Econometrics 1

Lecture 2

Statistical inference for Econometrics

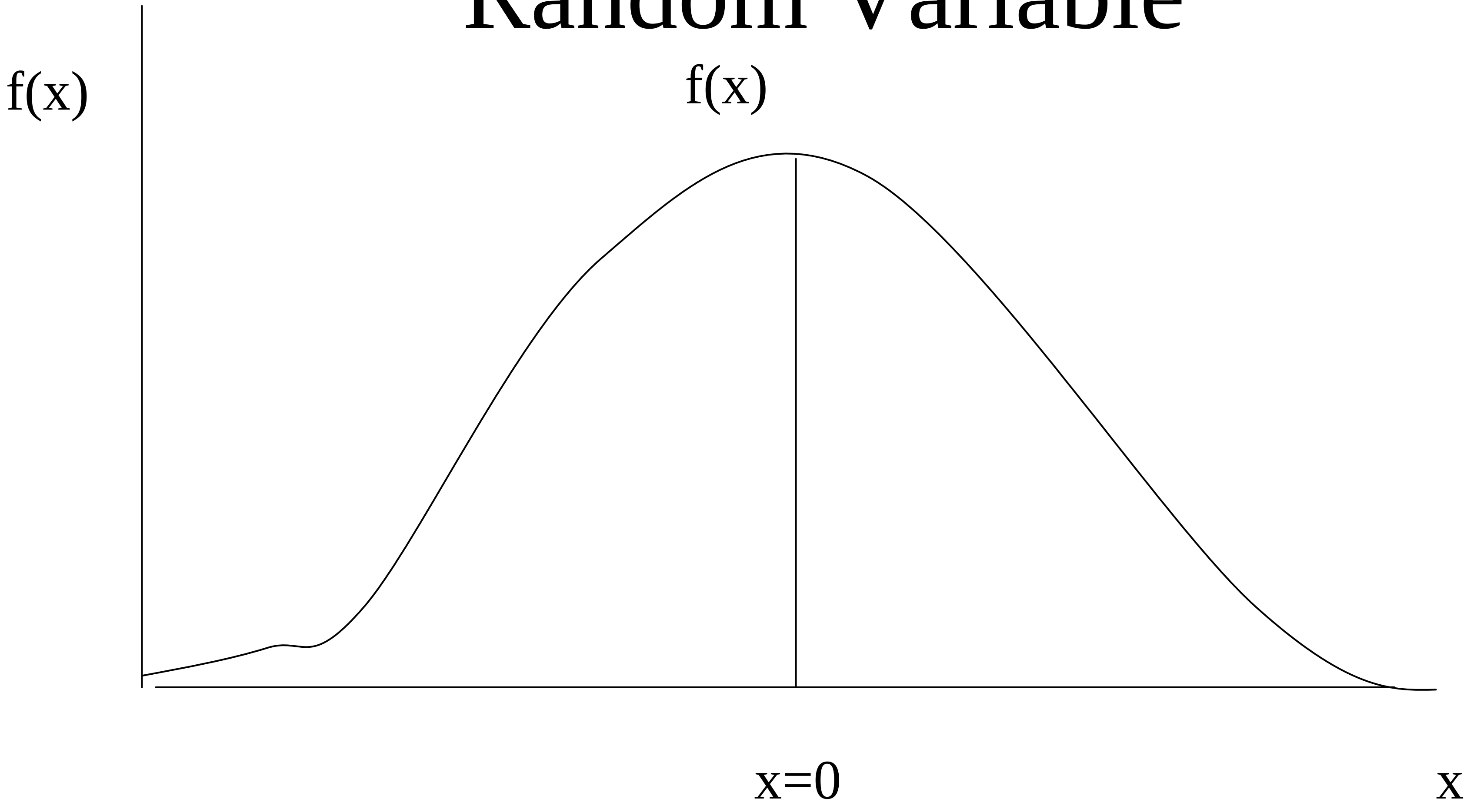
A brief review on statistical inference and random variables

Regression estimates are used to make statistical inferences about unknown values of random variables from the sample information.

Theory of statistical inference consists of methods on how to estimate parameters describing the distribution of the related random variable and to make a confidence interval of parameters and how to make decisions based on those estimates up to a certain level of confidence. These results are derived from statistical distributions.

In simple regression models we use normal distribution and various other distributions that derive from the normal distribution.

Probability Density Function of a Random Variable



Normal and Normal Related Distributions

- Normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})$$

- Log-normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}\frac{(\ln x-\mu)^2}{\sigma^2})$$

- χ^2 distribution

If $Z_1, Z_2, \ldots\ldots, Z_k$ are independent and normally distributed variables then

$Z = \sum_{i=1}^{k} Z_i^2$ is said to have a χ^2 distribution with k degrees of freedom.

- Students t-distribution

If Z_1 is normally distributed random variables and Z_2 has a χ^2 distribution with k dfs, then the following ratio is said to have a t-distribution :

$$t = \frac{Z_1}{\sqrt{Z_2 / k}}$$

- if Z_1 and Z_2 follow a χ^2 distribution with k1 and k2 dfs, then

$$F = \frac{\sqrt{Z_2 / k1}}{\sqrt{Z_2 / k2}}$$

A sample of the Shazam Program to Generate Normal or Uniform Distributions

```
Sample 1 100
genr x =nor(1)
genr y=20+nor(2)
genr T=time(0)
*dummy variable is positive when x is positive
genr D1=dum(x)
genr D2=dum(time(0)-10)
genr D3 =dum(x-y-1)
genr chi3 = nor(1)**2 +nor(1)**2 +nor(1)**2
genr chi5 = nor(1)**2 +nor(1)**2 +nor(1)**2+
nor(1)**2+nor(1)**2
genr f35 = (chi3/3)/(chi5/5)
print chi3 chi5 f35
plot x y
stat x y

OLS x y /rstat
if($DW .lt. 10)
auto x y

*sample 1 100
plot x y t
print d1 d2 d3
stop
gen1 lower =0
gen1 upper =1
integ x lower upper z= exp(-0.5*x**2)/sqrt(2*$pi)
print z
stop
```

Use of these distributions in hypothesis testing

Distribution of the dependent variable Y	Then the distribution of estimated parameter
$Y_t \sim N\left[\left(\beta_1 + \beta_1 X_1 + \beta_2 X_2\right), \sigma^2\right]$	$\hat{\beta}_k \sim N\left[\beta_k, \text{var}\left(\hat{\beta}_k\right)\right]$
Assuming that the error term is normally distributed $e_t \sim N\left(0, \sigma^2\right)$**.**	$z = \dfrac{\hat{\beta}_k - \beta}{\sqrt{\text{var}\left(\hat{\beta}_k\right)}} \sim N(0,1)$ $t = \dfrac{\hat{\beta}_k - \beta}{\sqrt{\text{var}\left(\hat{\beta}_k\right)}} \sim t_{(T-K)}$

Interval Estimation and the level of significance

$P\left(t_c \leq \frac{\hat{\beta}_k - \beta}{\sqrt{\text{var}\left(\hat{\beta}_k\right)}} \leq t_c \right) = 1 - \alpha$ $P\left(t \geq t_c\right) = \alpha/2$	$P\left(\hat{\beta}_k - t_c SE\left(\hat{\beta}_k\right) \leq \beta_k \leq \hat{\beta}_k + t_c SE\left(\hat{\beta}_k\right)\right) = 1 - \alpha$ $P\left[\hat{\beta}_k - t_c SE\left(\hat{\beta}_k\right), \beta_k + t_c SE\left(\hat{\beta}_k\right)\right]$

Measuring Goodness of Fit

$$Var(y_i) = \sum_i [y_i - \bar{y}]^2 = \sum_i [(\hat{y}_i - \bar{y}) + \hat{e}_i]^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i \hat{e}_i^2$$

$$+ 2\sum_i (\hat{y}_i - \bar{y})e_i$$

$$Var(y_i) = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i \hat{e}_i^2$$ For T observations and K explanatory variables

[Total variation] = [Explained variation] + [Residual variation]
df = T-1 K-1 T-K

$$1 = \frac{\sum (y_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} + \frac{\sum \hat{e}_i^2}{\sum (y_i - \bar{y})^2} = R^2 + (1 - R^2)$$

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}; \quad 0 \le R^2 \le 1$$

If the model does not contain an intercept parameter, then the measure R^2 given in (7.6.1) is no longer appropriate. The reason it is no longer appropriate is that, without an intercept term in the model,

$$Var(y_i) \ne \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i \hat{e}_i^2$$

The *F*-Distribution: Theory

If variables 1 and 2 are chi-square distributed ($\chi^2_{m_1}$ and $\chi^2_{m_2}$) their ratio has a F-distribution with m1 and m2 degrees of freedom.

If $V_1 \sim \chi^2_{m_1}$ and $V_2 \sim \chi^2_{m_2}$ and V_1 and V_2 are independent then

$$F = \frac{V_1 / m_1}{V_2 / m_2} \sim F_{(m_1, m_2)}$$

Hypothesis Testing

One-tailed hypothesis test	Two-tailed hypothesis test
$H_0 : \beta_k = 0$	$H_0 : \beta_k = 0$
$H_A : \beta_k \geq 0$	$H_A : \beta_k \neq 0$
Get the value of $\hat{\beta}_k$	Get the value of $\hat{\beta}_k$
Get the standard error of $\hat{\beta}_k$	Get the standard error of $\hat{\beta}_k$
Compute t-ratio	Compute t-ratio
$t = \frac{\hat{\beta}_k - \beta}{SE(\hat{\beta}_k)} \sim t_{(T-K)}$	$t = \frac{\hat{\beta}_k - \beta}{SE(\hat{\beta}_k)} \sim t_{(T-K)}$
Compare it with the critical value of t from the t-distribution table.	Compare it with the critical value of t from the t-distribution table.
• reject H_0 if $t \leq t_c$	• reject H_0 if the computed t-value is greater than or equal to t_c, or less than or equal to $-t_c$.

Type	Statistical Model	Slope	Elasticity
1. Linear	$y_t = \beta_1 + \beta_2 x_t + e_t$	β_2	$\beta_2 \frac{x_t}{y_t}$
2. Reciprocal	$y_t = \beta_1 + \beta_2 \frac{1}{x_t} + e_t$	$-\beta_2 \frac{1}{x_t^2}$	$-\beta_2 \frac{1}{x_t y_t}$
3. Log-Log	$\ln(y_t) = \beta_1 + \beta_2 \ln(x_t) + e_t$	$\beta_2 \frac{y_t}{x_t}$	β_2
4. Log-Linear (Exponential)	$\ln(y_t) = \beta_1 + \beta_2 x_t + e_t$	$\beta_2 y_t$	$\beta_2 x_t$
5. Linear-Log (Semi-log)	$y_t = \beta_1 + \beta_2 \ln(x_t) + e_t$	$\beta_2 \frac{1}{x_t}$	$\beta_2 \frac{1}{y_t}$
6. Log-inverse	$\ln(y_t) = \beta_1 - \beta_2 \frac{1}{x_t} + e_t$	$\beta_2 \frac{y_t}{x_t^2}$	$\beta_2 \frac{1}{x_t}$

Griffith, Hill, Judge 1993, p. 260.