woolRidge-vignette

Justin M Shea

2017-06-28

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

```
library(woolRidge)
```

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.

```
log_wage_model$coefficients
```

```
## (Intercept)        educ
##  0.58377267  0.08274437
```

```
summary(log_wage_model)$r.squared
```

```
## [1] 0.1858065
```

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.

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

Next, 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 log of
annual sales, and the log of the number of the employees, against the log of the scrape rate.

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

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

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

```
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_μ, 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:"

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

```
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)."

```
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.

```
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 $hrprice2$ 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
```

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

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 $log(\widehat{price})$ 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.

```
data(GPA1)
```

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

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

```
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.

```
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.

```
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.

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

FIGURE 9.1: Scatterplot of R&D intensity against firm sales
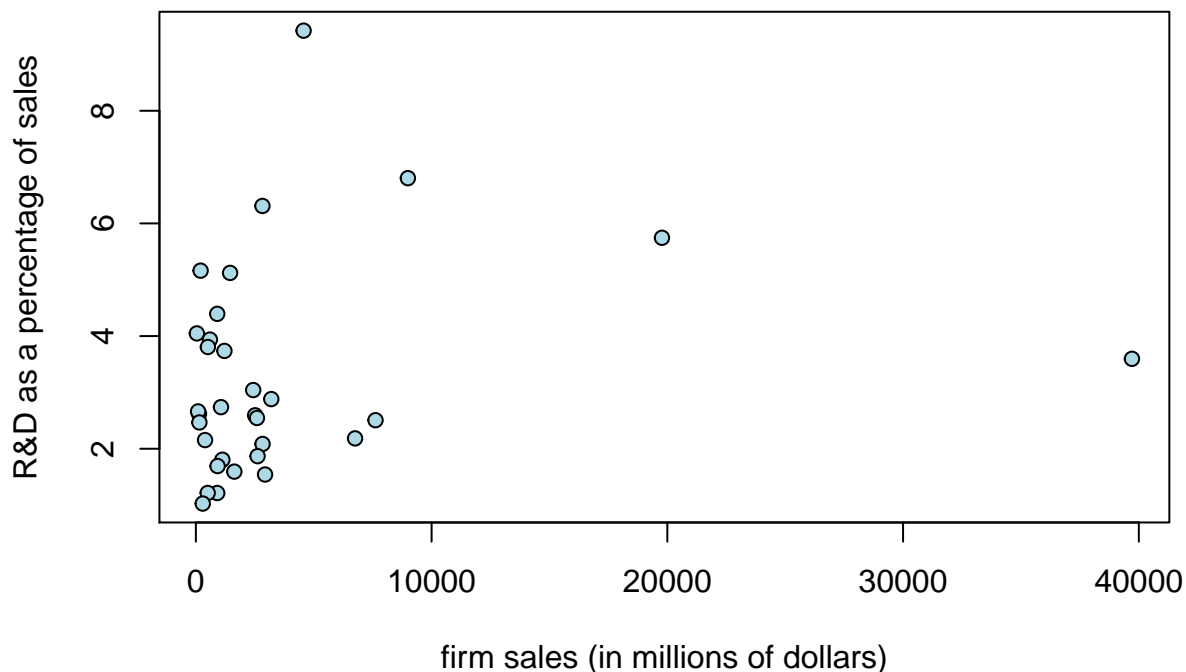
"In terms of sales, this firm is over twice as large as every other firm, so it might be a good idea to estimate the model without it. When we do this, we obtain:"

```
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:"

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

```
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.

```
# 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."

```
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:"

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

"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:"

```
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):"

```
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 R‑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 $n = 689$.

```
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 squaring the residuals contained in the **return_AR** model object and store the results in the variable named **return_mu2**. Then regress the **return_1** data also contained in the **return_AR** objects model matrix. Note the **data** arguement in the call to **lm** below.

```
return_mu2 <- return_AR$residuals^2

mu_squared_hat <- lm(return_mu2 ~ return_1, data = return_AR$model)

summary(mu_squared_hat)
```

```
##
## Call:
## lm(formula = return_mu2 ~ 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."