PS 1

ARE 212, Spring 2026

David Cai, Miriam Gold, Ella Moxley, Geoffrey Yip

2026-02-25

Set-up

```r
library(WDI)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(magrittr)
```

```
Attaching package: 'magrittr'

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

    set_names

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

    extract
```

```r
library(fs)


# Paths ==================
path <- "/Users/miriamgold/projects/ARE212_2026/ps1"
path_functions <- file.path(path, "functions")
path_data <- file.path(path, "data")

# Source custom functions
dir_walk(path_functions, source)

# Data setup ===============

iso_codes <-
  read_csv("https://gist.githubusercontent.com/radcliff/f09c0f88344a7fcef373/raw/2753c482ad09
```

```
Rows: 246 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (5): English short name lower case, Alpha-2 code, Alpha-3 code, Numeric ...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
wb_vars <- c("EN.GHG.CO2.MT.CE.AR5", "NY.GDP.MKTP.KD", "SP.POP.TOTL", "SH.XPD.GHED.GD.ZS")

countries <-
  iso_codes |>
  pull("Alpha-2 code")
```

Q1

```r
wdi_data_raw <-
  WDI(
    country = countries,
    indicator = wb_vars,
    start = 2010,
    end = 2010
  )
```

```
Warning in WDI(country = countries, indicator = wb_vars, start = 2010, end =
2010): Please use ISO-2, ISO-3, or World Bank regional codes. Some of the
country codes that you requested are invalid: AX, AI, AQ, BV, IO, CX, CC, CK,
FK, GF, TF, GP, GG, HM, VA, JE, MQ, YT, MS, NA, AN, NU, NF, PN, RE, BL, SH, PM,
GS, SJ, TW, TK, UM, WF, EH
```

Q2

```r
wdi_data_clean <-
  wdi_data_raw |>
  rename(
    CO2 = "EN.GHG.CO2.MT.CE.AR5",
    GDP = "NY.GDP.MKTP.KD",
    POP = "SP.POP.TOTL",
    GOV = "SH.XPD.GHED.GD.ZS"
  ) |>
  drop_na()
```

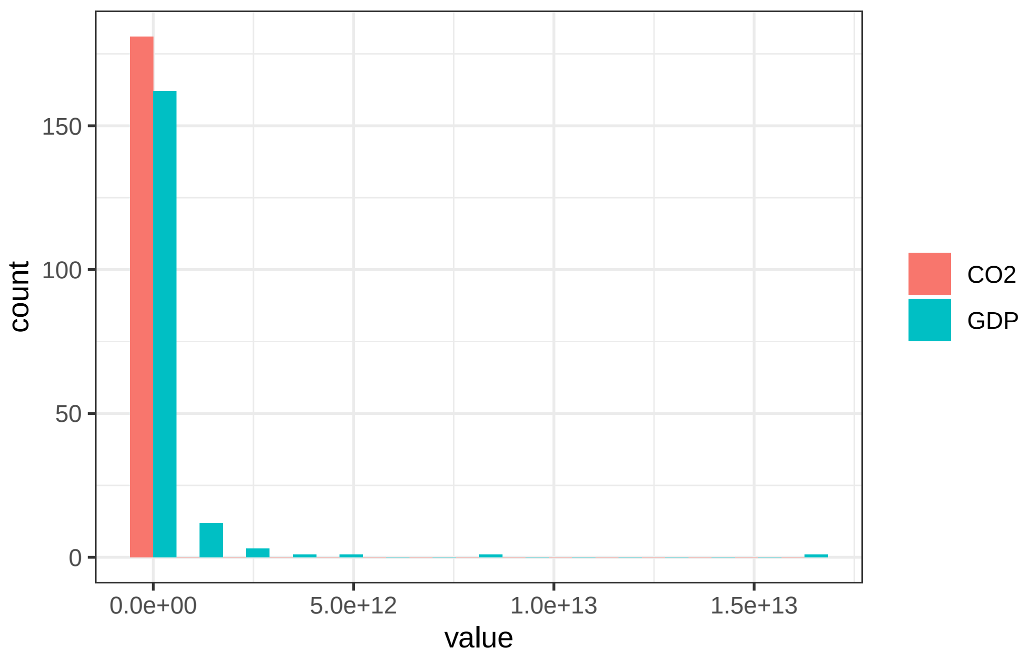
Q3

```r
wdi_summary_stats <-
  wdi_data_clean |>
   pivot_longer(cols = c("CO2", "GDP", "POP"), names_to = "var") |>
   group_by(var) |>
   summarise(
    mean = mean(value),
    sd = sd(value),
    min = min(value),
    max = max(value)
   )
print(wdi_summary_stats)
```

```
# A tibble: 3 x 5
  var             mean      sd        min      max
  <chr>          <dbl>   <dbl>      <dbl>    <dbl>
1 CO2             178. 8.15e 2          0  9.11e 3
2 GDP   353079197995. 1.42e12 30560853.  1.63e13
3 POP       37969801. 1.40e 8    10043   1.34e 9
```

Q4

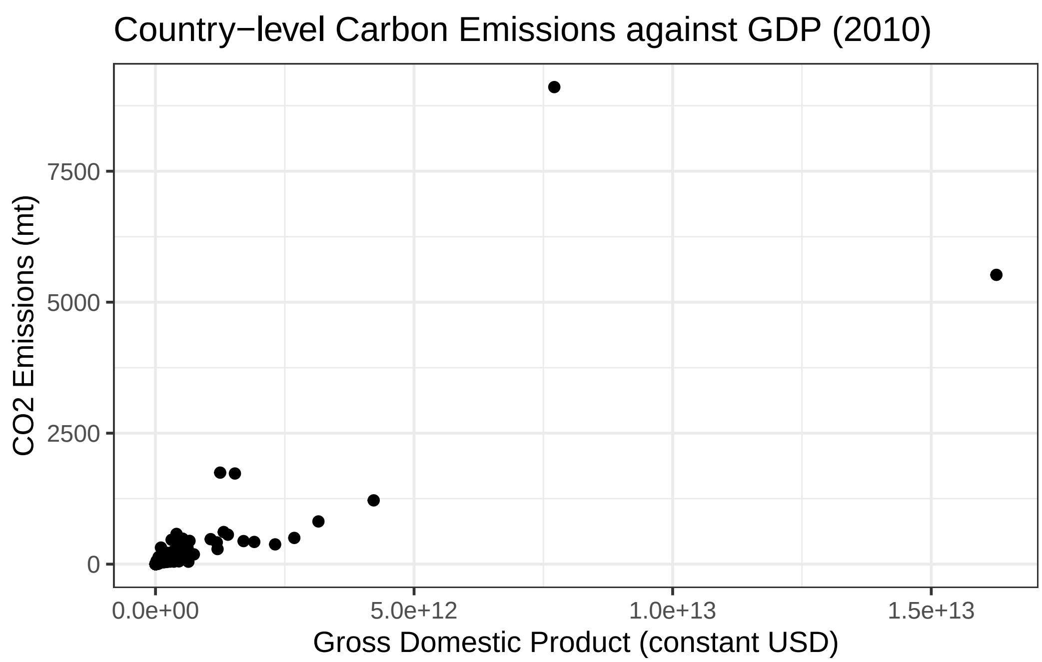
```
wdi_data_clean %>%
   pivot_longer(cols = c("CO2", "GDP", "POP"), names_to = "var") %>%
  filter(var %in% c("CO2", "GDP")) %>%
  ggplot(aes(x = value, fill = var)) +
  geom_histogram(bins = 15, position = "dodge") +
  scale_fill_discrete(name = NULL) +
  theme_bw()
```



Q5

```
wdi_data_clean %>%
  ggplot(aes(x = GDP, y = CO2)) +
  geom_point() +
  scale_x_continuous("Gross Domestic Product (constant USD)") +
  scale_y_continuous("CO2 Emissions (mt)") +
  labs(
    title = "Country-level Carbon Emissions against GDP (2010)"
```

```
  ) +
  theme_bw()
```



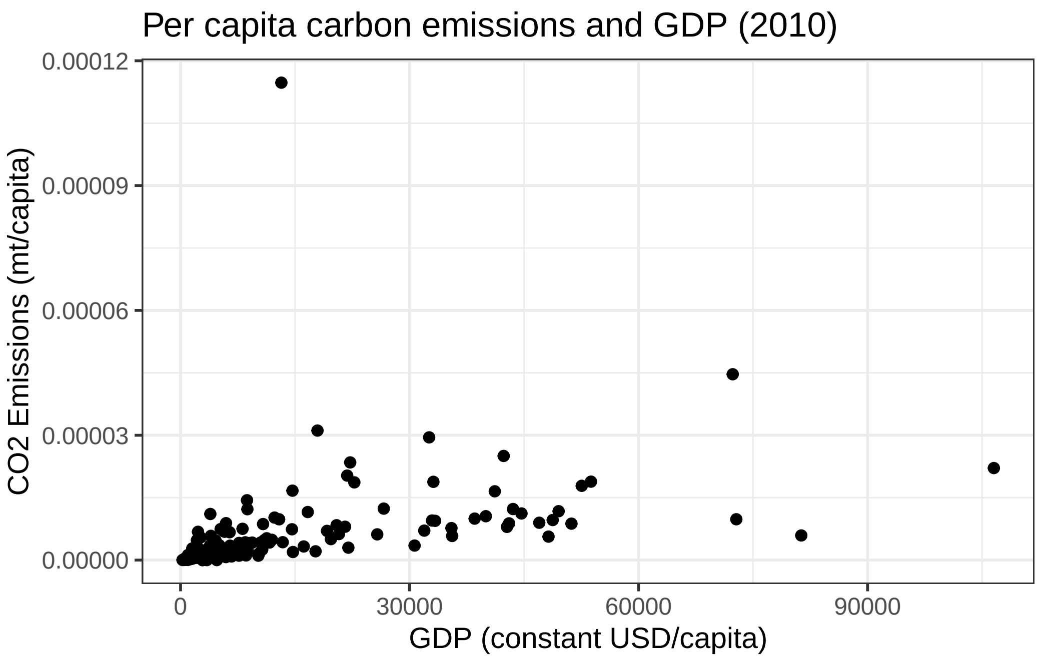
Q6/7

```
wdi_per_capita <-
  wdi_data_clean %>%
  mutate(
    CO2pc = CO2/POP,
    GDPpc = GDP/POP
  )
```

Q8

```
wdi_per_capita %>%
  ggplot(aes(x = GDPpc, y = CO2pc)) +
  geom_point() +
```

```
  scale_x_continuous("GDP (constant USD/capita)") +
  scale_y_continuous("CO2 Emissions (mt/capita)") +
  labs(
    title = "Per capita carbon emissions and GDP (2010)"
  ) +
  theme_bw()
```

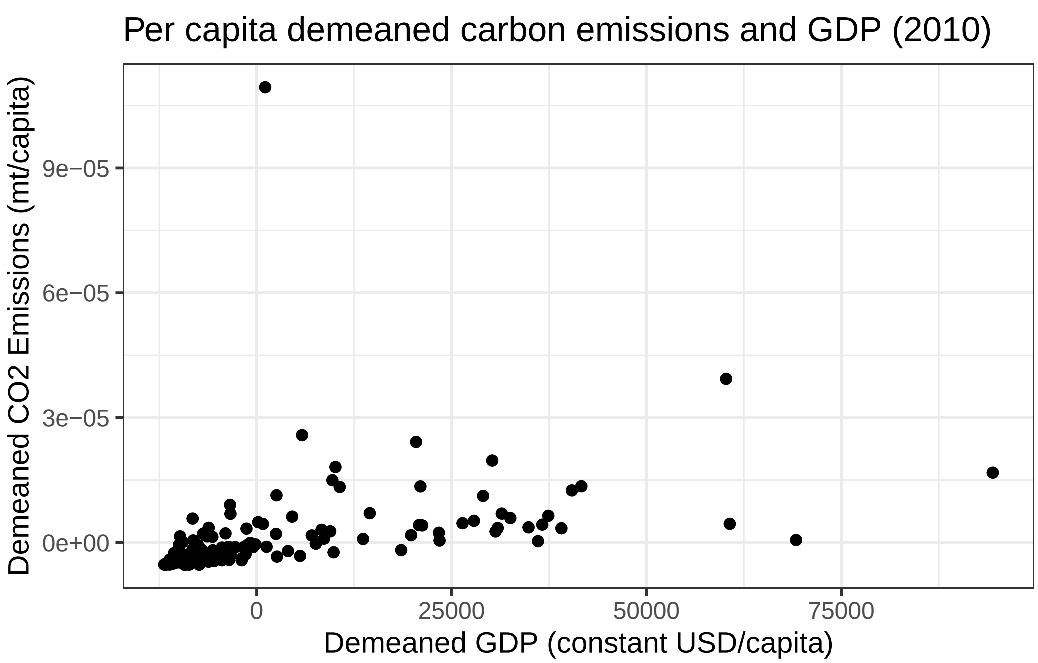


Q9

```
wdi_per_capita_demeaned <-
  wdi_per_capita %>%
  mutate(
    CO2pcdev = CO2pc-mean(CO2pc),
    GDPpcdev = GDPpc-mean(GDPpc)
  )
```

Q10

```
wdi_per_capita_demeaned %>%
  ggplot(aes(x = GDPpcdev, y = CO2pcdev)) +
  geom_point() +
  scale_x_continuous("Demeaned GDP (constant USD/capita)") +
  scale_y_continuous("Demeaned CO2 Emissions (mt/capita)") +
  labs(
    title = "Per capita demeaned carbon emissions and GDP (2010)"
  ) +
  theme_bw()
```



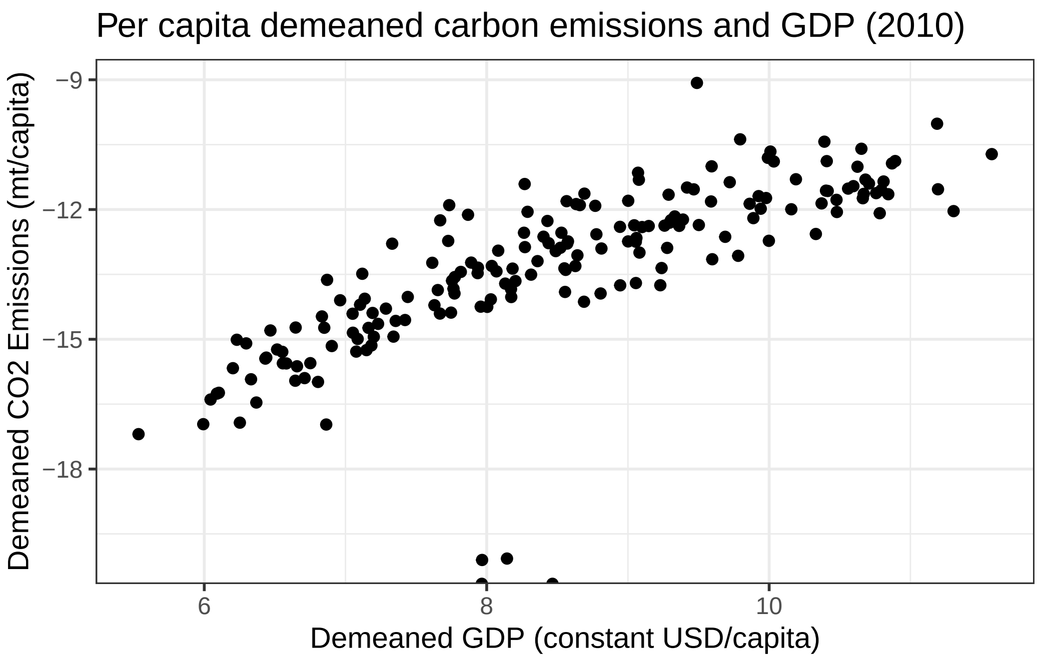
Q11

```
wdi_per_capita_log <-
  wdi_per_capita_demeaned %>%
  mutate(
    CO2pcln = log(CO2pc),
```

```
    GDPpcln = log(GDPpc)
  )
```

Q12

```
wdi_per_capita_log %>%
  ggplot(aes(x = GDPpcln, y = CO2pcln)) +
  geom_point() +
  scale_x_continuous("Demeaned GDP (constant USD/capita)") +
  scale_y_continuous("Demeaned CO2 Emissions (mt/capita)") +
  labs(
    title = "Per capita demeaned carbon emissions and GDP (2010)"
  ) +
  theme_bw()
```



Q13

```
  wdi_per_capita_log %>%
    write_csv(file = file.path(path_data, "are212_gold_ps1_wdi_clean.csv"))
```

Q14

We define a custom OLS regression function below

```
ols <- function(x, y, intercept = FALSE) {

  X <- as.matrix(x)
  Y <- as.matrix(y)

  if(intercept) {
    intercept_vector <- matrix(1, nrow(X))
    X <- cbind(intercept_vector, X)
  }


  N <- nrow(X)
  k <- ncol(X)
  b_ols <- solve(t(X)%*%X) %*% (t(X)%*%Y) # b = (X'X)^{-1}(X'Y)
  y_hat <- (X) %*% b_ols

  r2_uc <- (t(y_hat)%*%y_hat) / (t(Y)%*%Y)
  r2 <- t(y_hat-mean(Y))%*%(y_hat-mean(Y)) / t(Y-mean(Y))%*%(Y-mean(Y))
  r2 <- r2[[1,1]]
  r2_bar <- 1 - ((N-1)/(N-k))*(1-r2)

  e <- Y-y_hat
  s2 <- (t(e)%*%e)/(N-k)

  list(
    beta = b_ols,
    X = X,
    Y = Y,
    N = N,
    k = k,
    df = N-k,
    r2_uc = r2_uc,
    r2 = r2,
```

```
    r2_bar = r2_bar,
    y_hat = y_hat,
    e = e,
    s2 = s2
  )
}
```

Next, run each specification and report OLS coefficients

```
ols(
  wdi_per_capita_log$GDPpc,
  wdi_per_capita_log$CO2pc
)$beta
```

```
             [,1]
[1,] 3.195933e-10
```

```
ols(
  wdi_per_capita_log$GDPpc,
  wdi_per_capita_log$CO2pc*1000
)$beta # the coefficient has been multiplied by 1000)
```

```
             [,1]
[1,] 3.195933e-07
```

```
ols_co2_gdp <-
  ols(
    wdi_per_capita_log$GDPpc/1000,
    wdi_per_capita_log$CO2pc*1000
  )
ols_co2_gdp$beta
```

```
             [,1]
[1,] 0.0003195933
```

In the last regression, the coefficient is multiplied by a further factor of 1000

Q15

```r
## N = 197
n <- nrow(wdi_per_capita_log)

## Degrees of freedom: N-k
df <- n-1 # k=1 because there is no intercept

## beta
b <- ols_co2_gdp$beta

## R^2 uncentered
ols_co2_gdp$r2_uc
```

```
          [,1]
[1,] 0.3296782
```

```r
## R^2
ols_co2_gdp$r2
```

```
[1] 0.2979411
```

```r
## Adjusted R^2
ols_co2_gdp$r2_bar
```

```
[1] 0.2979411
```

```r
## s^2
ols_co2_gdp$s2
```

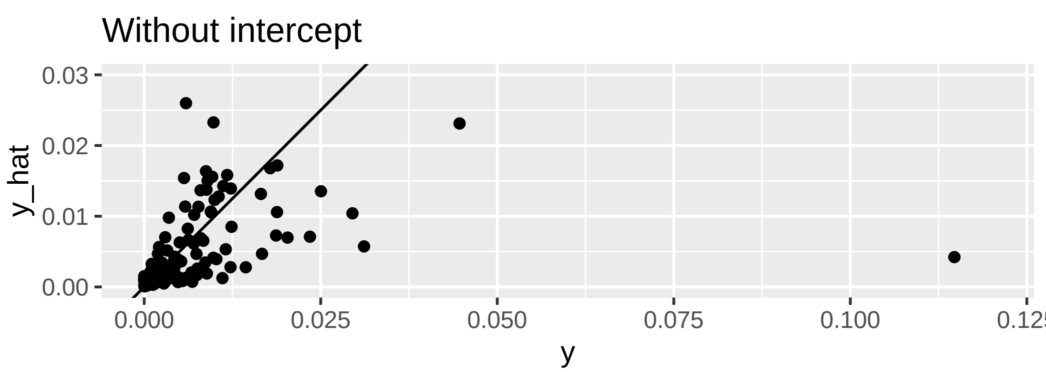
```
             [,1]
[1,] 9.257976e-05
```

```r
predicted_vs_actual <-
  data.frame(
    y_hat = ols_co2_gdp$y_hat,
    y = ols_co2_gdp$Y
)

predicted_vs_actual %>%
```

```r
  ggplot(aes(x = y, y = y_hat)) +
  geom_point() +
  geom_abline() +
  coord_fixed() +
  scale_y_continuous(limits = c(0, 0.03)) +
  scale_x_continuous(limits = c(0, 0.12)) +
  labs(
    title = "Without intercept"
  )
```

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



Q16

```r
ols_co2_gdp_intercept <-
  ols(
    wdi_per_capita_log$GDPpc/1000,
    wdi_per_capita_log$CO2pc*1000,
```

```r
    intercept = TRUE
  )

## beta
b_i <- ols_co2_gdp_intercept$beta

## R^2 uncentered
ols_co2_gdp_intercept$r2_uc
```

```
          [,1]
[1,] 0.3530937
```

```r
## R^2
ols_co2_gdp_intercept$r2
```

```
[1] 0.1835879
```

```r
## Adjusted R^2
ols_co2_gdp_intercept$r2_bar
```

```
[1] 0.179027
```

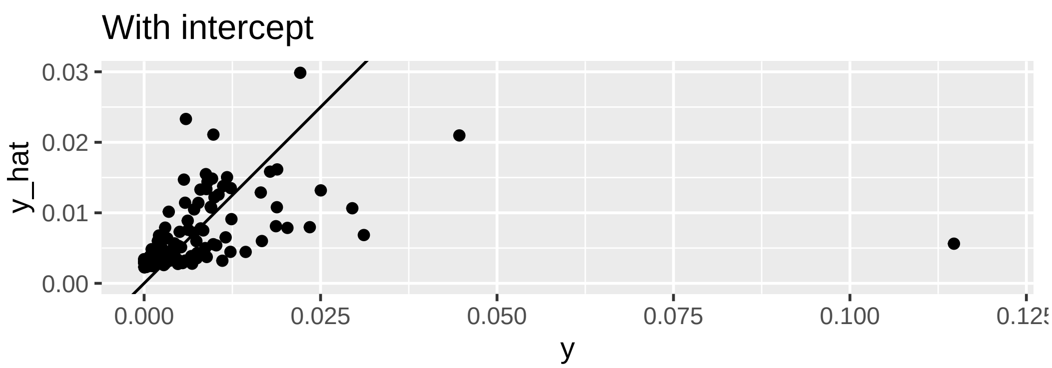
```r
## s^2
ols_co2_gdp_intercept$s2
```

```
             [,1]
[1,] 8.984494e-05
```

```r
predicted_vs_actual_intercept <-
  data.frame(
    y_hat = ols_co2_gdp_intercept$y_hat,
    y = ols_co2_gdp_intercept$Y,
    x = ols_co2_gdp_intercept$X[,2]
)

predicted_vs_actual_intercept %>%
  ggplot(aes(x = y, y = y_hat)) +
  geom_point() +
  geom_abline() +
```

```
  coord_fixed() +
  scale_y_continuous(limits = c(0, 0.03)) +
  scale_x_continuous(limits = c(0, 0.12)) +
  labs(
    title = "With intercept"
  )
```



Q17

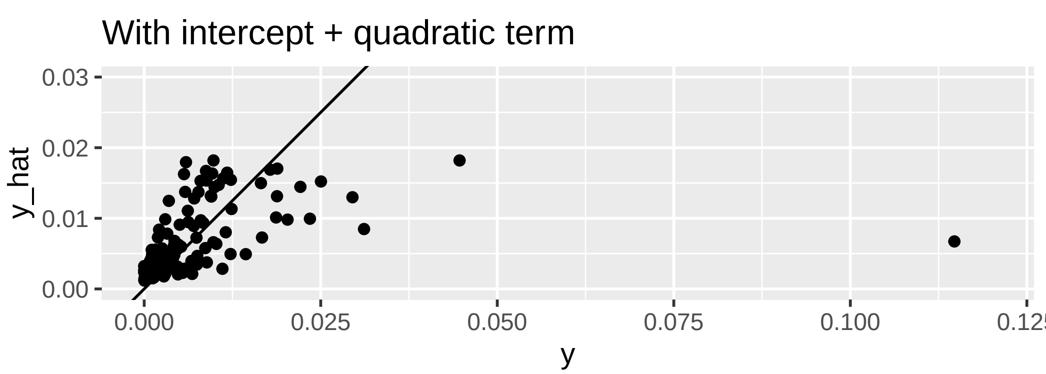
```
wdi_per_capita_quad <-
  wdi_per_capita_log %>%
  mutate(
    CO2pc_thou = CO2pc*1000,
    GDPpc_thou = GDPpc/1000,
    GDPpc_thou_2 = (GDPpc_thou^2)
  )

ols_co2_gdp_quad <-
  ols(
    x = wdi_per_capita_quad %>% select(GDPpc_thou, GDPpc_thou_2),
```

```
    y = wdi_per_capita_quad %>% select(CO2pc_thou),
    intercept = TRUE
  )

predicted_vs_actual_quad <-
  data.frame(
    y_hat = ols_co2_gdp_quad$y_hat,
    y = ols_co2_gdp_quad$Y,
    x = ols_co2_gdp_quad$X[,2]
) %>%
  rename(y_hat = 1, y = 2, x = 3)

predicted_vs_actual_quad %>%
  ggplot(aes(x = y, y = y_hat)) +
  geom_point() +
  geom_abline() +
  coord_fixed() +
  scale_y_continuous(limits = c(0, 0.03)) +
  scale_x_continuous(limits = c(0, 0.12)) +
  labs(
    title = "With intercept + quadratic term"
  )
```



Q18

```r
wdi_per_capita_demean <-
  wdi_per_capita_quad %>%
  mutate(
    CO2pc_thou_demean = CO2pc_thou - mean(CO2pc_thou),
    GDPpc_thou_demean = GDPpc_thou - mean(GDPpc_thou),
    GDPpc_thou_2_demean = GDPpc_thou_2 - mean(GDPpc_thou_2)
  )

ols_co2_gdp_demean <-
  ols(
    x = wdi_per_capita_demean %>% select(GDPpc_thou_demean, GDPpc_thou_2_demean),
    y = wdi_per_capita_demean %>% select(CO2pc_thou_demean),
    intercept = FALSE
  )

# They're the same!
ols_co2_gdp_quad$beta
```

```
               CO2pc_thou
             1.061127e-03
GDPpc_thou    4.718600e-04
GDPpc_thou_2 -3.248921e-06
```

```r
ols_co2_gdp_demean$beta
```

```
                    CO2pc_thou_demean
GDPpc_thou_demean        4.718600e-04
GDPpc_thou_2_demean     -3.248921e-06
```

Q19

```r
ols_co2_gdp_q19 <-
  ols(
    x = wdi_per_capita_quad %>% select(GDPpc_thou),
    y = wdi_per_capita_quad %>% select(CO2pc_thou),
```

```r
    intercept = TRUE
  )

resid_q19 <- ols_co2_gdp_q19$e

ols_1_gdp <-
  ols(
    x = wdi_per_capita_quad %>% select(GDPpc_thou),
    y = matrix(1, nrow = nrow(wdi_per_capita_quad)),
    intercept = FALSE
  )

ols_1_gdp_resid <- ols_1_gdp$e


ols_gdp2_gdp <-
  ols(
    x = wdi_per_capita_quad %>% select(GDPpc_thou),
    y = wdi_per_capita_quad %>% select(GDPpc_thou_2),
    intercept = FALSE
  )

ols_gdp2_gdp_resid <- ols_gdp2_gdp$e

resid_matrix <- cbind(ols_1_gdp_resid, ols_gdp2_gdp_resid)

ols_resid_on_resid <-
  ols(
    x = resid_matrix,
    y = resid_q19,
    intercept = FALSE
  )

ols_resid_on_resid$beta
```

```
                CO2pc_thou
             -1.132074e-03
GDPpc_thou_2 -3.248921e-06
```