Linear Regression continued

PSY 300

Regression with dummy variables

- Last time, we discussed a fictitious experiment in which we randomly assigned 500 people to a no-social-media treatment, and 500 people to a control group. We then ran the regression of life satisfaction (y) on treatment (T):

Regression with dummy variables

- Last time, we discussed a fictitious experiment in which we randomly assigned 500 people to a no-social-media treatment, and 500 people to a control group. We then ran the regression of life satisfaction (y) on treatment (T):

```
> fit <- lm(y ~ T, data=data)
> summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.04371  0.04375    92.43   <2e-16 ***
T           2.04231  0.06187    33.01   <2e-16 ***
```

The *t*-test equivalent

- We said this is the same as an independent-samples t-test. Let's check if this is true!

The *t*-test equivalent

- We said this is the same as an independent-samples t-test. Let's check if this is true!

```
> res<-t.test(y ~ T, data=data)
> res
Welch Two Sample t-test
data: y by T
t = -33.01, df = 997.34, p-value < 2.2e-16
alternative hypothesis: true difference in means is not
95 percent confidence interval:
-2.163721 -1.920900
sample estimates:
mean in group 0 mean in group 1
4.043709 6.086019
```

The *t*-test equivalent

- We said this is the same as an independent-samples t-test. Let's check if this is true!

```
> res<-t.test(y ~ T, data=data)
> res
Welch Two Sample t-test
data: y by T
t = -33.01, df = 997.34, p-value < 2.2e-16
alternative hypothesis: true difference in means is not
95 percent confidence interval:
-2.163721 -1.920900
sample estimates:
mean in group 0 mean in group 1
4.043709 6.086019
```

- This is the same result. Regression of an outcome on a treatment dummy is the same as an independent samples *t*-test comparing the treatment and control group.

Regression with clustered standard errors

- Consider an example where instead of assigning individual students to treatment, we assign entire classrooms to treatment. Maybe this is necessary for administrative reasons; or we are worried about spillovers if we were to randomize within classrooms. (Question to think about: in which direction would such spillovers bias the treatment effect?)

Regression with clustered standard errors

- Consider an example where instead of assigning individual students to treatment, we assign entire classrooms to treatment. Maybe this is necessary for administrative reasons; or we are worried about spillovers if we were to randomize within classrooms. (Question to think about: in which direction would such spillovers bias the treatment effect?)
- We still have 1000 participants, 500 treatment, 500 control; but now they are organized in classrooms. Each classroom has 10 students, so we have 100 classrooms in total. 50 classrooms are assigned to treatment, so all students in each of these classrooms receive treatment. 50 classrooms are assigned to control, so all students in each of these classrooms are in the control condition.

Regression with clustered standard errors

- Each of the 100 classrooms is a *cluster*. Because randomization is at the classroom level, and we measure outcomes at the student level, we have to adjust standard errors for clustering at that level.

Regression with clustered standard errors

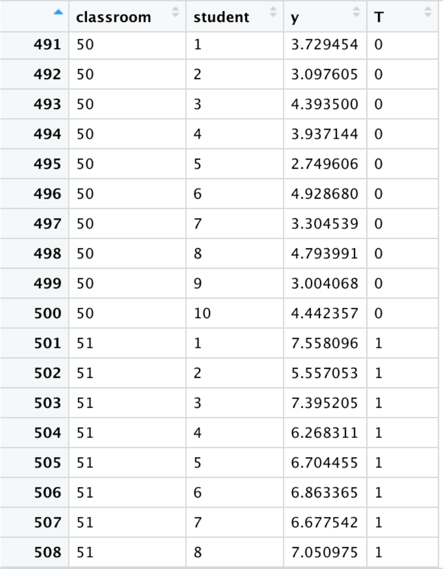
- Each of the 100 classrooms is a *cluster*. Because randomization is at the classroom level, and we measure outcomes at the student level, we have to adjust standard errors for clustering at that level.
    - (Remember the criterion for clustering: *we have to cluster our standard errors whenever randomization occurs at a level above the unit of observation*. Here, the level at which randomization occurs is the classroom: we are randomly assigning classrooms to treatment and control. The unit of observation is the student.)

Regression with clustered standard errors

- Each of the 100 classrooms is a *cluster*. Because randomization is at the classroom level, and we measure outcomes at the student level, we have to adjust standard errors for clustering at that level.
  - (Remember the criterion for clustering: *we have to cluster our standard errors whenever randomization occurs at a level above the unit of observation*. Here, the level at which randomization occurs is the classroom: we are randomly assigning classrooms to treatment and control. The unit of observation is the student.)
- The reason we need to cluster is because students within a classroom may be more similar to each other than they are to students in other classrooms. This is called "positive intra-cluster correlation."

Data

Here are some data from both the treatment and control groups:



	classroom	student	y	T
491	50	1	3.729454	0
492	50	2	3.097605	0
493	50	3	4.393500	0
494	50	4	3.937144	0
495	50	5	2.749606	0
496	50	6	4.928680	0
497	50	7	3.304539	0
498	50	8	4.793991	0
499	50	9	3.004068	0
500	50	10	4.442357	0
501	51	1	7.558096	1
502	51	2	5.557053	1
503	51	3	7.395205	1
504	51	4	6.268311	1
505	51	5	6.704455	1
506	51	6	6.863365	1
507	51	7	6.677542	1
508	51	8	7.050975	1

Results without clustering

```
> fit <- lm(y ~ T, data=data)
> summary(fit) # show results
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.96478 0.04506 87.99 <2e-16 ***
T 1.96121 0.06373 30.78 <2e-16 ***
```

Results with clustering

```
> install.packages("plm")
> install.packages("lmtest")
> library("plm")
> library("lmtest")
> model <- plm(y ~ T, data = data, model = "pooling")
> # compute the degrees of freedom and covariance matrix
> G <- length(unique(data$classroom))
> N <- length(data$classroom)
> df <- (G/(G - 1)) * (N - 1)/pm1$df.residual
> covariance <- df * vcovHC(model, type = "HC0", cluster
> # display the results
> coeftest(model, vcov = covariance)
t test of coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  3.96478    0.10841  36.572 < 2.2e-16 ***
T            1.96121    0.15292  12.825 < 2.2e-16 ***
```

Results with clustering

- In the results with clustering, the coefficient estimate is the same; but the standard errors are larger. This reflects that because of the positive correlation between students within classrooms, we have fewer “effective observations” in the sample.

Results with clustering

- In the results with clustering, the coefficient estimate is the same; but the standard errors are larger. This reflects that because of the positive correlation between students within classrooms, we have fewer "effective observations" in the sample.
- You don't have to understand all the code on the previous slide. Whenever you have to do this for your projects, ask us – we can help you with the code.

t-test equivalent

- t-tests cannot be adjusted for clustered standard errors.

t-test equivalent

- *t*-tests cannot be adjusted for clustered standard errors.
- But we can do something that will give us the same result: we can compute the average outcome of each cluster, and run a *t*-test on the averaged data!

t-test equivalent

- *t*-tests cannot be adjusted for clustered standard errors.

t-test equivalent

- *t*-tests cannot be adjusted for clustered standard errors.
- But we can do something that will give us the same result: we can compute the average outcome (e.g. life satisfaction) of each cluster, and run a *t*-test on the averaged data!

t-test equivalent

- t-tests cannot be adjusted for clustered standard errors.
- But we can do something that will give us the same result: we can compute the average outcome (e.g. life satisfaction) of each cluster, and run a t-test on the averaged data!
- Here: we would compute the mean GPA of each classroom. Now the unit of observation is the classroom, and the level of clustering is also the classroom. This means we don't have to cluster standard errors. So we can just run a t-test.

t-test equivalent

- *t*-tests cannot be adjusted for clustered standard errors.
- But we can do something that will give us the same result: we can compute the average outcome (e.g. life satisfaction) of each cluster, and run a *t*-test on the averaged data!
- Here: we would compute the mean GPA of each classroom. Now the unit of observation is the classroom, and the level of clustering is also the classroom. This means we don't have to cluster standard errors. So we can just run a *t*-test.

```
> t.test(y ~ T, data=data)
> t = -12.767, df = 97.997, p-value < 2.2e-16
```

t-test equivalent

- *t*-tests cannot be adjusted for clustered standard errors.
- But we can do something that will give us the same result: we can compute the average outcome (e.g. life satisfaction) of each cluster, and run a *t*-test on the averaged data!
- Here: we would compute the mean GPA of each classroom. Now the unit of observation is the classroom, and the level of clustering is also the classroom. This means we don't have to cluster standard errors. So we can just run a *t*-test.

  ```
  > t.test(y ~ T, data=data)
  > t = -12.767, df = 97.997, p-value < 2.2e-16
  ```

- Thus, the t-test on the collapsed data reproduces (reasonably exactly) the results of the regression with clustered standard errors.

Fixed effects and repeated measures

- Sometimes we can assign each unit of observation to both treatment and control.

Fixed effects and repeated measures

- Sometimes we can assign each unit of observation to both treatment and control.
- In our example: suppose we exposed half of the students first to the control condition, and then to the treatment condition. The other half get exposed first to the treatment condition, then to the control condition. We record their life satisfaction immediately after the completion of each condition.

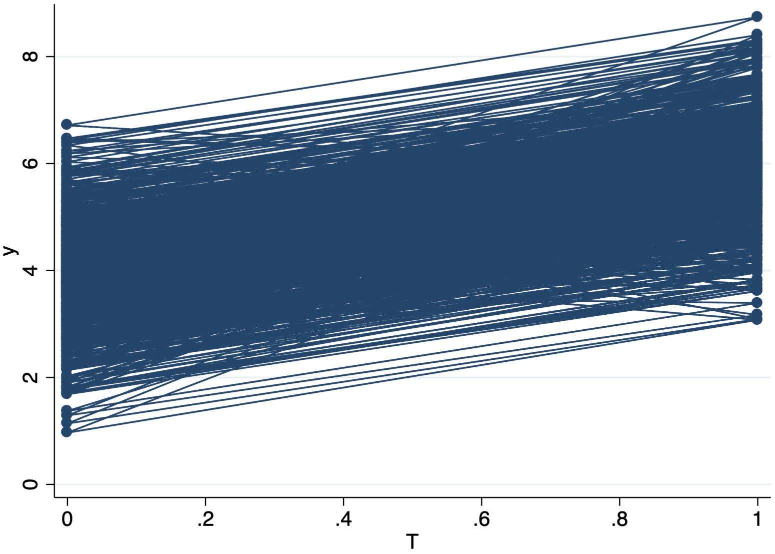
Fixed effects and repeated measures

- Sometimes we can assign each unit of observation to both treatment and control.
- In our example: suppose we exposed half of the students first to the control condition, and then to the treatment condition. The other half get exposed first to the treatment condition, then to the control condition. We record their life satisfaction immediately after the completion of each condition.
- This makes it possible to compare each student *to themselves* in the other condition. This can reduce variation and give us more precise standard errors. A good thing!

Fixed effects and repeated measures

- Sometimes we can assign each unit of observation to both treatment and control.
- In our example: suppose we exposed half of the students first to the control condition, and then to the treatment condition. The other half get exposed first to the treatment condition, then to the control condition. We record their life satisfaction immediately after the completion of each condition.
- This makes it possible to compare each student *to themselves* in the other condition. This can reduce variation and give us more precise standard errors. A good thing!
- This is also called "repeated measures" because we have more than one measure per participant (or other unit of observation).

Fixed effects and repeated measures



Fixed effects and repeated measures

- The graph on the previous slide shows the life satisfaction (y) of each participant under treatment (T=1) and control (T=0) conditions.

Fixed effects and repeated measures

- The graph on the previous slide shows the life satisfaction (y) of each participant under treatment (T=1) and control (T=0) conditions.
- There is considerable variability across participants in how high their average satisfaction is: some participants have high life satisfaction under control conditions, others have low life satisfaction. It looks like treatment increases life satisfaction by a similar amount from these baseline levels for many participants.

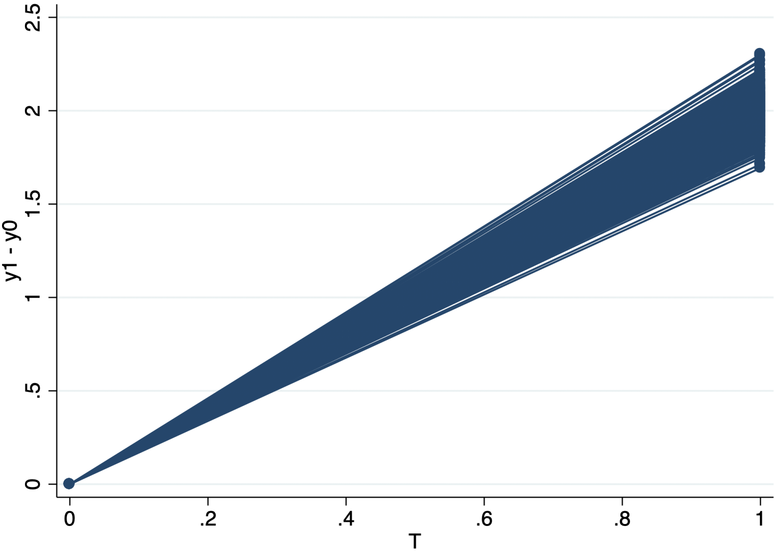
Fixed effects and repeated measures

- The graph on the previous slide shows the life satisfaction (y) of each participant under treatment (T=1) and control (T=0) conditions.
- There is considerable variability across participants in how high their average satisfaction is: some participants have high life satisfaction under control conditions, others have low life satisfaction. It looks like treatment increases life satisfaction by a similar amount from these baseline levels for many participants.
- How can we make this more precise? We can compare each participant against themselves over time.

Fixed effects and repeated measures

- The graph on the previous slide shows the life satisfaction (y) of each participant under treatment (T=1) and control (T=0) conditions.
- There is considerable variability across participants in how high their average satisfaction is: some participants have high life satisfaction under control conditions, others have low life satisfaction. It looks like treatment increases life satisfaction by a similar amount from these baseline levels for many participants.
- How can we make this more precise? We can compare each participant against themselves over time.
- The easiest way to do this is to subtract life satisfaction in the control condition from life satisfaction in the treatment condition for each participant

Fixed effects and repeated measures



Fixed effects and repeated measures

- The graph on the previous slide shows the life satisfaction (y) of each participant under treatment (T=1) and control (T=0) conditions, after subtracting life satisfaction under control conditions.

Fixed effects and repeated measures

- The graph on the previous slide shows the life satisfaction (y) of each participant under treatment (T=1) and control (T=0) conditions, after subtracting life satisfaction under control conditions.
- We have now reduced the variability considerably: We have removed the variability between participants under control conditions, and retained only the variability in the treatment effect. You can see visually that the dots are closer together.

Fixed effects and repeated measures

- In making this graph, we defined a new variable: the difference in life satisfaction between treatment and control conditions in each participant.

$$\Delta y_i = y_{i,T=1} - y_{i,T=0}$$

Fixed effects and repeated measures

- In making this graph, we defined a new variable: the difference in life satisfaction between treatment and control conditions in each participant.

$$\Delta y_i = y_{i,T=1} - y_{i,T=0}$$

- Testing whether there is a significant difference between treatment and control conditions can now be done by simply testing whether the mean of this new variable is different from zero:

$$t = \frac{estimator}{SE} = \frac{\bar{\Delta} y - 0}{\frac{SD(\Delta y_i)}{\sqrt{n}}}$$

Fixed effects and repeated measures

- In making this graph, we defined a new variable: the difference in life satisfaction between treatment and control conditions in each participant.

$$\Delta y_i = y_{i,T=1} - y_{i,T=0}$$

- Testing whether there is a significant difference between treatment and control conditions can now be done by simply testing whether the mean of this new variable is different from zero:

$$t = \frac{estimator}{SE} = \frac{\bar{\Delta y} - 0}{\frac{SD(\Delta y_i)}{\sqrt{n}}}$$

- **This is a repeated-measures t-test. (It's really a one-sample t-test in disguise: notice we're just comparing a single mean to zero.)**

Fixed effects and repeated measures

- Let's run a t-test in R that pretends the data are unpaired, i.e. we simply compare the treatment and control condition means without taking advantage of the fact that the data are paired.

Fixed effects and repeated measures

- Let's run a t-test in R that pretends the data are unpaired, i.e. we simply compare the treatment and control condition means without taking advantage of the fact that the data are paired.

```
> t.test(y ~ T, data=data)
Welch Two Sample t-test
data:  y by T
t = -43.118, df = 1998, p-value < 2.2e-16
```

Fixed effects and repeated measures

- Let's run a t-test in R that pretends the data are unpaired, i.e. we simply compare the treatment and control condition means without taking advantage of the fact that the data are paired.

  ```
  > t.test(y ~ T, data=data)
  Welch Two Sample t-test
  data:  y by T
  t = -43.118, df = 1998, p-value < 2.2e-16
  ```

- This result is highly significant, but it's wrong: it doesn't account for clustering; we have two observations per participant. In addition, we can make it even more precise by taking advantage of the paired nature of the data. Note the degrees of freedom are 1998, reflecting the fact that we have 2000 observations (1000 students, each measured twice).

Fixed effects and repeated measures

- Now let's compute the difference variable, Δy_i, and run a one-sample t-test comparing it to zero.

Fixed effects and repeated measures

- Now let's compute the difference variable, Δy_i, and run a one-sample *t*-test comparing it to zero.

```
> t.test(data$ym, data=data, mu=0)
One Sample t-test
data: data$ym
t = 651.17, df = 999, p-value < 2.2e-16
```

Fixed effects and repeated measures

- Now let's compute the difference variable, Δy_i, and run a one-sample *t*-test comparing it to zero.

```
> t.test(data$ym, data=data, mu=0)
One Sample t-test
data: data$ym
t = 651.17, df = 999, p-value < 2.2e-16
```

- This result is much more precise than the previous one: the t-statistic is much higher. Thus, taking advantage of the repeated-measures nature of the data has bought us a lot of precision.

Fixed effects and repeated measures

- Now let's compute the difference variable, Δy_i, and run a one-sample *t*-test comparing it to zero.

```
> t.test(data$ym, data=data, mu=0)
One Sample t-test
data: data$ym
t = 651.17, df = 999, p-value < 2.2e-16
```

- This result is much more precise than the previous one: the t-statistic is much higher. Thus, taking advantage of the repeated-measures nature of the data has bought us a lot of precision.
- The degrees of freedom are now 999, reflecting the fact that we have a single observation for each of the 1000 participants, reflecting their personal difference between treatment and control.

Regression equivalent

- The regression equivalent of a repeated-measures t-test uses "fixed effects". Without fixed effects:

$$y_i = \beta_0 + \beta_1 T_i + \varepsilon_i$$

Regression equivalent

- The regression equivalent of a repeated-measures t-test uses "fixed effects". Without fixed effects:

$$y_i = \beta_0 + \beta_1 T_i + \varepsilon_i$$

- With fixed effects:

$$y_{it} = \alpha_i + \beta_1 T_{it} + \varepsilon_{it}$$

Regression equivalent

- The regression equivalent of a repeated-measures t-test uses "fixed effects". Without fixed effects:

$$y_i = \beta_0 + \beta_1 T_i + \varepsilon_i$$

- With fixed effects:

$$y_{it} = \alpha_i + \beta_1 T_{it} + \varepsilon_{it}$$

- We have added an index "t" here to index the first and the second test for each participant.

Regression equivalent

- The regression equivalent of a repeated-measures t-test uses “fixed effects”. Without fixed effects:

$$y_i = \beta_0 + \beta_1 T_i + \varepsilon_i$$

- With fixed effects:

$$y_{it} = \alpha_i + \beta_1 T_{it} + \varepsilon_{it}$$

- We have added an index “t” here to index the first and the second test for each participant.
- The crucial addition is α_i: this is called the “individual-level fixed effect”.

Regression equivalent

$$y_{it} = \alpha_i + \beta_1 T_{it} + \varepsilon_{it}$$

- You can think of α_i as follows: it subtracts the individual-level mean of the outcome variable from each observation. So for each participant, it computes the average y of that participant under treatment and control conditions, and subtracts it from both the treatment and control values of that participant. For treatment effect estimation, this is equivalent to what we did above, i.e. subtracting the control group outcome.

Regression equivalent

- Let's run the regression in R without clustering at the subject level

Regression equivalent

- Let's run the regression in R without clustering at the subject level

```
> model = lm(y~T, data=data)
> summary(model)
Coefficients:
            Estimate   Std. Error t value Pr(>|t|)
(Intercept) 3.92873    0.03275    119.95  <2e-16 ***
T           1.99731    0.04632    43.12 <2e-16 ***
```

Regression equivalent

- Let's run the regression in R without clustering at the subject level

```
> model = lm(y~T, data=data)
> summary(model)
Coefficients:
            Estimate   Std. Error t value Pr(>|t|)
(Intercept) 3.92873    0.03275    119.95  <2e-16 ***
T           1.99731    0.04632    43.12 <2e-16 ***
```

- Again, this is wrong: it doesn't cluster standard errors appropriately; and it doesn't take into account the repeated-measures nature of the data.

Regression equivalent

- Now let's include individual-level fixed effects:

Regression equivalent

- Now let's include individual-level fixed effects:

```
> model = lm(y~T+factor(sid), data=data)
> summary(model)
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.334129   0.048522 109.932  < 2e-16 **
T                1.997306   0.003067 651.169  < 2e-16 **
```

Regression equivalent

- Now let's include individual-level fixed effects:

  ```
  > model = lm(y~T+factor(sid), data=data)
  > summary(model)
  Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
  (Intercept)      5.334129   0.048522 109.932  < 2e-16 **
  T                1.997306   0.003067 651.169  < 2e-16 **
  ```

- Note the much higher precision in identifying the treatment effect: t is now 651! Previously it was 43. So we have much more power with the inclusion of fixed effects.

Summary of t-test vs. regression

- Two-sample t-tests are regressions on a dummy:

$$y_i = \beta_0 + \beta_1 T_i + \varepsilon_i$$

- Repeated-measures t-tests are regressions on a dummy with individual-level fixed effects:

$$y_{it} = \alpha_i + \beta_1 T_{it} + \varepsilon_{it}$$