Problem Set / Data Exercise Example

Devin Judge-Lord

February 5, 2019

Imagine that you are provided a sample of data and asked to estimate the linear regression model $y_i = \alpha + \beta x_i + \epsilon_i$ (or, in equivilant notation, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$).

Let us say that these data contain 20 observations for two variables:

`Leg_Act` $\in \{-20, 40\}$ is the legislative activity of state assembly members, where -20 is no significant legislative activity and 40 is the maximum level of activity. This is the dependent variable, Y, with each observation being a y_i.

`terms` is the number of terms in office. This is your explanatory variable, X, with each observation being a x_i.

You have a number of tasks:

1. Plot the dependent variable against the explanatory variable.
2. Estimate the parameters α and β.
3. Compute the residuals (the difference between the observed values of the dependent variable and the predicted values from the estimated linear model (i.e. the distance of each observed x_i from the regression line).
4. Plot the residuals against the explanatory variable.
5. Correlate the observed values of the dependent variable Y (the vector of each y_i) with the predicted values $\hat{Y}$
6. Compare the square of this correlation (between the observed values of Y and predicted $\hat{Y}$) to the model R^2.
7. Test the null hypothesis that $\beta = 0$ against an alternative that $\beta \neq 0$.
8. Write a paragraph (double-spaced) interpreting the parameters and explaining the results of your hypothesis test.

But, for whatever reason, you want to do your problem set in R. R Markdown offers an easy way to do this without cutting and pasting. If you accidentally regressed X on Y rather than Y on X, fix the model and **pow**, your plots and estimates cited in your discussion are instantly corrected.

- Here is the RMarkdown template that made this pdf. Save it as a .Rmd file.
- Here is a pdf about writing in RMarkdown

But the data are in STATA!?! No problem. R can read .dta files.

In STATA, save the data generated by the `PS813_EX1` function with your seed:

```
net install PS813_EX1, from(https://faculty.polisci.wisc.edu/weimer/)

PS813_EX1 yourseed

save "EX1.dta"
```

Alternativly, run STATA in a chunk (R Markdown supports many languages!). First install Statamarkdown. Then, add a STATA setup chunk (just like our R setup chunk above) that allows STATA chunks: Instructions here.

Then load it into R with the `readstata13` package:

Note: R is looking for "EX1.dta" in a folder called "data" whereever this .Rmd files is saved

```
## Load your data, defining an R object called "d"
d <- readstata13::read.dta13(here("data/EX1.dta"))
glimpse(d)
```

```
## Observations: 20
## Variables: 2
## $ terms   <dbl> 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, ...
## $ Leg_Act <dbl> 2, 2, 9, -4, 7, 20, 9, 0, 11, 12, 0, 11, 16, 5, 9, 5, ...
```

Now on to the tasks:

In STATA, generate data with the `PS813_EX1` function:

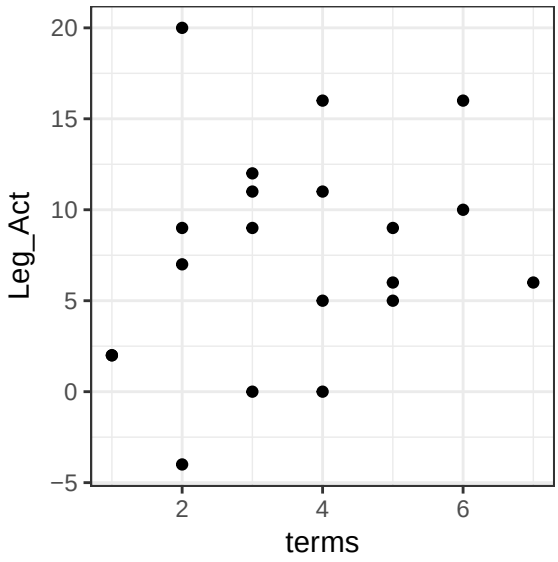
```
net install PS813_EX1, from(https://faculty.polisci.wisc.edu/weimer/)

PS813_EX1 yourseed

save "EX1.dta"
```

1. A plot of Legislative Activity by Terms in Office

```
## STATA: plot Leg_Act terms
## R:
ggplot(d, aes(y = Leg_Act, x = terms)) +
  geom_point()
```




```
## STATA: corr Leg_Act terms
## R:
corXY <- cor(d$Leg_Act, d$terms)
corXY
```

```
## [1] 0.2215867
```

The correlation between Legislative Activity and Terms in Office is 0.2215867

2. Estimating linear regression

```
## STATA: regress Leg_Act terms
## R:
model <- lm(d$Leg_Act ~ d$terms)
# summary(model)
alpha <- model$coefficients[1]
beta <- model$coefficients[2]
```

Regression coefficients: $\alpha = 4.7883212$ and $\beta = 0.7810219$

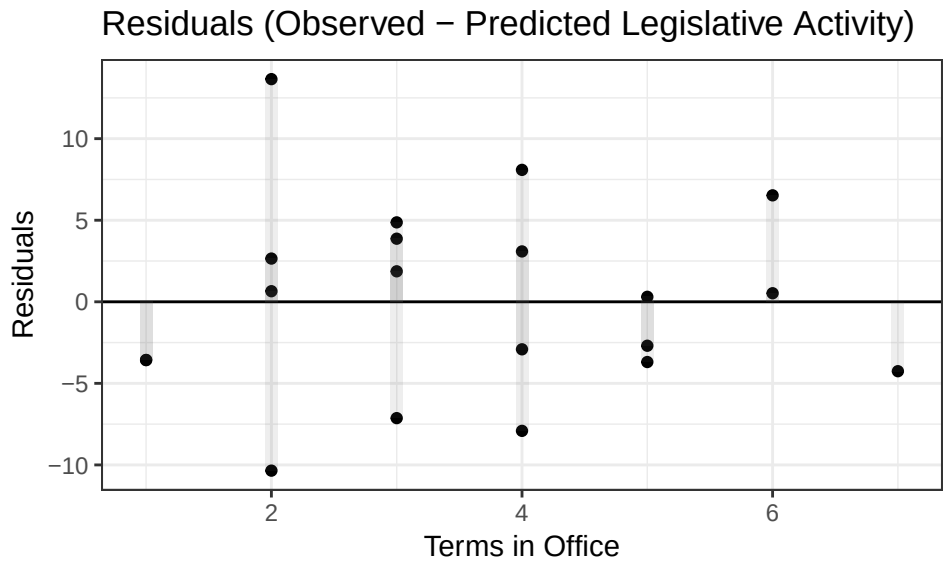
3. Computing residuals

```
## STATA: predict p_Leg_Act
## R:
d$p_Leg_Act <- predict(model)

## STATA: generate resid = Leg_Act - p_Leg_Act
## R:
d$resid <- d$Leg_Act - d$p_Leg_Act
```

4. Plot of Residuals

```
## STATA:  plot resid terms
## R:
ggplot(d) +
  aes(y = resid, x = terms) + # "aesthetics"
  geom_point() + # a layer of points
  ## to show how risiduals are the distance between an observation and the regression line:
  geom_hline(yintercept = 0) +
  geom_col(alpha = .1, width = .1, position = "dodge") +
  ## + labels:
  labs(title = "Residuals (Observed - Predicted Legislative Activity)",
       x = "Terms in Office",
       y = "Residuals")
```



5. $Cor(Y, \hat{Y})$

```
## STATA: corr Leg_Act p_Leg_Act
## R:
correlation <- cor(d$Leg_Act, d$p_Leg_Act)
```

$Cor(Y, \hat{Y}) = 0.2215867$

6. $Cor(Y, \hat{Y})^2$ vs. R^2.

```
## STATA: generate r2 =r(rho)*r(rho)
## R:
r2 <- summary(model)$r.squared
```

$R^2 = 0.0491007$

7. Hypothesis test

Lorem ipsum $\beta = 0$

Lorem ipsum $\beta \neq 0$

8. Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.