Empirical Project 3 | ECON 270

Group Members:

Shalin Luitel

Sweta Nanadakumar

Jahnavi Kalyan

Arnold Gyateng

1. Names of variables used for the project:

rate_black_population_2010
count_person_enrolledincollegeun
count_person_enrolledinschool
unemploymentrate_person
median_income_household
count_housingunit
genderincomeinequality_person_15
median_income_person
count_household_05orlessratiotop
count_household_nohealthinsuranc

Dropped variable “dcx6l3mg60hpe7c” using *drop dcx6l3mg60hpe7c*

2. Summary Statistics for kfr_pooled_p25 and variables used.

```
. sum kfr_pooled_p25 rate_black_population_2010 count_person_enrolledincollegeun count_person_enrolledinschool unemploymentrat
> e_person median_income_household count_housingunit genderincomeinequality_person_15 median_income_person count_household_05o
> rlessratiotop count_household_nohealthinsuranc
```

Variable	Obs	Mean	Std. Dev.	Min	Max
kfr_poole~25	1,259	.4132139	.0517072	.2128646	.6140298
rate_bl~2010	2,518	.1008264	.1444777	.0001027	.8338889
count_pers~n	2,518	.0466533	.0351332	.0070801	.4086202
count_pers~l	2,518	.234954	.047066	.0957636	.5243892
unemployme~n	3,220	.0341267	.017293	.005	.162
median_inc~d	2,518	3.512239	2.885716	.0197511	23.24831
count_hous~t	2,518	1.024281	.0822837	.6446485	1.933008
genderinc~15	3,220	.2277109	.0714873	-.424122	.5568318
median_inc~n	2,518	.8536498	.6967384	.0030308	5.529024
count_hous~p	2,518	.0271955	.0202923	.0008629	.1892992
count_hous~c	989	.2499479	.1313378	.0454211	1.283477

3. In addition to the predictors already in the training dataset inspect the result, and comment on what you find?

```
. regress kfr_pooled_p25 rate_black_population_2010 count_person_enrolledincollegeun count_person_enrolledinschool unemploymentrate_pers
> on median_income_household count_housingunit genderincomeinequality_person_15 median_income_person count_household_05orlessratiotop co
> unt_household_nohealthinsuranc
```

Source	SS	df	MS		
				Number of obs	= 498
				F(10, 487)	= 52.70
Model	.372561537	10	.037256154	Prob > F	= 0.0000
Residual	.344307598	487	.000706997	R-squared	= 0.5197
				Adj R-squared	= 0.5098
Total	.716869135	497	.001442393	Root MSE	= .02659

kfr_pooled_p25	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rate_black_population_2010	-.1384719	.0102434	-13.52	0.000	-.1585987	-.1183451
count_person_enrolledincollegeun	-.0815554	.0643576	-1.27	0.206	-.2080081	.0448974
count_person_enrolledinschool	.1655877	.0584825	2.83	0.005	.0506784	.2804969
unemploymentrate_person	.443496	.098501	4.50	0.000	.2499565	.6370355
median_income_household	.003821	.0055012	0.69	0.488	-.006988	.01463
count_housingunit	-.0276306	.0189981	-1.45	0.146	-.0649589	.0096977
genderincomeinequality_person_15	.1096356	.0285502	3.84	0.000	.0535389	.1657323
median_income_person	-.0291165	.0258405	-1.13	0.260	-.0798892	.0216562
count_household_05orlessratiotop	-1.263531	.1270193	-9.95	0.000	-1.513104	-1.013957
count_household_nohealthinsuranc	.0355552	.0120128	2.96	0.003	.011952	.0591585
_cons	.3994056	.0199089	20.06	0.000	.3602876	.4385237

In the regression analysis table for kfr_pooled_p25 vs datacommons variables, we see 498 total observations with an R-squared value of 0.52. Out of 10 variables used to predict kfr_pooled_p25, 6 variables are statistically significant for an alpha value of 0.05 which 4 variables fail to reject the null hypothesis of no relationship of variables with kfr_pooled_p25. We expect kfr_pooled_p25 value of 0.40 when all of the explanatory variables are 0, as shown by _cons, significant to alpha value of 0.05.

4. How well does your linear regression predict kfr_pooled_p25 in sample? You have to calculate the mean square error.

```
sum mse_forest if training == 1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mse_forest	498	.0001278	.0001909	3.07e-11	.0012536

The Mean square error is $1.28*10^{-4}$. The mean square error is very small (insignificant) in our model prediction. Therefore, this linear regression to predict kfr_pooled_p25 from the sample variables is robust with significant prediction strength.