0hv50:
Behavioral Research Methods 2

Dealing with data

Multiple regression (1)

(canvas.tue.nl)

Chris Snijders

c.c.p.snijders@gmail.com

Today's program

We went through

- Stata, Oncourse
- X→Y

- The general logic behind hypothesis testing (H_0, alpha, ...)

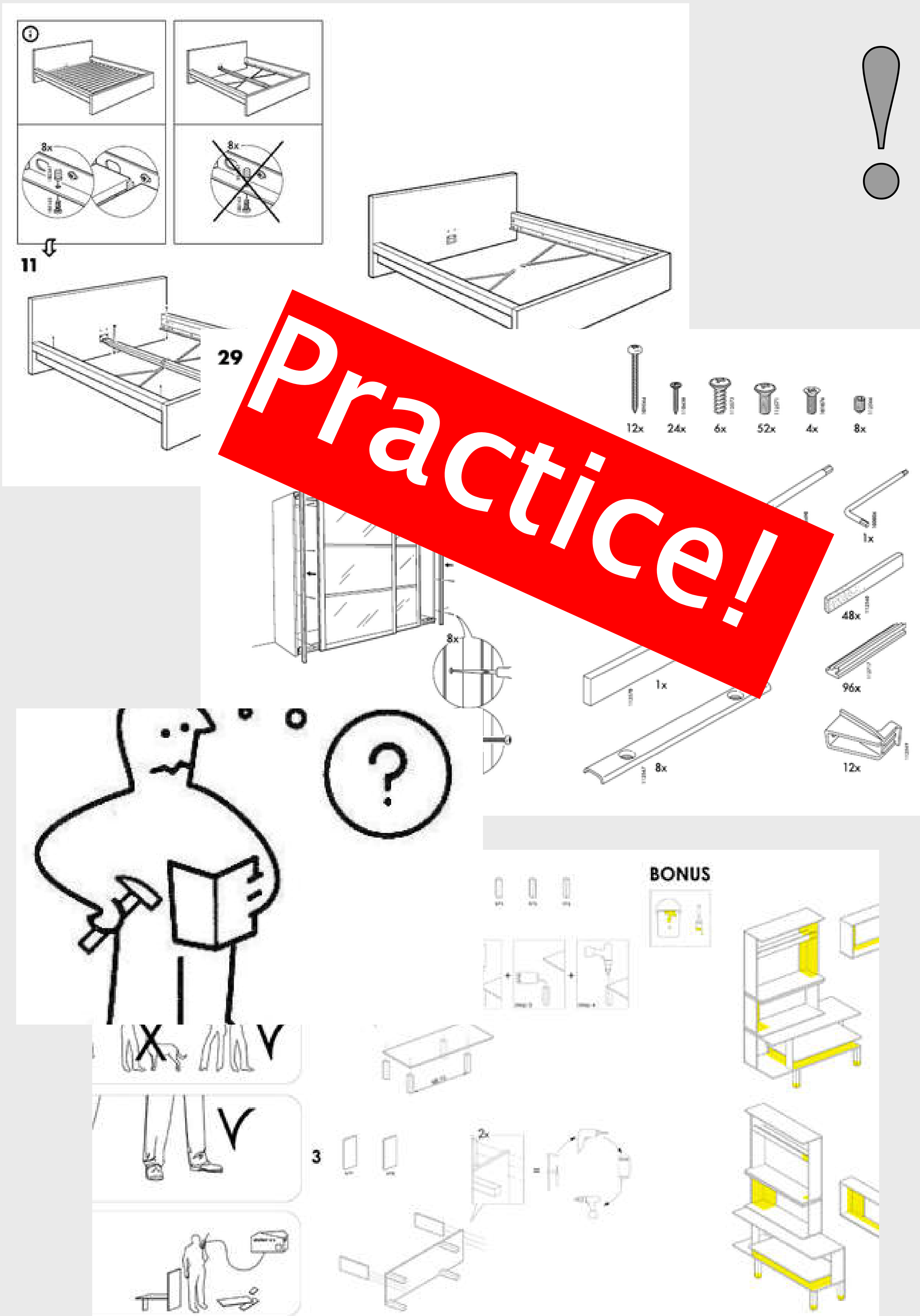
- CAT X CAT: chi2 + Fisher's exact
- CAT(2) X INT: ttest, ranksum, median
- INTERVAL x INTERVAL: pwcorr, reg
- (sample size determination)

And continue with…

Multiple regression
(1Y, more X's)

Stata commands

reg
predict
i.[var]
tab [var], gen(...)
test


Practice!
11
29
12x
24x
6x
52x
4x
8x
1x
48x
96x
12x
BONUS
3
2x

To Do (deadline: 2016)

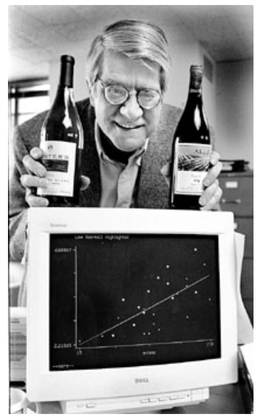
Assignment

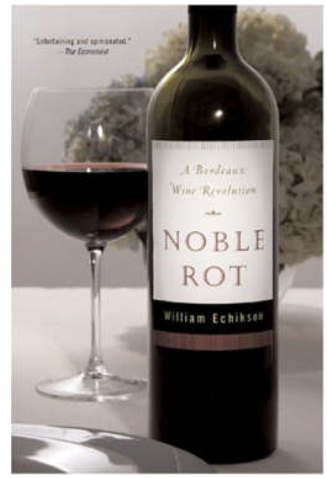
re-take the test (check the separate module), and score at least 31 out of 36

(unlimited attempts)

Wiki data: wine prices

Orley Ashenfelter, Princeton





Y = wine price

X1 = rainfall during the Oct–March

X2 = average summer temperature (Apr/Sept)

X3 = rainfall during the harvest time (Aug/Sept)

(X4 = the wine is a red wine)

(X5 = the type of grape: Pinot Noir / Syrah / Cabernet)

http://www.liquidasset.com/

THIS WEEK's WIKI:
Predict the value of a bottle of wine from rainfall and temperature data: multiple regression.

Multiple regression: what it is

- Y is an interval variable

(“the thing you are trying to predict”)

- X’s can be basically anything:
  - Interval variable
  - (Ordinal variable)
  - Categorical (2 categories)
  - Categorical ($>$2 categories)

  (“the things you use to predict Y with”)

  But: you have to know how to include them in the model!

AFTER TODAY
YOU SHOULD BE ABLE TO:

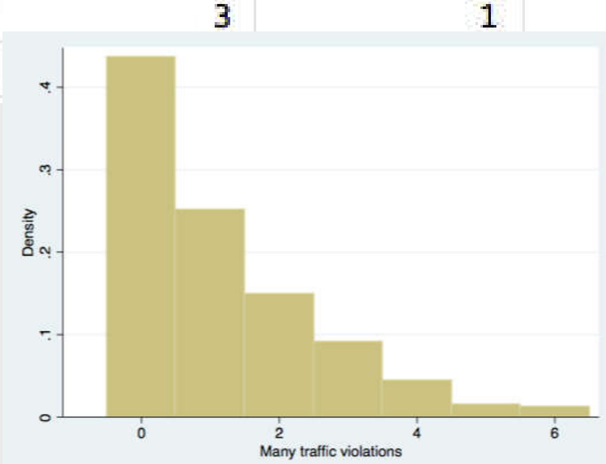
- RUN (SEQUENCES OF) MULTIPLE REGRESSION ANALYSIS

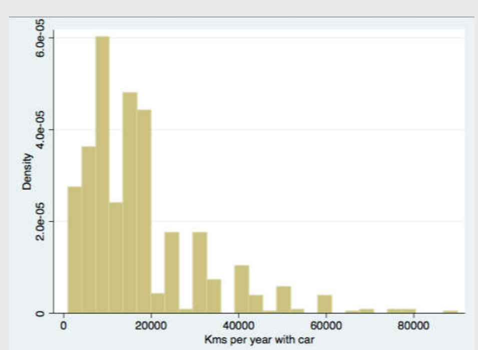
- … INCLUDING THOSE WITH CATEGORICAL VARIABLES

- … AND BEING ABLE TO INTERPRET THE OUTPUT

Main data file: traffic.dta

	dangerous	female	age	kmyear	RELIGION
1	0	1	51	2500	overig chri…
2	0	0	45	5000	geen
3	0	0	36	12000	boeddhistis…
4	0	0	46	15000	geen
5	0	1	41	2000	geen
6	.	0	23	20000	rooms-katho…
7	0	0	54	11000	samen-op-we…
8	0	0	53	6000	geen
9	6	0	25	45000	geen
10	1	0	23	6000	geen
11	0	0	22	2100	geen
12	1	0	22	5000	geen
13	0	0	35	12000	rooms-katho…
14	.	1	35	20000	geen
15	0	1	52	20000	rooms-katho…
16	1	0	61	15000	rooms-katho…
17	2	0	53	15000	rooms-katho…
18	2	0	52	40000	anders, nl.
19	5	0	28	10000	islamitisch
20	1	0	34	15000	geen
21	0	0	51	3000	geen
22	6	0	37	35000	geen
23	4	1	26	40000	geen
24	2	0	43	15000	geen
25	3	1	29	15000	geen
26			32	30000	geen





Multiple regression: predict Y from a set of X's

You have a target variable (Y) that you want to predict using predictor variables X_1 through X_n using:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n$$

where the b_i's have to be found in such a way that the estimated Y is close to the real Y.

Usually there are two reasons to want this:

- Predicting (example: weather / stock market)
- Understanding (example: see traffic-data)

Using this model is usually called

- Multiple regression analyses
- "Ordinary Least Squares" (OLS)

Some typical target variables ...

Y = creditworthiness

Y = likelihood to buy stuff / willingness to pay

Y = likelihood of (e.g., tax) fraud

Y = expected number of hamburgers

Y = “value” of job candidates

Y = social status

– – – – – – – – – – – – – – – – – –

Y = voltage as administered to other

Y = score on IQ test

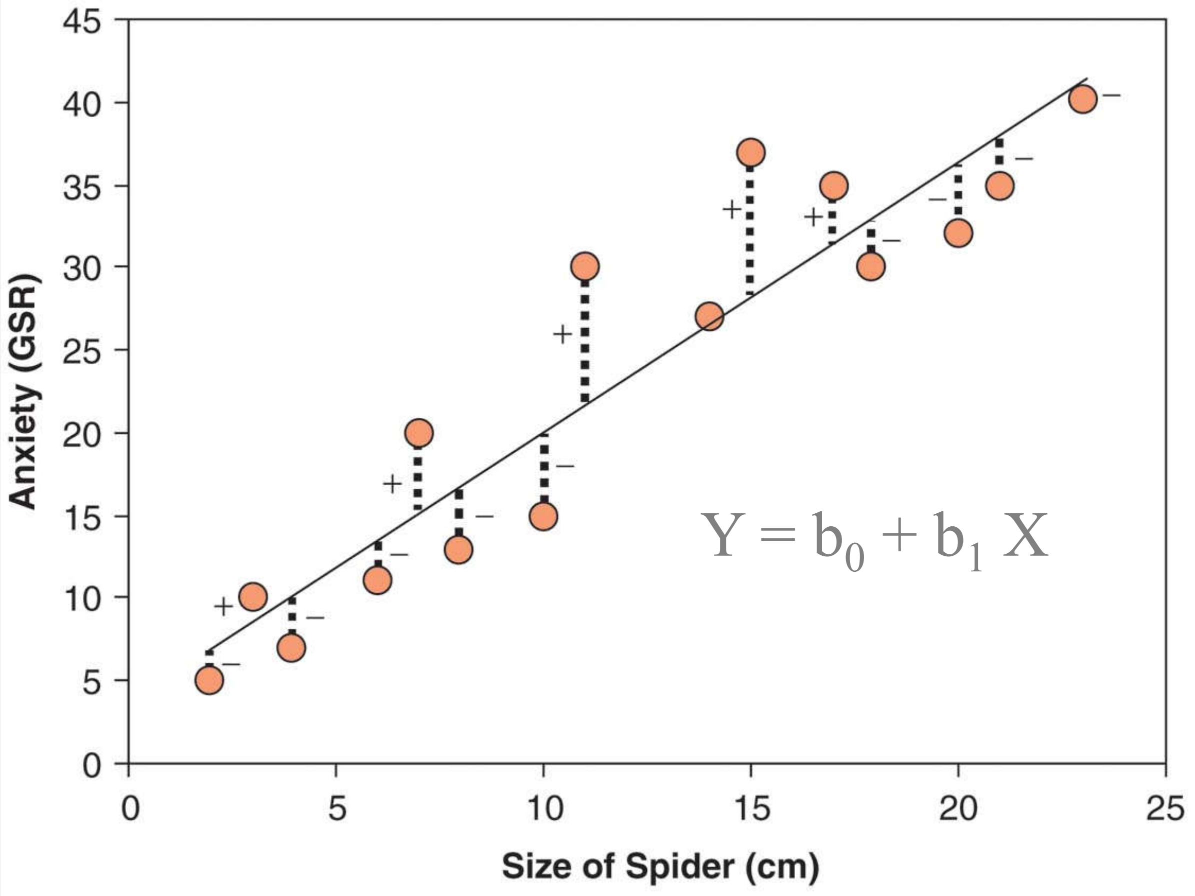
Y = …

Reminder

Simple regression:

Y is interval, and X is interval

Simple regression: Y and one X. Behind the scenes …



"Ordinary least squares":

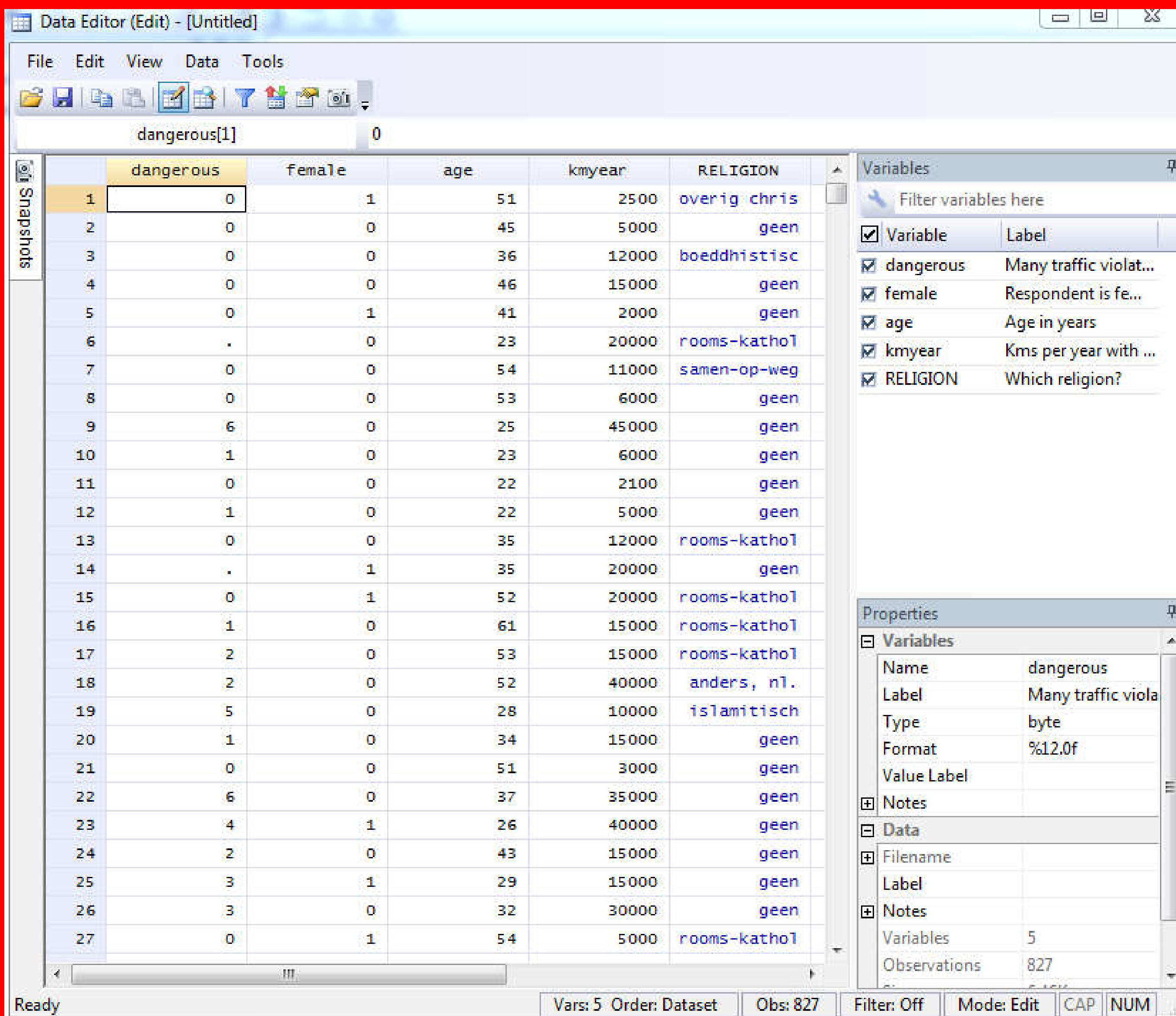
We define a concept of "wrongness", or deviation: it is the distance of the prediction to the real value, squared.

deviation = Σ (observed – model)^2

Choose the b's so that the deviance is minimized.

Today's data:

traffic.dta



	dangerous	female	age	kmyear	RELIGION
1	0	1	51	2500	overig chris
2	0	0	45	5000	geen
3	0	0	36	12000	boeddhistisc
4	0	0	46	15000	geen
5	0	1	41	2000	geen
6	.	0	23	20000	rooms-kathol
7	0	0	54	11000	samen-op-weg
8	0	0	53	6000	geen
9	6	0	25	45000	geen
10	1	0	23	6000	geen
11	0	0	22	2100	geen
12	1	0	22	5000	geen
13	0	0	35	12000	rooms-kathol
14	.	1	35	20000	geen
15	0	1	52	20000	rooms-kathol
16	1	0	61	15000	rooms-kathol
17	2	0	53	15000	rooms-kathol
18	2	0	52	40000	anders, nl.
19	5	0	28	10000	islamitisch
20	1	0	34	15000	geen
21	0	0	51	3000	geen
22	6	0	37	35000	geen
23	4	1	26	40000	geen
24	2	0	43	15000	geen
25	3	1	29	15000	geen
26	3	0	32	30000	geen
27	0	1	54	5000	rooms-kathol

Different kinds of X-vars

Y	X
dangerous	kmsperyear (INT)
dangerous	female (CAT-2)
dangerous	kmsperyear & female

Example: One Y, one (binary) X

Suppose I want to predict some target Y

For instance:

Y = number of regularly committed traffic violations (out of 7)   variable <dangerous>

My first guess: an important predictor is *gender*. Males are more reckless drivers so they will make more traffic violations

So X_1 = female, equal to 1 when the respondent is female and 0 otherwise, and the model with the best fit (this you get from Stata) is:

dangerous = 1.48 – 0.66 female

And this implies ...

$dangerous = 1.48 - 0.66\ female$

So my best estimate for females equals:

$dangerous = 1.48 - 0.66 * 1 = 0.82$

and for males we get

$dangerous = 1.48 - 0.66 * 0 = 1.48$

NOTE

Gender has two categories and:

1/ we do not label the variable <gender>, but choose a name that implies the direction of the coding

2/ we need only one variable, even though we have two categories

3/ as a prediction, this (obviously) totally sucks

2 categories, 1 dummy

Including both the variables MALE and FEMALE is in fact not only not helping, it is impossible:

ONLY FEMALE:

DANGEROUS

$= c0 + c1 \text{ FEMALE}$

BOTH MALE AND FEMALE:

DANGEROUS

$= b0 + b1 \text{ FEMALE} + b2 \text{ MALE}$

$= b0 + b1 \text{ FEMALE} + b2 (1 - \text{FEMALE})$

$= b0 + b2 + (b1 - b2) \text{ FEMALE}$

And we end up with an unidentified system:

(for instance (1,1,1) and (2,0,0) are the same model) .

(continued)

Possible additional argument: “We have an intervening variable here. Males tend to drive more kms per year. So the difference that you find is not because of gender differences, but because men drive more kms per year.”

[Solution 1] Split the data in two groups: <high mileage> and <low mileage>. Run simple regression analysis separately for both groups. This is possible, but has serious drawbacks. **Why?**

[Solution 2] Multiple regression: include <kms per year> as a second predictor.

$$\text{dangerous} = b_0 + b_1 \text{ male} + b_2 \text{ kmsperyear}$$

If we find that the b_1 variable is now much closer to zero, we have shown that it is not gender that shows the effect, but instead how often you drive (“explaining away the effect of gender”).

(this is one of the reasons why we want MULTIPLE regression: “explaining away”)

...and this is what we get

```
. reg dangerous female

      Source |       SS       df       MS              Number of obs =     720
-------------+------------------------------           F(  1,   718) =   44.61
       Model |  79.3955695     1  79.3955695           Prob > F      =  0.0000
    Residual |  1277.79887   718  1.77966417           R-squared     =  0.0585
-------------+------------------------------           Adj R-squared =  0.0572
       Total |  1357.19444   719  1.88761397           Root MSE      =   1.334

------------------------------------------------------------------------------
   dangerous |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.6641665    .099437    -6.68   0.000    -.8593884   -.4689445
       _cons |   1.482094    .070019    21.17   0.000     1.344627     1.61956
------------------------------------------------------------------------------
```

```
. reg dangerous female kmyear

      Source |       SS       df       MS              Number of obs =     720
-------------+------------------------------           F(  2,   717) =   35.36
       Model |  121.844139     2  60.9220693           Prob > F      =  0.0000
    Residual |  1235.35031   717  1.72294324           R-squared     =  0.0898
-------------+------------------------------           Adj R-squared =  0.0872
       Total |  1357.19444   719  1.88761397           Root MSE      =  1.3126

------------------------------------------------------------------------------
   dangerous |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.5050933   .1029546    -4.91   0.000    -.7072218   -.3029649
      kmyear |   .0000196   3.94e-06     4.96   0.000     .0000118    .0000273
       _cons |   1.072389   .1075154     9.97   0.000     .8613068    1.283472
------------------------------------------------------------------------------
```

The original effect of -0.66 diminished to -0.5
After inclusion of the [kmyear] variable.

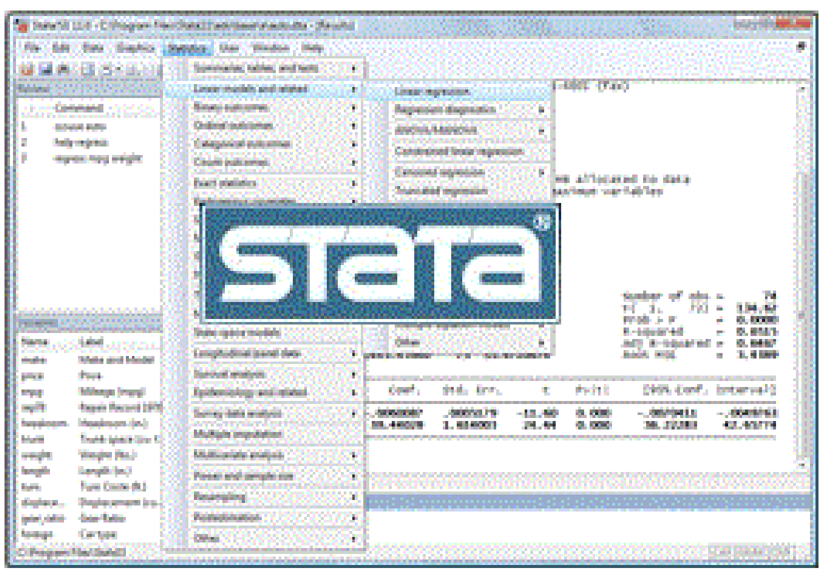
1. MR allows inclusion of more than 1 var
2. Estimated coefficients show net effects
   (“while controlling for other vars”)
3. Subsequent MR’s allow understanding of effects

Some background info on multiple regression

Any statistical software can run multiple regression

- Stata
- Alternatives for Stata (MiniTab, GLIM, SPSS, Statistica, Systat…)
- Several freeware packages (for instance *R*, PSPP)
- In Excel, straight away or using plug-ins (for instance *PopTools*, which is also freeware)

(We use Stata)



Why linear?

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n$$

Especially in the social sciences, **you often do not have a more precise equation** for the relation between X's and Y. Most of the time, we have an idea of the kind "if X increases, then Y is likely to increase", without any specific idea about the shape of the relation. A linear model is a good start.

Small print: and: even if you have a concrete non-linear equation, often you can find a linear approximation (using Taylor-expansion, for instance) that is good enough for all practical purposes.

Moreover, the equation is linear given the predictors, but the **predictors themselves can be non-linear**! So the linearity is not that restrictive anyway. For instance:

$$\text{dangerous} = b_0 + b_1 \text{kmspyear} + b_2 \text{kmspyear}^2$$

But this can't be estimated with multiple regression ...

$$y = \frac{b_0 + b_1 x_1 + b_2 \dfrac{\cos(b_3 + b_4 x_4)}{\log(\sqrt{b_5 + b_6 \sin(x_6)})}}{\int \sqrt{\arctan(b_7 + b_8 x_8)}}$$

(although it could be estimated using something called nonlinear regression)

Why is it beautiful …

[1] You can test hypotheses about effects of predictors on targets (Xs on Y), while taking into account possibly intervening factors

[2] It combines several “separate models” into a single analysis.

Y compared between two groups:
→ *t-test*

Y compared between three groups:
→ *anova*

Y compared between three groups and two treatments
→ *(blocked) anova*

Y predicted by an interval X
→ *correlation*

All of these (and more) can be done with multiple regression.

[3] more complicated methods are usually a logical consequence of multiple regression.

regression vs t-test

```
. reg dangerous female

      Source |       SS           df       MS      Number of obs   =       720
-------------+----------------------------------   F(1, 718)       =     44.61
       Model |  79.3955695         1  79.3955695   Prob > F        =    0.0000
    Residual |  1277.79887       718  1.77966417   R-squared       =    0.0585
-------------+----------------------------------   Adj R-squared   =    0.0572
       Total |  1357.19444       719  1.88761397   Root MSE        =     1.334

------------------------------------------------------------------------------
   dangerous |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.6641665    .099437    -6.68   0.000    -.8593884   -.4689445
       _cons |   1.482094    .070019    21.17   0.000     1.344627     1.61956
------------------------------------------------------------------------------
```

```
. ttest dangerous, by(female)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     363    1.482094    .0800777    1.525686    1.324618     1.63957
       1 |     357    .8179272    .0585151     1.10561    .7028484    .9330059
---------+--------------------------------------------------------------------
combined |     720    1.152778    .0512024    1.373905    1.052254    1.253302
---------+--------------------------------------------------------------------
    diff |            .6641665     .099437                .4689445    .8593884
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   6.6793
Ho: diff = 0                                     degrees of freedom =      718

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Notations / definitions

Notation: the OLS-estimator for Y, “Y hat”

$$\hat{Y} = \hat{b_0} + \hat{b_1} X_1 + \hat{b_2} X_2 + \ldots + \hat{b_n} X_n$$

is calculated by choosing values for b_i (“b_i hat”) so that

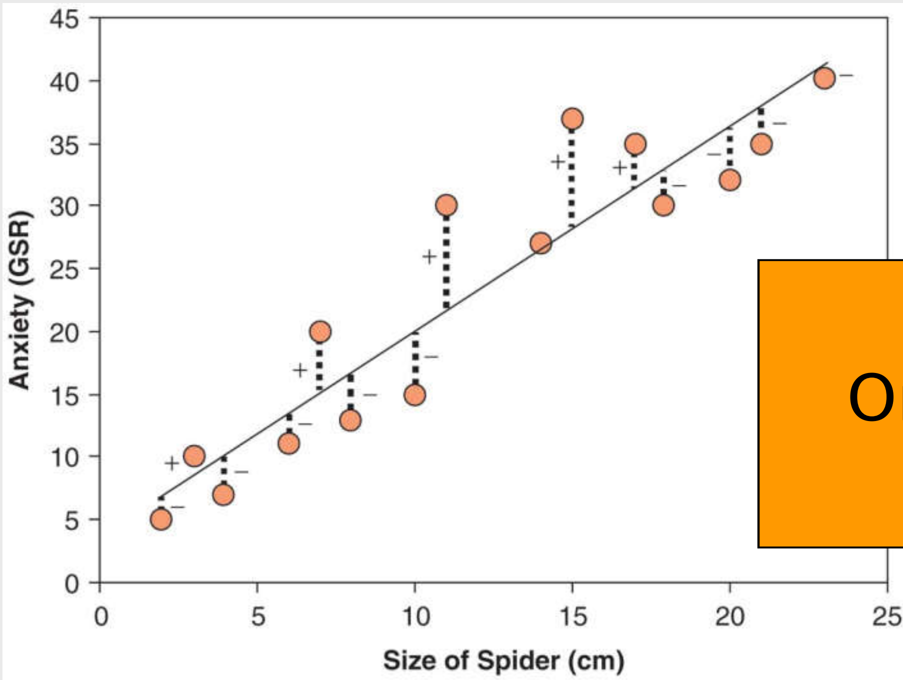
$$\text{deviance} = \sum_{\text{obs}} (Y_i - \hat{Y}_i)^2$$

is minimal, as with simple regression.
(= SSR sum of squared residuals)

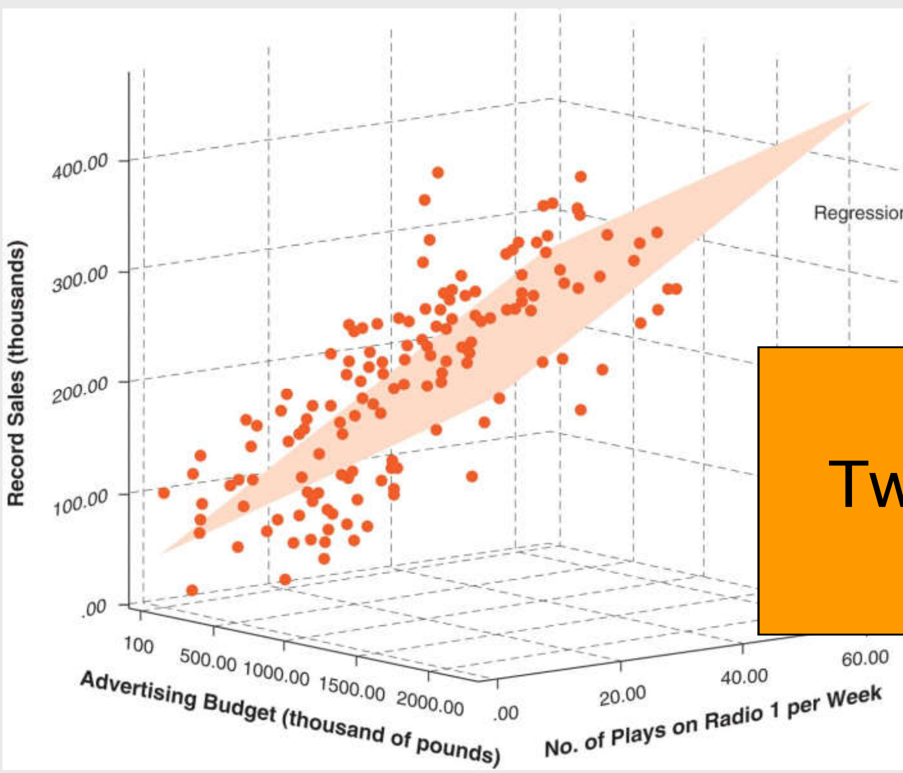
$$\text{error} = Y - \hat{Y}$$

And in principle, other measures of deviance are possible → different kinds of regression

Visually, this is ...



One X



Two Xs

Model fit

How well does the model fit?

Two ways to assess model fit

[1] Through the sum of squared errors (SS_R):

$$1 - \frac{SS_R(\text{full model})}{SS_T(\text{model with just } b_0)}$$

[2] Through correlation

$$(\text{correlation}(y, \hat{y}))^2$$

Note that in both cases $0 \leq \text{value} \leq 1$

And: [1] and [2] are **the same**, and called R^2

About R^2 and adjusted R^2

Intuitively: it is easier to get higher R^2 values when you have more predictor variables X.

Moreover, if you only have a handful of cases, your R^2 can be high just coincidentally.

To compare between different models (and data sets) we use "adjusted R^2", which takes into account the number of X's (*p*) and cases (*n*) you have used:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n-1)}{n-p-1}$$

$$R^2_{adj} = R^2 - (1 - R^2)\,\frac{p}{n-p-1}$$

Reminder: R^2 is *not* an absolute criterion, you can have a high R^2 but have learned nothing (and even a low R^2 and have learned something).

Let's check:

```
. reg dangerous female kmyear

      Source |       SS       df       MS              Number of obs =     720
-------------+------------------------------           F(  2,   717) =   35.36
       Model |  121.844139     2  60.9220693           Prob > F      =  0.0000
    Residual |  1235.35031   717  1.72294324           R-squared     =  0.0898
-------------+------------------------------           Adj R-squared =  0.0872
       Total |  1357.19444   719  1.88761397           Root MSE      =  1.3126

------------------------------------------------------------------------------
   dangerous |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.5050933   .1029546    -4.91   0.000    -.7072218   -.3029649
      kmyear |   .0000196   3.94e-06     4.96   0.000     .0000118    .0000273
       _cons |   1.072389   .1075154     9.97   0.000     .8613068    1.283472
------------------------------------------------------------------------------
```

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n-1)}{n-p-1}$$

$$R^2_{adj} = 1 - \frac{(1 - 0.0898)(720-1)}{720-2-1}$$

$$R^2_{adj} = 0.0872$$

In comes the statistics…

(and this only happens because we want to say something about the population)

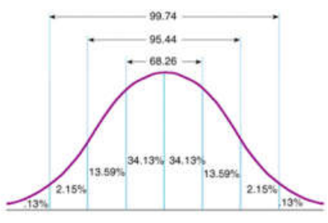
From sample to population

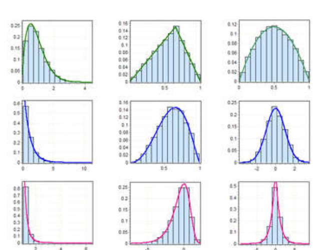
For several reasons, the best fitting values b-hat are not completely equal to their actual values in the population:

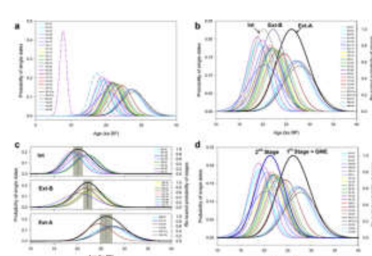
(NB only here the statistics comes in!)

[1] “Measurement error”
[2] “Sampling error”
[3] “Uncontrolled variance”

How can we say something about the value of the b_i in the population? We need some more assumptions …







Multiple regression:

$$y = b_0 + b_1 x_1 + \ldots + b_n x_n + \epsilon$$

with ϵ distributed as $N(0, \sigma^2)$

and ϵ does not depend on any x_i

And this implies that after running your multiple regression, you need to test whether these assumptions are met (more on those later).

For now:

And when this assumption is met…

Given that, you cannot only find best fitting values for b_i

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \ldots + \hat{b}_n x_n$$

but also test, for each coefficient
H_0: the coefficient in the population equals zero

Statistics programs give you:

- the values "b_i-hat"

and for each estimated coefficient

- a t-value (the "test statistic")
- a p-value (the estimated probability …)
- a (95%) confidence interval

As always, the p-value represents the probability to end up with the data that you have (or further away from H_0), given that H_0 holds.

Same rule: when $p<0.05$, we reject H_0.

Going through a regression table

```
. reg dangerous female

      Source |       SS       df       MS              Number of obs =     720
-------------+------------------------------           F(  1,   718) =   44.61
       Model |  79.3955695     1  79.3955695           Prob > F      =  0.0000
    Residual |  1277.79887   718  1.77966417           R-squared     =  0.0585
-------------+------------------------------           Adj R-squared =  0.0572
       Total |  1357.19444   719  1.88761397           Root MSE      =   1.334

------------------------------------------------------------------------------
   dangerous |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.6641665    .099437    -6.68   0.000    -.8593884   -.4689445
       _cons |   1.482094    .070019    21.17   0.000     1.344627     1.61956
------------------------------------------------------------------------------
```

Coefficients:
Dangerous = 1.48 – 0.66 female

Statistical tests
H0: coefficient=0

The other ones are related to the statistical tests.
(-0.66 +/- 1.96*0.099)

Sums of squares:
How far of with the model, compared to a base model

test–ing different H_0's

```
. reg dangerous female

      Source |       SS       df       MS              Number of obs =     720
-------------+------------------------------           F(  1,   718) =   44.61
       Model |  79.3955695     1  79.3955695           Prob > F      =  0.0000
    Residual |  1277.79887   718  1.77966417           R-squared     =  0.0585
-------------+------------------------------           Adj R-squared =  0.0572
       Total |  1357.19444   719  1.88761397           Root MSE      =   1.334

------------------------------------------------------------------------------
   dangerous |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.6641665    .099437    -6.68   0.000    -.8593884   -.4689445
       _cons |   1.482094    .070019    21.17   0.000     1.344627     1.61956
------------------------------------------------------------------------------
```

You could test for different H0's if you want:

```
. test female = -0.5

 ( 1)  female = -.5

       F(  1,   718) =    2.73
            Prob > F =    0.0992
```

Or, another form of test

```
. reg dangerous female kmyear

      Source |       SS       df       MS              Number of obs =     720
-------------+------------------------------           F(2, 717)     =   35.36
       Model |  121.844139     2  60.9220693           Prob > F      =  0.0000
    Residual |  1235.35031   717  1.72294324           R-squared     =  0.0898
-------------+------------------------------           Adj R-squared =  0.0872
       Total |  1357.19444   719  1.88761397           Root MSE      =  1.3126

------------------------------------------------------------------------------
   dangerous |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.5050933   .1029546    -4.91   0.000    -.7072218   -.3029649
      kmyear |   .0000196   3.94e-06     4.96   0.000     .0000118    .0000273
       _cons |   1.072389   .1075154     9.97   0.000     .8613068    1.283472
------------------------------------------------------------------------------
```

You could test for different H0's if you want:

```
. test female = kmyear

 ( 1)  female - kmyear = 0

       F(  1,   717) =   24.07
            Prob > F =    0.0000
```

Confidence intervals for the estimated coefficients

```
. reg dangerous female kmyear

      Source |       SS       df       MS              Number of obs =     720
-------------+------------------------------           F(  2,   717) =   35.36
       Model |  121.844139     2  60.9220693           Prob > F      =  0.0000
    Residual |  1235.35031   717  1.72294324           R-squared     =  0.0898
-------------+------------------------------           Adj R-squared =  0.0872
       Total |  1357.19444   719  1.88761397           Root MSE      =  1.3126

------------------------------------------------------------------------------
   dangerous |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.5050933   .1029546    -4.91   0.000    -.7072218   -.3029649
      kmyear |   .0000196   3.94e-06     4.96   0.000     .0000118    .0000273
       _cons |   1.072389   .1075154     9.97   0.000     .8613068    1.283472
------------------------------------------------------------------------------
```

- Confidence interval for coefficient of [female] is (–0.707, - 0.303)

$$-0.707 = -0.505 - 1.96 * 0.103$$

$$-0.303 = -0.505 + 1.96 * 0.103$$

... and the 1.96 is coming from the normal distribution.

Including different kinds of variables

(just categorical variables are a nuisance, the rest is easy)

So, once more …

$$y = b_0 + b_1 x_1 + \ldots + b_n x_n$$

- Y has to be an interval variable

- X can be basically anything:
  - Interval
  - (Ordinal)
  - Categorical (2 categories)
  - Categorical (>2 categories)

But: you have to know how to include a categorical variable in the model!

Including a categorical variable with more than 2 categories

Suppose you want to add [religion] as a predictor for the traffic violations.

Religion has 9 categories in the data.

```
. fre religion

religion — Which religion?
```

		Freq.	Percent
Valid	1 geen	477	57.68
	2 rooms-katholiek	164	19.83
	3 samen-op-weg (of protestantse kerk in nederland)	64	7.74
	4 overig christelijk	74	8.95
	5 islamitisch	9	1.09
	6 hindoeistisch	1	0.12
	7 boeddhistisch	3	0.36
	8 joods	2	0.24
	9 anders, nl.	33	3.99
	Total	827	100.00

Let’s reduce it to just the 5 largest categories:

1 – none

2 – roman catholics

3 – protestant

4 – other Christians

5 – all others

```
. recode religion (1=1)(2=2)(3=3)(4=4)(5 6 7 8 9=5), gen(reliL5)
(39 differences between religion and reliL5)
```

```
. tab reliL5

  RECODE of |
   religion |
     (Which |
 religion?) |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        477       57.68       57.68
          2 |        164       19.83       77.51
          3 |         64        7.74       85.25
          4 |         74        8.95       94.20
          5 |         48        5.80      100.00
------------+-----------------------------------
      Total |        827      100.00
```

```
. label var reliL5 "1=none/2=romancath/3=prot/4=othChris/5=allothers"

. tab reliL5

1=none/2=ro |
mancath/3=p |
rot/4=othCh |
ris/5=allot |
       hers |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        477       57.68       57.68
          2 |        164       19.83       77.51
          3 |         64        7.74       85.25
          4 |         74        8.95       94.20
          5 |         48        5.80      100.00
------------+-----------------------------------
      Total |        827      100.00
```

Including a categorical variable with more than 2 categories

Suppose you want to add [religion] as a predictor for the traffic violations.

Religion has **5** categories in our data.

What you do is: you create 5 dummy-variables:

$$\text{religion1} = \begin{cases} 1 \text{ if religion} = 1 \\ 0 \text{ otherwise} \end{cases}$$

etc.

Now you add **4** binary predictors to your regression equation! (one less than you have categories) WHY IS THAT?

This does give rise to some interpretation issues

What NOT to do

Adding a categorical variable "as is"

```
. reg dang female kmyear reliL5

      Source |       SS           df       MS      Number of obs   =       720
-------------+----------------------------------   F(3, 716)       =     23.54
       Model |  121.85827         3  40.6194234   Prob > F        =    0.0000
    Residual |  1235.33617       716  1.72532985   R-squared       =    0.0898
-------------+----------------------------------   Adj R-squared   =    0.0860
       Total |  1357.19444       719  1.88761397   Root MSE        =    1.3135

------------------------------------------------------------------------------
   dangerous |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.5047758   .1030856    -4.90   0.000    -.7071619   -.3023896
      kmyear |   .0000196   3.95e-06     4.96   0.000     .0000118    .0000273
      reliL5 |   .0036885   .0407554     0.09   0.928    -.0763259    .0837029
       _cons |   1.065379   .1325718     8.04   0.000     .8051033    1.325655
------------------------------------------------------------------------------
```

```
. tab reliL5

1=none/2=ro |
mancath/3=p |
rot/4=othCh |
ris/5=allot |
       hers |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        477       57.68       57.68
          2 |        164       19.83       77.51
          3 |         64        7.74       85.25
          4 |         74        8.95       94.20
          5 |         48        5.80      100.00
------------+-----------------------------------
      Total |        827      100.00
```

Stata won't tell you, but this is nonsense
(try interpreting the coefficient)

Creating “dummy-vars” in Stata (all ok)

```
tab reliL5, gen(r)


gen r1 = (reliL5==1)
gen r2 = (reliL5==2)
gen r3 = (reliL5==3)
gen r4 = (reliL5==4)
gen r5 = (reliL5==5)


forvalues i=1/5 {
      gen r`i’ = (reliL5==`i’)
}


xi i.reliL5
```

(nb this last one creates only 4 categories)

Adding categorical predictors

$$\text{danger} = b_0 + b_1\text{female} + b_2\text{kmsperyear} + \ldots$$

$$+c_2\text{reli}_2 + \ldots + c_5\text{reli}_5$$

```
reg dange fem km r2 r3 r4 r5
```

Source	SS	df	MS
Model	127.349847	6	21.2249746
Residual	1229.8446	713	1.72488723
Total	1357.19444	719	1.88761397

Number of obs	=	720
F(6, 713)	=	12.31
Prob > F	=	0.0000
R-squared	=	0.0938
Adj R-squared	=	0.0862
Root MSE	=	1.3133

dangerous	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.507454	.1034583	-4.90	0.000	-.7105733	-.3043346
kmyear	.0000196	3.94e-06	4.96	0.000	.0000118	.0000273
r2	-.0778179	.1279803	-0.61	0.543	-.3290812	.1734454
r3	-.1864143	.1804184	-1.03	0.302	-.5406292	.1678007
r4	.2031481	.1789272	1.14	0.257	-.1481391	.5544353
r5	-.0711879	.2253719	-0.32	0.752	-.5136597	.371284
_cons	1.090484	.1170245	9.32	0.000	.8607305	1.320238

(sidenote)

```
. reg dangerous female kmyear r1 r2 r3 r4 r5
note: r3 omitted because of collinearity
```

Source	SS	df	MS			
				Number of obs	=	720
				F(6, 713)	=	12.31
Model	127.349847	6	21.2249746	Prob > F	=	0.0000
Residual	1229.8446	713	1.72488723	R-squared	=	0.0938
				Adj R-squared	=	0.0862
Total	1357.19444	719	1.88761397	Root MSE	=	1.3133

dangerous	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.507454	.1034583	-4.90	0.000	-.7105733	-.3043346
kmyear	.0000196	3.94e-06	4.96	0.000	.0000118	.0000273
r1	.1864143	.1804184	1.03	0.302	-.1678007	.5406292
r2	.1085964	.2014052	0.54	0.590	-.2868218	.5040145
r3	0	(omitted)				
r4	.3895624	.2368902	1.64	0.101	-.0755233	.8546481
r5	.1152264	.2743848	0.42	0.675	-.4234724	.6539252
_cons	.9040701	.1910915	4.73	0.000	.5289009	1.279239

Adding categorical predictors

```
reg dange fem km r2 r3 r4 r5
```

dangerous	Coef.
female	-.507454
kmyear	.0000196
r2	-.0778179
r3	-.1864143
r4	.2031481
r5	-.0711879
_cons	1.090484

- Tell me [female], [kmyear] and the [reliL5] category and I will give you a prediction
- If female, then 0.5 lower score on [dangerous]
- If 10.000 km/year more, then 0.196 higher on [dangerous]

Adding categorical predictors

```
reg dange fem km r2 r3 r4 r5
```

dangerous	Coef.
female	-.507454
kmyear	.0000196
r2	-.0778179
r3	-.1864143
r4	.2031481
r5	-.0711879
_cons	1.090484

It's different for the dummy-variables …

Let's come up with predictions per religious category, say, for males who drive 10.000 kms per year:

```
r1:   1.09+0*-0.5 + 0.196 + 0
r2:   1.09+0*-0.5 + 0.196 - 0.0778
r3:   1.09+0*-0.5 + 0.196 - 0.1864
r4:   1.09+0*-0.5 + 0.196 + 0.2031
r5:   1.09+0*-0.5 + 0.196 - 0.0712
```

Adding categorical predictors

```
reg dange fem km r2 r3 r4 r5
```

dangerous	Coef.
female	-.507454
kmyear	.0000196
r2	-.0778179
r3	-.1864143
r4	.2031481
r5	-.0711879
_cons	1.090484

You indeed need only 4 (not 5).

The coefficients of the categories represent the difference between the given category and the one that you left out!

Let's come up with pred religious category, say, drive 10.000 kms per y

```
r1:   1.09+0*-0.5 + 0.196 + 0
r2:   1.09+0*-0.5 + 0.196 - 0.0778
r3:   1.09+0*-0.5 + 0.196 - 0.1864
r4:   1.09+0*-0.5 + 0.196 + 0.2031
r5:   1.09+0*-0.5 + 0.196 - 0.0712
```

Does it make a difference which category you leave out?

```
. reg dang female kmyear r2 r3 r4 r5

      Source |       SS           df       MS      Number of obs   =       720
-------------+----------------------------------   F(6, 713)       =     12.31
       Model |  127.349847         6  21.2249746   Prob > F        =    0.0000
    Residual |   1229.8446       713  1.72488723   R-squared       =    0.0938
-------------+----------------------------------   Adj R-squared   =    0.0862
       Total |  1357.19444       719  1.88761397   Root MSE        =    1.3133

------------------------------------------------------------------------------
   dangerous |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -.507454   .1034583    -4.90   0.000    -.7105733   -.3043346
      kmyear |   .0000196   3.94e-06     4.96   0.000     .0000118    .0000273
          r2 |  -.0778179   .1279803    -0.61   0.543    -.3290812    .1734454
          r3 |  -.1864143   .1804184    -1.03   0.302    -.5406292    .1678007
          r4 |   .2031481   .1789272     1.14   0.257    -.1481391    .5544353
          r5 |  -.0711879   .2253719    -0.32   0.752    -.5136597     .371284
       _cons |   1.090484   .1170245     9.32   0.000     .8607305    1.320238
------------------------------------------------------------------------------
```

```
. reg dang female kmyear r1 r3 r4 r5

      Source |       SS           df       MS      Number of obs   =       720
-------------+----------------------------------   F(6, 713)       =     12.31
       Model |  127.349847         6  21.2249746   Prob > F        =    0.0000
    Residual |   1229.8446       713  1.72488723   R-squared       =    0.0938
-------------+----------------------------------   Adj R-squared   =    0.0862
       Total |  1357.19444       719  1.88761397   Root MSE        =    1.3133

------------------------------------------------------------------------------
   dangerous |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -.507454   .1034583    -4.90   0.000    -.7105733   -.3043346
      kmyear |   .0000196   3.94e-06     4.96   0.000     .0000118    .0000273
          r1 |   .0778179   .1279803     0.61   0.543    -.1734454    .3290812
          r3 |  -.1085964   .2014052    -0.54   0.590    -.5040145    .2868218
          r4 |    .280966   .2001641     1.40   0.161    -.1120156    .6739476
          r5 |     .00663   .2429129     0.03   0.978      -.47028    .4835401
       _cons |   1.012667   .1450566     6.98   0.000     .7278774    1.297456
------------------------------------------------------------------------------
```

Does it make a difference which category you leave out?

```
. reg dang female kmyear r2 r3 r4 r5
```

Source	SS	df	MS
Model	127.349847	6	21.2249746
Residual	1229.8446	713	1.72488723
Total	1357.19444	719	1.88761397

Number of obs	=	720
F(6, 713)	=	12.31
Prob > F	=	0.0000
R-squared	=	0.0938
Adj R-squared	=	0.0862
Root MSE	=	1.3133

dangerous	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.507454	.1034583	-4.90	0.000	-.7105733	-.3043346
kmyear	.0000196	3.94e-06	4.96	0.000	.0000118	.0000273
r2	-.0778179	.1279803	-0.61	0.543	-.3290812	.1734454
r3	-.1864143	.1804184	-1.03	0.302	-.5406292	.1678007
r4	.2031481	.1789272	1.14	0.257	-.1481391	.5544353
r5	-.0711879	.2253719	-0.32	0.752	-.5136597	.371284
_cons	1.090484	.1170245	9.32	0.000	.8607305	1.320238

```
. reg dang female kmyear r1 r3 r4 r5
```

Source	SS	df	MS
Model	127.349847	6	21.2249746
Residual	1229.8446	713	1.72488723
Total	1357.19444	719	1.88761397

Number of obs	=	720
F(6, 713)	=	12.31
Prob > F	=	0.0000
R-squared	=	0.0938
Adj R-squared	=	0.0862
Root MSE	=	1.3133

dangerous	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.507454	.1034583	-4.90	0.000	-.7105733	-.3043346
kmyear	.0000196	3.94e-06	4.96	0.000	.0000118	.0000273
r1	.0778179	.1279803	0.61	0.543	-.1734454	.3290812
r3	-.1085964	.2014052	-0.54	0.590	-.5040145	.2868218
r4	.280966	.2001641	1.40	0.161	-.1120156	.6739476
r5	.00663	.2429129	0.03	0.978	-.47028	.4835401
_cons	1.012667	.1450566	6.98	0.000	.7278774	1.297456

Does it make a difference which category you leave out?

dangerous	Coef.
female	-.507454
kmyear	.0000196
r2	-.0778179
r3	-.1864143
r4	.2031481
r5	-.0711879
_cons	1.090484

dangerous	Coef.
female	-.507454
kmyear	.0000196
r1	.0778179
r3	-.1085964
r4	.280966
r5	.00663
_cons	1.012667

Difference between

r2 and r1 = −0.0778

r3 and r2 = −0.0778 – (−0.1864) = 0.1086

r4 and r3 = −0.186 – 0.203 = 0.3896 (left side)

r4 and r3 = −0.1086 – 0.281 = 0.3896 (right side)

Does it make a difference which category you leave out?

Answer:

For the model: NO

But you do see different values for the estimated coefficients of the dummy-variables.
This is because each coefficient says something about the difference between two categories.

r4 and r3 = −0.1086 – 0.281 = 0.3896 (right side)

```
. reg dang female kmyear r2 r3 r4 r5

      Source |       SS           df       MS      Number of obs   =       720
-------------+----------------------------------   F(6, 713)       =     12.31
       Model |  127.349847         6  21.2249746   Prob > F        =    0.0000
    Residual |   1229.8446       713  1.72488723   R-squared       =    0.0938
-------------+----------------------------------   Adj R-squared   =    0.0862
       Total |  1357.19444       719  1.88761397   Root MSE        =    1.3133

------------------------------------------------------------------------------
   dangerous |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -.507454   .1034583    -4.90   0.000    -.7105733   -.3043346
      kmyear |   .0000196   3.94e-06     4.96   0.000     .0000118    .0000273
          r2 |  -.0778179   .1279803    -0.61   0.543    -.3290812    .1734454
          r3 |  -.1864143   .1804184    -1.03   0.302    -.5406292    .1678007
          r4 |   .2031481   .1789272     1.14   0.257    -.1481391    .5544353
          r5 |  -.0711879   .2253719    -0.32   0.752    -.5136597     .371284
       _cons |   1.090484   .1170245     9.32   0.000     .8607305    1.320238
------------------------------------------------------------------------------
```

```
. reg dangerous female kmyear i.reliL5

      Source |       SS           df       MS      Number of obs   =       720
-------------+----------------------------------   F(6, 713)       =     12.31
       Model |  127.349847         6  21.2249746   Prob > F        =    0.0000
    Residual |   1229.8446       713  1.72488723   R-squared       =    0.0938
-------------+----------------------------------   Adj R-squared   =    0.0862
       Total |  1357.19444       719  1.88761397   Root MSE        =    1.3133

------------------------------------------------------------------------------
   dangerous |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -.507454   .1034583    -4.90   0.000    -.7105733   -.3043346
      kmyear |   .0000196   3.94e-06     4.96   0.000     .0000118    .0000273
             |
      reliL5 |
          2  |  -.0778179   .1279803    -0.61   0.543    -.3290812    .1734454
          3  |  -.1864143   .1804184    -1.03   0.302    -.5406292    .1678007
          4  |   .2031481   .1789272     1.14   0.257    -.1481391    .5544353
          5  |  -.0711879   .2253719    -0.32   0.752    -.5136597     .371284
             |
       _cons |   1.090484   .1170245     9.32   0.000     .8607305    1.320238
------------------------------------------------------------------------------
```

```
. reg dang female kmyear r2 r3 r4 r5
```

Source	SS	df	MS	
				720
				12.31
Model	127.349847	6	21.22497	0.0000
Residual	1229.8446	713	1.724887	0.0938
				0.0862
Total	1357.19444	719	1.887613	1.3133

dangerous	Coef.	Std. Err.	t			erval]
female	-.507454	.1034583	-4.90			043346
kmyear	.0000196	3.94e-06	4.96			000273
r2	-.0778179	.1279803	-0.61			734454
r3	-.1864143	.1804184	-1.03			678007
r4	.2031481	.1789272	1.14	0.257	-.1481391	.5544353
r5	-.0711879	.2253719	-0.32	0.752	-.5136597	.371284
_cons	1.090484	.1170245	9.32	0.000	.8607305	1.320238

You need to create dummy-variables first

```
. reg dangerous female kmyear i.reliL5
```

Source	SS	df	MS	
				720
				12.31
Model	127.349847	6	21.2249	0.0000
Residual	1229.8446	713	1.72488	0.0938
				0.0862
Total	1357.19444	719	1.88761	1.3133

dangerous	Coef.	Std. Err.	t			erval]
female	-.507454	.1034583	-4.90			043346
kmyear	.0000196	3.94e-06	4.96			000273
reliL5						
2	-.0778179	.1279803	-0.61			734454
3	-.1864143	.1804184	-1.03			678007
4	.2031481	.1789272	1.14	0.257	-.1481391	.5544353
5	-.0711879	.2253719	-0.32	0.752	-.5136597	.371284
_cons	1.090484	.1170245	9.32	0.000	.8607305	1.320238

No need to create dummies first, but you will not have dummy-variables in your data

The test-command (revisited)

```
. reg dang female kmyear r2 r3 r4 r5

      Source |       SS       df       MS              Number of obs =     720
-------------+------------------------------           F(6, 713)     =   12.31
       Model |  127.349847     6  21.2249746           Prob > F      =  0.0000
    Residual |   1229.8446   713  1.72488723           R-squared     =  0.0938
-------------+------------------------------           Adj R-squared =  0.0862
       Total |  1357.19444   719  1.88761397           Root MSE      =  1.3133

------------------------------------------------------------------------------
   dangerous |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -.507454   .1034583    -4.90   0.000    -.7105733   -.3043346
      kmyear |   .0000196   3.94e-06     4.96   0.000     .0000118    .0000273
          r2 |  -.0778179   .1279803    -0.61   0.543    -.3290812    .1734454
          r3 |  -.1864143   .1804184    -1.03   0.302    -.5406292    .1678007
          r4 |   .2031481   .1789272     1.14   0.257    -.1481391    .5544353
          r5 |  -.0711879   .2253719    -0.32   0.752    -.5136597     .371284
       _cons |   1.090484   .1170245     9.32   0.000     .8607305    1.320238
------------------------------------------------------------------------------
```

```
. test r3=r4

 ( 1)  r3 - r4 = 0

       F(  1,   713) =    2.70
            Prob > F =    0.1005
```

I made a mistake here; I should have added "=0"

```
. test r2=r3=r4=r5 = 0

 ( 1)  r2 - r3 = 0
 ( 2)  r2 - r4 = 0
 ( 3)  r2 - r5 = 0

       F(  3,   713) =    1.01
            Prob > F =    0.3891
```

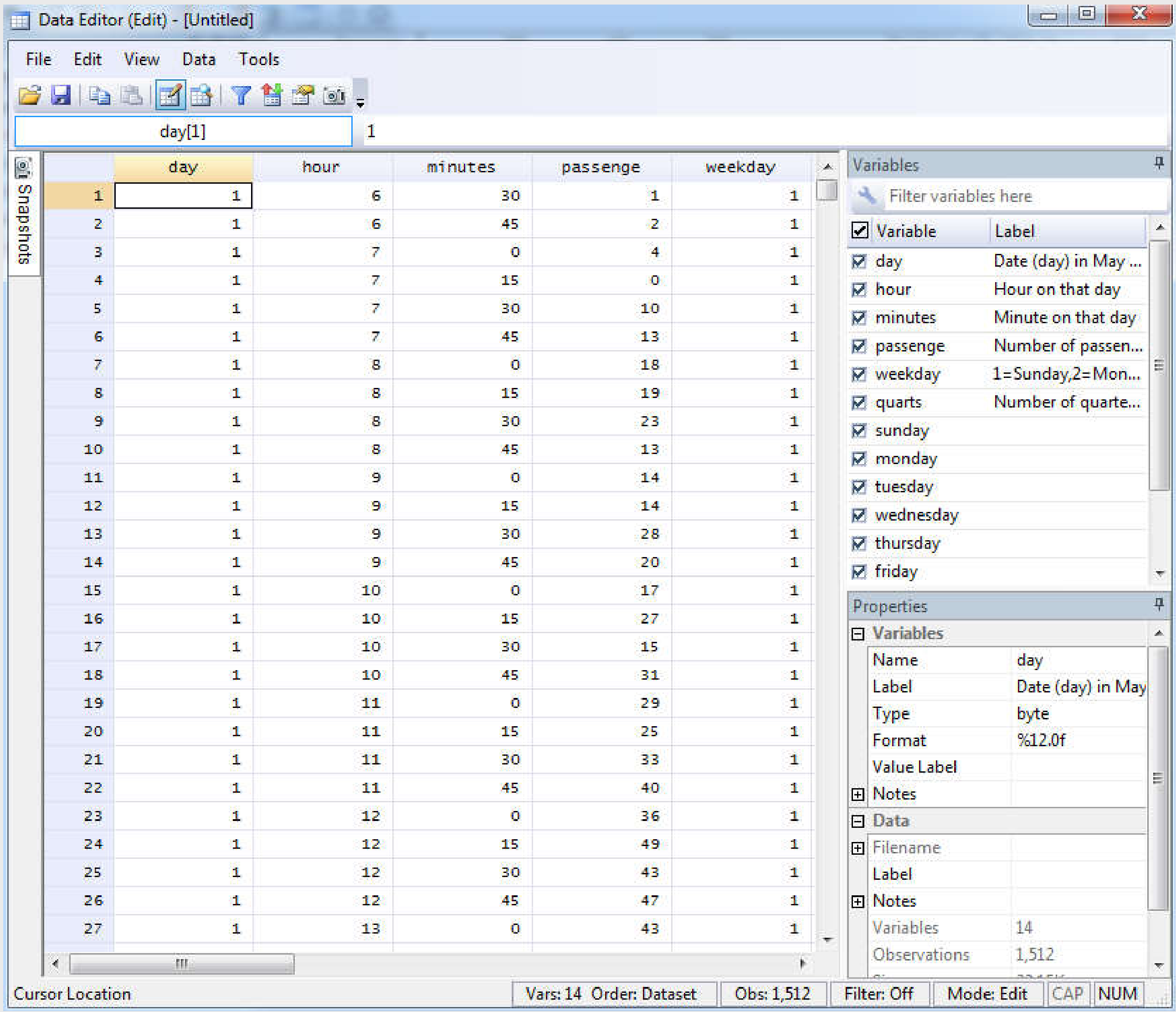
The do-file

```
clear                     // clear system
set more off              // Scroll until end of output automatically

use traffic               // Read in the data

// We need a convenience command that is not standard Stata here.
// type:

net install renvarlab

// This will install the command 'renvarlab'

renvarlab, lower          // creates lowercase variables, I prefer this

recode    religion (1=1)(2=2)(3=3)(4=4)(5 6 7 8 9=5), gen(reliL5)
label var reliL5 "1=none/2=romancath/3=prot/4=othChris/5=allothers"

tab reliL5, gen(r)

reg dang female kmyear r2 r3 r4 r5
test r2=r3=r4=r5

reg dang female kmyear r1 r3 r4 r5
test r3=r4
test r1=r3=r4=r5


reg dang female kmyear i.reliL5
```

Introducing WIKI data: airport passengers

Predict the number of passengers at an airport terminal …

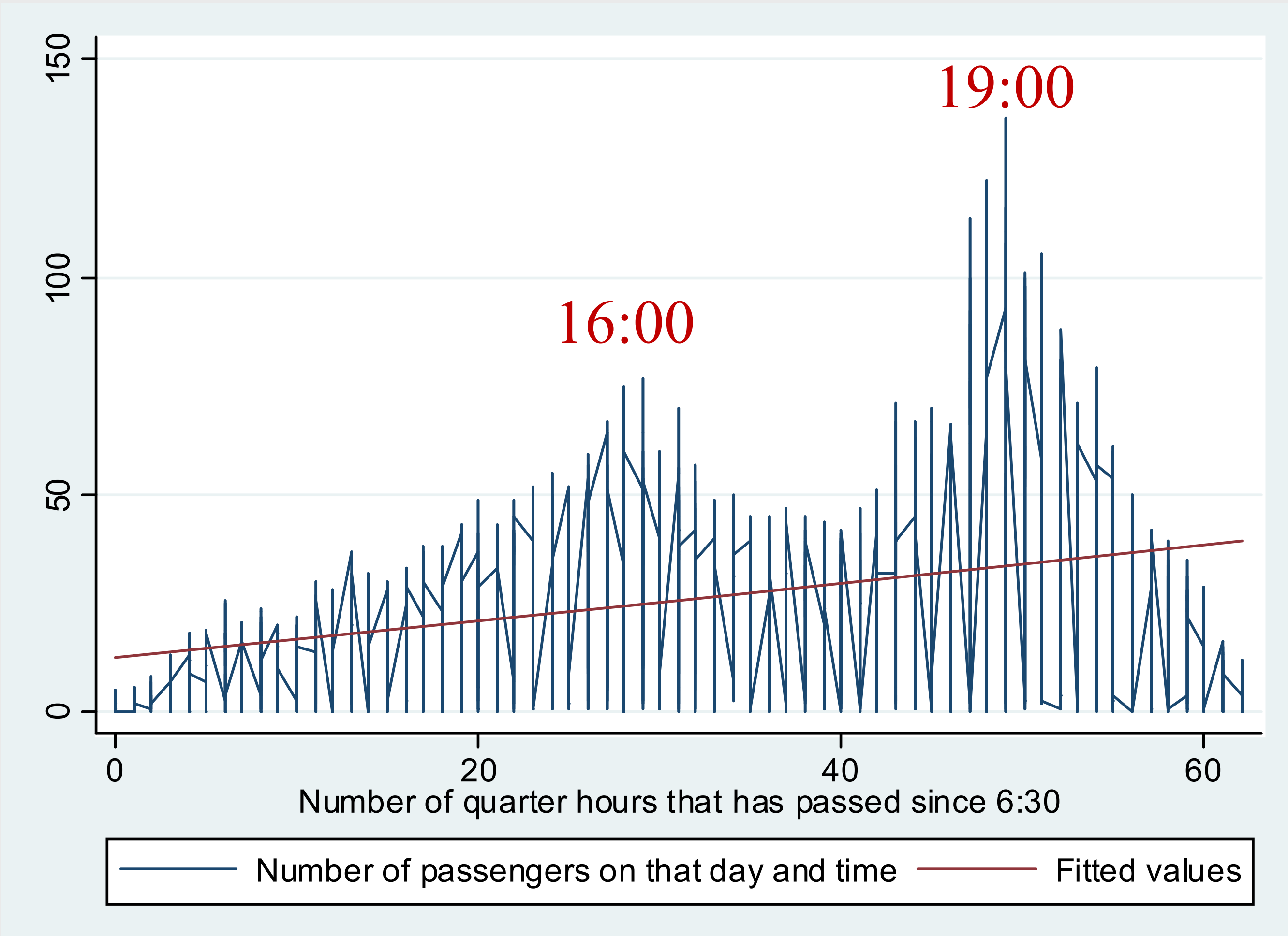
Number of passengers



day (as of May 2005)
hour, minutes, passenge, weekday, quarts

Predict **passenge** from the rest of the data.

Passengers by time of day



Red line shows the linear trend, but how can we improve over this?

What's up next?

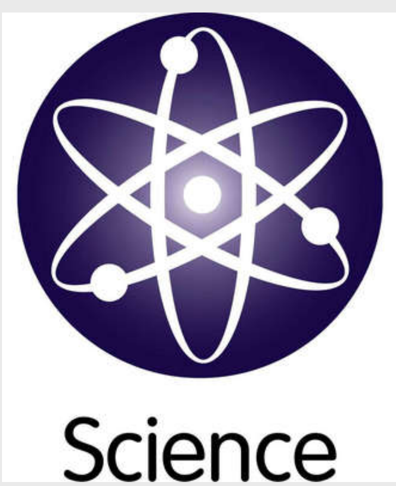
- Outliers
- Interaction effects and transformations of variables
- Multicollinearity
- Assumptions and their violations

Recap

- Simple regression can be run with non-INTERVAL X-variables as well
- Understanding what is going on can be based on a single regression OR on a succession of models
- Categorical variables need to be included as separate dummy-variables. You add as many dummy-variables to the model as there are categories, MINUS 1
- Measures of fit: R^2 and adjusted R^2
- Besides estimates for the coefficients, MR gives you a test of the base hypothesis that the coefficient equals zero
- You can get an overview of the differences between the categories of a categorical variable, by considering the different dummies

To Do

- Understand multiple regression
- **PRACTICE!** running regression analyses!
- Check out and add to the WIKIs
- Use other online material, for instance
  http://www.ats.ucla.edu/stat/Stata/output/reg_Stata(long).htm
  gives you annotated regression output



VS