

Problem Set 4

Ayush Pundyavana

```
knitr::opts_chunk$set(  
  echo = TRUE,  
  results = 'markup',  
  tidy = TRUE,  
  comment = NA,  
  width = 60, # wrap R output lines  
  max.print = 100 # limit huge outputs  
)  
options(width = 60) # ensures printed output wraps too
```

```
library(dplyr)
```

Warning: package 'dplyr' was built under R version 4.4.3

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

intersect, setdiff, setequal, union

```
library(ggplot2)
```

Warning: package 'ggplot2' was built under R version 4.4.3

```
library(sandwich)
```

Warning: package 'sandwich' was built under R version 4.4.3

```
library(lmtest)
```

Warning: package 'lmtest' was built under R version 4.4.3

Loading required package: zoo

Warning: package 'zoo' was built under R version 4.4.3

Attaching package: 'zoo'

The following objects are masked from 'package:base':

as.Date, as.Date.numeric

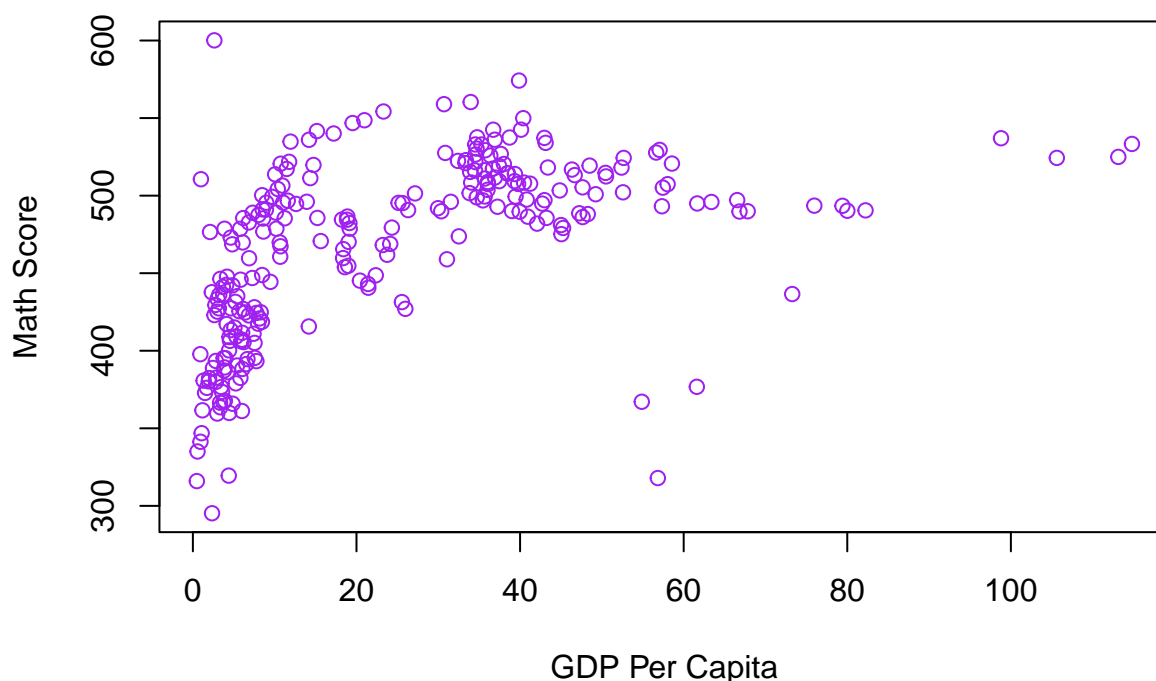
#Question 1

```
data1.1 <- read.csv("grades_and_temps.csv")  
data1.2 <- data1.1 %>% mutate(gdppc1k = round(gdppc/1000, 3))
```

1. Generate a scatter plot with GDP per capita in thousands of dollars (gdppc1k) on the x-axis and the average math score (math score) on the y-axis. Using visual inspection, do these variables seem to be positively correlated, negatively correlated, or not correlated at all? Does their relationship seem linear or non-linear?

```
plot(x = data1.2$gdppc1k, y = data1.2$math_score,  
     main = "Math Score based on GDP Per Capita ",  
     xlab = "GDP Per Capita",  
     ylab = "Math Score",  
     col = "purple"  
)
```

Math Score based on GDP Per Capita



```
cat("Using visual inspection of the scatterplot, the relationship between GDP  
Per Capita and Math Score seems non-linear since the rate of change in the  
Math score as GDP Per Capita increases seems to be non-constant.")
```

Using visual inspection of the scatterplot, the relationship between GDP Per Capita and Math Score seems non-linear since the rate of change in the Math score as GDP Per Capita increases seems to be non-constant.

2. Estimate the following two Equations:

$$\text{math score}_i = a + b \text{gdppc1k}_i + \varepsilon_i \quad (1) \quad \text{math score}_i = \alpha + \beta_1 \text{gdppc1k}_i + \beta_2 \text{gdppc1k2}_i + \varepsilon_i \quad (2)$$

Report the estimated coefficients \hat{b} , $\hat{\beta}_1$ and $\hat{\beta}_2$, along with their heteroskedasticity-robust standard errors for each equation. Determine whether the relationship between math scores and GDP per capita is linear (as opposed to quadratic).

```
#Linear model
lin_model <- lm(math_score ~ gdppc1k, data = data1.2)
vcov_m_g <- vcovHC(lin_model, type = "HC3")
coeftest(lin_model, vcov. = vcov_m_g)
```

t test of coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 432.9164      4.9696 87.1136 < 2.2e-16 ***
gdppc1k      1.4004      0.1621  8.6394 6.201e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
b_hat <- lin_model$coefficients[2]
b_hat_SE <- sqrt(diag(vcov_m_g))[2]

cat("b_hat:", b_hat, "\nb_hat SE:", b_hat_SE)
```

```
b_hat: 1.400427
b_hat SE: 0.1620984
```

```
#Quadratic model
quad_model <- lm(math_score ~ gdppc1k + I(gdppc1k*gdppc1k), data = data1.2)
vcov_m_g_g2 <- vcovHC(quad_model, type = "HC3")
coeftest(quad_model, vcov. = vcov_m_g_g2)
```

t test of coefficients:

```
              Estimate Std. Error t value
(Intercept)    412.0714864    5.8861464 70.0070
gdppc1k         3.4619840    0.4068620  8.5090
I(gdppc1k * gdppc1k) -0.0270478    0.0058909 -4.5915
              Pr(>|t|)
(Intercept)    < 2.2e-16 ***
gdppc1k        1.517e-15 ***
I(gdppc1k * gdppc1k) 6.917e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
beta1_hat <- lin_model$coefficients[2]
beta1_hat_SE <- sqrt(diag(vcov_m_g))[2]

beta2_hat <- lin_model$coefficients[3]
beta2_hat_SE <- sqrt(diag(vcov_m_g))[3]

cat("beta1_hat:", beta1_hat, "\nbeta1_hat SE:", beta1_hat_SE, "\n\nbeta2hat:",
    beta2_hat, "\nbeta2_hat SE:", beta2_hat_SE, "\n\n")
```

```
beta1_hat: 1.400427
beta1_hat SE: 0.1620984
```

```
beta2hat: NA
beta2_hat SE: NA
```

```
cat("The relationship between math scores and GDP per capita is non-linear
because beta2hat is statistically significant")
```

The relationship between math scores and GDP per capita is non-linear
because beta2hat is statistically significant

- Using your estimated coefficients from Equation (2) above, what is the expected value of the difference in math scores between a country with a GDP per capita of \$5,000 and a country with a GDP per capita of \$10,000?

```
pred_5k <- coef(quad_model)[1] + coef(quad_model)[2]*5 + coef(quad_model)[3]*5^2
pred_10k <- coef(quad_model)[1] + coef(quad_model)[2]*10 + coef(quad_model)[3]*10^2
diff_pred <- pred_10k - pred_5k

cat("Expected difference in math scores (10k vs 5k GDP per capita):",
    round(diff_pred, 3))
```

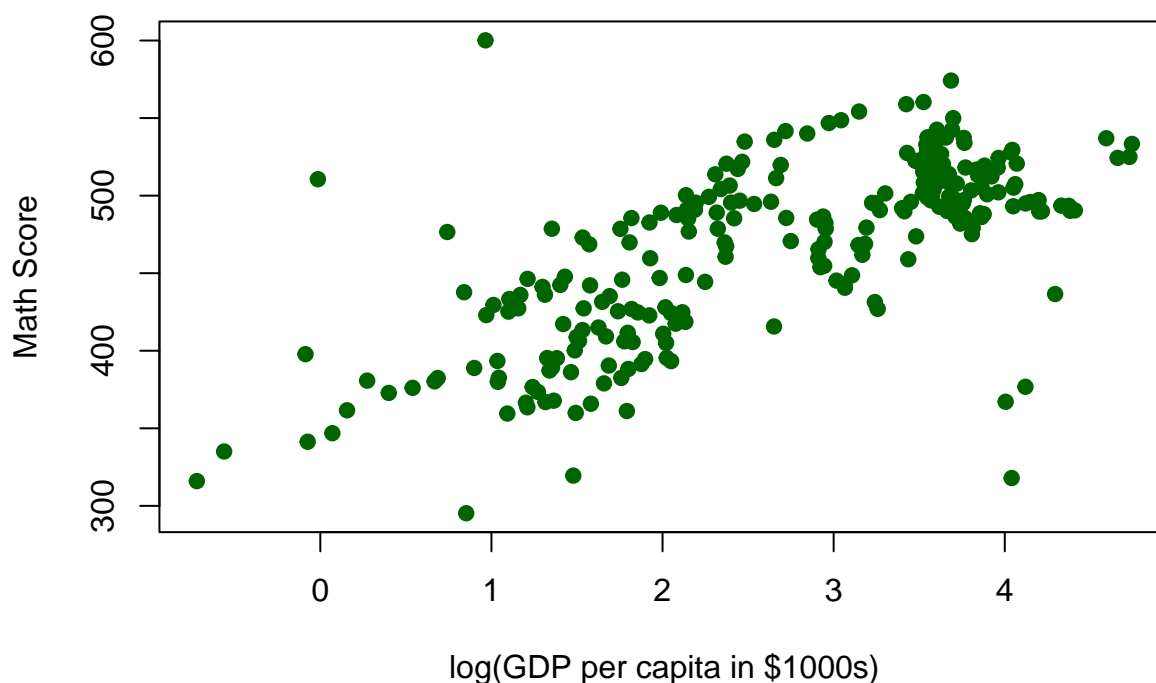
Expected difference in math scores (10k vs 5k GDP per capita): 15.281

- Generate a scatter plot with the math score on the y-axis and the (natural) logarithm of the GDP per capita (in thousands of dollars) on the x-axis. Based on visual inspection, is the relationship between these two transformed variables linear? Do you think that a linear- log specification will be able to better explain the relationship between these two variables compared to the variables in part 1?

```
data1.3 <- data1.2 %>% mutate(log_gdppc1k = log(gdppc1k))

plot(data1.3$log_gdppc1k, data1.2$math_score,
     main = "Math Score vs. log(GDP per capita)",
     xlab = "log(GDP per capita in $1000s)",
     ylab = "Math Score",
     col = "darkgreen", pch = 19)
```

Math Score vs. log(GDP per capita)



```
cat("Visual inspection shows a LINEAR relationship after taking logs, indicating
that the linear-log model might fit better than the raw-level model.")
```

Visual inspection shows a LINEAR relationship after taking logs, indicating that the linear-log model might fit better than the raw-level model.

5. Estimate the following two regression equations:

$$\text{math score}_i = \theta_1 + \theta_2 \ln(\text{gdppc1ki}) + u_i$$

$$\ln(\text{math score}_i) = \gamma_1 + \gamma_2 \ln(\text{gdppc1ki}) + v_i$$

Report the estimated intercept, slope, and their respective heteroskedasticity-robust standard errors for each equation. How do you interpret the intercept and the slope in each equation?

```
#Linear-log model
linlog_model <- lm(math_score ~ log_gdppc1k, data = data1.3)
vcov_li <- vcovHC(linlog_model, type = "HC3")
coeftest(linlog_model, vcov. = vcov_li)
```

t test of coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept) 374.3795      7.5512  49.579 < 2.2e-16 ***
log_gdppc1k  34.7885      2.5307  13.747 < 2.2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

#Extract individual values
theta1_hat <- coef(linlog_model)[1]
theta2_hat <- coef(linlog_model)[2]
theta1_se  <- sqrt(diag(vcovHC(linlog_model, type = "HC3")))[1]
theta2_se  <- sqrt(diag(vcovHC(linlog_model, type = "HC3")))[2]

#Report values
cat("theta1hat (Intercept):", theta1_hat,
    " theta1hatSE:", theta1_se, "\n")

```

```
theta1hat (Intercept): 374.3795 theta1hatSE: 7.55121
```

```

cat("theta2hat (Coefficient):", theta2_hat,
    " theta2hatSE:", theta2_se, "\n")

```

```
theta2hat (Coefficient): 34.78846 theta2hatSE: 2.530665
```

```

#Log-log model
data1.4 <- data1.3 %>% mutate(log_math = log(math_score))
loglog_model <- lm(log_math ~ log_gdppc1k, data = data1.4)
coeftest(loglog_model, vcov = vcovHC(loglog_model, type = "HC3"))

```

t test of coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.9291352  0.0178519  332.13 < 2.2e-16 ***
log_gdppc1k 0.0785827  0.0059399   13.23 < 2.2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

#Extract individual values neatly
gamma1_hat <- coef(loglog_model)[1]
gamma2_hat <- coef(loglog_model)[2]
gamma1_se  <- sqrt(diag(vcovHC(loglog_model, type = "HC3")))[1]
gamma2_se  <- sqrt(diag(vcovHC(loglog_model, type = "HC3")))[2]

```

```
#Report values
cat("gamma1hat (Intercept):", gamma1_hat,
    " gamma1hatSE:", gamma1_se, "\n")
```

```
gamma1hat (Intercept): 5.929135    gamma1hatSE: 0.01785189
```

```
cat("gamma2hat (Coefficient):", gamma2_hat,
    " gamma2hatSE:", gamma2_se, "\n")
```

```
gamma2hat (Coefficient): 0.07858271    gamma2hatSE: 0.00593991
```

```
cat("\nLinear-log model Interpretations:\n\n")
```

Linear-log model Interpretations:

```
cat("In the linear-log model, the slope (beta2hat) represents the change in
math score for a 1% change in GDP per capita.\n\n")
```

In the linear-log model, the slope (beta2hat) represents the change in math score for a 1% change in GDP per capita.

```
cat("In the log-log model, the slope (gamma2hat) (elasticity) shows the % change
in math score for a 1% change in GDP per capita.")
```

In the log-log model, the slope (gamma2hat) (elasticity) shows the % change in math score for a 1% change in GDP per capita.

6. Use the R-squared to determine which specification between the following pairs better explains the relationship between math scores and GDP per capita:

- a) Quadratic vs Linear-log, b) Linear-log vs Log-log. Explain.

```
#Obtaining R^2 values
R2_lin <- summary(lin_model)$r.squared
R2_quad <- summary(quad_model)$r.squared
R2_linlog <- summary(linlog_model)$r.squared
R2_loglog <- summary(loglog_model)$r.squared

cat("R_squared Linear:", round(R2_lin, 3),
    "\nR_squared Quadratic:", round(R2_quad, 3),
    "\nR_squared Linear-log:", round(R2_linlog, 3),
    "\nR_squared Log-log:", round(R2_loglog, 3))
```

R_squared Linear: 0.292
R_squared Quadratic: 0.415
R_squared Linear-log: 0.493
R_squared Log-log: 0.492

```
cat("\n\n(a) Quadratic vs Linear-log: The model with higher R2 explains more variance,  
but since they use different transformations, direct comparison must be cautious.\n\n")
```

(a) Quadratic vs Linear-log: The model with higher R^2 explains more variance, but since they use different transformations, direct comparison must be cautious.

```
cat("(b) Linear-log vs Log-log: Both have the same dependent variable transformation,  
so the higher R2 model better fits the data.")
```

(b) Linear-log vs Log-log: Both have the same dependent variable transformation, so the higher R^2 model better fits the data.

Question 2

```
data2.1 <- read.csv("DDCG_dataset.csv")
```

1. For the sample of countries with available data in 2005, compute the mean GDP per capita for those that have a democracy ($demo = 1$). Do the same for the sample of countries that do not have a democracy in 2005 ($demo = 0$). What is the difference? Can you reject the null hypothesis that the two means are equal at the 5% significance level (no regressions needed for this part)? Estimate a regression of GDP per capita on the democracy dummy, also restricted to year 2005. How do the estimated intercept and slope compare to your previous calculations?

```
# Restrict to 2005
data2_2005 <- data2.1 %>% filter(year == 2005)

#Get group means
means <- data2_2005 %>%
  group_by(dem) %>%
  summarise(mean_gdp = mean(gdp_capita, na.rm = TRUE))

means
```

```
# A tibble: 2 x 2
  dem mean_gdp
```

```

      <int>    <dbl>
1         0    4223.
2         1    9675.

```

```

#Difference in means
diff_means <- means$mean_gdp[means$dem == 1] - means$mean_gdp[means$dem == 0]
cat("Difference in mean GDP per capita (democracy - non-democracy):", round(diff_means,
  ↪ 3), "\n")

```

Difference in mean GDP per capita (democracy - non-democracy): 5452.018

```

#Two-sample t-test (unequal variances)
t_test <- t.test(gdp_capita ~ dem, data = data2_2005, var.equal = FALSE)
t_test

```

Welch Two Sample t-test

```

data:  gdp_capita by dem
t = -3.1665, df = 120.04, p-value = 0.001956
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -8860.969 -2043.067
sample estimates:
mean in group 0 mean in group 1
    4222.954      9674.972

```

```

p_val <- t_test$p.value

cat("\nInterpretation: We reject the null hypothesis that mean GDP per capita is
equal across statuses since p-value(", round(p_val, 3), ") is less than alpha(", 0.05,
  ↪ ")")

```

Interpretation: We reject the null hypothesis that mean GDP per capita is equal across statuses since p-value(0.002) is less than alpha(0.05)

```

reg_2005 <- lm(gdp_capita ~ dem, data = data2_2005)
coeftest(reg_2005, vcov = vcovHC(reg_2005, type = "HC3"))

```

t test of coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4223.0      1188.2  3.5540 0.0005271 ***
dem          5452.0      1736.3  3.1399 0.0020856 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

cat("Intercept (alpha) should equal mean(GDP | dem=0):",
    round(means$mean_gdp[means$dem == 0], 3), "\n")

```

Intercept (alpha) should equal mean(GDP | dem=0): 4222.954

```

cat("Slope (Beta_hat) should equal difference in means:",
    round(means$mean_gdp[means$dem == 1] - means$mean_gdp[means$dem == 0], 3), "\n")

```

Slope (Beta_hat) should equal difference in means: 5452.018

2. Throughout the rest of this question, we are going to focus on studying the relationship between the natural logarithm of GDP per capita and the democracy dummy:

$$\log(\text{gdp capita}) = \alpha + \beta \text{dem} + \varepsilon$$

Estimate the equation above. Report the estimated coefficients $\hat{\alpha}$ and $\hat{\beta}$, along with their corresponding heteroskedasticity-robust standard errors. Interpret $\hat{\beta}$. Is β statistically different from zero at the 5% significance level (report t-statistic and p-value)?

```

# Add log of GDP per capita
data2.2 <- data2.1 %>%
  mutate(log_gdp = log(gdp_capita))

model3 <- lm(log_gdp ~ dem, data = data2.2)
vcov <- vcovHC(model3, type = "HC3")
coeftest(model3, vcov. = vcov)

```

t test of coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.10971      0.12621 56.3312 < 2.2e-16 ***
dem          1.06652      0.16046  6.6468 1.033e-10 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```
# Extract estimates
alpha_hat <- coef(model3)[1]
beta_hat <- coef(model3)[2]
se_alpha <- sqrt(diag(vcovHC(model3, type = "HC3")))[1]
se_beta <- sqrt(diag(vcovHC(model3, type = "HC3")))[2]

t_beta <- beta_hat / se_beta

cat("alpha_hat =", round(alpha_hat, 3), " (SE =", round(se_alpha, 3), ")\n",
    "Beta_hat =", round(beta_hat, 3), " (SE =", round(se_beta, 3), ", t =", round(t_beta,
    ↵ 3), ")\n")
```

```
alpha_hat = 7.11 (SE = 0.126 )
Beta_hat = 1.067 (SE = 0.16 , t = 6.647 )
```

```
cat("\nInterpretation: Beta_hat represents the percentage difference in GDP per capita
between democracies and non-democracies. If Beta_hat is statistically significant
(|t| > 2), democracies have higher average log(GDP per capita).")
```

Interpretation: Beta_hat represents the percentage difference in GDP per capita between democracies and non-democracies. If Beta_hat is statistically significant ($|t| > 2$), democracies have higher average log(GDP per capita).

3. Do you think that the equation above can suffer from omitted variable bias (OVB)? List two omitted variables that can potentially lead to OVB when estimating Equation (5). For each of them, argue whether the estimate $\hat{\beta}$ is upward or downward biased (i.e., would $\hat{\beta}$ go down or up, if we included the omitted variable)?

```
cat("Possible omitted variables are education levels (more educated populations
--> higher GDP and higher likelihood of democracy), trade openness (trade boosts
income and correlates with democratic institutions)")
```

Possible omitted variables are education levels (more educated populations --> higher GDP and higher likelihood of democracy), trade openness (trade boosts income and correlates with democratic institutions)

```
cat("\n\nIf these are omitted, and both correlate positively with democracy and
GDP, then beta_hat is upward biased, meaning the democracy effect appears
larger than it really is.")
```

If these are omitted, and both correlate positively with democracy and GDP, then beta_hat is upward biased, meaning the democracy effect appears larger than it really is.

4.

```
model4 <- lm(log_gdp ~ dem + lp_bl + lh_bl, data = data2.2)
coeftest(model4, vcov = vcovHC(model4, type = "HC3"))
```

t test of coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.1858112	0.1833171	33.7438	< 2e-16 ***
dem	0.3822453	0.1651491	2.3145	0.02117 *
lp_bl	0.0082106	0.0037405	2.1951	0.02876 *
lh_bl	0.0942623	0.0102318	9.2127	< 2e-16 ***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
# Extract Beta_hat and SE
beta_hat_edu <- coef(model4)["dem"]
beta_se_edu <- sqrt(diag(vcovHC(model4, type = "HC3")))[2]

cat("With education controls:\nBeta_hat =", round(beta_hat_edu, 3),
    "| SE =", round(beta_se_edu, 3), "\n")
```

With education controls:

Beta_hat = 0.382 | SE = 0.165

```
cat("\nIf Beta_hat decreases after adding education controls, the earlier model
likely overstated democracy's effect, meaning the initial estimate was upward
biased.")
```

If Beta_hat decreases after adding education controls, the earlier model likely overstated democracy's effect, meaning the initial estimate was upward biased.

5. On top of lp bl and lh bl, also add ginv and tradewb as control variables in Equation (5). How does $\hat{\beta}$ change (report the standard error as well)? Provide an intuitive explanation. Does this change mean that the $\hat{\beta}$ was upward or downward biased in part 4?

```
model5 <- lm(log_gdp ~ dem + lp_bl + lh_bl + ginv + tradewb, data = data2.2)
coeftest(model5, vcov = vcovHC(model5, type = "HC3"))
```

t test of coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.73386870	0.28315581	20.2499	< 2.2e-16 ***
dem	0.43158752	0.15213068	2.8370	0.004799 **
lp_bl	0.00806050	0.00375963	2.1440	0.032672 *
lh_bl	0.09224219	0.00954171	9.6673	< 2.2e-16 ***
ginv	-0.00753655	0.00877757	-0.8586	0.391096
tradewb	0.00711982	0.00099077	7.1862	3.566e-12 ***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
# Extract updated Beta_hat and SE
beta_hat_econ <- coef(model5)["dem"]
beta_se_econ <- sqrt(diag(vcovHC(model5, type = "HC3")))[2]

cat("With education + economic controls:\nBeta_hat =", round(beta_hat_econ, 3),
    "\n| SE =", round(beta_se_econ, 3), "\n")
```

With education + economic controls:

Beta_hat = 0.432 | SE = 0.152

```
#compute changes
beta_change <- beta_hat_econ - beta_hat_edu
percent_change <- 100 * beta_change / beta_hat_edu

if (beta_change < 0) {
  cat("\n\nInterpretation:\n\n",
      "After adding ginv and tradewb, the democracy coefficient Beta_hat decreased.\n",
      "This indicates that the positive association between democracy and income was\n",
      "↪ partly explained\n",
      "by higher investment and trade levels typical of democratic countries.\n",
      "Therefore, the Beta_hat from part 4 was **upward biased** - omitting ginv and\n",
      "↪ tradewb made democracy appear\n",
      "to have a stronger effect on GDP per capita than it truly does once those factors\n",
      "↪ are controlled for.\n")
} else if (beta_change > 0) {
  cat("\n\nInterpretation:\n\n",
      "After adding ginv and tradewb, the democracy coefficient Beta_hat\n",
      "increase. This suggests that trade and investment were negatively correlated\n",
      "with democracy or positively correlated with GDP in a way that suppressed\n",
      "democracy's apparent effect. Thus, the Beta_hat from part 4 was downward biased.\n",
      "Omitting ginv and tradewb understated democracy's association with GDP per\n",
      "capita.\n")
}
```

Interpretation:

After adding `ginv` and `tradewb`, the democracy coefficient `Beta_hat` increase. This suggests that trade and investment were negatively correlated with democracy or positively correlated with GDP in a way that suppressed democracy's apparent effect. Thus, the `Beta_hat` from part 4 was downward biased. Omitting `ginv` and `tradewb` understated democracy's association with GDP per capita.

6. Using the estimates from part 5, can we reject the null hypothesis that the two control variables related to the economy are irrelevant in predicting GDP per capita (i.e., the coefficients on both `tradewb` and `ginv` are zero)?

```
#install.packages("car")  
library(car)
```

Warning: package 'car' was built under R version 4.4.3

Loading required package: carData

Warning: package 'carData' was built under R version 4.4.3

Attaching package: 'car'

The following object is masked from 'package:dplyr':

`recode`

```
# Perform robust Wald test (heteroskedasticity-robust F-test)  
joint_test <- linearHypothesis(model15,  
                               c("ginv = 0", "tradewb = 0"),  
                               vcov = vcovHC(model15, type = "HC3"))  
  
joint_test
```

Linear hypothesis test:

`ginv = 0`

`tradewb = 0`

Model 1: restricted model

Model 2: `log_gdp ~ dem + lp_bl + lh_bl + ginv + tradewb`

Note: Coefficient covariance matrix supplied.

```
Res.Df Df      F    Pr(>F)
1      381
2      379  2 25.821 3.069e-11 ***
```

Signif. codes:

```
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# --- Interpretation block ---
p_value <- joint_test[2, "Pr(>F)"]

if (p_value < 0.05) {
  cat("\nResult:\n",
      "We REJECT the null hypothesis (p-value =", round(p_value, 4), "). This
means ginv and tradewb are jointly significant predictors of
log(GDP per capita). Economic factors like investment and trade openness help
explain cross-country differences in income levels.\n")
} else {
  cat("\nResult:\n",
      "We FAIL TO REJECT the null hypothesis (p-value =", round(p_value, 4), ").\n",
      "This means there is no statistical evidence at the 5% level that ginv and tradewb
↪ are jointly relevant.\n",
      "Adding them does not significantly improve the model's ability to predict GDP per
↪ capita.\n")
}
```

Result:

We REJECT the null hypothesis (p-value = 0). This means ginv and tradewb are jointly significant predictors of log(GDP per capita). Economic factors like investment and trade openness help explain cross-country differences in income levels.