PSC 508

Jim Battista

Univ. at Buffalo, SUNY

Multiple regression
Multiple regression

“Easy” to find simple regression line $y = a + bx$

$yield = a + b(fertilizer)$
  - But what about sun? Rain?

$attitude = a + b(income)$

These other variables are “confounding variables”

A simple regression might give wrong results because it fails to take them into account
Say we’re looking at determinants of car mileage (in 1978) and find the following:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs =</th>
<th>74</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>536.541807</td>
<td>1</td>
<td>536.541807</td>
<td>F( 1, 72) =</td>
<td>20.26</td>
</tr>
<tr>
<td>Model</td>
<td>1906.91765</td>
<td>72</td>
<td>26.4849674</td>
<td>Prob &gt; F =</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>2443.45946</td>
<td>73</td>
<td>33.4720474</td>
<td>R-squared =</td>
<td>0.2196</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>73</td>
<td>33.4720474</td>
<td>Adj R-squared =</td>
<td>0.2087</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE =</td>
<td>5.1464</td>
</tr>
</tbody>
</table>

| mpg       | Coef.         | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|-----------|---------------|-----------|------|------|---------------------|
| price     | -.0009192     | .0002042  | -4.50| 0.000| -.0013263 -.0005121 |
| _cons     | 26.96417      | 1.393952  | 19.34| 0.000| 24.18538 29.74297  |

- What does this say about the effect of price on mileage?
- But is that the only thing that affects mileage? What else might?
Lots of things affect mpg
- Horsepower
- WEIGHT!!!

...and heavy cars tend to cost more

If we want to isolate the effect of price, we need to control for weight

How? This is very simple
- Add it as another independent variable to the regression
Okay, so we want to control for weight, and we find:

```
. reg mpg price weight
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 74</th>
<th>F( 2, 71) = 66.85</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1595.93249</td>
<td>2</td>
<td>797.966246</td>
<td>Prob &gt; F = 0.0000</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>847.526967</td>
<td>71</td>
<td>11.9369995</td>
<td>R-squared = 0.6531</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2443.45946</td>
<td>73</td>
<td>33.4720474</td>
<td>Adj R-squared = 0.6434</td>
<td></td>
</tr>
</tbody>
</table>

| mpg | Coef.  | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|-----|--------|-----------|-------|-----|------------------|
| price | -0.0000935 | 0.001627 | -0.57 | 0.567 | -0.000418 to 0.0002309 |
| weight | -0.0058175 | 0.0006175 | -9.42 | 0.000 | -0.0070489 to -0.0045862 |
| _cons  | 39.439666 | 1.621563 | 24.32 | 0.000 | 36.20635 to 42.67296 |

- What does this say about the effect of price on mileage now?
If we’d done it in R

```r
> newdata<-read.table('c:/temp/loadme.csv',sep="","header=T)
> attach(newdata)
> example<-lm(mpg~price+weight)
> summary(example)

Call:
  lm(formula = mpg ~ price + weight)

Residuals:
   Min     1Q Median     3Q    Max
-6.8678 -1.8560 -0.5006  0.8847 13.9328

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.944e+01  1.622e+00  24.322  < 2e-16 ***
price                 -9.351e-05  1.627e-04  -0.575   0.567
weight               -5.818e-03  6.175e-04  -9.421  3.94e-14 ***
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .  0.1  1

Residual standard error: 3.455 on 71 degrees of freedom
Multiple R-squared: 0.6531, Adjusted R-squared: 0.6434
F-statistic: 66.85 on 2 and 71 DF,  p-value: < 2.2e-16
```
Interpreting regression output

- Any competent package will give you:
  - “Coef” or “Coefficients”
  - Standard errors
  - Hypothesis testing w/ coefficients
    - For any $b$, $t = b/SE$
    - Degrees of freedom = $n - k - 1$ where $k$ is number of IVs
      NOT INCLUDING constant/intercept
Any competent package will give you

- $R^2$
  - Cheap, sloppy interpretation: percentage of variation in $Y$ that you’ve explained
  - What’s really low or high depends on topic
- F statistic, $p > F$
  - Tests hypothesis that all coefficients are zero
  - $H_0$ : All coefficients are zero
  - $H_A$ : Not all coefficients are zero
  - Can also use on subsets of variables
Datasets will typically have many variables available
Which should you include?
First: variables implied by your theory
Second: variables “required” by existing literature on the topic
That’s it
Don’t include extraneous crap to increase $R^2$
BUT including irrelevant variable better than excluding relevant one (inefficiency versus bias)
Important things to remember about data

- Always keep an unaltered copy of a downloaded or original dataset
  - Ideally as a csv or similar plaintext format
- If recoding a variable, recode to a new variable, not on top of itself
  - `gen incomeK=income/1000`
  - Not `replace income=income/1000`
- Save versions of dataset frequently as new files with descriptive filenames
- Best practice: work with scripts and logfiles
- Second best practice: work with command line and copy/paste sessions
- Bad practice: point and click UNLESS it saves command stream into output
I don’t care what you use
I can offer some help in Stata, less help with R, no help at all in other packages
Stata: costs money, extensible, good community, relatively simple command line operation, decent with data
R: free, extensible, better community, more complex operation for simple tasks, more of a pain for working with data
Extremely basic Stata

- **Getting data in**
  - Stata .dta file: just double click file or click “open” button
  - CSV: `insheet filename, comma`
  - Alternate CSV:
    1. Load CSV into spreadsheet (double click)
    2. `ctrl-a ctrl-c` to copy all
    3. Open “Data editor (edit)” in Stata
    4. Paste data

- **Basic regression**
  - `regress dv iv1 iv2 iv3`
Getting data in

CSV

1. `dataobject<-read.table(filename, sep="","header=T/F)`
2. `attach(dataobject)`
3. `attach()` not required but simplifies things unless using multiple datasets simultaneously

Other formats

- Typically use `library(foreign)` and then appropriate commands
- `read.dta`, `read.spss`, etc (see help files)

Basic regression

1. `outputobject<-lm(dv~iv1+iv2+iv3)`
2. `summary(outputobject)`