B4M36SAN
BE4M36SAN
Seminar/Tutorial

Alikhan Anuarbekov

Ali

anuarali@fel.cvut.cz

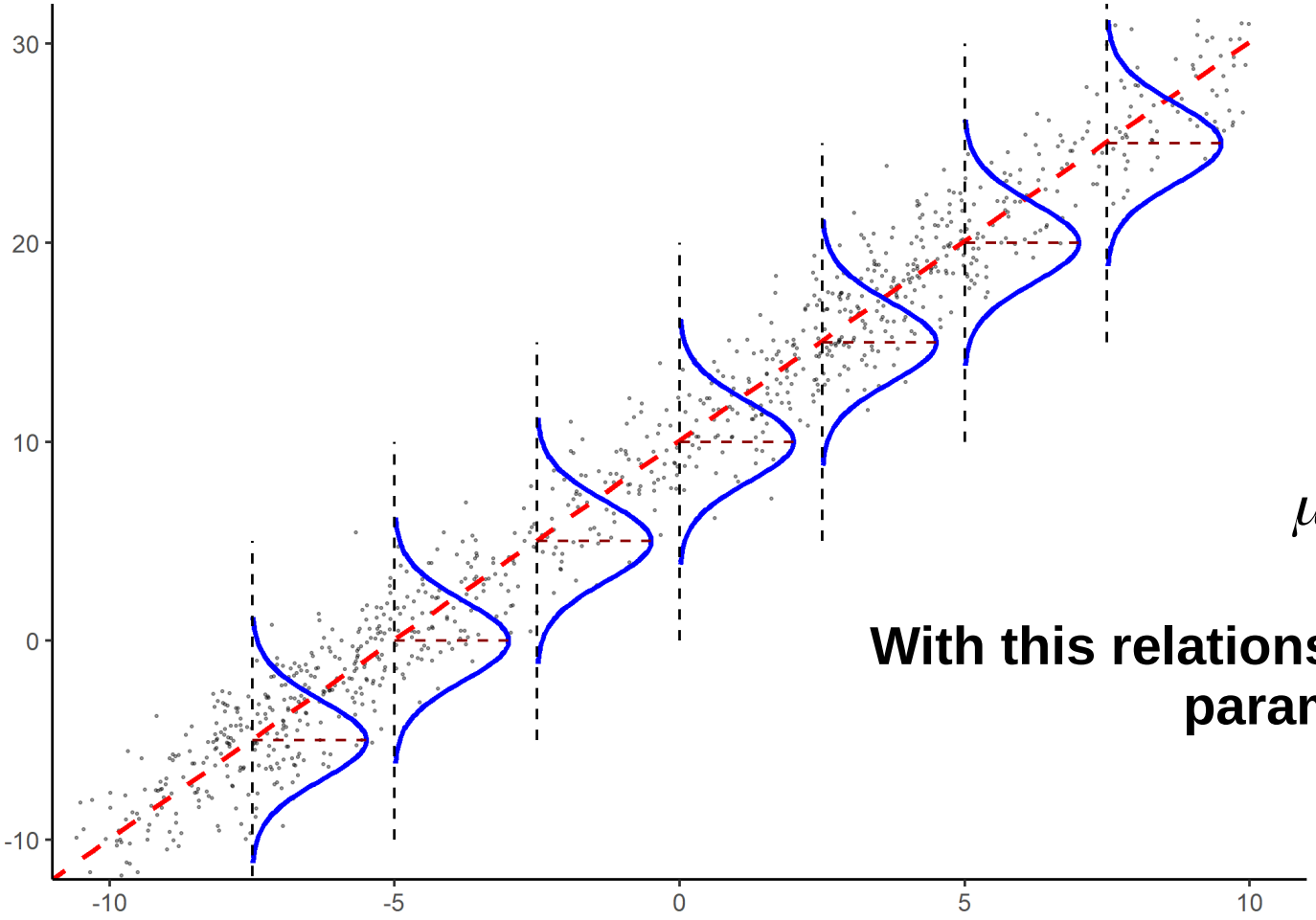
Big team project (Final assignment section in courseware**)**

- Start looking for 4 people team **NOW**
- Try to find a question/dataset that is interesting for you
- Look for previous year best projects for inspiration
- Ask question before the deadline for team creation (3.11)
- If no team on 3.11 (midterm week), then I will forcefully assign a team for you

Tips and recommendations:

- Do not focus on single question, try to combine two or more questions
  - Example: Not only look for medical question, but try to add socio-economy
- Try to find more than one dataset
  - Example: UNICEF + World Bank
- Start working early, at least to formulate a question

The linear regression is just a set of 1D distributions

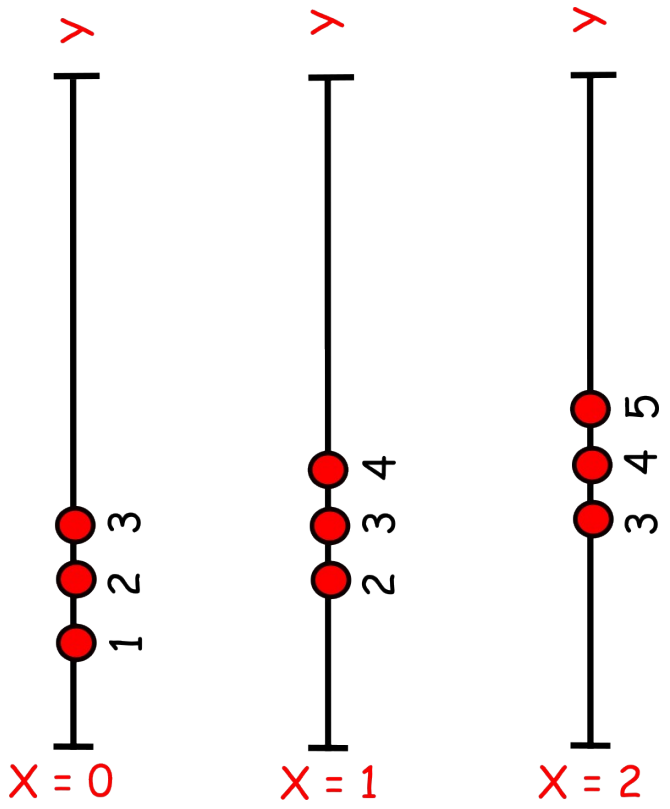


$$\mu_x = \beta_0 + X * \beta_1$$

With this relationship between mean parameters

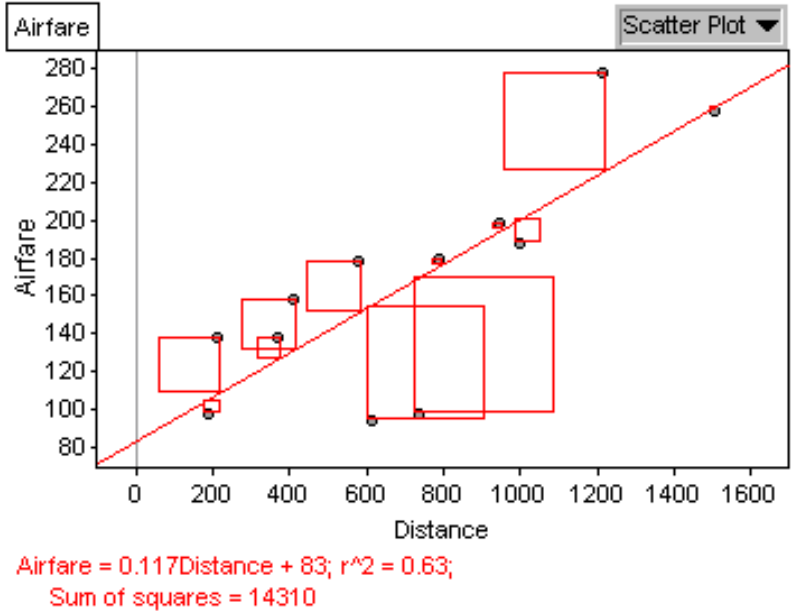
Practical assignment: manual computation of linreg by hand

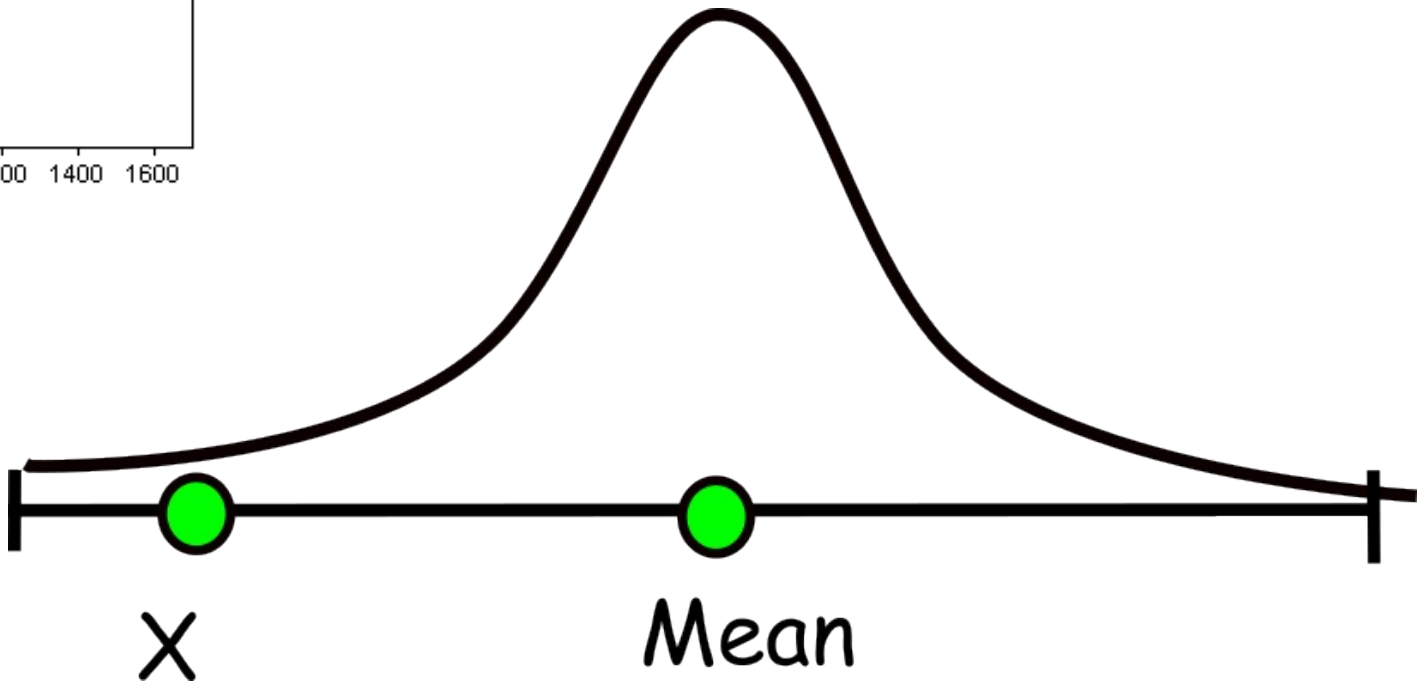
$$Linreg:$$
$$\beta_0 = 2$$
$$\beta_1 = 1$$
$$Baseline:$$
$$\beta_0 = 3$$



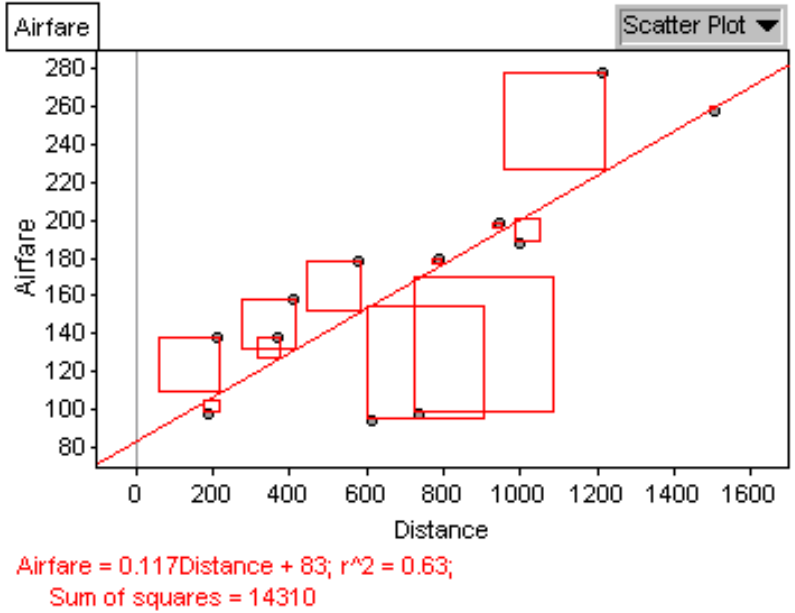
- Compute means for linreg model
- Compute means for baseline model
- Compute RSS
- Compute TSS
- Compute R^2

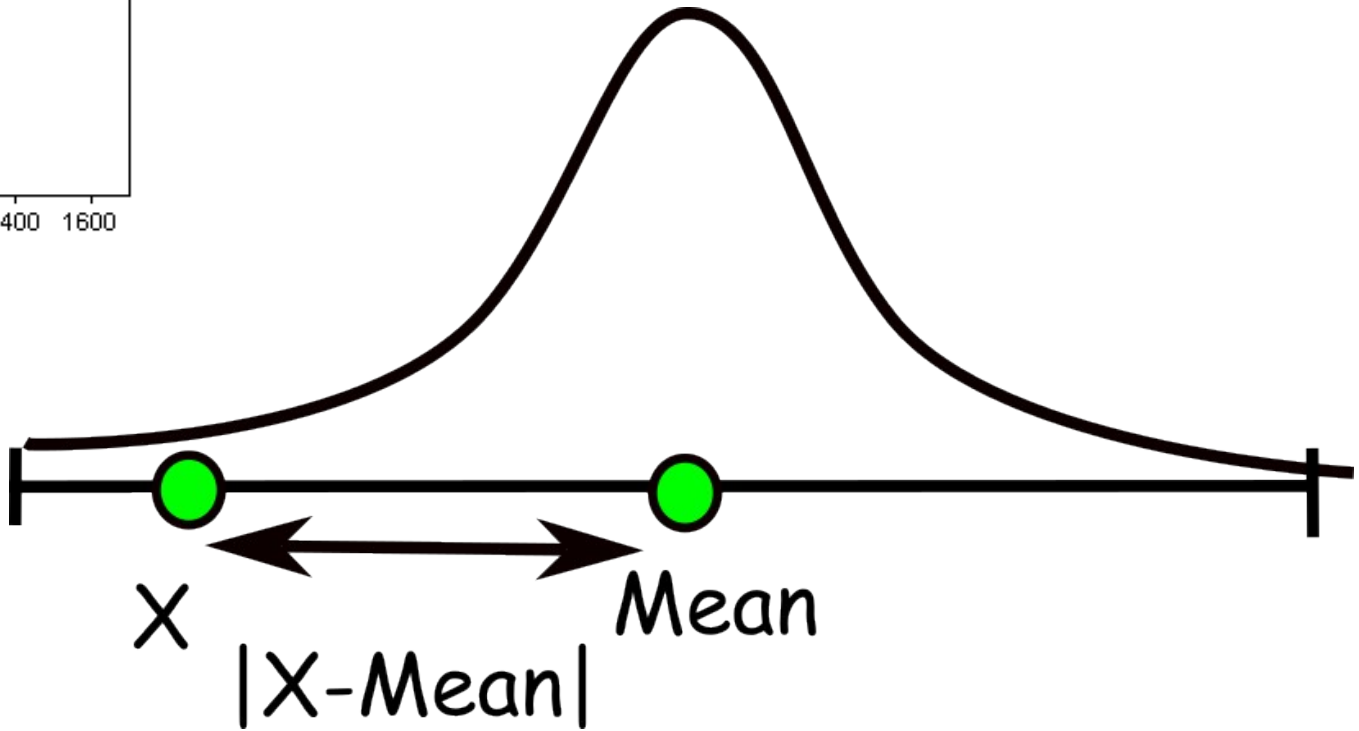
A little clarification of lectures: What are these boxes?



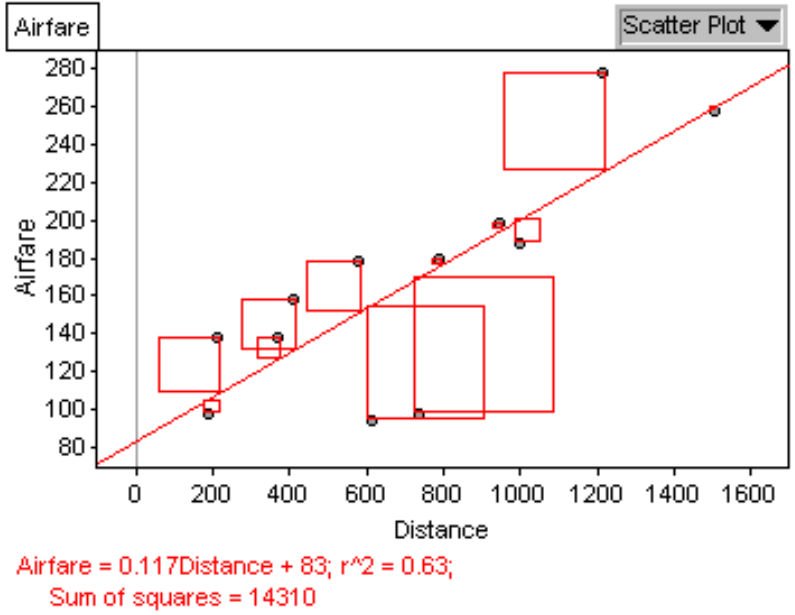


A little clarification of lectures: What are these boxes?

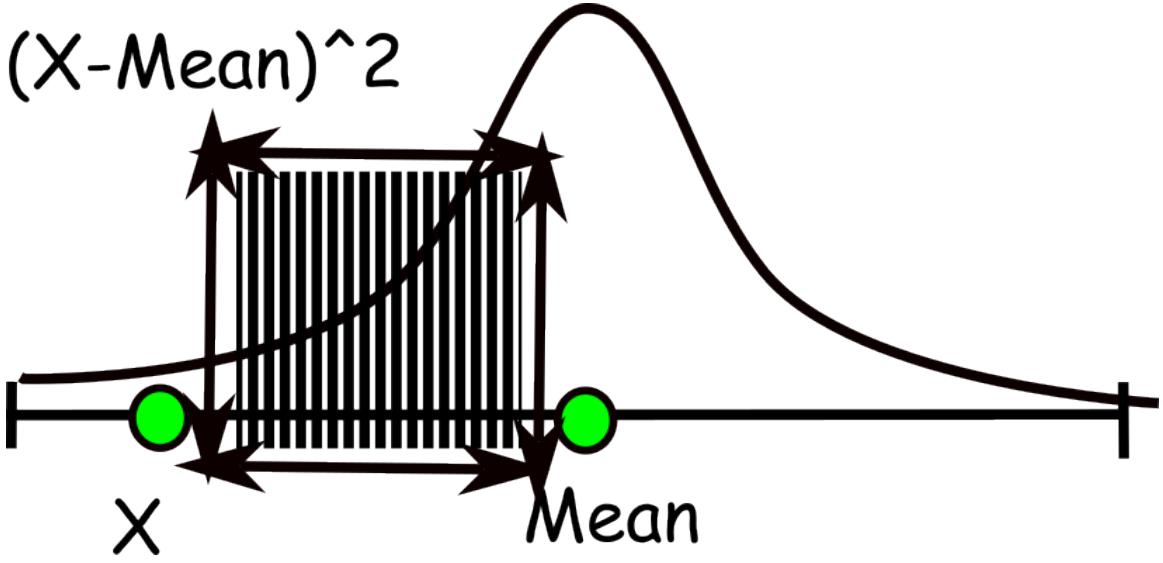




A little clarification of lectures: What are these boxes?



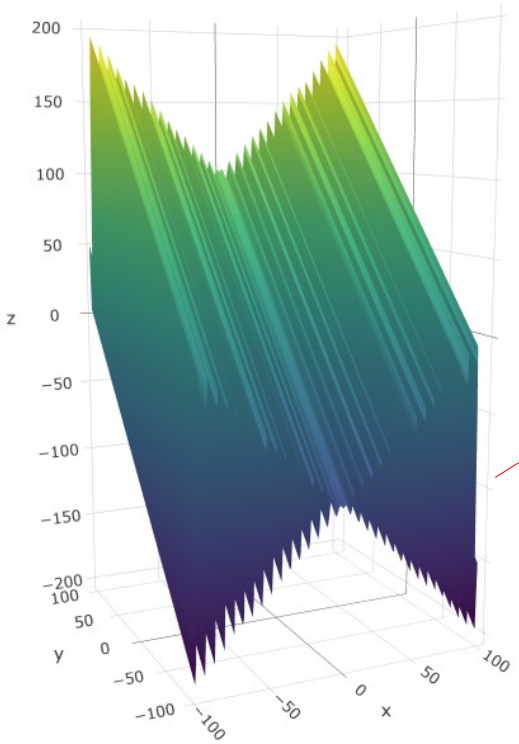
Geometrical visualization of amount of variance this particular point gives

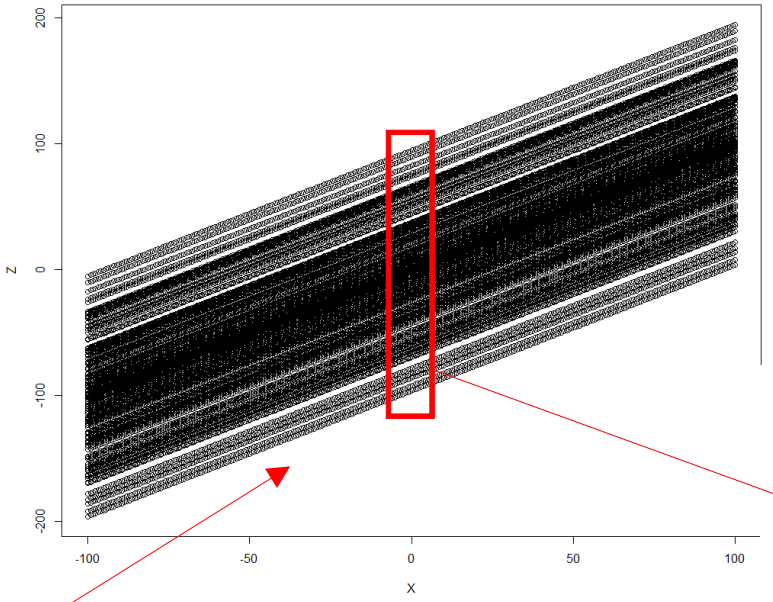


Now with the new topic

Remember the idea of randomness from hidden terms?

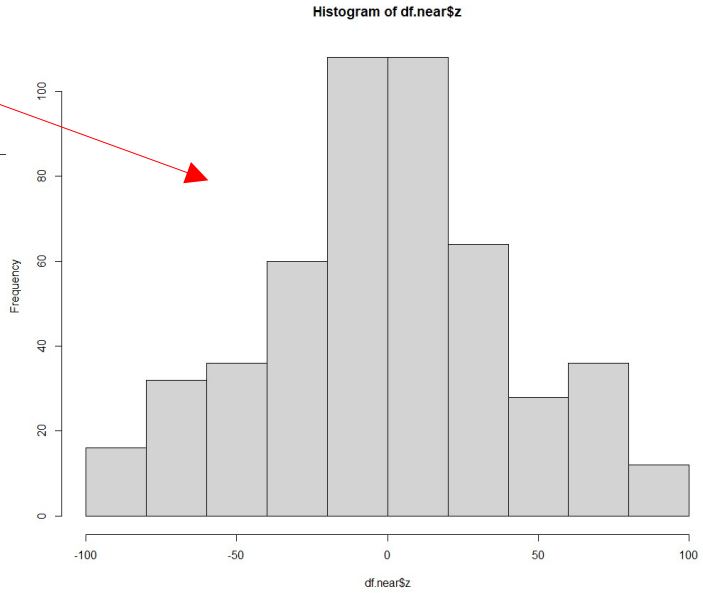
$$Z = X + Y * \sin(Y)$$

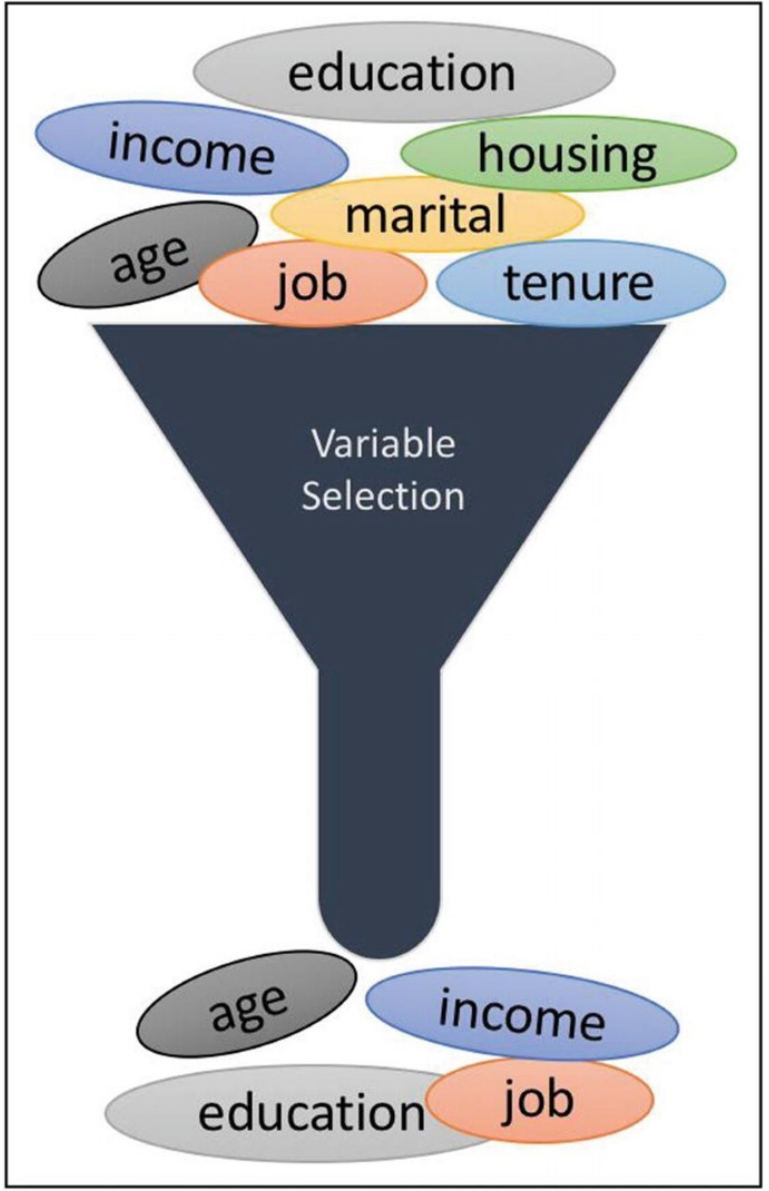




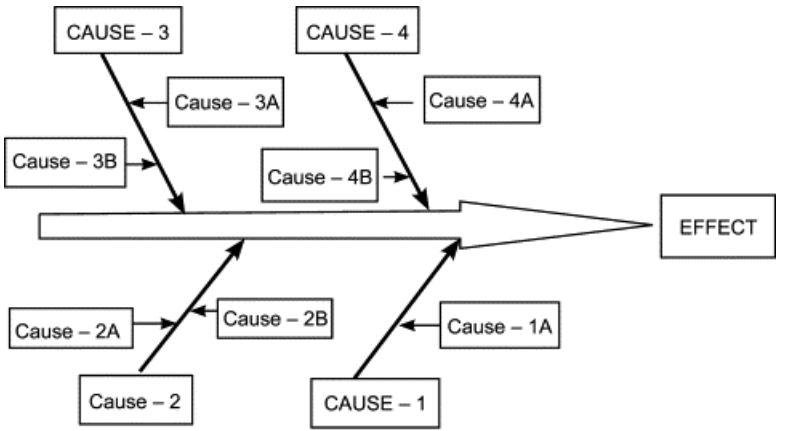
Hide Y axis

Look at some X-Z range:



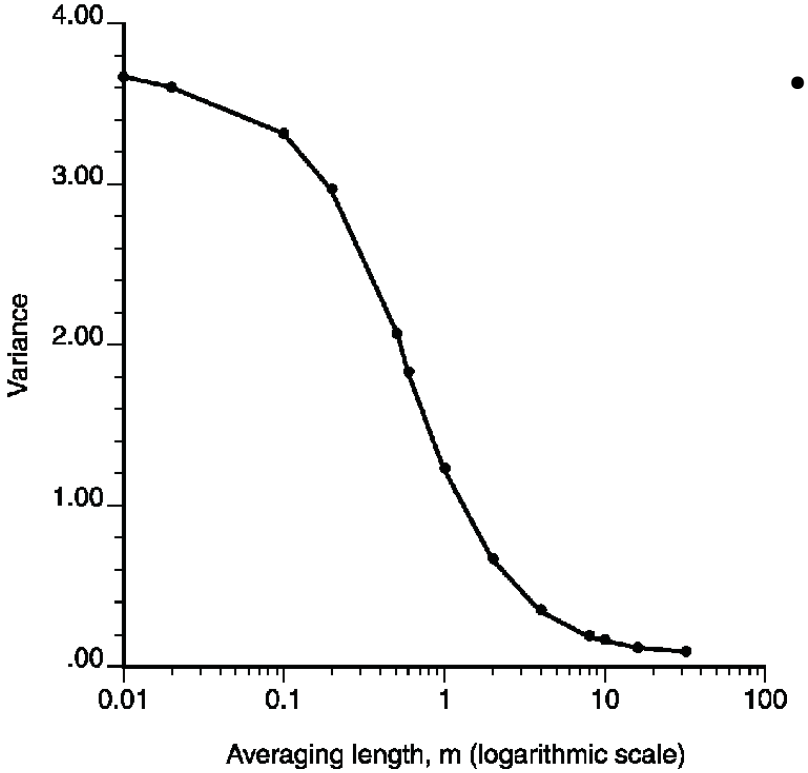


- Up until now the randomness was assumed to be as an absence of variables/terms and we wanted to use the rest to approximate the true formula
- However most of the time even measured variables are TOO MUCH to practically computations
- Even though they DO affect the outcome, the size of the effect is too low, we want to find **key factors**

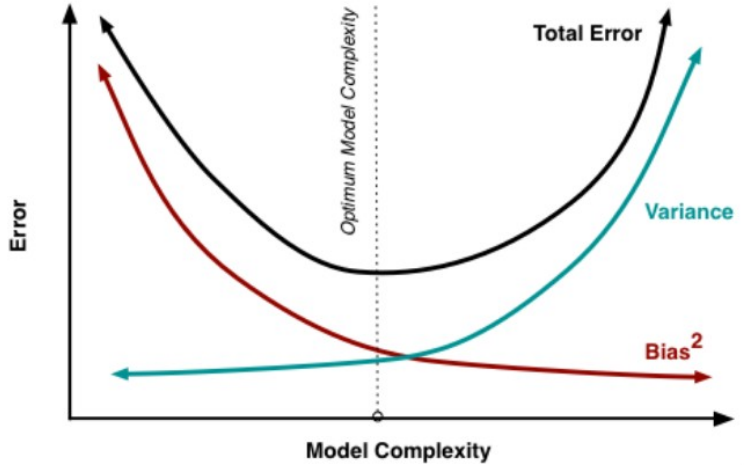


Cause-and-Effect Diagram

Variable selection

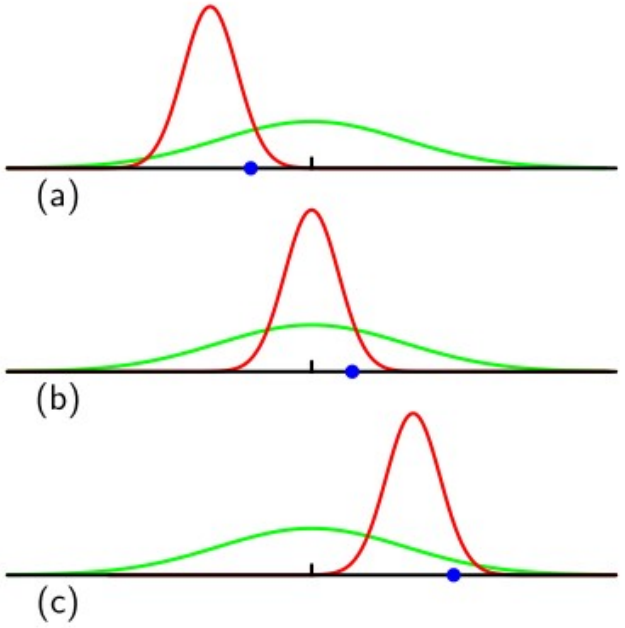


- You can think of this as following:
  - Addition of more variables makes the prediction more precise (smaller variance)
  - But decrease from 1.1% to 1.09% is too small and the computational complexity is spent for nothing
  - Moreover, with more variables we need more and more samples to maintain same variance
  - **Bias-variance Trade-off**



For example:

Figure 1.15 Illustration of how bias arises in using maximum likelihood to determine the variance of a Gaussian. The green curve shows the true Gaussian distribution from which data is generated, and the three red curves show the Gaussian distributions obtained by fitting to three data sets, each consisting of two data points shown in blue, using the maximum likelihood results (1.55) and (1.56). Averaged across the three data sets, the mean is correct, but the variance is systematically under-estimated because it is measured relative to the sample mean and not relative to the true mean.



https://stats.stackexchange.com/questions/136673/how-to-understand-that-mle-of-variance-is-biased-in-a-gaussian-distribution

Please open the today's activity .zip file and find the:

3	6.10.	JB, AA, JK	Shrinked linear regression	san_fs.zip, assignment1.zip, pres_3.pdf

LREG_artificial_irrelevant.R

You will find the activity (voluntary) questions at the end:

Summary

Irrelevant features cause overfitting and make our models work worse on unseen data. If having a finite/limited sample set the learning algorithm finds spurious relationships which increases variance and thus error. Removal of irrelevant features is crucial namely when dealing with a large number of them. (Proper) testing on unseen data (hold-out method, cross-validation) can help to detect overfitting and find out the optimal complexity of the model. ANOVA helps to decide the complexity from one run of the model only.

Further questions and tasks:

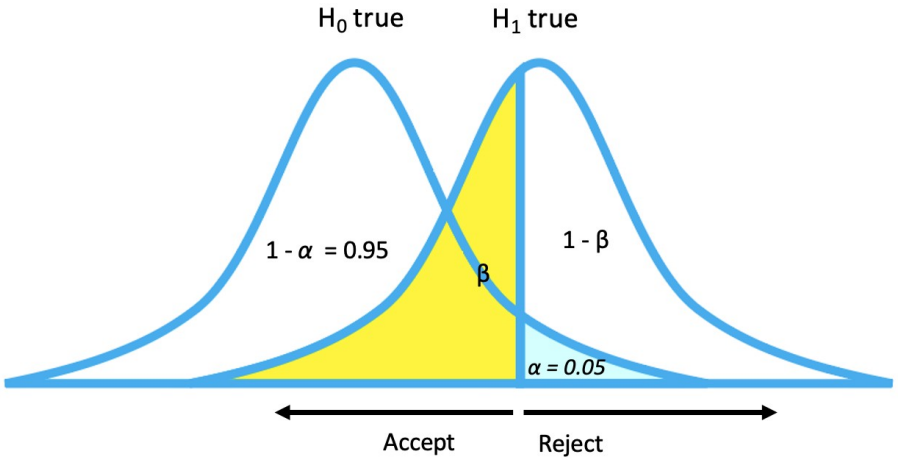
1. Show how the previously learned feature selection methods (p-values, stepwise selection, shrinkage) work in this case. Clearly demonstrate whether they work well/fail and explain why. Play with several different random generator seeds.
2. Describe step by step the correct way of model comparison procedure through cross-validation that was incorrectly implemented above. You can also implement the procedure, however, the implementation is optional only.

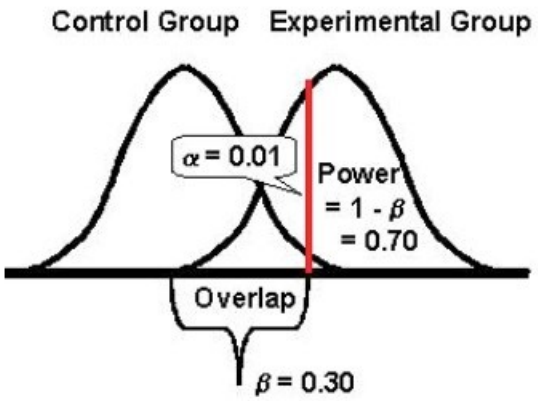
Deadline: 13.10.2025 (next seminar, voluntary, 1 point)

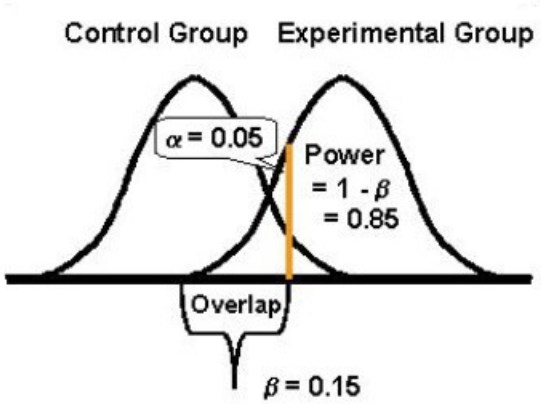
LREG_Boston_FS.R R code templates to be used (copy-pasted) in artificial_irrelevant

A recap of HW 1 (t-test): Type 1 vs Type 2 errors

By changing the “threshold” we can select needed ratio of
Type 1 to Type 2 errors



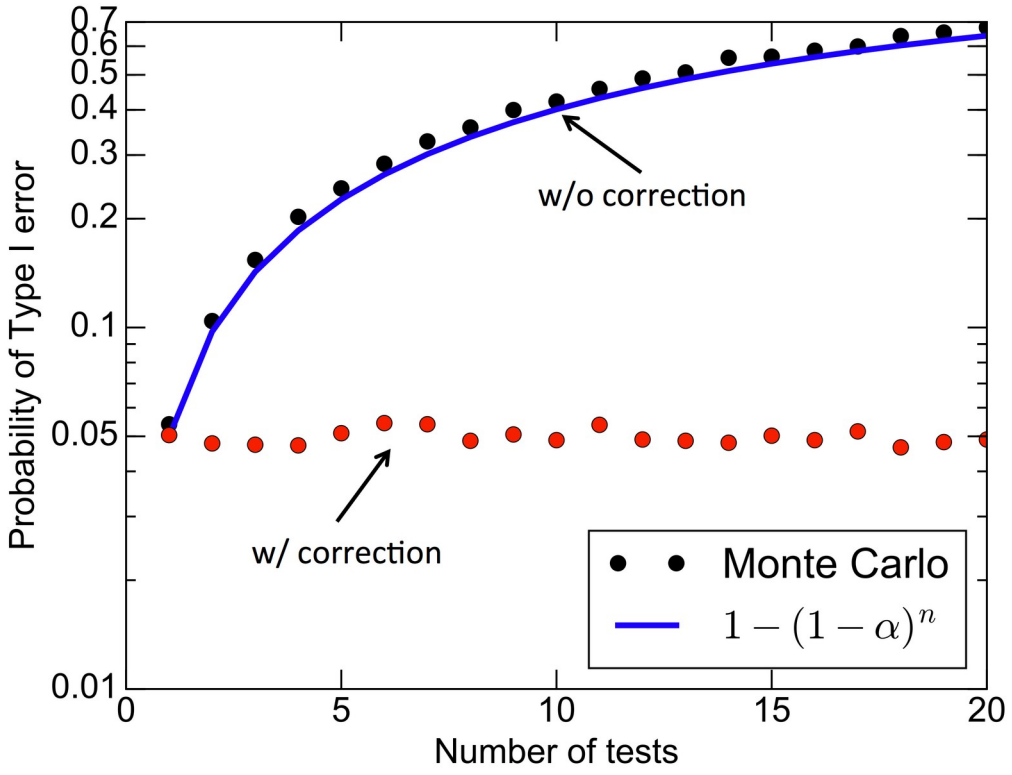




- To simplify the analysis we actually just fix one of errors, Type 1 = 0.05 = confidence
- After that we calculate the Type 2 and denote it as **power** of the test

https://youtu.be/OhDZlO8vtGw?si=yK76iwA0wftf2w4A&t=400

What is wrong with t-tests and how to overcome this issue?

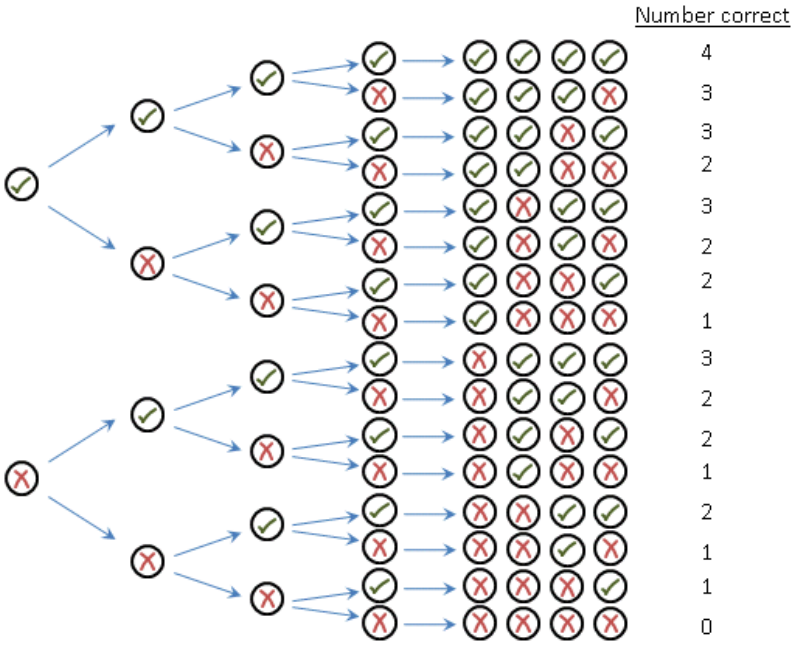


Multiple testing problem

If you still want to to t-test and p-values, you could perform **"confidence correction"** so that total Type 1 error is still 5% = 0.05
(for example setting alpha=0.0025 = 0.25%)

You should think of it this way:

- We have a fixed 5% error for 1 test
- Then with each additional test we have more combinations where at least one test will fail.
- But if even 1 variables fails, we have false subset of variables! Large Type 1!



Practical assignment: compute the adjusted p-value

Lets consider the linear regression problem:

- We have 3 variables: Y ~ X1, X2, X3
- We run the linear regression and compute t-test p-values:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1448     0.3816   0.380   0.7236
X1            0.3057     0.3201   0.955   0.3937
X2            0.1207     0.3914   0.308   0.7731
X3            1.4143     0.4130   3.424   0.0267 *
```

- We see one of them as significant (one star *), so it passes alpha=0.05
- But consider the issue of multiple testing from before:
  - We need **total_alpha = 0.05**
  - What will be the alpha per each variable?
  - Will any of X variables pass such corrected threshold?

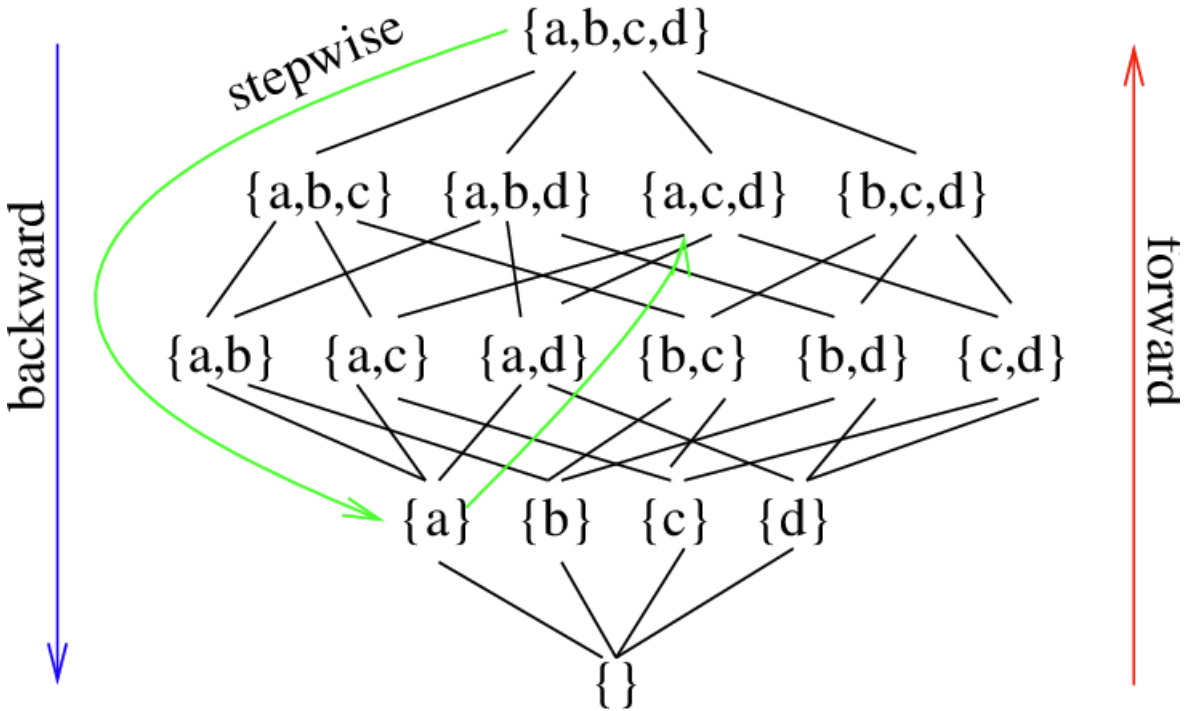
$$1 - (1 - \alpha)^n$$

We will almost always reject and get empty model

Is there an alternative?

Alternative solution = instead of looking at p-values locally, evaluate the entire model globally

Straightforward solution = just evaluate a subset of variable globally and change it. Check all subsets
A little **heuristic** = greedily check subsets based on score until it increases and stop afterwards



But how to evaluate a linear regression model?

Recall: objective value used in linear regression

$$RSS = e_1^2 + e_2^2 + \cdots + e_m^2 = \sum_{i=1}^{m} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- **Problem with directly using the value – increase between models has a scale that is data–dependent, e.g., we are not sure whether the difference is significant:**

  Try to guess – is X2 addition significant or not based on RSS decrease?

  - **RSS (X1+X2) = 9000 RSS (X1 only) = 10000**
    RSS(X1+X2) = 0.1 RSS (X1 only) = 10

Recall: objective value used in linear regression

$$RSS = e_1^2 + e_2^2 + \cdots + e_m^2 = \sum_{i=1}^{m} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- **Problem with directly using the value – increase between models has a scale that is data–dependent, e.g., we are not sure whether the difference is significant:**

  Try to guess – is X2 addition significant or not based on RSS decrease?

  - **RSS (X1+X2) = 9990 RSS (X1 only) = 10000 TSS = 10001**
    RSS(X1+X2) = 0.1 RSS (X1 only) = 10 TSS = 1000

    Whatever you assumed, you may be wrong

First option – objective value used in linear regression + normalization

- This is a way to measure how good we are in comparison to baseline
- RSS = our variance
- TSS = variance of baseline model
- Decrease of minimization value in percentage

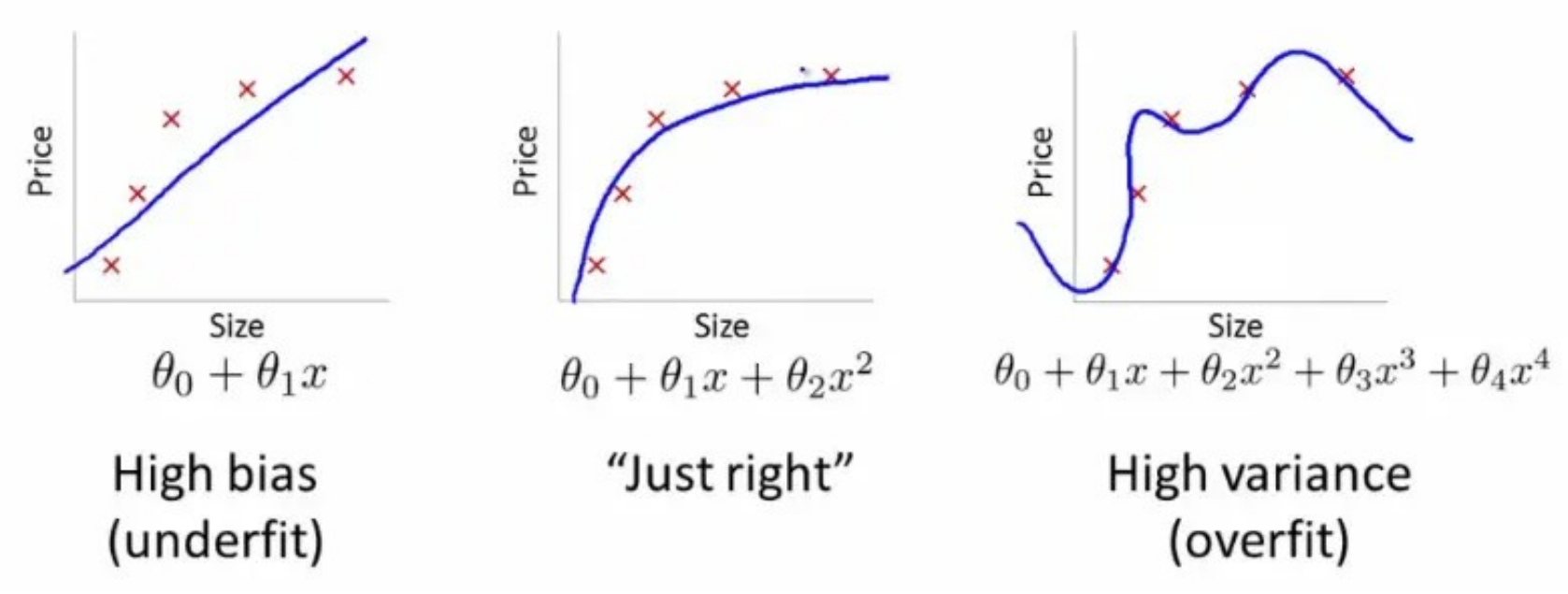
- **R-squared** gives the fraction of variance explained by the model

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

  - where $TSS = \sum_{i=1}^{m} (y_i - \bar{y})^2$ stands for the total sum of squares,
  - and $RSS = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$ stands for the residual sum of squares,

Problem: additional variables always decrease the RSS

$$RSS(\theta_1) > RSS(\theta_1+\theta_2) > RSS(\theta_1+\theta_2+\theta_3+\theta_4)$$



https://www.mlwithramin.com/blog/bias-variance-tradeoff

Use Adjusted R^2 or AIC/BIC, what is the difference?

The key difference between adjusted R^2 and AIC/BIC:

- **Adjusted R^2 is used only in Gaussian linear regression**
- **AIC/BIC can be used for nearly all models/algorithms**
  - We will show such advanced method in future seminars
- They are very similar, but adjusted R^2 tends to choose more complex models

$$R^2 = 1 - RSS/TSS$$

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

$$AIC = -2\ln(L) + 2k$$

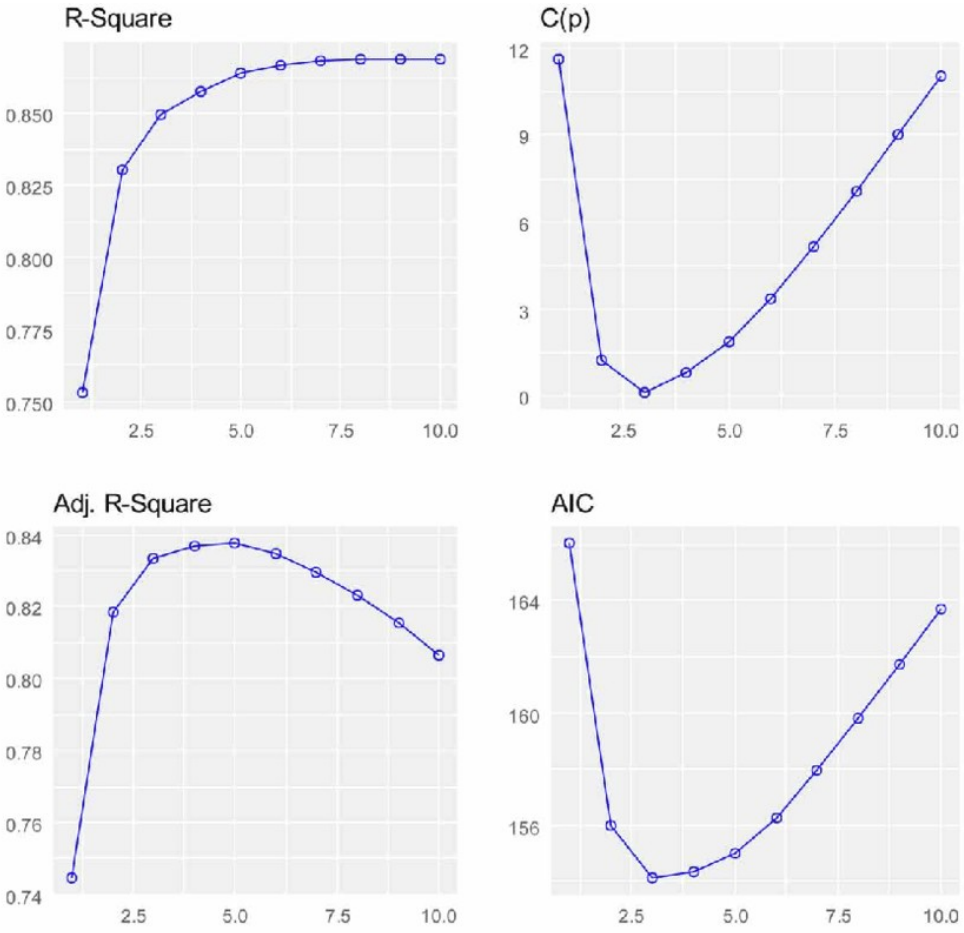


Figure 4

Again, the problem with objective value

$$RSS = e_1^2 + e_2^2 + \cdots + e_m^2 = \sum_{i=1}^{m} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- **Problem with directly using the value – increase between models has a scale that is data–dependent, e.g., we are not sure whether the difference is significant:**

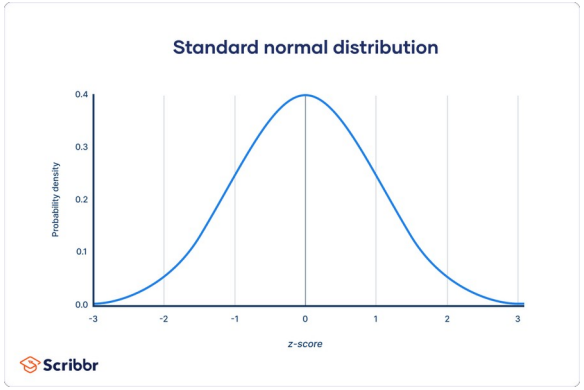
  Try to guess – is X2 addition significant or not based on RSS decrease?

  - **RSS (X1+X2) = 9000 RSS (X1 only) = 10000**
    RSS(X1+X2) = 0.1 RSS (X1 only) = 10

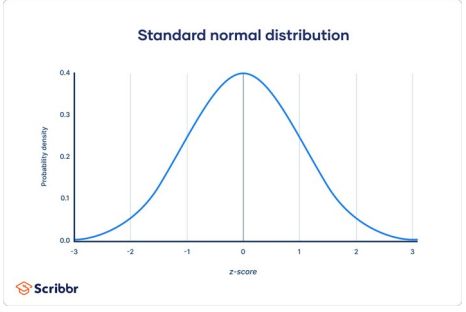
Solution 2: convert to formal hypothesis testing (F-statistics)!

Recap: where do statistics come from?

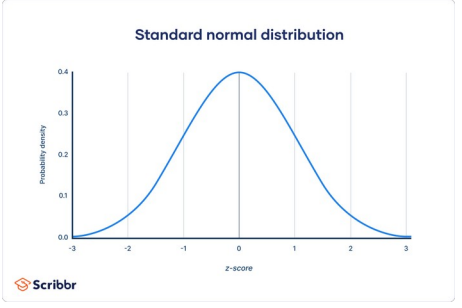
X1, X2, …., ,Xn ~



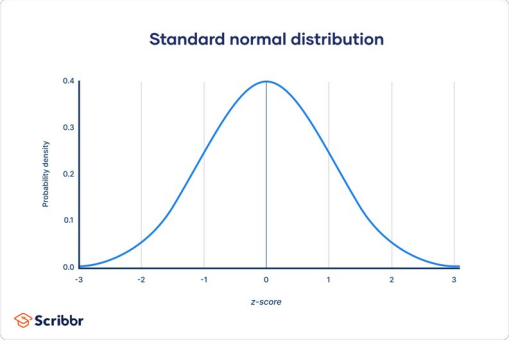
X1 + X2 + …. + Xn ~



+

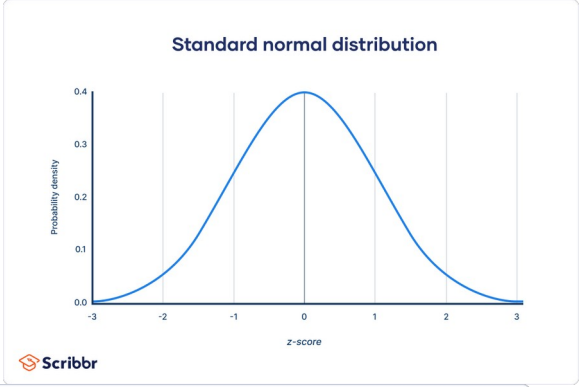


=

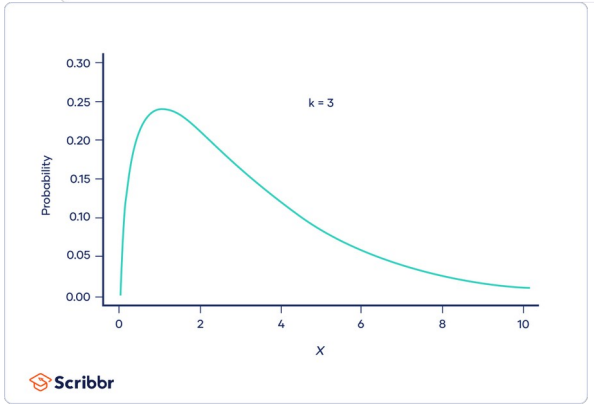


Recap: where do statistics come from?

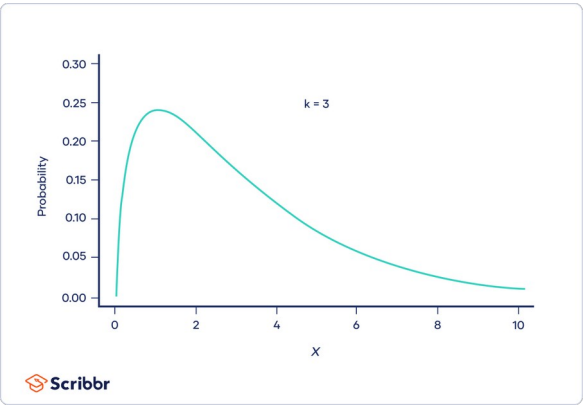
a, b, ……., w ~



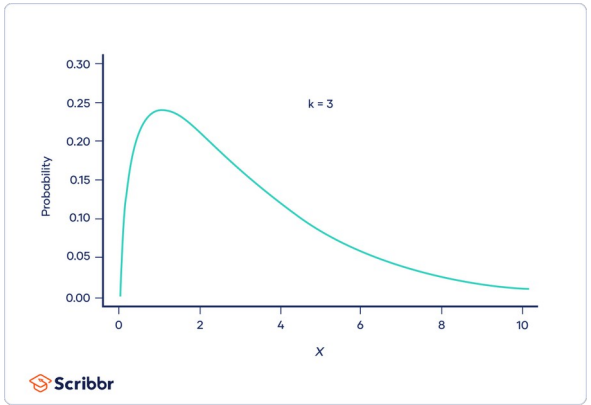
a^2, b^2, ……., w^2 ~



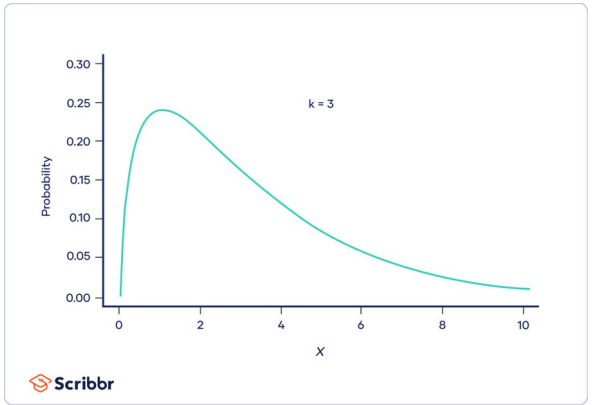
a^2 + b^2 + ……. + w^2 ~



+

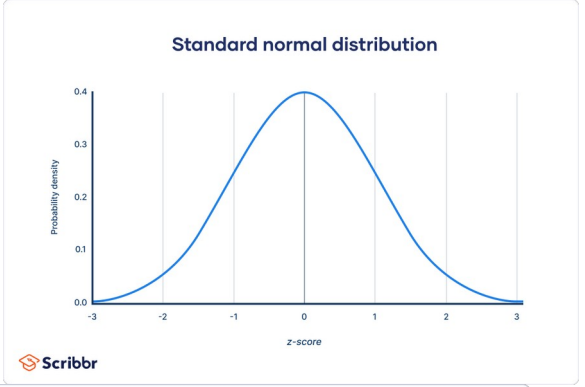


=

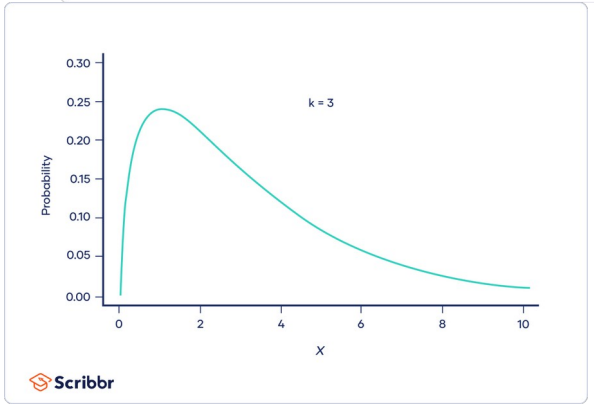


Recap: where do statistics come from?

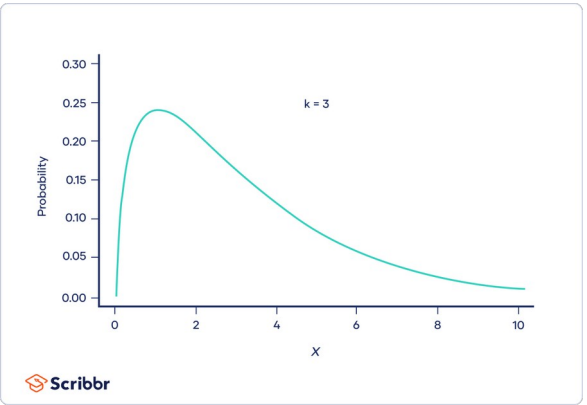
a, b, ……., w ~



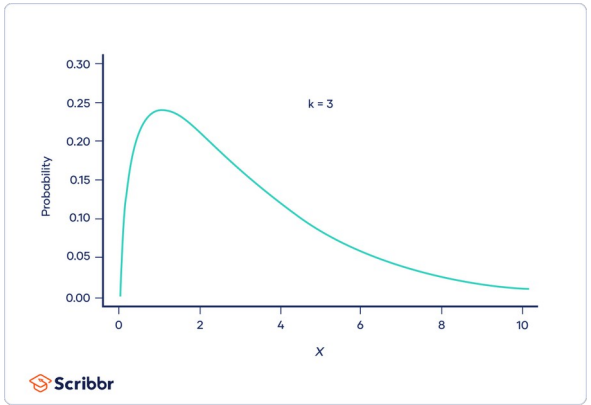
a^2, b^2, ……., w^2 ~



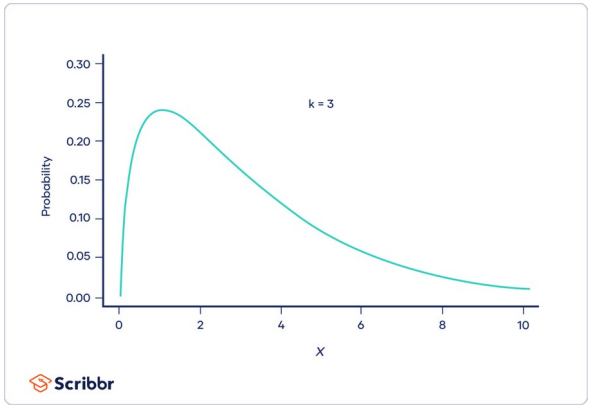
a^2 + b^2 + ……. + w^2 ~



+

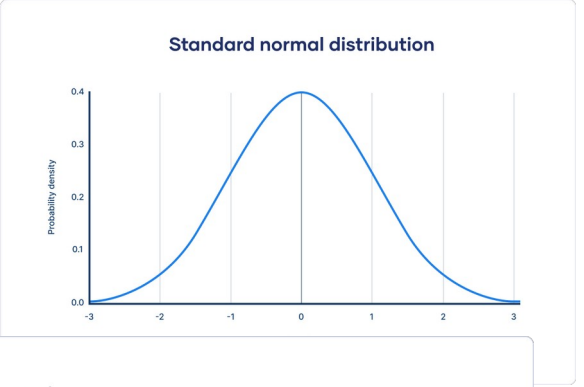


=



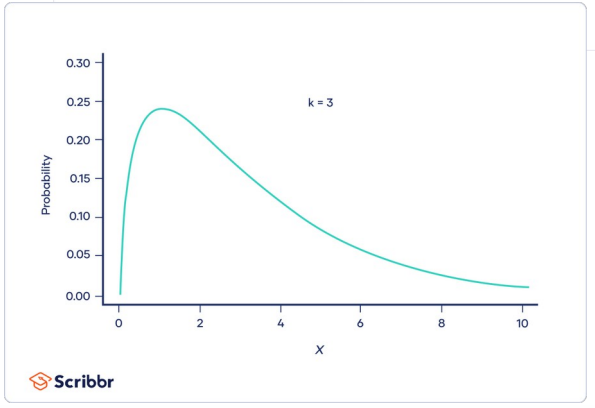
Recap: where do F-statistics come from?

$$x_1, x_2, \ldots, x_N, y_1, \ldots, y_M \sim$$

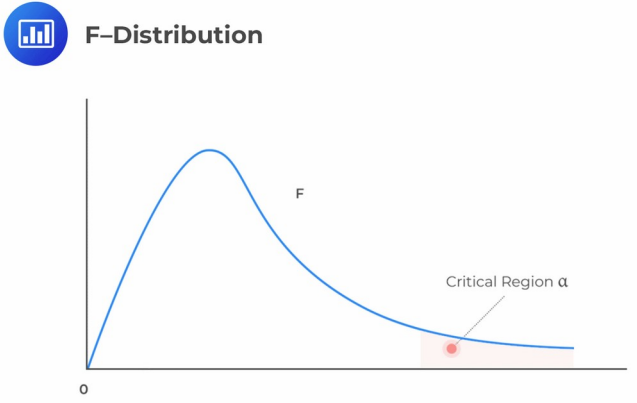


$$RSS_X = x_1^2 + x_2^2 + \ldots + x_N^2 \sim$$

$$RSS_Y = y_1^2 + y_2^2 + \ldots + y_M^2 \sim$$



$$\frac{RSS_x}{RSS_Y} = \frac{x_1^2 + x_2^2 + \ldots + x_N^2}{y_1^2 + y_2^2 + \ldots + y_M^2} \sim$$



Second option – objective value + hypothesis testing

$$F = \frac{s_1^2}{s_2^2}$$

$$where\ s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Only for **nested models:**
E.g. second model is addition of some variables

Model1: Y ~ X1 + X2
Model2: Y ~ X1 + X2 + X3 + X4+...

$$ANOVA\left(model_1, model_2\right) = \frac{F-stat_1}{F-stat_2} = \frac{S_1}{S_0} \div \frac{S_2}{S_0} = \frac{S_1}{S_2} = F-stat$$

Formally:

p-value < 0.05 ----------> model2 is better! (adjusted R^2 model2 >> adjusted R^2 model1)

p-value > 0.05 ----------> model1 is better! (adjusted R^2 model2 << adjusted R^2 model1)

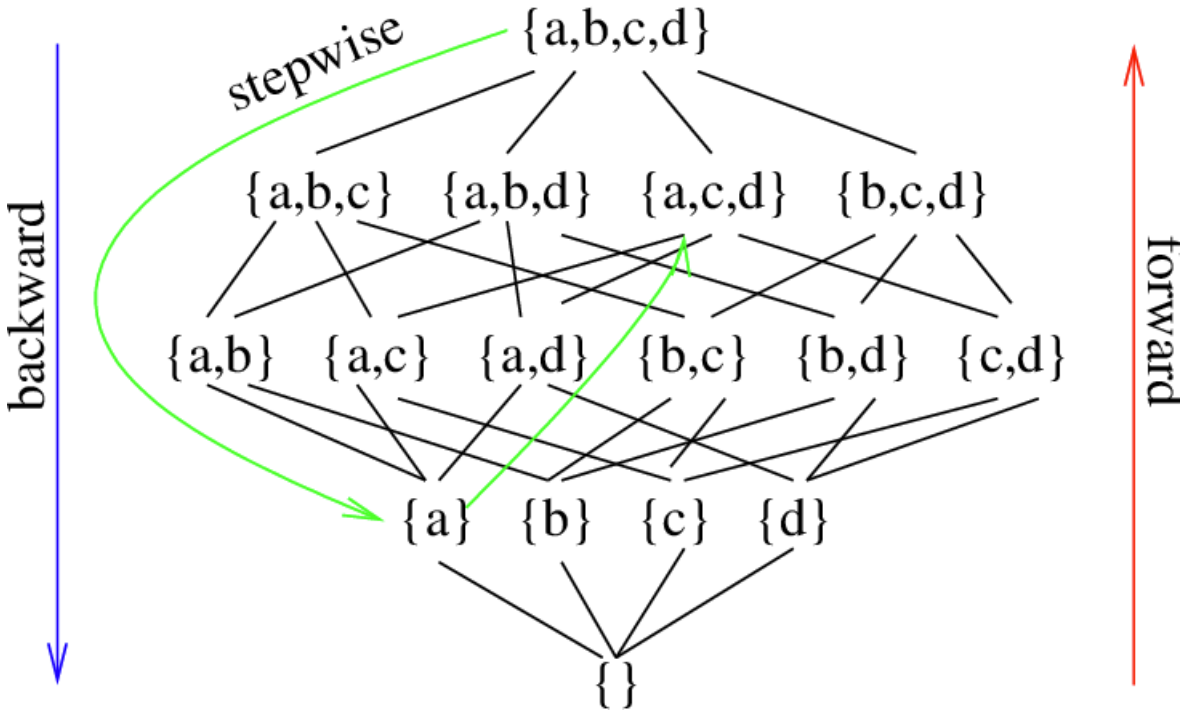
In R:

$$Y \sim X_1,$$
$$Y \sim X_1 + X_2,$$
$$Y \sim X_1 + X_2 + X_3, \ldots$$

$$= anova\left(model_1, model_2\right), anova\left(model_2, model_3\right), \ldots$$

Use these metrics in Forward/Backward stepwise

Straightforward solution = just evaluate a subset of variable globally and change it. Check all subsets
A little **heuristic** = greedily check subsets based on score until it increases and stop afterwards



Just score each subset and move one change at the time

But people realized that subset selection is not too good

Here are *some* of the problems with stepwise variable selection.

1. It yields R-squared values that are badly biased to be high.
2. The *F* and chi-squared tests quoted next to each variable on the printout do not have the claimed distribution.
3. The method yields confidence intervals for effects and predicted values that are falsely narrow; see Altman and Andersen (1989).
4. It yields *p*-values that do not have the proper meaning, and the proper correction for them is a difficult problem.
5. It gives biased regression coefficients that need shrinkage (the coefficients for remaining variables are too large; see Tibshirani [1996]).
6. It has severe problems in the presence of collinearity.
7. It is based on methods (e.g., *F* tests for nested models) that were intended to be used to test prespecified hypotheses.
8. Increasing the sample size does not help very much; see Derksen and Keselman (1992).
9. It allows us to not think about the problem.
10. It uses a lot of paper.

“All possible subsets” regression solves none of these problems.

Current modern solution: Shrikage with Ridge and LASSO

$$RSS = \sum_{\vec{x}} (y_i - x_1 \cdot \beta_1 - x_2 \cdot \beta_2 - \beta_0)^2 \quad \rightarrow \quad local\ optimum\,(estimate) : min\ RSS \rightarrow \quad \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_0$$

$$RSS \text{ and regularization} \quad \rightarrow \quad local\,optimum\,(estim) : min\,(RSS + Lasso) \quad \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_0$$

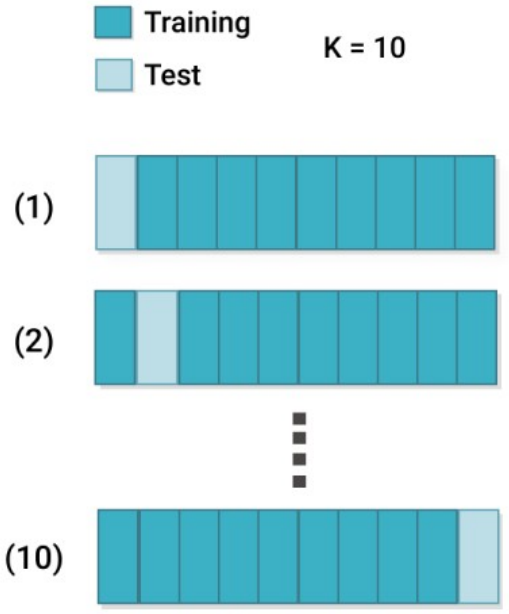
$$regularization = \lambda \sum \beta_i^2 (Ridge) \quad \text{or} \quad = \lambda \sum |\beta_i| (Lasso) \qquad Hyperparameter\ \lambda !$$

> **Ridge** (glmnet alpha = 0) typically **only makes parameters small**(~0), but has **better performance in case of correlation/multicollinearilty**

> **LASSO** (glmnet alpha = 1) typically **removes variables completely**(=0), but **is worse with correlation/multicol**

> Standardization of variables needed (normalize intervals so that all X are same, e.g. Age 1-40, Height 120-190)

> Learn lambda with Cross-Validation technique



Choose a set of lambdas: $\lambda \in [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}]$

1) Split train into 10 folds (or you can split inside for loop)

2) For every lambda λ :

For every possible validation folds:

2.1) min [RSS + λ * regulariz] = fit on 9/10 train folds

2.2) Compute MSE = RSS on 1/10 test fold

3) Average MSE test errors from 2.2) for every lambda

4) Choose the best lambda by validation score from 3)

> K-fold for small subset (can run multiple times, too little dataset = use processing power)

> Hold-out for large subset (if my model or data are huge, just run once)

More details:

https://medium.com/analytics-vidhya/regularization-and-cross-validation-how-to-choose-the-penalty-value-lambda-1217fa4351e5

- **Today a big homework will be presented now and the countdown starts**
  - **You will have ~ 2 weeks to complete it, the time is parallel to every-week small HW**

3	7.10.	JB, AA, JK	Shrinked linear regression	san_fs.zip, assignment1.zip, lab3.zip, Linear regression.pptx

Deadline: 20.10.2025 (two weeks, full points)
27.10.2025 (three weeks, half points)