# Lecture 4: Prediction, Goodness-of-Fit and Modelling Issues

Shuo Liu

UCLA Summer School Econ 103

July 2, 2017

# Outline

# Least Square Prediction

Assume we use sample data $\{(x_i, y_i)\}_{i=1}^{n}$ to estimate simple linear regression model:

$$y = \beta_1 + \beta_2 x + e \tag{1}$$

$\Longrightarrow$

$$\hat{y} = b_1 + b_2 x \tag{2}$$

- Assume $(x_0, y_0)$ is a data point **outside** the sample data, and given $x_0$, we want to use estimated model to predict $y_0$;
- We must assume that $y_0$ and $x_0$ are related to one another by the same regression model that describes our sample data;

$$y_0 = \beta_1 + \beta_2 x_0 + e_0 \tag{3}$$

where $e_0$ is a random error.

# Least Square Prediction

- It is intuitive that the least square (point) predictor of $y_0$ comes from the fitted line:

$$\hat{y_0} = b_1 + b_2 x_0 \tag{4}$$

- To define how well this predictor performs, we define the **forecast error**:

$$f = y_0 - \hat{y_0} = (\beta_1 + \beta_2 x_0 + e_0) - (b_1 + b_2 x_0) \tag{5}$$

and we have

$$E(f) = E(\beta_1 + \beta_2 x_0 + e_0) - E(b_1 + b_2 x_0) = (\beta_1 + \beta_2 x_0 + 0) - (\beta_1 + \beta_2 x_0) = 0 \tag{6}$$

what does (6) mean?

- If SR1-SR5 hold, $\hat{y_0}$ is also the best linear unbiased predictor (BLUP) of $y_0$.

# Least Square Prediction

- To provide more information on reliability of the predictor, we also need to get **variance of the forecast** $f$

$$Var(f) = Var(y_0 - \hat{y_0}) = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \tag{7}$$

The variance of forecast is smaller when:

1. the uncertainty in the random error $\sigma^2$ is smaller;
2. the sample size $n$ is larger;
3. the sum of squares of deviation from sample mean of explanatory variable $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is larger;
4. the value of $(x_0 - \bar{x})^2$ is small.

# Least Square Prediction

- Again, if we do not know $\sigma^2$, in practice, we use $\hat{\sigma}^2$:

$$\hat{Var}(f) = \hat{\sigma}^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \tag{8}$$

and the standard error of the forecast is:

$$\hat{Se}(f) = \sqrt{\hat{Var}(f)} \tag{9}$$

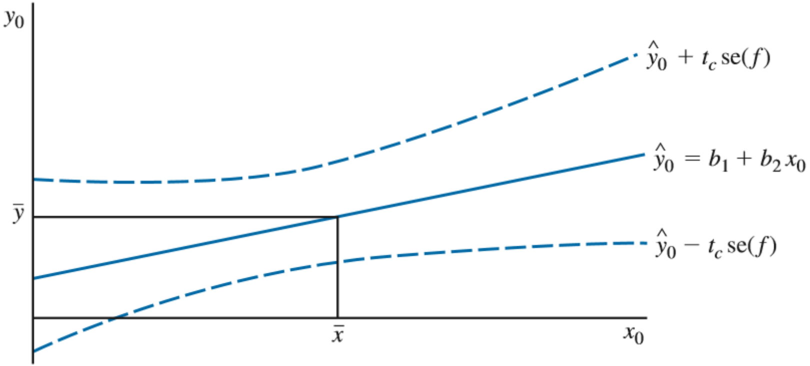
With **point predictor** and **standard error of forecast**, given $\alpha$, we can construct $100(1-\alpha)\%$ confidence interval as:

$$\left[ \hat{y_0} - t_c \hat{Se}(f), \hat{y_0} + t_c \hat{Se}(f) \right] \tag{10}$$

where

$$P(-t_c < t_{n-2} < t_c) = 1 - \alpha \tag{11}$$

# Least Square Prediction



# Least Square Prediction

The estimated variance of the forecast error is also:

$$\hat{Var}(f) = \hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

$$= \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{n} + (x_0 - \bar{x})^2 \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{n} + (x_0 - \bar{x})^2 \hat{Var}(b_2)$$

# Measure Goodness-of-Fit

- Besides "single hypothesis testing", we can also obtain the measure of goodness-of-fit to evaluate the model:
  1. whether the **variation** of explanatory variable $x$ "explain" as much as possible the **variation** of dependent variable $y$;
  2. whether the model fits the sample data well.
- **variation** means the "sum of squares of deviation from corresponding sample mean": $\sum_{i=1}^{n}(x-\bar{x})^2$
- **Theoretical Model**:

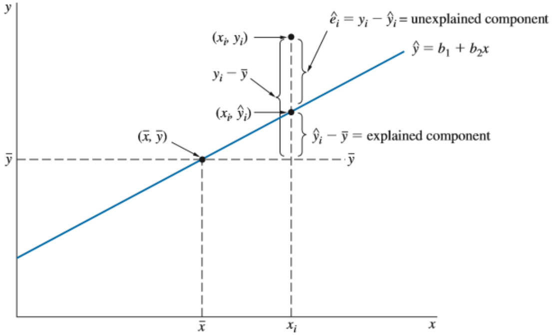
$$y_i = \underbrace{E(y_i)}_{\text{explainable}} + \underbrace{e_i}_{\text{unexplainable}} \tag{12}$$

  where $E(y_i)$ is the explainable/systematic part and $e_i$ is the unexplainable/unsystematic part.
- **Analogously in Fitted Model**:

$$y_i = \hat{y_i} + \hat{e_i} \tag{13}$$

Then we further have for each observed data point $(x_i, y_i)$:

$$\underbrace{y_i - \bar{y}}_{\text{total deviation}} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{explained component}} + \underbrace{\hat{e}_i}_{\text{unexplained component}} \tag{14}$$



# Measure Goodness-of-Fit

To further get **total variation**:

$$\begin{aligned}
\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{SST} &= \sum_{i=1}^{n}(\hat{y}_i-\bar{y}+\hat{e}_i)^2 \\
&= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}\hat{e_i}^2+2\sum_{i=1}^{n}\left[(\hat{y}_i-\bar{y})\,\hat{e}_i\right] \\
&= \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{SSR}+\underbrace{\sum_{i=1}^{n}\hat{e_i}^2}_{SSE}
\end{aligned}$$

- SST: total sum of squares, same as the sample variance of dependent variable $y$ that is to be explained $s_y^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$;
- SSR: sum of squares due to the regression, replacing observed $y_i$ with predicted $\hat{y}_i$;
- SSE: sum of squares due to error.

# Measure Goodness-of-Fit

Why do we have $2\sum_{i=1}^{n}\left[\left(\hat{y}_i - \bar{y}\right)\hat{e}_i\right] = 0$?

$$\begin{aligned}\sum_{i=1}^{n}\left[\left(\hat{y}_i - \bar{y}\right)\hat{e}_i\right] &= \sum_{i=1}^{n}\left[\left(b_1 + b_2 x_i - \bar{y}\right)\hat{e}_i\right] \\ &= b_1\sum_{i=1}^{n}\hat{e}_i + b_2\sum_{i=1}^{n}(x_i\hat{e}_i) - \bar{y}\sum_{i=1}^{n}\hat{e}_i\end{aligned}$$

- OLS estimation first order condition $\frac{\partial S(b_1,b_2)}{\partial b_2} = 0$ gives us:

$$\frac{\partial S(b_1,b_2)}{\partial b_2} = -2\sum_{i=1}^{n} x_i(y_i - b_1 - b_2 x_i) = -2\sum_{i=1}^{n}(x_i\hat{e}_i) = 0 \tag{15}$$

- OLS estimation first order condition $\frac{\partial S(b_1,b_2)}{\partial b_1} = 0$ gives us:

$$\frac{\partial S(b_1,b_2)}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - b_1 - b_2 x_i) = -2\sum_{i=1}^{n}\hat{e}_i = 0 \tag{16}$$

# Measure Goodness-of-Fit

Then to evaluate the model in the sense that **whether the (estimated) variation from regression "explains" large part of total (observed) variation**, we define **the coefficient of determination**:

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} \tag{17}$$

- The closer $R^2$ is to 1, the closer $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is to $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$, to closer sample value $y_i$ is to the fitted regression equation $\hat{y}_i$;
- What if $R^2 = 1$?
- What if $R^2$ is closer to 0?
- **Note** that in practice, to evaluate the model, we put less weight on $R^2$ than the "significance" of parameters.

# Measure Goodness-of-Fit

- More intuitively, we can interpret $R^2$ as: the proportion of the variation in $y$ about its mean that is explained by the regression model;
- $R^2$ is correlated with the **sample correlation coefficient**:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{18}$$

where

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{19}$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \tag{20}$$

$$s_{xy} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}} \tag{21}$$

# Measure Goodness-of-Fit

Two relationships between $R^2$ and $r_{xy}$:

- $r_{xy}^2 = R^2$;
- $R^2$ can also be computed as the square of the sample correlation coefficient between $y_i$ and $\hat{y}_i = b_1 + b_2 x_i$, as given fixed sample data, $b_1$ and $b_2$ are fixed.

Also we define **adjusted-$R^2$** (usually used to evaluate multi-regression model) as:

$$\bar{R}^2 = 1 - \frac{SSE/(n-K)}{SST/(n-1)} \tag{22}$$

where $K$ is number of population parameters in the linear model.

# Reporting the Results

If we do a regression, the key ingredients to report are:

- the OLS estimators;
- the standard errors of OLS estimators (or equivalently the t-values (the value of t statistic));
- an indication of statistical significance;
- the coefficient of determination $R^2$.

Food expenditure example, we have:

- $FOODEXP$: weekly food expenditure by a household of size 3, in dollars;
- $INCOME$: weekly household income, in \$100.

```
. reg food_exp income

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =   23.79
       Model |  190626.984     1  190626.984           Prob > F      =  0.0000
    Residual |  304505.176    38  8013.2941            R-squared     =  0.3850
-------------+------------------------------           Adj R-squared =  0.3688
       Total |   495132.16    39  12695.6964           Root MSE      =  89.517

------------------------------------------------------------------------------
    food_exp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   10.20964   2.093264     4.88   0.000     5.972052    14.44723
       _cons |     83.416   43.41016     1.92   0.062    -4.463279    171.2953
------------------------------------------------------------------------------
```

Report the result:

$$\underset{(se)}{\widehat{FOODEXP}} = \underset{(43.41)}{83.416} + \underset{(2.09)^{***}}{10.21}\ INCOME, \quad R^2 = 0.385 \tag{23}$$

# Reporting the Results

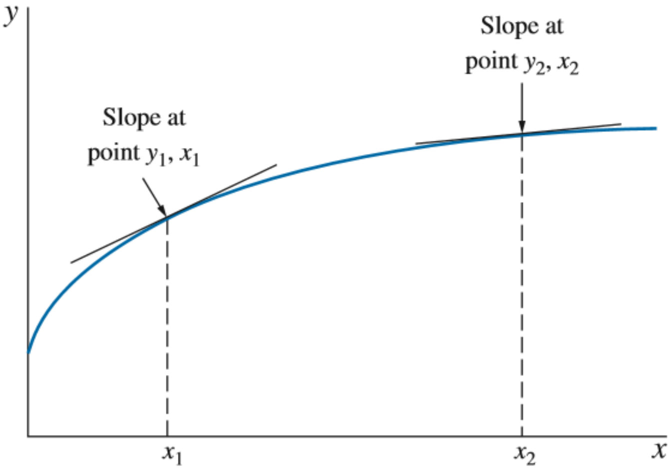
where

$$\begin{aligned} &* \text{ indicates significant at the 10\% level} \\ &** \text{ indicates significant at the 5\% level} \\ &*** \text{ indicates significant at the 1\% level} \end{aligned}$$

Based the output table, if $b_1 = 83.416$ and $b_2 = 10.21$, what are the values of the following items?

- $V\hat{a}r(b_1)$, $V\hat{a}r(b_2)$?
- $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{e_i}^2}{n-2}$?
- ...

# Choosing Functional Form




A nonlinear relationship between food expenditure and income.

# Choosing Functional Form

$$SLOPE = \frac{dy}{dx}, \quad \eta_{yx} = \frac{dy/y}{dx/x} = \frac{dy}{dx}\frac{x}{y} = SLOPE * \frac{x}{y} \tag{24}$$

**Table 4.1** **Some Useful Functions, their Derivatives, Elasticities and Other Interpretation**

| Name | Function | Slope = $dy/dx$ | Elasticity |
|---|---|---|---|
| **Linear** | $y = \beta_1 + \beta_2 x$ | $\beta_2$ | $\beta_2 \frac{x}{y}$ |
| **Quadratic** | $y = \beta_1 + \beta_2 x^2$ | $2\beta_2 x$ | $(2\beta_2 x)\frac{x}{y}$ |
| **Cubic** | $y = \beta_1 + \beta_2 x^3$ | $3\beta_2 x^2$ | $(3\beta_2 x^2)\frac{x}{y}$ |
| **Log-Log** | $\ln(y) = \beta_1 + \beta_2 \ln(x)$ | $\beta_2 \frac{y}{x}$ | $\beta_2$ |
| **Log-Linear** | $\ln(y) = \beta_1 + \beta_2 x$ | $\beta_2 y$ | $\beta_2 x$ |
| | or, a 1 unit change in $x$ leads to (approximately) a 100 $\beta_2$% change in $y$ | | |
| **Linear-Log** | $y = \beta_1 + \beta_2 \ln(x)$ | $\beta_2 \frac{1}{x}$ | $\beta_2 \frac{1}{y}$ |
| | or, a 1% change in $x$ leads to (approximately) a $\beta_2/100$ unit change in $y$ | | |

# Choosing Functional Form

Food expenditure example:

**Model 1**:

$$FOODEXP = \beta_1 + \beta_2 INCOME \tag{25}$$

$\Longrightarrow$

$$\underset{(se)}{\widehat{FOODEXP}} = \underset{(43.41)}{83.416} + \underset{(2.09)^{***}}{10.21}\, INCOME, \quad R^2 = 0.385 \tag{26}$$

How to interpret $b_2 = 10.21$?

**Model 2**:

$$FOODEXP = \beta_1 + \beta_2 ln(INCOME) \tag{27}$$

$\Longrightarrow$

$$\underset{(se)}{\widehat{FOODEXP}} = \underset{(84.24)}{-97.19} + \underset{(28.8)^{***}}{132.17}\, ln(INCOME), \quad R^2 = 0.357 \tag{28}$$

How to interpret $b_2 = 132.17$? How much will household additionally spend on food from an additional \$100 income? Is it still constant for households of all income levels?

Continue with **Model 2**:

$$\frac{d\widehat{FOODEXP}}{dln(INCOME)} = 132.17 = \frac{d\widehat{FOODEXP}}{dINCOME} INCOME \tag{29}$$

$\Longrightarrow$

$$\frac{d\widehat{FOODEXP}}{dINCOME} = \frac{\frac{d\widehat{FOODEXP}}{dln(INCOME)}}{INCOME} = \frac{132.17}{INCOME} \tag{30}$$

How about a household with \$2000 weekly income? remember the unit of INCOME is \$100.

# Choosing Functional Form

Up to now, we need to choose a functional form (evaluate whether the assumed model form is good or not):

- consistent with economic theory;
- population parameters are “significant” (significantly different from zero);
- fit the data well/explain large proportion of total variation;
- **Next: satisfy assumptions SR1-SR6**.

For simple linear regression model:

$$y = \beta_1 + \beta_2 x + e \tag{31}$$

We will mainly focus on the SR3, SR4 and SR6.

- **SR3(homoskedasticity)**: for each value of $x$, the conditional variance of the random error is

$$Var(e|x) = \sigma^2 \Longrightarrow Var(y|x) = \sigma^2$$

- **SR4(no serial correlation)**: the covariance between any pair of random errors,

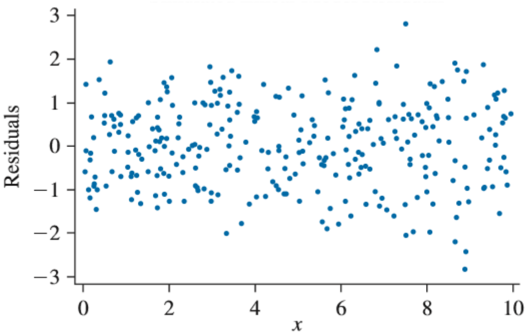
$$Cov(e_i, e_j) = 0 \quad \text{for all } i \neq j \Longrightarrow Cov(y_i, y_j) = 0 \quad \text{for all } i \neq j$$

- **SR6(normality)**:

$$e \sim N(0, \sigma^2) \Longrightarrow y|x \sim N(\beta_1 + \beta_2 x, \sigma^2)$$

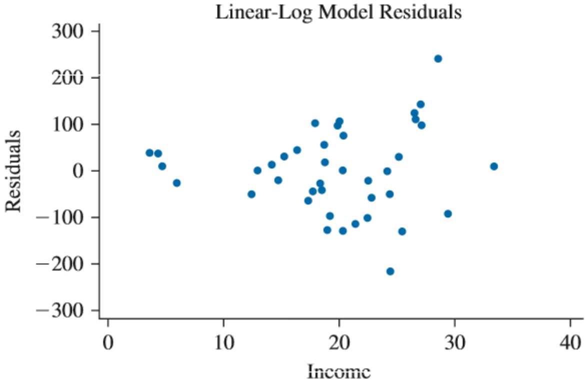
For testing **SR3(homoskedasticity)**, we can refer to diagnostic residual plots:

**random and homoskedastic residuals**:



# Choosing Functional Form

**Heteroskedastic residuals**:



# Choosing Functional Form

For testing **SR4(no serial correlation)**, we can use the obtained residual data $\{\hat{e_i}\}_{i=1}^{n}$ to do the regression:

$$\hat{e_i} = \beta_1 + \beta_2 \hat{e_{i-1}} + w_i \tag{32}$$

If $\beta_1$ and $\beta_2$ are significant, we can conclude residuals are serially correlated.

# Choosing Functional Form

For testing **SR6(normality)**, we can refer to **Jarque-Bera (JB) Test** to test normality (there are also many other formal tests):

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right) \tag{33}$$

where $n$ is sample size, $S$ is skewness, $K$ is kurtosis. (standard normal distribution has skewness as zero, kurtosis as 3)

- When residuals are normally distributed (SR6 applies), JB statistic $\sim \chi^2_{(2)}$
- We set $H_0$ : residuals are normally distributed, then we reject $H_0$ when the value of JB statistic exceeds a critical value of $\chi^2_{c(2)}$ (remember $\chi^2$ test is always right-tail test)
- For example: for $\alpha = 0.05$, critical value is 5.99; for $\alpha = 0.01$, critical value is 9.21; $JB = 6.21$; Then:
    1. since $6.21 > 5.99$, we reject SR6 at the 5% level of significance;
    2. since $6.21 < 9.21$, we can not reject SR6 at the 1% level of significance.
- Remember "significance level" can be interpreted as: probability of Type I error.