Part 1

Here I use a version of the Card Krueger minimum wage study data set. This has data for fast food restaurants in New Jersey (NJ) and Pennsylvania (PA), collected in 2 waves before and after the minimum wage was raised from $4.25 to $5.05 in NJ (it remained constant at $4.25 in PA). I am studying the changes in wages and employment, first focusing on between-state differences, then on differences at the store-level using various econometric methods.

First, I look at the state-wide changes. The following table shows the means of various variables of interest on a state level and the difference between state means.

pchwage is the percent change in wage. pchemp is arc-percent change in employment.

TABLE 1

nj	NJ	PA	NJ - PA
wage_st	4.612982	4.653636	-0.040654
wage_st2	5.082140	4.618788	0.463352
pchwage	0.107230	-0.004168	0.111399
emptot	20.678246	23.704545	-3.026300
emptot2	21.076316	21.825758	-0.749442
pchemp	0.022006	-0.032929	0.054935

We can see that starting wage in NJ grew faster than in PA. (starting wage actually declined in PA). NJ earlier lagged PA bit a little bit in starting wage. At wave 2, it has higer mean wage just above its new minimum wage.

We can also see that the employment gap in wave 2 decreased between NJ and PA. NJ trailed PA by -3.026. This reduced to -0.749442. This is beacuse NJ increases in empoyment while PA decreases.

First, I will be using the dummy variable NJ as an instrumental variable for PCHWAGE to predict PCHEMP. (NJ is 1 is a store is in NJ otherwise 0).

First Stage

$$PCHWAGE_i = \gamma_0 + \gamma_1 NJ_i + \epsilon_i$$

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0042	0.010	-0.428	0.669	-0.023	0.015
nj	0.1114	0.011	10.302	0.000	0.090	0.133

The coefficient $\hat{\gamma_1}$ on nj is 0.1114

Reduced Form

$$PCHEMP_i = \rho_0 + \rho_1 NJ_i + \phi_i$$

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0329	0.043	-0.757	0.449	-0.118	0.053
nj	0.0549	0.048	1.139	0.256	-0.040	0.150

The coefficient $\hat{\rho_1}$ on nj is 0.0549

Causal Model of interest

$$PCHEMP_i = \beta_0 + \beta_1 PCHWAGE_i + \mu_i$$

The IV results for the the above model.

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-0.0309	0.0455	-0.6786	0.4974	-0.1200	0.0583
pchwage	0.4931	0.4617	1.0680	0.2855	-0.4118	1.3981

$$\hat{\beta^{IV}} = \frac{\hat{\rho_1}}{\hat{\gamma_1}} = \frac{0.0549}{0.1114} = 0.49281867145421904$$

Therefore, the IV matches the 2SLS estimate.

The first stage shows us that NJ has 11.1399% excess change in wage. The reduced form captures the excess change in employment of NJ which is 0.0549. Thus the IV estimate is actually measuring the ratio of differce in change emoployment to difference in change in wage between NJ and PA. Rather than m asuring the causal impact of wages and employment at the store level. This is to do with NJ being a bad instrumental variable. Exclusion restriction is not satisfied as we know that NJ has had a minimum wage increase which is correlated with our wage covariate and possibly employment (the outcome variable).

I will again try to measure the same causal Model of interest

$$PCHEMP_i = \beta_0 + \beta_1 PCHWAGE_i + \mu_i$$

But this time using GAP as an instrumental variable. GAP is the proportional increase in the starting wages needed to meet the new minimum wage. Table 2 shows the results of the OLS, 2SLS and IV models.

TABLE 2

	OLS (7)	First Stage (8)	Reduced Form (9)	IV2SLS
const	-0.024	-0.002	-0.032	-0.031
	(0.026)	(0.003)	(0.028)	(0.028)
pchwage	0.418**			0.494**
	(0.208)			(0.238)
gap		1.041***	0.514**	
		(0.031)	(0.247)	
Observations	351	351	351	351
R^2	0.011	0.768	0.012	0.011
Adjusted R^2	0.009	0.768	0.009	0.008
Residual Std. Error	0.352	0.044	0.352	0.352
F Statistic	4.032**	1157.848***	4.326**	4.321**

Note:

The OLS estimate is very close to the IV estimate. The first state model measures the relationship between percent change in wage to percent chnage required to meet minimum wage. It stands to reason that employers will generally increase only to meet mimiume wage (and not much more). Hence the estimate is close to 1. Since π_1 is close to 1, the OLS estimate β_1 and the reduced form estimate δ_1 are very close to each other. Thus its more or less precisely estimated.

We can verify $\frac{\hat{\delta_1}}{\hat{\pi_1}} = \frac{0.514}{1.041} = 0.494 = \hat{\beta_1^{IV}}$

Checking a concern that stores in NJ and PA are not exactly the same.

Causal model of interest

$$PCHEMP_i = \beta_0 + \beta_1 PCHWAGE_i + \beta_2 NJ_i + \mu_i$$

Table 3 shows the results of the OLS, 2SLS and IV models.

TABLE 3

	OLS (10)	First Stage (11)	Reduced Form (12)	IV2SLS
const	-0.031	-0.004	-0.033	-0.031
	(0.043)	(0.005)	(0.043)	(0.043)
pchwage	0.396*			0.494*
	(0.238)			(0.285)
gap		1.031***	0.510*	
		(0.036)	(0.294)	
nj	0.011	0.004	0.002	-0.000
	(0.055)	(0.007)	(0.057)	(0.058)
Observations	351	351	351	351
R^2	0.012	0.769	0.012	0.011
Adjusted R^2	0.006	0.767	0.007	0.005
Residual Std. Error	0.352	0.044	0.352	0.352
F Statistic	2.030	577.840***	2.157	2.155

Note:

In the above table, we have controlled for a store's state. This is to test an assumption of whether NJ store are systematically different from PA. However, we can see that the coefficient estimates for a NJ indicator are insignificant on a 5% level. We also see that the point estimates for pchwage and gap are relastively similar to the model without account for a state indicator. With this, we can safely say that controlling for minimum wage, NJ and PA stores are realtively the same without any big systematic differneces.

Recreating TABLE 2 but with only the data from NJ

We can see that if we subsample to just NJ stores, the estimates for gap and pchwage stay relatively similar to table 3 (and by extension table 2). This makes sense as we saw that using NJ as a indicator for state had no significant effect on the estimates. Thus, if we account for the minimum wage changes in NJ, the employement and wage relationship at a NJ store level should look similar to that of a population of both NJ and PA.

Checking a model which accounts for local demand shocks.

Causal model of interest

$$PCHEMP_i = \beta_0 + \beta_1 PCHWAGE_i + \sum_{r=1}^{5} \beta_r^R Region_{ri} + \mu_i$$

Table 5 shows the results of the OLS, 2SLS and IV models.

TABLE 5

	OLS (13)	First Stage (14)	Reduced Form (15)	IV2SLS
const	-0.110*	0.005	-0.108	-0.110*
	(0.067)	(0.008)	(0.067)	(0.067)
pchwage	0.402*			0.489*
	(0.239)			(0.286)
gap		1.031***	0.504*	
		(0.036)	(0.295)	
southj	0.105	-0.006	0.093	0.096
	(0.083)	(0.011)	(0.085)	(0.084)
centralj	0.049	-0.005	0.037	0.040
	(0.086)	(0.011)	(0.088)	(0.088)
northj	0.104	-0.004	0.094	0.095
	(0.076)	(0.010)	(0.078)	(0.078)
shore	-0.049	-0.006	-0.051	-0.048
	(0.069)	(0.009)	(0.069)	(0.069)
pa2	0.137	-0.016	0.130	0.138
	(0.088)	(0.011)	(0.088)	(0.088)
Observations	351	351	351	351
R^2	0.023	0.771	0.024	0.023
Adjusted R^2	0.006	0.767	0.007	0.006
Residual Std. Error	0.352	0.044	0.352	0.352
F Statistic	1.366	192.517***	1.384	1.383

Note:

We can see here that even after controling for local demand shocks, the estimates for pcwage and gap have note really changed. So, we can ignore local demand shocks while trying to esitmate the impact of minimum wage.

Using all the possible control variables in the data set. Table 6 shows the results of the OLS, 2SLS and IV models.

```
Regressions with all covariates
roys dummy was not selected among the store dummies. bk, kfc and wendys w
ere selected
```

TABLE 6

	OLS	First Stage	Reduced Form	IV2SLS
const	-1.389**	0.017	-1.392**	-1.399**
	(0.653)	(0.084)	(0.653)	(0.654)
pchwage	0.349			0.421
	(0.255)			(0.309)
gap		1.030***	0.434	
		(0.041)	(0.319)	
Observations	323	323	323	323
R^2	0.066	0.776	0.066	0.066
Adjusted R^2	0.004	0.761	0.004	0.004
Residual Std. Error	0.352	0.045	0.352	0.352
F Statistic	1.067	52.254***	1.066	1.066

Note:

Again, we see that controlling for all these other factors does not affect the relationship between wage and employment. That is, factors like location, store brand, number of cash reigsters, etc do not affect or help explain the relationship between wage and employment.

Concerned with over-fitting the model, trying a “double ML” approach in which I use lasso models and cross-fitting to partial out the effects of the control variables when estimating the effect of GAP in the first stage and reduced form models.

```
Step 2:  cross-fit partialing-out double ML with lasso model and all cova
riates to prevent over-fitting
Causal OLS model
```

	coef	std err	t	P>|t|	2.5 %	97.5 %
pchwage	0.364843	0.212048	1.720566	0.08533	-0.050764	0.78045

```
First Stage
```

	coef	std err	t	P>|t|	2.5 %	97.5 %
gap	1.049386	0.034646	30.288972	1.601643e-201	0.981481	1.11729

Reduced Form

	coef	std err	t	P>|t|	2.5 %	97.5 %
gap	0.465177	0.259965	1.789381	0.073553	-0.044346	0.974699

TABLE 7

	Causal OLS	First Stage	Reduced Form
pchwage	0.376655	NaN	NaN
std error	0.034740	NaN	NaN
gap	NaN	1.049699	1.049699
std error	NaN	0.034740	0.034740

Part II

Using data from a paper by Card, Dobkin and Maestas studying how use of health services changes at age 65, as people in the US become (nearly) universally eligible for Medicare.

Mean values of covered, mcare, inhosp and sawdr for each age.

inhosp = 1 if had an overnight stay in hospital in last year

sawdr = 1 if had a visit with a doctor in last year

covered = 1 if any form of health insurance

mcare = 1 if have medicare insurance

variable age4 is measured in quarters

age4	covered	mcare	inhosp	sawdr
55.25	0.872848	0.047731	0.083790	0.800729
55.50	0.877742	0.041430	0.093165	0.799742
55.75	0.880615	0.044917	0.085962	0.810828
56.00	0.871956	0.042419	0.096622	0.797892
56.25	0.876936	0.048492	0.081566	0.805112
...	...	...	...	...
73.75	0.991077	0.942001	0.174745	0.894097
74.00	0.991747	0.951169	0.176187	0.890255
74.25	0.993464	0.949673	0.175932	0.901109
74.50	0.995935	0.942412	0.182373	0.918463
74.75	0.992100	0.953917	0.190633	0.913043

79 rows × 4 columns

Figure 2.1: Proportion covered by any form of health insurance

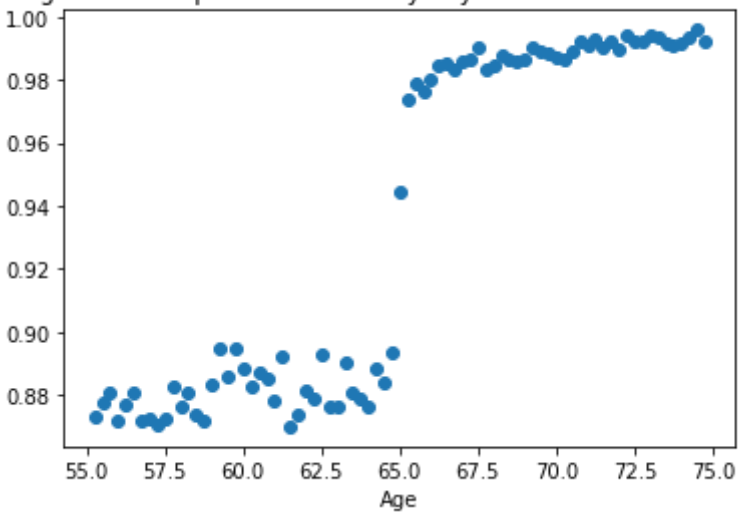


Figure 2.2: Proportion having have medicare insurance

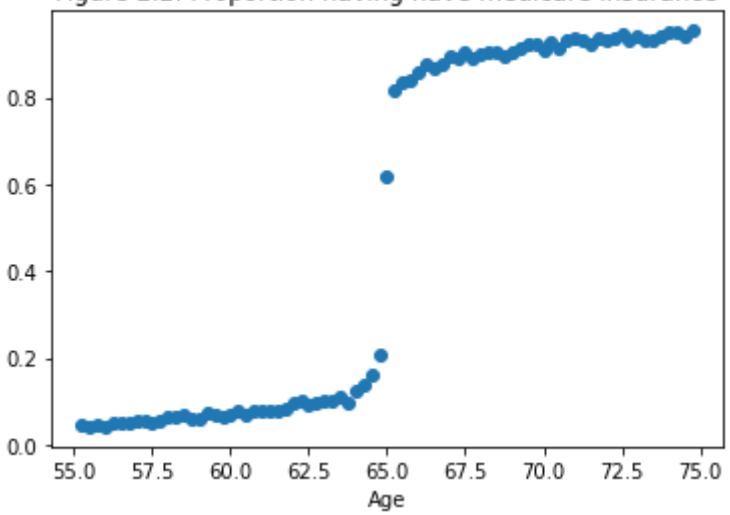
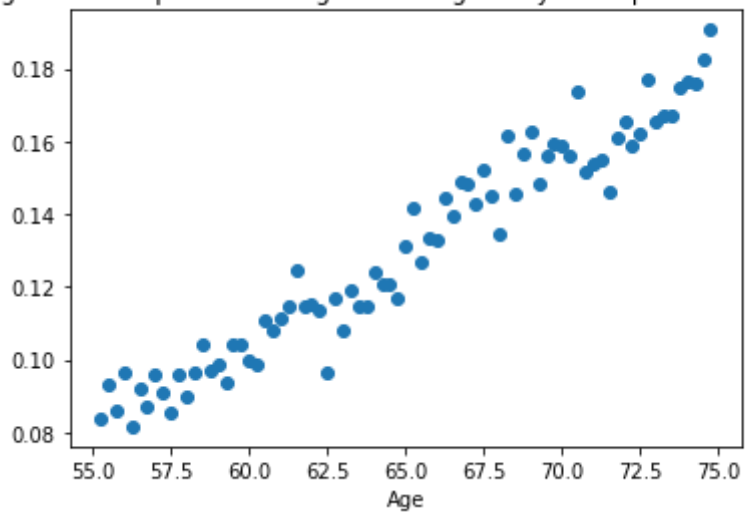
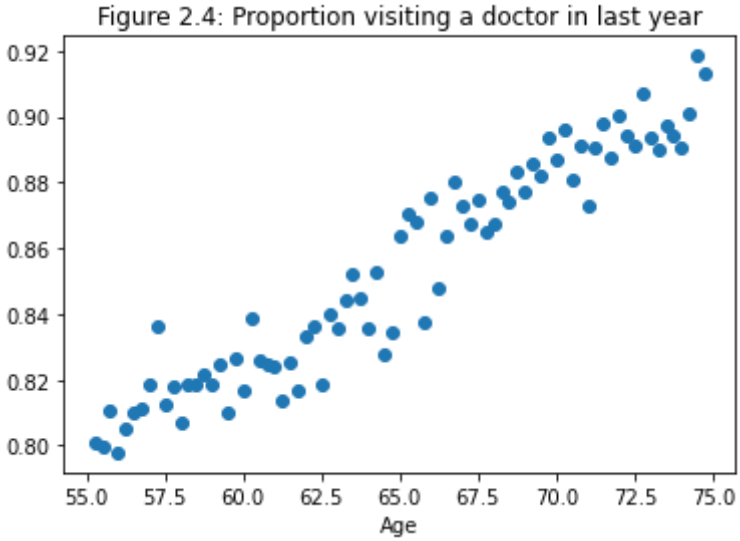


Figure 2.3: Proportion having an overnight stay in hospital in last year





Based on the figures, insurace coverage dramatically increases after age 65. We also see the proportion being covered by medicare increases even more dramatically at 65. This might indicate the people switch from private or other insurance to medicare. The jump at exactly 65 is interesting. These correspond to people who just turnes 65 (and less than a quarter past 65). This jump in the middle of the two dircerable slopes shows that there might be a lag in errolling in medicare insurance.

However, there is no visually discernable impact of turning 65 on the probability of visits to the doctor and the hospital. The variables for doctor and hopistal visits are mesearued over the year. Thus, there could be a smoothing out effect as age is measured over quarters. If the doctors and hospital visits were measured over 3 months, it might show a jump.

Adding new variables:

$r_i = age4 - 65$ as the running variable for an RD analysis (i.e. “recentered” age)

$z_i = 1[age4 \geq 65]$ - this is the dummy for passing the eligibility threshold. And define

$w_i = r_i \times z_i$ (this is variable r_z)

A first stage local linear model for coverage is:

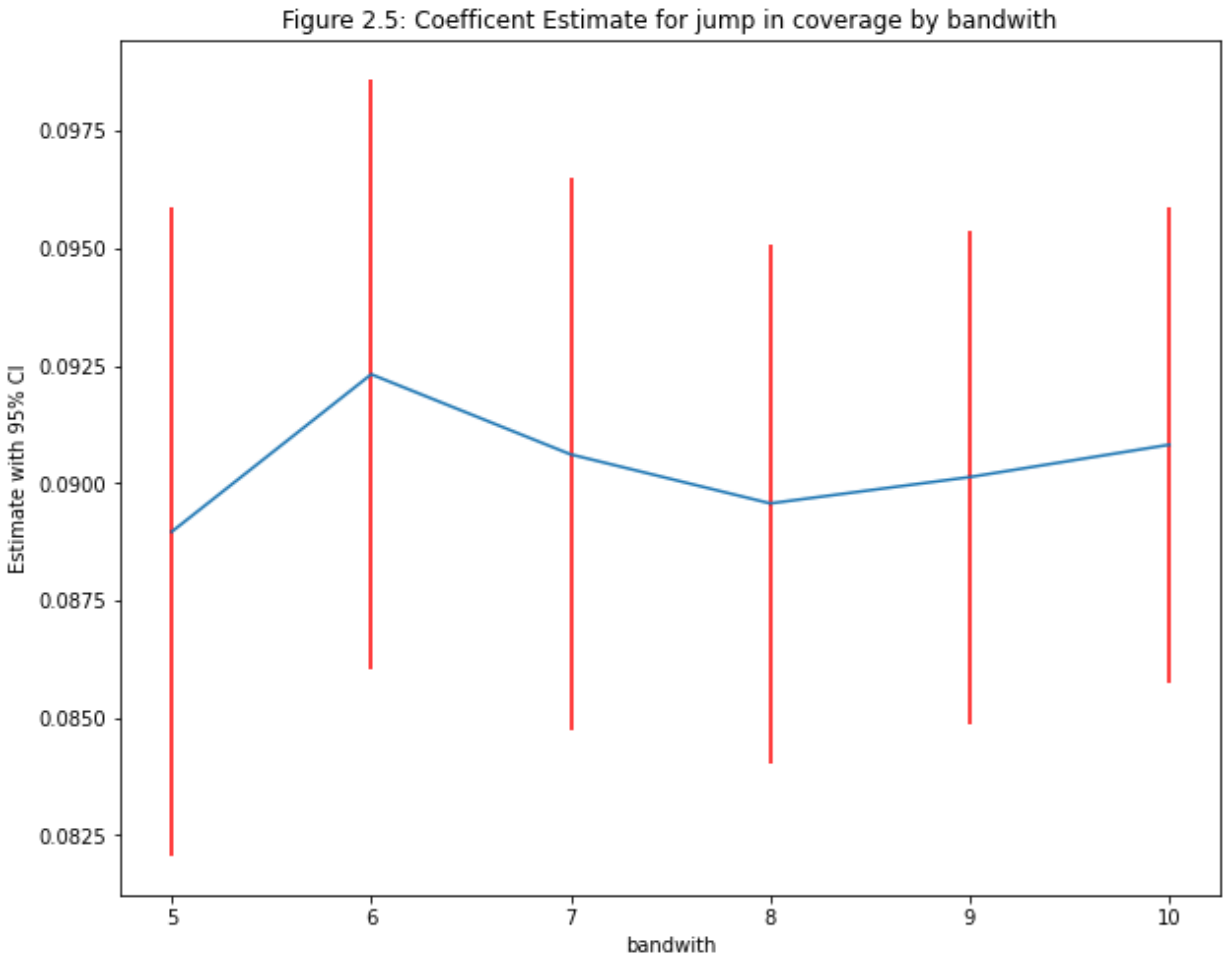
$$D_i = \pi_0 + \pi_1 z_i + \pi_2 r_i + \pi_3 w_i + \eta_i$$

In this model, the coefficient π_1 measures the “jump” in D at age 65, the slope of a local linear model for coverage to the left of the 65 cutoff is π_2 ,and the slope to the right is $\pi_2 + \pi_3$.

Running the above model with a bandwith of 10 years.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8860	0.002	482.769	0.000	0.882	0.890
z	0.0908	0.003	35.110	0.000	0.086	0.096
r	0.0011	0.000	3.422	0.001	0.000	0.002
r_z	0.0011	0.000	2.358	0.018	0.000	0.002

Re-estimating the first stage model using bandwidths of 5,6,7,8,9,10



Re-estimating the model with a bandwidth of 10 years, omitting the data for the people with age 65.0, i.e, people who jsut turned 65 and are less than 65 and a quarter in age.

Out[39]:

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8860	0.002	482.298	0.000	0.882	0.890
z	0.0945	0.003	35.510	0.000	0.089	0.100
r	0.0011	0.000	3.419	0.001	0.000	0.002
r_z	0.0005	0.000	1.073	0.283	-0.000	0.001

At 65, the probablity of having health insurace increases by about 9% points. At bandwidth of 10,

the estimate is 0.0908. This estimate seems to be robust to different bandwiths. If we omit the people who just tunred 65 (and less than a quarter past 65), the estimate increases to 0.0945. This is higher than if we include people who are 65. However, the 95% confidence intervals do overlap.

Using “discrete” running variables for a local quadratic model. Defining:

$r_i^2 = r_i \times r_i$

$w_i^2 = w_i \times w_i = r_i 2 \times z_i (\text{since} z_i^2 = z_i)$

The local quadratic model is:

$D_i = \pi_0 + \pi_1 z_i + \pi_2 r_i + \pi_3 w_i + \pi_4 r_i^2 + \pi_5 w_i^2 + \eta_i$

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8833	0.003	308.601	0.000	0.878	0.889
z	0.0871	0.004	22.434	0.000	0.079	0.095
r	-0.0005	0.001	-0.359	0.719	-0.003	0.002
r_z	0.0068	0.002	3.731	0.000	0.003	0.010
r2	-0.0001	0.000	-1.215	0.225	-0.000	9.12e-05
w2	-0.0003	0.000	-1.618	0.106	-0.001	6.07e-05

The estimate here is lower than the one in Figure 2.5 (with a bandwith of 10). This is less precise with a std. error of 0.004, compared to 0.003 in the one in figure 2.5

Checking the validity of the RD design: checking for discontinuity in the exogenous characteristics at the RD threshold

Model

$x_i = \rho_0 + \rho_1 z_i + \rho_2 r_i + \rho_3 w_i + \phi_i$

$x_i = college$

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1663	0.003	59.613	0.000	0.161	0.172
z	0.0023	0.004	0.589	0.556	-0.005	0.010
r	-0.0071	0.000	-15.177	0.000	-0.008	-0.006
r_z	0.0034	0.001	4.982	0.000	0.002	0.005

$x_i = \text{white non-hispanic}$

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.7490	0.003	237.939	0.000	0.743	0.755
z	-0.0018	0.004	-0.417	0.677	-0.011	0.007
r	0.0027	0.001	5.139	0.000	0.002	0.004
r_z	0.0022	0.001	2.796	0.005	0.001	0.004

$x_i = \text{hispanic}$

$x_i = \text{minority}$

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2212	0.003	73.467	0.000	0.215	0.227
z	0.0020	0.004	0.468	0.640	-0.006	0.010
r	-0.0021	0.001	-4.149	0.000	-0.003	-0.001
r_z	-0.0022	0.001	-2.973	0.003	-0.004	-0.001

Race based variables dont show significant discontinuity around 65. The estimate are close to zero and the 95% CI includes zero.

We are interested in causal models relating health insurance to the probability of visiting a doctor or a hospital stay.

The causal model:

$y_i = \beta_0 + \beta_1 D_i + \beta_2 r_i + \beta_3 w_i + u_i$

First we do a local linear model and a local quaratic model for the two outcome variables.

`local linear reduced form model for y = sawdr`

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8428	0.003	268.827	0.000	0.837	0.849
z	0.0172	0.004	3.931	0.000	0.009	0.026
r	0.0039	0.001	7.330	0.000	0.003	0.005
r_z	0.0008	0.001	1.103	0.270	-0.001	0.002

```
local linear reduced form model for y = inhosp
```

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1122	0.003	39.590	0.000	0.107	0.118
z	0.0131	0.004	3.317	0.001	0.005	0.021
r	0.0034	0.000	7.178	0.000	0.003	0.004
r_z	0.0003	0.001	0.502	0.616	-0.001	0.002

```
local reduced form model with quadratic specifications for y = sawdr
```

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8417	0.005	172.788	0.000	0.832	0.851
z	0.0176	0.007	2.686	0.007	0.005	0.031
r	0.0033	0.002	1.496	0.135	-0.001	0.008
r_z	0.0018	0.003	0.594	0.552	-0.004	0.008
r2	-5.859e-05	0.000	-0.277	0.782	-0.000	0.000
w2	1.877e-05	0.000	0.062	0.950	-0.001	0.001

```
local reduced form model with quadratic specifications for y = inhosp
```

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1090	0.004	24.760	0.000	0.100	0.118
z	0.0164	0.006	2.758	0.006	0.005	0.028
r	0.0016	0.002	0.818	0.413	-0.002	0.006
r_z	0.0021	0.003	0.757	0.449	-0.003	0.008
r2	-0.0002	0.000	-0.937	0.349	-0.001	0.000
w2	0.0002	0.000	0.683	0.495	-0.000	0.001

Next we using z_i as an instrumental variable for D_i where D is the indicator for health insurance coverage.

Table 2.2

	Local Linear sawdr	Local quadratic sawdr	Local Linear inhosp	Local quadratic inhosp
covered	0.182***	0.192***	0.139***	0.178***
	(0.046)	(0.070)	(0.042)	(0.065)
Observations	105,484	105,484	105,484	105,484
R^2	0.032	0.032	0.001	-0.004
Adjusted R^2	0.032	0.032	0.001	-0.004
Residual Std. Error	0.351	0.351	0.322	0.323
F Statistic	296.402***	177.925***	237.376***	141.798***

Note:

We can see that having healthcare insurance coverage is associated with a statistically significant increase in probablity of visiting a doctor or hospital. The estimate for probablity of visiting a doctor is quite robust to local linear versus local quadratic specifications. While the corresponding estimate for a hosiptal stay is a little different for local linear versus local quadratic specifications, the reuslts are still significant.

Checking Robustness

First, I will check for any discontinuity for another control variable: female.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5349	0.004	147.405	0.000	0.528	0.542
z	0.0035	0.005	0.676	0.499	-0.007	0.013
r	0.0011	0.001	1.873	0.061	-5.29e-05	0.002
r_z	0.0026	0.001	2.928	0.003	0.001	0.004

There seems to be no signifacent discontinuity.

Now I will add some control variables to the model from table 2.2 to see if these controls have any impact.

Table 2.3 (with controls)

	Local Linear sawdr	Local quadratic sawdr	Local Linear inhosp	Local quadratic inhosp
covered	0.181***	0.189***	0.139***	0.177***
	(0.045)	(0.070)	(0.042)	(0.064)
Observations	105,484	105,484	105,484	105,484
R^2	0.039	0.040	0.005	-0.001
Adjusted R^2	0.039	0.040	0.005	-0.001
Residual Std. Error	0.349	0.349	0.321	0.322
F Statistic	305.299***	229.053***	158.989***	118.734***

Note:

Here, we can see that adding there controls do not change the estimates much (or their significance). The estimates are robust.

Based on this, the causal effect of insurace coverage on seeing a doctor is 0.182 that for staying overnight at a hospital is 0.192.