wooldridge-vignette

Justin M Shea

Contents

Introduction

This vignette contains examples from every chapter of *Introductory Econometrics: A Modern Approach* by Jeffrey M. Wooldridge. Each example illustrates how to load data, build econometric models, and compute estimates with **R**.

Economics students new to both econometrics and **R** may find the introduction to both a bit challenging. In particular, the process of loading and preparing data prior to building one's first econometric model can present challenges. The `wooldridge` data package aims to lighten this task. It contains 105 data sets from *Introductory Econometrics: A Modern Approach*, and will load any set by typing its name into the `data()` function.

While the course companion site also provides publicly available data sets for Eviews, Excel, MiniTab, and Stata commercial software, **R** is the open source option. Furthermore, using **R** while building a foundation in econometrics, can become the first step in a student's journey toward using the most innovative new methods in statistical computing for handling larger, more modern data sets.

In addition, please visit the **Appendix** for sources on using R for econometrics. For example, an excellent reference is *"Using R for Introductory Econometrics"* by Florian Hess, written to compliment *Introductory Econometrics: A Modern Approach.* The full text can be viewed on the book website.

Now, install and load the `wooldridge` package and lets get started.

```r
install.packages("wooldridge")
```

```r
library(wooldridge)
```

Chapter 2: The Simple Regression Model

`Example 2.10`: A Log Wage Equation

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

Load the `wage1` data and check out the documentation.

```
data("wage1")
?wage1
```

These are data from the 1976 Current Population Survey, collected by Henry Farber when he and Wooldridge were colleagues at MIT in 1988.

Estimate a linear relationship between the *log of wage* and *education*.

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

Print the results. I'm using the `stargazer` package to print the model results in a clean and easy to read format. See the bibliography for more information.

```
stargazer(log_wage_model, single.row = TRUE, header = FALSE)
```

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)
Note:	*p<0.1; **p<0.05; ***p<0.01

Chapter 3: Multiple Regression Analysis: Estimation

`Example 3.2`: Hourly Wage Equation

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

Estimate the model regressing *education*, *experience*, and *tenure* against *log(wage)*. The `wage1` data should still be in your working environment.

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

Print the estimated model coefficients:

```
stargazer(hourly_wage_model,  single.row = TRUE, header = FALSE)
```

Table 2:

	Dependent variable:
	lwage
educ	0.092*** (0.007)
exper	0.004** (0.002)
tenure	0.022*** (0.003)
Constant	0.284*** (0.104)
Observations	526
R^2	0.316
Adjusted R^2	0.312
Residual Std. Error	0.441 (df = 522)
F Statistic	80.391*** (df = 3; 522)
Note:	*p<0.1; **p<0.05; ***p<0.01

Chapter 4: Multiple Regression Analysis: Inference

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

Load the `jtrain` data set and if you are using R Studio, `View` the data set.

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

From H. Holzer, R. Block, M. Cheatham, and J. Knott (1993), *Are Training Subsidies Effective? The Michigan Experience*, Industrial and Labor Relations Review 46, 625-636. The authors kindly provided the data.

```
?jtrain
View(jtrain)
```

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

```
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.

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

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

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

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

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

```
stargazer(linear_model,  single.row = TRUE, header = FALSE)
```

Table 3:

	Dependent variable:
	lscrap
hrsemp	-0.029 (0.023)
lsales	-0.962^{**} (0.453)
lemploy	0.761^{*} (0.407)
Constant	12.458^{**} (5.687)
Observations	29
R^2	0.262
Adjusted R^2	0.174
Residual Std. Error	1.376 (df = 25)
F Statistic	2.965^{*} (df = 3; 25)
Note:	*p<0.1; **p<0.05; ***p<0.01

Chapter 5: Multiple Regression Analysis: OLS Asymptotics

`Example 5.3`: Economic Model of Crime

From J. Grogger (1991), *Certainty vs. Severity of Punishment*, Economic Inquiry 29, 297-309. Professor Grogger kindly provided a subset of the data he used in his article.

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

$narr86$: 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.

```
data("crime1")
?crime1
```

Estimate the model.

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

Create a new variable `restricted_model_u` containing the residuals $\tilde{\mu}$ from the above regression.

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

Next, regress `pcnv`, `ptime86`, `qemp86`, `avgsen`, and `tottime`, against the residuals $\tilde{\mu}$ saved in `restricted_model_u`.

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

```
LM_u_model <- lm(restricted_model_u ~ pcnv + ptime86 + qemp86 + avgsen + tottime,
    data = crime1)
summary(LM_u_model)$r.square
```

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

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

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

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

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

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

The p-value is:

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

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

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

Chapter 6: Multiple Regression: Further Issues

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

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

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.

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

Load the `hprice2` data and view the documentation.

```
data("hprice2")
?hprice2
```

Data from *Hedonic Housing Prices and the Demand for Clean Air*, by Harrison, D. and D.L.Rubinfeld, Journal of Environmental Economics and Management 5, 81-102. Diego Garcia, a former Ph.D. student in economics at MIT, kindly provided these data, which he obtained from the book Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, by D.A. Belsey, E. Kuh, and R. Welsch, 1990. New York: Wiley.

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)
```

```
stargazer(housing_standard,  single.row = TRUE, header = FALSE)
```

Table 4:

	Dependent variable:
	scale(price)
scale(nox)	−0.340*** (0.044)
scale(crime)	−0.143*** (0.031)
scale(rooms)	0.514*** (0.030)
scale(dist)	−0.235*** (0.043)
scale(stratio)	−0.270*** (0.030)
Observations	506
R^2	0.636
Adjusted R^2	0.632
Residual Std. Error	0.606 (df = 501)
F Statistic	174.822*** (df = 5; 501)
Note:	*p<0.1; **p<0.05; ***p<0.01

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

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

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

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

Compare the results with the model from `example 6.1`.

```r
stargazer(housing_standard, housing_interactive, single.row = TRUE, header = FALSE)
```

Table 5:

	Dependent variable:	
	scale(price)	lprice
	(1)	(2)
scale(nox)	−0.340*** (0.044)	
scale(crime)	−0.143*** (0.031)	
scale(rooms)	0.514*** (0.030)	
scale(dist)	−0.235*** (0.043)	
scale(stratio)	−0.270*** (0.030)	
lnox		−0.902*** (0.115)
log(dist)		−0.087** (0.043)
rooms		−0.545*** (0.165)
I(roomsˆ2)		0.062*** (0.013)
stratio		−0.048*** (0.006)
Constant		13.385*** (0.566)
Observations	506	506
R^2	0.636	0.603
Adjusted R^2	0.632	0.599
Residual Std. Error	0.606 (df = 501)	0.259 (df = 500)
F Statistic	174.822*** (df = 5; 501)	151.770*** (df = 5; 500)
Note:		*p<0.1; **p<0.05; ***p<0.01

Chapter 7: Multiple Regression Analysis with Qualitative Information

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

This time, use the `hrprice1` data.

```
data("hprice1")
```

Data collected from the real estate pages of the Boston Globe during 1990. These are homes that sold in the Boston, MA area.

If you recently worked with `hprice2`, 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)
```

```
stargazer(housing_qualitative, single.row = TRUE, header = FALSE)
```

Table 6:

	Dependent variable:
	lprice
llotsize	0.168^{***} (0.038)
lsqrft	0.707^{***} (0.093)
bdrms	0.027 (0.029)
colonial	0.054 (0.045)
Constant	-1.350^{**} (0.651)
Observations	88
R^2	0.649
Adjusted R^2	0.632
Residual Std. Error	0.184 (df = 83)
F Statistic	38.378^{***} (df = 4; 83)
Note:	*p<0.1; **p<0.05; ***p<0.01

Chapter 8: Heteroskedasticity

Example 8.9: Determinants of Personal Computer Ownership

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

Christopher Lemmon, a former MSU undergraduate, collected these data from a survey he took of MSU students in Fall 1994. Load `gpa1` and create a new variable combining the `fathcoll` and `mothcoll`, into `parcoll`. This new column indicates if either parent went to college.

```
data("gpa1")
?gpa1
```

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

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

Calculate the weights and then pass them to the `weights` argument.

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

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

Compare the OLS and WLS model in the table below:

```
stargazer(GPA_OLS, GPA_WLS, single.row = TRUE, header = FALSE)
```

Table 7:

	Dependent variable:	
	PC	
	(1)	(2)
hsGPA	0.065 (0.137)	0.033 (0.130)
ACT	0.001 (0.015)	0.004 (0.015)
parcoll	0.221** (0.093)	0.215** (0.086)
Constant	−0.0004 (0.491)	0.026 (0.477)
Observations	141	141
R^2	0.042	0.046
Adjusted R^2	0.021	0.026
Residual Std. Error (df = 137)	0.486	1.016
F Statistic (df = 3; 137)	1.979	2.224*
Note:		*p<0.1; **p<0.05; ***p<0.01

Chapter 9: More on Specification and Data Issues

Example 9.8: R&D Intensity and Firm Size

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

From *Businessweek R&D Scoreboard*, October 25, 1991. Load the data and estimate the model.

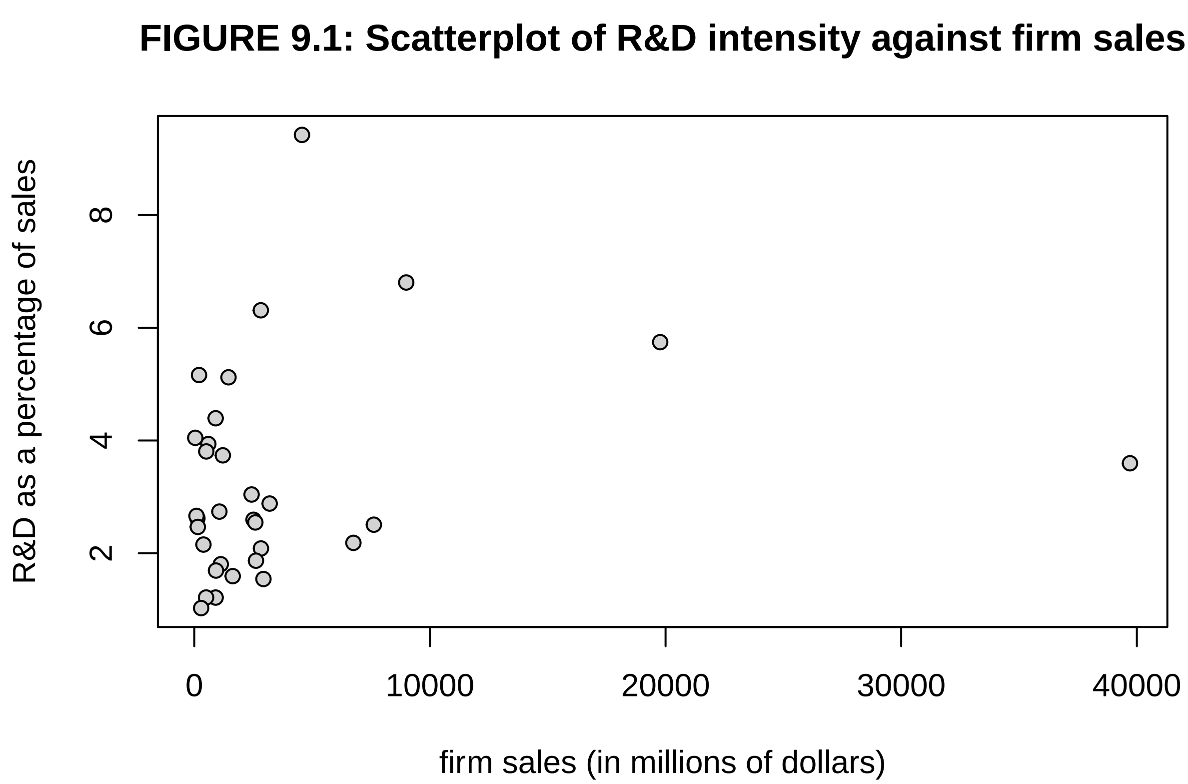
```
data("rdchem")
?rdchem
```

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

Plotting the data reveals the outlier on the far right of the plot, which will skew the results of our model.

```
plot_title <- "FIGURE 9.1: Scatterplot of R&D intensity against firm sales"
x_axis <- "firm sales (in millions of dollars)"
y_axis <- "R&D as a percentage of sales"

plot(rdintens ~ sales, pch = 21, bg = "lightgrey", data = rdchem, main = plot_title,
    xlab = x_axis, ylab = y_axis)
```



So, we can estimate the model without that data point to gain a better understanding of how `sales` and `profmarg` describe `rdintens` for most firms. We can use the `subset` argument of the linear model function to indicate that we only want to estimate the model using data that is less than the highest sales.

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

The table below compares the results of both models side by side. By removing the outlier firm, *sales* become a more significant determination of R&D expenditures.

```r
stargazer(all_rdchem, smallest_rdchem, single.row = TRUE, header = FALSE)
```

Table 8:

	Dependent variable:	
	rdintens	
	(1)	(2)
sales	0.0001 (0.00004)	0.0002** (0.0001)
profmarg	0.045 (0.046)	0.048 (0.044)
Constant	2.625*** (0.586)	2.297*** (0.592)
Observations	32	31
R^2	0.076	0.173
Adjusted R^2	0.012	0.114
Residual Std. Error	1.862 (df = 29)	1.792 (df = 28)
F Statistic	1.195 (df = 2; 29)	2.925* (df = 2; 28)
Note:		*p<0.1; **p<0.05; ***p<0.01

Chapter 10: Basic Regression Analysis with Time Series Data

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

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

Data from the *Economic Report of the President, 2004*, Tables B-64, B-73, and B-79.

```
data("intdef")
?intdef
```

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

```
stargazer(tbill_model, single.row = TRUE, header = FALSE)
```

Table 9:

	Dependent variable:
	i3
inf	0.606*** (0.082)
def	0.513*** (0.118)
Constant	1.733*** (0.432)
Observations	56
R^2	0.602
Adjusted R^2	0.587
Residual Std. Error	1.843 (df = 53)
F Statistic	40.094*** (df = 2; 53)
Note:	*p<0.1; **p<0.05; ***p<0.01

`Example 10.11`: Seasonal Effects of Antidumping Filings

C.M. Krupp and P.S. Pollard (1999), *Market Responses to Antidumpting Laws: Some Evidence from the U.S. Chemical Industry*, Canadian Journal of Economics 29, 199-227. Dr. Krupp kindly provided the data. They are monthly data covering February 1978 through December 1988.

```
data("barium")
?barium
```

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

Estimate a new model, `barium_seasonal` which accounts for seasonality by adding dummy variables contained in the data. Compute the `anova` between the two models.

```
barium_seasonal <- lm(lchnimp ~ lchempi + lgas + lrtwex + befile6 + affile6 +
    afdec6 + feb + mar + apr + may + jun + jul + aug + sep + oct + nov + dec,
    data = barium)
barium_anova <- anova(barium_imports, barium_seasonal)
```

```
stargazer(barium_imports, barium_seasonal, single.row = TRUE, header = FALSE)
```

```
stargazer(barium_anova, single.row = TRUE, header = FALSE)
```

Table 10:

	Dependent variable:	
	lchnimp	
	(1)	(2)
lchempi	3.117*** (0.479)	3.265*** (0.493)
lgas	0.196 (0.907)	−1.278 (1.389)
lrtwex	0.983** (0.400)	0.663 (0.471)
befile6	0.060 (0.261)	0.140 (0.267)
affile6	−0.032 (0.264)	0.013 (0.279)
afdec6	−0.565* (0.286)	−0.521* (0.302)
feb		−0.418 (0.304)
mar		0.059 (0.265)
apr		−0.451* (0.268)
may		0.033 (0.269)
jun		−0.206 (0.269)
jul		0.004 (0.279)
aug		−0.157 (0.278)
sep		−0.134 (0.268)
oct		0.052 (0.267)
nov		−0.246 (0.263)
dec		0.133 (0.271)
Constant	−17.803 (21.045)	16.779 (32.429)
Observations	131	131
R^2	0.305	0.358
Adjusted R^2	0.271	0.262
Residual Std. Error	0.597 (df = 124)	0.601 (df = 113)
F Statistic	9.064*** (df = 6; 124)	3.712*** (df = 17; 113)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 11:

Statistic	N	Mean	St. Dev.	Min	Max
Res.Df	2	118.500	7.778	113	124
RSS	2	42.545	2.406	40.844	44.247
Df	1	11.000		11	11
Sum of Sq	1	3.403		3.403	3.403
F	1	0.856		0.856	0.856
Pr(>F)	1	0.585		0.585	0.585

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

`Example 11.7`: Wages and Productivity

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

Data from the *Economic Report of the President, 1989*, Table B-47. The data are for the non-farm business sector.

```
data("earns")
?earns
```

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

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

```
stargazer(wage_time, wage_diff, single.row = TRUE, header = FALSE)
```

Table 12:

	Dependent variable:	
	lhrwage	diff(lhrwage)
	(1)	(2)
loutphr	1.640*** (0.093)	
t	−0.018*** (0.002)	
diff(loutphr)		0.809*** (0.173)
Constant	−5.328*** (0.374)	−0.004 (0.004)
Observations	41	40
R^2	0.971	0.364
Adjusted R^2	0.970	0.348
Residual Std. Error (df = 38)	0.029	0.017
F Statistic	641.224*** (df = 2; 38)	21.771*** (df = 1; 38)
Note:		*p<0.1; **p<0.05; ***p<0.01

Chapter 12: Serial Correlation and Heteroskedasticiy in Time Series Regressions

Example 12.4: Prais-Winsten Estimation in the Event Study

```r
data("barium")
barium_model <- lm(lchnimp ~ lchempi + lgas + lrtwex + befile6 + affile6 + afdec6,
    data = barium)
# Load the `prais` package, use the `prais.winsten` function to estimate.
library(prais)
barium_prais_winsten <- prais.winsten(lchnimp ~ lchempi + lgas + lrtwex + befile6 +
    affile6 + afdec6, data = barium)
```

```r
barium_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
```

```r
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
```

Example 12.8: Heteroskedasticity and the Efficient Markets Hypothesis

These are Wednesday closing prices of value-weighted NYSE average, available in many publications. Wooldridge does not recall the particular source used when he collected these data at MIT, but notes probably the easiest way to get similar data is to go to the NYSE web site, www.nyse.com.

$$return_t = \beta_0 + \beta_1 return_{t-1} + \mu_t$$

```
data("nyse")
?nyse
```

```
return_AR1 <-lm(return ~ return_1, data = nyse)
```

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

```
return_mu <- residuals(return_AR1)
mu2_hat_model <- lm(return_mu^2 ~ return_1, data = return_AR1$model)
```

```
stargazer(return_AR1, mu2_hat_model, single.row = TRUE, header = FALSE)
```

Table 13:

	Dependent variable:	
	return	return_mu^2
	(1)	(2)
return_1	0.059 (0.038)	-1.104^{***} (0.201)
Constant	0.180^{**} (0.081)	4.657^{***} (0.428)
Observations	689	689
R^2	0.003	0.042
Adjusted R^2	0.002	0.041
Residual Std. Error (df = 687)	2.110	11.178
F Statistic (df = 1; 687)	2.399	30.055^{***}
Note:		*p<0.1; **p<0.05; ***p<0.01

`Example 12.9`: ARCH in Stock Returns

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

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 estimate 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)
```

```
stargazer(arch_model, single.row = TRUE, header = FALSE)
```

Table 14:

	Dependent variable:
	mu2_hat
mu2_hat_1	0.337*** (0.036)
Constant	2.947*** (0.440)
Observations	688
R^2	0.114
Adjusted R^2	0.112
Residual Std. Error	10.759 (df = 686)
F Statistic	87.923*** (df = 1; 686)
Note:	*p<0.1; **p<0.05; ***p<0.01

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

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

Wooldridge collected these data from two sources, the 1992 *Statistical Abstract of the United States* (Tables 1009, 1012) and *A Digest of State Alcohol-Highway Safety Related Legislation*, 1985 and 1990, published by the U.S. National Highway Traffic Safety Administration.

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

```
data("traffic1")
?traffic1
```

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

```
stargazer(DD_model, single.row = TRUE, header = FALSE)
```

Table 15:

	Dependent variable:
	cdthrte
copen	−0.420** (0.206)
cadmn	−0.151 (0.117)
Constant	−0.497*** (0.052)
Observations	51
R^2	0.119
Adjusted R^2	0.082
Residual Std. Error	0.344 (df = 48)
F Statistic	3.231** (df = 2; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

Chapter 14: Advanced Panel Data Methods

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

In this section, we will estimate a linear panel modeg using the `plm` function from the `plm: Linear Models for Panel Data` package. See the bibliography for more information.

```
library(plm)
```

```
## Loading required package: Formula
```

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

```
stargazer(scrap_panel, single.row = TRUE, header = FALSE)
```

Table 16:

	Dependent variable:
	lscrap
d88	−0.080 (0.109)
d89	−0.247* (0.133)
grant	−0.252* (0.151)
grant_1	−0.422** (0.210)
Observations	162
R^2	0.201
Adjusted R^2	−0.237
F Statistic	6.543*** (df = 4; 104)
Note:	*p<0.1; **p<0.05; ***p<0.01

Chapter 15: Instrumental Variables Estimation and Two Stage Least Squares

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

T.A. Mroz (1987), *The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions*, Econometrica 55, 765-799. Professor Ernst R. Berndt, of MIT, kindly provided the data, which he obtained from Professor Mroz.

$$log(wage) = \beta_0 + \beta_1 educ + \mu$$

```
data("mroz")
?mroz
```

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

$$\widehat{educ} = \beta_0 + \beta_1 fatheduc$$

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.

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

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

```
library("AER")
wage_educ_IV <- ivreg(lwage ~ educ | fatheduc, data = mroz)
```

```
stargazer(wage_educ_model, fatheduc_model, wage_educ_IV, single.row = TRUE,
    header = FALSE)
```

Table 17:

	Dependent variable:		
	lwage	educ	lwage
	OLS	*OLS*	*instrumental variable*
	(1)	(2)	(3)
educ	0.109*** (0.014)		0.059* (0.035)
fatheduc		0.269*** (0.029)	
Constant	−0.185 (0.185)	10.237*** (0.276)	0.441 (0.446)
Observations	428	428	428
R^2	0.118	0.173	0.093
Adjusted R^2	0.116	0.171	0.091
Residual Std. Error (df = 426)	0.680	2.081	0.689
F Statistic (df = 1; 426)	56.929***	88.841***	
Note:			*p<0.1; **p<0.05; ***p<0.01

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

Data from M. Blackburn and D. Neumark (1992), *Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials*, Quarterly Journal of Economics 107, 1421-1436. Professor Neumark kindly provided the data, of which Wooldridge uses the data for 1980.

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

```
data("wage2")
?wage2
```

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

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

Again, estimate the model using the `ivreg` function in the `AER (Applied Econometrics with R)` package.

```
library("AER")
educ_sibs_IV <- ivreg(lwage ~ educ | sibs, data = wage2)
```

```
stargazer(educ_sibs_model, educ_sibs_IV, wage_educ_IV, single.row = TRUE, header = FALSE)
```

Table 18:

	Dependent variable:		
	educ	lwage	
	OLS	*instrumental variable*	
	(1)	(2)	(3)
sibs	-0.228^{***} (0.030)		
educ		0.122^{***} (0.026)	0.059^{*} (0.035)
Constant	14.139^{***} (0.113)	5.130^{***} (0.355)	0.441 (0.446)
Observations	935	935	428
R^2	0.057	−0.009	0.093
Adjusted R^2	0.056	−0.010	0.091
Residual Std. Error	2.134 (df = 933)	0.423 (df = 933)	0.689 (df = 426)
F Statistic	56.667^{***} (df = 1; 933)		
Note:			*p<0.1; **p<0.05; ***p<0.01

`Example 15.5`: Return to Education for Working Women

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

Use the `ivreg` function in the `AER (Applied Econometrics with R)` package to estimate.

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

Table 19:

	Dependent variable:
	lwage
educ	0.061* (0.031)
exper	0.044*** (0.013)
expersq	−0.001** (0.0004)
Constant	0.048 (0.400)
Observations	428
R^2	0.136
Adjusted R^2	0.130
Residual Std. Error	0.675 (df = 424)
Note:	*p<0.1; **p<0.05; ***p<0.01

Chapter 16: Simultaneous Equations Models

`Example 16.4:` INFLATION AND OPENNESS

Data from D. Romer (1993), *Openness and Inflation: Theory and Evidence*, Quarterly Journal of Economics 108, 869-903. The data are included in the article.

$$inf = \beta_{10} + \alpha_1 open + \beta_{11} log(pcinc) + \mu_1$$
$$open = \beta_{20} + \alpha_2 inf + \beta_{21} log(pcinc) + \beta_{22} log(land) + \mu_2$$

`Example 16.6:` INFLATION AND OPENNESS

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

```
data("openness")
?openness
```

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

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

Use the `ivreg` function in the `AER (Applied Econometrics with R)` package to estimate.

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

```
stargazer(open_model, inflation_IV, single.row = TRUE, header = FALSE)
```

Table 20:

	Dependent variable:	
	open	inf
	OLS	*instrumental variable*
	(1)	(2)
open		-0.337^{**} (0.144)
lpcinc	0.546 (1.493)	0.376 (2.015)
lland	-7.567^{***} (0.814)	
Constant	117.085^{***} (15.848)	26.899^{*} (15.401)
Observations	114	114
R^2	0.449	0.031
Adjusted R^2	0.439	0.013
Residual Std. Error (df = 111)	17.796	23.836
F Statistic	45.165^{***} (df = 2; 111)	
Note:		*p<0.1; **p<0.05; ***p<0.01

Chapter 17: Limited Dependent Variable Models and Sample Selection Corrections

`Example 17.3`: POISSON REGRESSION FOR NUMBER OF ARRESTS

```
data("crime1")
```

Sometimes, when estimating a model with many variables, defining a `model` object containing the formula makes for much cleaner code.

```
formula <- (narr86 ~ pcnv + avgsen + tottime + ptime86 + qemp86 + inc86 + black +
    hispan + born60)
```

Then, pass the `formula` object into the `lm` function, and define the `data` argument as usual.

```
econ_crime_model <- lm(formula, data = crime1)
```

To estimate the `poisson` regression, use the general linear model function `glm` and define the `family` argument as `poisson`.

```
econ_crim_poisson <- glm(formula, data = crime1, family = poisson)
```

Use the `stargazer` package to easily compare diagnostic tables of both models.

```
stargazer(econ_crime_model, econ_crim_poisson, single.row = TRUE, header = FALSE)
```

Table 21:

	Dependent variable:	
	narr86	
	OLS	*Poisson*
	(1)	(2)
pcnv	−0.132*** (0.040)	−0.402*** (0.085)
avgsen	−0.011 (0.012)	−0.024 (0.020)
tottime	0.012 (0.009)	0.024* (0.015)
ptime86	−0.041*** (0.009)	−0.099*** (0.021)
qemp86	−0.051*** (0.014)	−0.038 (0.029)
inc86	−0.001*** (0.0003)	−0.008*** (0.001)
black	0.327*** (0.045)	0.661*** (0.074)
hispan	0.194*** (0.040)	0.500*** (0.074)
born60	−0.022 (0.033)	−0.051 (0.064)
Constant	0.577*** (0.038)	−0.600*** (0.067)
Observations	2,725	2,725
R^2	0.072	
Adjusted R^2	0.069	
Log Likelihood		−2,248.761
Akaike Inf. Crit.		4,517.522
Residual Std. Error	0.829 (df = 2715)	
F Statistic	23.572*** (df = 9; 2715)	
Note:		*p<0.1; **p<0.05; ***p<0.01

Chapter 18: Advanced Time Series Topics

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

Data from *Economic Report of the President, 2004*, Tables B-42 and B-64.

```
data("phillips")
?phillips
```

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

Estimate the linear model in the usual way and note the use of the `subset` argument to define data equal to and before the year 1996.

```
unem_AR1 <- lm(unem ~ unem_1, data = phillips, subset = (year <= 1996))
```

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

```
unem_inf_VAR1 <- lm(unem ~ unem_1 + inf_1, data = phillips, subset = (year <= 1996))
```

Table 22:

	Dependent variable:	
	unem	
	(1)	(2)
unem_1	0.732*** (0.097)	0.647*** (0.084)
inf_1		0.184*** (0.041)
Constant	1.572*** (0.577)	1.304** (0.490)
Observations	48	48
R^2	0.554	0.691
Adjusted R^2	0.544	0.677
Residual Std. Error	1.049 (df = 46)	0.883 (df = 45)
F Statistic	57.132*** (df = 1; 46)	50.219*** (df = 2; 45)
Note:		*p<0.1; **p<0.05; ***p<0.01

Bibliography

Yves Croissant, Giovanni Millo (2008). *Panel Data Econometrics in R: The plm Package.* Journal of Statistical Software 27(2). URL www.jstatsoft.org/v27/i02/.

Marek Hlavac (2015). *stargazer: Well-Formatted Regression and Summary Statistics Tables.* R package version 5.2. https://CRAN.R-project.org/package=stargazer

Christian Kleiber and Achim Zeileis (2008). *Applied Econometrics with R.* New York: Springer-Verlag. ISBN 978-0-387-77316-2. URL https://CRAN.R-project.org/package=AER

Franz Mohr (2015). *prais: Prais-Winsten Estimation Procedure for AR(1) Serial Correlation.* R package version 0.1.1. https://CRAN.R-project.org/package=prais

R Core Team (2017). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Hadley Wickham and Winston Chang (2016). *devtools: Tools to Make Developing R Packages Easier.* R package version 1.12.0. https://CRAN.R-project.org/package=devtools

Hadley Wickham. *testthat: Get Started with Testing.* R package version 1.0.2. https://CRAN.R-project.org/package=testthat

Jeffrey M. Wooldridge (2013). *Introductory Econometrics: A Modern Approach.* Mason, Ohio :South-Western Cengage Learning.

Yihui Xie (2017). *knitr: A General-Purpose Package for Dynamic Report Generation in R.* R package version 1.16. https://CRAN.R-project.org/package=knitr

Appendix

Using R for Introductory Econometrics

This is an excellent open source complimentary text to "Introductory Econometrics" by Jeffrey M. Wooldridge and should be your number one resource. This excerpt from the book's website:

> This book introduces the popular, powerful and free programming language and software package R with a focus on the implementation of standard tools and methods used in econometrics. Unlike other books on similar topics, it does not attempt to provide a self-contained discussion of econometric models and methods. Instead, it builds on the excellent and popular textbook "Introductory Econometrics" by Jeffrey M. Wooldridge.

Hess, Florian. *Using R for Introductory Econometrics.* ISBN: 978-1-523-28513-6, CreateSpace Independent Publishing Platform, 2016, Dusseldorf, Germany.

url: https://urfie.net.

Applied Econometrics with R

From the publisher's website:

> This is the first book on applied econometrics using the R system for statistical computing and graphics. It presents hands-on examples for a wide range of econometric models, from classical linear regression models for cross-section, time series or panel data and the common non-linear models of microeconometrics such as logit, probit and tobit models, to recent semiparametric extensions. In addition, it provides a chapter on programming, including simulations, optimization, and an introduction to R tools enabling reproducible econometric research. An R package accompanying this book, AER, is available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=AER.

Kleiber, Christian and Achim Zeileis. *Applied Econometrics with R.* ISBN 978-0-387-77316-2, Springer-Verlag, 2008, New York. http://www.springer.com/us/book/9780387773162