# Handout 8

*Put your name here!*

In this handout, we will learn more about using the function `lm` and how to use residual and conditional plots.

### Topics and Concepts Covered

- `I()` to add new terms to a linear model
- `data` and `subset` options for `lm`
- `$res` and `$fit` to recover the fitted values and residuals in a linear model
- `coplot` for a *conditional plot*
- `pch` for plotting characters in a figure (1 = open circle, 19 = filled circle)

## Using `I` to Add New Terms to a Linear Model

We have been using `lm` to generate linear models. This function follows the syntax `lm(Y ~ X1 + X2 + X3` where $Y$ is the dependent variable and the independent variables are $X1$, $X2$, and $X3$. Let's say we save a linear model as

```
lm1 <- lm(Y ~ X1 + X2 + X3)
```

If we wanted to add X1$^2$ to the model we need to do the calculation *inside* the `I` function so that the new variable we make on the fly is computed before R thinks about interpreting the rest of the formula. Like this:

```
lm1 <- lm(Y ~ X1 + I(X1^2) + X2 + X3)
```

We could also accomplish through defining a new variable `X1.sq <- X1^2` and then write

```
lm1 <- lm(Y ~ X1 + X1.sq + X2 + X3)
```

Either method is acceptable.

In addition to squaring values, `I` can be used for other operators such as `+`, `-`, \, and `*`.

## Using the `data` and `subset` options for the `lm` command

So far, you have had to use the syntax `data.frame$variable` in order to reference `variable` from the object `data.frame` when running a linear model with the `lm` command. `lm` includes an option, `data`, that allows you to specify what dataset you are using and then you can call the variable directly. The syntax for `lm` is

```
lm(variable1 ~ variable2, data = data.frame)
```

As you have already seen in this course, there are many times when we want to only look at a subset of the data. Within the `lm` command we can subset the data using the `subset` option. The syntax for `subset` is

```
lm(variable1 ~ variable2, data = data.frame, subset = rows)
```

You can also use the `data` and `subset` options with the `plot` command.

## `$res` and `$fit` to Recover Fitted Values and Residuals

If we wanted to look at the residuals, we would type `lm1$res`. To get the fitted values we would type `lm1$fit`.

Alternatively there are functions `fitted` and `resid` that do this extraction for you. Often these are a bit clearer. Then to get the residuals you would type `resid(lm1)` and to get the fitted values you would type `fitted(lm1)`.

Again, either method is acceptable.

To plot residuals against fitted values—a useful diagnostic—we could either

```r
plot(lm1$fit, lm1$res,
     main = "Fitted vs. Residual Plot",
     xlab="Fitted Values", ylab = "Residuals")
```

or

```r
plot(fitted(lm1), resid(lm1),
     main = "Fitted vs. Residual Plot",
     xlab="Fitted Values", ylab = "Residuals")
```

R knows that this is a useful plot, so you can get it directly using `plot` on the linear model object like this:

```r
plot(lm1, which = 1) # which = 1 chooses the residuals vs fitted plot
```

In a classic paper Mankiw, Romer, and Weil (1992) investigate the determinants of long-term economic growth. In particular, they empirically explore the Solow growth model using cross-country data. The paper uses as its starting point the Solow growth model, which predicts both the steady-state GDP per capital of a country as a function of its savings rate and population growth, in addition to predicting the rate at which countries 'converge' to this steady-state. The paper demonstrates empirically that the Solow growth models the data well when both human capital and physical capital are considered.

In the `data` folder you will find the file `international_growth_data.csv`, which contains the data from the Mankiw, et al. study. First, we are going to read it in, and take a look at what variables it contains

```r
Data.Growth <- read.csv("data/international_growth_data.csv")
names(Data.Growth)
```

```
 [1] "number"                         "country"
 [3] "nonoil"                         "intermediate"
 [5] "oecd"                           "gdp_per_adult_1960"
 [7] "gdp_per_adult_1985"             "gdp_growth_1960to1985"
 [9] "working_age_pop_growth_1960to198" "investment_per_gdp"
[11] "schooling"
```

The Solow growth model makes some very specific predictions about the relationships among these variables. Specifically,

$$\begin{aligned}\log[\text{gdpperadult1985}_i] =& \alpha + \beta_k \log[\text{investmentpergdp}_i/100] \\ &+ \beta_P \log[\text{workingagepopgrowth}_i/100 + 0.05] + \epsilon_i\end{aligned}$$

The model predicts $\beta_k = \frac{1}{2}$, $\beta_p = -\frac{1}{2}$. Unfortunately, our data is defined in terms of the raw values, not logs. Therefore, in this example, we will need to log these variables. We can do this directly within the `lm` command.

```r
lm1 <- lm(log(gdp_per_adult_1985) ~ log(investment_per_gdp / 100) +
          log(working_age_pop_growth_1960to198 / 100 + .05), data = Data.Growth)

summary(lm1)
```

```

Call:
lm(formula = log(gdp_per_adult_1985) ~ log(investment_per_gdp/100) +
    log(working_age_pop_growth_1960to198/100 + 0.05), data = Data.Growth)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.99973 -0.49380 -0.04073  0.55369  3.02410 

Coefficients:
                                                 Estimate Std. Error
(Intercept)                                        9.6293     1.6496
log(investment_per_gdp/100)                        1.4780     0.1631
log(working_age_pop_growth_1960to198/100 + 0.05)  -0.4573     0.5837
                                                 t value Pr(>|t|)    
(Intercept)                                        5.837 6.08e-08 ***
log(investment_per_gdp/100)                        9.063 8.32e-15 ***
log(working_age_pop_growth_1960to198/100 + 0.05)  -0.784    0.435    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7938 on 104 degrees of freedom
  (14 observations deleted due to missingness)
Multiple R-squared:  0.4767,    Adjusted R-squared:  0.4667 
F-statistic: 47.37 on 2 and 104 DF,  p-value: 2.365e-15
```

These names are too long, so let's clean them up through creating new variables, instead of using `log` within the `lm` command:

```
log.dv <- log(Data.Growth$gdp_per_adult_1985)
log.invest <- log(Data.Growth$investment_per_gdp / 100)
log.pop <- log(Data.Growth$working_age_pop_growth_1960to198 / 100 + .05)
lm1 <- lm(log.dv ~ log.invest + log.pop, data = Data.Growth)
```

From looking at the output, the second coefficient is approximately $-0.5$, as expected. The first coefficient is not very close. In fact, we can calculate how far away it is:

```
summary(lm1)$coef
```

```
              Estimate Std. Error    t value     Pr(>|t|)
(Intercept)  9.6292513  1.6495650  5.8374487 6.082868e-08
log.invest   1.4779957  0.1630853  9.0627170 8.316214e-15
log.pop     -0.4573491  0.5836656 -0.7835807 4.350671e-01
```

```
beta <- summary(lm1)$coef[2,1]
beta
```

```
[1] 1.477996
```

```
sigma <- summary(lm1)$coef[2,2]
sigma
```

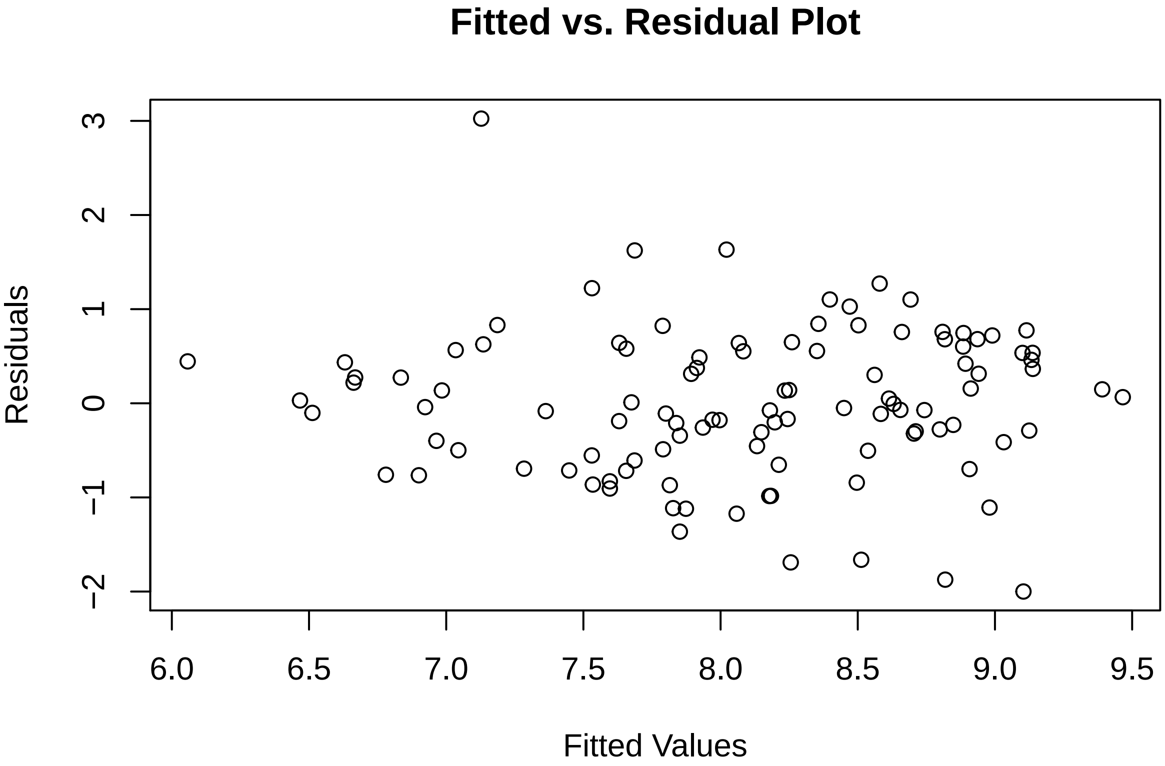
```
[1] 0.1630853
```

```
(beta - 0.5) / sigma
```

```
[1] 5.996836
```

The estimate coefficient is about 6 standard deviations from the value predicted by Solow's theory—a discrepancy not likely to occur due to chance (i.e., a 'significant difference').

```
plot(lm1$fit, lm1$res,
     main = "Fitted vs. Residual Plot",
     xlab = "Fitted Values", ylab = "Residuals")
```

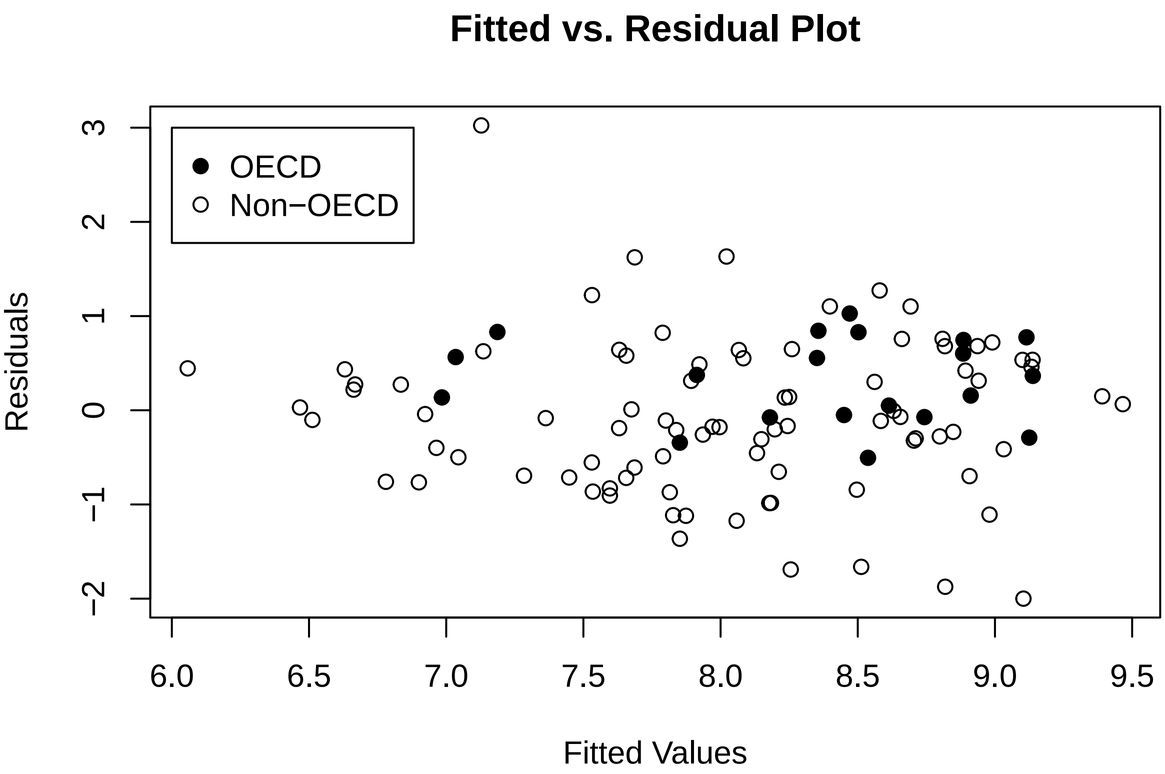


Simply by looking at the figure, there appears nothing wrong—no signs of 'fanning out' to indicate heteroskedasticity, and no curvature that would leave you worried about missing quadratic terms.

## Using `pch` to Vary Plotted Points on a Figure

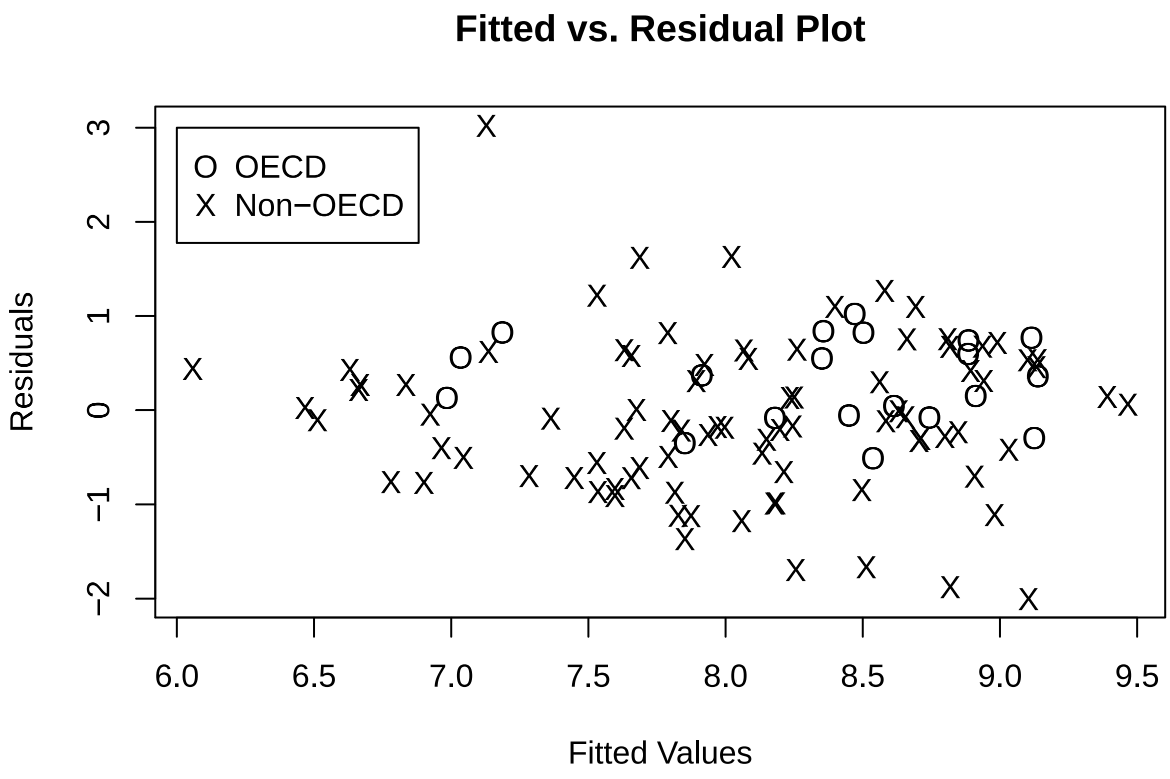
Now, Professor Solow is no slouch; neither is Professor Mankiw. In this section, we are going to examine the residual plots in order to identify an additional variable, `schooling`, that needs to be included in order to recover Solow's results. First, we are going to explore whether being in the OECD is a missing variable. If the OECD countries have residuals systematically above or below the residuals of non-OECD countries, we would argue that we need to add OECD to the regression. We can assess this in the fitted-residual plot, by plotting OECD countries with a solid dot and non-OECD countries with an open dot:

```
pch.oecd <- ifelse(Data.Growth$oecd, 19, 1)
plot(lm1$fit, lm1$res,
     main = "Fitted vs. Residual Plot",
     xlab = "Fitted Values", ylab = "Residuals",
     pch = pch.oecd)
legend(x = 6, y = 3,
       legend = c("OECD", "Non-OECD"),
       pch = c(19,1))
```



Since the OECD countries seem to have residuals that are indistinguishable from non-OECD countries, we can conclude that OECD is probably not a missing variable. Notice that `plot` takes the option `pch`, which stands for 'plotting character.' If we wanted to use X's and O's instead of solid and open points, we could have simply run

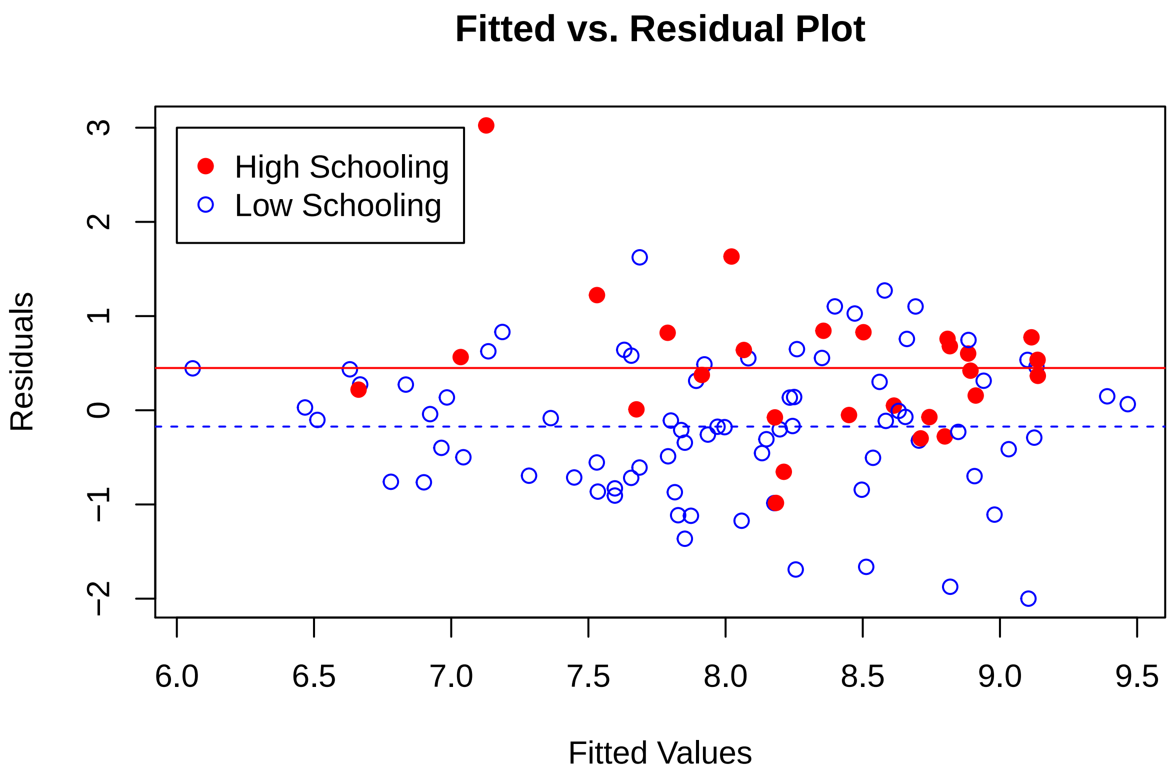
```
pch.oecd <- ifelse(Data.Growth$oecd, "O", "X")
plot(lm1$fit, lm1$res,
     main = "Fitted vs. Residual Plot",
     xlab = "Fitted Values", ylab = "Residuals",
     pch = pch.oecd)
legend(x = 6, y = 3,
       legend = c("OECD", "Non-OECD"),
       pch = c("O","X"))
```



which would return O's plotted for OECD countries and X's plotted for non-OECD countries.

Next, let's see if schooling is an omitted variable. For this figure, we are going to use solid, red dots for countries with high levels of schooling, and open, blue dots for countries with low levels of schooling. Finally, we are going to add lines at the mean residual for countries with high- and low-levels of schooling:

```
pch.schooling <- ifelse(Data.Growth$schooling > 8, 19, 1)
col.schooling <- ifelse(Data.Growth$schooling > 8, "red", "blue")
plot(lm1$fit, lm1$res,
     main = "Fitted vs. Residual Plot",
     xlab = "Fitted Values", ylab = "Residuals",
     pch = pch.schooling, col = col.schooling)
abline(h = mean(lm1$res[Data.Growth$schooling > 8], na.rm = TRUE), col = "red")
abline(h = mean(lm1$res[Data.Growth$schooling <= 8], na.rm = TRUE), lty = 2, col = "blue")
legend(x = 6, y = 3,
       legend = c("High Schooling", "Low Schooling"),
       pch = c(19,1),
       col = c("red","blue"))
```



As suggested by the residual plot, we could add the variable `log.schooling` to our regression:

```
Data.Growth$log.schooling <- log(Data.Growth$schooling)
lm2 <- lm(log.dv ~ log.invest + log.pop + log.schooling, data = Data.Growth)
summary(lm2)
```

```

Call:
lm(formula = log.dv ~ log.invest + log.pop + log.schooling, data = Data.Growth)

Residuals:
     Min       1Q   Median       3Q      Max
-1.39277 -0.39848  0.00573  0.39105  2.06954

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.94529    1.31397   5.286 7.15e-07 ***
log.invest     0.70293    0.15542   4.523 1.65e-05 ***
log.pop       -0.58055    0.45245  -1.283    0.202
log.schooling  0.68567    0.08161   8.401 2.73e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6129 on 102 degrees of freedom
  (15 observations deleted due to missingness)
Multiple R-squared:  0.6897,    Adjusted R-squared:  0.6806
F-statistic: 75.56 on 3 and 102 DF,  p-value: < 2.2e-16
```

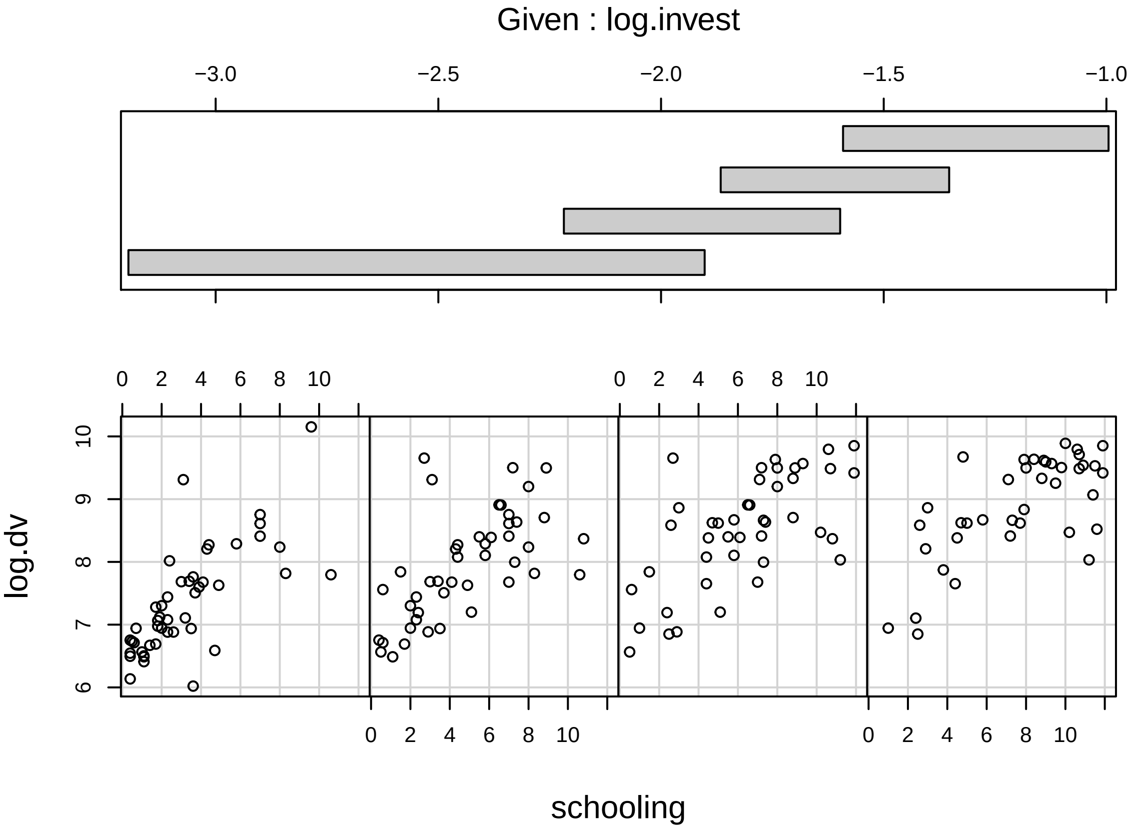
and when we do so, as predicted by theory, we find that the first two coefficients are far closer to their predicted values of $\beta_k = \frac{1}{2}$ an $\beta_p = -\frac{1}{2}$.

## `coplot` for Conditional Plots

Now, we just added the variable `log.schooling` to our linear regression. We are assuming, in this multivariate regression, that schooling has an additive, linear effect on the dependent variable. Of course, a reasonable scholar might argue that the effect of schooling varies systematically with investment and population levels: the returns to schooling might be larger in high-investment and high-population areas. When the effect of one variable varies systematically with another variable, we call this an *interaction effect*, a concept we will discuss in some detail throughout the course. For now, we are simply assessing whether the slope of schooling varies with the values of investment or population. To assess, we can use a *conditional plot*.

A conditional plot produces a figure with several panels. Each panel plots the dependent variable against an independent variable, within the range of a *conditioning variable*. This is perhaps better explained after you see one:

```
coplot(log.dv ~ schooling | log.invest,
       data = Data.Growth,
       rows = 1, columns = 4, number = 4)
```

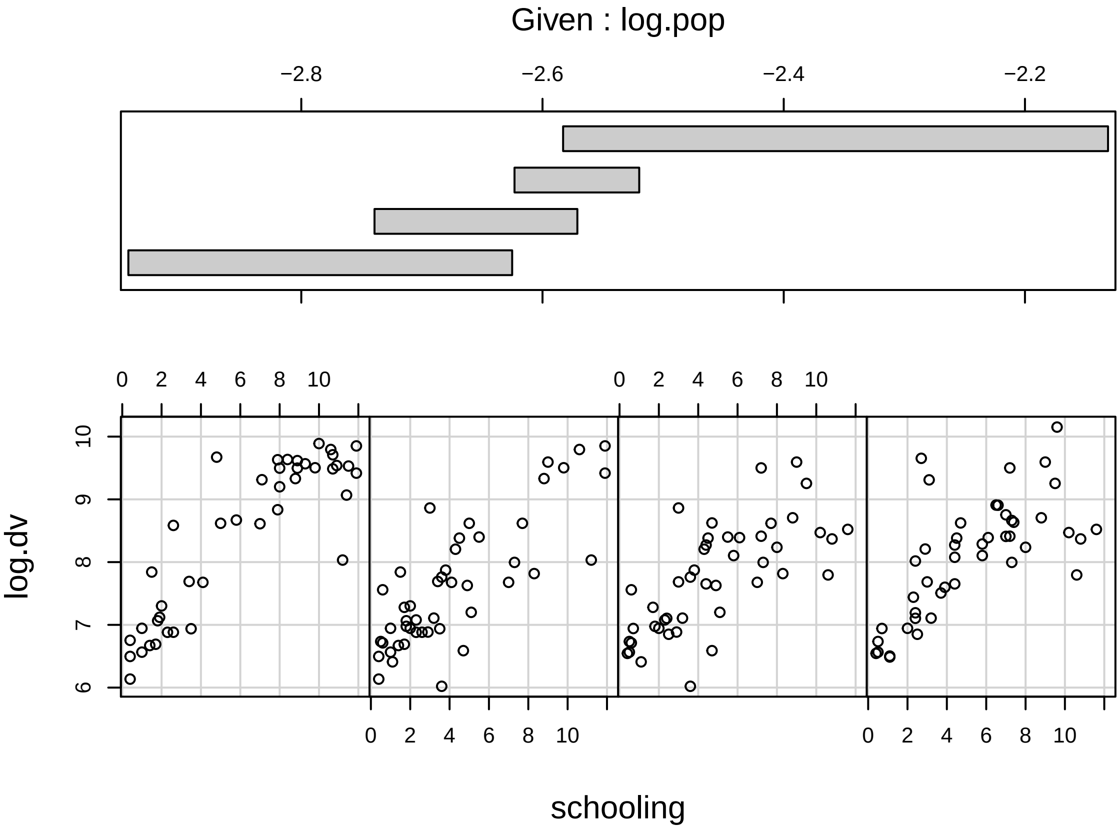


```
 Missing rows: 14, 16, 36, 44, 45, 66, 68, 72, 78, 81, 82, 91, 111, 114, 118
```

This figure shows GDP vs. schooling, for low levels of investment (left) through high levels of investment (right). Notice that the panel up top shows what range of `log.invest` each plot is showing us. The relationship between schooling an GDP seems linear, regardless of the level of investment. The options `rows`, `columns`, and `number` give the number of rows, number of columns, and total number of figures in the conditional plot. I prefer 1 x 3 or 1 x 4 figures, based off of how much data you have. The more data you have, the more subsets you can explore.

Next, let's look at the same figure, except now we are going to use `log.pop` as the conditioning variable:

```
coplot(log.dv ~ schooling | log.pop,
       data = Data.Growth,
       rows = 1, columns = 4, number = 4)
```



```
Missing rows: 14, 16, 36, 44, 45, 66, 68, 72, 78, 81, 82, 91, 111, 114, 118
```

From these figures, it appears that schooling is having an independent, additive effect on GDP.

## Precept Questions

We will revisit Mankiw's data next week. For now, let's consider the Ross data from Handout 2. The file `RossOilWomenIslamData.csv` can be found in the `data` folder.

### Question 1

First, plot female labor participation versus oil rents. Plot the Middle Eastern countries red and non-Middle Eastern countries blue. Plot 'high Muslim' countries, for which `islam > 0` with solid dots, and 'low Muslim' countries with open dots. Make sure to include a title, informative axis labels, and a legend. Please note that the `legend` command takes options `pch` and `col` just like `plot` does.

Which countries are driving the negative correlation between oil and female labor force participation? What does this imply for Ross's claim that oil, not Islam, is leading to a decrease in labor force participation?

### Question 2

Create a conditional plot with 1 row and 4 columns, as above. Plot labor participation versus proportion Muslim in each square, using oil rents as the conditioning variable. Plot the Middle Eastern countries in red, and the remainder in blue (`coplot` takes the option `col`, just as plot does).

Please add `coplot(..., panel = panel.smooth)` in order to add a ‘smooth’ curve to each figure in your conditional plot.

The figure should show a clear interaction between proportion Muslim, oil, and Middle Eastern countries. Where do we see a relationship between proportion Muslim and female labor participation and where not?

## References

Mankiw, N. G., D. Romer, and D. N. Weil. 1992. “A Contribution to the Empirics of Economic Growth.” *The Quarterly Journal of Economics* 107 (2): 407–37. https://doi.org/10.2307/2118477.