Model selection

& some other Miscelaneous

Mabel Carabali

EBOH, McGill University

updated: 2025-12-02

What do you think?

Table 3 Univariate and Multivariable Regression Analyses of the Association Between Heavy Alcohol Use and Admission Radiologic and Clinical Characteristics and Outcomes in Patients With ICH

	Regression analysis	Crude β-coefficients/OR (95% CI)	*p* Value	Adjusted β-coefficients/OR[a] (95% CI)	*p* Value
CT imaging characteristics					
ICH volume[b]	Linear regression	1.61[b] (1.19–2.18)	0.002	1.68 (1.17–2.41)	0.005
ICH location (nonlobar)	Logistic regression	1.84 (1.22–2.77)	0.003	2.01 (1.11–3.64)	0.021
Intraventricular extension	Logistic regression	2.07 (1.37–3.12)	<0.001	1.94 (1.02–3.71)	0.045
Clinical characteristics on admission and outcomes					
Platelets	Linear regression	−16.57 (−31.44 to −1.70)	0.029	−17.73 (−32.75 to −2.72)	0.021
MAP on admission	Linear regression	7.33 (2.53–12.13)	0.003	4.81 (0.06–9.56)	0.047
Length of hospital stay[b]	Linear regression	1.30 (1.09–1.54)	0.003	1.22 (1.02–1.45)	0.033
Loss of dependency at discharge	Logistic regression	2.37 (1.41–3.99)	0.001	1.92 (1.12–3.31)	0.018
In-hospital mortality	Logistic regression	0.65 (0.39–1.07)	0.092	0.71 (0.42–1.22)	0.219

Effects of Heavy Alcohol Use on Acute Intracerebral Hemorrhage and Cerebral Small Vessel Disease Hindsholm, M., et al.,. (2025). Neurology, 105 (11), doi: 10.1212/WNL.0000000000214348.

Expected Competencies

- Understand the difference between predictive and etiological epidemiology principles.
- Knows statistical assumptions for different regression models.
- Knows basic model fit or "goodness-of-fit" statistics.

Objectives

- To revise the principles and differences between predictive and etiological research.
- To revise the use of main "goodness-of-fit" statistics.
- To revise the overall framework for model parameterization and specification for prediction models.
- To revise variable parameterization and selection.

What is model selection?

- There are two important considerations when building a regression model:
  - The kind of outcome data you want to model (continuous, count, binary, etc.)
  - **The variables and interactions you include in your model**

- **Model selection** is a process that attempts to find the best model for a given purpose.

- The two main purposes of regression models are:
  - **Causal Inference:** To estimate the effect of one or more variables while adjusting for the possible confounding effects of other variables.
  - **Prediction:** To predict outcomes for a set of similar individuals.

Prediction: Examples

Not only weather...

- Credit scores
- Netflix recommendations
- Cardiovascular risk scores *
- Kidney function estimates *
- Predicting mortality

* Some, including the misstep "race correction", different form "race-adjustment".

Prediction Examples

Framingham cardiovascular disease (10-year risk) calculator

FRS Calculator

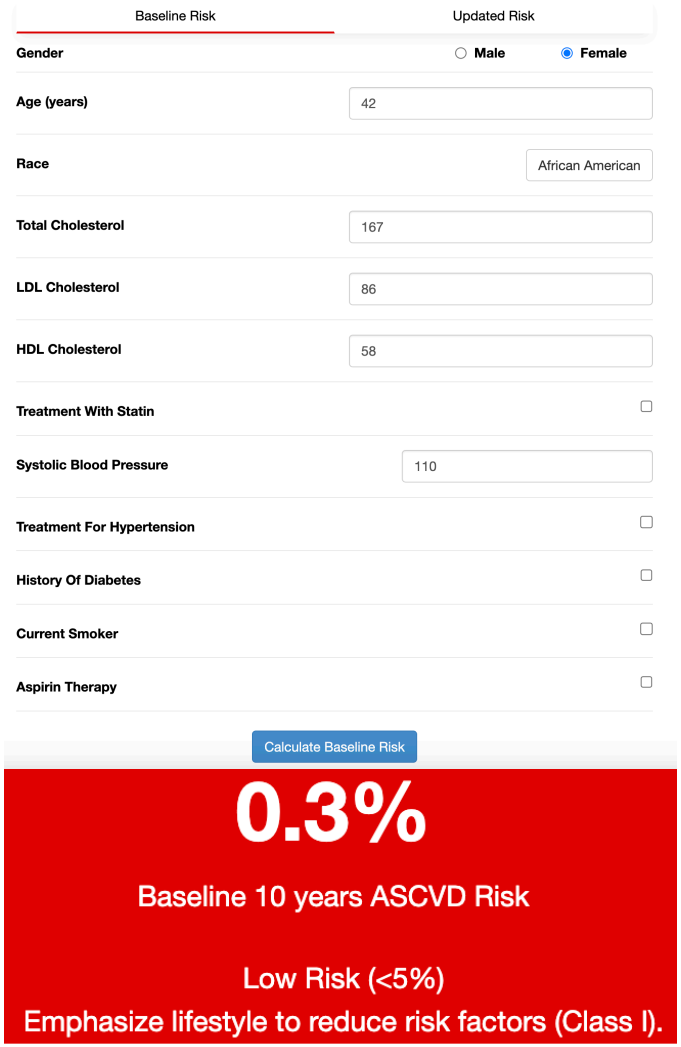
Age	42	years
Sex	Female	
HDL-C	1.52	mmol/L
Total-C	4.33	mmol/L
Systolic BP	110	mmHg
	No — BP Treated	
	No — Smoker	
	No — Diabetes	
	Yes — Fam Hx. of premature CVD	

Calculate

10-Year CVD Risk **3.0 % (Low-Risk)**

Heart Age **31**

Or AHA's CV Risk Calculator



Causal Inference vs. Prediction

Important: We can't interpret coefficients from prediction models as causal parameters

- Not appropriately controlled for confounding (because we aren't worried about this)
- This is often a pitfall that people fall into!

American Journal of Epidemiology
© The Author 2013. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
Vol. 177, No. 4
DOI: 10.1093/aje/kws412
Advance Access publication:
January 30, 2013

Commentary

The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients

Daniel Westreich* and Sander Greenland

* Correspondence to Dr. Daniel Westreich, Department of Obstetrics and Gynecology, Duke Global Health Institute, Duke University, DUMC 3967, Durham, NC 27710 (e-mail: daniel.westreich@duke.edu).

OK, but how do I build a model?

What is required to build a "Good" model?

Criteria:

- Is our research question causal or predictive?
  - This will determine the "philosophy" we use to build our model
- Sample size (Power and Precision)
- Choice of statistical model/distribution/link (Poisson, logistic, linear, etc.)
- Consideration of different sources of bias (for causal inference).
- What data do we have available to us?

Overall framework for model building / specification:

1) Variable specification (including assessment of number and parameterization)

2) Interaction assessment (including assessment of heterogeneity)*

3) Confounding assessment (including consideration of precision)*

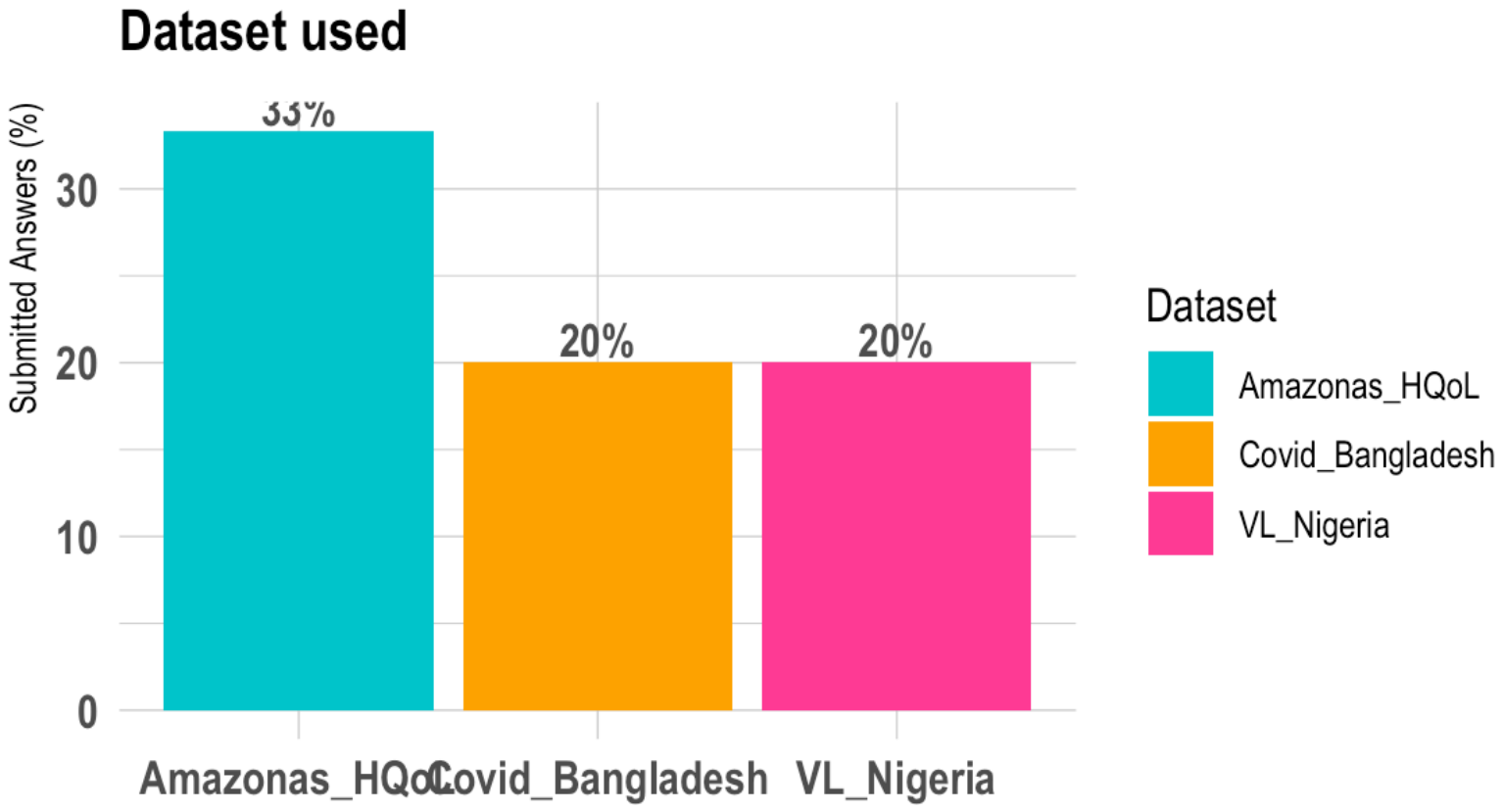
*Steps 1 and 2 can occur iteratively and depending on the research question investigated differently.

When do we say our model is "good"? - Variable specification

Especially when we have many predictor variables (with many possible interactions), it can be difficult to find a good model.

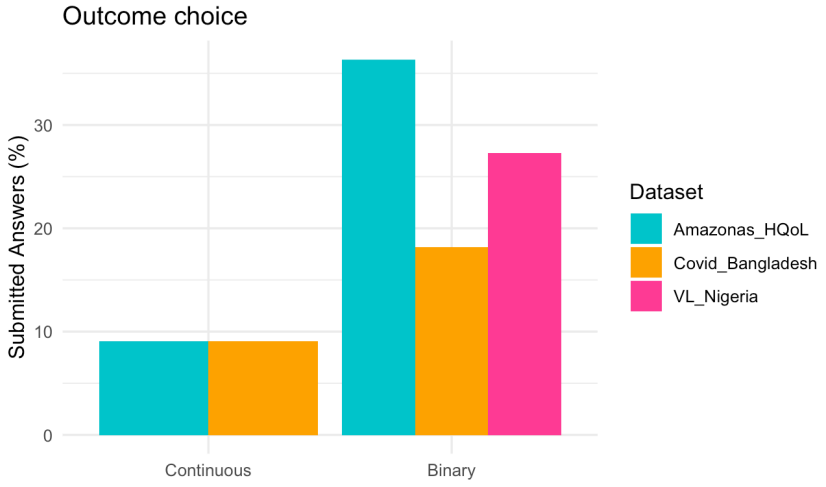
- Which main effects do we include?
- Which interactions do we include?
- With 6 variables, there are 64 potential models with just main effects!

What did we get from our "dream model"?

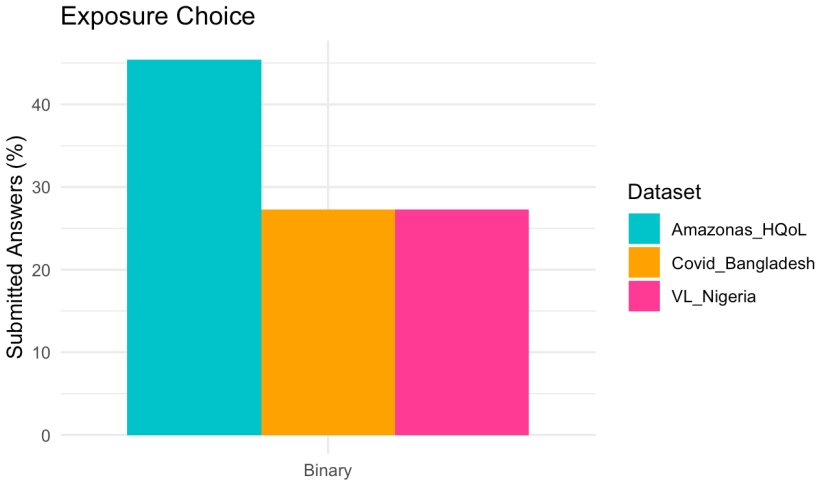


What did we get from our "dream model"?

Outcome



Exposures



When do we say our model is "good"? - Variable specification

- An active research problem in statistics
- Model selection procedures try to simplify this task.

Evaluating Model Fit

To implement a model selection procedure, we first need a criterion to compare models.

The goal is to select the model with the optimal value of the criterion.

To that end, we assess the: **Goodness-of-Fit Statistics and use some statistical model metrics**

Evaluating Model Fit: Recall

R-squared R^2

- R^2 represents the proportion of variance in the outcome explained by all the predictors in the model.
- **Criterion:** Choose the model with the largest R^2
- **Problem:** R^2 always increases with model size
- If only use R^2 means simply choosing the largest model: not very useful.
- Helpful when comparing models with the same number of parameters.

Adj. R-squared R^2

- Adjusted R^2 represents the proportion of variance in the outcome explained by all the predictors in the model, **penalized for the number of variables included in the model**.
- **Criterion:** Choose the model with the largest adjusted R^2.
- Here, the largest model is not necessarily the best model.

AIC = Akaike Information Criterion

$$AIC = -2ln(Likelihood) + 2p$$

- **The best model is the one with the smallest AIC**.
- The AIC is formed by two terms:
  - The likelihood: measure of fit
  - The penalty term: $2p$, accounts for adding more terms to the model.

The first term **always** decreases as more terms are added to the model, so $2p$ is needed for "balance".

- AIC can be used whenever we have a likelihood, so this generalizes to many statistical models.

BIC = Bayesian Information Criterion

$$BIC = -2ln(Likelihood) + pln(n)$$

- **The best model is the one with the smallest BIC**.
- AIC and BIC are very similar - only the last term changes
- BIC will always choose a model as small or smaller than the AIC (if using the same search strategy).

Selection Strategies - Prediction

- Now that we know how to evaluate model fit, we need to figure out how to find the model with the **best** fit.

- **Best subset:** Search all possible models and take the one with the highest R^2, or lowest MSE/AIC/BIC, etc.
    - Such searches are typically only feasible when you have less than 30 potential predictor variables.

- **Stepwise (forward, backward, or both) searches:** Useful when the number of potential predictor variables is large.

Illustration with the datasets used in 704 (1)

```
library(epib.704.data)
data("Amazonas_HQoL")
glimpse(Amazonas_HQoL)
```

```
## Rows: 1,179
## Columns: 47
## $ X        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,…
## $ ID       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,…
## $ morador  <int> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1,…
## $ FILTRO   <int> 3, 3, 2, 3, 3, 3, 2, 3, 2, 2, 3, 3, 3, 2, 3, 2, 3, 3, 3, 2, 3,…
## $ P72      <int> 1, 1, 1, 2, 1, 1, 1, 1, 3, 3, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1,…
## $ P73      <int> 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1,…
## $ P74      <int> 1, 1, 1, 2, 1, 1, 3, 1, 2, 3, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1,…
## $ P75      <int> 2, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 3, 2, 2, 1, 2, 2, 2,…
## $ P76      <int> 1, 1, 1, 2, 2, 1, 2, 1, 1, 3, 2, 2, 1, 1, 2, 1, 2, 1, 1, 1, 2,…
## $ P77      <int> 90, 80, 80, 50, 40, 80, 50, 70, 50, 10, 80, 60, 80, 80, 70, 80…
## $ P1       <int> 9, 6, 7, 3, 4, 2, 5, 2, 3, 1, 1, 3, 4, 2, 2, 4, 7, 2, 1, 1, 1,…
## $ P11      <int> 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 2,…
## $ P12      <int> 6, 3, 3, 3, 3, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 1, 1, 3, 1, 3,…
## $ P13      <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1,…
## $ P14      <int> 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1,…
## $ P15      <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1,…
## $ P16      <int> 3, 5, 4, 5, 4, 5, 6, 5, 2, 4, 3, 2, 2, 1, 4, 4, 3, 6, 3, 3, 1,…
## $ P17      <int> 2, 1, 1, 2, 2, 3, 1, 1, 1, 1, 1, 2, 0, 0, 0, 1, 1, 1, 1, 1, 1,…
## $ P18      <int> 3, 4, 4, 4, 3, 3, 4, 3, 3, 3, 3, 3, NA, NA, NA, 3, 2, 4, 4, 1,…
## $ P19      <int> 2, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 1,…
```

Regression subset selection: Illustration with the "Amazonas_HQoL" data

```
library(leaps) #Regression subset selection package

regfit_full.1 <- regsubsets(binQoL ~ .,
                      data = Amazonas_HQoL,
                      method = "exhaustive", really.big=T)
```

```
## Reordering variables and trying again:
```

Need to specify when >50 vars

```
reg_summary.1 <- summary(regfit_full.1)
summary(regfit_full.1)
```

```
## Subset selection object
## Call: regsubsets.formula(binQoL ~ ., data = Amazonas_HQoL, method = "exhaustive",
##     really.big = T)
## 58 Variables  (and intercept)
##               Forced in Forced out
## X                 FALSE      FALSE
## ID                FALSE      FALSE
## morador           FALSE      FALSE
## FILTRO            FALSE      FALSE
## P72               FALSE      FALSE
## P73               FALSE      FALSE
## P74               FALSE      FALSE
## P75               FALSE      FALSE
## P76               FALSE      FALSE
## P77               FALSE      FALSE
```

Illustration with the "Amazonas_HQoL" data

```
par(mfrow = c(1,2))

plot(reg_summary.1$adjr2, xlab = "Number of Variables", ylab = "Adjusted RSq", type = "l")
adj_r2_max.1 = which.max(reg_summary.1$adjr2)
points(adj_r2_max.1, reg_summary.1$adjr2[adj_r2_max.1], col ="red", cex = 2, pch = 20)

plot(reg_summary.1$bic, xlab = "Number of Variables", ylab = "BIC", type = "l")
bic_min.1 = which.min(reg_summary.1$bic)
points(bic_min.1, reg_summary.1$bic[bic_min.1], col = "red", cex = 2, pch = 20)
```

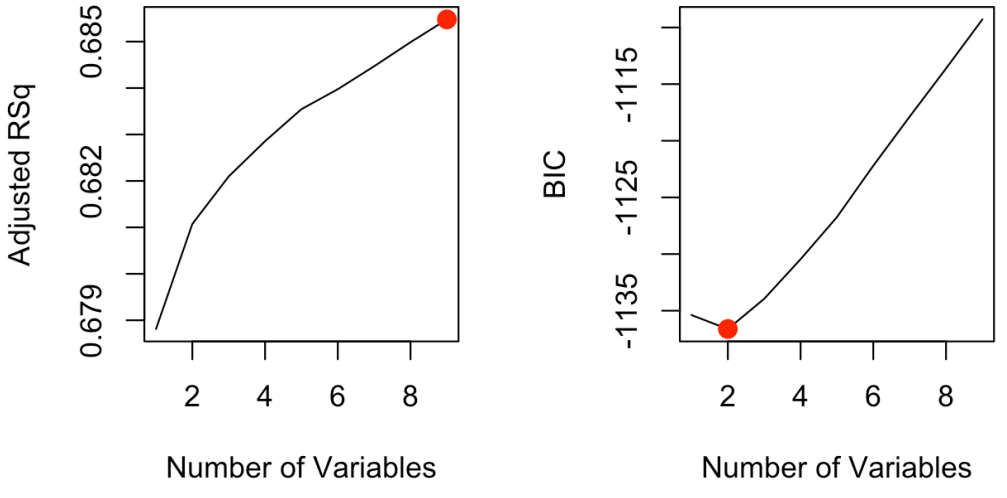


Illustration with the "Amazonas_HQoL" data

```
plot(regfit_full.1, scale = "adjr2",
     main= "'adjr2' for 'Amazonas_HQoL' data")
```

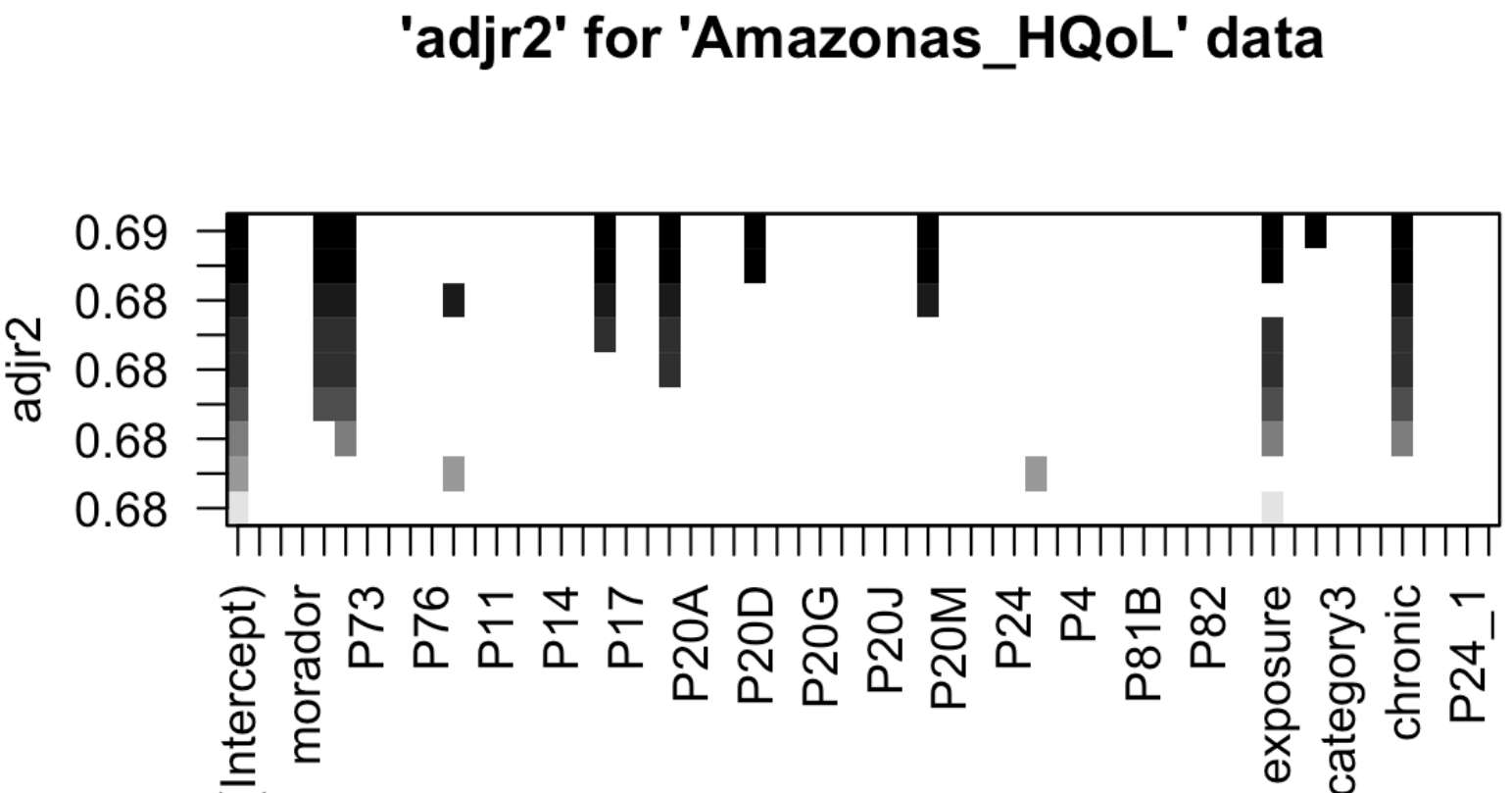
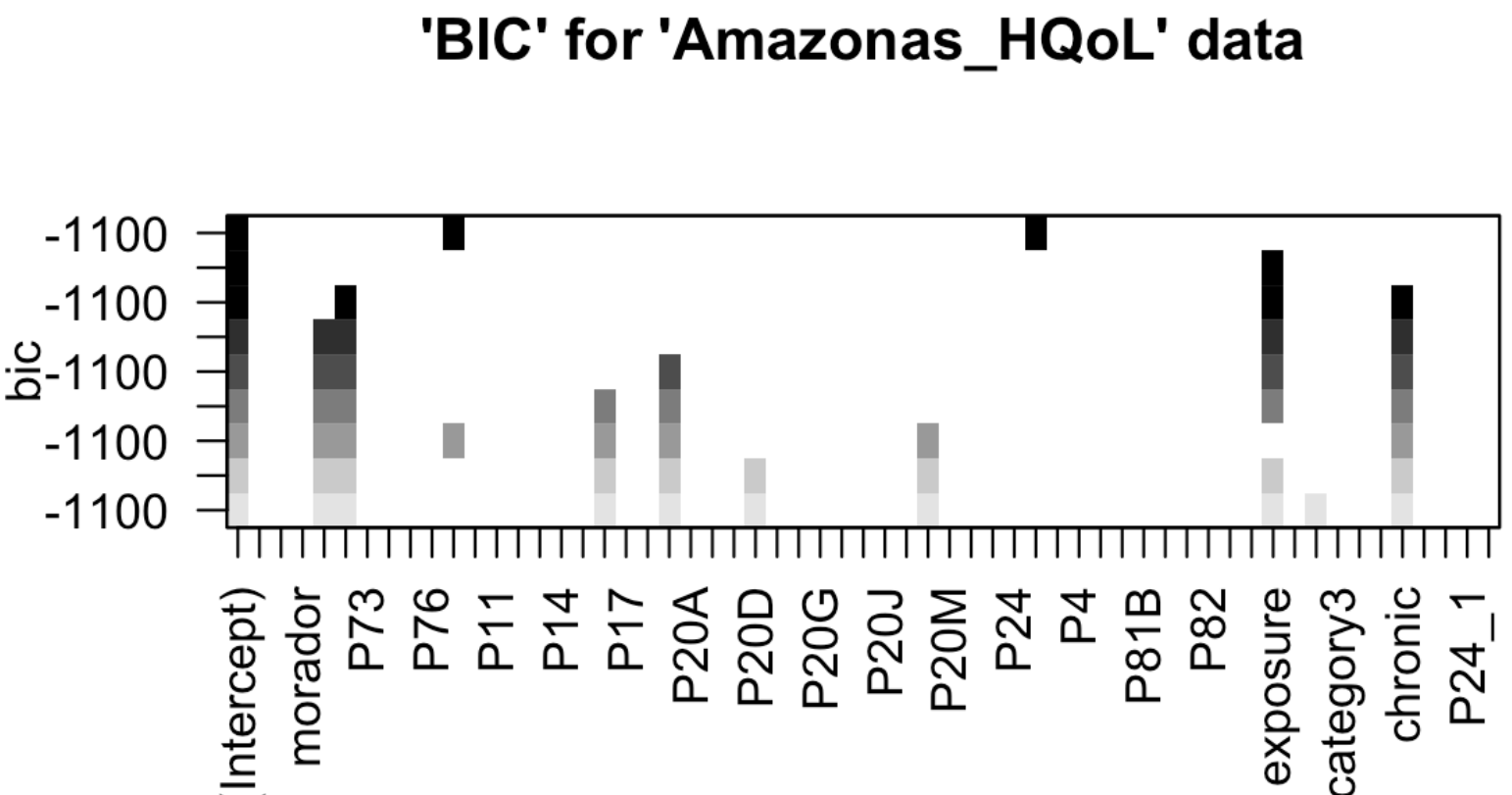


Illustration with the "Amazonas_HQoL" data

```
plot(regfit_full.1, scale = "bic",
     main= "'BIC' for 'Amazonas_HQoL' data")
```



What did we get from our "dream model"?

What dataset you used?	Write the equation of the model
Amazonas_HQoL	VAS ~ any_malaria + SES_Index_scaled + Age + Sex +ZONA
Amazonas_HQoL	"logit(P(Yi=1))=β0+β1Xi, when adding in the weights
Amazonas_HQoL	Prob(HRQoL_VeryHigh=1∣Malaria)]=β0+β1(Recent Malaria)
Amazonas_HQoL	"Model <- glm(binQoL ~ Status2, data = Data1, family = binomial(), weights = w.out1$weights)"
Amazonas_HQoL	"Model fitted: weighted logistic regression: log{(P(Y=0∣Matched Set)/P(Y=1∣Matched Set)} = B0 + Bexposure.Xexposure+ Bsex.Xsex +Bage.Xage + BSES.XSES + Bresidence.Xresidence

Illustration with the dataset used in 704 (2)

```
glimpse(Covid_Bangladesh)
```

```
## Rows: 2,502
## Columns: 27
## $ ex         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1…
## $ hhid       <int> 2, 2, 3, 3, 6, 6, 8, 8, 9, 9, 10, 10, 11, 12, 12, 13, 13, 14…
## $ memid      <int> 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 1, 2, 1, 2, 1, 2, 1, …
## $ gih_4      <chr> "Dhaka", "Dhaka", "Dhaka", "Dhaka", "Dhaka", "Dhaka", "Dhaka…
## $ village    <chr> "Duaripara", "Duaripara", "Duaripara", "Duaripara", "Duaripa…
## $ vdoses     <int> 2, 2, 2, 2, 1, 2, 1, 3, 3, 2, 2, 2, 2, 2, 2, 0, 2, 2, 2, 2, …
## $ dose2      <int> 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, …
## $ dose3      <int> 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ age_cat    <chr> "18-40 years old", "18-40 years old", "18-40 years old", "18…
## $ gender     <chr> "Male", "Female", "Female", "Male", "Female", "Male", "Male"…
## $ marital    <chr> "Married", "Married", "Married", "Separated/Divorced/Widow(e…
## $ educ       <chr> "No education", "Primary or less", "Primary or less", "Above…
## $ occup2     <chr> "Day labor", "Service", "Day labor", "Business or self-emplo…
## $ relation2  <chr> "HH head", "Spouse", "Spouse", "Others", "Spouse", "Others",…
## $ sesh_15_5  <int> 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, …
## $ sesh_15_7  <int> 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, …
## $ sesi_51    <chr> "No", "No", "Yes", "No", "No", "No", "No", "No", "No", "No",…
## $ chroill_1  <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No", …
## $ acill_26   <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No", "No",…
## $ mi_1       <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No", …
## $ gih_15     <int> 5, 5, 4, 4, 6, 6, 4, 4, 6, 6, 4, 4, 2, 3, 3, 11, 11, 4, 4, 2…
## $ income     <dbl> 10000, 10000, 20000, 20000, 30000, 30000, 16000, 16000, 4000…
## $ wbscore    <int> 12, 20, 76, 68, 80, 36, 52, 64, 32, 36, 40, 40, 20, 60, 80, …
```

Regression subset selection: Illustration with the "Covid_Bangladesh" data

```
library(leaps) #Regression subset selection package
Covid_Bangladesh1 <- Covid_Bangladesh %>% select(wbbin, exposure, age_cat, gender, marital, 

regfit_full.2 <- regsubsets(wbbin ~ .,
                      data = Covid_Bangladesh1,
                      method = "exhaustive")
```

```
reg_summary.2 <- summary(regfit_full.2)
summary(regfit_full.2)
```

```
## Subset selection object
## Call: regsubsets.formula(wbbin ~ ., data = Covid_Bangladesh1, method = "exhaustive")
## 20 Variables  (and intercept)
##                                                 Forced in Forced out
## exposure                                            FALSE      FALSE
## age_catAbove 40                                     FALSE      FALSE
## genderMale                                          FALSE      FALSE
## maritalSeparated/Divorced/Widow(er)/Unmarried       FALSE      FALSE
## educNo education                                    FALSE      FALSE
## educPrimary or less                                 FALSE      FALSE
## occup2Business or self-employed                     FALSE      FALSE
## occup2Day labor                                     FALSE      FALSE
## occup2Others                                        FALSE      FALSE
## occup2Service                                       FALSE      FALSE
## vdoses                                              FALSE      FALSE
## dose3                                               FALSE      FALSE
## chroill_1Yes                                        FALSE      FALSE
## income1                                             FALSE      FALSE
## wbscore                                             FALSE      FALSE
```

Illustration with the "Covid_Bangladesh" data

```
par(mfrow = c(1,2))

plot(reg_summary.2$adjr2, xlab = "Number of Variables", ylab = "Adjusted RSq", type = "l")
adj_r2_max.2 = which.max(reg_summary.2$adjr2)
points(adj_r2_max.2, reg_summary.2$adjr2[adj_r2_max.2], col ="red", cex = 2, pch = 20)

plot(reg_summary.2$bic, xlab = "Number of Variables", ylab = "BIC", type = "l")
bic_min.2 = which.min(reg_summary.2$bic)
points(bic_min.2, reg_summary.2$bic[bic_min.2], col = "red", cex = 2, pch = 20)
```

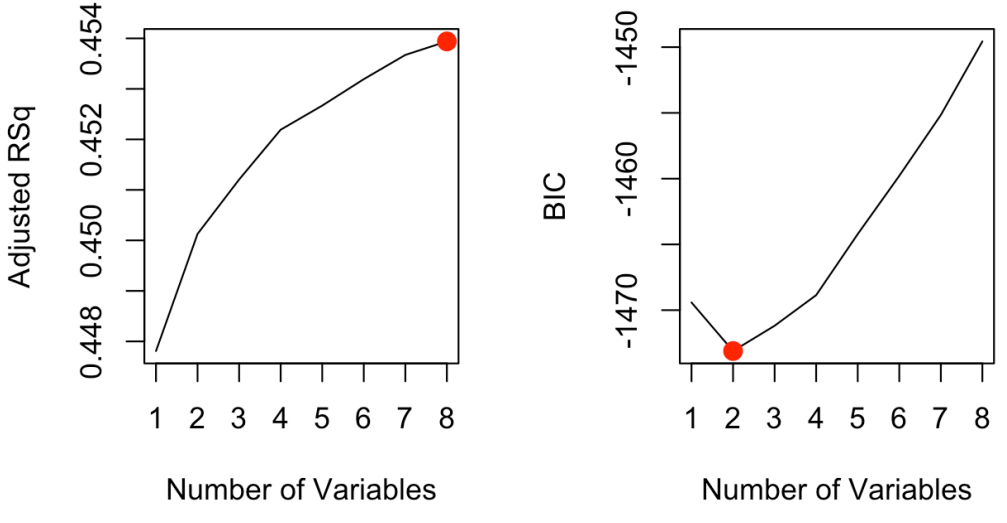


Illustration with the "Covid_Bangladesh" data

```
plot(regfit_full.2, scale = "adjr2",
     main= "'adjr2' for 'Covid_Bangladesh' data")
```

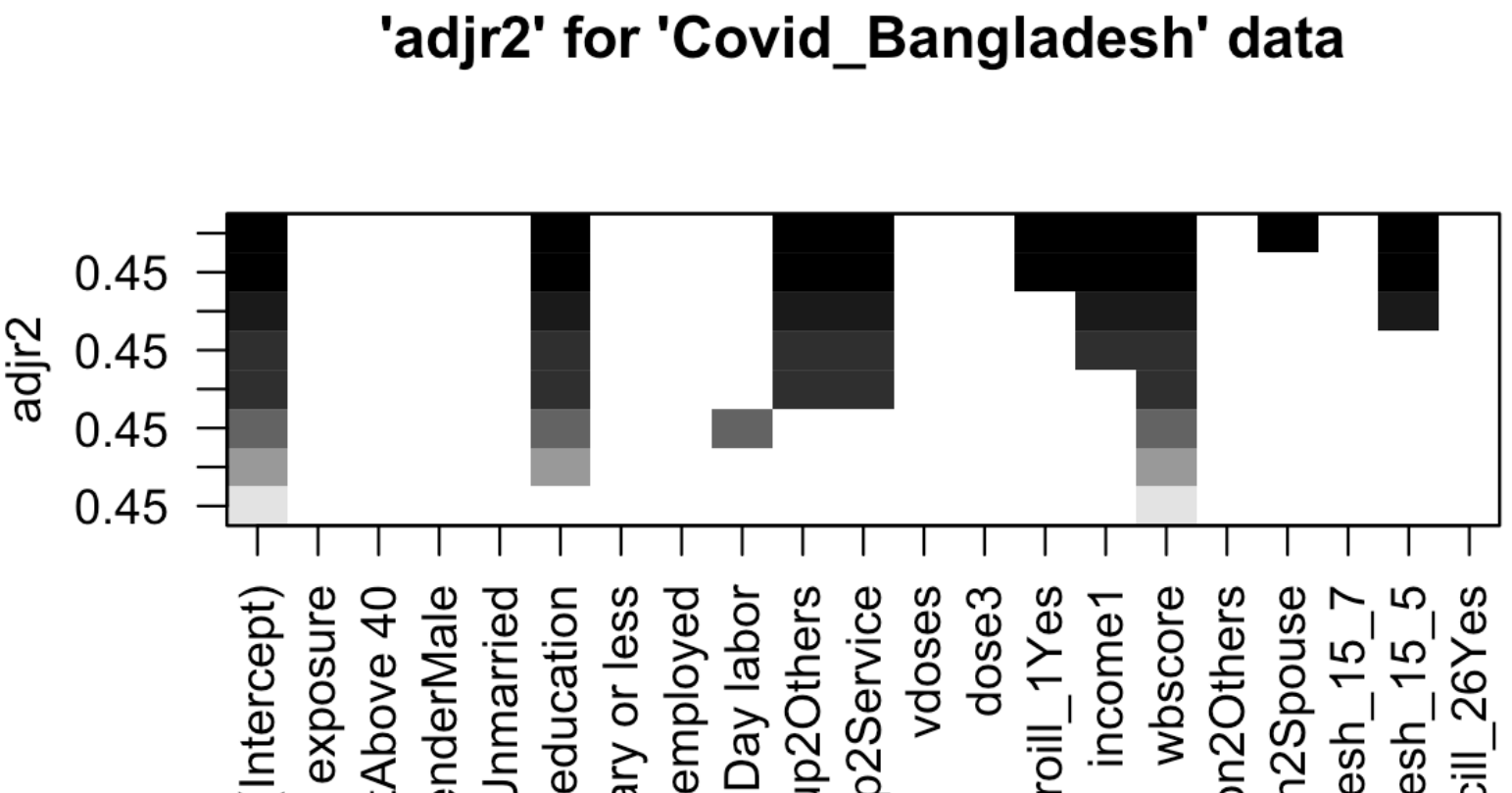
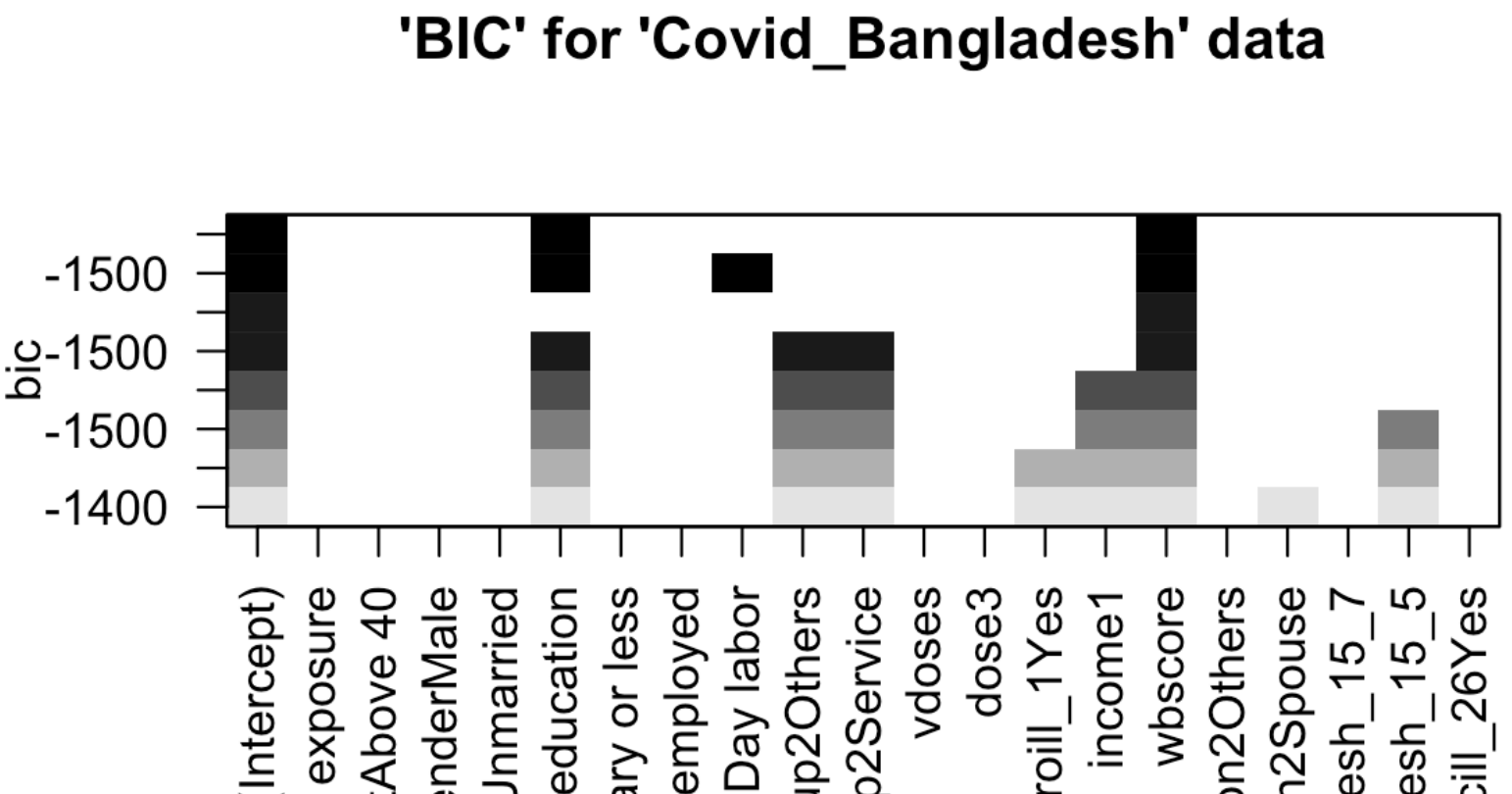


Illustration with the "Covid_Bangladesh" data

```
plot(regfit_full.2, scale = "bic",
     main= "'BIC' for 'Covid_Bangladesh' data")
```



What did we get from our "dream model"?

What dataset you used?	Write the equation of the model
Covid_Bangladesh	Mental score=B0+B1xCovid-19vaccine+B2xGender+B3xage + B4xarea + B5xEducation+ B6xChronic + B7xrelation_household+ B8xOccupation + B9xHousehold_size + B10xMigration + B11xAccess_TV + B12x Acces_smartphone.
Covid_Bangladesh	log(pi) = beta0 + betaXi
Covid_Bangladesh	"E (Yi|Xi) = B0 + B1 x Exp+ B2 x Sex + B3 x Age+ B4 x Chronic Illness + B5 x Occupation + B6 x Income"

Illustration with the dataset used in 704 (3)

```
glimpse(VL_Nigeria)
```

```
## Rows: 127,198
## Columns: 30
## $ S_No_                                  <int> 1, 2, 4, 5, 8, 9, 11, 12, 13, 15, 16,…
## $ State                                  <chr> "Adamawa", "Adamawa", "Adamawa", "Ada…
## $ SEX                                    <chr> "Female", "Male", "Male", "Male", "Fe…
## $ Age.group                              <chr> "20-29", "40+", "40+", "40+", "40+", …
## $ ART_Start                              <int> 2021, 2020, 2015, 2020, 2016, 2018, 2…
## $ Days_of_ARV_Refill                     <int> 180, 180, 180, 180, 180, 180, 180, 18…
## $ Current_Regimen_Line                   <chr> "Adult.1st.Line", "Adult.1st.Line", "…
## $ Current_ART_Regimen                    <chr> "TDF-3TC-DTG", "TDF-3TC-DTG", "TDF-3T…
## $ Functional_Status_at_last_Visit        <chr> "Working", "Working", "Working", "Wor…
## $ Clinical_Staging_at_Last_Visit         <chr> "Stage I", "Stage I", "Stage I", "Sta…
## $ VL.Result                              <dbl> NA, 38.1, 81.7, 419.0, 0.0, NA, 1617.…
## $ Date_of_current_viral_load__yyyy       <int> NA, 2021, 2020, 2022, 2019, NA, 2021,…
## $ Viral_Load_Indication                  <chr> "", "Routine Monitoring", "Routine Mo…
## $ Current_ART_Status                     <chr> "Active", "Active", "Active", "Active…
## $ ART_enrollment_setting                 <chr> "Community", "Facility", "Facility", …
## $ Enhanced_Adherence_Counselling         <chr> "No", "No", "No", "No", "No", "No", "…
## $ Date_of_commencement_of_EAC_yyyy       <chr> "", "", "", "", "", "", "2021", "", "…
## $ Number_of_EAC_Sessions_Completed       <int> NA, NA, NA, NA, NA, NA, 3, NA, 3, 3, …
## $ X                                      <int> NA, NA, NA, NA, NA, NA, 2021, NA, 202…
## $ Repeat_Viral_Load___Post_EAC_VL        <chr> "", "", "", "", "", "", "Yes", "", "Y…
## $ Date_of_Repeat_Viral_Load___Post       <int> NA, NA, NA, NA, NA, NA, 2022, NA, 202…
## $ Repeated.VL                            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, 26.1,…
## $ X.1                                    <int> NA, NA, NA, NA, NA, NA, NA, NA, 2022,…
```

Regression subset selection: Illustration with the "VL_Nigeria" data

```
library(leaps) #Regression subset selection package
VL_Nigeria1 <- VL_Nigeria %>% select(binSupress, exposure, Age.group, SEX, clinstage, funcsta
#Number_of_EAC_Sessions_Completed, Repeated.VL,

regfit_full.3 <- regsubsets(binSupress ~ .,
                      data = VL_Nigeria1,
                      method = "exhaustive")
```

```
## Reordering variables and trying again:
```

```
reg_summary.3 <- summary(regfit_full.3)
summary(regfit_full.3)
```

```
## Subset selection object
## Call: regsubsets.formula(binSupress ~ ., data = VL_Nigeria1, method = "exhaustive")
## 44 Variables  (and intercept)
##                                                                      Forced in
## exposure                                                                 FALSE
## Age.group10-19                                                           FALSE
## Age.group19-20                                                           FALSE
## Age.group20-29                                                           FALSE
## Age.group30-39                                                           FALSE
## Age.group40+                                                             FALSE
## SEXMale                                                                  FALSE
## clinstageStage II                                                        FALSE
## clinstageStage III                                                       FALSE
## clinstageStage IV                                                        FALSE
## funcstatBedridden                                                        FALSE
## funcstatAmbulatory                                                       FALSE
```

Illustration with the "VL_Nigeria" data

```
par(mfrow = c(1,2))

plot(reg_summary.3$adjr2, xlab = "Number of Variables", ylab = "Adjusted RSq", type = "l")
adj_r2_max.3 = which.max(reg_summary.3$adjr2)
points(adj_r2_max.3, reg_summary.3$adjr2[adj_r2_max.3], col ="red", cex = 2, pch = 20)

plot(reg_summary.3$bic, xlab = "Number of Variables", ylab = "BIC", type = "l")
bic_min.3 = which.min(reg_summary.3$bic)
points(bic_min.3, reg_summary.3$bic[bic_min.3], col = "red", cex = 2, pch = 20)
```

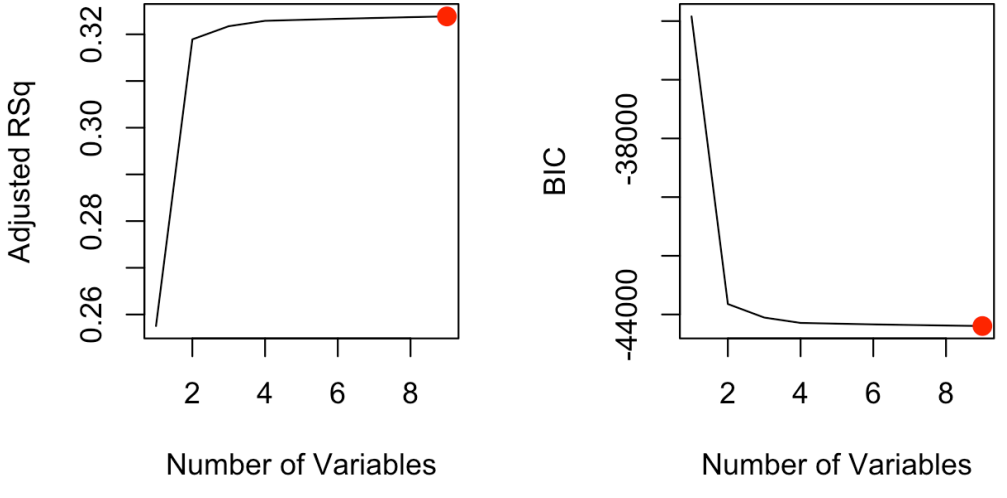


Illustration with the "VL_Nigeria" data

```
plot(regfit_full.3, scale = "adjr2",
     main= "'adjr2' for 'VL_Nigeria' data")
```

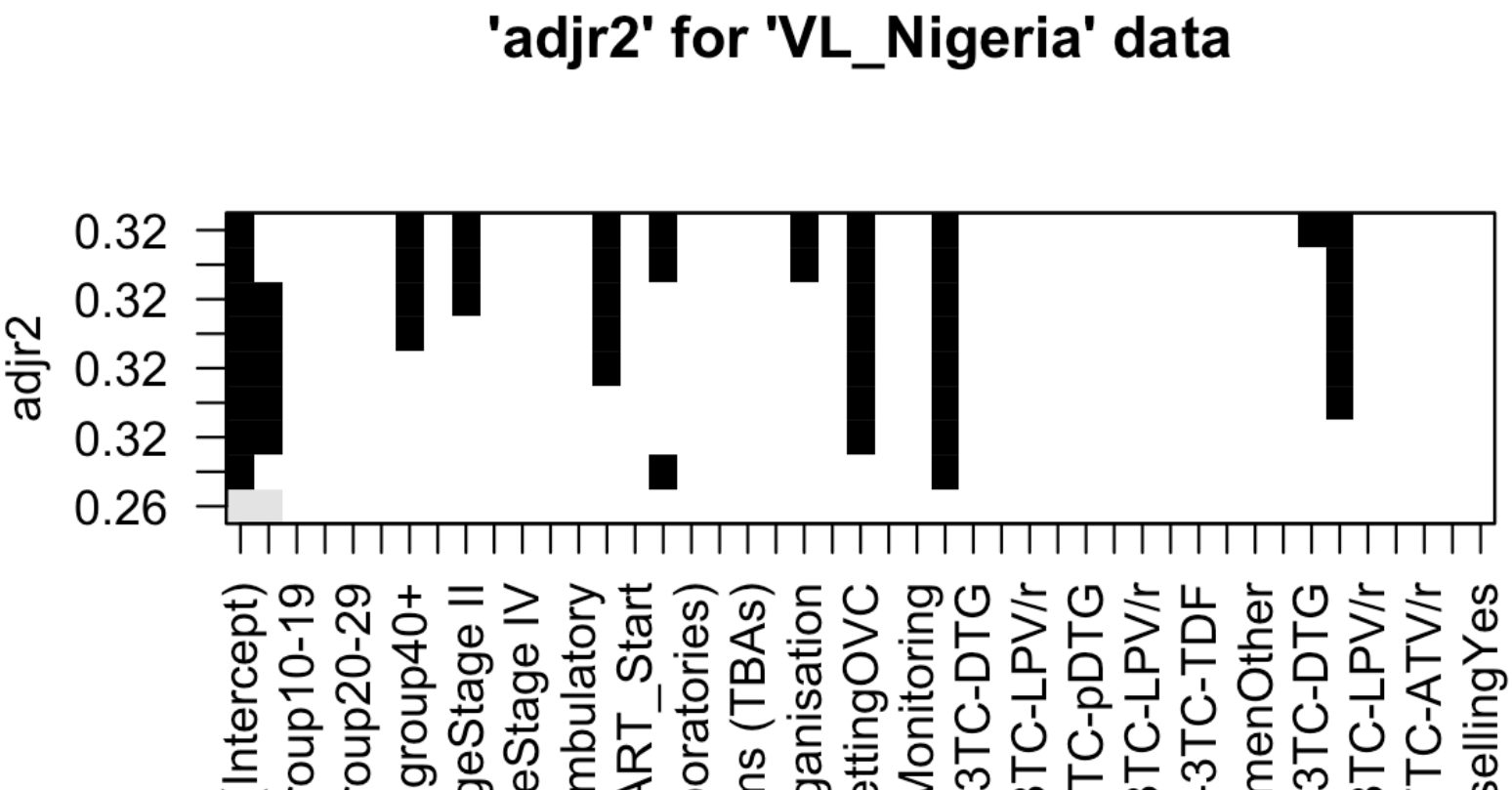
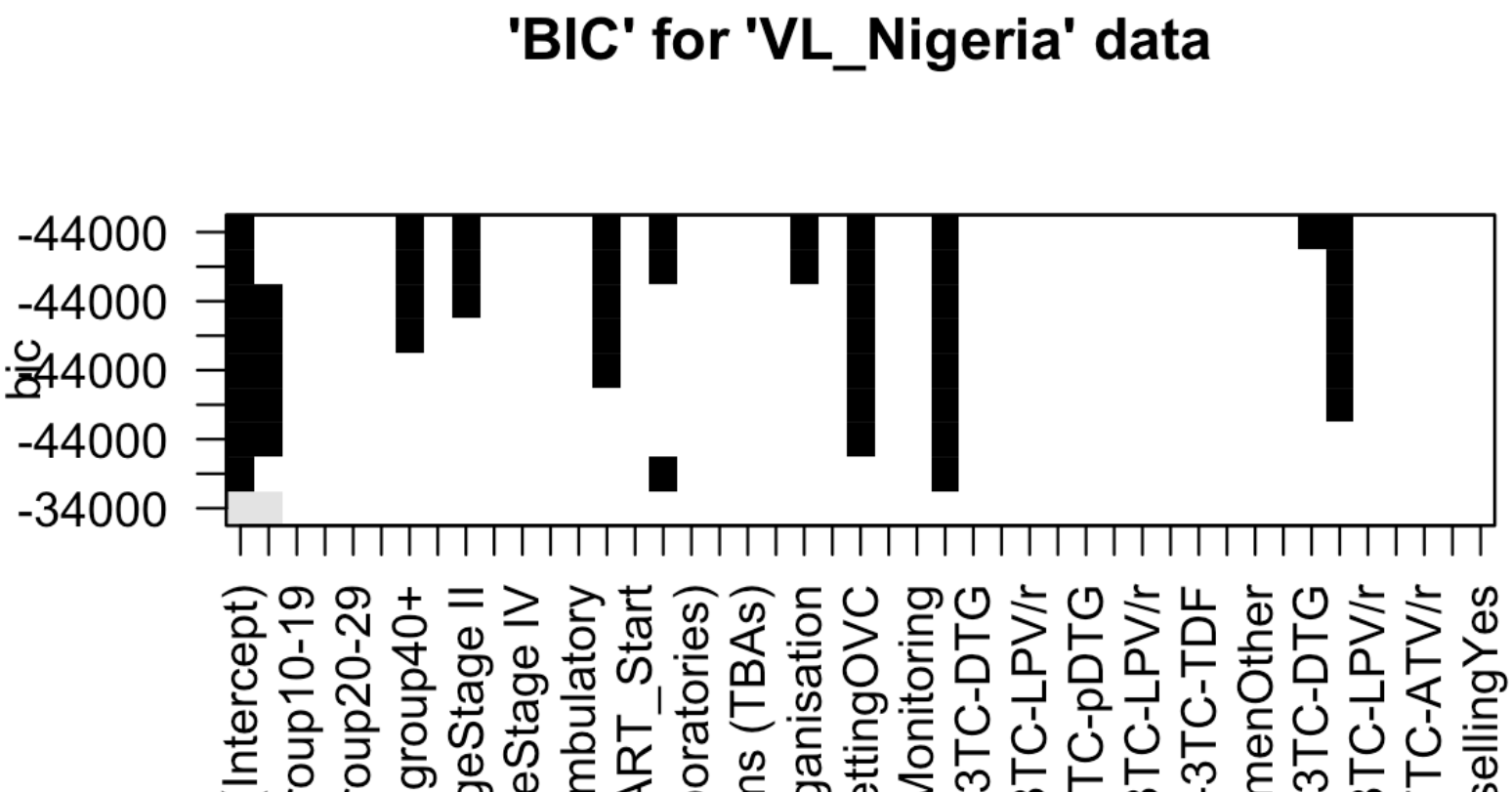


Illustration with the "VL_Nigeria" data

```
plot(regfit_full.3, scale = "bic",
     main= "'BIC' for 'VL_Nigeria' data")
```



What did we get from our "dream model"?

What dataset you used?	Write the equation of the model
VL_Nigeria	PS Model (Weight Generation): $logit(P\ (Exposure = 1)) = \beta 0 + \beta 1(Age.group) + \beta 2(SEX) + \beta 3(AR$
VL_Nigeria	log((P(Outcome=1|Exposure)/P(Outcome=0|Exposure))= log(Odds)= logit(P)=β0+β1xExposu
VL_Nigeria	$ViralLoadSupression = \beta_0 + \beta_{EAC}X_{EAC} + \beta_{age_2}X_{age2} + \beta_{age_3}X_{age3} + \beta_{age_4}X_{age4} +$

Illustration of selection with the 'VL_Nigeria' dataset

Subset selection object $\rightarrow$ automatically selects the "best" sets: regsubsets.formula(binSupress ~ ., data = VL_Nigeria, method = "exhaustive")

- 44 Variables/ parameter (and intercept) with 1 subsets of each size up to 9 variables
- Based on this, we would probably choose to include a model with 3-8 variables:
  - Likely exposure, age, SAB, clinstage, funcstat, TB_Screening_Outcome, Current_ART_Regimen depending on the choice of metrics...

Do you trust this approach? Any concerns??

Even the *"smartest"* software, program or technology will need, until now, a bit of guidance...

Selecting a pool of variables to choose from, may help

```
VL_Nigeria2 <- VL_Nigeria %>% select(binSupress, exposure, Age.group, SEX, clinstage,
                                     funcstat, TB_Screening_Outcome, ART_Start,
                                     Enhanced_Adherence_Counselling, ART_enrollment_setting)
#removing Current_ART_Regimen, State
```

Then re-formulating the selection

```
regfit_full.4 <- regsubsets(binSupress ~ .,
                      data = VL_Nigeria2,
                      method = "exhaustive")
```

```
## Reordering variables and trying again:
```

```
reg_summary.4 <- summary(regfit_full.4)
```

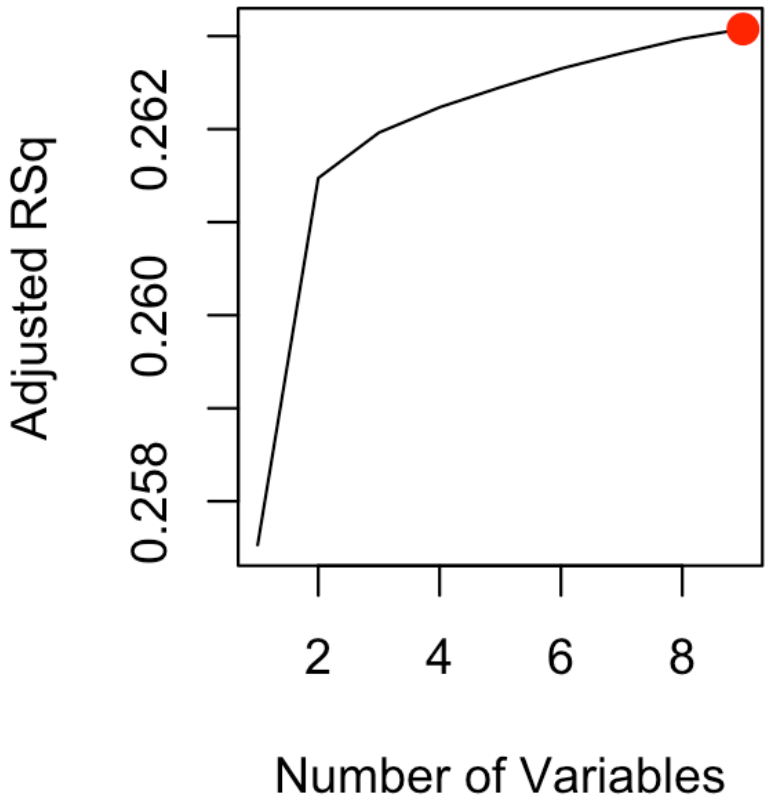
Illustration of selection with the "VL_Nigeria" data

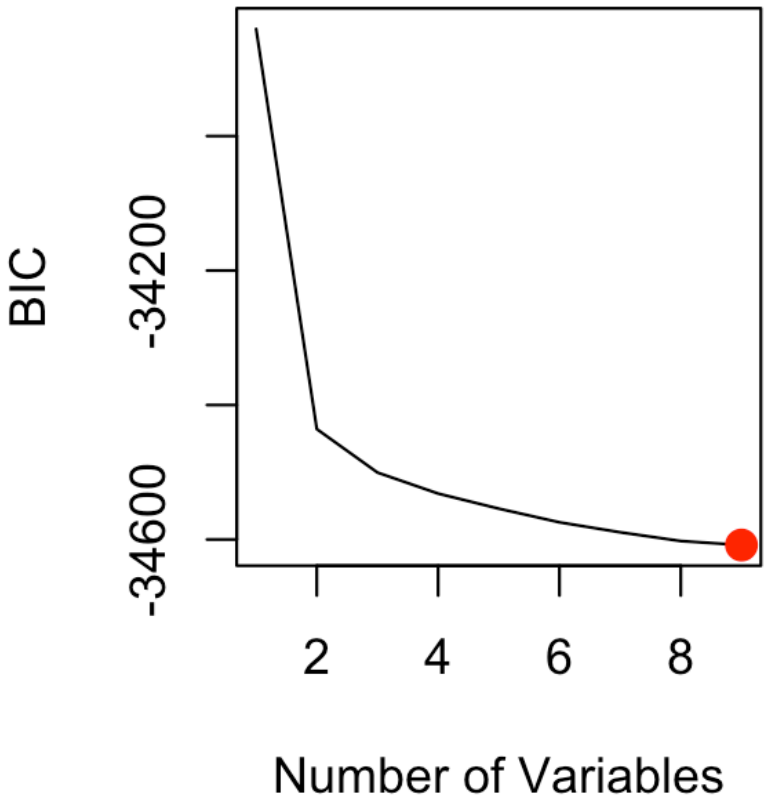
From the selected covariates, there are 9 variables and 24 parameters to be estimated.

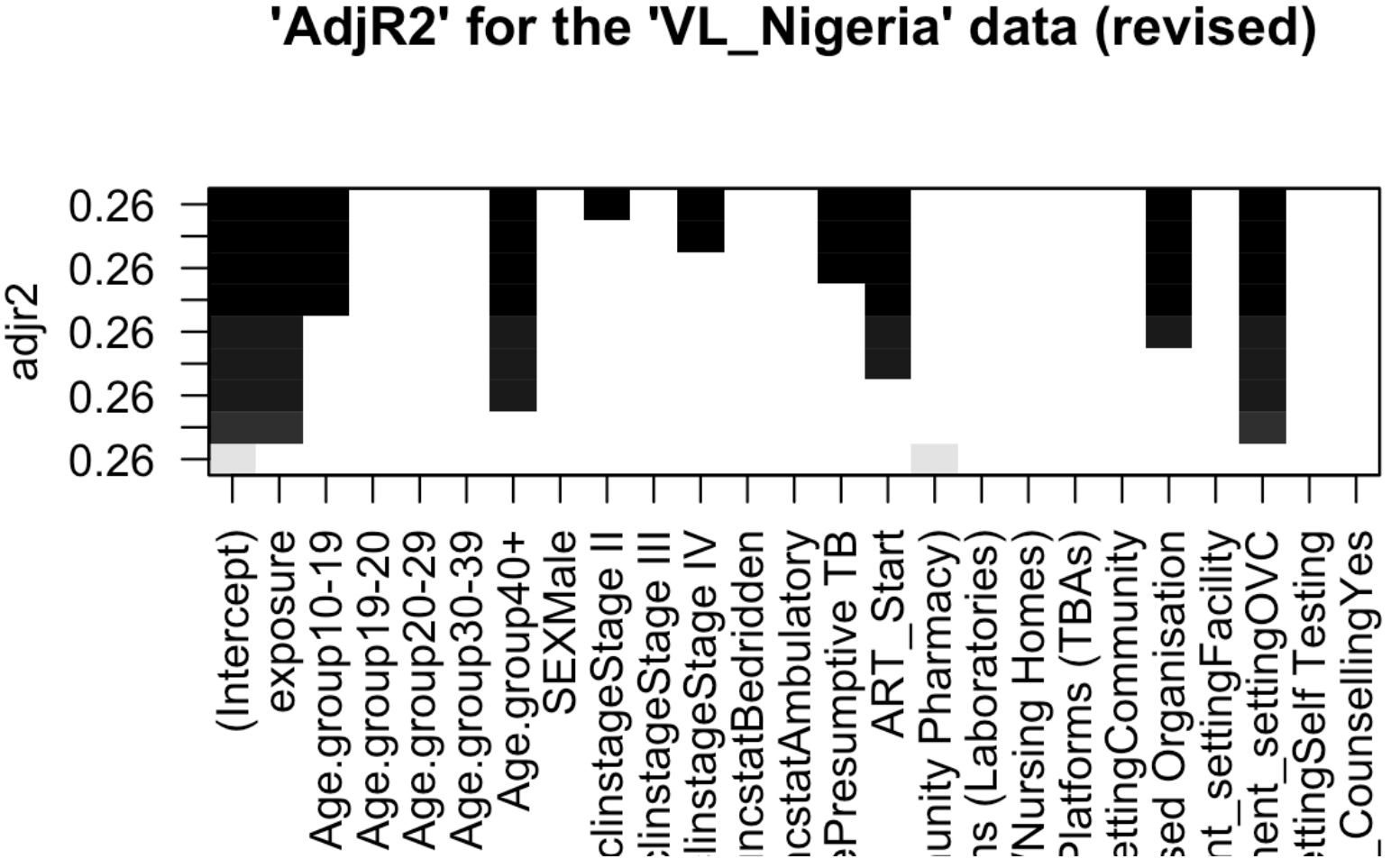
- The package proposes 9 models with several combinations of the parameters/covariates.

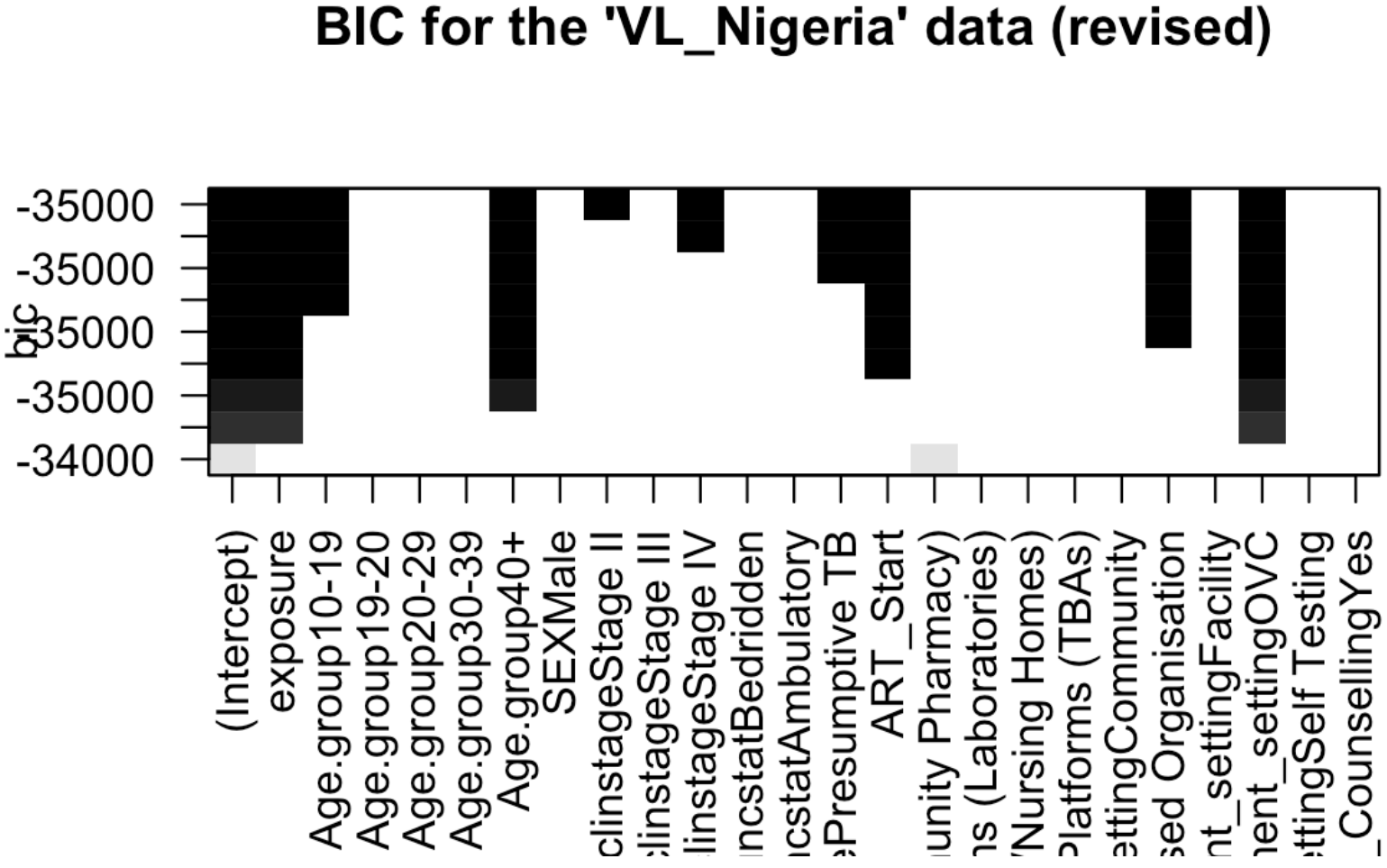
```
## Subset selection object
## Call: regsubsets.formula(binSupress ~ ., data = VL_Nigeria2, method = "exhaustive")
## 24 Variables  (and intercept)
##                                                                  Forced in
## exposure                                                             FALSE
## Age.group10-19                                                       FALSE
## Age.group19-20                                                       FALSE
## Age.group20-29                                                       FALSE
## Age.group30-39                                                       FALSE
## Age.group40+                                                         FALSE
## SEXMale                                                              FALSE
## clinstageStage II                                                    FALSE
## clinstageStage III                                                   FALSE
## clinstageStage IV                                                    FALSE
## funcstatBedridden                                                    FALSE
## funcstatAmbulatory                                                   FALSE
## TB_Screening_OutcomePresumptive TB                                   FALSE
## ART_Start                                                            FALSE
## ART_enrollment_settingClinical Platforms (Community Pharmacy)        FALSE
## ART_enrollment_settingClinical Platforms (Laboratories)              FALSE
## ART_enrollment_settingClinical Platforms (PHCs/Private Clinics/Nursing Homes)  FALSE
## ART_enrollment_settingClinical Platforms (TBAs)                      FALSE
```

Illustration of selection with the "VL_Nigeria" data






'AdjR2' for the 'VL_Nigeria' data (revised)
adjr2
0.26
0.26
0.26
0.26
0.26
(Intercept)
exposure
Age.group10-19
Age.group19-20
Age.group20-29
Age.group30-39
Age.group40+
SEXMale
clinstageStage II
clinstageStage III
linstageStage IV
ncstatBedridden
ncstatAmbulatory
ePresumptive TB
ART_Start
unity Pharmacy)
ns (Laboratories)
Nursing Homes)
Platforms (TBAs)
ettingCommunity
ed Organisation
nt_settingFacility
ent_settingOVC
ettingSelf Testing
_CounsellingYes


BIC for the 'VL_Nigeria' data (revised)
bic
-35000
-35000
-35000
-35000
-34000
(Intercept)
exposure
Age.group10-19
Age.group19-20
Age.group20-29
Age.group30-39
Age.group40+
SEXMale
clinstageStage II
clinstageStage III
linstageStage IV
ncstatBedridden
ncstatAmbulatory
ePresumptive TB
ART_Start
unity Pharmacy)
ns (Laboratories)
'Nursing Homes)
Platforms (TBAs)
ettingCommunity
sed Organisation
nt_settingFacility
ment_settingOVC
ettingSelf Testing
_CounsellingYes

Other Subset Selection

```
#Backwards selection
regsubsets(outcome ~ .,
           data = VL_Nigeria,
           method = "backward")

#Forwards selection
regsubsets(outcome ~ .,
           data = VL_Nigeria,
           method = "forward")

#Stepwise selection
regsubsets(outcome ~ .,
           data = cVL_Nigeria,
           method = "seqrep")
```

Still any decision should be made with caution!

Selection Strategies: Caution!

- Different model fit metrics (e.g., R^2, AIC, Mallow's Cp) can show different models as the "best".

- Can be a recipe for a type I error (chance significant findings) recall Ioannidis paper?
  - Particularly when the number of candidate predictors is large relative to the sample size.

- Automated selection procedures make it easy for researchers to ignore good practices for causal inference, like choosing variables based on a DAG.

- High risk of overfitting to your data set.
  - Selected variables may strongly discriminate among individuals in your data set, but may have less ability to do so on other data.

Overall framework for model specification:

1) Variable specification:

> - What is the universe of variables I would consider?
>   - There is a limited number available, some of the ones we would need, will not be.
>   - We need to think this through and choose.
>   - Leaving variables out is a strong assumption about that not having an effect ($\beta = 0$)

2) Interaction assessment (including assessment of heterogeneity)

3) Confounding assessment (including consideration of precision)

What did we get from our "dream model"?

Selection and Specification

> *"The approach is guided by several principles, including the adherence to a hierarchically defined initial (full) model, and a backward elimination strategy.*

Corollaries

- Collinearity, correlation (other hierarchical structures)
- Moderate or correct for your own Degrees of Freedom
- Multiple testing and Bonferroni corrections

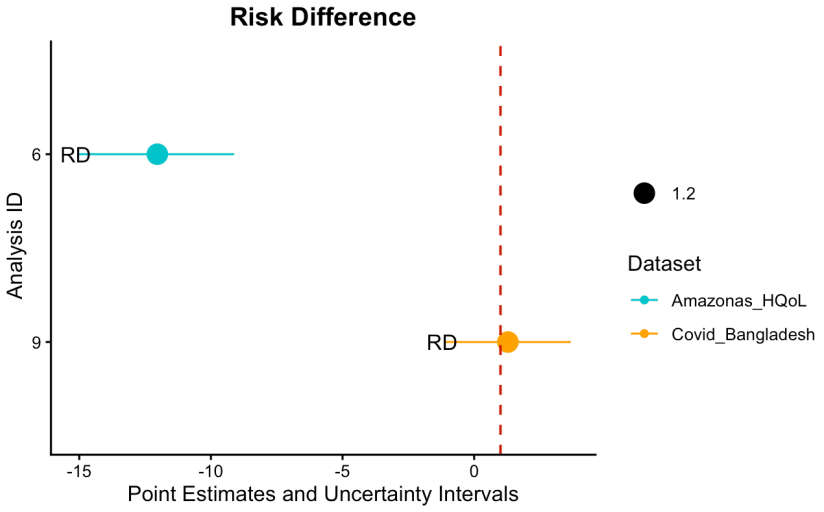
Summary

- **Prediction and inference are different types of problems**.
- If the goal is causal inference, we must think about confounding (and other biases!) and model interpretation is important.
- If the goal is prediction, model interpretation can be less important, and the choice of predictor variables is driven more by the data we have available and model evaluation metrics.
- Model selection strategies can be used for both, but much more care must be taken if you choose to use them for an inference question

“All models are wrong, but some are useful”

— George Box (1919-2013)

What did we get from our "dream model"?



Read it again:

“All models are wrong, but some are useful”

— George Box (1919-2013)

QUESTIONS?

COMMENTS?

RECOMMENDATIONS?

Endogeneity tweet

Resources:

- Greenland et al. 2016. "Outcome modelling strategies in epidemiology: traditional methods and basic alternatives." Int. J. Epidemiol.

- Steyerberg, Ewout W. 2019. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer, Cham.

- Harrell, Frank E., Jr. 2015. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer Series in Statistics. Cham: Springer International Publishing.

Causal Inference and Prediction: Different Goals!

- **Causal Inference**: estimating the effect of a variable on an outcome
  - Usually adjusted for confounding
  - Want a model with good confounder selection, determined via a DAG and substantive knowledge

- **Prediction**: predicting a future outcome using a set of covariates
  - We want predictions that are close to the actual value, but avoid overfitting
  - May prioritize measurable predictors over strong predictors
- **The two goals do NOT require the same model selection approach!**

Prediction??

- **Prediction** aims to anticipate some future outcome using a set of covariates
- Prediction models are typically built using data from an existing group of individuals, with the goal of being able to predict the outcome for future individuals.
- Here, we are concerned with getting the best model that gives "optimal" predictions for future subjects.
- A good predictive model doesn't necessarily tell you anything useful about how to intervene to change the outcome!
    - Age, sex, and prior hospitalization may be good predictors of heart failure, but we wouldn't stop admitting people to the hospital to prevent heart failure

Miscelaneous

Goodness-of-Fit Statistics and model metrics

- **R-squared** R^2 : % of variation in `Y` explained by `"predictors"` variables. The higher R^2, the better the model.
- **Root Mean Squared Error (RMSE)**: the average error performed by the model in predicting the outcome for an observation. RMSE = sqrt(MSE); The $\uparrow$ the RMSE, the better the model.
- **Residual Standard Error (RSE):** the model sigma, a variant of RMSE adjusted # predictors in the model. The $\downarrow$ RSE, the better the model.
- **Mean Absolute Error (MAE)**, measures the prediction error. MAE = mean(abs(observed-predicted)). Less sensitive to outliers compared to RMSE.

Goodness-of-Fit Statistics and model metrics

- **AIC:** (Akaike's Information Criteria) penalizes the inclusion of additional variables to a model.
- **AICc:** is a version of AIC corrected for small sample sizes.
- **BIC:** (or Bayesian information criteria) is a variant of AIC with a stronger penalty.
- **Mallows Cp:** A variant of AIC developed by Colin Mallows.
- **WAIC:** Widely (Watanabe) Application Criterion (Bayesian)
- **PSIS:** Pareto-Smoothed Importance Sampling (Bayesian)
- **DIC:** Deviance Information Criterion (Bayesian)

Evaluating Prediction Models

- The model fit in the **training data** (i.e., the data used to develop a predictive model) is not our primary interest
- What we are primarily interested in is the accuracy of our model predictions **when our model is applied to new data** that was not used as part of the model development process (i.e., **test data**)
- **A good model fit in the training data doesn't necessarily ensure a good ability to predict using other data**.

Paramaterization, Trends, Dose-Response?

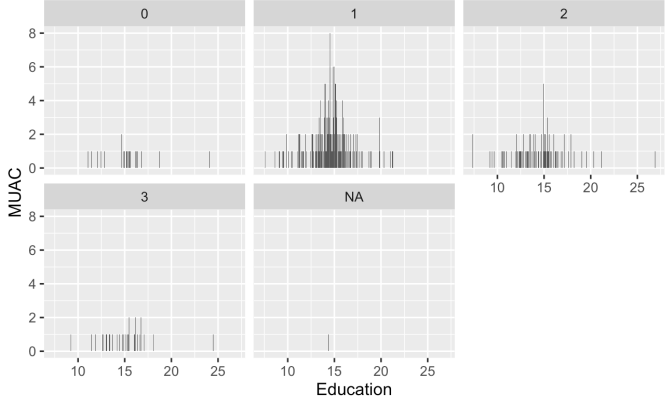
Recall this?

Table 1. Summary of Study Characteristics

Characteristic	N = 355[1]
covid_status	55 (16%)
adm_agemons	
Median (Q1, Q3)	30 (15, 63)
muac_average	
Median (Q1, Q3)	14.70 (13.40, 15.70)
caregiver_educ1	
None	21 (5.9%)
Primary	216 (61%)
Secondary	87 (25%)
Above secondary	30 (8.5%)
nutrition	
malnutrition	123 (35%)

Paramaterization, Trends, Dose-Response?

Let's consider Education of the caregiver and nutritional status measured by Mid Upper Arm Circumference (MUAC) in cm, as the continuous outcome.




```
## # A tibble: 5 × 2
##   caregiver_educ_n m.muac
##              <dbl>  <dbl>
## 1                0   15.2
## 2                1   14.6
## 3                2   14.7
## 4                3   15.2
## 5               NA   14.4
```

Paramaterization, Trends, Dose-Response?

How can we assess their relationship? **Using a continuous variable assumptions?**

```
muac1<- glm(muac_average ~ caregiver_educ_n, data = L25data)
round(j_summ(muac1, confint = T)$coeftable, 2)
```

```
##                    Est.  2.5% 97.5% t val.    p
## (Intercept)       14.67 14.11 15.22  52.15 0.00
## caregiver_educ_n -0.03 -0.39  0.33  -0.18 0.86
```

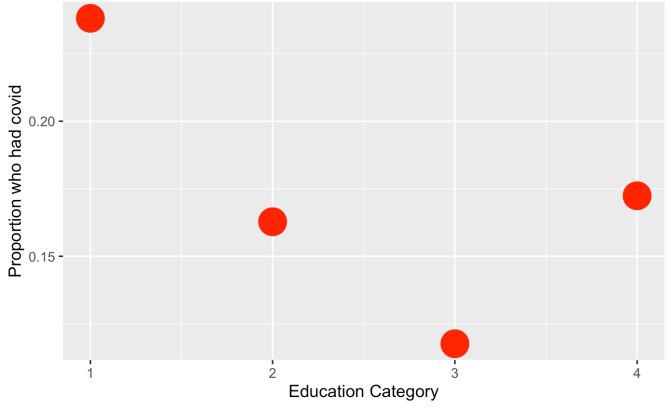
Using a categorical version of the variable assumptions?

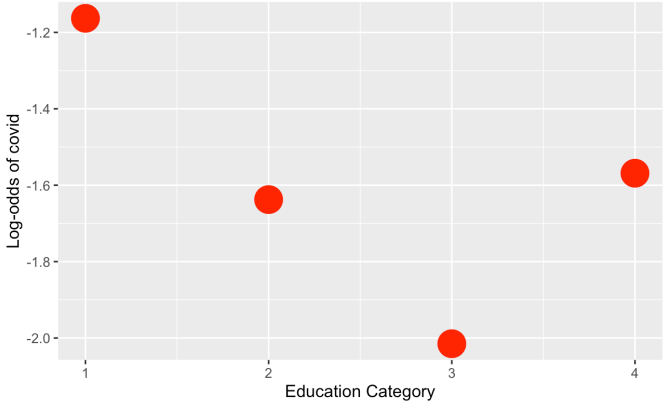
```
muac2<- glm(muac_average ~ factor(caregiver_educ_n), data = L25data)
round(j_summ(muac2, confint = T)$coeftable, 2)
```

```
##                                Est.  2.5% 97.5% t val.    p
## (Intercept)                   15.23 14.18 16.29  28.23 0.00
## factor(caregiver_educ_n)1 -0.65 -1.76  0.46  -1.15 0.25
## factor(caregiver_educ_n)2 -0.79 -1.97  0.39  -1.31 0.19
## factor(caregiver_educ_n)3 -0.24 -1.62  1.14  -0.35 0.73
```

Paramaterization, Trends, Dose-Response?

Let's consider COVID-19 status (disease yes/no) and Education, can we assess their relationship?





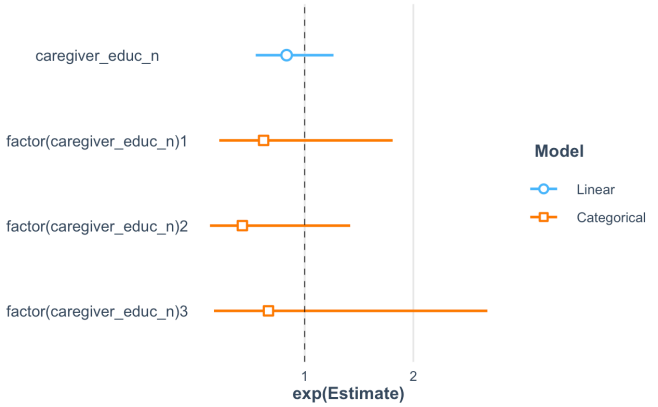
Clinical epidemiology of COVID-19 among hospitalized children in rural western Kenya

Paramaterization, Trends, Dose-Response?

Dichotomous Outcome

```
muac3<- glm(covid_status ~ caregiver_educ_n
            data = L25data, family= binomia
round(j_summ(muac3, confint = T)$coeftable,
```

```
##                    Est.  2.5% 97.5% z val.    p
## (Intercept)      -1.44 -2.05 -0.83  -4.64 0.00
## caregiver_educ_n -0.18 -0.60  0.24  -0.85 0.39
```




```
muac4<- glm(covid_status ~ factor(caregiver
            data = L25data, family= binomia
round(j_summ(muac4, confint = T)$coeftable,
```

```
##                              Est.  2.5% 97.5% z val.    p
## (Intercept)                -1.16 -2.17 -0.16  -2.27 0.02
## factor(caregiver_educ_n)1 -0.47 -1.54  0.59  -0.87 0.38
## factor(caregiver_educ_n)2 -0.85 -2.05  0.35  -1.39 0.16
## factor(caregiver_educ_n)3 -0.41 -1.80  0.99  -0.57 0.57
```

Paramaterization, Trends, Dose-Response?

What happens when the referent group has a very small sample size?

- OR estimates are all imprecise (since the imprecision for the referent group is propogated)
  - One way are this is to use incremental coding of the dummy variables

```
L25data$edu1<- L25data$edu2<-L25data$edu3<-0

L25data$edu1[L25data$caregiver_educ_n>=1]<- 1
L25data$edu2[L25data$caregiver_educ_n>=2]<- 1
L25data$edu3[L25data$caregiver_educ_n>=3]<- 1

L25data$edu1[is.na(L25data$caregiver_educ_n)==T]<- NA
L25data$edu2[is.na(L25data$caregiver_educ_n)==T]<- NA
L25data$edu3[is.na(L25data$caregiver_educ_n)==T]<- NA
```

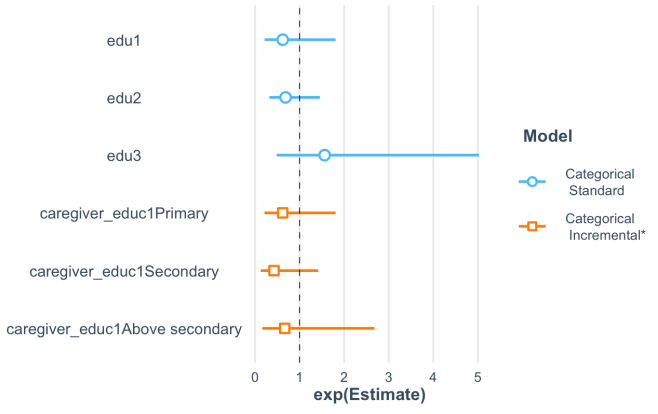
Clinical epidemiology of COVID-19 among hospitalized children in rural western Kenya

Paramaterization, Trends, Dose-Response?

Incremental categories*

```
muac5<- glm(covid_status ~ edu1+edu2+edu3,
            data = L25data,
            family= binomial(link = "logit"
round(j_summ(muac5, confint = T)$coeftable,
```

```
##                Est.  2.5% 97.5% z val.    p
## (Intercept) -1.16 -2.17 -0.16  -2.27 0.02
## edu1        -0.47 -1.54  0.59  -0.87 0.38
## edu2        -0.38 -1.13  0.38  -0.98 0.33
## edu3         0.45 -0.72  1.61   0.75 0.45
```



OR what we usually do...

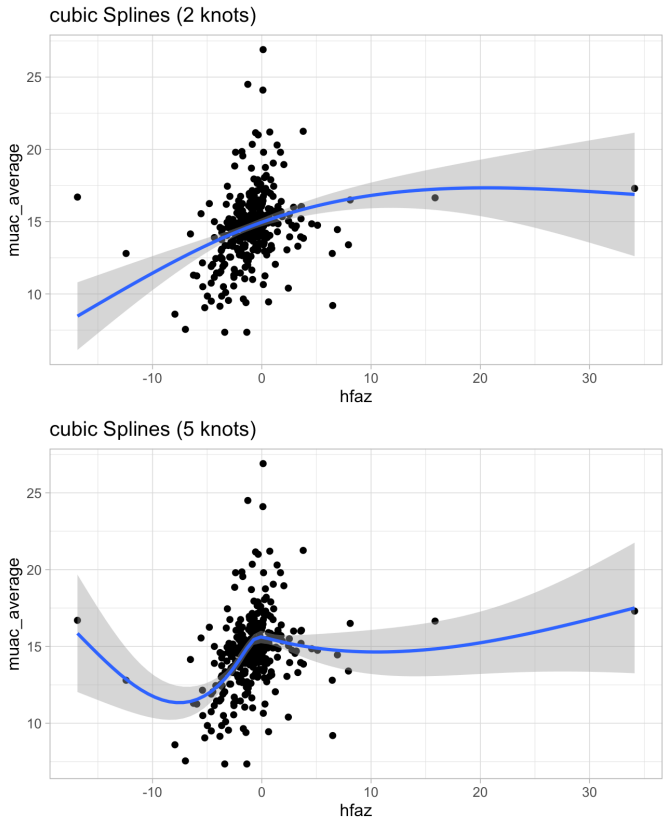
```
muac6<- glm(covid_status ~ caregiver_educ1,
            data = L25data, family= binomia
round(j_summ(muac6, confint = T)$coeftable,
```

```
##                                 Est.  2.5% 97.5% z val.    p
## (Intercept)                    -1.16 -2.17 -0.16  -2.27 0.02
## caregiver_educ1Primary         -0.47 -1.54  0.59  -0.87 0.38
## caregiver_educ1Secondary       -0.85 -2.05  0.35  -1.39 0.16
## caregiver_educ1Above secondary -0.41 -1.80  0.99  -0.57 0.57
```

*Incremental means, that the reference is the immediately previous category/group

Splines examples!

- Y= muac_average (MUAC in cms);
- X= hfaz (Height for age, Z-score)



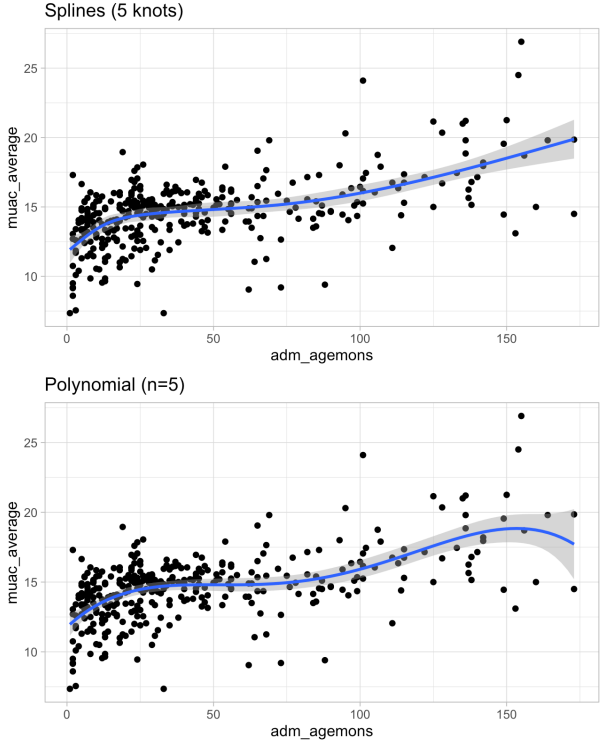

```
require(splines)
lmmod6.11a<- glm(muac_average~ breast_feedi
                 ns(hfaz, df=2), data = c
#round(j_summ(lmmod6.11a, confint = T)$coef
lmmod6.11b<- glm(muac_average~ breast_feedi
                 ns(hfaz, df=5), data = c
round(j_summ(lmmod6.11b, confint = T)$coeft
```

```
##                                Est.   2.5% 97.5% t
## (Intercept)                   12.95   9.55 16.35
## breast_feeding1nobreastfeed -0.08 -0.74  0.59  -
## adm_sexMale                    0.08 -0.34  0.49
## age_cat_new>=5 years           3.99   3.01  4.96
## age_cat_new12-23 months        1.58   0.66  2.50
## age_cat_new2-5 years           2.53   1.59  3.46
## age_cat_new6-11 months         1.24   0.26  2.22
## ns(hfaz, df = 5)1             -0.64 -3.91  2.63  -
## ns(hfaz, df = 5)2              0.03 -3.32  3.39
## ns(hfaz, df = 5)3              3.32   0.24  6.40
## ns(hfaz, df = 5)4             -2.67 -9.94  4.60  -
## ns(hfaz, df = 5)5              7.34   3.22 11.46
```

```
#attr(terms(lmmod6.11b), "predvars")
```

Last Splines examples!

- Y= muac_average (MUAC in cms);
- X= adm_agemons (Age in months)




```
lmmod6.12a<- glm(muac_average~ breast_feedi
                        ns(adm_agemons, df=5), d
#round(j_summ(lmmod6.12a, confint = T)$coef
lmmod6.12b<- glm(muac_average~ breast_feedi
                        poly(adm_agemons, 5, raw
round(j_summ(lmmod6.12b, confint = T)$coeft
```

```
##                                    Est.  2.5% 97.5%
## (Intercept)                       11.70 10.75 12.66
## breast_feeding1nobreastfeed       -0.29 -0.97  0.38
## adm_sexMale                        0.13 -0.30  0.56
## poly(adm_agemons, 5, raw = T)1     0.21  0.09  0.34
## poly(adm_agemons, 5, raw = T)2     0.00 -0.01  0.00
## poly(adm_agemons, 5, raw = T)3     0.00  0.00  0.00
## poly(adm_agemons, 5, raw = T)4     0.00  0.00  0.00
## poly(adm_agemons, 5, raw = T)5     0.00  0.00  0.00
```

```
#attr(terms(lmmod6.12b), "predvars")
```