Regression Model Building 3: Model comparison

Michael Otterstatter

BCCDC Biostats Session

Aug 16, 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

- And now focus our attention on
  - methods for comparing and selecting models

From last time…

- 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?

From last time…

- 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

Selected 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 Lymphogranuloma venereum 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)

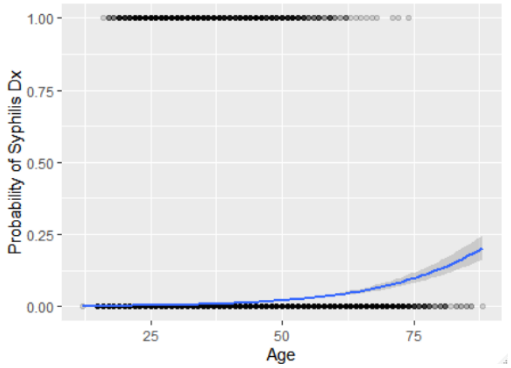
Building a model

- Although there are many approaches to model building, one always needs to
  - **First, visualise the data**: summary statistics, plots, etc.
  - **Then, choose a candidate model** (simple model, full model, etc.) as a starting point, assess fit, add or remove covariates
  - **Then, compare the fit of candidate models** against one another, by
    - generating predicted (‘fitted’) values or residuals (‘errors’) from the model and assessing relative fit
    - Examine ‘goodness-of-fit’ statistics (deviance, AIC, proportion variance explained, dispersion)

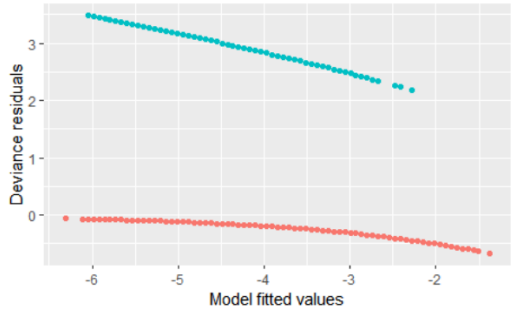
Model building: our starting model

- We started with an age-only model

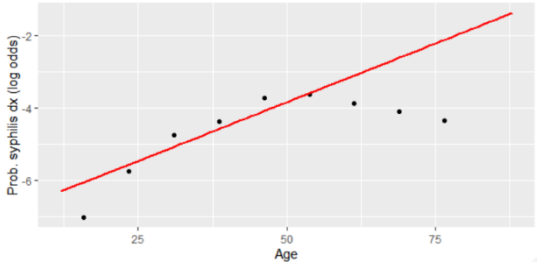
  Prob(syphilis dx) = patient age



- Then summarised model fit using residuals and deviance measures

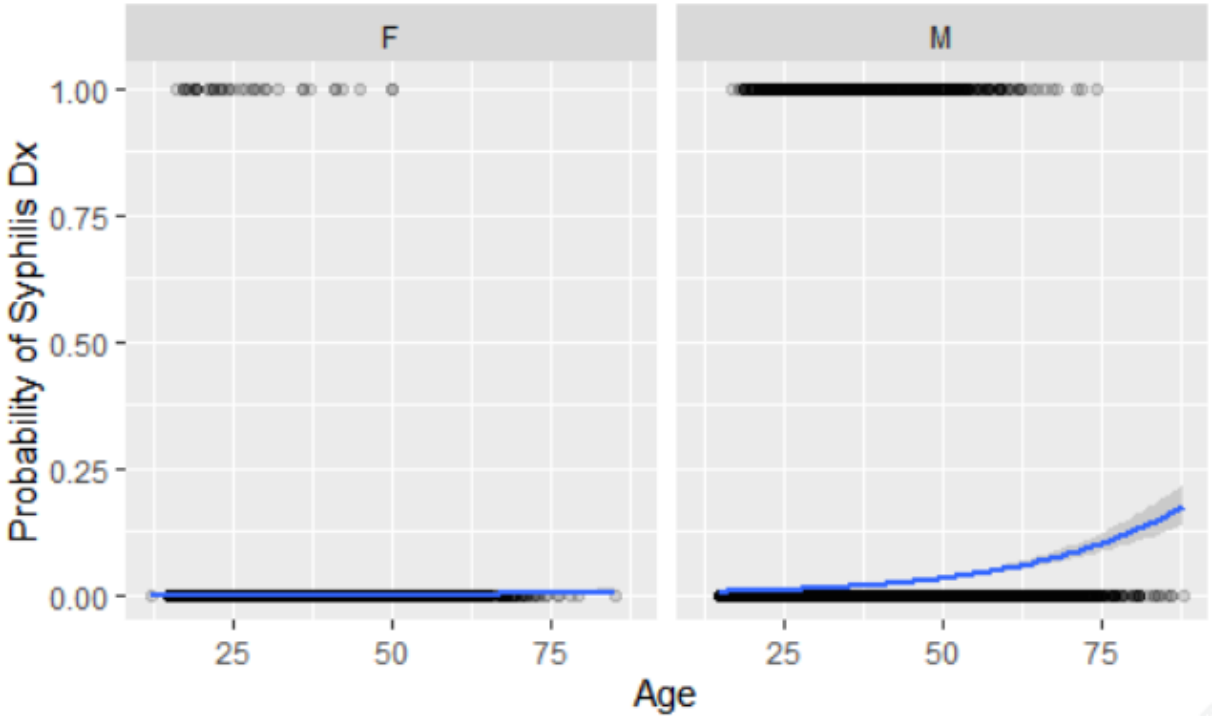


- But found this model only explained ≈6% of variation and assumption of linearity was questionable



Model building: adding covariates

- age by itself does not describe the probability of syphilis – let's add covariates into our model
  - add sex: *Prob(syphilis dx) = patient age + sex + age*sex*



Compare fit of new model

- Assess model fit: age and sex both appear as significant predictors, with higher probability of syphilis dx among men (but no interaction between age and sex)
- Model with sex and age is a better fit than model with just age (significant reduction in deviance)

```
Coefficients:
                            Estimate Std. Error z value    Pr(>|z|)
(Intercept)                -8.22467    0.47722 -17.234    << 0.001
earliest_age_yrs            0.02373    0.01660   1.429       0.153
gender_binM                 2.40398    0.48791   4.927    << 0.001
earliest_age_yrs:gender_binM 0.02491    0.01681   1.482       0.138

                            Deviance Resid.Df   Resid.Dev  Pr(>Chi)
NULL                                   132901     9938.7
earliest_age_yrs             553.54    132900      9385.1     << 0.001
gender_bin                   859.73    132899      8525.4     << 0.001
earliest_age_yrs:gender_bin   2.45     132898      8523.0        0.1175
```

Compare fit of new model

- Goodness-of-fit statistics also help compare models
  - Akaike information criterion (AIC)

```
> age_only_model$aic
[1] 9389.143

> age_sex_model$aic
[1] 8531.409

> age_sex_int_model$aic
[1] 8530.959
```

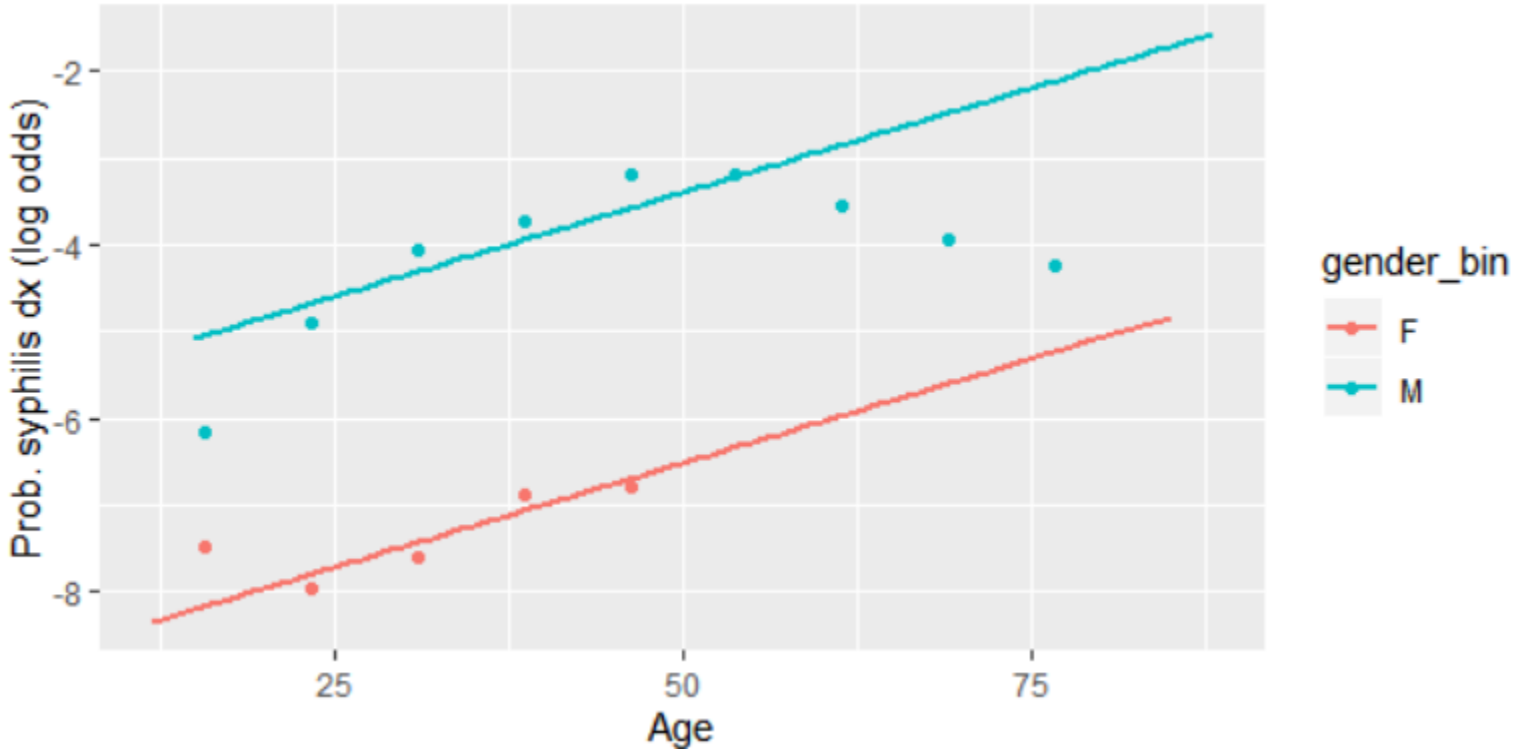
lower AIC, better model

similar AIC, similar models

  - Pseudo R^2 (Nagelkerke's): Proportion of deviance explained by age + sex model is ≈**14%,** compared to only ≈**6%** for age-only model

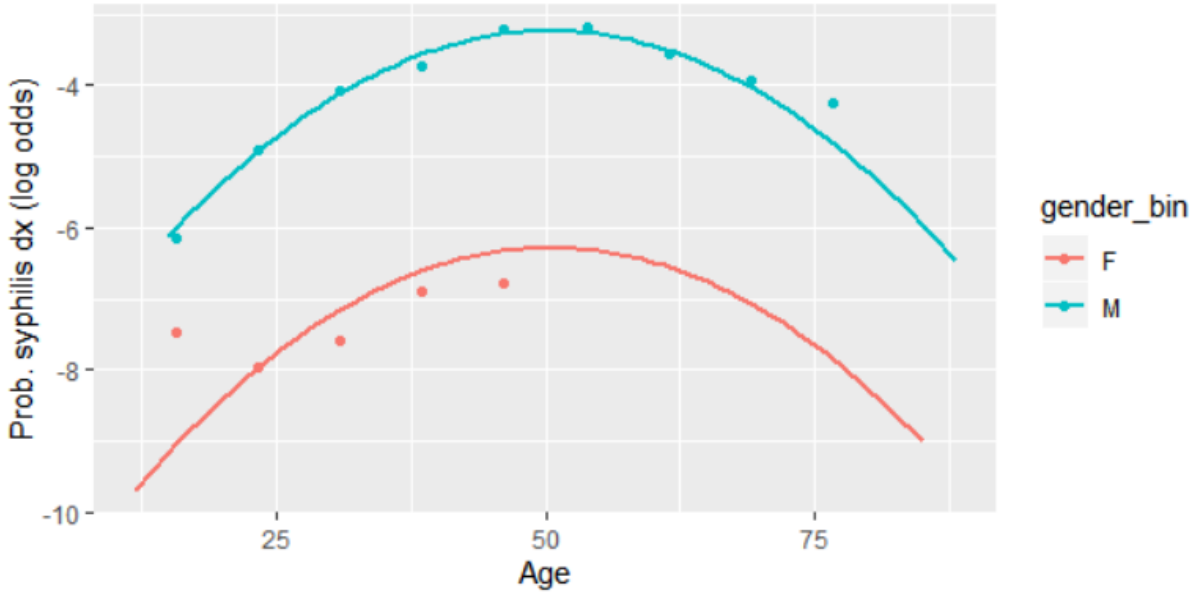
Compare fit of new model

- Assess model assumptions: linearity still questionable



Compare fit of new model

- By adding a quadratic term to our model, the fit with age is improved
  - Significant age^2 term (P << 0.001)
  - Overall reduction in model deviance (P << 0.001)
  - AIC is lower compared to age-sex model (8419 vs 8531)



Compare fit of new model

- Other aspects of model assumptions/fit
  - In logistic and Poisson regression, we assume variance is proportional to the mean (deviance/df ≈ 1.0)

```
# summary(age_sq_sex_model)
# ...
# Residual deviance: 8411.3 on 132898 degrees of freedom
```

  - When data appear over- (dev/df >> 1) or under- (dev/df << 1) dispersed, SE values are likely too narrow and we should consider a ‘quasi’ model in which the dispersion factor is estimated from the data

Compare fit of new model

- Fitting a 'quasibinomial' model only improves the estimates of the standard error, not the fit to the data per se

```
age_sq_sex_model <- glm(syph_dx ~ earliest_age_yrs + age_squared +
gender_bin, family = "binomial", data = analysis_data)

                   Estimate  Std. Error z value          Pr(>|z|)
(Intercept)      -12.1261140   0.3846859 -31.522        << 0.001 ***
earliest_age_yrs   0.2314716   0.0193865  11.940        << 0.001 ***
age_squared       -0.0022917   0.0002459  -9.322        << 0.001 ***
gender_binM        3.0427839   0.1648438  18.459        << 0.001 ***


age_sq_sex_model_quasi <- glm(syph_dx ~ earliest_age_yrs + age_squared +
gender_bin, family = "quasibinomial", data = analysis_data)

                   Estimate  Std. Error t value          Pr(>|t|)
(Intercept)      -12.1261140   0.4393457 -27.600        << 0.001 ***
earliest_age_yrs   0.2314716   0.0221412  10.454        << 0.001 ***
age_squared       -0.0022917   0.0002808  -8.162        << 0.001 ***
gender_binM        3.0427839   0.1882664  16.162        << 0.001 ***
```

Compare fit of several new models

- Through this iterative process, we can build and compare a series of models – for each, we
  - assess fit to the data (generating model summaries and plots of predicted values and residuals)
  - compare to simpler models (using deviance reduction, % of variation explained, and goodness-of-fit statistics)
  - examine model assumptions and adjust as appropriate (e.g., quadratic terms with relations are non-linear)

Considering the full model

- Starting with, or working our way up to, a full model will all (roughly 30) relevant covariates and interactions
  - we find many significant covariates (main effects and interaction terms
    - Full model explains ≈ 49% of variation in syphilis diagnosis
  - but also many terms (at least 12) that add little improvement to model fit

Considering the full model

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			120594	7908.8	
earliest_age_yrs	**1**	**534.10**	**120593**	**7374.7**	**<< 0.001**
hiv_atoc	**1**	**913.86**	**120592**	**6460.9**	**<< 0.001**
everlgv	**1**	**133.19**	**120591**	**6327.7**	**<< 0.001**
gender_bin	**1**	**551.08**	**120590**	**5776.6**	**<< 0.001**
surveillance_region_ha	**4**	**225.97**	**120586**	**5550.6**	**<< 0.001**
TotalCTGC	**1**	**844.96**	**120585**	**4705.7**	**<< 0.001**
post2011	**1**	**225.86**	**120584**	**4479.8**	**<< 0.001**
age_squared	**1**	**23.16**	**120583**	**4456.6**	**<< 0.001**
CTGC_squared	**1**	**178.86**	**120582**	**4277.8**	**<< 0.001**
earliest_age:hiv_atoc	**1**	**8.19**	**120581**	**4269.6**	**0.0042111**
earliest_age:everlgv	1	0.92	120580	4268.7	0.3384752
earliest_age:gender_bin	**1**	**6.67**	**120579**	**4262.0**	**0.0097825**
earliest_age:surv_region	4	4.62	120575	4257.4	0.3288628
earliest_age:TotalCTGC	**1**	**7.49**	**120574**	**4249.9**	**0.0062037**
earliest_age:post2011	1	3.51	120573	4246.4	0.0609164
hiv_atoc:everlgv	**1**	**18.08**	**120572**	**4228.3**	**<< 0.001**
hiv_atoc:gender	1	0.02	120571	4228.3	0.8783368
hiv_atoc:surveillance_region	**4**	**11.76**	**120567**	**4216.5**	**0.0192030**
hiv_atoc:CTGC	**1**	**7.20**	**120566**	**4209.3**	**0.0072782**
hiv_atoc:post2011	1	0.00	120565	4209.3	0.9602960
everlgv:gender	1	0.08	120564	4209.2	0.7745739
everlgv:surveillance_region	**3**	**17.37**	**120561**	**4191.9**	**0.0005944**
everlgv:CTGC	1	0.96	120560	4190.9	0.3273099
everlgv:post2011	1	0.40	120559	4190.5	0.5258297
gender_bin:surv_region	4	3.60	120555	4186.9	0.4631156
gender_bin:CTGC	1	1.19	120554	4185.7	0.2754139
gender_bin:post2011	**1**	**43.38**	**120553**	**4142.3**	**<< 0.001**
surveillance_region_ha:CTGC	4	5.93	120549	4136.4	0.2044090
surveillance_region:post2011	4	3.13	120545	4133.3	0.5361186
TotalCTGC:post2011	**1**	**27.72**	**120544**	**4105.6**	**<< 0.001**

Considering a reduced model

- We could manually prune the full model to include only those terms adding significant improvement to model fit
  - In this way, the reduced model has half as many covariates (17) as full model but still explains the same amount of variation in syphilis dx (~49%)

Considering a reduced model

	Df	Deviance	Resid. Df	Resid. Dev	P(>Chi)
NULL			120594	7908.8	
earliest_age_yrs	1	534.10	120593	7374.7	<< 0.001
hiv_atoc	1	913.86	120592	6460.9	<< 0.001
everlgv	1	133.19	120591	6327.7	<< 0.001
gender_bin	1	551.08	120590	5776.6	<< 0.001
surveillance_region_ha	4	225.97	120586	5550.6	<< 0.001
TotalCTGC	1	844.96	120585	4705.7	<< 0.001
post2011	1	225.86	120584	4479.8	<< 0.001
age_squared	1	23.16	120583	4456	<< 0.001
CTGC_squared	1	178.86	120582	4277.8	<< 0.001
earliest_age_yrs:hiv_atoc	1	8.19	120581	4269	0.0042111
earliest_age_yrs:gender	1	6.50	120580	4263.1	0.0107726
hiv_atoc:everlgv	1	18.73	120579	4244.4	<< 0.001
hiv_atoc:surv_region	4	12.04	120575	4232.3	0.0170443
hiv_atoc:TotalCTGC	1	10.11	120574	4222.2	0.0014774
everlgv:surveillance_region	3	17.38	120571	4204.8	<< 0.001
gender_bin:post2011	1	31.78	120570	4173.1	<< 0.001
TotalCTGC:post2011	1	33.46	120569	4139.6	<< 0.001

Next time…

- We can talk more about variable selection and automated vs manual approaches to model building