

(1) Regression through the origin

There are some occasions that we assume or '**force**' our estimation model without an intercept as

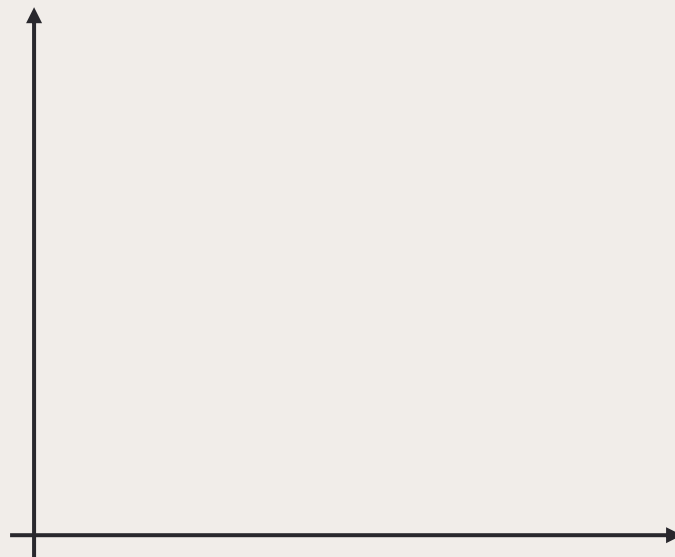
- $Y_i = \hat{\beta}_2 X_i + \hat{u}_i$

Obtaining the estimator from OLS,

- $\hat{\beta}_2 = \frac{\sum X_i Y_i}{\sum X_i^2}$

- $\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum X_i^2}$

- $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-1}$



Note that the d.f. is one unit less due to the disappearance of $\hat{\beta}_1$.

(1) Regression through the origin

Differences that should also be noted are

- $\sum \hat{u}_i X_i = 0$ but $\sum \hat{u}_i$ need not be zero

› r^2 can be negative, so we need another coefficient of determination, defined as **raw r^2**

- $raw\ r^2 = \frac{(\sum X_i Y_i)^2}{\sum X_i^2 \sum Y_i^2}$

Unless there is **very strong a priori expectation**, we should avoid zero intercept regression model and stick to the conventional intercept-present model, because it may lead to **specification error**.

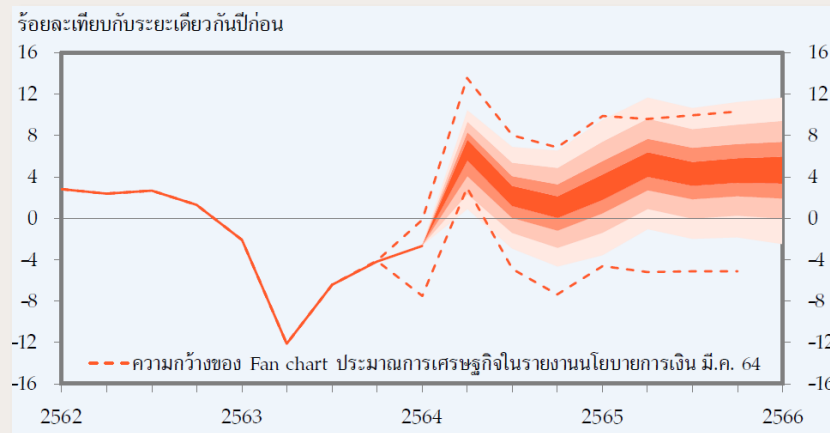
However, if $\hat{\beta}_1$ turns out to be statistically insignificant (from being zero), $\hat{\beta}_2$ is **a lot more precise** when estimated by the regression through the origin model.

(2) Out-of-sample prediction

Prediction is a very popular manner but note that the regression done so far is based on **historical data**. Prediction is most of the time inaccurate, especially looking forward into the future since there are many unknown uncertainties lying ahead.

For this topic, we will use the regression result to 'predict' Y value from an out-of-sample X_0 value of interest.

$$\circ \hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$$



ร้อยละ	2564			2565			2566	
	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1
> 18	0	0	0	0	0	0	0	0
16.0-18.0	0	0	0	0	0	0	0	1
14.0-16.0	0	0	0	0	1	1	1	1
12.0-14.0	1	0	0	1	3	2	2	3
10.0-12.0	7	0	0	3	6	4	5	6
8.0-10.0	23	2	2	6	12	9	10	11
6.0-8.0	28	6	5	12	19	16	17	16
4.0-6.0	16	17	12	18	21	21	21	19
2.0-4.0	14	28	20	21	18	21	19	18
0.0-2.0	9	22	24	18	11	14	13	13
(-2.0)-0.0	2	15	19	12	5	8	6	7
(-4.0)-(-2.0)	0	7	11	6	2	3	3	3
(-6.0)-(-4.0)	0	2	5	2	1	1	1	2
< -6	0	0	3	2	0	1	1	1

Source: Financial Report, June 2021, BOT

(2) Out-of-sample prediction

There are two types of prediction that we can make once we retrieved the estimators. The methods are different due to its assumption on the variance.

(1) Mean prediction: Providing that mean estimation follows this equation

$$\circ \hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$$

when $\hat{Y}_0$ is an estimator of $E(Y|X_0)$ while X_0 represents a value of interest. Let's consider an easier example here, given that

$$\circ \hat{Y}_i = -0.0144 + 0.7240X_i$$

If we are interested in out-of-sample $X_0 = 20$, then

$$\circ \hat{Y}_0 = -0.0144 + 0.7240(20) = 14.4656$$

(2) Out-of-sample prediction

Given that the variance of $\hat{Y}_0$ is, (no proof provided here)

$$\circ \text{ var}(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (x_i - \bar{X})^2} \right]$$

Again, we do not have the true value of σ^2 , so $\hat{\sigma}^2$ is replaced. We can then find the CI at the specific point of X_0 by

$$\circ \Pr \left[\hat{Y}_0 - \left(t_{\frac{\alpha}{2}} \cdot \text{se}_{\hat{Y}_0} \right) \leq Y_0 \leq \hat{Y}_0 + \left(t_{\frac{\alpha}{2}} \cdot \text{se}_{\hat{Y}_0} \right) \right] = 1 - \alpha$$

(2) Out-of-sample prediction

Example: Given that

- $\hat{Y}_i = -0.0144 + 0.7240X_i$ and $X_0 = 20, \hat{Y}_0 = 14.4656$
- $n = 13, \bar{X} = 12, \sum(x_i - \bar{X})^2 = 182, \hat{\sigma}^2 = 0.8936$

Step 1: Find the $\text{var}(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$

(2) *Out-of-sample prediction*

Step 2: find the $se_{\hat{y}_0}$

Step 3: find the 95% CI for $E(Y|X_0 = 20)$

Interpretation: the CI covers the mean value or 95 out of 100 times that the CI will cover true value $E(Y|X_0)$.

(2) Out-of-sample prediction

(2) Individual prediction: Contrast to the mean prediction, which estimates the variance around Y_0 , individual prediction focuses on forecasting error (fe), defined as

- $fe = \hat{Y}_0 - Y_0$

Therefore, we define the variance of this fe as

- $\text{var}(fe) = \text{var}(\hat{Y}_0 - Y_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$

Similarly, replacing the unknown σ^2 with the unbiased estimator $\hat{\sigma}^2$, we can derive the CI for Y_0 corresponding to X_0

- $\Pr \left[\hat{Y}_0 - \left(t_{\frac{\alpha}{2}} \cdot \text{se}_{fe} \right) \leq Y_0 \leq \hat{Y}_0 + \left(t_{\frac{\alpha}{2}} \cdot \text{se}_{fe} \right) \right] = 1 - \alpha$

(2) Out-of-sample prediction

Example: Given that

- $\hat{Y}_i = -0.0144 + 0.7240X_i$ and $X_0 = 20, \hat{Y}_0 = 14.4656$
- $n = 13, \bar{X} = 12, \sum(x_i - \bar{X})^2 = 182, \hat{\sigma}^2 = 0.8936$

Step 1: Find the $\text{var}(fe) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$

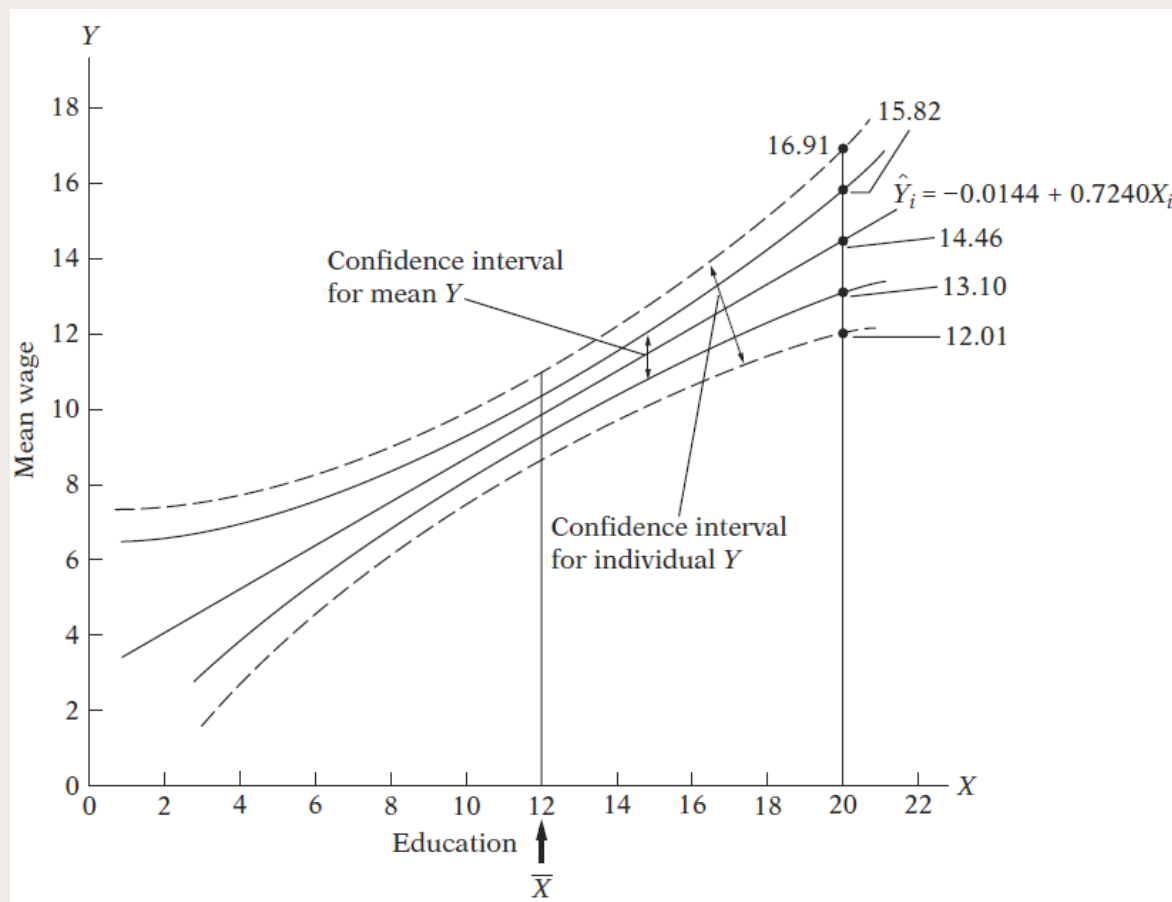
(2) *Out-of-sample prediction*

Step 2: Find the se_{fe}

Step 3: Find the 95% CI for Y_0 corresponding to $X_0 = 20$

(2) Out-of-sample prediction

Comparing CI band of mean and individual prediction.



(3) Data scaling

Sometimes when the result is reported, scaling can be difficult to make sense of. For example, if $\hat{\beta}_2 = 3.054e^{-15}$ which is not very effective for communication. Thus, data scaling can fix this **without affecting the result or the interpretation**. See the examples below.

Original dataset

Observation	X_i	Y_i
1	200,000	85,000
2	100,000	60,000
3	180,000	80,000
4	60,000	40,000
5	100,000	30,000
6	500,000	100,000
7	.	.
8	.	.
.	.	.

Transformed dataset

Observation	X_i	Y_i
1	200	85
2	100	60
3	180	80
4	60	40
5	100	30
6	500	100
7	.	.
8	.	.
.	.	.

(3) Data scaling

1. Scale both sides

From the table in the previous page, we transform both X_i and Y_i down by *dividing* all the values with 1,000. The unit of the data becomes **thousands** Baht. See the result of the regression below.

Original dataset
(Baht)

Source	SS	df	MS	Number of obs	=	52
				F(1, 50)	=	21.22
Model	1.0675e+11	1	1.0675e+11	Prob > F	=	0.0000
Residual	2.5157e+11	50	5.0314e+09	R-squared	=	0.2979
				Adj R-squared	=	0.2839
Total	3.5832e+11	51	7.0259e+09	Root MSE	=	70932

exp	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
inc	.2797096	.0607242	4.61	0.000	.1577414	.4016778
_cons	38687.86	16223.66	2.38	0.021	6101.68	71274.03

Transformed dataset
(thousands Baht)

Source	SS	df	MS	Number of obs	=	52
				F(1, 50)	=	21.22
Model	106751.785	1	106751.785	Prob > F	=	0.0000
Residual	251567.523	50	5031.35046	R-squared	=	0.2979
				Adj R-squared	=	0.2839
Total	358319.308	51	7025.86878	Root MSE	=	70.932

thexp	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
thinc	.2797096	.0607242	4.61	0.000	.1577414	.4016778
_cons	38.68786	16.22366	2.38	0.021	6.10168	71.27403

(3) *Data scaling*

1. Scale both sides

- For $\hat{\beta}_1$, when household income is zero, the intercept (autonomous consumption) is 38,687.86 Baht (38.68786 thousands Baht).
- For $\hat{\beta}_2$, when household income increases by 1 Baht (1 thousand Baht), on average household expenditure increases by 0.2797 Baht (0.2797 thousand Baht).

Notice that there is no difference when we interpret the coefficients.

On the other hand, if we multiply both sides with 1,000, how would it change the coefficient(s), standard error and the confidence interval.

(3) Data scaling

2. Scale Y_i

Now imagine we only scale the Y_i , dividing them with 1,000, the data table is displayed below.

Original dataset

Observation	X_i	Y_i
1	200,000	85,000
2	100,000	60,000
3	180,000	80,000
4	60,000	40,000
5	100,000	30,000
6	500,000	100,000
7	.	.
8	.	.
.	.	.

Transformed dataset

Observation	X_i	Y_i
1	200,000	85
2	100,000	60
3	180,000	80
4	60,000	40
5	100,000	30
6	500,000	100
7	.	.
8	.	.
.	.	.

(3) Data scaling

2. Scale Y_i

From the table in the previous page, the unit of expenditure becomes **thousands** Baht while the unit of income remains the same. See the result of the regression below.

Original dataset
(Baht)

Source	SS	df	MS	Number of obs	=	52
				F(1, 50)	=	21.22
Model	1.0675e+11	1	1.0675e+11	Prob > F	=	0.0000
Residual	2.5157e+11	50	5.0314e+09	R-squared	=	0.2979
				Adj R-squared	=	0.2839
Total	3.5832e+11	51	7.0259e+09	Root MSE	=	70932

exp	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
inc	.2797096	.0607242	4.61	0.000	.1577414	.4016778
_cons	38687.86	16223.66	2.38	0.021	6101.68	71274.03

Transformed dataset
(Y_i thousands Baht)

Source	SS	df	MS	Number of obs	=	52
				F(1, 50)	=	21.22
Model	106751.785	1	106751.785	Prob > F	=	0.0000
Residual	251567.523	50	5031.35046	R-squared	=	0.2979
				Adj R-squared	=	0.2839
Total	358319.308	51	7025.86878	Root MSE	=	70.932

thexp	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
inc	.0002797	.0000607	4.61	0.000	.0001577	.0004017
_cons	38.68786	16.22366	2.38	0.021	6.10168	71.27403

(3) *Data scaling*

2. Scale Y_i

- For $\hat{\beta}_1$, when household income is zero, the intercept (autonomous consumption) is 38.68786 thousands Baht.
- For $\hat{\beta}_2$, when household income increases by 1 Baht, on average household expenditure increases by 0.0002797 thousand Baht.

Notice that, again, there is no difference when we interpret the coefficients.

On the other hand, if we multiply Y_i with 1,000, how would it change the coefficient(s), standard error and the confidence interval.

(3) Data scaling

3. Scale X_i

Now imagine we only scale the X_i , dividing them with 1,000, the data table is displayed below.

Original dataset

Observation	X_i	Y_i
1	200,000	85,000
2	100,000	60,000
3	180,000	80,000
4	60,000	40,000
5	100,000	30,000
6	500,000	100,000
7	.	.
8	.	.
.	.	.

Transformed dataset

Observation	X_i	Y_i
1	200	85,000
2	100	60,000
3	180	80,000
4	60	40,000
5	100	30,000
6	500	100,000
7	.	.
8	.	.
.	.	.

(3) Data scaling

3. Scale X_i

From the table in the previous page, the unit of income becomes **thousands** Baht while the unit of expenditure remains the same. See the result of the regression below.

Original dataset
(Baht)

Source	SS	df	MS	Number of obs	=	52
				F(1, 50)	=	21.22
Model	1.0675e+11	1	1.0675e+11	Prob > F	=	0.0000
Residual	2.5157e+11	50	5.0314e+09	R-squared	=	0.2979
				Adj R-squared	=	0.2839
Total	3.5832e+11	51	7.0259e+09	Root MSE	=	70932

exp	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
inc	.2797096	.0607242	4.61	0.000	.1577414	.4016778
_cons	38687.86	16223.66	2.38	0.021	6101.68	71274.03

Transformed dataset
(X_i thousands Baht)

Source	SS	df	MS	Number of obs	=	52
				F(1, 50)	=	21.22
Model	1.0675e+11	1	1.0675e+11	Prob > F	=	0.0000
Residual	2.5157e+11	50	5.0314e+09	R-squared	=	0.2979
				Adj R-squared	=	0.2839
Total	3.5832e+11	51	7.0259e+09	Root MSE	=	70932

exp	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
thinc	279.7096	60.72423	4.61	0.000	157.7414	401.6778
_cons	38687.86	16223.66	2.38	0.021	6101.68	71274.03

(3) *Data scaling*

3. Scale X_i

- For $\hat{\beta}_1$, when household income is zero, the intercept (autonomous consumption) is 38,687.86 Baht.
- For $\hat{\beta}_2$, when household income increases by 1 thousand Baht, on average household expenditure increases by 279.7096 Baht.

Notice that, again, there is no difference when we interpret the coefficients.

On the other hand, if we multiply X_i with 1,000, how would it change the coefficient(s), standard error and the confidence interval.

(3) Data scaling

Summary

Scaling	Factor	$\hat{\beta}_1, \text{se}_{\hat{\beta}_1}, \text{CI}$	$\hat{\beta}_2, \text{se}_{\hat{\beta}_2}, \text{CI}$
Both Y_i and X_i	w	w	-
	$\frac{1}{w}$	$\frac{1}{w}$	-
Y_i	w	w	w
	$\frac{1}{w}$	$\frac{1}{w}$	$\frac{1}{w}$
X_i	w	-	$\frac{1}{w}$
	$\frac{1}{w}$	-	w

(4) *Functional forms*

1. Log-log model

Sometimes called **double-log**, or **log-linear** models, a log-log model takes a form of

- $Y = \beta_1 X_i^{\beta_2} e^{u_i}$; we can linearize the function by taking log on both sides.

We will find that the slope and elasticity has a very interesting properties, by differentiation.

- Slope
- Elasticity

(4) *Functional forms*

1. Log-linear model

Original dataset

Observation	X_i	Y_i
1	200,000	85,000
2	100,000	60,000
3	180,000	80,000
4	60,000	40,000
5	100,000	30,000
6	500,000	100,000
7	.	.
8	.	.
.	.	.

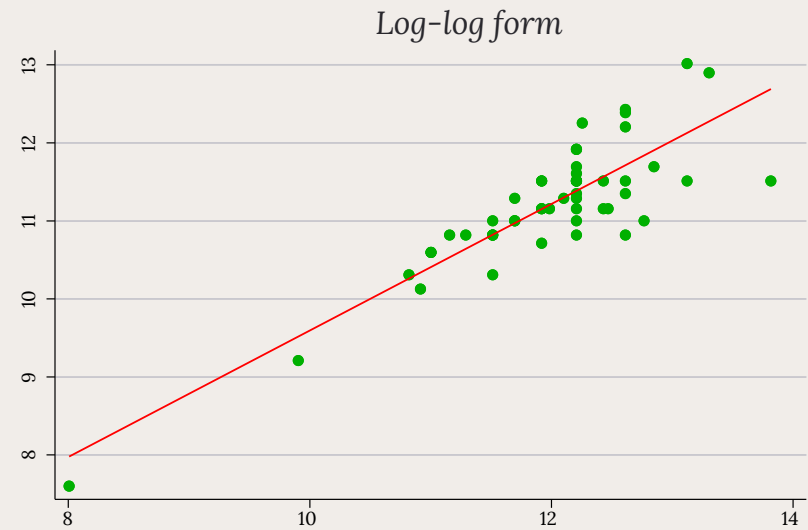
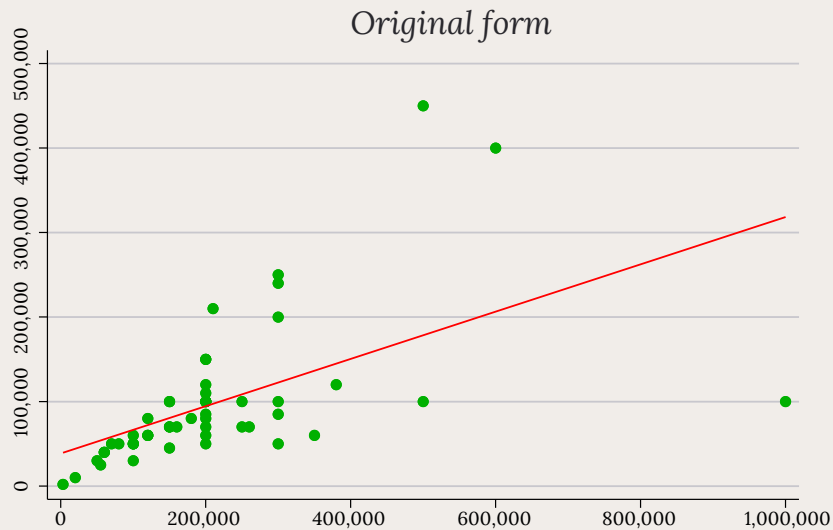
Transformed dataset

Observation	X_i	Y_i
1	12.20607	11.35041
2	11.51293	11.0021
3	12.10071	11.28978
4	11.0021	10.59663
5	11.51293	10.30895
6	12.12236	11.51293
7	.	.
8	.	.
.	.	.

3.3 Measuring the goodness of fit

(4) Functional forms

1. Log-linear model



3.8 Other topics

(4) *Functional forms*

1. Log-linear model

Original form

Source	SS	df	MS	Number of obs	=	52
				F(1, 50)	=	21.22
Model	1.0675e+11	1	1.0675e+11	Prob > F	=	0.0000
Residual	2.5157e+11	50	5.0314e+09	R-squared	=	0.2979
				Adj R-squared	=	0.2839
Total	3.5832e+11	51	7.0259e+09	Root MSE	=	70932

exp	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
inc	.2797096	.0607242	4.61	0.000	.1577414	.4016778
_cons	38687.86	16223.66	2.38	0.021	6101.68	71274.03

Log-log form

Source	SS	df	MS	Number of obs	=	52
				F(1, 50)	=	137.35
Model	26.2503087	1	26.2503087	Prob > F	=	0.0000
Residual	9.55577384	50	.191115477	R-squared	=	0.7331
				Adj R-squared	=	0.7278
Total	35.8060825	51	.702080049	Root MSE	=	.43717

ln_exp	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
ln_inc	.8111731	.069214	11.72	0.000	.6721526	.9501936
_cons	1.485608	.8318879	1.79	0.080	-.1852878	3.156504

(4) *Functional forms*

1. Log-linear model

To interpret this model, we should write the result down as an equation

- $\ln \widehat{consm}_i = 1.4856 + 0.8112 \ln inc_i$

- An increase by **one percent** of income is associated with expenditure increase of $\widehat{\beta}_2$ **percent**.

(4) *Functional forms*

2. Log-lin model

The two following models can be called semi-log model. The first one here is log-lin model which takes a form of

- $\ln Y = \beta_1 + \beta_2 X_i$

We will find the slope and elasticity again.

- Slope

- Elasticity

(4) *Functional forms*

2. Log-lin model

Original dataset

Observation	X_i	Y_i
1	200,000	85,000
2	100,000	60,000
3	180,000	80,000
4	60,000	40,000
5	100,000	30,000
6	500,000	100,000
7	.	.
8	.	.
.	.	.

Transformed dataset

Observation	X_i	Y_i
1	200,000	11.35041
2	100,000	11.0021
3	180,000	11.28978
4	60,000	10.59663
5	100,000	10.30895
6	500,000	11.51293
7	.	.
8	.	.
.	.	.

(4) Functional forms

2. Log-lin model

Original form

Source	SS	df	MS	Number of obs	=	52
				F(1, 50)	=	21.22
Model	1.0675e+11	1	1.0675e+11	Prob > F	=	0.0000
Residual	2.5157e+11	50	5.0314e+09	R-squared	=	0.2979
				Adj R-squared	=	0.2839
Total	3.5832e+11	51	7.0259e+09	Root MSE	=	70932

exp	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
inc	.2797096	.0607242	4.61	0.000	.1577414	.4016778
_cons	38687.86	16223.66	2.38	0.021	6101.68	71274.03

Log-lin form

Source	SS	df	MS	Number of obs	=	52
				F(1, 50)	=	21.94
Model	10.9184499	1	10.9184499	Prob > F	=	0.0000
Residual	24.8876325	50	.497752651	R-squared	=	0.3049
				Adj R-squared	=	0.2910
Total	35.8060825	51	.702080049	Root MSE	=	.70552

ln_exp	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
inc	2.83e-06	6.04e-07	4.68	0.000	1.62e-06	4.04e-06
_cons	10.60821	.1613665	65.74	0.000	10.2841	10.93233

(4) *Functional forms*

2. Log-lin model

To interpret $\hat{\beta}_1$, now we can assume that income is zero from this equation

- $\ln \widehat{consmp}_i = 10.60821 + 0.00000283 inc_i$

(4) *Functional forms*

2. Log-lin model

Now consider interpreting $\hat{\beta}_2$ from the same result.

- $\ln \widehat{consmp}_i = 10.60821 + 0.00000283inc_i$
- **Approximate measure:** an increase by **one unit of income** is associated with expenditure increase of **$100 \cdot \hat{\beta}_2$ percent**. (suitable for small value of $\hat{\beta}_2$)
- **Exact measure:** an increase by **one unit of income** is associated with expenditure increase of **$(100 \cdot e^{\hat{\beta}_2}) - 100$ percent**.

(4) *Functional forms*

2. Log-lin model

Here is a comparison between approximate and exact change.

$\widehat{\beta}_2$	$100 \cdot \widehat{\beta}_2$ (approximate)	$(100 \cdot e^{\widehat{\beta}_2}) - 100$ (exact)	Diff (approximate-exact)
0.005	0.5	0.5012521	-0.0012521
0.01	1	1.0050167	-0.0050167
0.015	1.5	1.5113065	-0.0113065
0.02	2	2.020134	-0.020134
0.025	2.5	2.5315121	-0.0315121
0.03	3	3.0454534	-0.0454534
0.035	3.5	3.5619709	-0.0619709
0.04	4	4.0810774	-0.0810774
0.045	4.5	4.602786	-0.102786
0.05	5	5.1271096	-0.1271096
0.055	5.5	5.6540615	-0.1540615
0.06	6	6.1836547	-0.1836547
0.065	6.5	6.7159024	-0.2159024
0.07	7	7.2508181	-0.2508181
0.075	7.5	7.7884151	-0.2884151
0.08	8	8.3287068	-0.3287068
.	.	.	.
.	.	.	.

(4) *Functional forms*

3. Lin-log model

The lin-log model takes a form of

- $Y = \beta_1 + \beta_2 \ln X_i$

We will find the slope and elasticity again.

- Slope

- Elasticity

(4) *Functional forms*

3. Lin-log model

Original dataset

Observation	X_i	Y_i
1	200,000	85,000
2	100,000	60,000
3	180,000	80,000
4	60,000	40,000
5	100,000	30,000
6	500,000	100,000
7	.	.
8	.	.
.	.	.

Transformed dataset

Observation	X_i	Y_i
1	12.20607	85,000
2	11.51293	60,000
3	12.10071	80,000
4	11.0021	40,000
5	11.51293	30,000
6	12.12236	100,000
7	.	.
8	.	.
.	.	.

(4) Functional forms

3. Lin-log model

Original form

Source	SS	df	MS	Number of obs	=	52
				F(1, 50)	=	21.22
Model	1.0675e+11	1	1.0675e+11	Prob > F	=	0.0000
Residual	2.5157e+11	50	5.0314e+09	R-squared	=	0.2979
				Adj R-squared	=	0.2839
Total	3.5832e+11	51	7.0259e+09	Root MSE	=	70932

exp	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
inc	.2797096	.0607242	4.61	0.000	.1577414	.4016778
_cons	38687.86	16223.66	2.38	0.021	6101.68	71274.03

Lin-log form

Source	SS	df	MS	Number of obs	=	52
				F(1, 50)	=	21.42
Model	1.0746e+11	1	1.0746e+11	Prob > F	=	0.0000
Residual	2.5086e+11	50	5.0172e+09	R-squared	=	0.2999
				Adj R-squared	=	0.2859
Total	3.5832e+11	51	7.0259e+09	Root MSE	=	70832

exp	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
ln_inc	51899.89	11214.45	4.63	0.000	29375.01	74424.76
_cons	-524013.9	134787.1	-3.89	0.000	-794741.8	-253286

(4) *Functional forms*

3. Lin-log model

Again, it does not make sense interpreting $\hat{\beta}_1$, so we are going to consider only $\hat{\beta}_2$ from this equation.

- $\widehat{consm}_i = -524,013.9 + 51,899.89 \ln inc_i$
- **Approximate measure:** an increase by **one percent of income** is associated with expenditure increase of $\frac{\hat{\beta}_2}{100}$ **unit**. (suitable for small value of change of independent variable)
- **Exact measure:** an increase by **one percent of income** is associated with expenditure increase of $\hat{\beta}_2 \cdot \ln(1.01)$ **unit**.

(4) *Functional forms*

4. Reciprocal model and log-reciprocal

These two model, respectively, take a form of

- $Y = \beta_1 + \beta_2(\frac{1}{X_i})$
- $\ln Y = \beta_1 - \beta_2(\frac{1}{X_i})$

Further details can be found in the textbook.

