# wooldRidge-vignette

*Justin M Shea*

An excellent approach to learning is to find an example from your textbook and then recreate it. Below are examples from every chapter and the syntax provided here should get you through most of the book.

Load the `wooldRidge` package to access data in the manner specified in each example.

```
library(wooldRidge)
library(stargazer)
library(xtable)
options(xtable.comment = FALSE)
```

## Chapter 2: The Simple Regression Model

### `Example 2.10`: A Log Wage Equation

From the text:

> " Using the wage1 data as in Example 2.4, but using log(wage) as the dependent variable, we obtain the following relationship:"

$$\widehat{log(wage)} = \beta_0 + \beta_1 educ$$

First, load the `wage1` data.

```
data(wage1)
```

Next, estimate a linear relationship between the log of *wage* and *education*.

```
log_wage_model <- lm(lwage ~ educ, data = wage1)
```

Finally, print the coefficients and $R^2$.

```
xtable(log_wage_model)
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 0.5838 | 0.0973 | 6.00 | 0.0000 |
| educ | 0.0827 | 0.0076 | 10.94 | 0.0000 |

```
stargazer(log_wage_model, omit.table.layout = "n")
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Jul 03, 2017 - 5:02:51 PM

Table 1:

| | *Dependent variable:* |
| --- | --- |
| | lwage |
| educ | $0.083^{***}$ |
| | (0.008) |
| Constant | $0.584^{***}$ |
| | (0.097) |
| Observations | 526 |
| $R^2$ | 0.186 |
| Adjusted $R^2$ | 0.184 |
| Residual Std. Error | 0.480 (df = 524) |
| F Statistic | $119.582^{***}$ (df = 1; 524) |

## Chapter 3: Multiple Regression Analysis: Estimation

### `Example 3.2`: Hourly Wage Equation

From the text:

> " Using the 526 observations on workers in 'wage1', we include *educ*(years of education), *exper*(years of labor market experience), and *tenure*(years with the current employer) in an equation explain log(*wage*)."

$$\widehat{log(wage)} = \beta_0 + \beta_1 educ + \beta_3 exper + \beta_4 tenure$$

Estimate the model regressing *education*, *experience*, and *tenure* against *log(wage)*.

```
hourly_wage_model <- lm(lwage ~ educ + exper + tenure, data = wage1)
```

Again, print the estimated model coefficients:

```
hourly_wage_model$coefficients
```

```
## (Intercept)        educ       exper      tenure
## 0.284359541 0.092028988 0.004121109 0.022067218
```

# Chapter 4: Multiple Regression Analysis: Inference

### `Example 4.7` Effect of Job Training on Firm Scrap Rates

From the text:

> " The scrap rate for a manufacturing firm is the number of defective items - products that must be discarded - out of every 100 produced. Thus, for a given number of items produced, a decrease in the scrap rate reflects higher worker productivity."
>
> "We can use the scrap rate to measure the effect of worker training on productivity. Using the data in jtrain, but only for the year 1987 and for non-unionized firms, we obtain the following estimated equation:"

First, load the `jtrain` data set.

```r
data("jtrain")
```

Next, create a logical index identifying which observations occur in 1987 and are non-union.

```r
index <- jtrain$year == 1987 & jtrain$union == 0
```

Next, subset the jtrain data by the new index. This returns a data.frame of `jtrain` data of non-union firms for the year 1987.

```r
jtrain_1987_nonunion <- jtrain[index,]
```

Now create the linear model regressing hrsemp(total hours training/total employees trained), the log of annual sales, and the log of the number of the employees, against the log of the scrape rate.

$$lscrap = \alpha + \beta_1 hrsemp + \beta_2 lsales + \beta_3 lemploy$$

```r
linear_model <- lm(lscrap ~ hrsemp + lsales + lemploy, data = jtrain_1987_nonunion)
```

Finally, print the complete summary statistic diagnostics of the model.

```r
summary(linear_model)
```

```
##
## Call:
## lm(formula = lscrap ~ hrsemp + lsales + lemploy, data = jtrain_1987_nonunion)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6301 -0.7523 -0.4016  0.8697  2.8273
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.45837    5.68677   2.191   0.0380 *
## hrsemp      -0.02927    0.02280  -1.283   0.2111
## lsales      -0.96203    0.45252  -2.126   0.0436 *
## lemploy      0.76147    0.40743   1.869   0.0734 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.376 on 25 degrees of freedom
##   (97 observations deleted due to missingness)
## Multiple R-squared:  0.2624, Adjusted R-squared:  0.1739
## F-statistic: 2.965 on 3 and 25 DF,  p-value: 0.05134
```

## Chapter 5: Multiple Regression Analysis: OLS Asymptotics

### `Example 5.3`: Economic Model of Crime

From the text:

> "We illustrate the Lagrange Multiplier ($LM$) statistics test by using a slight extension of the crime model from example 3.5."

$$narr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + \mu$$

$narr86$ : number of times arrested, 1986.

$pcnv$ : proportion of prior arrests leading to convictions.

$avgsen$ : average sentence served, length in months.

$tottime$ : time in prison since reaching the age of 18, length in months.

$ptime86$ : months in prison during 1986

$qemp86$ : quarters employed, 1986

Load the `crime1` data set containing arrests during the year 1986 and other information on 2,725 men born in either 1960 or 1961 in California.

```
data(crime1)
```

From the text:

> "We use the $LM$ statistic to test the null hypothesis that $avgsen$ and $tottime$ have no effect on $narr86$ once other factors have been controlled for. First, estimate the restricted model by regressing $narr86$ on $pcnv, ptime86$, and $qemp86$; the variables $avgsen$ and $tottime$ are excluded from this regression."

```
restricted_model <- lm(narr86 ~ pcnv + ptime86 + qemp86, data = crime1)
```

We obtain the residuals $\tilde{\mu}$ from this regression, 2,725 of them.

```
restricted_model_u <- restricted_model$residuals
```

Next, we run the regression of:

$$\tilde{\mu} = \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86$$

From the text:

> "As always, the order in which we list the independent variables is irrelevant.This second regression produces $R^2_\mu$, which turns out to be about 0.0015."

```
LM_u_model <- lm(restricted_model_u ~ pcnv + ptime86 + qemp86 + avgsen + tottime,
    data = crime1)

summary(LM_u_model)$r.square
```

```
## [1] 0.001493846
```

> "This may seem small, but we must multiple it by $n$ to get the $LM$ statistic:"

$$LM = 2,725(0.0015)$$

```r
LM_test <- nobs(LM_u_model) * 0.0015
LM_test
```

```
## [1] 4.0875
```

> "The 10% critical value in a chi-square distribution with two degrees of freedom is about 4.61 (rounded to two decimal places)."

```r
qchisq(1 - 0.10, 2)
```

```
## [1] 4.60517
```

> "Thus, we fail to reject the null hypothesis that $\beta_{avgsen} = 0$ and $\beta_{tottime} = 0$ at the 10% level."

The $p$-value is:

$$P(X_2^2 > 4.09) \approx 0.129$$

so we would reject the $H_0$ at the 15% level.

```r
1-pchisq(LM_test, 2)
```

```
## [1] 0.129542
```

## Chapter 6: Multiple Regression: Further Issues

**Example 6.1:' Effects of Pollution on Housing Prices, standardized.**

From the text:

> "We use the data *hrprice*2 to illustrate the use of beta coefficients. Recall that the key independent variable is *nox*, a measure of nitrogen oxide in the air over each community. One way to understand the size of the pollution effect-without getting into the science underling nitrogen oxide's effect on air quality-is to compute beta coefficients. The population equation is the level-level model:"

$$price = \beta_0 + \beta_1 nox + \beta_2 crime + \beta_3 rooms + \beta_4 dist + \beta_5 stratio + \mu$$

*price*: median housing price.

*nox*: Nitrous Oxide concentration; parts per million.

*crime*: number of reported crimes per capita.

*rooms*: average number of rooms in houses in the community.

*dist*: weighted distance of the community to 5 employment centers.

*stratio*: average student-teacher ratio of schools in the community.

The beta coefficients are reported in the following equation (so each variable has been converted to its $z$-score):"

$$\widehat{zprice} = \beta_1 znox + \beta_2 zcrime + \beta_3 zrooms + \beta_4 zdist + \beta_5 zstratio$$

First, load the `hrpice2` data.

```
data(hrpice2)
```

Next, estimate the coefficient with the usual `lm` regression model but this time, standardized coefficients by wrapping each variable with R's `scale` function:

```
housing_standard <- lm(scale(price) ~ 0 + scale(nox) + scale(crime) + scale(rooms) +
    scale(dist) + scale(stratio), data = hprice2)

housing_standard$coefficients
```

```
##     scale(nox)    scale(crime)   scale(rooms)    scale(dist) scale(stratio)
##     -0.3404460     -0.1432828      0.5138878     -0.2348385     -0.2702799
```

**`Example 6.2`: Effects of Pollution on Housing Prices, Quadratic Interactive Term**

We modify the housing model, adding a quadratic term in *rooms*:

$$log(price) = \beta_0 + \beta_1 log(nox) + \beta_2 log(dist) + \beta_3 rooms + \beta_4 rooms^2 + \beta_5 stratio + \mu$$

```
housing_interactive <- lm(lprice ~ lnox + log(dist) + rooms+I(rooms^2) + stratio, data = hprice2)

summary(housing_interactive)
```

```
##
## Call:
## lm(formula = lprice ~ lnox + log(dist) + rooms + I(rooms^2) +
##     stratio, data = hprice2)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04285 -0.12774  0.02038  0.12650  1.25272
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.385478   0.566473  23.630  < 2e-16 ***
## lnox        -0.901682   0.114687  -7.862 2.34e-14 ***
## log(dist)   -0.086781   0.043281  -2.005  0.04549 *
## rooms       -0.545113   0.165454  -3.295  0.00106 **
## I(rooms^2)   0.062261   0.012805   4.862 1.56e-06 ***
## stratio     -0.047590   0.005854  -8.129 3.42e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2592 on 500 degrees of freedom
## Multiple R-squared:  0.6028, Adjusted R-squared:  0.5988
## F-statistic: 151.8 on 5 and 500 DF,  p-value: < 2.2e-16
```

## Chapter 7: Multiple Regression Analysis with Qualitative Information

### `Example 7.4`: Housing Price Regression, Qualitative Binary variable

This time we use the `hrpice1` data.

```
data(hrpice1)
```

Having just worked with `hrpice2`, it may be helpful to view the documentation on this data set and read the variable names.

```
?hprice1
```

$$\widehat{log(price)} = \beta_0 + \beta_1 log(lotsize) + \beta_2 log(sqrft) + \beta_3 bdrms + \beta_4 colonial$$

Estimate the coefficients of the above linear model on the `hprice` data set.

```
housing_qualitative <- lm(lprice ~ llotsize + lsqrft + bdrms + colonial, data = hprice1)

summary(housing_qualitative)
```

```
##
## Call:
## lm(formula = lprice ~ llotsize + lsqrft + bdrms + colonial, data = hprice1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69479 -0.09750 -0.01619  0.09151  0.70228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.34959    0.65104  -2.073   0.0413 *
## llotsize     0.16782    0.03818   4.395 3.25e-05 ***
## lsqrft       0.70719    0.09280   7.620 3.69e-11 ***
## bdrms        0.02683    0.02872   0.934   0.3530
## colonial     0.05380    0.04477   1.202   0.2330
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1841 on 83 degrees of freedom
## Multiple R-squared:  0.6491, Adjusted R-squared:  0.6322
## F-statistic: 38.38 on 4 and 83 DF,  p-value: < 2.2e-16
```

Summary from the text:

> "All the variables are self-explanatory except *colonial*, which is a binary variable equal to one if the house is of the colonial style. What does the coefficient on *colonial* mean? For given levels of *lotsize*, *sqrt*, and *bdrms*, the difference in $\widehat{log(price)}$ between a house of colonial style and that of another style is 0.54. This means that colonial-style house is predicted to sell for about 5.4% more, holding other factors fixed."

## Chapter 8: Heteroskedasticity

### `Example 8.9`: Determinants of Personal Computer Ownership

> "We use the data in $GPA1$ to estimate the probability of owning a computer. Let $PC$ denote a binary indicator equal to unity if the student owns a computer, and zero otherwise. The variable $hsGPA$ is high school GPA, $ACT$ is achievement test score, and $parcoll$ is a binary indicator equal to unity if at least one parent attended college."
>
> "The equation estimated by OLS is:"

$$\widehat{PC} = \beta_0 + \beta_1 hsGPA + \beta_2 ACT + \beta_3 parcoll + \beta_4 colonial$$

Load the `gpa1` data and create a new variable combining the`fathcoll` and `mothcoll`, into one, `parcoll`. This new column indicates if any parent went to college, not just one or the other.

```r
data(GPA1)
```

```
## Warning in data(GPA1): data set 'GPA1' not found
```

```r
gpa1$parcoll <- as.integer(gpa1$fathcoll==1 | gpa1$mothcoll)
```

```r
GPA_OLS <- lm(PC ~ hsGPA + ACT + parcoll, data = gpa1)

summary(GPA_OLS)
```

```
##
## Call:
## lm(formula = PC ~ hsGPA + ACT + parcoll, data = gpa1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4915 -0.4494 -0.2437  0.5375  0.8223
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0004322  0.4905358  -0.001   0.9993
## hsGPA        0.0653943  0.1372576   0.476   0.6345
## ACT          0.0005645  0.0154967   0.036   0.9710
## parcoll      0.2210541  0.0929570   2.378   0.0188 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.486 on 137 degrees of freedom
## Multiple R-squared:  0.04153,    Adjusted R-squared:  0.02054
## F-statistic: 1.979 on 3 and 137 DF,  p-value: 0.1201
```

> "Just as with example 8.8, there are no striking differences between the usual and robust standard errors. Nevertheless, we also estimate the model by Weighted Least Squares or $WLS$. Because all of the $OLS$ fitted values are inside the unit interval, no adjustments are needed"

First, calculate the weights and then pass them to the same linear model.

```r
weights <- GPA_OLS$fitted.values * (1-GPA_OLS$fitted.values)

GPA_WLS <- lm(PC ~ hsGPA + ACT + parcoll, data = gpa1, weights = 1/weights)

summary(GPA_WLS)
```

```
##
## Call:
## lm(formula = PC ~ hsGPA + ACT + parcoll, data = gpa1, weights = 1/weights)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0015 -0.9029 -0.5576  1.0800  2.0429
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.026210   0.476650   0.055   0.9562
## hsGPA       0.032703   0.129882   0.252   0.8016
## ACT         0.004272   0.015453   0.276   0.7826
## parcoll     0.215186   0.086292   2.494   0.0138 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.016 on 137 degrees of freedom
## Multiple R-squared:  0.04644,    Adjusted R-squared:  0.02556
## F-statistic: 2.224 on 3 and 137 DF,  p-value: 0.08816
```

> "There are no important differences in the OLS and WLS estimates. The only significant explanatory variable is *parcoll*, and in both cases we estimate that the probability of *PC* ownership is about .22 higher if at least on parent attended college"

## Chapter 9: More on Specification and Data Issues

### `Example 9.8`: R&D Intensity and Firm Size

> "Suppose the R&D expenditures as a percentage of sales, $rdintens$, are realted to $sales$ (in millions) and profits as a percentage of sales, $profmarg$:"

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 profmarg + \mu$$

> "The $OLS$ equation using data on 32 chemical companies in $rdchem$ is"

Load the data, run the model, and apply the `summary` diagnostics function to the model.

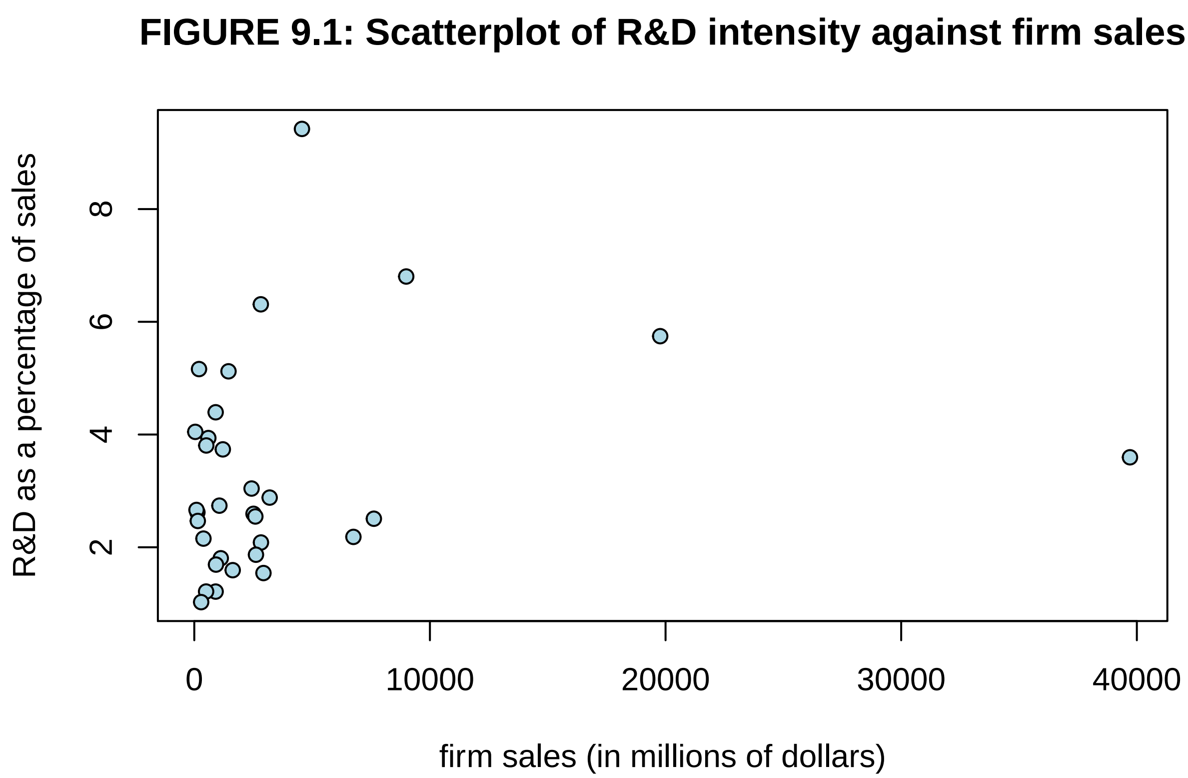
```r
data(rdchem)

all_rdchem <- lm(rdintens ~ sales + profmarg, data = rdchem)

summary(all_rdchem)
```

```
##
## Call:
## lm(formula = rdintens ~ sales + profmarg, data = rdchem)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2221 -1.1414 -0.6068  0.5008  6.3702
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.625e+00  5.855e-01   4.484 0.000106 ***
## sales       5.338e-05  4.407e-05   1.211 0.235638
## profmarg    4.462e-02  4.618e-02   0.966 0.341966
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.862 on 29 degrees of freedom
## Multiple R-squared:  0.07612,    Adjusted R-squared:  0.0124
## F-statistic: 1.195 on 2 and 29 DF,  p-value: 0.3173
```

Neither $sales$ nor $profmarg$ is statistically significant at even the 10% level in this regression.

Of the 32 firms, 31 have annual sales less than 20 billion. One firm has annual sales of almost 40 billions. Figure 9.1 shows how far this firm is from the rest of the sample.

```r
plot(rdintens ~ sales, pch = 21, bg = "lightblue", data = rdchem, main = "FIGURE 9.1: Scatterplot of R&
    xlab = "firm sales (in millions of dollars)", ylab = "R&D as a percentage of sales")
```




"In terms of sales, this firm is over twice as large as every other firm, so it might be a good idea to estimate the model without it. When we do this, we obtain:"

```
smallest_rdchem <- lm(rdintens ~ sales + profmarg, data = rdchem,
                      subset = (sales < max(sales)))
summary(smallest_rdchem)
```

```
##
## Call:
## lm(formula = rdintens ~ sales + profmarg, data = rdchem, subset = (sales <
##     max(sales)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0687 -1.1867 -0.7956  0.6486  6.0811
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.2968508  0.5918045   3.881 0.000577 ***
## sales       0.0001856  0.0000842   2.204 0.035883 *
## profmarg    0.0478411  0.0444831   1.075 0.291336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.792 on 28 degrees of freedom
## Multiple R-squared:  0.1728, Adjusted R-squared:  0.1137
## F-statistic: 2.925 on 2 and 28 DF,  p-value: 0.07022
```

## Chapter 10: Basic Regression Analysis with Time Series Data

### Example 10.2: Effects of Inflation and Deficits on Interest Rates

> "The data in INTDEF.RAW come from the 2004 Economic Report of the President (Tables B-73 and B-79) and span the years 1948 through 2003. The variable $i3$ is the three-month T-bill rate, $inf$ is the annual inflation rate based on the consumer price index (CPI), and $def$ is the federal budget deficit as a percentage of GDP. The estimated equation is:"

$$\widehat{i3} = \beta_0 + \beta_1 inf_t + \beta_2 def_t$$

```r
data("intdef")

tbill_model <- lm(i3 ~ inf + def, data = intdef)

summary(tbill_model)
```

```
##
## Call:
## lm(formula = i3 ~ inf + def, data = intdef)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9948 -1.1694  0.1959  0.9602  4.7224
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.73327    0.43197   4.012  0.00019 ***
## inf          0.60587    0.08213   7.376 1.12e-09 ***
## def          0.51306    0.11838   4.334 6.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.843 on 53 degrees of freedom
## Multiple R-squared:  0.6021, Adjusted R-squared:  0.5871
## F-statistic: 40.09 on 2 and 53 DF,  p-value: 2.483e-11
```

> "These estimates show that increases in inflation or the relative size of the deficit increase short-term interest rates, both of which are expected from basic economics. For example, a ceteris paribus one percentage point increase in the inflation rate increases i3 by .606 points. Both inf and def are very statistically significant, assuming, of course, that the CLM assumptions hold."

### Example 10.11: Seasonal Effects of Antidumping Filings

In *Example 10.5*, we used monthly data (in the file `BARIUM`) that have not been seasonally adjusted.

```r
# Example 10.5
data("barium")

lm(lchnimp ~ lchempi + lgas + lrtwex + befile6 + affile6 + afdec6, data = barium)
```

```
##
## Call:
## lm(formula = lchnimp ~ lchempi + lgas + lrtwex + befile6 + affile6 +
##     afdec6, data = barium)
##
## Coefficients:
```

```
## (Intercept)      lchempi         lgas       lrtwex      befile6
##   -17.80300      3.11719      0.19635      0.98302      0.05957
##     affile6       afdec6
##    -0.03241     -0.56524
```

> "Therefore, we should add seasonal dummy variables to make sure none of the important conclusions change. It could be that the months just before the suit was filed are months where imports are higher or lower, on average, than in other months."

```r
barium_seasonal <- lm(lchnimp ~ lchempi + lgas + lrtwex + befile6 + affile6 + afdec6 + feb + mar + apr +

barium_seasonal_hat <- lm(lchnimp ~ lchempi + lgas + lrtwex + befile6 + affile6 + afdec6, data = barium)

anova(barium_seasonal, barium_seasonal_hat)
```

```
## Analysis of Variance Table
##
## Model 1: lchnimp ~ lchempi + lgas + lrtwex + befile6 + affile6 + afdec6 +
##     feb + mar + apr + may + jun + jul + aug + sep + oct + nov +
##     dec
## Model 2: lchnimp ~ lchempi + lgas + lrtwex + befile6 + affile6 + afdec6
##   Res.Df    RSS  Df Sum of Sq      F Pr(>F)
## 1    113 40.844
## 2    124 44.247 -11   -3.4032 0.8559 0.5852
```

> "When we add the 11 monthly dummy variables as in 10.41 and test their joint significance, we obtain $p - value = 5.5852$, and so the seasonal dummies are jointly insignificant. In addition, nothing important changes in the estimates once statistical significance is taken into account. Krupp and Pollard (1996) actually used three dummy variables for the seasons (fall, spring, and summer, with winter as the base season), rather than a full set of monthly dummies; the outcome is essentially the same."

## Chapter 11: Further Issues in Using OLS with with Time Series Data

### `Example 11.7`: Wages and Productivity

"The variable *hrwage* is average hourly wage in the U.S. economy, and *outphr* is output per hour. One way to estimate the elasticity of hourly wage with respect to output per hour is to estimate the equation:"

$$\widehat{log(hrwage_t)} = \beta_0 + \beta_1 log(outphr_t) + \beta_2 t + \mu_t$$

"where the time trend is included because $log(hrwage)$ and $log(outphr)$ both display clear, upward, linear trends. Using the data in 'EARNS' for the years 1947 through 1987, we obtain:"

```r
data("earns")

wage_time <- lm(lhrwage ~ loutphr + t, data = earns)

summary(wage_time)
```

```
##
## Call:
## lm(formula = lhrwage ~ loutphr + t, data = earns)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.059230 -0.026151  0.002411  0.020322  0.051966
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.328454   0.374449  -14.23  < 2e-16 ***
## loutphr      1.639639   0.093347   17.57  < 2e-16 ***
## t           -0.018230   0.001748  -10.43 1.05e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02854 on 38 degrees of freedom
## Multiple R-squared:  0.9712, Adjusted R-squared:  0.9697
## F-statistic: 641.2 on 2 and 38 DF,  p-value: < 2.2e-16
```

"(We have reported the usual goodness-of-fit measures here; it would be better to report those based on the detrended dependent variable, as in Section 10.5.). The estimated elasticity seems too large: a 1% increase in productivity increases real wages by about 1.64%. Because the standard error is so small, the 95% confidence interval easily excludes a unit elasticity. U.S. workers would probably have trouble believing that their wages increase by more than 1.5% for every 1% increase in productivity."

"The regression results must be viewed with caution. Even after linearly detrending $log(hrwage)$, the first order autocorrelation is .967, and for detrended $log(outphr), \hat{p} = 0.945$. These suggest that both series have unit roots, so we reestimate the equation in first differences (and we no longer need a time trend):"

```r
wage_diff <- lm(diff(lhrwage) ~ diff(loutphr), data = earns)

summary(wage_diff)
```

```
##
## Call:
```

```
## lm(formula = diff(lhrwage) ~ diff(loutphr), data = earns)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.040921 -0.010165 -0.000383  0.007969  0.040329
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.003662   0.004220  -0.868    0.391
## diff(loutphr)  0.809316   0.173454   4.666 3.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01695 on 38 degrees of freedom
## Multiple R-squared:  0.3642, Adjusted R-squared:  0.3475
## F-statistic: 21.77 on 1 and 38 DF,  p-value: 3.748e-05
```

> "Now, a 1% increase in productivity is estimated to increase real wages by about 0.81%, and the estimate is not statistically different from one. The adjusted Râ€‘squared shows that the growth in output explains about 35% of the growth in real wages."

## Chapter 12: Serial Correlation and Heteroskedasticiy in Time Series Regressions

### `Example 12.4`: Prais-Winsten Estimation in the Event Study

> "Again using the data in BARIUM, we estimate the equation in Example 10.5 using iterated Prais-Winsten estimation."
>
> "The coefficients that are statistically significant in the Prais-Winsten estimation do not differ by much from the OLS estimates [in particular, the coefficients on $log(chempi)$, $log(rtwex)$, and $afdec6$]. It is not surprising for statistically insignificant coefficients to change, perhaps markedly, across different estimation methods.

First, run the linear model from example 10.5 and 10.11.

```
data("barium")
# Example 10.5
barium_linear_model <- lm(lchnimp ~ lchempi + lgas + lrtwex + befile6 + affile6 +
    afdec6, data = barium)

barium_linear_model
```

```
##
## Call:
## lm(formula = lchnimp ~ lchempi + lgas + lrtwex + befile6 + affile6 +
##     afdec6, data = barium)
##
## Coefficients:
## (Intercept)      lchempi         lgas       lrtwex      befile6
##   -17.80300      3.11719      0.19635      0.98302      0.05957
##     affile6       afdec6
##    -0.03241     -0.56524
```

Then load the `prais` package and use the `prais.winsten` function to estimate the same model. Print the names of both models to the console to compare the results of both.

```
library(prais)
barium_prais_winsten <- prais.winsten(lchnimp ~ lchempi + lgas + lrtwex + befile6 + affile6 + afdec6, da

barium_prais_winsten
```

```
## [[1]]
##
## Call:
## lm(formula = fo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01146 -0.39152  0.06758  0.35063  1.35021
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## Intercept -37.07771   22.77830  -1.628   0.1061
## lchempi     2.94095    0.63284   4.647 8.46e-06 ***
## lgas        1.04638    0.97734   1.071   0.2864
## lrtwex      1.13279    0.50666   2.236   0.0272 *
## befile6    -0.01648    0.31938  -0.052   0.9589
## affile6    -0.03316    0.32181  -0.103   0.9181
```

```
## afdec6     -0.57681    0.34199  -1.687   0.0942 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5733 on 124 degrees of freedom
## Multiple R-squared:  0.9841, Adjusted R-squared:  0.9832
## F-statistic:  1096 on 7 and 124 DF,  p-value: < 2.2e-16
##
##
## [[2]]
##        Rho Rho.t.statistic Iterations
##  0.2932171        3.483363          8
```

> "Notice how the standard errors in the second column are uniformly higher than the standard errors in column (1). This is common. The Prais-Winsten standard errors account for serial correlation; the $OLS$ standard errors do not. As we saw in Section 12.1, the OLS standard errors usually understate the actual sampling variation in the OLS estimates and should not be relied upon when significant serial correlation is present. Therefore, the effect on Chinese imports after the International Trade Commissionâ€™s decision is now less statistically significant than we thought."
>
> "Finally, an R-squared is reported he $PW$ estimation that is well below the R-squared for the $OLS$ estimation in this case. However, these R-squareds should not be compared. For $OLS$, the R-squared, as usual, is based on the regression with the untransformed dependent and independent variables. For $PW$, the R-squared comes from the final regression of the $transformed$ dendent variable on the transformed independent vari-ables. It is not clear what this $R^2$ actually measuring; nevertheless, it is traditionally reported."

**`Example 12.8`: Heteroskedasticity and the Efficient Markets Hypothesis**

> "In Example 11.4, we estimated the simple $AR(1)$ model:"
>
> $$return_t = \beta_0 + \beta_1 return_{t-1} + \mu_t$$
>
> "The EMH states that $\beta_1 = 0$. When we tested this hypothesis using the data in 'NYSE', we obtained $t_b 1 = 1.55$ with $n = 689$.

```r
data("nyse")

return_AR <-lm(return ~ return_1, data = nyse)

summary(return_AR)
```

```
##
## Call:
## lm(formula = return ~ return_1, data = nyse)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.261  -1.302   0.098   1.316   8.065
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.17963    0.08074   2.225   0.0264 *
## return_1     0.05890    0.03802   1.549   0.1218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.11 on 687 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.003481,   Adjusted R-squared:  0.00203
## F-statistic: 2.399 on 1 and 687 DF,  p-value: 0.1218
```

> "With such a large sample, this is not much evidence against the EMH. Although the EMH states that the expected return given past observable information should be constant, it says nothing about the conditional variance. In fact, the Breusch-Pagan test for heteroskedasticity entails regressing the squared $OLS$ residuals $\hat{\mu_t^2}$ on $return_{t-1}$""

$$\hat{\mu_t^2} = \beta_0 + \beta_1 return_{t-1} + residual_t$$

Calculated $\hat{\mu_t^2}$ by taking the residuals contained in the `return_AR` model object and store the results in the variable named `return_mu`. Then regress the `return_1` variable against the square of `return_mu`. Notice, we set data equal to the `return_AR` objects model matrix, which contains data free of leading missing values inherent to lagged variables.

```r
return_mu <- residuals(return_AR)

mu2_hat_model <- lm(return_mu^2 ~ return_1, data = return_AR$model)

summary(mu2_hat_model)
```

```
##
## Call:
## lm(formula = return_mu^2 ~ return_1, data = return_AR$model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -9.689  -3.929  -2.021   0.960 223.730
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.6565     0.4277  10.888  < 2e-16 ***
## return_1     -1.1041     0.2014  -5.482  5.9e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.18 on 687 degrees of freedom
## Multiple R-squared:  0.04191,    Adjusted R-squared:  0.04052
## F-statistic: 30.05 on 1 and 687 DF,  p-value: 5.905e-08
```

> "The $t$ statistic on $return_{t-1}$ is about -5.5, indicating strong evidence of heteroskedasticity. Because the coeffict on $return_{t-1}$ is negative, we have the interesting finding that volatility in stock returns is lower the previous return was high, and vice versa. Therefore, we have found what is common in many financial studies: the expected value of stock returns does not depend on past returns, but the variance of returns does."

### `Example 12.9`: ARCH in Stock Returns

> "In Example 12.8, we saw that there was heteroskedasticity in weekly stock returns. This heteroskedasticity is actually better characterized by the ARCH model in (12.50). If we compute the OLS residuals from (12.47), square these, and regress them on the lagged squared residual, we obtain:"

$$\hat{\mu_t^2} = \beta_0 + \hat{\mu_{t-1}^2} + residual_t$$

We still have `return_mu` in the working environment so we can use it to create $\hat{\mu_t^2}$, (`mu2_hat`) and $\hat{\mu_{t-1}^2}$ (`mu2_hat_1`). Notice the use `R`'s matrix subset operations to perform the lag operation. We drop the first observation of `mu2_hat` and squared the results. Next, we remove the last observation of `mu2_hat_1` using the subtraction operator combined with a call to the `NROW` function on `return_mu`. Now, both contain 688 observations and we can run a standard linear model.

```
mu2_hat  <- return_mu[-1]^2

mu2_hat_1 <- return_mu[-NROW(return_mu)]^2

arch_model <- lm(mu2_hat ~ mu2_hat_1)

summary(arch_model)
```

```
##
## Call:
## lm(formula = mu2_hat ~ mu2_hat_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.337  -3.292  -2.157   0.556 223.981
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.94743    0.44023   6.695 4.49e-11 ***
## mu2_hat_1    0.33706    0.03595   9.377  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.76 on 686 degrees of freedom
## Multiple R-squared:  0.1136, Adjusted R-squared:  0.1123
## F-statistic: 87.92 on 1 and 686 DF,  p-value: < 2.2e-16
```

> "The t statistic on $\hat{\mu_{t-1}^2}$ (mu2_hat_1) is over nine, indicating strong ARCH. As we discussed earlier, a larger error at time $t-1$ implies a larger variance in stock returns today.
>
> "It is important to see that, though the *squared OLS* residuals are autocorrelated, the *OLS* residuals themselves are not (as is consistent with the EMH). Regressing on $\hat{\mu_t}$ and $\hat{\mu_{t-1}}$ gives $\hat{p} = 0.0014$ with $t_{\hat{p}} = 0.038$.

## Chapter 13: Pooling Cross Sections across Time: Simple Panel Data Methods

### `Example 13.7`: Effect of Drunk Driving Laws on Traffic Fatalities

> "Many states in the United States have adopted different policies in an attempt to curb drunk driving. Two types of laws that we will study here are *open container laws* -which make it illegal for passengers to have open containers of alcoholic beverages -and *administrative per se laws* -which allow courts to suspend licenses after a driver is arrested for drunk driving but before the driver is convicted. One possible analysis is to use a single cross section of states to regress driving fatalities (or those related to drunk driving) on dummy variable indicators for whether each law is present. This is unlikely to work well because states decide, through legislative processes, whether they need such laws. Therefore, the presence of laws is likely to be related to the average drunk driving fatalities in recent years. A more convincing analysis uses panel data over a time period where some states adopted new laws (and some states may have repealed existing laws). The file TRAFFIC1 contains data for 1985 and 1990 for all 50 states and the District of Columbia. The dependent variable is the number of traffic deaths per 100 million miles driven (dthrte). In 1985, 19 states had open container laws while 22 states had such laws in 1990. In 1985, 21 states had per se laws; the number had grown to 29 by 1990. Using OLS after first differencing gives:"

$$\widehat{\Delta dthrte} = \beta_0 + \Delta open + \Delta admin$$

```r
data("traffic1")

DD_model <- lm(cdthrte ~ copen + cadmn, data = traffic1)

summary(DD_model)
```

```
##
## Call:
## lm(formula = cdthrte ~ copen + cadmn, data = traffic1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25261 -0.14337 -0.00321  0.19679  0.79679
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.49679    0.05243  -9.476 1.43e-12 ***
## copen       -0.41968    0.20559  -2.041   0.0467 *
## cadmn       -0.15060    0.11682  -1.289   0.2035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3435 on 48 degrees of freedom
## Multiple R-squared:  0.1187, Adjusted R-squared:  0.08194
## F-statistic: 3.231 on 2 and 48 DF,  p-value: 0.04824
```

> "The estimates suggest that adopting an open container law lowered the traffic fatality rate by 0.42, a nontrivial effect given that the average death rate in 1985 was 2.7 with a standard deviation of about 0.6. The estimate is statistically significant at the 5% level against a twosided alternative. The administrative per se law has a smaller effect, and its t statistic is only -1.29; but the estimate is the sign we expect. The intercept in this equation shows that traffic fatalities fell substantially for all states over the five-year period, whether or not there were any law changes. The states that adopted an open container law over this period saw a further drop, on average, in fatality rates."

“Other laws might also affect traffic fatalities, such as seat belt laws, motorcycle helmet laws, and maximum speed limits. In addition, we might want to control for age and gender distributions, as well as measures of how influential an organization such as Mothers Against Drunk Driving is in each state.”

## Chapter 14: Advanced Panel Data Methods

### `Example 14.1`: Effect of Job Training on Firm Scrap Rates

> "We use the data for three years, 1987, 1988, and 1989, on the 54 firms that reported scrap rates in each year. No firms received grants prior to 1988; in 1988, 19 firms received grants; in 1989, 10 different firms received grants. Therefore, we must also allow for the possibility that the additional job training in 1988 made workers more productive in 1989. This is easily done by including a lagged value of the grant indicator. We also include year dummies for 1988 and 1989. The results are given in Table below.

Install the `plm` package and check out the documentation. The model syntax for `plm` models is very similar to the linear model, with additional slots to further define various estimation methods.

```r
library(plm)

data("jtrain")

scrap_panel <- plm(lscrap ~ d88 + d89 + grant + grant_1, data = jtrain, index = c("fcode",
    "year"), model = "within", effect = "individual")

summary(scrap_panel)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = lscrap ~ d88 + d89 + grant + grant_1, data = jtrain,
##     effect = "individual", model = "within", index = c("fcode",
##         "year"))
##
## Balanced Panel: n=54, T=3, N=162
##
## Residuals :
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -2.286936 -0.112387 -0.017841  0.144272  1.426674
##
## Coefficients :
##          Estimate Std. Error t-value Pr(>|t|)
## d88     -0.080216   0.109475 -0.7327  0.46537
## d89     -0.247203   0.133218 -1.8556  0.06634 .
## grant   -0.252315   0.150629 -1.6751  0.09692 .
## grant_1 -0.421590   0.210200 -2.0057  0.04749 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    32.25
## Residual Sum of Squares: 25.766
## R-Squared:      0.20105
## Adj. R-Squared: -0.23684
## F-statistic: 6.54259 on 4 and 104 DF, p-value: 9.7741e-05
```

> "We have reported the results in a way that emphasizes the need to interpret the estimates in light of the unobserved effects model, (14.4). We are explicitly controlling for the unobserved, time-constant effects in $\alpha_i$. The time-demeaning allows us to estimate the $\beta_j$, but (14.5) is not the best equation for interpreting the estimates.

> "Interestingly, the estimated lagged effect of the training grant is substantially larger than the

contemporaneous effect: job training has an effect at least one year later. Because the dependent variable is in logarithmic form, obtaining a grant in 1988 is predicted to lower the firm scrap rate in 1989 by about 34.4% [exp(-0.422)-1 = -0.344]; the coefficient on $grant_1$ is significant at the 5% level against a twosided alternative. The coefficient $grant$ is significant at the 10% level, and the size of the coefficient is hardly trivial. Notice the *df* is obtained as N(T-1) - k = 54(3-1)-4 = 104"

"The coefficient on $d89$ indicates that the scrap rate was substantially lower in 1989 than in the base year, 1987, even in the absence of job training grants. Thus, it is important to allow for these aggregate effects. If we omitted the year dummies, the secular increase in worker productivity would be attributed to the job training grants. The diagnostic results above shows that, even after controlling for aggregate trends in productivity, the job training grants had a large estimated effect."

"Finally, it is crucial to allow for the lagged effect in the model. If we omit $grant_1$, then we are assuming that the effect of job training does not last into the next year. The estimate on $grant$ when we drop $grant_1$ is -0.082 $t = -0.65$; this is much smaller and statistically insignificant."

## Chapter 15: Instrumental Variables Estimation and Two Stage Least Squares

### Example 15.1: Estimating the Return to Education for Married Women

> "We use the data on married working women in *mroz* to estimate the return to education in the simple regression model"
>
> $$log(wage) = \beta_0 + \beta_1 educ + \mu$$
>
> "For comparison, we first obtain the *OLS* estimates:"

```r
data("mroz")

wage_educ_model <- lm(lwage ~ educ, data = mroz)

summary(wage_educ_model)
```

```
##
## Call:
## lm(formula = lwage ~ educ, data = mroz)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.10256 -0.31473  0.06434  0.40081  2.10029
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1852     0.1852  -1.000    0.318
## educ          0.1086     0.0144   7.545 2.76e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.68 on 426 degrees of freedom
##   (325 observations deleted due to missingness)
## Multiple R-squared:  0.1179, Adjusted R-squared:  0.1158
## F-statistic: 56.93 on 1 and 426 DF,  p-value: 2.761e-13
```

> "The estimate for $\beta_1$ implies an almost 11% return for another year of education."
>
> "Next, we use father's education $fatheduc$ as an instrumental variable for *educ*. We have to maintain that $fatheduc$ is uncorrelated with $\mu$. The second requirement is that *educ* and $fatheduc$ are correlated. We can check this very easily using a simple regression of *educ* on $fatheduc$, using only the working women in the sample:"
>
> $$\widehat{educ} = \beta_0 + \beta_1 fatheduc$$

We run the typical linear model, but notice the use of the `subset` argument. `inlf` is a binary variable in which a value of 1 means they are "In the Labor Force". By sub-setting the `mroz` data.frame by observations in which `inlf==1`, only working women will be in the sample.

```r
fatheduc_model <- lm(educ ~ fatheduc, data = mroz, subset = (inlf==1))

summary(fatheduc_model)
```

```
##
## Call:
```

```
## lm(formula = educ ~ fatheduc, data = mroz, subset = (inlf ==
##     1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4704 -1.1231 -0.1231  0.9546  5.9546
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.23705    0.27594  37.099   <2e-16 ***
## fatheduc     0.26944    0.02859   9.426   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.081 on 426 degrees of freedom
## Multiple R-squared:  0.1726, Adjusted R-squared:  0.1706
## F-statistic: 88.84 on 1 and 426 DF,  p-value: < 2.2e-16
```

> "The $t$ statistic on $fatheduc$ is 9.42, which indicates that *educ* and $fatheduc$ have a statistically significant positive correlation. In fact, $fatheduc$ explains about 17% of the variation in *educ* in the sample. Using fatheduc as an $IV$ for *educ* gives:"

In this section, we will perform an **Instrumental-Variable Regression**, using the `ivreg` function in the `AER (Applied Econometrics with R)` package.

```r
library("AER")

wage_educ_IV <- ivreg(lwage ~ educ | fatheduc, data = mroz)

summary(wage_educ_IV, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = lwage ~ educ | fatheduc, data = mroz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0870 -0.3393  0.0525  0.4042  2.0677
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.44110    0.44610   0.989   0.3233
## educ         0.05917    0.03514   1.684   0.0929 .
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   1 426     88.84  <2e-16 ***
## Wu-Hausman         1 425      2.47   0.117
## Sargan             0  NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6894 on 426 degrees of freedom
## Multiple R-Squared: 0.09344, Adjusted R-squared: 0.09131
## Wald test: 2.835 on 1 and 426 DF,  p-value: 0.09294
```

"The $IV$ estimate of the return to education is 5.9%, which is barely more than one half of the OLS. This suggests that the $OLS$ estimate is too high and is consistent with omitted ability bias. But we should remember that these are estimates from just one sample: we can never know whether 0.109 is above the true return to education, or whether 0.059 is closer to the true return to education. Further, the standard error of the $IV$ estimate is two and one-half times as large as the $OLS$ standard error this is expected, for the reasons we gave earlier. The 95% confidence interval for using $OLS$ is much tighter than that using the $IV$. In fact, the $IV$ confidence interval actuay contains the $OLS$ estimate. Therefore, although the differences between 15.15 and 15.17 are practically large, we cannot say whether the difference is statistically significant. We will show how to test this in Section 15.5."

**`Example 15.2`: Estimating the Return to Education for Men**

"We now use $wage2$ data to estimate the return to education for men. We use the variable *sibs*, or number of siblings, as an instrument for *educ*. These are negatively correlated, as we can verify from a simple regression:"

$$\widehat{educ} = \beta_0 + sibs$$

```r
data("wage2")

educ_sibs_model <- lm(educ ~ sibs, data = wage2)

summary(educ_sibs_model)
```

```
##
## Call:
## lm(formula = educ ~ sibs, data = wage2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.139 -1.683 -0.683  1.931  6.140
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.13879    0.11314 124.969  < 2e-16 ***
## sibs        -0.22792    0.03028  -7.528 1.22e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.134 on 933 degrees of freedom
## Multiple R-squared:  0.05726,    Adjusted R-squared:  0.05625
## F-statistic: 56.67 on 1 and 933 DF,  p-value: 1.215e-13
```

"This equation implies that every sibling is associated with, on average, about 0.23 less of a year of education. If we assume that sibs is uncorrelated with the error term in 15.14, then the IV estimator is consistent. Estimating equation 15.14 from example 15.1, using *sibs* as an IV for *educ* gives:"

$$\widehat{log(wage)} = \beta_0 + educ$$

In this section, we will perform an **Instrumental-Variable Regression**, using the `ivreg` function in the `AER (Applied Econometrics with R)` package.

```r
library("AER")

educ_sibs_IV <- ivreg(lwage ~ educ | sibs, data = wage2)

summary(educ_sibs_IV, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = lwage ~ educ | sibs, data = wage2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85429 -0.26950  0.04223  0.29276  1.31038
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.13003    0.35517  14.444  < 2e-16 ***
## educ         0.12243    0.02635   4.646 3.86e-06 ***
##
## Diagnostic tests:
##                  df1 df2 statistic  p-value
## Weak instruments   1 933    56.667 1.22e-13 ***
## Wu-Hausman         1 932     6.733  0.00961 **
## Sargan             0  NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4233 on 933 degrees of freedom
## Multiple R-Squared: -0.009174,   Adjusted R-squared: -0.01026
## Wald test: 21.59 on 1 and 933 DF,  p-value: 3.865e-06
```

> "For comparison, the OLS estimate of $\beta_1$ is 0.059 with a standard error of 0.006. Unlike in the previous example, the $IV$ estimate is now much higher than the $OLS$ estimate. While we do not know whether the difference is statistically significant, this does not mesh with the omitted ability bias from $OLS$. It could be that *sibs* is also correlated with ability: more siblings means, on average, less parental attention, which could result in lower ability. Another interpretation is that the $OLS$ estimator is biased toward zero because of measurement error in *educ*. This is not entirely convincing because, as we discussed in Section 9.3, educ is unlikely to satisfy the classical errors-in-variables model."

**Example 15.5: Return to Education for Working Women**

> "We estimate equation 15.40 using the data in *mroz*. First, we test $H_0 : \pi_3 = 0, \pi_4 = 0$ in 15.41using an $F$ test. The result is $F = 55.40$, and $p - value = 0.0000$. As expected, *educ* is partially correlated with parents education."

> "When we estimate 15.40 by 2SLS, we obtain, in equation form,"

$$\widehat{log(wage)} = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2$$

```r
library(AER)
data("mroz")

wage_educ_exper_IV <- ivreg(lwage ~ educ + exper + expersq | exper + expersq +
    motheduc + fatheduc, data = mroz)
```

```r
summary(wage_educ_exper_IV, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = lwage ~ educ + exper + expersq | exper + expersq +
##     motheduc + fatheduc, data = mroz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0986 -0.3196  0.0551  0.3689  2.3493
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0481003  0.4003281   0.120  0.90442
## educ         0.0613966  0.0314367   1.953  0.05147 .
## exper        0.0441704  0.0134325   3.288  0.00109 **
## expersq     -0.0008990  0.0004017  -2.238  0.02574 *
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   2 423    55.400  <2e-16 ***
## Wu-Hausman         1 423     2.793  0.0954 .
## Sargan             1  NA     0.378  0.5386
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6747 on 424 degrees of freedom
## Multiple R-Squared: 0.1357,  Adjusted R-squared: 0.1296
## Wald test: 8.141 on 3 and 424 DF,  p-value: 2.787e-05
```

> "The estimated return to education is about 6.1%, compared with an *OLS* estimate of about 10.8%. Because of its relatively large standard error, the 2SLS estimate is barely statistically significant at the 5% level against a two-sided alternative."

## Chapter 16: Simultaneous Equations Models

### `Example 16.4:` INFLATION AND OPENNESS

> "Romer (1993) proposes theoretical models of inflation that imply that more"open" countries should have lower inflation rates. His empirical analysis explains average annual inflation rates (since 1973) in terms of the average share of imports in gross domestic product since 1973 - which is his measure of openness. In addition to estimating the key equation by OLS, he uses instrumental variables. While Romer does not specify both equations in a simultaneous system, he has in mind a two-equation system:"
>
> $$inf = \beta_{10} + \alpha_1 open + \beta_{11} log(pcinc) + \mu_1$$
> $$open = \beta_{20} + \alpha_2 inf + \beta_{21} log(pcinc) + \beta_{22} log(land) + \mu_2$$
>
> "where *pcinc* is 1980 per capita income, in U.S. dollars, assumed to be exogenous, and *land* is the land area of the country in square miles, also assumed to be exogenous. The first equation is the one of interest, with the hypothesis that $\alpha < 0$. More open economies have lower inflation rates."
>
> "The second equation reflects the fact that the degree of openness might depend on the average inflation rate, as well as other factors. The variable $log(pcinc)$ appears in both equations, but $log(land)$ is assumed to appear only in the second equation. The idea is that, ceteris paribus, a smaller country is likely to be more open, so $\beta_{22} < 0$."
>
> "Using the identification rule that was stated earlier, the first equation is identified, provided $\beta_{22} \neq 0$. The second equation is *not* identified because it contains both exogenous variables. Be we are interested in the first equation.

### `Example 16.6:` INFLATION AND OPENNESS

> "Before we estimate the first equation in 16.4 using the data in *openness*, we check to see whether *open* has sufficient partial correlation with the proposed $IV$, $log(land)$. The reduced form regression is:"
>
> $$\widehat{open} = \beta_0 + \beta_1 log(pcinc) + \beta_2 log(land)$$

```r
data("openness")

open_model <-lm(open ~ lpcinc + lland, data = openness)

summary(open_model)
```

```
##
## Call:
## lm(formula = open ~ lpcinc + lland, data = openness)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.907  -8.843  -3.109   6.057  82.792
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 117.0845    15.8483   7.388 2.97e-11 ***
## lpcinc        0.5465     1.4932   0.366    0.715
## lland        -7.5671     0.8142  -9.294 1.51e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 17.8 on 111 degrees of freedom
## Multiple R-squared:  0.4487, Adjusted R-squared:  0.4387
## F-statistic: 45.17 on 2 and 111 DF,  p-value: 4.451e-15
```

> "The $t$ statistic on $log(land)$ is over nine in absolute value, which verifies Romer's assertion that smaller countries are more open. The fact that $log(pcinc)$ is so insignificant in this regression is irrelevant."
>
> "Estimating the first equation using $log(land)$ as an $IV$ for $open$ gives:"

$$\widehat{inf} = \beta_0 + \beta_1 open + \beta_2 log(pcinc)$$

```r
library(AER)

inflation_IV <- ivreg(inf ~ open + lpcinc | lpcinc + lland, data = openness)

summary(inflation_IV)
```

```
##
## Call:
## ivreg(formula = inf ~ open + lpcinc | lpcinc + lland, data = openness)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.686 -10.176  -5.857   2.912 184.875
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.8993    15.4012   1.747   0.0835 .
## open         -0.3375     0.1441  -2.342   0.0210 *
## lpcinc        0.3758     2.0151   0.187   0.8524
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.84 on 111 degrees of freedom
## Multiple R-Squared: 0.03088, Adjusted R-squared: 0.01341
## Wald test:  2.79 on 2 and 111 DF,  p-value: 0.06574
```

The coefficient on open is statistically significant at about the 1% level against a one sided alternative of $\alpha_1 < 0$. The effect is economically important as well: for every percentage point increase in the import share of GDP, annual inflation is about 1/3 of a percentage point lower. For comparison, the OLS estimate is -0.215, $se = 0.095$.

## Chapter 18: Advanced Time Series Topics

### `Example 18.8`: FORECASTING THE U.S. UNEMPLOYMENT RATE

> "We use the $PHILLIPS$ DATA, but only for the years 1948 through 1996, to forecast the U.S. civilian unemployment rate for 1997. We use two models. The first is a simple AR(1) model for $unem$:"

$$\widehat{unemp}_t = \beta_0 + \beta_1 unem_{t-1}$$

> "In a second model, we add inflation with a lag of one year:"

$$\widehat{unemp}_t = \beta_0 + \beta_1 unem_{t-1} + \beta_2 inf_{t-1}$$

```r
data("phillips")

library(dynlm)

phillips <- ts(phillips, start = 1948)

unem_AR1 <- dynlm(unem ~ unem_1, data = phillips, end = 1996)
unem_inf_VAR1 <- dynlm(unem ~ unem_1 + inf_1, data = phillips, end = 1996)
```

```r
stargazer(unem_AR1, unem_inf_VAR1, keep.stat=c("n","adj.rsq","ser"))
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Jul 03, 2017 - 5:02:53 PM

Table 2:

| | *Dependent variable:* | |
|---|---|---|
| | unem | |
| | (1) | (2) |
| unem_1 | 0.732*** | 0.647*** |
| | (0.097) | (0.084) |
| inf_1 | | 0.184*** |
| | | (0.041) |
| Constant | 1.572*** | 1.304** |
| | (0.577) | (0.490) |
| Observations | 48 | 48 |
| Adjusted $R^2$ | 0.544 | 0.677 |
| Residual Std. Error | 1.049 (df = 46) | 0.883 (df = 45) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

"The lagged inflation rate is very significant in the second model ($t \approx 4.5$), and the adjusted R-squared much higher than that from the first. Nevertheless, this does not necessarily mean that the second equation will produce a better forecast for 1997. All we can say so far is that, using the data up through 1996, a lag of inflation helps to explain variation in the unemployment rate."

"To obtain the forecasts for 1997, we need to know *unemployment* and *inflation* in 1996. These are 5.4 and 3.0, respectively. Therefore, the forecast of $unem_{1997}$ from the first equation is $1.572 + .732(5.4)$, or about 5.52. The forecast from the second equation is $1.304 + 0.647(5.4) + 0.184(3.0)$, or about 5.35. The actual civilian unemployment rate for 1997 was 4.9, so both equations overpredict the actual rate. The second equation does provide a somewhat better forecast."

"We can easily obtain a 95% forecast interval. When we regress $unem_1$ on $(unem_{t-1} - 5.4)$ and $(inf_{t-1} - 3.0)$, we obtain 5.35 as the intercept - which we already computed as the forecast - and $se(\hat{f}_n) = 0.137$. Therefore, because $\hat{\sigma} = 0.883$, we have $se(\hat{e_{n+1}}) = [(0.137)^2 + (0.883)^2]^{1/2} \approx 0.894$. The 95% forecast interval of $\hat{f}_n \pm 1.96 * se(\hat{e_{n-1}})$ is $5.35 \pm 1.96(0.894)$, or about [3.6, 7.1]. This is a wide interval, and the realized 1997 value, 4.9, is well within the interval. As expected, the standard error of $\mu_{n+1}$, which is .883, is a very large fraction of $se(\hat{e_{n-1}})$"