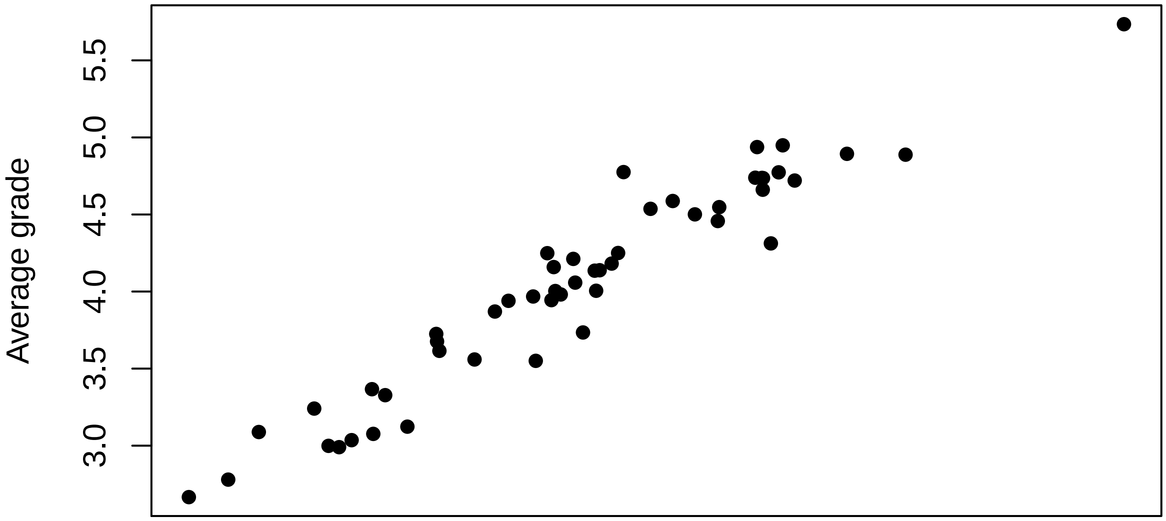
Lab 8

Wiktor Soral

April 25th 2017

Linear regression



- Suppose you observe a sample of average grades from students of some faculty

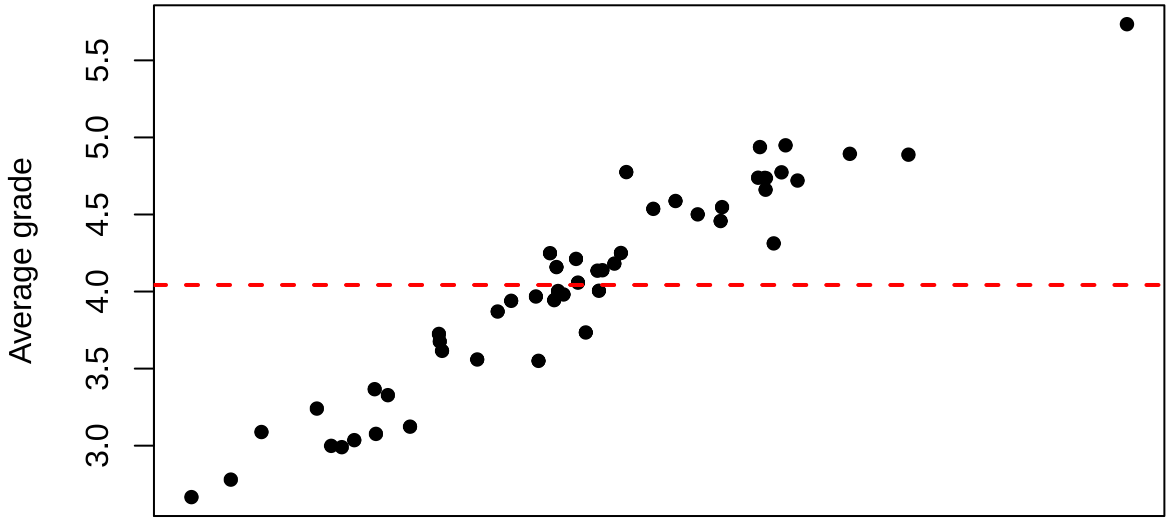
Linear regression

- Lets say that you would like to predict average grade, y, for other students of that faculty
- When making predictions we take into account associated error of prediction: $\sum_{i=1}^{N} |\hat{y}_i - \tilde{y}_i|$ where $\hat{y}_i$ is predicted value of y for each person, and $\tilde{y}_i$ is true value of y for each person
- Note that we never know $\tilde{y}_i$, but we would like to guess $\hat{y}_i$ reasonably close to the true value
- Which value for the $\hat{y}_i$ we should take?

Linear regression

- Without any additional information, we know that the value $\hat{y}_i$ that will minimize our error in predicting $\tilde{y}_i$ is equal to expected value of y, $E(y)$, i.e. arithmetic mean of y in a sample

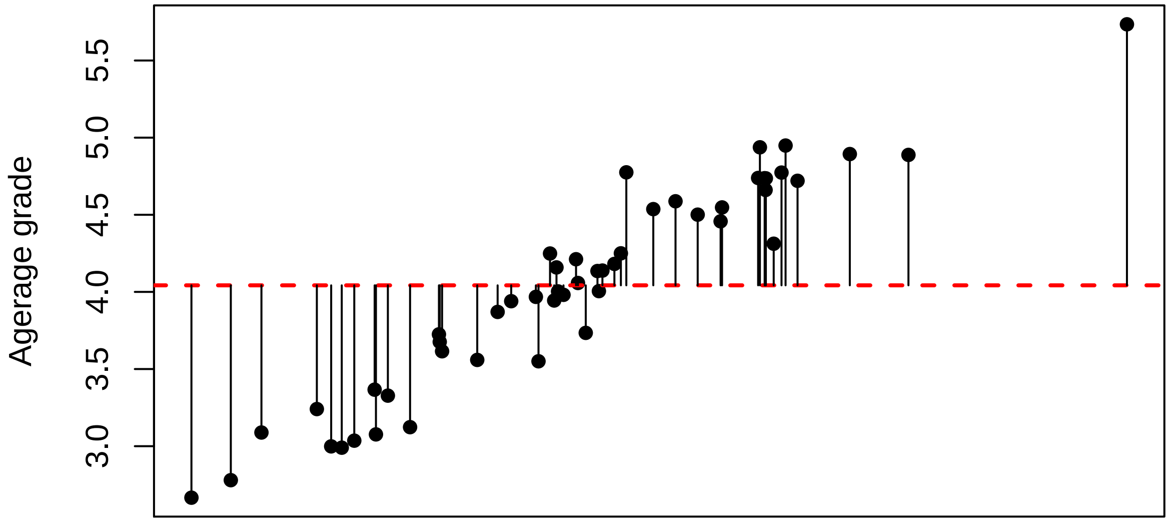
Linear regression



Linear regression

- Note that if we set $\hat{y_i}$ to $E(y)$ we are still making some error:
- $\sum_{i=1}^{N} |y_i - \hat{y_i}|$

Linear regression

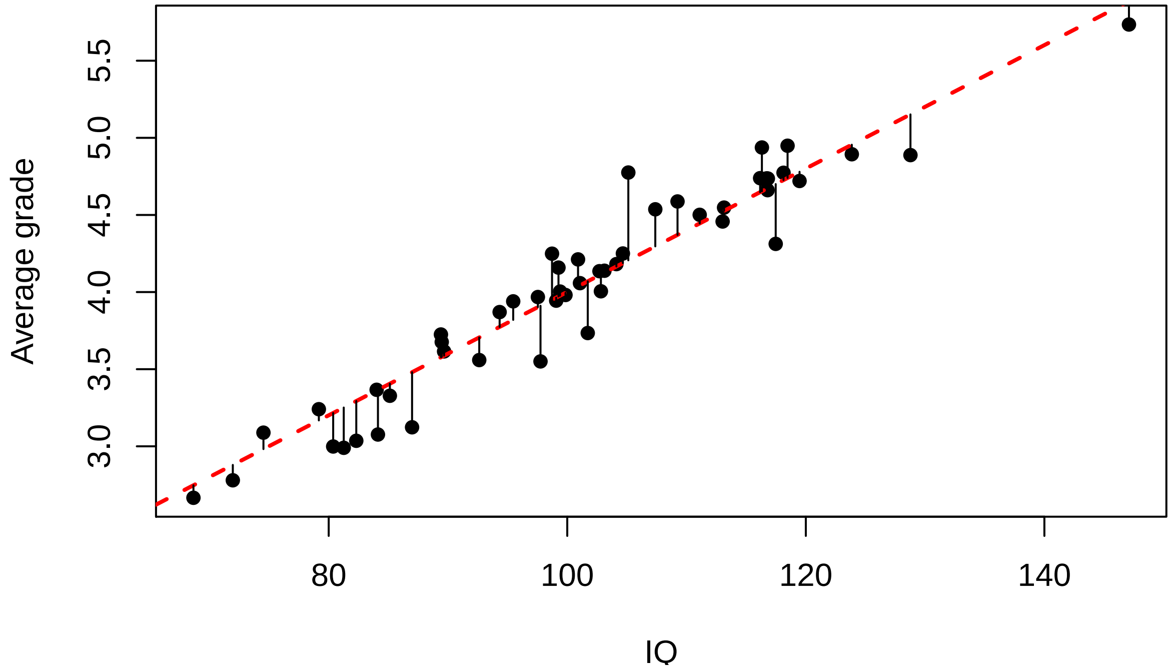


Linear regression

- With some additional information surely we can do better, i.e. if we know IQ for each individual we could make better predictions of their average grades
- If we assume that average grade is proportional to the IQ of each individual we can write:
- $grade = a + b * IQ$
- b is regression weight, i.e. how much average grade will increase/decrease if we change IQ by 1 unit

- a is intercept, i.e. average grade for individuals with IQ equal to 0 (not very useful here)

Linear regression



- Compare the lengths of vertical lines with plot where we use mean to predict y. Do you see any improvement?

Ordinary least squares (OLS)

- In OLS we want to set values for b and a that will minimize the sum of squares of residual values (i.e. length of vertical lines):
- $\sum_{i=1}^{N}(y_i - \hat{y_i})^2$ where
- $\hat{y_i} = a + b * X$
- Exact derivation requires some knowledge of calculus and matrix algebra, but ultimately it results in nice analytical solution (i.e. simple formula for a and b)

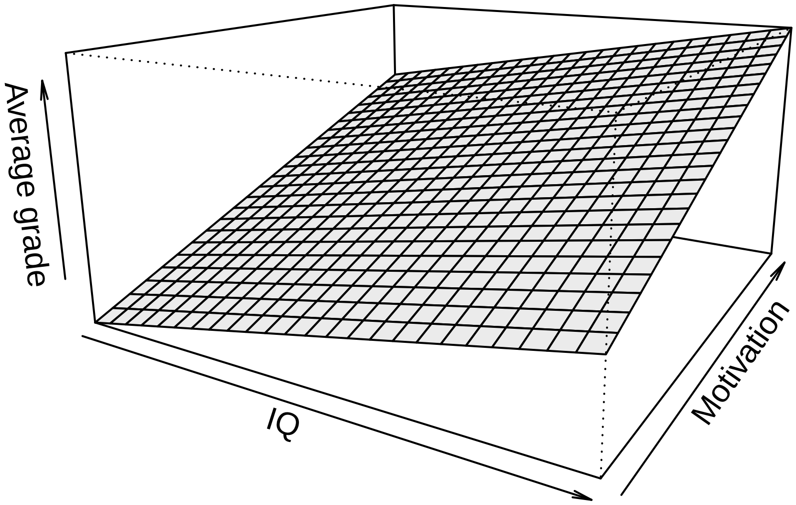
Assumptions in OLS

- Correct specification: we should use OLS only if we are sure of linear relation between predictor and outcome variables
- Residuals should have Normal distribution with mean equal to 0
- For every value of outcome variable residuals should have similar variance: homoscedasticity (recall homogeneity assumption in ANOVA)
- Residuals should be independently distributed

Multiple regression

- Lets say that we know that average grade results not only from IQ of each individual, but also from a level of motivation of each person.
- If we add another predictor to our equation we can hopefully improve our predictions
- $grade = a + b_1 * IQ + b_2 * motivation$
- In multiple regression we are trying to fit not a line, but rather a plane or hyperplane defined by our coefficients
- However the general rule of OLS is the same

Multiple regression



Multiple regression - no multicollinearity

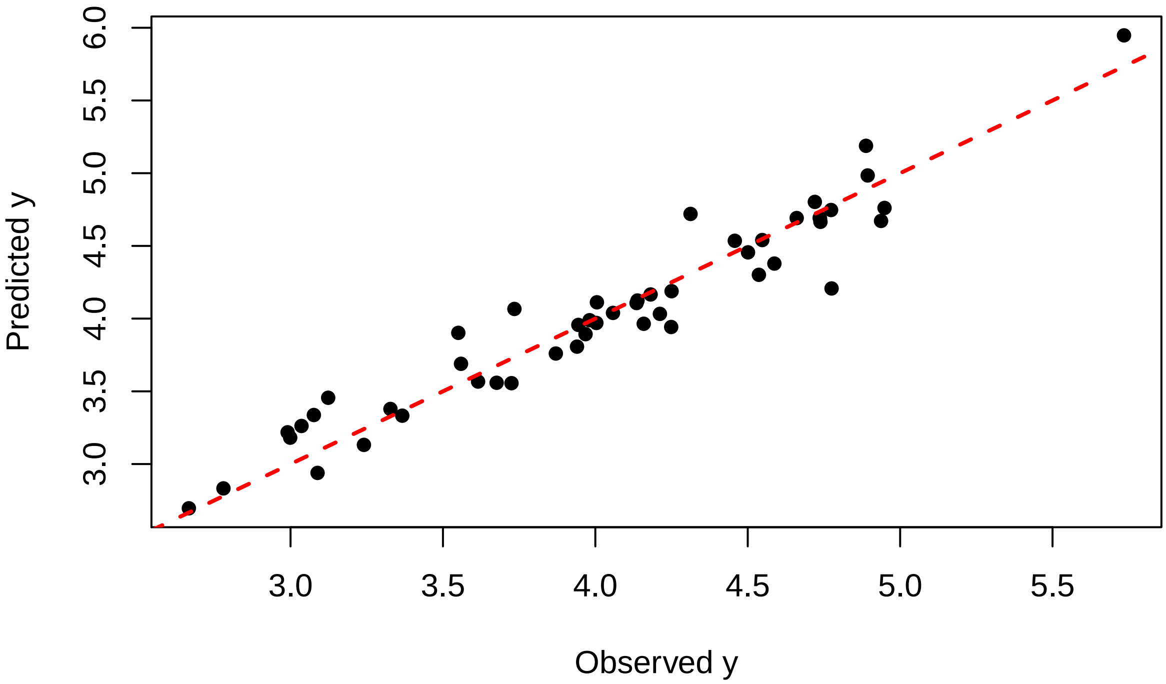
- When we use multiple regression we usually assume that predictors are not redundant - there is no multicollinearity
- Suppose you would like to predict average grade and you would use results from 2 IQ tests as predictors
- Usually correlation between results of different IQ tests is quite high (we neglect concepts as multiple intelligences for simplicity)
- Therefore using additional test results as predictors would not increase our predictive power
- Worse yet, it would inflate uncertainty associated with each predictors, i.e. our algorithm would have problems with discriminating variances explained by each predictor

Model fit and R^2

- Suppose we have build our predictive model. Does it really make better predicitions than simple model with only mean of average grades? Maybe the better prediction is a result of pure coincidence?
- We can test this by checking the ratio of variance explained by model to variance that is unexplained - note the analogy to ANOVA
- In fact we are usually using F-test to test model fit

Model fit and R^2

- To check goodness of fit we can also look at the correlation of predicted and observed values of the outcome variable, R



Model fit and R^2

- Note that if we square correlation coefficient, R, we will obtain coefficient of determination, R^2, i.e. proportion of variance of x_1 explained by variable x_2
- The same logic applies to multiple regression, where R^2 denotes the variance of the outcome variable explained by model
- According to Cohen's suggestions when we want assess the effect size of the model fit
- $R^2 = 0.02$ indicates small effect size
- $R^2 = 0.13$ indicates medium effect size
- $R^2 = 0.26$ indicates large effect size

Hierarchical regression

- Often we will have a large number of predictors that and we will not want to include them into the model at the same time
- E.g. first we would like to check only for demographic variable (gender, year of study), then we would like to include cognitve skills (intelligence), and only at the end we would like to include motivational factors
- Such approach is called hierarchical regression
- It allows for testing whether adding additional theoretically relevant variable improves our predictive power
- E.g. whether after accounting for the level of intelligence, motivational factors still predict average grade - note the similarity of this approach to ANCOVA

Hierarchical regression

- In hierarchical regression we are assessing model fit not by comparing it to the null model (with only mean), but to model fit in previous step
- We are assessing ΔR^2, i.e. change in goodness of fit