

# 36-402 Homework 2

*Eu Jing Chua*

*eujingc*

*January 29, 2019*

## Question 1

Table 1: Coefficients & Std. Error of linear model

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.03525	0.00665	-5.30037	0.00000
underval	0.00476	0.00218	2.18614	0.02898
log(gdp)	0.00630	0.00079	7.96591	0.00000

Since the coefficient of  $\log(gdp)$  is significantly positive, this model does not seem to support the idea of “catching-up” as countries with higher GDP have a higher economic growth rate. However, it does support the idea that under-valuing a currency boosts economic growth as the coefficient of *underval* is quite significantly positive (assuming 5% significance level), indicating a positive *underval* index, which represents undervaluing, leads to higher economic growth.

## Question 2

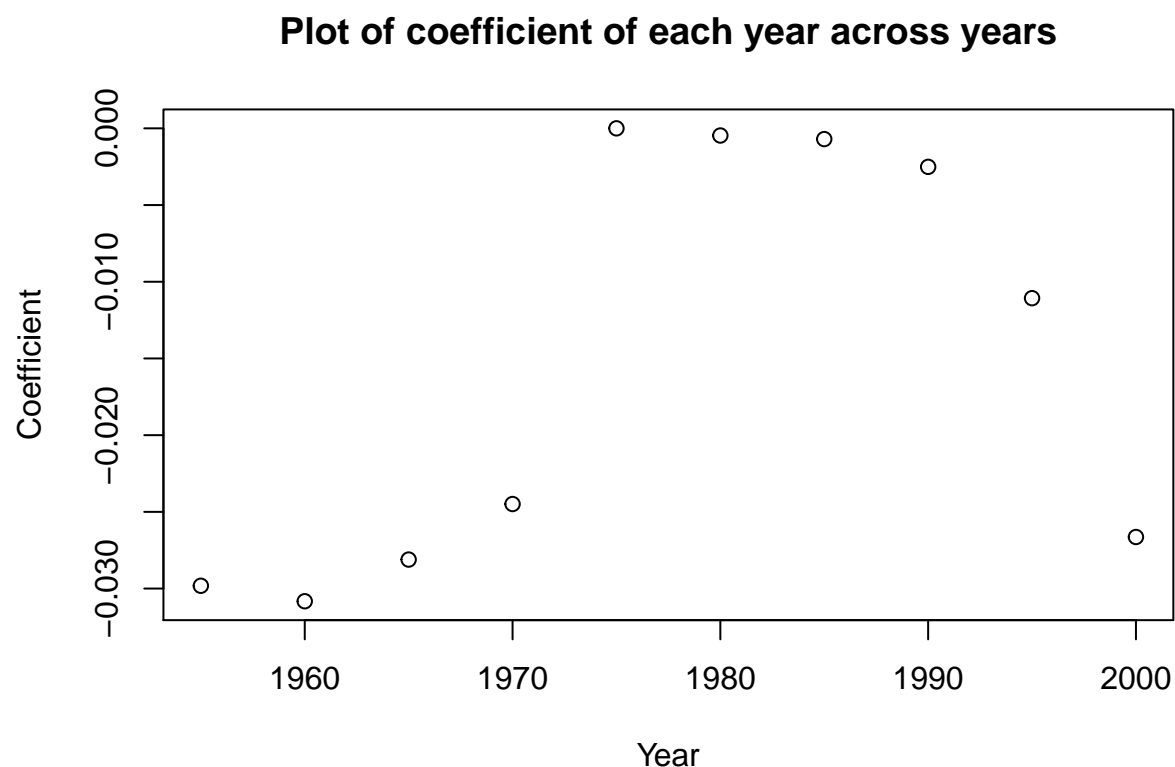
**Q2 a)**

Table 2: Coefficients & Std. Error of linear model

	Estimate	Std. Error	t value	Pr(> t )
underval	0.01361	0.00290	4.69667	0
log(gdp)	0.02892	0.00317	9.13254	0

**Q2 b)** It is more appropriate to use `factor(year)` as there are only 10 unique years that are 5 years apart. As such, it might be more appropriate to model year as a discrete variable rather than a continuous one. Modelling this way, we will have a slope for each 5-year interval rather than a single slope for each increment of year.

**Q2 c)**



**Q2 d)**

Since the coefficient of  $\log(gdp)$  is positive, this model does not seem to support the idea of “catching-up” as countries with higher GDP have a higher economic growth rate. However, it does support the idea that under-valuing a currency boosts economic growth as the coefficient of *underval* is positive, indicating a positive *underval* index, which represents undervaluing, leads to higher economic growth.

### Question 3

**Q3 a)**

Table 3:  $R^2$  values for each linear model

	Model 1	Model 2
$R^2$	0.04855	0.42924
Adj. $R^2$	0.04709	0.33214

**Q3 b)**

Table 4:  $M\hat{S}E$  of linear models by LOOCV

	x
Model 1	0.0010303

	x
Model 2	0.0009528

The second linear model seems to perform better on predictions, having an estimated MSE that is 7.53% lower than the first model.

### Q3 c)

Since the lowest count of any country is 2, doing 5-fold cross validation has a high chance of resulting in the training 4-folds missing some countries because the 2 rows are in the testing 1-fold that is left out. This would create models that leave out some countries by chance and fail to predict with a new category it has not seen in training.

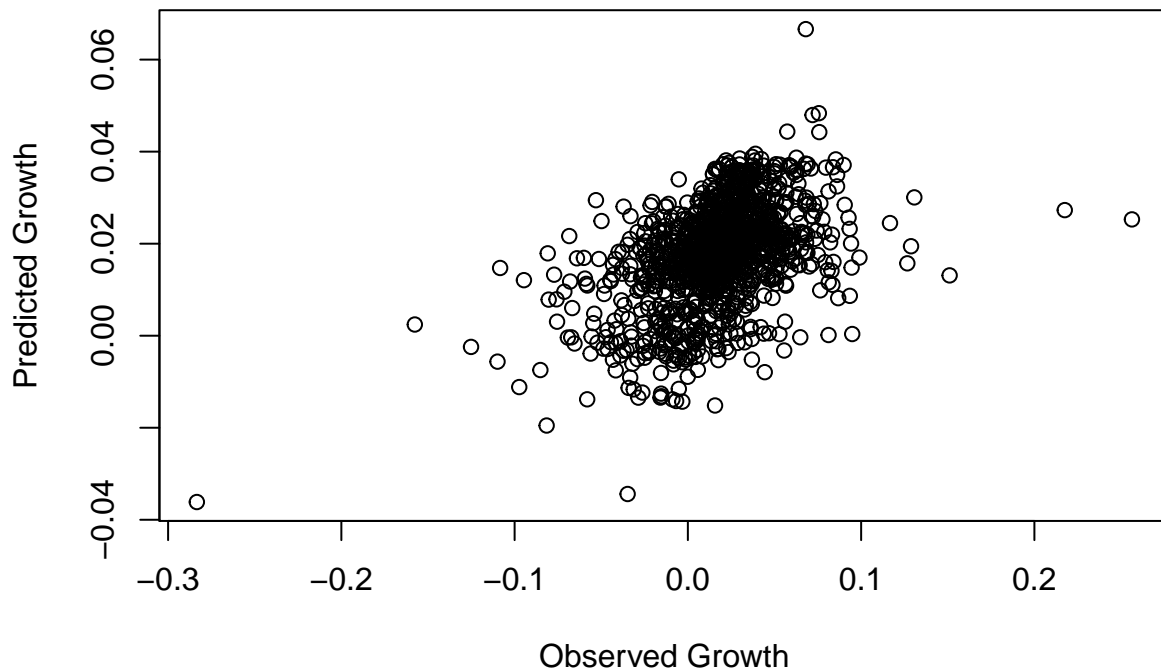
## Question 4

### Q4 a)

There are no coefficients to report for kernel regression, as it is a non-parametric smoothing method that smooths the data with a kernel. The smoothing is only controlled by the choice of the kernel and the related bandwidth for the kernel, without any coefficients on the predictors.

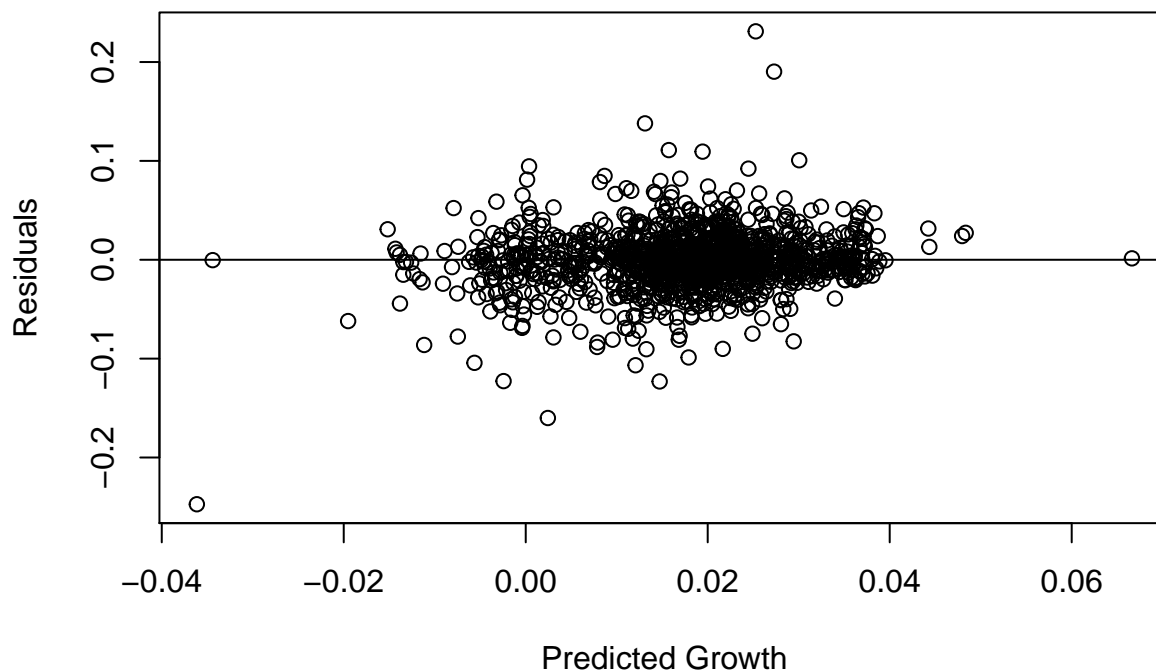
### Q4 b)

**Plot of Predicted Growth against Observed Growth**



### Q4 c)

### Plot of Residuals against Predicted Growth



The points should be scattered around a flat line at 0 if the model was right, as we assume  $\mathbb{E}[\epsilon] = 0$ . In this case, it would seem that the residuals indeed are roughly scattered around a flat line at 0.

Q4 d)

Table 5:  $\hat{MSE}$  of the regressions via cross validation

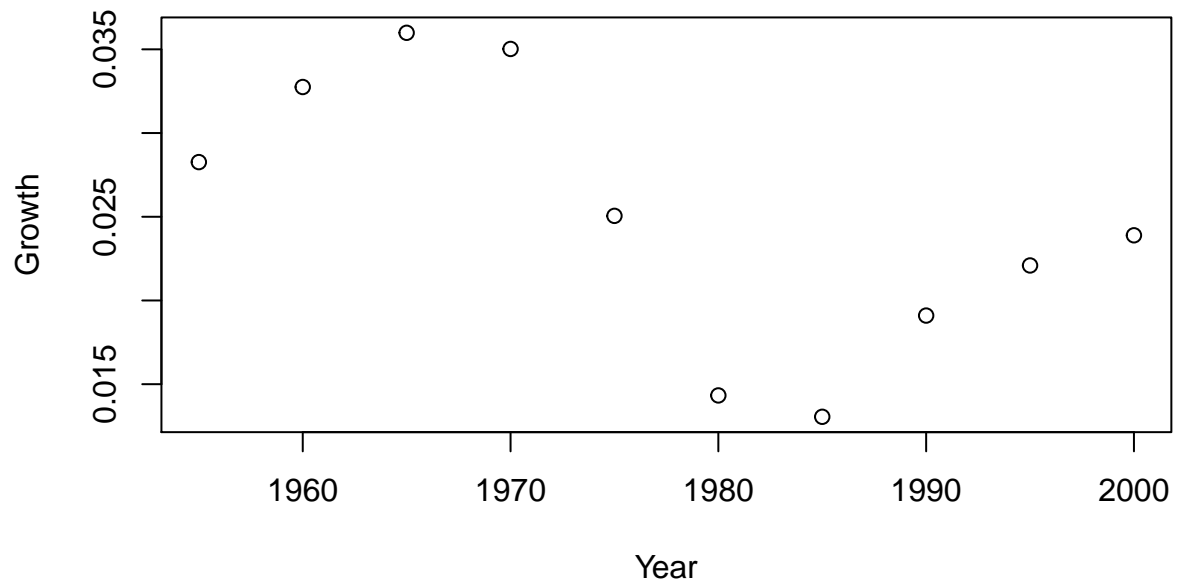
	x
Linear Regression	0.0009688
Kernel Regression	0.0009481

As seen from above, kernel regression has a lower estimated MSE than a linear model with the same covariates. Hence, the kernel regression is better in generalizing and predicting better than the linear model.

## Question 5

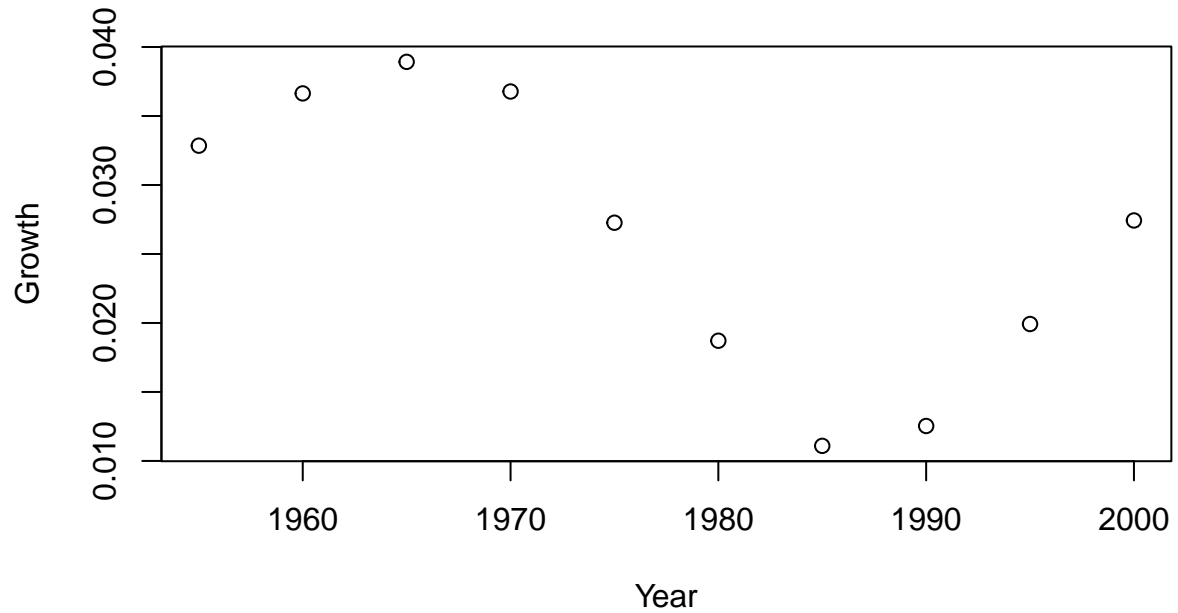
Q5 a)

**Predicted growth with gdp = 20000, underval = 0**



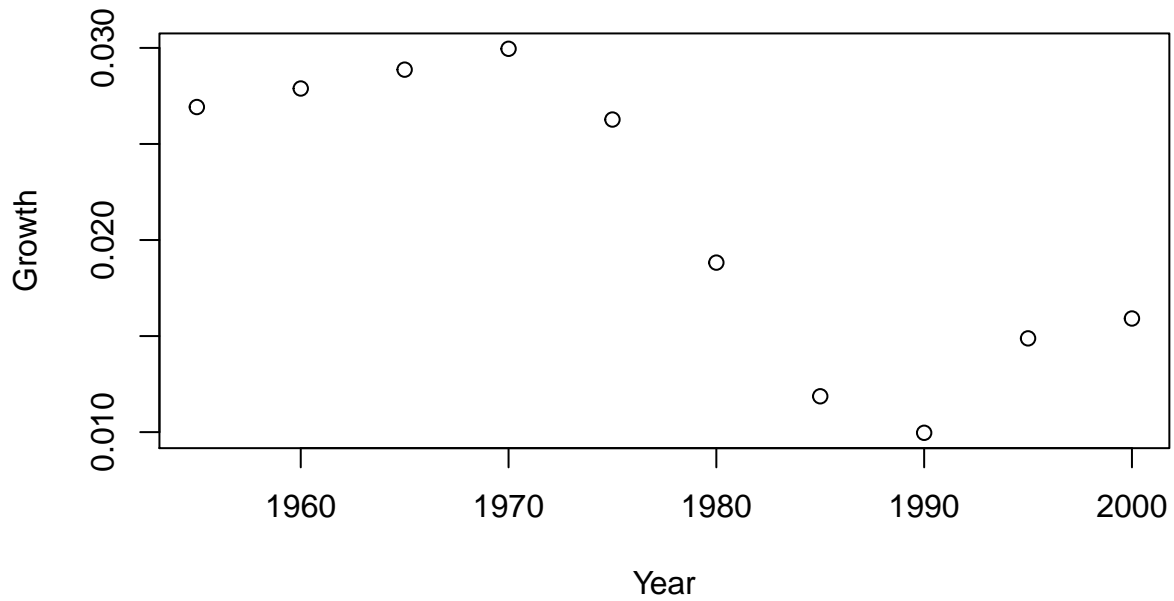
Q5 b)

**Predicted growth with gdp = 20000, underval = +0.5**



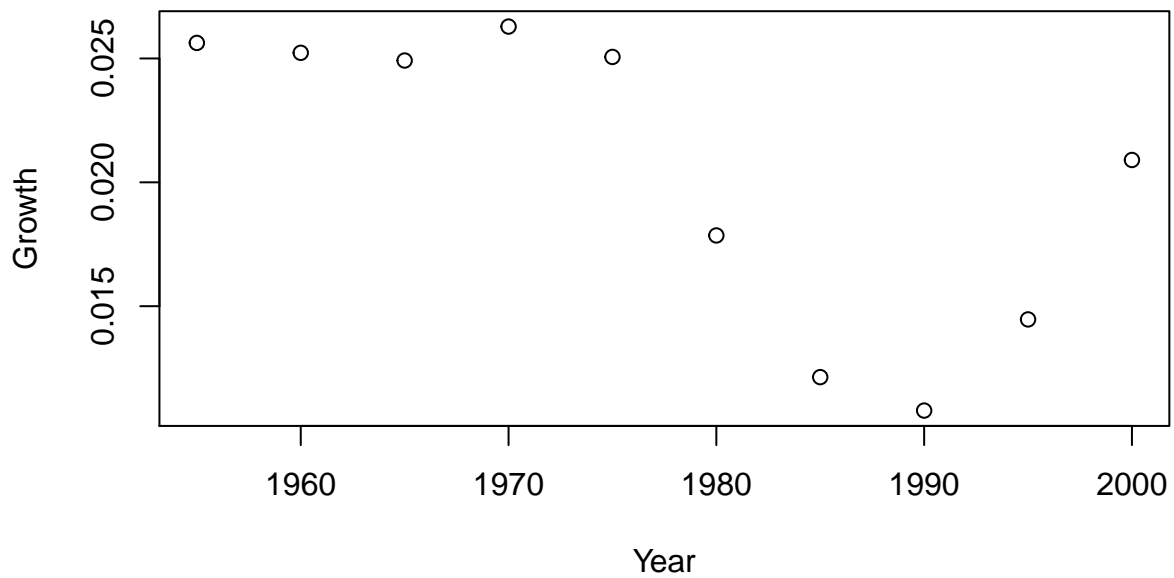
Q5 c)

**Predicted growth with gdp = 3000, underval = 0**



Q5 d)

**Predicted growth with gdp = 3000, underval = +0.5**

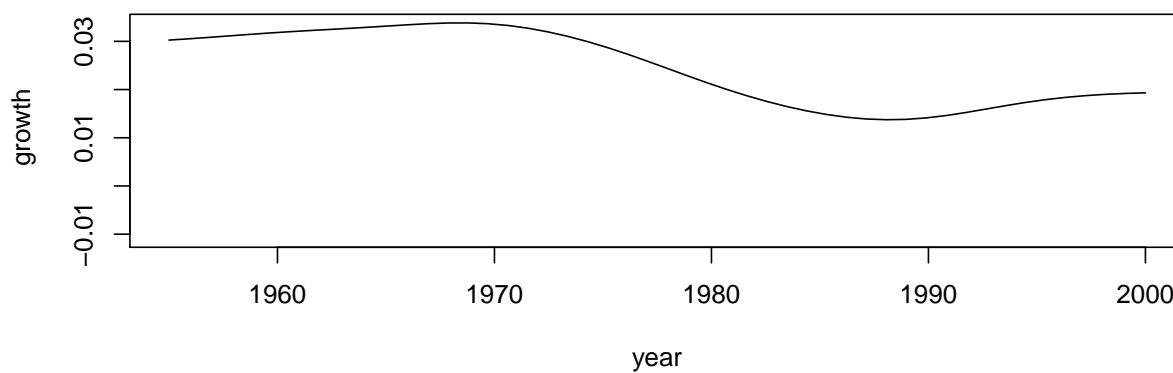
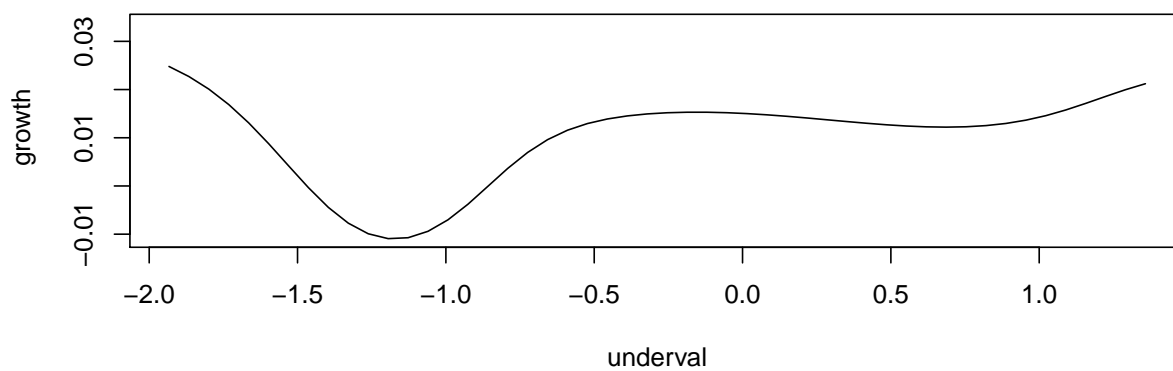
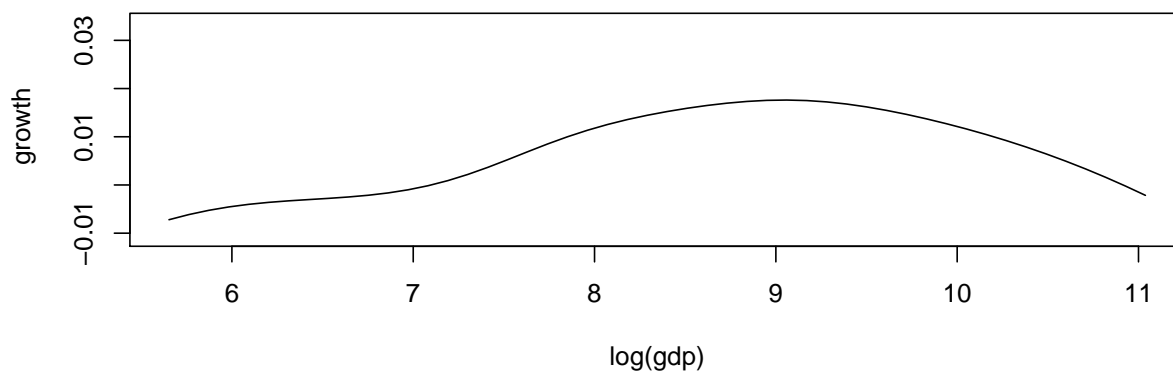


Q5 e)

By comparing the graphs with initial  $GDP = 20000$ , we can see that by increasing under-valuation from 0 to 0.5, the predicted values increased from 1955 to 1970 but for  $GDP = 3000$ , increasing under-valuation

from 0 to 0.5 decreased the predicted values from 1955 to 1970. The graphs are not parallel across and the effects not similar when we change initial GDP and under-valuation, hence there should be some interaction between these two variables.

Q5 f)



From the plots of growth against each variable, we can see that both GDP and under-valuation have strong relationships with growth. For the plot of growth against  $\log(GDP)$ , there seems to be significant variation from varying  $\log(GDP)$  from 7 to 11. For the plot of growth against under-valuation, there also seems to be significant variation from varying under-valuation from -2.0 to -0.5. Finally, there is also some relationship, a weaker one than the previous two, between growth and year. For the plot of growth against year, there seems to be a slightly significant variation from varying year from 1970 to 1985.

## Question 6

$$\text{Optimism} = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n (Y'_i - \hat{\mu}(x_i))^2 \right] - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\mu}(x_i))^2 \right] \quad (1)$$

$$= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[ (Y'_i - \hat{\mu}(x_i))^2 \right] - \mathbb{E} \left[ (Y_i - \hat{\mu}(x_i))^2 \right] \quad (2)$$

$$= \frac{1}{n} \sum_{i=1}^n \text{Var} [Y'_i - \hat{\mu}(x_i)] + \mathbb{E} [Y'_i - \hat{\mu}(x_i)]^2 - \text{Var} [Y_i - \hat{\mu}(x_i)] - \mathbb{E} [Y_i - \hat{\mu}(x_i)]^2 \quad (3)$$

$$= \frac{1}{n} \sum_{i=1}^n \text{Var} [Y'_i] + \text{Var} [\hat{\mu}(x_i)] - 2\text{Cov} [Y'_i, \hat{\mu}(x_i)] - \text{Var} [Y_i] - \text{Var} [\hat{\mu}(x_i)] + 2\text{Cov} [Y_i, \hat{\mu}(x_i)] \quad (4)$$

$$+ (\mathbb{E} [Y'_i] - \mathbb{E} [\hat{\mu}(x_i)])^2 - (\mathbb{E} [Y_i] - \mathbb{E} [\hat{\mu}(x_i)])^2 \quad (5)$$

$$= \frac{1}{n} \sum_{i=1}^n 2\text{Cov} [Y_i, \hat{\mu}(x_i)], \text{ as } \text{Var} [Y'_i] = \text{Var} [Y_i], \mathbb{E} [Y'_i] = \mathbb{E} [Y_i], \text{ and } \text{Cov} [Y'_i, \hat{\mu}(x_i)] = 0 \quad (6)$$

$$= \frac{2}{n} \sum_{i=1}^n \text{Cov} \left[ Y_i, \sum_{j=1}^n w(x_j, x_i) Y_j \right] \quad (7)$$

$$= \frac{2}{n} \sum_{i=1}^n w(x_i, x_i) \text{Var} [Y_i], \text{ as } \text{Cov} [Y_i, Y_j] = 0 \forall i \neq j \quad (8)$$

$$= \frac{2\sigma^2}{n} \text{tr}(\mathbf{w}) = \frac{2\sigma^2}{n} df(\hat{\mu}) \quad (9)$$

## Question 7

$$\hat{h}_{CV} = h_{opt} O(n^{-1/10}) + 2h_{opt} \quad (10)$$

$$= O(n^{-1/5}) O(n^{-1/10}) + O(n^{-1/5}) \quad (11)$$

$$= O(n^{-3/10}) + O(n^{-1/5}) \quad (12)$$

$$= O(n^{-1/5}) \quad (13)$$

$$MSE(\hat{h}_{CV}) - \sigma^2(x) = O(\hat{h}_{CV}^4) + O((nh_{CV})^{-1}) \quad (14)$$

$$= O((n^{-1/5})^4) + O(((n^{-1/5}))^{-1}) \quad (15)$$

$$= O(n^{-4/5}) + O(n^{-4/5}) \quad (16)$$

$$= O(n^{-4/5}) \quad (17)$$