Logistic regression

Daniel Fuller
Associate Professor
Department of Community Health and Epidemiology

UNIVERSITY OF SASKATCHEWAN

Hierarchy of Study Design (mine)

- Systematic review + meta analysis
- Randomized designs (RCT)
- Natural experiments
- Cohort studies
- Case control studies
- Cross-sectional & ecological studies

Study design and regression

- All study designs can and do use regression to estimate associations (effects)
- It's crucial to consider the design when thinking about regression.
  - An Odds Ratio (OR) from one study is not the same as one from another study.

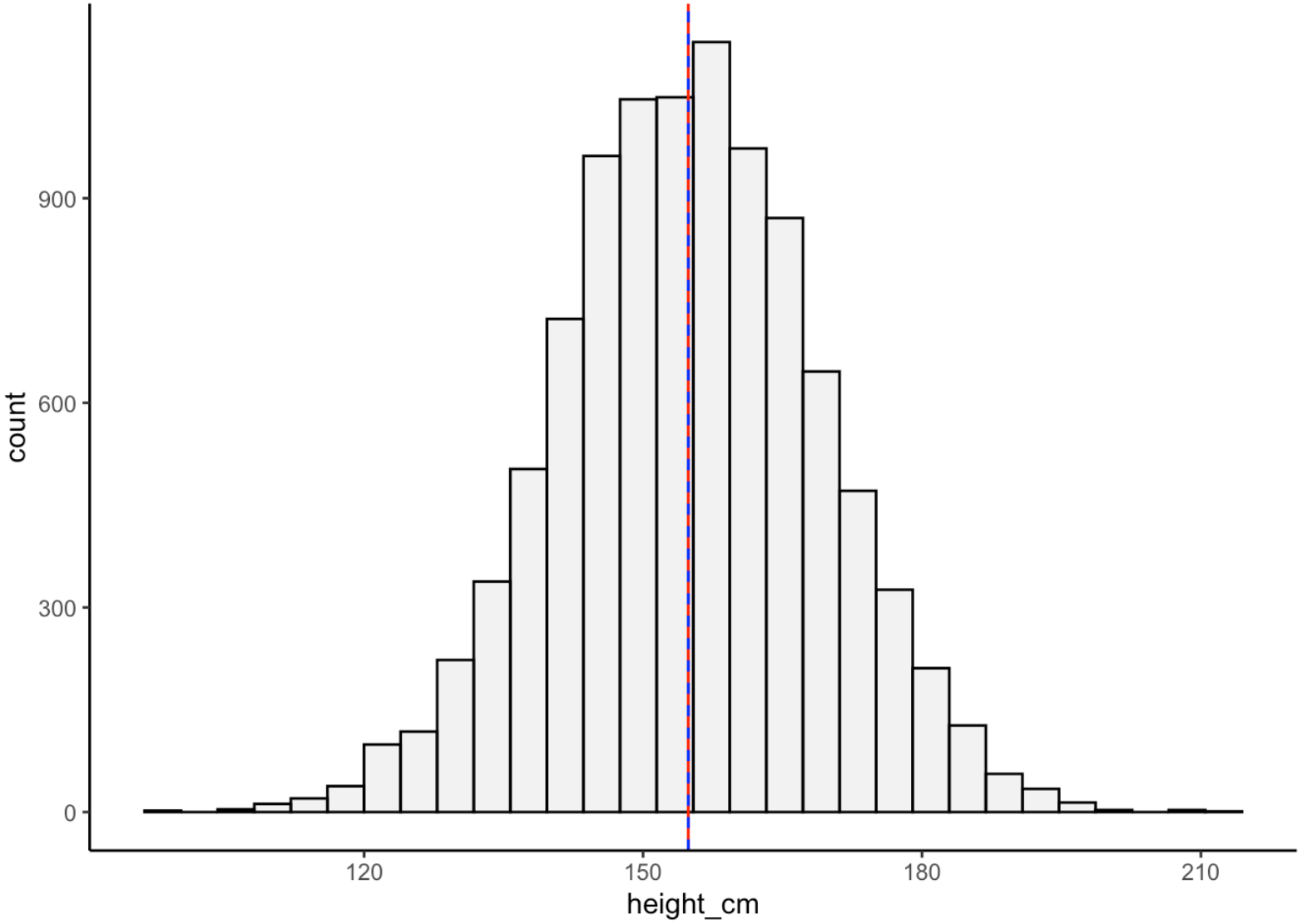
Regression in general

- Create a model that fits the data
- Common model types are named because of the distribution of the outcome variable
    - Or assumptions required for the models distribution

Common Distributions

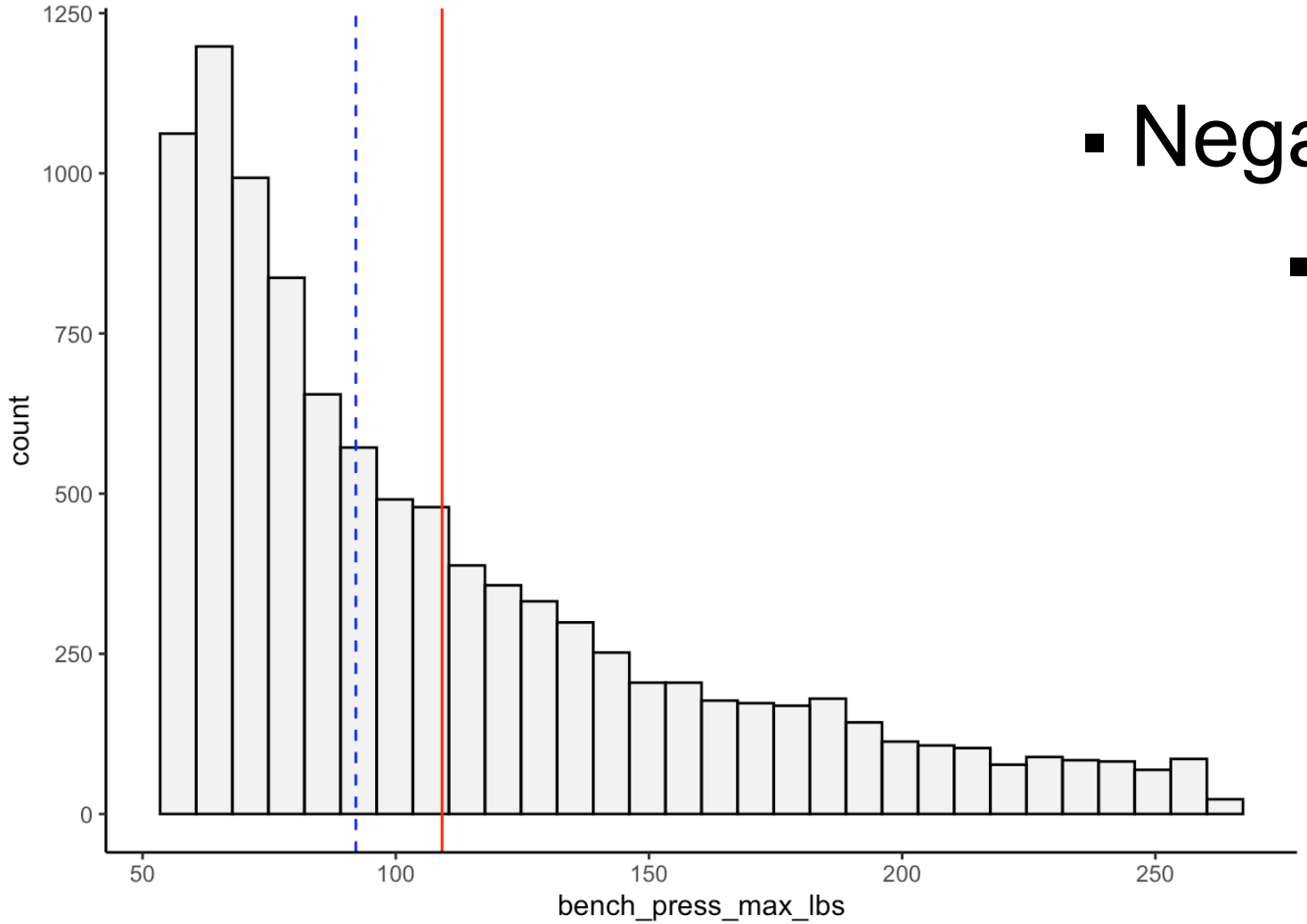
- Normal or Gaussian
  a) Bell curve
- Uniform
- Count type
  a) Poisson and negative binomial
- Bernoulli
  a) Yes or no
- Many, many, more
  a) https://en.wikipedia.org/wiki/List_of_probability_distributions

Normal (Gaussian) distribution



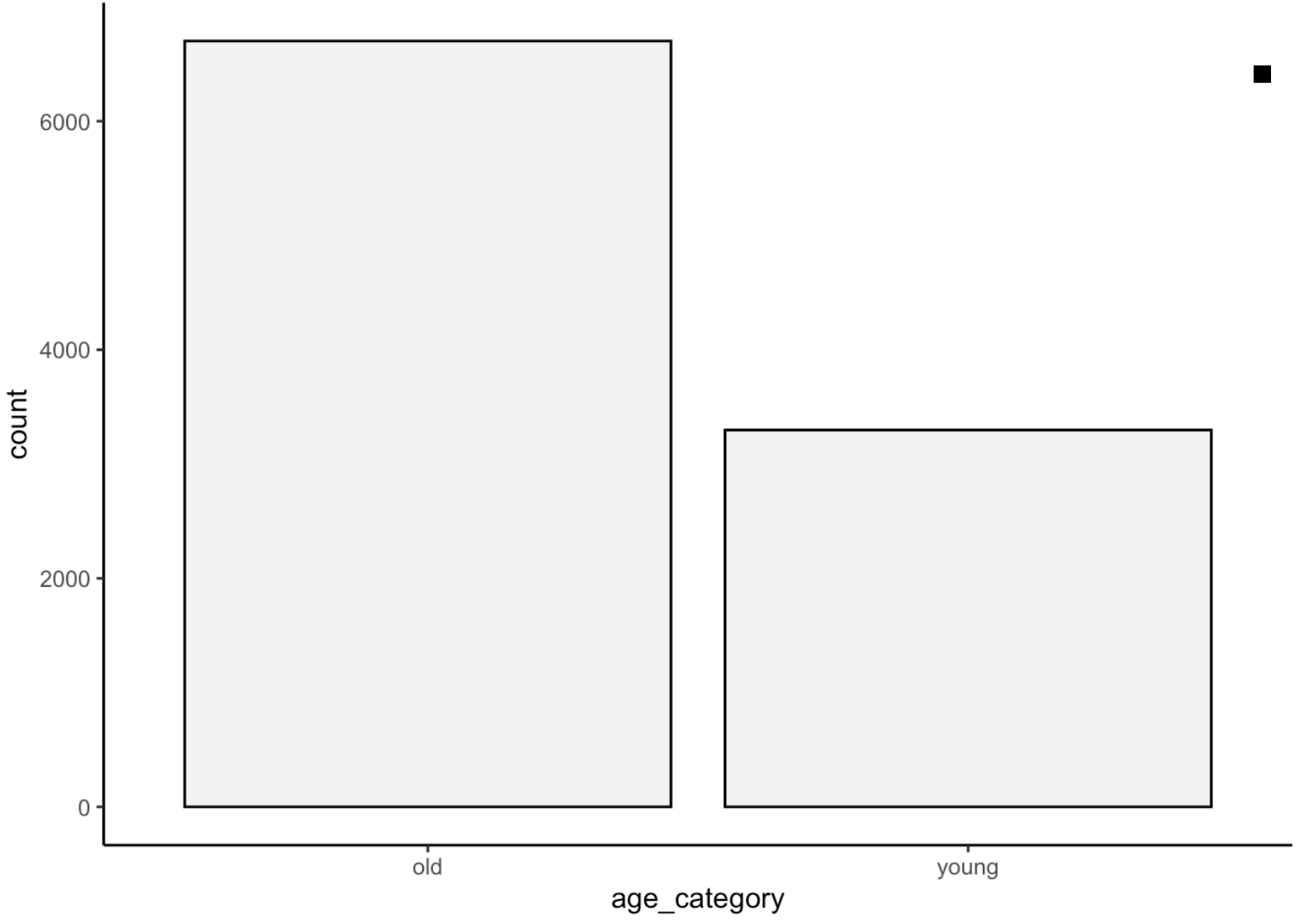
- Linear regression

Count type distribution



- Poisson models
- Negative binomial models
- Zero inflated models

Bernoulli distribution



- Logistic regression
  - Probit regression
    - Tobit regression

UNIVERSITY OF SASKATCHEWAN

Logistic regression

Logistic regression

- Dichotomous outcome
- Modelling logit transformation of the probability of the outcome as a linear function of a set of predictor variables
  a) Log odds of the outcome
- This transformation leads to the logistic model
- Can compute the odds of disease and odds ratio for disease that is associated with the presence of factor X (predictor)
- Can be extended to multiple predictors

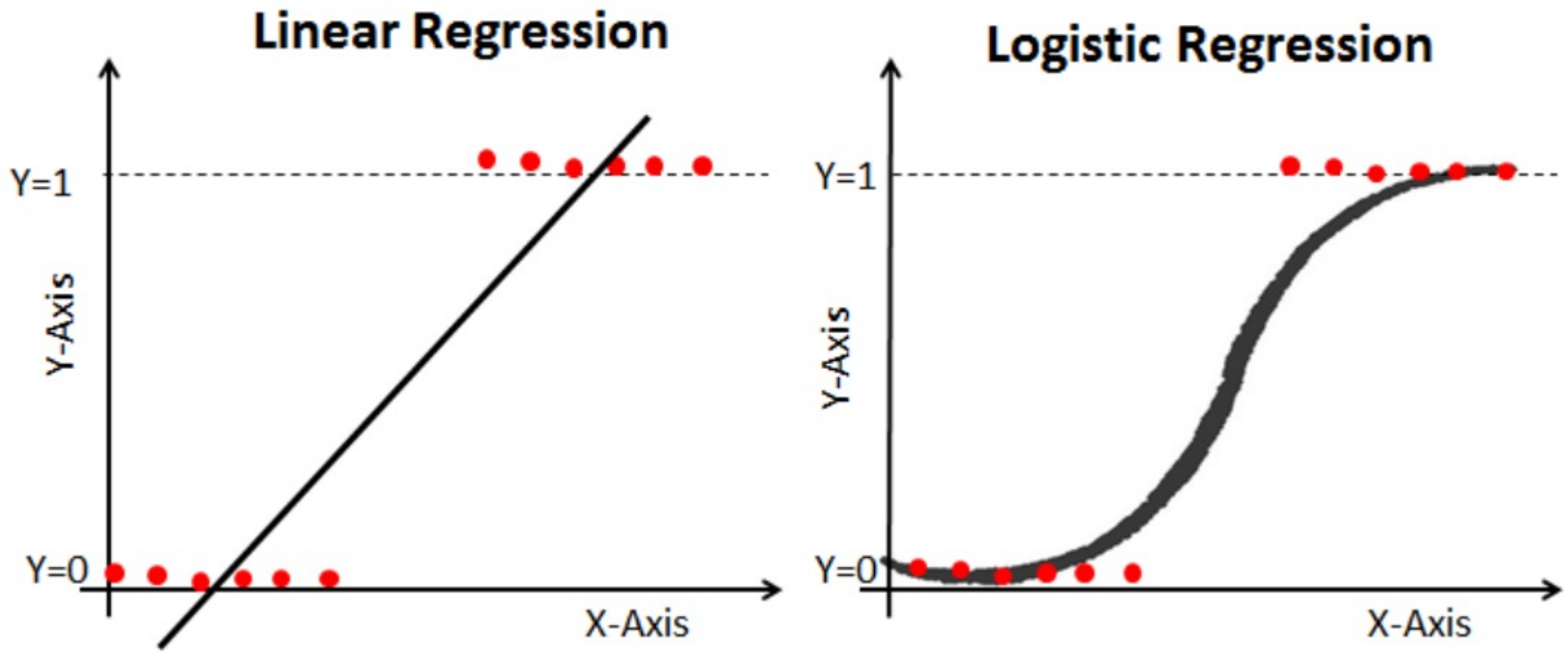
Fitting a logistic model

- Set of parameter values that are most likely to have produced the observed data
  a) Maximum likelihood (ML) estimates
  b) In reality, these will be very, very small
  c) Because of this (and because the estimation is easier) software packages usually work with the log likelihood
- Parameter estimate that gives you the maximum likelihood of having the binary outcome values 0/1

Typically for rare(ish) outcomes

- Distribution of 1/0 (cases/non cases) should be 5-15% for one and the inverse for the other.
- If the cases are too common logistic regression will overestimate the effect size
  a) Probit regression
  b) Tobit regression

- Software package side note
  a) Make sure you are modelling the right level of your outcome variable
  b) Some take 1 as the reference others take 0 as the reference by default


Linear Regression
Y=1
Y-Axis
Y=0
X-Axis
Logistic Regression
Y=1
Y-Axis
Y=0
X-Axis

Likelihood ratios

- To determine overall significance of the model = likelihood ratio test
  a) Compares the likelihood of the ‘full’ model (predictors added) with the ‘null’ model (no predictors)
     - Analogous to F-test in linear regression
  b) For model comparison, the models must be nested (predictors in the simpler model must be a subset of those in the full model)
- To determine significance of individual predictors = Wald test

Model interpretation

- The coefficients in a logistic regression model represent the amount the logit of the probability of the outcome changes with a unit increase in the predictor
  a) Not easy to interpret!
- Therefore – convert the coefficients to odds ratios
  a) Exponentiation
- Continuous predictors: OR represents the factor by which the odds of disease are multiplied for 1-unit change in the predictor
- Categorical predictor: as with linear regression, convert to dummy variables – coefficient represents the effect of that level to baseline

Logistic regression: small group discussion

- We are interested in how variables, such as GRE score, grade point average and ‘prestige’ of the undergrad institution effect admission to graduate school
- Outcome: admit (binary variable: admit/don’t admit)
- Predictors: gre, gpa (continuous variables) and rank (1 to 4; 1 is ‘highest’ prestige)
- Interpret the predictors
- Does this group of predictors reliably predict the dependent variable? How do you know?

```
logit admit gre gpa i.rank

Iteration 0:   log likelihood = -249.98826
Iteration 1:   log likelihood = -229.66446
Iteration 2:   log likelihood = -229.25955
Iteration 3:   log likelihood = -229.25875
Iteration 4:   log likelihood = -229.25875

Logistic regression                               Number of obs   =        400
                                                  LR chi2(5)      =      41.46
                                                  Prob > chi2     =     0.0000
Log likelihood = -229.25875                       Pseudo R2       =     0.0829

------------------------------------------------------------------------------
       admit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |   .0022644    .001094     2.07   0.038     .0001202    .0044086
         gpa |   .8040377   .3318193     2.42   0.015     .1536838    1.454392
             |
        rank |
          2  |  -.6754429   .3164897    -2.13   0.033    -1.295751   -.0551346
          3  |  -1.340204   .3453064    -3.88   0.000    -2.016992   -.6634158
          4  |  -1.551464   .4178316    -3.71   0.000    -2.370399   -.7325287
             |
       _cons |  -3.989979   1.139951    -3.50   0.000    -6.224242   -1.755717
------------------------------------------------------------------------------
```

Source: IDRE

```
logit , or

Logistic regression                                     Number of obs   =        400
                                                        LR chi2(5)      =      41.46
                                                        Prob > chi2     =     0.0000
Log likelihood = -229.25875                             Pseudo R2       =     0.0829

------------------------------------------------------------------------------
       admit | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |   1.002267   .0010965     2.07   0.038      1.00012    1.004418
         gpa |   2.234545   .7414652     2.42   0.015     1.166122    4.281877
             |
        rank |
          2  |   .5089309   .1610714    -2.13   0.033     .2736922    .9463578
          3  |   .2617923   .0903986    -3.88   0.000     .1330551    .5150889
          4  |   .2119375   .0885542    -3.71   0.000     .0934435    .4806919
------------------------------------------------------------------------------
```

Confounding and interaction

As with linear regression

- Confounding: ‘substantial’ (typically >10%) change in coefficient
  a) Statistical assessment – don’t forget to also include confounding variables from your DAG regardless of statistical assessment results
- Interaction: Cross-product term

Logistic regression assumptions

- Independence
  a) Observations are independent from one another
- Linearity
  a) Any predictor measured on a linear scale is assumed to have a linear relationship with the outcome
- Regression model diagnostics:
  a) A whole topic in and of itself

Studying

- So many primers on logistic regression!
  a) Harris JK. Primer on binary logistic regression. Fam Med Community Health. 2021 Dec;9(Suppl 1):e001290. https://doi.org/10.1136%2Ffmch-2021-001290
  b) https://github.com/jenineharris/logistic-regression-tutorial

Goal of model building

- Best fit and parsimonious
- “Simplest model using the minimum number of parameters needed to explain a given phenomenon”

  - Raykov & Marcoulides, 1999

UNIVERSITY OF SASKATCHEWAN

Model-building

- Focus here is on **statistical models** – but there are others
- History in epidemiology:
  a) William Farr, 1840
- Decide on the goal of the model building
  a) Predict vs causal
- Multivariable model building
  a) Multivariate vs multivariable? https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3518362/
  b) Incorporate both statistical and subject matter knowledge
- Balance best fit with parsimony
  a) Definition of ‘best fit’ and parsimonious will depend on the goal
  b) For example, in machine learning, you may not want the most parsimonious

Objectives of model-building

- In general, regression modelling has one of two broad objectives:
  1. Best fitting model to predict future observations
     - Details of model may not matter (such as effect of specific predictors) but need to exclude any variable with a questionable relationship with the outcome
  2. Understand the relationship(s) (potentially causal) between one or more predictors and outcome of interest
     - Requires most precise coefficient estimates possible for variables of interest; careful attention to confounding and interaction
     - Most often the goal of epidemiological models

Model-building: Steps

1. Specify the maximum model to be considered i.e.) identify the outcome and full set of predictors
2. Specify the criterion/criteria to be used in selecting the variables to be included in the model
3. Specify the strategy for applying the criterion/criteria
4. Conduct the analyses
    1. Forward or backward step by step approach
    2. Stepwise forward or backward
    3. Best subset method.
5. Evaluate reliability of model chosen

Model building

- Is an **art** and a **science**
  a) Both non-mathematical and mathematical
- Draw causal diagram
- Know your data, research question, study design, hypothesis, goal
- Iterative process
  a) Fitting, assessing, comparing, diagnosing issues, fitting, assessing, comparing, diagnosing…
  b) Potential interaction, confounding, violating model assumptions, ‘best fit’,
- Parsimony, best fit and interpretation

Model assumptions

Linear regression

- Validity and representativeness
- Additivity and linearity
- Independence of errors
- Equal variance of errors
- Normality of errors

Logistic regression

- Validity and representativeness
- Additivity and linearity
- Independence of errors

How to assess?

To assess:

- Plot outcome vs fitted values
- Plot residuals vs fitted values
- Plot residuals vs continuous covariates
- Plot residuals vs time/spatial/cluster dependent covariates
- QQ plots of the residuals

Other

- Multicollinearity of predictors
  a) Variance inflation factor
- Influence measures of particular observations/covariate patterns
- For binary and survival outcomes: calibration and discrimination

UNIVERSITY OF SASKATCHEWAN

Paper Review

- Lavoie K, Gosselin-Boucher V, Stojanovic J, Gupta S, Gagné M, Joyal-Desmarais K, Séguin K, Gorin SS, Ribeiro P, Voisard B, Vallis M, Corace K, Presseau J, Bacon S; iCARE Study Team. Understanding national trends in COVID-19 vaccine hesitancy in Canada: results from five sequential cross-sectional representative surveys spanning April 2020-March 2021. BMJ Open. 2022 Apr 5;12(4):e059411. https://doi.org/10.1136/bmjopen-2021-059411

- What is the outcome variables
  a) How was it created?
- What are the predictor variables
  a) Why were they included in the model
  b) What are their associations with the outcome
- What is the study design?
- Is the outcome common or not?