Review: The Bivariate Regression Model

- Bivariate Regression analysis has a clear model for two variables in mind:
  - One variable is given ***exogenously***. This variable is called the independent variable and denoted by X.
    In theory, the independent variable can be controlled and its measurement values can be replicated.
  - The response, ***endogenous***, or dependent variable Y is linked by a linear function to the independent variable

    $$Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$$

    where ε_i is a random disturbance term.
- Regression traces the conditional distribution of *Y* given any fixed value of *X*. If we assume linearity and equal distributions, we get $\mu_{y_i|x_{i1}} \equiv E(y_i \mid x_i) = \beta_0 + \beta_1 \cdot x_i$

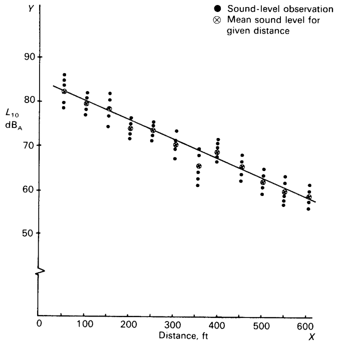

FIGURE 13-5. Graph for Example 13-3.

Fig 6.1

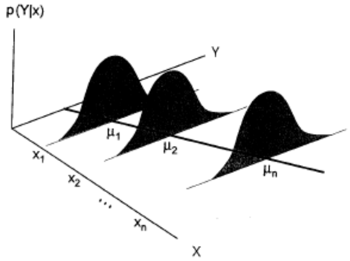

Figure 6.1. The assumptions of linearity, constant variance, and normality in simple regression. The figure shows the conditional population distributions of Y given X for several values of the independent variable, labeled $x_1, x_2, \ldots, x_n$. The conditional means of Y given X are denoted $\mu_1, \mu_2, \ldots, \mu_n$.

Review: Notational Considerations

- For the *i*-th observation the ***population model*** is $y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$.
  None of the parameters β_0, β_1 nor the error term ε_i are directly observable.
  The parameter β_0 is called the intercept and the parameter β_1 is the slope.
  The parameters β_0 and β_1 are not indexed by the observations *i*. Thus they are constant across all observations.
  The random error term ε_i is also called ***disturbance***. It is directly associated to each observation by the index *i.*

- For the ***sample estimates*** the predicted value of the model is $\hat{y}_i = b_0 + b_1 \cdot x_i$ with the ***residual*** $e_i = y_i - \hat{y}_i$.

- *Note* that some books use α instead of β_0 for the population intercept and a instead of b_0 for the estimated intercept (for instance, Burt and Barber).

- Note some people also write $\hat{\beta}_0$ and $\hat{\beta}_1$ for the estimated parameters and $\hat{\varepsilon}_i$ for the residuals.

- Bivariate regression is defined by two parameters *K* = 2: the intercept β_0 and slope β_1.

Review: Key Linear Regression Properties

- By design regression residuals are always linearly uncorrelated with the predicted value and the independent variables:
    1. $Cor(\mathbf{e}, \hat{\mathbf{y}}) = 0$, and
    2. $Cor(\mathbf{e}, \mathbf{x}_k) = 0$
- The mean of the regression residuals is always zero: $\frac{1}{n} \cdot \sum_{i=1}^{n} e_i = 0$
- The predicted regression line (bivariate case) or surface (multivariate case) always will go through the means of the independent and dependent variables.

Key Assumptions of Regression Analysis

1. Linear Relationship: the relationship between independent variable and the dependent variable is ***linear*** (or can be transformed to linearity)
2. Non-randomness of *X*: The independent variables *X* are in theory deterministic variables. *Note: should they for some reason be influenced by randomness, then we need at least assume that they are uncorrelated with the uncertainty of the endogenous variable Y.*
3. Disturbances: the disturbances have ***identical distributions***, with (a) zero mean and (b) constant variance, for every value of the independent variables *X*
4. Independence: Disturbances are in general assumed to be independent (uncorrelated) among each other.

5. Normality: The additional assumption of *normality* of the disturbances allows exact statistical significance testing in the estimated regression model.

- Note:
  - The independence and identical distribution assumption of the disturbances is abbreviated by ***i.i.d.*** (**i**ndependently **i**dentically **d**istributed)
  - Only the disturbances are required to be normal i.i.d. Neither *Y* nor *X* need to follow a normal distribution.
  - However, a ***joint normal distribution*** is highly desirable because it ensures ***linearity*** in the relationship between the dependent and the independent variables.

Review: Coefficient of Determination: R^2 and Adjusted R^2_{adj}

- The goodness of fit measure is defined as $R^2 \equiv \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$
- It measures what ***percentage of the total variation in the dependent variable is explained*** by the regression model.
- The adjusted goodness of fit takes the degrees of freedom into account because, the more variables we enter into the regression equation, the better the fit of the model will be (recall the perfect fit of the regression through two points)

$$R^2_{adj} \equiv 1 - \frac{RSS/(n-K)}{TSS/(n-1)}$$

When *n* is large relative to *K* then the difference between the adjusted and the ordinary R^2 becomes negligible.

- Remember: Burt, Barber and Rigby define the abbreviations for the explained sum of squares due to the regression model as RSS and the residual sum of squares differently as ESS.
  The standard definitions in the literature are reversed:
  - ***RSS*** are the residual sum of squares, and
  - ***ESS*** are the explained sum of squares.

Review: Root Mean Square Error (Standard Error of Estimate)

- The root mean square error measures the standard deviation of the residuals:

$$s_e = \sqrt{Var(e)} = \sqrt{\frac{RSS}{n-K}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-K}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)_i^2}{n-K}}$$

  (Note: Burt and Barber use the notation $s_{Y|X}$)

- In bivariate regression analysis we lose $K = 2$ degrees freedom because the regression line is constrained to go through the point $(\bar{x}, \bar{y})$.
  Alternative explanation: We need two points to define a line.

- In multiple regression analysis we would lose as many degrees of freedom as the regression model as parameters to estimate.

Unknown Regression Parameters

- The ***variations from sample to sample*** make the estimated regression parameters b_0 and b_1 ***random variables***, which exhibit a distribution with an associated ***standard error***.

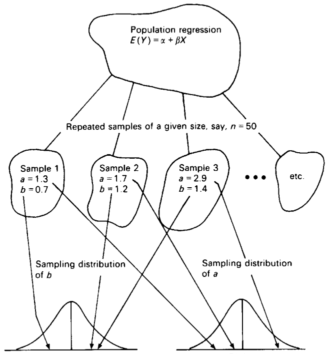

FIGURE 14-4. The sampling procedure in a regression model.

Standard errors of the regression estimates

- Assuming ***i.i.d.*** disturbances, the estimated regression parameters b_0 and b_1 will be approximately normal distributed with $b_0 \sim N(\beta_0, \sigma^2_{b_0})$ and $b_1 \sim N(\beta_1, \sigma^2_{b_1})$.
- The square roots of the estimated parameter variances $\sqrt{\hat{\sigma}^2_{b_0}} = SE_{b_0}$ and $\sqrt{\hat{\sigma}^2_{b_1}} = SE_{b_1}$ are called ***standard errors*** of the estimated regression coefficients.
- The standard error SE_{b_1} depends on the level of spread of the independent variable X around its mean, i.e., TSS_X.

  As TSS_X increases the estimated regression line becomes more precise (i.e., smaller standard errors).

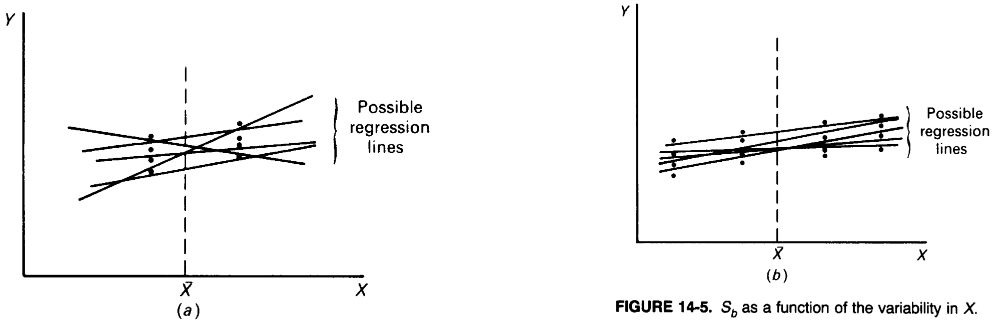

FIGURE 14-5. S_b as a function of the variability in X.

Testing Regression Coefficients and Confidence Intervals

- Usually we are not interested whether an intercept differs significantly from zero. That is, the hypothesis $H_0: b_0 = 0$ is irrelevant.
- Since we work with an estimate for standard error of b_k, the distribution of the statistics $t = \frac{b_1 - \beta_k}{SE_{b_k}}$ follows a *t-distribution* with $df = n - K$.

  This *t*-statistic resembles a *z*-transformed variable around the *hypothetical* $\beta_k = E(b_k)$.
- If an exogenous variable X_k does not explain any variation in the endogenous variable Y then the slope estimate b_k does not differ significantly from zero.
  - The zero hypothesis in this case is $H_0 : \beta_k = 0$ and the alternative hypothesis is $H_1 : \beta_k \neq 0$ Therefore, the test-statistic reduces to $t = \left(b_k - 0\right) / SE_{b_k} = b_k / SE_{b_k}$

- If a theory suggests that the regression parameter is expected to be positive (or negative, respectively), then a one-sided test becomes appropriate. That is, for example, $H_0 : \beta_k \leq 0$ and $H_1 : \beta_k > 0$.
- An alternative test statistic in ***bivariate*** regression analysis is based on the *F*-test
$$F = \frac{ESS / (K-1)}{RSS / (n-k)},$$
which follows under the zero hypothesis $H_0 : \beta_1 = 0$ a *F*-distribution with $df_1 = K - 1$ and $df_2 = n - K$.
In ***multiple*** regression with $K - 1$ independent variables the null hypothesis becomes $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{K-1} = 0$ against $H_1 : \beta_k \neq 0$ for at least one $\beta_k \in \{\beta_1, \beta_2, \ldots, \beta_{K-1}\}$.
- A $1-\alpha$ confidence interval, e.g., 95% interval, around the hypothetical value β_k is given by
$$\Pr(\underset{-}{t_{\alpha/2,df}} < \frac{b_k - \beta_k}{SE_{b_k}} < \underbrace{t_{1-\alpha/2,df}}_{+}) = 1 - \alpha$$
$$\Leftrightarrow \Pr(b_k \underbrace{-SE_{b_k} \cdot \underbrace{t_{1-\alpha/2,df}}_{+}}_{-} < \beta_k < b_k \underbrace{-SE_{b_k} \cdot \underset{-}{t_{\alpha/2,df}}}_{+}) = 1 - \alpha$$
$$\Leftrightarrow \Pr(b_k - SE_{b_k} \cdot t_{1-\alpha/2,df} < \beta_k < b_k + SE_{b_k} \cdot t_{1-\alpha/2,df}) = 1 - \alpha$$
<u>Consequently</u>, the null hypothesis $H_0 : \beta_k = 0$ cannot be rejected at the significance level α if the value zero for β_k is within the confidence interval.

Confidence and Predicting Intervals in Bivariate Regression

- The ***predicted regression line*** $\hat{Y} = b_0 + b_1 \cdot X$ and the *ith* ***predicted value*** $\hat{Y}_i = b_0 + b_1 \cdot X_i$ are random because they are based on the estimates of b_0 and b_1, which themselves are random variables.

- There are two confidence intervals for ***the regression model***:

  [a] One for the ***predicted regression line***.

  [b] One for an ***individual point prediction*** of $Y_{i,\text{point}}$.

  Both become wider (the uncertainty is increasing) as X_i deviates from its mean, i.e., $X_i - \bar{X}$

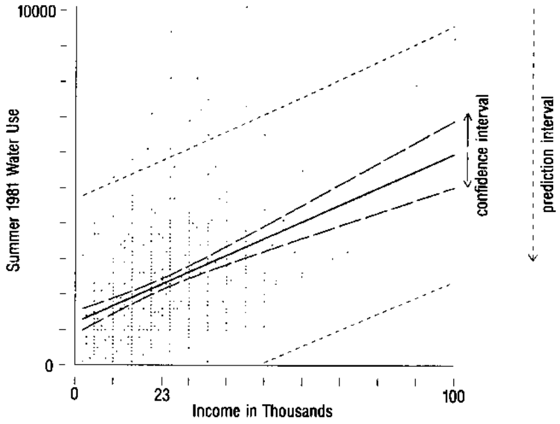

Figure 2.7 Confidence and prediction intervals around regression line.

- There is always higher uncertainty for in predicting an individual point than in predicting the regression line.

- The larger the sample size *n* gets the narrower the confidence intervals will become.

Distribution of the Residuals

- The standardized residuals $z(e_i) = \frac{e_i - E(e_i)}{s_e \cdot \sqrt{1-h_{ii}}}$ with $E(e_i) = 0$ are ***approximately standard normal distributed*** $z(e_i) \sim N(0,1)$. The term $\sqrt{1-h_{ii}}$ is an adjustment factor related to how far X_i is from its mean $\bar{X}$. Remember: Each regression line has the pivot $(\bar{X}, \bar{Y})$.
- The ***studentized residuals*** $z_{student}(e_i) = \frac{e_i - E(e_i)}{s_e[-i] \cdot \sqrt{1-h_{ii}}}$, where $s_e[-i]$ is the root means square error calculated excluding the *i*th observation, are exactly t-distributed with $df = n - K - 1$
- The studentized residuals are preferred to test whether an observed y_i deviates significantly from the regression line.

Statistical Inference on the Problems with Regression

- Problems:
  - ***Omitted relevant variable***. ⇨ The estimate regression parameters may become *biased*.
  - ***Non-linear relationship***. ⇨ Potentially misspecified model. Perhaps the residuals indicate autocorrelation or a variable should be added in its square form.
  - ***Heteroscedasticity***. ⇨ The variance of the regression residuals changes systematically with either an independent variable or some other external factor.

- ○ ***Autocorrelation***. ⇨ The disturbances are no longer independent between pairs of observations.
- ○ ***Non-normal disturbances***. ⇨ Test statistics, based on the normal assumption, become unreliable for small samples.
- ○ ***Influential cases***. ⇨ Regression analysis is not resistant to outliers. Large squared residuals $\hat{e}_i^2 = \left(y_i - \hat{y}_i\right)^2$ have a strong impact on the ordinary least squares estimation (a squared large distance becomes even larger).

- Some of problems can be identified by an inspection of the regression residuals. Recall, that the residuals $\hat{e}_i$ are ***uncorrelated*** with the predicted values $\hat{Y}_i = b_0 + b_1 \cdot X_i$. The same holds for the exogenous variable X_i and the regression residuals $\hat{e}_i$.
- Therefore, a ***scatterplot of the residuals*** against either the ***predicted value*** or an ***independent variable*** should <u>not</u> show a pattern.

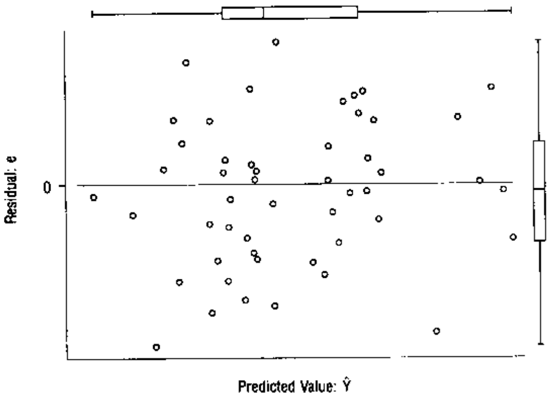

Figure 2.10 "All clear" e-versus-$\hat{Y}$ plot (artificial data).

- Some violations are indicated in the plots below:

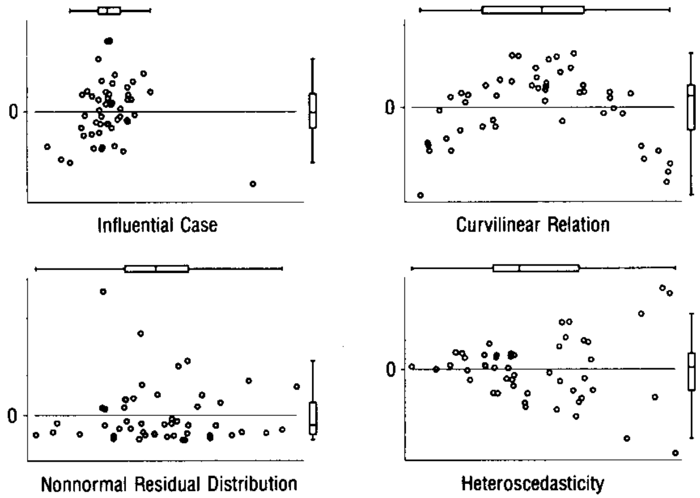

Figure 2.11 Examples of trouble seen in e-versus-$\hat{Y}$ plots (artificial data).