Section 4: Difference-in-Differences and Effects of Medicaid Expansion

Ian McCarthy | Emory University
Econ 470 & HLTH 470

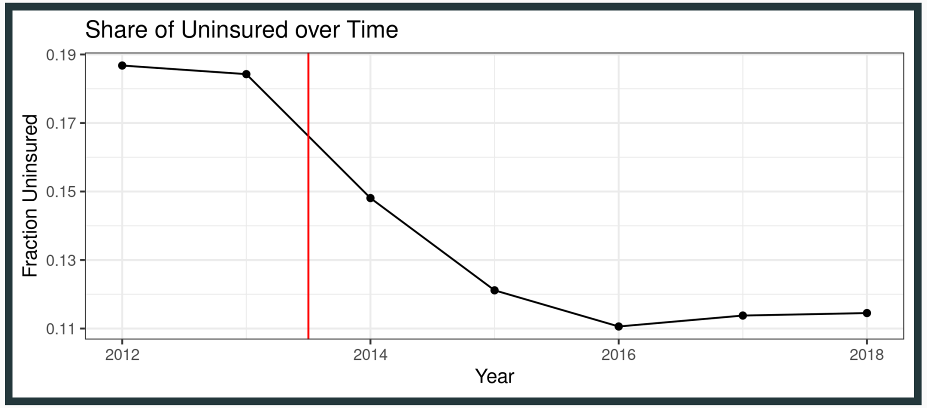
Table of contents

Background on the Affordable Care Act

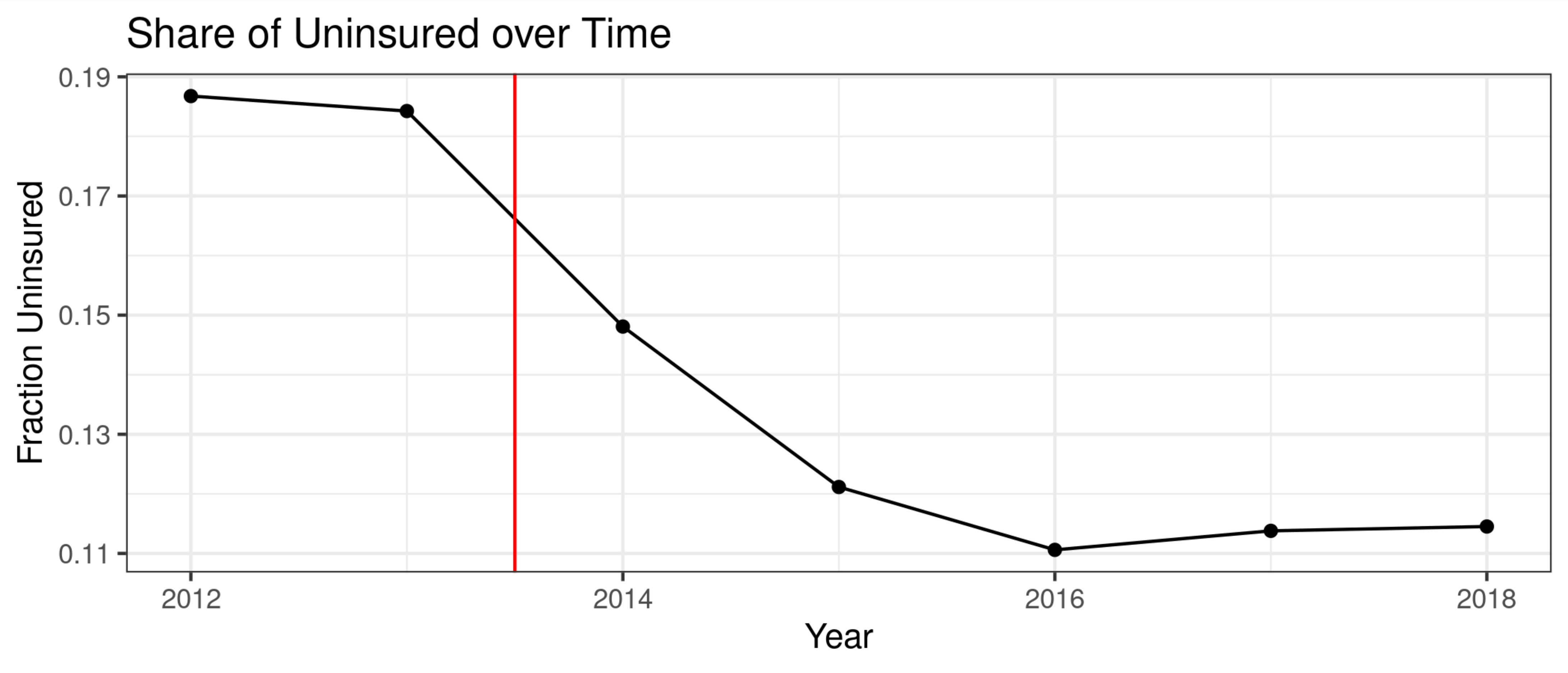
Affordable Care Act

Background

1. What percent of people are uninsured?



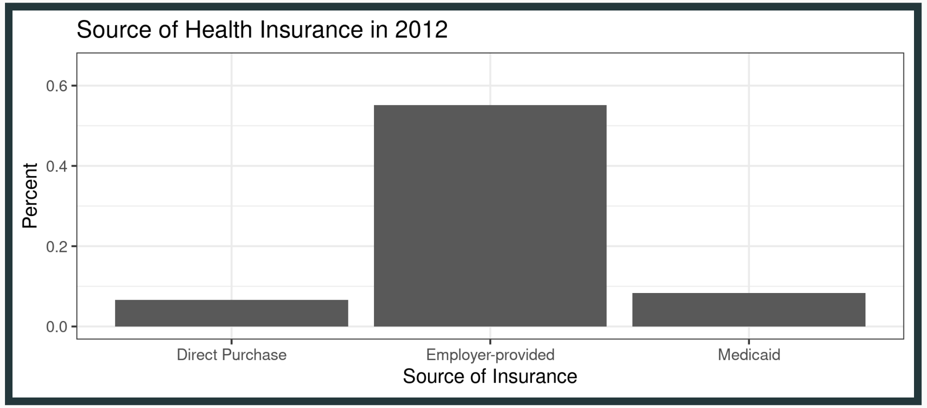
What percent of people are uninsured?



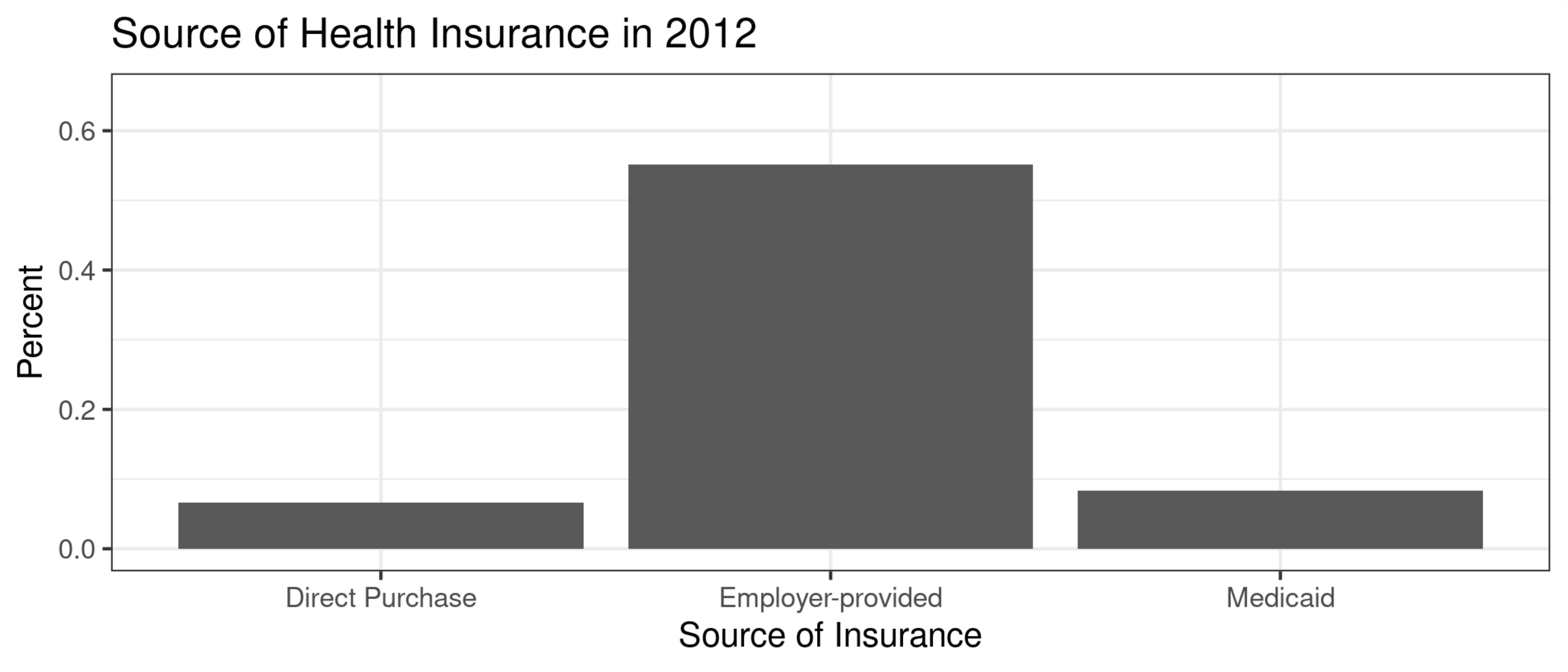
Background

1. What percent of people are uninsured?

2. How do people get health insurance?



How do people get health insurance?



Employer provided insurance

The U.S. still relies heavily on private insurance provided by employers.

Any thoughts on why?

Employer provided insurance

1. Stabalization act of 1942 (wages frozen but not benefits)

2. Tax exclusion for insurance expenditures (1954)

How did the ACA change things?

1. Create health insurance exhanges
    - Individual mandate (since set to $0)
    - Premium and cost-sharing subsidies (some unpaid by Trump administration)
    - Insurance subsidies (removed before intended)
    - Decision assistance
    - Minimum benefits and community ratings

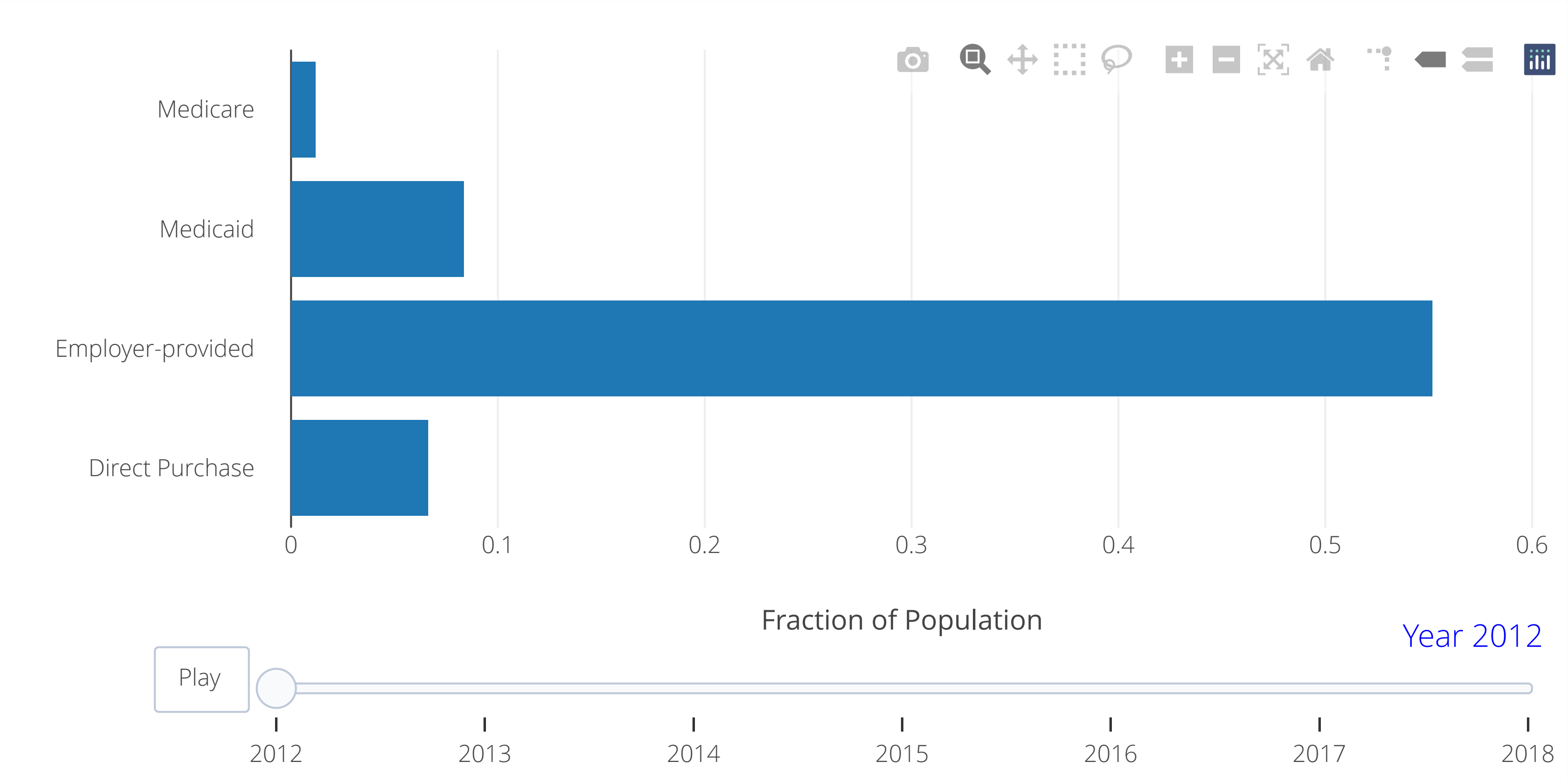
2. Stay on parent's plan to 26

How did the ACA change things?

3. Medicaid Expansion
    - Originally tied to federal funding
    - Made voluntary by supreme court ruling
    - Higher initial federal match rate, decreasing over time

4. Pay-for-performance measures
    - Hospital value-based purchasing
    - Hospital readmission reduction
    - Medicare Advantage quality improvement program
    - Bundled payments and ACOs (related)

Change in Insurance Type over Time



Main points

1. Large reduction in uninsured population following ACA
2. Biggest gains going to direct purchase (exchanges) and Medicaid (expansion)

But what amount of extra insurance is *due to* Medicaid expansion? In other words, who got insurance through Medicaid that wouldn't have gotten it otherwise?

What does the literature say

The *Kaiser Family Foundation* has some great info on this...

- KFF Medicaid Coverage
- KFF Report on ACA Expansion
- Health Insurance and Mortality (not what we're discussing here but still important)

Difference-in-Differences

Setup

Want to estimate $E[Y_1(1) - Y_0(1)|W = 1]$

	Post-period	Pre-period
Treated	`$E(Y_1(1)|W = 1)$`	`$E(Y_0(0)|W = 1)$`
Control	`$E(Y_0(1)|W = 0)$`	`$E(Y_0(0)|W = 0)$`

Problem: We don't see $E[Y_0(1)|W = 1]$

Setup

	Post-period	Pre-period
Treated	$E(Y_1(1)\|W = 1)$	$E(Y_0(0)\|W = 1)$
Control	$E(Y_0(1)\|W = 0)$	$E(Y_0(0)\|W = 0)$

Strategy 1: Estimate $E[Y_0(1)|W = 1]$ using $E[Y_0(0)|W = 1]$ (before treatment outcome used to estimate post-treatment)

Setup

	Post-period	Pre-period
Treated	`$E(Y_1(1)\|W=1)$`	`$E(Y_0(0)\|W=1)$`
Control	`$E(Y_0(1)\|W=0)$`	`$E(Y_0(0)\|W=0)$`

Strategy 2: Estimate $E[Y_0(1)|W=1]$ using $E[Y_0(1)|W=0]$ (control group used to predict outcome for treatment)

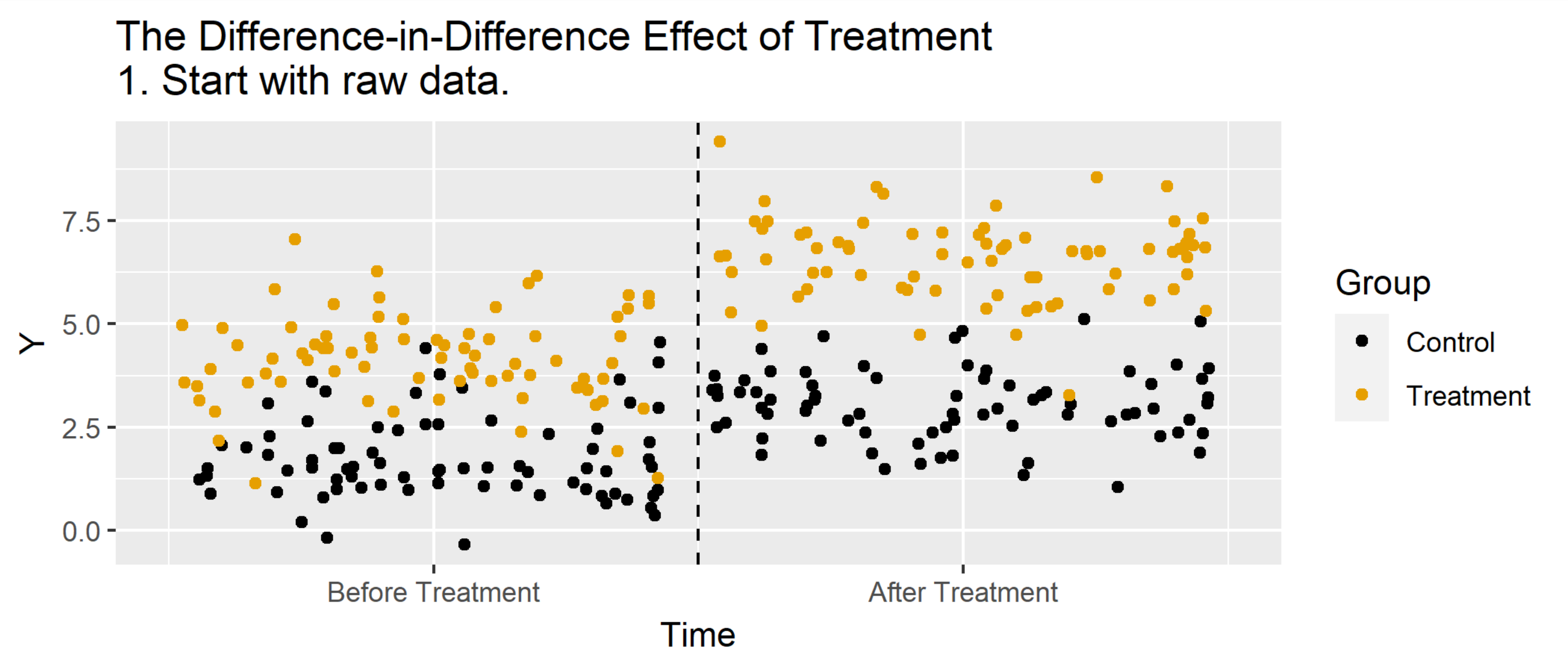
Setup

	Post-period	Pre-period
Treated	`$E(Y_1(1)\|W = 1)$`	`$E(Y_0(0)\|W = 1)$`
Control	`$E(Y_0(1)\|W = 0)$`	`$E(Y_0(0)\|W = 0)$`

Strategy 3: DD estimate…

Estimate $E[Y_1(1)|W = 1] - E[Y_0(1)|W = 1]$ using $E[Y_0(1)|W = 0] - E[Y_0(0)|W = 0]$ (pre-post difference in control group used to predict difference for treatment group)

Animations!



The Difference-in-Difference Effect of Treatment
1. Start with raw data.
Y
7.5
5.0
2.5
0.0
Before Treatment
After Treatment
Time
Group
Control
Treatment

Estimation

Key identifying assumption is that of *parallel trends*

$$E[Y_0(1) - Y_0(0)|W = 1] = E[Y_0(1) - Y_0(0)|W = 0]$$

Estimation

Sample means:

$$
\begin{aligned}
E[Y_1(1) - Y_0(1)|W = 1] = \quad & (E[Y(1)|W = 1] - E[Y(1)|W = 0]) \\
& - (E[Y(0)|W = 1] - E[Y(0)|W = 0])
\end{aligned}
$$

Estimation

Regression:

$$Y_i = \alpha + \beta W_i + \lambda 1(Post) + \delta W_i \times 1(Post) + \varepsilon$$

	After	Before	After - Before
Treated	$\alpha + \beta + \lambda + \delta$	$\alpha + \beta$	$\lambda + \delta$
Control	$\alpha + \lambda$	α	λ
Treated - Control	$\beta + \delta$	β	δ

Simulated data

```
N <- 5000
dd.dat <- tibble(
  w = (runif(N, 0, 1)>0.5),
  time_pre = "pre",
  time_post = "post"
)

dd.dat <- pivot_longer(dd.dat, c("time_pre","time_post"), values_to="time") %>%
  select(w, time) %>%
  mutate(t=(time=="post"),
         y.out=1.5+3*w + 1.5*t + 6*w*t + rnorm(N*2,0,1))
```

Mean differences

```
dd.means <- dd.dat %>% group_by(w, t) %>% summarize(mean_y = mean(y.out))
knitr::kable(dd.means, col.names=c("Treated","Post","Mean"), format="html")
```

Treated	Post	Mean
FALSE	FALSE	1.522635
FALSE	TRUE	3.002374
TRUE	FALSE	4.515027
TRUE	TRUE	12.004623

Mean differences

In this example:

- $E[Y(1)|W=1] - E[Y(1)|W=0]$ is 9.0022495
- $E[Y(0)|W=1] - E[Y(0)|W=0]$ is 2.9923925

So the ATT is 6.0098571

Regression estimator

```
dd.est <- lm(y.out ~ w + t + w*t, data=dd.dat)
summary(dd.est)
```

```
##
## Call:
## lm(formula = y.out ~ w + t + w * t, data = dd.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0038 -0.6674  0.0047  0.6609  3.6135
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.52263    0.01970   77.28   <2e-16 ***
## wTRUE        2.99239    0.02795  107.07   <2e-16 ***
## tTRUE        1.47974    0.02786   53.10   <2e-16 ***
## wTRUE:tTRUE  6.00986    0.03953  152.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9881 on 9996 degrees of freedom
## Multiple R-squared:  0.9433,    Adjusted R-squared:  0.9433
## F-statistic: 5.543e+04 on 3 and 9996 DF,  p-value: < 2.2e-16
```

Insurance Data and Medicaid Expansion

Data sources

We'll use two main data sources here:

1. Data on which states expanded Medicaid (and when
    - Available from *Kaiser Family Foundation*
2. Data on insurance status and source of health insurance by state
    - Available from the *American Community Survey*
    - These data can be tricky to work with due to their size, but there are some handy tricks in `R`

Data sources

Code and links available at the Insurance Access GitHub repository

Medicaid Expansion

- Directly downloaded from KFF website
- Just a raw .csv file

Insurance status and source

- Data from the American Community Survey
- CPS data also available but questions changed in 2014
- Easiest way to access ACS data is through a Census API and the `acs` package...details on the *GitHub* repo

Describing the data

First let's take a look at the final dataset

```
head(ins.dat %>% arrange(year, State))
```

```
## # A tibble: 6 × 20
##   State       year adult_pop ins_employer ins_direct ins_medicare ins_medicaid
##   <chr>      <int>     <dbl>        <dbl>      <dbl>        <dbl>        <dbl>
## 1 Alabama     2012   2937335      1528419     180043        56890       190312
## 2 Alaska      2012    460946       222769      15608         2027        28177
## 3 Arizona     2012   3866694      1867954     263076        41042       428972
## 4 Arkansas    2012   1761365       871970     106277        39157       114012
## 5 California  2012  23798381     12015639    1824564       180861      2275053
## 6 Colorado    2012   3270163      1801613     303179        27254       213045
## # … with 13 more variables: uninsured <dbl>, expand_ever <lgl>,
## #   date_adopted <date>, expand_year <dbl>, expand <lgl>, perc_private <dbl>,
## #   perc_public <dbl>, perc_ins <dbl>, perc_unins <dbl>, perc_employer <dbl>,
## #   perc_medicaid <dbl>, perc_medicare <dbl>, perc_direct <dbl>
```

Summary stats

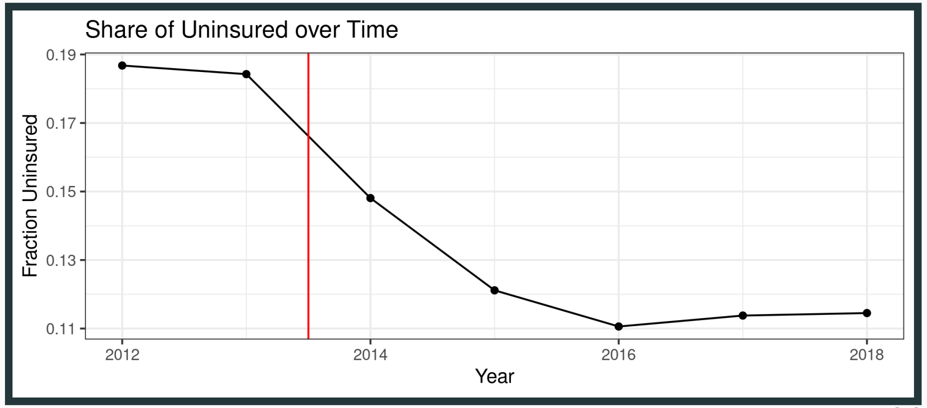
And now for some basic summary stats (pooling all years):

```
stargazer(as.data.frame(ins.dat %>% select(perc_unis, perc_direct, perc_medicaid)), type="html")
```

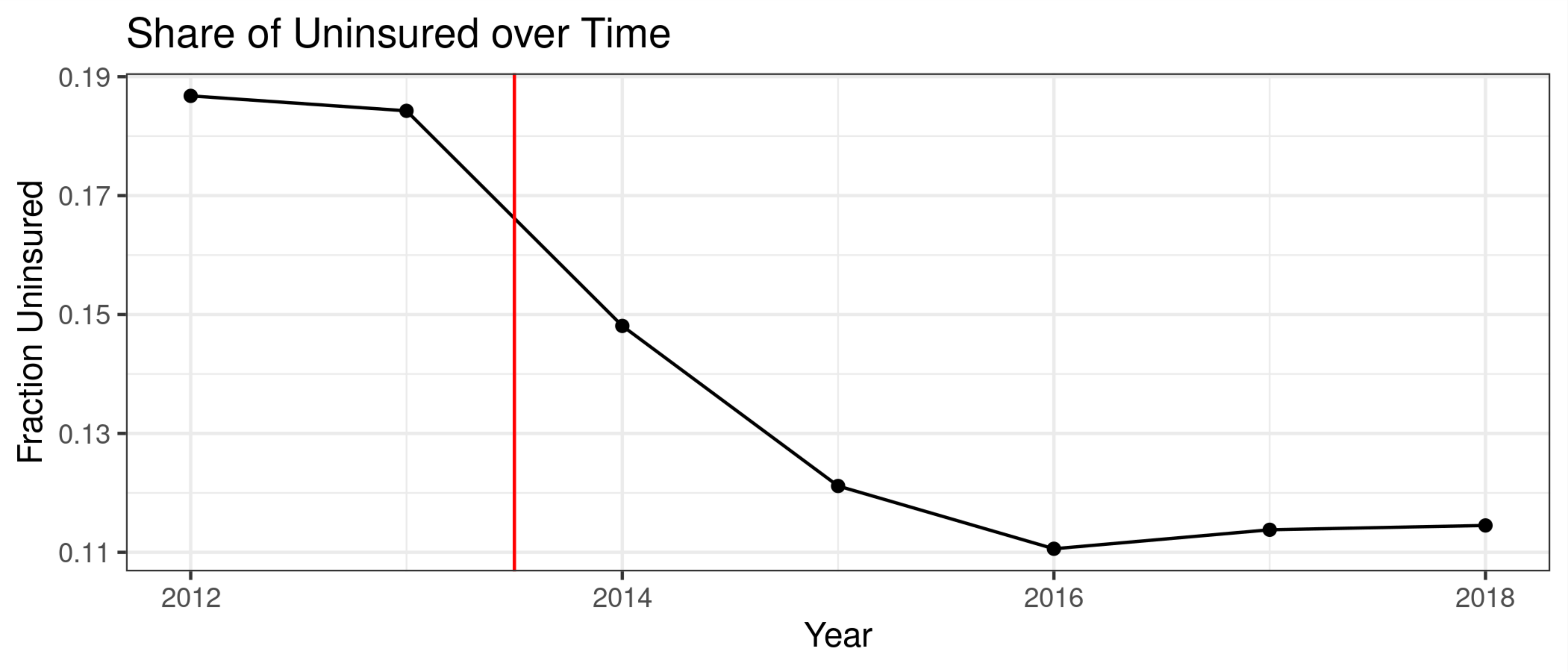
```
## Error: Can't subset columns that don't exist.
## x Column `perc_unis` doesn't exist.
```

Uninsurance over time

```r
ins.dat %>% group_by(year) %>% summarize(mean=mean(perc_unins)) %>%
  ggplot(aes(x=year,y=mean)) + geom_line() + geom_point() + theme_bw() +
  labs(
    x="Year",
    y="Fraction Uninsured",
    title="Share of Uninsured over Time"
  ) +
  geom_vline(xintercept=2013.5, color="red")
```

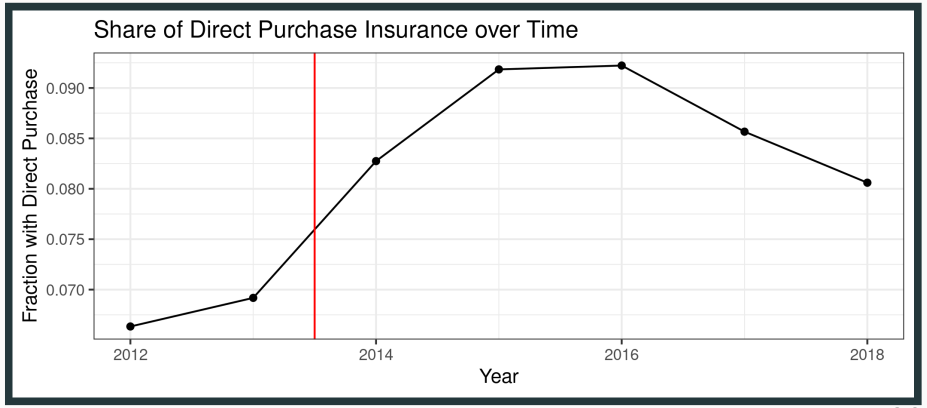


Uninsurance over time

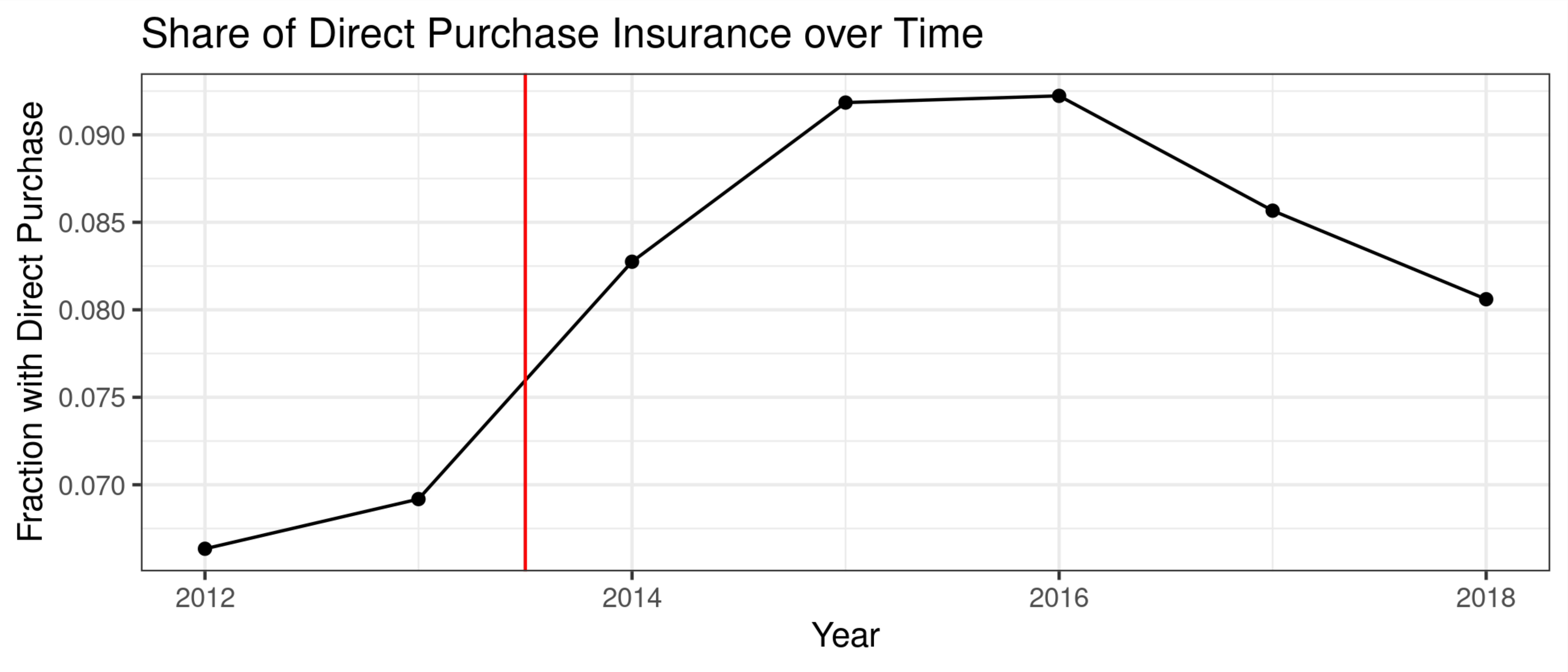


Direct purchase over time

```r
ins.dat %>% group_by(year) %>% summarize(mean=mean(perc_direct)) %>%
  ggplot(aes(x=year,y=mean)) + geom_line() + geom_point() + theme_bw() +
  labs(
    x="Year",
    y="Fraction with Direct Purchase",
    title="Share of Direct Purchase Insurance over Time"
  ) +
  geom_vline(xintercept=2013.5, color="red")
```

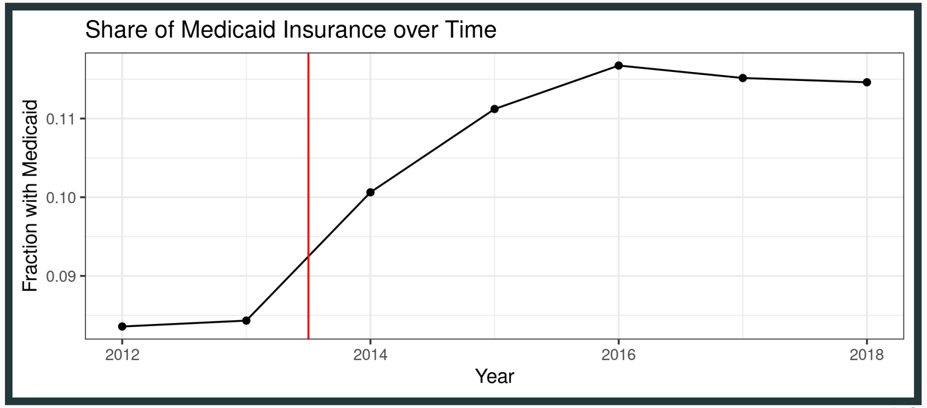


Direct purchase over time

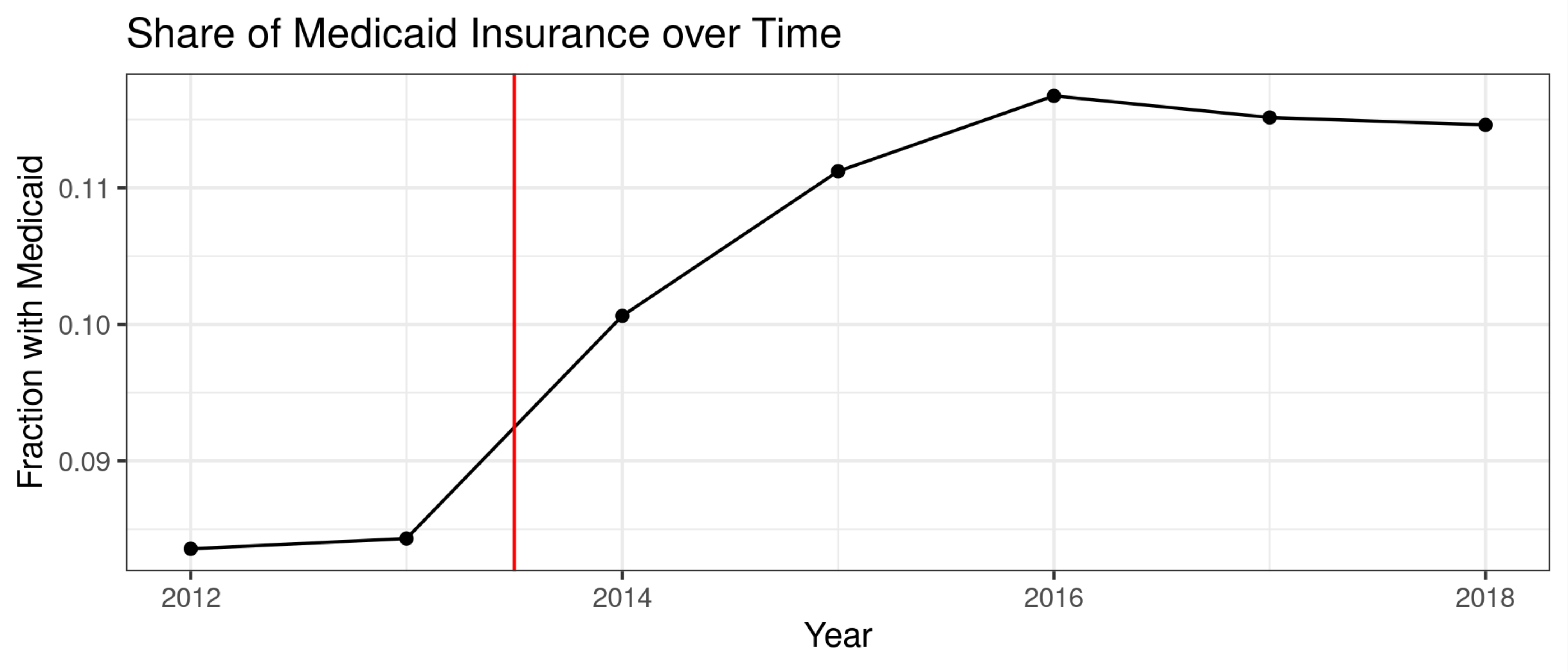


Medicaid over time

```r
ins.dat %>% group_by(year) %>% summarize(mean=mean(perc_medicaid)) %>%
  ggplot(aes(x=year,y=mean)) + geom_line() + geom_point() + theme_bw() +
  labs(
    x="Year",
    y="Fraction with Medicaid",
    title="Share of Medicaid Insurance over Time"
  ) +
  geom_vline(xintercept=2013.5, color="red")
```



Medicaid enrollment over time



Medicaid expansion and health insurance?

Key issue

What is the causal effect of Medicaid expansion?

- Clearly affects insurance markets- but Medicaid enrollment partially crowds out private insurance

Research design

Use pre/post and expansion/non-expansion states to identify effect of Medicaid expansion:

$$y_{it} = \alpha + \beta \times 1(Post) + \gamma \times 1(Expand) + \delta \times 1(Post) \times 1(Expand) + \varepsilon$$

Regression results

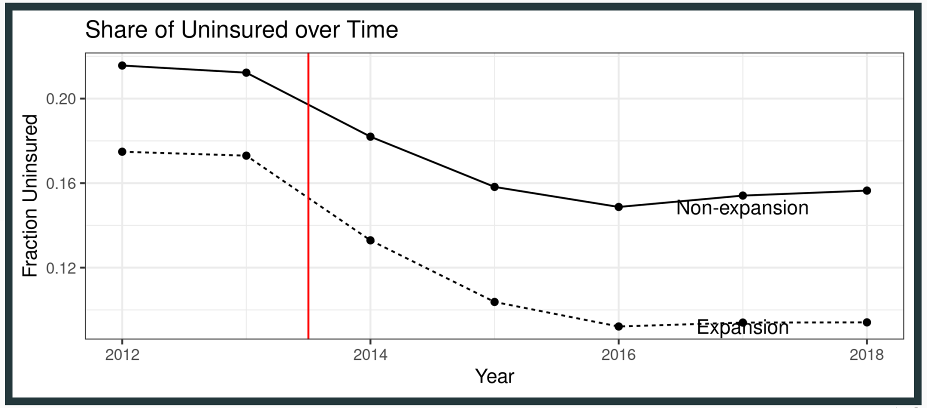
```
ins.dat.2014 <- ins.dat %>% mutate(post = (year>=2014), treat=post*expand_ever) %>% filter(is.na(expand_year) | expand_y
dd.ins.reg <- lm(perc_unins ~ post + expand_ever + post*expand_ever, data=ins.dat.2014)
summary(dd.ins.reg)
```

```
##
## Call:
## lm(formula = perc_unins ~ post + expand_ever + post * expand_ever,
##     data = ins.dat.2014)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.115667 -0.027106 -0.006804  0.027765  0.117597
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.213965   0.007180  29.799  < 2e-16 ***
## postTRUE                 -0.054068   0.008496  -6.364 7.22e-10 ***
## expand_everTRUE          -0.046326   0.009166  -5.054 7.48e-07 ***
## postTRUE:expand_everTRUE -0.018403   0.010845  -1.697   0.0908 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04187 on 304 degrees of freedom
##   (7 observations deleted due to missingness)
```

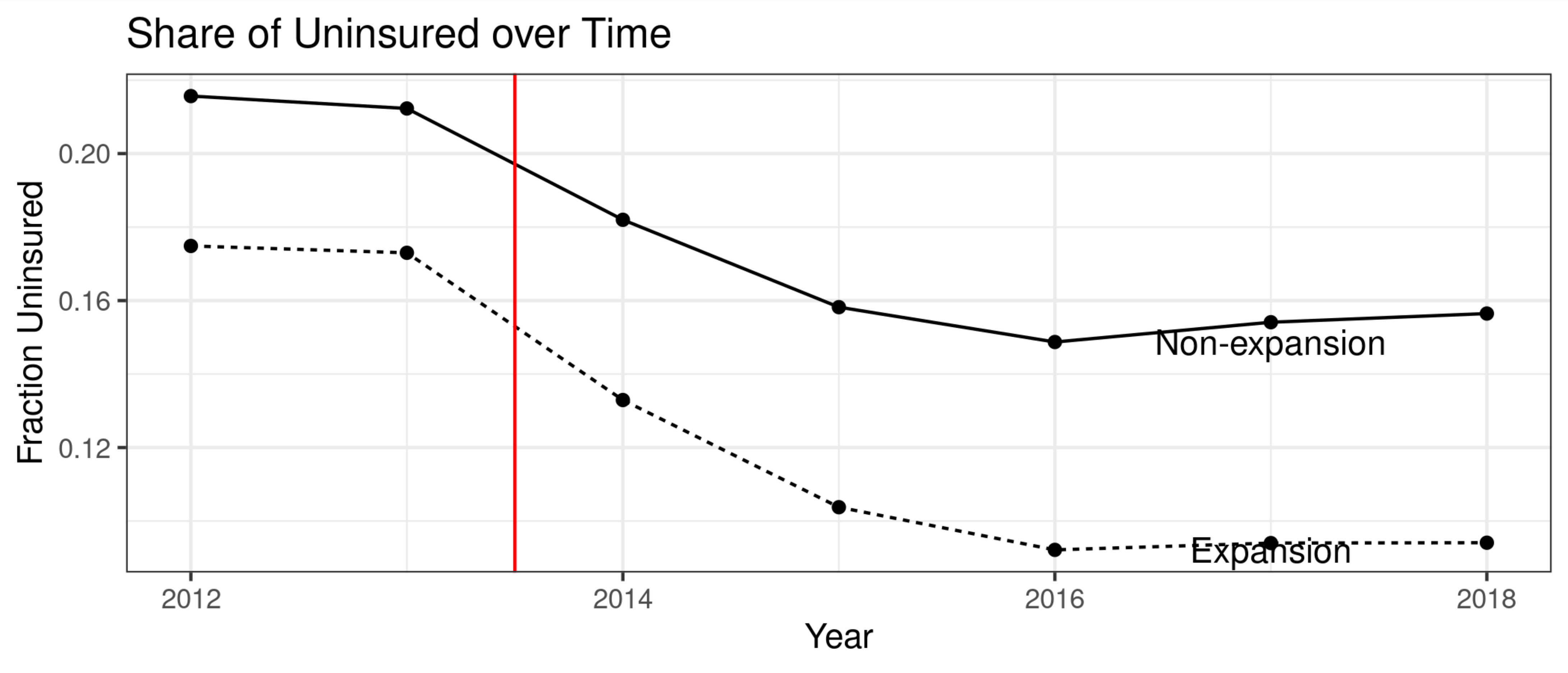
Checking pre-trends

First just plot seperately by group:

```r
ins.plot.dat <- ins.dat %>% filter(!is.na(expand_ever)) %>%
  group_by(expand_ever, year) %>% summarize(mean=mean(perc_unins))

ggplot(data=ins.plot.dat, aes(x=year,y=mean,group=expand_ever,linetype=expand_ever)) +
  geom_line() + geom_point() + theme_bw() +
  geom_vline(xintercept=2013.5, color="red") +
  geom_text(data = ins.plot.dat %>% filter(year == 2016),
            aes(label = c("Non-expansion","Expansion"),
                x = year + 1,
                y = mean)) +
  guides(linetype=FALSE) +
  labs(
    x="Year",
    y="Fraction Uninsured",
    title="Share of Uninsured over Time"
  )
```



Checking pre-trends



Event study

This is poorly named:

- In finance, even study is just an *interrupted time series*
- In economics, we usually have a treatment/control group *and* a break in time

Event study

First create all of the treatment/year interactions:

```
event.dat <- ins.dat.2014 %>%
  mutate(expand_2012 = expand_ever*(year==2012),
         expand_2013 = expand_ever*(year==2013),
         expand_2014 = expand_ever*(year==2014),
         expand_2015 = expand_ever*(year==2015),
         expand_2016 = expand_ever*(year==2016),
         expand_2017 = expand_ever*(year==2017),
         expand_2018 = expand_ever*(year==2018))
```

Event study

Second, run regression with full set of interactions and group/year dummies:

```
event.ins.reg <- lm(perc_unins ~ expand_2012 + expand_2014 +
                      expand_2015 + expand_2016 + expand_2017 +
                      expand_2018 + factor(year) + factor(State), data=event.dat)
point.est <- as_tibble(c(event.ins.reg$coefficients[c("expand_2012","expand_2014","expand_2015",
                                            "expand_2016","expand_2017","expand_2018")]),
                       rownames = "term")
ci.est <- as_tibble(confint(event.ins.reg)[c("expand_2012","expand_2014","expand_2015",
                                    "expand_2016","expand_2017","expand_2018"),],
                    rownames = "term")
```

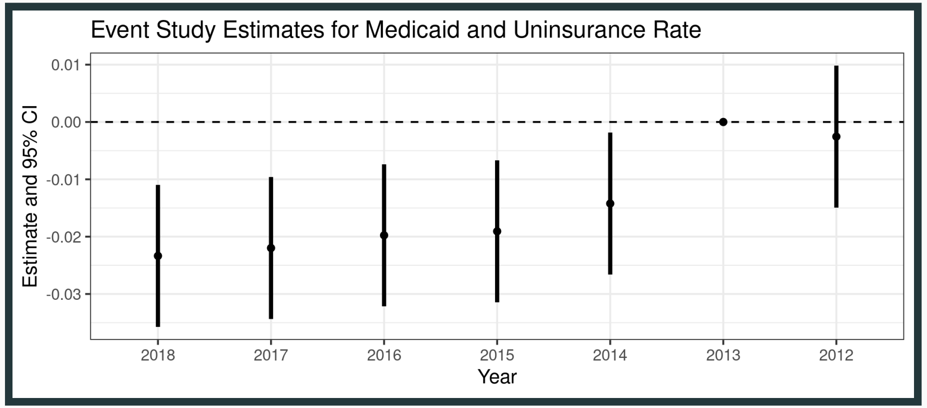
Event study

Third, organize results into a new dataset:

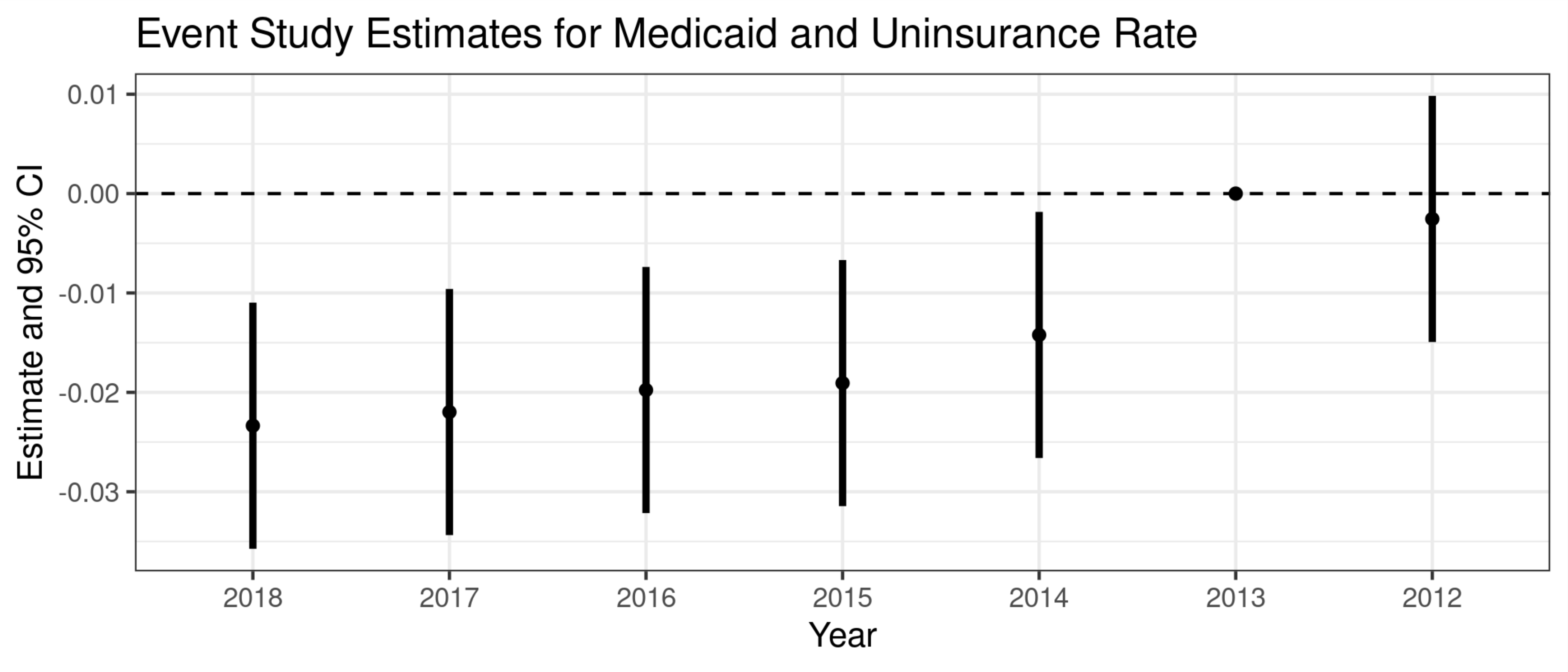
```
point.est <- point.est %>% rename(estimate = value)
ci.est <- ci.est %>% rename(conf.low = `2.5 %`, conf.high = `97.5 %`)
new.row <- tibble(
  term = "expand_2013",
  estimate = 0,
  conf.low = 0,
  conf.high = 0,
  year = 2013
)

event.plot.dat <- point.est %>%
  left_join(ci.est, by=c("term")) %>%
  mutate(year = c(2012, 2014, 2015, 2016, 2017, 2018)) %>%
  bind_rows(new.row) %>%
  arrange(year)
```

Event study

Finally, plot coefficients and confidence intervals



Event study


Event Study Estimates for Medicaid and Uninsurance Rate
Estimate and 95% CI
0.01
0.00
-0.01
-0.02
-0.03
2018
2017
2016
2015
2014
2013
2012
Year

Fixed Effects?

Recall our original regression specification:

$$y_{it} = \alpha + \beta \times 1(Post) + \gamma \times 1(Expand) + \delta \times 1(Post) \times 1(Expand) + \varepsilon$$

This is a special case of a general fixed effects estimator:

$$y_{it} = \alpha + \delta W_{it} + \gamma_i + \gamma_t + \varepsilon,$$

where γ_i and γ_t denote a set of coefficients on state (\$i\$) and year (\$t\$) dummy variables (or fixed effects).

Fixed Effects?

In R, we can estimate the fixed effects specification using the `felm` command (among others), which is part of the `lfe` package. Intuitively, the treatment dummy is now captured by γ_i and the pre/post dummy is captured by γ_t. For small datasets, we can estimate γ_i and γ_t directly. For large datasets, the "fixed effects" estimators will "remove" those variables by first differencing or mean differencing each variable in the regression.

Equivalence

DD is just a special case of the fixed effects approach.

```
summary(lm(perc_unins ~ post + expand_ever + post*expand
```

```
##
## Call:
## lm(formula = perc_unins ~ post + expand_ever + post * expand_ever,
##     data = ins.dat.2014)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.115667 -0.027106 -0.006804  0.027765  0.117597
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.213965   0.007180  29.799  < 2e-16 ***
## postTRUE                 -0.054068   0.008496  -6.364 7.22e-10 ***
## expand_everTRUE          -0.046326   0.009166  -5.054 7.48e-07 ***
## postTRUE:expand_everTRUE -0.018403   0.010845  -1.697   0.0908 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(felm(perc_unins ~ treat | factor(State) + factor
```

```
##
## Call:
##    felm(formula = perc_unins ~ treat | factor(State) + factor
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.042349 -0.007307 -0.000520  0.007342  0.039814
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## treat -0.018403   0.003702  -4.971 1.22e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 0.01429 on 257 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared(full model): 0.9507   Adjusted R-squared:
```