**Metodologie vědecké práce - statistika**

**Lubor Homolka**



**2018**

OBSAH

[OBSAH 2](#_Toc1030087)

[1 Analýza síly testu 3](#_Toc1030088)

[1.1 T-test – A priori analýza 3](#_Toc1030089)

[1.2 T-test – Kalibrační analýza 5](#_Toc1030090)

[2 Regresní ANALÝZA 7](#_Toc1030091)

[2.1 Log modely 7](#_Toc1030092)

[2.2 Regresní analýza – Logistická regrese 10](#_Toc1030093)

[2.2.1 Logistická regrese – model s nulovým členem 10](#_Toc1030094)

[2.2.2 Logistická regrese – model s kategoriální proměnnou 11](#_Toc1030095)

[2.2.3 Logistická regrese – model s metrickou proměnnou 13](#_Toc1030096)

[3 Testování statistických hypotéz 16](#_Toc1030097)

[3.1 Test středních hodnot – t-test 16](#_Toc1030098)

[3.2 Frekventistická a Bayesovská regresní analýza 18](#_Toc1030099)

[3.2.1 Frekventistický model 19](#_Toc1030100)

[3.2.2 Bayesův model 19](#_Toc1030101)

[4 analýza časových řad 21](#_Toc1030102)

[4.1 Korelace časových řad 21](#_Toc1030103)

[4.1.1 Problém nestacionárních časových řad 21](#_Toc1030104)

[4.1.2 Měření korelace mezi časovými řadami 26](#_Toc1030105)

[4.2 Stacionarita časových řad 27](#_Toc1030106)

[4.2.1 Deterministická (trendová) stacionarita 29](#_Toc1030107)

[4.2.2 Stochastická (unit root) stacionarita 30](#_Toc1030108)

[Seznam zdrojů 38](#_Toc1030109)

# 1 Analýza síly testu

## T-test – A priori analýza

Z literární rešerše vyplývá, že průměrná hodnota výdajů první zákaznické skupiny vydá na spotřebu v průměru 500 Kč (směrodatná odchylka 30 Kč). Průměrná hodnota výdajů druhé skupiny je 515 Kč. Druhá skupina je ovšem více heterogenní – výdaje kolísaly se směrodatnou odchylkou 40 Kč. Je možné prokázat rozdíl mezi skupinami (druhá skupina vydává více)? Kolik pozorování byste potřebovali, pokud je vyžadováno:

1. Síla testu 0.8
2. Tolerance k chybě I. Druhu je 0.05
3. Je možné použít *t-test*

**Řešení**: V první řadě je nutné odhadnout*effect size*. Tuto hodnotu je možné vypočítat v software GPower. Jako vstupní parametry zadáme očekáváné hodnoty průměrných hodnot a též hodnoty variability.



Obrázek 1 Výpočet effect size, Welschův *t-test. Zdroj: Vlastní*

Size effect je poměrně malý, pouze 0.424. Dle Cohenovy škály se tedy jedná o středně velký efekt.Ve druhém kroku tento odhad využijeme pro odhad velikosti vzorku.



Obrázek 2 Analýza síly testu. Zdroj: Vlastní

Vzhledem k tomu, že používáme analýzu síly testu před sběrem dat, **volíme apriorní analýzu** (červený rámeček v Obrázku 2). Zelený rámeček obsahuje požadované parametry citlivosti analýzy. Vzhledem k tomu, že chceme dokázat vyšší výdaje druhé skupiny, zvolíme *One tail* analýzu. Pokud bychom očekávali, že nebude možné získat odpovědi v obou skupinách zákazníků, změnili bychom i *Allocation ratio*. Modrý rámeček poskytuje výstupní informace o minimální hodnotě t-testu (1.656), ale hlavně o minimální velikosti vzorku. V tomto případě minimálně 70 respondentů v každé skupině. Této hodnoty musí být dosaženo náhodným výběrem. Jiný způsob, například *quota sampling*, by totiž nesplňoval podmínku reprezentativnosti. Pravděpodobnost výběru respondenta v každé skupině musí být stejný.

## T-test – Kalibrační analýza

Pokračujme s předchozím příkladem. Zjistili jste, že nejste schopni sesbírat více než 50 + 50 odpovědí. Proveďte citlivostní analýzu předcházejícího příkladu. Vzhledem k tomu, že hodnota Effect size vychází z předchozího výzkumu, není možné ji změnit. Jaké možnosti tedy nyní máte?

1. Snížení síly testu



Obrázek 3 Citlivost minimální velikosti vzorku na sílu testu . Zdroj: Vlastní

Obrázek 3 odkazuje na nelineární povahu vztahu. Velikost minimálního vzorku roste rychleji se zvyšující požadovanou hodnotou síly testu. Pokud by tým výzkumníků dokázal oslovit maximálně 100 respondentů, museli by se spokojit s úrovní síly testu 0.675.

1. Zvýšení hladiny významnosti

Obdobně jako v předchozí analýze síly testu, pokles úrovně hladiny významnosti není lineární. V případě dotázaných 100 respondentů a zafixování síly testu ve výši 0.8 by hladina významnosti vzrostla na 0.1. Pravděpodobnost toho, že by tým výzkumníků našel neexistující efekt by se tedy zvýšila nad běžně akceptovatelnou hodnotu 5%.



Obrázek 4 Citlivost minimální velikosti vzorku na hodnotu . Zdroj: Vlastní

Pokud by bylo možné získat pouze 100 respondentů, hladina významnosti by klesla na 10 %. To by znamenalo zvýšení pravděpodobnosti nálezu neexistujícího efektu.

# Regresní ANALÝZA

## Log modely

Příklad: Máme za úkol popsat elasticitu změny proměnné na změnu . Data je možné znázornit v rozptylovém grafu:

Obrázek 5 Rozptylový graf. Levé okno znázorňuje data na originální škále, pravé okno na logaritmické škále. Zdroj: Vlastní

Oblíbeným typem modelu, který analyzuje nerovnoměrně rozdělená data, je model s logaritmicky transformovanými proměnnými. Pro situaci výše by se mohl hodit například model:

Tento model je lineární v parametrech (mezi regresními koeficienty je aditivní vztah), proto je možné regresní parametry odhadnout metodou nejmenších čtverců.

Výsledek regresního modelu tak, jak ho poskytuje statistický software R.

Call:

lm(formula = lnY ~ lnX)

Model jsme odhadli standardním přístupem metody nejmenších čtverců. Do funkce jsme již vložili zlogaritmované hodnoty. Prokládáme tedy regresní funkci hodnotami, které jsou uvedeny v pravém okně Obrázku 5. Tento přístup je vhodný i z toho důvodu, že hlavní masa bodů je blíže sobě a nedochází tedy k velkému ovlivnění tzv. pákového efektu (leverage effect) hodnot vzdálenějších od hlavního datového shluku. Na druhé straně tato transformace „poškodí“ pozorování s vysokou hodnotou, protože logaritmická transformace není 1:1 transformací (jako transformace z cm na palce). Log 10 je 1, tzn. 10 % z původní hodnoty. Log 100 = 2, tzn. 2 % z původní hodnoty. Z tohoto důvodu se též někdy využívám jiného odhadu regresní funkce, a to vážené metody nejmenších čtverců. Interpretace je ovšem stejné bez ohledu na zvolenou techniku odhadu, proto budeme pokračovat s jednoduchým odhadem dle metody nejmenších čtverců.

Následující tabulka ukazuje popisnou statistiku reziduí. Je žádoucí, aby hodnoty měly nulovou střední hodnotu a nebyly zešikmené. Pokud by některý z těchto předpokladů byl porušen, potom by model systematicky nadhodnocoval, nebo podhodnocoval realitu.

Rezidua:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Min | 1Q | Median | 3Q | Max |
| -0.73647 | -0.15589 | -0.04212 | 0.16543 | 0.68556 |

Další tabulka obsahuje regresní koeficienty (Estimate), hodnotu standardní chyby (Std. Error), která měří to, s jakou „přesností“ byl odhadnut regresní koeficient, poměr t, který se stanoví jako podíl odhadu k chybě. Větší hodnota znamená, že podíl signal/noise je vysoký, a tedy námi identifikovaný efekt je prokazatelný. Poslední sloupec obsahuje p-hodnoty t-testu. Hodnoty menší než námi zvolená hladina alpha znamenají, že máme důkaz pro tvrzení obsažené v alternativní hypotéze. V tomto případě, že regresní koeficient populační regresní funkce je nenulový, a tedy že odpovídající proměnná je statisticky průkazná (ovlivňuje závisle proměnnou).

Koeficienty regresní funkce:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |  |
| (Intercept) | 2.7521 | 0.0424 | 64.903 | < 0.01 | \*\*\* |
| lnX | 0.4558 | 0.0511 | 8.913 | < 0.01 | \*\*\* |

Oba regresní parametry jsou tedy statisticky průkazné. Chyba odhadu prvního je téměř 65krát menší než samotný odhad.

Prvotním úkolem bylo interpretovat výsledky modelu jako míry elasticity. Výsledky regresních koeficientů je možné interpretovat vždy k danému procentnímu nárůstu /poklesu . Napříkad o 25 % povede ke zvýšení o , jak bylo ukázáno v přednáškových materiálech, což odpovídá nárůstu o 10.7 %. Nelinearita vztahu je znázorněna na dalším grafu.



Obrázek 6 Nelinearita efektu jednotkové změny. Zdroj: Vlastní

Ačkoliv lineární regresní funkce ( má tvar lineární funkce (), která má přímou interpretaci, interpretace se díky logaritmické transformaci zásadně změnila. Poslední částí výstupu R je seznam charakteristik kvality regresního modelu.

|  |  |
| --- | --- |
| Residual standard error: | 0.2976 on 48 degrees of freedom |
| Multiple R-squared: 0.6233 | Adjusted R-squared: 0.6155 |
| F-statistic: 79.44 on 1 and 48 DF | p-value: < 0.01 |

Model byl odhadnut dobře. F-test indikuje, že všechny proměnné v modelu kromě nulového členu jsou nenulové (v tomto případě, kdy máme pouze jednu nezávisle proměnnou a nulový člen, je ekvivalentní testovému kritériu t-testu^2). Model odhaduje přibližně 62% rozptylu hodnot závisle proměnné.

## Regresní analýza – Logistická regrese

V následujícím příkladu máme *vyrovnaný počet pozitivních a negativních výsledků*. Úlohou je odhadnout pravděpodobnost toho, že nastane pozitivní hodnota (pozitivní ve smyslu dobrý výsledek). Vysvětlující proměnné jsou dvě: dichotomní proměnná indikující, zda byla použita nový, nebo starý postup, a metrická proměnná x měřící např. tlak (standardizované hodnoty).



Obrázek 7 Vizualizace příslušnosti hodnot do třídy na základě členění do skupin a spojité proměnné x. Zdroj: Vlastní

### Logistická regrese – model s nulovým členem

Začněme jednoduchým případem logistického regresního modelu s nulovým členem, bez další vysvětlující proměnné. Vzhledem k tomu, že máme balancovaná data, informace obsažená v nulovém členu hovoří o podílu hodnot z pozitivní skupiny (y=1) na celkovém počtu pozorování. Vzhledem k tomu, že logistická regrese je případem funkce nelineární v parametrech, není možné použít standardní funkce využívající metody nejmenších čtverců pro odhad regresních koeficientů. Z tohoto důvodu využijeme funkci glm, která umožňuje odhad celé řady tzv. generalised linead models, zevšeobecněných modelů. Níže je uveden výpis ze software R:

|  |
| --- |
| Call: |
| glm(formula = y ~ 1, family = binomial(link = "logit"), data = dataReg) |

Regresní koeficient a odhad standardní chyby doplněný o inferenční test významnosti regresního parametru.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z-value | Pr(>|z|) |
| (Intercept) | 0.0000 | 0.2828 | 0 | 1 |

Regresní model není možné odhadnout bez chyby odhadu parametru, protože se hodnoty překrývají. Tzn., i když máme perfektní odhad 0.000, tento odhad není bezchybný.

Hodnota estimate je hodnotou a tedy nemá přímou interpretaci. Pro výpočet je tedy nutné použít logistickou regresní funkci . V tomto případě by tedy bylo výsledkem , což přesně odpovídá podílů pozitivních výsledků. Pokud by bylo více pozitivních výsledků, hodnota estimate by byla kladná.

### Logistická regrese – model s kategoriální proměnnou

V následujícím příkladu rozšíříme originální model o kategoriální proměnnou. Analyzujeme to, zda nová metoda dosahuje lepších výsledků (má více pozitivních hodnot). Z Obrázku 7 je zřejmé, že nová metoda označená zelenou barvou má proporcionálně více úspěšných pozorování (y=1), než stará metoda. Tento výsledek ale není 100%, protože existují i negativní výsledky při použití nové metody.

Z celkového počtu 50 měření bylo 30 dle nového postupu. Z těchto 30 měření bylo 23 hodnot pozitivních (76.7 %). Šance pozitivního výsledku tak je . U starého postupu jsme zjistili 2/20, šance je tedy . Šance, že při využitím nového postupu dostaneme pozitivní výsledek, je tak krát vyšší. Tomuto číslu říkáme poměr šancí (*odds ratio*).

Ve výstupech statistického software bývají zapsány šance pouze jednou hodnotu, byť by se mělo zapisovat jako 1:1 (tato šance bývá zapsána jako 1). Této situaci odpovídá situace, kdy první skupina (např. užití starého výrobního postupu) má stejnou šanci něco udělat (splnit normu), jako druhá skupina (nový výrobní postup). Pokud je šance 1.5, potom má firma, která využívá nový výrobní postup 1.5 vyšší šanci, že splní normu.

Interpretace šancí menší než 1 je složitější, protože je nutné „otočit“ její vliv. Pokud by nový výrobní postup byl horší než starý, potom by šance mohla být např. 0.8. Tomu by odpovídala interpretace: Firmy, které využívají starý přístup, mají 1.25 krát větší šanci splnit normu, než firmy využívající nový postup. K převodu je možné použít vztahu , v grafu:



Obrázek 8 Grafické znázornění převodu poměru šancí. Zdroj: Vlastní

|  |
| --- |
| Call: |
| glm(formula = y ~ group, family = binomial(link = "logit"), data = dataReg) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z-value | Pr(>|z|) |
| (Intercept) | -2.1972 | 0.7453 | -2.948 | <0.01 |
| groupNew | 3.3868 | 0.8613 | 3.932 | <0.01 |

Oba regresní koeficienty jsou statisticky průkazné (nejsou nulové). Stejně jako v předchozím příkladu nejsou výsledky přímo interpretovatelné. Vzhledem k tomu, že vysvětlující proměnnou je skupina New, tak interpretace nulového členu odkazuje na starý postup.

Hodnotu estimate nulového členu odpovídá: . Hodnota tedy obsahuje informaci o , kterou je možné získat inverzní funkcí k logaritmické funkci: .

Výpočet interpretovatelné hodnoty koeficientu groupNew je obdobná:. Interpretace této hodnoty, která se vztahuje k referenční skupině, je uvedena výše v textu.

Pokud by nás zajímala šance, při použití nového postupu dosáhneme pozitivního výsledku, postupujeme následovně: .

### Logistická regrese – model s metrickou proměnnou

|  |
| --- |
| Call: |
| glm(formula = y ~ x, family = binomial(link = "logit"), data = dataReg) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z-value | Pr(>|z|) |
| (Intercept) | -0.018 | 0.367 | -0.049 | 0.961 |
| x | 2.162 | 0.609 | 3.549 | <0.01 |

V případě, kdy hodnota je rovna 0, potom je pravděpodobnost toho, že nastane pozitivní výsledek . Se zvyšující se hodnotou roste i pravděpodobnost úspěchu. Obdobně jako v předchozím modelu když přírůstek (dříve: když skupina = New, nyní: když metrická proměnná nabývá hodnoty 1). Při výpočtu ale nesmíme opomenout hodnotu nulového členu, stejně jako v případě jednoduchého lineárního modelu.

Jaká je tedy pravděpodobnost toho, že nastane pozitivní výsledek, když je:



Z výše uvedených hodnot a z grafu níže si můžete ověřit, že vztah je nelineární. Není tedy možné interpretovat hodnotu 2.162 stejně jako u obecného lineárního modelu. Jednotková změna neodpovídá vždy stejné změně závisle proměnné.



Obrázek 9 Odhad pravděpodobnosti příslušnosti do první skupiny. Zdroj: Vlastní

### Ukázka využití logistické regrese ve vědeckém článku

V článku od Moomen at. Al (2018) autoři analyzují faktory, které mají vliv na nehodovost v kamionové dopravě v USA. Z abstraktu článku je možné vyčíst:

*“This paper investigated the factors influencing truck crashes on downgrades; an attempt to fill in some of the research gaps. An empirical analysis of factors affecting truck crashes on two-lane downgrade roadways in Wyoming was carried out using a binary logistic regression technique.”*

První odstavec článku popisuje metody sběru a zdroje data. Autoři se ve své analýze omezili na stát Wyoming a silnice se specifickým profilem

*“The prevalence of mountainous terrain and mountain passes in western United States means states like Wyoming have relatively high incidences of downgrade truck crash occurrence.”*

Studovaný fenomén – nehodovost kamionů – je analyzován ve specifickém kontextu. Je tedy otázkou, jaká je zevšeobecnitelnost poznatků, a zda tedy má smysl využívat inferenční statistiku.

Autoři použili logistickou regresní analýzu z následujícího důvodu:

*“The binary nature of the dependent variable (truck crashes or otherwise) makes logistic regression suitable for the analysis. Mathematically, the logistic regression model is flexible and intuitive which results in meaningful interpretations (Hosmer and Lemeshow, 2000).”*

V časti Metod autoři zdůrazňují rozdíly v technikách odhadu regresní funkce:

*“Unlike linear regression which uses the method of least squares to derive parameter estimates, logistic regression estimates parameters by the maximum likelihood method. The maximum likelihood estimation method yields parameter estimates which maximize the probability of obtaining the values observed in a set of data.”*

První odstavec by mohl být považován za nadbytečný, protože poukazuje na všeobecně známý poznatek. Navíc, využití metody maximální věrohodnosti je obecně považováno za korektní postup, který je implicitně implantován ve většině statistických programů. Diskuse o metodách odhadů je možné vynechat, pokud nedochází k nějaké specifické úpravě, případně pokud by datová omezení mohla představovat problém pro věrohodný odhad (typicky „populární“ GMM přístup na malých vzorcích dat).

Autoři dále pokračují s matematickým popisem regresních modelů. Vyzdvihují skutečnosti, že:

*“Like any other regression model, logistic regression modeling involves assessment of the significance of variables in the model. In logistic regression, significance testing is achieved with the Wald test. The Wald test statistic is defined as beta/SE. The hypothesis of Wald test H0:beta=0 states that the probability of success is independent of x. Though the Wald test is adequate for larger samples, the more powerful likelihood-ratio test is preferred for sample sizes often used in practice (Agresti, 2007).”*

V literatuře bylo vymezeno několik typů inferenčních testů. Vzhledem k podstatě testu a druhu dat, není možné použít standardní t-test, jak je tomu v případě jednoduché regresní analýzy. Navrhovaný Waldův test využívá vypočítanou věrohodnot modelu v případě plně identifikovaného modelu. Plně identifikovaný model využívá všechny proměnné. Poté byl stanoven omezený (restricted) model bez proměnné, která je předmětem inferenčního testování. Srovnání věrohodností podílem rozhodne, zda začlenění proměnné statisticky průkazně zlepšilo vysvětlující schopnost modelu.

Výběr proměnných byl proveden jak na základě datově-analytického přístupu, tak expertního odhadu. Jednalo se zejména o redukci počtu proměnných z důvodu nestability odhadu.

*“A subset of twenty-three explanatory variables was selected from the CARE database for this study. The subset of selected variables was done with a view to build the most parsimonious model that explains the occurrence of truck crashes on downgrades. The rationale for minimizing the number of variables in the model was to produce a numerically stable model which can be easily generalized.”*

Tento přístup je mírně kontroverzní, protože přímo vede k problému s vynechanou proměnnou (omitted variable). Existuje několik analytických metod, které umožní pracovat s velkým počtem vysvětlujících proměnných. V prvním kroku je důležité se podívat na korelační matici proměnných. Pokud existuje více korelovaných proměnných, faktorová analýza umožní identifikovat společné faktory. Ty je poté možné vyjádřit pomocí numerických skór (např. Bartlettova skóre). Další možností je použít tzv. shrinkage přístupů, jako je lasso nebo ridge regrese.

*“The subset of variables was classified into six categories: driver characteristics, environmental factors, temporal factors, crash characteristics, traffic characteristics, and geometric features. Driver age and gender were noted for all vehicles involved in the crashes. Environmental factors which included weather, lighting and road condition were of important interest (Table 1).*

*Driver age was categorized into three groups to reflect the proportion of truck drivers involved in crashes based on results of the LTCCS (Federal Motor Carrier Safety Administration (FMCSA), 2006). These groupings are young aged drivers (25 years or younger), middle aged drivers (26-66 years), and old aged drivers (66 years and above).”*

Všechny proměnné byly přímo pozorovatelné, tudíž nebylo nutné ověřovat validitu a reliabilitu proměnných, jak je tomu v případě analýzy latentních proměnných.

Autoři dále pokračovali s popisnou statistickou v sekci Kalibrace modelu a výsledky. Zařazení tématu kalibrace obvykle bývá v sekci metod. Autoři začali kapitolu následovně:

*“A binary logistic regression model was fitted to the data. The variable selection method enables the user to specify the way in which the independent variables are entered into the model. In the stepwise selection procedure, statistically insignificant variables are removed from the model before adding a significant variable. The addition or removal of each independent variable is listed as a separate step and at each step a new model is fitted. The procedure is stopped when no more independent variables can be added or removed from the stepwise model. The variables retained in the model form the final model.”*

Tento text vysvětluje proces výběru proměnných, která již ovšem byla nastíněna v úvodu. Výběr proměnných pomocí dopředného výběru (stepwise) byla použita. Tento přístup je běžně využíván, i když obvykle bývá doplněn o zpětnou (backwards) metodu, za účelem identifikování klíčových proměnných. Jeho nevýhodou je ovšem selektivnost (nemusí existovat jednoznačné řešení) a výpočetní náročnost, zejména v případě velkého počtu proměnných. Jedná se totiž o kombinatorický problém spojený s testováním statistických hypotéz o prokazatelném zlepšení/zhoršení kvality modelu.

Další odstavec popisuje indikátory kvality modelu:

*“The Hosmere Lemeshow test was used as a standard test for goodness-of-fit of the logistic regression. The prediction ability of the model was evaluated using the area under the receiver operating characteristic (ROC) curve (Hosmer and Lemeshow, 2000).”*

Waldův test byl popsán v předchozí sekci. Celkový test, který se ve své podstatě liší od F-testu z klasické lineární regresní analýzy, je dále představen. Další indikátory, které vycházejí z prediktivní schopnosti modelu a nikoli z inferenčního statistického rámce jsou uvedeny na koci sekce. Následuje interpretace výsledků:

*“The negative coefficient of weather (b= -0.6412) indicates that the probability of being involved in a truck crash on downgrades in clear weather is lower compared to adverse weather conditions. The odds ratio suggests the probability of a truck crash occurring on a downgrade in clear conditions is about 47% lower compared to adverse conditions.”*

Interpretace efektu počasí je se odvíjí od negativního znaménka regresního koeficientu. Ten znamená, že v případě dobrého počasí (x=1) dochází k poklesu pravděpodobnosti nehody. Naopak je tomu v případě dobrých světelných podmínek. Překvapivě, dobré světelné podmínky dle poznatků autorů spíše vedou k nehodovosti. Šance, že kamion bude účastníkem nehody je o 61 % vyšší v případě dobrých světelných podmínek.

*“The results suggest a higher probability of truck crashes is associated with lighted compared to unlighted conditions (b = 0.4755). The odds ratio of 1.609 indicates that the risk of a downgrade truck crash is 61% higher in lighted conditions compared to unlighted conditions.”*

Tento článek nevyužil žádnou metrickou proměnnou.

# Testování statistických hypotéz

Na následujícím příkladu jsou demonstrovány tři přístupy k testování statistických hypotéz. Výzkumník začínající s výzkumem chce zjistit, zda existuje rozdíl mezi dvěma typy manažerů v souvislosti s jejich postoji k určitému tématu, měřené metrickou proměnnou. Mějte referenční skupinu manažerů „Type A“ a kontrastní skupinu „Type B“. Provedeme šetření mezi 100 manažery a získáme skóre (Value) znázorněné v boxplotu níže:



Obrázek 10 Boxploty dvou skupin znázorňující průměrné a mediánové hodnoty. Zdroj: Vlastní

Horizontální čárkovaná linka odkazuje na průměrnou hodnotu, plná linka na medián. Otázka zní, zda manažerský typ A se ve svém postoji liší od typu B. Popisná statistika vzorku říká, že průměrná hodnota Type A je 5.03, u Type B je to 5.35. Směrodatná odchylka hodnot byla 0.926, resp. 0.905.

## Test středních hodnot – t-test

Základní t-test spoléhá na stejné hodnoty rozptylu analyzovaných skupin. Tuto podmínku je nutné testovat před provedením testu (např. pomocí F testu). Druhou možností je provést Welchův t-test s *pooled* rozptylem.

|  |  |
| --- | --- |
| **Welch Two Sample t-test** | |
| data: data$typeB and data$typeA | |
| t = 2.2496, df = 97.951, p-value = 0.02671 | |
| alternative hypothesis: true difference in means is not equal to 0 | |
| 95 percent confidence interval: | |
| 0.049 0.775 | |
| sample estimates: | |
| mean of x | mean of y |
| 5.446 | 5.034 |

Z výsledků vyplývá, že existuje na hladině významnosti 5 % statisticky průkazný rozdíl mezi typy manažerů. Výsledky mohou být statisticky významné i u malých efektů, a to zejména z důvodu velkého počtu pozorování. Výzkumník by považoval za zajímavé, kdyby skutečný efekt byl rozdíl alespoň bodu oproti druhému manažerskému typu.

Tento problém je možné vyřešit aplikací dvou jednostranných testů (two-one sided tests, TOST) za účelem zjištění ekvivalence požadovaného minimálního effect size (smallest effect size of interest, SESOI) a skutečného effect size, který je ovšem pouze možný odhadnout. V našem případě je SESOI .



Obrázek 11 Srovnání intervalu spolehlivosti statistického testu a tzv. TOST intervalu. Zdroj: Vlastní

Předchozí obrázek znázorňuje bodový odhad efektu, dva překrývající se intervaly spolehlivosti a ekvivalenční meze. Interval spolehlivosti označený tenkou čarou je interval sloužící k testování nulové hypotézy, která tvrdí, že skutečný efekt je roven 0. Na základě tohoto intervalu zamítáme nulovou hypotézu a předpokládáme, že rozdíl je pozitivní, tedy manažerský typ B dosahuje vyšších hodnot. Druhý (kratší) interval je TOST interval. Vzhledem k tomu, že tento interval, který je konstruován na základě dvou jednostranných intervalů (1. test testuje, zda je efekt větší než 0, druhý že je nižší než 0), zasahuje do regionu SESOI, potom nemůžeme tvrdit, že efekt je dostatečně velký, aby mělo smysl se tímto výzkumem zabývat. Výše uvedená ilustrace je příkladem toho, že statisticky významný výsledek může být z věcného pohledu nevýznamný.

## Frekventistická a Bayesovská regresní analýza

Uvažujme stejná data. Výzkumník ale chce prokázat, že rozdíl je větší než 0.2.

Dvouvýběrový t-test je možné zapsat jako lineární model s dichotomní nezávislou proměnnou. Testován je poté odpovídající koeficient.

### Frekventistický model

V klasickém (frekventistickém) regresním přístupu je proveden t-test regresního koeficientu, případně stanoven interval spolehlivosti.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |  |
| (Intercept) | 5.0344 | 0.1295 | 38.88 | <0.01 | \*\*\* |
| TypeB | 0.4120 | 0.1831 | 2.25 | 0.0267 | \* |

Interval spolehlivosti pro koeficient TypeB je přibližně . Interpretace tohoto koeficientu není intuitivní: tento interval je jedním z mnoha intervalů, z nichž 5 % obsahuje skutečnou hodnotu efektu. Test, který ověřuje, zda je efekt vyšší než 0.2, je nutné provést individuálně. Zde je výstup z R vycházející ze srovnání dvou regresních modelů (saturated - restricted).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Resid. Df | RSS | Sum of SQ | F | p-val |
| Model 1 | 99 | 83.300 |  |  |  |
| Model 2 | 98 | 82.176 | 1.1237 | 1.34 | 0.2498 |

Vzhledem k tomu, že p-hodnota je vyšší než 5%, nezamítáme H0 a nemáme důkaz, že je hodnota jiná než 0.2.

### Bayesův model

V Bayesovském přístupu je vypočítán *credible* interval (který buď obsahuje testovanou nulovou hodnotu, nebo ne). Dále je možné vypočítat tzv. Bayesův faktor, který srovnává plochu posteriorního rozdělení koeficientu před testovanou a za testovanou hodnotou.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Est.Error | l-95% CI | u-95% CI |
| (Intercept) | 5.03 | 0.13 | 4.78 | 5.28 |
| TypeB | 0.42 | 0.18 | 0.06 | 0.77 |

Odhady regresních koeficientů jsou téměř identické s frekventistickém modelem. Na rozdíl od předchozího modelu ale není uvedena hodnota t-testu a p-value. Namísto toho je uveden dolní a horní mez intervalu. S pravděpodobností 0.95 je skutečná hodnota v tomto intervalu. Formálně je možné testovat, zda je hodnota efektu vyšší než 0.2:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Hypothesis | Estimate | Est.Error | CI.Lower | Evid.Ratio |
| (TypeB)-(0.2) > 0 | 0.22 | 0.18 | -0.08 | 7.3 |

Nulová hypotéza je říká, že efekt je menší než 0.2, alternativní (předmět naší analýzy), že je efekt vyšší. Je přibližně 7.3 krát pravděpodobnější že alternativní hypotéza je pravdivá, než nulová hypotéza.

# analýza časových řad

Analýza časových řad je odvětvím statistické analýzy která zahrnuje celou řadu pod-disciplín. Během analýzy časových řad obvykle využíváme přístupy vizualizace, rozkladu řady do krátkodobých, dlouhodobých a periodických složek, identifikace strukturálních změn, tvorbě predikcí, testování kauzality či významnosti proměnných. Existuje několik přístupů k analýze časových řad. Některé z nich je možné převzít z klasické analýzy cross-sectional dat. Lineární regresní analýza s časovým indexem jako s vysvětlující proměnnou je jedním z nich. Další přístupy jsou jedinečné pro analýzu časových řad, protože využívají autokorelovanosti hodnot. Tato kapitola tedy nemůže být systematickým přístupem k analýze časových řad, protože neexistuje obecný a doporučovaný přístup k analýze. Namísto toho tato je tato kapitola věnována dílčím problémům, které je obvykle nutné řešit před samotnou analýzou časových řad.

## Korelace časových řad

Korelace dvou a více proměnných je častým zájmem analytiků a výzkumníků. Korelace je ovšem koncept, který navzdory své zdánlivé jednoduchosti výpočtu a interpretace na cross-sectional datech, trpí celou řadou problémů, když je aplikovaná na časové řady. Korelační koeficient může být dobrým indikátorem asociace mezi stacionárními časovými řadami. Dalším problémem je testování statistických hypotéz. Těmto problémům jsou věnovány následující kapitoly.

### Problém nestacionárních časových řad

Problematiku nestacionarity časových řad si demonstrujme pomocí empirického příkladu v software R. Pomocí kódu níže generujme 100 hodnot z každé náhodné veličiny, která je rozdělena dle normálního rozdělení.

|  |
| --- |
| set.seed(123) # zajistí reprodukovatelnost výstupu |
| randomData1 <- rnorm(100) # první náhodná řada 100 hodnot |
| randomData2 <- rnorm(100) # druhá náhodná řada 100 hodnot |

Následující obrázek znázorňuje průběh dvou řad. Vzhledem k tomu, že data byla generována náhodně, neočekáváme a ani nevidíme žádnou společnou vazbu mezi daty. Dle obrázku lze soudit že korelace mezi řadami je velmi malá.



Obrázek 12 Vývoj dvou časových řad. Zdroj: Vlastní zpracování

To je možné ověřit formálně pomocí Pearsonova korelačního koeficientu. P-hodnota testu je ovšem stanovena špatně, jak je vysvětleno v následující kapitole. Samotná hodnota korelačního koeficientu je stanovena korektně.

|  |
| --- |
| cor.test(randomData1, randomData2) |
| Pearson's product-moment correlation  data: randomData1 and randomData2  t = -0.49095, df = 98, p-value = 0.6246  alternative hypothesis: true correlation is not equal to 0  95 percent confidence interval:  -0.2435805 0.1483291  sample estimates:  cor  -0.04953215 |

Korelační koeficient je odhadnut ve výši -0.0495 (prakticky žádná korelace mezi řadami). Nyní uvažujme vstup třetí proměnné, která ovlivní stejnoměrně obě proměnné. Efekt této proměnné se projeví rostoucím trendem:

|  |
| --- |
| ts0 <- 1 + 0.1\*seq(100) # modrá přímka, společný trend  ts1 <- 1 + 0.1\*seq(100) + randomData1 # modrá + random1data = červená  ts2 <- 1 + 0.1\*seq(100) + randomData2 # modrá + random2data = černá |

Předchozí kód generoval hodnoty na základě jednoduchého regresního modelu. Všechny proměnné je nyní možné znázornit graficky:



Obrázek 13 Nové časové řady se společným trendem. Zdroj: Vlastní zpracování

Pokud bychom nyní provedli stejný test jako v minulém případě, obdrželi bychom následující výsledky:

|  |
| --- |
| cor.test(ts1, ts2) |
| Pearson's product-moment correlation  data: ts1 and ts2  t = 21.085, df = 98, p-value < 2.2e-16  alternative hypothesis: true correlation is not equal to 0  95 percent confidence interval:  0.8620513 0.9353196  sample estimates:  cor  0.9051989 |

Nyní jsme obdrželi velmi silnou korelaci mezi proměnnými, o kterých ale víme, že jsou generovány nezávisle na sobě. Jedná se o velmi běžný problém tzv. zdánlivé korelace/regrese. Tento problém může vyvstat během analýzy nestacionárních dat. Tomuto problému je ovšem možné předejít několika způsoby. Hlavní myšlenkou je odstranění společných vzorů, jako je například dlouhodobý trend, cyklické výkyvy, případně sezónnost. Jako výchozí bod analýzy je obvykle provádění odstranění trendu metodami jednoduché regresní analýzy. Ne vždy ovšem tento přístup pomůže, a proto je nutné zvolit další formu transformace. V tomto kontextu se často používá tzv. pre-whiteningu časových řad (slovo whitening odkazuje ma white noise reziduální složky). Prewhitening je proces, ve kterém analytický model (nejčastěji autoregresivní AR model) je vytvořen na první časové řadě. Poté je aplikován na druhou řadu, čímž se odstraní společný vzor. Rezidua z obou modelů jsou poté analyzovány za účelem zjištění skutečné korelace.

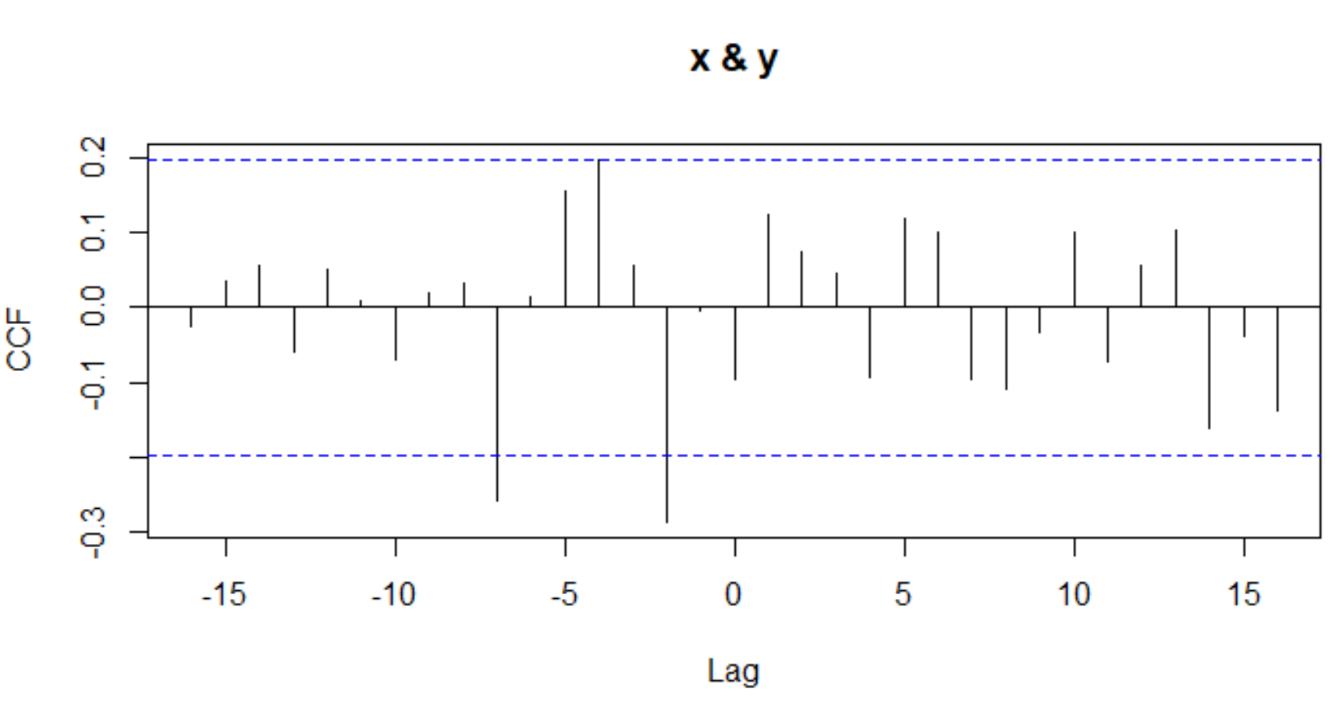
Pokračujme s předchozím příkladem:

|  |
| --- |
| library(forecast) # umožní použití pokročilých metod  library(tsa) |

Kód v následujícím okně umožní odhadnout parametry autoregresivního modelu na základě minimalizace BIC. Identifikovaný model má následující vlastnosti:

|  |
| --- |
| selectARIMA <- auto.arima(ts1, seasonal=FALSE) # identifikuje model  summary(selectARIMA) # vypíše detaily o modelu  Series: ts1  ARIMA(2,1,0) with drift  Coefficients:  ar1 ar2 drift  -0.6834 -0.4862 0.0977  s.e. 0.0891 0.0898 0.0474  sigma^2 estimated as 1.063: log likelihood=-142.38  AIC=292.77 AICc=293.19 BIC=303.15  Training set error measures:  ME RMSE MAE MPE MAPE MASE  Training set 0.008660619 1.010388 0.764617 -6.988041 20.19891 0.7634066  ACF1  Training set -0.06707982 |

ARIMA(2,1,0) model s driftem znamená, že časová řada není stacionární, ale její řád je 1. Diferencování časové řady ji převede do stacionární podoby. Současná hodnota je vysvětlena zpožděnými hodnotami až dvě období dozadu. Tento model je poté aplikován na druhou časovou řadu. Korelace mezi řadami je možné zachytit pomocí cross-correlation grafu.



Obrázek 14 Graf Cross correlation dvou časových řad. Tato analýza na očištěných datech neukazuje na významnou korelaci v okamžiku změny (lag 0). Zdroj: Vlastní zpracování

Z obrázku 14 vyplývá, že korelace hodnot ve stejný čas je mírně záporná, ale statisticky neprůkazná. Tato analýza chybně identifikovala korelaci mezi časovými řadami ve smyslu zpožděného efektu o 3, resp. 7 období.

### Měření korelace mezi časovými řadami

I když je stanovená hodnota Pearsonova korelačního koeficientu v pořádku, inferenční test testující významnost nadhodnocuje skutečnost. Pearsonova statistika je totiž navržena tak, aby analyzovala na sobě nezávislá pozorování, jak je tomu v cross-sectional datech. Tento předpoklad je ovšem porušen v kontextu časových řad, kde se jedná o sekvenci hodnot stejného jevu seřazeného chronologicky. Samotný test se poté domnívá, že má k dispozici více pozorování, než ve skutečnosti má. Efektivní velikost vzorku je nižší.

Korelace časových řad se obvykle neprovádí pouze s ohledem na jeden okamžik v čase. Je možné že dochází ke zpožděnému efektu reakce. Z tohoto důvodu se korelace znázorňuje pomocí tzv. cross-correlations, jak je vidět např. na obrázku 14. Hodnota 0 na ose x značí situaci korelaci ve stejném časovém období. Hodnota Lag 1 odkazuje na korelaci mezi současnou hodnotou první časové řady a hodnotou druhé řady posunutou o jedno období do budoucna. Pokud bychom uvažovali vztah dvou proměnných zachycených v časové řadě: finanční stimul, například formou zvýšených FDI a míru nezaměstnanosti v určitém měsíci. Záporná hodnota korelace s lagem 2 znamená, že existuje statisticky průkazný efekt snížení nezaměstnanosti, ke kterému dochází 2 měsíce po zvýšení FDI.

## Stacionarita časových řad

Časová řada, stejně jako průřezová cross-sectional data může být popsána nástroji popisné statistiky. Mezi základní nástroje popisné statistiky patří tzv. momenty hodnot. Prvním momentem je průměrná hodnota, druhým rozptyl, následuje šikmost a špičatost a další méně užívané míry. Pouze některé z nich ovšem mají správnou interpretační schopnost v případě nestacionárních dat. Příkladem momentu, který má smysl analyzovat pouze na stacionárních datech je rozptyl. Míra rozptylu značí míru fluktuace hodnot kolem střední hodnoty. Je možné říci, že čím větší je rozptyl, tím větší je neurčitost v datech, protože hodnoty kolísají. Budoucí hodnoty časové řady, kterou je charakteristická vysokou variabilitou, je obtížné predikovat. Podívejte se na Obrázky 12 a 13. Hodnoty ve druhém grafu jsou totožné jako v prvním, jen jsou upraveny o deterministický trend. Predikce budoucí hodnoty je v obou případech stejně obtížná. Přesto, směrodatná odchylka ve druhém grafu je přibližně třikrát vyšší než v grafu prvním.

Důležitou podmínkou nutnou k tomu, aby výsledky byly korektní, je tzv. Stacionarita časové řady. Výše zvolený příklad s rozptylem je pouze jedním z mnoha důvodů. Nestacionární časové řady například mají tu vlastnost, že důsledek šoku se v čase stále prohlubuje/zhoršuje. Z makroekonomie například známe efekty multiplikátorů, které předpokládají že zvýšení výdajů (např. stavba nové dálnice) má největší vliv na GDP v daném období. Ve druhém období jsou peníze opět použity domácnostmi, které je dostaly formou mezd. Tyto peníze ale nejsou použity celé a část odpovídající meznímu sklonu k úsporám se nevrátí do ekonomiky. Nestacionární časová řady odporuje této úvaze. Pokud má časová řada konstantní 1) průměrnou hodnotu, 2) směrodatnou odchylku a 3) kovarianci mezi dvěma vybranýma hodnota vzdálenými libovolnou výši zpoždění, potom hovoříme o tzv. slabé stacionaritě. Silná stacionarita vyžaduje, aby všechny momenty časové řady byly konstantní napříč celou historií řady. Rozeznáváme dva druhy stacionarity:

1. Deterministická stacionarita
2. Stochastická stacionarita (unit root)

Obecným přístupem využívaným v analýze časových řad je práce s transformovanými hodnotami časové řady, nikoliv hodnotami na původní škále (tzv. level data), pokud jsou tato data nestacionární. Pokud odstranění trendové složky, například proložením dat metodou jednoduché regresní analýzy) zajistí stacionární rezidua, hovoříme o deterministické stacionaritě. Bohužel, tento jednoduchý přístup v ekonomických časových řadách většinově nefunguje. Pokud transformace pomocí diferencí hodnot dosáhneme stacionarity, potom hovoříme o stochastické (unit root stationarity). V případě, že stacionarity lze dosáhnout provedením pouze prvních diferencí, hovoříme o integrované časové řadě prvního řádu. Tato řada se obyčejně zapisuje jako I(1), nebo přímo ve specifikaci modelu, např. AR**I**MA(0,1,0). V některých případech je nutné diferencovat diferencovanou časovou (atd…) řadu za účelem dosažení stacionarity. V tom případě hovoříme o I(k) kde k je počet diferencí nutných l zajištění stacionarity. Samotný název unit root vychází z matematické terminologie a odkazuje na požadavek, aby kořen charakteristické rovnice byl menší než 1. Pokud je roven 1, případně vyšší, časová řada není stacionární. Typicky by model popisující tuto časovou řadu vypadal následovně: a model časové řady by rostl do nekonečna velmi rychle. Těmto časovým řadám se říká tzv. explozivní časové řady. Přímou implikací tohoto modelu je skutečnost, že starší hodnoty mají výrazně vyšší vliv na současné hodnoty, než nejnovější pozorování.

V následujících podkapitolách budou představeny způsoby, jak zjistit, zda je časová řada stacionární, deterministicky stacionární, nebo stochasticky stacionární.

### Deterministická (trendová) stacionarita

Jak již bylo zmíněno výše, tento typ stacionarity není v ekonomických aplikacích běžný. Přesto je zajímavé testovat, zda by odstranění trendu nevyřešilo problém s nestacionaritou. Diferencování je zdánlivě jednodušší proces odstranění trendu, nicméně při něm dochází ke ztrátě jednoho pozorování v případě I(1). První hodnotu totiž není možné diferencovat. K testování deterministické stacionarity byl navržen KPSS test. Testujme tedy časovou řadu z Obrázku 12. První možností KPSS testu je testování, zda hodnoty kolísají kolem průměrné hodnoty (trend je vyjádřen jako ):

|  |
| --- |
| > summary(ur.kpss(ts1, type=”mu”))  #######################  # KPSS Unit Root Test #  #######################  Test is of type: mu with 4 lags.  Value of test-statistic is: 2.0569  Critical value for a significance level of:  10pct 5pct 2.5pct 1pct  critical values 0.347 0.463 0.574 0.739 |

Protože testová statisticka 2.0569 je větší než kritická hodnota 0.463 zamítáme nulovou hypotézu o tom, že časová řada je stacionární. Odečtením průměrné hodnoty od všech hodnot nezaručí, že časová řada bude stacionární. V následujícím kroku proveďme tzv. KPSS-tau test:

|  |
| --- |
| > summary(ur.kpss(ts1, type = c("tau")) )  #######################  # KPSS Unit Root Test #  #######################  Test is of type: tau with 4 lags.  Value of test-statistic is: 0.0591  Critical value for a significance level of:  10pct 5pct 2.5pct 1pct  critical values 0.119 0.146 0.176 0.216 |

Nyní je testová statistika 0.0591 nižší než kritická hodnota 0.146. Pokud odstraníme trend ve tvaru , transformovaná časová řada bude stacionární. Nepodařilo se nám totiž zamítnout nulovou hypotézu, která tvrdí, že časová řada je stacionární.

### Stochastická (unit root) stacionarita

Stochastickou stacionaritu je možné testovat pomocí několika testů. Testem s největší flexibilitou i sílou testu je Augmented Dickey-Fuller test (ADF test). Tento test umožňuje kombinací tří nastavení:

1. Test unit root
2. Test unit root a driftu (přítomnost konstantního zvýšení)
3. Test unit root a zároveň deterministického trendu

Funkcionalitu testu si ověříme na uměle generované časové řadě, která obsahuje unit root a deterministický trend.

|  |
| --- |
| set.seed(123)  err <- rnorm(100) # generuje rezidua dle normálního rozdělení  trend <- seq(100) # časový index od 1 do 100  tsURtrend <- rep(0, 100) # objekt, do kterého se uloží generované hodnoty  for(i in seq(2, 100)){ # iterativně dopočítá hodnoty  tsURtrend[i] <- 0.5\*trend[i] + tsURtrend[i-1] + err[i]  } |

Tato časová řada má následující průběh.

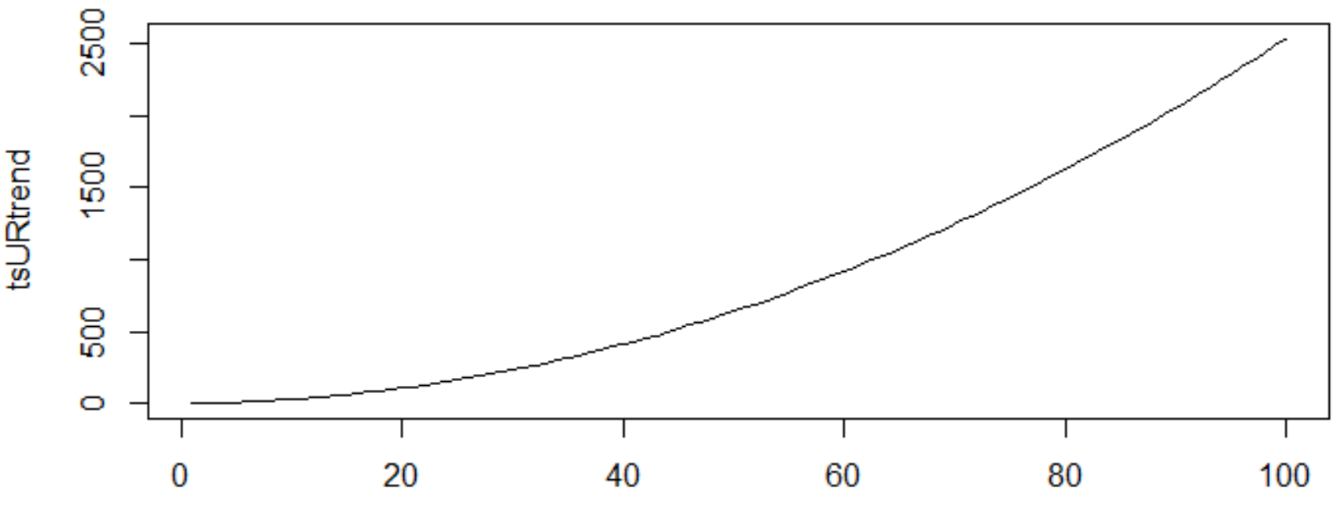


Image 15 Časová řada obsahující jak unit root, tak deterministický trend. Zdroj: Vlastní

Již z grafu je patrné, že časová řada není stacionární. Typ stacionarity ale není možné odhadnout. Pomůže proložení trendovou funkcí odstranit nestacionaritu?

Aplikujme KPSS-tau test

|  |
| --- |
| > summary(ur.kpss(tsURtrend, type = c("tau")) )  #######################  # KPSS Unit Root Test #  #######################  Test is of type: tau with 4 lags.  Value of test-statistic is: 0.5172  Critical value for a significance level of:  10pct 5pct 2.5pct 1pct  critical values 0.119 0.146 0.176 0.216 |

Testová statistika je větší než kritická hodnota. To znamená, že i po odstranění trednu časová řada nebude stacionární. Musíme tedy provést jinou transformaci. Provedeme sérii ADF testů, abychom zjistili, co přesně způsobuje nestacionaritu.

Začněme jednoduchým ADF testem testující pouze přítomnost unit-rootu (type=“none“).

|  |
| --- |
| summary(ur.df(y=tsURtrend, type = "none"))  ###############################################  # Augmented Dickey-Fuller Test Unit Root Test #  ###############################################  Test regression none  Call:  lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)  Residuals:  Min 1Q Median 3Q Max  -3.0761 -0.7504 -0.0163 0.8614 3.5547  Coefficients:  Estimate Std. Error t value Pr(>|t|)  **z.lag.1 -0.0003800** 0.0004923 -0.772 0.442  z.diff.lag 1.0276416 0.0191604 53.634 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1.352 on 96 degrees of freedom  Multiple R-squared: 0.9979, Adjusted R-squared: 0.9979  F-statistic: 2.331e+04 on 2 and 96 DF, p-value: < 2.2e-16  Value of test-statistic is: -0.7718  Critical values for test statistics:  1pct 5pct 10pct  tau1 -2.6 -1.95 -1.61 |

Parametr z.lag.1 který se rovná -0.0003800 je centrem našeho zájmu. Pokud se populační parametr tohoto parametru (v literatuře běžně označovaný jako gamma) rovná nule, potom časová řada obsahuje unit root. Bohužel, vzhledem k povaze nestacionárních dat není možné využít odhadu standardních chyb a tudíž i stanovených p-value. Z toho důvodu o platnosti nulové hypotézy musíme rozhodnout na základě kritických hodnot, které jsou uvedeny na posledním řádku výstupu z programu R. Je nutné srovnat testové kritérium -0.7718 k odpovídajícímu testovému kritériu ve výši -1.95. Nulová hypotéza tvrdí, že gamma=0, tedy že data nejsou stacionární, protože obsahují unit root. Všimněte si, že kritické hodnoty se jsou orientovány k mínus nekonečnu, jak klesá požadovaná hodnota alpha, na které je test prováděn. Vzhledem k tomu, že testové kritérium nespadá do kritického oboru, nezamítáme nulovou hypotézu. Nemůžeme tedy tvrdit, že máme důkaz o tom, že je časová řada unit-root stacionární. První diference tedy nevyřeší problém. Z KPSS testu víme, že samotné odstranění trendu též nevyřeší nestacionaritu. Rozšiřme tedy analýzu o začlenění konstantního přírůstku, tzv. driftu (type = "drift").

|  |
| --- |
| > summary(ur.df(y=tsURtrend, type = "drift"))  ###############################################  # Augmented Dickey-Fuller Test Unit Root Test #  ###############################################  Test regression drift  Call:  lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)  Residuals:  Min 1Q Median 3Q Max  -2.7225 -0.8184 -0.1263 0.7330 3.5545  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 1.1575954 0.4189646 2.763 0.00688 \*\*  z.lag.1 0.0011866 0.0007404 1.603 0.11233  z.diff.lag 0.9338435 0.0386763 24.145 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1.307 on 95 degrees of freedom  Multiple R-squared: 0.9918, Adjusted R-squared: 0.9917  F-statistic: 5761 on 2 and 95 DF, p-value: < 2.2e-16  Value of test-statistic is: 1.6026 4.1355  Critical values for test statistics:  1pct 5pct 10pct  tau2 -3.51 -2.89 -2.58  phi1 6.70 4.71 3.86 |

Výsledkem testu již není pouhý jeden řádek, protože testujeme unit root a drift. První statistika se nazývá tau2 a je analogická prvnímu případu. Protože hodnota 1.6026 nespadá do kritického oboru, nezamítáme hypotézu o unit-root. Druhá statistika testuje společnou hypotézu že po odstranění unit root diferencováním a driftu dosáhneme stacionarity. Testová statistika se nazývá phi1 a testuje, zda gamma = drift = 0. Vzhledem k tomu, že testová statistika 4.1355 nepřekročila kritickou hodnotu 4.71, nemůžeme tuto hypotézu také zamítnout. V tuto chvíli nám již zbývá poslední možnost, a to kombinace unit root a deterministického trendu.

|  |
| --- |
| > summary(ur.df(y=tsURtrend, type = "trend"))  ###############################################  # Augmented Dickey-Fuller Test Unit Root Test #  ###############################################  Test regression trend  Call:  lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)  Residuals:  Min 1Q Median 3Q Max  -2.42464 -0.60909 -0.00321 0.69992 2.13650  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 0.7947716 0.2981214 2.666 0.00904 \*\*  z.lag.1 0.0006863 0.0005253 1.306 0.19459  tt 0.5130963 0.0522186 9.826 4.31e-16 \*\*\*  z.diff.lag -0.0568259 0.1044551 -0.544 0.58771  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.923 on 94 degrees of freedom  Multiple R-squared: 0.996, Adjusted R-squared: 0.9958  F-statistic: 7736 on 3 and 94 DF, p-value: < 2.2e-16  Value of test-statistic is: 1.3064 37.7129 50.8503  Critical values for test statistics:  1pct 5pct 10pct  tau3 -4.04 -3.45 -3.15  phi2 6.50 4.88 4.16  phi3 8.73 6.49 5.47 |

Tau 3 má stejnou interpretaci jakou předchozí tau. Phi 2 testuje, zda gamma = drift = trend = 0. Můžeme tuto hypotézu zamítnout, protože testové kritérium 37.7129 > 4.88. Víme tedy, že alespoň jedna z výše uvedených složek způsobuje nestacionaritu. Podívejme se tedy na poslední test phi3, který testuje gamma = trend = 0. Opět zamítáme hypotézu (50.8503 > 5.47) a vidíme, že alespoň jeden z uvedených důvodů způsobuje stacionaritu. Vzhledem k tomu, že samostatný KPSS test ani jednoduchý ADF test (tau1) neodhalili jako jednoznačnou příčinu deterministický trend, respektive unit root, problémem musí být vzájemná kombinace těchto faktorů. Odstranění trendu a diferencování časové řady vyřeší problém s nestacionaritou.

# Seznam zdrojů

Lakens, D., Scheel A. M. & Isager P. M. Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science* [online]. 2018, **1**(2), 259-269 DOI: 10.1177/2515245918770963.

Moomen, M., Rezapour, M., & Ksaibati, K. (2018). An investigation of influential factors of downgrade truck crashes: A logistic regression approach. *Journal of Traffic and Transportation Engineering (English Edition)*. [DOI> 10.1016/j.jtte.2018.03.005](https://doi.org/10.1016/j.jtte.2018.03.005)