ID529: Modeling workflows, The RMarkdown Version

Jarvis Chen

2023-01-15

Here, we illustrate how the `gtsummary`, `sjPlot`, `stargazer`, and `ggstatsplot` packages can be used in an RMarkdown document to generate pretty tables and figures. We show a few more options for adjusting things like table captions to make your tables and figures more informative.

This example uses the NHANES dataset and the linear and logistic regression models that we looked at on Day 4.

Setting up the data and fitting linear and logistic regression models to use as examples

```
# Set up the analytic dataset
df <- NHANES  %>%
  # Remember that we have to restrict to people 25 and above
  filter(Age>=25)  %>%
  # recoding of the variables we're going to use
  mutate(agecat = case_when(
      Age < 35 ~ "25-34",
      35 <= Age & Age < 45 ~ "35-44",
      Age >= 45 & Age < 55 ~ "45-54",
      Age >= 55 & Age < 65 ~ "55-64",
      Age >= 65 & Age < 75 ~ "65-74",
      Age >= 75 ~ "75+"),
    # We want College Grad to be the reference category for education, so we'll
    # re-order the factor so that it is reversed from the way it came in the NHANES dataset
    Education = factor(Education,
                       levels=rev(levels(NHANES$Education))),
    # Here we collapse Hispanic and Mexican into the Hispanic category
    racecat = factor(case_when(
      Race1 %in% c("Hispanic", "Mexican") ~ "Hispanic",
      Race1 %in% c("Asian", "Other") ~ "Other Non-Hispanic",
      Race1 == "Black" ~ "Black Non-Hispanic",
      Race1 == "White" ~ "White Non-Hispanic"),
      levels = c("White Non-Hispanic", "Black Non-Hispanic", "Hispanic", "Other Non-Hispanic"))
  ) %>%
  # select just variables we are going to use in the analysis
  select(ID, SurveyYr, Gender, Age, agecat, Education, racecat, BPSysAve, SmokeNow)


# Knowing that there are differing amounts of missing data in the different variables,
# it would be better if we defined our analytic dataset based non-missing data on all of
```

```
# variables we know we are going to include in our analysis.
# NOTE: There is a substantial amount of missing data, so complete case analysis could
# yield biased results if the data are not Missing Completely At Random!

df_completecase <- df %>%
  filter(!is.na(BPSysAve) & !is.na(agecat) & !is.na(Gender) & !is.na(racecat))
```

```
# Fit the linear regression models of interest
# relating BPSysAve to Education, and adjusting
# for covariates: age category, gender, race category, and the interaction
# of race and gender

lm_model1 <- lm(BPSysAve ~ factor(Education),
                data=df_completecase)
lm_model2 <- lm(BPSysAve ~ factor(Education) + factor(agecat) + Gender,
                data=df_completecase)
lm_model3 <- lm(BPSysAve ~ factor(Education) + factor(agecat) + Gender + factor(racecat),
                data=df_completecase)

lm_model4 <- lm(BPSysAve ~ factor(Education) + factor(agecat) + interaction(Gender,factor(racecat)),
                 data=df_completecase)

# Fit the logistic regression models of interest
# relating the same covariates to SmokeNow

logistic_model1 <- glm(SmokeNow ~ factor(Education),
               family=binomial(link=logit),
               data=df_completecase)

logistic_model2 <- glm(SmokeNow ~ factor(Education) + factor(agecat) + Gender,
               family=binomial(link=logit),
               data=df_completecase)

logistic_model3 <- glm(SmokeNow ~ factor(Education) + factor(agecat) + Gender + factor(racecat),
               family=binomial(link=logit),
               data=df_completecase)

logistic_model4 <- glm(SmokeNow ~ factor(Education) + factor(agecat) + interaction(Gender,factor(raceca
                family=binomial(link=logit),
                data=df_completecase)
```

Note that if you simply call `broom::tidy` and use the default chunk options in RMarkdown, the results of calling `broom::tidy` will appear in your knitted output.

```
broom::tidy(lm_model4)
```

```
## # A tibble: 17 x 5
##    term                                estimate std.error statistic   p.value
##    <chr>                                  <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)                          109.        0.598   183.    0
##  2 factor(Education)Some College          2.89      0.512     5.65  1.69e-  8
##  3 factor(Education)High School           3.41      0.572     5.97  2.46e-  9
##  4 factor(Education)9 - 11th Grade        2.67      0.677     3.94  8.31e-  5
##  5 factor(Education)8th Grade             3.68      0.929     3.96  7.58e-  5
##  6 factor(agecat)35-44                    4.06      0.616     6.59  4.83e- 11
```

```
##  7 factor(agecat)45-54                        6.64      0.614    10.8   4.70e- 27
##  8 factor(agecat)55-64                       12.7       0.649    19.6   8.15e- 83
##  9 factor(agecat)65-74                       15.8       0.736    21.5   3.14e- 99
## 10 factor(agecat)75+                         23.8       0.832    28.7   6.50e-170
## 11 interaction(Gender, factor(racecat))m~     3.70      0.477     7.74  1.11e- 14
## 12 interaction(Gender, factor(racecat))f~     4.79      0.903     5.30  1.18e-  7
## 13 interaction(Gender, factor(racecat))m~     7.73      0.952     8.12  5.69e- 16
## 14 interaction(Gender, factor(racecat))f~    -0.716     0.904    -0.792 4.28e-  1
## 15 interaction(Gender, factor(racecat))m~     4.38      0.858     5.11  3.40e-  7
## 16 interaction(Gender, factor(racecat))f~    -3.03      1.08     -2.81  4.95e-  3
## 17 interaction(Gender, factor(racecat))m~     4.30      1.12      3.82  1.33e-  4
```

Summary tables using gtsummary

We can integrate calls to `gtsummary::tbl_regression` into our RMarkdown document and get pretty tables to appear in our knitted document.

```
# Creating Pretty Tables ------------------------------------------------------

tbl_lm_model1 <-
  tbl_regression(lm_model1, label = list('factor(Education)' ~ 'Education')) %>%
  bold_labels() %>%
    modify_caption("**Table 1:** Model 1 estimates showing crude associations of educational categories

# We can also add some model fit statistics to the output
# by using add_glance_table()
tbl_lm_model1_glance <-
  tbl_regression(lm_model1, label = list('factor(Education)' ~ 'Education')) %>%
  bold_labels() %>%
  add_glance_table() %>%
  modify_caption("**Table 1:** Model 1 estimates showing crude associations of educational categories wi
tbl_lm_model1_glance
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

```
# What if I want to compare my models

# We can set the gtsummary theme so that the table is formatted
# e.g. here we format to the JAMA journal format
set_gtsummary_theme(theme_gtsummary_journal("jama"))
```

```
## Setting theme `JAMA`
## Setting theme `JAMA`
```

```
# First we format each of the models using tbl_regression

# Note for this first one that I am showing how to integrate this
# into a workflow where you start with the analytic data frame,
# pipe it into lm() and then pipe the results into
# tbl_regression
tbl_lm_model1 <- df_completecase %>%
  lm(BPSysAve ~ factor(Education),
```

Table 1: **Table 1:** Model 1 estimates showing crude associations of educational categories with average systolic blood pressure

Characteristic	**Beta**	**95% CI**	**p-value**
___Education___			
College Grad	—	—	
Some College	3.8	2.7, 4.8	<0.001
High School	5.1	3.9, 6.4	<0.001
9 - 11th Grade	4.7	3.3, 6.1	<0.001
8th Grade	7.9	6.1, 9.8	<0.001
R^2	0.019		
Adjusted R^2	0.018		
Sigma	17.3		
Statistic	30.4		
p-value	<0.001		
df	4		
Log-likelihood	-26,991		
AIC	53,994		
BIC	54,035		
Deviance	1,886,669		
Residual df	6,319		
No. Obs.	6,324		

```
    data=.) %>%
  tbl_regression(intercept=TRUE,
                 label = list('factor(Education)' ~ 'Education'))

tbl_lm_model2 <- lm_model2 %>%
  tbl_regression(intercept=TRUE,
                 label = list('factor(Education)' ~ 'Education',
                              'factor(agecat)' ~ 'Age category'))

tbl_lm_model3 <- lm_model3 %>%
  tbl_regression(intercept=TRUE,
                 label = list('factor(Education)' ~ 'Education',
                              'factor(agecat)' ~ 'Age category',
                              'factor(racecat)' ~ 'Racialized group'))

tbl_lm_model4 <- lm_model4 %>%
  tbl_regression(intercept=TRUE,
                 label = list('factor(Education)' ~ 'Education',
                              'factor(agecat)' ~ 'Age category',
                              'interaction(Gender, factor(racecat))' ~ 'Gender X Racialized group'),
                 )


# Now that each of the models has been formatted, I can use tbl_merge to
# put the models together to be shown side-by-side
tbl_merge_ex1 <-
  tbl_merge(
    tbls = list(tbl_lm_model1,
                tbl_lm_model2,
```

Table 2: Table 2: Comparison of linear model results

Characteristic	**Beta** **(95% CI)**	**p-value**	**Beta** **(95% CI)**	**p-value**	**Beta**
(Intercept)	119 (118 to 119)	<0.001	109 (108 to 110)	<0.001	109 (1
Education					
College Grad	—		—		
Some College	3.8 (2.7 to 4.8)	<0.001	3.2 (2.2 to 4.2)	<0.001	2.9 (1
High School	5.1 (3.9 to 6.4)	<0.001	3.9 (2.7 to 5.0)	<0.001	3.4 (2
9 - 11th Grade	4.7 (3.3 to 6.1)	<0.001	3.2 (1.9 to 4.5)	<0.001	2.7 (1
8th Grade	7.9 (6.1 to 9.8)	<0.001	3.7 (1.9 to 5.4)	<0.001	3.6 (1
Age category					
25-34			—		
35-44			4.0 (2.8 to 5.2)	<0.001	4.1 (2
45-54			6.7 (5.5 to 7.9)	<0.001	6.7 (5
55-64			13 (11 to 14)	<0.001	13 (1
65-74			16 (14 to 17)	<0.001	16 (1
75+			24 (22 to 25)	<0.001	24 (2
Gender					
female			—		
male			4.0 (3.2 to 4.8)	<0.001	4.1 (3
Racialized group					
White Non-Hispanic					
Black Non-Hispanic					4.4 (3
Hispanic					0.03 (-
Other Non-Hispanic					-1.3 (-2
Gender X Racialized group					
female.White Non-Hispanic					
male.White Non-Hispanic					
female.Black Non-Hispanic					
male.Black Non-Hispanic					
female.Hispanic					
male.Hispanic					
female.Other Non-Hispanic					
male.Other Non-Hispanic					

```
                tbl_lm_model3,
                tbl_lm_model4),
    # the tab_spanner argument specifies the headings at the top of the table
    # that span multiple columns
    tab_spanner = c("**Model 1**", "**Model 2**", "**Model 3**", "**Model 4**")
  ) %>%
  modify_caption("Table 2: Comparison of linear model results")

tbl_merge_ex1
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Summary tables using `sjPlot`

`sjPlot::tab_model` also works seamlessly with RMarkdown, although annoyingly it seems complicated to add a table caption (title).

```
# Pretty tabular output using sjPlot ----------------------------------------------
# tab_model can also print multiple models at once, which are printed side by side.
tab_model(lm_model2, lm_model3)
```

BP Sys Ave

BP Sys Ave

Predictors

Estimates

CI

p

Estimates

CI

p

(Intercept)

109.18

108.08 – 110.28

<0.001

108.99

107.84 – 110.13

<0.001

Education [Some College]

3.17

2.17 – 4.18

<0.001

2.88

1.88 – 3.88

<0.001

Education [High School]

3.85

2.74 – 4.96

<0.001

3.40

2.28 – 4.52

<0.001

Education [9 - 11thGrade]

3.24

1.94 – 4.55

<0.001

2.68

1.35 – 4.00

<0.001

Education [8th Grade]

3.65

1.95 – 5.36

<0.001

3.63

1.81 – 5.45

<0.001

factor(agecat)35-44

4.01

2.80 – 5.23

<0.001

4.08

2.87 – 5.29

<0.001

factor(agecat)45-54

6.66

5.46 – 7.86

<0.001

6.67

5.46 – 7.87

<0.001

factor(agecat)55-64

12.74

11.47 – 14.01

<0.001

12.71

11.43 – 13.98

<0.001

factor(agecat)65-74

15.86

14.42 – 17.29

<0.001

15.89

14.45 – 17.34

<0.001

agecat [75+]

23.61

22.00 – 25.22

<0.001

23.86

22.23 – 25.49

<0.001

Gender [male]

4.01

3.23 – 4.79

<0.001

4.06

3.28 – 4.84

<0.001

racecat [BlackNon-Hispanic]

4.44

3.15 – 5.73

<0.001

racecat [Hispanic]

0.03

-1.23 – 1.30

0.958

racecat [OtherNon-Hispanic]

-1.29

-2.82 – 0.24

0.099

Observations

6324

6324

R2 / R2 adjusted

0.183 / 0.182

0.190 / 0.188

```
# Note that for generalized linear models,  instead of Estimates
# the column is labeled Odds Ratios (for logistic regression)
tab_model(logistic_model2, logistic_model3)
```

Smoke Now

Smoke Now

Predictors

Odds Ratios

CI

p

Odds Ratios

CI

p

(Intercept)

0.76

0.59 – 0.97

0.030

0.68

0.52 – 0.88

0.004

Education [Some College]

2.31

1.84 – 2.90

<0.001

2.36

1.88 – 2.98

<0.001

Education [High School]

2.72

2.13 – 3.47

<0.001

2.81

2.19 – 3.62

<0.001

Education [9 - 11thGrade]

4.85

3.73 – 6.35

<0.001

5.01

3.82 – 6.61

<0.001

Education [8th Grade]

3.30

2.30 – 4.75

<0.001

3.79

2.58 – 5.57

<0.001

factor(agecat)35-44

0.66

0.52 – 0.85

0.001

0.68

0.53 – 0.87

0.003

factor(agecat)45-54

0.51

0.40 – 0.64

<0.001

0.51

0.40 – 0.65

<0.001

factor(agecat)55-64

0.41

0.32 – 0.53

<0.001

0.41

0.32 – 0.53

<0.001

factor(agecat)65-74

0.18

0.13 – 0.24

<0.001

0.18

0.13 – 0.24

<0.001

agecat [75+]

0.06

0.04 – 0.09

<0.001

0.06

0.04 – 0.09

<0.001

Gender [male]

0.93

0.79 – 1.09

0.372

0.92

0.78 – 1.08

0.316

racecat [BlackNon-Hispanic]

1.77

1.35 – 2.34

<0.001

racecat [Hispanic]

0.76

0.58 – 1.00

0.055

racecat [OtherNon-Hispanic]

2.62

1.82 – 3.81

<0.001

Observations

2899

2899

R2 Tjur

0.148

0.164

Table summaries using `stargazer`

The `stargazer` package also can be used to generate pretty tables. Note that we had to specify `results='asis'` in the code chunk (see the Rmd file).

```
stargazer(lm_model1, lm_model2, lm_model3, lm_model4,
          header=FALSE,
          type='html',
          title="Table 2: Comparison of models for average systolic blood pressure in relation to educa
```

Table 2: Comparison of models for average systolic blood pressure in relation to education

Dependent variable:

BPSysAve

(1)

(2)

(3)

(4)

factor(Education)Some College

3.750***

3.175***

2.879***

2.889***

(0.559)

(0.511)

(0.512)

(0.512)

factor(Education)High School

5.140***

3.851***

3.396***

3.414***

(0.619)

(0.567)

(0.572)

(0.572)

factor(Education)9 - 11th Grade

4.688***

3.243***

2.677***

2.668***

(0.727)

(0.666)

(0.678)

(0.677)

factor(Education)8th Grade

7.921***

3.653***

3.634***

3.678***

(0.941)

(0.870)

(0.929)

(0.929)

factor(agecat)35-44

4.015***

4.081***

4.056***

(0.617)

(0.616)

(0.616)

factor(agecat)45-54

6.660***

6.667***

6.642***

(0.614)

(0.614)

(0.614)

factor(agecat)55-64

12.742***

12.707***

12.708***

(0.647)

(0.650)

(0.649)

factor(agecat)65-74

15.856***

15.894***

15.840***

(0.731)

(0.736)

(0.736)

factor(agecat)75+

23.610***

23.857***

23.839***

(0.820)

(0.832)

(0.832)

Gendermale

4.010***

4.062***

(0.398)

(0.397)

factor(racecat)Black Non-Hispanic

4.440***

(0.658)

factor(racecat)Hispanic

0.034

(0.647)

factor(racecat)Other Non-Hispanic

-1.290*

(0.781)

interaction(Gender, factor(racecat))male.White Non-Hispanic

3.696***

(0.477)

interaction(Gender, factor(racecat))female.Black Non-Hispanic

4.787***

(0.903)

interaction(Gender, factor(racecat))male.Black Non-Hispanic

7.725***

(0.952)

interaction(Gender, factor(racecat))female.Hispanic

-0.716

(0.904)

interaction(Gender, factor(racecat))male.Hispanic

4.379***

(0.858)

interaction(Gender, factor(racecat))female.Other Non-Hispanic

-3.030***

(1.078)

interaction(Gender, factor(racecat))male.Other Non-Hispanic

4.297***

(1.124)

Constant

118.563***

109.178***

108.987***

109.178***

(0.391)

(0.563)

(0.583)

(0.598)

Observations

6,324

6,324

6,324

6,324

R2

0.019

0.183

0.190

0.191

Adjusted R2

0.018

0.182

0.188

0.189

Residual Std. Error

17.279 (df = 6319)

15.773 (df = 6313)

15.713 (df = 6310)

15.707 (df = 6307)

F Statistic

30.376*** (df = 4; 6319)

141.603*** (df = 10; 6313)

113.754*** (df = 13; 6310)

92.940*** (df = 16; 6307)

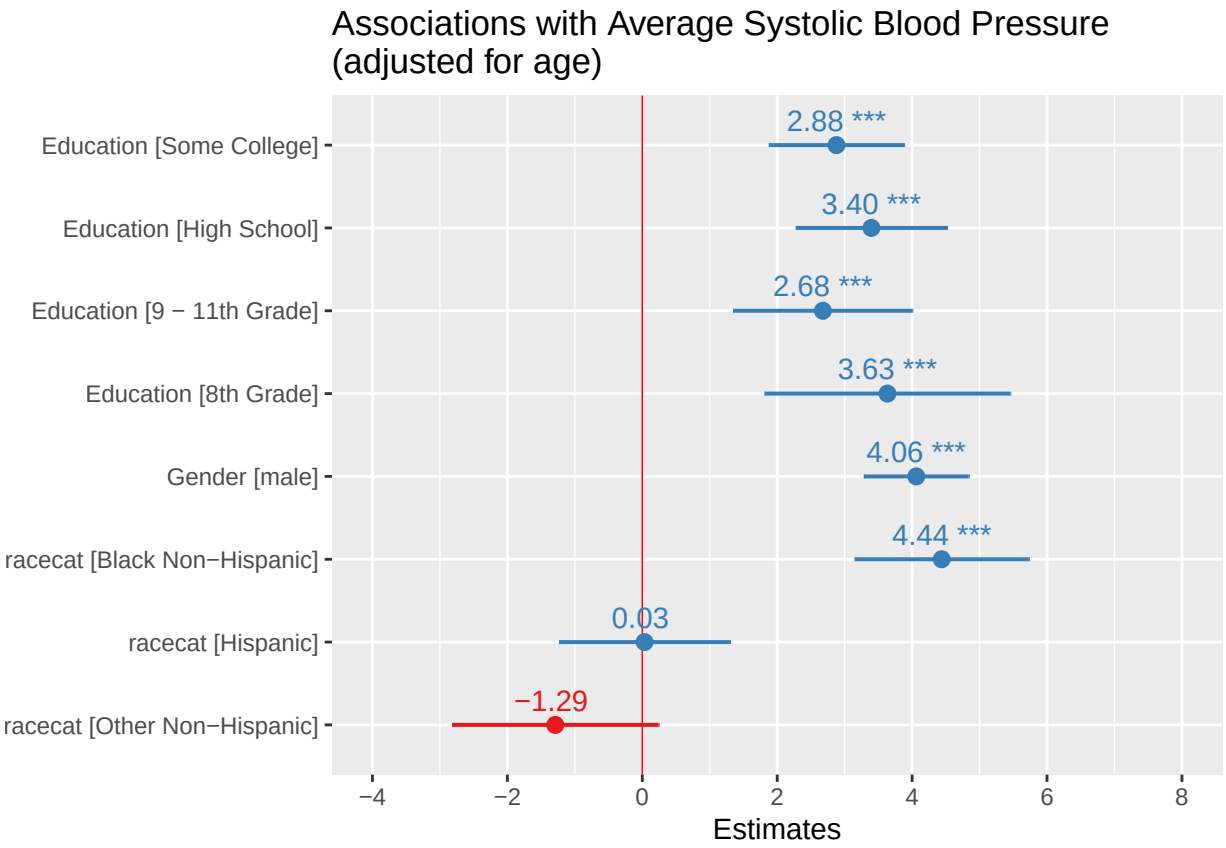
Note:

p<0.1; ***p<0.05;*** p<0.01

Plotting model results using `sjPlot`

```
# Plotting models ------------------------------------------------------------


# We can also add value labels and a title
plot_model(lm_model3,
           show.values=TRUE, value.offset=0.3,
           title="Associations with Average Systolic Blood Pressure (adjusted for age)",
           vline.color="red",
           terms = c("factor(Education)Some College",
                     "factor(Education)High School",
                     "factor(Education)9 - 11th Grade",
                     "factor(Education)8th Grade",
                     "Gendermale",
                     "factor(racecat)Black Non-Hispanic",
                     "factor(racecat)Hispanic",
                     "factor(racecat)Other Non-Hispanic"))
```



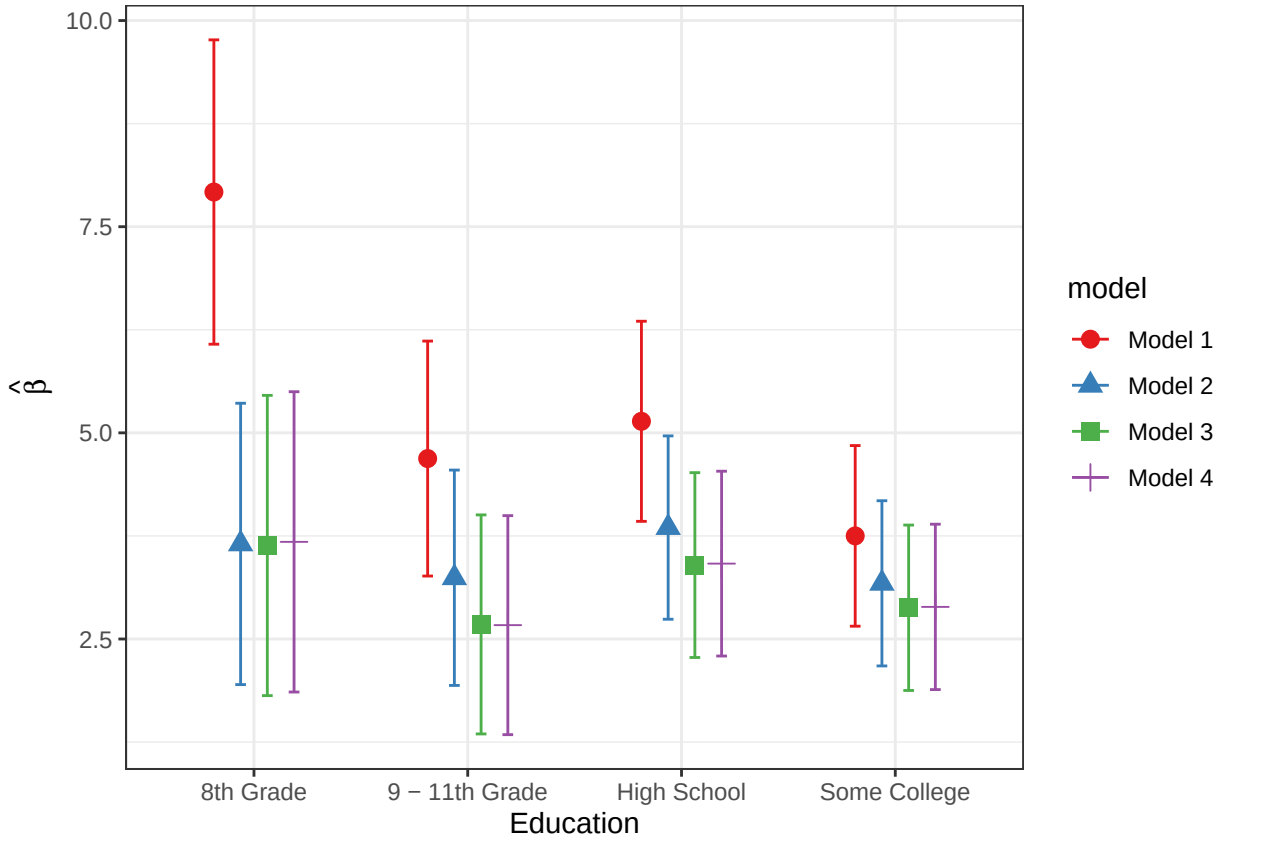
Example using `broom::tidy` and `ggplot`

```
# Comparing models visually ------------------------------------------------

# We want to compare estimates of the education effect in the crude and adjusted models
# Here, I show an example of using broom::tidy to extract the model estimates,
# stacking them together in a tibble,
# filtering out just the education terms,
# and piping the tibble into ggplot in order to plot the estimates.

# Extract the education effects from each model and combine in a tibble
lm_education_estimates <- bind_rows(broom::tidy(lm_model1, conf.int=TRUE) %>%
                                      mutate(model = "Model 1"),
                                    broom::tidy(lm_model2, conf.int=TRUE) %>%
                                      mutate(model = "Model 2"),
                                    broom::tidy(lm_model3, conf.int=TRUE) %>%
                                      mutate(model = "Model 3"),
                                    broom::tidy(lm_model4, conf.int=TRUE) %>%
                                      mutate(model = "Model 4")) %>%
  # here, we use stringr::str_detect to detect the entries
  # where term includes the string 'Education'
  filter(stringr::str_detect(term, "Education")) %>%
  # here, we use the separate() function to pull out the category labels
  # from term so that we can have nice labeling in the plot
  separate(col=term, sep=17, into=c("term", "category"), convert=TRUE)
```

```r
# Use ggplot to plot the point estimates and 95% CIs
# Note that we are differentiating the models by color AND by the shape of the plotting symbol
ggplot(lm_education_estimates, aes(x=category, y=estimate, color=model, shape=model)) +
  # position=position_dodge() is specified so that the estimates are side by side rather than
  # plotted on top of one another
    geom_point(position=position_dodge(0.5), size=3) +
  # geom_errorbar allows us to plot the 95% confidence limits
    geom_errorbar(aes(ymin=conf.low, ymax=conf.high), position=position_dodge(0.5), width=0.2) +
  # scale_color_brewer allows me to control the colors for plotting the different models
    scale_color_brewer(palette="Set1") +
    labs(x="Education", y=expression(hat(beta))) +
    theme_bw()
```



Plotting using `ggstatsplot`

We can use the `ggcoefstats()` function from the `ggstatsplot` package as well.

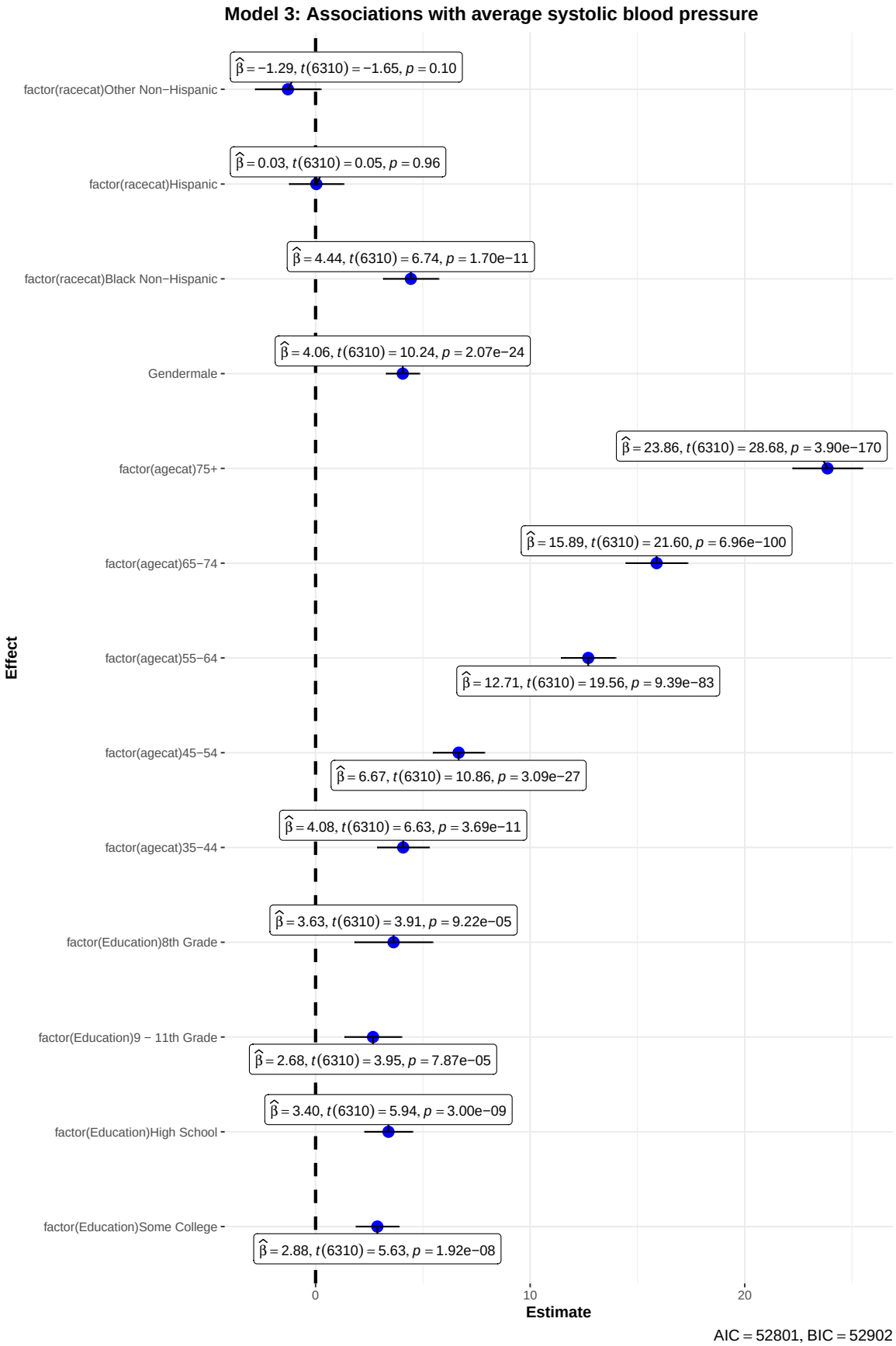
A few things to note here:

- `sort='none'` specifies no sorting of the effects so they appear in the order that they are specified in the `lm()` model call. But note that they are being plotted from the bottom of the plot going up (rather than from the top going down as in `sjPlot::plot_model`)
- To make this figure readable, we adjusted the `fig.width` and `fig.height` chunk options (see Rmd), otherwise everything was too squished.

- `stats.labels` gives us these helpful labels with the actual $\hat{\beta}$ estimates. `ggcoefstats` is using the `ggrepel` package so that these labels repel away from each other and from the data points. Additional arguments can be passed to `ggrepel::geom_label_repel()` via the `stats.label.args` argument.

```
ggcoefstats(lm_model3,
            exclude.intercept = TRUE,
            stats.labels = TRUE,
            xlab = "Estimate",
            ylab = "Effect",
            title = "Model 3: Associations with average systolic blood pressure",
            sort = "none")
```

```
## Number of labels is greater than default palette color count.
## * Select another color `palette` (and/or `package`).
```


Model 3: Associations with average systolic blood pressure
β̂ = −1.29, t(6310) = −1.65, p = 0.10
factor(racecat)Other Non−Hispanic
β̂ = 0.03, t(6310) = 0.05, p = 0.96
factor(racecat)Hispanic
β̂ = 4.44, t(6310) = 6.74, p = 1.70e−11
factor(racecat)Black Non−Hispanic
β̂ = 4.06, t(6310) = 10.24, p = 2.07e−24
Gendermale
β̂ = 23.86, t(6310) = 28.68, p = 3.90e−170
factor(agecat)75+
β̂ = 15.89, t(6310) = 21.60, p = 6.96e−100
factor(agecat)65−74
β̂ = 12.71, t(6310) = 19.56, p = 9.39e−83
factor(agecat)55−64
β̂ = 6.67, t(6310) = 10.86, p = 3.09e−27
factor(agecat)45−54
β̂ = 4.08, t(6310) = 6.63, p = 3.69e−11
factor(agecat)35−44
β̂ = 3.63, t(6310) = 3.91, p = 9.22e−05
factor(Education)8th Grade
β̂ = 2.68, t(6310) = 3.95, p = 7.87e−05
factor(Education)9 − 11th Grade
β̂ = 3.40, t(6310) = 5.94, p = 3.00e−09
factor(Education)High School
β̂ = 2.88, t(6310) = 5.63, p = 1.92e−08
factor(Education)Some College
Effect
0
10
20
Estimate
AIC = 52801, BIC = 52902