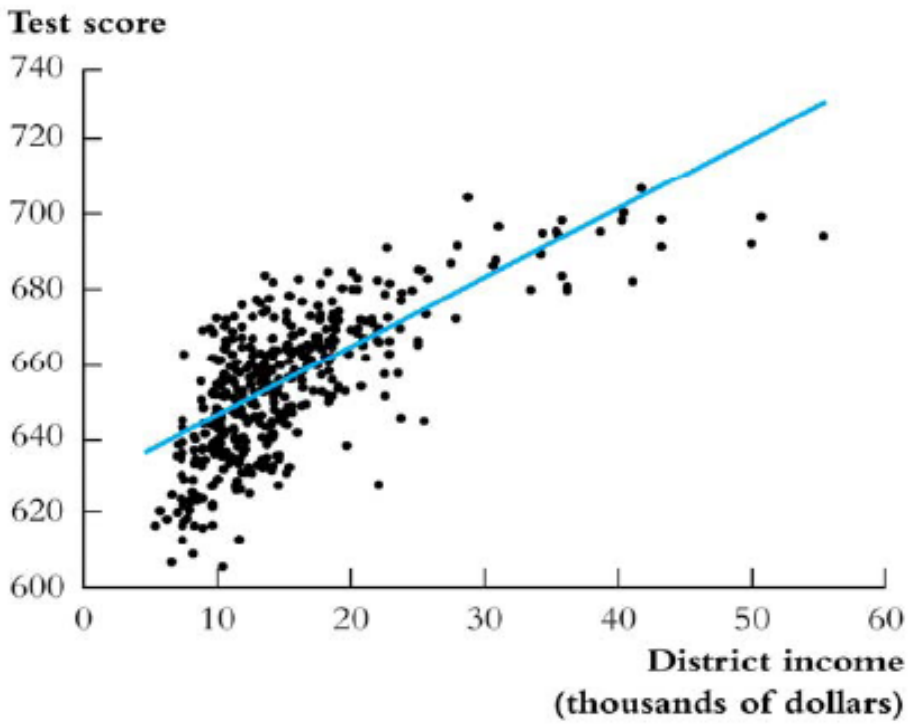
Lecture notes to Stock and Watson chapter 8

Nonlinear regression

Tore Schweder

September 2009

Example: TestScore – Income relation, linear or nonlinear?



General problem

Fully nonlinear regression:

$$Y = f(X_1, X_2, \cdots, X_k; \beta) + u$$

for given parameter vector β $f(x_1, \cdots, x_k; \beta)$ is a specific (nonlinear) function of $(x_1, \cdots, x_k)$.

Estimate β from data $(Y_i, X_{i1}, \cdots, X_{ik})$ $i = 1, \cdots, n$ (assuming the sample being random, the random terms u_i having mean zero, and having finite fourth moment)

Effect calculation:

$$\begin{aligned} \Delta Y &= f(X_1 + \Delta X_1, X_2, \cdots, X_k; \beta) - f(X_1, X_2, \cdots, X_k; \beta) \\ \widehat{\Delta Y} &= f(X_1 + \Delta X_1, X_2, \cdots, X_k; \widehat{\beta}) - f(X_1, X_2, \cdots, X_k; \widehat{\beta}) \end{aligned}$$

Simplified method, only one X

linear regression of one or more nonlinear functions of X on Y

$$Y = \beta_0 + \beta_1 X + \beta_2 f_2(X) + \cdots + \beta_k f_k(X) + u$$

Standard assumptions for a sample (Y_i, X_i) $i = 1, \cdots, n$

- Estimate $\beta_0, \beta_1, \cdots \beta_k$ by linear regression of $(X_1 = X, X_2 = f_2(X), \cdots, X_k = f_k(X))$ on Y
- What is required of the functions f_j $j = 2, \cdots, k$ to avoid perfect collinearity?

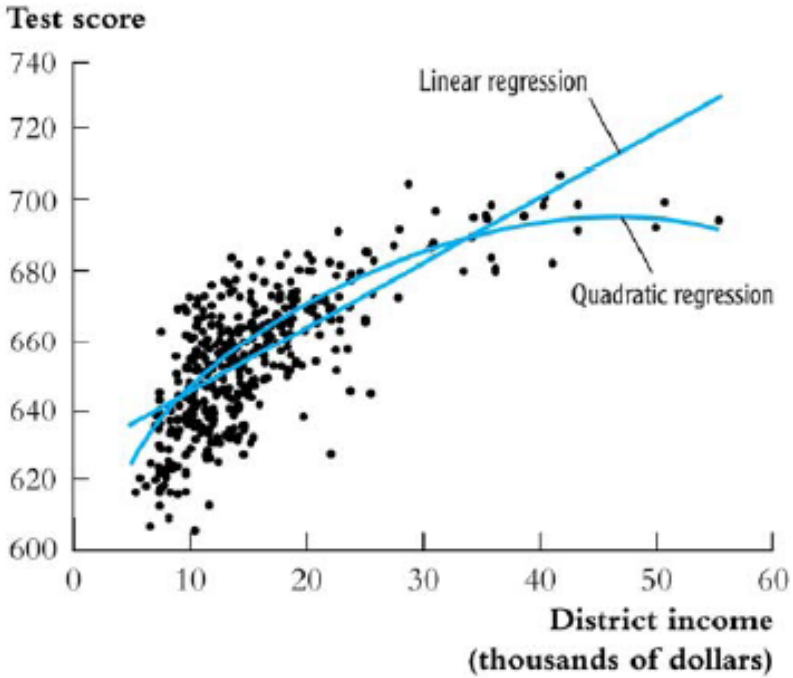
Polynomial regression, Example

$$TestScore = \beta_0 + \beta_1 Income + \beta_2 (Income)^2 + u$$

```
generate avginc2 = avginc*avginc;       Create a new regressor
reg testscr avginc avginc2, r;

Regression with robust standard errors                Number of obs =     420
                                                      F(  2,   417) =  428.52
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5562
                                                      Root MSE      =  12.724

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |   3.850995   .2680941    14.36   0.000      3.32401    4.377979
     avginc2 |  -.0423085   .0047803    -8.85   0.000     -.051705   -.0329119
       _cons |   607.3017   2.901754   209.29   0.000     601.5978    613.0056
------------------------------------------------------------------------------
```



- Reasonable to expect an optimal District income, and declining Test scores for richer districts?
- Cubic regression?, linear-log regression?

Example: linear-log regression

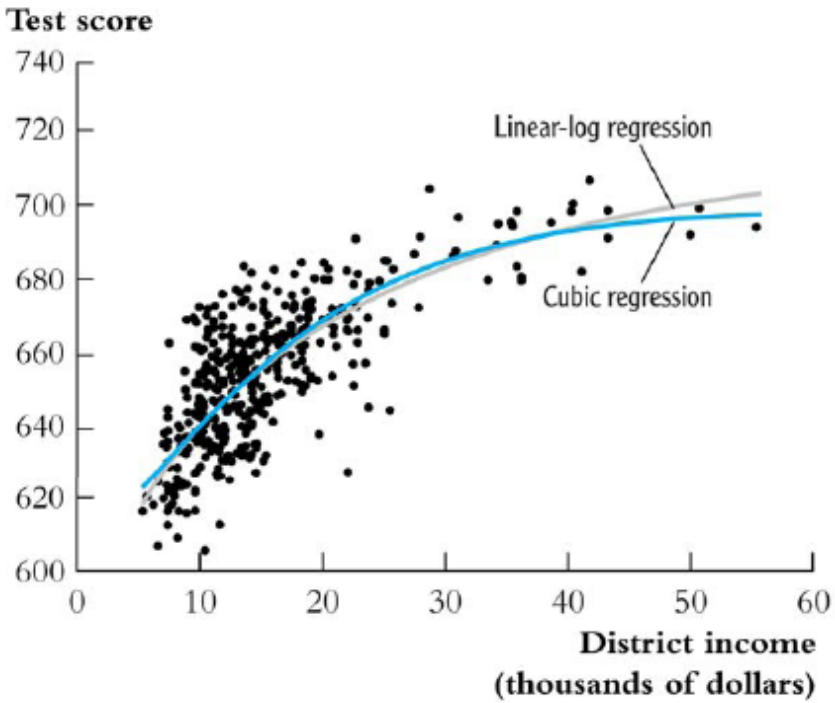
```
. gen lnIncome=ln(avginc)

. reg  testscr  lnIncome,r

Linear regression                       Number of obs  =     420
                                        F(  1,   418)  =  679.70
                                        Prob > F       =  0.0000
                                        R-squared      =  0.5625
                                        Root MSE       =  12.618

                        Robust
testscr       Coef.     Std. Err.  t        P>t  [95% Conf.     Interval]

lnIncome    36.41968    1.396943  26.07    0.000     33.67378  39.16559
_cons      557.8323     3.83994   145.27   0.000     550.2843  565.3803
```


Test score
740
720
700
680
660
640
620
600
0
10
20
30
40
50
60
Linear-log regression
Cubic regression
District income
(thousands of dollars)

Log-log regression

$$\begin{aligned} \ln(Y) &= \beta_0 + \beta_1 \ln(X) + u \\ Y &= e^{\beta_0} X^{\beta_1} e^{u} \end{aligned}$$

- β_1 is the expected elasticity, $\widehat{\beta_1}$ is the estimated elasticity
- Effect estimates on log scale: $\widehat{\beta_1} = \frac{d}{dY} \ln(Y) / \frac{d}{dX} \ln(X)$
- Effect estimates on original scale: $\Delta Y = \widehat{\beta_1} \frac{Y}{X} \Delta X$
- The estimated elasticity of Test score on District income is $\widehat{\beta_1} = .055$

Effect calculations for Test score by District income

Expected increase in Test score by one unit increase in average district income (1000$)

	5-6	25-26	45-46	50-51
quadratic	3.38	1.69	0.00	-0.42
cubic	4.03	1.47	0.56	0.59
linear-log	6.64	1.43	0.80	0.72
log-log	6.27	1.47	0.85	0.77

Which model to choose?

- The model must allow estimation of the desired quantity
  - log-log is good if you are desperate for an invariant elasticity
- It must fit the data
- It should not be more complex than necessary
- If only fit and complexity matters, use AIC or BIC as information criteria
  - STAT command `estat ic`

Example Test score District income

After each regression, ask for `estat ic`. Low value of AIC and BIC is good! Only models with the response `testscr` on the same scale can be compared.

Response variable $Y =$ `testscr`

Model	AIC	BIC
linear	3373.1	3381.2
quadratic	3331.4	3343.5
cubic	3331.3	3347.5
log	3323.4	3331.5

The linear-log model is preferred!
BIC penalize complexity more than AIC!

Dummy variables and interactions

```
. gen HiEL= el_pct>10
. gen lnInc_HiEL=  lnIncome*HiEL
. reg  testscr    lnIncome HiEL    lnInc_HiEL,r

Linear regression               Number of obs  =     420
                                F(  3,   416)  =  387.22
                                Prob > F  =  0.0000
                                R-squared =  0.6918
                                Root MSE  =  10.616

                                Robust
testscr       Coef.      Std. Err.      t    P>t  [95% Conf.     Interval]

lnIncome       29.3605   1.605297     18.29  0.000      26.20499 32.516
HiEL          -29.54333   7.253457   -4.07  0.000|    -43.80133 -15.28534
lnInc_HiEL     5.923055   2.754631     2.15  0.032      .5083248 11.33779
_cons          583.1411    4.58449   127.20  0.000     574.1295  592.1528
```

AIC=3180.2

Residual plot

```
predict r, residuals
twoway (scatter r avginc), by(HiEL)
```

Figure: Residuals from the linear-log regression with interaction:
$testscr = \beta_0 + \beta_1 \log(avginc) + \beta_2 HiEL + \beta_3 HiEL \cdot \log(avginc)$

Problem in polynomial regression - near collinearity?

Example polynomial regression of degree 4:

$X_i = i/25 \; i = 1, \cdots, 25 = n. \;\; k = 4$

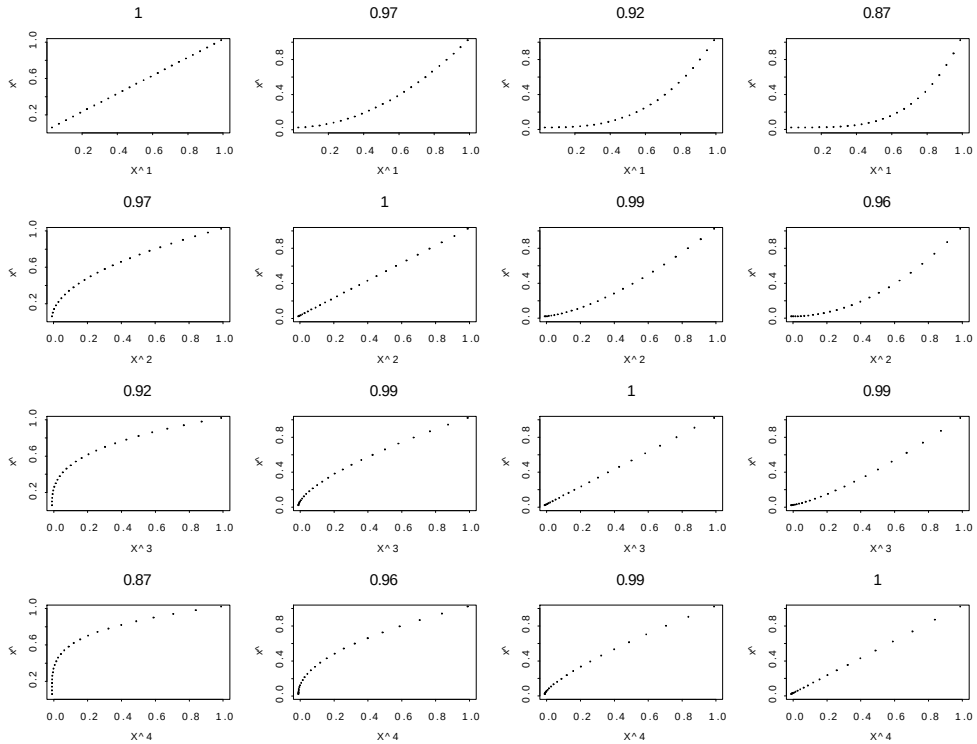


Figure: plots of X^i vr. X^j Correlation given on top of each plot.

Remedy: work with orthogonal polynomials!

Example cont

Let $OX_2 = X^2 - \widehat{X^2}$ where $\widehat{X^2}$ is the fitted values from the regression

$$X^2 = \alpha_{02} + \alpha_{12}X + u_2$$

, $OX_3 = X^3 - \widehat{X^3}$ where $\widehat{X^3}$ is the fitted values from the regression

$$X^3 = \alpha_{03} + \alpha_{13}X + \alpha_{13}X^2 + u_3$$

etc.

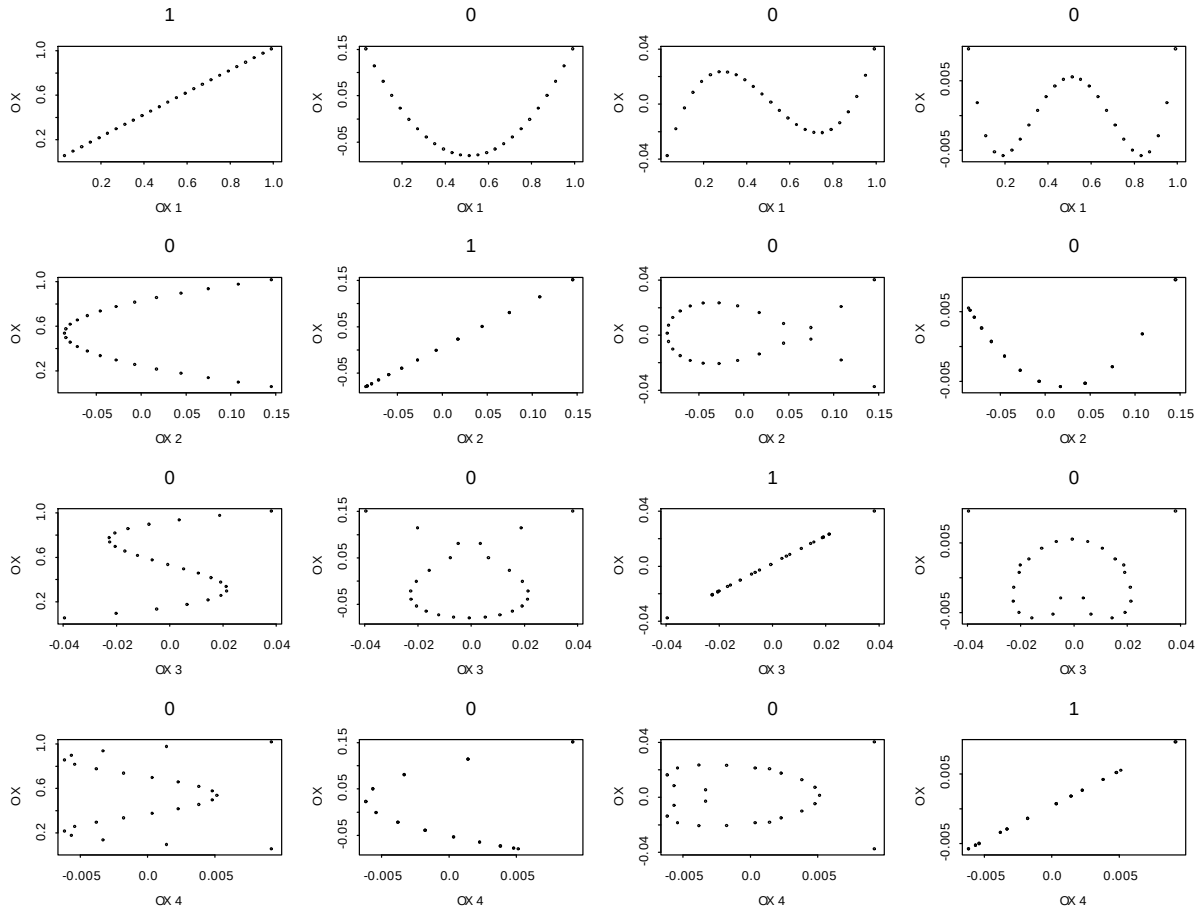


Figure: Orthogonal covariates plotted against each other, correlations over each graph.

Making the orthogonal variables for District income

```
. reg   avginc2 avginc,r
. predict OInc2, residuals
. reg   avginc3 avginc  avginc2,r
. predict OInc3, residuals
```

```
.  corr avginc avginc2 avginc3
(obs=420)

     avginc  avginc2     avginc3

avginc    1.0000
avginc2   0.9592   1.0000
avginc3   0.8717   0.9727     1.0000
```

```
.  corr avginc OInc2 OInc3
(obs=420)

     avginc    OInc2     OInc3

avginc    1.0000
OInc2     -0.0000   1.0000
OInc3     -0.0000  -0.0000    1.0000
```

Comparing direct polynomial regression and equivalent polynomial regression on orthogonal variables

```
. reg  testscr   avginc avginc2  avginc3,r
Linear regression               Number of obs  =     420
                                R-squared =  0.5584
                                Root MSE  =  12.707

                                Robust
testscr       Coef.   Std. Err.      t  P>t  [95% Conf.     Interval]

avginc    5.018677   .7073504     7.10  0.000     3.628251  6.409104
avginc2   -.0958052   .0289537    -3.31 0.001     -.152719  -.0388913
avginc3    .0006855   .0003471     1.98 0.049     3.27e-06  .0013677
_cons     600.079   5.102062   117.61   0.000     590.0499  610.108

. reg  testscr   avginc  OInc2 OInc3,r
Linear regression               Number of obs  =     420
                                R-squared =  0.5584
                                Root MSE  =  12.707

                                Robust
testscr       Coef.   Std. Err.      t  P>t  [95% Conf.     Interval]

avginc     1.87855    .068793    27.31  0.000     1.743324  2.013775
OInc2   -.0423085   .0048888    -8.65   0.000     -.0519183 -.0326986
OInc3    .0006855   .0003471     1.98   0.049     3.27e-06  .0013677
_cons    625.3836   1.328025   470.91   0.000     622.7731  627.9941
```

- Do SW: 7.8