Empirical example: return to education **ARE212**: Section 04

The purpose of this section is to estimate the returns to education using `R`. There is nothing valid about the results found in this section; but the empirical application gives us a chance to explore categorical dummies and the `ggplot` package. First, as always, we load the required libraries.

```
library(foreign)
library(ggplot2)
library(xtable)
```

We can then read the wage data directly from the online repository for the supplementary data sets for the Wooldridge (2002) text. You will need an internet connection. We only need the `wage`, `educ`, and `age` variables, and we omit all observations with missing observations.

```
f <- "http://fmwww.bc.edu/ec-p/data/wooldridge/wage2.dta"
data <- read.dta(f)
data <- data[ , c("wage", "educ", "age")]
data <- na.omit(data)
```

A quick visualization reveals the distribution of wages in the data set:

```
hist(data$wage, xlab = "wage", main = "", col = "grey", border = "white")
```

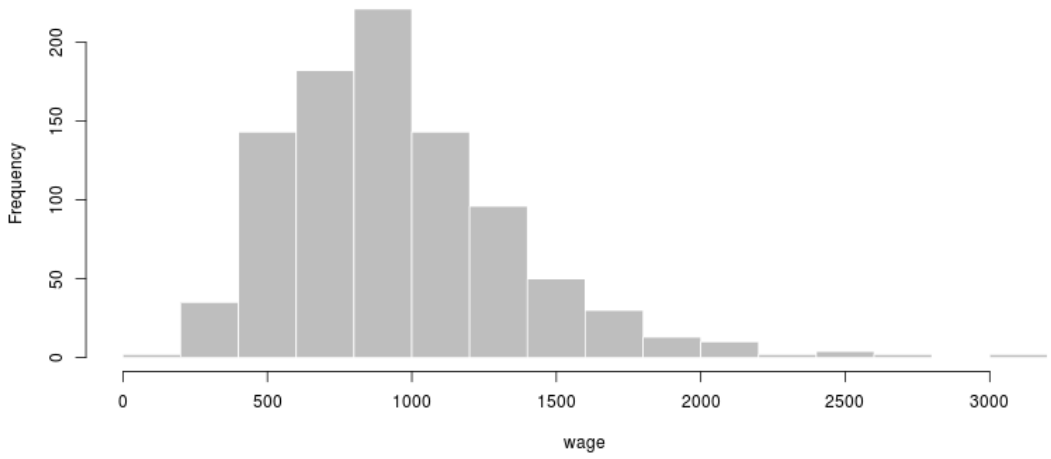

Figure 1: Loess regression

Roughly following page 38 of the lecture notes, we create a rough measure of educational attainment from the `educ` variable.

```
e1 <- ifelse(data$educ %in% 1:12, 1, 0)
e2 <- ifelse(data$educ %in% 13:14, 1, 0)
e3 <- ifelse(data$educ %in% 15:16, 1, 0)
e4 <- ifelse(data$educ %in% 17:18, 1, 0)
```

The categorical education variables sum to one, and the `lm()` function will force-drop one of the variables. Note that the intercept in this regression reflects the mean wage of the `e4` class. The other coefficients reflect the relative wages of the other three classes.

```
(m1 <- lm(wage ~ 1 + e1 + e2 + e3 + e4, data = data))
```

```
Call:
lm(formula = wage ~ 1 + e1 + e2 + e3 + e4, data = data)

Coefficients:
(Intercept)           e1           e2           e3           e4
    1196.96      -350.46      -229.97       -90.51           NA
```

Suppose we want to estimate the premium on education, relative to the least educated class. We then specifiy the regression:

```
xtable(m2 <- lm(wage ~ 1 + e2 + e3 + e4, data = data))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	846.4948	17.4837	48.42	0.0000
e2	120.4929	34.8322	3.46	0.0006
e3	259.9565	32.5528	7.99	0.0000
e4	350.4640	42.6787	8.21	0.0000