Correlation and Regression

ML

1/22/2020

Set Up

Load the packages

```
library(tidyverse)
library(tidycensus)
library(pander)
```

Load the data. This is a dataset I put together merging several different datasets from Opportunity Insights.

```
county_mobility <- read.csv("https://raw.githubusercontent.com/mjclawrence/soci1001/master/data/county_
```

Correlations

Most of our engagement with data so far has focused on disributions of one variable. It can also be helpful to know how two (or more) variables tend to move together. Correlation coefficients measure such associations or relationships between variables.

Let's start with a basic correlation between `absolute_upward_mobility` and `social_capital_90` (the composite social capital measure Chetty et al and Weber et al use). We will use the `cor()` function to find correlations, and will need to add the `use = "complete"` option since some counties may be missing values on one or both of these measures.

```
cor(county_mobility$social_capital_90,
    county_mobility$absolute_upward_mobility,
    use = "complete")
```

```
## [1] 0.5444131
```

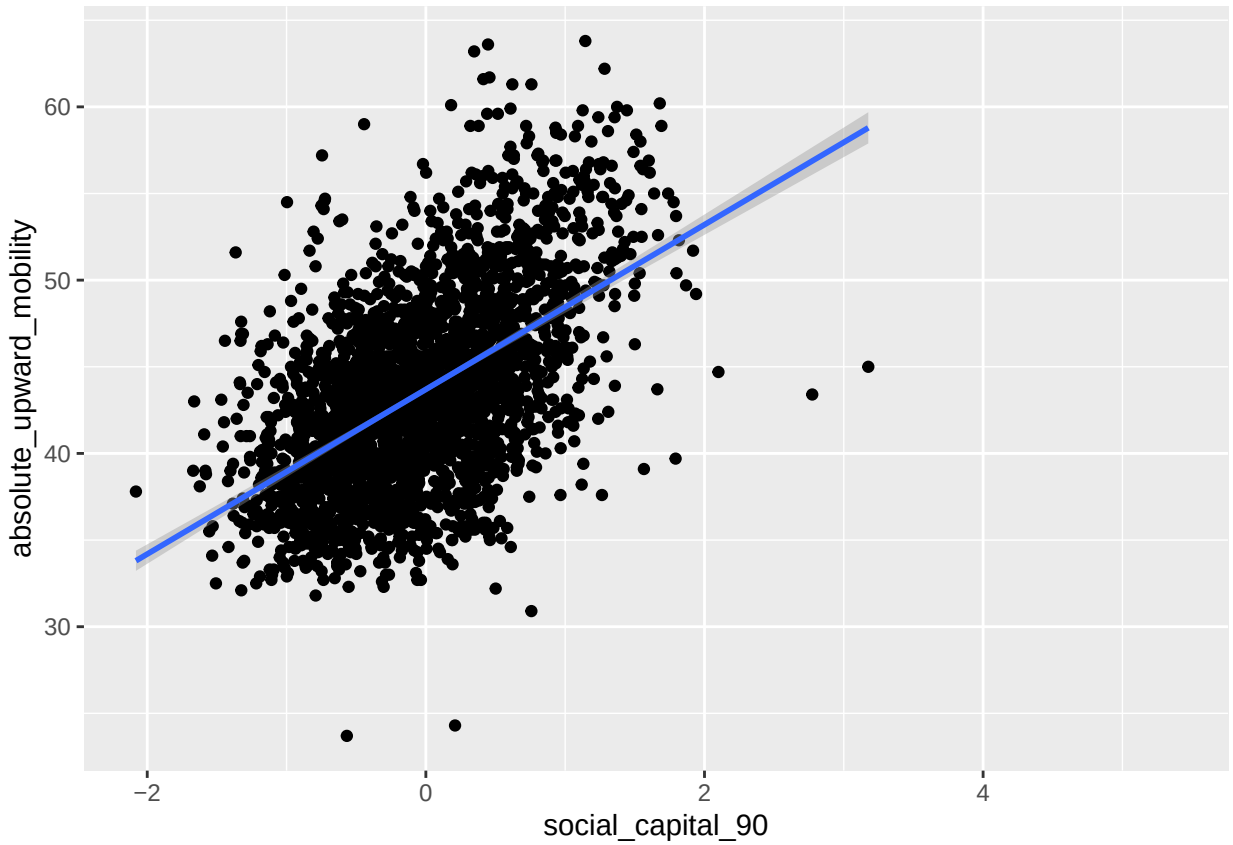
How would you describe this correlation?

The assocation between social capital and mobility may be easier to understand if we plot each county's social capital value (on the x axis) against each county's mobility value (on the y axis).

```
ggplot(county_mobility, aes(x = social_capital_90,
                            y = absolute_upward_mobility)) +
  geom_point() +  # for a scatterplot
  geom_smooth(method = "lm") # for a line of best fit
```

```
## Warning: Removed 395 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 395 rows containing missing values (geom_point).
```



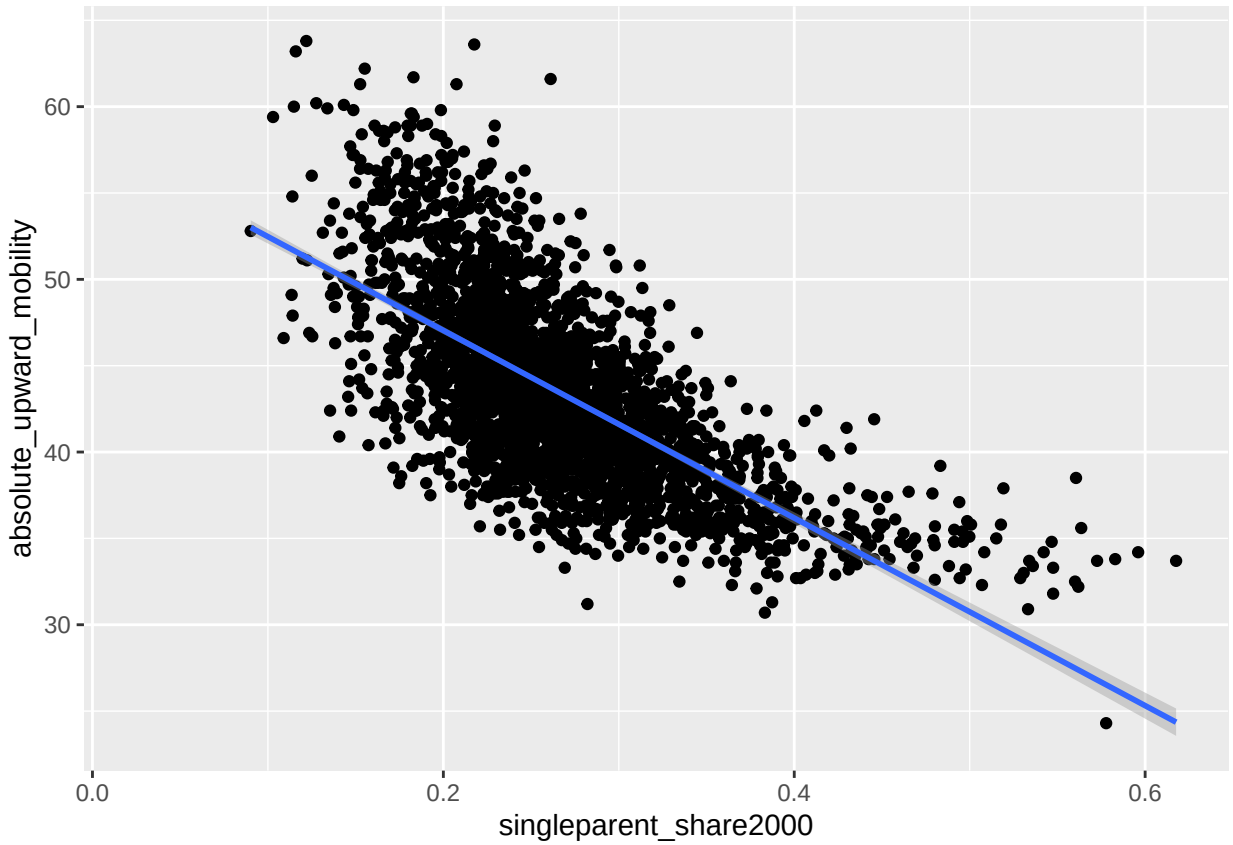
Find the correlation between mobility and `singleparent_share2000`. Make a scatterplot showing the association.

REPLACE THIS LINE WITH YOUR CODE

```
cor(county_mobility$singleparent_share2000,
    county_mobility$absolute_upward_mobility,
    use = "complete")
```

```
## [1] -0.6753921
```

```
ggplot(county_mobility, aes(x = singleparent_share2000,
                            y = absolute_upward_mobility)) +
  geom_point() + geom_smooth(method = "lm")
```



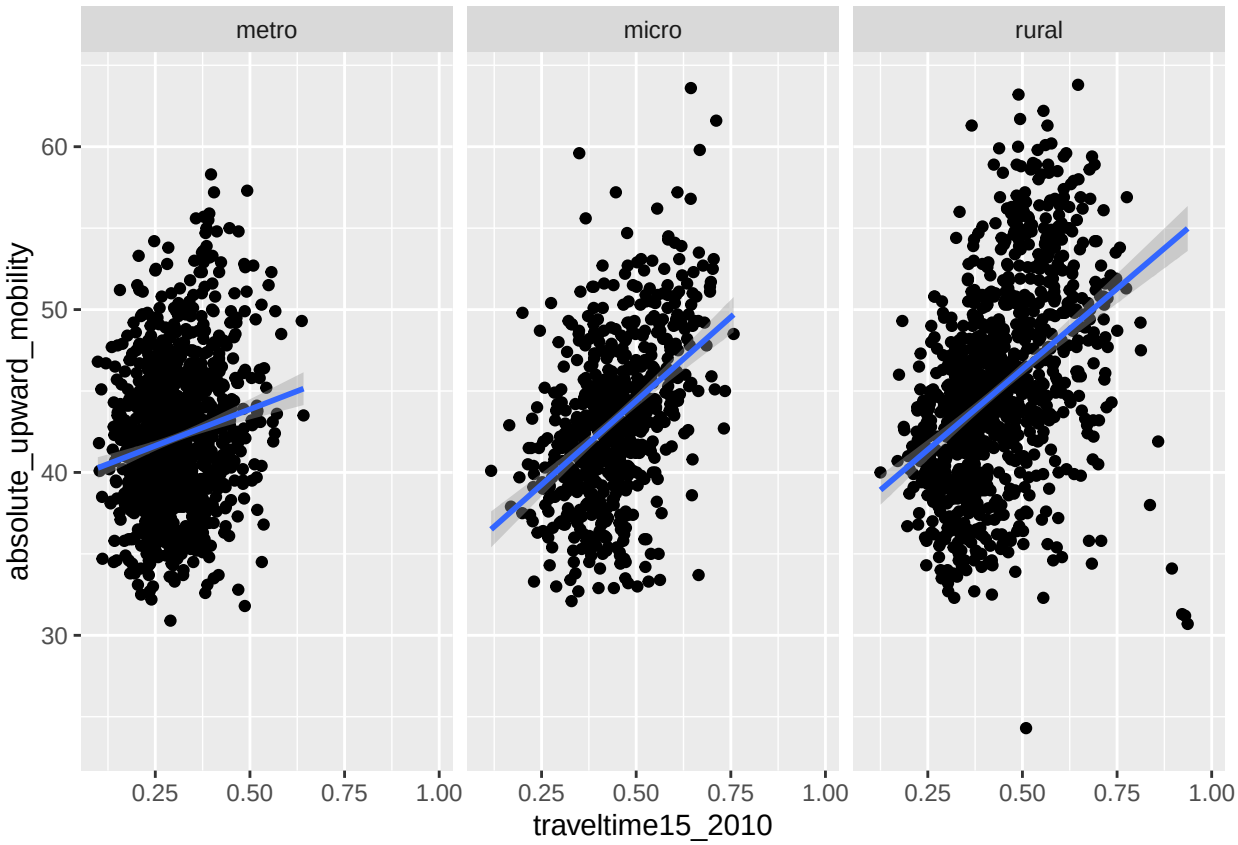
How could we see plots separately for metro, micro, and rural areas (categories of the `type`) variable? Try it with the association etween `traveltime15_2010` and `absolute_upward_mobility`.

REPLACE THIS LINE WITH YOUR CODE

```
county_mobility %>%
  ggplot(aes(x = traveltime15_2010, y = absolute_upward_mobility)) +
  geom_point() + geom_smooth(method = "lm") +
  facet_wrap(~type)
```

```
## Warning: Removed 376 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 376 rows containing missing values (geom_point).
```



We might be interested in seeing the correlations between mobility and all the variables considered the "big five" predictors of mobility. How can we summarize those correlations together? The `GGally` package has some neat tools to help visualize correlations. Install and load the package.

```
#install.packages("GGally") # Put a hashtag in front of this line after installing
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:pander':
##
##     wrap
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

We want to use the big five variables, so let's pull their names into a vector so we can reference them as a group.

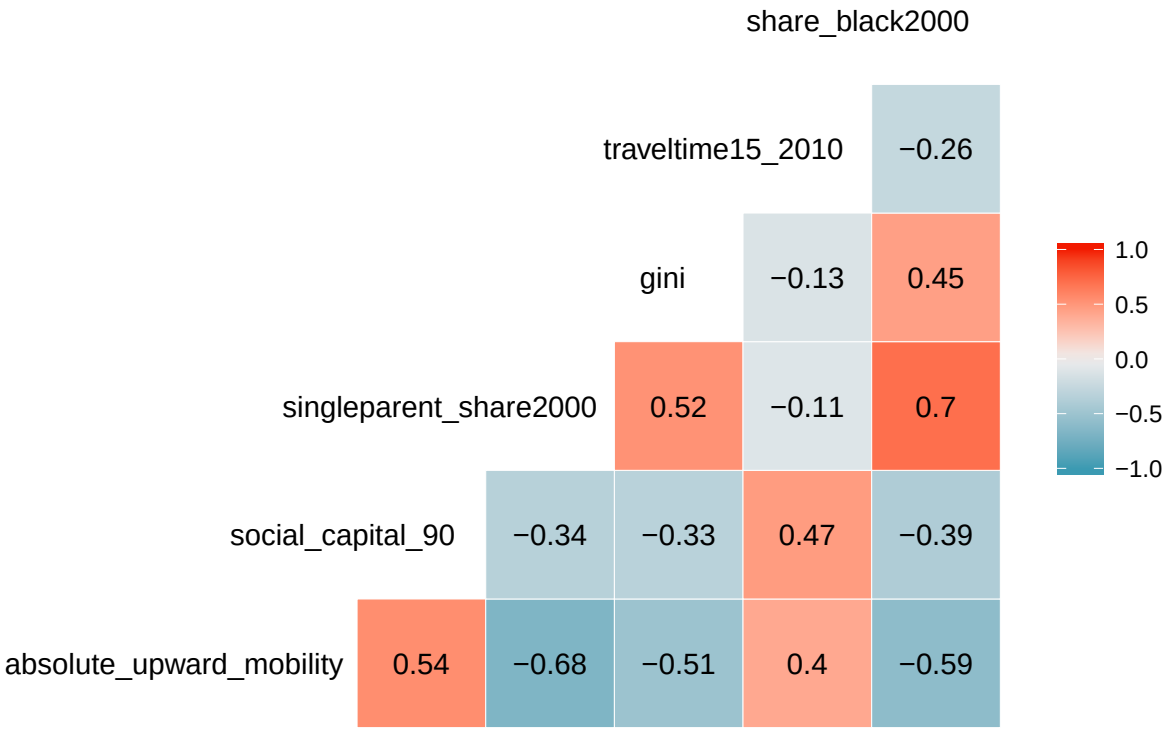
```
big_five_vars <- c("social_capital_90", "singleparent_share2000",
                   "gini", "traveltime15_2010",
                   "share_black2000")
```

And create a new data frame that only has the mobility variable and the big five variables.

```
big_five_correlations <- county_mobility %>%
  select(absolute_upward_mobility, big_five_vars)
```

We can use the `ggcorr()` funcion from the `GGally` package to visualize the correlations among all these variables.

```
ggcorr(big_five_correlations, # the data frame
       palette = "RdBu", # the color palette; this is the default
       label = TRUE, # show the correlation coefficient
       label_round = 2, # round the coefficient to two places
       hjust = .85, # move the variable labels away from the plot
       layout.exp = 2, # expand the layout of the plot
       method = c("pairwise", "pearson")) # type of correlations we want
```



How do you interpret these coefficients?

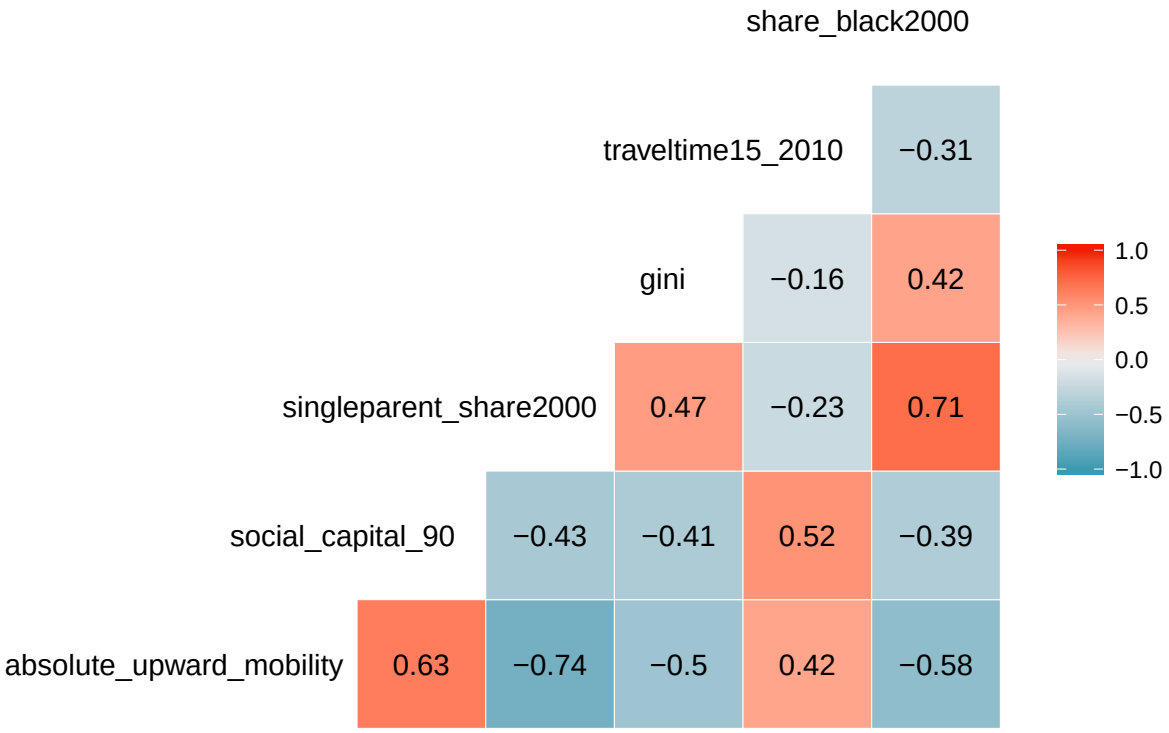
How can we look at this separately by type of county?

REPLACE THIS LINE WITH YOUR CODE

```
county_mobility %>%
  select(absolute_upward_mobility, big_five_vars, type) %>%
  filter(type == "rural") %>%
  ggcorr(palette = "RdBu", # the color palette; this is the default
       label = TRUE, # show the correlation coefficient
       label_round = 2, # round the coefficient to two places
       hjust = .85, # move the variable labels away from the plot
       layout.exp = 2, # expand the layout of the plot
       method = c("pairwise", "pearson")) # type of correlations we want
```

```
## Warning in ggcorr(., palette = "RdBu", label = TRUE, label_round = 2, hjust
```

```
## = 0.85, : data in column(s) 'type' are not numeric and were ignored
```



Here are the individual variables that make up the 1990 social capital index. There are descriptions of the individual variables at this site.

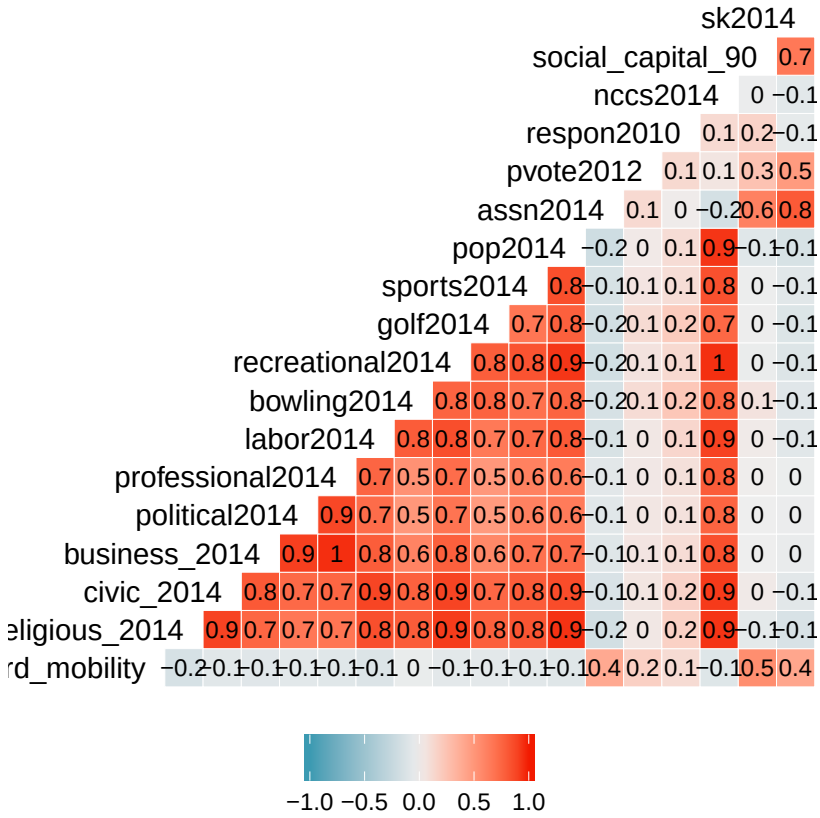
```
social_capital_vars <- c("religious_2014", "civic_2014", "business_2014",
                         "political2014", "professional2014", "labor2014",
                         "bowling2014", "recreational2014", "golf2014",
                         "sports2014", "pop2014", "assn2014", "pvote2012",
                         "respon2010", "nccs2014", "social_capital_90",
                         "sk2014")
```

How are they correlated with mobility?

REPLACE THIS LINE WITH YOUR CODE

```
county_mobility %>%
  select(absolute_upward_mobility, social_capital_vars, type) %>%
  ggcorr(palette = "RdBu",
         label = TRUE,
         label_round = 1,
         label_size = 3, # This is new; default is 4
         hjust = 1,
         layout.exp = 3,
         method = c("pairwise", "pearson"),
         legend.position = "bottom")
```

```
## Warning in ggcorr(., palette = "RdBu", label = TRUE, label_round = 1,
## label_size = 3, : data in column(s) 'type' are not numeric and were ignored
```



Our data frame has many other variables we have not looked at yet. Explore them in groups by type of county. What's the strongest positive correlation? What's the strongest negative correlation?

```
county_mobility %>%
  select(5:49, -big_five_vars, -social_capital_vars) %>%
  ggcorr(palette = "RdBu",
         label = TRUE,
         label_round = 1,
         hjust = 1,
         layout.exp = 3,
         method = c("pairwise", "pearson"),
         label_size = 3)
```

```
## Warning in ggcorr(., palette = "RdBu", label = TRUE, label_round = 1, hjust
## = 1, : data in column(s) 'type' are not numeric and were ignored
```

Regression

How can we do simple linear regressions in R? To see how regression works, find the average upward mobility for each county type.

REPLACE THIS LINE WITH YOUR CODE

```
avg_mobility <- county_mobility %>%
  group_by(type) %>%
  summarize(mean_mobility = mean(absolute_upward_mobility, na.rm=TRUE))

avg_mobility
```

```
## # A tibble: 3 x 2
##   type  mean_mobility
##   <fct>         <dbl>
## 1 metro          42.1
## 2 micro          43.2
## 3 rural          45.1
```

Regression will give us the same exact information. We'll save the results from our models in an object. We'll use the `lm()` function for our linear models. The basic syntax is `y ~ x, data =`.

Regress mobility on county type.

```
model1 <- lm(absolute_upward_mobility ~ type,
             data = county_mobility,
```

```
            na.action = na.exclude) # This is necessary to get predictions

summary(model1)
```

```
##
## Call:
## lm(formula = absolute_upward_mobility ~ type, data = county_mobility,
##     na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3726  -3.5403  -0.3403   3.1597  20.4169
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.1403     0.1572 268.153  < 2e-16 ***
## typemicro     1.0428     0.2635   3.957 7.78e-05 ***
## typerural     2.9322     0.2293  12.787  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.297 on 2766 degrees of freedom
##   (372 observations deleted due to missingness)
## Multiple R-squared:  0.05644,    Adjusted R-squared:  0.05576
## F-statistic: 82.72 on 2 and 2766 DF,  p-value: < 2.2e-16
```

The intercept estimate is the average mobility in metro areas (the "reference category"). The micro estimate is the amount by which average mobility is higher in micro areas than metro areas. The rural estimate is the amount by which average mobility is higher in rural areas than metro areas. The stars on the far right of the table tell us that these differences are statistically significant.

Are there still differences in average mobility across these types if we *control for* the proportion of children growing up in each commuting zone who end up leaving as adults (the `leave_cz` variable)?

We can control for additional variables by adding them to our model.

```
model2 <- lm(absolute_upward_mobility ~ type + leave_cz,
             data = county_mobility,
             na.action = na.exclude) # This is necessary to get predictions
```

Review this model and interpret it.

REPLACE THIS LINE WITH YOUR CODE

```
summary(model2)
```

```
##
## Call:
## lm(formula = absolute_upward_mobility ~ type + leave_cz, data = county_mobility,
##     na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.9272  -3.1715  -0.0462   2.9953  17.8361
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.8244     0.3711  91.144  < 2e-16 ***
## typemicro    -0.9695     0.2526  -3.837 0.000127 ***
## typerural     0.2668     0.2358   1.132 0.257848
## leave_cz     23.8922     0.9847  24.263  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.797 on 2761 degrees of freedom
##   (376 observations deleted due to missingness)
## Multiple R-squared:  0.2235, Adjusted R-squared:  0.2227
## F-statistic: 264.9 on 3 and 2761 DF,  p-value: < 2.2e-16
```

Finally, let's use an interaction model to test whether the association between `absolute_upward_mobility` and `leave_cz` varies by `type`. This model tests if the *slopes* are different for each type of county. To add an interaction, replace the plus sign in the previous model with an asterisk (since we are taking the product of each county type and its value for `leave_cz`).

```
model3 <- lm(absolute_upward_mobility ~ type * leave_cz,
             data = county_mobility,
             na.action = na.exclude)
```

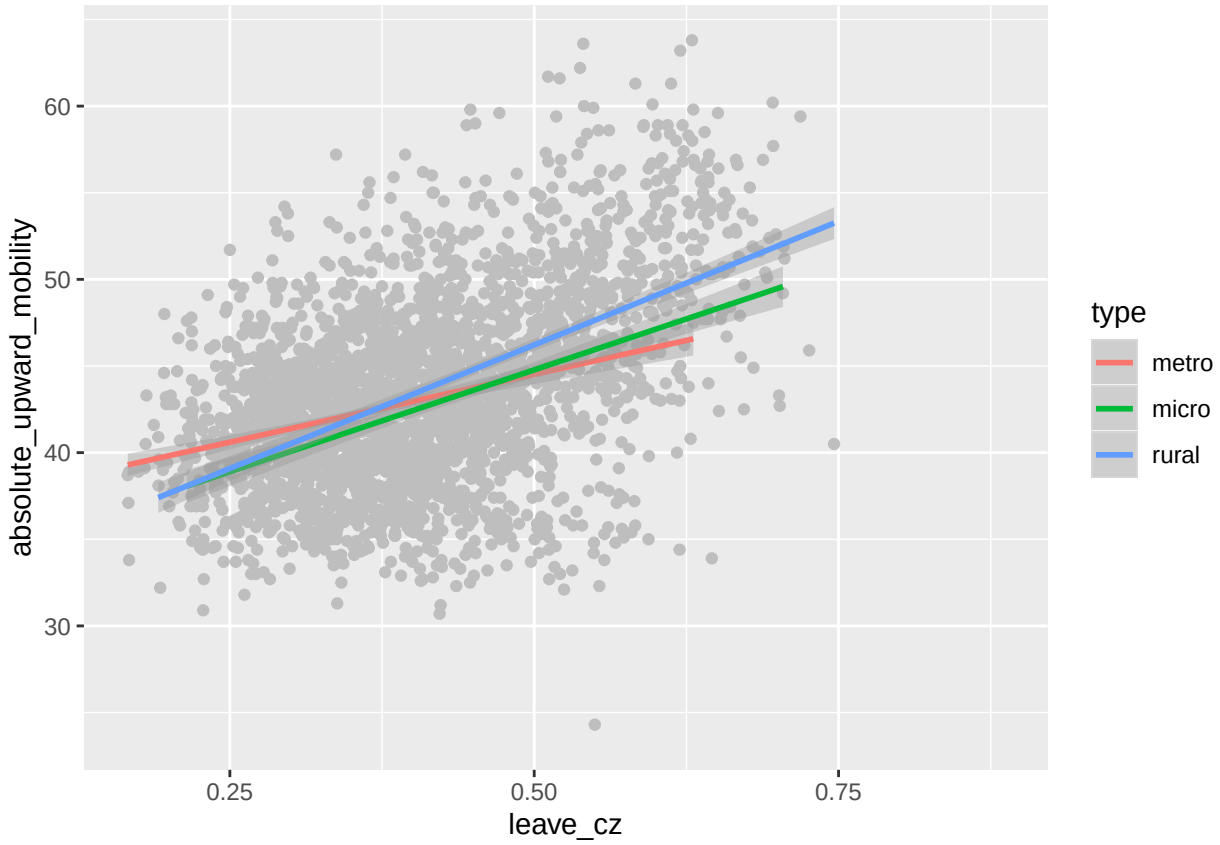
```
summary(model3)
```

```
##
## Call:
## lm(formula = absolute_upward_mobility ~ type * leave_cz, data = county_mobility,
##     na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.3393  -3.0780  -0.0301   3.0011  17.8756
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          36.6888     0.6598  55.602  < 2e-16 ***
## typemicro            -3.6758     1.1555  -3.181  0.00148 **
## typerural            -4.7375     0.9252  -5.120 3.26e-07 ***
## leave_cz             15.6626     1.8517   8.459  < 2e-16 ***
## typemicro:leave_cz    7.8638     2.8371   2.772  0.00561 **
## typerural:leave_cz   12.8713     2.3023   5.591 2.48e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.772 on 2759 degrees of freedom
##   (376 observations deleted due to missingness)
## Multiple R-squared:  0.2322, Adjusted R-squared:  0.2308
## F-statistic: 166.9 on 5 and 2759 DF,  p-value: < 2.2e-16
```

Let's visualize these interactions. GGplot will default to an interaction model if you add a `color =` option to your aesthetic map. But mute the colors of your points to be able to see the different lines.

```
county_mobility %>%
  ggplot(aes(x = leave_cz, y = absolute_upward_mobility,
             color = type)) +
```

```
  geom_point(color = "grey") +
  geom_smooth(method = "lm")
```



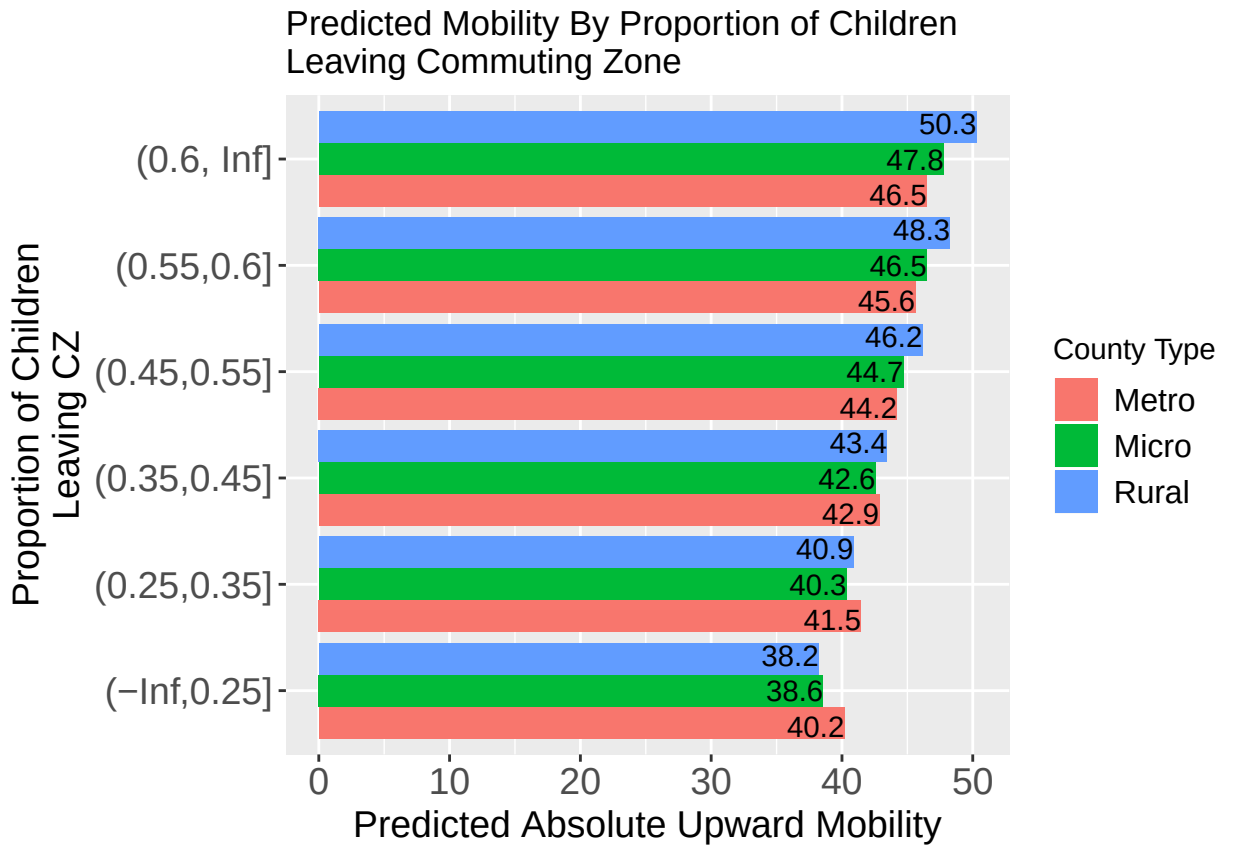
DELETE Predictions

```
county_mobility <- county_mobility %>%
  mutate(leave_cat = cut(leave_cz,
                         breaks = c(-Inf, .25, .35, .45, .55, .60, Inf)))

predictions <- county_mobility %>%
  mutate(pr_mobility = fitted(model3)) %>%
  filter(!is.na(leave_cat)) %>%
  group_by(type, leave_cat) %>%
  summarize(mean_pr_mobility = mean(pr_mobility, na.rm=TRUE))

predictions %>%
  ggplot(aes(x = leave_cat, y = mean_pr_mobility,
             fill = type, label = round(mean_pr_mobility,1))) +
  geom_col(position = "dodge") +
  geom_text(position = position_dodge(width = 1), hjust = 1, size = 4) +
  coord_flip() +
  theme(axis.text = element_text(size = 14),
        axis.title = element_text(size = 14),
        legend.text = element_text(size = 12)) +
```

```
  labs(y = "Predicted Absolute Upward Mobility",
       x = "Proportion of Children\nLeaving CZ",
       title = "Predicted Mobility By Proportion of Children\nLeaving Commuting Zone") +
  scale_fill_discrete(name = "County Type",
                      labels = c("Metro", "Micro", "Rural"))
```



Extra: Getting Regression Output Out Of R

Several packages offer ways to present regression output in publication-ready formats. We'll use `stargazer`. Install it and load it.

```
#install.packages("stargazer")
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

There are many ways to customize stargazer output. Here are some of the most useful. Perhaps most importantly, note that the opening line of the code chunk must include `results = 'asis'` for the markdown output to be formatted in a way to knit correctly.

Table 1: Absolute Upward Mobility

	OLS Models		
	1	2	3
Micropolitan County	1.043^{***}	$-.970^{***}$	-3.676^{**}
	(.264)	(.253)	(1.156)
Rural County	2.932^{***}	.267	-4.737^{***}
	(.229)	(.236)	(.925)
Proportion Leaving CZ		23.892^{***}	15.663^{***}
		(.985)	(1.852)
Micropolitan X Proportion Leaving			7.864^{**}
			(2.837)
Rural X Proportion Leaving			12.871^{***}
			(2.302)
Constant	42.140^{***}	33.824^{***}	36.689^{***}
	(.157)	(.371)	(.660)
Observations	2,769	2,765	2,765
R^2	.056	.224	.232
Notes:	$^{*}P < .05$		
	$^{**}P < .01$		
	$^{***}P < .001$		