36-707 Multiple Regression Example

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).

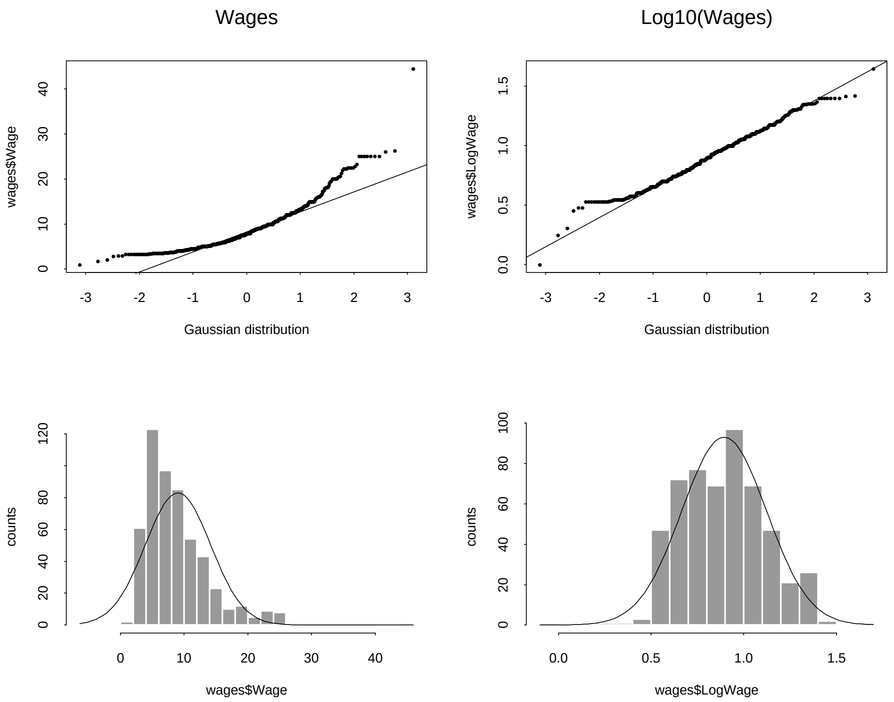
AGE: Age (years).

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

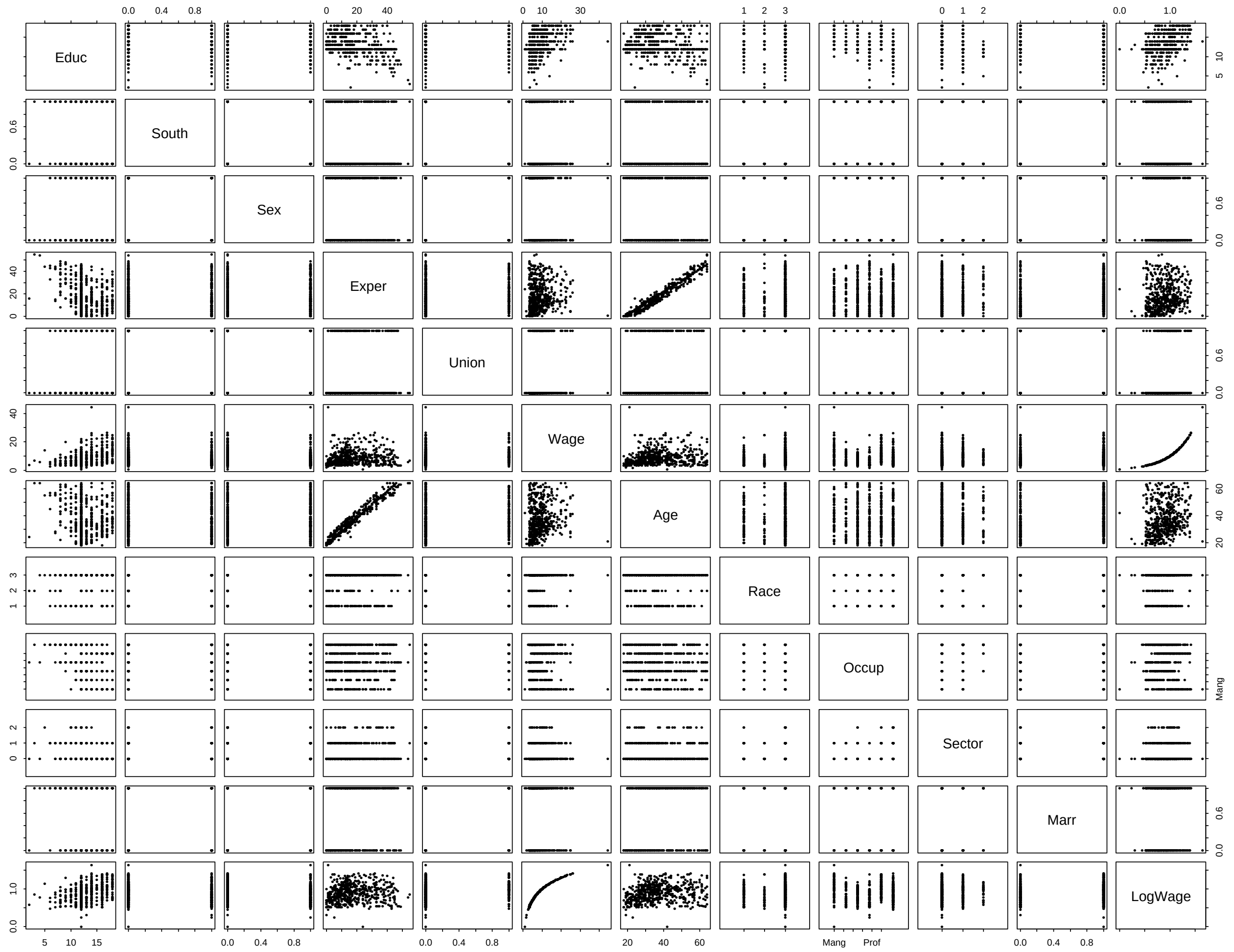
OCCUPATION: 0=Other, 1=Mngmnt, 2=Sales, 3=Clerical, 4=Service, 5=Prfssnl.

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

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

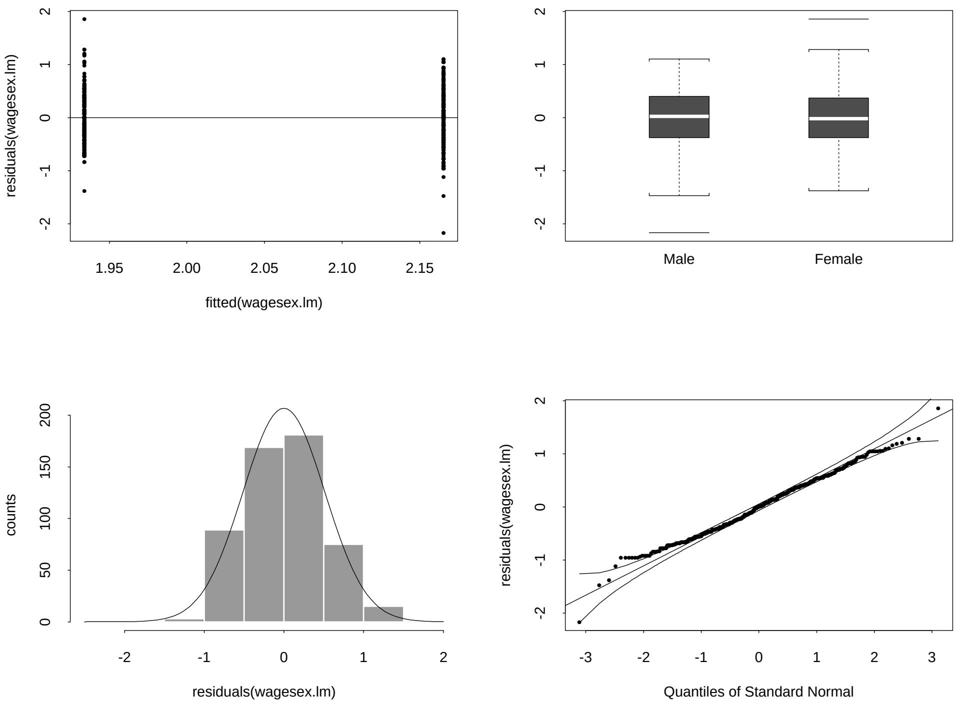


```
> wages <- read.table("wages.dat",header=T)
> wages <- cbind(wages,LogWage=log(wages$Wage)
> pairs(wages)
```

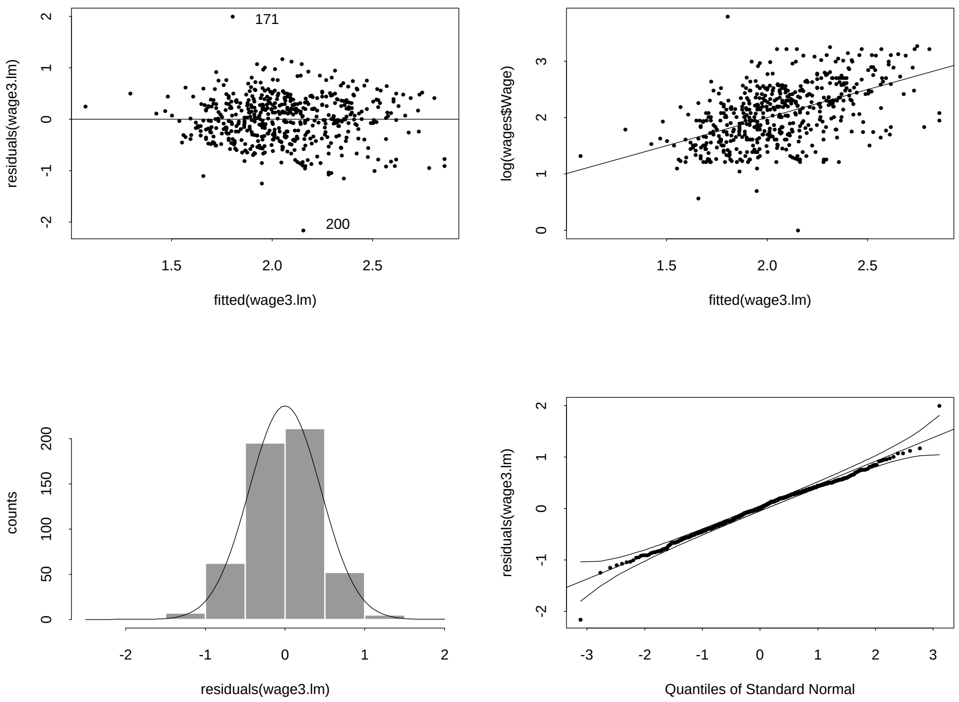



```
> round(cor(wages[,c("Educ","South","Sex","Exper","Union","Age",
+ "Marr","LogWage")]),3)
          Educ  South    Sex  Exper  Union    Age   Marr LogWage
   Educ  1.000 -0.140  0.002 -0.353 -0.024 -0.150 -0.036   0.380
  South -0.140  1.000 -0.021 -0.007 -0.086 -0.039  0.007  -0.172
    Sex  0.002 -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
```

```
> wage1.lm _ lm( log(Wage) ~ Sex, data=wages)
> summary(wage1.lm)

Coefficients:
                Value Std. Error  t value Pr(>|t|)
(Intercept)   2.1653   0.0303    71.4106   0.0000
        Sex  -0.2312   0.0448    -5.1658   0.0000

Residual standard error: 0.5155 on 532 degrees of freedom
Multiple R-Squared: 0.04776
F-statistic: 26.69 on 1 and 532 degrees of freedom, the p-value is 3.39e-07

> par(mfrow=c(2,2))
> plot(fitted(wage1.lm),residuals(wage1.lm))
> abline(h=0)
> boxplot(split(residuals(wage1.lm),wages$Sex),names=c("Male","Female"))
> ndhist(residuals(wage1.lm))  # copy "ndhist.q" and "qqenv.q" from the
> qqenv(residuals(wage1.lm))   # data area of the class webpages to your
>                              # own dir, then say e.g. source("ndhist.q")
```

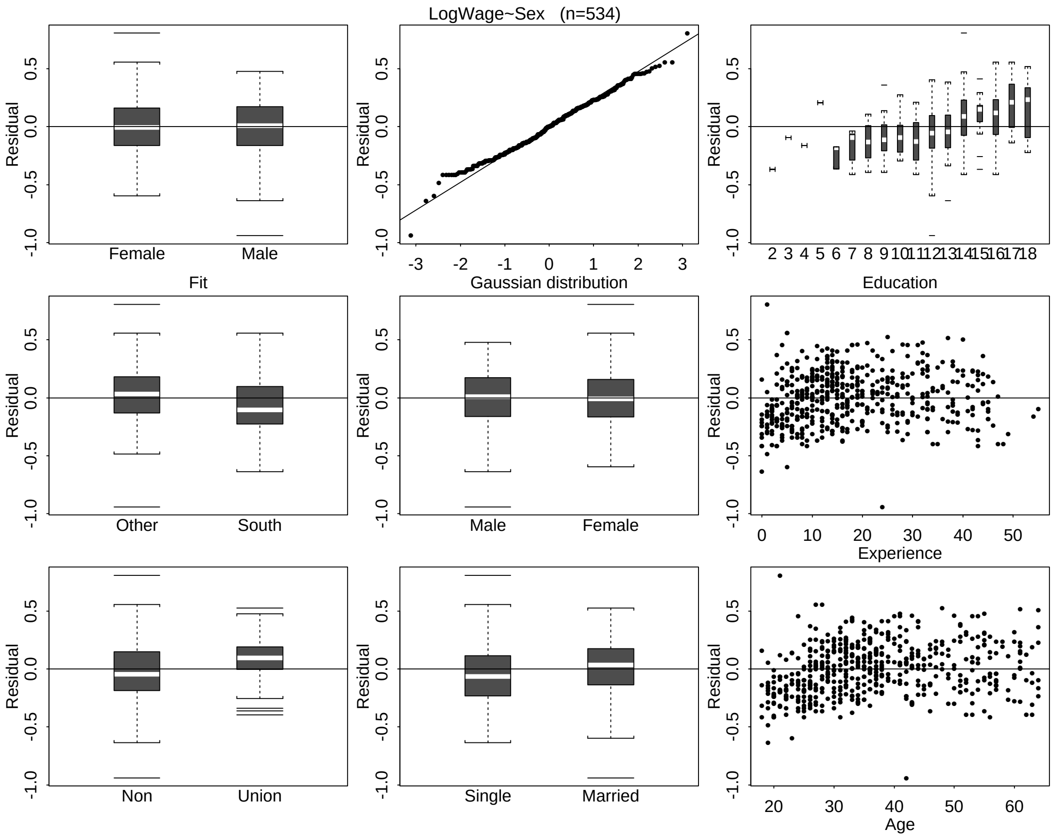


```
> wage3.lm _ lm(log(Wage) ~ Educ + Sex + Exper, data=wages)
> summary(wage3.lm)

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   0.6749   0.1205     5.6006   0.0000
       Educ   0.0980   0.0080    12.2372   0.0000
        Sex  -0.2559   0.0394    -6.4946   0.0000
      Exper   0.0127   0.0017     7.4661   0.0000

> par(mfrow=c(2,2))
> plot(fitted(wage3.lm),residuals(wage3.lm))
> abline(h=0)
> identify(fitted(wage3.lm),residuals(wage3.lm),labels=row.names(wages))
> plot(fitted(wage3.lm),log(wages$Wage))
> abline(0,1)
> ndhist(residuals(wage3.lm))
> qqenv(residuals(wage3.lm))
```

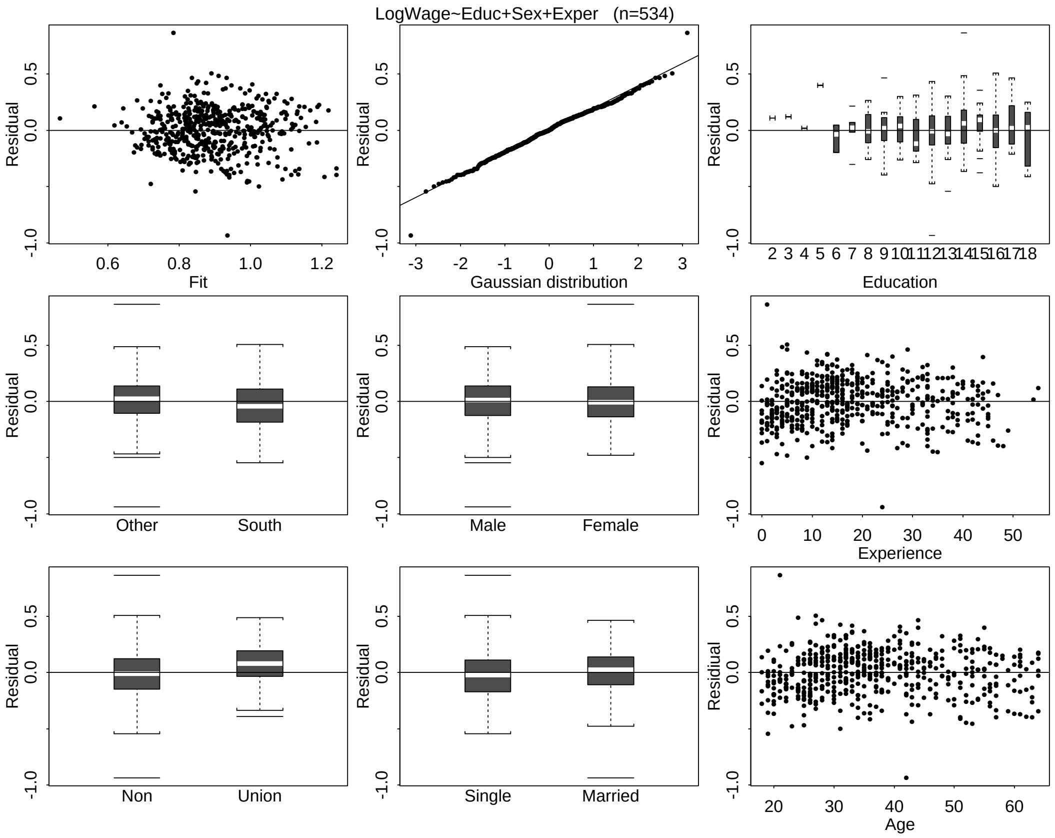



```
> wages[c(200,171),]
    Educ South Sex Exper Union Wage Age Race Occup Sector Marr LogWage
200   12     0   0    24     0  1.0  42    3 Manag      0    1 0.00000
171   14     0   1     1     0 44.5  21    3 Manag      0    0 1.64836
```

Comparing residuals of simple regression $\log(\mathrm{Wage}) = \beta_0 + \beta_1 \mathrm{Sex} + \epsilon \ldots$



$\ldots$ with residuals of multiple regression $\log(\mathrm{Wage}) = \beta_0 + \beta_1 \mathrm{Educ} + \beta_2 \mathrm{Sex} + \beta_3 \mathrm{Exper} + \epsilon$



More on the summary of a mulitple regression: F and R^2

```
> wage3.lm _ lm(log(Wage) ~ Educ + Sex + Exper, data=wages)
> summary(wage3.lm)
Call: lm(formula = log(Wage) ~ Educ + Sex + Exper, data = wages)
Residuals:
    Min      1Q   Median     3Q   Max
 -2.155 -0.3076 0.007084 0.3096 1.992

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   0.6749   0.1205     5.6006   0.0000
       Educ   0.0980   0.0080    12.2372   0.0000
        Sex  -0.2559   0.0394    -6.4946   0.0000
      Exper   0.0127   0.0017     7.4661   0.0000

Residual standard error: 0.4523 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
```

The F Statistic

$$F = \frac{SS_{Reg}/k}{RSS/(n-k-1)}$$

which is distributed as $F_{k,n-k-1}$ when the model is $y_i = \alpha + \epsilon_i$, with $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

- $SS_{Reg} = \sum(\hat{y}_i - \overline{y})^2$ can be obtained by summing the sums of squares in an ANOVA table for the regression (see below). Similarly for its df, k.
- $RSS = \sum(y_i - \hat{y}_i)^2$ is easy to get from Splus output (see below). Similarly for its df, $n - k - 1$.

```
> anova(wage3.lm)
          Df Sum of Sq  Mean Sq  F Value        Pr(F)
     Educ   1   21.4807 21.48068 105.0042 0.000000e+00
      Sex   1    7.1408  7.14079  34.9064 6.192308e-09
    Exper   1   11.4034 11.40337  55.7432 3.420000e-13
Residuals 530  108.4220  0.20457

> ( (21.4807 + 7.1408 +  11.4034)/(1 + 1 + 1) ) / ( 108.422/530 )
[1] 65.218
> 1-pf(65.218,3,530)   # look at help(pf)!
[1] 0
```

Multiple R^2

$$R^2 = r^2_{y,\hat{y}} = r_{y,X\hat{\beta}},$$

the *squared correlation* between the y's and the least-squares estimates $\hat{y} = X\hat{\beta}$.

- Can also calculate as

$$R^2 = \frac{\sum_i (\hat{y}_i - \overline{y})^2}{\sum_i (y_i - \overline{y})^2} = \frac{SS_{Reg}}{SYY} = 1 - \frac{RSS}{SYY}$$

- Multiple R^2 is the *maximum correlation* between y and *any possible* linear combination of the x's.

- *Whenever* we add more x variables, *we increase* R^2.

Adjusted R^2: Replace SS's with MS's

$$R^2_{Adj} = 1 - \frac{RSS/(n-k-1)}{SYY/(n-1)} = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2) = R^2 - \left(\frac{k}{n-k-1}\right)(1-R^2)$$

- Attempts to "penalize" for complicated models; R^2_{Adj} need not increase as the number of x variables increases.

- No particularly good distribution theory.

- Sometimes used to compare "non-nested" regressions, as well as regressions fitted on different data sets. *But there are other statistics that work better for this—more later!*

For the `wages` data set, there isn't much of an adjustment, since $n = 534$ is large, compared to $k \leq 3$, the number of regressors:

```
> attach(wages)
> R.summary _ function(call) {
+       tmp <- summary(lm(call)); R2 <- tmp$r.squared
+       k <- tmp$df[1] - 1; n <- tmp$df[2] + tmp$df[1]
+       R2a <- 1 - (n - 1)/(n - k - 1) * (1 - R2)
+       return(data.frame(call = paste(as.character(call)[c(2, 1, 3)],
+           collapse = ""), R2 = R2, R2adj = R2a)[1,  ]) }
> rbind( R.summary(LogWage ~ Educ), R.summary(LogWage ~ Educ + Sex),
+ R.summary(LogWage ~ Educ + Sex + Exper) )
                        call        R2     R2adj
1               LogWage ~ Educ 0.1447029 0.1430952
2         LogWage ~ Educ + Sex 0.1928062 0.1897660
3 LogWage ~ Educ + Sex + Exper 0.2696241 0.2654899
```