Problem set 3

Jingyuan Zhou

5/13/2017

Regression diagnostics

```
data <- na.omit(data_all)
data$index <-c(1: nrow(data))

rd_mod <- lm(biden ~ age + female + educ, data = data)
tidy(rd_mod)
```

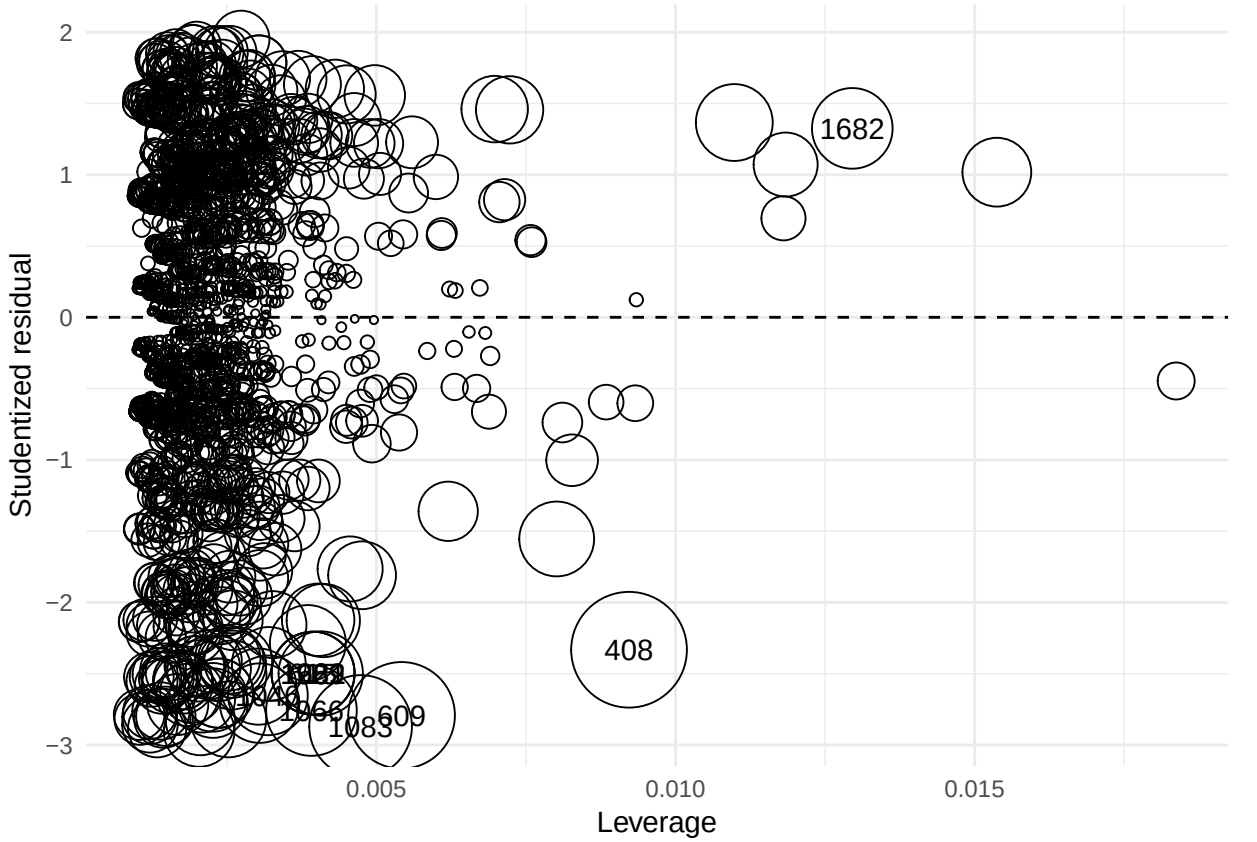
```
##          term estimate std.error statistic  p.value
## 1 (Intercept)  68.6210    3.5960     19.08 4.34e-74
## 2         age   0.0419    0.0325      1.29 1.98e-01
## 3      female   6.1961    1.0967      5.65 1.86e-08
## 4        educ  -0.8887    0.2247     -3.96 7.94e-05
```

```
glance(rd_mod)
```

```
##   r.squared adj.r.squared sigma statistic  p.value df logLik   AIC   BIC
## 1    0.0272        0.0256  23.2      16.8 8.88e-11  4  -8240 16491 16518
##   deviance df.residual
## 1   967076        1803
```

1. Test the model to identify any unusual and/or influential observations. Identify how you would treat these observations moving forward with this research. Note you do not actually have to estimate a new model, just explain what you would do. This could include things like dropping observations, respecifying the model, or collecting additional variables to control for this influential effect.

```
# add key statistics
b_augment <- data %>%
  mutate(hat = hatvalues(rd_mod),
         student = rstudent(rd_mod),
         cooksd = cooks.distance(rd_mod))

# draw bubble plot
ggplot(b_augment, aes(hat, student)) +
  geom_hline(yintercept = 0, linetype = 2) +
  geom_point(aes(size = cooksd), shape = 1) +
  geom_text(data = b_augment %>%
              arrange(-cooksd) %>%
              slice(1:10),
            aes(label = index)) +
  scale_size_continuous(range = c(1, 20)) +
  labs(x = "Leverage",
       y = "Studentized residual") +
  theme(legend.position = "none")
```



By combining all three variables into a "bubble plot", we can visualize all three variables simultaneously. Each observation's leverage (h_i) is plotted on the axis. Each observation's discrepancy (i.e. Studentized residual) is plotted on the axis. Each symbol is drawn proportional to the observation's Cook's D_i

The bubble plot shows: 1. The size/color of the symbols is proportional to Cook's D, which is in turn a multiplicative function of the square of the Studentized residuals (Y axis) and the leverage (X axis), so observations farther away from and/or have higher values of will have larger symbols. 2. The plot tells us whether the large influence of an observation is due to high discrepancy, high leverage, or both

1682th observation has high leverage and low discrepancy 408th observation has relatively high leverage and high discrepancy 1066th, 609th and 1083th observations have low leverage but very high discrepancy

```
out <- data %>%
  filter((index %in% c(1682, 408)))
out
```

```
##   biden female age educ dem rep index
## 1    20      1  58    4   0   1   408
## 2   100      0  85    3   1   0  1682
```

```
data_omit <- data %>%
  filter(!(index %in% c(1682, 408)))

rd_mod_omit <- lm(biden ~ age + female + educ, data = data_omit)
tidy(rd_mod_omit)
```

```
##          term estimate std.error statistic  p.value
## 1 (Intercept)  68.9085    3.6092     19.09 3.75e-74
## 2         age   0.0404    0.0325      1.24 2.14e-01
```

```
## 3      female   6.2909    1.0958      5.74 1.10e-08
## 4        educ  -0.9078    0.2263     -4.01 6.28e-05
```

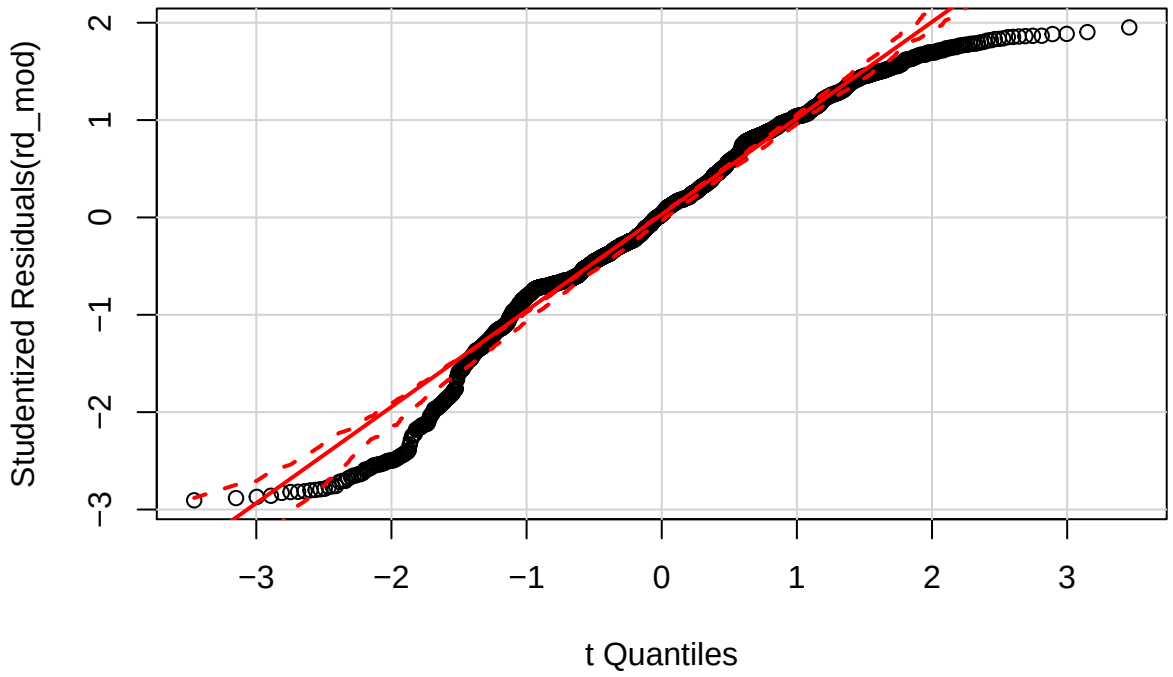
```
glance(rd_mod_omit)
```

```
##   r.squared adj.r.squared sigma statistic  p.value df logLik   AIC   BIC
## 1    0.0279        0.0263  23.1      17.3 4.81e-11  4  -8229 16467 16495
##   deviance df.residual
## 1   963255        1801
```

Just by looking at these two outliers, it's hard to explain "why is it so strange?". Since we only have two outliers here, it doesn't seem like out model is wrong. After fitting a new model by removing the two observations, we do observe a significant decrease of the estimate of the coefficient of "female". This indicates that we should remove them.

2. Test for non-normally distributed errors. If they are not normally distributed, propose how to correct for them.

```
car::qqPlot(rd_mod)
```



The above figure is a quantile-comparison plot, graphing for each observation its studentized residual on the yy axis and the corresponding quantile in the tt-distribution on the xx axis. The dashed lines indicate 95% confidence intervals calculated under the assumption that the errors are normally distributed. If any observations fall outside this range, this is an indication that the assumption has been violated. Clearly, here that is the case.

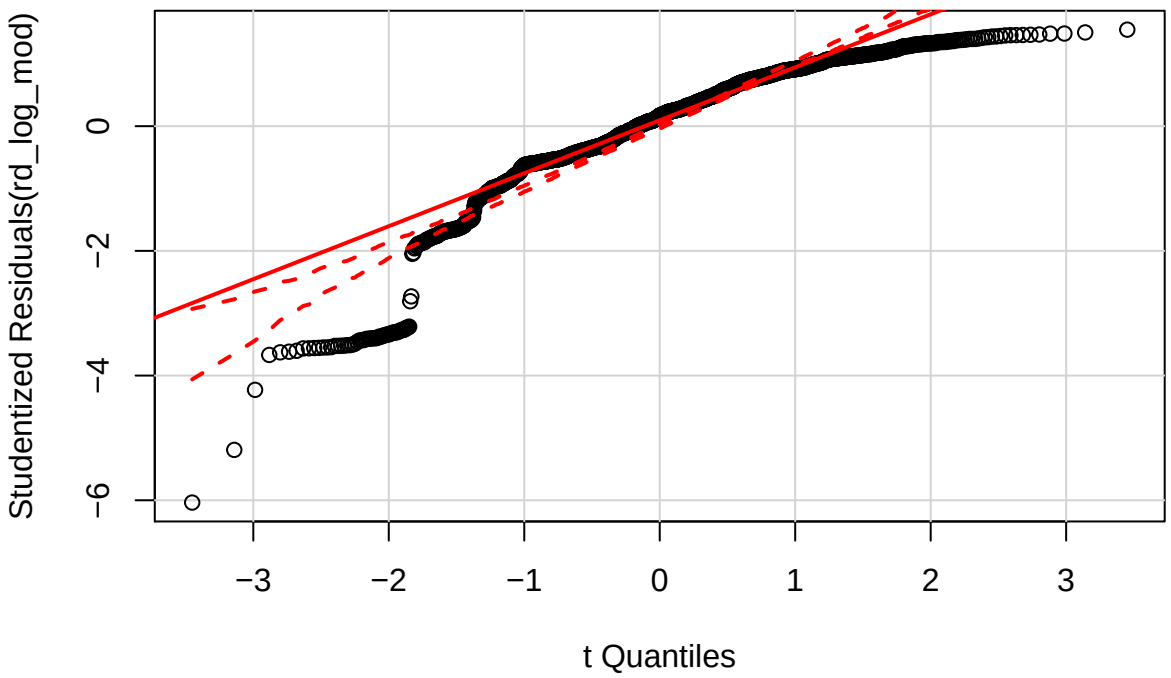
To fix the non-normally distributed errors, we can use log transformation to correct them.

```
data_omit$biden_log = log(data_omit$biden)

data_omit <- data_omit %>%
  filter(!(biden_log %in% c(-Inf, Inf)))
data_omit <- na.omit(data_omit)
```

```
rd_log_mod <- lm(biden_log ~ age + female + educ, data = data_omit)
tidy(rd_log_mod)
```

```
##          term  estimate std.error statistic  p.value
## 1 (Intercept)  4.197646  0.063646     65.95 0.00e+00
## 2         age  0.000719  0.000576      1.25 2.12e-01
## 3      female  0.083874  0.019394      4.32 1.61e-05
## 4        educ -0.013986  0.003982     -3.51 4.56e-04
```

```
car::qqPlot(rd_log_mod)
```



However, as we can observe from the transformed quantile-comparison plot, even though it's more symmetric than before, it does not entirely solve this problem.

3. Test for heteroscedasticity in the model. If present, explain what impact this could have on inference.

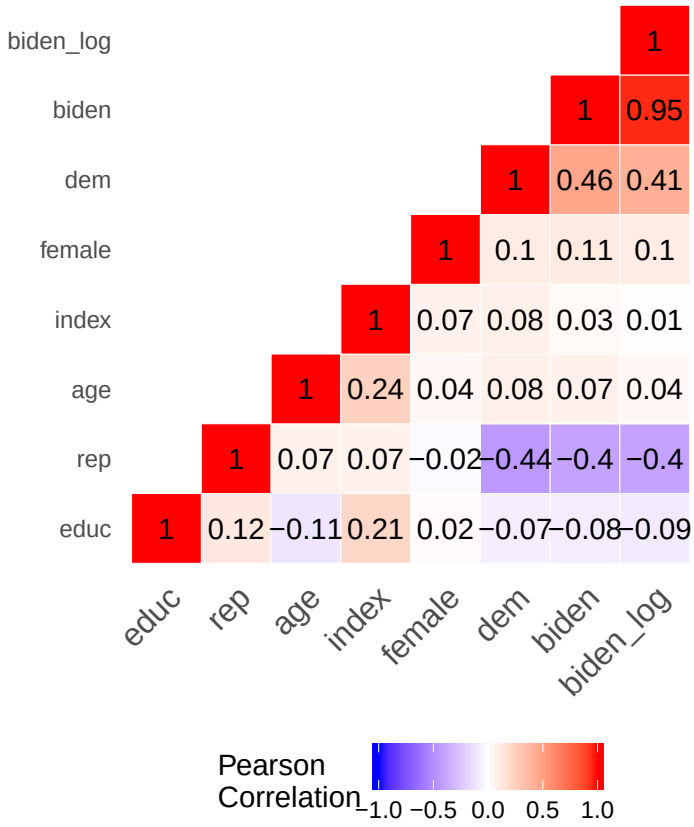
```
bptest(rd_mod)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  rd_mod
## BP = 20, df = 3, p-value = 5e-05
```

We observe from the Breusch-Pagan test that p-value is significantly smaller than 0.05. This shows that we have enough evidence to reject the null hypothesis. Thus, heteroscedasticity is present. By assuming the variance is constant, we substantially over or underestimate the actual response of biden score as covariates increase.

4. Test for multicollinearity. If present, propose if/how to solve the problem.

```
cormat_heatmap <- function(data){
  # generate correlation matrix
  cormat <- round(cor(data), 2)

  # melt into a tidy table
  get_upper_tri <- function(cormat){
    cormat[lower.tri(cormat)]<- NA
    return(cormat)
  }

  upper_tri <- get_upper_tri(cormat)

  # reorder matrix based on coefficient value
  reorder_cormat <- function(cormat){
    # Use correlation between variables as distance
    dd <- as.dist((1-cormat)/2)
    hc <- hclust(dd)
    cormat <-cormat[hc$order, hc$order]
  }

  cormat <- reorder_cormat(cormat)
  upper_tri <- get_upper_tri(cormat)

  # Melt the correlation matrix
  melted_cormat <- reshape2::melt(upper_tri, na.rm = TRUE)

  # Create a ggheatmap
  ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
    geom_tile(color = "white")+
    scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                         midpoint = 0, limit = c(-1,1), space = "Lab",
                         name="Pearson\nCorrelation") +
    theme_minimal()+ # minimal theme
    theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                     size = 12, hjust = 1))+
    coord_fixed()

  # add correlation values to graph
  ggheatmap +
    geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
    theme(
      axis.title.x = element_blank(),
      axis.title.y = element_blank(),
      panel.grid.major = element_blank(),
      panel.border = element_blank(),
      panel.background = element_blank(),
      axis.ticks = element_blank(),
      legend.position = "bottom")
}

cormat_heatmap(select_if(data_omit, is.numeric))
```



There's not much collinearity in the data because all correlations are strictly less than 0.5.

Interaction term

```
it_mod <- lm(biden ~ age*educ, data = data_omit)
tidy(it_mod)
```

```
##          term estimate std.error statistic  p.value
## 1 (Intercept)  43.4865    8.7262      4.98 6.87e-07
## 2         age   0.5706    0.1569      3.64 2.85e-04
## 3        educ   1.2995    0.6523      1.99 4.65e-02
## 4    age:educ  -0.0378    0.0119     -3.18 1.51e-03
```

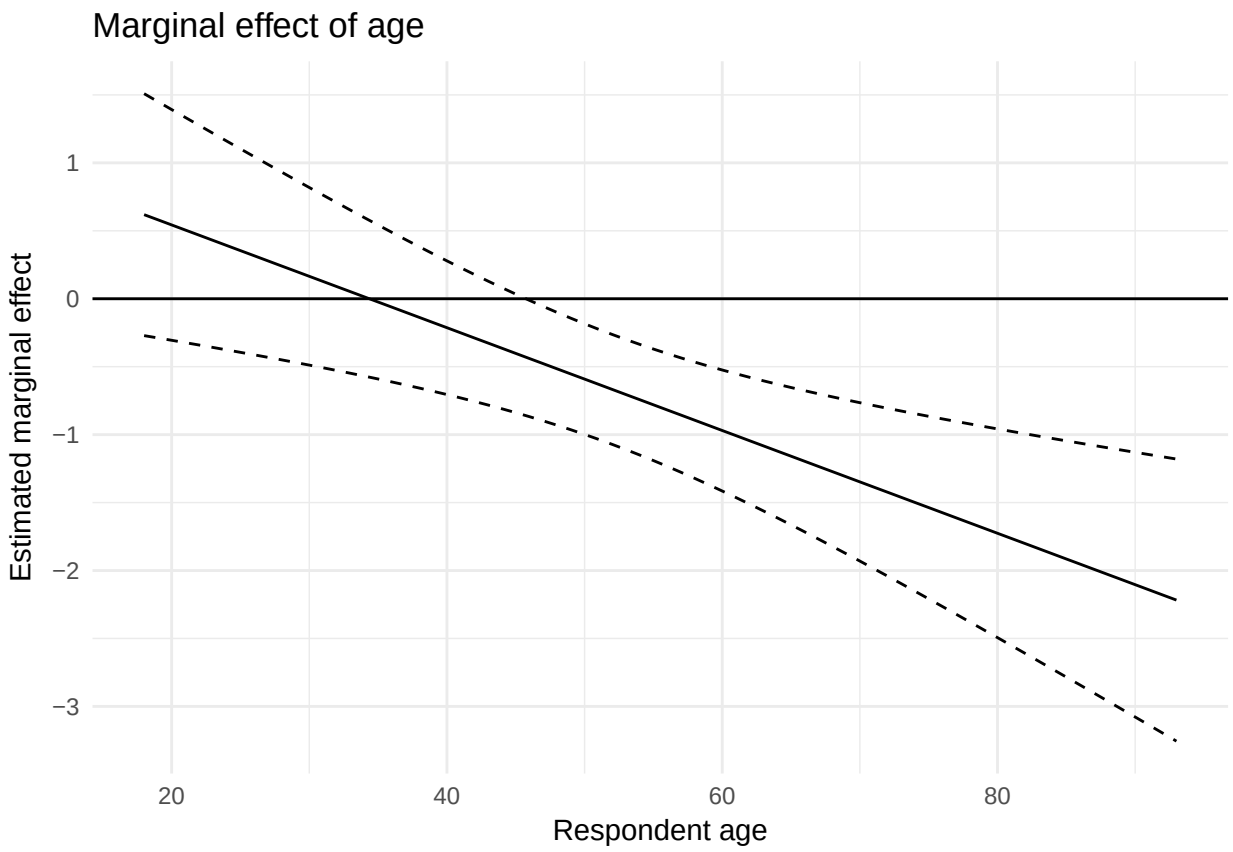
```
glance(it_mod)
```

```
##   r.squared adj.r.squared sigma statistic  p.value df logLik   AIC   BIC
## 1     0.017        0.0153  20.8        10 1.47e-06  4  -7796 15602 15629
##   deviance df.residual
## 1   758612        1746
```

1. Evaluate the marginal effect of age on Joe Biden thermometer rating, conditional on education. Consider the magnitude and direction of the marginal effect, as well as its statistical significance.

```
# function to get point estimates and standard errors
# model - lm object
# mod_var - name of moderating variable in the interaction
instant_effect <- function(model, mod_var){
  # get interaction term name
  int.name <- names(model$coefficients)[[which(str_detect(names(model$coefficients), ":"))]]

  marg_var <- str_split(int.name, ":")[[1]][[which(str_split(int.name, ":")[[1]] != mod_var)]]

  # store coefficients and covariance matrix
  beta.hat <- coef(model)
  cov <- vcov(model)

  # possible set of values for mod_var
  if(class(model)[[1]] == "lm"){
    z <- seq(min(model$model[[mod_var]]), max(model$model[[mod_var]]))
  } else {
    z <- seq(min(model$data[[mod_var]]), max(model$data[[mod_var]]))
  }

  # calculate instantaneous effect
  dy.dx <- beta.hat[[marg_var]] + beta.hat[[int.name]] * z

  # calculate standard errors for instantaeous effect
  se.dy.dx <- sqrt(cov[marg_var, marg_var] +
                     z^2 * cov[int.name, int.name] +
                     2 * z * cov[marg_var, int.name])

  # combine into data frame
  data_frame(z = z,
             dy.dx = dy.dx,
             se = se.dy.dx)
}

# point range plot
# instant_effect(it_mod, "age") %>%
#   ggplot(aes(z, dy.dx,
#              ymin = dy.dx - 1.96 * se,
#              ymax = dy.dx + 1.96 * se)) +
#   geom_pointrange() +
#   geom_hline(yintercept = 0, linetype = 2) +
#   labs(title = "Marginal effect of age",
#        #subtitle = "By respondent conservatism",
#        x = "Respondent age",
#        y = "Estimated marginal effect")

# line plot
instant_effect(it_mod, "age") %>%
```

```
  ggplot(aes(z, dy.dx)) +
  geom_line() +
  geom_line(aes(y = dy.dx - 1.96 * se), linetype = 2) +
  geom_line(aes(y = dy.dx + 1.96 * se), linetype = 2) +
  geom_hline(yintercept = 0) +
  labs(title = "Marginal effect of age",
       #subtitle = "By respondent conservatism",
       x = "Respondent age",
       y = "Estimated marginal effect")
```




```
linearHypothesis(it_mod, "age + age:educ")
```

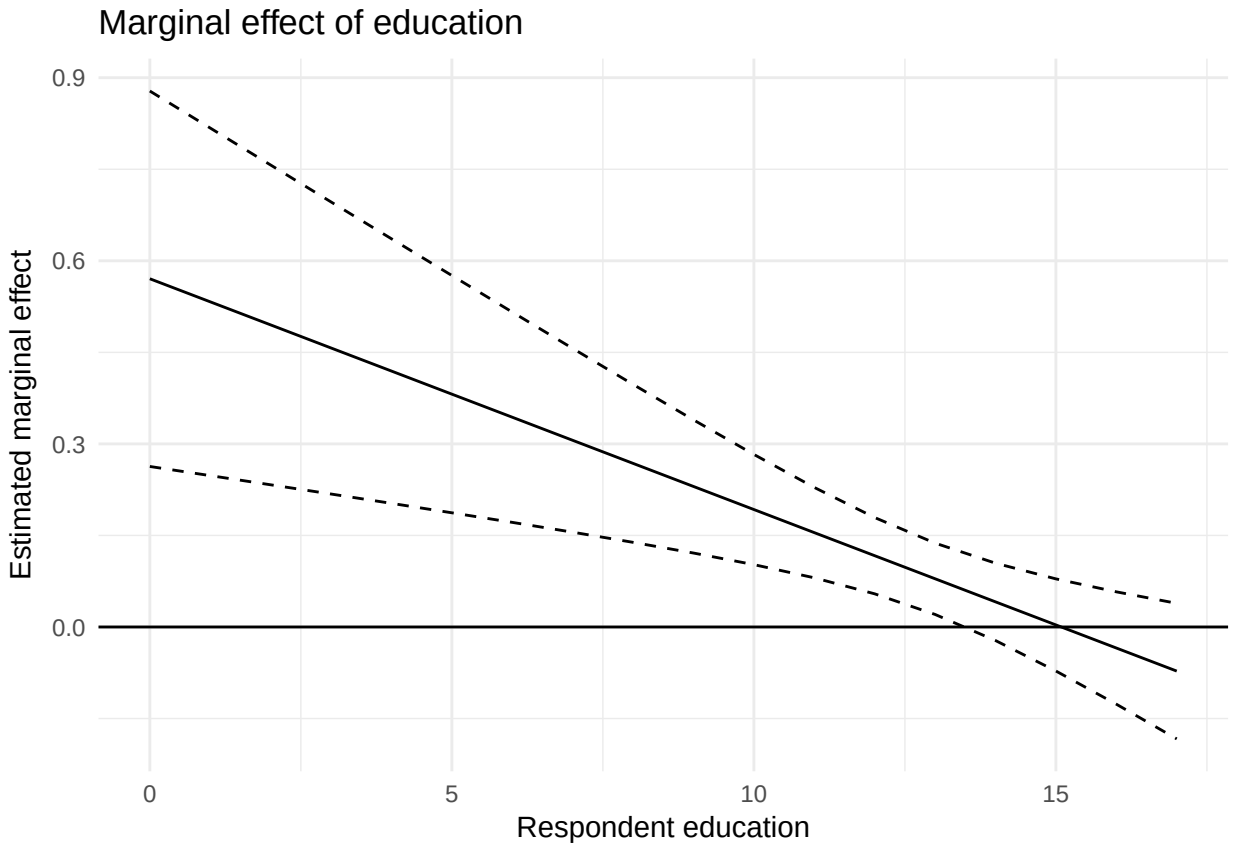
```
## Linear hypothesis test
##
## Hypothesis:
## age  + age:educ = 0
##
## Model 1: restricted model
## Model 2: biden ~ age * educ
##
##   Res.Df    RSS Df Sum of Sq    F  Pr(>F)
## 1   1747 764455
## 2   1746 758612  1      5843 13.4 0.00025 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the result of hypothesis testing, we can see that the marginal effect of age is statistically significant. As

age of respondent increase, the marginal effect of age decreases from 0.6 to almost -2.1. The 95% confidence interval is shown in the graph.

2. Evaluate the marginal effect of education on Joe Biden thermometer rating, conditional on age. Consider the magnitude and direction of the marginal effect, as well as its statistical significance.

```
instant_effect(it_mod, "educ") %>%
  ggplot(aes(z, dy.dx)) +
  geom_line() +
  geom_line(aes(y = dy.dx - 1.96 * se), linetype = 2) +
  geom_line(aes(y = dy.dx + 1.96 * se), linetype = 2) +
  geom_hline(yintercept = 0) +
  labs(title = "Marginal effect of education",
       #subtitle = "By respondent conservatism",
       x = "Respondent education",
       y = "Estimated marginal effect")
```




```
linearHypothesis(it_mod, "educ + age:educ")
```

```
## Linear hypothesis test
##
## Hypothesis:
## educ  + age:educ = 0
```

```
##
## Model 1: restricted model
## Model 2: biden ~ age * educ
##
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1   1747 760295
## 2   1746 758612  1      1683 3.87  0.049 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the result of hypothesis testing, we can see that the marginal effect of education is statistically significant. As education level of respondent increase, the marginal effect of age decreases from 0.6 to almost -0.1. The 95% confidence interval is shown in the graph.

Missing Data

This time, use multiple imputation to account for the missingness in the data. Consider the multivariate normality assumption and transform any variables as you see fit for the imputation stage. Calculate appropriate estimates of the parameters and the standard errors and explain how the results differ from the original, non-imputed model.

```
data_all$index <-c(1: nrow(data_all))
data_all_omit <- data_all %>%
  filter(!(index %in% c(1682, 408)))
#md_mod <- lm(log(biden) ~ age + female + educ, data = data_all)
#tidy(md_mod)
#glance(md_mod)

md_out <- amelia(as.data.frame(data_all_omit), m = 5)
```

```
## -- Imputation 1 --
##
##   1  2  3  4  5  6
##
## -- Imputation 2 --
##
##   1  2  3  4  5
##
## -- Imputation 3 --
##
##   1  2  3  4  5
##
## -- Imputation 4 --
##
##   1  2  3  4  5  6
##
## -- Imputation 5 --
##
##   1  2  3  4  5  6
```

```
models_imp <- data_frame(data = md_out$imputations) %>%
  mutate(model = map(data, ~ lm(biden ~ age + female + educ,
                                data = .x)),
         coef = map(model, tidy)) %>%
```

```
  unnest(coef, .id = "id")

mi.meld.plus <- function(df_tidy){
  # transform data into appropriate matrix shape
  coef.out <- df_tidy %>%
    select(id:estimate) %>%
    spread(term, estimate) %>%
    select(-id)

  se.out <- df_tidy %>%
    select(id, term, std.error) %>%
    spread(term, std.error) %>%
    select(-id)

  combined.results <- mi.meld(q = coef.out, se = se.out)

  data_frame(term = colnames(combined.results$q.mi),
             estimate.mi = combined.results$q.mi[1, ],
             std.error.mi = combined.results$se.mi[1, ])
}

# compare results
tidy(rd_mod) %>%
  left_join(mi.meld.plus(models_imp)) %>%
  select(-statistic, -p.value)
```

```
##          term estimate std.error estimate.mi std.error.mi
## 1 (Intercept)  68.6210    3.5960      65.588       3.4584
## 2         age   0.0419    0.0325       0.056       0.0288
## 3      female   6.1961    1.0967       5.600       0.9976
## 4        educ  -0.8887    0.2247      -0.706       0.2246
```

Comparing the new model with the original model, we observe that there is slight increase in the estimate of the coefficient of age, decrease in the estimate of the coefficent of female and increase in the estimate of the coefficient of education. Changes in standard error can almost be ignored. This might due to the fact that there weren't many missing values in this dataset.