Regression Model Building 1: Overview

Michael Otterstatter

BCCDC Biostats Session

July 19, 2019

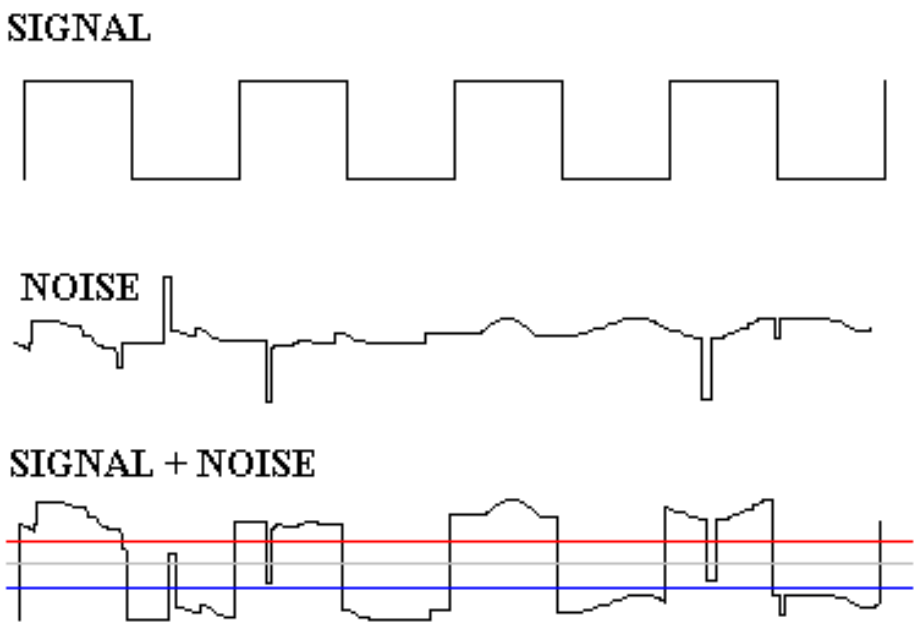
Session overview

- In this session we will discuss
  - Key components in the model building process
  - Models as a tool for exploring and describing data

Background

- Recall, data can be thought of as

observations = signal + noise



Underlying pattern of interest

Other sources of variation

What we actually observe

Background

- Statistical models try to understand the relationship of signal and noise that generates data

observations = signal + noise

$$\overbrace{Y} = \overbrace{\alpha + \beta_1 X_1} + \overbrace{\varepsilon}$$

- If we understand this relationship, we can
  - succinctly <u>describe</u> patterns
  - <u>explain</u> observed patterns
  - <u>predict</u> future patterns

Model building

- Building a regression model requires careful thought throughout, not simply a ‘cookbook’ activity of following predefined steps

- In general, you must consider and decide
  - What is the **purpose of my model** (describe, explain, predict)?
  - What **type of model** is appropriate for my purpose and data (ordinary linear, generalized linear, etc.)?
  - What is the **best fit model** for my data?

Regression model basics

An **ordinary linear model** represents patterns in our data with a straight line, plus unexplained (error) variation

intercept parameter

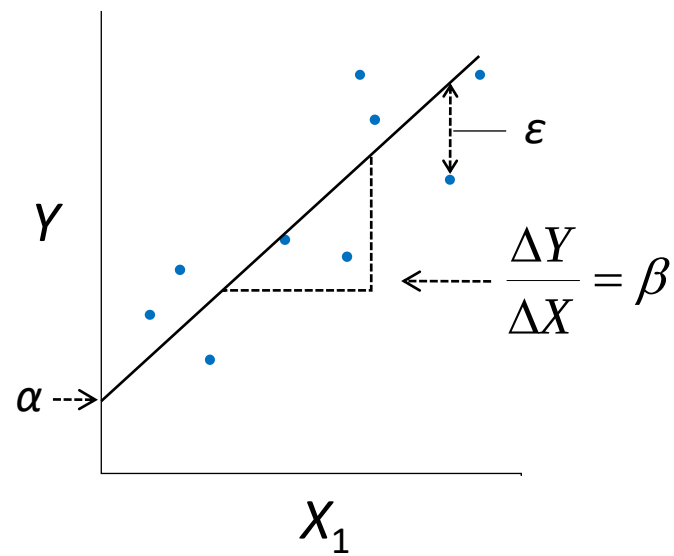
predictor variable

$$Y = \alpha + \beta_1 X_1 + \varepsilon$$

response variable

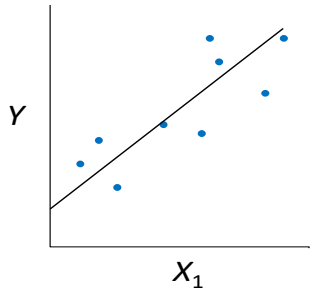
slope parameter

error 'residual'

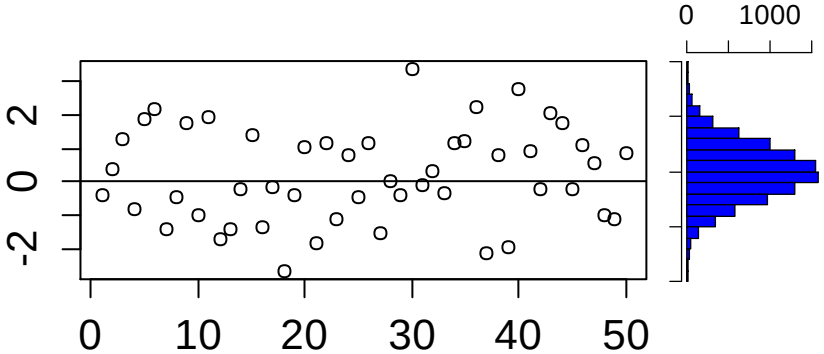


Ordinary linear model assumptions

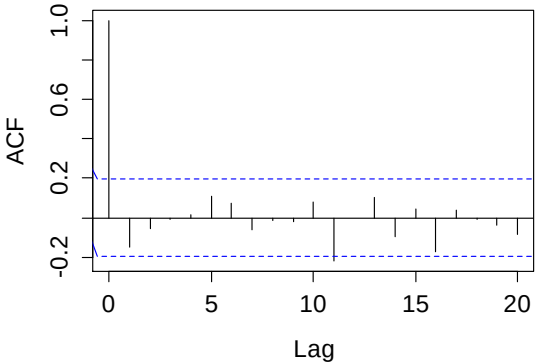
- Relationship between response and predictor(s) is **linear**



- Errors are **normally distributed** and have **constant variance**



- Errors are **independent** of one another



Regression model basics

- **Generalized linear models** (GLIMs) expand on ordinary linear models in two ways:
  1. Response variable *Y* assumed to have a distribution from the *exponential family*
     - Normal (ordinary linear regression, ANOVA, etc.):
     - Binomial (logistic regression), Poisson
     - Others (gamma, negative binomial, multinomial, inverse Gaussian, etc.)

  2. The expected value of the response variable (μ_i) is related to a linear equation of predictors through a **link function** (g)

$$g(\mu_i) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

Regression model basics

- In Binomial (logistic) regression, a **logit link** is typically used

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

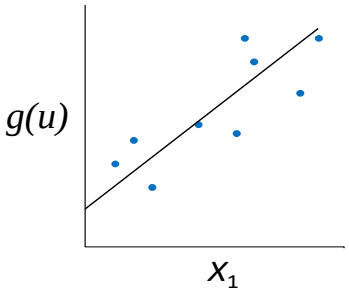
- In Poisson regression, a **log link*** is typically used

$$\log(\mu_i) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

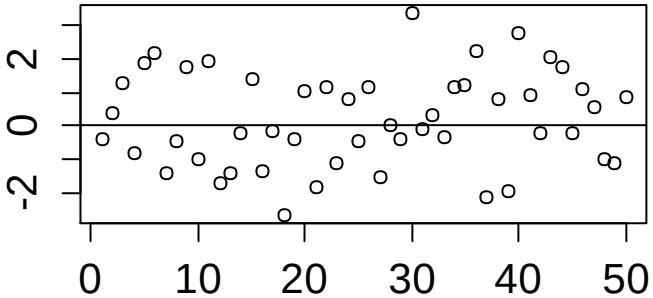
* natural (base *e*) logarithm

GLIM assumptions

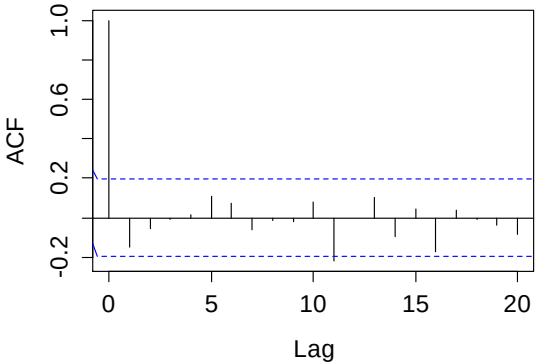
- Relationship between *transformed response* $g(\mu_i)$ and predictor(s) is **linear**



- *Residuals* should be evenly distributed and have **constant variance** (but not assumed to be normally distributed)



- Residuals are **independent** of one another



Descriptive modeling

- In this session we focus on *descriptive modeling*
- Having decided on the aim of our model (to describe), we turn our attention to
  - What **type of model** is appropriate for describing my data (ordinary linear, generalized linear, etc.)?
  - What is the **best fit model** for my data?

Descriptive modeling

- *Descriptive modeling* aims to
  - summarise or represent data in a compact manner
  - capture associations between dependent and independent variables
  - generate hypotheses (but not test hypotheses)

- Different from
  - *explanatory modeling*: hypothesis testing - based on underlying causal theory
  - *predictive modeling*: model as a tool for predicting new observations

Our data

- As an example, we consider individual-level clinic data from STI sentinel surveillance (provided by Clinical Prevention Services, BCCDC)
- Chlamydia and gonorrhea diagnoses (2006-17) were linked to infectious syphilis diagnoses (up to 12-months after)
- Patient-level information is based on case report forms and linkage to HIV surveillance data
- Our interest is to describe the associations between syphilis diagnosis and the patient characteristics

What type of model?

- The outcome of interest (syphilis dx: yes or no) is binary – *what type of model would be an appropriate choice?*

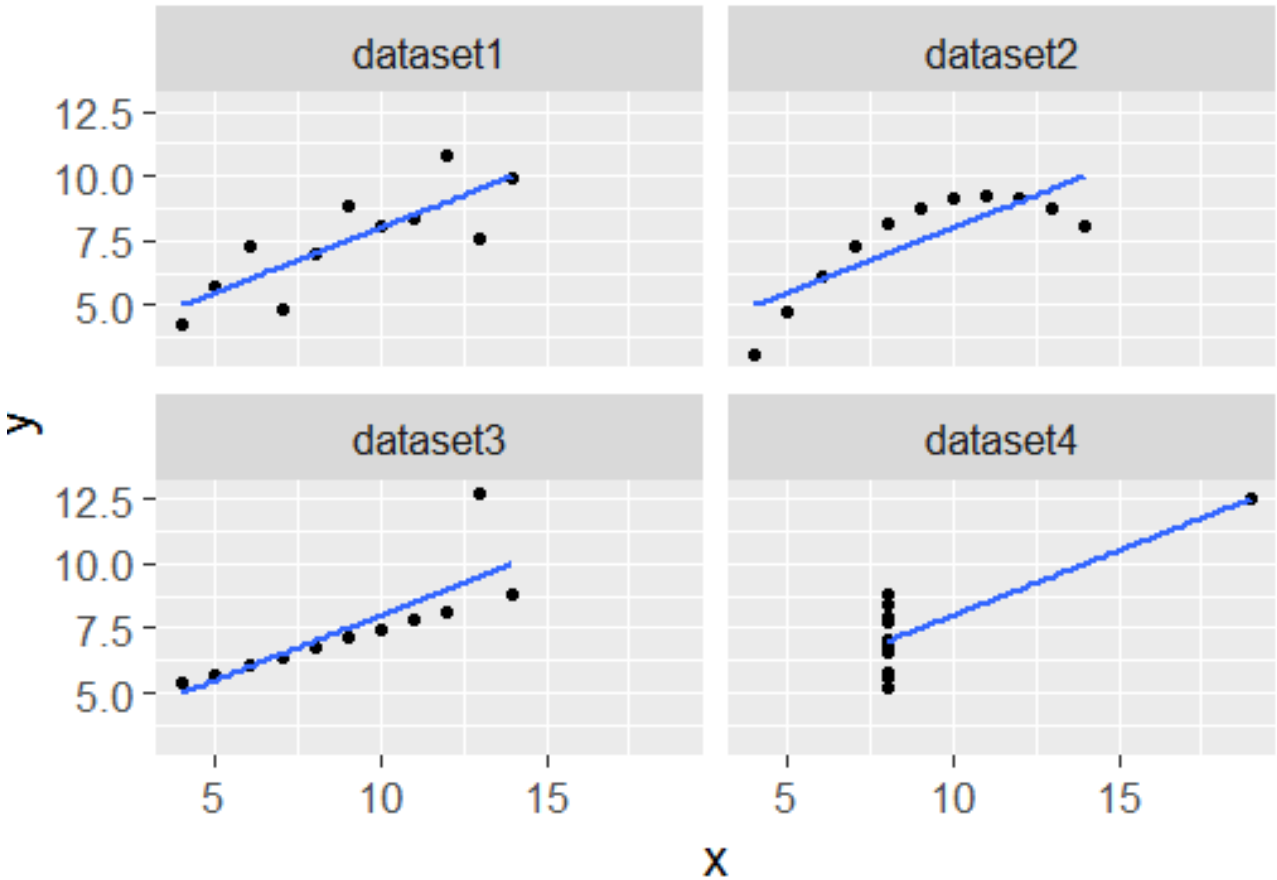
What type of model?

- The outcome of interest (syphilis dx: yes or no) is binary – *what type of model would be an appropriate choice?*

- We use binomial (logistic) regression as an illustrative example
  - In the accompanying R script, we analyse the probability of syphilis diagnosis in association with patient-level covariates (demographics, previous STI diagnoses)

Visualising your data and model

- Recall Anscombe's quartet: 4 different data sets having identical regression fits!



Visualising the data

- Use standard descriptive statistics (frequencies, mean ± SD, etc.) to summarise data
- Plots are generally most helpful for seeing and communicating patterns
- Use univariate tests (chi-square, t-test, etc.) for seeing simple patterns in data

Visualising data: descriptive summaries

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

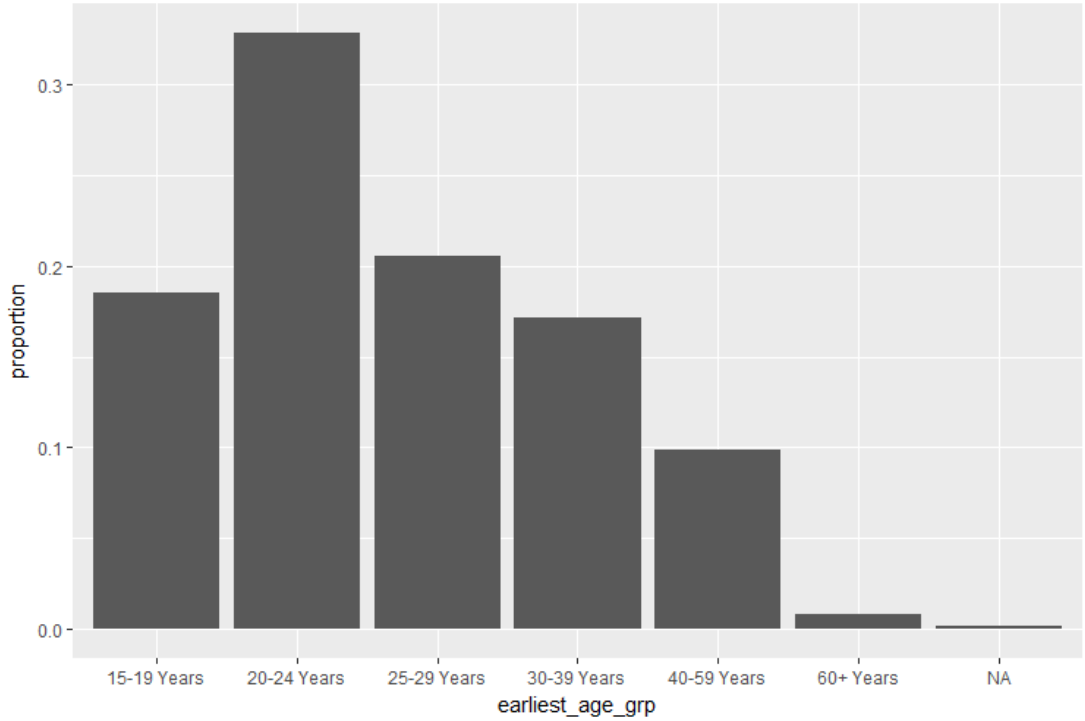
Go to file/function Addins

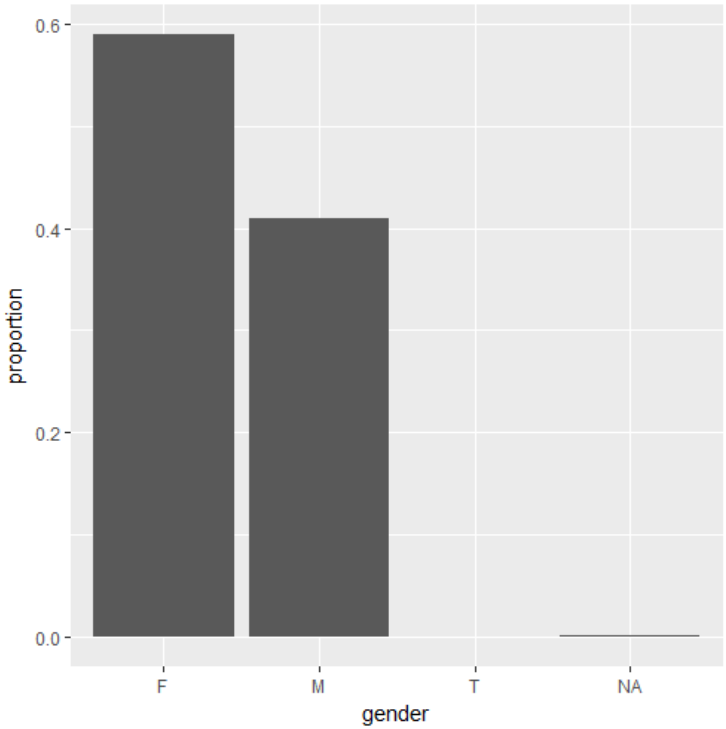
Source

Console //phsabc/root/BCCDC/Groups/Analytics_Resources/Training/Biostats/Sessions/Jul 19 2019 - model building 1/

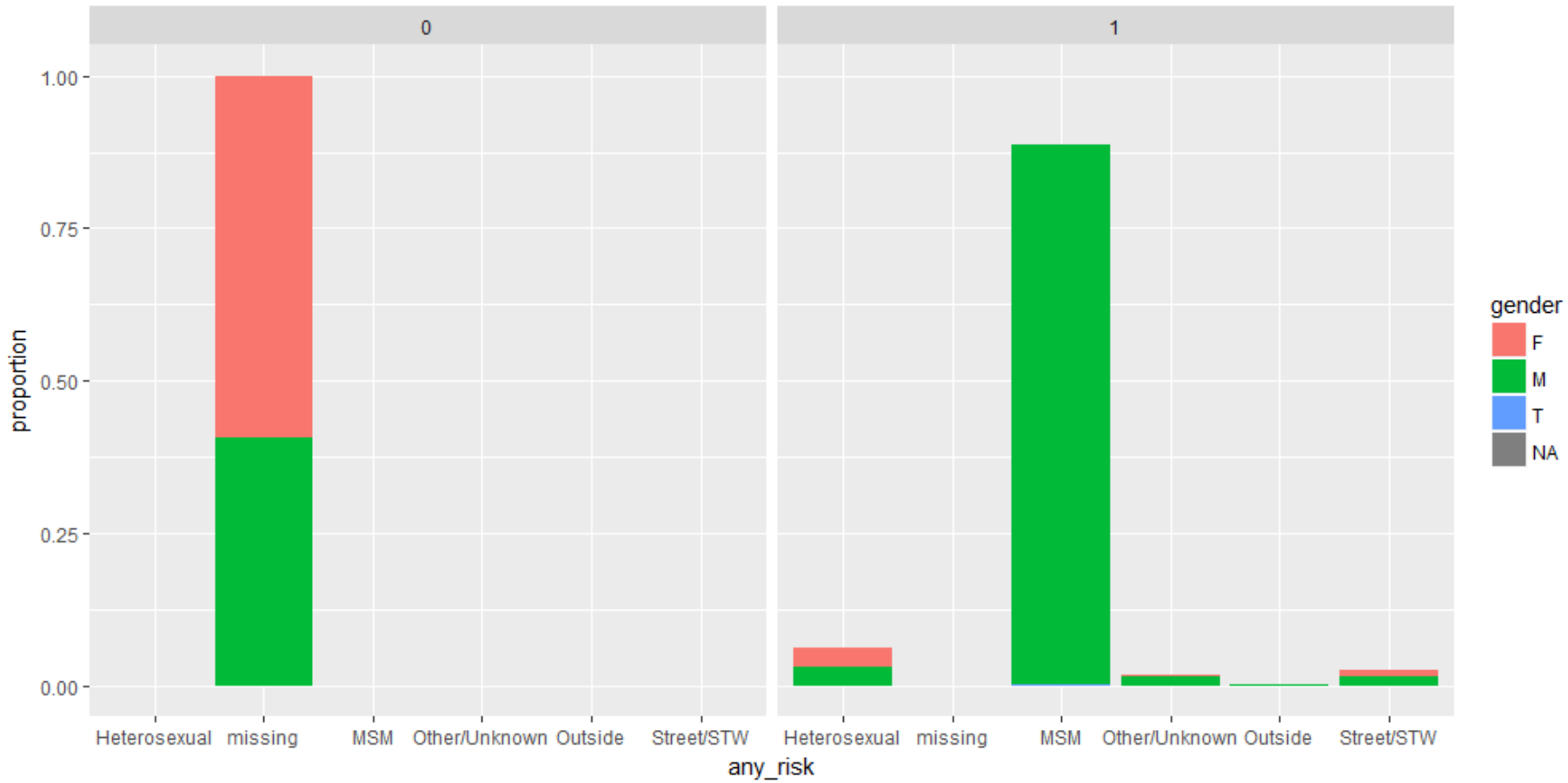
```
> # numbers of cases and non-cases
> clean_data %>% group_by(syph_dx) %>% summarise(unique.patients = n_distinct(patient_master_key))
# A tibble: 2 x 2
  syph_dx unique.patients
    <int>           <int>
1       0          132446
2       1             819
>
>
> # univariate frequency tables (case counts) by key covariates
> clean_data %>% count(syph_dx) %>% mutate(proportion = prop.table(n))
# A tibble: 2 x 3
  syph_dx      n proportion
    <int>  <int>      <dbl>
1       0 132446 0.99385435
2       1    819 0.00614565
>
>
> clean_data %>% count(gender) %>% mutate(proportion = prop.table(n))
# A tibble: 4 x 3
  gender     n   proportion
   <chr> <int>        <dbl>
1      F 78538 0.5893370352
2      M 54564 0.4094398379
3      T    57 0.0004277192
4   <NA>   106 0.0007954076
>
>
> clean_data %>% count(earliest_age_grp) %>% mutate(proportion = prop.table(n))
# A tibble: 7 x 3
  earliest_age_grp     n  proportion
             <chr> <int>       <dbl>
1      15-19 Years 24742 0.185660151
2      20-24 Years 43785 0.328555885
3      25-29 Years 27365 0.205342738
4      30-39 Years 22833 0.171335309
5      40-59 Years 13197 0.099028252
6        60+ Years  1126 0.008449330
7             <NA>   217 0.001628335
>
```

Visualising data: simple plots





Visualising data: simple plots



Visualising data: summary statistics

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Source

Console //phsabc/root/BCCDC/Groups/Analytics_Resources/Training/Biostats/Sessions/Jul 19 2019 - model building 1/

```
> # run chi-square tests of independence on selected covariates
> var_list <- c("hiv_atoc", "everlgv", "surveillance_region_ha", "earliest_age_grp", "gender", "ctgc_cat", "post2011")
>
> # previous chlamydia/gonorrhea dx
> clean_data %>%
+   select(var_list) %>%
+   summarise_all(funs(chisq.test(., clean_data$ctgc_cat, simulate.p.value = TRUE)$p.value))
# A tibble: 1 x 7
      hiv_atoc      everlgv surveillance_region_ha earliest_age_grp       gender     ctgc_cat     post2011
         <dbl>        <dbl>                  <dbl>            <dbl>        <dbl>        <dbl>        <dbl>
1 0.0004997501 0.0004997501           0.0004997501     0.0004997501 0.0004997501 0.0004997501 0.0004997501
>
```

Building a model

- There are various approaches to model building
  - begin with ‘full model’, containing all relevant covariates, then possibly remove covariates to achieve a better fit
  - begin with simple model (e.g., only one covariate) and build by iteratively adding covariates and assessing model fit
- Regardless of approach, it is essential to assess the fit of candidate models against the observed data
  - generate predicted (‘fitted’) values from model and compare against data
  - generate residuals (‘errors’) from model and assess fit

Building a model

- As an example, consider the age-only model:

P(syphilis dx) = age group

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

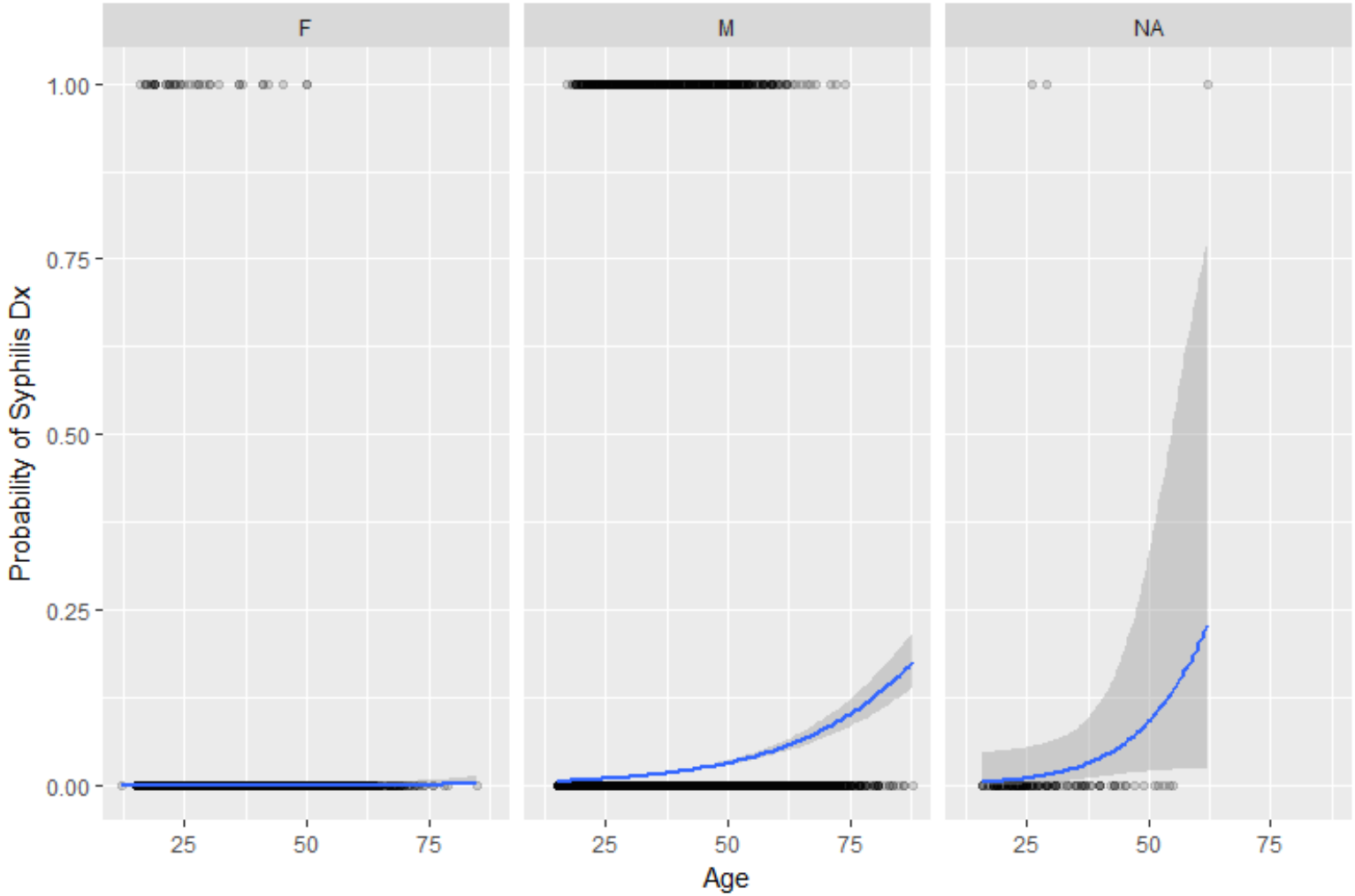
Go to file/function Addins

Source

Console //phsabc/root/BCCDC/Groups/Analytics_Resources/Training/Biostats/Sessions/Jul 19 2019 - model building 1/

```
> # plot probability of syph_dx as a function of patient age
> clean_data %>%
+   ggplot(aes(x = earliest_age_yrs, y = syph_dx)) +
+   geom_point(alpha = 0.15) +
+   geom_smooth(method = "glm", method.args = list(family = "binomial")) +
+   facet_wrap(~ gender_bin) +
+   xlab("Age") + ylab("Probability of Syphilis Dx")
Warning messages:
1: Removed 217 rows containing non-finite values (stat_smooth).
2: Removed 217 rows containing missing values (geom_point).
>
```

Building a model



Building a model

- Continued next time with **assessing model fit** and running **model comparisons**!