Voluntary assignment - example solution with some comments

Jaromír Mazák & Aleš Vomáčka

1. 7. 2021

# How to work with this document

This is only an example solution. It does not mean that your different solution cannot be perfectly adequate, too. The text on grey background is the R code we used to produce our solution. It also includes additional comments which may be of interest.

This is a *short version* which gives you an idea about the extent and depth we expected from this submission. We also produced a *long version* with more detailed consideration of the whole procedure.

(Note: Detailed assignment instructions were provided in a separate document and are not repeated here.)

# STEP 1: Preparation

## Load libraries

# Always keep all libraries used in the script at the top of the script.   
  
library(tidyverse) # for core tidyverse packages including dplyr and ggplot2  
library(here) # for smart path referencing  
library(patchwork) # for arranging graphs next to each other  
library(ggeffects) # for ggpredict, i.e. for plotting marginal effects  
  
options(scipen = 999) # turn off scientific notaiton (produces easier to read tables)

## Make sure you know and keep in mind what the task is

RQ: Is the accuracy of polls (variable *error*) different depending on the methodology of data collection (variable *methodology*)?

## Data import

polls = read.csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/pollster-ratings/raw-polls.csv")  
  
polls %>% count(methodology) %>% arrange(desc(n)) # we only analyse "IVR", "Live Phone", "Online" becuase (a) they are most frequent, (b) the combination catgories would be harder to interpret.

## methodology n  
## 1 Live Phone 5603  
## 2 IVR 2518  
## 3 Online 1586  
## 4 IVR/Online 529  
## 5 IVR/Live Phone 152  
## 6 Online/Text 86  
## 7 Live Phone/Online 84  
## 8 IVR/Online/Text 82  
## 9 IVR/Text 59  
## 10 Text 30  
## 11 Mail 22  
## 12 IVR/Live Phone/Text 7  
## 13 Live Phone/Text 6  
## 14 IVR/Live Phone/Online 4  
## 15 Live Phone/Online/Text 4  
## 16 Face-to-face 3  
## 17 IVR/Live Phone/Online/Text 1

polls = subset(polls, methodology == c("IVR", "Live Phone", "Online"))  
  
polls = as\_tibble(polls)  
  
# You can look at the variables description here - need to uncomment below:  
# browseURL("https://github.com/fivethirtyeight/data/tree/master/pollster-ratings")

# STEP 2: Variable selection

### Dependent variable:

* *error* (“Absolute value of the difference between the actual and polled result.”)

### Main independent variable:

* *methodology* (“Methodology used to conduct this poll. One or more of the following values: … only relevant values for us are *IVR*, i.e. Interactive voice response, otherwise known as automated polls or”robopolls"; *Live Phone*, i.e., Live telephone interviews, may or may not include calls to cell phones ; and *Online*, i.e., Poll conducted by Internet; generally this mean by web browser, or application-based polling of mobile phones … see the filtering in data import section)

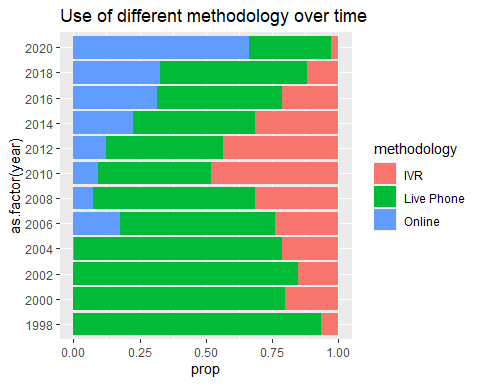
### Other independent variables:

Dependent and main independent variables are given. One important task was “selecting the controls”. Our modeling exercise is about inference: We want to decide which methodology seems to produce the best results. Therefore, we should control for all variables we suspect to be confounders (confounding the true effect of methodology) and avoid controlling for those we suspect may be colliders (colliders are variables which are causally affected by both the independent and the dependent variable).

**List of what we believe are potential confounders:**

* *year* - polling standards may have changed since 1998, so it makes good sense to include it
* *type\_simple* - it differentiates the following US elections: president primary, president general, senators general, house of representatives general, and governors general election. We think it makes good sense to expect that different election types are easier or more difficult to predict. So we should control for it.
* *polldate* and *electiondate* - by themselves, these variables are not very interesting. But it could be very important to consider the difference between the two, i.e. the gap between the polling and the actual election as polling further away from the elections is expected to be less precise in terms of predicting the actual result. We will need to compute a new variable for that.
* *samplesize* - sample size is important for each individual poll’s estimate (specifically, its precision), so we need to include it in the model. Otherwise, if any of the polling methods is associated with different sample sizes than the other two, the results could just be reflecting the different precision of different sample sizes, not the different perfomance of each method.
* *interaction between year and methodology* - Maybe each methodology developed differently over the years in terms of quality. While the same could be said for any other predictor as well (one could argue, e.g., that samplesize may have different effect on error over the years), we find this argument especially compelling for methodology since we see a big shift in methodology composition over time (see the plot below). Online was only born in the period we are looking at and IVR grew to relevance and then became insignificant again. It seems theoretically plausible that the interaction between year and methodology is worth exploring. On the contrary, there seems to be no good tehreotical reason why sample size should have different effect over time, so we are comfortable ignoring that interaction. Also remember that one needs to have a lot of statistical power to detect interactions. So we usually just look at those where we have a good reason to find difference.

# First plot methodology by year (only for years with 50+ polls) to see the difference. Maybe we could make live phone the reference category as it is most stable across the years.  
polls %>% group\_by(year) %>% mutate(year\_n = length(year)) %>% filter(year\_n > 50) %>%   
 group\_by(year, methodology) %>%   
 summarise(n = n()) %>%   
 mutate(prop = n/sum(n)) %>%   
 ggplot(aes(y=as.factor(year), x=prop, fill=methodology))+  
 geom\_col(position = "fill") +  
 labs(title = "Use of different methodology over time")



\*Note that is is fine to consider other confounders as well, as long as you have good theoretical reason to believe they are confounders. It is not like there is only one adequate solution.

**However, we think you should definitely NOT include the following, as they may be colliders:**

* *rightcall* - whether or not the poll called the outcome correctly is likely influenced by the method (provided that any of the three methods really is better or worse, but we assume it is possible or we would not run this analysis). Yet it is also likely influenced by the error (polls with smaller errors are expected to call the outcome correctly more often). When both the dependent and the main independent variable may be reasonably argued to cause a third variable, this third variable should not be included in the model (should not be controlled for).
* *bias* - you should also NOT include this variable in the model for similar reasons as rightcall. Again, polling method can likely influence bias. And bias can influence the value of error - indeed, error is computationally directly derived from the the bias.

# STEP 3: Variables description

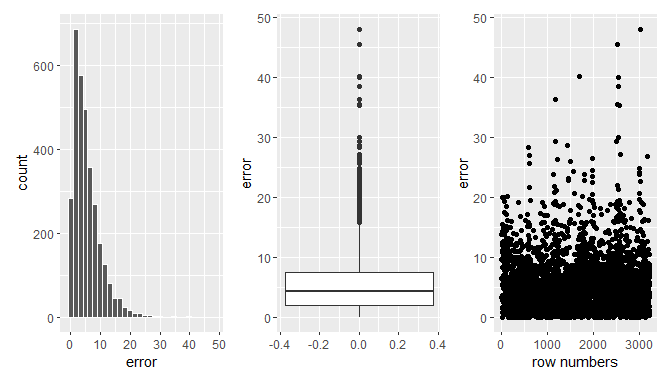
Before we start modeling, we should always pay attention to getting to know our variables. We can do either tabular or graphical description of our variables. Or both.

### Dependent variable

sum(is.na(polls$error)) # no missing values here

## [1] 0

error\_h <-   
 polls %>%   
 ggplot(aes(x=error)) +  
 geom\_histogram(color = "white")  
  
error\_b <-  
 polls %>%   
 ggplot(aes(y=error)) +  
 geom\_boxplot()  
  
error\_p <-  
 polls %>%   
 ggplot(aes(y=error, x=seq\_along(error)))+  
 geom\_point()+  
 labs(x="row numbers")  
  
error\_h + error\_b + error\_p # this patchwork functionality enables to put multiple plot together



As we see above, the *error* variable is skewed with large majority of polls having error between 0 and 10 percentage points. Some errors are extreme. Someone could consider leaving out the cases with extreme errors. That can be fine, but we will proceed with all observations (assuming extreme errors can happen as part of life rather than being glitches in the data).

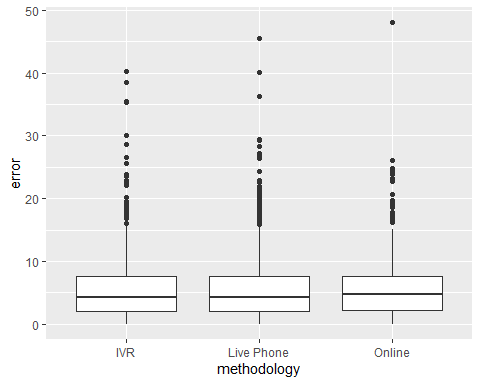
### Main independent variable

In the outputs below, we first look at the frequencies of methodology types in our data. We then look both visually and analytically at the bivariate relationship between the dependent and the main independent variable (i.e., without controlling for anything). There does not seem to be any significant difference among methodologies in the resulting error. However, this is no reason to stop the analysis. Some differences may be “hidden” due to confounders. So we still need to conduct the full analysis with controls to make sure.

polls %>% count(methodology) # no missing values here

## # A tibble: 3 x 2  
## methodology n  
## <chr> <int>  
## 1 IVR 844  
## 2 Live Phone 1876  
## 3 Online 511

# first, we can look at the raw differences of errors across different methodologies without controlling for anything...  
  
# ... visually  
  
polls %>% ggplot(aes(x=methodology, y=error))+  
 geom\_boxplot()



# ... or analytically  
  
m1 <- lm(error ~ methodology, data = polls)  
summary(m1, digits=3)

##   
## Call:  
## lm(formula = error ~ methodology, data = polls)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.613 -3.463 -1.153 2.053 42.427   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.510438 0.166886 33.019 <0.0000000000000002 \*\*\*  
## methodologyLive Phone -0.007389 0.200950 -0.037 0.971   
## methodologyOnline 0.122556 0.271757 0.451 0.652   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.848 on 3228 degrees of freedom  
## Multiple R-squared: 9.279e-05, Adjusted R-squared: -0.0005267   
## F-statistic: 0.1498 on 2 and 3228 DF, p-value: 0.8609

### Other independent variables

Before we look at the individual independent variables, we will check missingness for all of them at one go. This is where your dplyr skills come in handy.

polls %>% select(year, type\_simple, polldate, electiondate, samplesize, partisan) %>%   
 summarise(across(.cols = everything(),  
 .fns = ~sum(is.na(.))))

## # A tibble: 1 x 6  
## year type\_simple polldate electiondate samplesize partisan  
## <int> <int> <int> <int> <int> <int>  
## 1 0 0 0 0 0 0

Oh, this is such non problematic data! No missing values! Thank you, Ales, for finding this dataset.

Next, we start looking at both univariate distribution of each independent variable (it helps us detect where we can expect problems) and bivariate distribution against the dependent variable (it helps as predict which of the variables we control for are likely to have big effect in the final model). However, the bivariate analysis is only indicative, it is no substitute for diagnostics of the final model.

We think the best way to do this is visually. All the code to produce the visual analysis follows. Notice that the variable time\_gap is not part of the dataset, it needs to be constructed.

# year variable  
  
year\_uni <-  
 polls %>% ggplot(aes(x=year))+  
 geom\_bar()  
  
year\_bi <-  
 polls %>% ggplot(aes(x=year, y=error))+  
 geom\_point()+  
 geom\_smooth()  
  
# type\_simple  
  
type\_uni <-  
 polls %>% ggplot(aes(x=type\_simple))+  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))  
  
type\_bi <-   
 polls %>% ggplot(aes(x=type\_simple, y=error))+  
 geom\_boxplot()+  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))  
  
# time\_gap - we first need to calculate time gap between the polls and the actual elections   
  
polls$polldate %>% str() # the variables are character, not date, so we have to transform them before we can subtract them from each other

## chr [1:3231] "10/13/1998" "10/13/1998" "10/13/1998" "10/14/1998" ...

# first, always check the transformations does what you want  
polls %>%   
 mutate(polldate\_d = as.Date(polldate, format = "%m/%d/%Y"),  
 electiondate\_d = as.Date(electiondate, format = "%m/%d/%Y")) %>%   
 select(polldate, polldate\_d, electiondate, electiondate\_d) %>%   
 head()

## # A tibble: 6 x 4  
## polldate polldate\_d electiondate electiondate\_d  
## <chr> <date> <chr> <date>   
## 1 10/13/1998 1998-10-13 11/3/1998 1998-11-03   
## 2 10/13/1998 1998-10-13 11/3/1998 1998-11-03   
## 3 10/13/1998 1998-10-13 11/3/1998 1998-11-03   
## 4 10/14/1998 1998-10-14 11/3/1998 1998-11-03   
## 5 10/14/1998 1998-10-14 11/3/1998 1998-11-03   
## 6 10/14/1998 1998-10-14 11/3/1998 1998-11-03

# only then proceed with the transformation   
  
polls <-   
 polls %>%   
 mutate(polldate\_d = as.Date(polldate, format = "%m/%d/%Y"),  
 electiondate\_d = as.Date(electiondate, format = "%m/%d/%Y"),  
 time\_gap = electiondate\_d - polldate\_d,  
 time\_gap = as.numeric(time\_gap))  
  
time\_gap\_uni <-  
 polls %>% ggplot(aes(y=time\_gap, x=""))+  
 geom\_boxplot()+  
 labs(x = "")  
  
time\_gap\_bi <-  
 polls %>%   
 ggplot(aes(x=time\_gap, y=error))+  
 geom\_point()+  
 geom\_smooth()  
  
# samplesize  
  
# oh, there is extreme skew here...  
samplesize\_uni\_raw <-  
 polls %>% ggplot(aes(y=samplesize, x=""))+  
 geom\_boxplot()+  
 labs(x = "")  
  
# we shoudl consider some transformation such log  
samplesize\_uni\_log <-  
 polls %>% ggplot(aes(y=log10(samplesize),x=""))+  
 geom\_boxplot()+  
 labs(x = "")  
  
samplesize\_bi <-  
 polls %>% ggplot(aes(x=log10(samplesize), y=error))+  
 geom\_point()+  
 geom\_smooth()  
  
  
# partisan   
  
partisan\_uni <-  
 polls %>% ggplot(aes(x=partisan))+  
 geom\_bar()  
  
partisan\_bi <-  
 polls %>% ggplot(aes(x=partisan, y = error))+  
 geom\_boxplot()

The plot composition below has the following logic: one independent variable is displayed in one row. Mostly, there is the univariate distribution to the left and bivariate to the right. For sample size, there is also the log-tranformed graph as the distribution is otherwise too extremely skewed to see anything. The bivarite graph for sample size also uses this log-transformed sample size.

For the *year* variable, it seems that the number of polls per year very much depends on how many and how important elections took place each year, i.e. the election cycle (plot A). However, there does not seem to be any eye-catching relationship between what year the polls took place and what the error was in terms of significant reduction or increase in error over time (plot B).

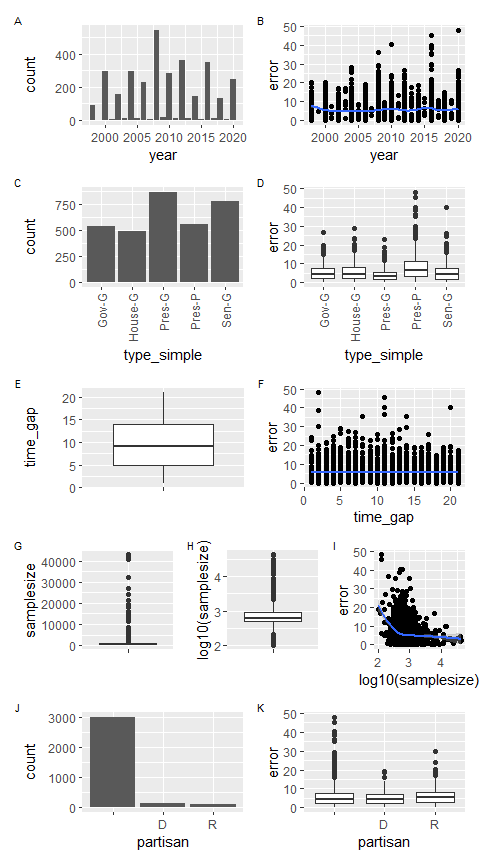
There is between ca 450 and 850 observations for each election type (considering the five types we decided to refer to, plot C). The type of elections seems to be associated with the error at least to some extent, for example, Presidential primaries tend to have the largest error, whereas Presidential general elections the smallest (plot D).

Noteworthy, time gap in all the polls in the sample is small, between 0 and 20 days or so. In addition, there seems to be no big relationship between the time gap between the polls and actual elections on the one hand and the error on the other hand (F)

Sample size unsurprisingly plays some role, but in accordance with statistical theory, the relationship with error is not linear (plot I). The error is significantly bigger for very small samples, but goes down quickly and levels off. This could be a hint that we should use some tool to model non-linearity, but not necessarily, we cannot know ahead what will happen in the multidimensional world of multiple linear regression.

Partisan polls are rare (plot J) and their effect on error seems to be small if any. The error for Republican commissioned polls seems slightly higher than the rest of the polls (plot K).

(year\_uni + year\_bi) / (type\_uni + type\_bi) / (time\_gap\_uni + time\_gap\_bi) /   
 (samplesize\_uni\_raw + samplesize\_uni\_log + samplesize\_bi) /  
 (partisan\_uni + partisan\_bi) +  
 plot\_annotation(tag\_levels = 'A') &   
 theme(plot.tag = element\_text(size = 8))



# STEP 4: Model presentation and interpretation

In this short version of the solution, we just run one model based on our theoretical consideration in the section on variable selection. Since the model includes an interaction, it is more useful to run it on centered predictors. However, since we also want to use plots of marginal effects for the presentation of our model, it is useful to also run the same model on raw (non-centered variables) - plotting of marginal effects becomes more straightforward.

The code below first performs the necessary transformations. Then, there is the summary of the centered model. We notice some statistically significant relationships (other than statistically significant differences can probably be disregarded straight away as we have a really big sample), but we also know not to spend too much time looking in the table: a more complex model is much more easily interpreted using marginal effect plots. We still may notice that the interaction coefficients for IVR and Online methodology with year are positive (0.15 and 0.13 respectively) and statistically significant, which means that both these methodologies get relatively worse in predicting the election outcome compared to the reference category (Live Phone). Note that this does not mean they are necessarily getting worse per se. It might be that they are getting better, but not as quickly as Live Phones. We don’t know this simply from looking at the coefficients. This is what we mean by the words ‘relatively worse’.

We actually do not need to agonize too much about the other coefficients in the tables, they are just controls instrumental for answering our main research question. Yet it may be interesting to notice that there is relatively bigger error in the Presidental primaries and, in contrast, smaller error in Presidential general elections. In addition, time gap, does, in fact, increases the error (it did not look so in the bivariate analysis), other things in the model being equal: polls taken 1 day further away from the election produce 0.06 percentage points worse estimates than comparable polls taken 1 day closer to the elections. In other words, 10 days difference correspondes to 0.6 percentage point difference in error (within the period we look at, this is, of course, something we should not be tempted to generalize beyond - it is unlikely that polls 1000 days from the election will have 60 % error on average). Sample size decreases error.

# Changing reference category in methodology.   
  
polls <-  
 polls %>% mutate(methodology = fct\_relevel(methodology, "Live Phone")) # set as first level, i.e. reference category, since it is the most prevalent category significant across the whole time period we analyze  
  
  
# Centering and logging sample size.  
  
polls <-  
 polls %>% mutate(  
 year\_c = year - mean(year),  
 time\_gap\_c = time\_gap - mean(time\_gap),  
 samplesize\_log = log10(samplesize),  
 samplesize\_log\_c = log10(samplesize) - mean(log10(samplesize)))  
  
model <- lm(error ~ methodology\*year + type\_simple + time\_gap + samplesize\_log + partisan, data = polls)  
model\_c <- lm(error ~ methodology\*year\_c + type\_simple + time\_gap\_c + samplesize\_log\_c + partisan, data = polls)  
  
  
# we only run summary on the centered model, because that is the one where we can meaningfully interpret the main effects  
summary(model\_c)

##   
## Call:  
## lm(formula = error ~ methodology \* year\_c + type\_simple + time\_gap\_c +   
## samplesize\_log\_c + partisan, data = polls)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.111 -3.049 -0.915 2.103 37.036   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.95114 0.21598 22.924 < 0.0000000000000002 \*\*\*  
## methodologyIVR 0.30834 0.19678 1.567 0.117230   
## methodologyOnline 0.63038 0.35034 1.799 0.072062 .   
## year\_c -0.02733 0.01769 -1.545 0.122533   
## type\_simpleHouse-G 0.19426 0.29237 0.664 0.506466   
## type\_simplePres-G -1.07129 0.25608 -4.183 0.00002949160624 \*\*\*  
## type\_simplePres-P 2.94085 0.29149 10.089 < 0.0000000000000002 \*\*\*  
## type\_simpleSen-G 0.02777 0.25829 0.108 0.914395   
## time\_gap\_c 0.06087 0.01532 3.974 0.00007231078010 \*\*\*  
## samplesize\_log\_c -2.57015 0.37231 -6.903 0.00000000000609 \*\*\*  
## partisanD -0.06240 0.42864 -0.146 0.884263   
## partisanR 0.78185 0.45483 1.719 0.085711 .   
## methodologyIVR:year\_c 0.14711 0.03888 3.784 0.000157 \*\*\*  
## methodologyOnline:year\_c 0.12672 0.04785 2.649 0.008121 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.596 on 3217 degrees of freedom  
## Multiple R-squared: 0.1045, Adjusted R-squared: 0.1009   
## F-statistic: 28.89 on 13 and 3217 DF, p-value: < 0.00000000000000022

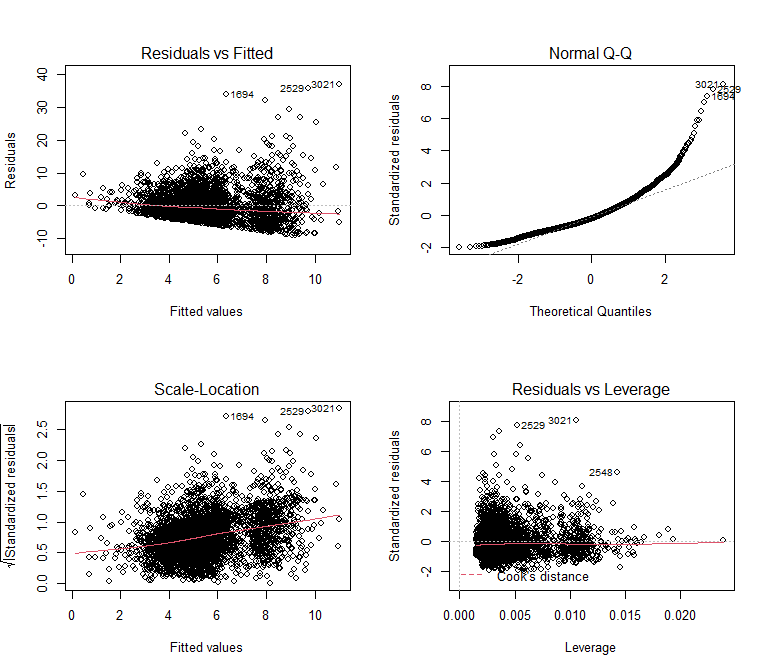
Before we look at the marginal effect plot for interpretation which is easier to grasp, let’s check the assumptions using diagnostic plots.

# STEP 5: Model fit and diagnostics

We already could assess model fit in the summary above. With R-squared of 0.1, we know the predictive power of the model is modest, but not negligible. Knowing the values of all variables in the model, we predict ca 10 % of the variance in error across the models.

Next, we should see the diagnostic plot.

par(mfrow = c(2,2)) #adjusts how plot are displayed in the plots pane of R  
# you can reset the above to default by either restarting R Studio or running par(mfrow = c(1,1))  
plot(model)



par(mfrow = c(1,1))

How can we interpret them?

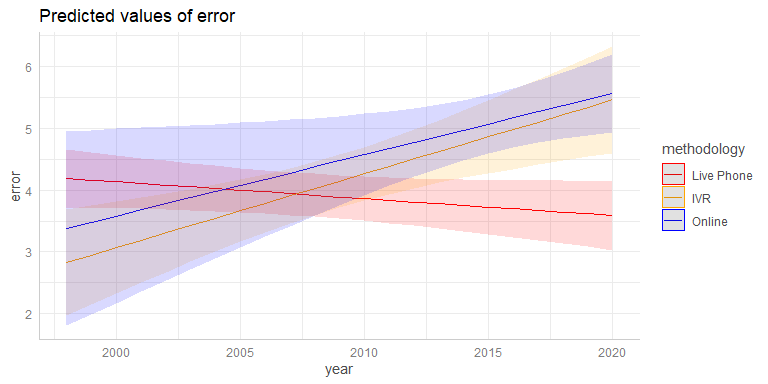
* the first (residuals vs fitted) is mainly used to assess linearity of the modeled relationships (can we reasonably model the dependent variable as a linear combination of our predictors?) - since the red line does not deviate much from the dotted and is straight rather than wiggly, we can say that our specification of the model does not violate the linearity assumption too much.
* the second (normal Q-Q) is used to assess the normality of residuals. We see some serious departure from normality (points should be around the dotted line if the assumption holds), so this assumption is violated. However, remember that this assumption is actually comparatively least important and is only relevant if we need to make predictions on individual level (for individual polls) or when we use small sample sizes for inference. In case we are interested in predicting individual cases, we could try to make separate groups for different sample sizes or some transformations until we are satisfied. Since we are not interested in such exercise here, we will just let it be.
* the third (scale-location) is used to assess the assumption of homoscedasticity. Is the variance of residuals similar around across different levels of fitted values? This is a bit hard to say as the number of observations is very different across different levels of fitted values. Still it looks to us that there is or could be some funneling out of the errors indicating heteroscedasticity. Practically, this means that the p-values for the regression coefficients may be biased (i.e. incorrect p-values, incorrect confidence intervals). (If you want to, you can learn about robust standard errors and use them. We did not cover robust standard errors in the classes, but we provide some material on the course web page. However, since the sample size is large, we are not much worried about probably just a small inaccuracy in p-values.) Especially given out intention to mainly intepret the model visually.
* the forth (residuals vs leverage) shows values with large Cook’s distance. With our sample size, it is quite unlikely we would see any significant distortion of the estimated coefficients due to a few influential cases, but we can still have a look at those which exert most influence. Anyway, the good news is that the cases with big leverage are close to the dotted line - do not exert much influence on the model estimates.

# STEP 6: Visual model presentation and interpretation

First, we just look at marginal effects of different methodologies. Remember, marginal effects are expected values of the dependent variable for given values of selected independent variable while other independent variables are held constant.

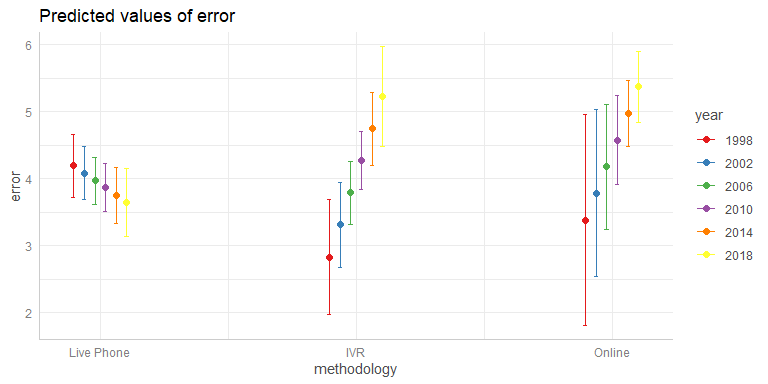
The plot gives a fairly straightforward idea of what is going on: It is not like one of the methodologies has always been better, but there seems to be significant difference in their performance today: Live Phone methodology has produced slightly decreasing errors over time, while both Online methodology and IVR have experienced increase in errors. We should also remind the reader that online actually only appeared in 2006, so anything before that on the plot is nonsensical extrapolation. See, there is value in taking things slowly and producing many different exploratory charts to understand the data underlying a model. Going straight to modeling may be compelling, but it is prone to misinterpretations.

model\_marginal\_effect <- plot(ggpredict(model, terms = c("year", "methodology")), colors = c("red", "orange", "blue"))   
  
model\_marginal\_effect



Alternatively, you could prefer to plot methodology on x axis (as the main independent variable) and show year with color. In that case, it is advisable to specify in code for which values of year you want to display the results. We get a similar conclusion - if choosing your methodology one today, Live Phone seems preferable all other things being equal.

model\_marginal\_effect\_alt <- plot(ggpredict(model, terms = c("methodology", "year [1998, 2002, 2006, 2010, 2014, 2018]")))   
  
model\_marginal\_effect\_alt



But what if you are on a budget and can either have 300 people interviewed using live phone or 1000 using an online survey? There are multiple ways to do this, but we can also stay in the framework of ggpredict. From the plot below, we see that predicted error for 300 respondents in a Live Phone survey is over 8 while it is below 8 for Online with 1000 respondents. In addition, the online survey has narrow confidence intervals, so all the more reason to opt for it under such conditions.

# we used log sample size, so we need to specify values of sample size as logs, too.  
model\_marginal\_effect\_ss <- plot(ggpredict(model, terms = c("methodology", "samplesize\_log [log10(300), log10(1000)]", "year [2020]")))   
  
model\_marginal\_effect\_ss

