Hypothesis Testing

It is useful in economics to use hypothesis testing to test the significance of an economic variable in influencing other variables (such as the significance of price on sales)

In the last section we discussed the basis of **forming a best linear unbiased estimator**, however **one further condition** is required of a model before using it in **a regression + testing hypotheses**:

Error is normally distributed, $\varepsilon \sim \mathrm{N}(0, \sigma^2)$

This means that the error distribution has a **mean, μ, of 0 and a variance of σ^2**, with all errors:

- Following a **bell-curve shape**
- Falling **close to the mean**
- Falling within a **certain range**
- Having **an equal likelihood of being below 0 than above it**

It is important for **error to be normally distributed** so that the data estimates, which naturally contain unavoidable error, are **also normally distributed**

The data estimates must be **normally distributed in order to be tested with a hypothesis**, as we will see later in this section when using a t distribution

Constructing a Hypothesis Test

Once this assumption is met, the hypothesis **proceeds in the following steps**:

1. State the **hypothesis**
2. State the **significance level**
3. Find **critical value**
4. **Calculate + collect required data**
5. Test data **against hypothesis** using significance level + **state statistical result**

1: State the Hypothesis

This comes from the **economic question of interest, usually testing the significance of the relationship between two economic variables, upon which a regression line has been drawn**

Process involves **putting forward a null hypothesis (H_0)** where a parameter is said to have a certain value, followed by introducing an **alternative hypothesis (H_1), contradicting the null hypothesis**, and this is what you are testing for, for example:

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

This above hypothesis is testing to see **if an x variable is significant in influencing a y variable** (if β = 0, there is **no slope in the regression line** and **hence no significant relationship between the x and y variables**)

This example is also a **two-tailed hypothesis test**, which is the **preferred method over a one-tailed test** as we are examining deviations **above + below** predicted value of β = 0 (i.e. we are testing for **both a negative + positive correlation**)

One-tailed tests typically take the form of $\boldsymbol{H_1: \beta > 0}$ **or** $\boldsymbol{H_1: \beta < 0}$ to test for **positive or negative correlation** respectively

If the variables in question were **price and sales**, it would be better to **use a one-tailed test** in most cases where $\boldsymbol{H_1: \beta < 0}$, since **using standard economic theory**, we can reasonably assume that if there was going to be a relationship between the two variables, it would **certainly be negative**

2: State the Significance Level

The significance level, α, gives the **percentage margin of error that is accepted**, where random chance is **thought unlikely to explain the result within this margin**

There are **four possible results** of any hypothesis test:

- **Reject H_0 correctly** (e.g. there is correlation)
- **Reject H_0 incorrectly** (**type I error** – e.g. concluding there is correlation when there isn't)
- **Fail to reject H_0 correctly** (e.g. there is no correlation)
- **Fail to reject H_0 incorrectly** (**type II error** – e.g. concluding there is no correlation when there is)

The probability of a type I error is **equal to the significance level**, α, which is the probability of **incorrectly rejecting the null hypothesis**

It's important to note that the significance level is **chosen in advance** before the test is carried out, with **5% being a typical choice**, seeing **incorrect results 5% of the time**

A **smaller significance level**, such as 1%, does mean a **lower proportion of type I errors**, however due to the **inevitable error in any model**, this may see all models **fail the test despite being good enough to predict actual results** most of the time (more type II errors)

3: Find Critical Value

Once the significance level is chosen, statistics tables are used to find the **critical value**

For instance, a **5% significance level** may give a value of **1.99** whilst a more **stringent 1% significance level** may give a value of **2.62 – hence the critical regions are ±1.99 and ±2.62 respectively**

If the value calculated from test statistic falls **within the critical region**, i.e. higher or lower than **1.99 or -1.99 respectively**, then we have **enough evidence to reject H_0** and suggest there is **correlation since β is far enough from 0** in this two-tailed test

By rejecting H_0, however, there is a **5% chance (the significance level) that we have incorrectly accepted H_1 due to random error**

4: Calculating + Collecting Required Data

Before **calculating + collecting data**, an estimator is first required to **proxy the population data's variance (σ^2) and its sample distribution**, which are both unknown, to get our test statistic

An **unbiased estimator of variance** is shown mathematically as follows:

$$\underline{\sigma}^2 = \frac{\sum \varepsilon_i{}^2}{n-2}$$

This says that the estimated variance is equal to **the residual sum of squares (RSS = $\sum {\varepsilon_i}^2$)** divided by the **degrees of freedom**, which is equal to **the number of observations in the model (n) minus the number of parameters in the model (2)**

If the error is **distributed normally**, as assumed earlier, then on **repeated sampling we can show the following, using the unbiased estimator of variance**:

$$\frac{(n-2)\underline{\sigma^2}}{\sigma^2} \sim {\chi^2}_{(n-2)}$$

This means that the **degrees of freedom multiplied by the estimated variance divided by the actual variance** approximately follows a **Chi-squared (χ^2) distribution**

The **ratio of a standard normal distribution to the square root of a χ^2 distribution** is the **t distribution** with the **same degrees of freedom as the χ^2 distribution**

NOTE that a t distribution roughly follows the same shape as a standard normal distribution, and is used because we do not know the population variance to formulate a standard normal distribution to generate the test statistic (however a t distribution may be used if we use the sample variance estimator)

T distribution **has larger tails** and has a **flatter peak** to account for fact population variance is **unknown**

If **sample variance is used as an estimator**, then on **repeated sampling** the following holds to calculate the test statistic, which will be compared to the critical value:

$$\frac{(\underline{\beta} - \beta)}{\underline{Se(\beta)}} \sim t_{(n-2)}$$

Here, Se(β) is the **estimated (from the sample) standard error (Se) of β**

The standard error of β has a very **complicated formula**, so using **statistical software** to calculate this is advised (i.e. **Toolpak on excel**)

This result allows for hypothesis tests to be run on β **without requiring knowledge of any population parameters** (just estimated sample parameters)

If H_0 is $\beta = 0$, then this result simplifies to:

$$\frac{\underline{\beta}}{\underline{Se(\beta)}} \sim t_{(n-2)}$$

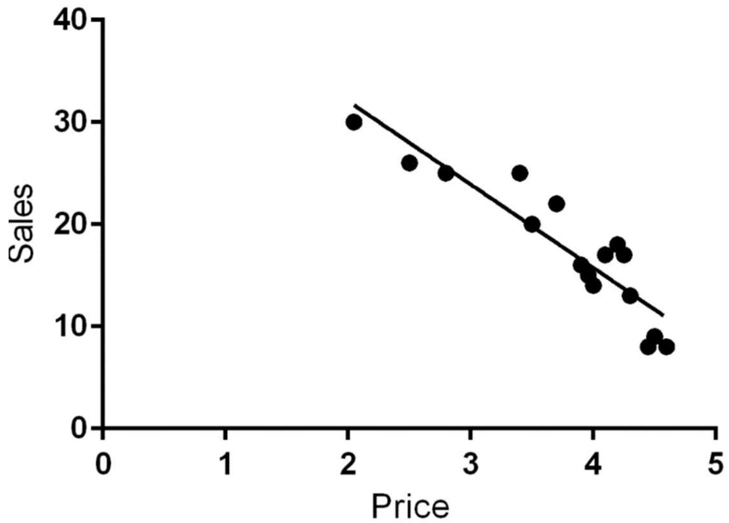
Similarly if H_0 was $\beta = 2$, then this result simplifies to:

$$\frac{\underline{\beta} - 2}{\underline{Se(\beta)}} \sim t_{(n-2)}$$

Now let's apply our knowledge to a piece of sample data, which shows the **number of sales (y) of a product at different prices (x):**

1. y = 30, x = 2.05
2. y = 26, x = 2.50
3. y = 25, x = 2.80
4. y = 25, x = 3.40
5. y = 22, x = 3.70
6. y = 16, x = 3.90
7. y = 14, x = 4.00
8. y = 15, x = 3.96
9. y = 17, x = 4.10
10. y = 18, x = 4.20
11. y = 17, x = 4.25
12. y = 13, x = 4.30
13. y = 9, x = 4.50
14. y = 8, x = 4.45
15. y = 8, x = 4.60
16. y = 20, x = 3.50

Once we perform the **OLS method to this data to form a regression line (either by hand or using a data software tool such as Toolpak on excel)**, we can **infer the following results**:



$$y = -8.1615x + 48.400$$

$$R^2 = 0.83799$$

$$RSS = 107.157$$

$$Se(\beta) = 0.95907$$

In this example, $R^2 = 0.83799$, suggesting the regression line **explains roughly 84% of the changes in sales**

The residual sum of squares, which as previously discussed is the **summation of the squares of each individual error term**, is 107.157, and hence using the aforementioned equation, $\underline{\boldsymbol{\sigma}}^2 = \frac{\sum \varepsilon_i{}^2}{n-2}$, we can calculate the sample variance as $\underline{\boldsymbol{\sigma}}^2 = \frac{\mathbf{107.157}}{\mathbf{16-2}} = \mathbf{7.654}\ldots$, with the degrees of freedom **being 14 (since number of observations, n, is 16)**

The sample variance here acts as a **proxy for the entire population variance**, although **not very accurate since it is only based off 16 samples**

Taking the **root of the sample variance** gives the **sample standard deviation,** $\underline{\boldsymbol{\sigma}} = \mathbf{2.767}$, and this is **how far each sales estimate is on average from the best fit line** (either above or below it)

We can use the values of β and Se(β) for the test statistic, using H_0 as β = 0:

$$\frac{(\underline{\beta} - \beta)}{\underline{Se(\beta)}} \sim t_{(n-2)} = \frac{-8.1641}{0.95907} \sim t_{(n-2)} = -8.5125 \sim t_{(14)}$$

5: Test Against Hypothesis

The data result is **compared to the 5% significance level critical value** (since we are using 5% in this hypothesis test) for n – 2 = 14 degrees of freedom to assess the test

Critical values of *t* for two-tailed tests

Significance level (α)

Degrees of freedom (*df*)	.2	.15	.1	.05	.025	.01	.005	.001
1	3.078	4.165	6.314	12.706	25.452	63.657	127.321	636.619
2	1.886	2.282	2.920	4.303	6.205	9.925	14.089	31.599
3	1.638	1.924	2.353	3.182	4.177	5.841	7.453	12.924
4	1.533	1.778	2.132	2.776	3.495	4.604	5.598	8.610
5	1.476	1.699	2.015	2.571	3.163	4.032	4.773	6.869
6	1.440	1.650	1.943	2.447	2.969	3.707	4.317	5.959
7	1.415	1.617	1.895	2.365	2.841	3.499	4.029	5.408
8	1.397	1.592	1.860	2.306	2.752	3.355	3.833	5.041
9	1.383	1.574	1.833	2.262	2.685	3.250	3.690	4.781
10	1.372	1.559	1.812	2.228	2.634	3.169	3.581	4.587
11	1.363	1.548	1.796	2.201	2.593	3.106	3.497	4.437
12	1.356	1.538	1.782	2.179	2.560	3.055	3.428	4.318
13	1.350	1.530	1.771	2.160	2.533	3.012	3.372	4.221
14	1.345	1.523	1.761	2.145	2.510	2.977	3.326	4.140
15	1.341	1.517	1.753	2.131	2.490	2.947	3.286	4.073
16	1.337	1.512	1.746	2.120	2.473	2.921	3.252	4.015
17	1.333	1.508	1.740	2.110	2.458	2.898	3.222	3.965
18	1.330	1.504	1.734	2.101	2.445	2.878	3.197	3.922
19	1.328	1.500	1.729	2.093	2.433	2.861	3.174	3.883
20	1.325	1.497	1.725	2.086	2.423	2.845	3.153	3.850
21	1.323	1.494	1.721	2.080	2.414	2.831	3.135	3.819
22	1.321	1.492	1.717	2.074	2.405	2.819	3.119	3.792
23	1.319	1.489	1.714	2.069	2.398	2.807	3.104	3.768
24	1.318	1.487	1.711	2.064	2.391	2.797	3.091	3.745
25	1.316	1.485	1.708	2.060	2.385	2.787	3.078	3.725
26	1.315	1.483	1.706	2.056	2.379	2.779	3.067	3.707
27	1.314	1.482	1.703	2.052	2.373	2.771	3.057	3.690
28	1.313	1.480	1.701	2.048	2.368	2.763	3.047	3.674
29	1.311	1.479	1.699	2.045	2.364	2.756	3.038	3.659
30	1.310	1.477	1.697	2.042	2.360	2.750	3.030	3.646
40	1.303	1.468	1.684	2.021	2.329	2.704	2.971	3.551
50	1.299	1.462	1.676	2.009	2.311	2.678	2.937	3.496
60	1.296	1.458	1.671	2.000	2.299	2.660	2.915	3.460
70	1.294	1.456	1.667	1.994	2.291	2.648	2.899	3.435
80	1.292	1.453	1.664	1.990	2.284	2.639	2.887	3.416
100	1.290	1.451	1.660	1.984	2.276	2.626	2.871	3.390
1000	1.282	1.441	1.646	1.962	2.245	2.581	2.813	3.300
Infinite	1.282	1.440	1.645	1.960	2.241	2.576	2.807	3.291

The critical value (found in statistics tables for these parameters) **is ±2.145**, whilst the **test statistic, as calculated above, is -8.5125**

Since the test result is **below the negative critical value**, we have **sufficient evidence to reject the null hypothesis** and suggest that the **slope (β) of the regression line is not 0**, **and there is a correlation between price and sales,** with at least **95% confidence** that this conclusion **has not been made due to random error**