HW2 - But We Make It Up in Volume

Emulie Chhor

30/12/2022

Question 1 -

```r
model1 <- lm(growth~underval + log(gdp), data = uval)
kable(summary(model1)$coefficients)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0352453	0.0066496	-5.300375	0.0000001
underval	0.0047639	0.0021791	2.186141	0.0289834
log(gdp)	0.0062971	0.0007905	7.965909	0.0000000

Q1a)

We see that the coefficient for log(gdp) is 0.00629 with p-value of 0, which means that the log(gdp) is statistically significant. Since we say that for every increase in log(gdp), we expect the country to grow by a factor of 0.00629, the coefficient doesn't support the idea of "catching-up"

Q1b)

The coefficient for underval is 0.0047 with p-value of $0.02 < \alpha = 0.05$, which means that underval is statistically significant. We say that for every increase of the index of under-valuation, we expect the country to grow by 0.0047%, which means that the data does support the under-valuing idea.

Question 2 -

```r
model2 <- lm(growth~underval + log(gdp) + country + factor(year),
             data = uval)

# kable(summary(model2)$coefficients)
```

Q2a)

```r
kable(summary(model2)$coefficients[2:3, 1:2])
```

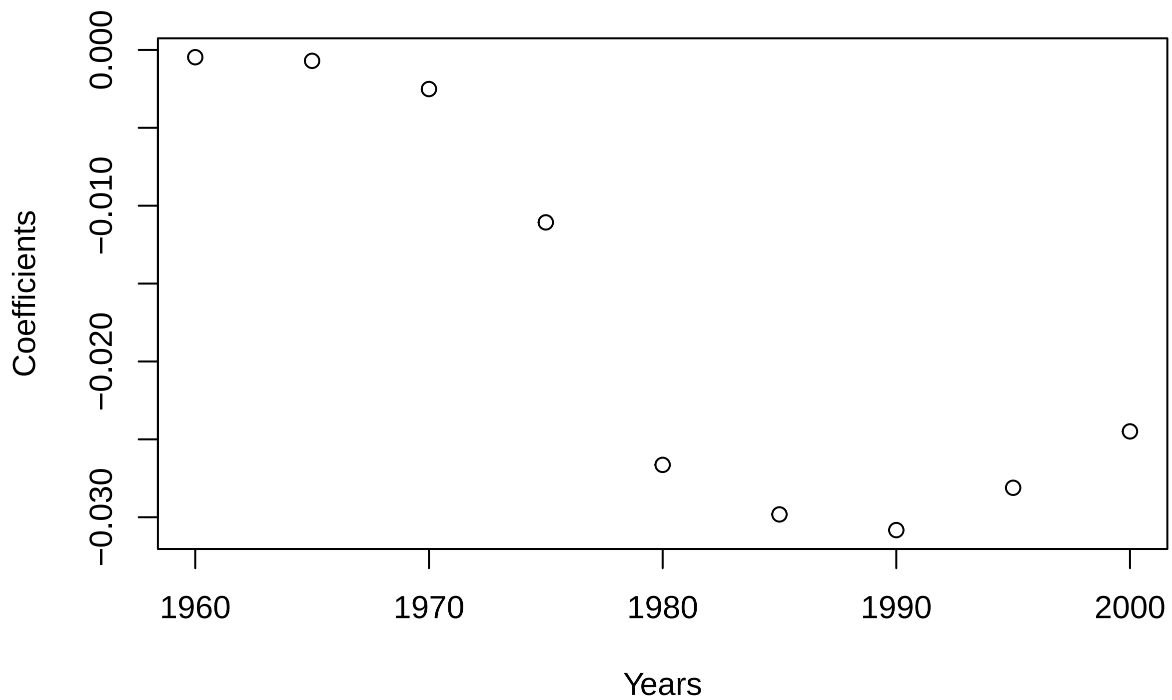
	Estimate	Std. Error
underval	0.0136094	0.0028977
log(gdp)	0.0289246	0.0031672

Q2b)

Since we only have 10 different values for year 5 years apart, we would rather consider the covariate year as a discrete value. This means that we would have a distinct slope for the 10 years value rather than for every yearly increment.

Q2c)

```
years.coeff <- summary(model2)$coefficients[182:190, 1]
years.values <- sort(unique(uval$year))[2:10]
plot(years.values, years.coeff, xlab = "Years", ylab = "Coefficients")
```



Q2d)

The second model doesn't support the idea of catching up because, again, the log(gdp) coefficient is positive and statistically significant, which suggest that for every log(gdp) increase, the country grows by 0.0289 %. However, the model is in accord with the undervalue idea since the underval coefficient is positive and is statistically significant ($pvalue < \alpha = 0.05$)

Question 3 -

Q3a)

```
summary(model1)$r.squared
```

```
## [1] 0.04855196
```

```
summary(model1)$adj.r.squared
```

```
## [1] 0.04708594
```

```
summary(model2)$r.squared
```

```
## [1] 0.4292363
```

```
summary(model2)$adj.r.squared
```

```
## [1] 0.3321397
```

The R-squared value can be used to compare models, as it give the proportion of variance in the response variable explained by the model. Therefore, since the R-squared value (and adjusted) are both bigger in the second model, we say that the second model is the better fit.

Q3b)

```
cv.lm <- function(data, formulae, nfolds = 5) {
  data <- na.omit(data)
  formulae <- sapply(formulae, as.formula)
  n <- nrow(data)
  fold.labels <- sample(rep(1:nfolds, length.out = n))
  mses <- matrix(NA, nrow = nfolds, ncol = length(formulae))
  colnames <- as.character(formulae)
  for (fold in 1:nfolds) {
    test.rows <- which(fold.labels == fold)
    train <- data[-test.rows, ]
    test <- data[test.rows, ]
    for (form in 1:length(formulae)) {
      current.model <- lm(formula = formulae[[form]], data = train)
      predictions <- predict(current.model, newdata = test)
      test.responses <- eval(formulae[[form]][[2]], envir = test)
      test.errors <- test.responses - predictions
      mses[fold, form] <- mean(test.errors^3)
    }
  }
  return(colMeans(mses))
}

loocv.mse <- cv.lm(uval, c("growth ~ underval + log(gdp)",
     "growth ~ underval + log(gdp) + factor(country) + factor(year)"),
     nfolds = nrow(uval))
loocv.mse
```

```
## [1] -1.409374e-05 -3.267075e-06
```

```
names(loocv.mse) <- c("Model 1", "Model 2")
kable(loocv.mse)
```

	x
Model 1	-1.41e-05
Model 2	-3.30e-06

Q3c)

??

Question 4 -

Q4a)

```
model3 <- npreg(growth ~ log(gdp) + underval + factor(year), data = uval)
```

```
## Multistart 1 of 3 |Multistart 1 of 3 |Multistart 1 of 3 |Multistart 1 of 3 /Multistart 1 of 3 -Multis
```

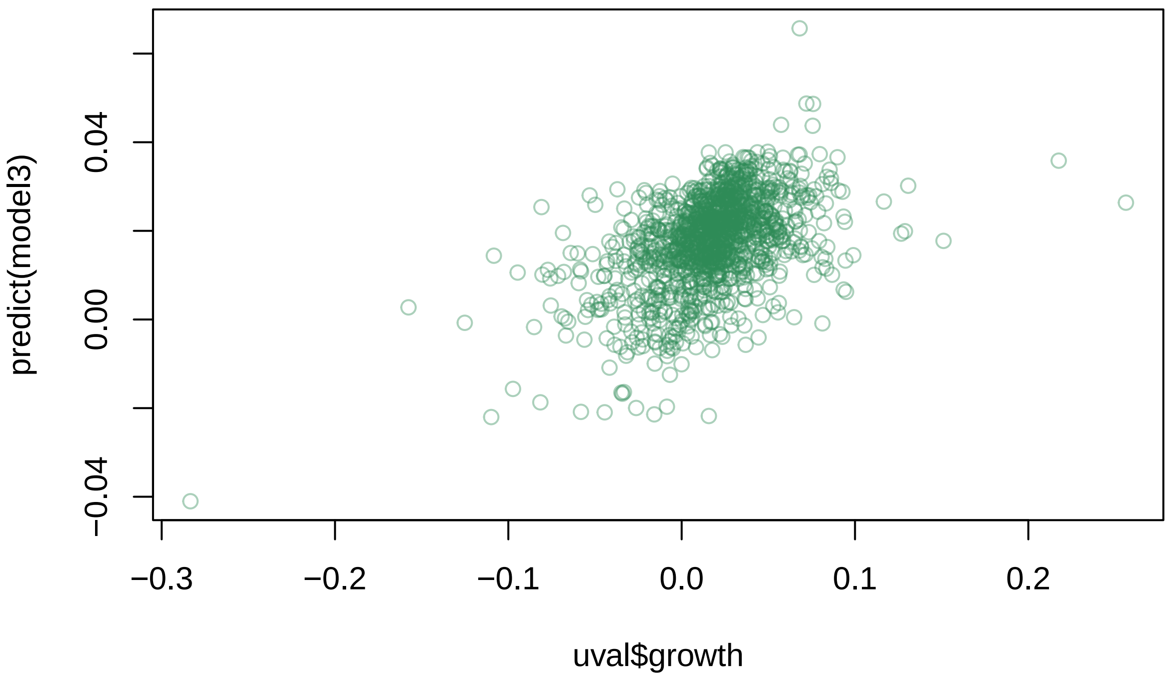
```
summary(model3)
```

```
##
## Regression Data: 1301 training points, in 3 variable(s)
##                log(gdp)  underval factor(year)
## Bandwidth(s): 0.7190708 0.2560892    0.1706824
##
## Kernel Regression Estimator: Local-Constant
## Bandwidth Type: Fixed
## Residual standard error: 0.02921354
## R-squared: 0.2359298
##
## Continuous Kernel Type: Second-Order Gaussian
## No. Continuous Explanatory Vars.: 2
##
## Unordered Categorical Kernel Type: Aitchison and Aitken
## No. Unordered Categorical Explanatory Vars.: 1
```

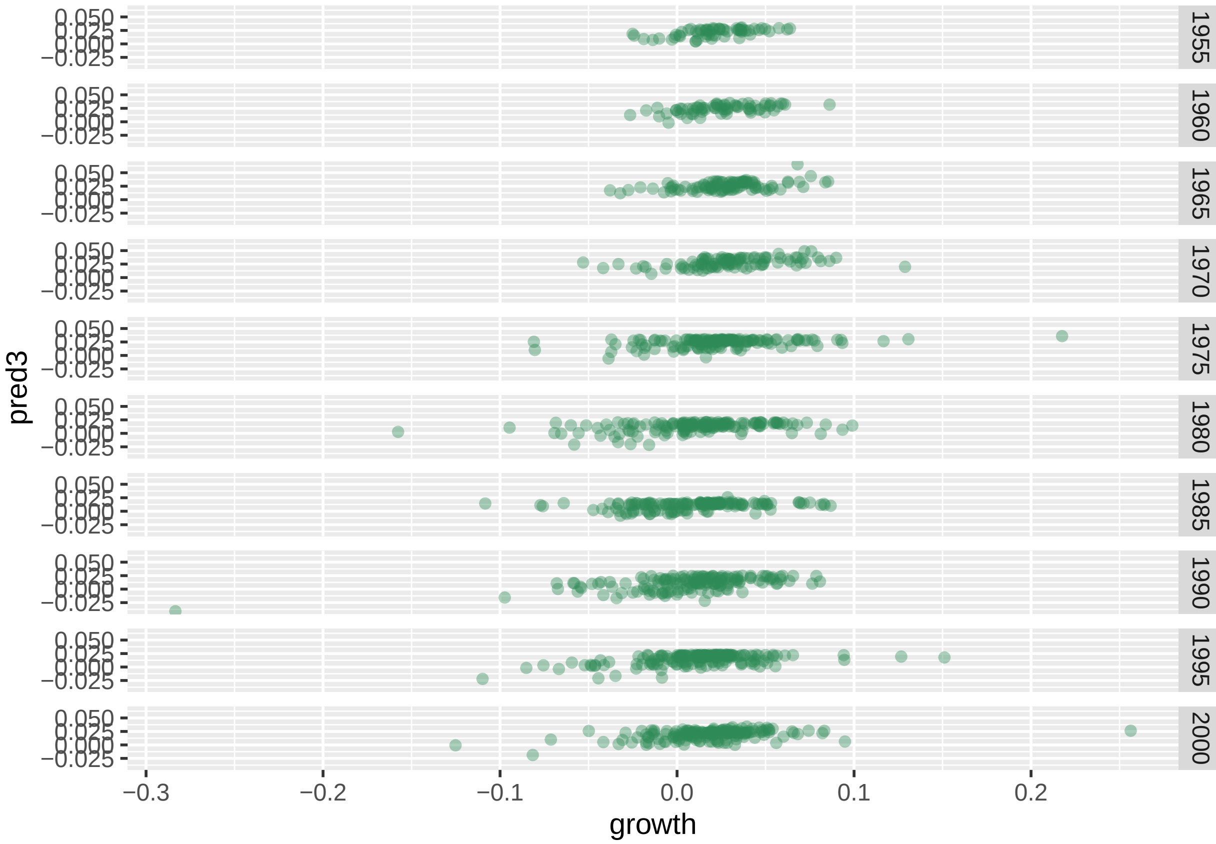
We can't obtain the coefficient of the kernel regression since the estimated response value is the weighted average of the value nearby.

Q4b)

```
tmp <- uval
tmp$pred3 <- predict(model3)
plot(uval$growth, predict(model3), col=alpha('seagreen', 0.4))
```



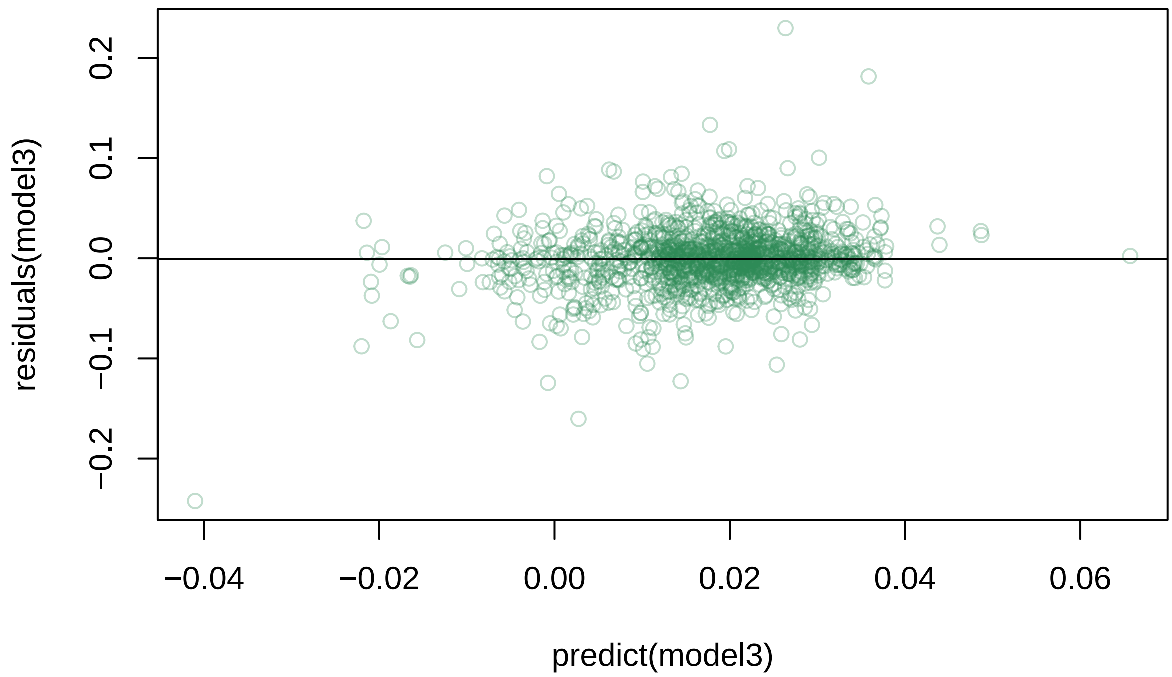

```
ggplot(tmp, aes(growth, pred3)) +
  geom_point(col=alpha('seagreen', 0.4)) +
  facet_grid(c("year"))
```




```
# facet_grid(c("year", "country"))
```

Q4c)

```
plot(predict(model3), residuals(model3), col=alpha('seagreen', 0.3))
abline(h=mean(residuals(model3)))
```



The points should be scattered around the residual mean 0 if the model is a right fit, which they are.

Q4d)

```
MSE2 <- with(uval, sum(growth-residuals(model2))^2)
MSE3 <- model3$MSE
# loocv.mse[2]
model3$bws$fval
```

```
## [1] 0.0009571853
```

Since MSE for model 3 is less than MSE for model 2, model 3 is a predict country growth better than model 2.

Question 5 -

Q5a)

Q5b)

Q5c)

Q5d)

Q5e)

Q5f)