UNIVERSITÀ
DEGLI STUDI DELLA
TUSCIA

SAPIENZA
UNIVERSITÀ DI ROMA

Statistics for Business and decision making

Dr. Ilaria Benedetti

12. Multiple regression model

Outline

- *Introduction*
- *Estimating the Parameters of the Multiple Regression Model*
- *Sampling Properties of the Least Squares Estimators*
- *Interval Estimation*
- *Hypothesis Testing*
- *Measuring Goodness-of-fit*
- *Indicator variables*
- *Omitted variables and Collinearity*
- *Heteroskedasticity*

Introduction

In most economic models there are two or more explanatory variables

- *When we turn an economic model with more than one explanatory variable into its corresponding econometric model, we refer to it as a* ***multiple regression model***
- *Most of the results we developed for the simple regression model can be extended naturally to this general case*

REAL ECONOMIC DATA: we begin with a model used to explain sales revenue for a fast-food hamburger chain with outlets in small U.S. cities.

We call Big Andy's to disguise the real identity

- *Important decisions made by the management of Big Andy's include its pricing policy for different products and how much to spend on advertising.*
- *To assess the effect of different price structures and different levels of advertising expenditure, Big Andy's sets different prices, and spends varying amounts on advertising, in different cities*
- *Let's set up an economic model in which sales revenue depends on one or more explanatory variables*

Real Economic data

- 1) We initially hypothesize that sales revenue is linearly related to price and advertising expenditure

- The economic model is:

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT$$

- where SALES represents monthly sales revenue in a given city, PRICE represents price in that city.
- ADVERT is monthly advertising expenditure in that city.
- Both SALES and ADVERT are measured in terms of thousands of dollars. Because sales in bigger cities will tend to be greater than sales in smaller cities, we focus on smaller cities with comparable populations.
- Since a hamburger outlet sells a number of products—burgers, fries, and shakes—and each product has its own price, management has constructed a single price index PRICE, measured in dollars and cents, that describes overall prices in each city.

Interpreting the coefficients

- β_2 is the change in monthly sales *SALES* ($1000) when the price index *PRICE* is increased by one unit ($1), and advertising expenditure *ADVERT* is held constant

$$\beta_2 = \frac{\Delta SALES}{\Delta PRICE\ (ADVERT \text{ held constant})}$$

$$= \frac{\partial SALES}{\partial PRICE}$$

- The sign of β_2 could be positive or negative. If an increase in price leads to an increase in sales revenue, then $\beta_2 > 0$, and the demand for the chain's products is price-inelastic.
- Conversely, a price-elastic demand exists if an increase in price leads to a decline in revenue, in which case $\beta_2 < 0$. Thus, knowledge of the sign of β_2 provides information on the price elasticity of demand. The magnitude of β_2 measures the amount of change in revenue for a given price change.

Interpreting the coefficients

- Similarly, β_3 is the change in monthly sales *SALES* ($1000) when the advertising expenditure is increased by one unit ($1000), and the price index *PRICE* **is held constant**

$$\beta_3 = \frac{\Delta SALES}{\Delta ADVERT \ (PRICE \text{ held constant})}$$

$$= \frac{\partial SALES}{\partial ADVERT}$$

- The parameter β_3 describes the response of sales revenue to a change in the level of advertising expenditure. We expect the sign of β_3 to be positive.
- Whether or not the increase in revenue is sufficient to justify the added advertising expenditure, as well as the added cost of producing more hamburgers, is another question.
- With $\beta_3 < 1$, an increase of $1,000 in advertising expenditure will yield an increase in revenue that is less than $1,000. For $\beta_3 > 1$, it will be greater. Thus, in terms of the chain's advertising policy, knowledge of β_3 is very important.

Data description

City	*SALES* $1,000 units	*PRICE* $1 units	*ADVERT* $1,000 units
1	73.2	5.69	1.3
2	71.8	6.49	2.9
3	62.4	5.63	0.8
4	67.4	6.22	0.7
5	89.3	5.02	1.5
.	.	.	.
.	.	.	.
.	.	.	.
73	75.4	5.71	0.7
74	81.3	5.45	2.0
75	75.0	6.05	2.2
	Summary statistics		
Sample mean	77.37	5.69	1.84
Median	76.50	5.69	1.80
Maximum	91.20	6.49	3.10
Minimum	62.40	4.83	0.50
Std. Dev.	6.49	0.52	0.83

Data description with R

```
# import andy data

library(readxl)
andy <- read_excel("andy.xlsx")
```

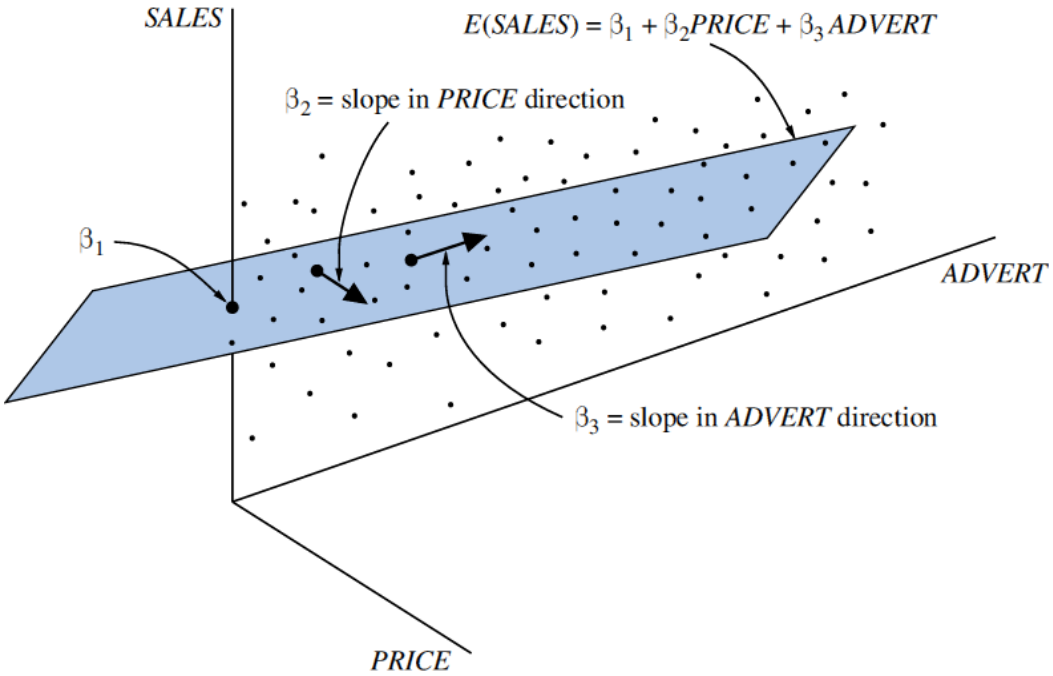
```
> summary(andy)
     sales            price           advert
 Min.   :62.40   Min.   :4.830   Min.   :0.500
 1st Qu.:73.20   1st Qu.:5.220   1st Qu.:1.100
 Median :76.50   Median :5.690   Median :1.800
 Mean   :77.37   Mean   :5.687   Mean   :1.844
 3rd Qu.:82.20   3rd Qu.:6.210   3rd Qu.:2.700
 Max.   :91.20   Max.   :6.490   Max.   :3.100
```

The econometric model is:

$$SALES = E(SALES) + e = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + e$$

- To allow for a difference between observable sales revenue and the expected value of sales revenue, we add a random error term, e = *SALES - E(SALES)*

$$E(SALES) = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT$$



Multiple Regression model

- In a general multiple regression model, a dependent variable *y* is related to a number of explanatory variables x_2, x_3, …, x_K through a linear equation that can be written as:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \mathrm{K} + \beta_K x_K + e$$

- A single parameter, call it β_k, measures the effect of a change in the variable x_k upon the expected value of *y*, all other variables held constant

$$\beta_k = \frac{\Delta E(y)}{\Delta x_k}\bigg|_{\text{other xs held constant}} = \frac{\partial E(y)}{\partial x_k}$$

- The parameter β1 is the intercept term;
- The equation for sales revenue can be viewed as a special case of the above equation where *K* = 3, *y* = *SALES*, $x_1 = 1$, x_2 = *PRICE* and x_3 = *ADVERT*

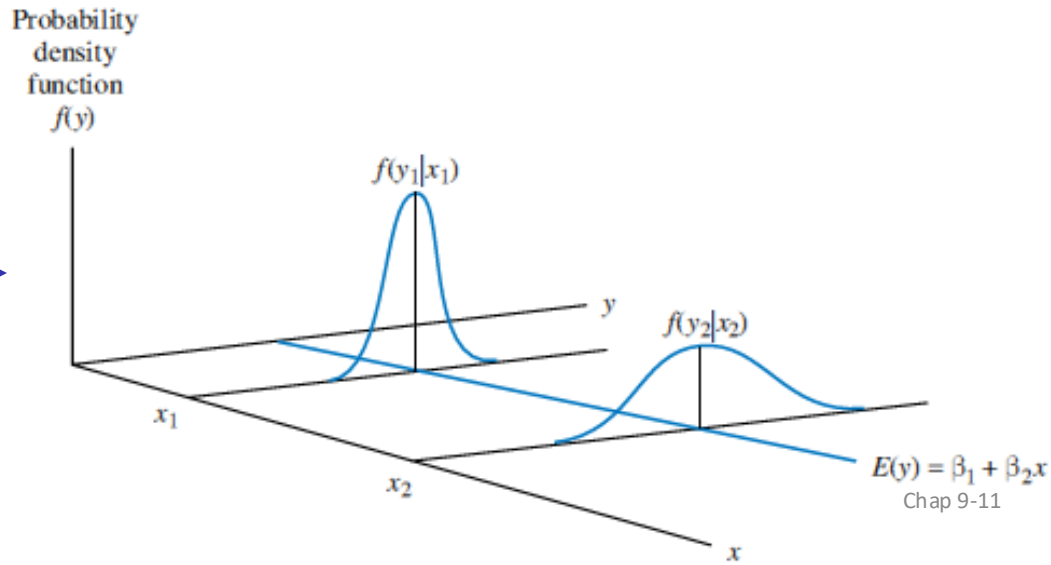
$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

Assumptions

- To make the econometric model complete, assumptions about the probability distribution of the random errors need to be made
- We make the following assumptions:
  - $E(e) = 0$
  - $var(e) = \sigma^2$

    Errors with this property are said to be ***homoskedastic***
  - $cov(e_i, e_j) = 0$
  - $e \sim N(0, \sigma^2)$

Heterosckedastic errors →



Assumptions

The statistical properties of *y* follow from those of *e*:

$E(y) = \beta_1 + \beta_2 x_2 + \beta_3 x_3$

$\text{var}(y) = \text{var}(e) = \sigma^2$

$\text{cov}(y_i, y_j) = \text{cov}(e_i, e_j) = 0$

$y \sim N[(\beta_1 + \beta_2 x_2 + \beta_3 x_3), \sigma^2]$

We make two assumptions about the explanatory variables:

1. The explanatory variables are *not random variables*. We are assuming that the values of the explanatory variables are known to us prior to our observing the values of the dependent variable.
2. *Any one of the explanatory variables is not an exact linear function of the others* .This assumption is equivalent to assuming that no variable is redundant. If this assumption is violated – a condition called **exact collinearity** - the least squares procedure fails

Assumptions

MR1. $y_i = \beta_1 + \beta_2 x_{i2} + \mathrm{L} \ + \beta_K x_{iK} + e_i, \ i = 1, \mathrm{K}, N$

MR2. $E(y_i) = \beta_1 + \beta_2 x_{i2} + \mathrm{L} \ + \beta_K x_{iK} \Leftrightarrow E(e_i) = 0$

MR3. $\mathrm{var}(y_i) = \mathrm{var}(e_i) = \sigma^2$

MR4. $\mathrm{cov}(y_i, y_j) = \mathrm{cov}(e_i, e_j) = 0$

MR5. The values of each x_{tk} are not random and are not exact linear functions of the other explanatory variables

MR6. $y_i \sim N\left[(\beta_1 + \beta_2 x_{i2} + \mathrm{L} \ + \beta_K x_{iK}), \sigma^2\right] \Leftrightarrow e_i \sim N(0, \sigma^2)$

Estimating the Parameters of the Multiple Regression Model

- We will discuss estimation in the context of the model including 2 explanatory variables, which we repeat here for convenience, with *i* denoting the *i*th observation

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

➢ This model is simpler than the full model, yet all the results we present carry over to the general case with only minor modifications

Estimating the Parameters of the Multiple Regression Model

- Mathematically we minimize the sum of squares function $S(\beta_1, \beta_2, \beta_3)$, which is a function of the unknown parameters, given the data:

$$S(\beta_1,\beta_2,\beta_3) = \sum_{i=1}^{N}\left(y_i - E(y_i)\right)^2$$

$$= \sum_{i=1}^{N}\left(y_i - \beta_1 - \beta_2 x_{i2} - \beta_3 x_{i3}\right)^2$$

Formulas for b_1, b_2, and b_3, obtained by minimizing the above equation, are estimation procedures, which are called the **least squares estimators** of the unknown parameters

- **In general, since their values are not known until the data are observed and the estimates calculated, the least squares estimators are random variables**

Estimating the Parameters of the Multiple Regression Model
Example

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	118.9136	6.3516	18.7217	0.0000
PRICE	−7.9079	1.0960	−7.2152	0.0000
ADVERT	1.8626	0.6832	2.7263	0.0080

$R^2 = 0.4483$ $SSE = 1718.943$ $\hat{\sigma} = 4.8861$ $s_y = 6.48854.$

Estimates along with their standard errors and the equation's R^2 are typically reported in equation format as:

$$\widehat{SALES} = 118.91 - 7.908 PRICE + 1863 ADVERT \qquad R^2 = 0.448$$
$$(se) \qquad (6.35) \quad (1.096) \qquad (0.683)$$

Estimating the Parameters of the Multiple Regression Model - Example with R

```
> mod1 <- lm(sales~price+advert, data=andy)
> print(mod1) #only coefficients and intercepts

Call:
lm(formula = sales ~ price + advert, data = andy)

Coefficients:
(Intercept)        price       advert
    118.914       -7.908        1.863
```

$$\widehat{SALES} = 118.91 - 7.908 PRICE + 1863 ADVERT$$

```
> summary(mod1) #all the table

Call:
lm(formula = sales ~ price + advert, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max
-13.4825  -3.1434  -0.3456   2.8754  11.3049

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 118.9136     6.3516  18.722  < 2e-16 ***
price        -7.9079     1.0960  -7.215 4.42e-10 ***
advert        1.8626     0.6832   2.726  0.00804 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.886 on 72 degrees of freedom
Multiple R-squared:  0.4483,    Adjusted R-squared:  0.4329
F-statistic: 29.25 on 2 and 72 DF,  p-value: 5.041e-10
```

Estimating the Parameters of the Multiple Regression Model – output example with excel

Regression						
R multiplo	0,670					
R2	0,448					
R2 adjsted	0,433					
Standard error of the regression	4,886		estimates of σ^2	23,874		
obs	75,000					
	Coefficienti	*Errore standard*	*Stat t*	*p-value*	*L1*	*L2*
constant	118,9136	6,3516	18,7217	0,0000	106,2519	131,5754
price	-7,9079	1,0960	-7,2152	0,0000	-10,0927	-5,7230
advert	1,8626	0,6832	2,7263	0,0080	0,5007	3,2245
ANOVA						
	gdl	*SQ*	*MQ*	*F*	*p-value*	
regression	2,00	1396,539	698,269	29,248	0,000	
error	72,00	1718,943	23,874			
total	74,00	3115,482				

Estimating the Parameters of the Multiple Regression Model

Interpretations of the results:

1. The negative coefficient on *PRICE* suggests that demand is price elastic; we estimate that, with advertising held constant, an increase in price of $1 will lead to a fall in monthly revenue of $7,908

2. The coefficient on advertising is positive; we estimate that with price held constant, an increase in advertising expenditure of $1,000 will lead to an increase in sales revenue of $1,863

3. The estimated intercept implies that if both price and advertising expenditure were zero the sales revenue would be $118,914.
   - Clearly, this outcome is not possible; a zero price implies zero sales revenue.
   - In this model, as in many others, it is important to recognize that the model is an approximation to reality in the region for which we have data.
   - Including an intercept improves this approximation even when it is not directly interpretable

Estimating the Parameters of the Multiple Regression Model

ESTIMATION OF THE ERROR VARIANCE σ^2

- We need to estimate the **error variance**, σ^2
  - We can think of σ^2 as the expectation or population mean of the squared errors e^2_i

$$\sigma^2 = \text{var}(e_i) = E\left(e_i^2\right)$$

  - But, the squared errors are unobservable, so we develop an estimator for σ^2 based on the squares of the least squares residuals:

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \left(b_1 + b_2 x_{i2} + b_3 x_{i3}\right)$$

Estimating the Parameters of the Multiple Regression Model

- An estimator for σ^2 that uses the information from

  and has good statistical properties is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{e}_i^2}{N - K}$$

where *K* is the number of β parameters being estimated in the multiple regression model.

➢For the hamburger chain example:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{75} \hat{e}_i^2}{N - K} = \frac{1718.943}{75 - 3} = 23.874$$

Estimating the Parameters of the Multiple Regression Model

- Note that the **sum of squared errors** is:

$$SSE = \sum_{i=1}^{N} \hat{e}_i^2 = 1718.943$$

- Also, note that the standard error of the regression or the root mse

$$\hat{\sigma} = \sqrt{23.874} = 4.8861$$

❖Both quantities typically appear in the output from your computer software

- Different software refer to it in different ways.

Estimating the Parameters of the Multiple Regression Model

SSE and variance of the error term with R

```
> andy$res<-(mod1$residuals)
> head(andy) #visualize the first six units in the dataset
# A tibble: 6 x 4
  sales price advert     res
  <dbl> <dbl>  <dbl>   <dbl>
1  73.2  5.69    1.3  -3.14
2  71.8  6.49    2.9  -1.19
3  62.4  5.63    0.8 -13.5
4  67.4  6.22    0.7  -3.63
5  89.3  5.02    1.5   7.29
6  70.3  6.41    1.3  -0.346
> SSE<-sum(andy$res^2)
> SSE
[1] 1718.943
```

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{75} \hat{e}_i^2}{N-K} = \frac{1718.943}{75-3} = 23.874$$

Estimating the Parameters of the Multiple Regression Model

SSE and variance of the error term with R

```
> # variance of the error term
> df=mod1$df.residual
> df
[1] 72
> var_e=SSE/df
> var_e
[1] 23.87421
> se_e=sqrt(var_e)
> se_e
[1] 4.886124
```

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{75} \hat{e}_i^2}{N-K} = \frac{1718.943}{75-3} = 23.874$$

Sampling Properties of the Least Squares Estimators

In a general context, the least squares estimators (b_1, b_2, b_3) are random variables;

- they take on different values in different samples, and their values are unknown until a sample is collected and their values computed.

- The sampling properties of a least squares estimator tell us how the estimates vary from sample to sample.

- They provide a basis for assessing the reliability of the estimates.

Sampling Properties of the Least Squares Estimators

THE GAUSS–MARKOV THEOREM:

For the multiple regression model, if assumptions MR1–MR5 hold, then the least squares estimators are the **best linear unbiased estimators (*BLUE*)** of the parameters.

- If the errors are not normally distributed, then the least squares estimators are approximately normally distributed in large samples
  - What constitutes ‘‘large’’ is tricky
  - It depends on a number of factors specific to each application
  - Frequently, ***N – K* = 50 will be large enough**

Sampling Properties of the Least Squares Estimators

THE DISTRIBUTION OF THE LEAST SQUARES ESTIMATORS

Consider the general for of a multiple regression model:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \mathrm{L} + \beta_K x_{iK} + e_i$$

If we add assumption MR6, that the random errors e_i have **normal probability distributions**, then the dependent variable y_i is normally distributed:

$$y_i \sim N\left[(\beta_1 + \beta_2 x_{i2} + \mathrm{L} + \beta_K x_{iK}), \sigma^2\right] \Leftrightarrow e_i \sim N(0, \sigma^2)$$

Since the least squares estimators are linear functions of dependent variables, it follows that the least squares estimators are also normally distributed:

$$b_k \sim N\left[\beta_k, \mathrm{var}(b_k)\right]$$

Sampling Properties of the Least Squares Estimators

THE DISTRIBUTION OF THE LEAST SQUARES ESTIMATORS

- We can now form the standard normal variable *Z*:

$$z = \frac{b_k - \beta_k}{\sqrt{\text{var}(b_k)}} \sim N(0,1), \text{ for } k = 1, 2, \ldots, K$$

- Replacing the variance of b_k with its estimate:

$$t = \frac{b_k - \beta_k}{\sqrt{\widehat{var}(b_k)}} = \frac{b_k - \beta_k}{\text{se}(b_k)} \sim t_{(N-K)}$$

- Notice that the number of degrees of freedom for *t*-statistics is *N* - *K*

Interval Estimations

- We write the general expression for a 100(1-α)% confidence interval as:

$$\left(b_k - t_{(1-\alpha/2,\, N-K)} \times \text{se}(b_k),\ b_k + t_{(1-\alpha/2,\, N-K)} \times \text{se}(b_k)\right)$$

- For the hamburger example, we need:

$$P(-t_c < t_{(72)} < t_c) = .95$$

- Using $t_c = 1.993$, we can rewrite it as:

$$P\left(-1.993 \leq \frac{b_2 - \beta_2}{\text{se}(b_2)} \leq 1.993\right) = .95$$

- Rearranging, we get:

$$P\left[b_2 - 1.993 \times \text{se}(b_2) \leq \beta_2 \leq b_2 + 1.993 \times \text{se}(b_2)\right] = .95$$

$$\left[b_2 - 1.993 \times \text{se}(b_2),\ b_2 + 1.993 \times \text{se}(b_2)\right]$$

Interval Estimations
Example

- Using our data, we have b_2 = -7.908 and $se(b_2)$ = 1.096, so that:

$$(-7.9079 - 1.993 \times 1.096, -7.9079 + 1.993 \times 1.096) = (10.093, -5.723)$$

This interval estimate suggests that decreasing price by $1 will lead to an increase in revenue somewhere between $5,723 and $10,093.

  - In terms of a price change whose magnitude is more realistic, a 10-cent price reduction will lead to a revenue increase between $572 and $1,009

- For advertising, we get:

$$(1.8626 - 1.9935 \times 0.6832, 1.8626 + 1.9935 \times 0.6832) = (0.501, 3.225)$$

We estimate that an increase in advertising expenditure of $1,000 leads to an increase in sales revenue of between $501 and $3,225

This interval is a relatively wide one; it implies that extra advertising expenditure could be unprofitable (the revenue increase is less than $1,000) or could lead to a revenue increase more than three times the cost of the advertising

Interval Estimations - Example with R

```
> summary(mod1)

Call:
lm(formula = sales ~ price + advert, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max
-13.4825  -3.1434  -0.3456   2.8754  11.3049

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 118.9136     6.3516  18.722 < 0.0000000000000002 ***
price        -7.9079     1.0960  -7.215      0.000000000442 ***
advert        1.8626     0.6832   2.726             0.00804 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.886 on 72 degrees of freedom
Multiple R-squared:  0.4483,    Adjusted R-squared:  0.4329
F-statistic: 29.25 on 2 and 72 DF,  p-value: 0.0000000005041
```

$$\left[b_2 - 1.993 \times \mathrm{se}(b_2),\ b_2 + 1.993 \times \mathrm{se}(b_2)\right]$$

```
> confint(mod1)
                2.5 %     97.5 %
(Intercept) 106.251852 131.575368
price       -10.092676  -5.723032
advert        0.500659   3.224510
```

Confidence interval interpretation??

Hypothesis Testing

- We need to ask whether the data provide any evidence to suggest that y is related to each of the explanatory variables
  - If a given explanatory variable, say x_k, has no bearing on y, then $\beta_k = 0$
  - Testing this null hypothesis is sometimes called a **test of significance** for the explanatory variable x_k

Null Hypotesis

$$H_0 : \beta_k = 0$$

Alternative Hypotesis

$$H_1 : \beta_k \neq 0$$

Statistics test

$$t = \frac{b_k}{\text{se}(b_k)} \sim t_{(N-K)}$$

$$t_c = t_{(1-\alpha/2,\, N-K)} \text{ and } -t_c = t_{(\alpha/2,\, N-K)}$$

Hypothesis Testing example

- For our hamburger example, we can conduct a test that sales revenue is related to price:

1. The null and alternative hypotheses are:

$$H_0 : \beta_2 = 0 \text{ and } H_1 : \beta_2 \neq 0$$

2. The test statistic, if the null hypothesis is true, is:

$$t = b_2 / \text{se}(b_2) \sim t_{(N-K)}$$

3. Using a 5% significance level (α=.05), and 72 degrees of freedom, the critical values that lead to a probability of 0.025 in each tail of the distribution are:

$$t_{(.975,72)} = 1.993 \text{ and } t_{(.025,72)} = -1.993$$

```
> summary(mod1)

Call:
lm(formula = sales ~ price + advert, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max
-13.4825  -3.1434  -0.3456   2.8754  11.3049

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 118.9136     6.3516  18.722 < 0.0000000000000002 ***
price        -7.9079     1.0960  -7.215        0.000000000442 ***
advert        1.8626     0.6832   2.726               0.00804 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.886 on 72 degrees of freedom
Multiple R-squared:  0.4483,    Adjusted R-squared:  0.4329
F-statistic: 29.25 on 2 and 72 DF,  p-value: 0.0000000005041

> confint(mod1)
                 2.5 %     97.5 %
(Intercept) 106.251852 131.575368
price       -10.092676  -5.723032
advert        0.500659   3.224510
```

Hypothesis Testing example

4. The computed value of the *t*-statistic is:

$$t = \frac{-7.908}{1.096} = -7.215$$

and the *p*-value from software is:

$$P\left(t_{(72)} > 7.215\right) + P\left(t_{(72)} < -7.215\right) = 2 \times (2.2 \times 10^{-10}) = 0.000$$

5. Since -7.215 < -1.993, we reject H0: β2 = 0 and conclude that there is evidence from the data to suggest sales revenue depends on price
   - Using the p-value to perform the test, we reject H0 because 0.000 < 0.05 .

Hypothesis Testing example

- Similarly, we can conduct a test that sales revenue is related to advertising expenditure:
  1. The null and alternative hypotheses are:

$$H_0 : \beta_3 = 0 \text{ and } H_1 : \beta_3 \neq 0$$

  2. The test statistic, if the null hypothesis is true, is:

$$t = b_3 / \text{se}(b_3) \sim t_{(N-K)}$$

  3. Using a 5% significance level (α=.05), and 72 degrees of freedom, the critical values that lead to a probability of 0.025 in each tail of the distribution are:

$$t_{(.975,72)} = 1.993 \text{ and } t_{(.025,72)} = -1.993$$

Hypothesis Testing example

4. The computed value of the *t*-statistic is:

$$t = \frac{1.8626}{0.6832} = 2.726$$

and the *p*-value from software is:

$$P\left(t_{(72)} > 2.726\right) + P\left(t_{(72)} < -2.726\right) = 2 \times 0.004 = 0.008$$

5. Since 2.726 > 1.993, we reject H_0: $\beta_3 = 0$: the data support the conjecture that revenue is related to advertising expenditure
   - Using the *p*-value to perform the test, we reject H_0 because 0.008 < 0.05 .

Sampling Properties of the Least Squares Estimators

THE DISTRIBUTION OF THE LEAST SQUARES ESTIMATORS

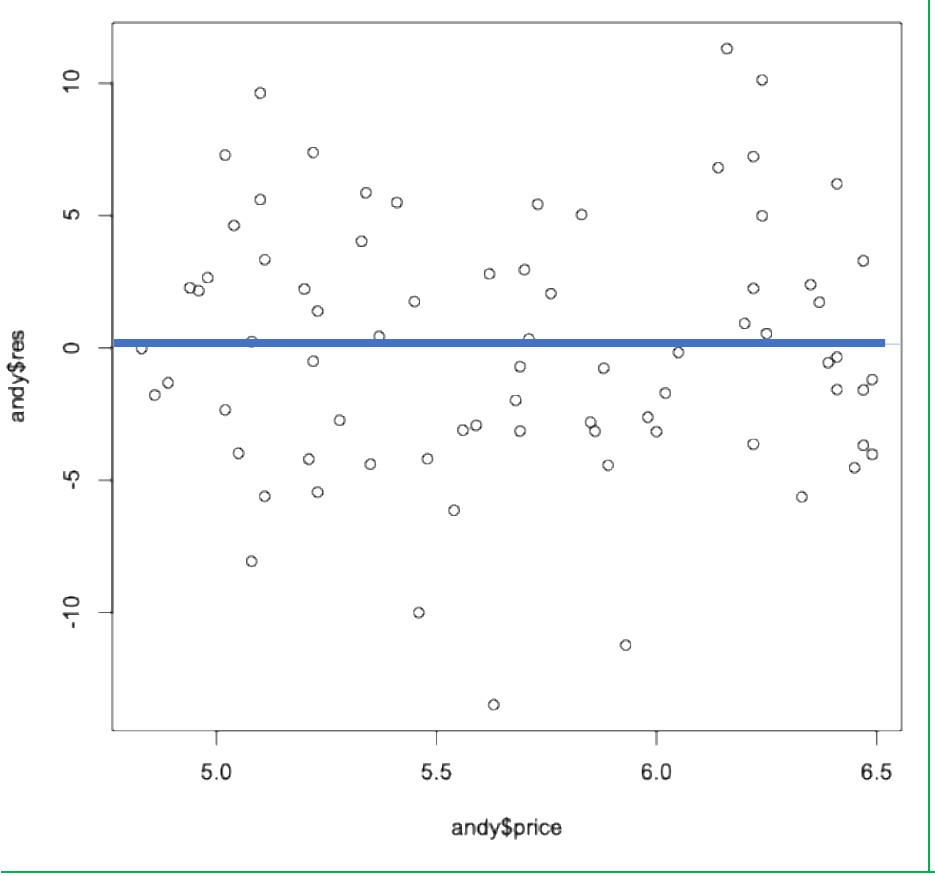
- What happens if the errors are not normally distributed?
  - Then the least squares estimator will not be normally distributed
  - They will, however, be approximately true in large samples
- We can check the distribution of the residuals using:
  - **A histogram**
  - **Formal statistical test**
    - A good one is the **Jarque–Bera test** for normality

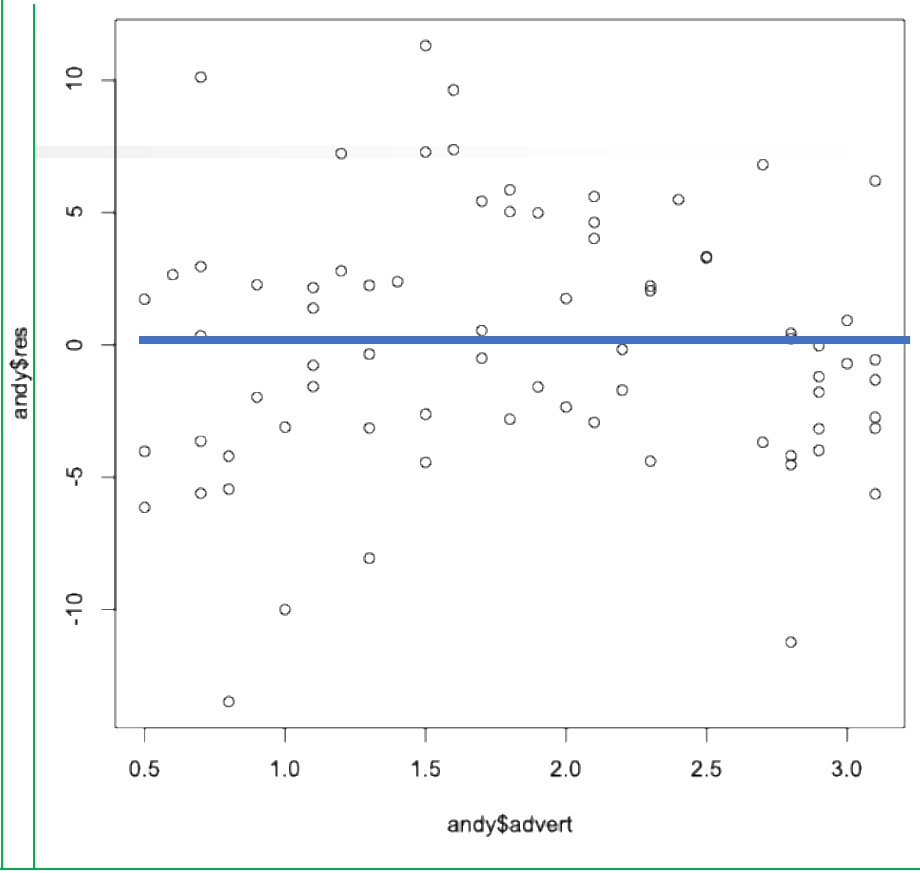
$$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

N = sample size
S = skewness
K = kurtosis

When the residuals are normally distributed, the Jarque–Bera statistic has a chi-squared distribution with two degrees of freedom

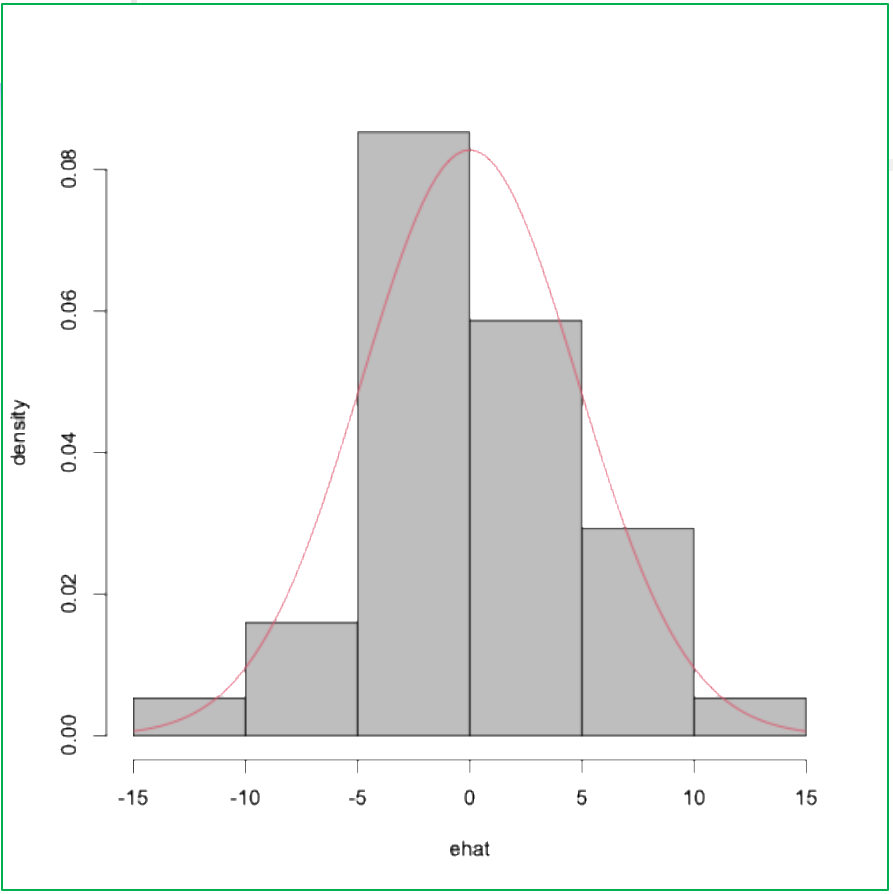
THE DISTRIBUTION OF THE RESIDUALS





Homoskedasticity assumption????

THE DISTRIBUTION OF THE RESIDUALS




```
################## homoskedasticity and normality of

plot(andy$price, andy$res)
plot(andy$advert, andy$res)

hist(andy$res)

ebar <- mean(andy$res)
sde <- sd(andy$res)
#hist for the residuals
hist(andy$res, col="grey", freq=FALSE, main="",
     ylab="density", xlab="ehat")
#normal curve with mean and sd
curve(dnorm(x, ebar, sde), col=2, add=TRUE,
      ylab="density", xlab="ehat")
```

```
> jarque.bera.test(andy$res) #(in package 'tseries')

        Jarque Bera Test

data:  andy$res
X-squared = 0.15896, df = 2, p-value = 0.9236
```

Joint Hypothesis Testing

- The *F*-statistic determines what constitutes a large reduction or a small reduction in the sum of squared errors

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/\left(N-K\right)}$$

where ***J* is the number of restrictions**, *N* is the number of observations and ***K* is the number of coefficients** in the unrestricted model

- **If the null hypothesis is true**, then the statistic *F* has what is called an *F*-distribution with *J* numerator degrees of freedom and *N* - *K* denominator degrees of freedom
- **If the null hypothesis is not true**, then the difference between SSE_R and SSE_U becomes large
  - The restrictions placed on the model by the null hypothesis significantly reduce the ability of the model to fit the data

```
> summary(mod1)

Call:
lm(formula = sales ~ price + advert, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max
-13.4825  -3.1434  -0.3456   2.8754  11.3049

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 118.9136     6.3516  18.722 < 0.0000000000000002 ***
price        -7.9079     1.0960  -7.215        0.000000000442 ***
advert        1.8626     0.6832   2.726               0.00804 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.886 on 72 degrees of freedom
Multiple R-squared:  0.4483,    Adjusted R-squared:  0.4329
F-statistic: 29.25 on 2 and 72 DF,  p-value: 0.0000000005041

> confint(mod1)
                2.5 %     97.5 %
(Intercept) 106.251852 131.575368
price       -10.092676  -5.723032
advert        0.500659   3.224510
```

Goodness of Fit

- In the multiple regression model the R^2 is relevant and the same formulas are valid, but now we talk of the proportion of variation in the dependent variable explained by all the explanatory variables included in the linear model
- The coefficient of determination is

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$$= 1 - \frac{SSE}{SST}$$

$$= 1 - \frac{\sum_{i=1}^{N}\hat{e}_i^{\,2}}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Goodness of Fit

- The predicted value of *y* is:

$$\hat{y}_i = \mathrm{b}_1 + \mathrm{b}_2 x_{i2} + b_3 x_{i3} + \mathrm{L} \; + b_K x_{iK}$$

- Recall that:

$$s_y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2} = \sqrt{\frac{SST}{N-1}}$$

- Then: $SST = (N-1)s_y^2$

Goodness of Fit

For the hamburger example:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \hat{e}_i^{\,2}}{\sum_{i=1}^{N} (y_i - \bar{y})^2} 1 - \frac{1718.943}{3115.482} = 0.448$$

Interpretation

- 44.8% of the variation in sales revenue is explained by the variation in price and by the variation in the level of advertising expenditure
- In our sample, 55.2% of the variation in revenue is left unexplained and is due to variation in the error term or to variation in other variables that implicitly form part of the error term

Goodness of Fit

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

$$\sum_{i=1}^{N}(y_i-\bar{y})^2 \neq \sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{N}\hat{e}_i^{\,2}$$

so *SST ≠ SSR + SSE*

- Under these circumstances it does not make sense to talk of the proportion of total variation that is explained by the regression
  - When your model does not contain a constant, it is better not to report R2
    - Even if your computer displays one

Model selection

A model could be misspecified if:

- ***we have omitted important variables***
- ***included irrelevant ones***
- ***chosen a wrong functional form***
- ***have a model that violates the assumptions of the multiple regression model***

There are three main model selection criteria:

1. ***The adjusted coefficient of determination R2***
2. ***The Akaike information criterion AIC***
3. ***Shwarz criteria SC (Bayesian information criterion BIC)***

A common feature of the criteria we describe is that they are suitable only for comparing models with the same dependent variable, not models with different dependent variables like y and ln(y)

Model selection

- The problem is that R^2 can be made large by adding more and more variables, even if the variables added have no justification

  Algebraically, it is a fact that as variables are added the sum of squared errors *SSE* goes down, and thus R^2 goes up

  If the model contains N - 1 variables, then R^2 =1

  An alternative measure of goodness of fit called the adjusted-R^2, denoted as $\overline{R^2}$:

$$\overline{R^2} = 1 - \frac{SSE/(N-K)}{SST/(N-1)}$$

- The **Akaike information criterion** (***AIC***) is given by: $AIC = \ln\left(\frac{SSE}{N}\right) + \frac{2K}{N}$
- **Schwarz criterion** (***SC***), also known as the **Bayesian information criterion** (***BIC***) is given by: $SC = \ln\left(\frac{SSE}{N}\right) + \frac{K\ln(N)}{N}$

```
> mod1 <- lm(sales~price+advert, data=andy)
> smod1 <- summary(mod1)
> adjRsq <- smod1$adj.r.squared #adjusted R2
> adjRsq
[1] 0.4329316
> AIC(mod1)
[1] 455.739
> BIC(mod1)
[1] 465.009
```

Indicator variable

An indicator variable is a binary variable that takes the values zero or one; it is used to represent a non quantitative characteristic, such as gender, race, or location.

For example, in the data file utown.dat we have a sample of 1000 observations on house prices (PRICE, in thousands of dollars) in two neighborhoods.

One neighborhood is near a major university and called University Town. Another similar neighborhood, called Golden Oaks, is a few miles away from the university. The indicator variable of interest is

$$UTOWN = \begin{cases} 1 & \text{house is in University Town} \\ 0 & \text{house is in Golden Oaks} \end{cases}$$

Indicator variable

Consider the model:

$$PRICE = \beta_1 + \beta_2 SQFT + e$$

β_2 is the value of an additional square foot of living area and β_1 is the value of the land alone

How do we account for location, which is a qualitative variable?

- Indicator variables are used to account for qualitative factors in econometric models
- They are often called **dummy, binary or dichotomous** variables, because they take just two values, usually one or zero, to indicate the presence or absence of a characteristic or to indicate whether a condition is true or false
- They are also called **dummy variables**, to indicate that we are creating a numeric variable for a qualitative, non-numeric characteristic
- We use the terms indicator variable and dummy variable interchangeably
- Generally, we define an indicator variable D as:

$$D = \begin{cases} 1 & \text{if characteristic is present} \\ 0 & \text{if characteristic is not present} \end{cases}$$

- So, to account for location, a qualitative variable, we would have:

$$D = \begin{cases} 1 & \text{if property is in the desirable neighborhood} \\ 0 & \text{if property is not in the desirable neighborhood} \end{cases}$$

Indicator variable

Adding our indicator variable to our model:

$$PRICE = \beta_1 + \delta D + \beta_2 SQFT + e$$

If our model is correctly specified, then:

$$E(PRICE) = \begin{cases} (\beta_1 + \delta) + \beta_2 SQFT & \text{when } D = 1 \\ \beta_1 + \beta_2 SQFT & \text{when } D = 1 \end{cases}$$

Adding an indicator variable causes a parallel shift in the relationship by the amount δ

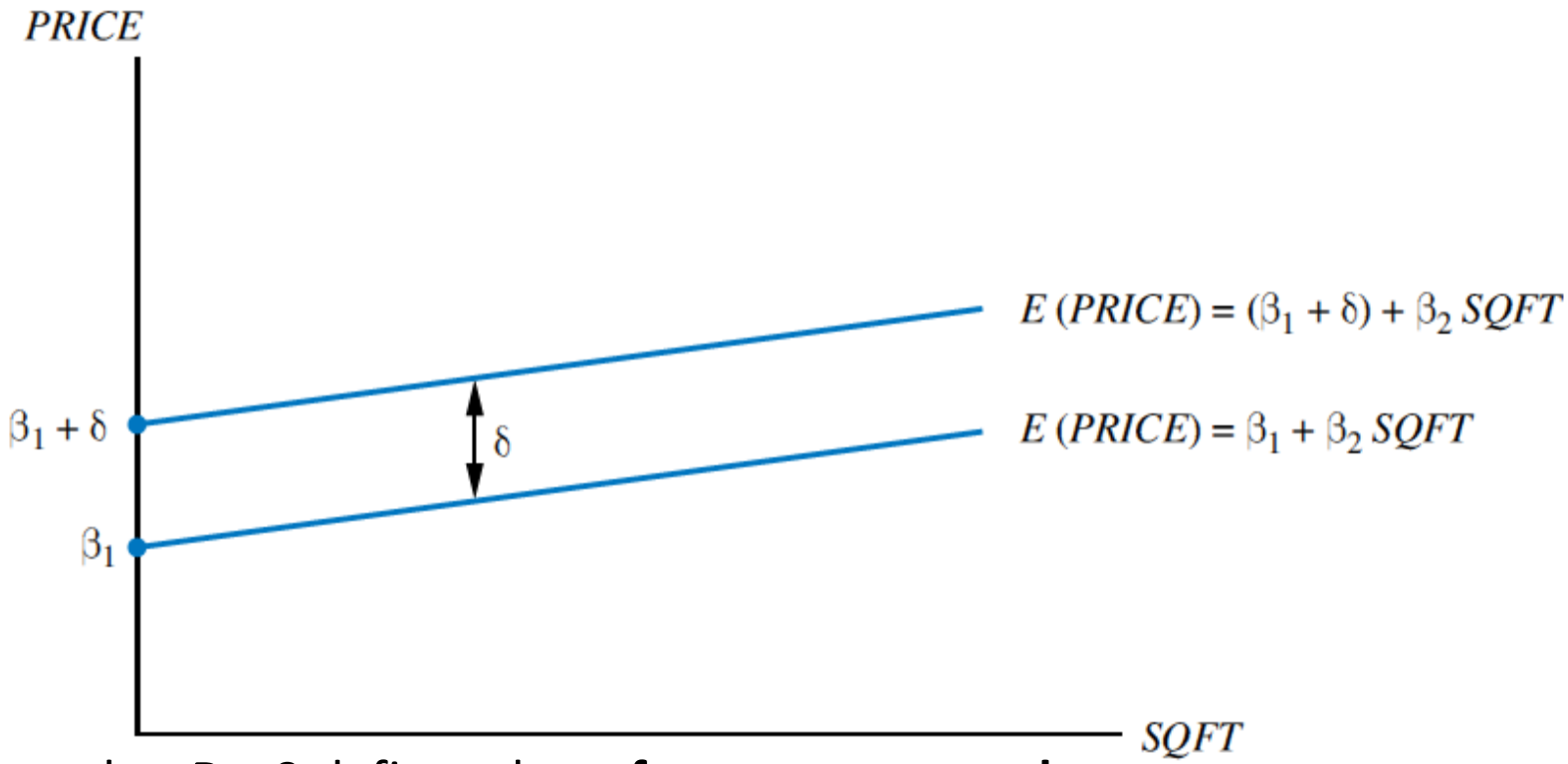
An indicator variable like D that is incorporated into a regression model to capture a shift in the intercept as the result of some qualitative factor is called an intercept indicator variable, or an intercept dummy variable

The least squares estimator's properties are not affected by the fact that one of the explanatory variables consists only of zeros and ones

D is treated as any other explanatory variable.

We can construct an interval estimate for *D*, or we can test the significance of its least squares estimate

Indicator variable



- The value $D = 0$ defines the **reference group**, or **base group**
  - We could pick any base
  - For example: $$LD = \begin{cases} 1 & \text{if property is not in the desirable neighborhood} \\ 0 & \text{if property is in the desirable neighborhood} \end{cases}$$
  - Then our model would be: $$PRICE = \beta_1 + \lambda LD + \beta_2 SQFT + e$$

Indicator variable – example with R

Obs: 1000 observations

price house price, in $1000
sqft square feet of living area, in 100's
age house age, in years
utown =1 if close to university
pool =1 if house has pool
fplace =1 if house has fireplace

```
> summary(utown)
     price            sqft             age            utown    pool     fplace
 Min.   :134.3   Min.   :20.03   Min.   : 0.000   0:481   0:796   0:482
 1st Qu.:215.6   1st Qu.:22.83   1st Qu.: 3.000   1:519   1:204   1:518
 Median :245.8   Median :25.36   Median : 6.000
 Mean   :247.7   Mean   :25.21   Mean   : 9.392
 3rd Qu.:278.3   3rd Qu.:27.75   3rd Qu.:13.000
 Max.   :345.2   Max.   :30.00   Max.   :60.000
```

```
> utown$utown <- as.factor(utown$utown)
> utown$pool <- as.factor(utown$pool)
> utown$fplace <- as.factor(utown$fplace)
>
> mod4 <- lm(price~sqft+utown+age+pool+fplace, data=utown)
> summary(mod4)

Call:
lm(formula = price ~ sqft + utown + age + pool + fplace, data = utown)

Residuals:
    Min      1Q  Median      3Q     Max
-47.971 -10.411   0.198  10.438  44.759

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.91188    4.28937   1.611 0.107410
sqft         8.31832    0.16717  49.759  < 2e-16 ***
utown1      60.19623    0.97153  61.960  < 2e-16 ***
age         -0.19299    0.05157  -3.743 0.000193 ***
pool1        4.35257    1.20526   3.611 0.000320 ***
fplace1      1.39881    0.97681   1.432 0.152452
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.33 on 994 degrees of freedom
Multiple R-squared:  0.8686,    Adjusted R-squared:  0.8679
F-statistic:  1314 on 5 and 994 DF,  p-value: < 2.2e-16
```

Model Specification

In any econometric investigation, choice of the model is one of the first steps

- What are the important considerations when choosing a model?
- What are the consequences of choosing the wrong model?
- Are there ways of assessing whether a model is adequate?

It is possible that a chosen model may have important **variables omitted**

- Our economic principles may have overlooked a variable, or lack of data may lead us to drop a variable even when it is prescribed by economic theory

Model Specification – example with R

edu_inc.def

faminc he we kl6 xtra_x5 xtra_x6

obs: 428 subsample of Mroz 1975 data including families with working wives

faminc = Family income in 2006 dollars
= [husband's hours worked in 1975 * husbands hourly wage
+ wife's hours worked in 1975 * wife's hourly wage]*3.78
(The multiplier 3.78 is used to convert 1975 dollars to 2006 dollars.)
he = Husband's educational attainment, in years
we = Wife's educational attainment, in years
kl6 = Number of children less than 6 years old in household
xtra_x5 = an artifically generated variable used to illustrate the effect of irrelevant variables.
xtra_x6 = a second artifically generated variable used to illustrate the effect of irrelevant variables.

Model Specification – example with R

EXAMPLE (edu_inc dataset)

$$\widehat{FAMINC} = -5534 + 3132HEDU + 4523WEDU$$
$$(\text{se}) \quad (11230) \quad (803) \quad (1066)$$
$$(p\text{-value}) \quad (0.622) \quad (0.000) \quad (0.000)$$

If we incorrectly omit wife's education:

$$\widehat{FAMINC} = -26191 + 5155HEDU$$
$$(\text{se}) \quad (8541) \quad (658)$$
$$(p\text{-value}) \quad (0.002) (0.000)$$

Model Specification

- Omitting *WEDU* leads us to overstate the effect of an extra year of education for the husband by about $2,000
  - Omission of a relevant variable (defined as one whose coefficient is nonzero) leads to an estimator that is biased
  - This bias is known as **omitted-variable bias**

Write a general model as:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

Omitting x_3 is equivalent to imposing the restriction $\beta_3 = 0$

It can be viewed as an example of imposing an incorrect constraint on the parameters

Model Specification

- You to think that a good strategy is to include as many variables as possible in your model.
  - Doing so will not only complicate your model unnecessarily, but may also inflate the variances of your estimates because of the presence of **irrelevant variables**.

$$\widehat{FAMINC} = -7759 + 3340\,HEDU + 5869\,WEDU - 14200\,KL6 + 889\,X_5 - 1067\,X_6$$

(se)	(11195)	(1250)	(2278)	(5044)	(2242) (1982)
$(p\text{-value})$	(0.500)	(0.008)	(0.010)	(0.005)	(0.692) (0.591)

- The inclusion of irrelevant variables ***has reduced the precision of the estimated coefficients*** for other variables in the equation

Model Specification

Some points for choosing a model:

1. Choose variables and a functional form on the basis of your theoretical and general understanding of the relationship
2. If an estimated equation has coefficients with unexpected signs, or unrealistic magnitudes, they could be caused by a misspecification such as the omission of an important variable
3. One method for assessing whether a variable or a group of variables should be included in an equation is to perform significance tests
4. Consider various model selection criteria
5. The adequacy of a model can be tested using a general specification test known as RESET

There are three main model selection criteria:

1. R^2
2. *AIC*
3. *SC* (*BIC*)

Model Specification

A model could be misspecified if:

- we have omitted important variables
- included irrelevant ones
- chosen a wrong functional form
- have a model that violates the assumptions of the multiple regression model

Collinearity and Insignificance

In general, whenever there are one or more exact linear relationships among the explanatory variables, then the condition of exact collinearity exists

- In this case the least squares estimator is not defined
- We cannot obtain estimates of β_k's using the least squares principle

The effects of this imprecise information are:

1. When estimator standard errors are large, it is likely that the usual t-tests will lead to the conclusion that parameter estimates are not significantly different from zero
2. Estimators may be very sensitive to the addition or deletion of a few observations, or to the deletion of an apparently insignificant variable
3. Accurate forecasts may still be possible if the nature of the collinear relationship remains the same within the out-of-sample observations

Collinearity and Insignificance

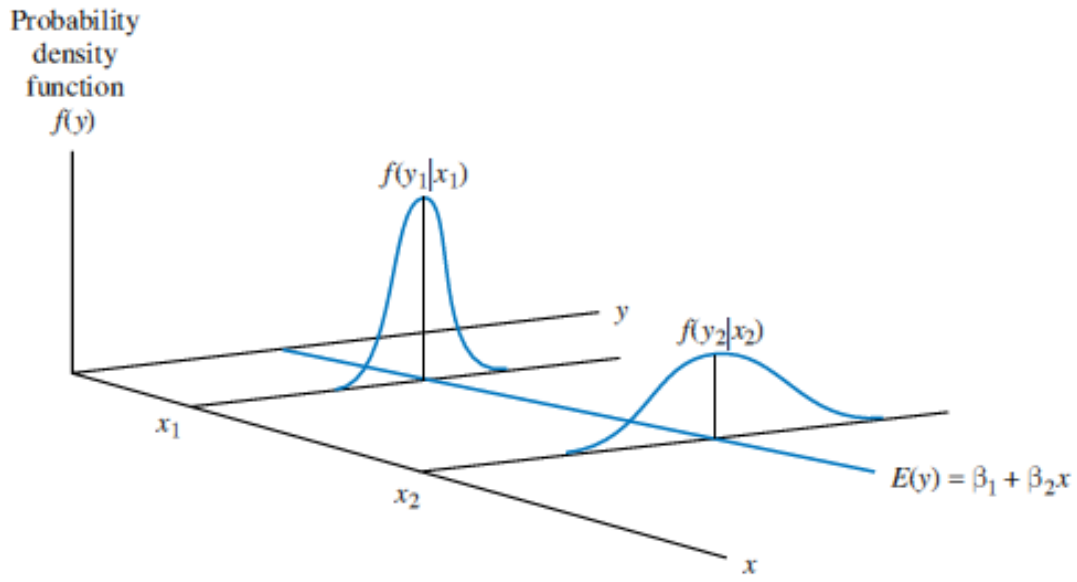
One simple way to detect collinear relationships is to use sample correlation coefficients between pairs of explanatory variables

- These sample correlations are descriptive measures of linear association
- However, in some cases in which collinear relationships involve more than two of the explanatory variables, the collinearity may not be detected by examining pairwise correlations

Try an auxiliary model: $x_2 = a_1 x_1 + a_3 x_3 + \mathrm{L} + a_K x_K + error$

If R2 from this artificial model is high, above 0.80, say, the implication is that a large portion of the variation in x2 is explained by variation in the other explanatory variables

Heteroskedasticity

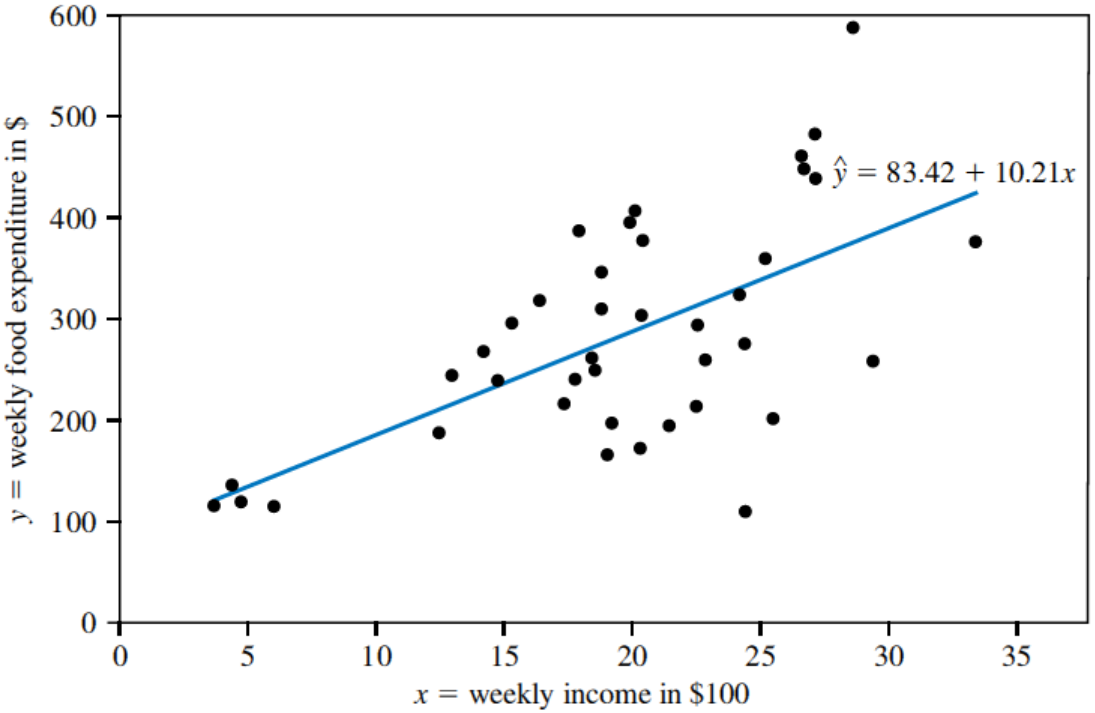


The variance of errors increases as x increases. The assumptions about error:

$E(e_i) = 0 \qquad \text{var}(e_i) = \sigma^2 \qquad \text{cov}(e_i, e_j) = 0$ ⟹ $\text{var}(y_i) = \text{var}(e_i) = h(x_i)$

➢ Heteroschedasticity is frequent in cross-sectional data

Heteroskedasticity



The residual graph provides information on the presence of heteroschedasticity

Heteroskedasticity

Two implications:

1. The least squares estimator is still a linear and correct estimator but it is no longer the best (i.e. minimum variance) estimator
2. The standard errors of the least squares estimator calculated using the formula valid in the case of homoschedasticity are wrong:

$$\text{var}(b_2) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Confidence intervals and hypothesis tests based on these standard errors may therefore be misleading.

$$y_i = \beta_1 + \beta_2 x_i + e_i \qquad \text{var}(e_i) = \sigma_i^2$$

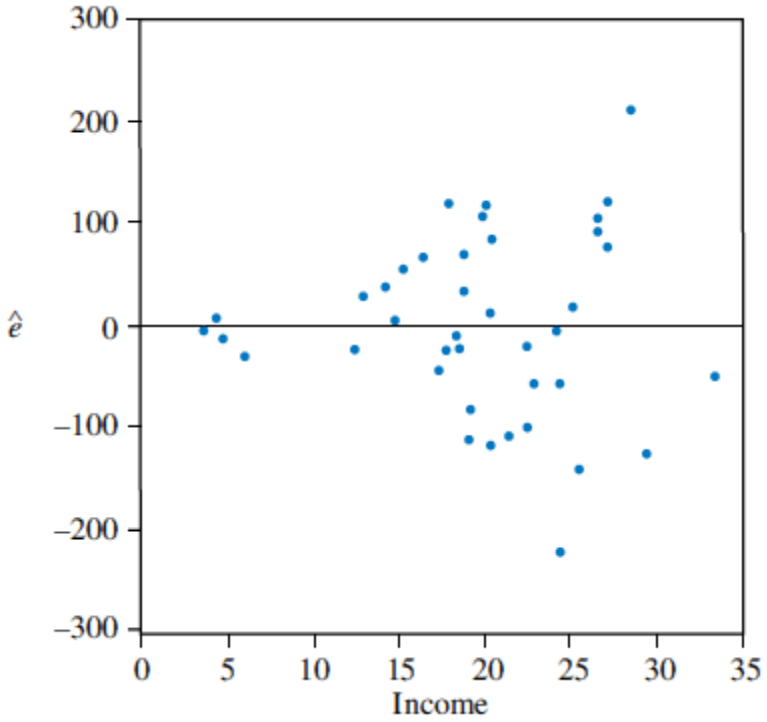
$$\text{var}(b_2) = \sum_{i=1}^{N} w_i^2 \sigma_i^2 = \frac{\sum_{i=1}^{N}\left[(x_i - \bar{x})^2 \sigma_i^2\right]}{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^2}$$

Heteroskedasticity

How to detect heteroskedasticity

1. Graphs of residuals (with respect to each independent variable and with respect to predicted y)

In R: `rstandard()`. Use stardardised residuals



Heteroskedasticity

2. Test

Since R2 measures goodness of fit it can be used as a test statistic knowing that:

Breusch-Pagan test

$$\chi^2 = N \times R^2 \sim \chi^2_{(S-1)}$$

Where:

n = sample size
R2 = R2(Coefficient of Determination) of the regression of squared residuals from the original regression.
k = number of independent variables.
The test statistic approximately follows a chi-square distribution.

- The null hypothesis for this test is that the error variances are all equal (**homoskedasticity)**
- The alternate hypothesis is that the error variances are not equal. More specifically, as Y increases, the variances increase (or decrease).

Heteroskedasticity

The test statistic for the Breusch-Pagan-Godfrey test is:
N * R2 (with k degrees of freedom)

Where:

n = sample size
R2 = R2(Coefficient of Determination) of the regression of squared residuals from the original regression.
k = number of independent variables.
The test statistic approximately follows a chi-square distribution.

- The null hypothesis for this test is that the error variances are all equal.
- The alternate hypothesis is that the error variances are not equal. More specifically, as Y increases, the variances increase (or decrease).

```
> #Heteroskedasticity Tests
> # The Breusch-Pagan heteroskedasticiy test
> summary(mod1)

Call:
lm(formula = sales ~ price + advert, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max
-13.4825  -3.1434  -0.3456   2.8754  11.3049

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 118.9136     6.3516  18.722  < 2e-16 ***
price        -7.9079     1.0960  -7.215 4.42e-10 ***
advert        1.8626     0.6832   2.726  0.00804 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.886 on 72 degrees of freedom
Multiple R-squared:  0.4483,    Adjusted R-squared:  0.4329
F-statistic: 29.25 on 2 and 72 DF,  p-value: 5.041e-10

> library(lmtest)
> bptest(mod1)

        studentized Breusch-Pagan test

data:  mod1
BP = 2.5722, df = 2, p-value = 0.2763
```

Accept Ho → residuals are Homoskedastic