POL 345: Quantitative Analysis and Politics

Precept Handout 3

Week 4 (Verzani Chapter 3: 3.1-3.4)

Remember to complete the entire handout and submit the precept questions to the Blackboard DropBox 24 hours before precept. In this handout, we cover the following new materials:

- Graphing multivariate data with `plot()` and `boxplot()`
- Plotting with character variables using `pch`
- Calculating correlation using `cor()`
- Fitting linear regression models using `lm()`
- Summarizing linear regression models using `summary()`
- Obtaining model coefficients using `coef()`
- Adding best fit lines to scatter plots using `abline()`
- Obtaining residuals using `resid()`
- Obtaining fitted values using `fitted()`
- Creating residual plots using `plot()`

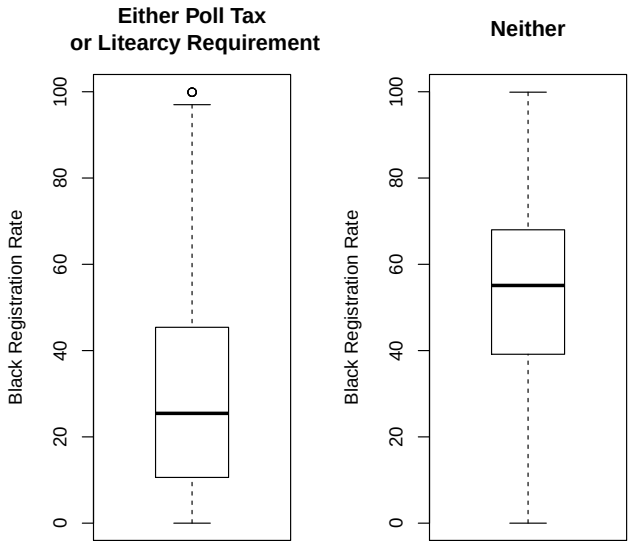
1 More Graphs

In addition to histograms, **R** can make various graphs. Here, we cover `plot()` (generates scatterplots and trend plots) and `boxplot()` (generates boxplots with median and IQR). Note that the commands we learned in the previous week such as `main` and `xlab` will be applicable to these functions. Similarly, the functions such as `points()` and `lines()` can be used to add additional features to these graphs.

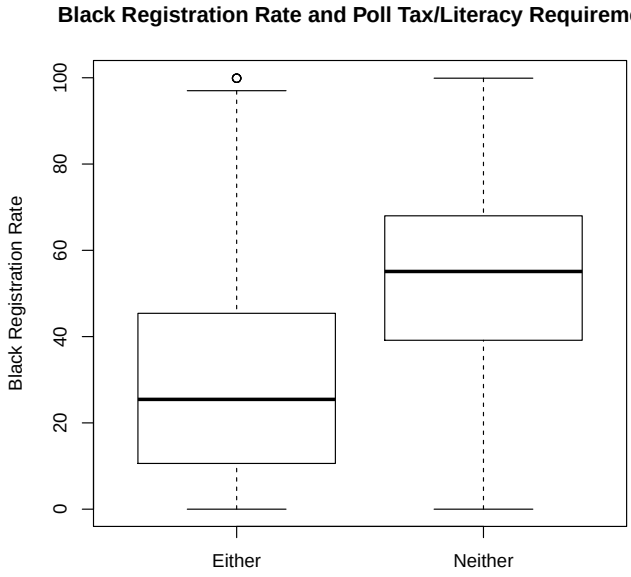
1.1 Boxplots

The function `boxplot()` will produce a **boxplot** figure, which was covered in the lecture.

```
> reg <- read.table("registration.txt", header=TRUE)
> par(mfrow = c(1, 2))
> ## the same information used for both graphs
> ylab <- "Black Registration Rate"
> ylim <- c(0, 100)
> ## \n for a change of line
> boxplot(reg$pBlackReg[reg$polltax == 1 | reg$litreq == 1], ylim = ylim,
+         ylab = ylab, main = "Either Poll Tax\n or Litearcy Requirement")
> boxplot(reg$pBlackReg[reg$polltax == 0 & reg$litreq == 0],
+         ylim = ylim, ylab = ylab, main = "Neither")
```



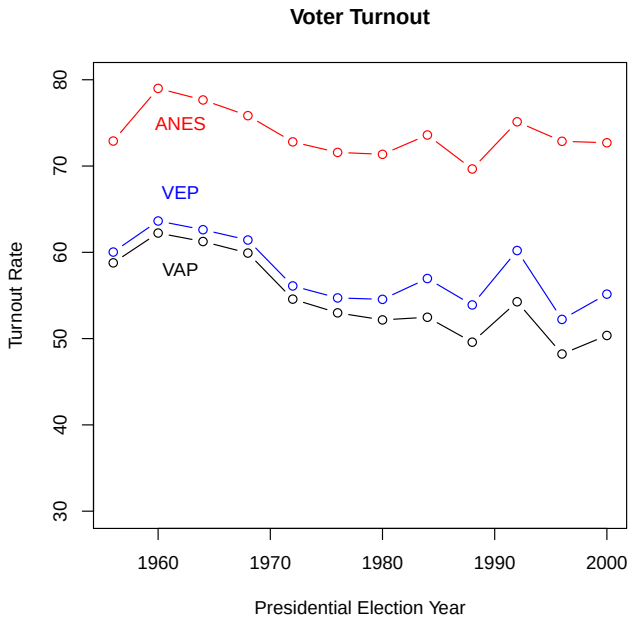

```
> ## two boxplots can be combined
> boxplot(reg$pBlackReg[reg$polltax == 1 | reg$litreq == 1],
+         reg$pBlackReg[reg$polltax == 0 & reg$litreq == 0], ylim = ylim,
+         ylab = ylab, names = c("Either", "Neither"),
+         main = "Black Registration Rate and Poll Tax/Literacy Requirement")
```



1.2 Trend Plots

The function `plot(X, Y)` will produce the trend plot where `X` is a vector for time such as years and `Y` is a corresponding vector of numeric values. To display data as lines (points) use `type = "l"` (`type = "p"`). Setting `type = "b"` will show both lines and points. See `?plot` for more options.

```
> ## ANES data from Precept 1
> ANES <- read.table("ANES.txt", header = TRUE)
> ## Generate VEP and VAP turnout rates as new variables
> ANES$VAPtr <- (ANES$total - ANES$overseas) / ANES$VAP * 100
> ANES$VEPtr <- (ANES$total - ANES$overseas) /
+   (ANES$VAP - ANES$felons - ANES$noncit) * 100
> ## Extract presidential elections as done in Precept 1
> n.obs <- dim(ANES)[1]
> pres <- ANES[seq(from = 1, to = n.obs, by = 2), ]
> ## plotting
> plot(pres$year, pres$VAPtr, type = "b", ylim = c(30, 80), xlim = c(1956, 2000),
+      main = "Voter Turnout", xlab = "Presidential Election Year",
+      ylab = "Turnout Rate")
> lines(pres$year, pres$VEPtr, type = "b", col = "blue") # Add VEP line
> lines(pres$year, pres$ANES, type = "b", col = "red") # Add ANES line
> text(1962, 75, "ANES", col = "red")
> text(1962, 67, "VEP", col = "blue")
> text(1962, 58, "VAP", col = "black")
```



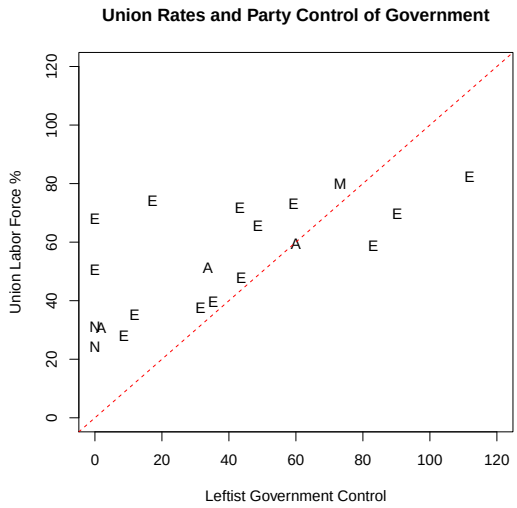
1.3 Scatter Plots

Here, we learn scatter plots, a simple graphical way to display two variables at one time using `plot(x, y, ...)`. As an example, we use the data from an exchange between Wallterstein and Stephens (*American Political Science Review*, 1991) about the relationship between union membership and government structure (the extent to which leftist parties controlled the government). The data set is available as `union.csv` on BlackBoard. The variables are as follows:

Variable	Description
country	Country
region	Country region:
	Asia and Australia, Europe, Middle East, North America
union	Percentage of workers who belong to a union
left	Extent to which parties of the left have controlled government
size	Size of the labor force
concentration	Measure of economic concentration in top four industries

The input `pch` for `plot()` can take a chracter vector, in which case the first letter of each element of the vector will be plotted.

```
> union <- read.csv("union.csv", header=TRUE) # load union data
> union$region <- as.character(union$region) # coerce to character for plotting
> plot(union$left, union$union, pch = union$region, xlim = c(0, 120), ylim = c(0, 120),
+      xlab = "Leftist Government Control", ylab = "Union Labor Force %",
+      main = "Union Rates and Party Control of Government")
> abline(0, 1, lty = 2, col = "red") # add 45 degree line
```



2 Correlation

The upward sloping data cloud in the above scatter plot shows a positive correlation between union membership and leftist government control. To compute the correlation coefficient, we can use the function `cor(X, Y)`, which takes in two numeric vectors, `X` and `Y`, and returns their sample correlation.

```
> cor(union$union, union$left)

[1] 0.6779826
```

3 Linear Regression

- The function `lm(Y ~ X, data = Z)` regresses a variable `Y` on a variable `X` taken from the data frame `Z`. Below, we regress union on the extent to which leftist parties have controlled the government. Recall that the slope coefficient represents the average change in `Y` given a one unit change in `X`.

```
> fit.1 <- lm(union ~ left, data = union)
> fit.1

Call:
lm(formula = union ~ left, data = union)

Coefficients:
(Intercept)         left
    39.8841       0.3764
```

- Applying `summary()` to the regression output will produce a detailed summary. In the lectures, we only covered `Residuals`, `Estimate` and `Multiple R-squared`. Don't worry about the others for now but make sure you know how to interpret them.

```
> fit.1 # short summary

Call:
lm(formula = union ~ left, data = union)

Coefficients:
(Intercept)         left
    39.8841       0.3764

> summary(fit.1) # long summary

Call:
lm(formula = union ~ left, data = union)

Residuals:
    Min      1Q  Median      3Q     Max
-15.384 -10.269  -3.558  10.808  28.216

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.88406    4.81269   8.287 1.48e-07 ***
left         0.37639    0.09619   3.913  0.00102 **
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

Residual standard error: 14.16 on 18 degrees of freedom
Multiple R-squared: 0.4597,        Adjusted R-squared: 0.4296
F-statistic: 15.31 on 1 and 18 DF,  p-value: 0.001019
```

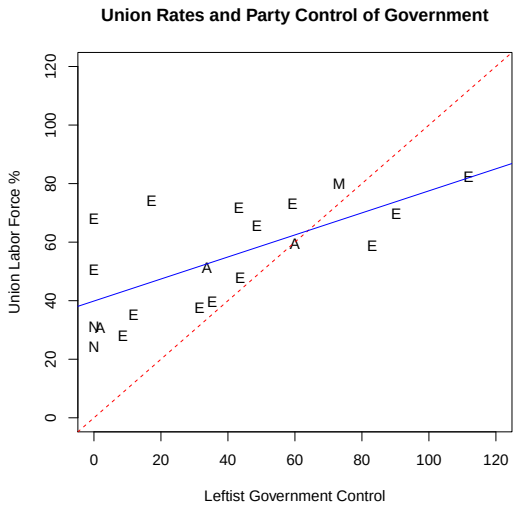
- Applying the function `coef()` to your fitted regression model will return both the intercept and the slope coefficient estimates as a numerical vector.

```
> coef(fit.1)

(Intercept)        left
 39.8840609   0.3763868
```

- You can easily add the fitted regression line to the scatter plot above using `abline()`.

```
> abline(0, 1, lty = 2, col = "red") # add 45 degree line
> abline(fit.1, col = "blue")  # add regression line
```



- The function `resid()` yields the residuals from a linear regression, while the function `fitted()` yields the fitted values from a linear regression. Look for patterns of systematic errors and heteroskedasticity.

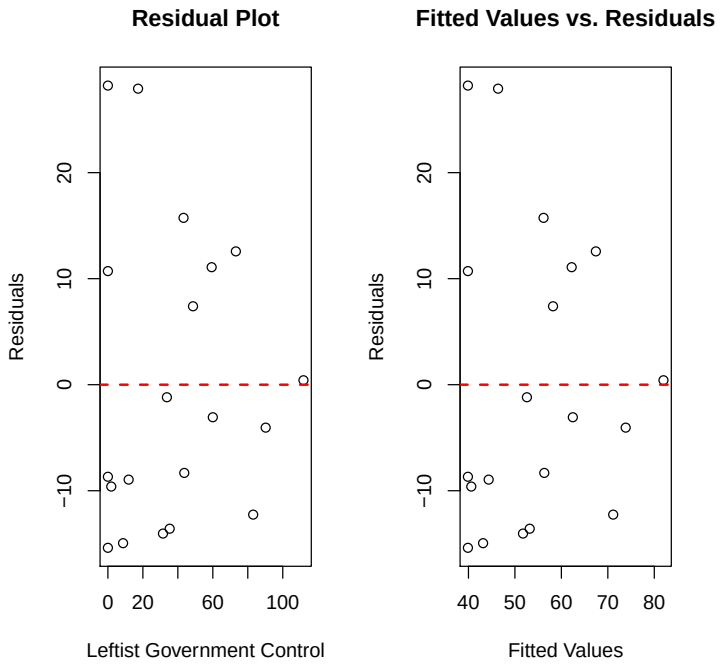
```
> resid(fit.1) # residuals for all observations

         1          2          3          4          5          6          7          8
 0.420834  12.575713  27.923266  11.084907  15.737208  -4.049210  28.215939   7.397191
        11         12         13         14         15         16         17         18
-1.183353  10.715939  -8.320875 -13.581808 -14.040247  -8.951773  -8.684061  -9.606724

> fitted(fit.1) # fitted values for all observations

       1        2        3        4        5        6        7        8        9        1
81.97917 67.42429 46.37673 62.21509 56.16279 73.84921 39.88406 58.20281 62.46727 71.1542
      14       15       16       17       18       19       20
53.18181 51.74025 44.35177 39.88406 40.60672 43.14733 39.88406

> par(mfcol = c(1, 2))
> plot(union$left, resid(fit.1), xlab = "Leftist Government Control",
+      ylab = "Residuals", main = "Residual Plot")
> abline(h = 0, col = "red", lty = 2, lwd = 2) # Adds a zero line
> plot(fitted(fit.1), resid(fit.1), xlab = "Fitted Values", ylab = "Residuals",
+      main = "Fitted Values vs. Residuals")
> abline(h = 0, col = "red", lty = 2, lwd = 2)
```



4 Precept Questions

In preparation for precept, please answer the following questions and submit your code following the instruction given in the syllabus.

1. Download the data file `florida.txt` from Blackboard and load it into R. This data set contains the 1996 and 2000 Presidential election results for Florida counties. Calculate the correlation between Perot's and Buchanan's votes with all observations. Repeat this calculation without Palm Beach.

2. Generate a scatterplot of Perot's 1996 votes (`perot96` on the x-axis) against Buchanan's 2000 votes (`buch00` on the y-axis). Regress Buchanan's 2000 votes on Perot's 1996 votes. Report the estimated coefficients. Add two regression lines to the scatterplot. For the first, add a solid line based on the regression of Buchanan's 2000 votes on Perot's 1996 votes. Next, add a dashed line based on this regression excluding Palm Beach county (for this observation, the `county` variable is `PalmBeach`). Submit your graph as a pdf file to Blackboard.