Issues and wrong conclusions in statistical testing

Lauri Tarpila

Domast Student Seminar

2021

Introduction

I am trying to show that assumptions cannot be validated using statistic testing

In addition I show that if we claim to be able to validate assumptions we also end up in a situation where we cannot validate assumptions

Because of these I suggest that science would not be considered as an objective and justified point of view by itself but rather as a tool for "unscientific" ideas.

Structure

First I am going to show how I earlier thought how statistical testing worked.

After that I show you what are the wrong conclusion in that.

Finally I will show the how the same problem arises even if we think that we can verify models by assigning them some "usefulness" values.

Linear Model

To calculate linear model s regression coefficient (or correlation coefficient)

- 1.1. Observations independent and from the same distribution $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^{1\times p})$
- 1.2. Finite variance of Y and X
- 1.3. Connection of Y and X is linear $Y = X\beta + \epsilon$
- 1.4 Orthogonality of X and ϵ ($Cov(X, \epsilon) = 0$)
- 1.5 $Var(\epsilon_i) = \sigma^2$
- 1.6 If $X \in \mathbb{R}^{n\times p}$ then r(X)=p

Every vector can be written as a sum of orthogonal vectors. Now because $X \perp \epsilon$ AND 1.3 is true we can develop an estimator for β by using the projection of Y to Col(X). The project matrix to Col(X) is $P = X(X^TX)^{-1}X^T$.

Linear Model

To calculate linear model's regression coefficient (or correlation coefficient)

- 1.1. Observations independent and from the same distribution $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^{1\times p})$
- 1.2. Finite variance of Y and X
- 1.3. $Y = \beta x + \epsilon$
- 1.4 Orthogonality of X and ϵ $(Cov(X, \epsilon) = 0)$
- 1.5 $Var(\epsilon_i) = \sigma^2$
- 1.6 If $X \in \mathbb{R}^{n\times p}$ then r(X)=p

Every vector can be written as a sum of orthogonal vectors. Now because $X \perp \epsilon$ AND 1.3 is true we can develop an estimator for β by using the fact that $X\beta$ is a projection of Y to Col(X). The project matrix for Col(X) is $P = X(X^TX)^{-1}X^T$.

$$PY = X(X^TX)^{-1}X^TY = X(X^TX)^{-1}X^T(X\beta + \epsilon) =$$

$$X(X^TX)^{-1}X^TX\beta + X(X^TX)^{-1}X^T\epsilon = X\beta + 0 = X\beta$$

So from here

$$\beta = (X^TX)^{-1}X^TY$$

Projection

Every vector can be written as a sum of orthogonal vectors. In this case Y can be written as a sum of $X\beta$ and ϵ because they are orthogonal. Now because orthogonality AND because 1.3 is true we can develop an estimator for β by using the fact that $X\beta$ is a projection of Y to Col(X). The project matrix for Col(X) is $P = X(X^TX)^{-1}X^T$.

$$PY = X(X^TX)^{-1}X^TY = X(X^TX)^{-1}X^T(X\beta + \epsilon) =$$

$$X(X^TX)^{-1}X^TX\beta + X(X^TX)^{-1}X^T\epsilon = X\beta + 0 = X\beta$$

So from here

$$\beta = (X^TX)^{-1}X^TY$$

When using data to estimate β we write

$$\hat{\beta} = (X_n^TX_n)^{-1}X_n^TY_n = \beta + (X_n^TX_n)^{-1}X_n^T\epsilon_n$$

Normality

$$\beta = (X^T X)^{-1} X^T Y$$

When using data to estimate β we write

$$\hat{\beta} = (X_n^T X_n)^{-1} X_n^T Y_n = \beta + (X_n^T X_n)^{-1} X_n^T \epsilon_n$$

According to Central Limit Theorem we get that the sum of rows in ϵ_n converges to a normal distribution so,

$$X_n^T \epsilon_n \to N(0, X_n X_n^T \sigma^2)$$

Consistency

According to Central Limit Theorem the sum of independent random variables converges asymptotically to a normal distribution. So,

$$X_n^T \epsilon_n \rightarrow N(0, X_n X_n^T \sigma^2)$$

And then

$$(X_n^T X_n)^{-1} X_n^T \epsilon_n \rightarrow N(0, (X_n^T X_n)^{-1} (X_n X_n^T)^{-1} X_n X_n^T \sigma^2) =$$

$$N(0, (X_n^T X_n)^{-1} sigma^2)$$

So

$$\hat{\beta} \sim N(\beta, \sigma^2 (X_n^T X_n)^{-1})$$

Because $(X_n^T X_n)$ converges to $X^T X$ and because of full column rank of X the matrix $X^T X$ is invertible. In addition Law of Large Numbers state that $X_n^T \epsilon_n$ converges to $E[X^T \epsilon] = Cov(X, \epsilon)$ which we assumed to be zero. In conclusion

$$(X_n^T X_n)^{-1} X_n^T \epsilon_n \to (X^T X)^{-1} E[X^T \epsilon] = 0$$

So

$$\hat{\beta} = \beta + (X_n^T X_n)^{-1} X_n^T \epsilon_n \to \beta + (X^T X)^{-1} E[X^T \epsilon] = \beta + 0 = \beta$$

estimator $\hat{\beta}$ is consistent.

Hypothesis testing

$$\hat{\beta} \sim N(\beta, \sigma^2 (X_n^T X_n)^{-1})$$

- 1.1. Observations independent and from the same distribution $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^{1 \times p})$
- 1.2. Finite variance of Y and X
- 1.3. $Y = \beta x + \epsilon$
- 1.4 Orthogonality of X and e ($Cov(X, e) = 0$)
- 1.5 $Var(\epsilon_i) = \sigma^2$
- 1.6 If $X \in \mathbb{R}^{n \times p}$ then r(X)=p

Now we have showed the needed assumption for a consistent estimator for β. Now we would set null and alternative hypothesis and choose a statistical test for it.

$$H_0 : \beta = 0 \;\; H_1 : \beta \neq 0$$

The statistical test used is T-test. It calculates the probability that the test measure is greater or equal than the test measures value in that data set when the Null hypothesis is true.

Now we have showed the needed assumption for a consistent estimator for β. Now we would set null and alternative hypothesis and choose a statistical test for it.

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

The statistical test used is T-test. It gives us a p-value which calculates the probability that the test measure is greater or equal than the test measures value in that data set when the Null hypothesis is true.

$$T = \frac{\hat{\beta}_j - 0}{S\sqrt{d_{jj}}},$$

where S^2 on unbiased estimator for variance σ^2 and d_{jj} is the element from the row j and column j from the matrix $(X_n^T X_n)^{-1}$.

And when all the linear model assumptions are true then T follows T-distribution and it follows the the zero centered distribution when $\beta_j = 0$.

$$T = \frac{\hat{\beta}_j - 0}{S\sqrt{d_{jj}}},$$

where S^2 on unbiased estimator for variance σ^2 and d_{jj} is the element from the row j and column j from the matrix $(X_n^T X_n)^{-1}$.

And when all the linear model assumptions are true then T follows T-distribution and it follows the the zero centered distribution when $\beta_j = 0$.

Now we would choose a significance for the p-value which the test gives us. If the test's p-value is less than the significance level we reject the Null hypothesis. Otherwise we stick with the Null hypothesis.

$$T = \frac{\hat{\beta}_j - 0}{S\sqrt{d_{jj}}},$$

Now we would choose a critical value for the T test measure. If the test measure value exceeds the significance level we reject the Null hypothesis. Otherwise we stick with the Null hypothesis.

I earlier thought that rejecting the Null hypothesis would somehow validate the assumptions made in the model. In addition I feel that it is thought that not rejecting the Null hypothesis somehow gives evidence that the linearity assumptions are not truth. I will now try to explain why I don't think like this.

$$T = \frac{\hat{\beta}_j - 0}{S\sqrt{d_{jj}}},$$

The test gives us the p-value which "calculates the probability that the test measure is greater or equal than the test measures value in that data set when the Null hypothesis is true".

$$\text{p-value } = P(T > t | H_0 \text{ is true})$$

And then according to the p-value we reject or stick with the Null hypothesis. So this procedure could be written as

If H_0 TRUE and p-value $>$ critical p-value
$\rightarrow$ stick with H_0
$\rightarrow$ no validation to linearity assumption

And then according to the p-value we reject or stick with the Null hypothesis. So this procedure could be written as

If H_0 TRUE and p-value $>$ critical p-value
$\rightarrow$ stick with H_0
$\rightarrow$ no validation to linearity assumption

Because all the assumptions are needed to create a consistent estimator for β when H_0 is true then all the assumption have to be true also. Even if one assumptions (or alternative assumptions for 1.1.,1.2 and 1.5) does not hold $\hat{\beta}$ is not a consistent estimator and so the T-test measure does not follow a T-distribution and the p-value does not give reliable results.

Because all the assumptions are needed to create a consistent estimator for β when H_0 is true then all the assumption have to be true also. Even if one assumptions does not hold $\hat{\beta}$ is not a consistent estimator and so the T-test measure does not follow a T-distribution and the p-value does not give reliable results.

Because of this the previous formula should be written as

If all the assumptions are true (including linearity) and
$\beta = 0$ and p-value $>$ critical p-value
$\rightarrow$ stick with H_0
$\rightarrow$ no validation to linearity assumption

Because of this the previous formula should be written as

If all the assumptions are true (including linearity) TRUE and
$\beta = 0$ and p-value $>$ critical p-value
$\rightarrow$ stick with H_0
$\rightarrow$ no validation to model assumptions (including linearity)

However it is not logical to assume linearity and then conclude that we should not validate linearity. Linearity is assumed true so concluding that it would not be truth would be mathematically incorrect.

Verifying Models

I have tried to show that we cannot verify the assumptions using statistical testing. But what would happen if we said that we could?

We could also verify the assumptions by using some property of the model as the criteria. For example, a model created using some assumptions could minimize the squared error of the model's predictions and the observations better than other models. If the p value of the regression coefficient $\hat{\beta}$ is zero we then would conclude that the linear model's assumptions are correct over the assumptions of the model without the linear connection ($Y = \epsilon$)

And similarly in this situation when we would compare two model's abilities to minimize the squared error (accuracies) we would use variance analysis to determine the difference of the accuracies of the models. The model' assumptions with a statistically significant higher accuracy compared to the other would be then be

Verifying Models

We could also verify the assumptions by using some property of the model as the criteria. For example, a model created using some assumptions could minimize the squared error of the model's predictions and the observations better than other models. If the p value of the regression coefficient $\hat{\beta}$ is zero we then would conclude that the linear model's assumptions are correct over the assumptions of the model without the linear connection ($Y = \epsilon$).

We would however need an other criteria to verify the criteria we used to verify the assumptions. For example we could use the feeling of happiness to choose minimizing squared error over minimizing the absolute error of the predictions and observations. Even so using another criteria for example the feeling of joy would also yield results and without an even other criteria we could not justify the using of happiness criteria over joy criteria.

Example

We would however need an other criteria to verify the criteria we used to verify the assumptions. For example we could use the feeling of happiness to choose minimizing squared error over minimizing the absolute error of the predictions and observations. Even so using another criteria for example the feeling of joy would also yield results and without an even other criteria we could not justify the using of happiness criteria over joy criteria.

Example: Scientist try to conclude should they use squared error or absolute error as the minimizing criteria. They choose to calculate a happiness index $[0, 100]$ from each country which uses squared error and which uses absolute error. Then they use variance analysis to compare the index values of the countries which use ϵ^2 and which use $|\epsilon|$ as the minimization criteria.

Example

Example: Scientist try to conclude should they use absolute error or squared error as the minimizing criteria. They choose to calculate a happiness index $[0, 100]$ from each country which uses absolute error and which uses squared error .

In this study they code the countries which use absolute error as "1" and countries with squared error as "2". So they will have n_1 ($Y_1 1 \ldots Y_1 n_1$) data points from countries which use absolute error and n_2 ($Y_2 1 \ldots Y_2 n_2$)from countries which use squared error. The mean of the "1" is μ_1 and "2" is μ_2. They use variance analysis which is a type of linear model to compare the distributions of "1" and "2".

Assumptions for this model are as in the linear model

- 1.1. Observations independent and from the same distribution $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^{1\times p})$
- 1.2. Finite variance of Y (and X)
- 1.3. $Y_{ji} = \mu_j + \epsilon_{ji}$
- 1.4 Orthogonality of X and e ($Cov(X, e) = 0$)
- 1.5 $Var(\epsilon_{ji}) = \sigma^2$
- 1.6 If $X \in \mathbb{R}^{n\times p}$ then r(X)=p

Model written in matrix form

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \end{bmatrix}$$

Assumptions for this model are as in the linear model

- 1.1. Observations independent and from the same distribution $(y_i, x_i) \in \mathbb{R}^n \times \mathbb{R}^{n\times p})$
- 1.2. Finite variance of Y (and X)
- 1.3. $Y_{ji} = \mu_j + \epsilon_{ji}$
- 1.4 Orthogonality of X and e ($Cov(X, e) = 0$)
- 1.5 $Var(\epsilon_{ji}) = \sigma^2$
- 1.6 If $X \in \mathbb{R}^{n\times p}$ then r(X)=p

Again like in the previous Linear Model example we create a estimator for μ using the projection to Col(X).

$$\hat{\mu} = (X_n^T X_n)^{-1} X_n^T Y_n = \mu + (X_n^T X_n)^{-1} X_n^T \epsilon_n \text{ (1.2,1.3,1.6)}$$

Again the $\hat{\mu}$ is normally distributed (1.1,1.5)

$$\hat{\mu} = \mu + (X_n^T X_n)^{-1} X_n^T \epsilon_n \to N(\mu, (X_n^T X_n)^{-1} (X_n X_n^T)^{-1} X_n X_n^T \sigma^2) =$$

$$N(\mu, (X_n^T X_n)^{-1} sigma^2)$$

and consistent (1.4)

$$\hat{\mu} = \mu + (X_n^T X_n)^{-1} X_n^T \epsilon_n \to \mu + (X^T X)^{-1} E[X^T \epsilon] = \mu + 0 = \mu$$

Assumptions for this model are as in the linear model

- 1.1. Observations independent and from the same distribution $(y_i, x_i) \in \mathbb{R}^n \times \mathbb{R}^{n \times p})$
- 1.2. Finite variance of Y (and X)
- 1.3. $Y_{ji} = \mu_j + \epsilon_{ji}$
- 1.4 Orthogonality of X and e ($Cov(X, e) = 0$)
- 1.5 $Var(\epsilon_{ji}) = \sigma^2$
- 1.6 If $X \in \mathbb{R}^{n \times p}$ then r(X)=p

and consistent (1.1,1.2,1.4)

$$\hat{\mu} = \mu + (X_n^T X_n)^{-1} X_n^T \epsilon_n \rightarrow \mu + (X^T X)^{-1} E[X^T \epsilon] = \mu + 0 = \mu$$

This time we would create a test measure from the difference of μ_1 and μ_2 and we know that the difference of normal distribution is normally distributed

$$\hat{\mu_1} - \hat{\mu_2} \sim N(\mu_1 - \mu_2, \frac{\sigma^2}{n_2} + \frac{\sigma^2}{n_2})$$

Then our Hypothesis would be

$$H_0 : \mu_1 = \mu_2 \;\; H_1 : \mu_1 \neq \mu_2$$

Assumptions for this model are as in the linear model

- 1.1. Observations independent and from the same distribution $(y_i, x_i) \in \mathbb{R}^n \times \mathbb{R}^{n \times p})$
- 1.2. Finite variance of Y (and X)
- 1.3. $Y_{ji} = \mu_j + \epsilon_{ji}$
- 1.4 Orthogonality of X and e ($Cov(X, e) = 0$)
- 1.5 $Var(\epsilon_{ji}) = \sigma^2$
- 1.6 If $X \in \mathbb{R}^{n \times p}$ then r(X)=p

and consistent (1.1,1.2,1.4)
Then our Hypothesis would be

$$H_0 : \mu_1 = \mu_2 \; H_1 : \mu_1 \neq \mu_2$$

Again the T-test measure will be T-distributed because $\mu_1 - \mu_2$ is normalli distributed.

Then we would calculate the T-test value and compare the p-value to the significance level. If the p-value exceeds the significance level we would reject the H_0 and otherwise we stick with the H_0.

Assumptions for this model are as in the linear model

- 1.1. Observations independent and from the same distribution $(y_i, x_i) \in \mathbb{R}^n \times \mathbb{R}^{n \times p})$
- 1.2. Finite variance of Y (and X)
- 1.3. $Y_{ji} = \mu_j + \epsilon_{ji}$
- 1.4 Orthogonality of X and e $(Cov(X, e) = 0)$
- 1.5 $Var(\epsilon_{ji}) = \sigma^2$
- 1.6 If $X \in \mathbb{R}^{n \times p}$ then r(X)=p

and consistent (1.1,1.2,1.4)
Then our Hypothesis would be

$$H_0 : \mu_1 = \mu_2 \ H_1 : \mu_1 \neq \mu_2$$

The Scientist wanted to conclude that which minimization criteria (absolute error or squared error) yields more happiness. So they want to make a choice according to the result that is there a difference between the happiness indexes or not. However the chain of deduction remains similar than in the previous linear model that you cannot calculate the difference without assuming that there is a difference. Because of this the result will not yield an answer whether or not the happiness levels differ or not.

Scientists want to make a choice according to the result that is there a difference between the happiness indexes or not. However the chain of deduction remains similar than in the previous linear model that you cannot calculate the difference without assuming that there is a difference. Because of this the result will not yield an answer whether or not the happiness levels differ or not.

Again the p value of this test will give a probability for the test measure to be as great as it is or greater assuming that the Null hypothesis is true. Because we need all the assumptions to be true in order to make a reliable test the test cannot give us information about the assumptions.

If all the assumptions are true (including linearity) and
$\mu = 0$ and p-value $>$ critical p-value
$\rightarrow$ stick with H_0
$\rightarrow$ no validation to model assumptions (including linearity)

My idea was to show that we need to make "unscientific" assumptions first and then use science to that. For example we assume that using squared error gives us the truth and then use science to minimize squared error.

The "unscientific" name is misleading because in order to talk about unscience we need to define science and in that high level we have not yet defined it.

In my opinion this interpretation of science is most consistent because then we don't have any issues with unprovable assumptions.

Thank you

- 1.1. Observations independent and from the same distribution $(y_i, x_i) \in \mathbb{R}^n \times \mathbb{R}^{n \times p})$
- 1.2. Finite variance of Y (and X)
- 1.3. $Y_{ji} = \mu_j + \epsilon_{ji}$
- 1.4 Orthogonality of X and e ($Cov(X, e) = 0$)
- 1.5 $Var(\epsilon_j i) = \sigma^2$
- 1.6 If $X \in \mathbb{R}^{n \times p}$ then r(X)=p

Model written in matrix form

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \end{bmatrix}$$

And then because

$$\hat{\mu}_i \sim N(\mu_i, \frac{\sigma^2}{n_1})$$

$$\hat{\mu}_1 - \hat{\mu}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2})$$

- 1.1. Observations independent
- 1.2. Finite variance of Y (and X)
- 1.3. $Y_{ji} = \mu_j + \epsilon_{ji}$
- 1.3 and 1.2 Orthogonality of X and e ($Cov(X, e) = 0$)

$$\hat{\mu_1} - \hat{\mu_2} \sim N(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2})$$

And then we have the Null hypothesis as $H_0 : \mu_1 = \mu_2$ and we can calculate the p value as before in the previous linear model.

Again regardless what the p-value is we have already assumed that the distribution of happiness index values differ. We can only calculate the difference of the values if we assume that there is a difference. Hence even with the happiness criteria we cannot conclude if the squared error would yield different results that the absolute error criteria without assuming that they do.

Sample frame title

It is good to go through the situations where it looks like that we are able to give support to the assumptions.

- 1.1. if $\hat{\beta} = 0 \rightarrow Y \neq \beta x + \epsilon$ is FALSE
- 1.2 what happens when $\hat{\beta}$ statistically significant but linearity and orthogonality are false
- 1.3. Ambigiousity of models if we say that we can estimate the correctness of the assumptions

1.1. if $\hat{\beta} = 0 \rightarrow Y \neq \beta x + \epsilon$ is FALSE

The statistical test used is T-test. It calculates the probability that the test measure is greater or equal than the test measures value in that data set when the Null hypothesis is true.

Null hypothesis for the j_{th}component is that $\beta_j = c$

$$T = \frac{\hat{\beta}_j - c}{S\sqrt{d_{jj}}},$$

where S^2 on unbiased estimator for variance σ^2 and d_{jj} is the element from the row j and column j from the matrix $(X_n^T X_n)^{-1}$. And when all the linear model assumptions are true then T follows T-distribution and it follows the the zero centered distribution when $\beta_j = 0$.

1.1. if $\hat{\beta} = 0 \rightarrow Y \neq \beta x + \epsilon$ is FALSE

$$T = \frac{\hat{\beta}_j - c}{S\sqrt{d_{jj}}},$$

where S^2 on unbiased estimator for variance σ^2 and d_{jj} is the element from the row j and column j from the matrix $(X_n^T X_n)^{-1}$.

We are not testing the assumptions but whether or not the value of the $\hat{\beta}$ is close to zero. All the other assumptions have to be true in the Null hypothesis because otherwise the $\hat{\beta}$ could not be used to estimate β and the test would not work.

Maybe more intuitive way of thinking is that because of our assumptions we know that X has a linear connection to Y and we are checking is the connection so close to zero that a random fluctuation could easily produce a similar change.

Sample frame title

- 1.2 what happens when $\hat{\beta}$ statistically significant but linearity and orthogonality are false

If X is not orthogonal to ϵ and X does not have a linear connection, but the observations are independent the $\hat{\beta}$ still converges to something. We can write

$$\hat{\beta} \to \bar{\beta} \ n \to \infty$$

Then we can alter the previous assumptions and write like this

- 1.1. Observations independent and from the same distribution
- 1.2. Finite variance of Y and X
- 1.3. $Y = \bar{\beta}x + \epsilon_2$
- 1.3 and 1.2 Orthogonality of X and ϵ_2 $(Cov(X, \epsilon_2) = 0)$

It is clear that the assumptions are different so we are not looking at the same thing as in the beginning so again we are not investigating the linearity $Y = \beta x + \epsilon$ but something else.