Elements of Econometrics.
Lecture 6.
Variables Transformation in Regression Analysis.

FCS, 2022-2023

LINEARITY AND NONLINEARITY

Linear in variables and parameters:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u$$

Linear in parameters, nonlinear in variables:

$$Y = \beta_1 + \beta_2 X_2^2 + \beta_3 \sqrt{X_3} + \beta_4 \log X_4 + u$$

$$Z_2 = X_2^2, \quad Z_3 = \sqrt{X_3}, \quad Z_4 = \log X_4$$

$$Y = \beta_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + u$$

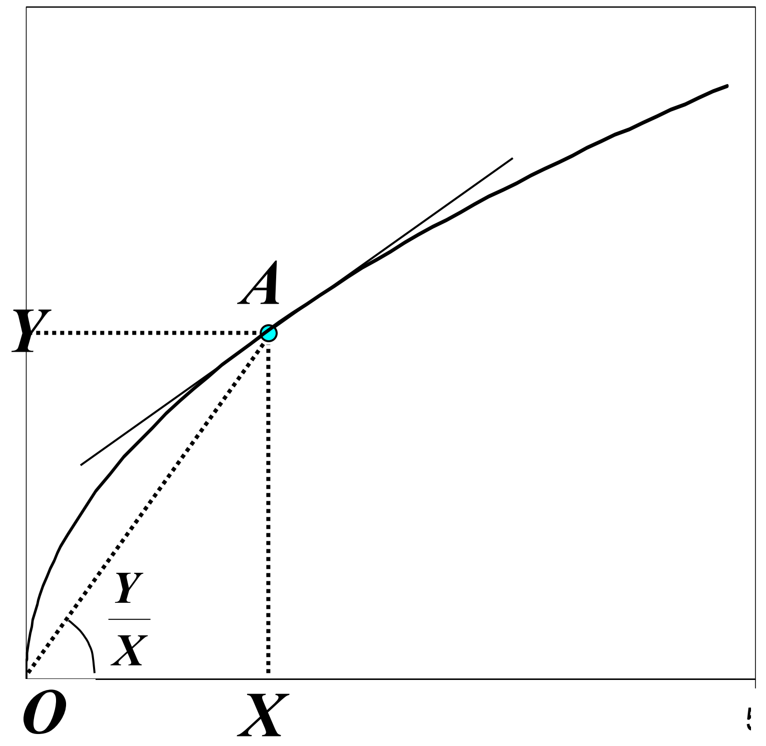
Nonlinear in parameters:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_2 \beta_3 X_4 + u$$

This model is nonlinear in parameters and can not be linearised by appropriate transformations. Some others can be linearised (for example, by taking logarithms).

ELASTICITIES AND LOGARITHMIC MODELS

Definition:
The elasticity of Y with respect to X is the proportional change in Y per proportional change in X.



elasticity

$$= \frac{dY/Y}{dX/X} = \frac{dY/dX}{Y/X}$$

$$= \frac{\text{slope of the tangent at } A}{\text{slope of } OA}$$

The elasticity at any point on the curve is the ratio of the slope of the tangent at that point to the slope of the line joining the point to the origin.

ELASTICITIES AND LOGARITHMIC MODELS

$$Y = \beta_1 X^{\beta_2}$$

$$\frac{dY}{dX} = \beta_1 \beta_2 X^{\beta_2 - 1}$$

$$\frac{Y}{X} = \frac{\beta_1 X^{\beta_2}}{X} = \beta_1 X^{\beta_2 - 1}$$

elasticity $= \dfrac{dY/dX}{Y/X} = \dfrac{\beta_1 \beta_2 X^{\beta_2 - 1}}{\beta_1 X^{\beta_2 - 1}} = \beta_2$

Hence we obtain the expression for the elasticity. This simplifies to β_2 and is therefore constant.

ELASTICITIES AND LOGARITHMIC MODELS

$$Y = \beta_1 X^{\beta_2} \qquad\qquad logY = \beta'_1 + \beta_2 logX$$

$$\begin{aligned} \log Y &= \log \beta_1 X^{\beta_2} \\ &= \log \beta_1 + \log X^{\beta_2} \\ &= \log \beta_1 + \beta_2 \log X \end{aligned}$$

$$Y' = \beta_1' + \beta_2 X'$$ **where** $$\begin{aligned} Y' &= \log Y\,, \\ X' &= \log X \\ \beta_1' &= \log \beta_1 \end{aligned}$$

The constant term will be an estimate of log β_1. To obtain an estimate of β_1, you calculate exp(b_1'), where b_1' is the estimate of β_1'. (This assumes that you have used natural logarithms, that is, logarithms to base *e*, to transform the model.)

Elasticities: Double Logarithmic Function

Chief Executive Officer (CEO) salary and firm sales

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u$$

Natural logarithm of CEO salary

Natural logarithm of his/her firm's sales

This changes the interpretation of the regression coefficient:

$$\beta_1 = \frac{\Delta \log(salary)}{\Delta \log(sales)} = \frac{\frac{\Delta salary}{salary}}{\frac{\Delta sales}{sales}}$$

<u>Percentage</u> change in salary if sales increase by 1%

Logarithmic changes are always percentage changes

$$\widehat{\log}(salary) = 4.822 + 0.257 \log(sales)$$

$+1\%\ sales \rightarrow +.257\%\ salary$

The double *log* form means a constant elasticity model

Elasticities: Cobb-Douglas Production Function

Cobb-Douglas Production Function: $Y = A \cdot K^{\alpha} \cdot L^{\beta}$

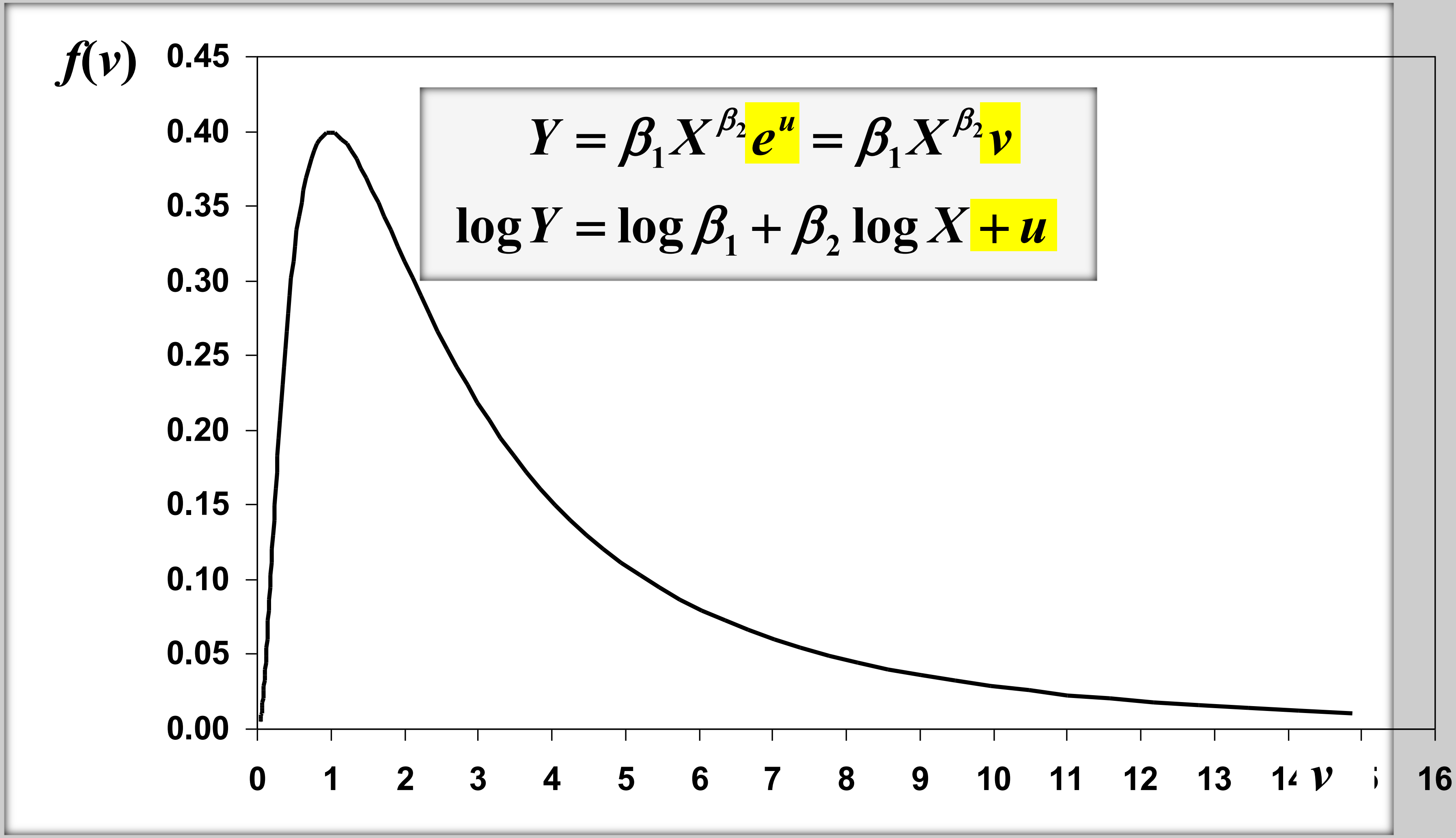
$$\ln Y = \ln A + \alpha \ln K + \beta \ln L$$

$$\frac{dY_t}{Y_t} = \alpha \cdot \frac{dK_t}{K_t} + \beta \cdot \frac{dL_t}{L_t}$$

$$e_L = \left(\frac{\partial Y}{\partial L}\right) : \left(\frac{Y}{L}\right) = \left(\frac{\partial Y}{Y}\right) : \left(\frac{\partial L}{L}\right) = \frac{\partial \ln Y}{\partial \ln L} = \beta \approx \left(\frac{\Delta Y}{Y}\right) : \left(\frac{\Delta L}{L}\right)$$

$$e_K = \left(\frac{\partial Y}{\partial K}\right) : \left(\frac{Y}{K}\right) = \left(\frac{\partial Y}{Y}\right) : \left(\frac{\partial K}{K}\right) = \frac{\partial \ln Y}{\partial \ln K} = \alpha \approx \left(\frac{\Delta Y}{Y}\right) : \left(\frac{\Delta K}{K}\right)$$

THE DISTURBANCE TERM IN LOGARITHMIC MODELS



$$Y = \beta_1 X^{\beta_2} e^{u} = \beta_1 X^{\beta_2} v$$

$$\log Y = \log \beta_1 + \beta_2 \log X + u$$

For the regression results in a linearised model to have the desired properties, the disturbance term in the transformed model should be additive and it should satisfy the regression model conditions. For the logarithmic model, this will be the case if *v* has a lognormal distribution, shown above.

SEMILOGARITHMIC MODELS

$$Y = \beta_1 e^{\beta_2 X} \qquad logY = \beta'_1 + \beta_2 X$$

$$\frac{dY}{dX} = \beta_1 \beta_2 e^{\beta_2 X} = \beta_2 Y \qquad \frac{dY/Y}{dX} = \beta_2$$

$$\frac{\Delta Y/\Delta X}{Y} \approx \beta_2 \qquad log\,Y = log\,\beta_1\,e^{\beta_2 X} = log\,\beta_1 + log\,e^{\beta_2 X} = \beta_1' + \beta_2 X$$

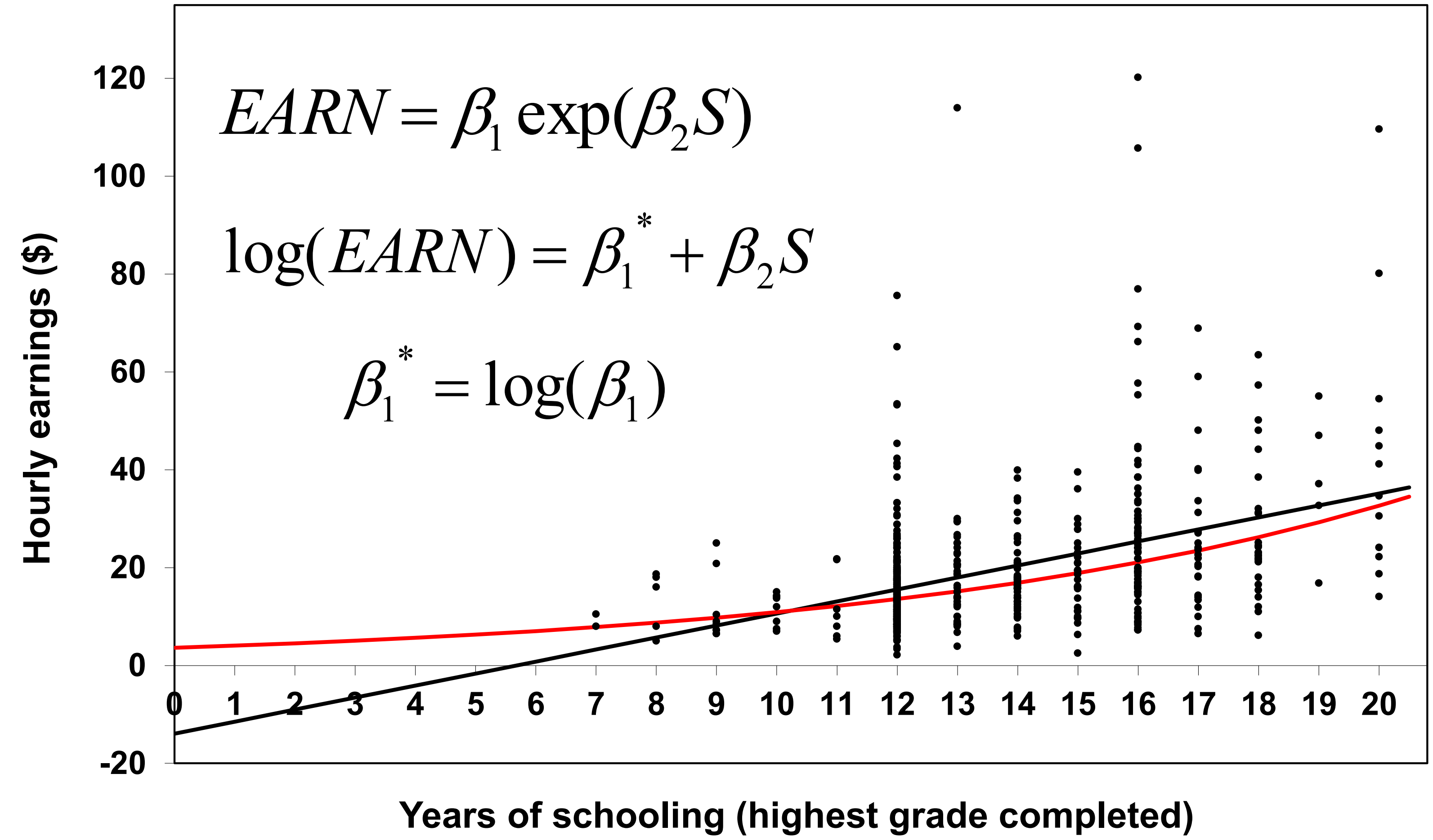
β_2 shows the relative change in Y per unit of change of X

$$Y = \beta_1 + \beta_2\,log\,X$$

$$\frac{dY}{dX} = \beta_2/X \qquad dY = \beta_2 \frac{dX}{X} \qquad \Delta Y \approx \beta_2 \frac{\Delta X}{X}$$

β_2 shows the change in Y per unit of relative change of X

EARNINGS FUNCTION: SEMILOGARITHMIC MODEL



The slope coefficient of the semi-logarithmic specification has a simple interpretation and the specification does not give rise to nonsensical predictions outside the data range.

COMPARING LINEAR AND LOGARITHMIC SPECIFICATIONS

$$Y = \beta_1 + \beta_2 X + u$$

$$\log Y = \beta_1 + \beta_2 X + u$$

Zarembka scaling : $\boldsymbol{Y^* = Y}$ **/ geometric mean of** $\boldsymbol{Y}$

$$e^{\frac{1}{n}\sum \log Y_i} = e^{\frac{1}{n}\log(Y_1 Y_2 \ldots Y_n)}$$

$$= e^{\log(Y_1 Y_2 \ldots Y_n)^{\frac{1}{n}}}$$

$$\log Y^* = \beta_1' + \beta_2' X + u$$

$$Y^* = \beta_1' + \beta_2' X + u$$

The residual sums of squares are now directly comparable since *X*~log(1+*X*) for small *X*. The specification with the smaller *SSR* therefore provides the better fit.

COMPARING LINEAR AND LOGARITHMIC SPECIFICATIONS

$$EARNINGS = \beta_1 + \beta_2 ASVABC + \beta_3 S + u$$

$$EARNINGS = \beta_1 + \beta_2 ASVABC + \beta_3 \log(S) + u$$

Dependent Variable: EARNINGS
Method: Least Squares
Date: 10/17/18 Time: 22:41
Sample: 1 500
Included observations: 500

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.032117	3.123368	0.330450	0.7412
ASVABC	1.361713	0.621336	2.191587	0.0289
S	1.190864	0.216750	5.494189	0.0000

R-squared	0.118648	Mean dependent var	18.43730
Adjusted R-squared	0.115102	S.D. dependent var	12.04802
S.E. of regression	11.33346	Akaike info criterion	7.699378
Sum squared resid	63838.32	Schwarz criterion	7.724666
Log likelihood	-1921.844	Hannan-Quinn criter.	7.709301
F-statistic	33.45331	Durbin-Watson stat	2.039146
Prob(F-statistic)	0.000000		

Dependent Variable: EARNINGS
Method: Least Squares
Date: 10/17/18 Time: 22:51
Sample: 1 500
Included observations: 500

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-23.59722	7.972461	-2.959841	0.0032
ASVABC	1.418354	0.625970	2.265851	0.0239
LOG(S)	15.77606	3.020258	5.223416	0.0000

R-squared	0.113770	Mean dependent var	18.43730
Adjusted R-squared	0.110204	S.D. dependent var	12.04802
S.E. of regression	11.36478	Akaike info criterion	7.704898
Sum squared resid	64191.68	Schwarz criterion	7.730186
Log likelihood	-1923.224	Hannan-Quinn criter.	7.714821
F-statistic	31.90122	Durbin-Watson stat	2.042039
Prob(F-statistic)	0.000000		

Data set EAWE22. The linear specification 1 with smaller *SSR* provides better fit.

COMPARING LINEAR AND LOGARITHMIC SPECIFICATIONS

$$\log(EARNINGS) = \beta_1 + \beta_2 ASVABC + \beta_3 S + u$$

$$\log(EARNINGS) = \beta_1 + \beta_2 ASVABC + \beta_3 \log(S) + u$$

Dependent Variable: LOG(EARNINGS)
Method: Least Squares
Date: 10/17/18 Time: 22:43
Sample: 1 500
Included observations: 500

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.842506	0.139198	13.23657	0.0000
ASVABC	0.060392	0.027691	2.180934	0.0297
S	0.063085	0.009660	6.530677	0.0000

R-squared	0.149786	Mean dependent var	2.762770
Adjusted R-squared	0.146364	S.D. dependent var	0.546684
S.E. of regression	0.505095	Akaike info criterion	1.477841
Sum squared resid	126.7950	Schwarz criterion	1.503129
Log likelihood	-366.4602	Hannan-Quinn criter.	1.487764
F-statistic	43.77927	Durbin-Watson stat	2.074625
Prob(F-statistic)	0.000000		

Dependent Variable: LOG(EARNINGS)
Method: Least Squares
Date: 10/17/18 Time: 23:00
Sample: 1 500
Included observations: 500

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.439703	0.354454	1.240508	0.2154
ASVABC	0.059226	0.027830	2.128082	0.0338
LOG(S)	0.872960	0.134280	6.501045	0.0000

R-squared	0.149177	Mean dependent var	2.762770
Adjusted R-squared	0.145753	S.D. dependent var	0.546684
S.E. of regression	0.505276	Akaike info criterion	1.478557
Sum squared resid	126.8858	Schwarz criterion	1.503844
Log likelihood	-366.6392	Hannan-Quinn criter.	1.488480
F-statistic	43.57012	Durbin-Watson stat	2.078074
Prob(F-statistic)	0.000000		

The specification 1 with smaller *SSR* provides better fit. But the difference is very small.

COMPARING LINEAR AND LOGARITHMIC SPECIFICATIONS

$$EARNINGS = \beta_1 + \beta_2 ASVABC + \beta_3 S + u$$

$$\log(EARNINGS) = \beta_1 + \beta_2 ASVABC + \beta_3 S + u$$

genr earnings1=earnings/exp(@mean(log(earnings)))

Dependent Variable: EARNINGS1
Method: Least Squares
Date: 10/11/16 Time: 21:24
Sample: 1 500
Included observations: 500

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.065144	0.197137	0.330450	0.7412
ASVABC	0.085947	0.039217	2.191587	0.0289
S	0.075163	0.013681	5.494189	0.0000

R-squared	0.118648	Mean dependent var	1.163701
Adjusted R-squared	0.115102	S.D. dependent var	0.760431
S.E. of regression	0.715330	Akaike info criterion	2.173837
Sum squared resid	254.3135	Schwarz criterion	2.199125
Log likelihood	-540.4592	F-statistic	33.45331
Durbin-Watson stat	2.039146	Prob(F-statistic)	0.000000

Dependent Variable: LOG(EARNINGS1)
Method: Least Squares
Date: 10/11/16 Time: 21:24
Sample: 1 500
Included observations: 500

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.920265	0.139198	-6.611182	0.0000
ASVABC	0.060392	0.027691	2.180934	0.0297
S	0.063085	0.009660	6.530677	0.0000

R-squared	0.149786	Mean dependent var	2.30E-16
Adjusted R-squared	0.146364	S.D. dependent var	0.546684
S.E. of regression	0.505095	Akaike info criterion	1.477841
Sum squared resid	126.7950	Schwarz criterion	1.503129
Log likelihood	-366.4602	F-statistic	43.77927
Durbin-Watson stat	2.074625	Prob(F-statistic)	0.000000

The Semi-logarithmic specification 2 with smaller *SSR* provides better fit.
Is it significantly better?

COMPARING LINEAR AND LOGARITHMIC SPECIFICATIONS: BOX-COX TEST

$$\chi^2(1) = \frac{n}{2}\log\frac{SSR_2}{SSR_1} =$$

$$= \frac{500}{2}\log\frac{254.3}{126.8} = 75.56 > 10.83 = {\chi_{crit,0.1\%}}^2(1)$$

Hence *Ho* (no significant difference in the quality of two models) is rejected

The Semi-logarithmic specification 2 provides significantly better fit.

QUADRATIC EXPLANATORY VARIABLES

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_2^2 + u$$

$$\frac{\mathrm{d}Y}{\mathrm{d}X_2} = \beta_2 + 2\beta_3 X_2 \quad \text{-} \quad \text{changing marginal effect}$$

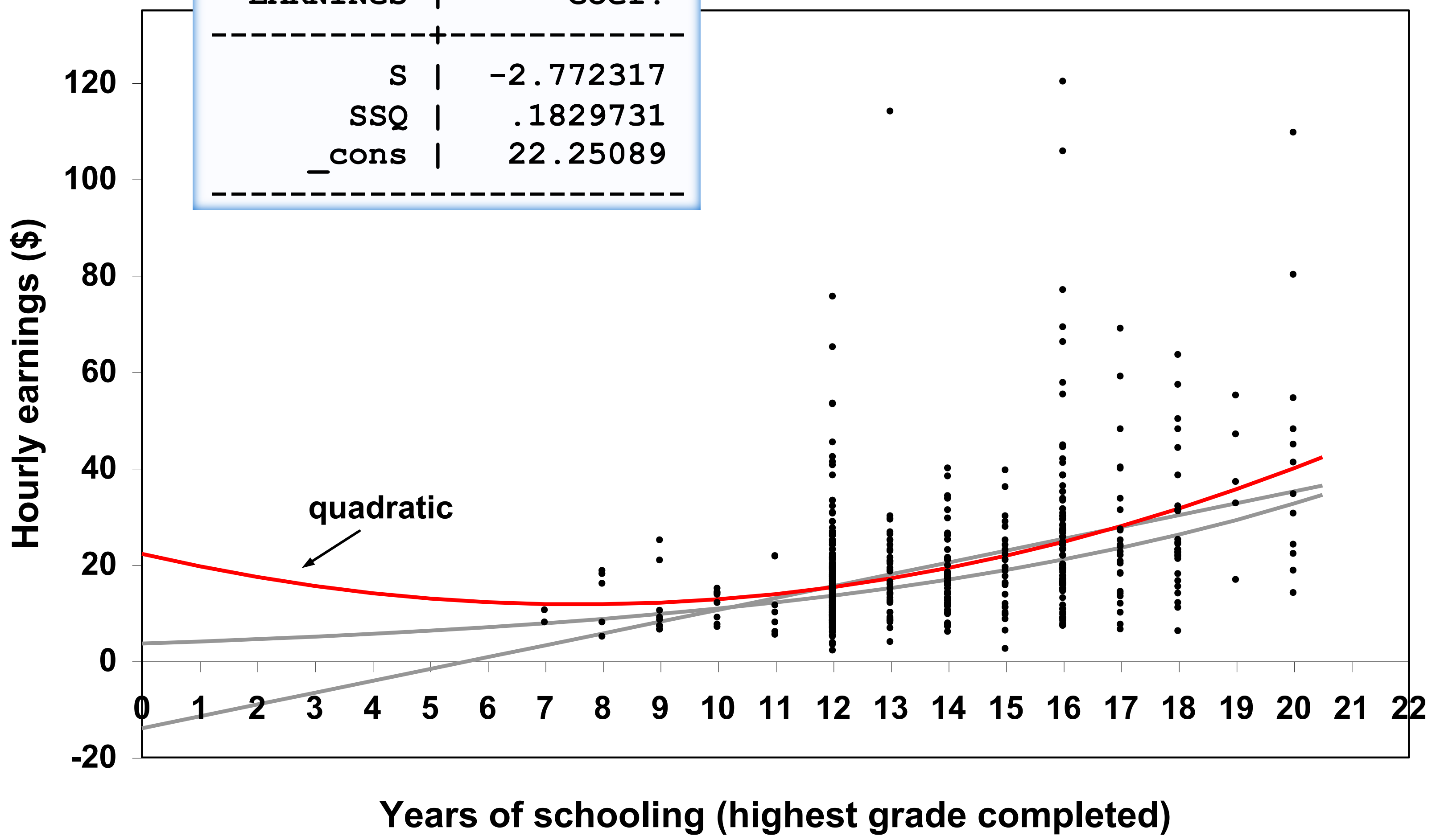
$$Y = \beta_1 + (\beta_2 + \beta_3 X_2) X_2 + u$$

$$\frac{\mathrm{d}Y}{\mathrm{d}X_2} = 0 \quad \Rightarrow \quad \beta_2 = -2\beta_3 X_2 \quad \Rightarrow X_2 = \frac{-\beta_2}{2\beta_3} \quad - \quad \text{min } \mathit{or} \text{ max}$$

The coefficients β_2 and β_3 determine together the position of the regression line/
The impact of a unit change in X_2 on Y, (β_2 + 2$\beta_3 X_2$), is a linear function of X_2.

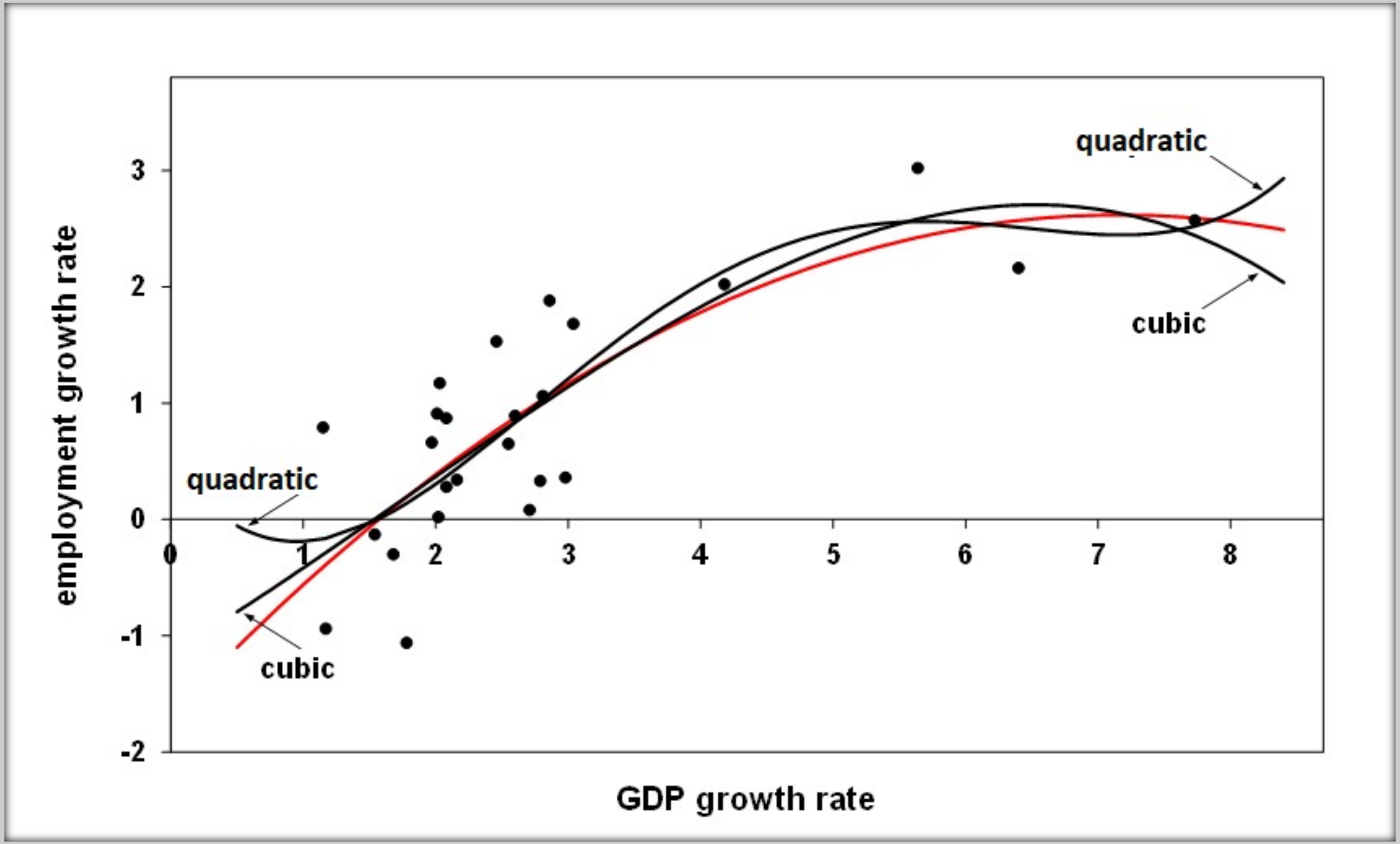
QUADRATIC EXPLANATORY VARIABLES

```
------------------------
EARNINGS |       Coef.
---------+--------------
       S |   -2.772317
     SSQ |    .1829731
   _cons |    22.25089
------------------------
```



Quadratic specification does not differ much from the semi-logarithmic ones within the sample range.

POLYNOMIAL EXPLANATORY VARIABLES



Diminishing marginal effects are standard in economic theory, justifying quadratic specifications, but economic theory seldom suggests that a relationship might sensibly be represented by higher-order polynomial.

INTERACTIVE EXPLANATORY VARIABLES

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 X_3 + u$$

$$X_2^* = X_2 - \bar{X}_2 \qquad X_3^* = X_3 - \bar{X}_3$$

$$X_2 = X_2^* + \bar{X}_2 \qquad X_3 = X_3^* + \bar{X}_3$$

$$Y = \beta_1 + \beta_2 (X_2^* + \bar{X}_2) + \beta_3 (X_3^* + \bar{X}_3) + \beta_4 (X_2^* + \bar{X}_2)(X_3^* + \bar{X}_3) + u$$

$$\beta_1^* = \beta_1 + \beta_2 \bar{X}_2 + \beta_3 \bar{X}_3 + \beta_4 \bar{X}_2 \bar{X}_3 \qquad \beta_2^* = \beta_2 + \beta_4 \bar{X}_3$$

$$\beta_3^* = \beta_3 + \beta_4 \bar{X}_2$$

$$Y = \beta_1^* + \beta_2^* X_2^* + \beta_3^* X_3^* + \beta_4 X_2^* X_3^* + u$$

The coefficients of X^*_2 and X^*_3 show the marginal effects of the variables when the other variable is at its sample mean.

RAMSEY'S RESET TEST OF FUNCTIONAL MISSPECIFICATION

$$Y = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + u$$

$$\hat{Y} = b_1 + \sum_{j=2}^{k} b_j X_j$$

$\hat{Y}^2$: **Add to regression specification**
Test its coefficient

If the *t* statistic for the coefficient of $\hat{Y}^2$ is significant, this indicates that some kind of nonlinearity may be present.

Cobb-Douglas Production Function with Technical Progress

Cobb-Douglas Production Function with technical progress (logarithmic and semi-logarithmic terms):

$$Y_t = A \cdot {K_t}^{\alpha} \cdot {L_t}^{\beta} \cdot e^{\gamma t} \cdot v_t$$

$$\ln Y_t = \ln A + \alpha \ln K_t + \beta \ln L_t + \gamma t + u_t$$

$$\frac{dY_t}{Y_t} = \alpha \cdot \frac{dK_t}{K_t} + \beta \cdot \frac{dL_t}{L_t} + \gamma \cdot dt + du_t$$

$$\frac{\Delta Y_t}{Y_t} = \alpha \cdot \frac{\Delta K_t}{K_t} + \beta \cdot \frac{\Delta L_t}{L_t} + \gamma + w_t \quad (dt = 1)$$

$$or \quad y_t = \alpha \cdot k_t + \beta \cdot l_t + \gamma + w_t - \quad in \quad growth \quad rates$$

Cobb-Douglas Production Function

$$\frac{\Delta Y_t}{Y_t} = \alpha \cdot \frac{\Delta K_t}{K_t} + \beta \cdot \frac{\Delta L_t}{L_t} + \gamma + w_t =$$
$$+ \left(K \cdot \frac{MPK}{Y}\right) \cdot \frac{\Delta K_t}{K_t} + \left(L \cdot \frac{MPL}{Y}\right) \cdot \frac{\Delta L_t}{L_t} + \gamma + w_t$$

$$e_K = \left(K \cdot \frac{MPK}{Y}\right) = \frac{rK}{Y}; \quad e_L = \left(L \cdot \frac{MPL}{Y}\right) = \frac{wL}{Y}$$

$$USSR,\ 1928 - 1987 \quad \hat{Y} = 0.82 \cdot K^{0.40} \cdot L^{0.60} \cdot e^{0.011t}$$

$$\alpha + \beta = 1 \quad - \text{ constant returns to scale}$$

$$\frac{Y}{L} = A \cdot \left(\frac{K}{L}\right)^{\alpha} e^{\gamma t} \qquad \ln\left(\frac{Y}{L}\right) = \ln A + \alpha \ln\left(\frac{K}{L}\right) + \gamma t$$

CES (Constant Elasticity of Substitution) Production Function

$$\textit{Elasticity of Substitution}: \quad \sigma_{LK} = \frac{d\ln(K/L)}{d\ln(Y'_L/Y'_K)}$$

$$\text{Marginal Rate of Substitution} \quad MRS_{KL} = -\frac{dK}{dL} = \frac{Y'_L}{Y'_K}$$

$$\textit{CES Function}: \quad Y = A \cdot (u \cdot K^{-\rho} + (1-u) \cdot L^{-\rho})^{-n/\rho} e^{\gamma t}$$

$$\rho \geq -1 \qquad n > 0 \qquad A > 0 \qquad 0 < u < 1 \qquad \sigma = \frac{1}{1+\rho}$$

$\rho = -1 \quad \Rightarrow \textit{function with linear isoquants} \qquad \rho \to 0 \quad \Rightarrow \textit{Cobb} - \textit{Douglas Function } (\sigma = 1)$

$\rho \to \infty \quad \Rightarrow \textit{Leontiev function}$

$$\ln\left(\frac{Y}{L}\right) = \ln A - \left(\frac{1}{\rho}\right) \cdot \ln\left[u \cdot \left(\frac{K}{L}\right)^{-\rho} + (1-u)\right] + \gamma \cdot t$$

$$\textit{USSR},\ 1928 - 1987 \quad \hat{Y} = 0.966 \cdot (0.4074 \cdot K^{-3.03} + 0.5926 \cdot L^{-3.03})^{-1/3.03} \cdot e^{0.0252t}$$

$$\sigma = \frac{1}{1+\rho} \approx 0.25$$

Non-linear Estimation

$$\hat{u}_i = Y_i - f(b, X_i) \qquad \{b_j\} - \ parameters\ to\ estimate$$

$$Non-Linear\ Least\ Squares\ (NLS):\quad F = \sum_{i}^{\Sigma} (Y_i - f(b, X_i))^2 \to \min$$

$$-2 \sum_i (Y_i - f(b, X_i)) \cdot f'_{bj}(b, X_i) = 0 \quad - \quad first\ order\ conditions$$

$$j = 1, \dots, k$$

$$Estimated\ by\ iterative\ procedures$$

CES function (constant returns to scale)

$$\ln\left(\frac{Y}{L}\right) = \ln A - \left(\frac{1}{\rho}\right) \cdot \ln\left[u \cdot \left(\frac{K}{L}\right)^{-\rho} + (1-u)\right] + \gamma \cdot t$$

In *EViews*:

NLS LYL = c(1) + (1/c(2)) * log(c(3) * KL^c(2) + (1 − c(3)) + c(4) * t