36-401 Class Notes and Exercise for 9/19/00

Howard Seltman

Introduction to Multiple Regression (RwG ch. 3)

Measure $K-1$ covariates on $i = 1, \ldots, n$ subjects.

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{K-1} X_{i,K-1} + \epsilon_i \quad \Leftrightarrow \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\text{with } \boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}) \quad \Rightarrow \quad \mathbf{Y} \sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$$

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B} \qquad \mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e}$$

Example: Description: The wages.dat datafile contains 534 observations on 11 variables sampled from the Current Population Survey of 1985.

EDUCATION: Number of years of education.

SOUTH: Southern Region (1=Person lives in South, 0=Person lives elsewhere).

SEX: Indicator variable for sex (1=Female, 0=Male).

EXPERIENCE: Number of years of work experience.

UNION: Union membership (1=Union member, 0=Not union member).

WAGE: Wage (dollars per hour).

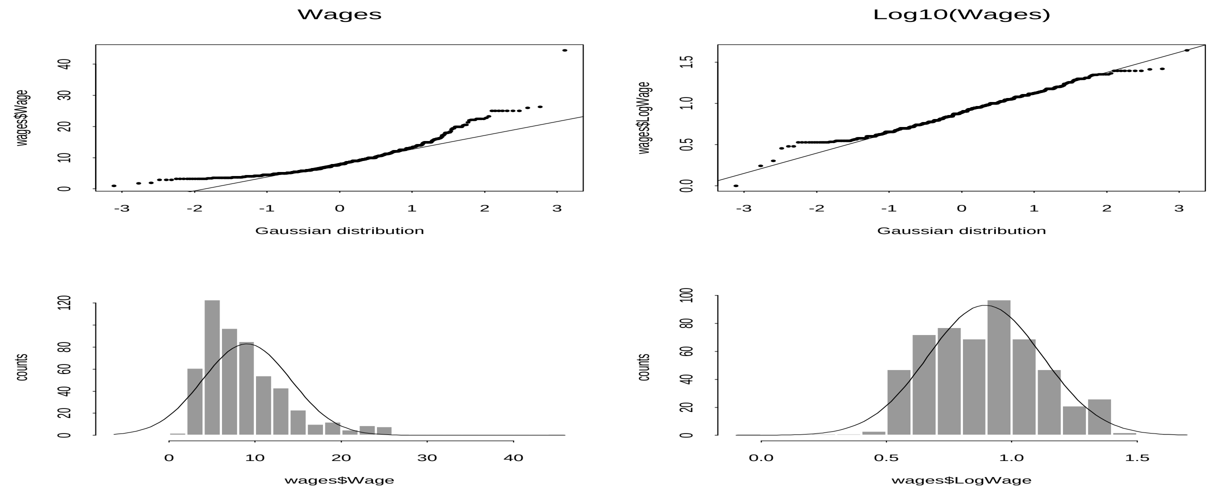
AGE: Age (years).

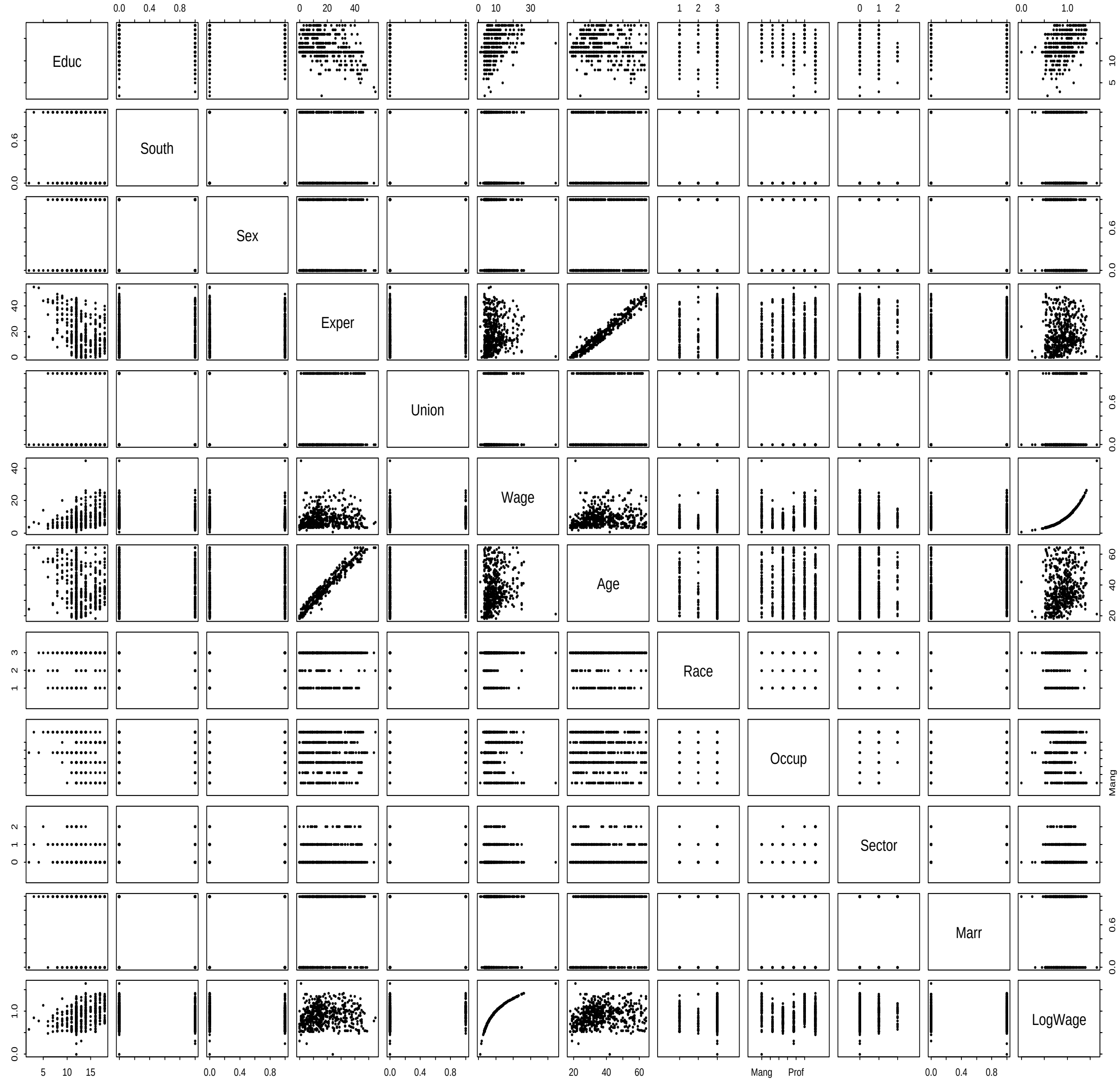
RACE: Race (1=Other, 2=Hispanic, 3=White).

OCCUPATION: 0=Other, 1=Mngmnt, 2=Sales, 3=Clerical, 4=Service, 5=Prfssnl.

SECTOR: Sector (0=Other, 1=Manufacturing, 2=Construction).

MARR: Marital Status (0=Unmarried, 1=Married).





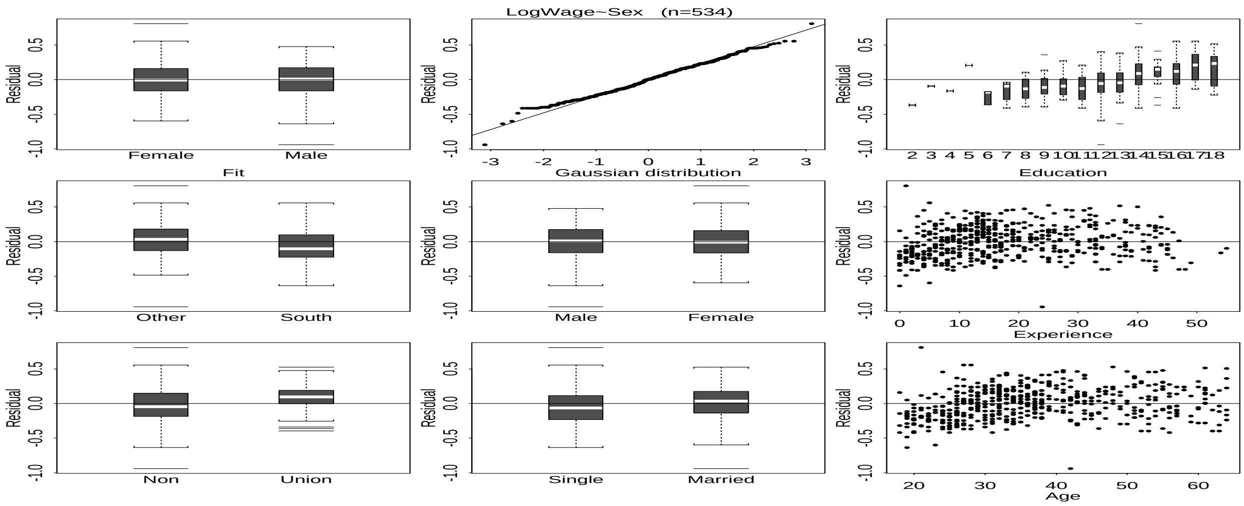

```
> round(cor(wages[,c("Educ","South","Sex","Exper","Union","Age","Marr","LogWage")]),3)
          Educ  South    Sex  Exper  Union    Age   Marr LogWage
   Educ  1.000 -0.140  0.002 -0.353 -0.024 -0.150 -0.036   0.380
  South -0.140  1.000 -0.021 -0.007 -0.086 -0.039  0.007  -0.172
    Sex  0.002 -0.021  1.000  0.075 -0.157  0.079  0.011  -0.219
  Exper -0.353 -0.007  0.075  1.000  0.118  0.978  0.271   0.108
  Union -0.024 -0.086 -0.157  0.118  1.000  0.119  0.093   0.208
    Age -0.150 -0.039  0.079  0.978  0.119  1.000  0.279   0.198
   Marr -0.036  0.007  0.011  0.271  0.093  0.279  1.000   0.136
LogWage  0.380 -0.172 -0.219  0.108  0.208  0.198  0.136   1.000
```

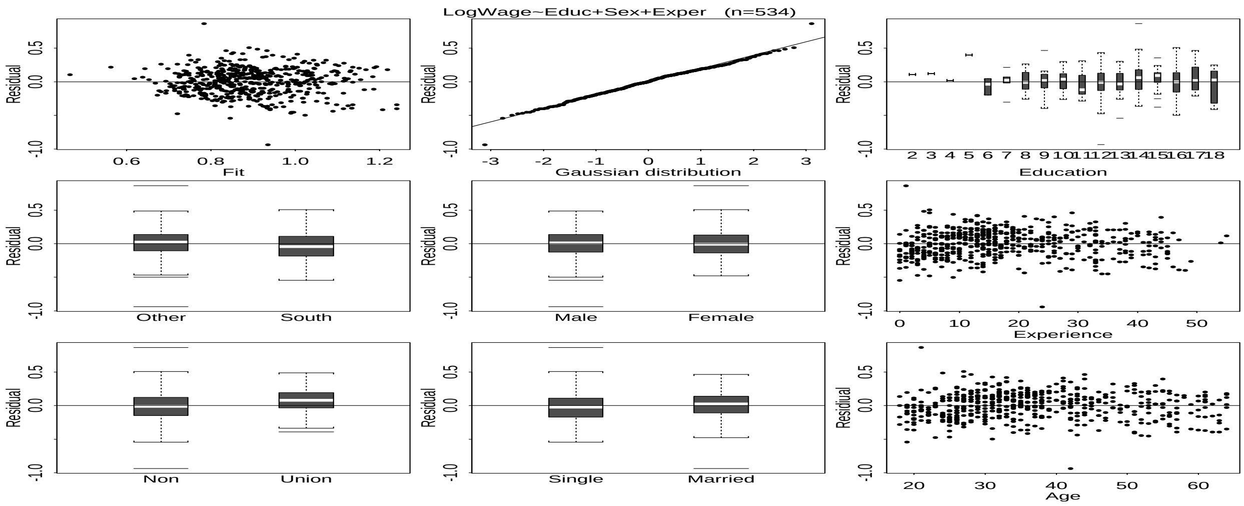
```
> summary(lm(LogWage~Educ+Sex+Exper, wages))
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   0.2931   0.0523     5.6006   0.0000
       Educ   0.0426   0.0035    12.2372   0.0000
        Sex  -0.1112   0.0171    -6.4946   0.0000
      Exper   0.0055   0.0007     7.4661   0.0000

Residual standard error: 0.1964 on 530 degrees of freedom
Multiple R-Squared: 0.2696
F-statistic: 65.22 on 3 and 530 degrees of freedom, the p-value is 0

Correlation of Coefficients:
      (Intercept)    Educ     Sex
 Educ -0.9494
  Sex -0.1032     -0.0306
Exper -0.5449      0.3538 -0.0812
```


```
> wages[c(200,171),]
    Educ South Sex Exper Union Wage Age Race Occup Sector Marr LogWage
200   12     0   0    24     0  1.0  42    3 Manag      0    1 0.00000
171   14     0   1     1     0 44.5  21    3 Manag      0    0 1.64836
> sd(wages$LogWage)
[1] 0.2291955
```


LogWage~Sex (n=534)
Residual
Female
Male
Fit
Gaussian distribution
Education
Other
South
Male
Female
Experience
Non
Union
Single
Married
Age


LogWage~Educ+Sex+Exper (n=534)
Residual
Fit
Gaussian distribution
Education
Other
South
Male
Female
Experience
Non
Union
Single
Married
Age

Breakout Session for 9/19/00

```
> summary(wages[,c("LogWage","Educ","Sex","Exper")])
    LogWage            Educ            Sex              Exper      
 Min.   :0.0000   Min.   : 2.00   Min.   :0.0000   Min.   : 0.00  
 1st Qu.:0.7202   1st Qu.:12.00   1st Qu.:0.0000   1st Qu.: 8.00  
 Median :0.8910   Median :12.00   Median :0.0000   Median :15.00  
 Mean   :0.8943   Mean   :13.02   Mean   :0.4588   Mean   :17.82  
 3rd Qu.:1.0510   3rd Qu.:15.00   3rd Qu.:1.0000   3rd Qu.:26.00  
 Max.   :1.6480   Max.   :18.00   Max.   :1.0000   Max.   :55.00  
```

1. What are the regression equations for LogWage for males and for females? What is the 95% CI for the Sex coefficient; how do you interpret it?

2. What are the (substantive, not statistical) interpretations of the Intercept and the Education slope?

3. Calculate the importance of each of the three covariates in two ways. For Education and Experience, first determine the change in wage (not LogWage) due to a change in that covariate from the median to the third quartile, holding everything else constant. Second calculate the change in wage due to a change in the covariate from median to minimum value. Do the analogous calculation for Sex. Why is this a good way to evaluate importance?

4. If you have time, devise a plot to present the regression results in an informative way.