Lecture 6: Variable Selection

Prof. Sharyn O'Halloran

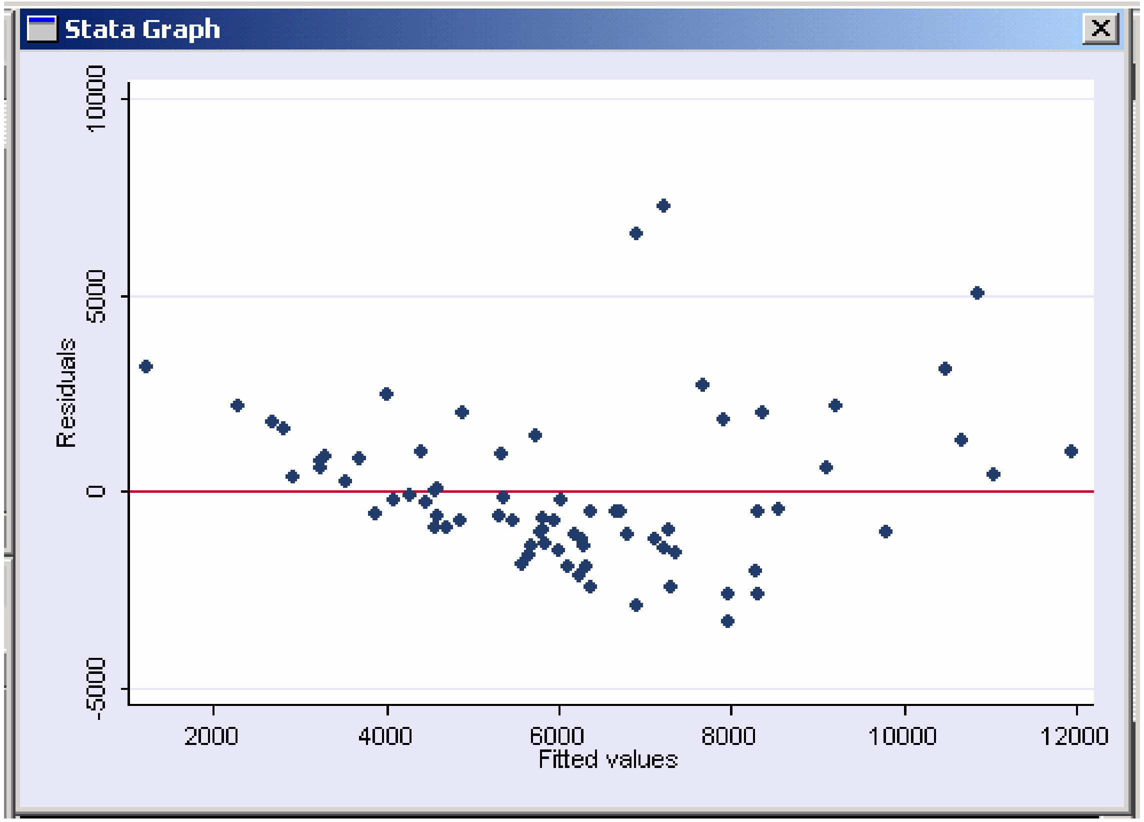
Sustainable Development U9611

Econometrics II

Regression Diagnostics: Review

- After estimating a model, we want to check the entire regression for:
  - Normality of the residuals
  - Omitted and unnecessary variables
  - Heteroskedasticity
- We also want to test individual variables for:
  - Outliers
  - Leverage
  - Influence
  - Collinearity
  - Functional form

Look at Residuals: rvfplot



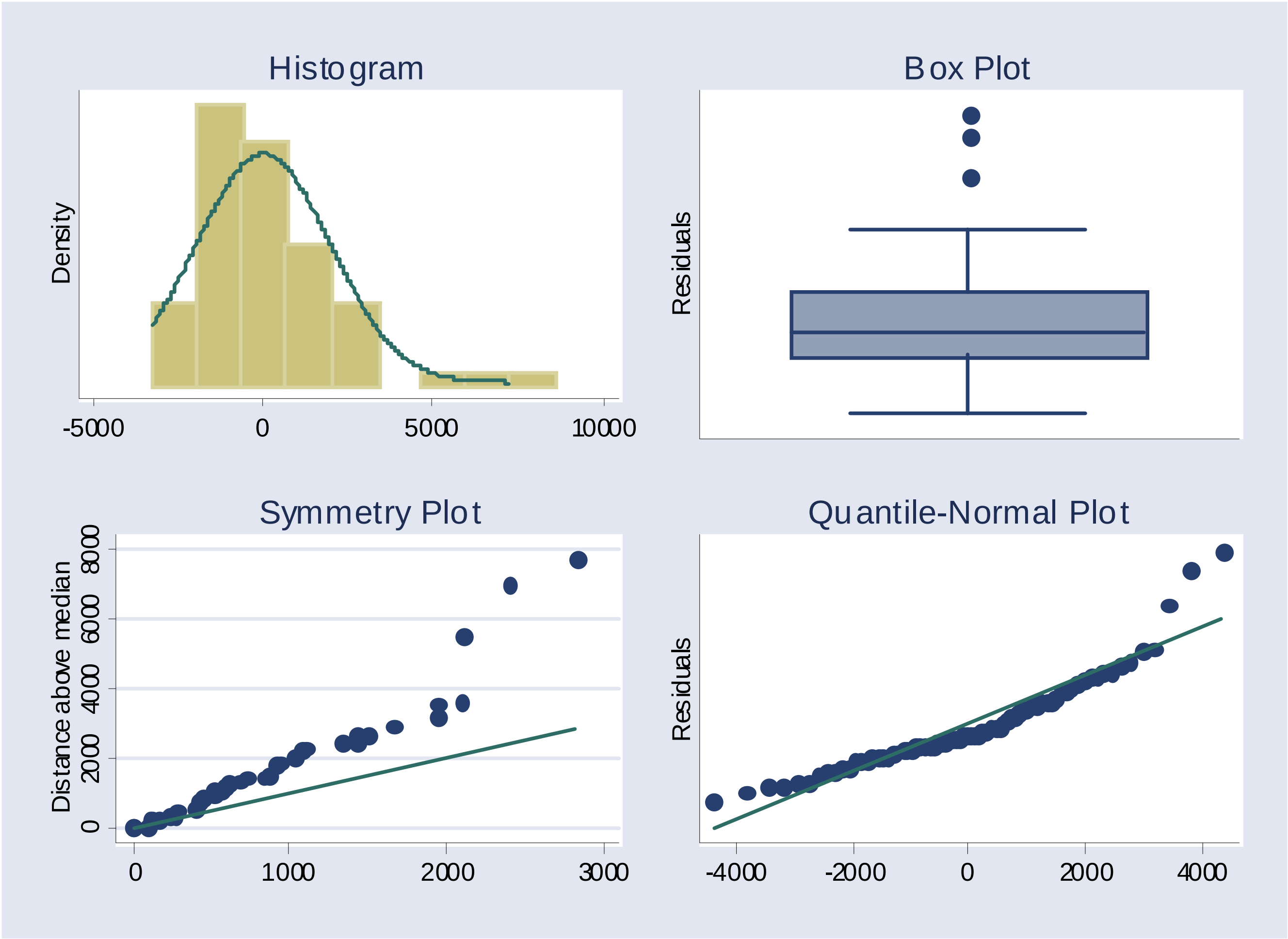

```
reg price weight mpg forXmpg foreign
rvfplot, plotr(lcol(black)) yline(0)
```

First, examine the residuals e_i vs. $\hat{Y}$.

Any pattern in the residuals indicates a problem.

Here, there is an obvious U-shape & heteroskedasticity.

Check Residuals for Normality



Residual plots also indicate non-normality

Stata Commands: imtest

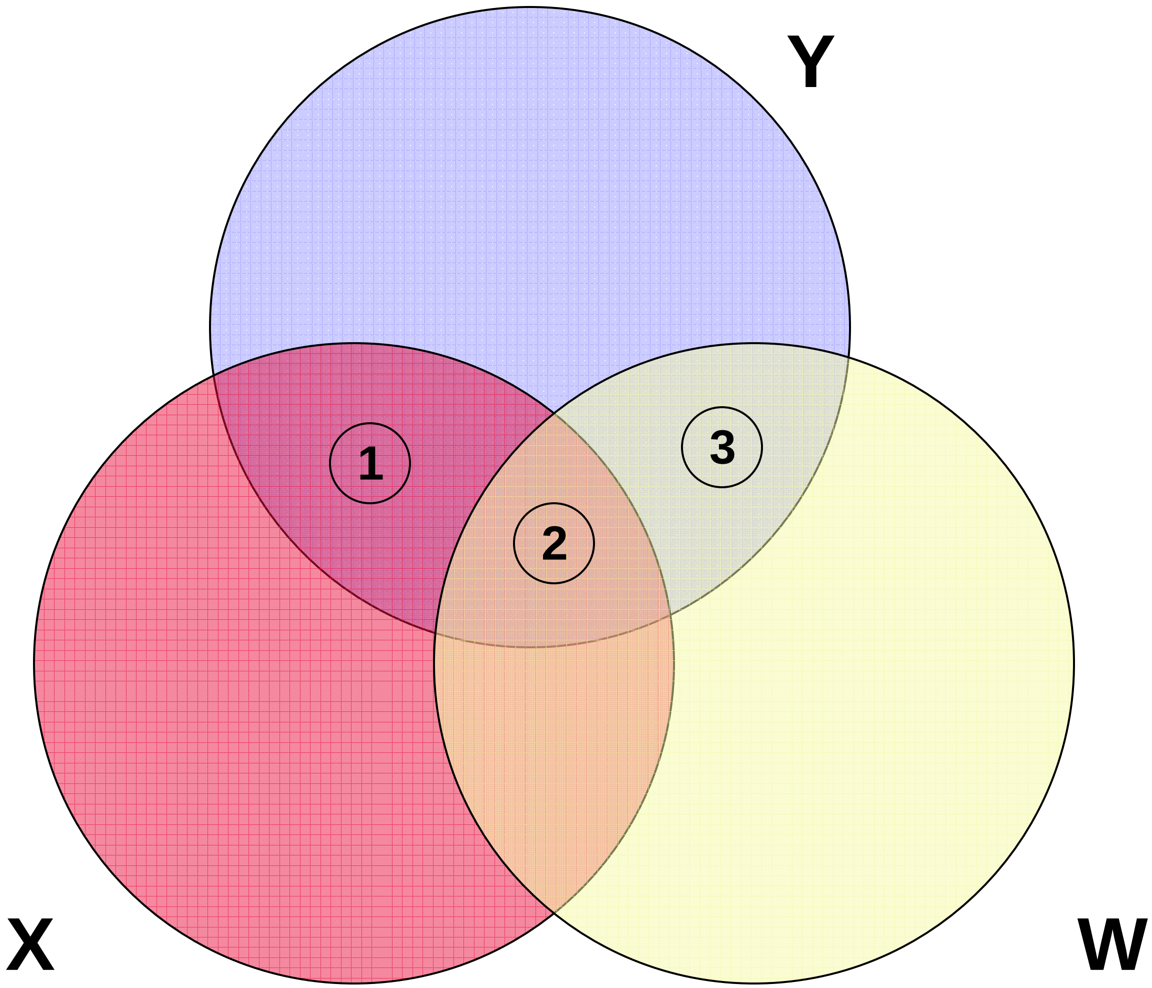
```
. imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      18.86     10    0.0420
            Skewness |      11.69      4    0.0198
            Kurtosis |       2.33      1    0.1273
---------------------+-----------------------------
               Total |      32.87     15    0.0049
---------------------------------------------------
```

The command “imtest” stands for “information matrix.” Here, we see heteroskedasticity and skewness.

Omitted Variables

- Omitted variables are variables that significantly influence Y and so should be in the model, but are excluded.
- Questions:
  - Why are omitted variables a problem?
  - How can we test for them?
  - What are the possible fixes?
- Let's check the Venn diagram…

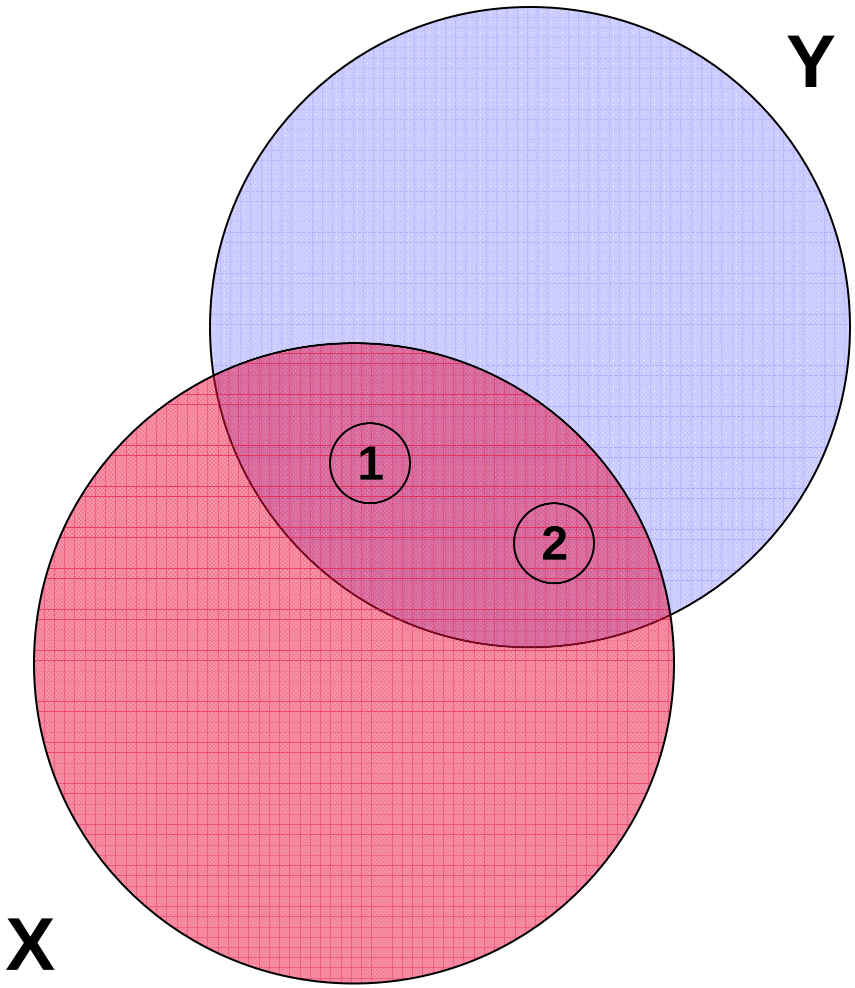
Omitted Variables



Y is determined by X and W, but we omit W from the regression.

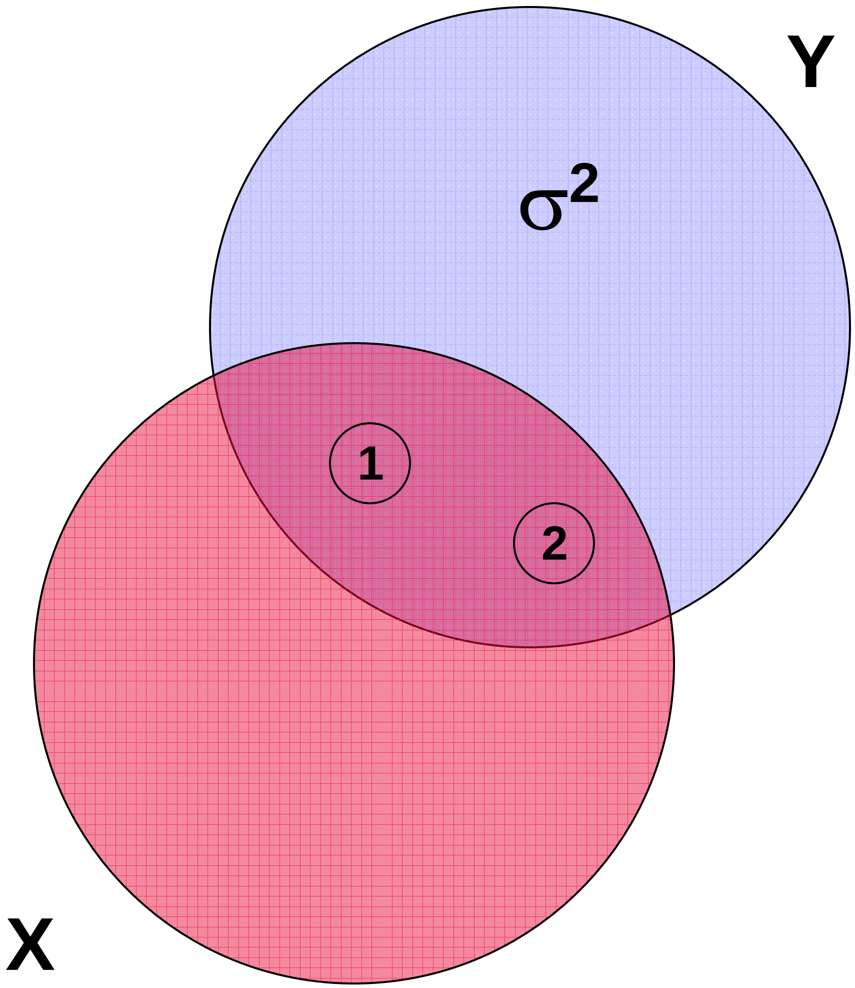
Here, the impact of X on Y is area 1, and the impact of W is area 3.

Omitted Variables



By omitting W, we now estimate the impact of X on Y by areas 1 and 2, rather than just area 1.

Omitted Variables

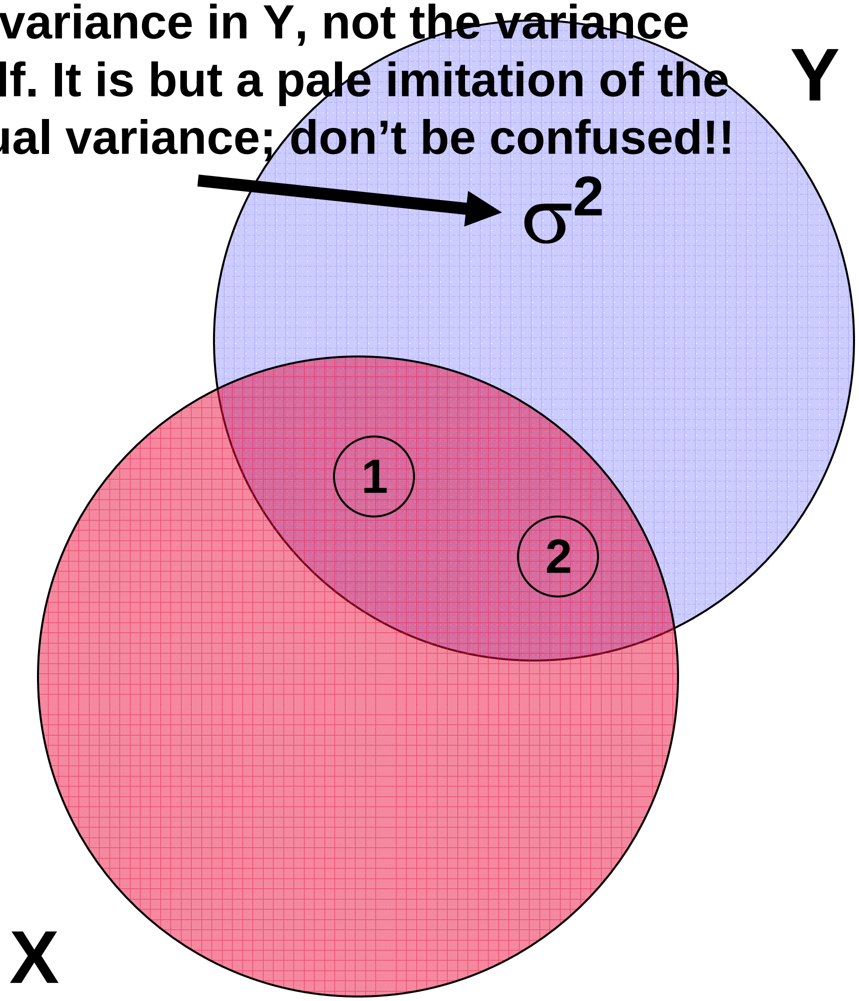


This means that:

1. The estimate of β_1 is biased (since area 2 actually belongs to W as well as X).
2. The variance of β_1 is reduced (since it's estimated by areas 1 and 2).
3. The unexplained variance for Y (σ^2) increases.

Omitted Variables

This is only a representation of the variance in Y, not the variance itself. It is but a pale imitation of the actual variance; don't be confused!!



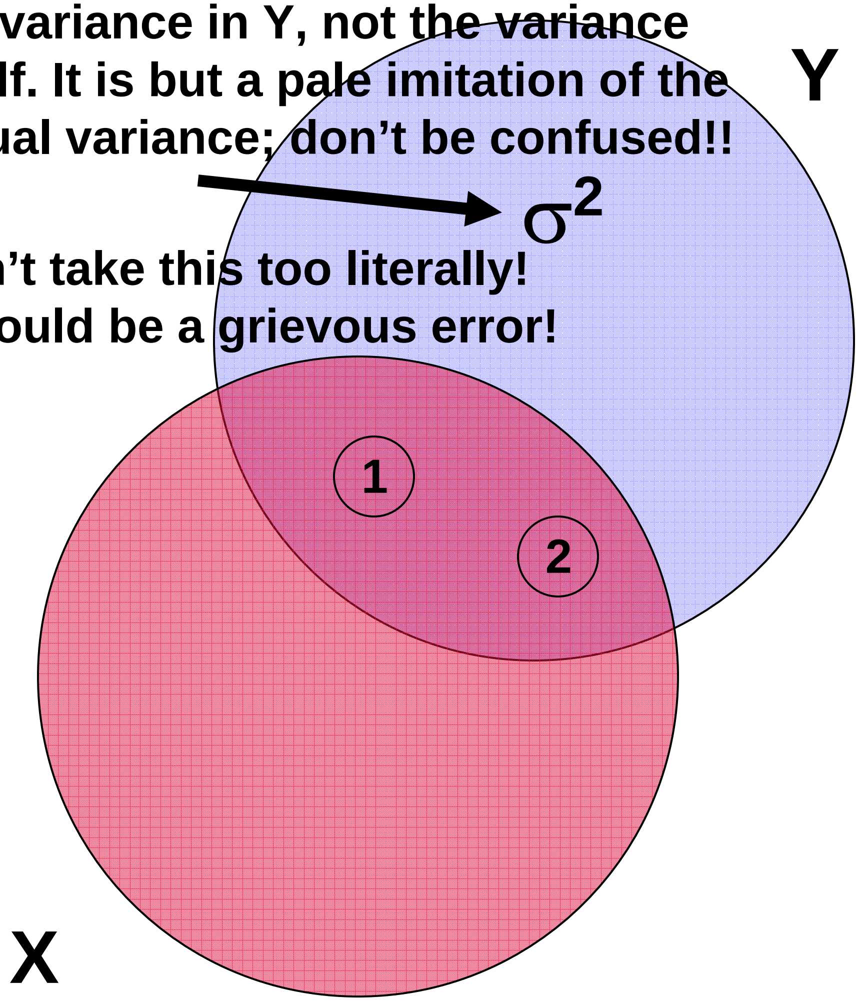
This means that:

1. The estimate of β_1 is biased (since area 2 actually belongs to W as well as X).
2. The variance of β_1 is reduced (since it's estimated by areas 1 and 2).
3. The unexplained variance for Y (σ^2) increases.

Omitted Variables

This is only a representation of the variance in Y, not the variance itself. It is but a pale imitation of the actual variance; don't be confused!!

Don't take this too literally! It would be a grievous error!



This means that:

1. The estimate of β_X is biased (since area 2 actually belongs to W as well as X).
2. The variance of β_X is reduced (since it's estimated by areas 1 and 2).
3. The unexplained variance for Y (σ^2) increases.

Stata Command: ovtest

```
. ovtest

Ramsey RESET test using powers of the fitted
values of price
       Ho:  model has no omitted variables
                 F(3, 66) =      7.77
                 Prob > F =      0.0002

. hettest

Breusch-Pagan / Cook-Weisberg test for
heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of price

         chi2(1)      =     6.50
         Prob > chi2  =   0.0108
```

Stata Command: ovtest

```
. ovtest

Ramsey RESET test using powers of the fitted
values of price
       Ho:  model has no omitted variables
                  F(3, 66) =      7.77
                  Prob > F =      0.0002

. hettest

Breusch-Pagan / Cook-Weisberg test for
heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of price

         chi2(1)      =     6.50
         Prob > chi2  =   0.0108
```

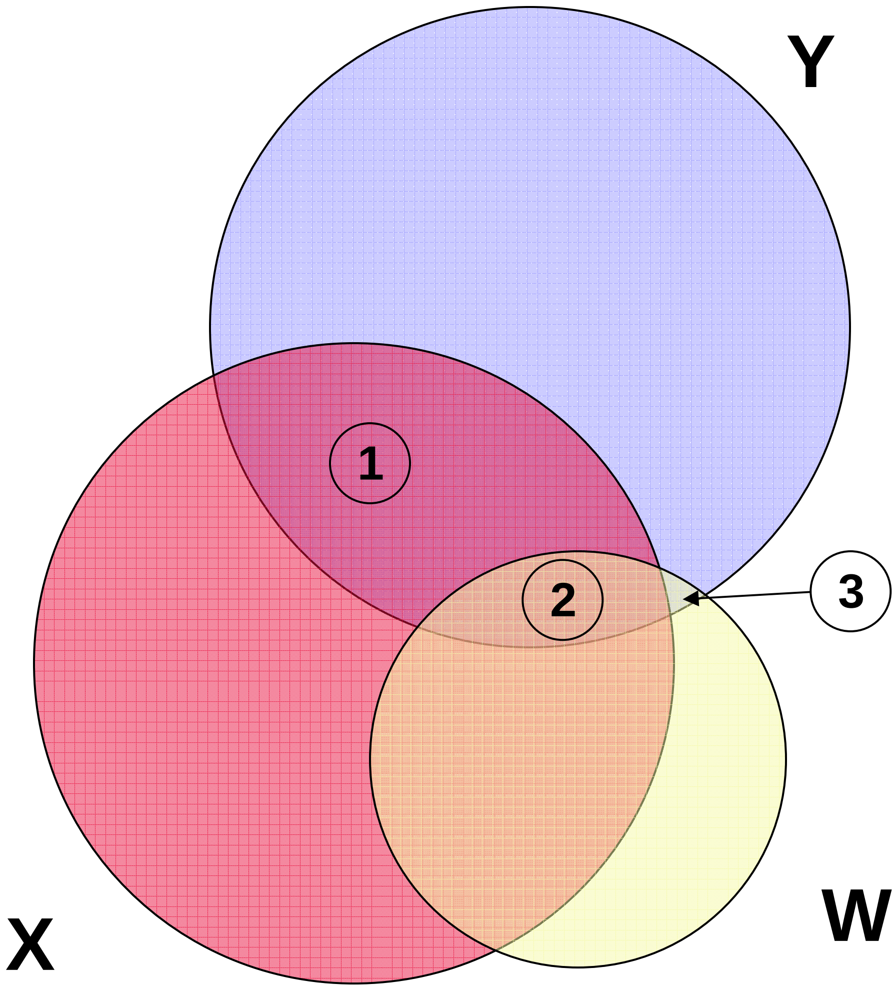
Evidence of omitted vars

Stata Command: ovtest

```
. ovtest

Ramsey RESET test using powers of the fitted
values of price
       Ho:  model has no omitted variables
                  F(3, 66) =      7.77
                  Prob > F =      0.0002

. hettest

Breusch-Pagan / Cook-Weisberg test for
heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of price

         chi2(1)      =     6.50
         Prob > chi2  =   0.0108
```

Evidence of omitted vars and non-constant variance, as before

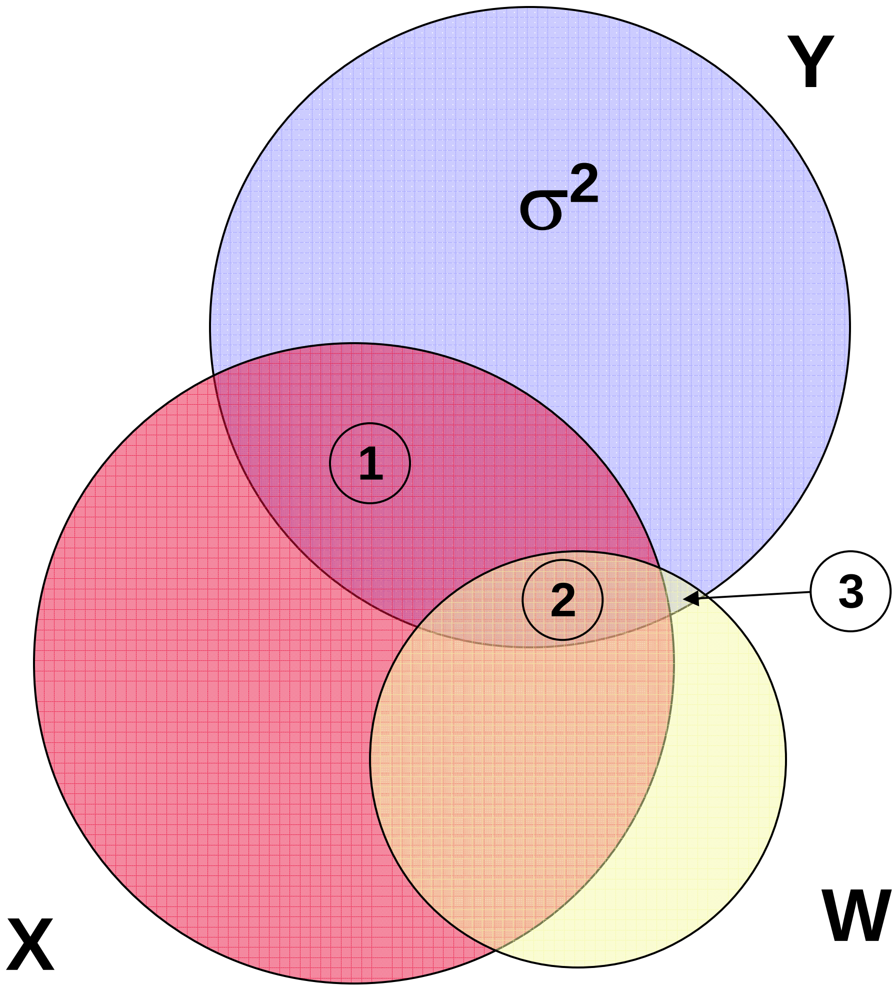
Including Unnecessary Variables



Here, variable W adds little on its own to explaining variation in Y (area 3).

Any explanatory power is due to its correlation with X (area 2).

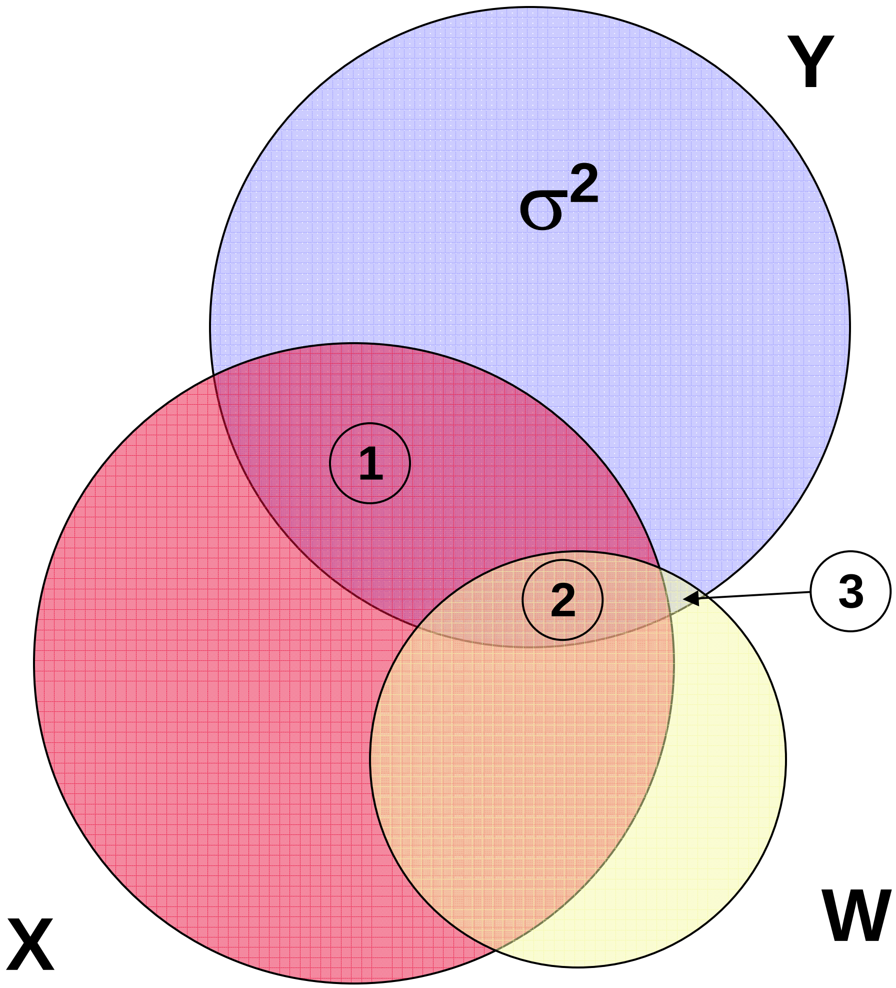
Including Unnecessary Variables



This means that:

1. The estimate of β_X is unbiased (since area 2 actually belongs only to X).
2. The variance of β_X is increased (since area 2 is removed).
3. The unexplained variance for Y (σ^2) is essentially the same (since area 3 is so small).

Including Unnecessary Variables



Solution: omit W from the regression.

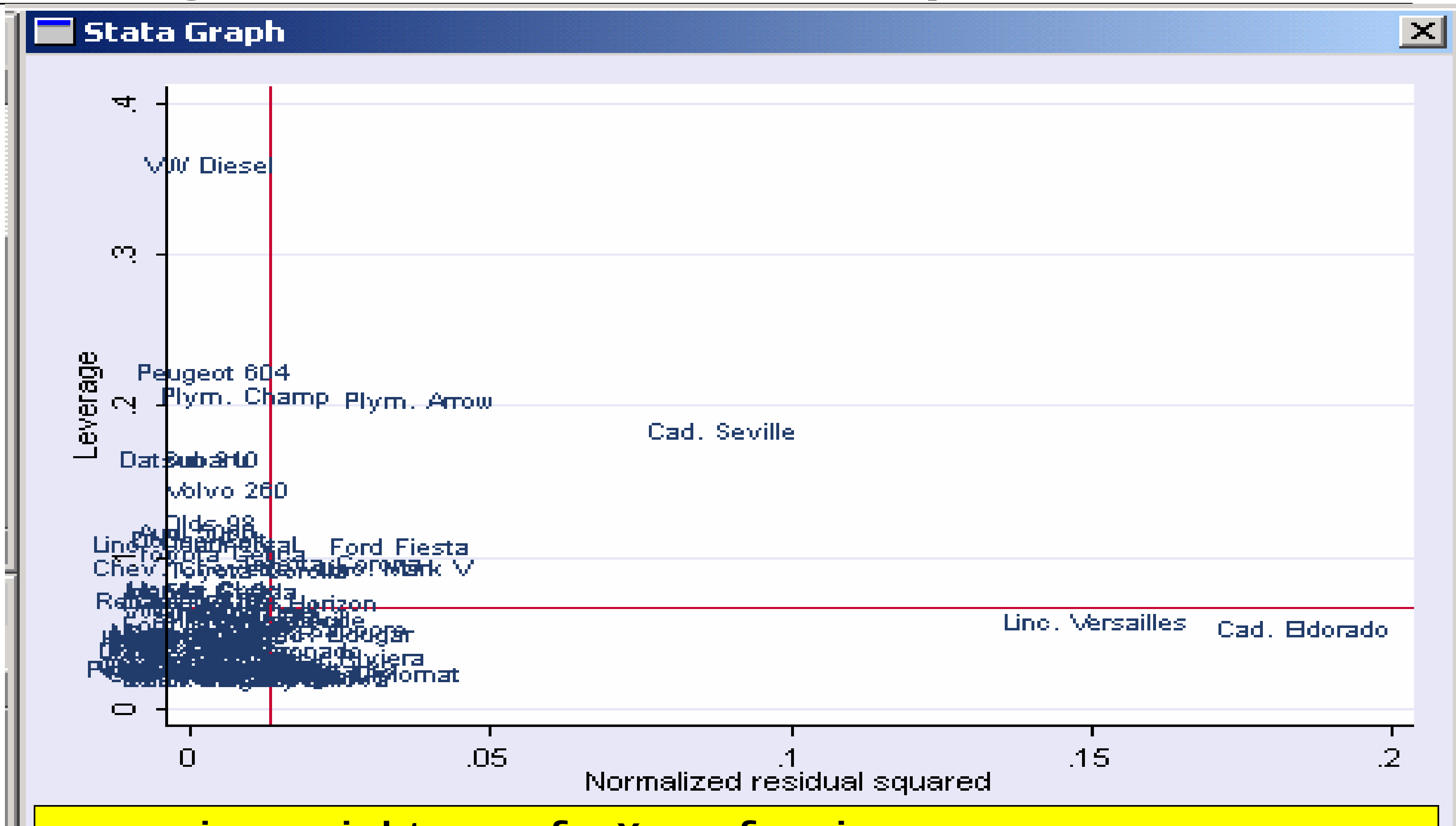
This is why we remove insignificant variables from regression equations.

Note: This is similar to multicollinearity: the more variables added to the model, the more uncertainty there is in estimating β_X.

Checking Individual Variables

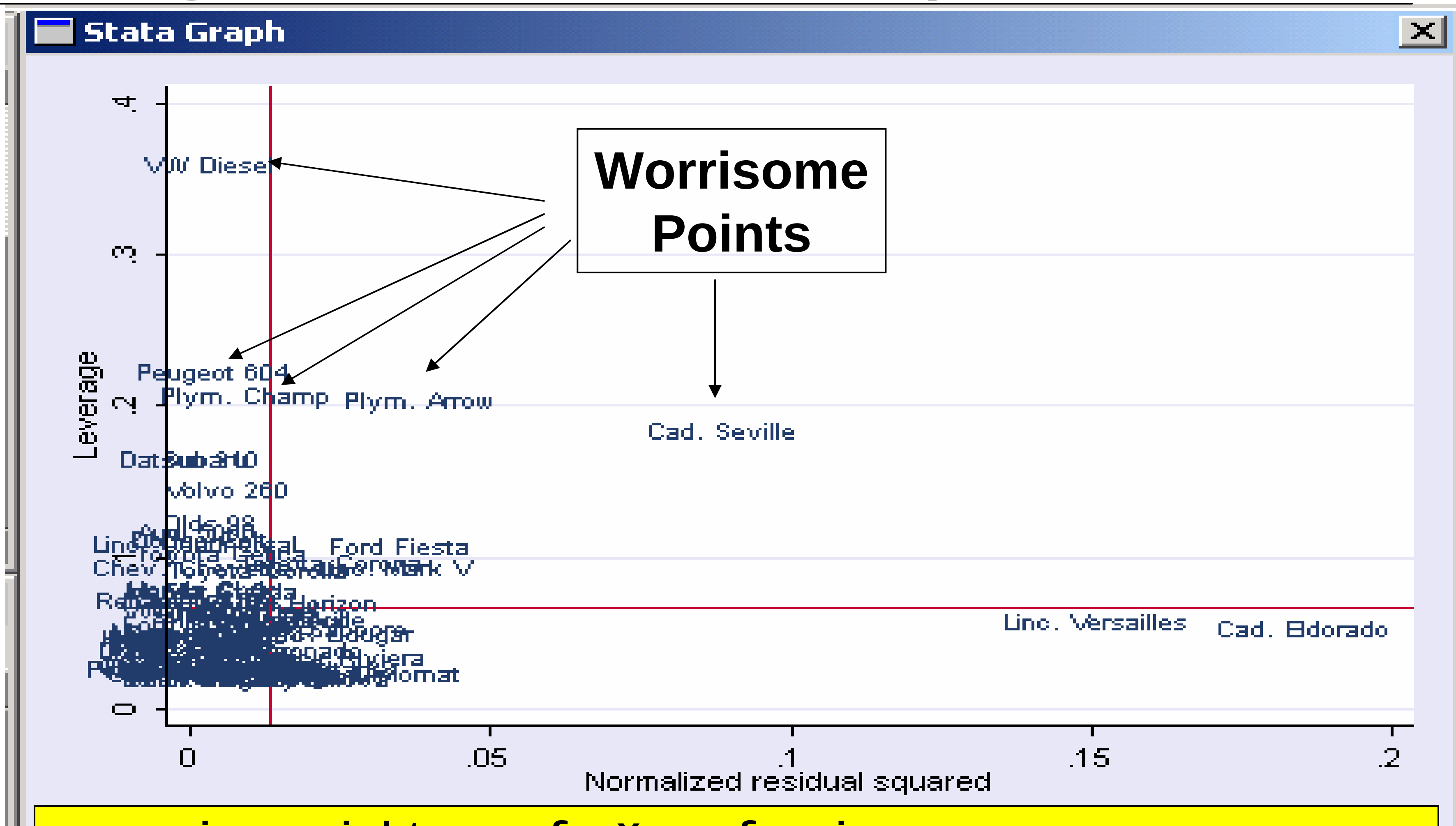
- If the diagnostics on the regression as a whole show potential problems, move to
  - Checking observations for:
    - Leverage
    - Outliers
    - Influence
  - Analyzing the contributions of individual variables to the regression:
    - Avplots
    - Cprplots

Diagnostic Plots: lvr2plot



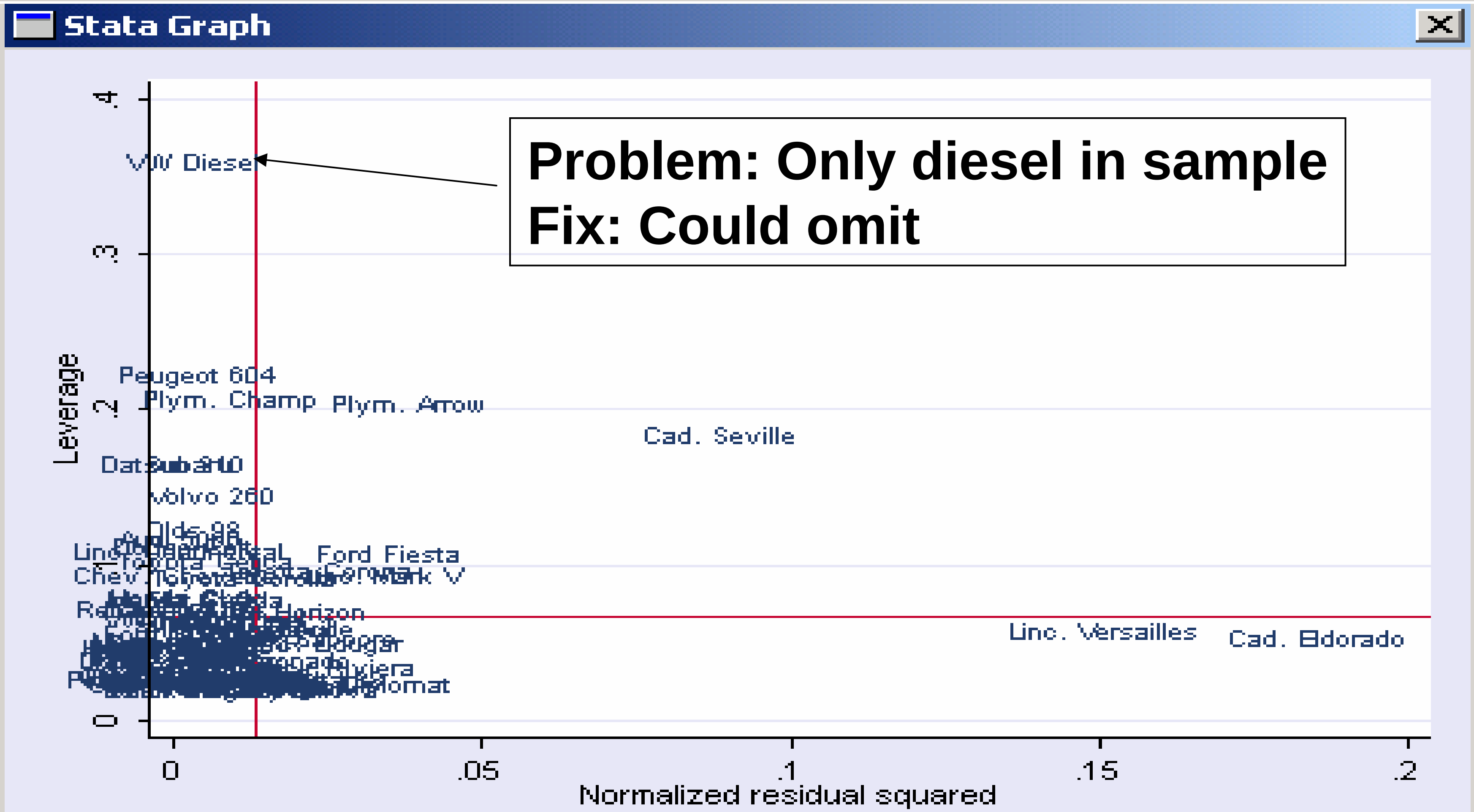

```
reg price weight mpg forXmpg foreign
lvr2plot, mlab(make) mlabp(0) m(none) mlabsize(small)
```

Diagnostic Plots: lvr2plot



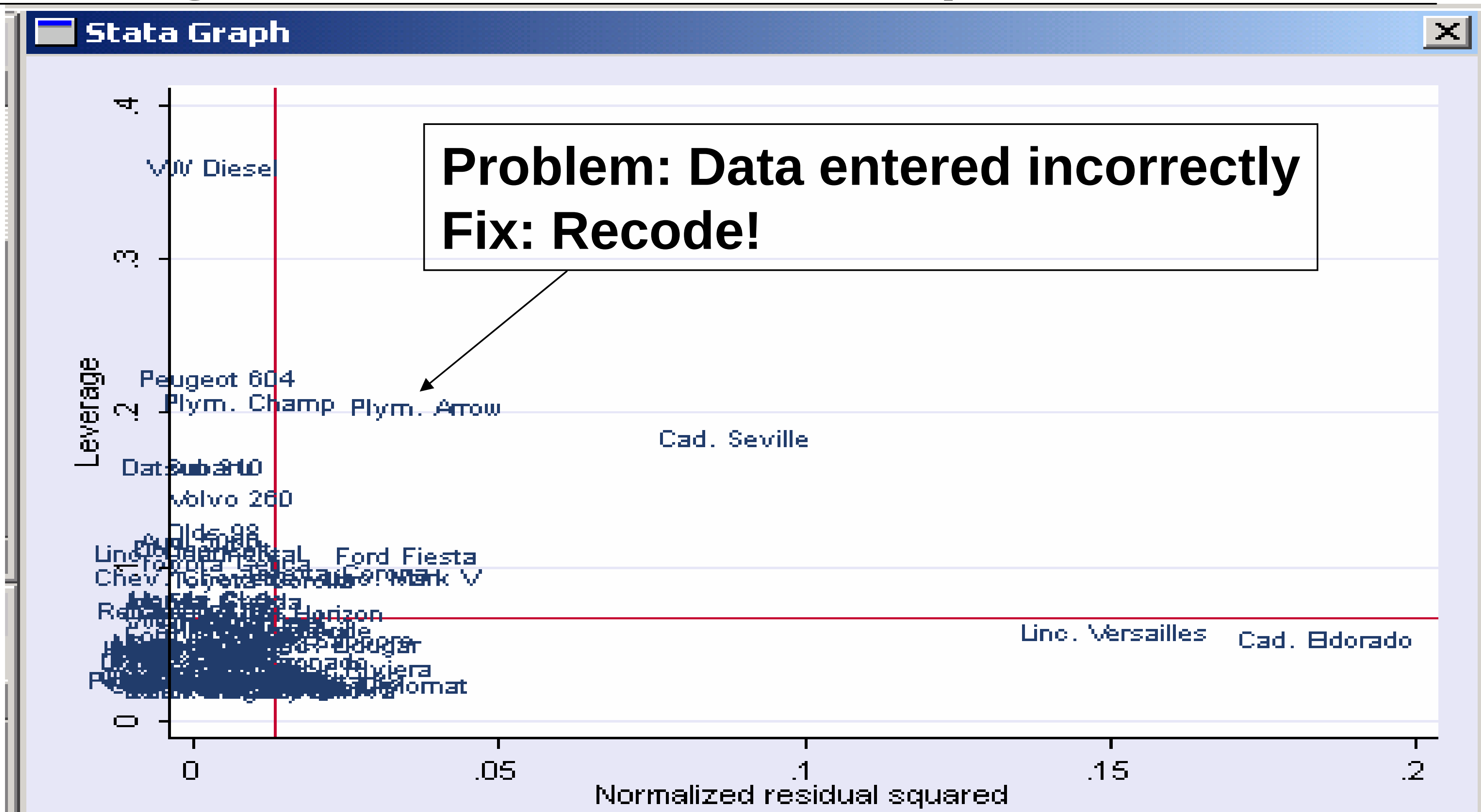

```
reg price weight mpg forXmpg foreign
lvr2plot, mlab(make) mlabp(0) m(none) mlabsize(small)
```

Diagnostic Plots: lvr2plot



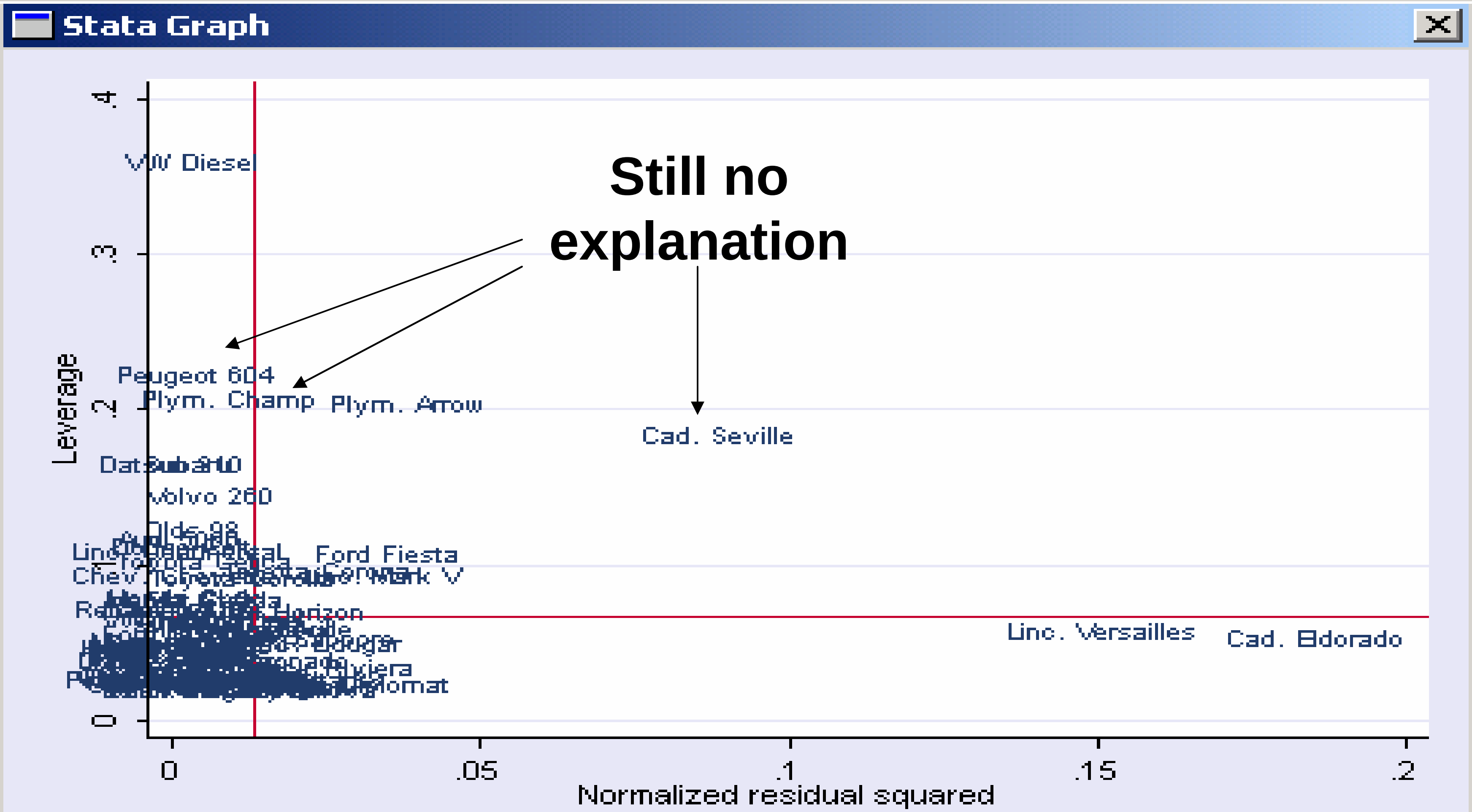

```
reg price weight mpg forXmpg foreign
lvr2plot, mlab(make) mlabp(0) m(none) mlabsize(small)
```

Diagnostic Plots: lvr2plot



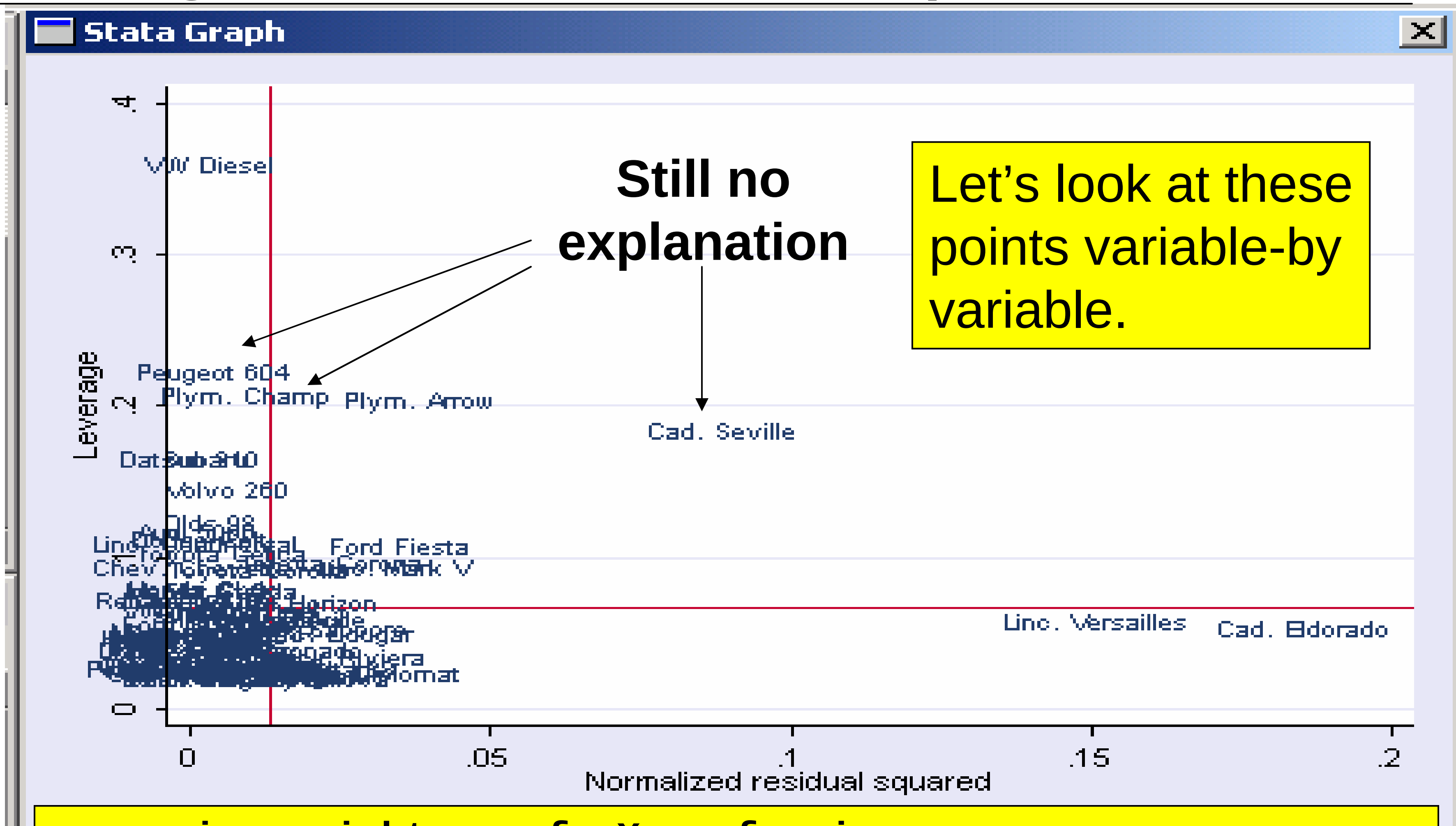

```
reg price weight mpg forXmpg foreign
lvr2plot, mlab(make) mlabp(0) m(none) mlabsize(small)
```

Diagnostic Plots: lvr2plot




```
reg price weight mpg forXmpg foreign
lvr2plot, mlab(make) mlabp(0) m(none) mlabsize(small)
```

Diagnostic Plots: lvr2plot



Let's look at these points variable-by variable.

```
reg price weight mpg forXmpg foreign
lvr2plot, mlab(make) mlabp(0) m(none) mlabsize(small)
```

Stata Commands: avplot

- Say the original model is:

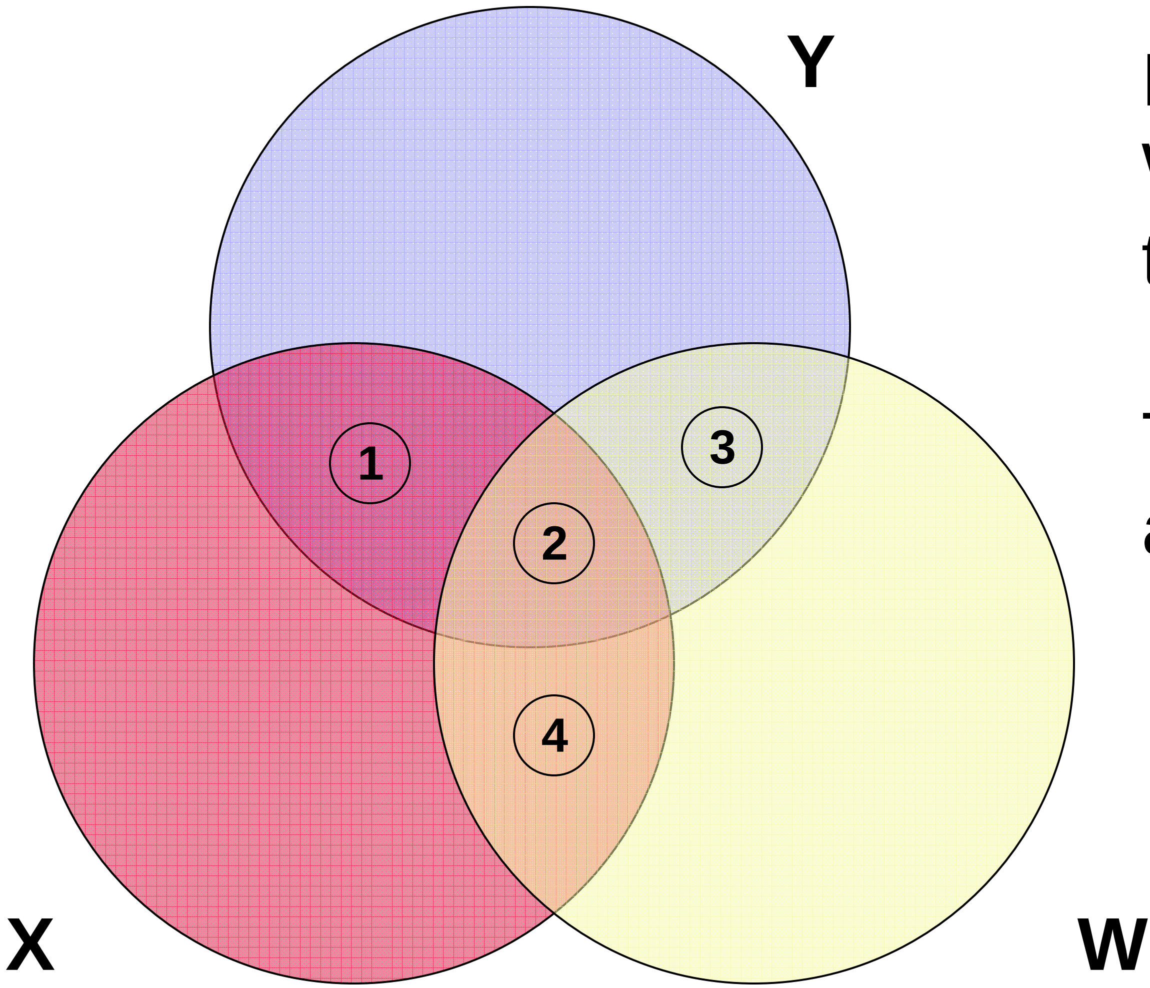
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

- We would like to graph the relation between Y and a single regressor x_1.
  - Can't do this directly, as we can with only one independent variable (too many dimensions).
- Added variable plots have the property that:
  1. There is a 1-to-1 correspondence btwn. Y_i & x_{1i}.
  2. A regression of Y on x_1 has the same slope and standard error as in the multiple regression.
  3. The "outlierliness" of each observation is preserved.

Stata Commands: avplot

- To obtain the avplot for x_1:
  1. Regress Y on x_2 and x_3 and calculate the residual; call this $e(Y|x_2,x_3)$
  2. Regress x_1 on x_2 and x_3 and calculate the residual; call this $e(x_1|x_2,x_3)$
  3. The avplot is then $e(Y|x_2,x_3)$ vs. $e(x_1|x_2,x_3)$
- The avplot thus provides a view of the relationship between Y and x_1 with the effects of x_2 and x_3 “taken out” of both.
- The slope coefficient in the avplot is the same as in the multiple regression. Why?

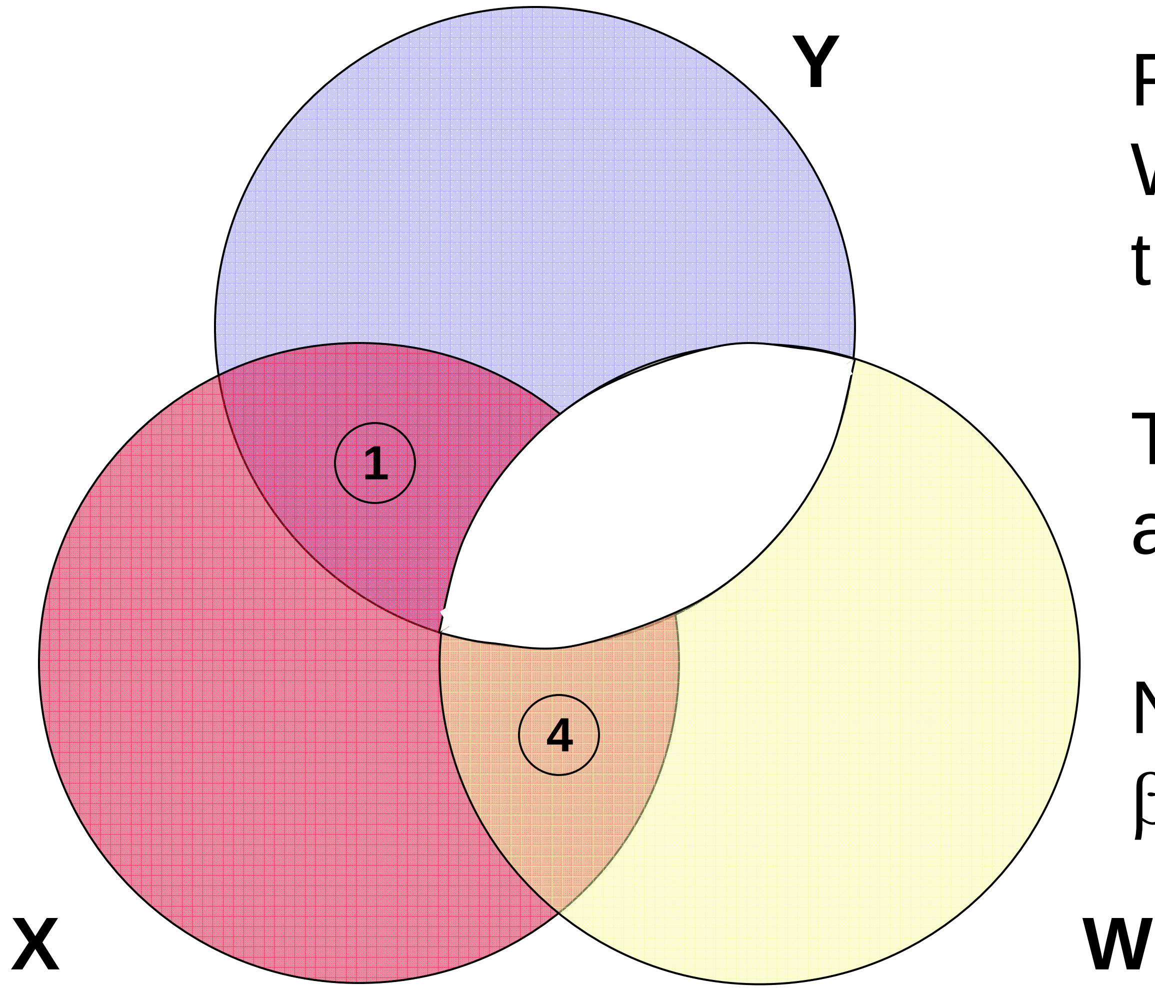
Example: Two Variables



Regress Y on just W first and take the residual.

This takes out areas 2 and 3

Example: Two Variables

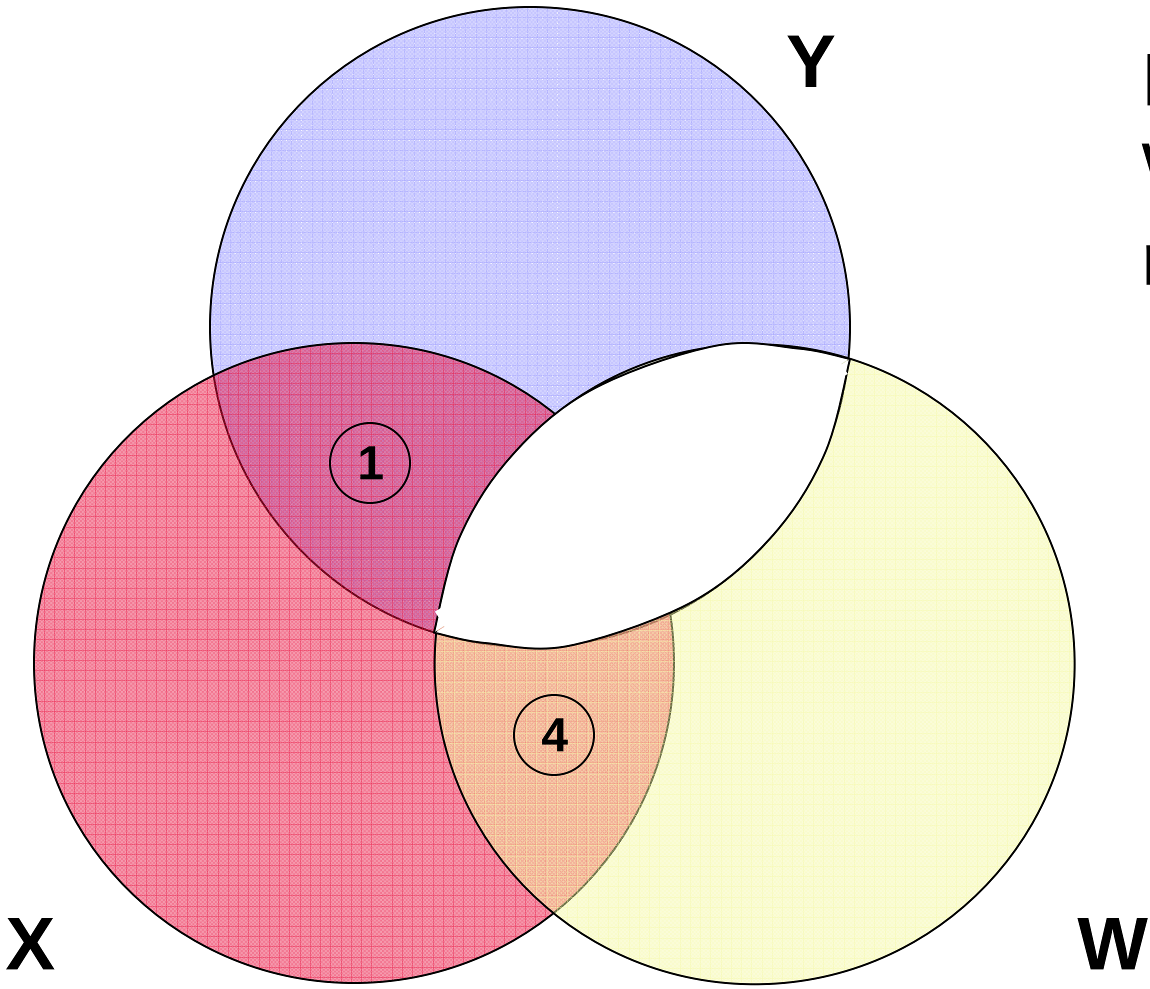


Regress Y on just W first and take the residual.

This takes out areas 2 and 3

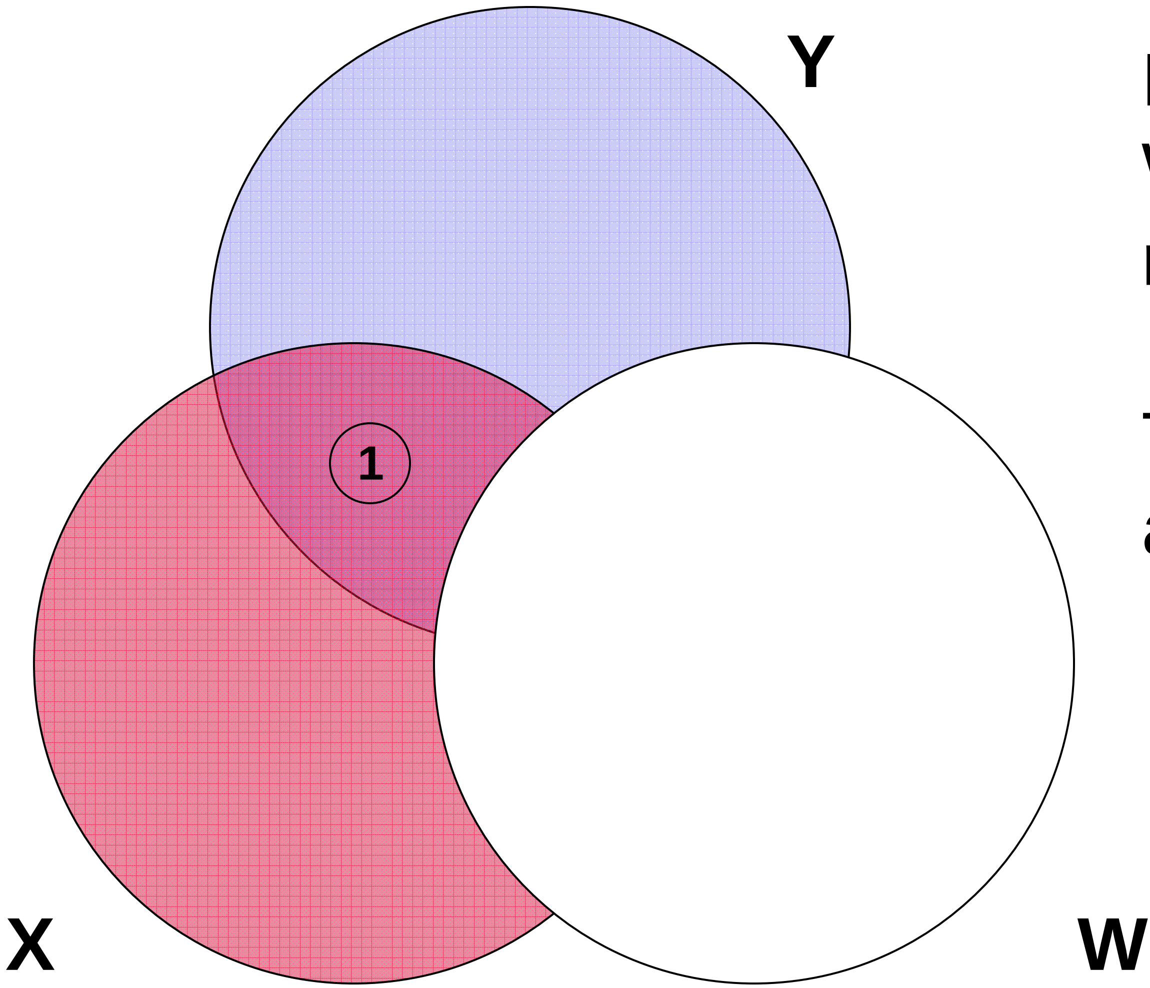
Note: estimate of β_W will be biased.

Example: Two Variables



Now regress X on W and take the residual.

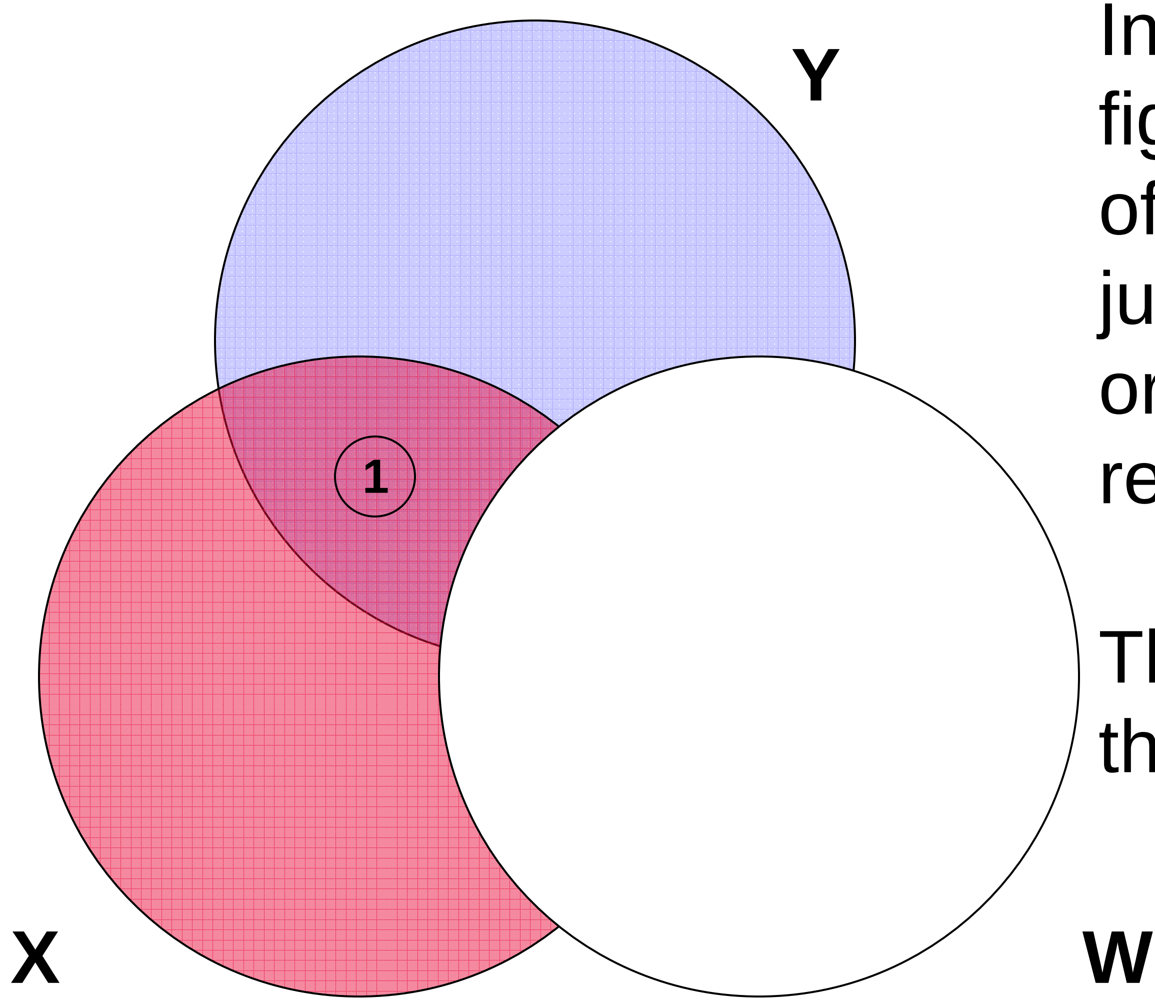
Example: Two Variables



Now regress X on W and take the residual.

This takes out area 4 as well.

Example: Two Variables

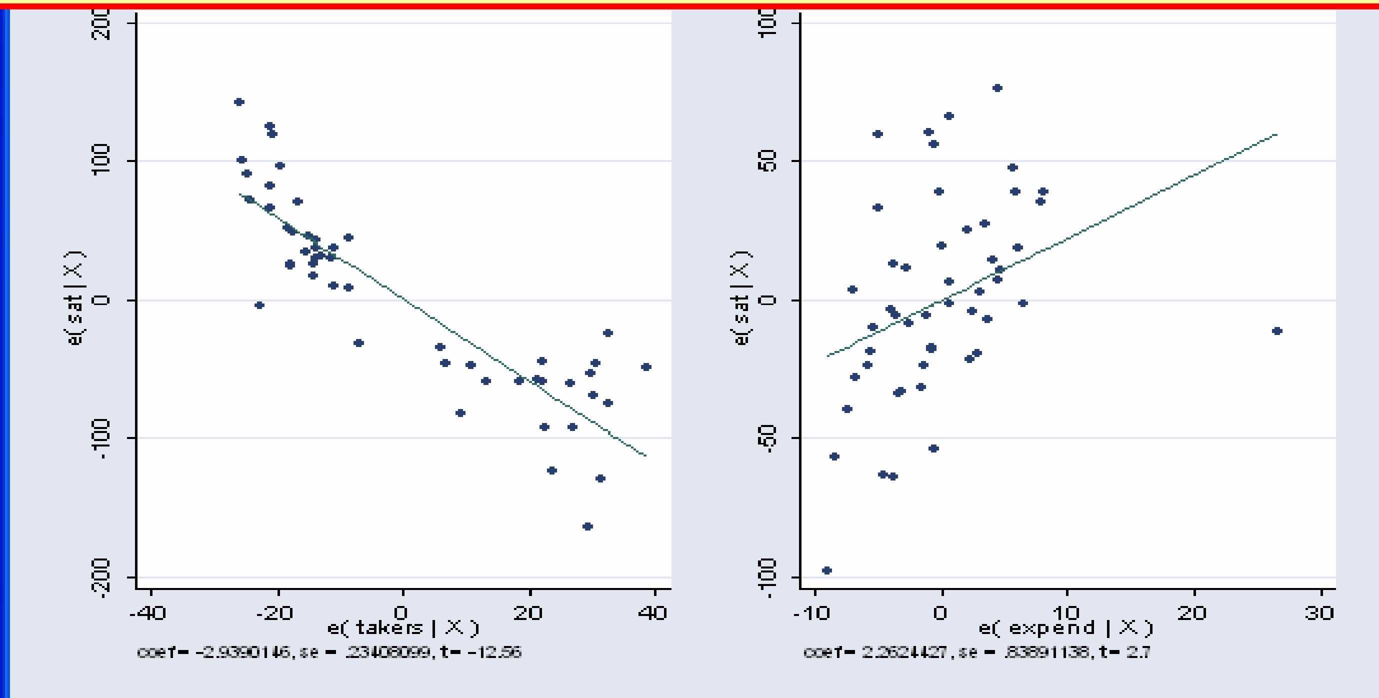


In the resulting figure, the overlap of Y and X is area 1, just as in the original multivariate regression!

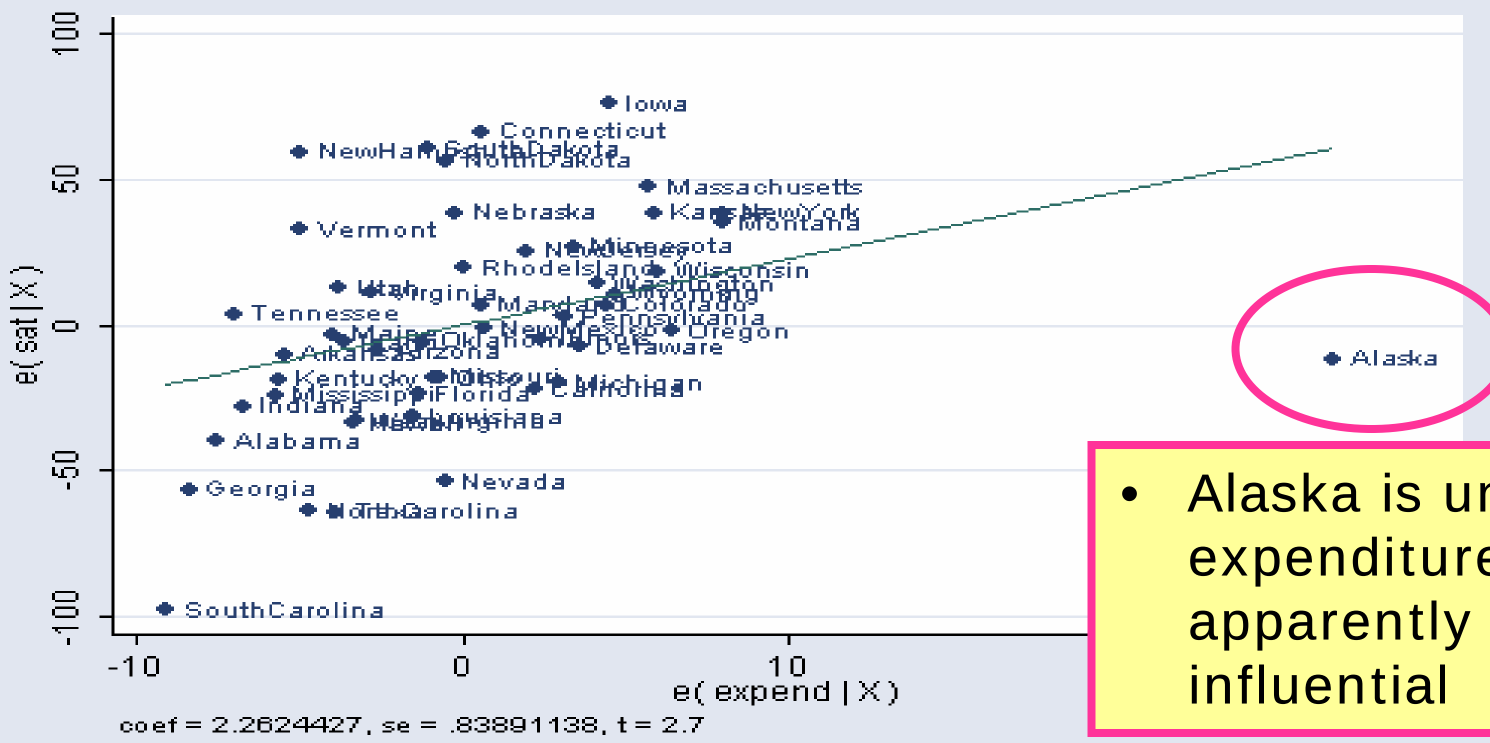
That's why we get the same coefficient

Added variable plots: example

- Is the state with largest expenditure influential?
- Is there an association of expend and SAT, after accounting for takers?

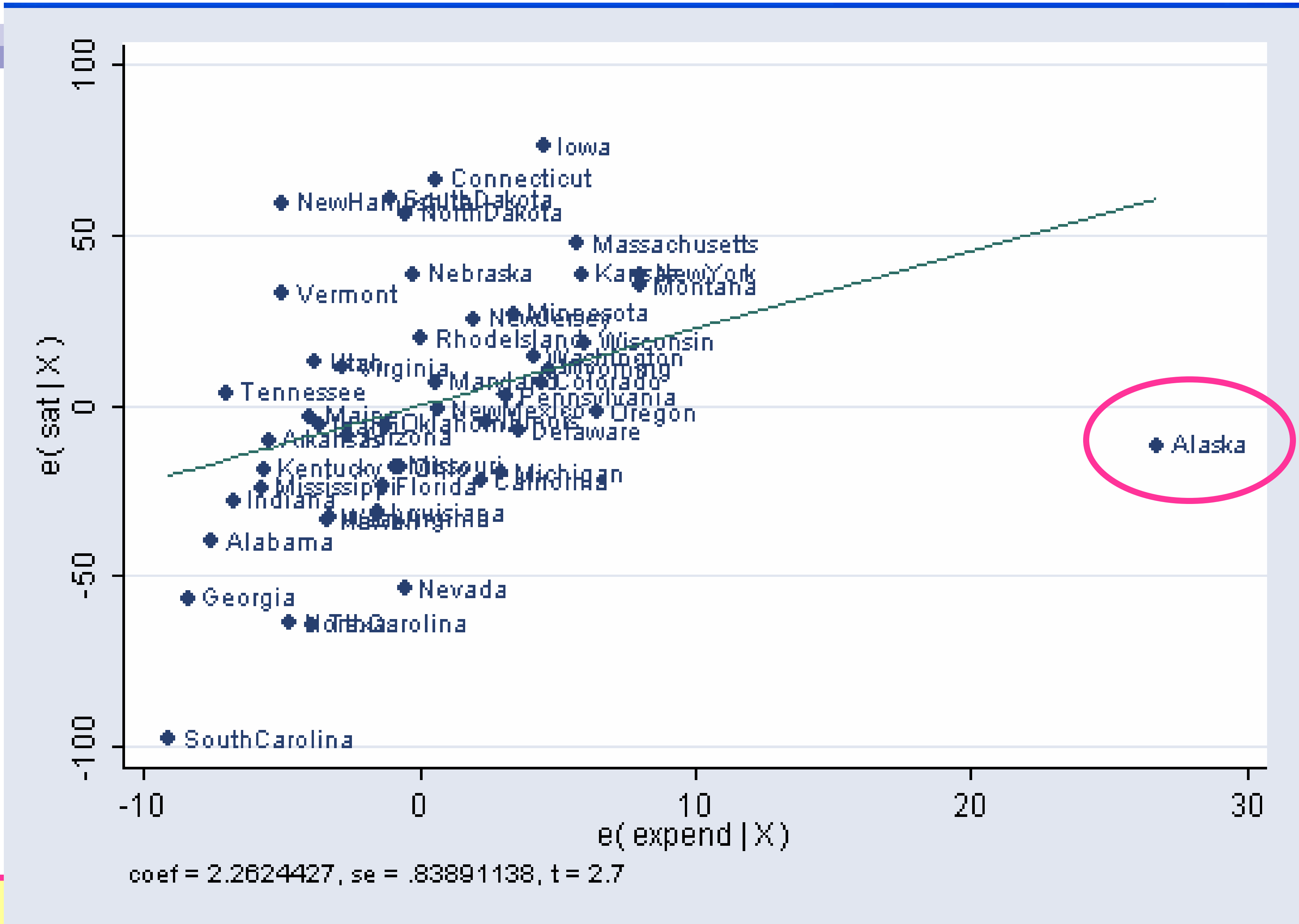



```
Stata R
Notes:
      1.
. log us                                                        ies\c
> txt\LO
      I                                                         ies\c
> txt\LO
  log ty
 opened

. inshee                                                        studi
> ase tx
(8 vars,

. reg sat takers expend

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =   79.23
       Model |  189732.978     2  94866.489            Prob > F      =  0.0000
    Residual |  56277.8419    47  1197.40089           R-squared     =  0.7712
-------------+------------------------------           Adj R-squared =  0.7615
       Total |   246010.82    49  5020.62898           Root MSE      =  34.603

------------------------------------------------------------------------------
         sat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      takers |  -2.939015    .234081   -12.56   0.000    -3.409925   -2.468104
      expend |   2.262443   .8389114     2.70   0.010     .5747707    3.950115
       _cons |   973.0426    19.1239    50.88   0.000     934.5703    1011.515
------------------------------------------------------------------------------

. avplots
```



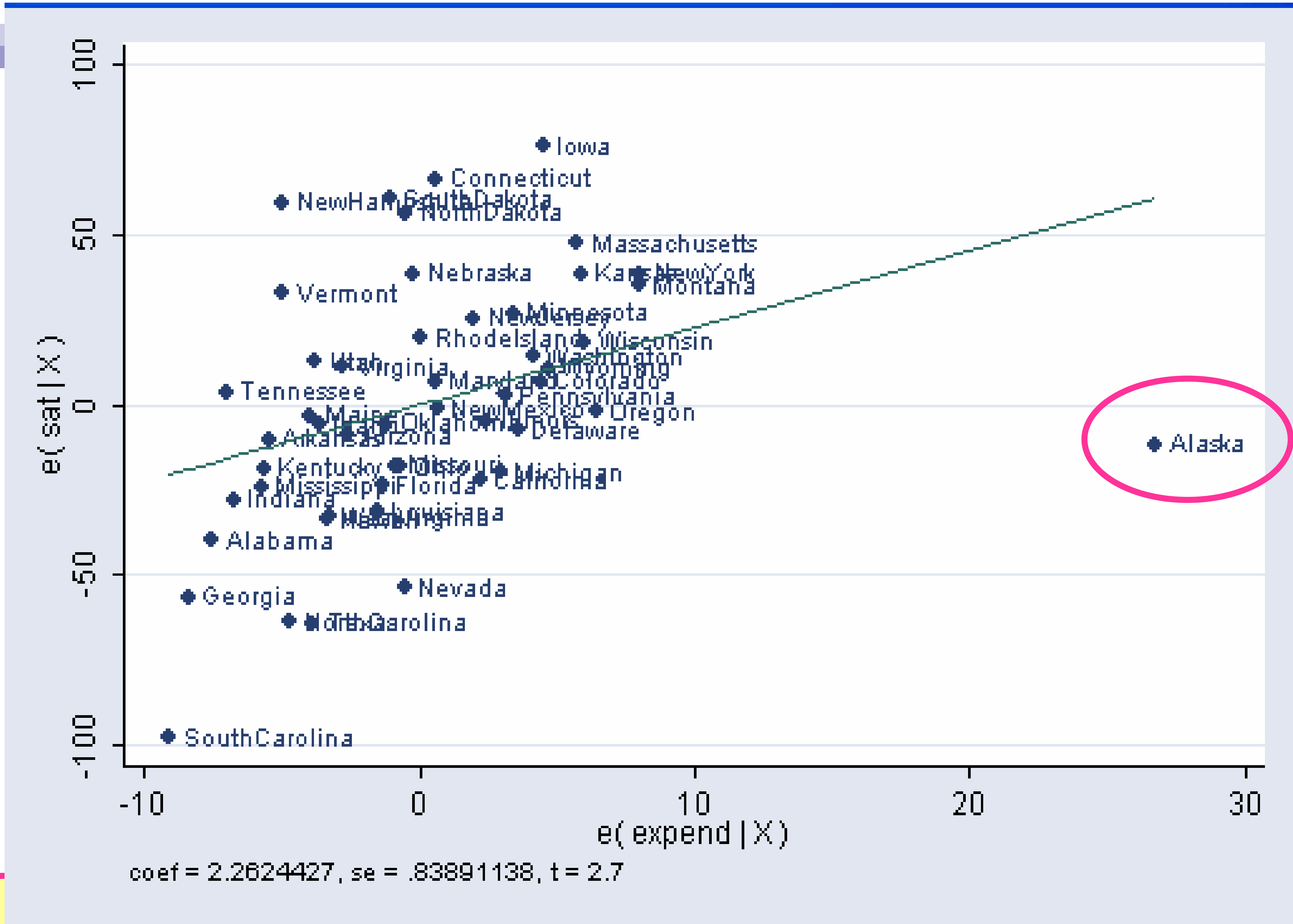
- Alaska is unusual in its expenditure, and is apparently quite influential

```
reg sat takers expend

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =   79.23
       Model |  189732.978     2  94866.489            Prob > F      =  0.0000
    Residual |  56277.8419    47  1197.40089           R-squared     =  0.7712
-------------+------------------------------           Adj R-squared =  0.7615
       Total |   246010.82    49  5020.62898           Root MSE      =  34.603

------------------------------------------------------------------------------
         sat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      takers |  -2.939015    .234081   -12.56   0.000    -3.409925   -2.468104
      expend |   2.262443   .8389114     2.70   0.010     .5747707    3.950115
       _cons |   973.0426    19.1239    50.88   0.000     934.5703    1011.515
------------------------------------------------------------------------------

avplots

avplot expend, mlabel(state)
```



X-axis: residuals after regression $expend_i = b_0 + b_1 * takers_i$

Y-axis: residuals after regression $SAT_i = b_0 + b_1 * takers_i$



After accounting for % of students who take SAT, there is a positive association between expenditure and mean SAT scores.

Component Plus Residual Plots

- We'd like to plot *y* versus x_2 but with the effect of x_1 subtracted out;

  i.e. plot $y - \beta_0 - \beta_1 x_1$ versus x_2

- To calculate this, get the *partial residual for* x_2:

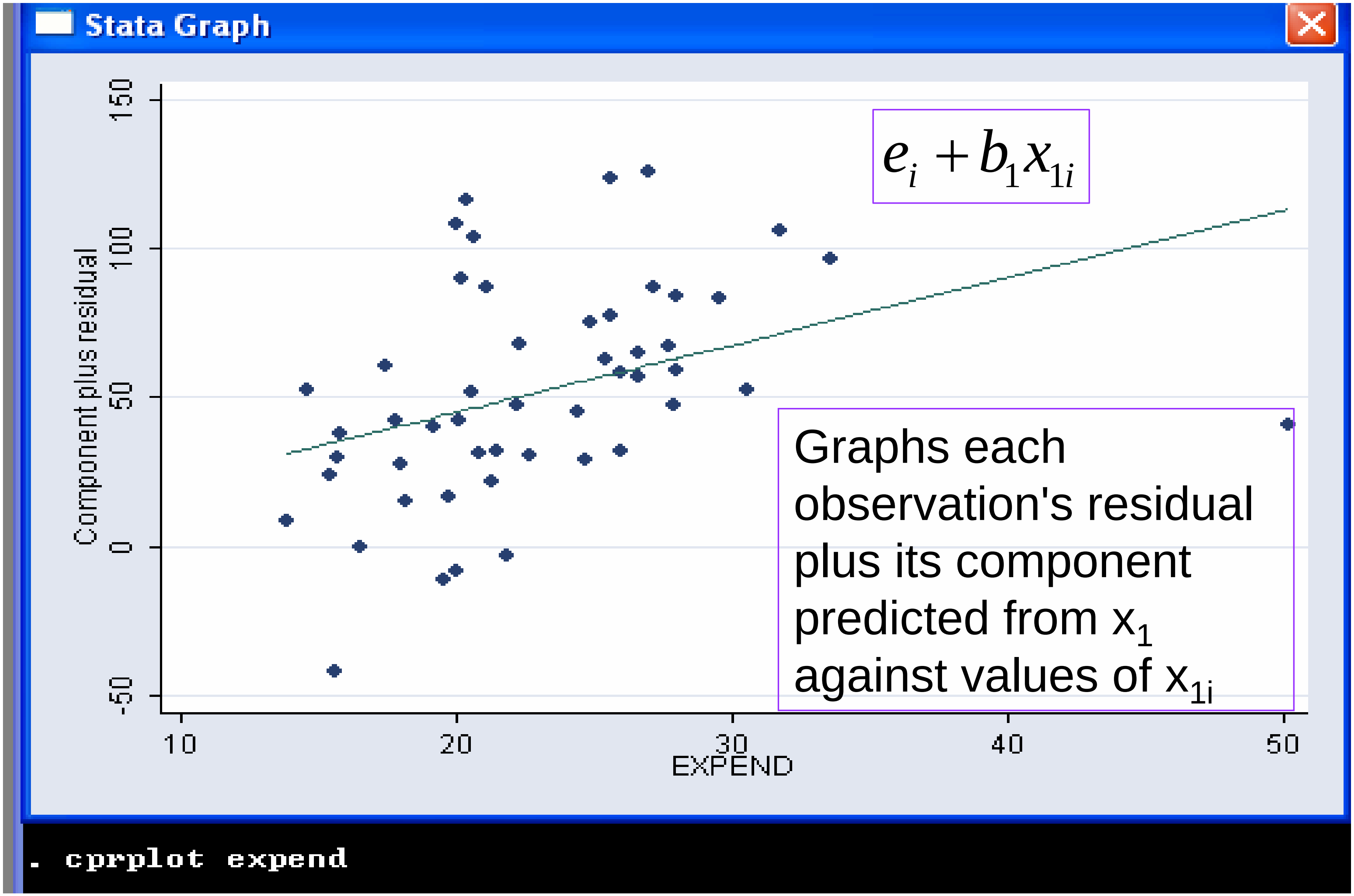
  a. Estimate $\beta_0, \beta_1, \text{and } \beta_2$ in $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

  b. Use these results to calculate $y - \beta_0 - \beta_1 x_1 = \beta_2 x_2 + \varepsilon_i$

  c. Plot this quantity vs. x_2.

- Whereas the avplots are better for detecting outliers, cprplots are better for determining functional form.

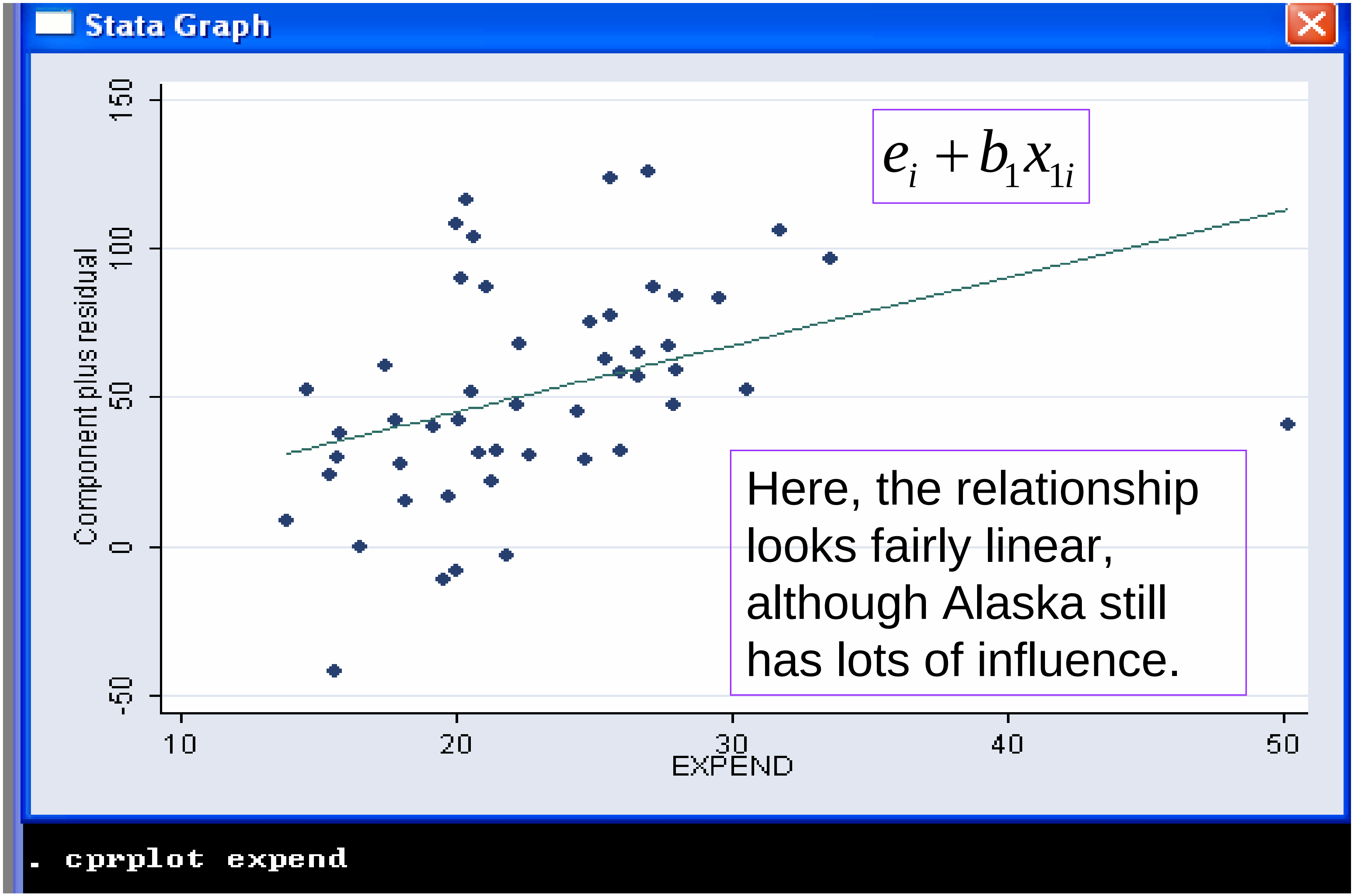
Graph cprplot x_1



$$e_i + b_1 x_{1i}$$

Graphs each observation's residual plus its component predicted from x_1 against values of x_{1i}

Graph cprplot x_1



Regression Fixes

- If you detect possible problems with your initial regression, you can:
  1. Check for mis-coded data
  2. Divide your sample or eliminate some observations (like diesel cars)
  3. Try adding more covariates if the ovtest turns out positive
  4. Change the functional form on Y or one of the regressors
  5. Use robust regression

Robust Regression

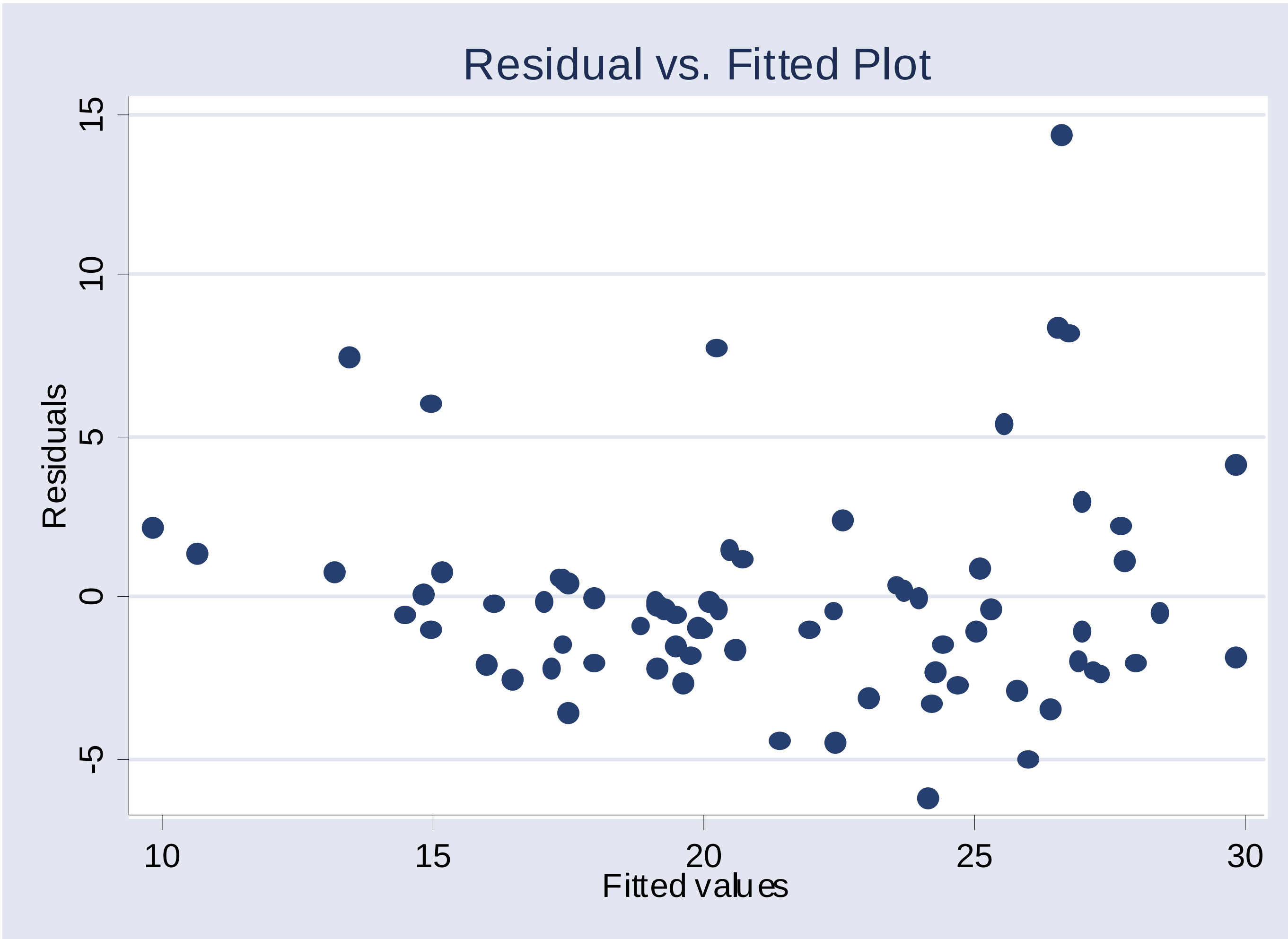
- This is a variant on linear regression that downplays the influence of outliers
  1. First performs the original OLS regression
  2. Drops observations with Cook's distance > 1
  3. Calculates weights for each observation based on their residuals
  4. Performs weighted least squares regression using these weights
- Stata command: "rreg" instead of "reg"

Robust Regression: Example

```
. reg mpg weight foreign

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  2,    71) =   69.75
       Model |   1619.2877     2  809.643849           Prob > F      =  0.0000
    Residual |  824.171761    71   11.608053           R-squared     =  0.6627
-------------+------------------------------           Adj R-squared =  0.6532
       Total |  2443.45946    73  33.4720474           Root MSE      =  3.4071

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0065879   .0006371   -10.34   0.000    -.0078583   -.0053175
     foreign |  -1.650029   1.075994    -1.53   0.130      -3.7955    .4954422
       _cons |    41.6797   2.165547    19.25   0.000     37.36172    45.99768
------------------------------------------------------------------------------
```

- This is the original regression
- Run the usual diagnostics

Robust Regression: Example



Robust Regression: Example

```
. reg mpg weight foreign

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  2,    71) =   69.75
       Model |   1619.2877     2  809.643849           Prob > F      =  0.0000
    Residual |  824.171761    71   11.608053           R-squared     =  0.6627
-------------+------------------------------           Adj R-squared =  0.6532
       Total |  2443.45946    73  33.4720474           Root MSE      =  3.4071

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0065879   .0006371   -10.34   0.000    -.0078583   -.0053175
     foreign |  -1.650029   1.075994    -1.53   0.130      -3.7955    .4954422
       _cons |    41.6797   2.165547    19.25   0.000     37.36172    45.99768
------------------------------------------------------------------------------
```

- rvfplot shows heterskedasticity
- Also, fails a hettest

Robust Regression: Example

```
. rreg mpg weight foreign, genwt(w)

   Huber iteration 1:  maximum difference in weights = .80280176
   Huber iteration 2:  maximum difference in weights = .2915438
   Huber iteration 3:  maximum difference in weights = .08911171
   Huber iteration 4:  maximum difference in weights = .02697328
Biweight iteration 5:  maximum difference in weights = .29186818
Biweight iteration 6:  maximum difference in weights = .11988101
Biweight iteration 7:  maximum difference in weights = .03315872
Biweight iteration 8:  maximum difference in weights = .00721325

Robust regression estimates                           Number of obs =      74
                                                      F(  2,    71) =  168.32
                                                      Prob > F      =  0.0000

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0063976    .0003718   -17.21   0.000     -.007139   -.0056562
     foreign |  -3.182639     .627964    -5.07   0.000    -4.434763   -1.930514
       _cons |   40.64022    1.263841    32.16   0.000      38.1202    43.16025
------------------------------------------------------------------------------
```

Robust Regression: Example

```
. rreg mpg weight foreign, genwt(w)

   Huber iteration 1:  maximum difference in weights = .80280176
   Huber iteration 2:  maximum difference in weights = .2915438
   Huber iteration 3:  maximum difference in weights = .08911171
   Huber iteration 4:  maximum difference in weights = .02697328
Biweight iteration 5:  maximum difference in weights = .29186818
Biweight iteration 6:  maximum difference in weights = .11988101
Biweight iteration 7:  maximum difference in weights = .03315872
Biweight iteration 8:  maximum difference in weights = .00721325

Robust regression estimates                            Number of obs =      74
                                                       F(  2,    71) =  168.32
                                                       Prob > F      =  0.0000

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0063976   .0003718   -17.21   0.000     -.007139   -.0056562
     foreign |  -3.182639    .627964    -5.07   0.000    -4.434763   -1.930514
       _cons |   40.64022   1.263841    32.16   0.000      38.1202    43.16025
------------------------------------------------------------------------------
```

- Note: Coefficient on foreign changes from -1.65 to -3.18

Robust Regression: Example

```
. rreg mpg weight foreign, genwt(w)

   Huber iteration 1:  maximum difference in weights = .80280176
   Huber iteration 2:  maximum difference in weights = .2915438
   Huber iteration 3:  maximum difference in weights = .08911171
   Huber iteration 4:  maximum difference in weights = .02697328
Biweight iteration 5:  maximum difference in weights = .29186818
Biweight iteration 6:  maximum difference in weights = .11988101
Biweight iteration 7:  maximum difference in weights = .03315872
Biweight iteration 8:  maximum difference in weights = .00721325

Robust regression estimates                            Number of obs =      74
                                                       F(  2,    71) =  168.32
                                                       Prob > F      =  0.0000

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.       t     P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0063976   .0003718    -17.21    0.000     -.007139   -.0056562
     foreign |  -3.182639    .627964     -5.07    0.000    -4.434763   -1.930514
       _cons |   40.64022   1.263841     32.16    0.000      38.1202    43.16025
------------------------------------------------------------------------------
```

- This command saves the weights generated by rreg

Robust Regression: Example

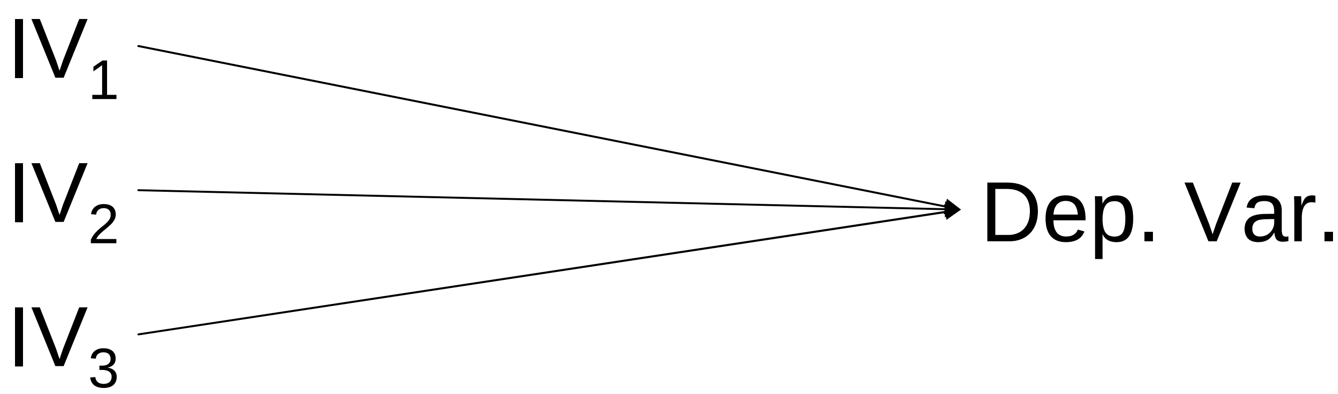
```
. sort w

. list make mpg weight w if w<.467, sep(0)

     +------------------------------------------+
     | make                mpg   weight       w |
     |------------------------------------------|
  1. | Subaru               35    2,050       0 |
  2. | VW Diesel            41    2,040       0 |
  3. | Datsun 210           35    2,020       0 |
  4. | Plym. Arrow          28    3,260  .04429567 |
  5. | Cad. Seville         21    4,290  .08241943 |
  6. | Toyota Corolla       31    2,200  .10443129 |
  7. | Olds 98              21    4,060  .28141296 |
     +------------------------------------------+
```

- This shows that three observations were dropped by the rreg, including the VW Diesel

Theories, Tests and Models

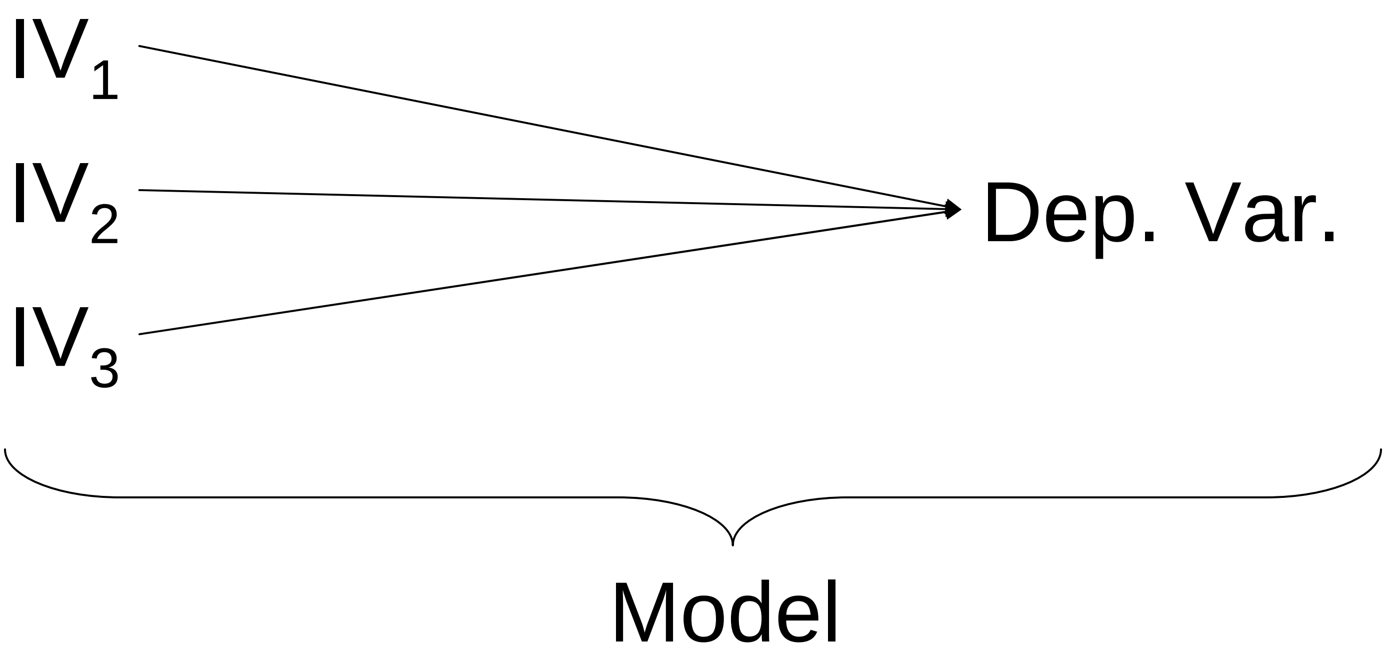
- Question: Which variables should we include on the RHS of our estimation?
- This is a fundamental question of research design and testing.
  - NOT merely a mechanical process.
- So let's first review some basic concepts of theory-driven research.
  - Theory-driven = based on a model of the phenomenon of interest, formal or otherwise.
  - We always do this, somehow, in our research

Theories, Tests and Models



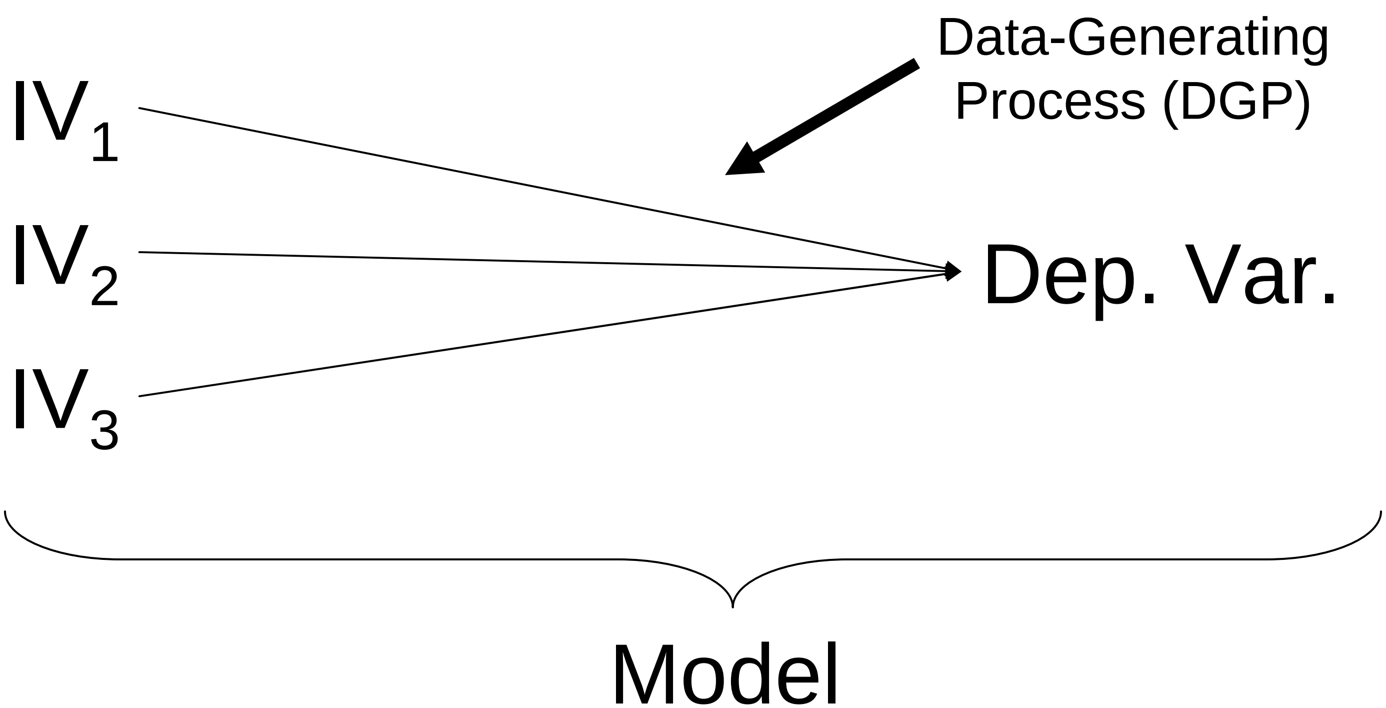
- Say we have a theory predicting IV_1, IV_2, and IV_3 all affect dependent variable Y.

Theories, Tests and Models



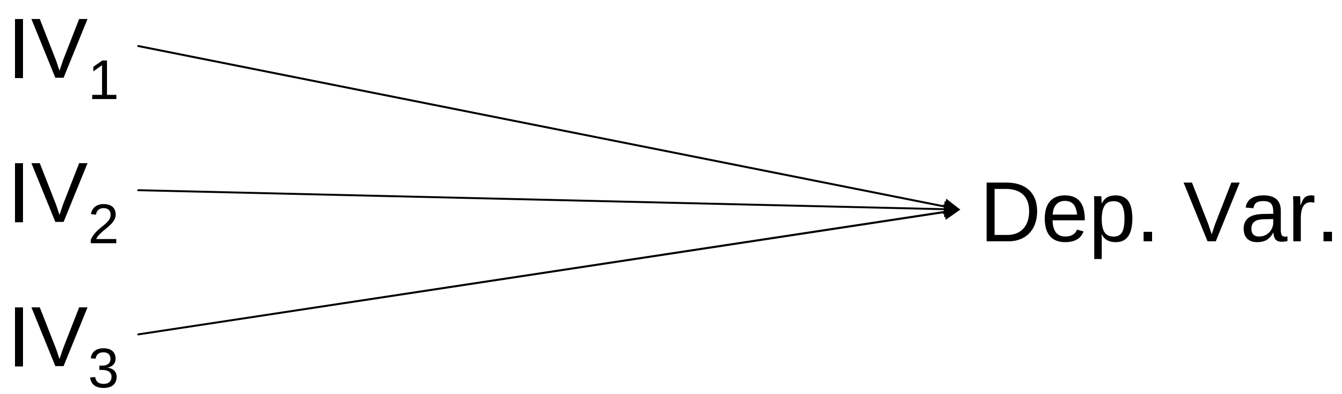
- Say we have a theory predicting IV_1, IV_2, and IV_3 all affect dependent variable Y.

Theories, Tests and Models



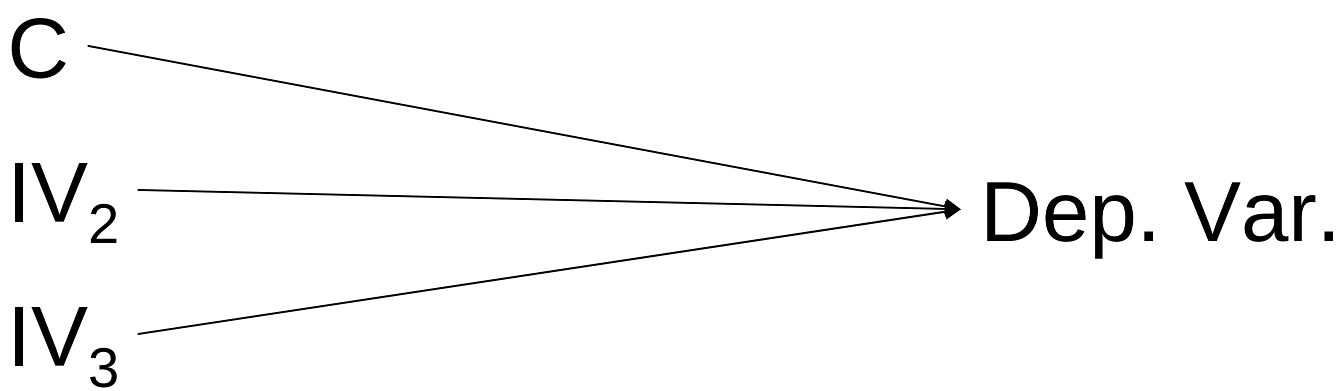
- Say we have a theory predicting IV_1, IV_2, and IV_3 all affect dependent variable Y.

Theories, Tests and Models



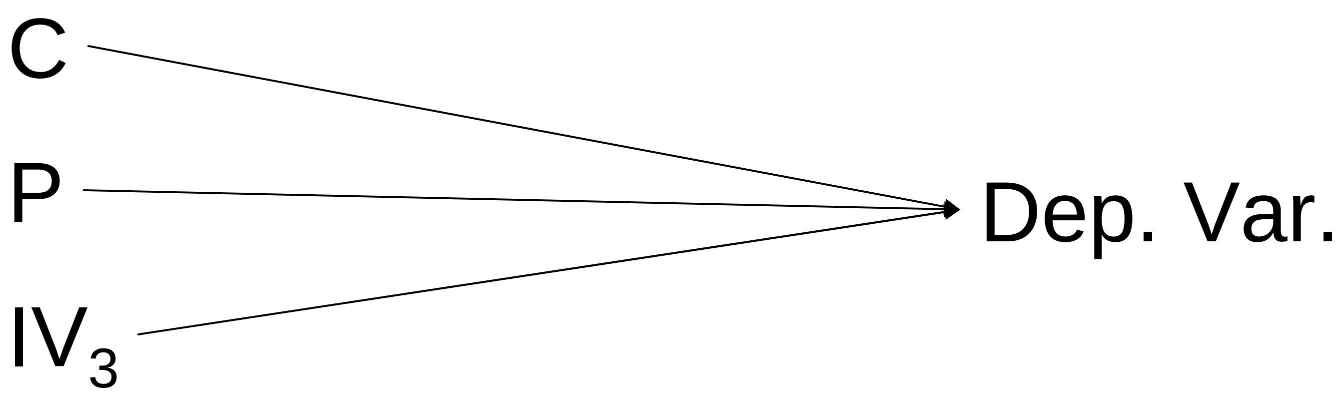
- For instance:

Theories, Tests and Models



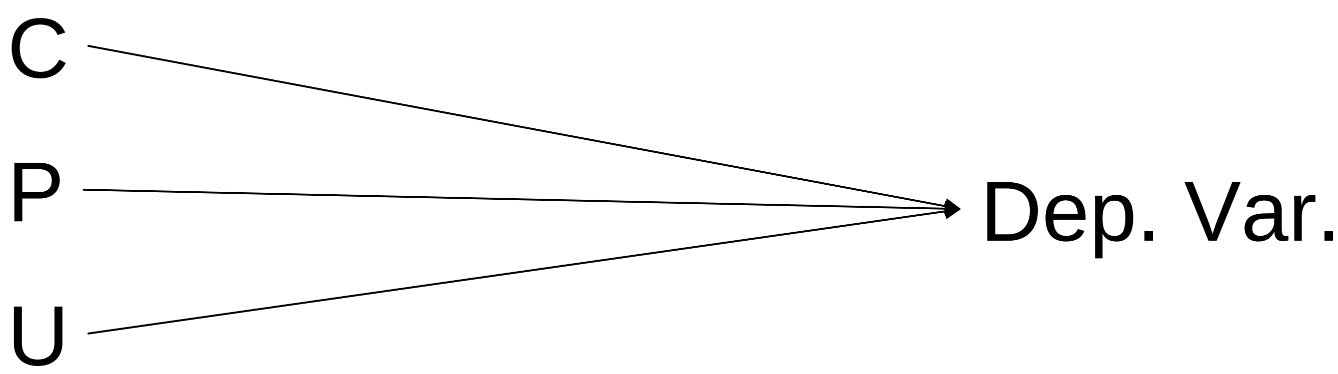
- For instance:
  1. Congressional Committee ideal points,

Theories, Tests and Models



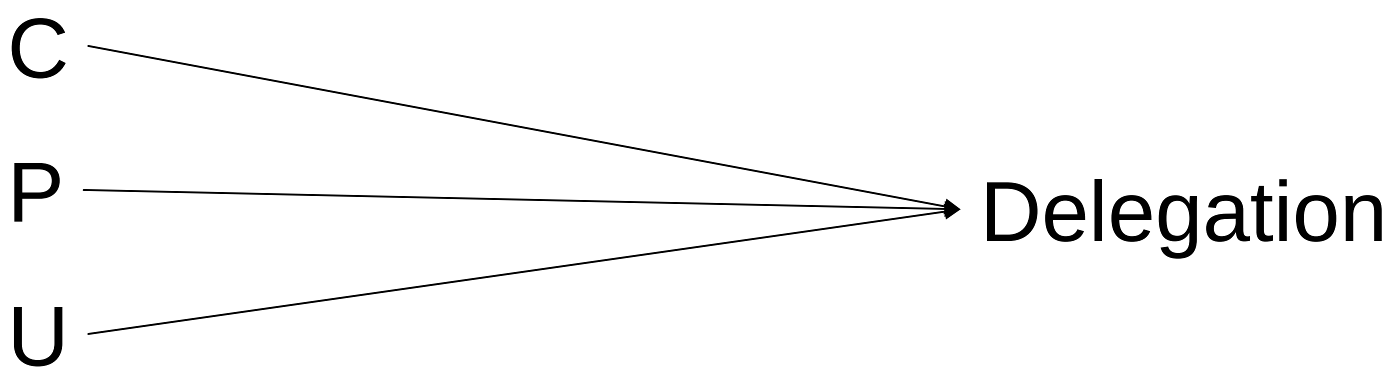
- For instance:
  1. Congressional Committee ideal points,
  2. The President's ideal point, and

Theories, Tests and Models



- For instance:
  1. Congressional Committee ideal points,
  2. The President's ideal point, and
  3. Uncertainty in the policy area

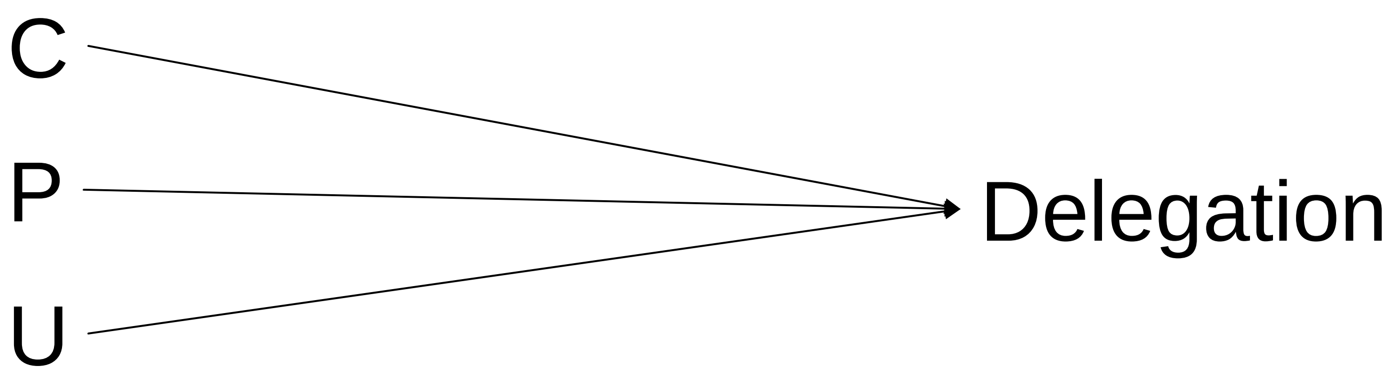
Theories, Tests and Models



- For instance:
  1. Congressional Committee ideal points,
  2. The President's ideal point, and
  3. Uncertainty in the policy area

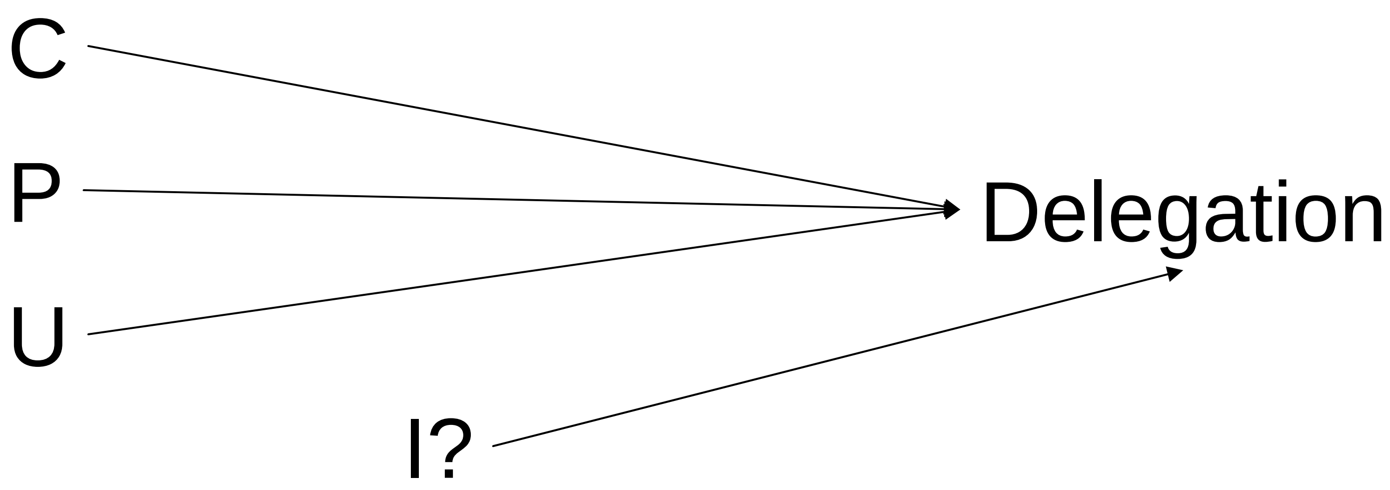
  ➔ Affect Delegation to the executive.

Theories, Tests and Models



- Question: What if another variable is suggested that might also impact Y?

Theories, Tests and Models



- Question: What if another variable is suggested that might also impact Y?
- For instance, the Interest group environment surrounding the issue.

Theories, Tests and Models

- Remember: our original regression variables came from a particular <u>model</u> of our subject.
- There are, generally, three options when new variables are suggested:
  1. Re-solve the model with the new variable(s) included as well;
  2. Assume the model is a complete Data Generating Process (DGP) and ignore other potential factors;
  3. Treat the model as a partial DGP.
- We will consider each in turn.

Add New Variables to the Model

- We could expand our model to include the new variable(s).
  - Formal models: re-solve the equilibrium
  - Qualitative models: re-evaluate the predicted effects
- In the long run, though, this is not feasible
  - There will always be more factors that might affect the phenomenon of interest
  - Don’t want to take the position that you can’t have a theory of anything without having a theory of everything

Treat the Model as a Complete DGP

- This means to just look at the impact of the variables suggested by your theory on the dependent variable
- This is unrealistic, but an important first cut at the problem
- You can do this with
  - Naturally occuring data
  - Experimental data
- Either way, this is a direct test of the theory

Treat the Model as a Partial DGP

- This is the modal response – just add the new variables to the estimation model
- But it has some shortcomings:
  - You lose the advantages of modelling
  - You're not directly testing your theory any more, but rather your theory plus some conjectures.
  - Have to be careful about how the error term enters the equation – where does ε come from?
- If you do this, best to use variables already known to affect the dependent variable.

How To Handle Many Regressors

- With this in mind, how can we handle a situation where we have many potential indep. variables?
- Two good reasons for seeking a subset of these:
  - General principle: smaller is better (Occam's razor)
  - Unnecessary terms add imprecision to inferences
- Computer assisted tools
  - Fit of all possible models (include or exclude each *X*) Compare with these statistics:
    - Cp, AIC, or BIC
  - Stepwise regression (search along favorable directions)
- But don't expect a BEST or a TRUE model or a law of nature

Objectives when there are many X's (12.2.1)

- **Assessment** of one X, after accounting for many others
  - Ex: Do males receive higher salaries than females, after accounting for legitimate determinants of salary?
  - Strategy: first find a good set of X's to explain salary; then see if the sex indicator is significant when added
- **Fishing for association**; i.e. what are the important X's?
  - The trouble with this: we can find several subsets of X's that explain Y; but that doesn't imply importance or causation
  - Best attitude: use this for hypothesis generation
- **Prediction**
  - This is a straightforward objective
  - Find a useful set of X's; no interpretation required

Loss of precision due to multicollinearity

- **Review**: variance of L.S. estimator of slope in simple reg. =

Variance about the regression → σ^2

$$\frac{\sigma^2}{(n-1)s_x^2}$$

Sample variance of X → s_x^2

- **Fact**: variance of L.S. estimator of coef. of X_j in mult. reg. =

$$\frac{\sigma^2}{(n-1)s_j^2(1-R_j^2)}$$

Sample variance of X_j → s_j^2

R^2 in the regression of X_j on the other X's in model → R_j^2

Implications of Multicollinearity

- So variance of an estimated coef. will tend to be larger if there are other X's in the model that can predict X_j.
- The S.E. of prediction will also tend to be larger if there are unnecessary or redundant X's in the model.
- The tradeoff for adding more regressors is:
  - You explain more of the variance in Y, but
  - You estimate the impact of all other variables on Y using less information

Implications of Multicollinearity

- Multicollinearity:
  - The situation in which $s_j^2(1-R_j^2)$ is small for one or more j's (usually characterized by highly correlated X's).
- Strategy:
  - There isn't a real need to decide whether multicollinearity is or isn't present, as long as one tries to find a subset of x's that adequately explains μ(Y), without redundancies.
- "Good" subsets of x's:
  - (a) lead to a small $\hat{\sigma}^2$
  - (b) with as few x's as possible (Criteria Cp, AIC, and BIC formalize this)

Strategy for dealing with many X's

1. Identify objectives; identify relevant set of X's
2. Exploration:
   - matrix of scatterplots; correlation matrix; Residual plots after fitting tentative models
3. Resolve transformation and influence before variable selection
4. Computer-assisted variable selection:
   - Best:
     - Compare all possible subset models using either Cp, AIC, or BIC; find some model with a fairly small value
   - Next best:
     - Use sequential variable selection, like stepwise regression (this doesn't look at all possible subset models, but may be more convenient with some statistical programs)

Sequential Variable Selection

1. Forward selection

   a. Start with no X's "in" the model St 412/512 page 98.

   b. Find the "most significant" additional X (with an F-test).

   c. If its p-value is less than some cutoff (like .05) add it to the model (and re-fit the model with the new set of X's).

   d. Repeat (b) and (c) until no further X's can be added.

2. Backward elimination

   a. Start with all X's "in" the mode.l

   b. Find the "least significant" of the X's currently in the model.

   c. If it's p-value is greater than some cutoff (like .05) drop it from the model (and re-fit with the remaining x's).

   d. Repeat until no further X's can be dropped.

Sequential Variable Selection (cont.)

3. Stepwise regression

a) Start with no X's "in" St 412/512 page 99

b) Do one step of forward selection

c) Do one step of backward elimination

d) Repeat (b) and (c) until no further X's can be added or dropped

4. Notes

a) Add and drop factor indicator variables as a group

b) Don't take p-values and CI's for selected variables seriously—because of serious data snooping (not a problem for objectives 1 and 3)

c) A drawback: the product is a single model. This is deceptive.

- Think not: "here is the best model."
- Think instead: "here is one, possibly useful model."

Criteria for Comparing Models

Some function of the mean square of residuals → $f(\hat{\sigma}^2)$

Some function of the number of β's in the model → $g(p)$

Criterion to minimize	=	$f(\hat{\sigma}^2)$	+	$g(p)$
Cp	=	$\frac{(n-p)(\hat{\sigma}^2 - \hat{\sigma}^2_{full})}{\hat{\sigma}^2_{full}}$	+	p
BIC	=	$n\log(\hat{\sigma}^2)$	+	$p\log(n)$
AIC	=	$n\log(\hat{\sigma}^2)$	+	$2p$

Idea: favor models with small mean square of residuals ↑

but penalize for too many x's ↑

Comparing Models (continued)

1. The proposed criteria: Mallow's Cp Statistic, Schwarz's Bayesian Information Criterion (BIC or SBC), and Akaike's Information Criterion (AIC)

2. The idea behind these is the same, but the theory for arriving at the trade-off between small $\hat{\sigma}^2$ and small p differs

3. My opinion: there's no way to truly say that one of these Criteria is better than the others

4. Computer programs: Fit all possible models; report the best 10 or so according to the selected criteria

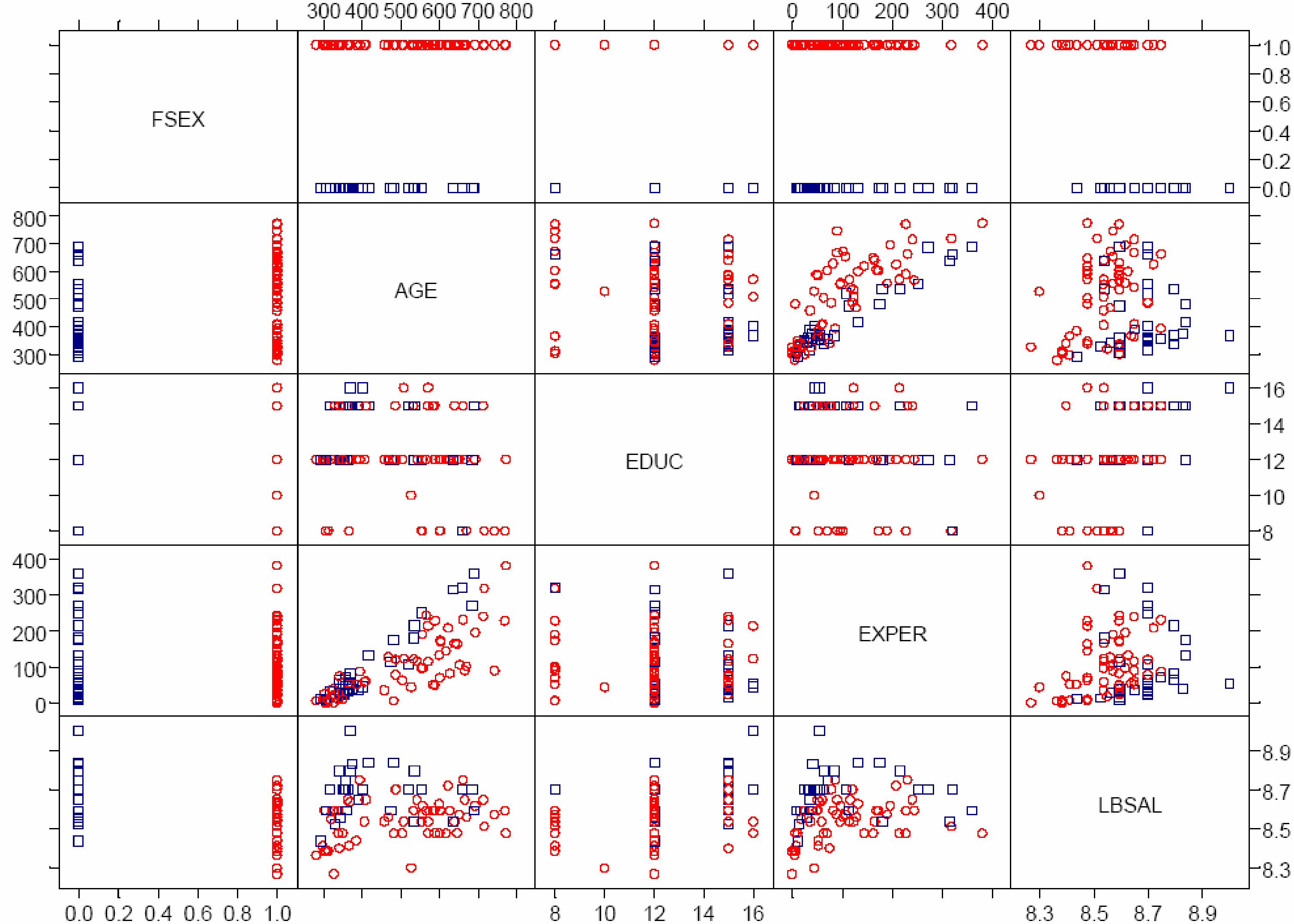
5. Note: one other criteria: $\hat{\sigma}^2$ + 0 (sometimes used; but isn't As good). An equivalent criterion is - $R^2_{Adjusted}$ (see Sect.10.4.1)

Cross Validation (12.6.4)

- If tests, CIs, or prediction intervals are needed after Variable selection and if n is large, maybe try:
- Cross validation
  - Randomly divide the data into 75% for model construction and 25% for inference
  - Perform variable selection with the 75%
  - Refit the same model (don't drop or add anything) on the remaining 25% and proceed with inference using that fit.

Example: Sex Discrimination

Matrix of scatterplots; response= log beginning salary (LBSAL)



Example: Sex Discrimination

2. Correlation matrix

	FSEX	AGE	EDUC	EXPER	LN.BSAL
FSEX	1.00	0.26	-0.33	-0.02	-0.54
AGE	0.26	1.00	-0.23	0.80	0.06
EDUC	-0.33	-0.23	1.00	-0.10	0.41
EXPER	-0.02	0.80	-0.10	1.00	0.19
LN.BSAL	-0.54	0.06	0.41	0.19	1.00

3. Notes from plot 1: possible curvature with respect to age, experience, education (also evident in partial residual plots)

4. Main idea: (1) find a model explaining Log(BSAL) as a function of AGE, EDUC, and EXPER; (2) add in FEMALE indicator and possible interactions

5. Since there are only 3 X's in task (1) and since there is evidence of curvature, try fitting a full 2nd order model (with squared and interaction terms), then manually apply backward elimination—but delete insignificant X^2 and interaction terms *first*

Example: Sex Discrimination

Fit all

Pr(>|t|)

(Intercept)	0.0000
AGE	0.5202
EDUC	0.4592
EXPER	0.0155
I(AGE^2)	0.2246
I(EXPER^2)	0.6947
I(EDUC^2)	0.7242
EXPER:EDUC	0.2968
EXPER:AGE	0.1169
EDUC:AGE	0.4112

Drop EDUC:AGE

Pr(>|t|)

(Intercept)	0.0000
AGE	0.0486
EDUC	0.7506
EXPER	0.0000
I(AGE^2)	0.0540
I(EXPER^2)	0.8225
I(EDUC^2)	0.5782
EXPER:EDUC	0.0051
EXPER:AGE	0.0336

Drop $EXPER^2$

Pr(>|t|)

(Intercept)	0.0000
AGE	0.0316
EDUC	0.7398
EXPER	0.0000
I(AGE^2)	0.0228
I(EDUC^2)	0.5815
EXPER:EDUC	0.0045
EXPER:AGE	0.0001

Drop $EDUC^2$

Pr(>|t|)

(Intercept)	0.0000
AGE	0.0245
EDUC	0.0000
EXPER	0.0000
I(AGE^2)	0.0167
EXPER:EDUC	0.0029
EXPER:AGE	0.0001

6. Now add the FEMALE indicator and possible interactions

Fit all

Pr(>|t|)

(Intercept)	0.0000
AGE	0.1221
EDUC	0.0026
EXPER	0.0001
I(AGE^2)	0.0888
FSEX	0.9362
EXPER:EDUC	0.0297
EXPER:AGE	0.0007
FSEX:AGE	0.8157
FSEX:EDUC	0.2686
FSEX:EXPER	0.9514

Drop FSEX:EXPER

Pr(>|t|)

(Intercept)	0.0000
AGE	0.1119
EDUC	0.0022
EXPER	0.0000
I(AGE^2)	0.0781
FSEX	0.8775
EXPER:EDUC	0.0271
EXPER:AGE	0.0004
FSEX:AGE	0.4052
FSEX:EDUC	0.2627

Drop FSEX:AGE

Pr(>|t|)

(Intercept)	0.0000
AGE	0.1034
EDUC	0.0005
EXPER	0.0000
I(AGE^2)	0.0455
FSEX	0.6340
EXPER:EDUC	0.0140
EXPER:AGE	0.0002
FSEX:EDUC	0.1870

Drop FSEX:EDUC

Pr(>|t|)

(Intercept)	0.0000
AGE	0.0692
EDUC	0.0001
EXPER	0.0000
I(AGE^2)	0.0274
FSEX	0.0000
EXPER:EDUC	0.0173
EXPER:AGE	0.0001

STOP

7.

	COEF	SE	t-STAT	
FSEX	-0.1155	0.0235	-4.9189	(coefficient of FSEX in final model)

Example: Sex Discrimination

8. **Summary: There is overwhelming evidence that the distribution of beginning salary for females was less than that for males, after accounting for the effects of education, age, and previous experience (p-value < .0001). It is estimated that the median salary for females is 89% that of males, after accounting for these other variables (95% confidence interval: 85% to 93% as much).**

9. **Same approach with automatic routines (output next page)**

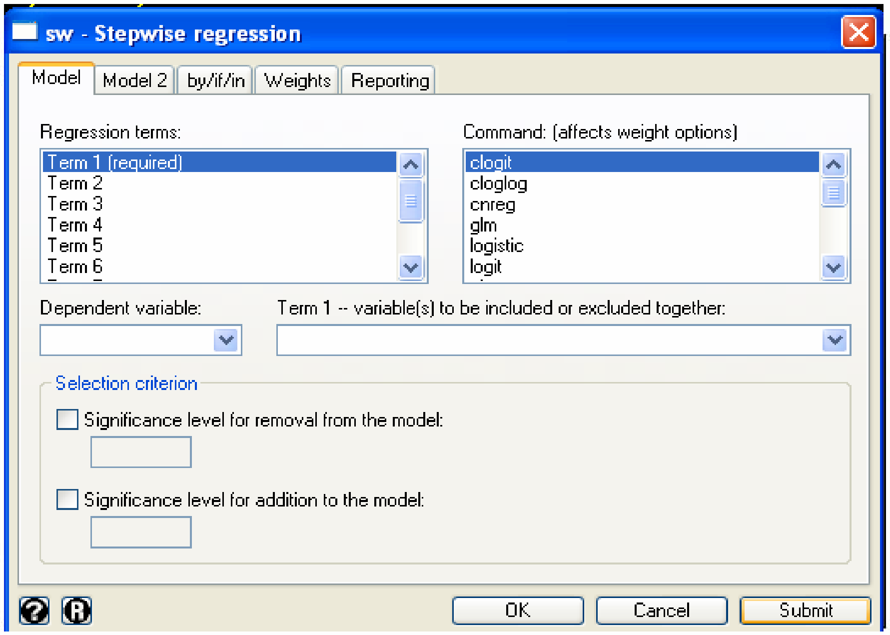
    a. **Automatic backward elimination: 89%**

    b. **Automatic forward selection: 87%**

    c. **Automatic stepwise regression: 89%**

10. **Same approach using all possible subsets and Cp criterion:**

Example: Sex Discrimination



Results from backward elmination (Stepping Direction: backward)

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	8.5381	0.2534	33.6989	0.0000
age	-0.0026	0.0011	-2.2901	0.0245
educ	0.0388	0.0076	5.0766	0.0000
exper	0.0073	0.0013	5.4292	0.0000
I(age^2)	0.0000	0.0000	2.4419	0.0167
age:exper	0.0000	0.0000	-4.2653	0.0001
educ:exper	-0.0002	0.0001	-3.0614	0.0029

①

Results from forward selection (Stepping Direction: Forward)

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	8.4645	0.0684	123.7843	0.0000
I(educ^2)	0.0008	0.0002	3.9628	0.0002
exper	0.0022	0.0005	4.3038	0.0000
I(exper^2)	0.0000	0.0000	-3.8374	0.0002
age	-0.0003	0.0001	-2.0326	0.0451

②

Results from stepwise (Stepping Direction: Both)

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	8.5381	0.2534	33.6989	0.0000
age	-0.0026	0.0011	-2.2901	0.0245
educ	0.0388	0.0076	5.0766	0.0000
exper	0.0073	0.0013	5.4292	0.0000
I(age^2)	0.0000	0.0000	2.4419	0.0167
age:exper	0.0000	0.0000	-4.2653	0.0001
educ:exper	-0.0002	0.0001	-3.0614	0.0029

③

Estimated coef. of FSEX when added into models
1 and 3: -.116 (SE = .023)
2: -.139 (SE = .025)

Female median as % of male median (accounting for other x's):
1 and 3: 89%; 2: 87%

Example: Sex Discrimination

All possible regressions in Stata using commands window

```
> age.ex <- age*exper
> ed.ex <- educ*exper
> age2 <- age*age
> exper2 <- exper*exper
> educ2 <- educ*educ
> x <- cbind(age,educ,exper,age2,exper2,educ2,age.ed,age.ex,ed.ex)
> r <- leaps(x,lbsal,nbest=1)
> terms <- r$label
> Cp  <- round(r$Cp,1)   # Round to one digit to the right of decimal pt.
> print(cbind(terms,Cp))
```

Combine all possible x's into a matrix, using cbind

The "leaps" command computes all possible regression models, with "nbest=1" it retains the model with the smallest Cp for each possible # of betas.

```
     Terms in Model                                          Cp Statistic
[1,] "educ2"                                                  "31.4"
[2,] "exper,educ2"                                            "25.9"
[3,] "exper,educ2,age.ex"                                     "13.9"
[4,] "exper,educ2,age.ex,ed.ex"                               "7.0"
[5,] "educ,exper,age2,age.ed,age.ex"                          "3.6"
[6,] "age,exper,age2,educ2,age.ex,ed.ex"                      "4.8"
[7,] "age,educ,exper,age2,age.ed,age.ex,ed.ex"                "6.3"
[8,] "age,educ,exper,age2,exper2,age.ed,age.ex,ed.ex"         "8.1"
[9,] "age,educ,exper,age2,exper2,educ2,age.ed,age.ex,ed.ex"   "10.0"
```

Choose a model with small Cp. How about # 5?

When FSEX is added into model 5, it's coefficient is -0.113 (SE = 0.024) (same conclusion as in the summary above)

a. **Note: the different variable selection methods lead to essentially the same coefficient for FSEX**

b. **There would also be similar *predictions* from the "good" models**