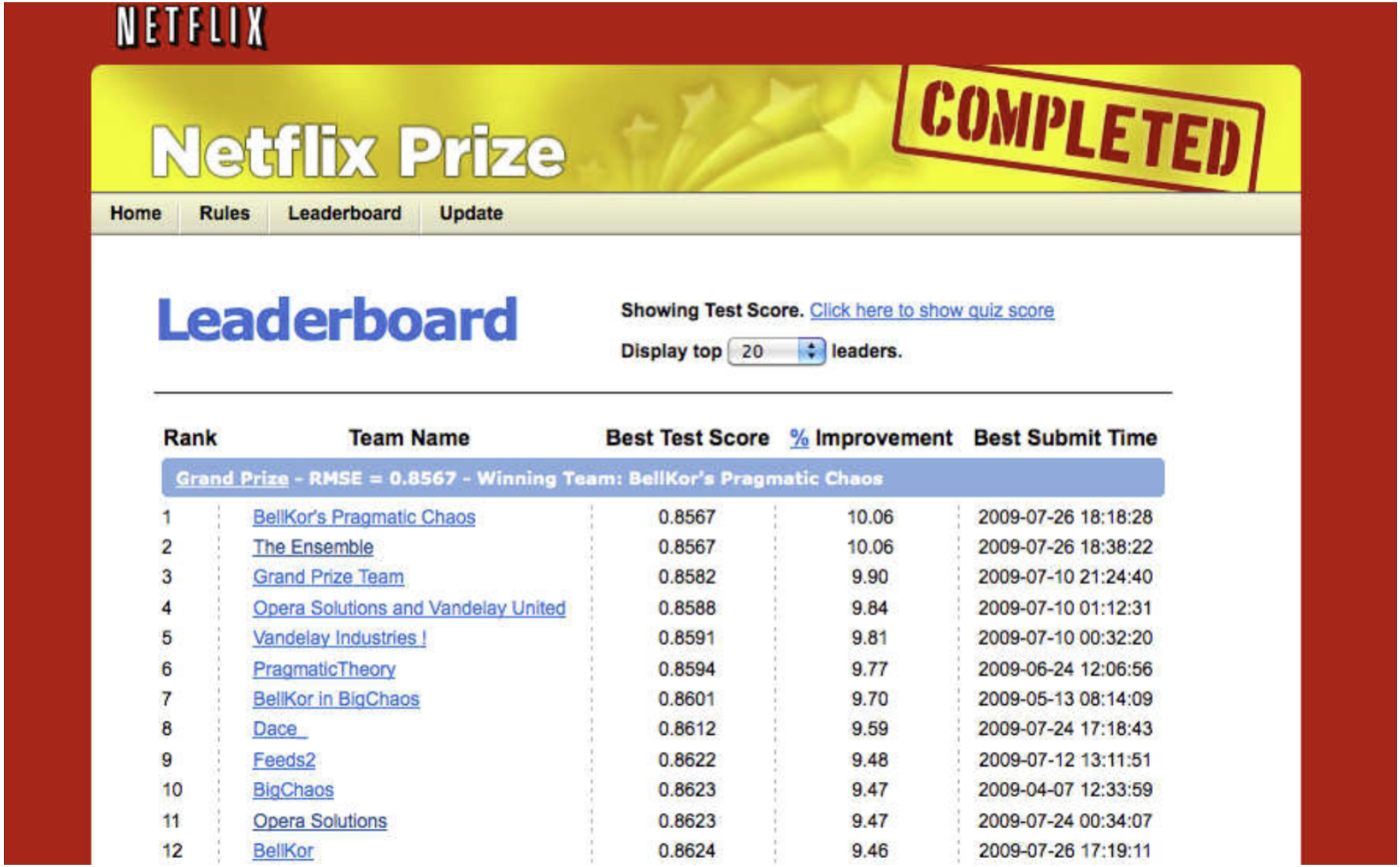
- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working [simpler methods often perform as well as fancier ones!]
- This is an exciting research area, having important applications in science, industry and finance.
- Statistical learning is a fundamental ingredient in the training of a modern data scientist.

Unsupervised Learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective is more fuzzy — find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- difficult to know how well your are doing.
- different from supervised learning, but can be useful as a pre-processing step for supervised learning.

The Netflix Prize

- competition started in October 2006. Training data is ratings for 18, 000 movies by 400, 000 Netflix customers, each rating between 1 and 5.
- training data is very sparse— about 98% missing.
- objective is to predict the rating for a set of 1 million customer-movie pairs that are missing in the training data.
- Netflix's original algorithm achieved a root MSE of 0.953. The first team to achieve a 10% improvement wins one million dollars.
- is this a supervised or unsupervised problem?



- BelKor's Pragmatic Chaos wins by a small margin over The Ensemble.

Statistical Learning vs. Machine Learning

- Machine learning arose as a subfield of Artificial Intelligence.
- Statistical learning arose as a subfield of Statistics.
- There is much overlap — both fields focus on supervised and unsupervised problems:
- • Machine learning has a greater emphasis on large scale applications and prediction accuracy.
- • Statistical learning emphasizes models and their interpretability, and precision and uncertainty.
- But the distinction has become more and more blurred, and there is a great deal of “cross-fertilization”.
- Machine learning has the upper hand in Marketing!

Schedule

Session	Topic	Literature / Textbook Chapter
1	Review of Fundamtental Statistical Concepts	Gordon Ch. 1&2
2	Multivariate Linear Regression I	Field et al. (2012), Ch. 7
3	Multivariate Linear Regression II	Field et al. (2012), Ch. 7
4	Analysis of Variance	Field et al. (2012), Ch. 10
5	Analysis of Covariance	Field et al. (2012), Ch. 11
6	Mixed designs	Field et al. (2012), Ch. 14
7	Generalized linear models: logistic regression	Field et al. (2012), Ch. 8
8	Multinomial logistic regression	Field et al. (2012), Ch. 8
9	Generalized linear models: Poisson regression	Field et al. (2012), Ch. 18
10	Exploratory Factor Analysis I	Field et al. (2012), Ch. 17
11	Exploratory Factor Analysis II	Field et al. (2012), Ch. 17
12	Repeated measures and panel data	Field et al. (2012), Ch. 13
13	Multilevel models I	Field et al. (2012), Ch. 19
14	Multilevel models II	Field et al. (2012), Ch. 19
	Final exam date to be determined.	

Tutorials

Given by TAs: Maria Ilie & Lilian Lommel

Time and Venue: TBD

Will provide some help and guidance in using R, support for the weekly exercises plus additional support to master content of class

Webpage on Teamwork:

https://teamwork.jacobs-university.de:8443/confluence/x/ggDbB

There will be weekly homework assignments!

Assessment

Weekly homework assignments: total of 6, parts of the assignment are to be graded, 8 points per week, a maximum of 40 points can be gained, i.e. the best five homework assignments will be counted.

Homework assignments can be done in groups of up to 3 students: stable group formation over the semester.

Homework assignments count for 40 %.

Final exam: counts for 60%.

Comprises practical component working with R!

will take place on October 11, 2016:

Group 1: 17:15 – 18:00, Group 2: 18:15 – 19:00 (random assignment)

- Questions with short, plain-text answers
- Answers require computations using R/Rcmdr/R Studio

Written exam part

- October 20, 2016: 13:00 – 14:00
- mix of questions with short, plain-text answers and some multiple-choice questions

Psychology Experiment Participation

- You can earn bonus points by participating in psychology experiments/surveys
- For each hour of experiment participation you can earn 1% bonus to your total grade
- A maximum of 5 % can be earned (i.e. a total of 5 hours participation)
- Durations of experiments will be rewarded proportionally (in steps of quarter of an hour)
- You are in charge of keeping track of your experiment participation!
- Participation sheet will be uploaded on campus net!
- Signature of experimenter is needed!
- Participation sheets must be handed in by October 31, 2016.

- **EXPERIMENT #**
- Name of experimenter:_____
  Signature of experimenter:_____
- Title of study:_____
- Date of participation _____ Duration (in hours): _____

Make-up Policy

- 1st Make-up for Practical component of Final exam:
  - October 19, 2016: 13:00 – 13:45
- 2nd Make-up for Practical component of Final exam:
  - November 08, 2016: 16:00 – 16:45
- Make-up for written part of Final Exam:
  - November 08, 2016: 17:00 – 18:00

Learning material

No designated Textbook:

Primary Reading: Andy Field, Jeremy Miles, and Zoe Field: *Discovering Statistics Using R*. Sage, London, 2012.

Accompanying webpage for this book

YouTube videos with Andy Field

Almost any other textbook and many internet resources.

Cartoon Guide for Statistics by Larry Gonick and Woollcott Smith

Teamwork page with R tutorials and glossary (H. Nida, Class of 2015)

Campusnet: I will upload each week: (please remind me in case I forget)

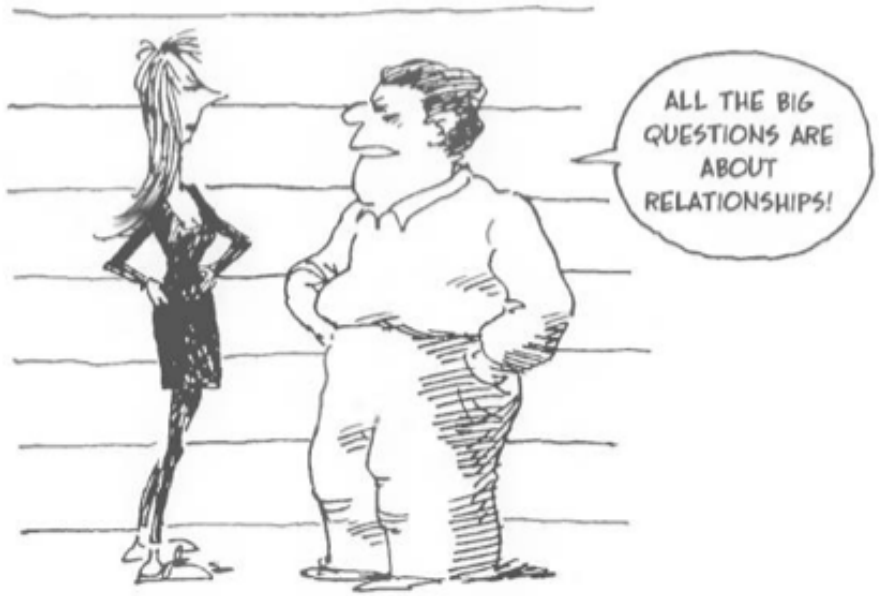
- lecture slides
- R code used in lecture
- pdf with (commented) R output
- homework assignments
- solution of homework assignments
- R script for homework assignment solution

Audit requirements

- For a successful audit, you have to participate in at least ten lectures.
- Audits, please collect my signature after each lecture

Statistical models

- Statistical models
  - The mean as a model – “naïve model”
  - Regression analysis



©The Cartoon Guide to Statistics

Statistical models

The general idea of data summary and statistical modeling is to reduce complexity.

The most relevant formula in statistics that you have to remember is the following:

$$\text{response}_i = \text{model} + \text{error}_i$$

In words: the individual responses (measurements, scores, outcomes) are the sum of a model and individual error terms.

Naïve model or NULL model

Now the question is, how can we develop resonable models.

In general, there is always the most simplistic model, often called the naive model. The naive model completely ignores systematic differences between employees and hence uses a measure of central tendency for modeling the salary and puts all variability into the error component.

$$\text{response}_i = \text{mean} + \text{error}_i$$

Example: Harris Trust & Savings Bank

In 1965 and 1968 the US government issued executive orders and regulations prohibiting discrimination against minorities and women by Government contractors. On the basis of these regulations, the Department of Treasury filed a complaint charging the Bank with violations of the Executive Order. The Government's complaint charged Harris with engaging in various employment practices which discriminated against women and minorities and with failing to take affirmative action to eradicate the present effects of past discrimination. The first hearing occurred in 1979, but the trial lasted through several re-openings until 1986. As it read in the file about the trial against Harris, both parties brought forth statistical as well as testimonial evidence. In the course of the hearings, each party provided studies by different statisticians as circumstantial evidence. Mainly the studies were based on statistical methods such as regression models and comparison of means. Since those studies lead to different results, the parties continuously challenged each other's statistics as suffering from coding errors, mischaracterizations of employees, and incomplete data.

Roberts, H.V. (1985), Koellner et al. (2002)

HARRIS BANK®

- Data set:
  - N = 474,
  - Salary in 1977
  - Salary at time of hire
  - Age
  - Seniority (time since first hired by Harris)
  - Work Experience (prior to hire by Harris)
  - Education level (years of education)
  - Job category
  - Minority (Race)
  - Sex
- Goal: Determine whether salary in 1977 was systematically lower for females and ethnic minorities controlling for education, work experience, etc.
- Method: Comparing means

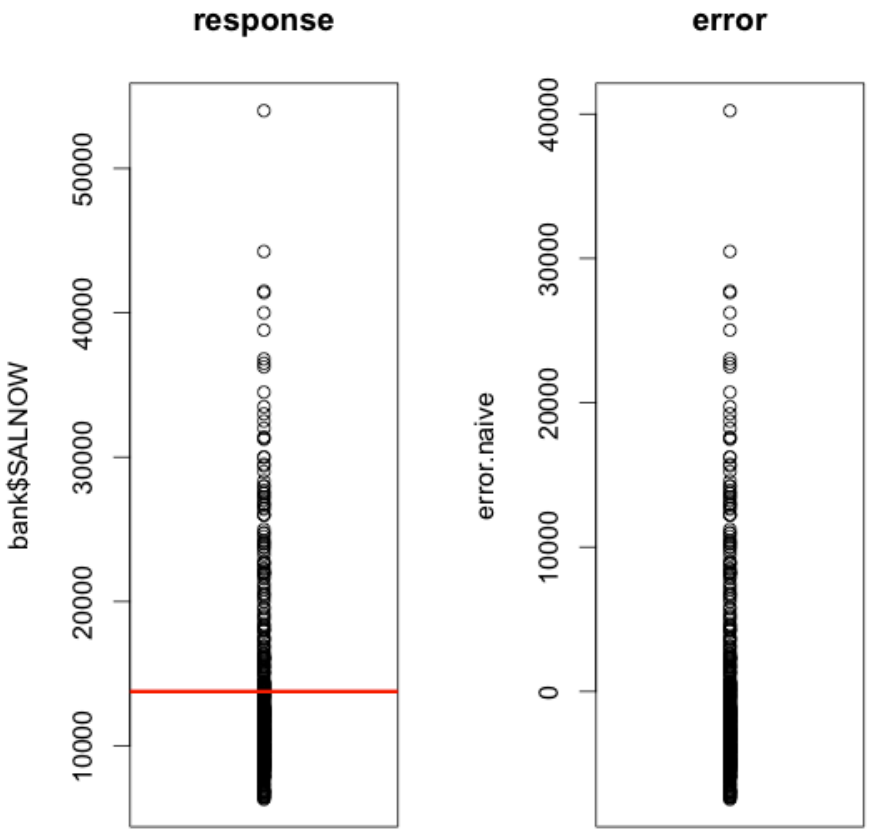
Null model for Harris Bank Data

For our specific example of bank salaries this means

$$\text{salary}_i = \text{mean}(\text{salary}) + \text{error}_i$$

$\text{Salary}_i = 13767.83 + \text{error}_i$

mean	13767.83
standard deviation	6830.27



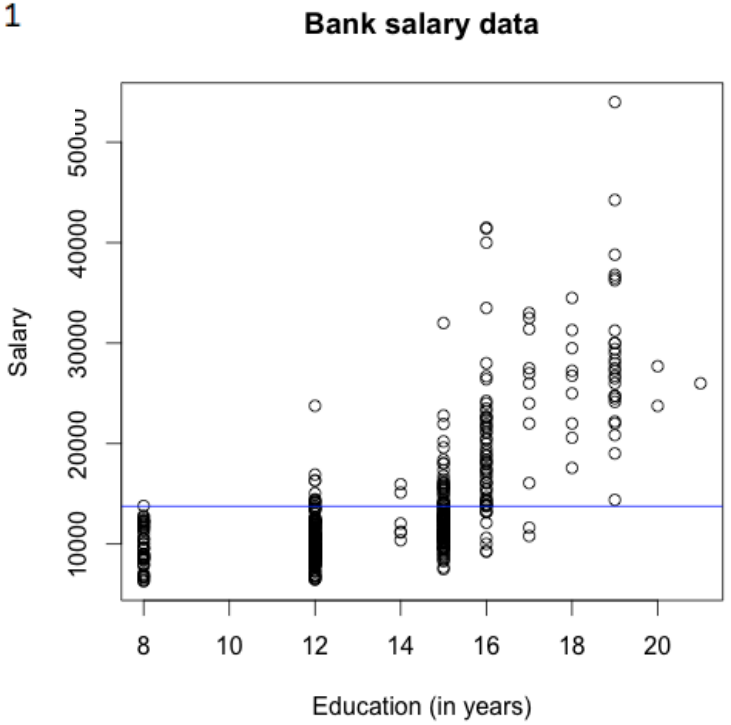
The mean as regression model

```
Call:
lm(formula = SALNOW ~ 1, data = bank)

Residuals:
   Min     1Q Median     3Q    Max
 -7468  -4168  -2218   1007  40232

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13767.8      313.7   43.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6830 on 473 degrees of freedom
```

Estimated intercept in the naïve model
= mean of response

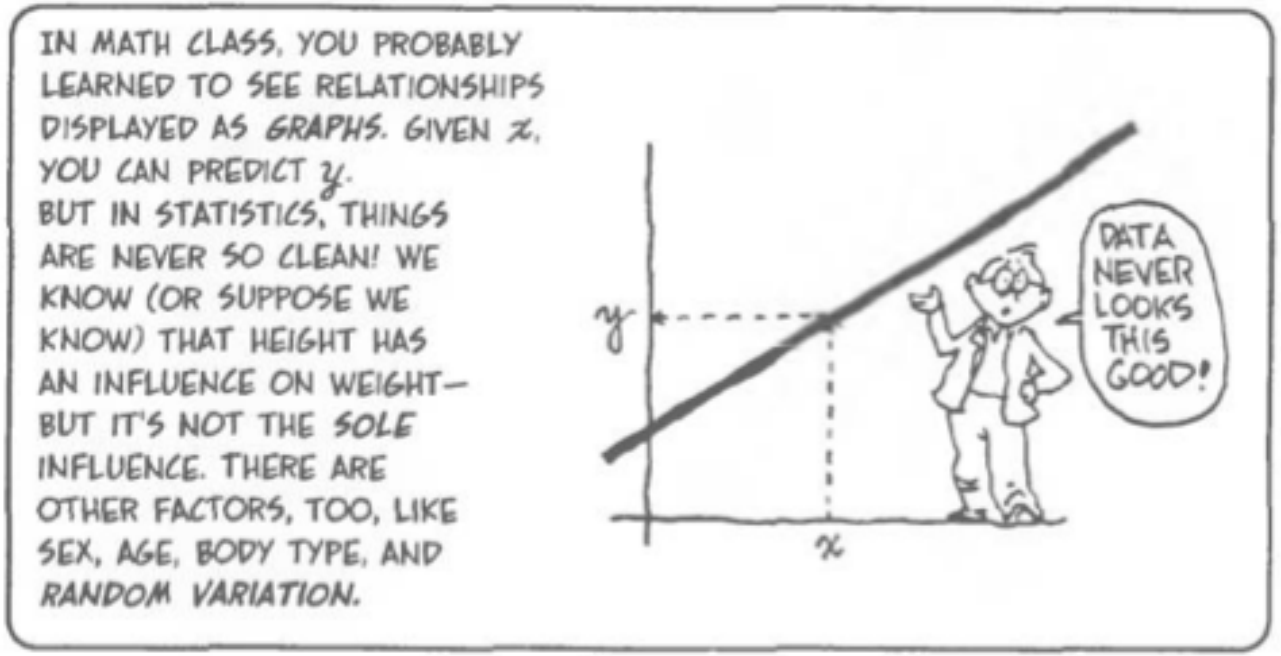
Residual standard error in the naïve model
= standard deviation of response

Standard error of intercept
= residual standard error / sqrt(n)
(sample size n=474)



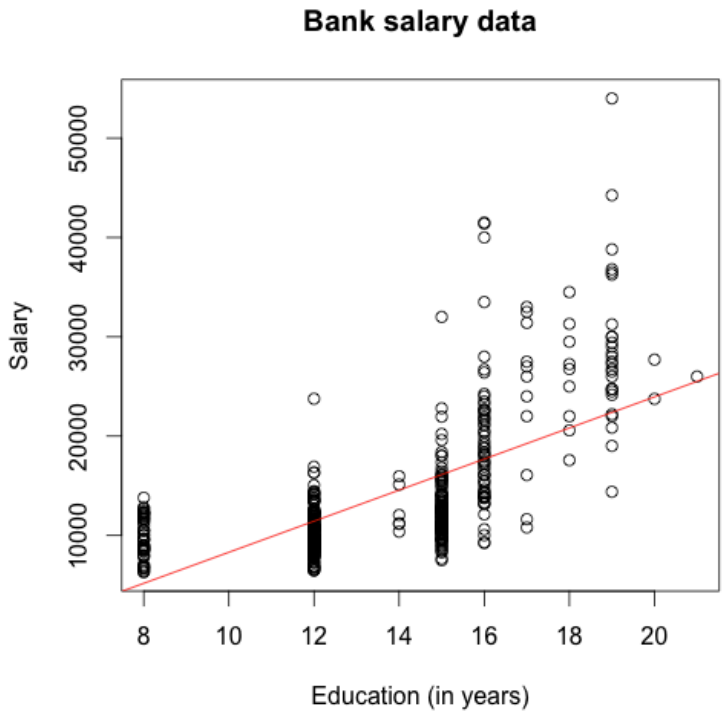
Terminology

Dependent Variable	Independent Variable
Output	Input
Response	Predictor
Regressand	Regressor
Y	X
	Explanatory variable



Regression of Salary on Education

- Statistical model: straight line
- Salary = - 7332.47 + 1563.96 * Education
- Interpretation:
  - each year of education brings 1563.96 units more income
  - Zero education, you have to pay for being allowed to work (unrealistic)
  - You need at least five years of education to get paid for your work (more realistic)
- Questions?
  - Is this model better than the naïve model?
  - How can we test this?
  - How good is the model in general?
  - Do we have a benchmark?
  - Does predictor have an impact on response?



Regression of Salary on Education

- Linear regression output
- lm(SALNOW ~EDLEVEL, data=bank)

```
Call:
lm(formula = SALNOW ~ EDLEVEL, data = bank)

Residuals:
   Min     1Q Median     3Q    Max
 -8627  -3284  -1001   2351  31617

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7332.47    1128.76  -6.496  2.1e-10 ***
EDLEVEL      1563.96      81.82  19.115  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5133 on 472 degrees of freedom
Multiple R-squared: 0.4363, Adjusted R-squared: 0.4351
F-statistic: 365.4 on 1 and 472 DF,  p-value: < 2.2e-16
```

intercept
slope

Next to them their
standard errors

p-values
t-values to the left

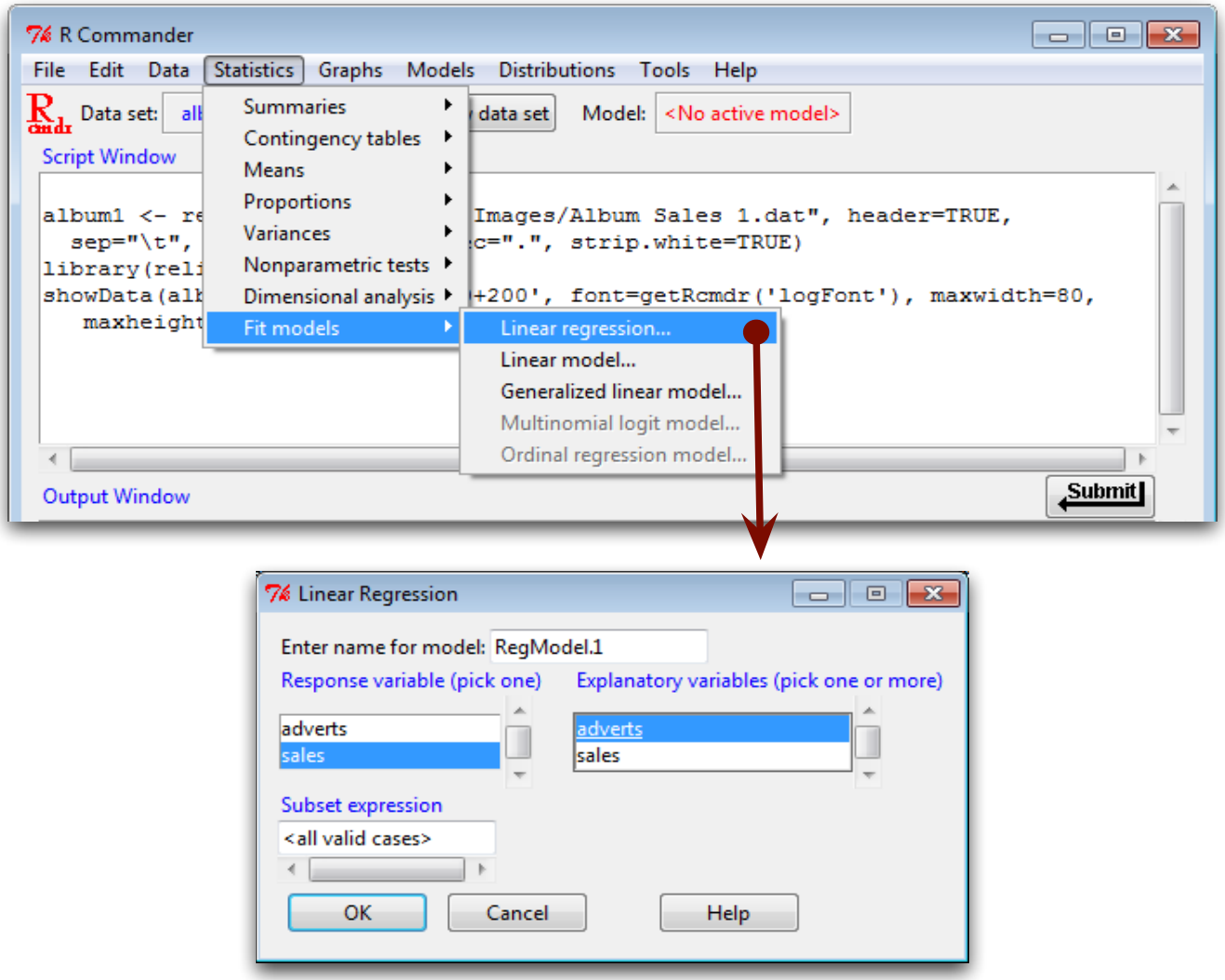
Model fit

comparison to null (naïve) model

Reading a regression output from R

- We look at:
  - p-value (Pr(>|t|)): to decide whether impact of predictor is statistically significant (standard comparison $p<0.05$?)
  - t-value: tells the same story as p-value (rule of thumb comparison, $|t| > 2$, for large samples you can even go for 1.96)
    - t = estimate / s.e.(estimate)
  - estimates: unstandardized coefficients (i.e. regression coefficients aka intercept and slope)
  - Adjusted R-squared to see how good the model fits the data ($0 \leq$ adj. R-squared ≤ 1)
  - Multiple R-squared is called the coefficient of determination and is the square of the correlation coefficient between the observed response and the fitted response

Simple linear regression using R commander



The t-test as regression model:

Gender difference for Harris Bank Data

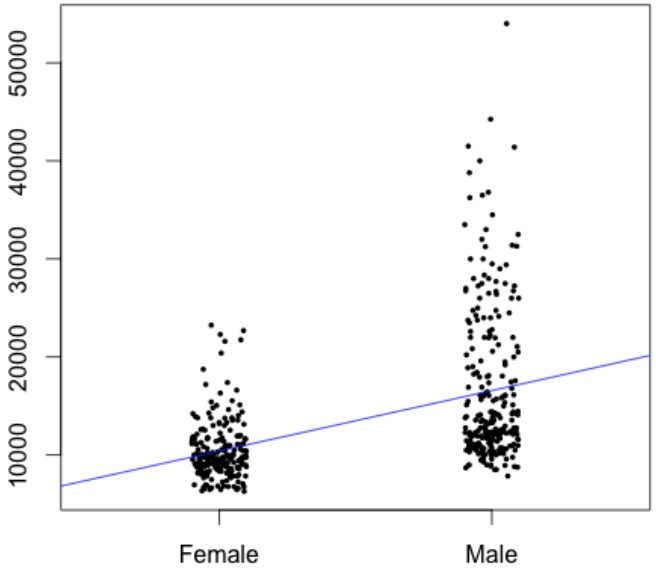
	Male	Female
Mean	16576.71	10412.77
Standard deviation	7799.69	3023.21

```
 Two Sample t-test

data:  bank$SALNOW by bank$SEX
t = -10.9452, df = 472, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7270.561 -5057.329
sample estimates:
mean in group Female   mean in group Male
            10412.77             16576.71
```

The t-test as regression model:

Gender difference for Harris Bank Data

```
 Two Sample t-test

data:  bank$SALNOW by bank$SEX
t = -10.9452, df = 472, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7270.561 -5057.329
sample estimates:
mean in group Female   mean in group Male
            10412.77             16576.71


Call:
lm(formula = SALNOW ~ SEX, data = bank)

Residuals:
   Min     1Q Median     3Q    Max
 -8717  -4022  -1293   1479  37423

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10412.8      415.5   25.06   <2e-16 ***
SEXMale       6163.9      563.2   10.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6106 on 472 degrees of freedom
Multiple R-squared: 0.2024, Adjusted R-squared: 0.2007
F-statistic: 119.8 on 1 and 472 DF,  p-value: < 2.2e-16
```



The t-test as regression model:

Gender difference for Harris Bank Data

Changing reference category:

```
bank$SEX <- relevel(bank$SEX,ref="Male")

bank.gender.f<-lm(SALNOW~SEX, data=bank)

summary(bank.gender.f)
```

```
Call:
lm(formula = SALNOW ~ SEX, data = bank)

Residuals:
   Min     1Q Median     3Q    Max
 -8717  -4022  -1293   1479  37423

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16576.7      380.2   43.60   <2e-16 ***
SEXFemale    -6163.9      563.2  -10.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6106 on 472 degrees of freedom
Multiple R-squared:  0.2024,    Adjusted R-squared:  0.2007
F-statistic: 119.8 on 1 and 472 DF,  p-value: < 2.2e-16
```

ref = “Male”

```
Call:
lm(formula = SALNOW ~ SEX, data = bank)

Residuals:
   Min     1Q Median     3Q    Max
 -8717  -4022  -1293   1479  37423

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10412.8      415.5   25.06   <2e-16 ***
SEXMale       6163.9      563.2   10.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6106 on 472 degrees of freedom
Multiple R-squared: 0.2024, Adjusted R-squared: 0.2007
F-statistic: 119.8 on 1 and 472 DF,  p-value: < 2.2e-16
```

ref = “Female”

Models are identical!!

Hypothesis tests

- To decide whether predictor is related to response we use hypothesis tests to check whether the estimate for the regression coefficient (slope) differs from 0 given the “natural fluctuation” of the estimate as measured by its standard error.
- This is a decision under uncertainty, since we want to infer from the sample to the whole population
- To assess our confidence in the correctness of the decision we look at likelihood of taking a wrong decision under the assumption that regression coefficient actually equals 0.
- This likelihood can either be derived via theoretical considerations (distributional assumption for population, making use of Central Limit Theorem, direct derivation from test statistic) or via resampling.
- Null hypothesis: slope = 0
- Alternative hypothesis: slope ≠ 0

Assumptions

- Theoretical derivations above are based on some assumptions
  - Linearity of relationship
  - Homoscedasticity, i.e. homogeneity of variance
  - Normality of residuals
    - Mean = 0
    - Constant variance (= homoscedasticity)
    - Symmetry
  - Independence of observations (cases)

- Various graphical and numerical tests and checks for this
  - E.g. Q-Q Plot (for normality)
  - Levene's test (for homoscedasticity)
  - Kolomogorov-Smirnov-Test (for normality)
  - Shapiro-Wilks-test (for normality)
- Transformations to meet assumptions

Transformations

- Logarithmic transformation
  - reduces skew
  - linearizes
  - widely used for monetary/economic variables such as GDP etc.

- Square root transformation
  - reduces skew
  - stabilizes variance
  - linearizes

- Reciprocal transformation
  - reduces skew
  - linearizes

- Be aware that you then compare on a different scale!
  - E.g. geometric mean vs. arithmetic mean
  - multiplicative vs. additive effects
  - For monotonic transformations general interpretations remain the same

The Least Squares Principle

a	intercept, constant term
b	slope
$\hat{y}_i$	fitted value, fit
e_i	residual

$$y_i = a + bx_i + e_i \qquad \text{all observed points}$$

$$\hat{y}_i = a + bx_i \qquad \text{all fitted points}$$

We have

$$e_i = y_i - \hat{y}_i.$$

The Least Squares Principle

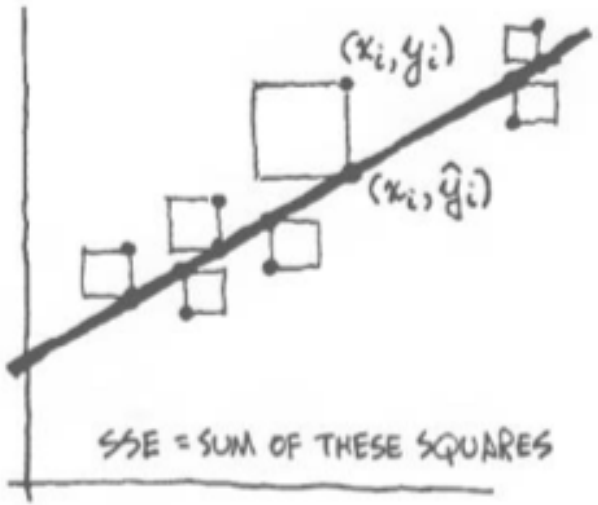
Least-Squares-Principle: Choose a and b to minimize sum of squared residuals, i.e. $\sum_{i=1}^{n} e_i^2$.

LS-Principle

$$\min_{a,b} \sum_{i=1}^{n} e_i^2 =$$

$$\min_{a,b} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 =$$

$$\min_{a,b} \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$



The Least Squares Principle

Solutions for Least-squares-problem:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{COV[X,Y]}{VAR[X]} = r\frac{s_y}{s_x}$$

$$a = \bar{y} - b\bar{x}$$

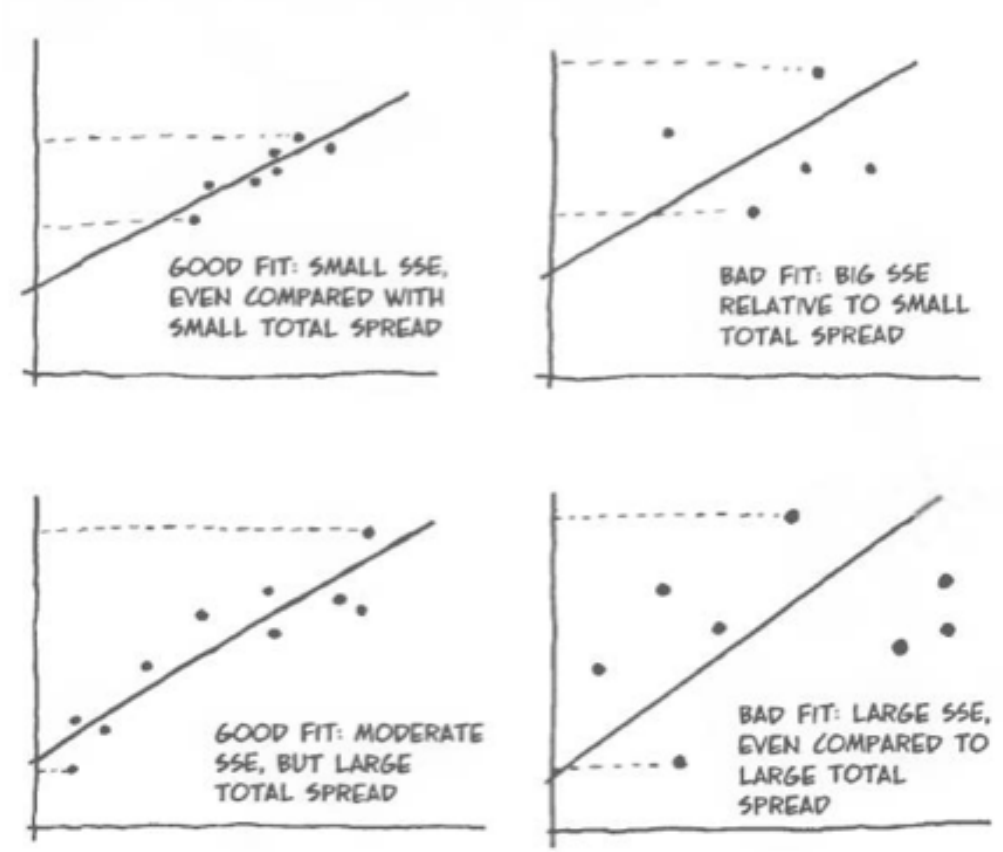
Regression line always runs through $(\bar{x}, \bar{y})$.

Linear Regression: Good model fit

- Assumptions should be met
- residuals should be small

$$y = a + bx + e$$

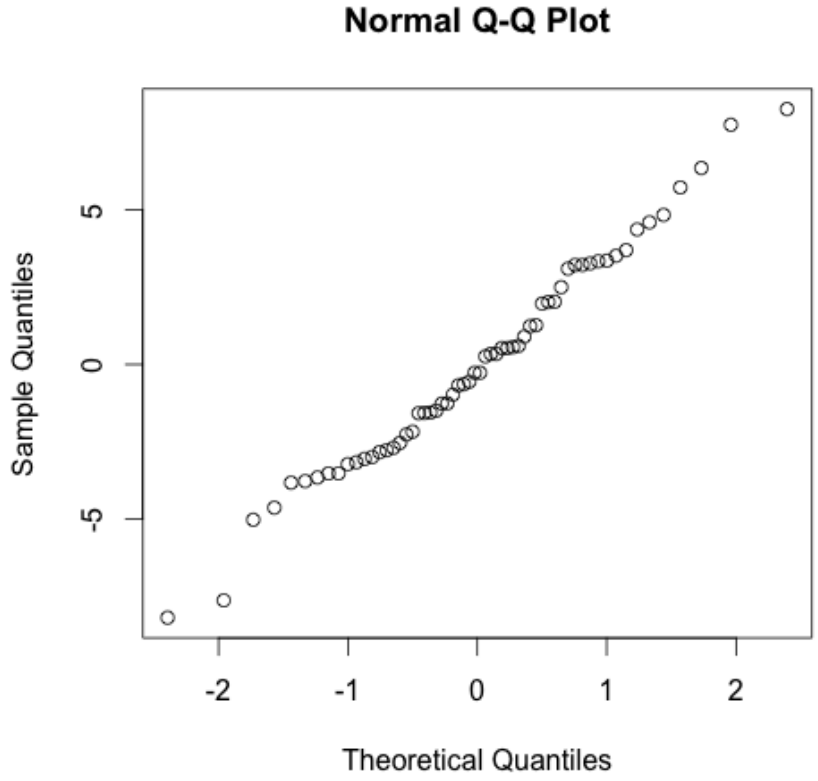
$$\texttt{observed} = \texttt{fitted} + \texttt{residual}$$



©The Cartoon Guide to Statistics

Model quality: Checking Assumptions

- Normality
- Q-Q plot of residuals



Model quality: Variance explained

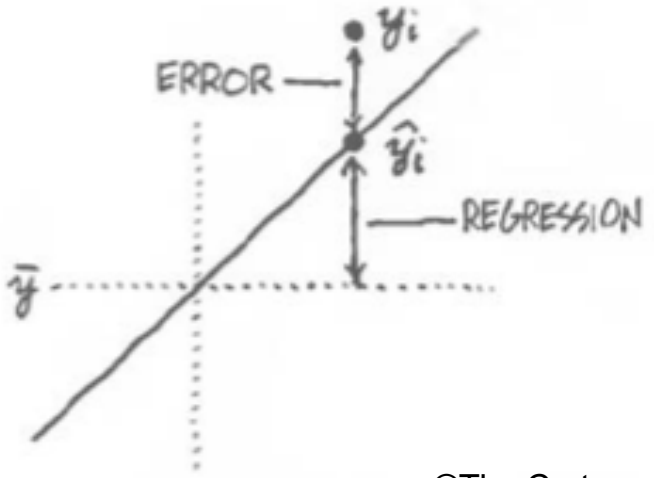
R^2 : general measure to assess quality of a model:

$$R^2 = \frac{VAR[Y] - VAR[e]}{VAR[Y]}$$

- proportion of variability explained by linear regression
- squared correlation coefficient

"R^2 adjusted" corrects R^2 by the number of parameters estimated, thus, also takes complexity of a model into account

Model quality: Variance explained



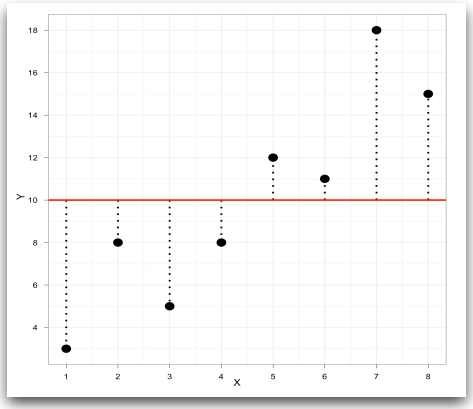
SOURCE OF VARIABILITY	SUM OF SQUARES
REGRESSION	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
ERROR	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
TOTAL	$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

Total Sum of Squares = Sum of Squares due to Regression

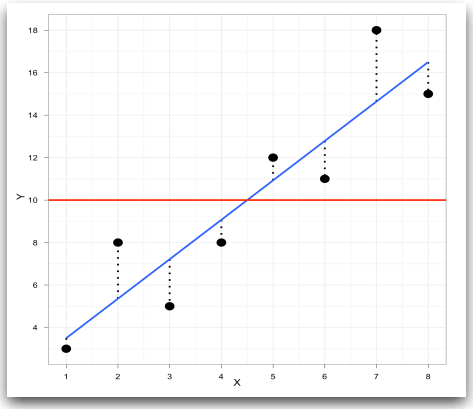
+

Sum of Squares due to error

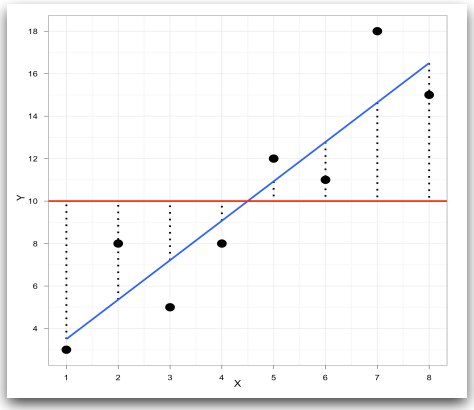
Model quality: Variance explained



SS_T uses the differences between the observed data and the mean value of Y



SS_R uses the differences between the observed data and the regression line



SS_M uses the differences between the mean value of Y and the regression line

Total Sum of Squares

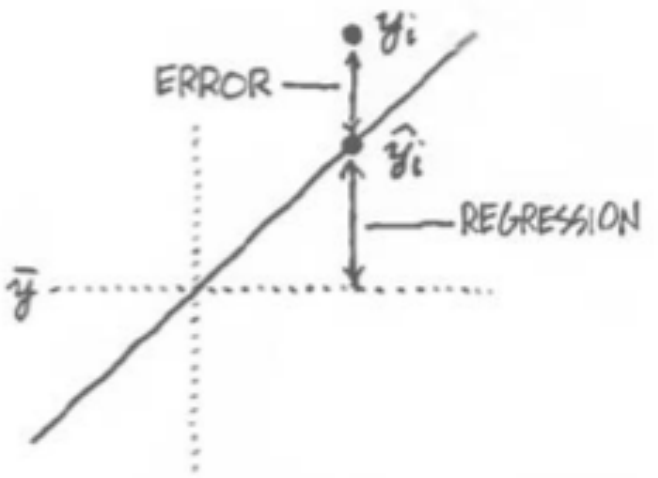
=

Sum of Squares due to Regression

+

Sum of Squares due to error

Model quality: Variance explained



SOURCE OF VARIABILITY	SUM OF SQUARES
REGRESSION	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
ERROR	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
TOTAL	$SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

©The Cartoon Guide to Statistics

standard error of the estimate:

$$s_{\hat{y}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Also called: residual standard error

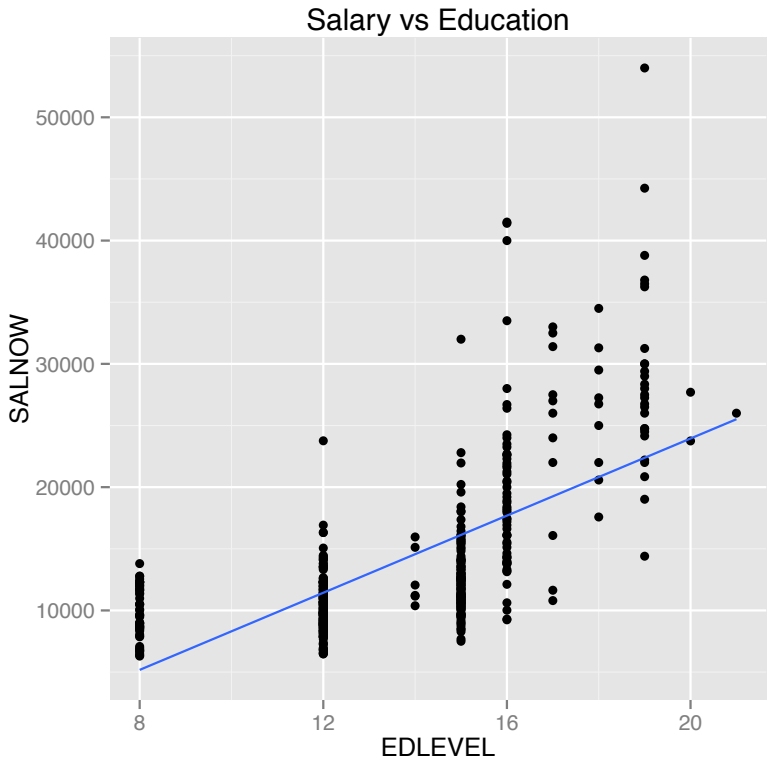
Multiple Regression

We are familiar with the simple linear regression model:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

$$i = 1, \ldots, n,$$

But, usually, there is more than just one cause!



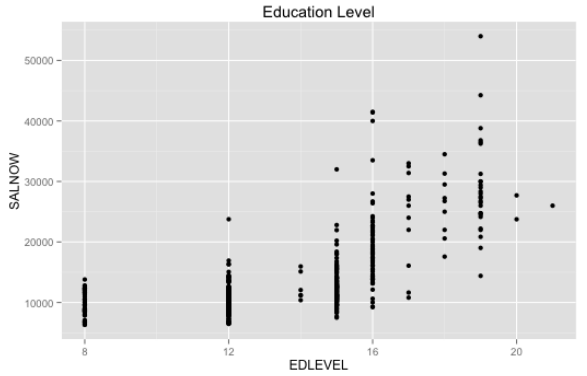
HARRIS BANK®

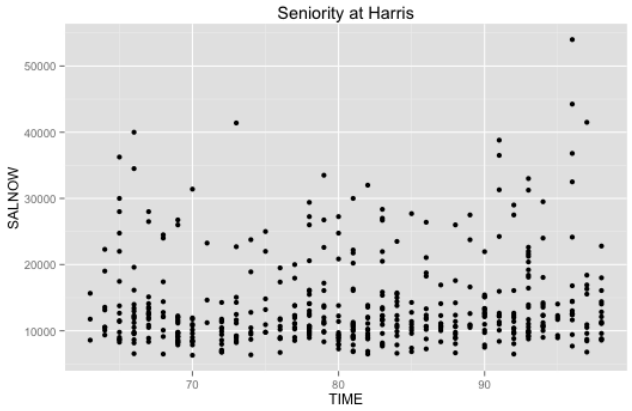
- Data set:
  - N = 474,
  - Salary in 1977
  - Salary at time of hire
  - Age
  - Seniority (time since first hired by Harris, measured in months)
  - Work Experience (prior to hire by Harris, measured in years)
  - Education level (years of education)
  - Job category
  - Minority (Race)
  - Sex
- Goal: Determine whether salary in 1977 was systematically lower for females and ethnic minorities controlling for education, work experience, etc.
- Method: Multiple linear regression

Example: Harris Bank Data

- Y: SALNOW (salary in 1977)
- X_1: education level (years of education)
- X_2: gender
- X_3: seniority (time since first hired by Harris)

```
Call:
lm(formula = SALNOW ~ EDLEVEL + SEX + TIME, data = bank)

Residuals:
    Min      1Q  Median      3Q     Max
-8931.6 -3084.6  -734.9  2238.2 30840.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -8563.93    2079.29  -4.119 4.50e-05 ***
EDLEVEL      1354.05      83.42  16.232  < 2e-16 ***
SEXMale      3337.63     483.22   6.907 1.61e-11 ***
TIME           27.70      22.40   1.237    0.217
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4889 on 470 degrees of freedom
Multiple R-squared:  0.4908,    Adjusted R-squared:  0.4876
F-statistic:   151 on 3 and 470 DF,  p-value: < 2.2e-16
```

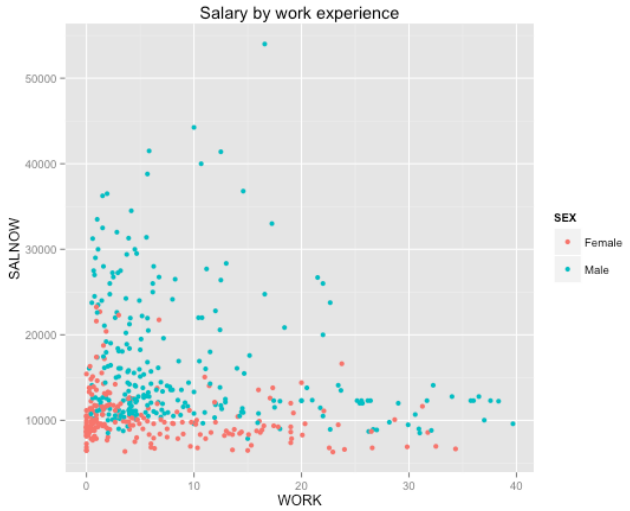




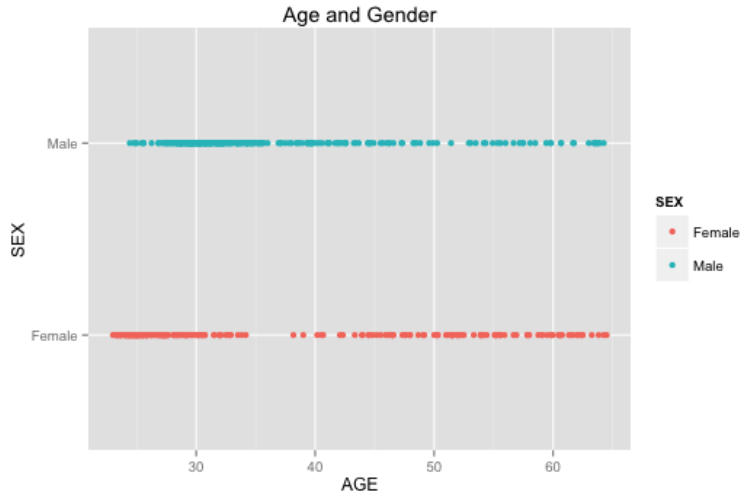
Example: Harris Bank Data

- *Y*: SALNOW (salary in 1977)
- X_2: gender
- X_4: Age
- X_5: Work Experience (prior to hire by Harris, measured in years)

```
Call:
lm(formula = SALNOW ~ AGE + WORK + SEX, data = bank)

Residuals:
   Min     1Q Median     3Q    Max
 -8935  -3625  -1235   1429  38250

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10136.38    1349.15   7.513 2.94e-13 ***
AGE            38.04      41.49   0.917  0.35960
WORK         -181.27      56.81  -3.191  0.00151 **
SEXMale      6733.08     590.78  11.397  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5997 on 470 degrees of freedom
Multiple R-squared:  0.2341,    Adjusted R-squared:  0.2292
F-statistic: 47.88 on 3 and 470 DF,  p-value: < 2.2e-16
```



Example: Harris Bank Data



Relationship between age and gender as well as between age, work experience and gender looks interesting

Is there a good explanation why there are no (so few) females in the age range between 35 and 40?

Why have females aged 40+ less work experience than males at the same age?

Example: Harris Bank Data

- Y: SALNOW (salary in 1977)
- X_1: education level (years of education)
- X_2: gender
- X_3: seniority (time since first hired by Harris)
- X_4: Age
- X_5: Work Experience (prior to hire by Harris, measured in years)
- X_6: job category
- X_7: minority (belongs to minority group)

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                1678.99    1795.40   0.935  0.35020
SEXMale                    2165.48     383.74   5.643 2.92e-08 ***
TIME                         40.57      15.77   2.573  0.01039 *
AGE                         -18.73      25.06  -0.748  0.45513
EDLEVEL                     507.84      75.24   6.750 4.46e-11 ***
WORK                        -45.81      34.36  -1.333  0.18306
JOBCATCollegeTrainee       9162.07     654.37  14.001  < 2e-16 ***
JOBCATExempt              11107.61     719.63  15.435  < 2e-16 ***
JOBCATMBATrainee          11762.99    1574.72   7.470 4.06e-13 ***
JOBCATOffice               -270.99     418.27  -0.648  0.51738
JOBCATSecurity             2551.18     805.75   3.166  0.00165 **
JOBCATTechnical           21548.11    1447.44  14.887  < 2e-16 ***
MINORITYMinority           -803.43     388.57  -2.068  0.03923 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3339 on 461 degrees of freedom
Multiple R-squared:  0.7671,    Adjusted R-squared:  0.7611
F-statistic: 126.6 on 12 and 461 DF,  p-value: < 2.2e-16
```

Example: Harris Bank Data

- Y: SALNOW (salary in 1977)
- X_1: education level (years of education)
- X_2: gender
- X_3: seniority (time since first hired by Harris)
- X_4: Age
- X_5: Work Experience (prior to hire by Harris, measured in years)
- X_6: job category
- X_7: minority (belongs to minority group)

```
Analysis of Variance Table

Response: SALNOW
           Df     Sum Sq    Mean Sq  F value    Pr(>F)
SEX         1 4466965285 4466965285 400.7442 < 2.2e-16 ***
TIME        1   65081115   65081115   5.8386   0.01607 *
AGE         1  349871557  349871557  31.3880 3.639e-08 ***
EDLEVEL     1 5964115492 5964115492 535.0578 < 2.2e-16 ***
WORK        1   13138481   13138481   1.1787   0.27819
JOBCAT      6 6021194768 1003532461  90.0298 < 2.2e-16 ***
MINORITY    1   47655363   47655363   4.2753   0.03923 *
Residuals 461 5138617209   11146675
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Harris Bank Data

- Challenges:
  - Which variables to include in the model?
  - Which model formula to use?
    - Only main effects?
    - Which interactions?
    - Only two-way or also higher-order interactions?
  - How good is the resulting model?
    - Can we do better?
    - What is the price of building a better model?
  - What is our general goal?
    - Describing discrimination?
    - Proving discrimination?
    - Predicting future salaries?
    - Explaining salary composition?

Regression: Main effects and interactions

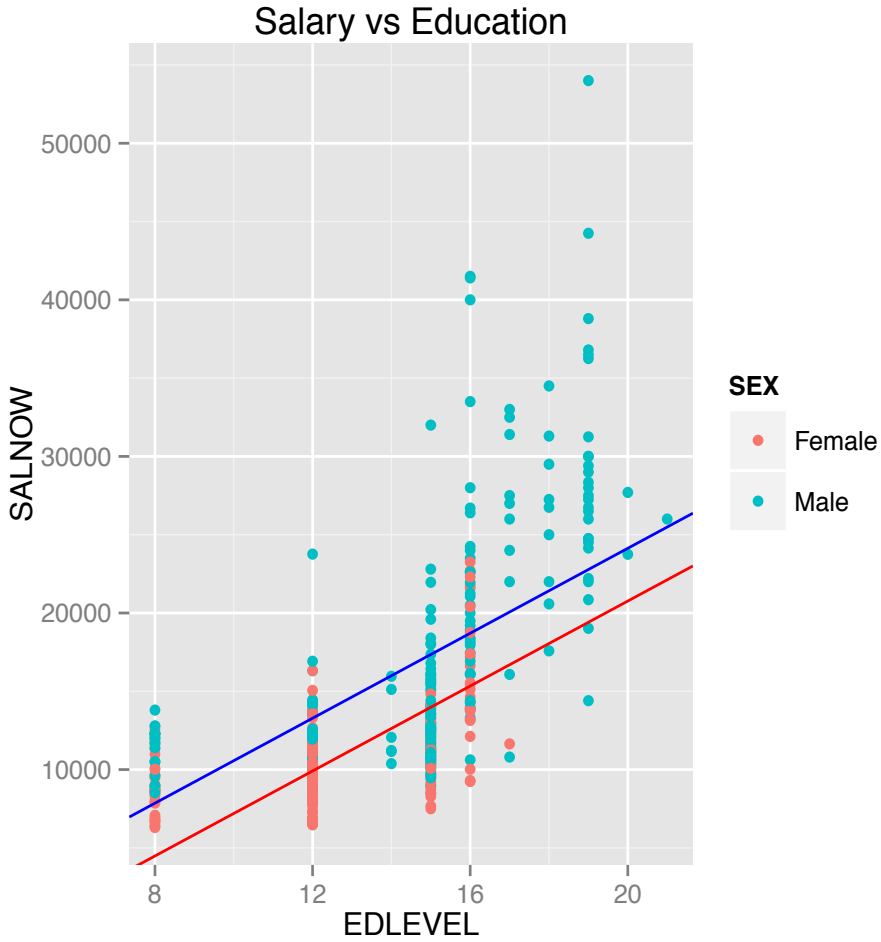
- Main effects are the effects on one variable holding all others constant (ceteris paribus)
- Interaction effects in the broad sense are effects not operating separately

- In statistical modeling it means that the effect of some predictor on the response is not constant, but depends on the level of some other predictor (e.g. salary increase due to education may be different for males and females)
- Computation results in two separate regression coefficients (main effect plus interaction). Interpretation always has to use the two components together.

Example: Harris Bank Data

- Y: SALNOW (salary in 1977)
- X_1: education level (years of education)
- X_2: gender

```
Call:
lm(formula = SALNOW ~ EDLEVEL + SEX, data = bank)

Residuals:
    Min      1Q  Median      3Q     Max
-9263.0 -3077.3  -783.3  2054.7 31223.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6369.78    1084.52  -5.873 8.08e-09 ***
EDLEVEL      1356.67      83.44  16.259  < 2e-16 ***
SEXMale      3369.38     482.81   6.979 1.02e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4892 on 471 degrees of freedom
Multiple R-squared:  0.4892,    Adjusted R-squared:  0.487
F-statistic: 225.5 on 2 and 471 DF,  p-value: < 2.2e-16
```



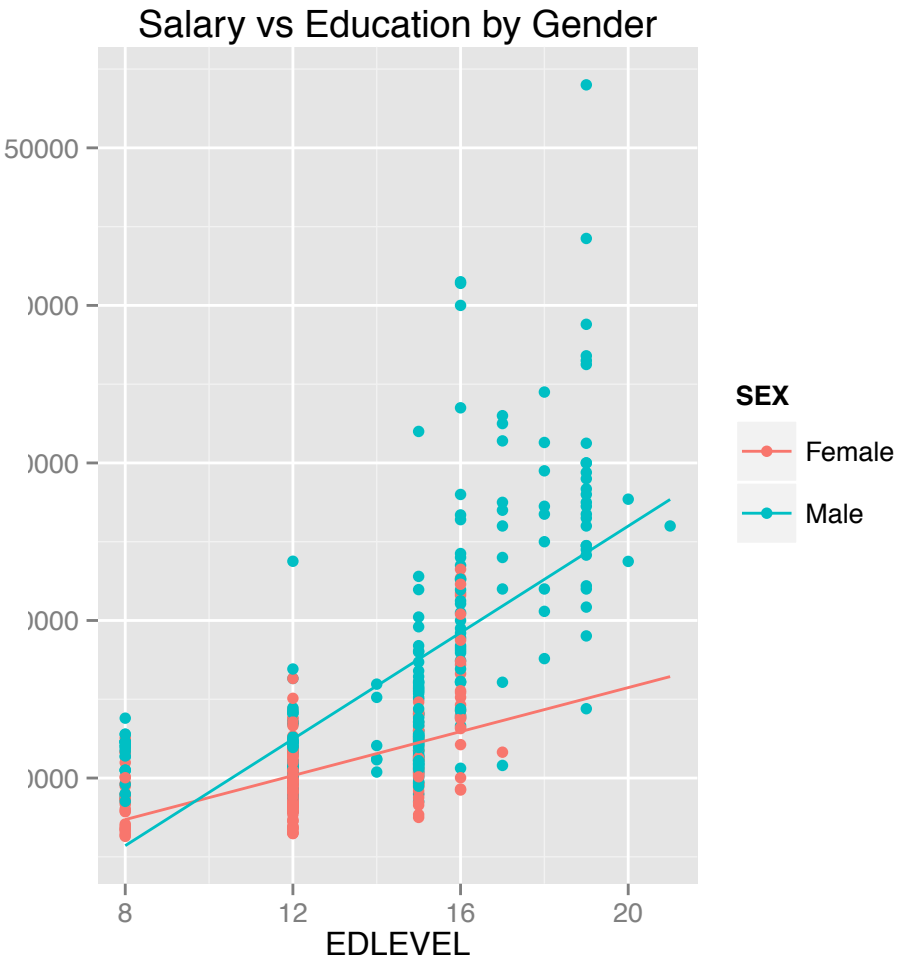
Gender yields an additional intercept

“varying intercepts model”

Example: Harris Bank Data

- *Y*: SALNOW (salary in 1977)
- X_1: education level (years of education)
- X_2: gender

```
Call:
lm(formula = SALNOW ~ EDLEVEL * SEX, data = bank)

Residuals:
     Min       1Q   Median       3Q      Max
-10120.7  -2441.4   -567.6   1697.0  29698.5

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       1775.0     1750.7   1.014    0.311
EDLEVEL            698.3      139.1   5.020 7.36e-07 ***
SEXMale          -9591.4     2279.1  -4.209 3.08e-05 ***
EDLEVEL:SEXMale    992.2      170.8   5.810 1.15e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4730 on 470 degrees of freedom
Multiple R-squared:  0.5234,    Adjusted R-squared:  0.5203
F-statistic:   172 on 3 and 470 DF,  p-value: < 2.2e-16
```



two different regression lines, one for each gender modeled as interaction between continuous and categorical predictor

“varying slopes model”

Linear models in R: model operators

formula	meaning
`Y` $\sim$ `X`	Y is modelled by X
`X`$_1$ + `X`$_2$	main effects of X_1 and X_2
`X`$_1$ − `X`$_2$	with X_1, without X_2
`X`$_1$: `X`$_2$	interaction of X_1 and X_2
`X`$_1$ * `X`$_2$	short for `X`$_1$ + `X`$_2$ + `X`$_1$: `X`$_2$
`X`$_1$ %*in*% `X`$_2$	X_1 nested in X_2
`X`$_1$/`X`$_2$	short for `X`$_1$ + `X`$_2$%*in*%`X`$_1$
`X ^ n`	all interactions up to order n
`f(X`$_1$ + `X`$_2$`)`	within functions:
	+ − /*: arithmetic operatiors
`I(X)`	identity: enables
	arithmetic operations
`Y` $\sim$.	Y using all available predictors
	in `data.frame`

Linear models in R: generic functions

`print(obj)`	model overview
`summary(obj)`	model summary
`coef(obj)`	model coefficients
`resid(obj)`	residuals
`fitted(obj)`	fitted values
`predict(obj,newdata= NewData)`	prediction for NewData
`deviance(obj)`	residual deviance
`anova(obj)`	sequential ANOVA-table
`plot(obj)`	some diagnostic plots
`obj$df.residual`	residual df

Summary

- Intro to class
- Recap of linear regression and fundamental statistical concepts
- Multiple linear regression