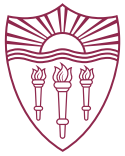


SUPERVISED LEARNING

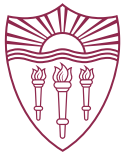
REGRESSION

Kristina Lerman

USC Information Sciences Institute

DSCI 552 – Spring 2021

Topics this week

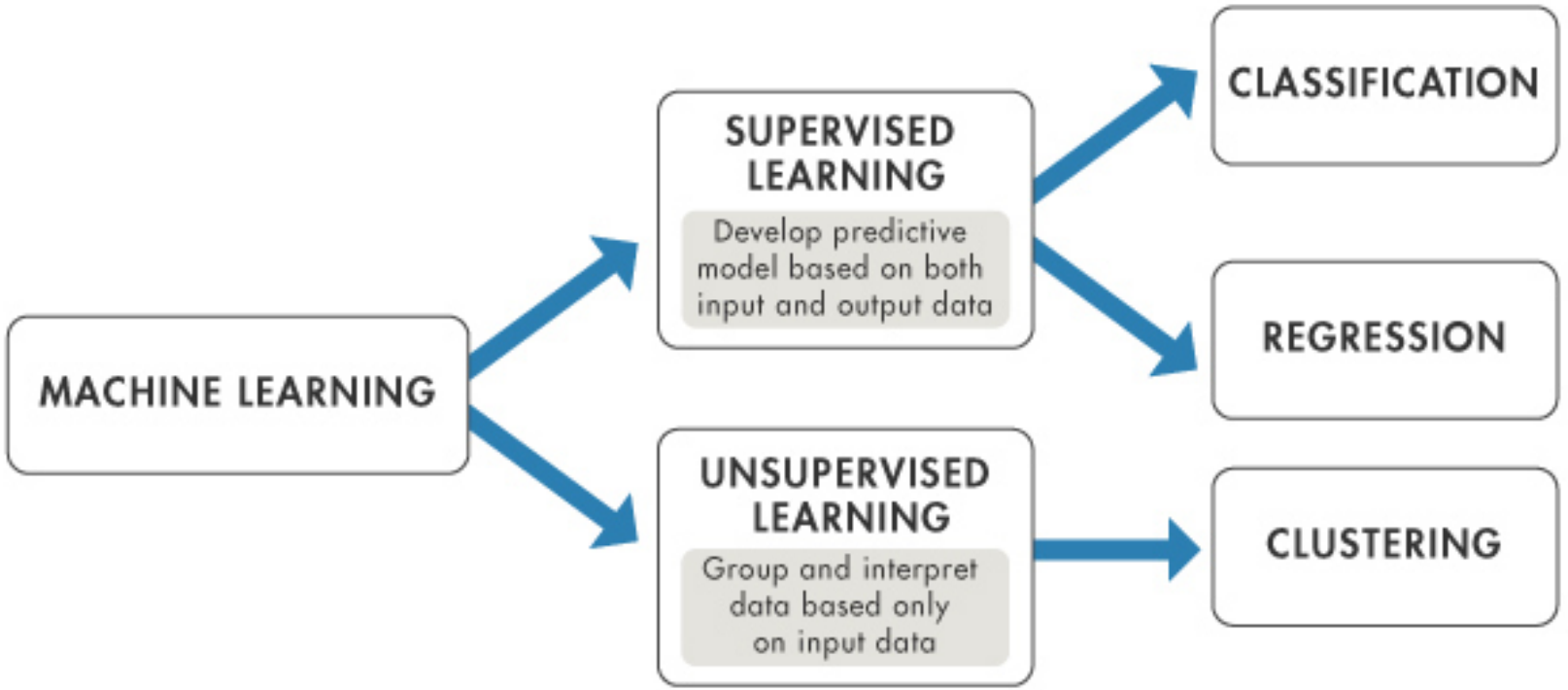


- Reminders:
  - Quiz 1 due today
  - Workplan due Thursday

- Linear Regression
- Mixed Effects Models
  - Demo: Python notebook for politeness study
- Regularization

Supervised vs unsupervised learning

Based on sample/training data and their given class labels or categories, is it possible to train a model that generalizes over unseen data to decide what class the sample belongs to?

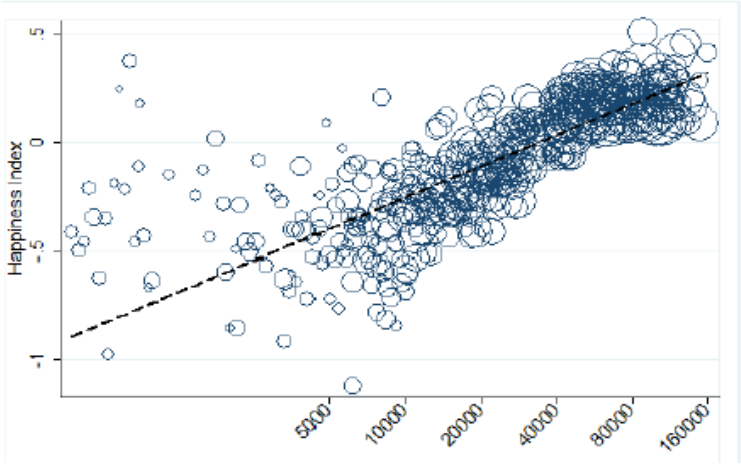


Output:**label**
- Binary
- Categorical

Output:
response/outcome
- Binary
- Real

Source: DeepAI.org

Why regression?



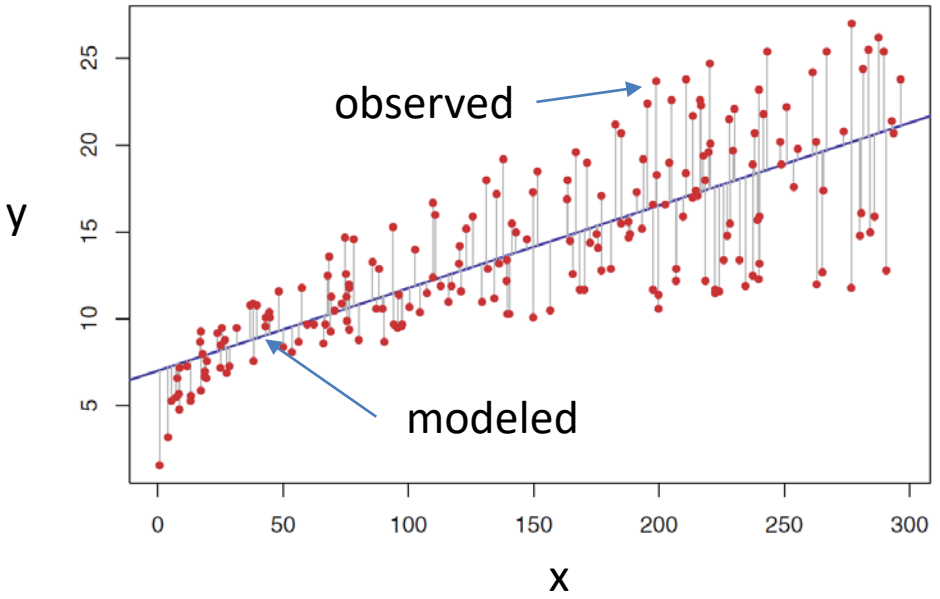
Source: US General Social Survey

- (Linear) regression models data as a line
  - Increasing independent variable x increases (or decreases) the outcome y
- Powerful and flexible tool for understanding the world
  - Creates an **interpretable** model that **explains** the data

Linear regression

- Parametric model: $\mathbf{y} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{\beta}_0$
  - **X**: observed features
  - **y**: observed response (outcome)
  - Model parameters β_0 and β_1 estimated from data
- Prediction: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Residual error: $e_i = y_i - \hat{y}_i$
- Residual sum of squares (RSS)

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$



Mathematical intuition: OLS

- RSS: $\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots$$

- Choose parameters that minimize RSS. *Ordinary least squares* coefficient estimates

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

$$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$$

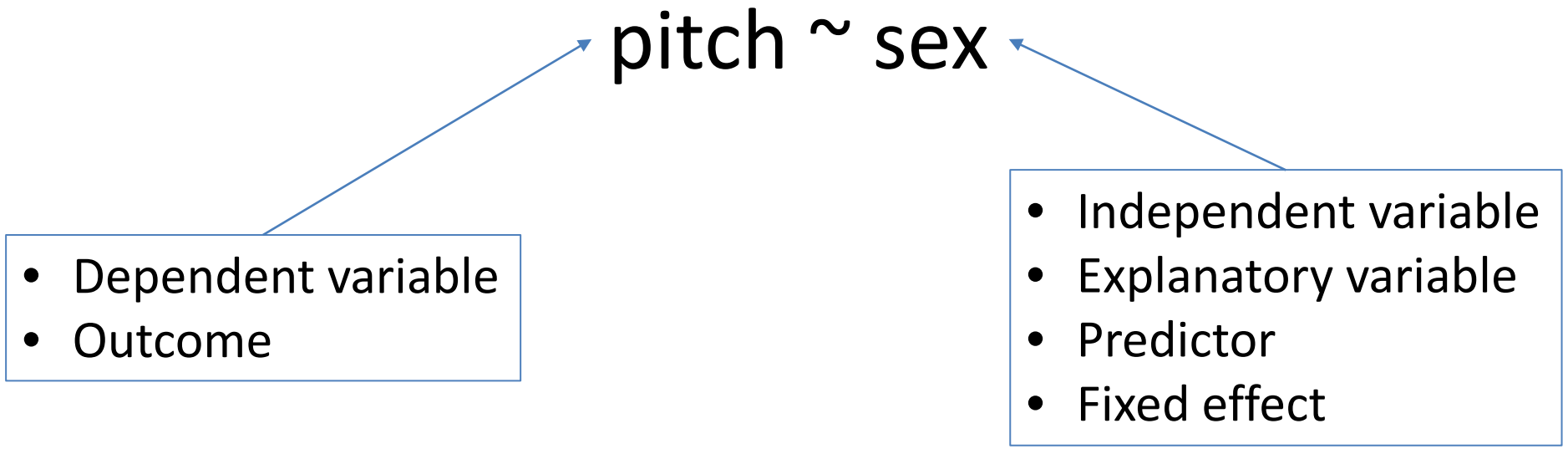
- are sample means

Creating a model

- *RQ: How much does the voice pitch of males and females differ?*
- Collect data
  - Population of men and women saying the word “mama”
  - Measure pitch, Bigger Hz → higher pitch
- Could the differences arise purely by chance?
  - Perhaps men and women have similar pitch
  - Experimenter just got unlucky
- Regression
  - Typical values for the pitch of men/women
  - Confidence about how likely these values are

Subject	Sex	Voice.Pitch
1	Female	233 Hz
2	Female	204 Hz
3	Female	242 Hz
4	Male	130 Hz
5	Male	112 Hz
6	Male	142 Hz

A simple model



A simple model

- Many other unmeasurable factors could affect pitch (culture, personality, age, nerves, …) – variations among individuals

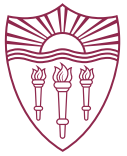
$$pitch \sim sex + \varepsilon$$

- Dependent variable
- Outcome
- Response

- Independent variable
- Explanatory variable
- Predictor
- Feature
- Fixed effect

- Error term
- Random factors

Fitting model to the data



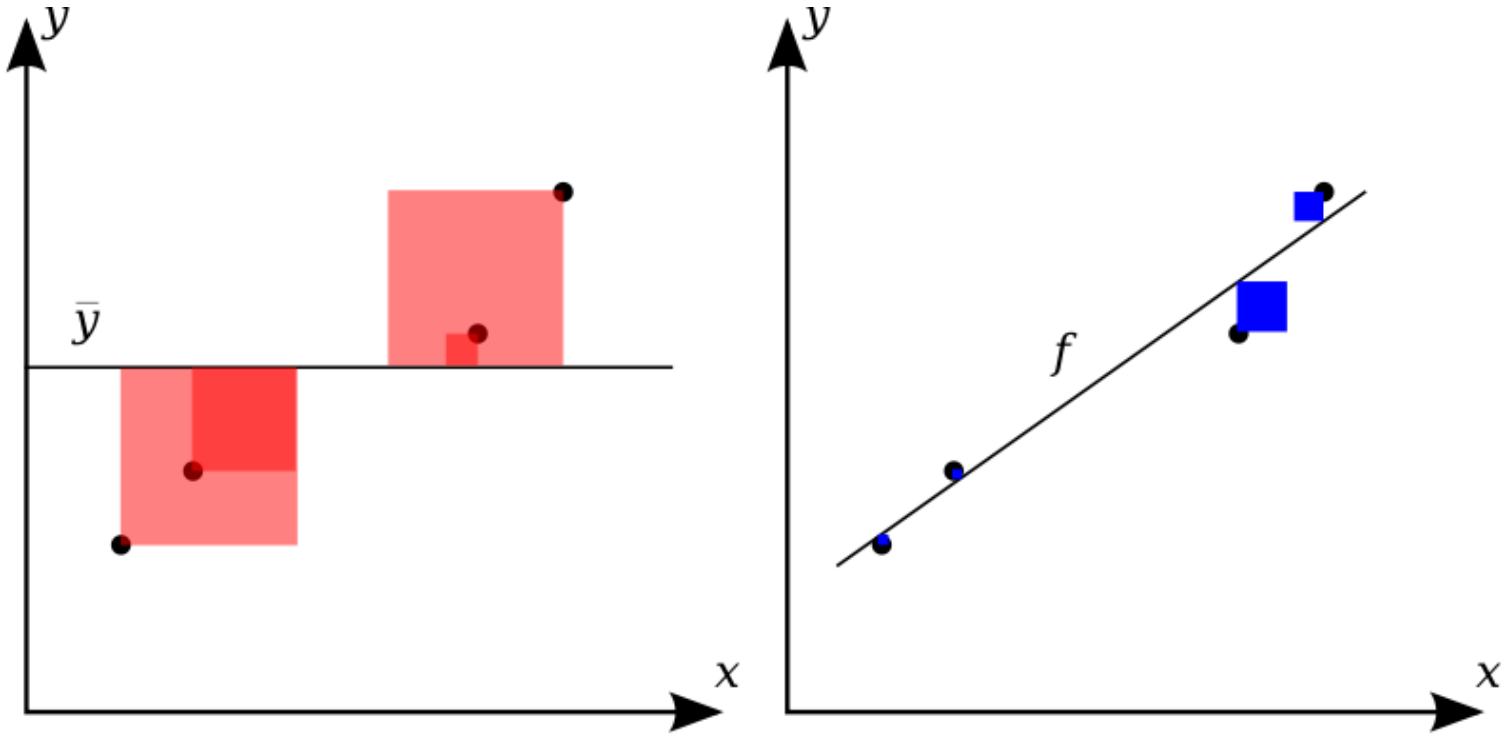
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   226.33      10.18  22.224 2.43e-05 ***
sexmale       -98.33      14.40  -6.827  0.00241 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.64 on 4 degrees of freedom
Multiple R-squared: 0.921,
  Adjusted R-squared: 0.9012
F-statistic: 46.61 on 1 and 4 DF,  p-value: 0.002407
```

R-squared reflects how much variance in data is accounted for by differences between males and females.

R2 = 92% of variance explained. Closer to 1 is better, but realistically, it won’t be that high.

Coefficient of determination – R2

- How much of the variation in y is explained by the variation in x?



$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Significance – p-value

- Could a null hypothesis, rather than regression, explain the data?
- Null hypothesis: sex has no effect on pitch

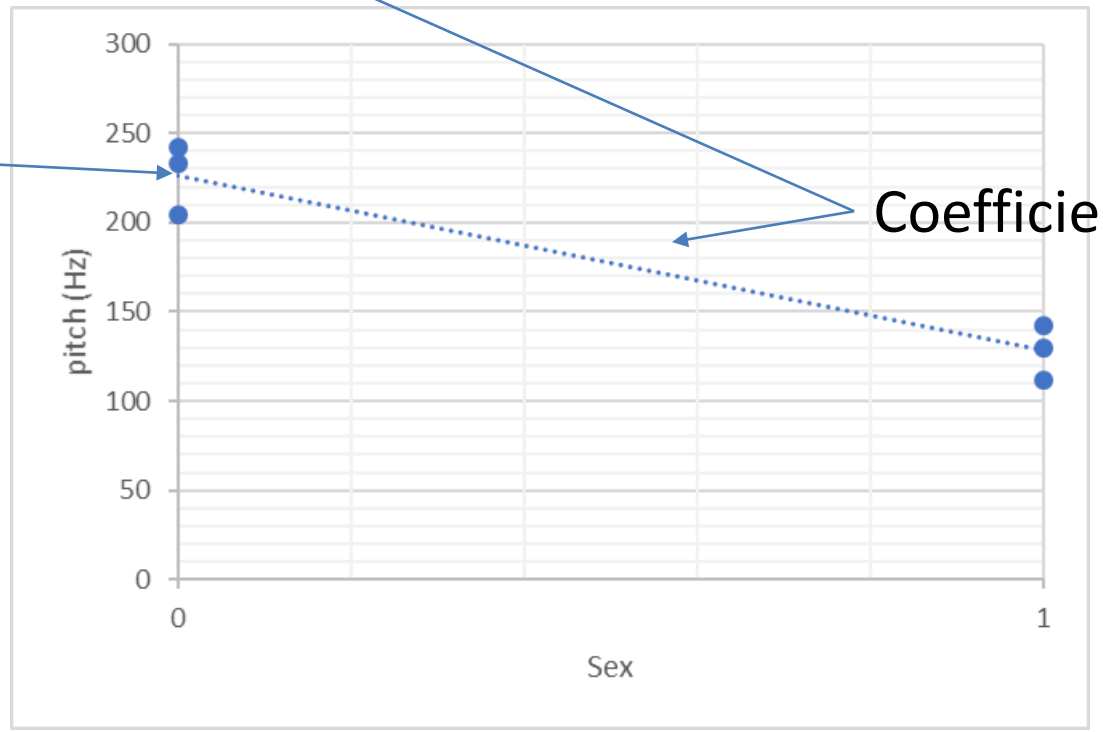
- P-value: probability our data could be observed given that the null hypothesis is true
  - With p-value = 0.002, this probability is very low
  - Then the alternative hypothesis “Sex affects pitch” is more likely.

  → The regression result is statistically significant

Fitting a model to data

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   226.33      10.18  22.224 2.43e-05 ***
sexmale       -98.33      14.40  -6.827  0.00241 **
```

intercept

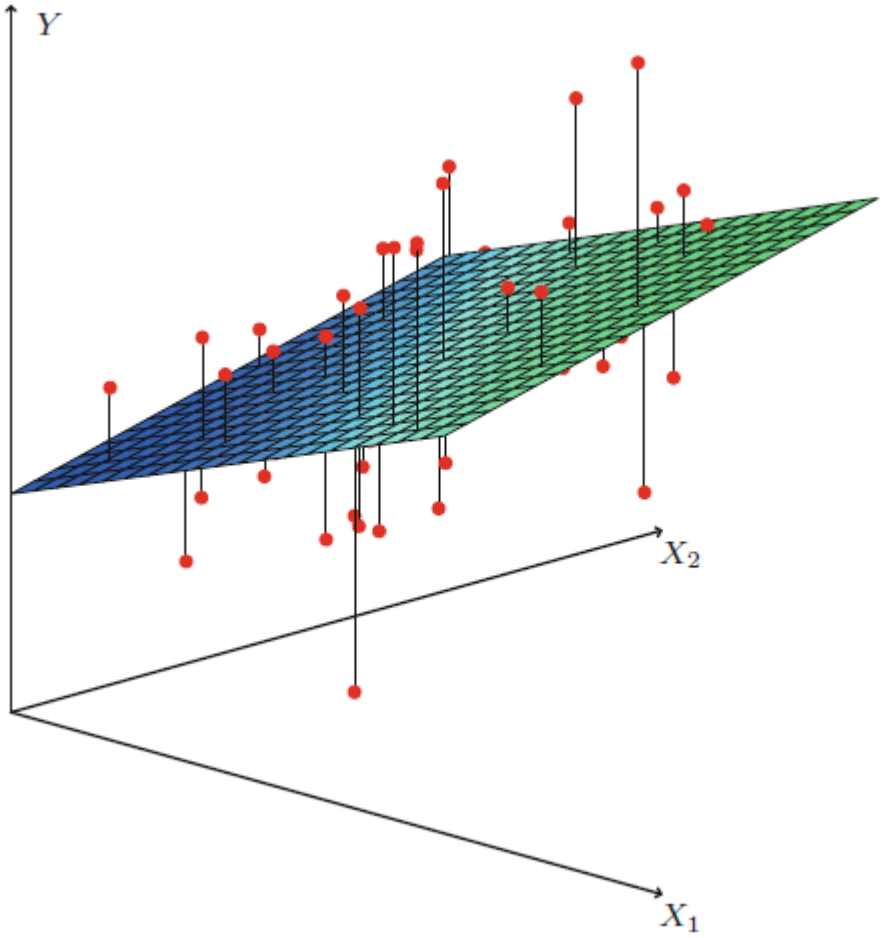


Coefficient (slope)

Multiple regression model

$$\text{pitch} \sim \text{sex} + \text{age} + \text{dialect} + \varepsilon$$

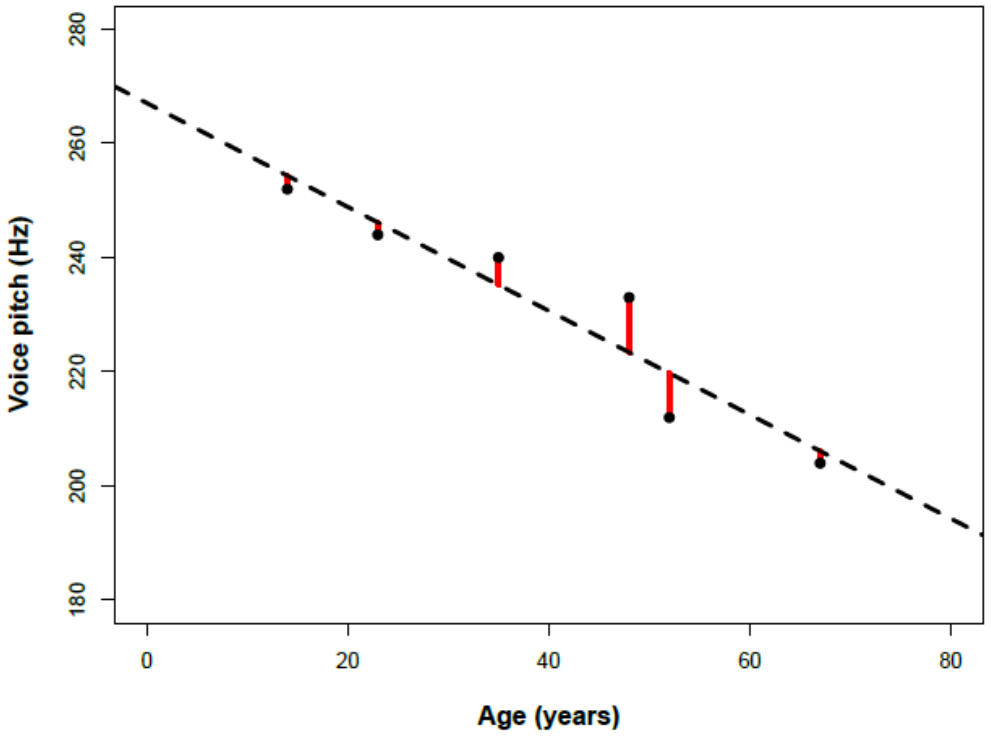
Linear multi-dimensional model



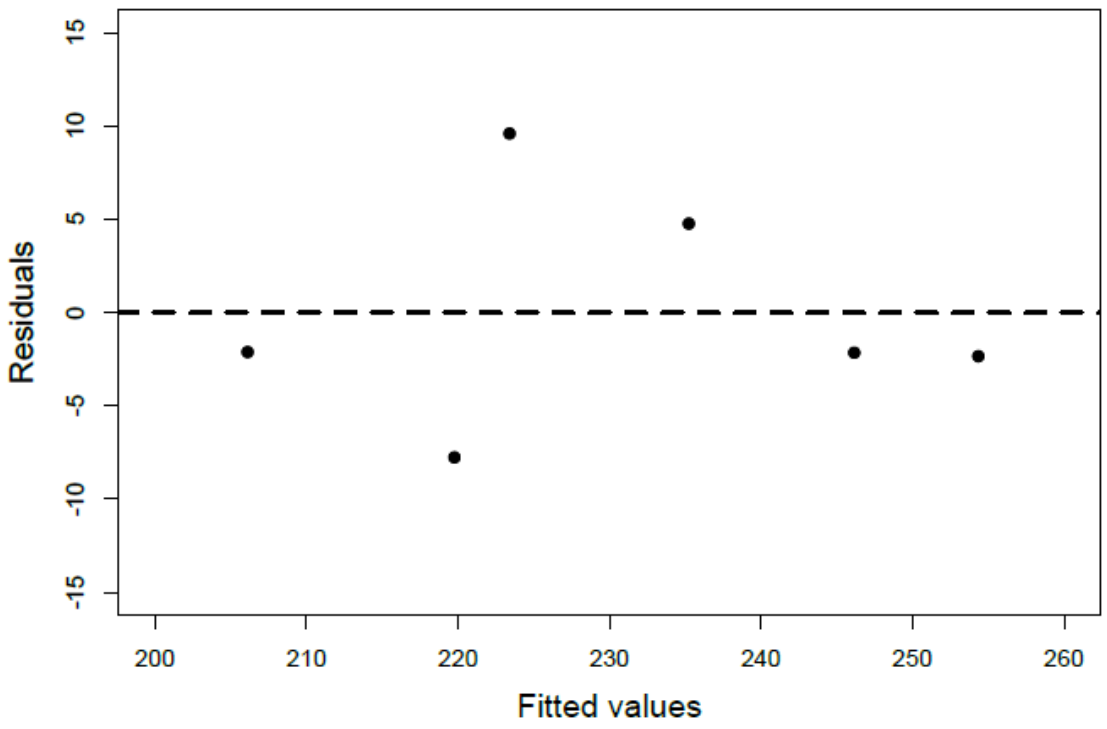
Assumptions of linear models

- Linearity – dependent variable is assumed to be a linear combination of the independent variables/features

Errors (residuals)



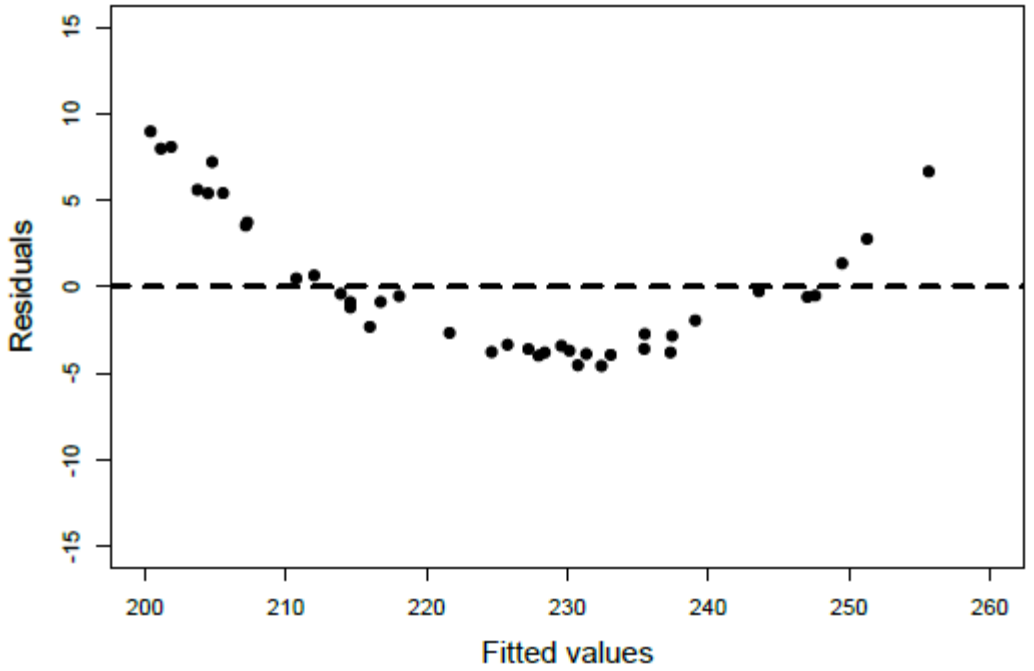
Residual plot



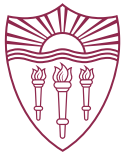
Check the residuals!

- Did you miss an important fixed effect that interacts with whatever fixed effects you already have in your model?
- Perform a nonlinear transformation of your outcome, e.g., by taking the log-transform.
- Perform a nonlinear transformation of the fixed effects. E.g., if age were related to pitch in a U-shaped way, then you could add age and age2 (age-squared) as predictors.
- If you're seeing stripes in the residual plot, then you're most likely dealing with some kind of categorical data – use a different class of models, such as logistic models

If your residuals look like this, the linearity assumption is violated



Assumptions of linear models: absence of collinearity



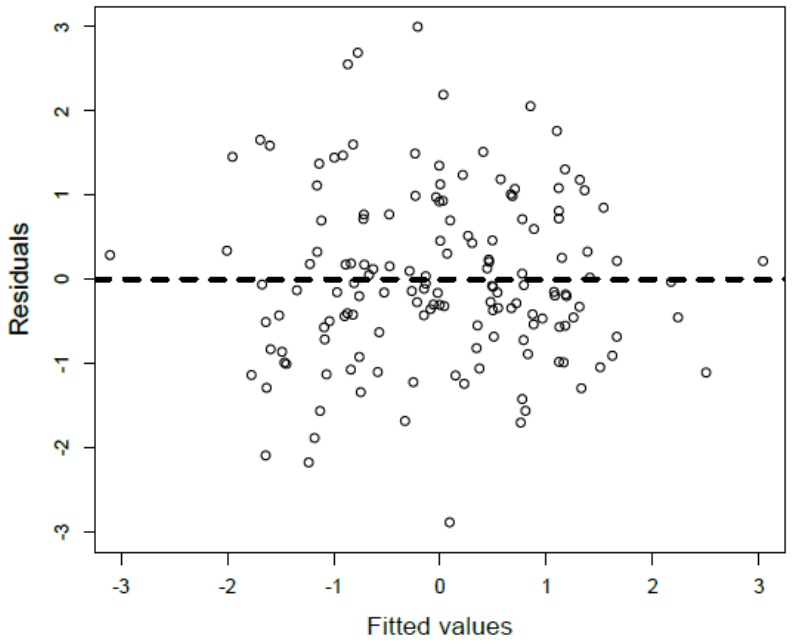
- If two fixed effects are correlated with each other, they are **collinear**.

- intelligence ~ syllables per second + words per second
  - Fixed effects are correlated
- Different linear combinations of fixed effects can produce the same response
  - Cannot interpret coefficients
- Use PCA, feature selection, etc. to choose a smaller set of explanatory fixed effects

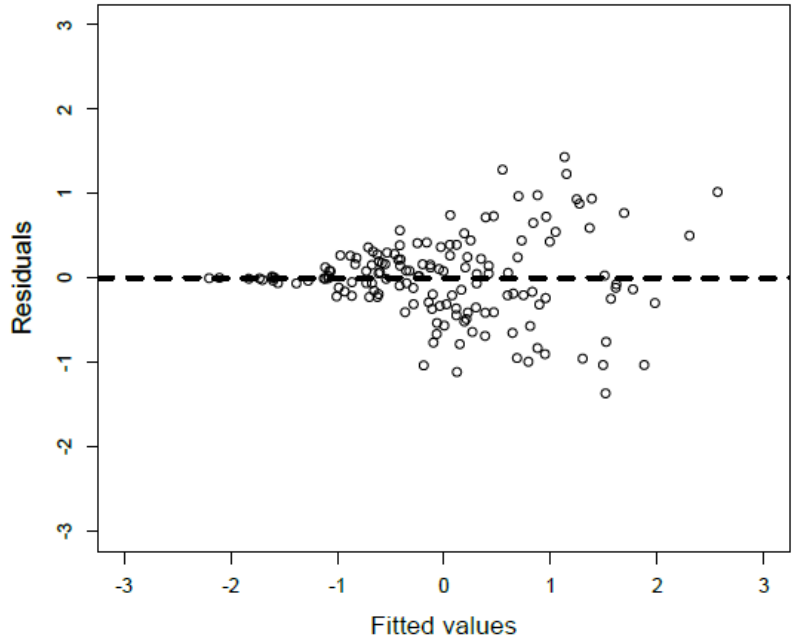
Homoskedasticity

- Unequal variances: Does the variance of the data stage the same across the range of predicted values (homoscedasticity) or does it change (heteroskedasticity).
  - Consider log-transforming the data.

homoscedastic



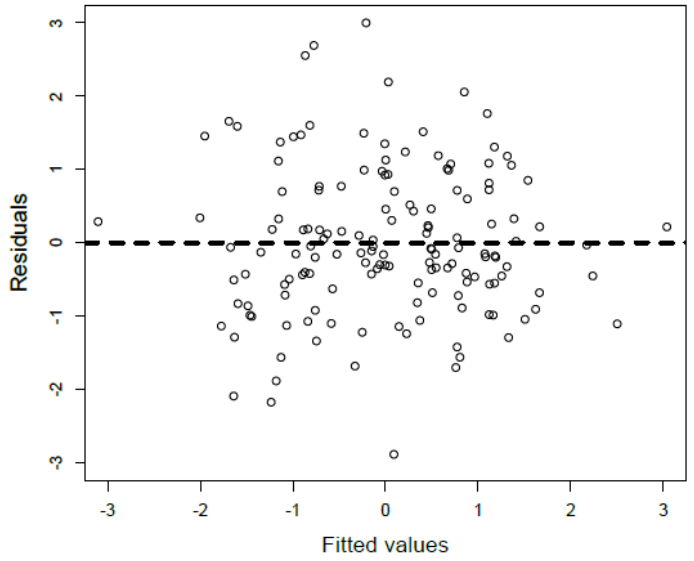
heteroscedastic



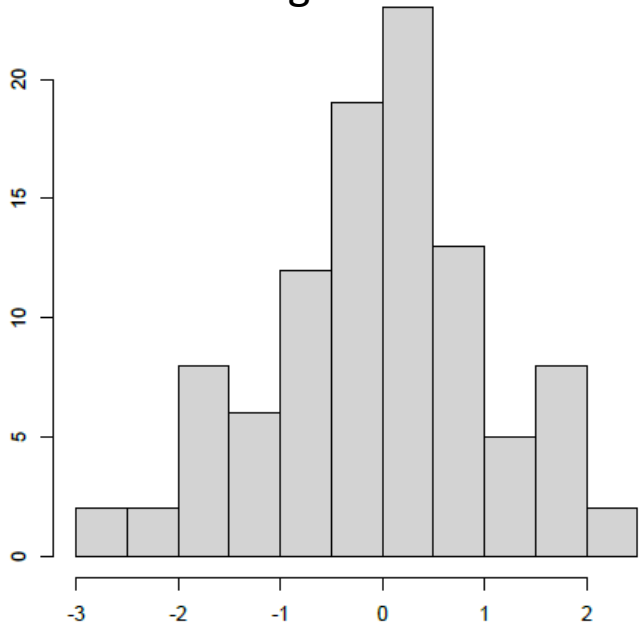
Normality of residuals

- Does the distribution of residuals look normal?
- Least important: linear models are robust to violations of normality

Residuals plot



Histogram of residuals



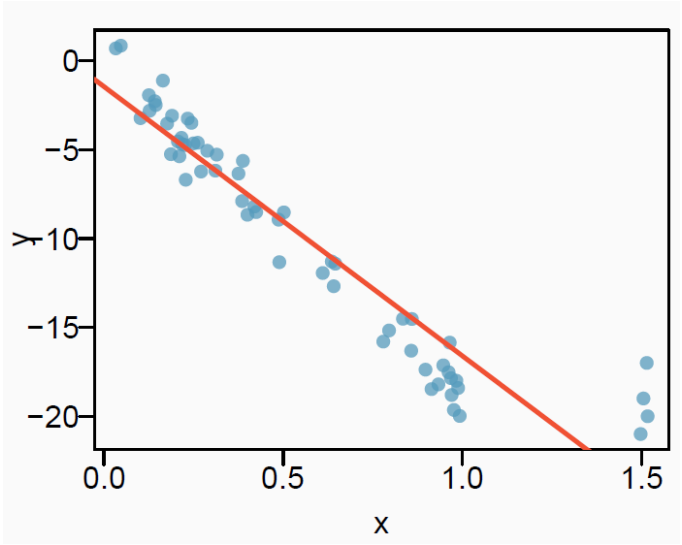
Outliers

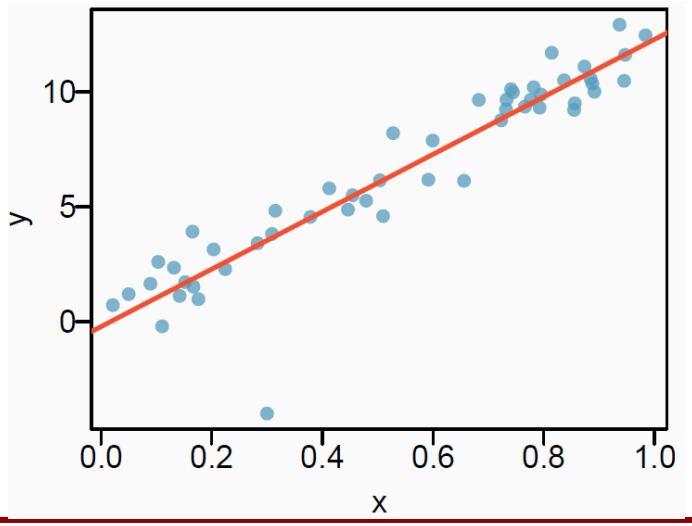
- Influential points: does leaving one data point out substantially change regression coefficients?
- Outliers that lie away from the center of the cloud in the x-direction are called high leverage points.
- A point is influential if including or excluding the point would considerably change the slope of the regression line.

- Do not exclude them from analysis, unless there is an obvious error with the data

Types of outliers

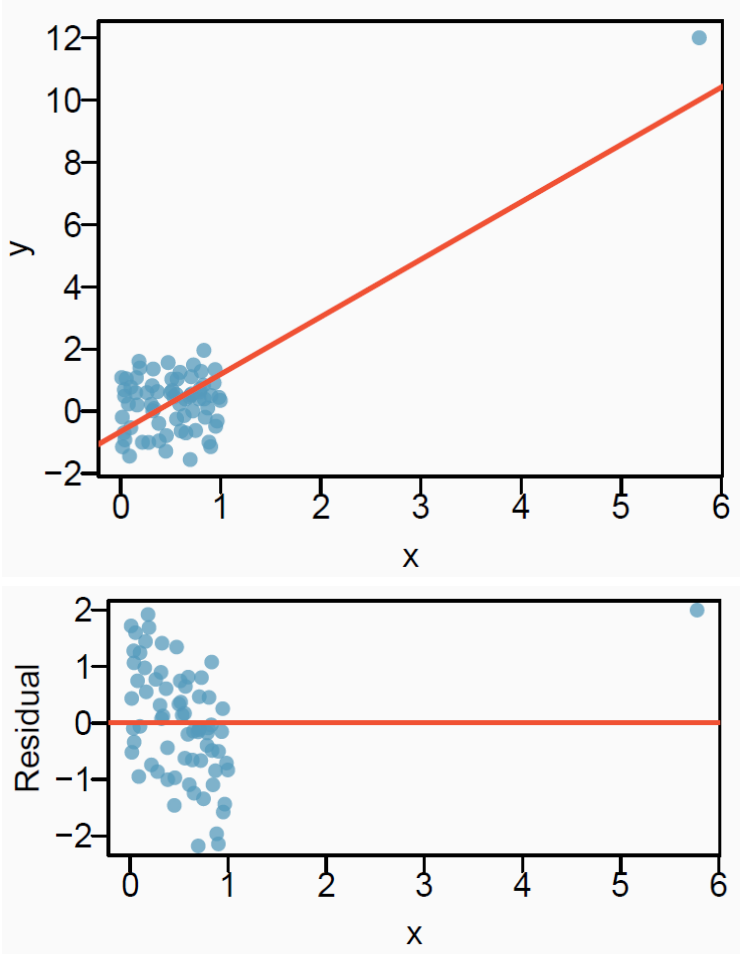
- How do outliers affect the regression line in this plot?
  - To answer, consider what the regression line would be without the outliers.
  - (top) Outliers pull the regression line away from the observations in the larger group of data.
  - (bottom) This outlier does not influence the regression line

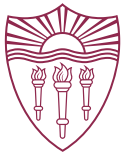




Types of outliers: influential points

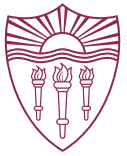
- How do outliers affect the regression line in this plot?
  - Without the outlier, there is no observable relationship between x and y.
- Influential points: does leaving one data point out substantially change regression coefficients?
- Do not exclude them from analysis, unless there is an obvious error with the data





MIXED EFFECTS MODELS

Independence assumption



- Most important assumption for linear models
- Each data point is independent of others
  - i.e., comes from a different subject

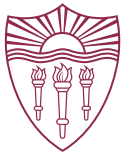
Study 1

Subject	Sex	Voice.Pitch
1	Female	233 Hz
2	Female	204 Hz
3	Female	242 Hz
4	Male	130 Hz
5	Male	112 Hz
6	Male	142 Hz

Study 2

Subject	Age	Voice.Pitch
1	14	252 Hz
2	23	244 Hz
3	35	240 Hz
4	48	233 Hz
5	52	212 Hz
6	67	204 Hz

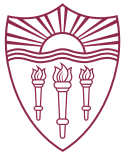
Independence assumption is often violated!



- More complex research questions require collecting multiple responses from the same subject
- But, multiple responses from the same subject cannot be regarded as independent from each other
  - Every person has a slightly different voice pitch, and this idiosyncratic factor will affect all responses from the same subject, making them inter-dependent rather than independent

subject	gender	scenario	frequency
F1	F	1	213.3
F1	F	1	204.5
F1	F	2	285.1
F1	F	2	259.7
F1	F	3	203.9
F1	F	3	286.9
F3	F	1	229.7
F3	F	1	237.3
F3	F	2	236.8
F3	F	2	251
F3	F	3	267
F3	F	3	266
M4	M	1	110.7
M4	M	1	123.6
M4	M	2	229
M4	M	2	114.9
M4	M	3	112.2

But, independence assumption is often violated



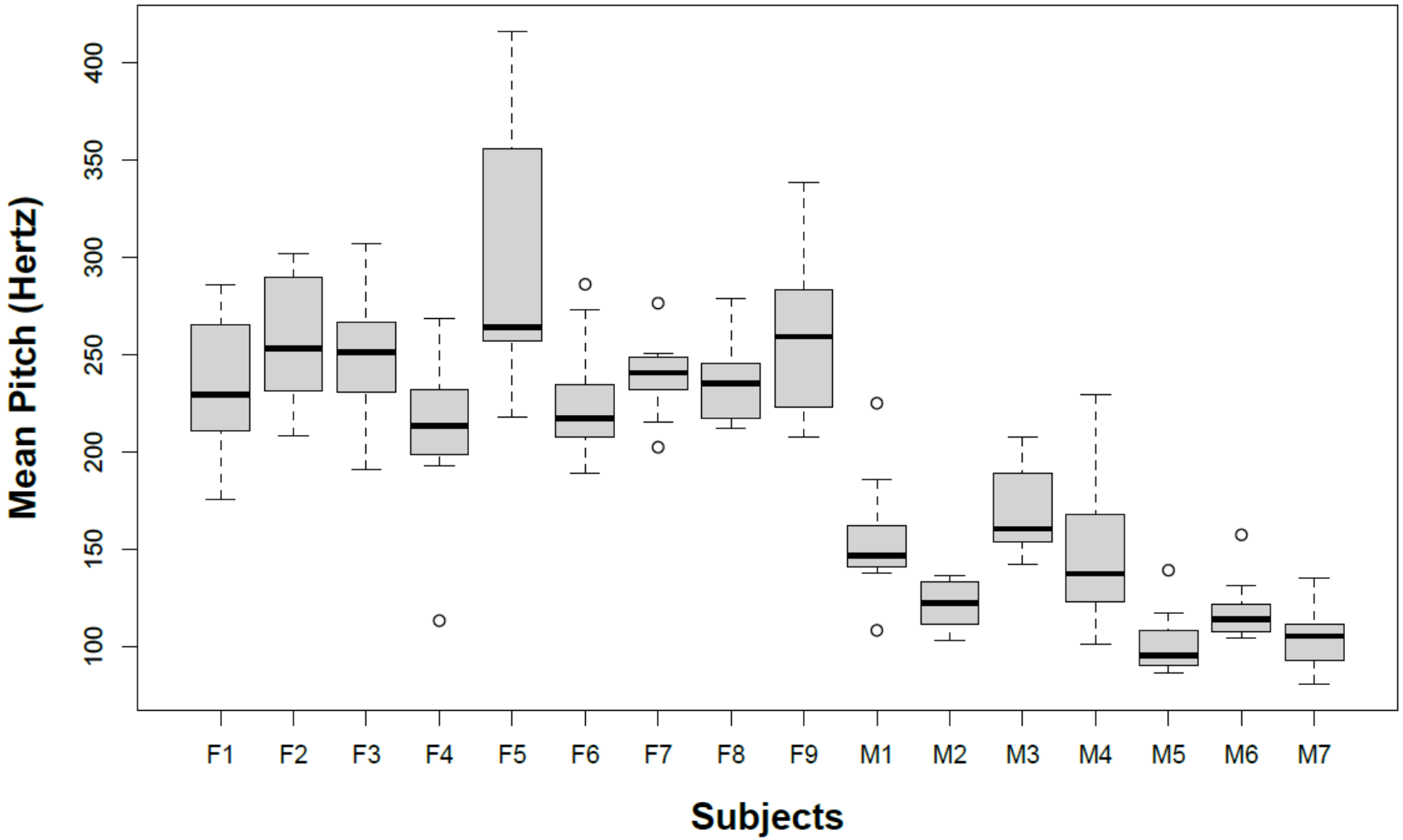
- Research question: Does politeness affect pitch?

  $$\text{pitch} \sim \text{politeness} + \text{sex} + \varepsilon$$

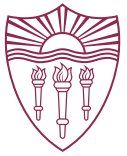
- Each subject gives multiple responses: polite and informal response
- Multiple responses from the same subject cannot be regarded as independent from each other
  - Every person has a slightly different voice pitch, and this idiosyncratic factor will affect all responses from the same subject, making them inter-dependent rather than independent
- Add a random effect
  - This allows us to resolve this non-independence by assuming a different "baseline" pitch for each subject.

Lots of individual variation

mixed_effect_on_politness_data.ipynb



Modeling individual differences with random effects



- Model individual differences by assuming different ***random intercepts*** for each subject.
  - Each subject is assigned a different intercept value, and the mixed model estimates these intercepts.
- Mixed model adds one or more **random effects** to the fixed effects model.
  - These random effects give structure to the error term "ε".
  - In the model, each "subject" becomes a random effect, and this characterizes idiosyncratic variation that is due to individual differences.

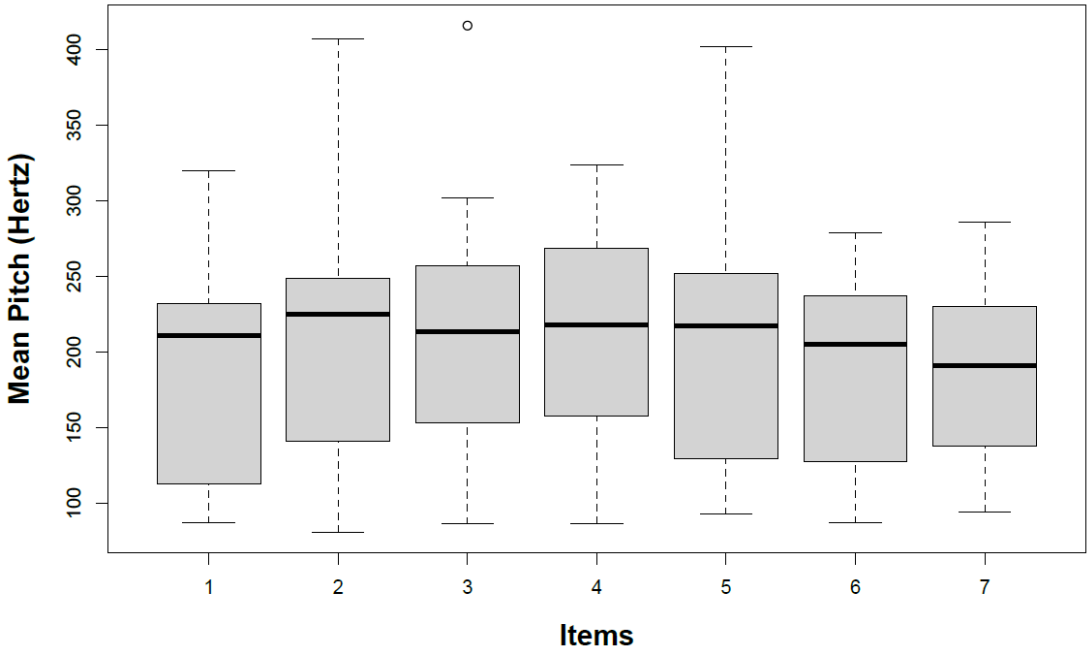
Mixed effects model

$$\text{pitch} \sim \text{politeness} + \text{sex} + (1|\text{subject}) + \varepsilon$$

- "(1|subject)" means a different intercept for each subject"
  - "1" stands for the intercept.
  - Formula tells the model to expect multiple responses per subject, and these responses will depend on each subject's baseline level.
  - This resolves the non-independence that stems from having multiple responses by the same subject.
- Error term "ε" captures remaining "random" differences between different utterances from the same subject.

Modeling multiple dependencies

- Systematic per-item variation
  - Some utterances (items) may have a higher pitch not explained by politeness and subject, but due to another factor that affects the voice pitch of all subjects (e.g., embarrassment)
  - Not accounting for this, violates the independence assumption



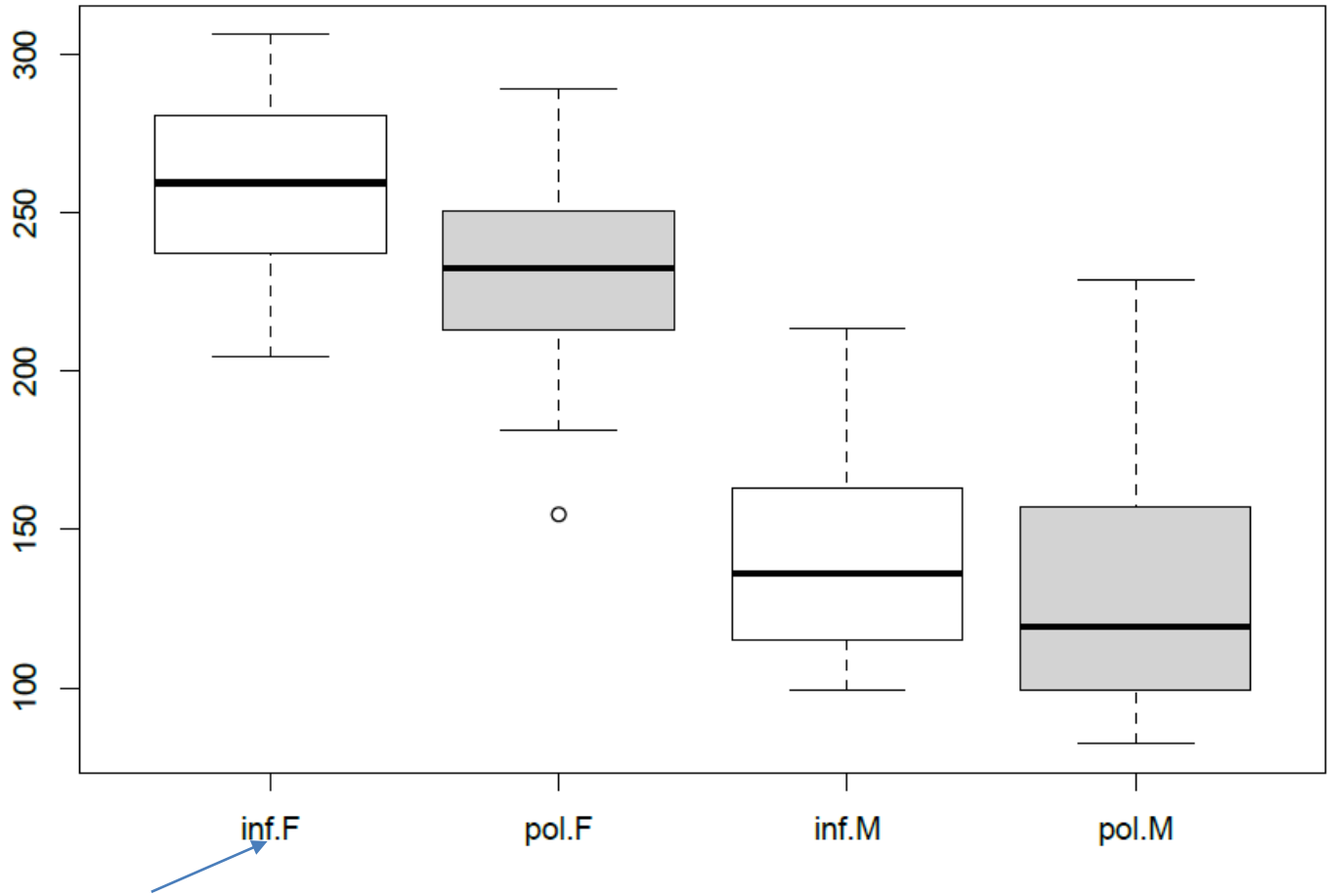
Multiple mixed effects model

$$\text{pitch} \sim \text{politeness} + \text{sex} + (1|\text{subject}) + (1|\text{item}) + \varepsilon$$

- The model knows there are multiple responses per subject and per item
  - 1|subject : Different intercepts for different subjects
  - 1|item: Different intercepts for different items.
- We now “resolved” these non-independencies and accounted for per-subject and per-item variation in overall pitch levels.

Illustration on “Politeness” data

pitch ~ attitude + sex + (1|subject) + (1|item) + ε



Informal speech by females

Polite speech by males

Random effects

pitch ~ attitude + gender + (1|subject) + (1|scenario) + ε

```
Random effects:
 Groups   Name        Variance Std.Dev.
 scenario (Intercept) 205.2    14.33
 subject  (Intercept) 417.0    20.42
 Residual             637.4    25.25
```

- Gender explains much of the between-subject variability in pitch. Without explicitly modeling gender, the subject variance is much higher.
- Residual - ε term – is the variability that is not due to “item” or “subject”

Fixed effects

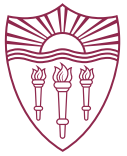
pitch ~ attitude + gender + (1|subject) + (1|scenario) + ε

```
Fixed effects:
             Estimate Std. Error t value
(Intercept)  256.847     13.827  18.576
attitudepol  -19.722      5.547  -3.555
genderM     -108.517     17.572  -6.176
```

- The coefficient “attitudepol” is the slope for the categorical effect of politeness.
  - Minus 19.695 means going from “informal” to “polite” utterances decreases the pitch by -19.695 Hz.

  →Polite speech has lower pitch than informal speech
- Coefficient of “genderM” is negative
  - Males have lower pitch than females

Statistical significance of mixed effects models



- Variety of opinions about the best approach
- Likelihood ratio test
  - Probability of observing the data you collected given the model you learned.
- The logic of the likelihood ratio test is to compare the likelihood of two models with each other.
  - *Null model*: The model *without* the factor that you're interested in

    pitch ~ gender + (1|subject) + (1|scenario) + ε
  - *Full model*: *with* the factor that you're interested in.

    pitch ~ attitude + gender + (1|subject) + (1|scenario) + ε

Likelihood ratio test

```
Data: politeness
Models:
politeness.null: frequency ~ gender + (1 | subject) + (1 | scenario)
politeness.model: frequency ~ attitude + gender + (1 | subject) + (1 | scenario)
                 Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
politeness.null   5 816.72 828.81 -403.36   806.72
politeness.model  6 807.10 821.61 -397.55   795.10 11.618      1  0.0006532 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- perform the likelihood ratio test using a standard package (eg, anova)

- Report the effect in a paper as follows:
  - “… politeness affected pitch ($\chi 2(1)=11.62$, $p=0.00065$), lowering it by about 19.7 Hz ± 5.6 (standard errors) …”

Random slopes vs random intercepts

pitch ~ attitude + gender + (1|subject) + (1|scenario) + ε

- Different intercept for each subject and each item
  - Different baselines for each subject
- But the same coefficients: Fixed effects of gender and attitude are the same for all subjects and items
- Need a model with random slopes to allow subjects to have individualized responses to fixed effects

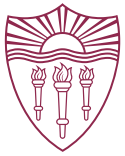
```
$scenario
  (Intercept) attitudepol   genderM
1    243.4859   -19.72207 -108.5173
2    263.3592   -19.72207 -108.5173
3    268.1322   -19.72207 -108.5173
4    277.2546   -19.72207 -108.5173
5    254.9319   -19.72207 -108.5173
6    244.8015   -19.72207 -108.5173
7    245.9618   -19.72207 -108.5173

$subject
   (Intercept) attitudepol   genderM
F1    243.3684   -19.72207 -108.5173
F2    266.9443   -19.72207 -108.5173
F3    260.2276   -19.72207 -108.5173
M3    284.3536   -19.72207 -108.5173
M4    262.0575   -19.72207 -108.5173
M7    224.1292   -19.72207 -108.5173

attr(,"class")
[1] "coef.mer"
```

Mixed effects with random slopes

pitch ~ attitude + gender + (1+attitude|subject) + (1+attitude|scenario) + ε

- Coefficient for the effect of politeness (“attitudepol”) is different for each subject and item
- despite individual variation, there is also consistency in how politeness affects voice: pitch tends to go down when speaking politely

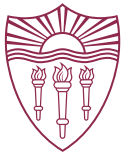
```
$scenario
  (Intercept) attitudepol   genderM
1    245.2603   -20.43832 -110.8021
2    263.3012   -15.94386 -110.8021
3    269.1432   -20.63361 -110.8021
4    276.8309   -16.30132 -110.8021
5    256.0579   -19.40575 -110.8021
6    246.8605   -21.94816 -110.8021
7    248.4702   -23.55752 -110.8021

$subject
   (Intercept) attitudepol   genderM
F1    243.8053   -20.68245 -110.8021
F2    266.7321   -19.17028 -110.8021
F3    260.1484   -19.60452 -110.8021
M3    285.6958   -17.91950 -110.8021
M4    264.1982   -19.33741 -110.8021
M7    227.3551   -21.76744 -110.8021

attr(,"class")
[1] "coef.mer"
```



Take 5

BREAK

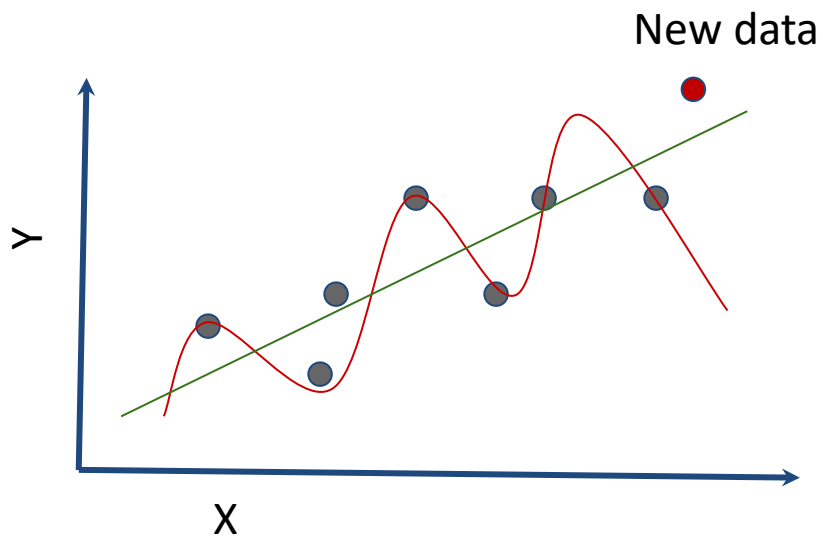
Pitfalls and best practices



- Overfitting
- Information leakage
- Noise
- Feature engineering
  - Feature transformation
    - E.g., when features have a large range, take log transform
  - Feature selection

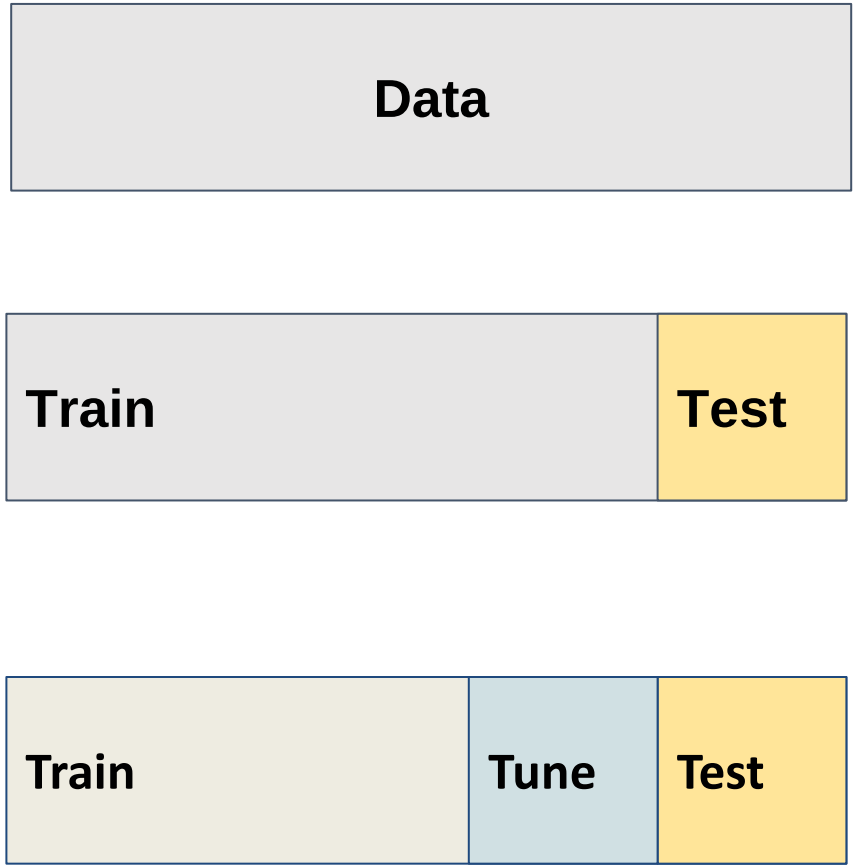
Overfitting

- Avoid overfitting
  - It may be tempting to create an “optimal” model
  - But, a complex model that performs well on training data, may not **generalize**
  - It may have learned to specialize to existing data
    - reduce parameters (simplify the model)
  - Signs of overfitting: very high R2 on training data



Information leakage

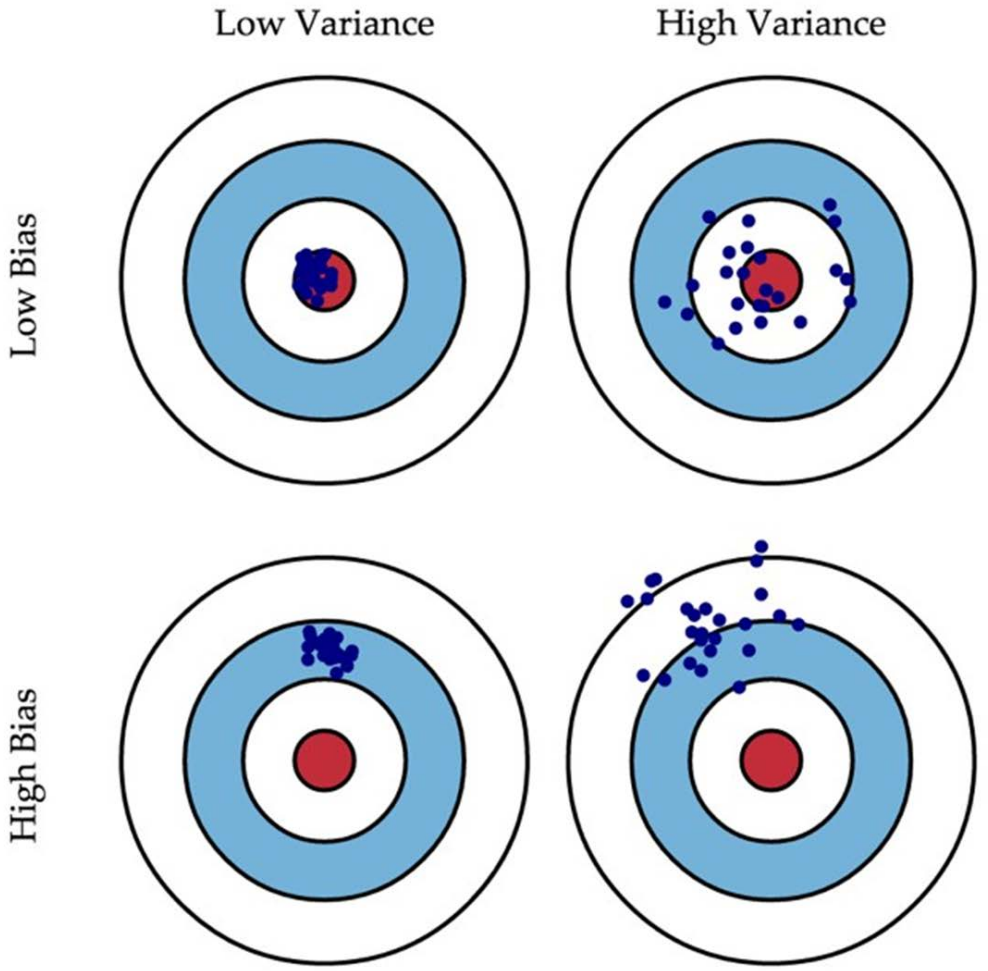
- Avoid information leakage
  - Never test on **the same** data used for training
  - 5-fold cross validation
    - Train on a random 80% of data, test on 20%
    - Average results over 5 random splits
  - Leave one out (for small data)
    - Train on N-1 data points, test on 1 point
  - For hyperparameter tuning
    - Train on 3 folds, validate on 1 fold, test on 1 fold



Noise, bias, and variance

If there is lots of noise in the data, repeatedly sampling the data may yield different models.

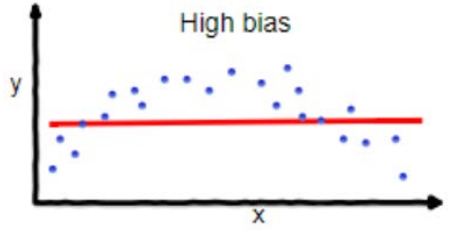
The best case is if we always hit the bull's eye. There are two sources of error: Bias and Variance, which we both want to minimize.

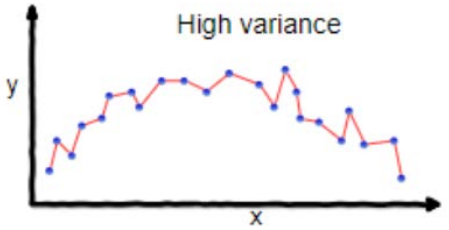


Bias-variance tradeoff

- Bias
  - difference between average prediction of the model and true value
  - Model underfits the data, oversimplifies the model
  - Could also be due to **systematic errors**

- Variance
  - variability of model prediction for a given feature value
  - Repeated sampling of population results in different data
  - Model overfits, does not generalize to test data





Bias-variance tradeoff

- MSE = Mean squared error at a point x :
- MSE = $Err(x) = \mathrm{E}[\left(Y - \hat{f}(x)\right)^2]$

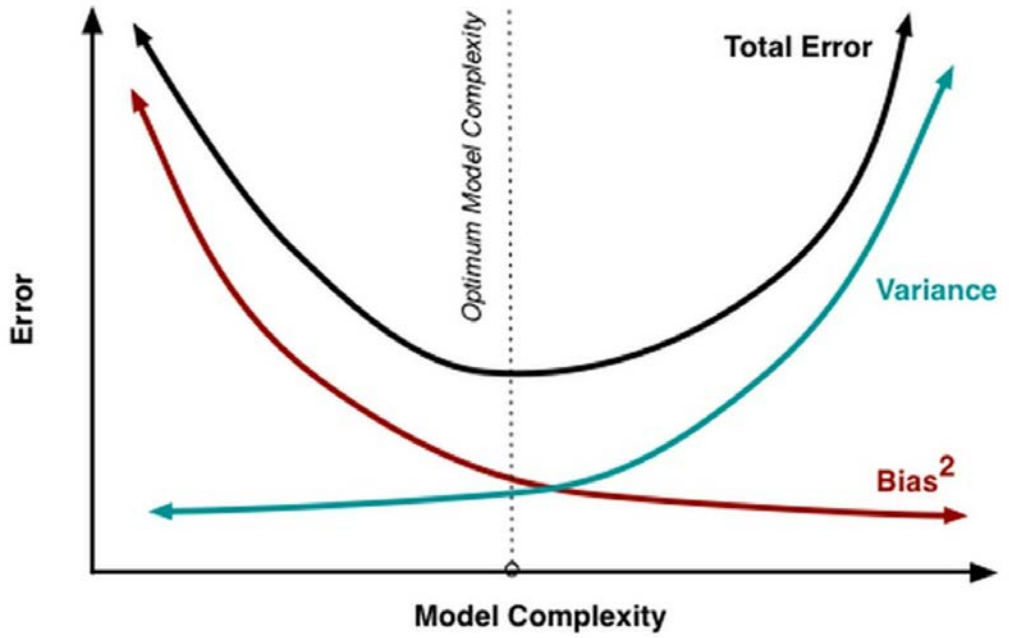
$$Err(x) = \left(E[\hat{f}\,(x)] - f(x)\right)^2 + E\left[\left(\hat{f}\,(x) - E[\hat{f}\,(x)]\right)^2\right] + \sigma_e^2$$

$$Err(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

Bias-variance tradeoff

Unfortunately, it is not always possible to minimize both variance and bias at the same time. In general, bias is reduced if we add more and more parameters to a model and make it more complex.

However, the more complex the model becomes the more variance we introduce in the model. In its core the problem alludes to over- and under-fitting.



Feature transformation

- Example (credit: M. Pazzani)
- Create a model of male faculty salary at UCI. Use the model to predict female faculty salary.

Salary	YoB	YoH	YoD	Gender
89,990	1960	2004	1991	M
72,660	1965	1998	1997	M
87,125	1963	1993	1991	F
67,500	1979	2003	2003	M
78,900	1973	1999	1998	F
102,500	1952	1980	1989	M

- Model y=Salary as a function of X:
  - Year of birth (YoB)
  - Year of hire (YoH)
  - Year of degree (YoD)

  - Salary = 3,420,751 – 293*YoB – 808*YoD – 593*YoH

Feature transformation

- **Salary = 3,420,751 – 293*YoB – 808*YoD – 593*YoH**

- Instead, model Salary as a function of
  - Years since birth (Age)
  - Years since degree (YsD)
  - Years since hire (YsH)

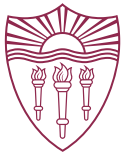
  Salary = 48,623 + 293*Age + 808*YsD + 593*YsH
- Use Years above Thirty instead of Age

  Salary = 64,418 + 293*YaT + 808*YsD + 593*YsH

Feature selection

- Feature engineering is necessary with regression, and many other models as well
  - Reduce the number of colinear features by eliminating un-informative features
  - Many methods: VIF, PCA, factor analysis, …
- **Variance Inflation Factor** - quantifies the severity of multicollinearity and measures how much the variance of an estimated regression coefficient is increased because of collinearity.
  - Rule of thumb – calculate VIF for each feature and eliminate features with VIF>5.
- **Minimum redundancy maximum relevance** (mRMR) – identified features that are highly correlated with the outcome (relevance), but uncorrelated with each other (redundancy)

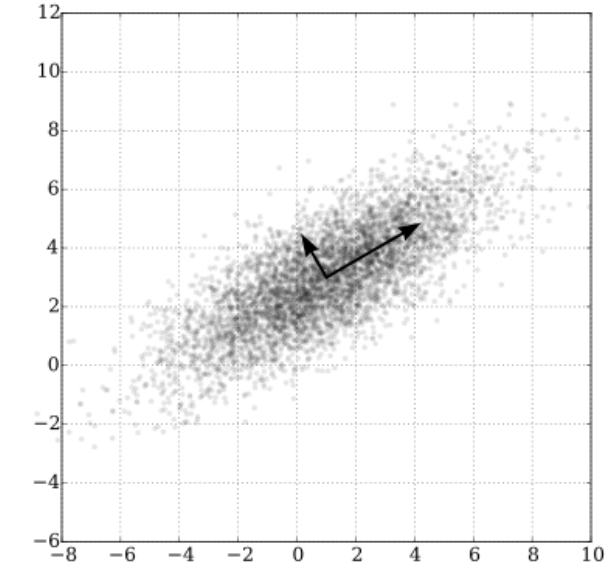
Forward/Backward Feature Selection

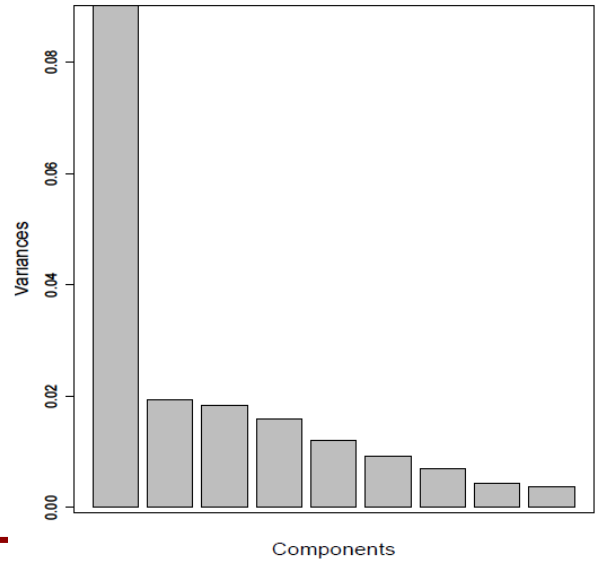


- **Forward Selection**: Forward selection is an iterative method, starting with no features in the model
  - In each iteration, add the best performing feature to the model
    - select the feature with the lowest p-value
  - Continue until adding a new feature does not improve the performance
- **Backward Elimination**: starting with all the features
  - In each iteration, remove the least significant feature
    - feature with the largest insignificant p-value
  - Continue until all features with insignificant p-values are removed

Principal Component Analysis - PCA

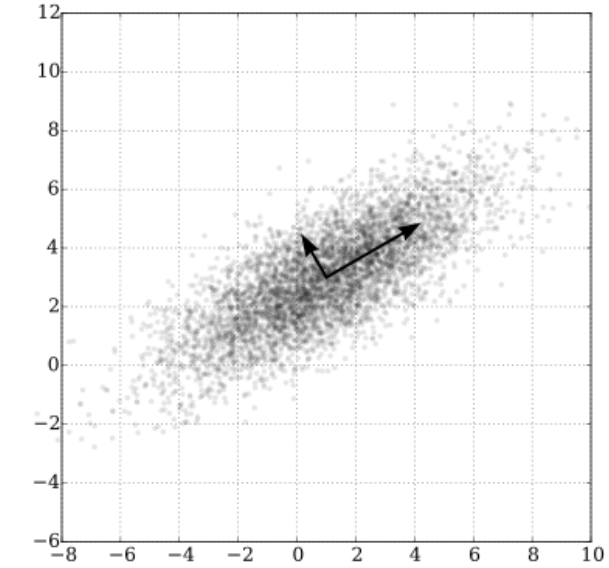
- Unsupervised method that identifies the internal structure of high-dimensional data that best explains its variance
- Embeds the data in a new lower-dimensional space, such that the greatest variance lies on the first coordinate (dimension), the second greatest variance on the second dimension, etc.

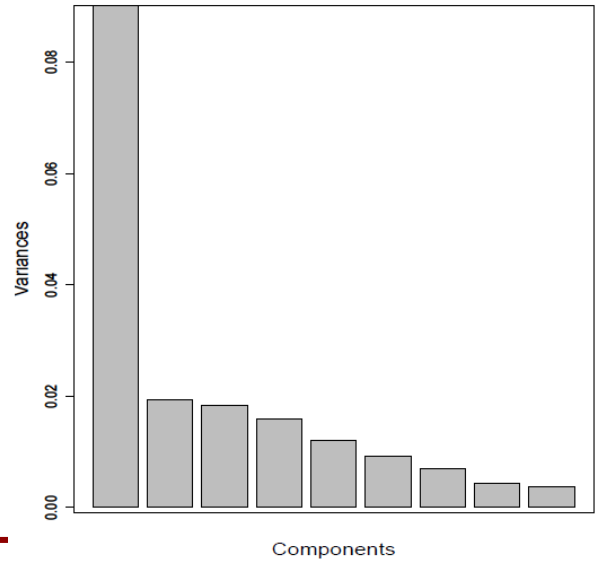




PCA

- Each component is a linear combination of the existing features
  - Not interpretable
  - Use instead of features in linear models
  - Computation:
    - deep mathematics with eigen decomposition
    - Efficient algorithms
- How many components to use?
  - Ignore less significant components
  - Information loss



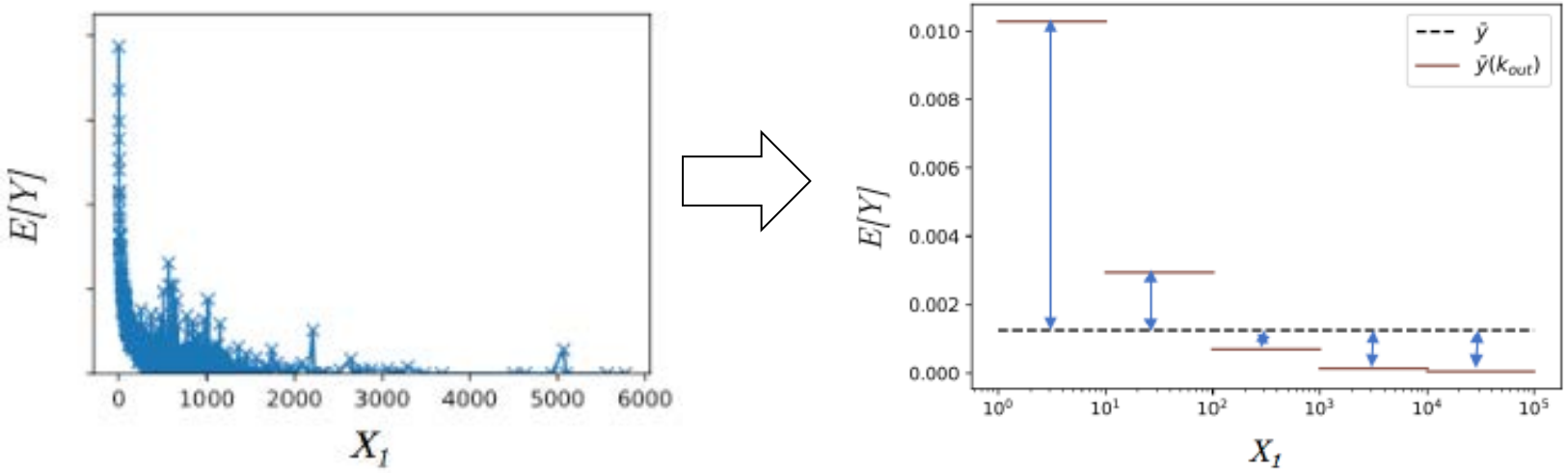


Non-linear feature selection methods

- Structured Sum of Squares Decomposition (S3D)
  - Successively picks features that collectively "best" explain an outcome variable
  - Creates a non-linear model of data

$$R^2 = \frac{\sum_{g=1}^{G} N_g \left(\bar{y}_g - \hat{y}\right)^2}{\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

$$\boldsymbol{Y = f(X_1)}$$

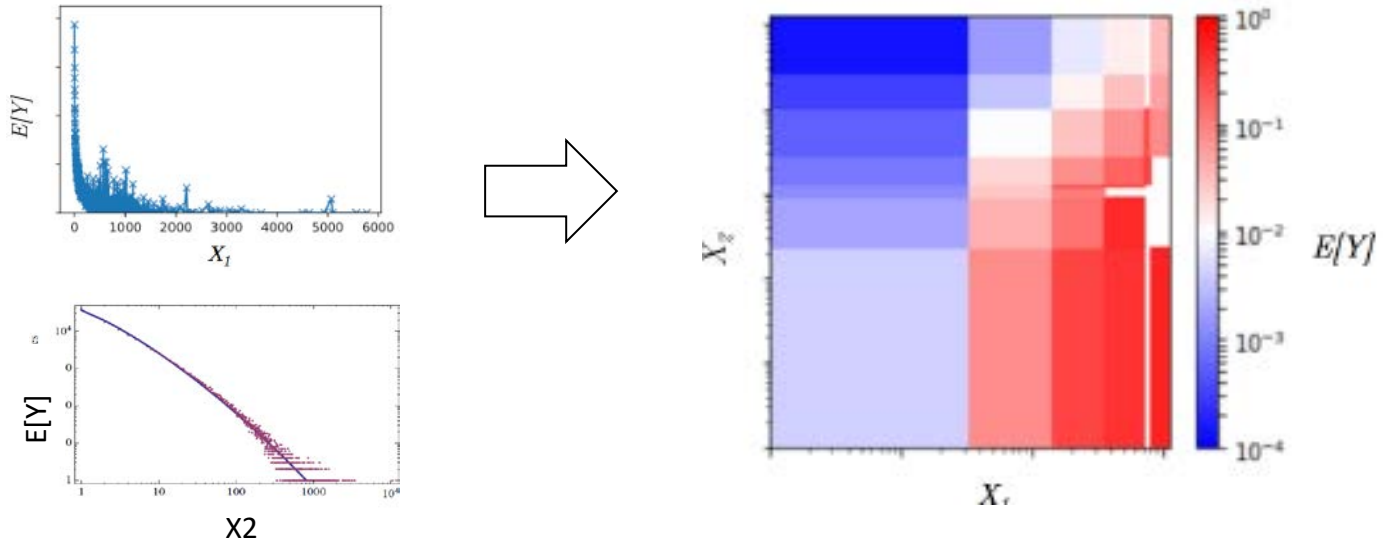


Structured Sum of Squares Decomposition (S3D)

https://github.com/peterfennell/S3D

- Successively picks features that collectively "best" explain an outcome variable
  - Similar to mRMR
- Creates a non-linear model of data useful for visualization and prediction

$$R^2 = \frac{\sum_{g=1}^{G} N_g \left(\bar{y}_g - \hat{y}\right)^2}{\sum_{i=1}^{N} (y_i - \hat{y})^2} \qquad \boldsymbol{Y = f(X_1, X_2)}$$

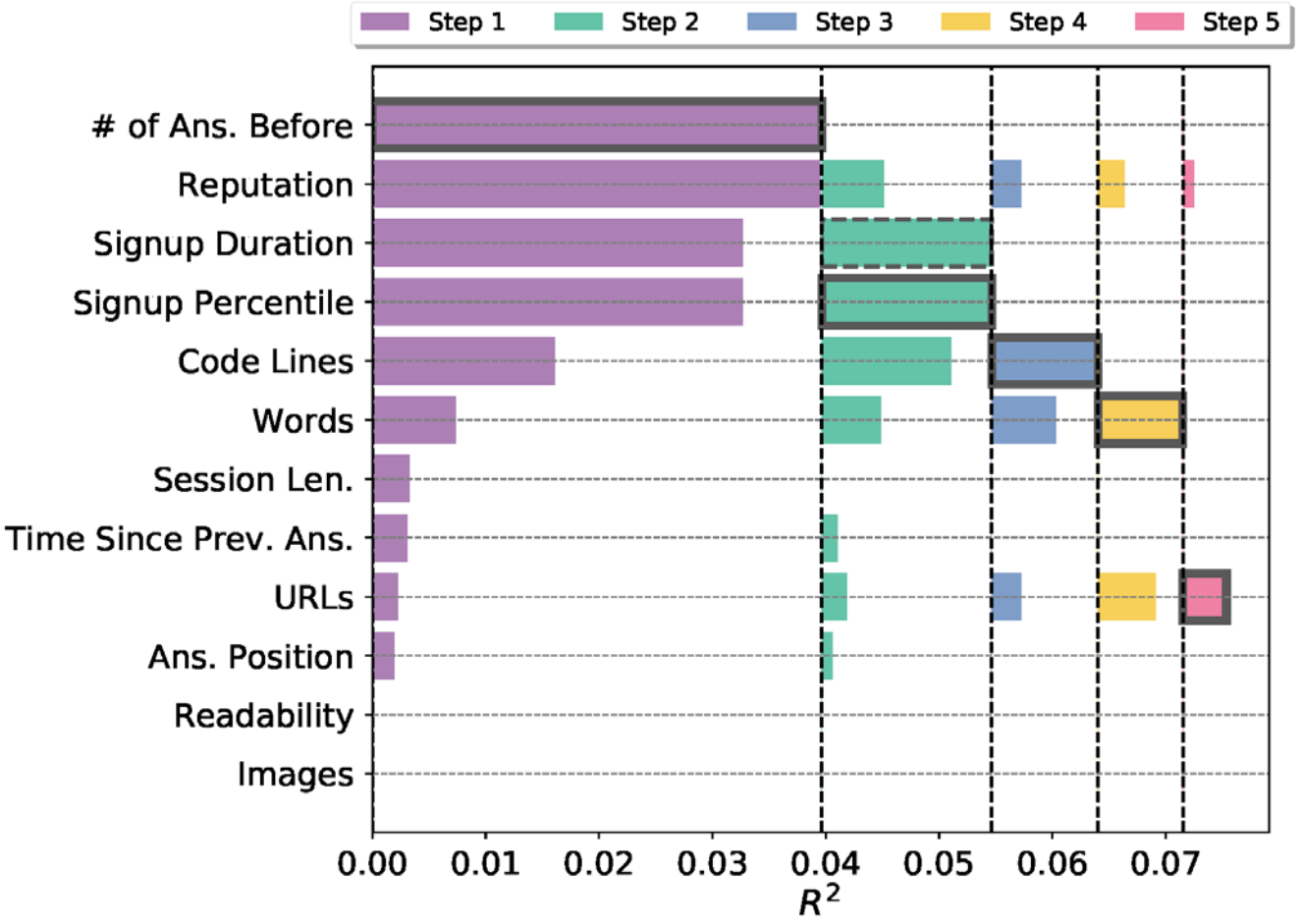


Color gives the average value of outcome in each bin.

← Nonlinear approximation of data

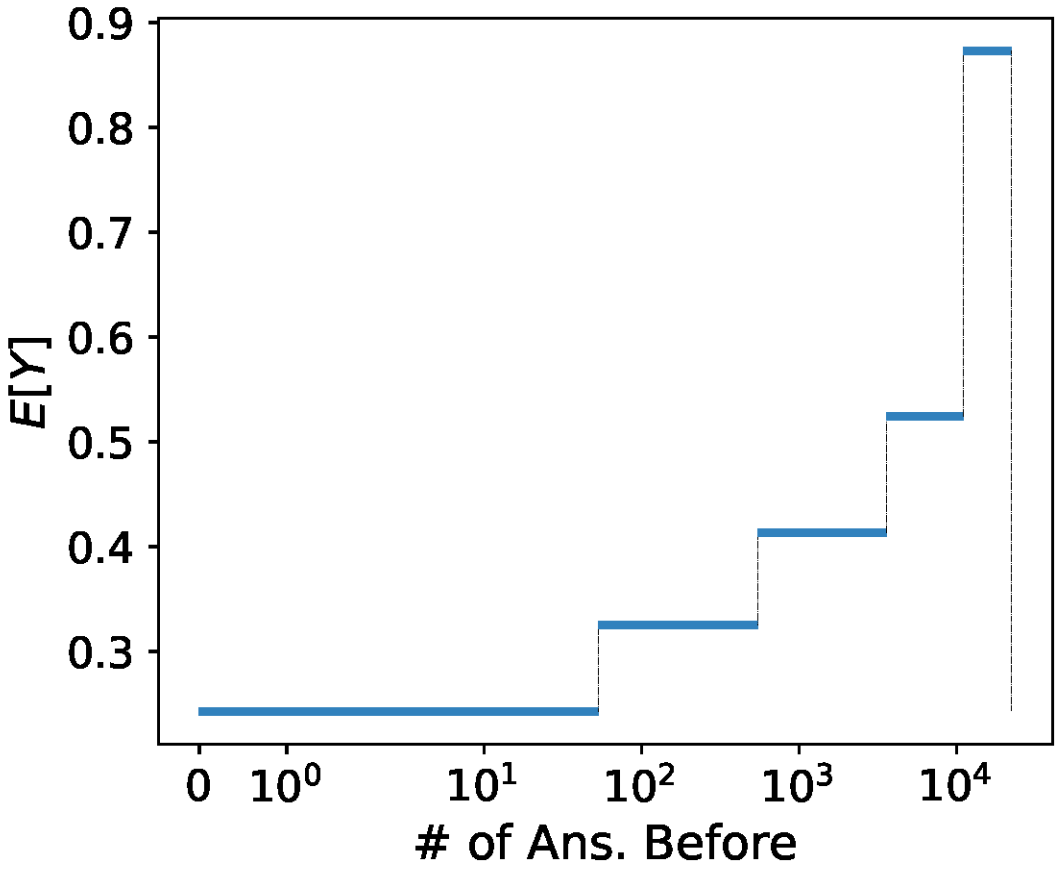
S3D illustration on StackOverflow data

Identifying important features explaining whether user's answer is accepted as best answer to the question (Y=binary outcome)



Visualizing data: 1st Important Feature

***# Answers Before**:* more experienced users (who had answered more questions on StackOverflow) are more likely to have their new answers accepted



2nd Important Feature

Signup percentile: rank of user's tenure on StackOverflow. Older users who had written the same number of answers are less likely to have a new answer accepted as best answer.