Linear Modelling in Stata
Session 6: Further Topics in Linear Modelling

Mark Lunt
Centre for Epidemiology Versus Arthritis
University of Manchester

24/11/2020
This Week

- Categorical Variables
  - Comparing outcome between groups
  - Comparing slopes between groups (Interactions)

- Confounding

- Variable Selection

- Other considerations
  - Polynomial Regression
  - Transformation
  - Regression through the origin
None of the linear model assumptions mention the distribution of $x$.

Can use $x$-variables with any distribution

This enables us to compare different groups
Dichotomous Variable

Let \( x = 0 \) in group A and \( x = 1 \) in group B.

Linear model equation is \( \hat{Y} = \beta_0 + \beta_1 x \)

In group A, \( x = 0 \) so \( \hat{Y} = \beta_0 \)

In group B, \( x = 1 \) so \( \hat{Y} = \beta_0 + \beta_1 \)

Hence the coefficient of \( x \) gives the mean difference between the two groups.
Dichotomous Variable Example

- $x$ takes values 0 or 1
- $Y$ is normally distributed with variance 1, and mean 3 if $x = 0$ and 4 if $x = 1$.
- We wish to test if there difference in the mean value of $Y$ between the groups with $x = 0$ and $x = 1$. 

### Dichotomous Variable: Stata output

```stata
. regress Y x
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>9.86319435</td>
<td>1</td>
<td>9.86319435</td>
<td>F( 1, 38) = 10.97</td>
</tr>
<tr>
<td>Residual</td>
<td>34.1679607</td>
<td>38</td>
<td>.89915686</td>
<td>Prob &gt; F = 0.0020</td>
</tr>
<tr>
<td>Total</td>
<td>44.031155</td>
<td>39</td>
<td>1.12900398</td>
<td>R-squared = 0.2240</td>
</tr>
</tbody>
</table>

|                         | Coef. | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|-------------------------|-------|-----------|-------|-------|---------------------|
| x                       | .9931362 | .2998594 | 3.31  | 0.002 | .3861025 - 1.60017  |
| _cons                   | 3.0325 | .2120326  | 14.30 | 0.000 | 2.603262 - 3.461737 |
Dichotomous Variables and the T-Test

- Differences in mean between two groups usually tested for with t-test.
- Linear model results are *exactly* the same.
- Linear model assumptions are *exactly* the same.
  - Normal distribution in each group
  - Same variance in each group
- A t-test is a special case of a linear model.
- Linear model is far more versatile (can adjust for other variables).
. ttest Y, by(x)

Two-sample t test with equal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20</td>
<td>3.0325</td>
<td>0.2467866</td>
<td>1.103663</td>
<td>2.515969 3.54903</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>4.025636</td>
<td>0.1703292</td>
<td>0.7617355</td>
<td>3.669133 4.382139</td>
</tr>
<tr>
<td>combined</td>
<td>40</td>
<td>3.529068</td>
<td>0.1680033</td>
<td>1.062546</td>
<td>3.189249 3.868886</td>
</tr>
<tr>
<td>diff</td>
<td>40</td>
<td>-0.9931362</td>
<td>0.2998594</td>
<td></td>
<td>-1.60017 -0.3861025</td>
</tr>
</tbody>
</table>

```
t = -3.3120
degrees of freedom = 38
```

Ha: diff < 0  Ha: diff != 0  Ha: diff > 0
Pr(T < t) = 0.0010  Pr(|T| > |t|) = 0.0020  Pr(T > t) = 0.9990
Categorical Variable with Several Categories

- What can we do if there are more than two categories?
- Cannot use \( x = 0, 1, 2, \ldots \).
- Instead we use “dummy” or “indicator” variables.
- If there are \( k \) categories, we need \( k - 1 \) indicators.
Three Groups: Example

<table>
<thead>
<tr>
<th>Group</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$Y$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

- $\beta_0 = \hat{Y}$ in group A
- $\beta_1 = \text{difference between } \hat{Y} \text{ in group A and } \hat{Y} \text{ in group B}$
- $\beta_2 = \text{difference between } \hat{Y} \text{ in group A and } \hat{Y} \text{ in group C}$
### Three Groups: Stata Output

```
. regress Y x1 x2

Source | SS        | df | MS
--------+-----------+----+------------------
Model   | 37.1174969| 2  | 18.5587485
Residual| 62.8970695| 57 | 1.10345736
--------+-----------+----+------------------
Total   | 100.014566| 59 | 1.69516214

Number of obs = 60
F( 2, 57) = 16.82
Prob > F = 0.0000
R-squared = 0.3711
Adj R-squared = 0.3491
Root MSE = 1.0505

Y | Coef.     | Std. Err. | t     | P>|t| | [95% Conf. Interval]
----+-----------+------------+-------+-------+-------------------------
  x1 | 1.924713  | .3321833  | 5.79  | 0.000 | 1.259528 2.589899
  x2 | 1.035985  | .3321833  | 3.12  | 0.003 | .3707994 1.701171
_cons| 3.075665  | .2348891  | 13.09 | 0.000 | 2.605308 3.546022
```
In the previous example, groups B and C both compared to group A.

Can we compare groups B and C as well?

In group B, \( \hat{Y} = \beta_0 + \beta_1 \)

In group C, \( \hat{Y} = \beta_0 + \beta_2 \)

Hence difference between groups is \( \beta_1 - \beta_2 \)

Can use \texttt{lincom} to obtain this difference, and test its significance.
The `lincom` Command

- `lincom` is short for linear combination.
- It can be used to calculate linear combinations of the parameters of a linear model.
- Linear combination = \( a_j \beta_j + a_k \beta_k + \ldots \)
- Can be used to find differences between groups (Difference between Group B and Group C = \( \beta_1 - \beta_2 \))
- Can be used to find mean values in groups (Mean value in group B = \( \beta_0 + \beta_1 \)).
Stata Output from \texttt{lincom}

\begin{verbatim}
. lincom x1 - x2
( 1)  x1 - x2 = 0

|          Y | Coef.    | Std. Err. |      t |   P>|t| |   [95% Conf. Interval] |
|-----------|----------|-----------|--------|--------|------------------------|
| (1)       | 0.8887284| 0.3321833 | 2.68   | 0.010  | 0.2235428 - 1.553914   |

. lincom _cons + x1
( 1)  x1 + _cons = 0

|          Y | Coef.    | Std. Err. |      t |   P>|t| |   [95% Conf. Interval] |
|-----------|----------|-----------|--------|--------|------------------------|
| (1)       | 5.0003781| 0.2348891 | 21.29  | 0.000  | 4.5300207 - 5.4707362  |
\end{verbatim}
Generating dummy variables can be tedious and error-prone.

Stata can do it for you.

Identify categorical variables by adding “i.” to the start of their name.

For example, suppose that the variable `group` contains the values “1”, “2” and “3” for the three groups in the previous example.
### Stata Output with a Factor Variable

```stata
.regress Y i.group
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>37.1174969</td>
<td>2</td>
<td>18.5587485</td>
</tr>
<tr>
<td>Residual</td>
<td>62.8970695</td>
<td>57</td>
<td>1.10345736</td>
</tr>
<tr>
<td>Total</td>
<td>100.014566</td>
<td>59</td>
<td>1.69516214</td>
</tr>
</tbody>
</table>

| Y         | Coef.   | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|-----------|---------|-----------|-----|-----|----------------------|
| group     |         |           |     |     |                      |
| 2         | 1.924713| 0.3321833 | 5.79| 0.000| 1.259528 - 2.589899  |
| 3         | 1.035985| 0.3321833 | 3.12| 0.003| 0.3707994 - 1.701171 |
| _cons     | 3.075665| 0.2348891 | 13.09| 0.000| 2.605308 - 3.546022  |
Using factor variables with `lincom`

```
. lincom 2.group - 3.group
( 1) 2.group - 3.group = 0

               Y |     Coef.  Std. Err.   t    P>|t|     [95% Conf. Interval]
-------------+-----------------------------------------
(1) |    .8887284   .3321833  2.68  0.010     .2235428   1.553914

. lincom _cons + 2.group
( 1) 2.group + _cons = 0

               Y |     Coef.  Std. Err.   t    P>|t|     [95% Conf. Interval]
-------------+-----------------------------------------
(1) |     5.000378   .2348891  21.29  0.000     4.530021   5.470736
```
Differences in mean between more than two groups usually tested for with ANOVA.

Linear model results are *exactly* the same.

Linear model assumptions are *exactly* the same.

ANOVA is a special case of a linear model.

Linear model is far more versatile (can adjust for other variables).
So far, we have only seen either continuous or categorical predictors in a linear model.

No problem to mix both.

E.g. Consider a clinical trial in which the outcome is strongly associated with age.

To test the effect of treatment, need to include both age and treatment in linear model.

Once upon a time, this was called Analysis of Covariance (ANCOVA)
Example Clinical Trial: simulated data
. regress Y treat

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>26.5431819</td>
<td>1</td>
<td>26.5431819</td>
<td>F( 1, 38) = 2.86</td>
</tr>
<tr>
<td>Residual</td>
<td>352.500943</td>
<td>38</td>
<td>9.27634061</td>
<td>Prob &gt; F = 0.0989</td>
</tr>
<tr>
<td>Total</td>
<td>379.044125</td>
<td>39</td>
<td>9.71908013</td>
<td>R-squared = 0.0700</td>
</tr>
</tbody>
</table>

| Y | Coef.  | Std. Err. | t    | P>|t|  | [95% Conf. Interval] |
|---|--------|-----------|------|------|------------------------|
| treat | 1.629208 | .9631376 | 1.69 | 0.099 | -.3205623 - 3.578978 |
| _cons | 4.379165 | .6810411 | 6.43 | 0.000 | 3.00047 - 5.757861 |
Observed and predicted values from linear model ignoring age.
. regress Y treat age

    Source |        SS     df       MS
-------------+----------------------
       Model |  354.096059     2   177.04803
    Residual |  24.9480658   37    .674272049
-------------+----------------------
      Total |  379.044125    39   9.71908013

Number of obs =  40
F(  2,    37) = 262.58
       Prob > F =  0.0000
       R-squared =  0.9342
       Adj R-squared =  0.9306
    Root MSE =  .82114

    Y |     Coef.   Std. Err.     t    P>|t|    [95% Conf. Interval]
-------------+-----------------------------------------------
     treat |    1.238752    .2602711     4.76  0.000     .7113924    1.766111
      age |   -.5186644    .0235322    -22.04  0.000    -.5663453   -.4709836
       _cons |   20.59089    .7581107     27.16  0.000    19.05481    22.12696

- Age explains variation in Y
- This reduces RMSE (estimate of $\sigma$)
- Standard error of coefficient = $\frac{\sigma}{\sqrt{ns_x}}$
Observed and predicted values from linear model including age
Interactions

- In previous example, assumed that the effect of age was the same in treated and untreated groups.
- I.e. regression lines were parallel.
- This may not be the case.
- If the effect of one variable varies accord to the value of another variable, this is called “interaction” between the variables.
- Don’t assume that an effect differs between two groups because it is significant in one, not in the other.
Consider the clinical trial in the previous example

Suppose treatment reverses the effect of aging, so that $\hat{Y}$ is constant in the treated group.

Thus the difference between the treated and untreated groups will increase with increasing age.

Need to fit different intercepts and different slopes in the two groups.
Clinical trial data with predictions assuming equal slopes

![Clinical trial data graph]

- Placebo
- Active Treatment
Regression Equations

- Need to fit the two equations

\[ Y = \begin{cases} 
\beta_{00} + \beta_{10} \times \text{age} + \epsilon & \text{if } \text{treat} = 0 \\
\beta_{01} + \beta_{11} \times \text{age} + \epsilon & \text{if } \text{treat} = 1 
\end{cases} \]

- These are equivalent to the equation

\[ Y = \beta_{00} + \beta_{10} \times \text{age} + (\beta_{01} - \beta_{00}) \times \text{treat} + (\beta_{11} - \beta_{10}) \times \text{age} \times \text{treat} + \epsilon. \]

- I.e. the output from stata can be interpreted as

  - _cons The intercept *in the untreated group* (treat == 0)
  - age The slope with age *in the untreated group*
  - treat The difference in intercept between the treated and untreated groups
  - treat#c.age The difference in slope between the treated and untreated groups
. regress Y i.treat age i.treat#c.age

Source | SS       | df  | MS           | Number of obs = 40
       |----------|-----|--------------|-------------------
Model  | 563.762012 | 3   | 187.920671 | F( 3, 36) = 173.38
Residual | 39.0189256 | 36  | 1.08385904 | Prob > F = 0.0000
Total  | 602.780938  | 39  | 15.4559215 | R-squared = 0.9353
        |           |     |             | Adj R-squared = 0.9299
        |           |     |             | Root MSE = 1.0411

| Y         | Coef.  | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|-----------|--------|-----------|-------|-------|----------------------|
| 1.treat   | -8.226356 | 1.872952 | -4.39 | 0.000 | -12.02488 -4.427833 |
| age       | -0.4866572 | 0.0412295 | -11.80 | 0.000 | -0.5702744 -0.40304  |
| treat#c.age | .4682374 | .0597378 | 7.84  | 0.000 | .3470836 .5893912   |
| _cons     | 19.73531 | 1.309553 | 15.07 | 0.000 | 17.07942 22.39121   |
**lincom** can be used to calculate the slope in the treated group:

```
. lincom age + 1.treat#c.age
```

( 1)  age + 1.treat#c.age = 0

| Y  | Coef.  | Std. Err. | t    | P>|t|   | [95% Conf. Interval] |
|----|--------|-----------|------|-------|----------------------|
| (1)| -0.0184198 | 0.0432288 | -0.43| 0.673 | -0.1060919 - 0.0692523 |

Can also be used to calculate intercept in treated group. However, this is not interesting since

- We are unlikely to be be interested in subjects of age 0
- The youngest subjects in our sample were 20, so we are extrapolating a long way from the data.
Interactions: Predictions from Linear Model

- Categorical Variables
- Confounding
- Variable Selection
- Other Considerations
- Dichotomous Variables
- Multiple Categories
- Categorical & Continuous
- Interactions

Graph showing the relationship between age and treatment response for Placebo and Active Treatment.
Treatment effect at different ages

```
. lincom 1.treat + 20*1.treat#c.age
( 1) 1.treat + 20*1.treat#c.age = 0

+-----------------------------------------------+
|        Y       |     Coef. |   Std. Err. |      t    |    P>|t| |  [95% Conf. Interval] |
|----------------+------------+-------------+----------+--------+------------------------|
| (1)            |   1.13839  |    .727983  |  1.56    |  0.127 |          -.3380261     |  2.61481    |
|----------------+------------+-------------+----------+--------+------------------------|

. lincom 1.treat + 40*1.treat#c.age
( 1) 1.treat + 40*1.treat#c.age = 0

+-----------------------------------------------+
|        Y       |     Coef. |   Std. Err. |      t    |    P>|t| |  [95% Conf. Interval] |
|----------------+------------+-------------+----------+--------+------------------------|
| (1)            |  10.50314  |   .6378479  | 16.47    |  0.000 |          9.209524      |  11.79676   |
```
The `testparm` Command

- Used to test a number of parameters simultaneously
- **Syntax:** `testparm varlist`
- Test $\beta = 0$ for all variables in `varlist`
- Produces a $\chi^2$ test on $k$ degrees of freedom, where there are $k$ variables in `varlist`.
Stata used to use a different syntax for categorical variables
Still works, but new method is preferred
You may still see old syntax in existing do-files

<table>
<thead>
<tr>
<th>Prefix</th>
<th>New syntax</th>
<th>Old Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>none required</td>
<td>Numeric</td>
<td>String or numeric</td>
</tr>
<tr>
<td>#</td>
<td>^</td>
<td>*</td>
</tr>
<tr>
<td>No</td>
<td>help fvvarlist</td>
<td>help xi</td>
</tr>
</tbody>
</table>

Old and new syntax for categorical variables
A linear model shows association.

It does not show *causation*.

Apparent association may be due to a third variable which we haven’t included in model.

Confounding is about causality, and knowledge of the mechanisms are required to decide if a variable is a confounder.
### Confounding Example: Fuel Consumption

```
. regress mpg foreign

Source | SS       df   MS          Number of obs = 74
--------+-----------------------------------------------
Model   | 378.153515 1  378.153515  F(  1,  72) = 13.18
Residual | 2065.30594 72  28.6848048  Prob > F = 0.0005
        |-----------------------------------------------
Total   | 2443.45946 73  33.4720474  R-squared = 0.1548
        |-----------------------------------------------
------------------------------------------------------------------------------
mpg | Coef.  Std. Err.      t    P>|t|     [95% Conf. Interval]
--------+-------------------------------------------------------------
foreign |  4.945804   1.362162    3.631 0.001     2.230384    7.661225
    _cons |  19.82692    .7427186  26.695 0.000    18.34634    21.30751
------------------------------------------------------------------------------
```

---

Categorical Variables

Confounding

Variable Selection

Other Considerations
Confounding Example: Weight and Fuel Consumption

The diagram illustrates the relationship between weight (lbs.) and mileage (mpg) for US and non-US vehicles. The data points show a negative correlation between weight and mileage, indicating that heavier vehicles tend to have lower fuel efficiency. The distinction between US and non-US vehicles is represented by two different symbols, helping to highlight the differences in performance between the two categories.
Confounding Example: Controlling for Weight

```
. regress mpg foreign weight

Source |      SS   df  MS
---------+---------------------
Model | 1619.2877    2 809.643849
Residual | 824.171761   71 11.608053
---------+---------------------
Total | 2443.45946   73 33.4720474

Number of obs =  74
F(  2,    71) = 69.75
Prob > F =  0.0000
R-squared =  0.6627
Adj R-squared =  0.6532
Root MSE =  3.4071

mpg |      Coef.    Std. Err.     t    P>|t|     [95% Conf. Interval]
---------+---------------------------------------------
foreign |  -1.650029   1.075994    -1.533  0.130   -3.7955    .4954421
weight |  -.0065879   .0006371   -10.340  0.000   -.0078583  -.0053175
_cons |       41.6797   2.165547    19.247  0.000     37.36172   45.99768
```
What is Confounding?

- What you see is not what you get
- \( \hat{Y} = \beta_0 + \beta_1 x \)
- Two groups differing in \( x \) by \( \Delta x \) will differ in \( Y \) by \( \beta_1 \Delta x \)
- If we change \( x \) by \( \Delta x \), what happens to \( \hat{Y} \)?
- If it changes by \( \beta_1 \Delta x \), no confounding
- If it changes by anything else, there is confounding
Path Variables vs. Confounders

Weight is a path variable

Weight is a confounder

Foreign → mpg

Weight ➔ mpg

Foreign ➔ mpg

Weight ➔ mpg
Identifying a Confounder

- Is a cause of the outcome irrespective of other predictors
- Is associated with the predictor
- Is not a consequence of the predictor

Weight is associated with mpg

This association does not depend on where the car was designed

But is weight a path variable?
Identifying a Confounder

- Is a cause of the outcome irrespective of other predictors
- Is associated with the predictor
- Is not a consequence of the predictor
- Weight is associated with mpg
- This association does not depend on where the car was designed
- But is weight a path variable?
  - Foreign designers produce smaller cars in order to get better fuel consumption: path variable
Identifying a Confounder

- Is a cause of the outcome irrespective of other predictors
- Is associated with the predictor
- Is not a consequence of the predictor

Weight is associated with mpg
This association does not depend on where the car was designed

But is weight a path variable?
- Foreign designers produce smaller cars in order to get better fuel consumption: path variable
- Size is decided for reasons other than fuel consumption: confounder
In theory, adding a confounder to a regression model is sufficient to adjust for confounding. Then parameters for other variables measure the effects of those variables when confounder does not change. 

This assumes
- Confounder measured perfectly
- Linear association between confounder and outcome

If either of the above are not true, there will be residual confounding.
May wish to reduce the number of predictors used in a linear model.
- Efficiency
- Clearer understanding

Several suggested methods
- Forward selection
- Backward Elimination
- Stepwise
- All subsets
Forward Selection

- Choose a significance level $p_e$ at which variables will enter the model.
- Fit each predictor in turn.
- Choose the most significant predictor.
- If its significance level is less than $p_e$, it is selected.
- Now add each remaining variable to this model in turn, and test the most significant.
- Continue until no further variables are added.
Backward Elimination

- Starts with all predictors in model.
- Removes the least significant.
- Repeat until all remaining predictors significant at chosen level $p_r$.
- Has the advantage that all parameters are adjusted for the effect of all other variables from the start.
- Can give unusual results if there are a large number of correlated variables.
Stepwise Selection

- Combination of preceding methods.
- Variables are added one at a time.
- Each time a variable is added, all the other variables are tested to see if they should be removed.
- Must have $p_r > p_e$, or a variable could be entered and removed on the same step.
All Subsets

- Can try every possible subset of variables.
- Can be hard work: 10 predictors = 1023 subsets.
- Need a criterion to choose best model.
- Adjusted $R^2$ is possible, there are others.
- Not implemented in stata.
Problems with Variable Selection

- **Significance Levels**
  - Hypotheses tested are not independent.
  - Variables chosen for testing not randomly selected.
  - Hence significance levels not equal to nominal levels.
  - Less of a problem in large samples.

- **Differences in Models Selected**
  - Models chosen by different methods may differ.
  - If variables are highly correlated, choice of variable becomes arbitrary.
  - Choice of significance level will affect models.
  - Need common sense.
Variable Selection in Stata

- **Command** `sw regress` is used for forwards, backwards and stepwise selection.
- **Option** `pe` is used to set significance level for inclusion.
- **Option** `pr` is used to set significance level for exclusion.
- **Set** `pe` for forwards, `pr` for backwards and both for stepwise regression.
- The `sw` command does not work with factor variables, so the old `xi:` syntax must be used.
### Variable Selection in Stata: Example 1

**Command:**

```stata
. sw regress weight price hdroom trunk length turn displ gratio, pe(0.05)
```

**Output:**

```
p = 0.0000 < 0.0500 adding length
p = 0.0000 < 0.0500 adding displ
p = 0.0015 < 0.0500 adding price
p = 0.0288 < 0.0500 adding turn

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 74</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>41648450.8</td>
<td>4</td>
<td>10412112.7</td>
<td>F( 4, 69) = 293.75</td>
</tr>
<tr>
<td>Residual</td>
<td>2445727.56</td>
<td>69</td>
<td>35445.3269</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>44094178.4</td>
<td>73</td>
<td>604029.841</td>
<td>Adj R-squared = 0.9413</td>
</tr>
</tbody>
</table>

```

**Coefficients:**

| weight | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|--------|-------|-----------|---|-----|----------------------|
| length | 19.38601 | 2.328203 | 8.327 | 0.000 | (14.74137, 24.03064) |
| displ  | 2.257083 | .467792 | 4.825 | 0.000 | (1.323863, 3.190302) |
| price  | .0332386 | .0087921 | 3.781 | 0.000 | (.0156989, .0507783) |
| turn   | 23.17863 | 10.38128 | 2.233 | 0.029 | (2.468546, 43.88782) |
| _cons  | -2193.042 | 298.0756 | -7.357 | 0.000 | (-2787.687, -1598.398) |
```
Variable Selection in Stata: Example 2

```
. sw regress weight price hdroom trunk length turn displ gratio, pr(0.05)
p = 0.6348 >= 0.0500 removing hdroom
p = 0.5218 >= 0.0500 removing trunk
p = 0.1371 >= 0.0500 removing gratio

Source | SS       df       MS                      Number of obs = 74
---------+------------------------------------------------------------
Model    | 41648450.8  4    10412112.7                 Prob > F      = 0.0000
Residual | 2445727.56  69   35445.3269                 R-squared    = 0.9445
---------+------------------------------------------------------------
Total    | 44094178.4  73   604029.841                 Adj R-squared = 0.9413
---------+------------------------------------------------------------

weight | Coef. Std. Err. t P>|t| [95% Conf. Interval]
---------+--------------------------------------------------
price   | 0.0332386  0.0087921    3.781 0.000   0.0156989   0.0507783
turn    | 23.17863   10.38128    2.233 0.029    2.468546   43.88872
displ   | 2.257083   .467792    4.825 0.000    1.323863   3.190302
length  | 19.38601   2.328203    8.327 0.000   14.74137   24.03064
_cons   | -2193.042  298.0756   -7.357 0.000   -2787.687  -1598.398
```

Categorical Variables
Confounding
Variable Selection
Other Considerations
Polynomial Regression

- If association between $x$ and $Y$ is non-linear, can fit polynomial terms in $x$.
- Keep adding terms until the highest order term is not significant.
- Parameters are meaningless: only entire function has meaning.
- Fractional polynomials and splines can also be used.
If $Y$ is not normal or has non-constant variance, it may be possible to fit a linear model to a transformation of $Y$.

Interpretation becomes more difficult after transformation.

Log transformation has a simple interpretation.

- $\log(Y) = \beta_0 + \beta_1 x$
- when $x$ increases by 1, $\log(Y)$ increases by $\beta_1$,
- $Y$ is multiplied by $e^{\beta_1}$

Transforming $x$ is not normally necessary unless the problem suggests it.
You may know that if \( x = 0, \ y = 0 \).

Stata can force the regression line through the origin with the option `nocons`.

However

\( R^2 \) is calculated differently and cannot be compared to conventional \( R^2 \).

If we have no data near the origin, should not force line through the origin.

May obtain a better fit with a non-zero intercept if there is measurement error.