Regression Model Building 2: Model assessment

Michael Otterstatter

BCCDC Biostats Session

Aug 2, 2019

Session overview

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

From last session…

- 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?

Descriptive modeling

- Here we focus on *descriptive modeling*, which 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

Our data

Variables available for modeling building:

- **syph_dx** - Patient had a syphilis diagnosis during the study period (yes/no)
- **earliest_age_grp** - patient age groups (15-19, 20-24, 25-29, 30-39, 40-59, 60+ years)
- **hiv_atoc** - Patient had HIV at the time of syphilis diagnosis (yes/no)
- **everlgv** – diagnosis with lgv anytime (lifetime or within study period)
- **gender_bin** – Patient sex categories (M, F, NA)
- **surveillance_region_ha** - Patient's Health Authority of residence
- **ctgc_cat** - Number of chlamydia or gonorrhea diagnoses patient had during study period (1-2, 3-4, 5+)
- **post2011** - Chlamydia/gonorrhea diagnosis was after 2011 (yes/no)

What type of model?

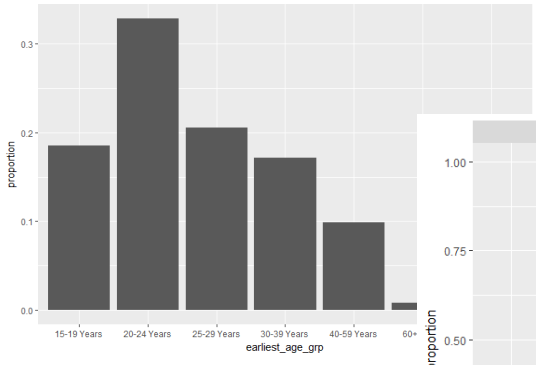
- The outcome of interest (syphilis dx: yes or no) is binary – *what type of model would be an appropriate choice?*
  - We use binomial (logistic) regression as an illustrative example
  - In the accompanying R script, we analyse the probability of syphilis diagnosis in association with patient-level covariates (demographics, previous STI diagnoses)

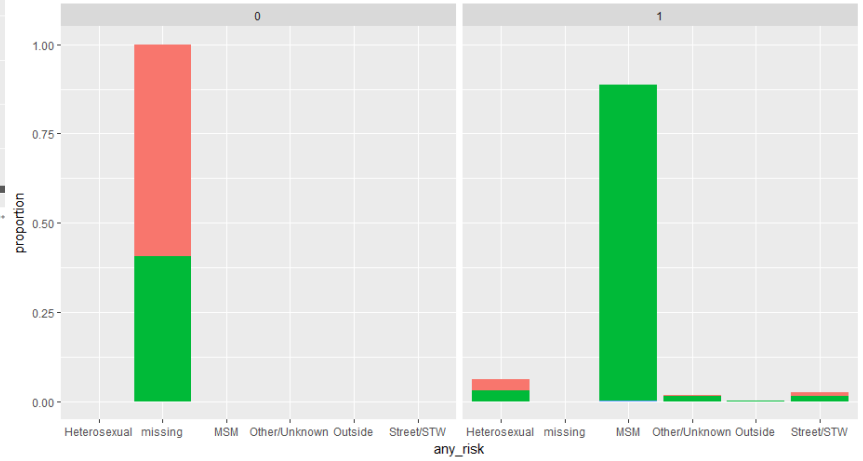
Building a model

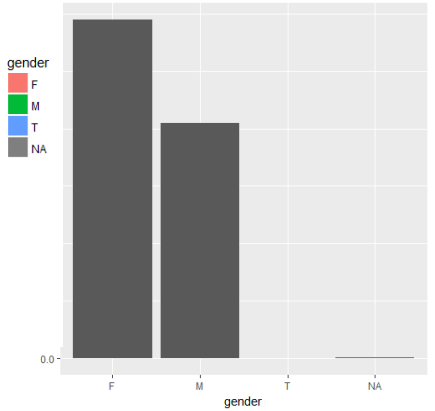
- Although there are many approaches to model building, one always needs to
  - **visualise the data**: summary statistics, plots, etc.
  - **choose a candidate model** (simple model, full model, etc.) as a starting point, assess fit, add or remove covariates
  - **compare the fit of candidate models** against one another and against the observed data
    - generate predicted ('fitted') values and/or residuals ('errors') values from model and assess fit
    - Examine goodness-of-fit statistics (AIC, BIC, dispersion)

Visualise the data

- We looked at that last session…







Choose a candidate model

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

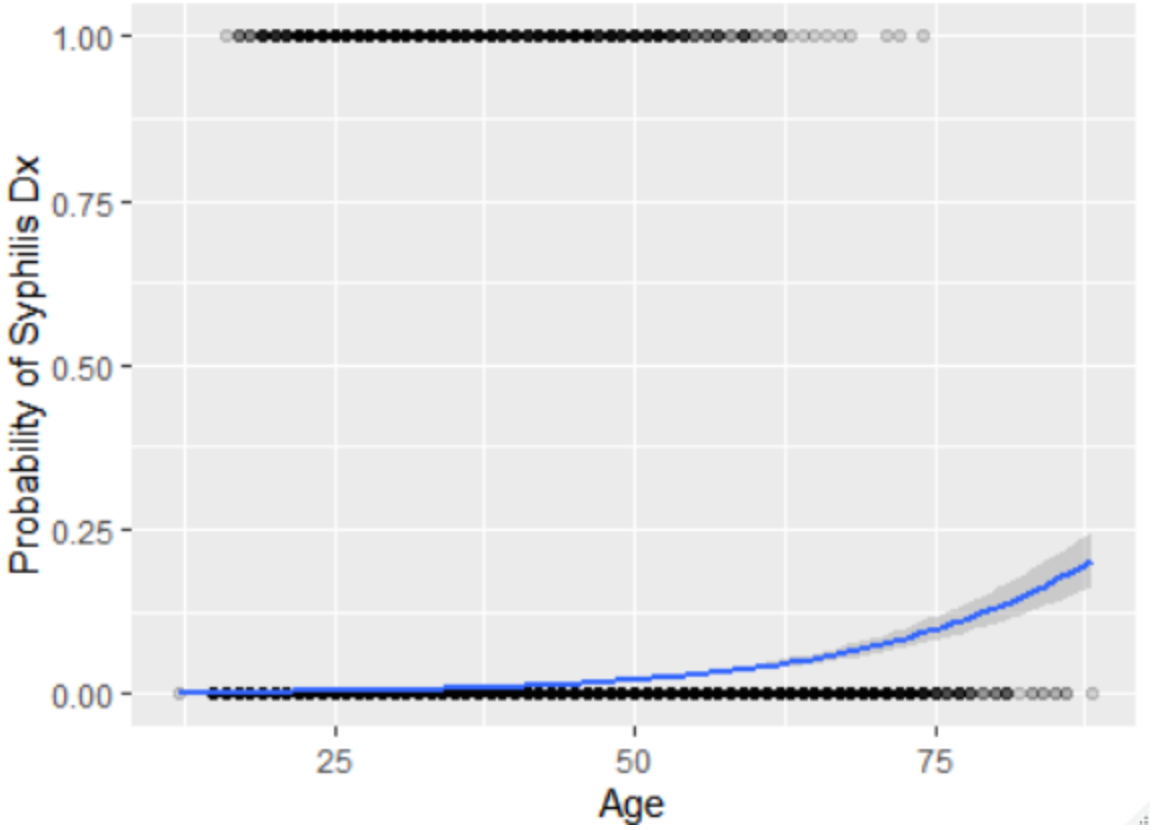
  Prob(syphilis dx) = patient age

```
age_only_model <- glm(syph_dx ~ earliest_age_yrs, family =
"binomial", data = analysis_data)
```

- Now what??
  - Often models are run without much attention to whether or not they are a good fit
  - Always consider visualisation – what does my model fit look like? Is it a good fit to the underlying data?

Choose a candidate model

- **Visualise the model fit**: plot probability of syph_dx as a function of patient age



Choose a candidate model

- **Assess model fit**: summarise model fit, residuals and deviance (measure of goodness of fit in generalized linear models)

```
summary(age_only_model) #

anova(age_only_model, test="Chisq")
```

Choose a candidate model

- **Assess model fit**: age is a significant predictor, with increased probability of syphilis dx with increasing age

```
# Coefficients:
#                     Estimate Std. Error z value       Pr(>|z|)
#   (Intercept)      -7.090126   0.095191  -74.48       << 0.001***
#   earliest_age_yrs  0.064993   0.002465   26.37       << 0.001***



#                   Df Deviance Resid. Df Resid. Dev    Pr(>Chi)
# NULL                             132901     9938.7
# earliest_age_yrs  1   553.54    132900     9385.1     << 0.001***
```

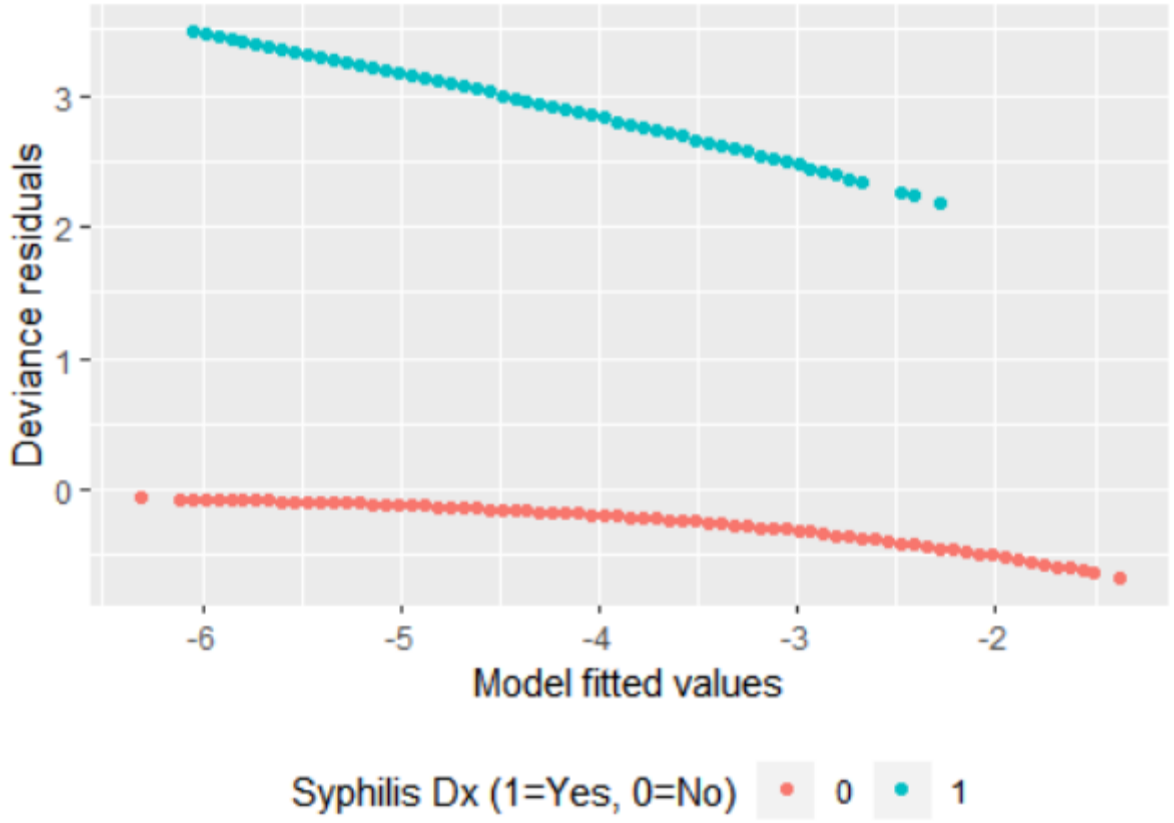
Choose a candidate model

- **Assess model fit**: but the distribution of residuals is quite skewed, suggesting a poor model fit

```
# Deviance Residuals:
#     Min       1Q   Median       3Q      Max
# -0.6727  -0.1081  -0.0890  -0.0781   3.4793
#
# Residual deviance: 9385.1  on 132900  degrees of freedom
 # AIC: 9389.1
```

Choose a candidate model

- **Assess model fit**: plotting the distribution of residuals can help assess model fit



Choose a candidate model

- **Assess model fit**: formal goodness-of-fit statistics can also help assess model fit
  - For example, Nagelkerke's **pseudo R^2**

$$R^2 = \frac{1 - \left\{ \frac{L(M_{Intercept})}{L(M_{Full})} \right\}^{2/N}}{1 - L(M_{Intercept})^{2/N}}$$

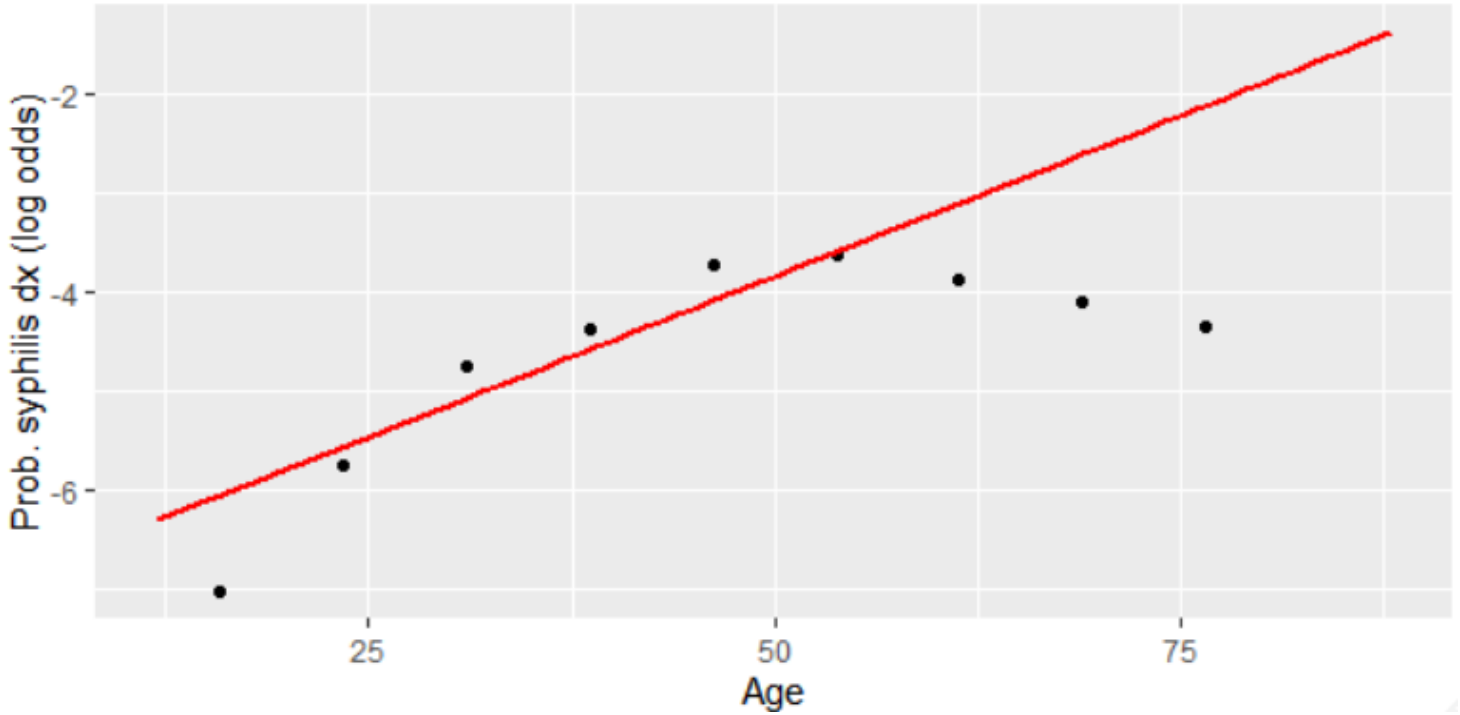
```
(1-exp( ((age_only_model$deviance)-(age_only_model$null.deviance)) / (age_only_model$df.null+1) )) /
  (1-exp( -(age_only_model$null.deviance) / (age_only_model$df.null+1) ))
```

```
[1] 0.05768332
```

- Proportion of deviance explained by model, similar to traditional R^2 in simple linear models
  - In this case, only 6% of deviance explained by age-only model

Choose a candidate model

- **Assess model assumptions**: check linearity



- Calculate observed proportions (as log odds) and compare with linear model fit – does the actual relationship look linear?

Choose a candidate model

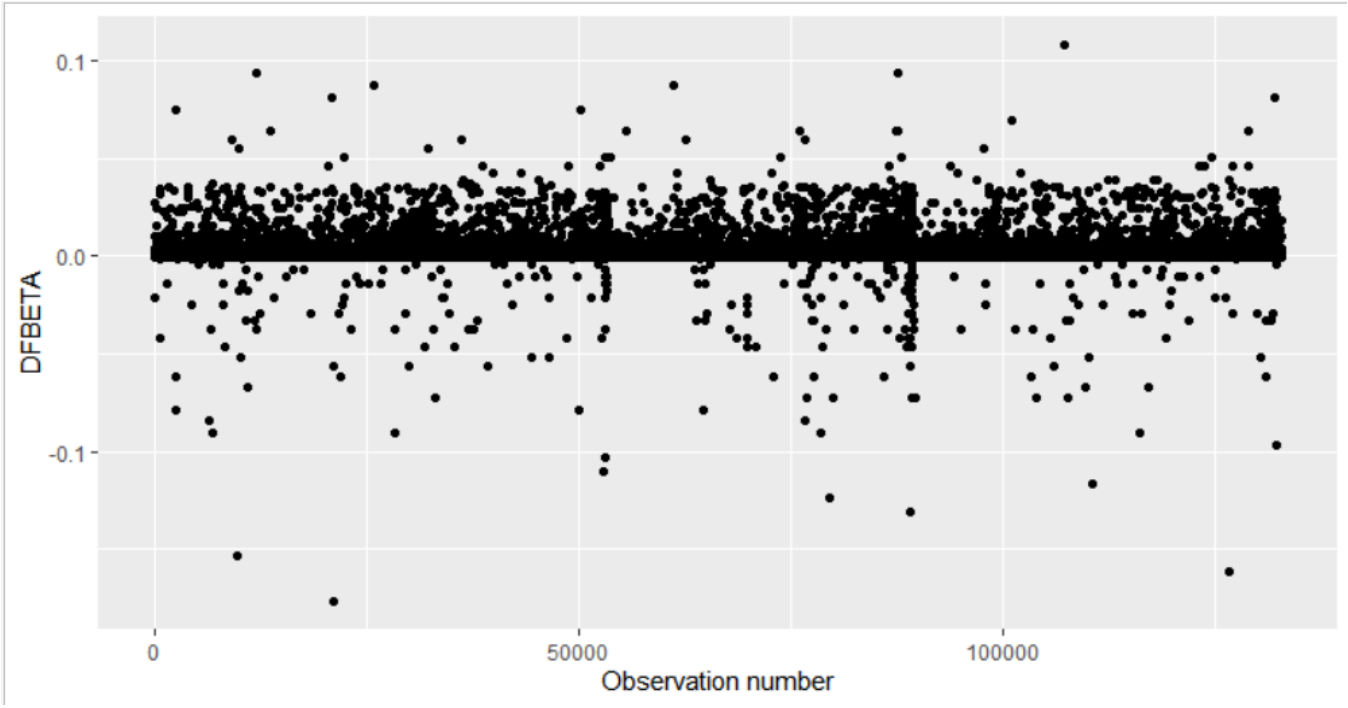
- **Assess model assumptions**: check outliers
  - In looking for unusual observations, it is helpful to calculate 'influence measures', which indicate impact of each data point on the overall model fit

```
inf.measures <- influence.measures(age_only_model)

as.data.frame(inf.measures$infmat) %>%
  rowid_to_column() %>%
  ggplot() + geom_point(aes(x = rowid, y = dfb.1_))
```

Choose a candidate model

- **Assess model assumptions**: check outliers
  - In looking for unusual observations, it is helpful to calculate 'influence measures', which indicate impact of each data point on the overall model fit



Next time

- Comparing the fit of differing models…