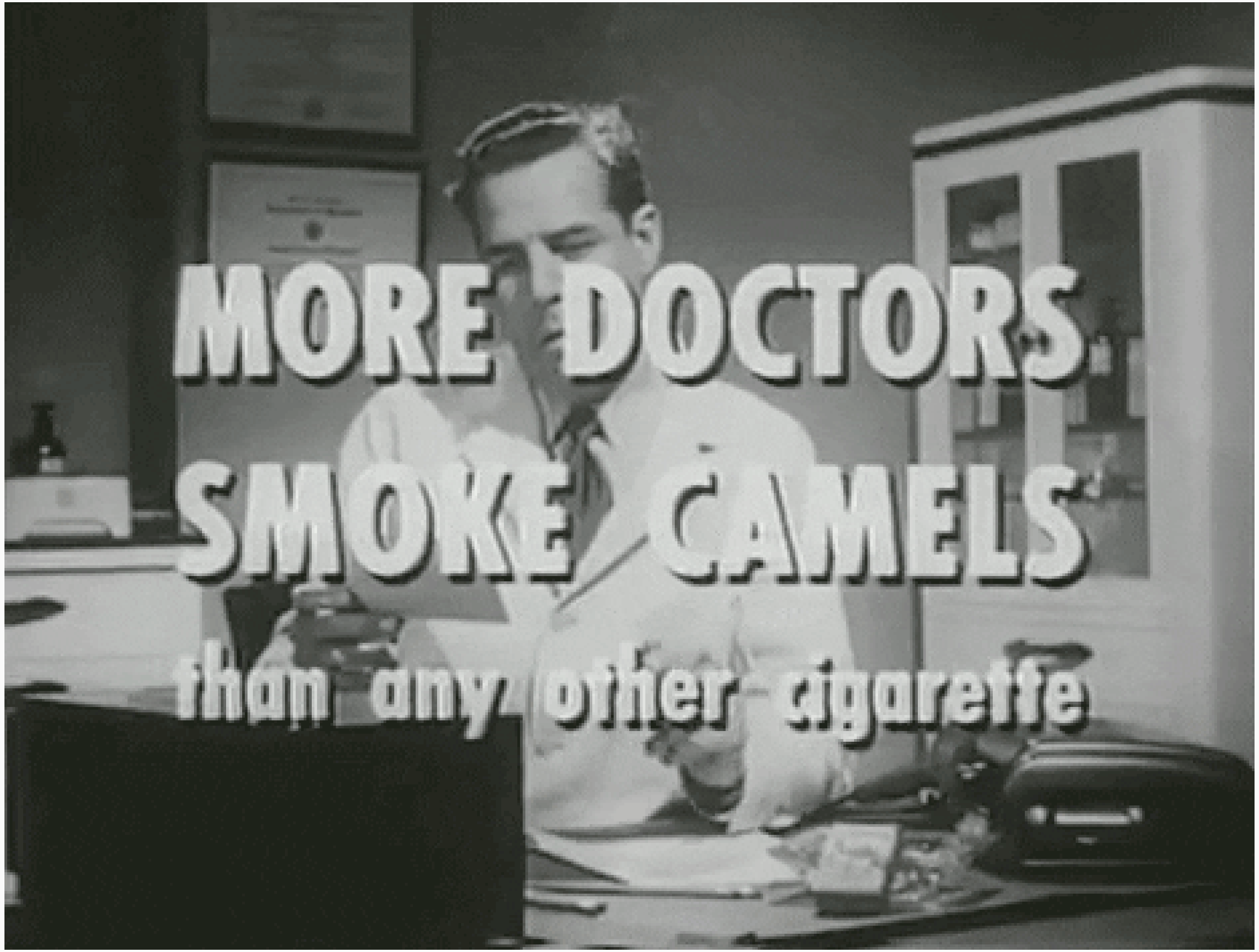
Section 2: Demand for Cigarettes and Instrumental Variables

Ian McCarthy | Emory University
Econ 470 & HLTH 470

Table of contents

Background on Cigarettes and Pricing

History of Smoking



History of Smoking

- Widespread smoking began in late 1800s
- Lung cancer becoming more common after 1930s
- First evidence of link in 1950s
- Surgeon general's report in 1964
- Very important in causal inference! (Section 5.1.1 of Causal Inference Mixtape)

Why it matters

1. Extreme public health concerns
    - Lung cancer prevalence
    - Fetal and baby health
2. Economic questions
    - Is it an information problem?
    - Externalities (second-hand smoke)
    - Moral hazard due to insurance

In our case

We want to focus on estimating demand for cigarettes. By this, I mean estimating price elasticity of demand.

We'll show that standard OLS isn't going to do this very well.

Cigarette Data

The Data

- Data from CDC Tax Burden on Tobacco
- Visit GitHub repository for other info: Tobacco GitHub repository
- Supplement with CPI data, also in GitHub repo.

Summary stats

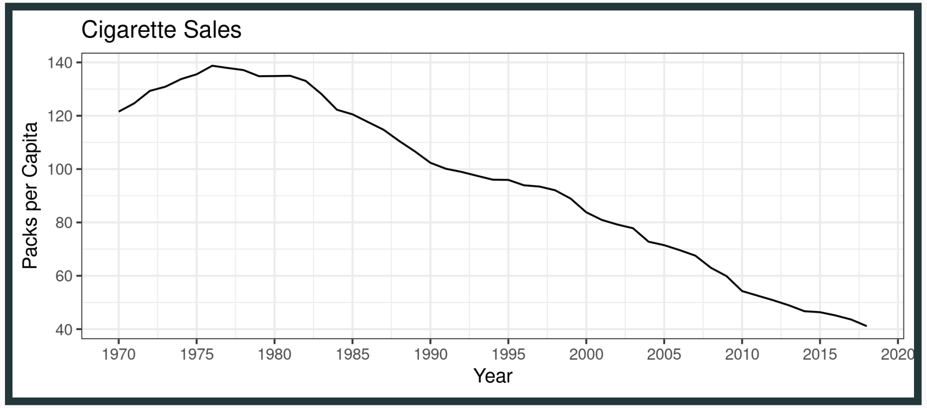
We're interested in cigarette prices and sales, so let's focus our summaries on those two variables

```
stargazer(as.data.frame(cig.data %>% select(sales_per_capita, price_cpi, cost_per_pack)), type="html")
```

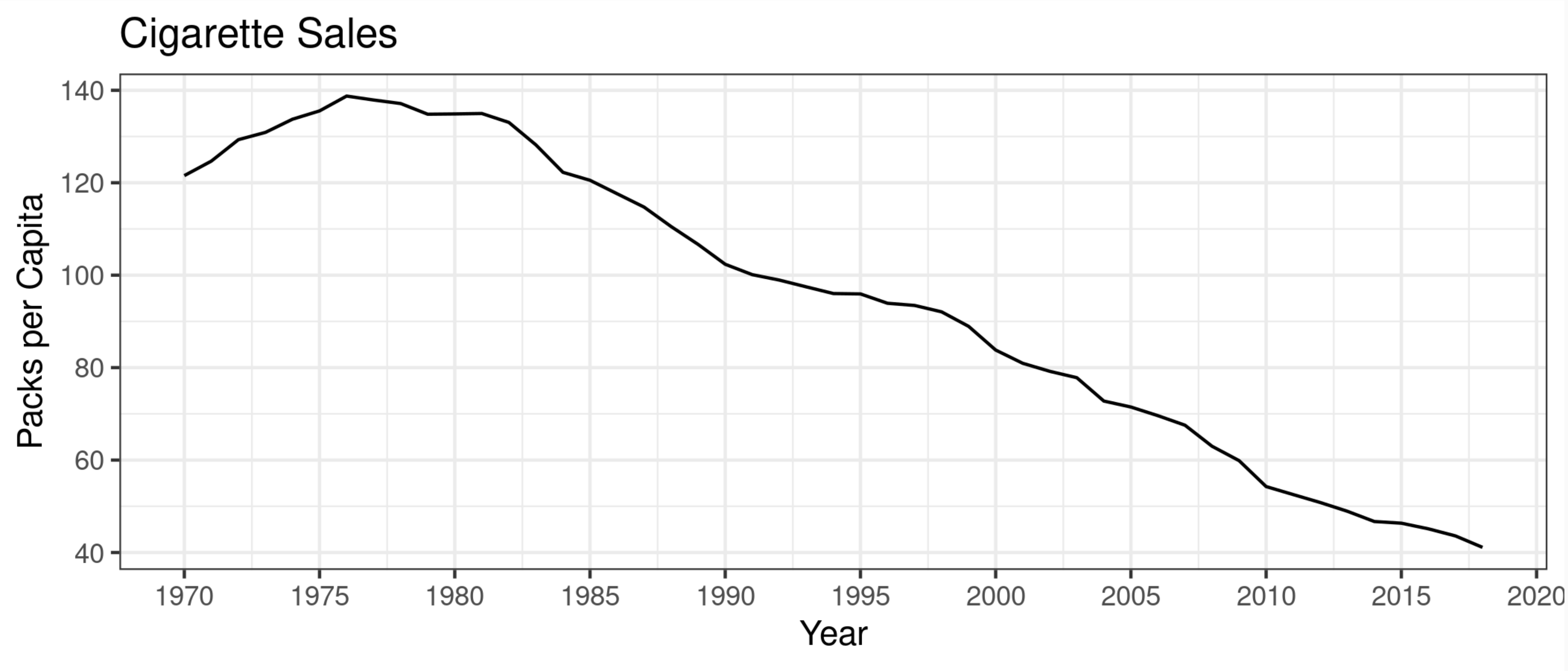
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
sales_per_capita	2,499	95.150	41.133	12.500	63.050	122.400	296.200
price_cpi	2,499	3.396	1.641	1.307	2.088	4.520	9.651
cost_per_pack	2,499	2.678	2.238	0.287	0.780	4.237	10.376

Cigarette Sales

```
cig.data %>%
  ggplot(aes(x=Year,y=sales_per_capita)) +
  stat_summary(fun.y="mean",geom="line") +
  labs(
    x="Year",
    y="Packs per Capita",
    title="Cigarette Sales"
  ) + theme_bw() +
  scale_x_continuous(breaks=seq(1970, 2020, 5))
```

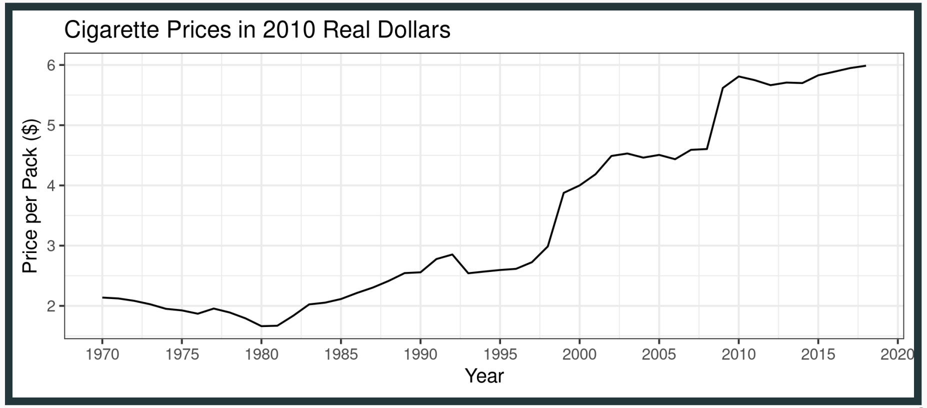


Cigarette Sales

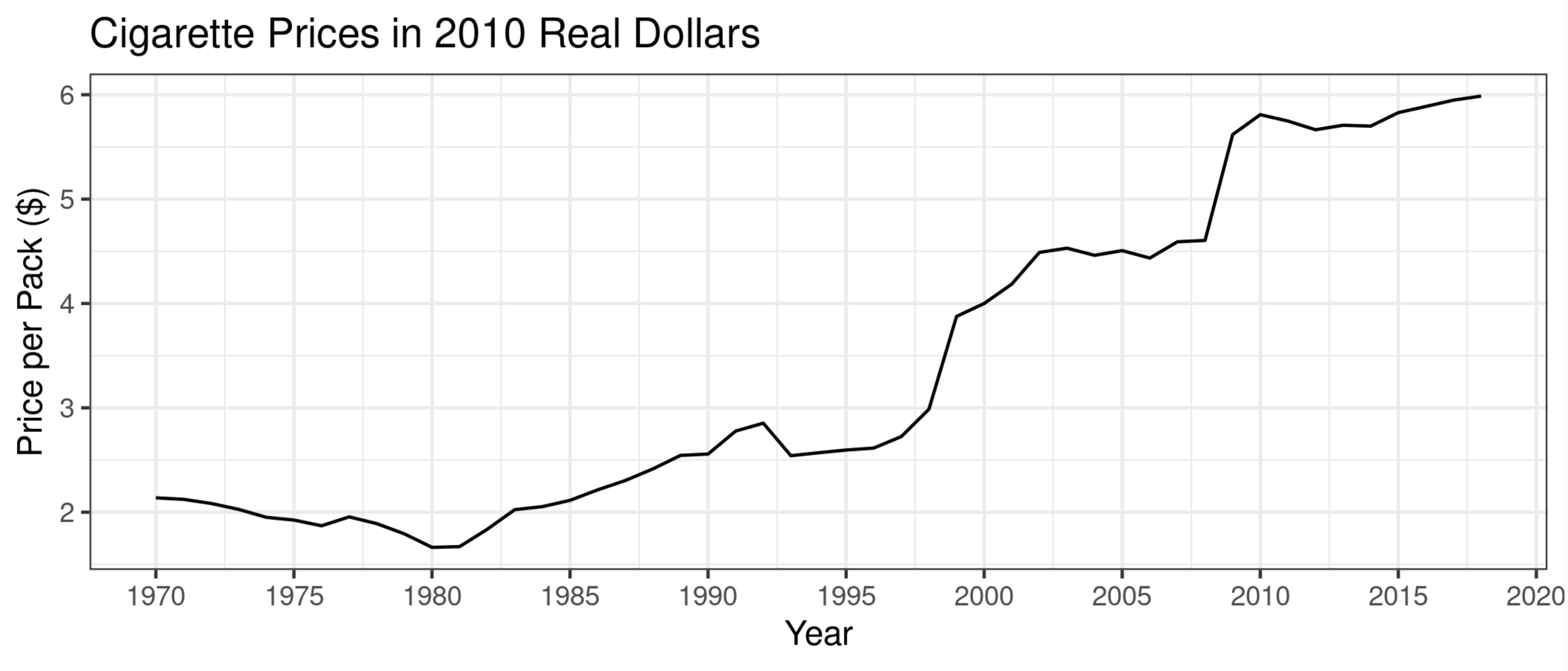


Cigarette Prices

```
cig.data %>%
  ggplot(aes(x=Year,y=price_cpi)) +
  stat_summary(fun.y="mean",geom="line") +
  labs(
    x="Year",
    y="Price per Pack ($)",
    title="Cigarette Prices in 2010 Real Dollars"
  ) + theme_bw() +
  scale_x_continuous(breaks=seq(1970, 2020, 5))
```



Cigarette Prices



Instrumental Variables

What is instrumental variables

Instrumental Variables (IV) is a way to identify causal effects using variation in treatment particpation that is due to an *exogenous* variable that is only related to the outcome through treatment.

Why bother with IV?

Two reasons to consider IV:

1. Selection on unobservables
2. Reverse causation

Either problem is sometimes loosely referred to as *endogeneity*

Simple example

- $y = \beta x + \varepsilon(x)$,
  where $\varepsilon(x)$ reflects the dependence between our observed variable and the error term.

- Simple OLS will yield
  $\frac{dy}{dx} = \beta + \frac{d\varepsilon}{dx} \neq \beta$

What does IV do?

- The regression we want to do:
  $y_i = \alpha + \delta D_i + \gamma A_i + \epsilon_i$,
  where D_i is treatment (think of schooling for now) and A_i is something like ability.

- A_i is unobserved, so instead we run:
  $y_i = \alpha + \beta D_i + \epsilon_i$

- From this "short" regression, we don't actually estimate δ. Instead, we get an estimate of
  $\beta = \delta + \lambda_{ds}\gamma \neq \delta$,
  where λ_{ds} is the coefficient of a regression of A_i on D_i.

Intuition

IV will recover the "long" regression without observing underlying ability

IF our IV satisfies all of the necessary assumptions.

More formally

- We want to estimate
  $$E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$
- With instrument Z_i that satisfies relevant assumptions, we can estimate this as
  $$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]}$$
- In words, this is effect of the instrument on the outcome ("reduced form") divided by the effect of the instrument on treatment ("first stage")

Derivation

Recall "long" regression: $Y = \alpha + \delta S + \gamma A + \epsilon$.

$$
\begin{aligned}
COV(Y,Z) &= E[YZ] - E[Y]E[Z] \\
&= E[(\alpha + \delta S + \gamma A + \epsilon) \times Z] - E[\alpha + \delta S + \gamma A + \epsilon)]E[Z] \\
&= \alpha E[Z] + \delta E[SZ] + \gamma E[AZ] + E[\epsilon Z] \\
&\quad - \alpha E[Z] - \delta E[S]E[Z] - \gamma E[A]E[Z] - E[\epsilon]E[Z] \\
&= \delta(E[SZ] - E[S]E[Z]) + \gamma(E[AZ] - E[A]E[Z]) \\
&\quad + E[\epsilon Z] - E[\epsilon]E[Z] \\
&= \delta C(S,Z) + \gamma C(A,Z) + C(\epsilon, Z)
\end{aligned}
$$

Derivation

Working from $COV(Y,Z) = \delta COV(S,Z) + \gamma COV(A,Z) + COV(\epsilon, Z)$, we find

$$\delta = \frac{COV(Y,Z)}{COV(S,Z)}$$

if $COV(A,Z) = COV(\epsilon, Z) = 0$

IVs in practice

Easy to think of in terms of randomized controlled trial...

Measure	Offered Seat	Not Offered Seat	Difference
Score	-0.003	-0.358	0.355
% Enrolled	0.787	0.046	0.741
Effect			0.48

Angrist *et al.*, 2012. "Who Benefits from KIPP?" *Journal of Policy Analysis and Management.*

What is IV *really* doing

Think of IV as two-steps:

1. Isolate variation due to the instrument only (not due to endogenous stuff)
2. Estimate effect on outcome using only this source of variation

In regression terms

Interested in estimating δ from $y_i = \alpha + \beta x_i + \delta D_i + \varepsilon_i$, but D_i is endogenous (no pure "selection on observables").

Step 1: With instrument Z_i, we can regress D_i on Z_i and x_i,
$D_i = \lambda + \theta Z_i + \kappa x_i + \nu$,
and form prediction $\hat{D}_i$.

Step 2: Regress y_i on x_i and $\hat{D}_i$,
$y_i = \alpha + \beta x_i + \delta \hat{D}_i + \xi_i$

Derivation

Recall $\hat{\theta} = \frac{C(Z,S)}{V(Z)}$, or $\hat{\theta}V(Z) = C(Y,Z)$. Then:

$$\begin{aligned}
\hat{\delta} &= \frac{COV(Y,Z)}{COV(S,Z)} \\
&= \frac{\hat{\theta}C(Y,Z)}{\hat{\theta}C(S,Z)} = \frac{\hat{\theta}C(Y,Z)}{\hat{\theta}^2 V(Z)} \\
&= \frac{C(\hat{\theta}Z,Y)}{V(\hat{\theta}Z)} = \frac{C(\hat{S},Y)}{V(\hat{S})}
\end{aligned}$$

In regression terms

But in practice, *DON'T* do this in two steps. Why?

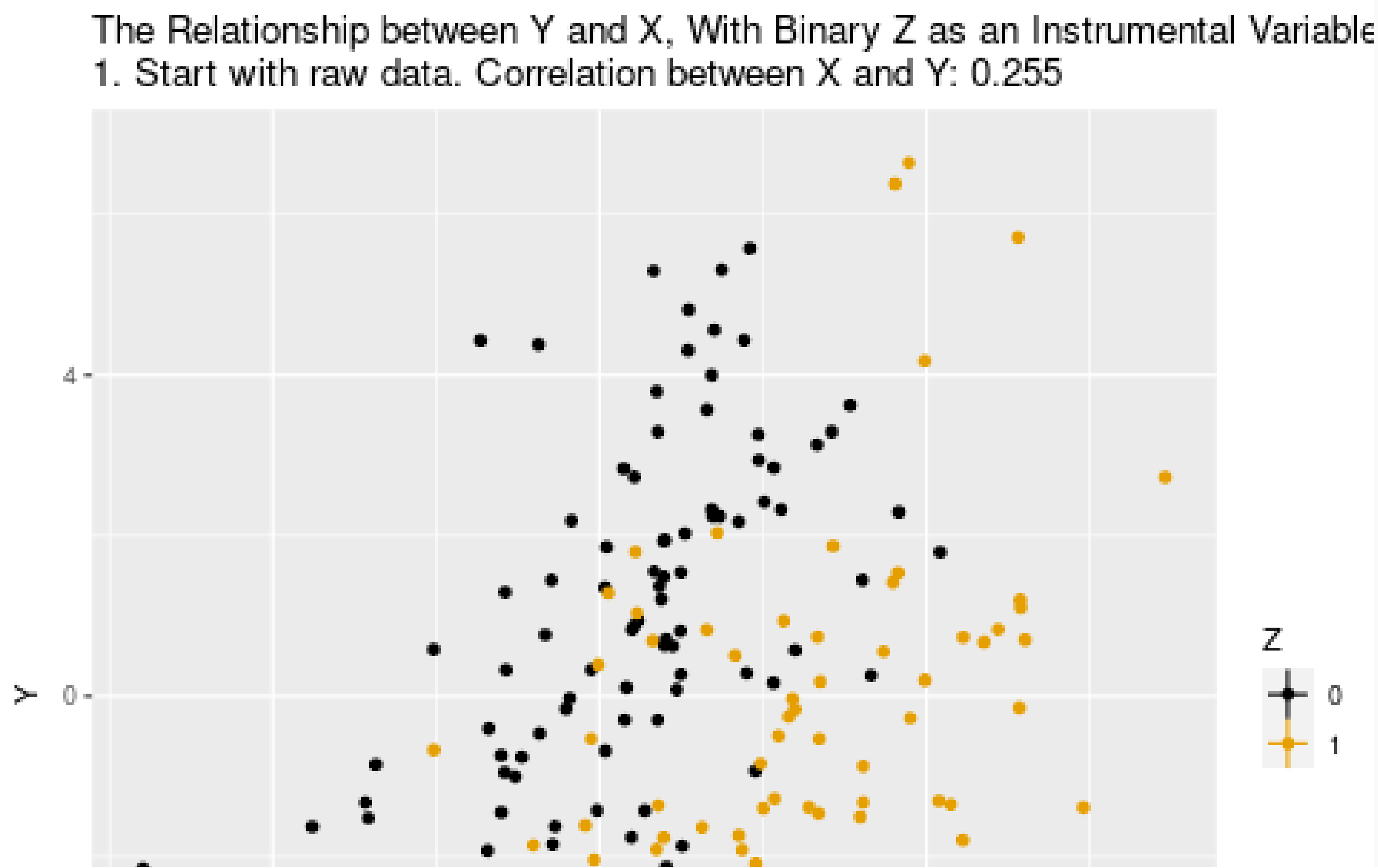
Because standard errors are wrong...not accounting for noise in prediction, $\hat{D}_i$.
The appropriate fix is built into most modern stats programs.

Key IV assumptions

1. *Exclusion:* Instrument is uncorrelated with the error term
2. *Validity:* Instrument is correlated with the endogenous variable
3. *Monotonicity:* Treatment more (less) likely for those with higher (lower) values of the instrument

Assumptions 1 and 2 sometimes grouped into an *only through* condition.

Animation for IV



Simulated data

```r
n <- 5000
b.true <- 5.25
iv.dat <- tibble(
  z = rnorm(n,0,2),
  eps = rnorm(n,0,1),
  d = (z + 1.5*eps + rnorm(n,0,1) >0.25),
  y = 2.5 + b.true*d + eps + rnorm(n,0,0.5)
)
```

- endogenous `eps`: affects treatment and outcome
- `z` is an instrument: affects treatment but no direct effect on outcome

Results with simulated data

Recall that the *true* treatment effect is 5.25

```
##
## Call:
## lm(formula = y ~ d, data = iv.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8475 -0.7007 -0.0197  0.7043  3.8847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.07350    0.02002   103.6   <2e-16 ***
## dTRUE        6.16139    0.02951   208.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.04 on 4998 degrees of freedom
## Multiple R-squared:  0.8971,    Adjusted R-squared:  0.8971
## F-statistic: 4.359e+04 on 1 and 4998 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## ivreg(formula = y ~ d | z, data = iv.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.25908 -0.76874  0.02717  0.74829  4.36714
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.48509    0.02992   83.05   <2e-16 ***
## dTRUE        5.26741    0.05492   95.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 1.132 on 4998 degrees of freedom
## Multiple R-Squared: 0.8783,    Adjusted R-squared: 0.8782
## Wald test:  9200 on 1 and 4998 DF,  p-value: < 2.2e-16
```

Checking instrument

- Check the 'first stage'

```
##
## Call:
## lm(formula = d ~ z, data = iv.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18377 -0.33100 -0.02694  0.34211  1.04655
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.457369   0.005720   79.96   <2e-16 ***
## z           0.145662   0.002859   50.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4045 on 4998 degrees of freedom
## Multiple R-squared:  0.3418,    Adjusted R-squared:  0.3417
## F-statistic:  2596 on 1 and 4998 DF,  p-value: < 2.2e-16
```

- Check the 'reduced form'

```
##
## Call:
## lm(formula = y ~ z, data = iv.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6854 -2.0943 -0.0718  2.0937  9.5522
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.05943    0.03960  127.78   <2e-16 ***
## z            0.86070    0.01975   43.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 2.8 on 4998 degrees of freedom
## Multiple R-squared:  0.2754,    Adjusted R-squared:  0.2753
## F-statistic:  1900 on 1 and 4998 DF,  p-value: < 2.2e-16
```

Two-stage equivalence

```
step1 <- lm(d ~ z, data=iv.dat)
d.hat <- predict(step1)
step2 <- lm(y ~ d.hat, data=iv.dat)
summary(step2)
```

```
##
## Call:
## lm(formula = y ~ d.hat, data = iv.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0433 -2.2201 -0.1163  2.2259  8.3417
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.48509    0.07554    32.9   <2e-16 ***
## d.hat        5.26741    0.13863    38.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.857 on 4998 degrees of freedom
## Multiple R-squared:  0.2241,    Adjusted R-squared:  0.224
## F-statistic:  1444 on 1 and 4998 DF,  p-value: < 2.2e-16
```

Do we need IV?

- Let's run an "augmented regression" to see if our OLS results sufficiently different than IV

```
d.iv <- lm(d ~ z, data=iv.dat)
d.resid <- residuals(d.iv)
haus.test <- lm(y ~ d + d.resid, data=iv.dat)
summary(haus.test)
```

```
##
## Call:
## lm(formula = y ~ d + d.resid, data = iv.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9224 -0.6581  0.0013  0.6715  3.8962
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.48509    0.02616   94.99   <2e-16 ***
## dTRUE        5.26741    0.04801  109.70   <2e-16 ***
## d.resid      1.35823    0.05918   22.95   <2e-16 ***
```

Testing exclusion

- Exclusion restriction says that your instrument does not directly affect your outcome
- Potential testing ideas:
  - "zero-first-stage" (subsample on which you know the instrument does not affect the endogenous variable)
  - augmented regression of reduced-form effect with subset of instruments (overidentified models only)

Testing exogeneity

- Only available in over-identified models
- Sargan or Hansen's J test (null hypothesis is that instruments are correlated with residuals)

Testing strength of instruments

Single endogenous variable

- F-test of instruments (rule of thumb critical value of 10)
- Partial R^2

Many endogenous variables

- More complicated

Why we care about instrument strength

Recall our schooling and wages equation,

$$y = \beta S + \epsilon$$

. Bias in IV can be represented as:

$$Bias_{IV} \approx \frac{Cov(S,\epsilon)}{V(S)} \frac{1}{F+1} = Bias_{OLS} \frac{1}{F+1}$$

- Bias in IV may be close to OLS, depending on instrument strength
- **Bigger problem:** Bias could be bigger than OLS if exclusion restriction not *fully* satisfied
- Over-reliance on "rules of thumb", as seen in Anders and Kasy (2019)

LATE and IV Interpretation

- With monotonicity assumption (all those affected by the instrument are affected in the same "direction")
- In the face of **heterogeneous treatment effects**, IV provides a "Local Average Treatment Effect"
- LATE: Effect of treatment among those affected by the instrument (compliers only)
- Why does this matter? Let's discuss Medicaid expansion in Oregon

Estimating Demand for Cigarettes

Naive estimate

Clearly a strong relationship between prices and sales. For example, just from OLS:

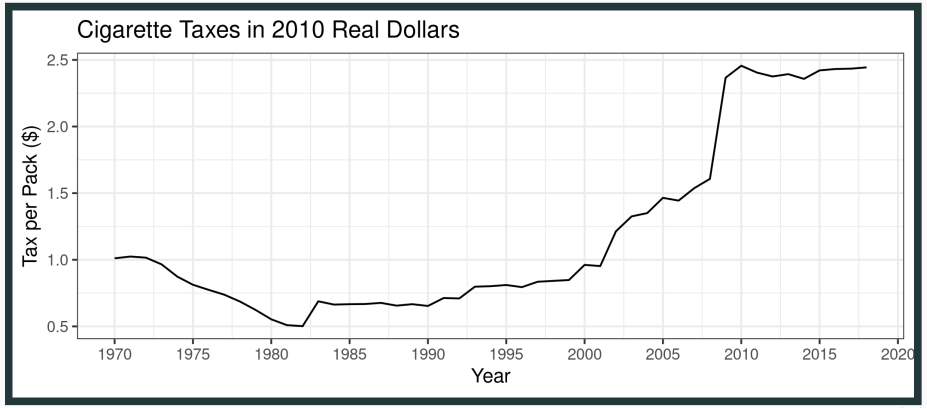
```
##
## Call:
## lm(formula = ln_sales ~ ln_price, data = cig.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23899 -0.17057  0.02239  0.18605  1.13866
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.689838   0.007209  650.55   <2e-16 ***
## ln_price    -0.420307   0.006464  -65.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3073 on 2497 degrees of freedom
## Multiple R-squared:  0.6287,    Adjusted R-squared:  0.6285
## F-statistic:  4228 on 1 and 2497 DF,  p-value: < 2.2e-16
```

Is this causal?

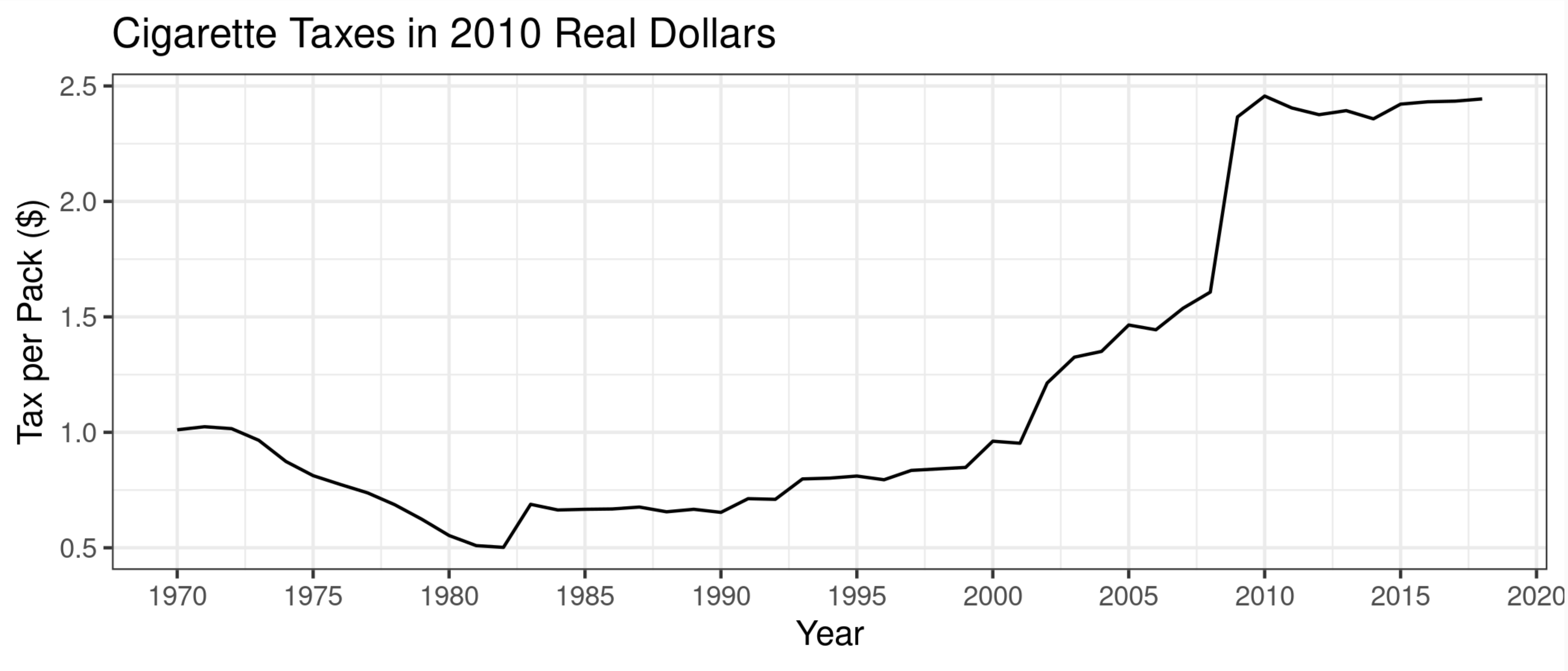
- But is that the true demand curve?
- Aren't other things changing that tend to reduce cigarette sales?

Tax as an IV

```r
cig.data %>%
  ggplot(aes(x=Year,y=total_tax_cpi)) +
  stat_summary(fun.y="mean",geom="line") +
  labs(
    x="Year",
    y="Tax per Pack ($)",
    title="Cigarette Taxes in 2010 Real Dollars"
  ) + theme_bw() +
  scale_x_continuous(breaks=seq(1970, 2020, 5))
```



Tax as an IV


Cigarette Taxes in 2010 Real Dollars
Tax per Pack ($)
2.5
2.0
1.5
1.0
0.5
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year

IV Results

```
##
## Call:
## ivreg(formula = ln_sales ~ ln_price | total_tax_cpi, data = cig.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24595 -0.23048  0.02863  0.23548  1.30999
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.805691   0.009703  495.29   <2e-16 ***
## ln_price    -0.619142   0.011128  -55.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3608 on 2497 degrees of freedom
## Multiple R-Squared: 0.488,    Adjusted R-squared: 0.4878
## Wald test:  3096 on 1 and 2497 DF,  p-value: < 2.2e-16
```

Two-stage equivalence

```
step1 <- lm(ln_price ~ total_tax_cpi, data=cig.data)
pricehat <- predict(step1)
step2 <- lm(ln_sales ~ pricehat, data=cig.data)
summary(step2)
```

```
##
## Call:
## lm(formula = ln_sales ~ pricehat, data = cig.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10960 -0.17805  0.01867  0.18697  1.14907
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.805691   0.008195  586.41   <2e-16 ***
## pricehat    -0.619142   0.009399  -65.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3048 on 2497 degrees of freedom
## Multiple R-squared:  0.6348,    Adjusted R-squared:  0.6346
## F-statistic:  4339 on 1 and 2497 DF,  p-value: < 2.2e-16
```

Different specifications

	Log Sales per Capita					
	OLS			IV		
	(1)	(2)	(3)	(4)	(5)	(6)
Log Price	-0.953***	-0.921***	-1.213***	-1.072***	-1.036***	-1.523***
	(0.012)	(0.008)	(0.034)	(0.014)	(0.010)	(0.041)
State FE	No	Yes	Yes	No	Yes	Yes
Year FE	No	No	Yes	No	No	Yes
Observations	2,499	2,499	2,499	2,499	2,499	2,499

Note:

Test the IV

	Log Price			Log Sales		
	First Stage			Reduced Form		
	(1)	(2)	(3)	(4)	(5)	(6)
Tax per Pack	0.444***	0.474***	0.187***	-0.476***	-0.491***	-0.284***
	(0.006)	(0.006)	(0.002)	(0.007)	(0.006)	(0.007)
State FE	No	Yes	Yes	No	Yes	Yes
Year FE	No	No	Yes	No	No	Yes
Observations	2,499	2,499	2,499	2,499	2,499	2,499

Note:

Summary

1. Most elasticities of around -0.25% to -0.37%

2. Much larger elasticities when including year fixed effects

3. Perhaps not too outlandish given more recent evidence: NBER Working Paper.

Some other IV issues

1. IV estimators are biased. Performance in finite samples is questionable.
2. IV estimators provide an estimate of a Local Average Treatment Effect (LATE), which is only the same as the ATT under some conditions or assumptions.
3. What about lots of instruments? The finite sample problem is more important and we may try other things (JIVE).

   The National Bureau of Economic Research (NBER) has a great resource here for understanding instruments in practice.