DSCI-531: FAIRNESS IN ARTIFICIAL INTELLIGENCE

SUPERVISED LEARNING

Kristina Lerman

Spring 2025

whoami

- Ashwin Rao
  - Teaching Assistant for DSCI531
  - Office Hours: Monday 10AM-11AM PT
  - Location: https://usc.zoom.us/j/9026240639
- 5th year PhD student in Computer Science
  - Study political polarization, misinformation and social networks
- Free time: Trying to learn more about human civilizations & evolution.

CSCI 599: Evolution
Units: 4.0
Spring 2025 — MonWed — 2:00-3:50PM

Location: Online via Zoom

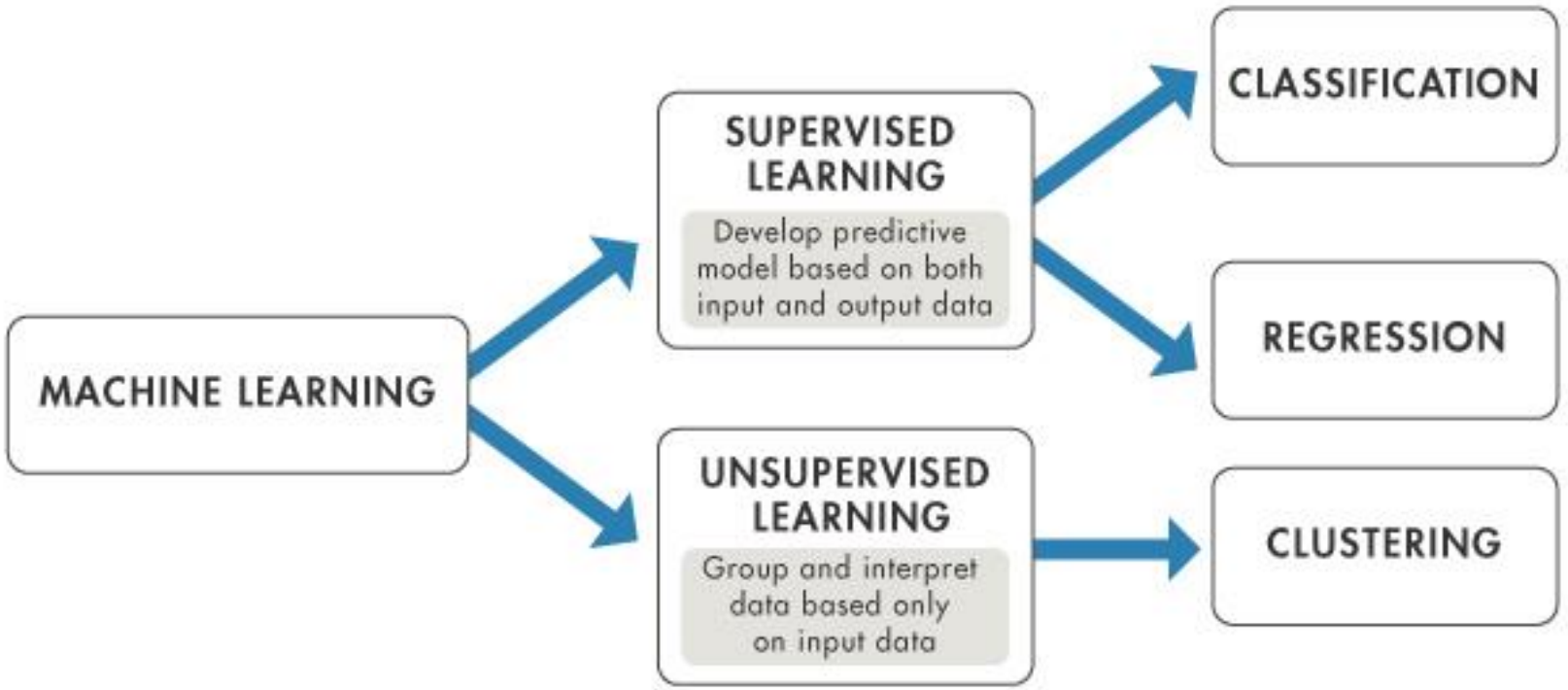
Instructor: Professor Leonard M. Adleman
Office: Remote
Office Hours: TBD
Contact Info: adleman@usc.edu

Lecture Outline

1. Supervised vs Unsupervised Learning
2. Linear Regression
3. Assumptions of Linear Regression
4. Logistic Regression
5. Mixed Effects Regression
6. Best Practices

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?



Source: DeepAI.org

Output:**label**
- Binary
- Categorical

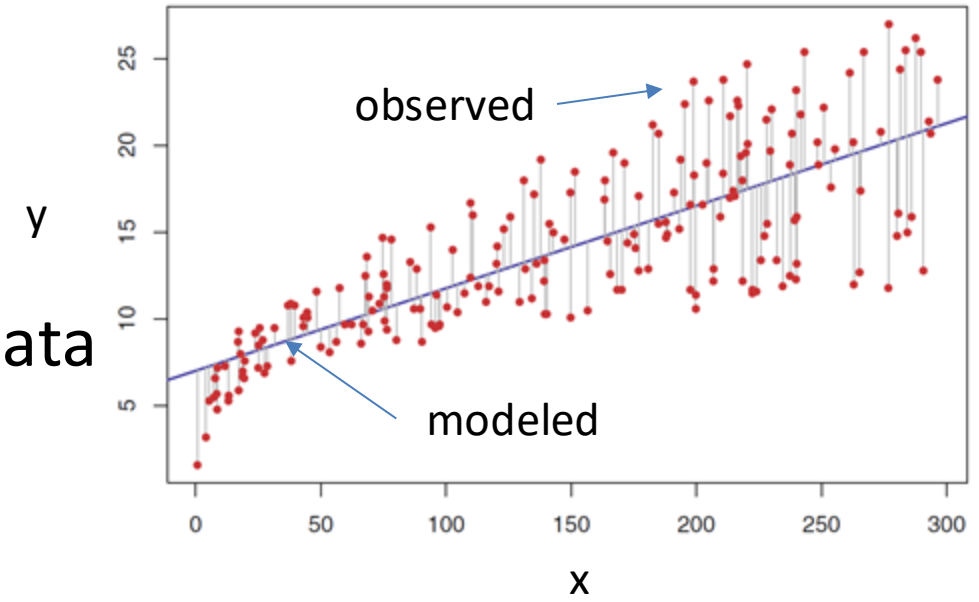
Output:
response/outcome
- Binary
- Real

Linear Regression

Linear regression

- Parametric model: $\mathbf{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x}$
  - **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) for *n* data

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



Linear Regression - Residual sum of squares (RSS)

- 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

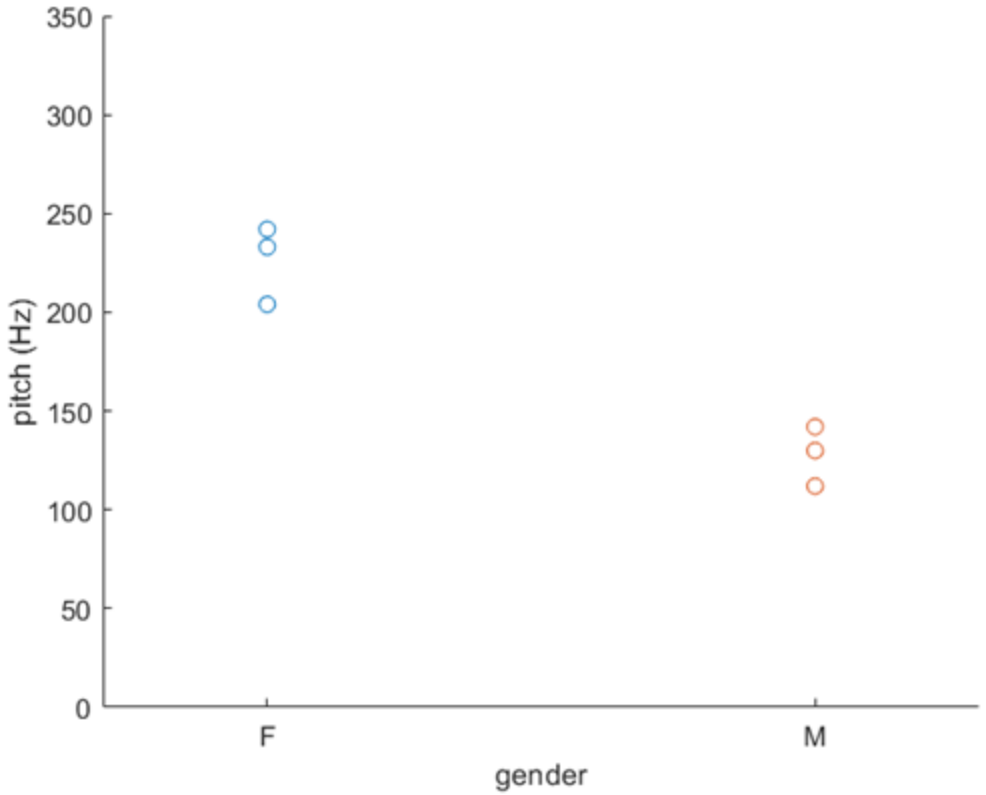
Voice Pitch Example: Creating a model

- *RQ: How much does the voice pitch of males and females differ?*
- Collect data
  - Measure pitch, Bigger Hz, higher pitch

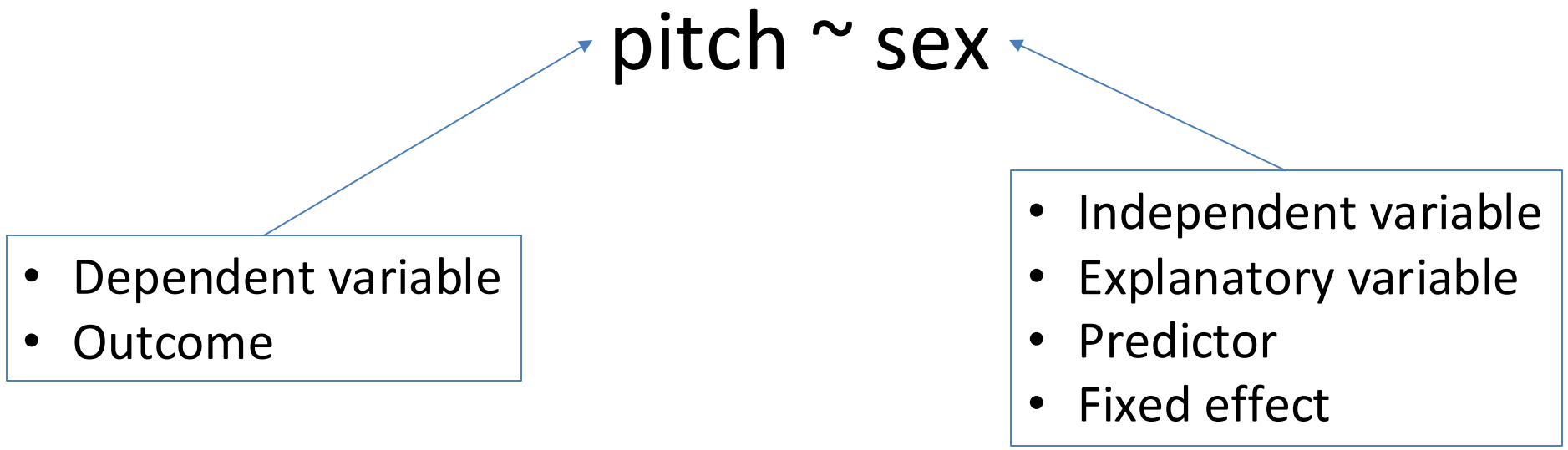
Subject	Sex	Voice.Pitch
1	Female	233 Hz
2	Female	204 Hz
3	Female	242 Hz
4	Male	130 Hz
5	Male	112 Hz
6	Male	142 Hz

Voice Pitch Example: Creating a model

- 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



A simple model



A simple model

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

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

- 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 *(R2)* 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.

Fitting model to the data

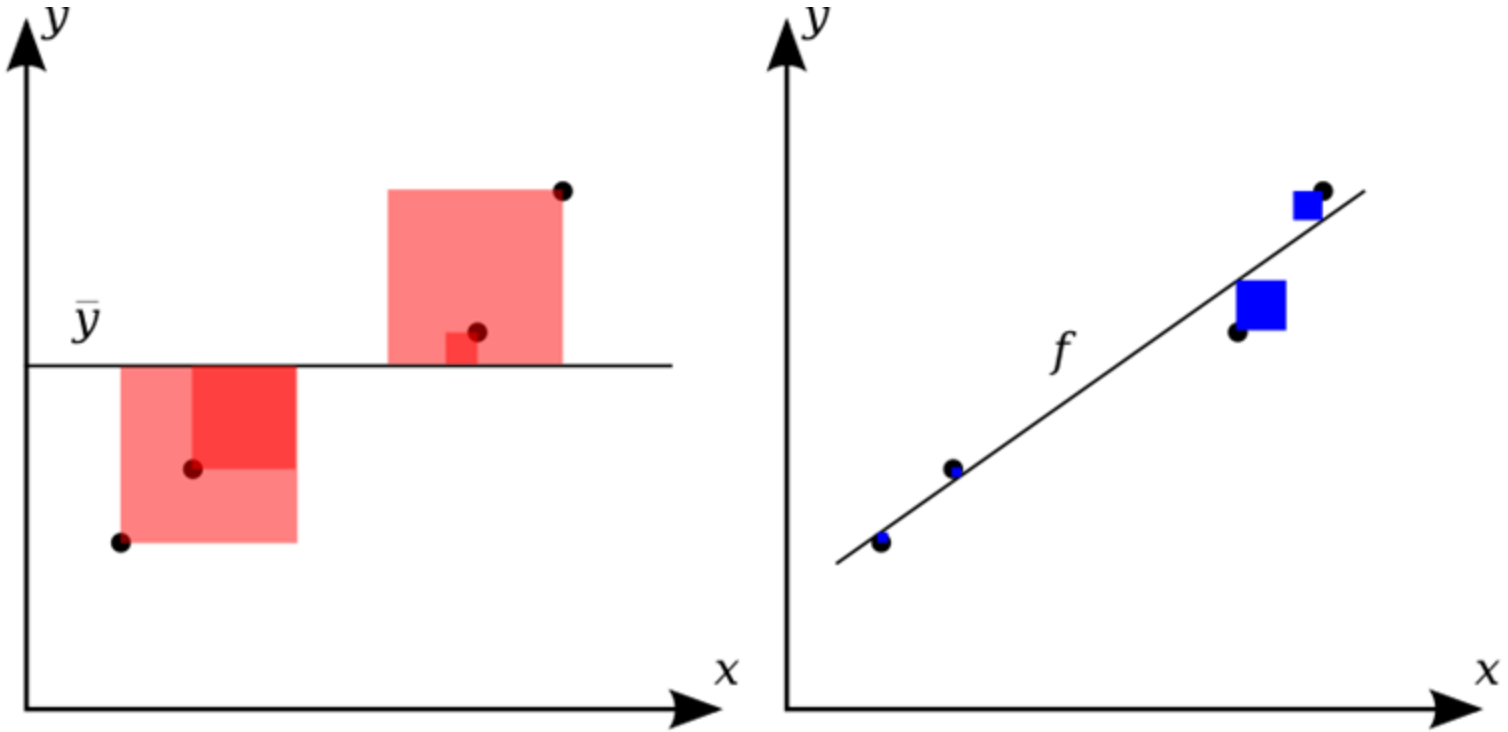
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   226.33      10.18  22.224 2.43e-05 ***
sexmale       -98.33      14.40  -6.827  0.00241 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

$$R^2_{\text{adj}} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

Adjusted R-squared *(R2)* controls for the complexity of the model: reflects how much variance in data is accounted for by the model considering how many variables (k) it uses.

Coefficient of determination – R2

- How much of the variation in y is explained by the variation in x? E.g., how much of the variation in pitch is explained by differences between males and females?



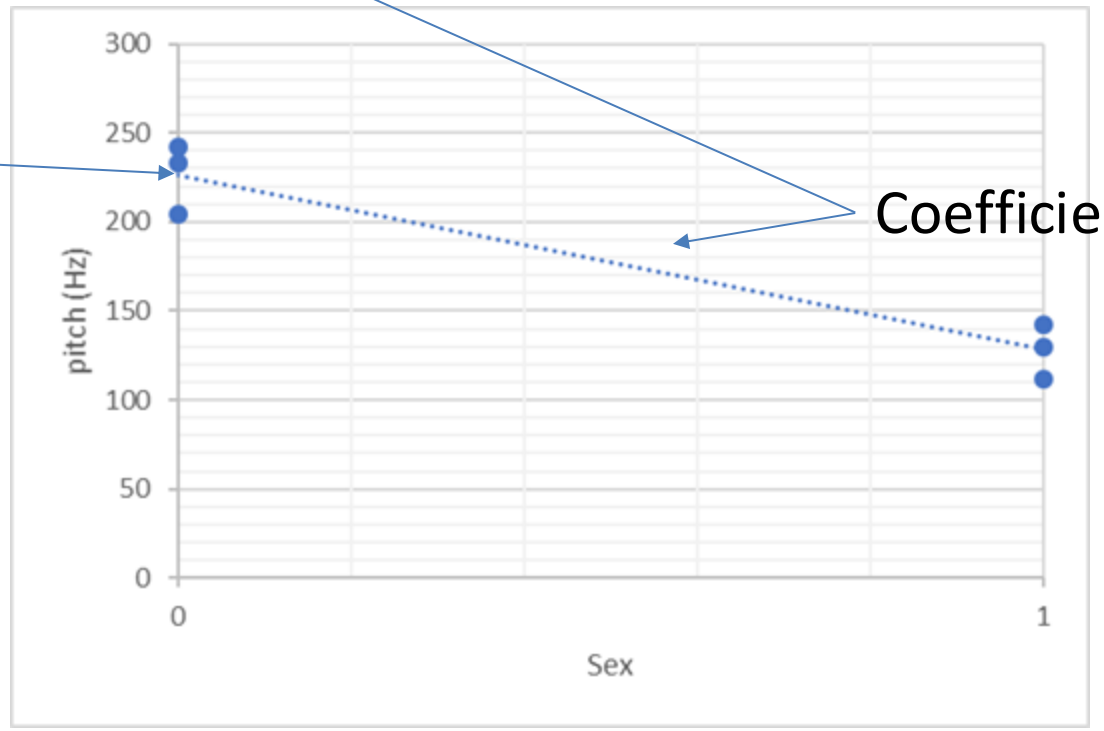
$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

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

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)

Significance – p-value

```
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 **
```

- 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

Multiple regression model (Multivariate regression)

Let's say there's another variable age

Subject	Sex	Voice.Pitch	Age
1	Female	233 Hz	14
2	Female	204 Hz	48
3	Female	242 Hz	23
4	Male	130 Hz	32
5	Male	112 Hz	54
6	Male	142 Hz	21

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

Multiple regression model

- How to interpret?
- Intercept – Predicted voice pitch for a female subject
- Age – For every year increase in age, pitch decreases by approximately 0.943 Hz, holding sex constant.
- Sex – Holding age constant, men have 91hz lower pitch than women.

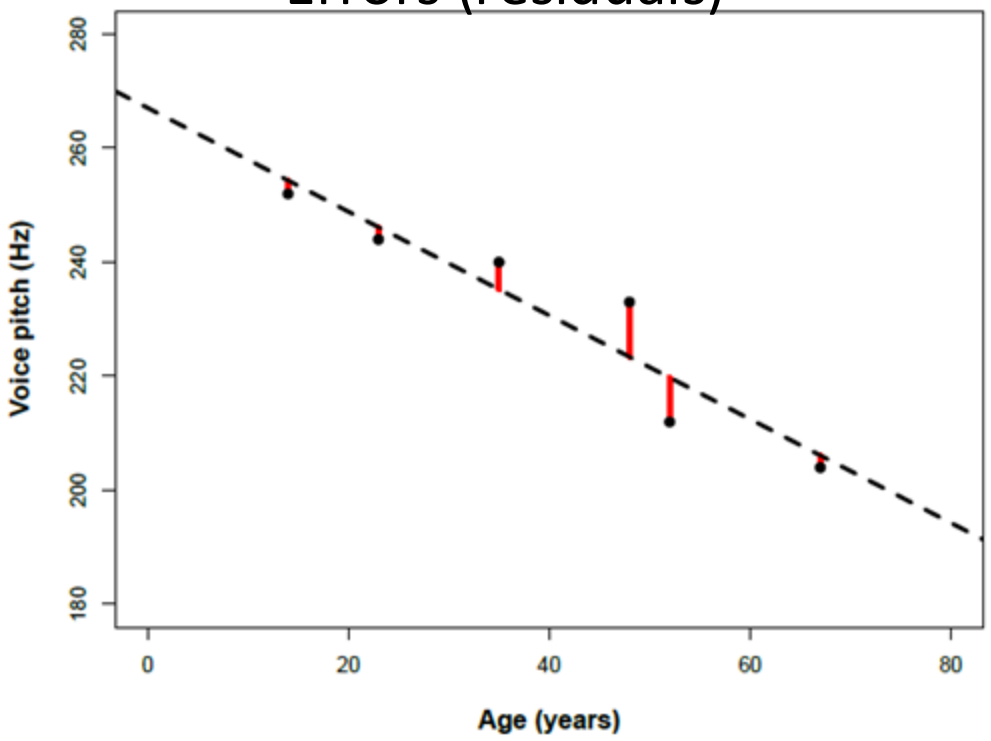
```
lm(formula = pitch ~ age + sex, data = my.df)

Residuals:
      1       2       3       4       5       6
-7.1681 -4.0966 11.2647 -1.4587  1.2934  0.1652

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 253.3740     8.2735  30.625 7.65e-05 ***
age          -0.9433     0.2375  -3.972 0.028533 *
sexmale     -91.7304     6.8528 -13.386 0.000901 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.142 on 3 degrees of freedom
Multiple R-squared:  0.9874,    Adjusted R-squared:  0.979
F-statistic: 117.3 on 2 and 3 DF,  p-value: 0.001419
```

ASSUMPTIONS OF LINEAR REGRESSION

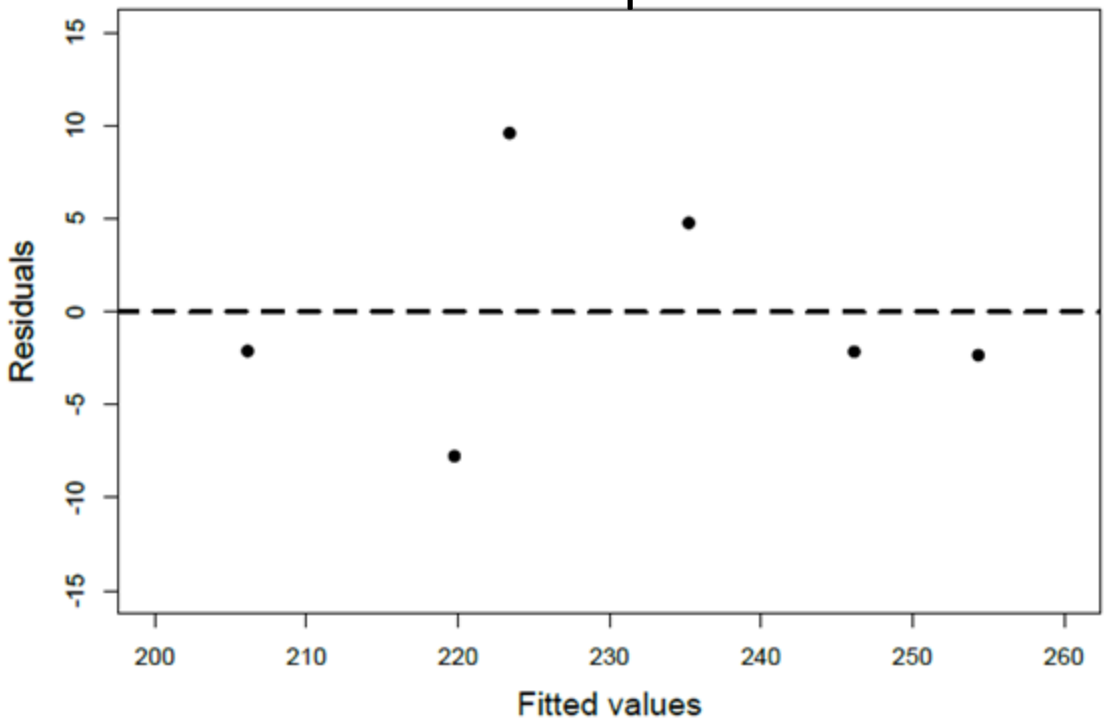
Linear Relationship

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

Errors (residuals)

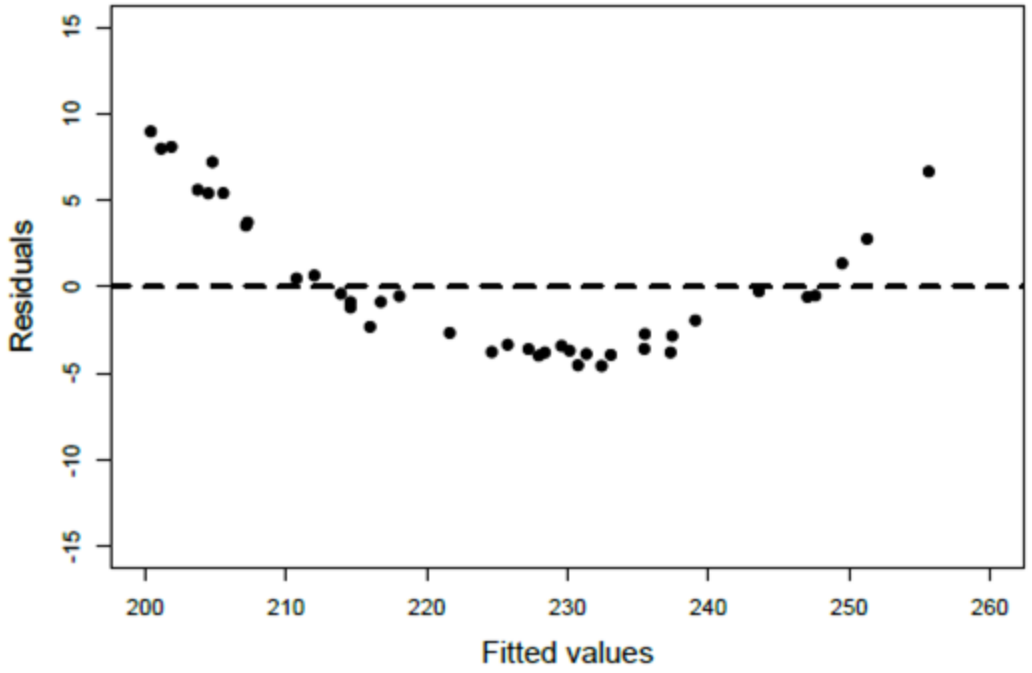


Residual plot



How to handle non-linearity?

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

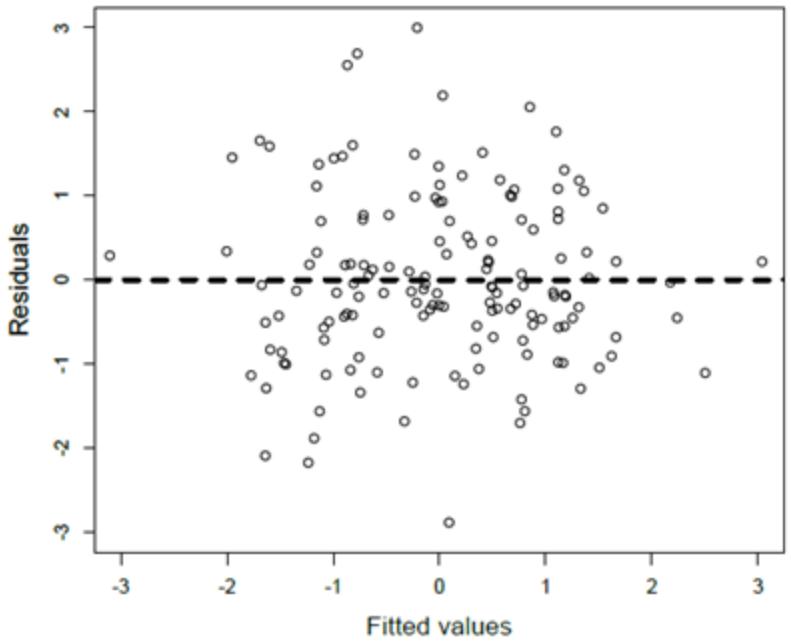


- Perform a nonlinear transformation of your outcome, e.g., by taking the log-transform.
- Perform a nonlinear transformation of the fixed effects. E.g., if age were related to pitch in a U-shaped way, then you could add age and age2 (age-squared) as predictors.
- If you're seeing stripes in the residual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

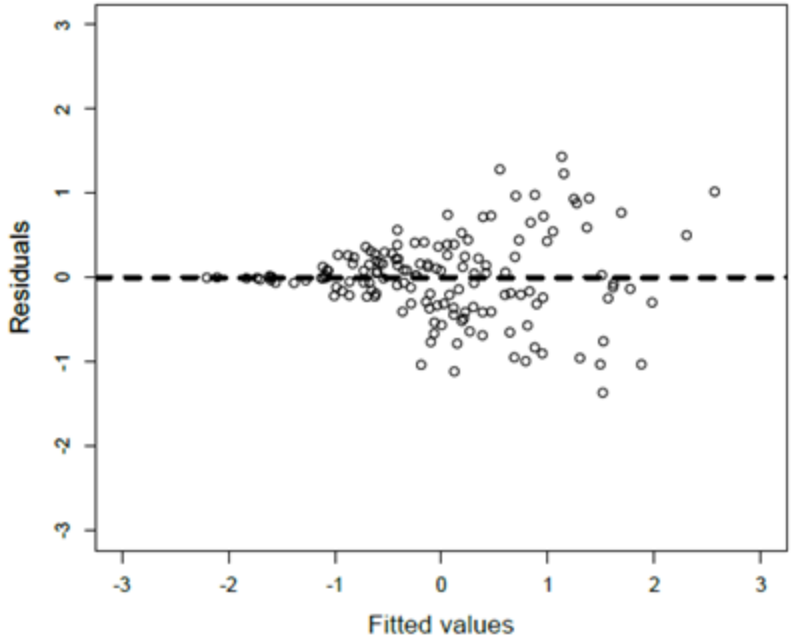
Homoskedasticity

- Unequal variances: Does the variance of the errors stay the same across all values of an independent variable (homoscedasticity) or does it change (heteroskedasticity).
  - Consider log-transforming the data.

homoscedastic



heteroscedastic



Assumptions of linear models: absence of collinearity

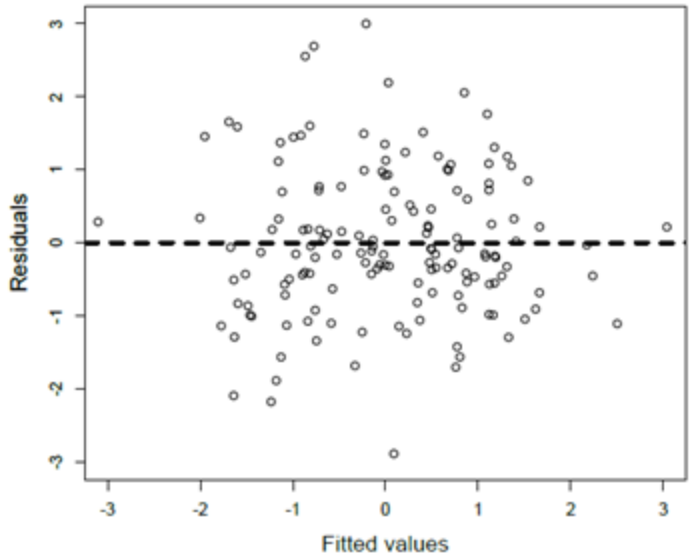
- If two fixed effects are correlated with each other, they are **collinear**.

- 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

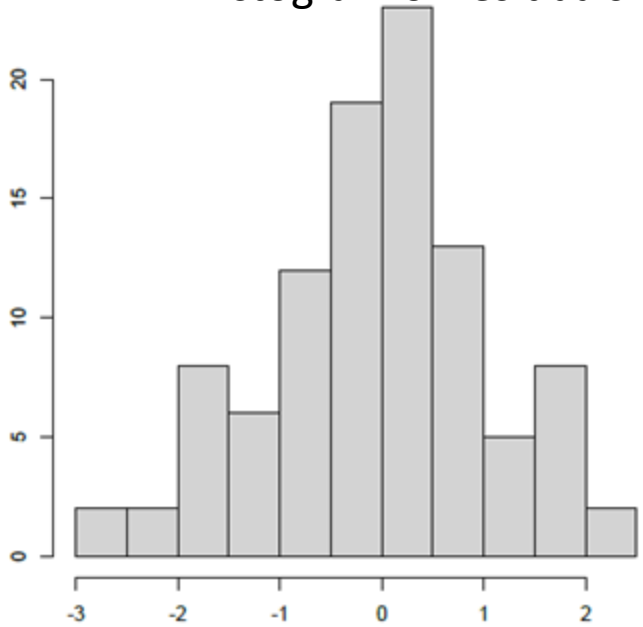
Normality of residuals – Least important

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

Residuals plot



Histogram of residuals



Independence assumption – Most important

- 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

What if outcome variable is binary?

Blood Alcohol Level	Speech Error
0.03	no
0.01	no
0.07	no
0.04	no
0.15	yes
0.09	yes
0.21	yes
0.18	yes
…	…

What assumptions are violated?

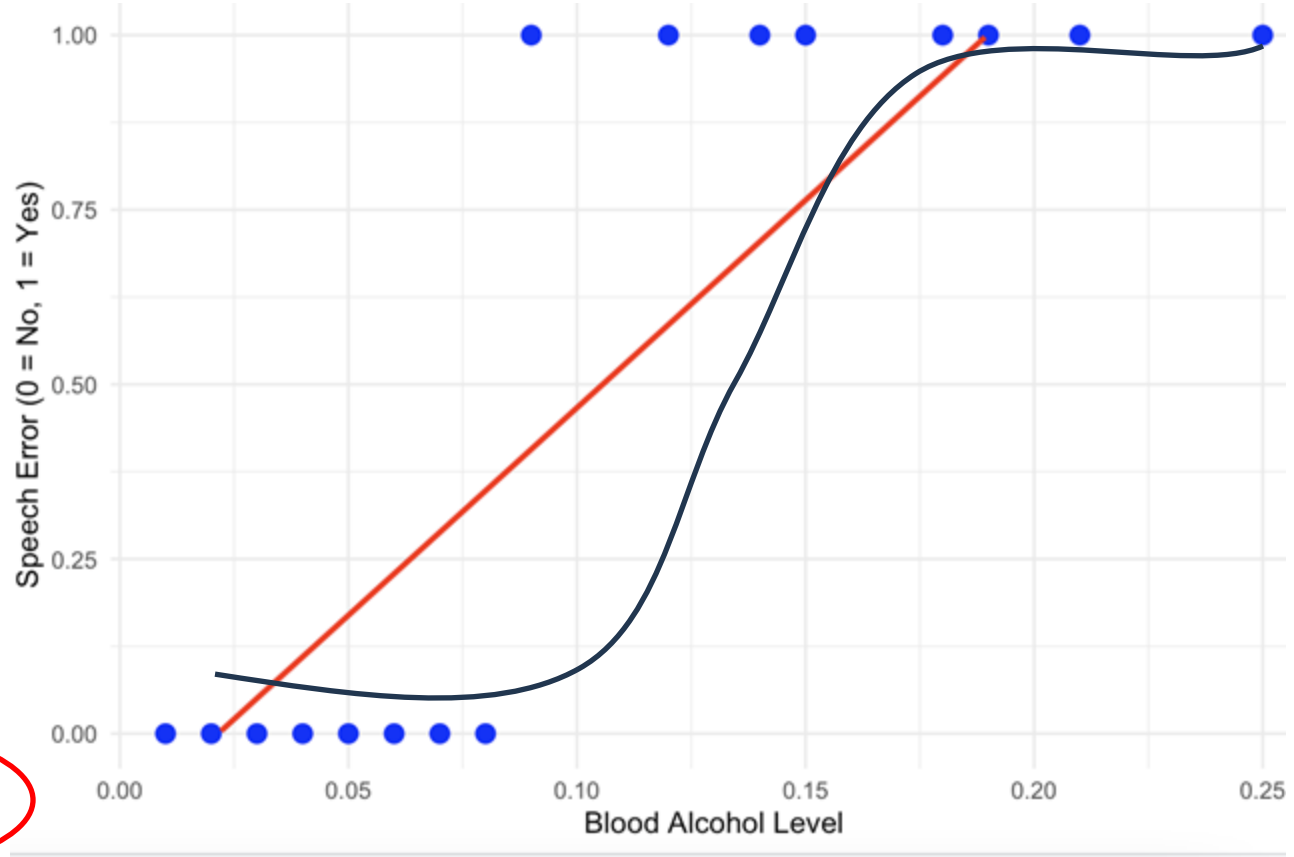
- **Linearity**: Non-linear relationship between predictors and outcome. Value of outcomes does not linearly increase with value of predictor.
- **Homoskedasticity**: Residuals for different values of the predictor vary.
  - Variance for binary variable: p(1-p) where p is probability of Y=1.
  - When p is close to 0/1, variance ~0.
  - When p=0.5, variance is 0.25

Logistic regression

- Instead of a linear fit, a curve would fit better?
- Sigmoid function: Constrains values between 0 and 1.

$$\sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+\exp(-z)}$$

- We'll just compute $\beta_0 + \beta_1 * x$ as earlier and then pass it through the sigmoid function.



This value is unbounded. We need a function to map it to 0 and 1. Sigmoid can map [-inf, +inf] to [0,1].

Idea behind logistic regression

- So, what does sigmoid gives us?
  - The probability of success (Y=1) given covariate X. $P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$
- Odds of success is defined as: $\text{Odds} = \frac{P(Y=1|X)}{1-P(Y=1|X)}$
- Which simplifies to: $\text{Odds} = e^{\beta_0+\beta_1 X}$
- Taking the natural logarithm: $\log(\text{Odds}) = \log\left(e^{\beta_0+\beta_1 X}\right) = \beta_0 + \beta_1 X$

Linear combination of predictors is equal to the log odds of success

An example

Model: am ~ mpg + wt

- am : 0 (automatic), 1(manual)
- mpg: miles per gallon
- wt: weight of car

```
Call:
glm(formula = am ~ wt + mpg, family = binomial, data = mtcars)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  25.8866    12.1935   2.123   0.0338 *
wt           -6.4162     2.5466  -2.519   0.0118 *
mpg          -0.3242     0.2395  -1.354   0.1759
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> mtcars
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE        16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL        17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC       15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
```

How do you interpret the coefficients?

Interpreting Coefficients

- Intercept: hard to interpret. Cars don't weigh 0 or have 0 horsepower.
  - But a baseline
- As weight increases by one unit, the log-odds of manual transmission decrease by ~8, holding mpg constant.
- As mpg increases by one unit, the log odds of it being manual reduce by 0.32, holding weight constant

```
Call:
glm(formula = am ~ wt + mpg, family = binomial, data = mtcars)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  25.8866    12.1935   2.123   0.0338 *
wt           -6.4162     2.5466  -2.519   0.0118 *
mpg          -0.3242     0.2395  -1.354   0.1759
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\log\left(\frac{P(am=1)}{P(am=0)}\right) = \beta_0 + \beta_1 \cdot wt + \beta_2 \cdot mpg$$

MIXED EFFECTS MODELS

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	Item/utterance	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

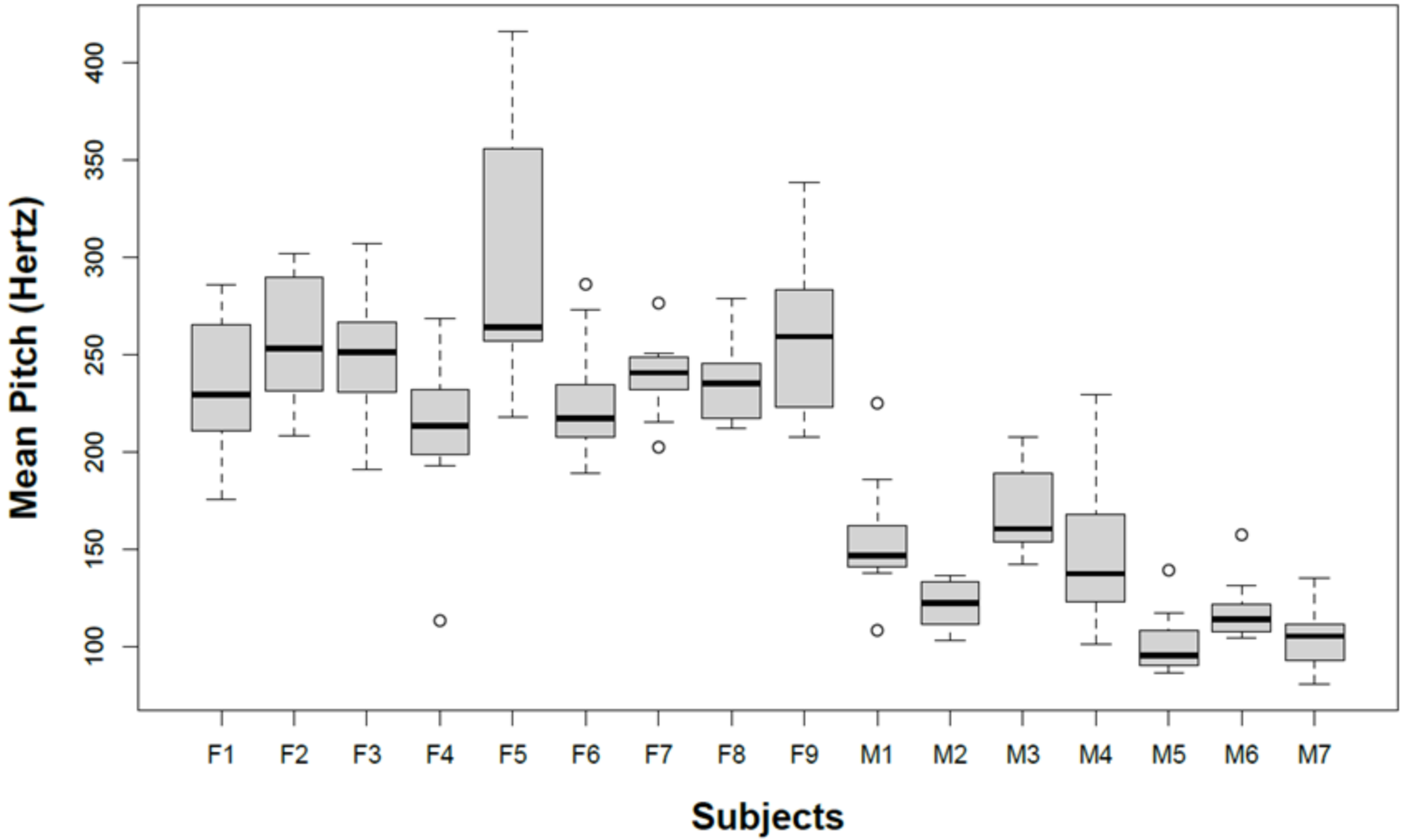
Independence assumption is often violated

- **Research question: Does politeness affect pitch**?

  pitch ~ politeness + sex + ε

- Each subject gives multiple responses: **polite** and **informal**
- 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



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.
  - 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.

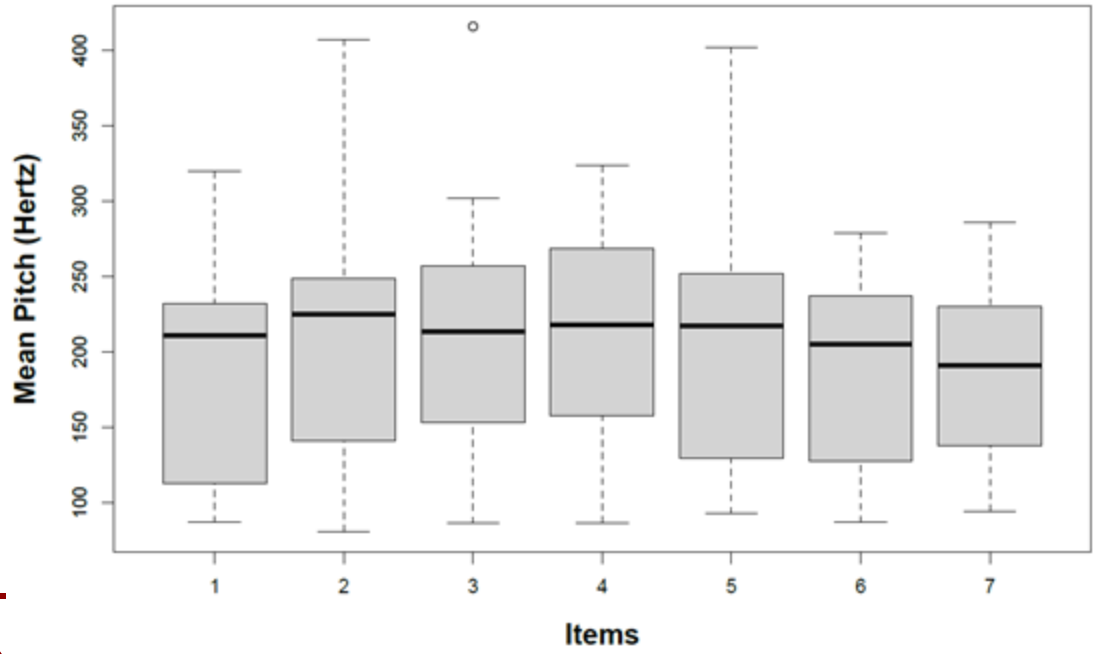
Mixed effects model

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

- "(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.

Making it more complex- 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



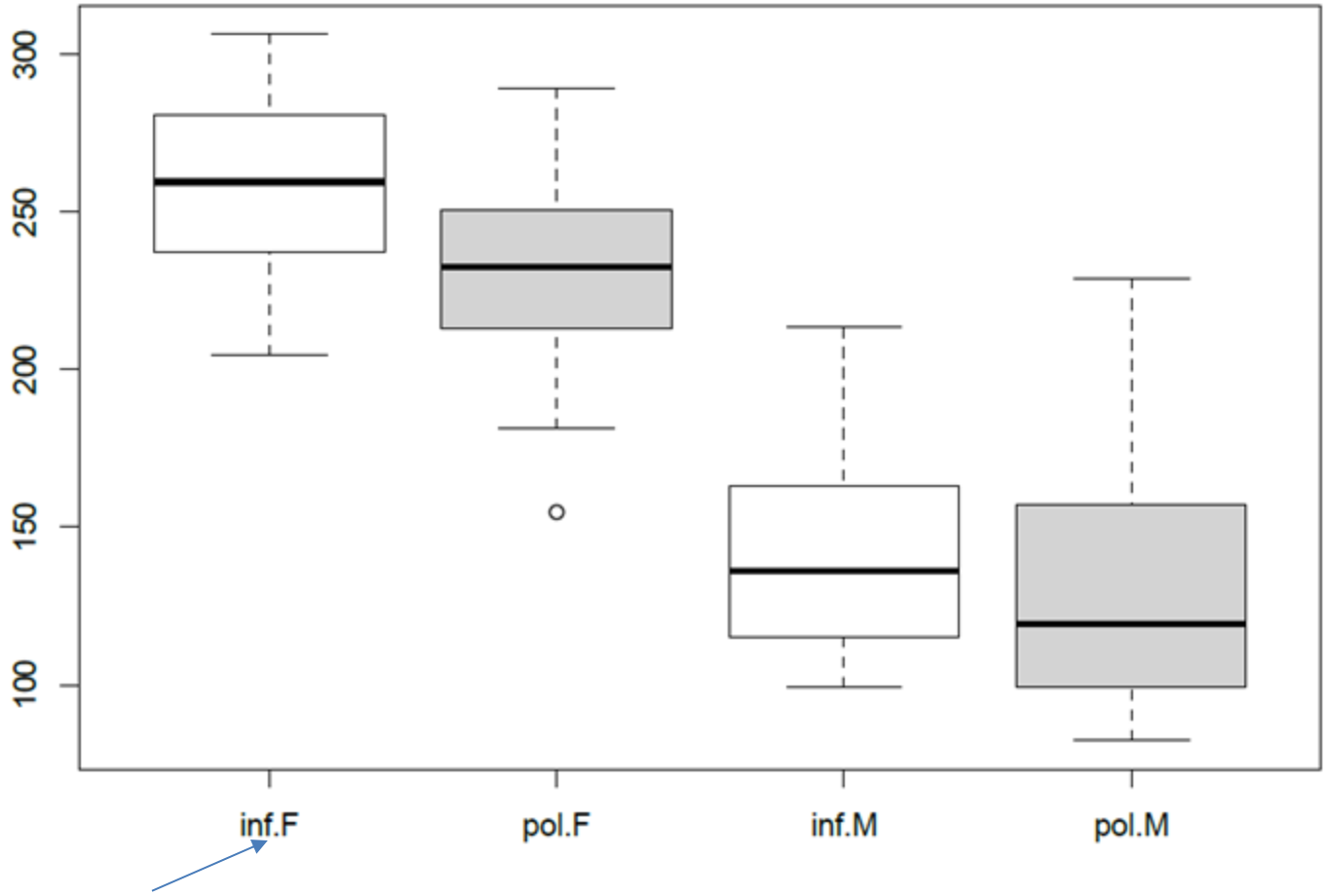
Multiple mixed effects model

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

- 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 ~ politeness + sex + (1|subject) + (1|item) + ε



Informal speech by females

Polite speech by males

Random effects

pitch ~ politeness + 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
```

- Tells the amount of variance between subjects and between scenarios/items.
- Residual - ε term – is the variability that is not due to “item” or “subject”

Fixed effects

pitch ~ politeness + 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 ~ politeness + 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"
```


Best Practice
1
2
3

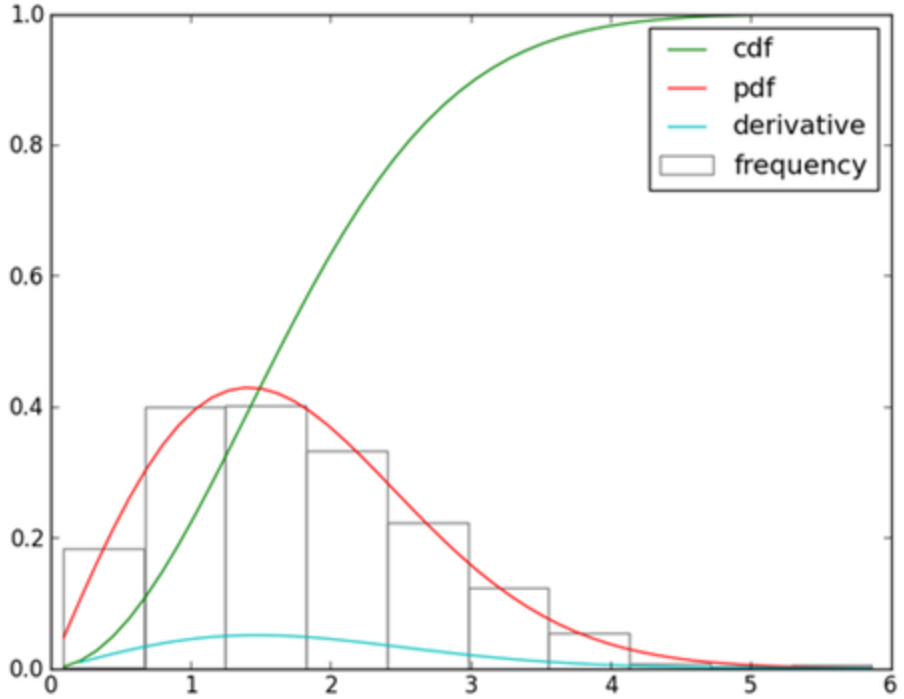
Best practices for data analysis

- Exploratory Data Analysis
- Data cleaning
  - Missing data
  - Outliers
- Feature engineering
  - Feature standardization
    - E.g., when features have a large range, take log transform
  - Correlated features
    - Feature selection
    - Regularization
- Modeling considerations
  - Overfitting
  - Information leakage
  - Finding the optimal model

Exploratory data analysis

All data analysis starts with exploratory data analysis and visualization

- Scatterplots to assess relationships
- Correlations – Pearson, Spearman?
- Data distributions:
  - **Histogram**: visualizes the frequency distribution of data values
  - **PDF** (Probability density function): Continuous representation of the likelihood of a random variable takes a particular range of values.
  - **CDF** (Cumulative Distribution Function): represents the probability that a random variable takes on a value less than or equal to a given value



Example: Distribution of dice roll

Data cleaning

- Common issues in tabular data
  1. Duplicate records
     - deduplicate
  2. Outliers
     - If due to an error (e.g., non-physical measurements such as negative counts), delete
  3. Missing values.

Data cleaning: outliers

Outliers deviate significantly from the majority of the data and can affect analysis. Understanding the causes of outliers is crucial for determining whether they should be removed or retained

- Mistakes in data collection and measurement errors
  - remove
- Natural variability and rare events
  - Retain and transform extremely large or small values
- Sampling errors: E.g., oversampling extreme values.
  - Resampling (in some cases)
- Methods to detect outliers:
  - Visualization: Box plots, scatter plots.
  - Statistical methods: IQR (Interquartile Range). [Q3 – Q1]
  - Thumb rule: <Q1-1.5*IQR and >Q3+1.5*IQR

Missing values

- Types of missingness
  - Missing at Random (MAR).
  - Missing Not at Random (MNAR)
  - May introduce systematic errors or bias
- Handling missing data
  - Deletion
  - Imputation:
    - Mean, median, mode replacement.
    - Predictive imputation.

Feature transformation

Transform features to make more interpretable model

- Example: Create a model of male faculty salary at UCI. Use the model to predict female faculty salary.
- Model Salary as a function of X:
  - **Salary = 3,420,751 – 293*YoB – 808*YoD – 593*YoH**
    - Year of birth (YoB)
    - Year of hire (YoH)
    - Year of degree (YoD)
- Instead, model Salary as a function of
  - **Salary = 48,623 + 293*Age + 808*YsD + 593*YsH**
    - Years since birth (Age)
    - Years since degree (YsD)
    - Years since hire (YsH)

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

Standardization

- Z-score is a data standardization technique that rescales data to have a mean of 0 and a standard deviation of 1.
  - Converts data points into their respective z-scores, which measure how many standard deviations (sigma) a value is away from its mean in the dataset.
  - z=(x-mean)/sigma
- Advantages
  - Puts all data on a comparable scale (normalization)
  - Helps identify outliers
  - Prepares data for machine learning algorithms

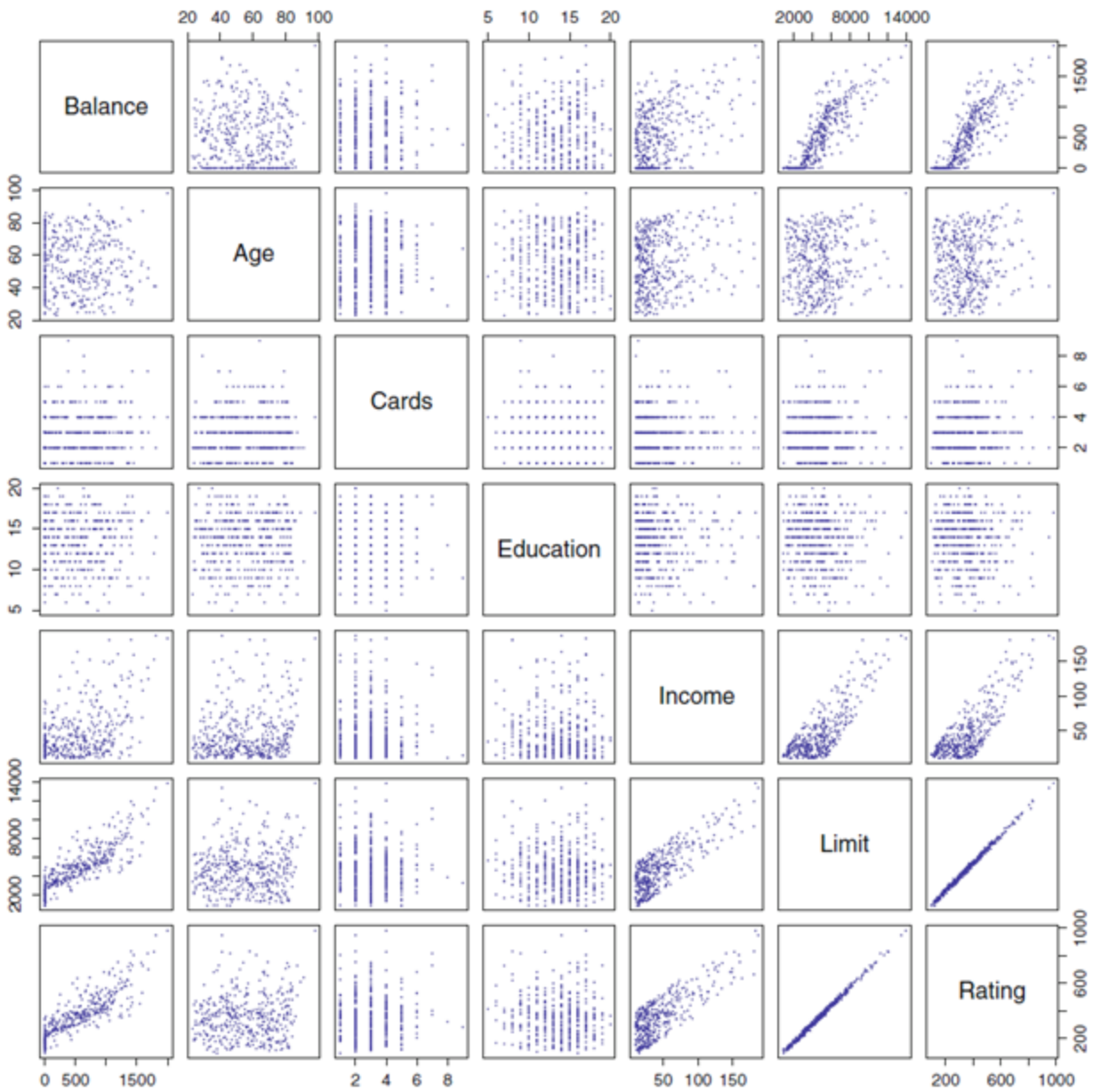
Feature Selection

- Subset Selection
  - Identify a subset of features most related to the outcome; fit a model using OLS on the reduced set.
  - Methods: forward feature selection, MRMR, …
- Shrinkage (Regularization)
  - Involves shrinking the estimated coefficients toward zero relative to the OLS estimates; has the effect of reducing variance and performs variable selection.
  - Methods: ridge regression, lasso
- Dimension Reduction
  - Involves projecting the p predictors into a M-dimensional subspace, where $M < p$, and fit the linear regression model using the M projections as predictors.
  - Methods: principal component analysis (PCA), principal components regression, partial least squares

Multi-collinearity (Credit Card Data)

<u>Scatter plots</u>

- Outcome:
  - Credit card **balance**
- Features:
  - **Age**,
  - number of **Cards**,
  - years of **Education**,
  - **Income**,
  - credit **Limit**
  - credit **Rating**,



Handling Multi-collinearity

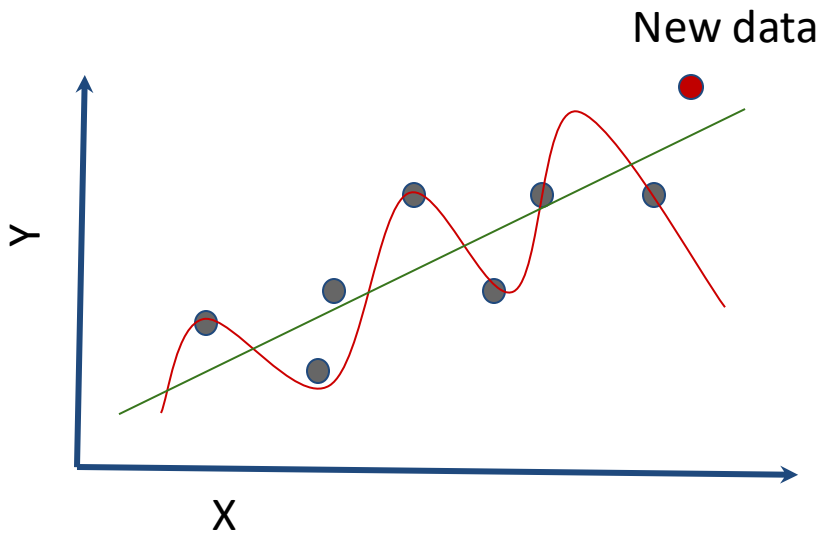
- Reduce the number of colinear features by eliminating un-informative features
- **Variance Inflation Factor** - quantifies the severity of multicollinearity and measures how much the variance of an estimated regression coefficient is increased because of collinearity.
  - Rule of thumb – calculate 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)

Choosing the Optimal Model

- The most complex model (using most predictors/features) will have the lowest **training error**: the smallest RSS and the largest R^2
- Alternatively, choose a model with low **test error.** *Remember that training error is usually a poor estimate of test error*.
  - Estimate test error on held-out data or using a cross-validation approach.
  - Thus, RSS and R^2 are not suitable for selecting the *best* model among a collection of models with different numbers of predictors.

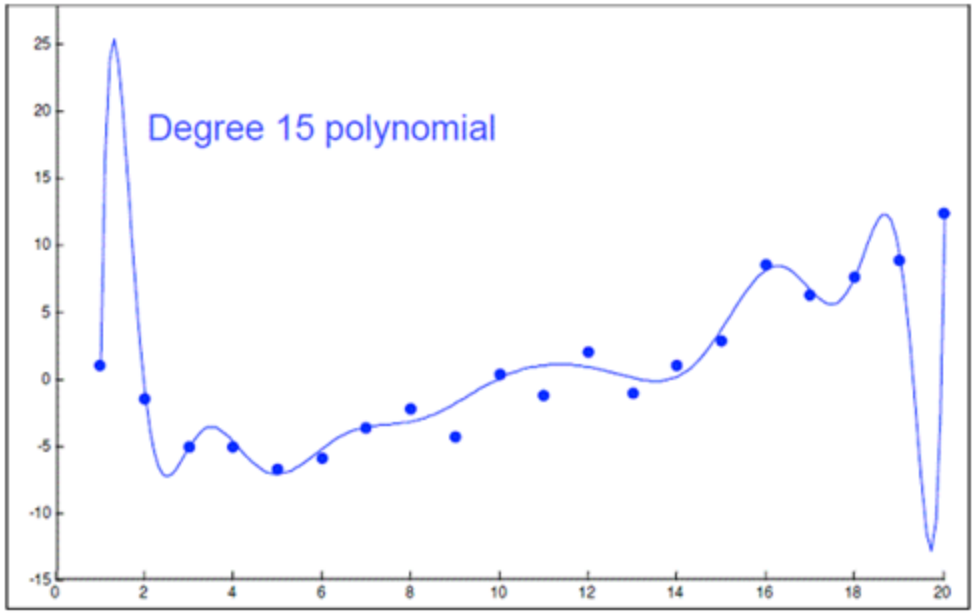
Overfitting

- 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



Overfitting

- Carefully selected features can improve model accuracy, but adding too many can lead to overfitting.
  - Overfitted models describe random error or noise instead of any underlying relationship.
  - They generally have poor predictive performance on test data.

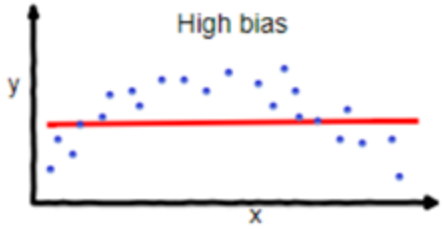


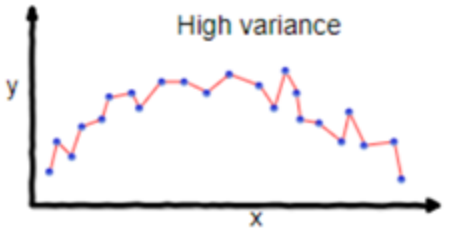
- For instance, we can use a 15-degree polynomial function to fit the following data so that the fitted curve goes nicely through the data points.
- However, a brand new dataset collected from the same population may not fit this particular curve well at all.

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

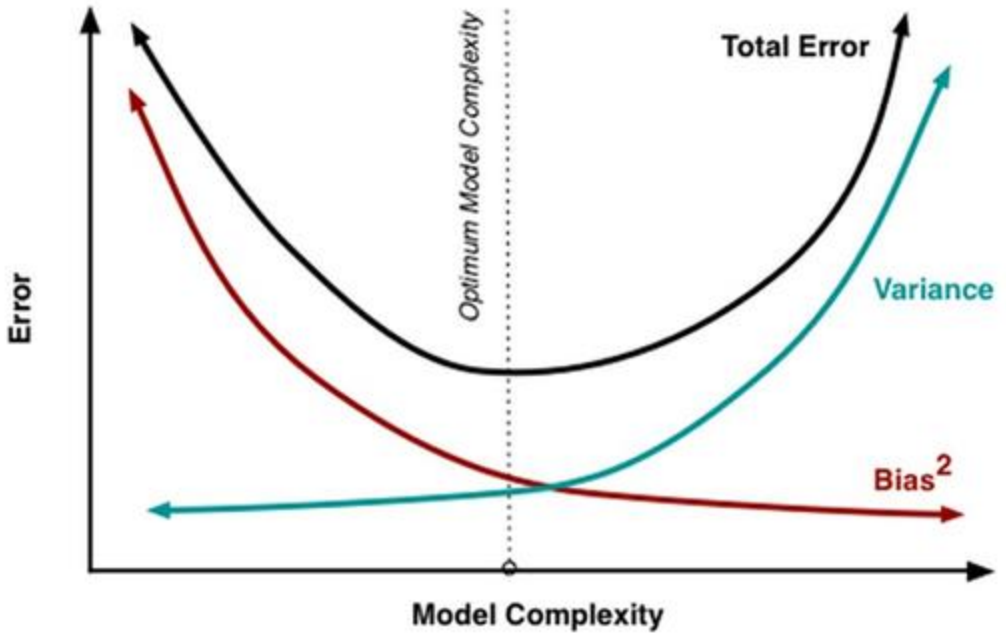




Bias-variance tradeoff

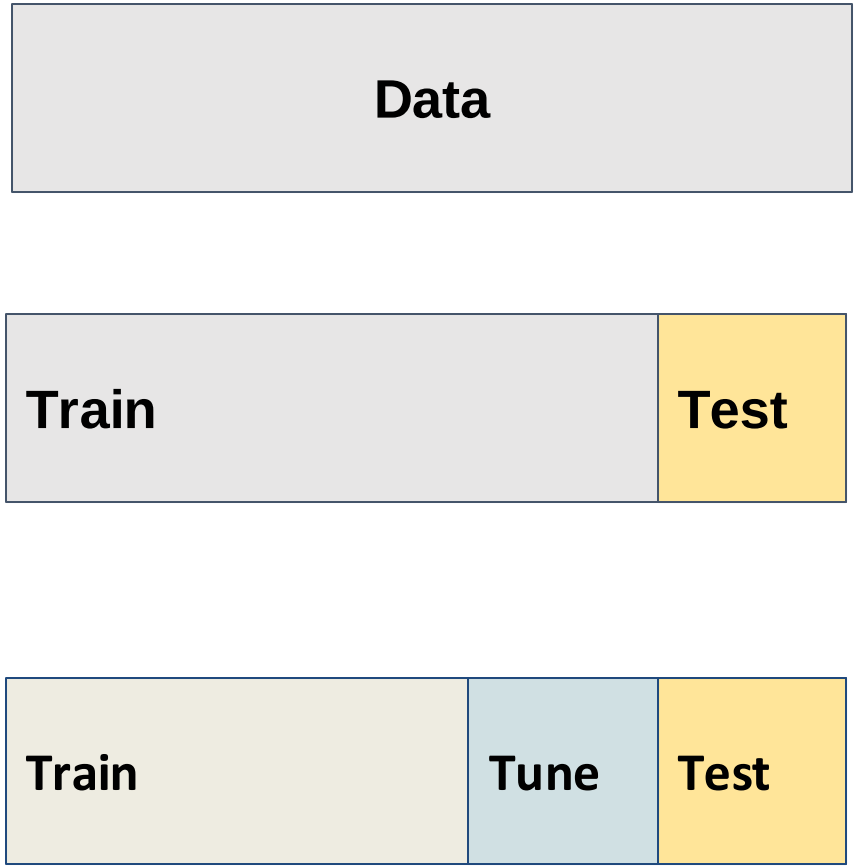
Unfortunately, it is not always possible to balance both variance and bias at the same time. In general, bias is reduced if we add more and more parameters to a model and make it more complex.

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.



Information leakage

- Avoid information leakage
  - Never test on **the same** data used for training
  - Cross validation, e.g., 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

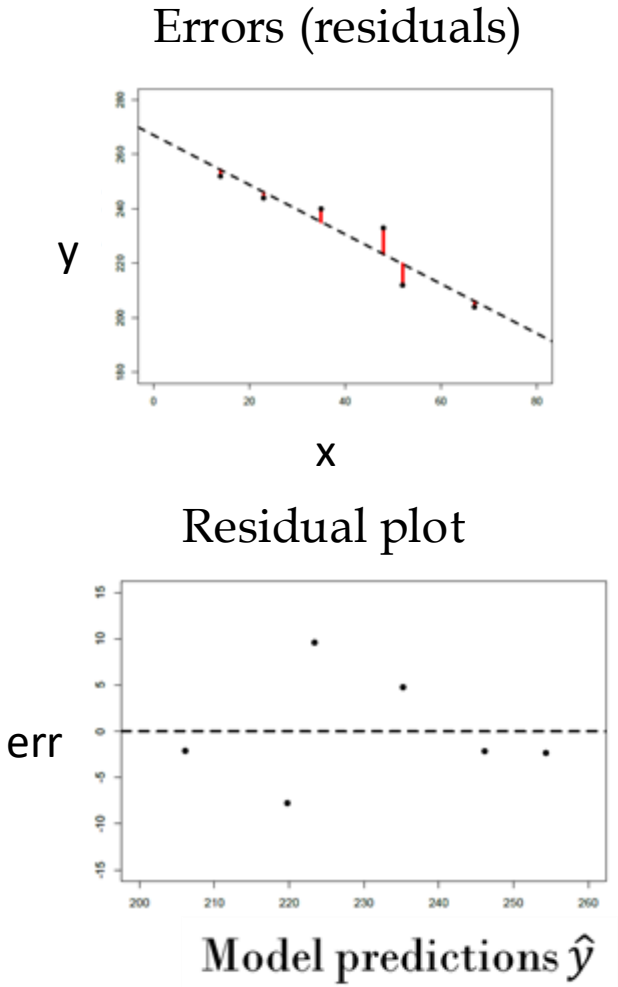


Model Interpretability

- Large number of predictors – many are redundant or have little effect.
- Including such irrelevant variable leads to unnecessary complexity.
- Leaving these variables in the model makes it harder to see the effect of the important variables.
- The model would be easier to interpret by removing the unimportant variables, e.g., setting their coefficients to zero.

Linear models: Summary

- Outcome as a weighted sum of features: flexible, computable and interpretable models
  - Linearity (check residuals) □ transform features
  - Homoscedasticity (check residuals) □ consider transforming data
  - Normality (check residuals): p-values may be invalid
  - Correlated features □ dimensionality reduction
  - Independence of observations □ mixed effects models
  - Outliers
  - Interpretable



Readings

- https://bodowinter.com/tutorial/bw_LME_tutorial1.pdf
- https://bodowinter.com/tutorial/bw_LME_tutorial2.pdf
- http://bigdatasummerinst.sph.umich.edu/wiki2019/images/d/db/LogisticRegression-slides.pdf

Logistics

- Homework 1 will be released tonight.
- Due 4PM January 27, 2025
- No class on January 20th.