Statistics II
Multiple Regression

Çağrı Çöltekin

ideas/examples/slides from
John Nerbonne & Hartmut Fitz

University of Groningen, Dept of Information Science

university of groningen

April 24, 2013

Some reminders

- Computer exercises:
  - The first exercise to be done this week. Deadline in two weeks.
  - We will have two weeks of break.
- Quizzes:
  - The first quiz is already on Nestor, the second one will be available today.
  - You can try them as many times as you like, but you need to do them in a two-week time window.
  - Note: less than 60% will count as 0.

Scheduling problems: the exam

Exam was scheduled at June 21 Friday at 10:00. However, it seems to conflict with some people. New alternatives:

June 2013

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
(17)	18	19	(20)	(21)	22	23
24	25	26	27	28	29	30

- June 17 Monday, 19.00–21.00
- June 20 Thursday

Correlation

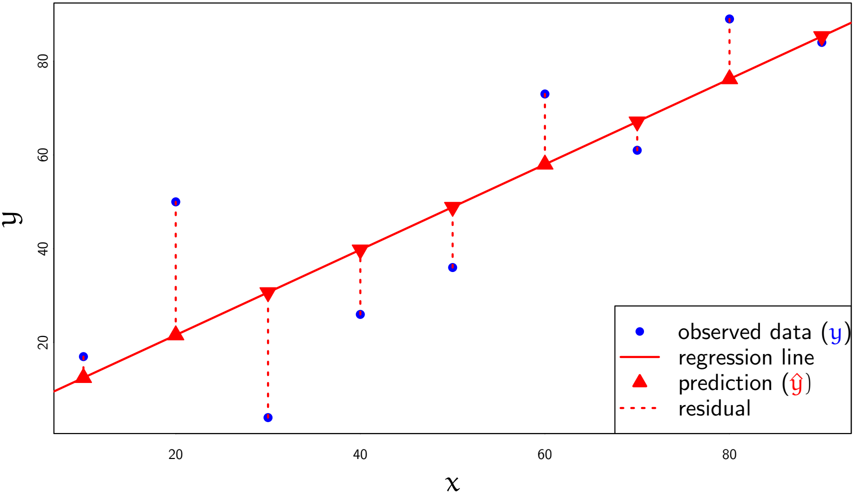
- The correlation coefficient (r) is a standardized symmetric measure of covariance between two variables.
- The correlation coefficient ranges between -1 and 1.
- Correlation and regression are strongly related.
- The most common correlation coefficient is Pearson's r, which assumes a linear relationship between two variables.
- When this assumption is not correct, non-parametric alternatives Spearman's ρ or Kendall's τ can be used.
- Correlation is not causation!

Simple regression

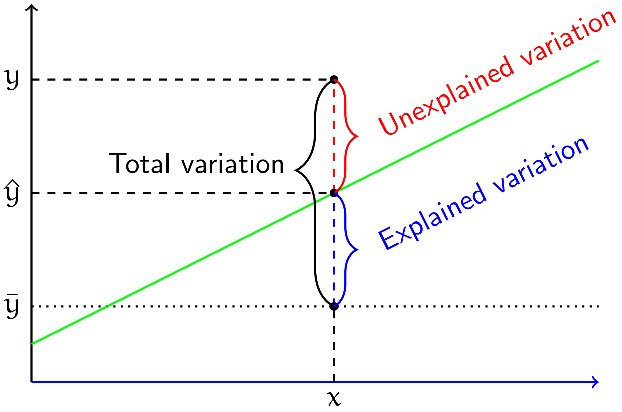
$$y_i = a + bx_i + e_i$$

- y is the *response* (or outcome, or dependent) variable. The index i represent each unit observation/measurement (sometimes called a 'case').
- x is the *predictor* (or explanatory, or independent) variable.
- a is the intercept.
- b is the slope of the regression line.
- $a + bx$ is the *deterministic* part of the model (we sometimes use $\hat{y}$).
- e is the residual, error, or the variation that is not accounted for by the model. Assumed to be (approximately) normally distributed with 0 mean (e_i are assumed to be i.i.d).

The regression line



Variation explained by regression



Total variation	=	Unexplained variation	+	Explained variation
$y - \bar{y}$	=	$y - \hat{y}$	+	$\hat{y} - \bar{y}$

Estimation and interpretation of regression

- The most common method of estimation is the 'least-squares regression', which minimizes the square of the residuals.
- Intercept (a) is the value y takes when $x = 0$.
- Slope (b) is the change in y when x changes 1 unit.
- Coefficient of determination (r^2) represent ratio of variance of y explained by x.
- Individual t-tests for coefficients indicates whether estimate is
- F-test indicates statistical significance of the overall model performance.

Regression analysis step by step

1. Collect/check your data: cases should be independent.
2. Fit your model (let the computer do it).
3. Check assumptions or problem indications:

linearity	scatter plot of 'y vs. x' or 'residuals vs. fitted'.
normality	(of residuals!) histogram, Q-Q (or P-P) plot.
constant variance	(of residuals!) 'residuals vs. fitted' plot.
outliers	scatter plot of 'y vs. x' together with regression line, residual histogram or box plot.
influential cases	scatter plot of 'y vs. x', 'residuals vs. fitted', or more specialized statistics like *Cook's distance*.

4. Interpret your results:
   - Model parameters (coefficients): intercept and slope estimates.
   - Model fit: coefficient of determination (r^2).
   - Generalizability of the estimates: F-test for the model, and t-tests for the coefficients.
   - Prediction: confidence intervals for regression line (expected value of the response variable), and future observations.

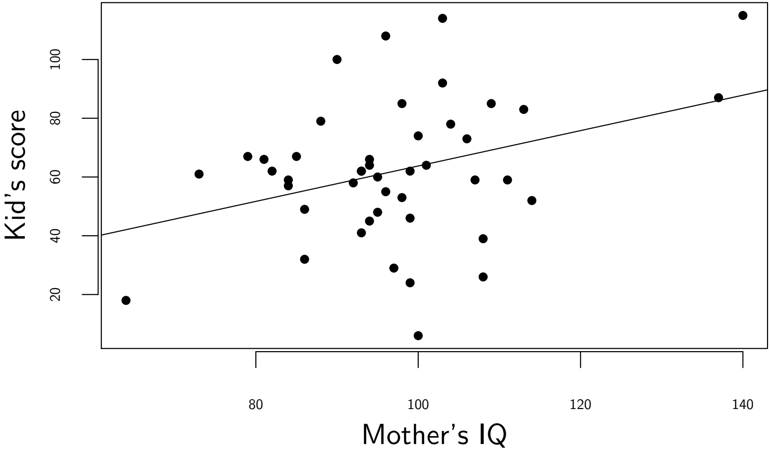
Regression example: 1. the data

Case	Kid's Score	Mom's IQ
1	109	91
2	99	102
3	96	88
...		
43	108	101
44	110	78
45	97	67

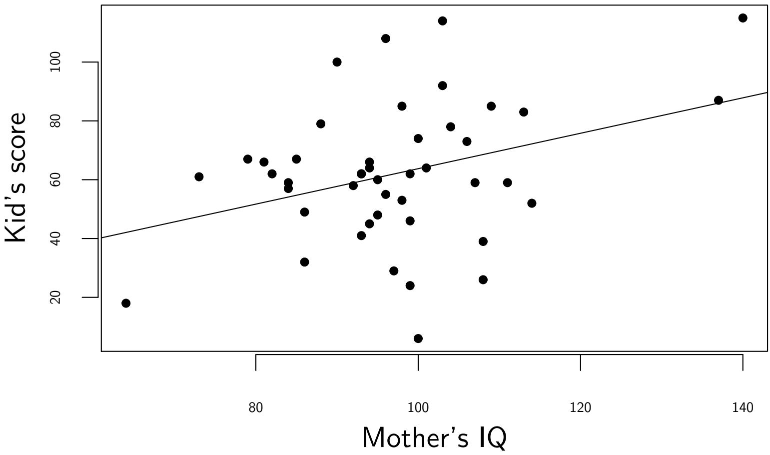
Not many assumptions here:

- Cases are independent.
- Both predictor and the response variables are numeric (not strictly, more on this later).

Regression example: 2. plot your data



Regression example: 2. plot your data



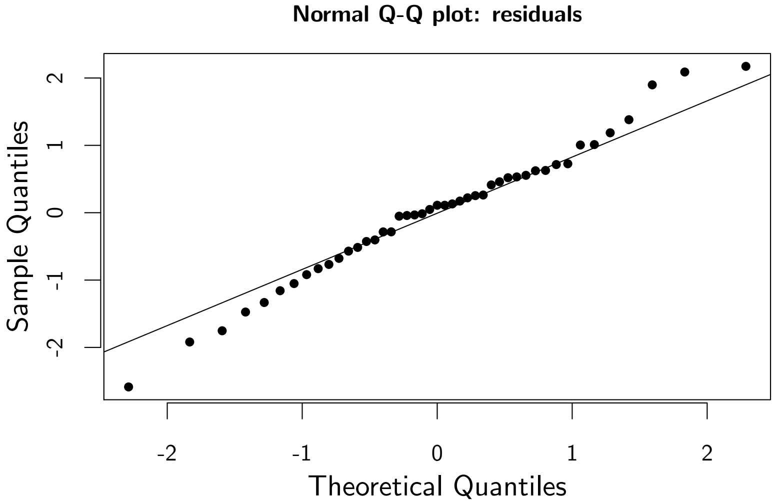
- Are there any non-linear patterns?
- Are there outliers or influential observations?

Regression example: 3. fit your model

```
lm(formula = kid.score ~ mother.iq)
Residuals:
   Min      1Q  Median      3Q     Max
-57.749 -12.737 2.467 12.286 48.444
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.5174   24.2375  0.145    0.885
mother.iq    0.6023    0.2471  2.437    0.019 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 22.59 on 43 degrees of freedom
Multiple R-squared: 0.1214,  Adjusted R-squared: 0.101
F-statistic: 5.941 on 1 and 43 DF, p-value: 0.019
```

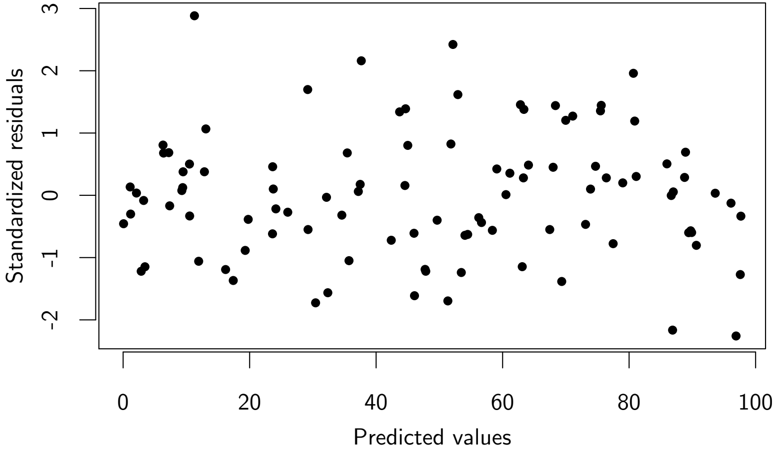
... but before drawing conclusions...

Regression example: 4. check residuals for normality

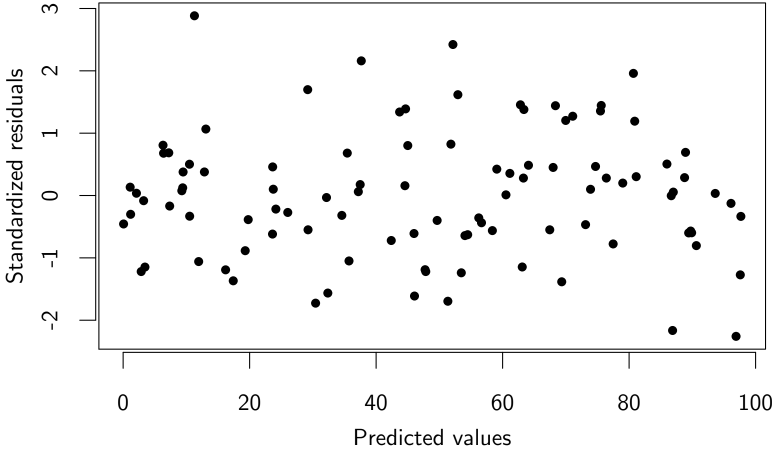


- Are residuals distributed approximately normally?

Regression example: 5. residuals vs. predicted



Regression example: 5. residuals vs. predicted



- Are there any patterns, e.g., non-linearity?
- Is the variance of residuals constant?
- Are there outliers?

Regression example: 6. what does the model say?

```
lm(formula = kid.score ~ mother.iq)
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.5174   24.2375  0.145   0.885
mother.iq    0.6023    0.2471  2.437   0.019 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 22.59 on 43 degrees of freedom
Multiple R-squared: 0.1214,  Adjusted R-squared: 0.101
F-statistic: 5.941 on 1 and 43 DF, p-value: 0.019
```

$b = 0.6$ Expected score difference between two children whose mother's IQ differs one unit.

Regression example: 6. what does the model say?

```
lm(formula = kid.score ~ mother.iq)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.5174   24.2375  0.145   0.885
mother.iq    0.6023    0.2471  2.437   0.019 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 22.59 on 43 degrees of freedom
Multiple R-squared: 0.1214,  Adjusted R-squared: 0.101
F-statistic: 5.941 on 1 and 43 DF, p-value: 0.019
```

$b = 0.6$ Expected score difference between two children whose mother's IQ differs one unit.

$r^2 = 0.12$ Mother's IQ explains 12% of the variation in test scores.

Regression example: 6. what does the model say?

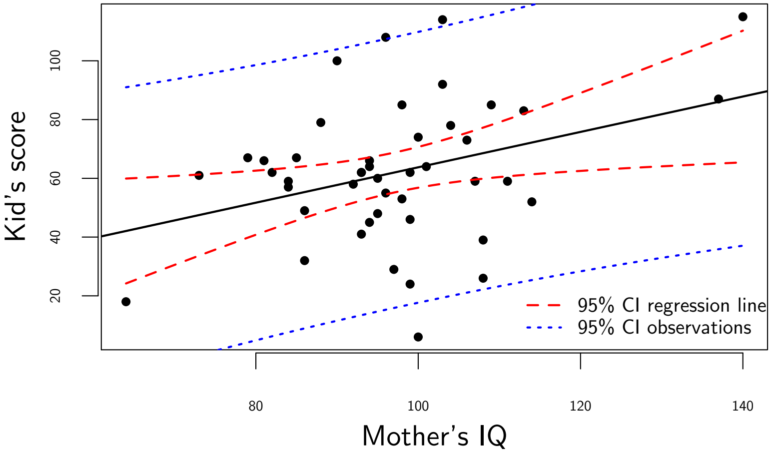
```
lm(formula = kid.score ~ mother.iq)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.5174   24.2375  0.145   0.885
mother.iq    0.6023    0.2471  2.437   0.019 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 22.59 on 43 degrees of freedom
Multiple R-squared: 0.1214,  Adjusted R-squared: 0.101
F-statistic: 5.941 on 1 and 43 DF, p-value: 0.019
```

$b = 0.6$ Expected score difference between two children whose mother's IQ differs one unit.

$r^2 = 0.12$ Mother's IQ explains 12% of the variation in test scores.

$p = 0.02$ Given the sample size, probability of finding b value that far from 0 (two-tailed t-test with null hypothesis $b = 0$).

Regression example: 7. prediction



Note: prediction error is not the same everywhere.

Multiple regression: motivating examples

Often we want to predict a (numeric) variable based on more than one (numeric) predictors. Examples:

- university performance dependent on general intelligence, high school grades, education of parents,...
- income dependent on years of schooling, school performance, general intelligence, income of parents,...
- level of language ability of immigrants depending on
  - leisure contact with natives
  - age at immigration
  - employment-related contact with natives
  - professional qualification
  - duration of stay
  - accommodation

Data for multiple regression

One response variable (y), k predictors (x_1 to x_k), and n data points (observations or cases).

Case	response	predictors		
1	y_1	$x_{1,1}$	...	$x_{1,k}$
2	y_2	$x_{2,1}$	...	$x_{2,k}$
...				
n	y_n	$x_{n,1}$	...	$x_{n,k}$

Multiple regression: formulation

$$y_i = \underbrace{a + b_1 x_{i,1} + b_2 x_{2,i} + \ldots + b_k x_{k,i}}_{\hat{y}} + e_i$$

a is the intercept (as before).

$b_{1..k}$ are the coefficients of the respective predictors.

e is the error term (residual).

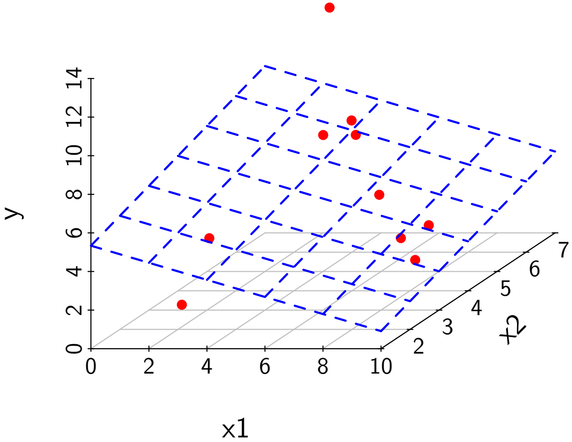
It is a generalization of simple regression with some additional power and complexity.

Multiple regression: issues and difficulties

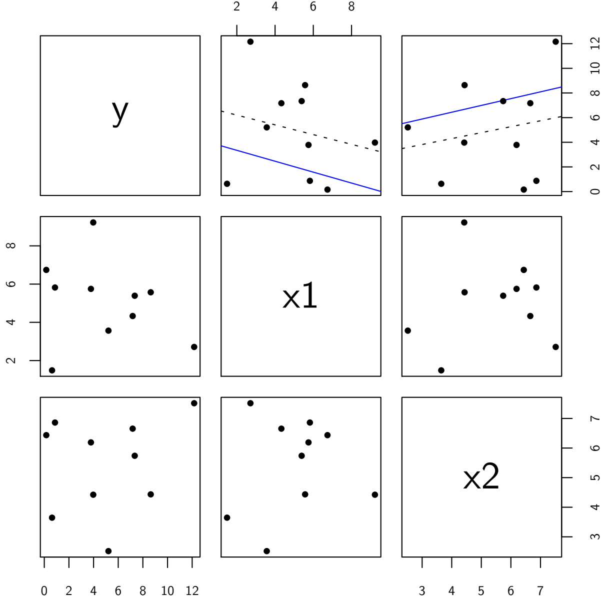
Multiple regression shares all aspects/assumptions of simple regression, and

- Visual inspection of the data becomes more difficult.
- Multicollinearity causes problems in estimation and interpretation of multiple-regression models.
- Suppression is another possibility, where combination of predictors are more useful than individual predictors.
- Overfitting, occurs when there are large number of predictors.
- Model selection (finding a model that fits the data well, but not more complex than necessary) is important.

Visualizing regression with two predictors



Pairwise scatter plots



Least-squares regression for multiple predictors

As in simple regression, we try to minimize SS_R

$$SS_R = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (a + b_1 x_{i,1} + \ldots + b_k x_{i,k}))^2$$

The parameter values (a, b_1, ..., b_k) that minimize the above expression can, again, be calculated analytically (if $n > k$).

Model fit: partitioning the variance

Similar to simple regression, we can partition the variance (sums of squares) as,

$$\begin{array}{ccccc} \text{Total variance} & = & \text{Explained variance} & + & \text{Unexplained variance} \\ \sum_i (y_i - \bar{y}_i)^2 & = & \sum_i (\hat{y}_i - \bar{y}_i)^2 & + & \sum_i (y_i - \hat{y}_i)^2 \\ SS_T & = & SS_M & + & SS_R \end{array}$$

$$\text{multiple-r}^2 = \frac{SS_M}{SS_T}$$

- Like in single regression, we interpret multiple-r^2 as the ratio of variance explained by the model.

Inference for multiple regression

Inference also follows single regression, we test significance of the model based on the F statistic distributed with $F(k, n-k-1)$.

$$F = \frac{MS_M}{MS_R}$$

This is significance test for at least one non-zero b value. The null hypothesis is

$$H_0 : b_1 = b_2 = \ldots = b_k = 0$$

As before, the estimates of the individual coefficients (a and $b_{1..k}$) are tested for significance using t-test.

An example multiple regression

We extend last week's example: we want to predict children's cognitive development based on their mother's IQ, and the amount of time they spend in front of TV. The data:

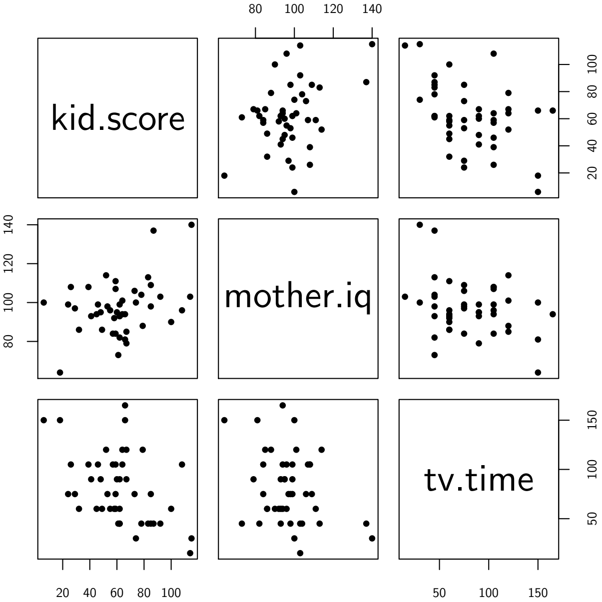
Case	Kid's Score	Mom's IQ
1	109	91
2	99	102
3	96	88
...		
43	108	101
44	110	78
45	97	67

An example multiple regression

We extend last week's example: we want to predict children's cognitive development based on their mother's IQ, and the amount of time they spend in front of TV. The data:

Case	Kid's Score	Mom's IQ	TV time (min/day)
1	109	91	45
2	99	102	90
3	96	88	150
...			
43	108	101	120
44	110	78	75
45	97	67	45

Always plot your data



Regression coefficients

```
lm(formula = kid.score ~ mother.iq + tv.time)
Coefficients:
(Intercept)   mother.iq     tv.time
    42.9056      0.4078     -0.2530
```

How to interpret it?

Intercept (a) Test score of a kid whose mother has $IQ = 0$, and who does not watch any TV at all.

$b_{mother.iq}$ Change in the test score when Mother's IQ is increased one unit, while keeping TV time constant.

$b_{tv.time}$ Change in the test score when increasing TV time one unit (minute) while keeping Mother's IQ constant.

Model fit

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 42.90562 26.94569  1.592  0.1188
mother.iq    0.40781  0.24186  1.686  0.0992 .
tv.time     -0.25302  0.09384 -2.696  0.0100 *
---
Residual standard error: 21.11 on 42 degrees of freedom
Multiple R-squared:  0.251, Adjusted R-squared:  0.2154
F-statistic: 7.039 on 2 and 42 DF, p-value: 0.00231
```

multiple-r^2 Is percentage of variation explained by the model.

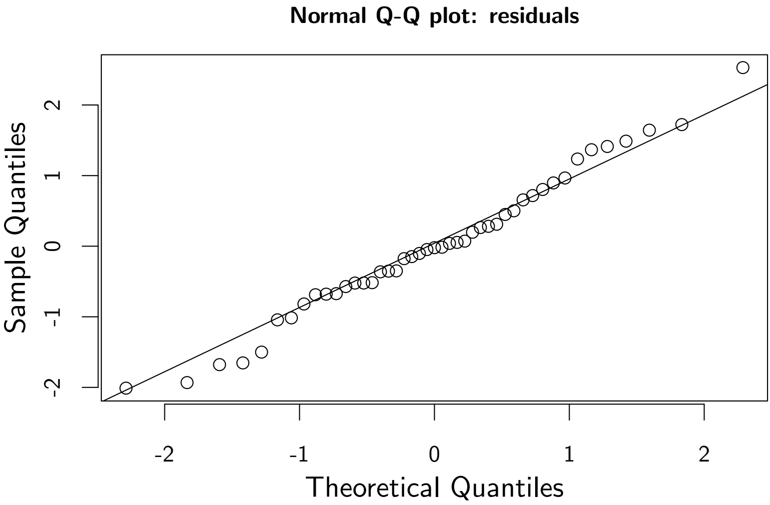
adjusted-r^2 Adding more predictors increase multiple-r^2. Adjusted-r^2 (or $\bar{r}^2$)corrects for by-chance increase due to more predictors. $\bar{r}^2 = 1 - \left[\frac{n-1}{n-k-1} \times (1 - r^2)\right]$.

Inference

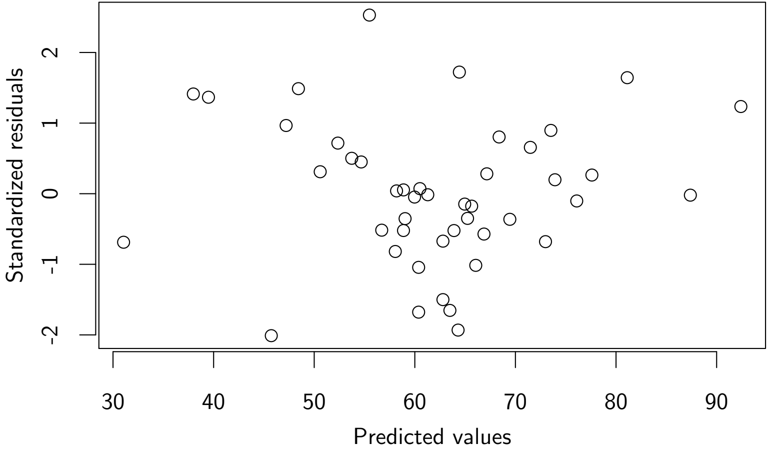
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 42.90562 26.94569 1.592 0.1188
mother.iq    0.40781  0.24186 1.686 0.0992 .
tv.time     -0.25302  0.09384 -2.696 0.0100 *
---
Residual standard error: 21.11 on 42 degrees of freedom
Multiple R-squared: 0.251,   Adjusted R-squared: 0.2154
F-statistic:  7.039 on 2 and 42 DF, p-value:  0.00231
```

- T-tests for predictors show significance of the coefficient estimates.
- F-test indicates the significance of the complete model.

Diagnostics: normality of the residuals



Diagnostics: predicted vs. residuals graph



More diagnostics: outliers and influential cases

- Influential observations affect the regression line (or surface)
- Outliers are easy to spot on a scatter plot for single predictor.
- Not all outliers are influential, an outlier is more likely to be influential if it is at the extreme values of predictors.
- One (of many) statistics that are used for detecting influential cases is Cook's distance, which measures the effect of removing a case from the regression estimation.
- The values for large (above 1) Cook's distance are a cause of concern.

Which predictors to include: model selection

Given two predictors $(x_1, x2)$ and a response variable (y), our options are:

$y_i = a + e_i$	the null model, or the 'model of the mean' (note that $a = \bar{y}$).
$y_i = a + b_1 x_{i,1} + e_i$	y depends only on x_1
$y_i = a + b_2 x_{i,2} + e_i$	y depends only on x_2
$y_i = a + b_1 x_{i,1} + b_2 x_{i,2} + e_i$	both x_1 and x_2 affect the outcome variable.

Model selection: the model fit

Everything being equal, we want the model that explains the data at hand the best (higher r^2).
For our example:

predictor	r^2	F-test (p value)	t-test (p-value)
Mother's IQ	0.12	0.0100	0.019
TV time	0.20	0.0021	0.002
Mother's IQ & TV time	0.25	0.0023	0.100 0.010

Model selection: the model fit

Everything being equal, we want the model that explains the data at hand the best (higher r^2).
For our example:

predictor	r^2	F-test (p value)	t-test (p-value)
Mother's IQ	0.12	0.0100	0.019
TV time	0.20	0.0021	0.002
Mother's IQ & TV time	0.25	0.0023	0.100 0.010

Things to note

- r^2's do not sum up.
- Significance drops with multiple predictor estimates.

Which model is the best?

We prefer models with high model fit (high r^2). However

- r^2 is a measure of how well your data fits to the current sample, we want to develop models that are useful beyond the sample at hand.
- Adding more predictors increase model fit.
- If you have as many predictors as data points, you have a *saturated* model.
- The model selection process is a balance between a model that fits well to the data and a model that is simpler (fewer parameters).

Which model is the best?

We prefer models with high model fit (high r^2). However

- r^2 is a measure of how well your data fits to the current sample, we want to develop models that are useful beyond the sample at hand.
- Adding more predictors increase model fit.
- If you have as many predictors as data points, you have a *saturated* model.
- The model selection process is a balance between a model that fits well to the data and a model that is simpler (fewer parameters).

Everything should be made as simple as possible, but no simpler.

Stepwise methods

Ideally, model selection should be based on your theories about the problem.

- You can compare two models using an F-test (as we compare our model to the null model).

$$F = \frac{MS_{m_1}}{MS_{m_2}}$$

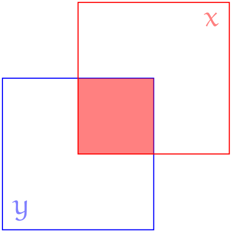
- You can also use more general statistics like 'Akaike information criterion' (AIC).
- Once you have a way to compare two models, you can also ask computer to search for the best model using stepwise methods.

Multicollinearity

Multicollinearity is a problem associated with multiple predictors explaining same portion of the variance in the response variable.

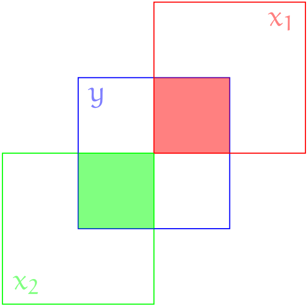
- In case of perfect multicollinearity (when one of the predictors is predicted by others perfectly) regression line cannot be estimated.
- Ideal case is when there is no multicollinearity: this rarely happens.
- High correlation between predictors is a sign of multicollinearity.
- High multicollinearity causes uncertain estimates of the coefficients.

Multicollinearity: visualization



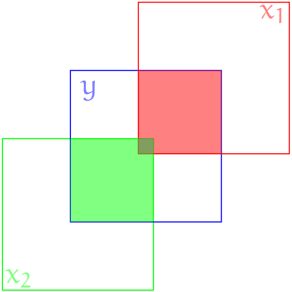
- Single regression $y = a + bx + e$.
- Filled area: r^2, variance of y by x, or square of the Pearson's r (correlation coefficient).

Multicollinearity: visualization



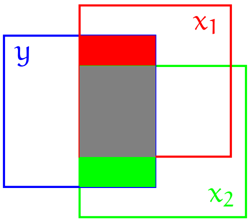
- Multiple regression
  $y = a + b_1 x_1 + b_2 x_2 + e$.
- No multicollinearity.
- Filled areas:
  - red: $r^2_{x_1} = 0.25$, due to x_1
  - green: $r^2_{x_2} = 0.25$, due to x_2
  - Total $r^2 = 0.50$, due to model.

Multicollinearity: visualization



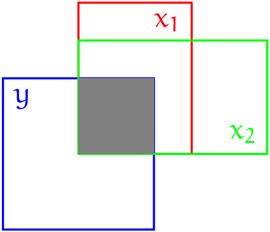
- Multiple regression
  $y = a + b_1 x_1 + b_2 x_2 + e$.
- Small/mild multicollinearity.
- Filled areas:
  - red: $r^2_{x_1} = 0.36$, due to x_1
  - green: $r^2_{x_2} = 0.36$, due to x_2
  - gray: $r^2_{x_1, x_2} = 0.04$, due to both variables.
  - Total $r^2 = 0.68$ (not 0.72), due to model.

Multicollinearity: visualization



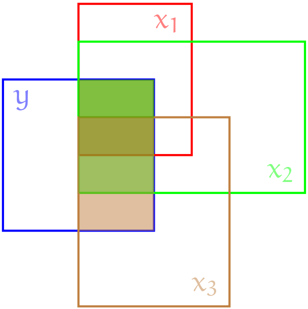
- Multiple regression
  $y = a + b_1 x_1 + b_2 x_2 + e$.
- Small/mild multicollinearity.
- Filled areas:
  - red+gray: $r^2_{x_1} = 0.4$, due to x_1
  - green+gray: $r^2_{x_2} = 0.4$, due to x_2
  - gray: $r^2_{x_1,x_2} = 0.3$, due to both variables.
  - Total $r^2 = 0.5$ (not 0.8), due to model.

Multicollinearity: visualization



- Multiple regression
  $y = a + b_1 x_1 + b_2 x_2 + e$.
- Perfect multicollinearity.
- Regression parameters cannot be estimated in this case.
- Some software will return an error, some will drop one of the predictors.

Multicollinearity: visualization



- Multiple regression $y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + e$.
- Another example of perfect multicollinearity with 3 variables.
- All explanation x_2 provides is also explained by combination of x_1 and x_3.

Multicollinearity: how to detect it?

- High pairwise correlation is an indication, but not a sufficient one.
- No/small increase in r^2 in the combined model with respect to individual predictors is another indication.
- Variance-inflation factor (VIF) statistics.
  - For each predictor, x_j, fit a regression model, $x_j = a + \ldots + x_{j-1} + x_{j+1} + \ldots x_k$
  - Calculate the r_j^2 for the model.
  - VIF statistics for j^{th} is,

$$\mathrm{VIF}_j = \frac{1}{1 - r_j^2}$$

  - Interpretation of VIF is also not straightforward.
  - Values over 5 (or 10 for some) is a case for concern.

Suppression

Another possibility in multiple regression is called suppression.
Consider the following hypothetical example:

- We do a language test with time limit. We'd like to know how multilingualism affects the task.
- We find multilingualism to be negatively correlated with the test score (negative regression coefficient).
- We also add 'speed' as a variable, which turns the negative effect to positive.

How can this happen?

Suppression

Another possibility in multiple regression is called suppression.
Consider the following hypothetical example:

- We do a language test with time limit. We'd like to know how multilingualism affects the task.
- We find multilingualism to be negatively correlated with the test score (negative regression coefficient).
- We also add 'speed' as a variable, which turns the negative effect to positive.

How can this happen?

- Multi-linguals are in fact better.
- But they are also slow at this task. They cannot finish the test, so they get bad scores.
- Adding speed to the regression allows us to find the correct effect of the multilingualism in the task.

Summary: multiple regression

$$y_i = \underbrace{a + b_1 x_{i,1} + b_2 x_{2,i} + \ldots + b_k x_{k,i}}_{\hat{y}} + e_i$$

- Multiple regression is a generalization of the simple regression, where we predict the outcome using multiple predictors.
- Multicollinearity causes problems in estimation and interpretation of multiple-regression models.
- Model selection (finding a model that fits the data well, but not more complex than necessary) is important.

Summary and Next week

Today:

- A review of Regression & correlation
- Multiple regression

Next lecture:

- Single-factor ANOVA (sections 7.11–7.12 & Ch.10)

Note: next lecture is in two weeks (on May 8).