Regressions in R

When to use Regression Analysis?

When to use Regression Analysis?

- Trying to identify causation
- Correlation vs causation
  - Height vs. weight
  - Get taller gain weight!
  - Spurious correlation

When to use Regression Analysis?

- Regression analysis is used to describe the relationship between:
  - A single response variable Y and
  - One or more predictor variable X1, X2, … Xn
    - If n = 1 i.e. 1 independent variable then it's just a simple regression
    - n > 1 then multivariate regression
- What conditions must the response variable meet for OLS?

When to use Regression Analysis?

- Regression analysis is used to describe the relationship between:
  - A single response variable Y and
  - One or more predictor variable X1, X2, … Xn
    - If n = 1 i.e. 1 independent variable then it's just a simple regression
    - n > 1 then multivariate regression
- What conditions must the response variable meet for OLS?
  - Continuous! but …
- What conditions must the predictor variables meet?

When to use Regression Analysis?

- Regression analysis is used to describe the relationship between:
  - A single response variable Y and
  - One or more predictor variable X1, X2, … Xn
    - If n = 1 i.e. 1 independent variable then it's just a simple regression
    - n > 1 then multivariate regression
- What conditions must the response variable meet?
  - Continuous! but …
- What conditions must the predictor variables meet?
  - None! These variables can be continuous, discrete, or categorical

Steps to take prior to analysis?

Steps to take prior to analysis?

- Check for:
    - Missing values
    - Outliers
    - Asymmetric distributions
    - Clustering
    - Unexpected patterns
- Numerical Summaries
    - Mean, min, max, variance etc.
    - Correlations
- Graphical Summaries
    - Scatter plots
    - Histograms
    - Boxplots

OLS Regression

- The Lahman package is an R package containing extensive statistics for baseball.
- First let's install the package and then load the library. As always once a package is downloaded you do not need to run the install.packages() command again.
- You can get a list of the data frames contained in the package by typing LahmanData

OLS Regression

- Load the following data frames from the package using the data() function
  - Salaries
  - Batting
  - Teams
- What variables do the data sets have in common?

OLS Regression

- I don't know much about baseball but lets see if we can put together a model to predict player salaries
- Find some documentation for the data sets
    - What are the AB and R variables in the Batting data frame?
    - What are the G, W, L, Division Winner, World Series Winner variables?

OLS Regression

- Create a new data frame (using dplyr) teams_small
  - Create a new columns frac_won
  - Take only the variables: playerID, lgID, teamID, name, Rank, G, frac_won
- Create a new data frame batting_small
  - Create a new column BA which is the number of hits /number at bats
  - Take only the variables playerID, yearID, teamID, lgID, BA, HR
- Now join all three data sets together

OLS Regression

- What are the dimensions of our data?
- What are some summary statistics for relevant variables?
- Let's make a scatter plot of
  - salary vs. BA
  - salary vs HR
- Why might these scatter plots not be the most informative?
- What is the correlation between BA and HR?
  - What problems can arise if your independent variables are highly correlated

OLS Regression

- What are the dimensions of our data?
- What are some summary statistics for relevant variables?
- Let's make a scatter plot of
  - salary vs. BA
  - salary vs HR
- Why might these scatter plots not be the most informative?
  - Inflation!
- How could we control for this in our regression without finding data on inflation?

OLS Regression

- My amazing model

$$\begin{aligned}\text{Salary}_{i,j,t} = \beta_0 + \beta_1 \text{Home Runs}_{i,j,t} + \beta_2 \text{Batting Average}_{i,j,t} \\ \beta_3 \text{Win Fraction}_{j,t} + \beta_4 \text{Games}_{j,t}\end{aligned}$$

- What do you think?
  - What variables might we be missing?
  - Any predictions for an R squared value?

OLS Regression

- lm(…)
- What are the arguments to the lm() function in R?

OLS Regression

- lm(…)
- What are the arguments to the lm() function in R?
- lm(formula, data, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, ...)

OLS Regression

```
# Multiple Linear Regression Example
fit <- lm(y ~ x1 + x2 + x3, data=mydata)
summary(fit) # show results
```

OLS Regression

- What is the structure of an lm object?
- What happens when you run summary on the linear model object?

OLS Regression

```
# Other useful functions
coefficients(fit) # model coefficients
confint(fit, level=0.95) # CIs for model parameters
fitted(fit) # predicted values
residuals(fit) # residuals
anova(fit) # anova table
vcov(fit) # covariance matrix for model parameters
influence(fit) # regression diagnostics
```

Table 1: Basic Baseball Salary Model

	Dependent variable:
	salary
HR	112,017.000***
	(2,882.950)
BA	−1,436,937.000***
	(222,402.200)
frac_won	157,854.100***
	(47,483.470)
G	29,219.110***
	(2,781.747)
Constant	−3,573,441.000***
	(518,772.300)
Observations	19,243
R^2	0.085
Adjusted R^2	0.085
Residual Std. Error	3,448,576.000 (df = 19238)
F Statistic	448.302*** (df = 4; 19238)
Note:	*p<0.1; **p<0.05; ***p<0.01

- What does the R squared value tell us?
- How seriously should we take the results of this model?
- What variable could we add that will probably make a huge improvement in the results?
- Statistical significance vs. economic significance

OLS Regression

- Residuals

OLS Regression

Table 1:

	Dependent variable:
	salary
HR	103,954.600***
	(2,682.381)
BA	−679,287.600***
	(206,898.900)
frac_won	470,362.600***
	(53,207.920)
G	−28,730.750
	(50,388.520)
as.factor(yearID)1986	−15,319.760
	(208,764.500)
as.factor(yearID)2015	4,499,609.000***
	(203,116.700)
Constant	2,494,775.000
	(8,142,606.000)
Observations	19,243
R^2	0.215
Adjusted R^2	0.214
Residual Std. Error	3,196,460.000 (df = 19208)
F Statistic	155.049*** (df = 34; 19208)
Note:	*p<0.1; **p<0.05; ***p<0.01

- How do we feel about the model results now?

OLS Regression

- Log salary variable

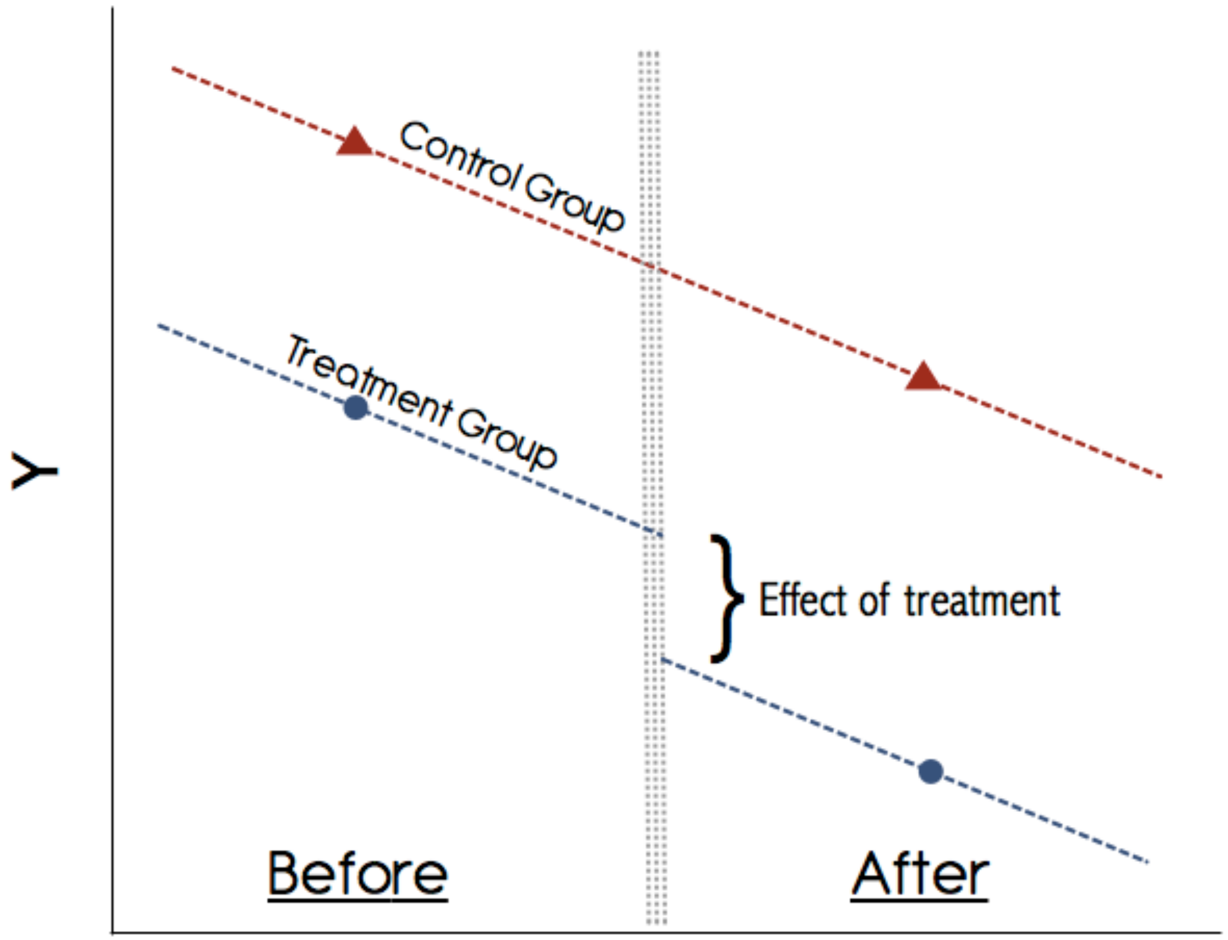
Diff in Diff Regressions

- What is a difference in difference regression?

Diff in Diff Regressions

- What is a difference in difference regression? (from NBER)
  - The simplest set up is one where outcomes are observed for two groups for two time periods.
  - One of the groups is exposed to a treatment in the second period but not in the first period.
  - The second group is not exposed to the treatment during either period.
- How does this work to remove potential bias?

DIFF IN DIFF



Control Group
Treatment Group
Y
Effect of treatment
Before
After

Diff in Diff Regressions

- Observing the same units within a group in each time period the average gain of the control group is subtracted from the the average gain of the treatment group
- Removes potential bias from permanent difference or time trends

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \times dB + \mu$$

Diff in Diff Regressions

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \times dB + \mu$$

- y is the outcome of interest
- d2 is a dummy variable for the second time period.
- The dummy variable dB captures possible differences between the treatment and control groups prior to the policy change.
- The time period dummy, d2, captures aggregate factors that would cause changes in y even in the absence of a policy change.
- The coefficient of interest is delta one multiplies the interaction term, d2 * dB, which is the same as a dummy variable equal to one for those observations in the treatment group in the second period.

Diff in Diff Regressions

- *Traffic Congestion and Infant Health – Janet Currie and Reed Walker*
- "We exploit the introduction of electronic toll collection, (E-ZPass), which greatly reduced both traffic congestion and vehicle emissions near highway toll plazas. We show that the introduction of E-ZPass reduced prematurity and low birth weight among mothers within 2km of a toll plaza by 10.8% and 11.8% respectively relative to mothers 2-10km from a toll plaza. There were no immediate changes in the characteristics of mothers or in housing prices near toll plazas that could explain these changes. The results are robust to many changes in specification and suggest that traffic congestion contributes significantly to poor health among infants."

Diff in Diff Regressions

- Does this immediately sound like an economics question?

Diff in Diff Regressions

- Does this immediately sound like an economics question?
  - “First, there is increasing evidence of the long-term effects of poor health at birth on future outcomes. For example, low birth weight has been linked to future health problems and lower educational attainment”
  - “The debate over the costs and benefits of emission controls and traffic congestion policies could be significantly impacted by evidence that traffic congestion has a deleterious effect on fetal health.”
  - “Second, the study of newborns overcomes several difficulties in making the connection between pollution and health because, unlike adult diseases that may reflect pollution exposure that occurred many years ago, the link between cause and effect is immediate”

Diff in Diff Regressions

- introduction of electronic toll collection, (E-ZPass)
- reduced both traffic congestion and vehicle emissions near highway toll plazas
- reduced prematurity and low birth weight among mothers within 2km of a toll plaza by 10.8% and 11.8% respectively **relative** to mothers 2-10km from a toll plaza.
- no immediate changes in the characteristics of mothers or in housing prices near toll plazas that could explain these changes.

Diff in Diff Regressions

- Who is the control group?
- Who is the treatment group?

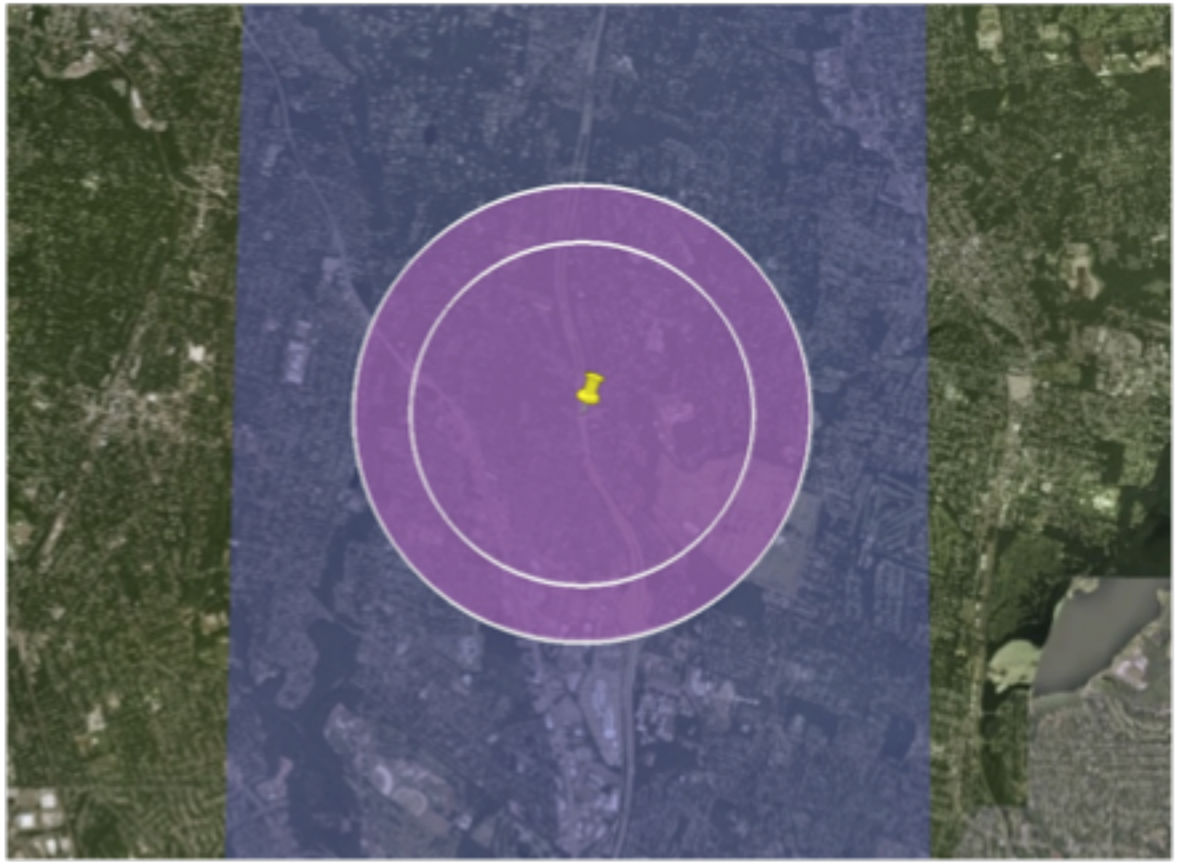
Diff in Diff Regressions

- "We compare the infant health outcomes of those living near an electronic toll plaza before and after implementation of E-ZPass to those living near a major highway but further away from a toll plaza. Specifically, we compare mothers within 2 kilometers of a toll plaza to mothers who are between 2 and 10 km from a toll plaza but still within 3 kilometers of a major highway before and after the adoption of E-ZPass in New Jersey and Pennsylvania"

$$Outcome_{it} = a + \beta_1 EZPass_{it} + \beta_2 Close_{it} + \beta_3 Plaza_{it} + \beta_4 EZPass_{it} * Close_{it}$$

$$+ \beta_5 Year + \beta_6 Month + \beta_7 X_{it} + \beta_8 Distance_{it} + e_{it},$$

Diff in Diff Regressions



Diff in Diff Regressions

- The data used in this paper is proprietary so I have had to generate random values but let’s see if we can replicate the methodology
  - What variables do we need to create?