Indítsuk el a *Statist.exe* programot!

A gyakorlatokhoz használt adatfile-ok a filek.zip könyvtárban vannak.

*2\*click* Statist.exe icon (a program bejelentkezésekor a képernyőn egy adatfile-t látunk).

Nyissuk meg ebben az alkönyvtárban a pulse00.sta adatfile-t:

*click* File>Open és a Windows-ban megszokott módon keressük meg a gyakorlatok számára létrehozott könyvtárban a kívánt file-t.

Nézzük végig a menüket (vezesse rájuk a cursor-t egyenként, erre legördülnek: File, Edit, View, Insert, Format, Statistics, Graphs, Tools, Data, Window, Help).

A menüpontok jelentős része már jól ismert a Windows-alkalmazásokból. Legfontosabb lesz számunkra a Statistics menü, ez tartalmazza a rendelkezésre álló statisztikai eljárásokat, a “Graphs” pedig a választható diagram típusokat.

A Statistica 6 program az elemzések eredményeit kétféle formában kezelheti: „Workbook” (Munkafüzet) vagy „Individual Window” (külön-külön munkalap). A felhasználó dönti el, hogy melyik üzemmódban használja a programot. Ebben a segédletben az elemzések eredményeit külön-külön kezeltük, ezért a bonyodalmak elkerülése érdekében ellenőrizzük, hogy a gyakorlaton futtatott program is ebben az üzemmódban működik-e.

*Tools>Option>Output Manager*

Ellenőrizzük, hogy az ablak bal fölső részében található „Individual Window” felirat előtti kis körben látható-e egy fekete pont! Amennyiben nem látható, akkor *clikk* a körre, vagy az *Individual Window* feliratra. (Vigyázat, a *clikk* váltja az üzemmódot, tehát a második *clikk* kikapcsol, és így tovább …)

Térjünk vissza a beolvasott adatokhoz (PULSE00.STA).

**Példa**

A hallgatók a tanteremben egyszerű kísérletet végeztek: mindenki megmérte a pulzusát (PULSE1), majd pénzfeldobással eldöntötték, hogy ki fut helyben 1 percig (RAN: 1 igen, 2 nem), futott, akinek kellett, utána ismét mindenki megmérte a pulzusát (PULSE2). Néhány egyéb adatot is rögzítettek: dohányzik-e (SMOKES: 1 igen, 2 nem), nem (SEX: 1 férfi, 2 nô), testmagasság (HEIGHT), testsúly (WEIGHT), életmód mozgásgazdagsága (ACTIVITY: 1 gyenge, 2 közepes, 3 jelentôs).

Készítsük el a PULSE1 változó Box-Whisker Plot ábráját.

Az ábrát legegyszerűbben az egér jobb oldali gombjával megnyitható menük (flying menu) segítségével készíthetjük el. Hajtsuk végre az alábbiakat:

* Kattintsunk a PULSE1 változó valamelyik cellájára (bal gombbal): *click*, majd a jobb gombbal: *jobb click*.

Megnyílik a leggyakoribb feladatok listáját tartalmazó menü.

* A menü harmadik sora: *Graphs of Input Data*. Mutassunk rá a cursor-ral!

A megnyíló ablakban a fontosabb diagramok listáját látjuk.

* A harmadik sor: *Box-Whisker PULSE1*. Mutassunk rá!

Négyféle módon készíthetjük el az ábrát, ezeket kínálja fel választásra a megnyíló menü.

* Válasszuk az elsôt: *Median-/Quart./Range*. *click* (a bal gombbal).

**Kérdés**: Mit lát az ábrán? Mekkora a medián, az első és harmadik kvartilis (a 25% ill. 75% valószínűséghez tartozó érték), a legkisebb és legnagyobb érték? Szimmetrikus az eloszlás?

Most diagramot (scatterplot) készítünk a testsúly (weight) és a magasság (height) összefüggéséről, és ki is számítjuk a két változó közötti korrelációs együtthatót.

* Statistics>Basic Statistics/Tables
* *click* Correlation matrices *click* OK (erre megjelenik a Product-Moment and Partial Correlations feliratú párbeszéd-ablak, ahol megadhatjuk, mely változókra akarunk számításokat végezni). A Quick fülön maradunk.
* *click* One variable list ....
* Kijelöljük a két változót. A height változóra kattintunk, majd a bal gombot lenyomva tartva lehúzzuk a weight változóra. Alul az ablakban megjelenik a kijelölt változók sorszáma (6-7). *click* OK
* *click* Summary:Correlation matrix

**Kérdés**: Mennyi a testsúly és testmagasság közötti korrelációs együttható? Mit mond ennek alapján, szignifikáns a korreláció?

Vegyük észre, hogy a Product-Moment and ... feliratú párbeszéd-ablak kicsinyített (ikon) formában a képernyő aljára került.

* *click* az ikonra, így visszakapjuk a párbeszéd-ablakot az eredeti méretben.

Most készítsük el az ábrát, ehhez váltsunk át a Quick melletti *Advanced/plot* fülre!

* *click* 2D scatterplots(bal oldalt, középen)*.* Megjelenik egy párbeszéd-ablak a változók kiválasztásához.

vízszintes tengely (horizontal): *click* height;

függôleges tengely (vertical): *click* weight. *click* OK.

**Kérdés**: Mit lát az ábrán? Milyen az összefüggés a testsúly és testmagasság között? Lehet azt mondani, hogy a magasabb ember biztosan súlyosabb is? Mekkora a 180 pnd súlyú emberek magasságának becsült értéke?

**Mennyiségi adatok statisztikai jellemzőinek számítása, grafikus megjelenítése**

* *Statistics>Basic Statistics/Tables*

*click* Descriptive Statistics *click* OK, erre megjelenik egy párbeszéd-ablak, melyben a Variables gombra kattintva megadhatjuk, hogy melyik változóra kívánunk elemzést végezni (legyen ez a PULSE1). Maradjunk a Quick fülön, és kattintsunk a Summary: Descriptive Statistics gombra!

Az *Advanced* fülre áttérve sok egyéb statisztikai mutatót is kérhetünk (pl. medián, ferdeség stb.)

Készítsünk hisztogramot a PULSE1 oszlop adataiból. Ha az elôírás szerint haladtunk eddig, a képernyő alján látható a Descriptive Statistics párbeszéd-ablak lekicsinyítve. Kattintsunk rá, visszakapjuk az eredeti alakban.

* A *Histograms* gombra kattintva megkapjuk a hisztogramot.

**Kérdés**: Mekkora az átlag és a szórás? Mekkora a gyakorisága a mintában a 60 és 70 közötti pulzusszámnak? Milyennek gondolja az eloszlást?

Figyeljük meg és értelmezzük az intervallumok beosztását, változtassuk azt meg: Az ábra megjelenésekor a képernyő aljára került Desciptive Statistics-ra kattintva (*Desciptive Statistics*) ismét megjelenik a párbeszéd ablak. Itt térjünk át a *Normality* fülre, majd a Histograms gomb alatt a Categorization címszó adta lehetőségek közül az ablakban az osztályok számát (Number of intervals) 10-rôl növeljük 20-ra (a bal gombbal a nyílra kattintva eggyel, a jobb gombbal kattintva tízzel változtatható a szám).

Készítsünk “torta” diagramot az ACTIVITY változó adatainak bemutatására!

 (*Graphs > 2D Graphs > Pie Charts)*, maradjunk a Quick fülön

Variables: *ACTIVITY;*

Graph type Pie Chart - Counts

Váltsunk át az Advanced fülre

Pie Legend: *Text and value;*

Shape: Ellipse (kattintsunk a mellette lévő körbe)

*>OK*

**Kérdés**: Hányszor fordul elő az 1, a 2 ill. a 3 aktivitás?

**Egy változó, több csoport**

Készítsünk box-plot-ot a WEIGHT oszlopra, a SEX szerinti csoportosításban (a testsúly adatok a weight oszlopban vannak, a sex változó azonos sorbeli értéke adja meg, hogy férfi vagy nő testsúlyáról van éppen szó).

*Graphs> 2D Graphs>Box Plots* megnyílik a 2D Box Plots párbeszéd-ablak (itt állítjuk be a változókat), maradjunk a Quick fülön.

*Variables:*

Dependent variable: Weight (egyszerre több is lehetne)

Grouping variable: SEX

*>OK*

**Kérdés**: Mit mondhatunk a testsúly eloszlásáról a nőknél és a férfiaknál?

**Kétmintás t-próba**

Hasonlítsuk össze a WEIGHT változónak a két nem szerinti csoportokra vonatkozó várható értékeit, elôször, hogy egyenlôk-e

A két egymástól független minta adatai két módon adhatók meg a program számára,

* “t-test, independent, by groups”: minden mintaelem értéke azonos oszlopban (változóban) van, egy másik változó vele egy sorban lévő cellájában pedig megadjuk a minta kódját (szám vagy betű). Az adatok megadásához használt egyik változó tehát magukat a mintaelemeket, a másik változó pedig a csoportosításhoz szükséges kódokat tartalmazza. Az adatok sorrendje tetszőleges, de természetesen azonos mindkét változóban. (Erre az adatmegadási módra látunk példát a pulse00.sta adatállományban, ahol a pulse1, pulse2, weight és height mintaadatokhoz tartoznak a csoportosításhoz szükséges kódokat tartalmazó sex, ran, smoke és activity változók.)
* “t-test, independent, by variables”: a két minta adatai külön oszlopban (változóban) szerepelnek, a változók hossza (az adatokat tartalmazó cellák száma) a feladat természetéből következően különböző lehet.

Most a két minta egyetlen oszlopban van, az adatok egy része az egyik mintához, másik része a másik mintához tartozik a SEX változó értéke szerint.:

 *Statistics>Basic Statistics/Tables>t-test, independent, by groups>OK*

*Variables*. A megjelenô dialog boxban Dependent variables: WEIGHT. Grouping variable: SEX. Használjuk a SEX változó mindkét értékét (automatikus).

* A *Quick* fülön: *Summary:* *T-test*.

**Kérdés**: Mennyi a nők és a férfiak átlagos testsúlya a mintában? Mekkora a nők és a férfiak testsúlyának szórása? Mi a p jelentése? Elfogadja a nullhipotézist? Mi itt a nullhipotézis? Mi az F-próba nullhipotézise?

 A megjelenő scrollsheet első felében a t-próba eredményeit látjuk: a SEX=1-hez és a SEX=2-höz tartozó minták átlaga (mean), a különbségükre vonatkozó próbastatisztika értéke (t-value) és annak meghaladási valószínűsége (p-value) ami itt nagyon kicsi, <<0,05 (az elsőfajú hiba elkövetésének valószínűsége, ha a próbastatisztika értéke jelöli ki az elfogadási/elutasítási tartomány határát).

 A táblázat további oszlopaiban az F-próba eredményeit találjuk: a SEX=1-hez (Group 1) és SEX=2-höz (Group 2) tartozó mintaelemek számát, a korrigált tapasztalati szórás értékeit, valamint az F próbastatisztikát és annak meghaladási valószínűségét (p variancs). Az F próbastatisztika a két csoport tapasztalati szórásnégyzetének hányadosa (a táblázatbeli Std.Dev értékek négyzetének hányadosa), értéke 1.942291, a hozzá tartozó p érték 0.04046, ami azt jelenti, hogy a nullhipotézis érvényessége (azaz a varianciák azonossága) esetén ekkora valószínűséggel kaphatunk ilyen, vagy ennél nagyobb próbastatisztika értéket. Esetünkben p<0.05, tehát a nullhipotézist elutasítjuk.

Mivel az F-statisztikai próba alapján el kell utasítanunk azt a feltételezést, hogy a két minta azonos varianciájú sokaságból származik, ismételjük meg a statisztikai próbát, de most ne feltételezzük a varianciák azonosságát!

* *click* a képernyő alján található “T-Test for Indep…” ikonra.

Az *Options* fülön: *click* a “*t-test with separate variance estimates*” előtti négyzeten (vagyis nem feltételezzük a varianciák azonosságát). *Summary* (jobb fölső sarokban). Ekkor a következő számítási eredményt kapjuk (az utolsó számjegyeket néhány helyen elhagytuk, ez nem befolyásolja a levonható következtetést):

A táblázatban a hetedik, nyolcadik és kilencedik oszlop a varianciák különbözőségének figyelembevételével számított próbastatisztika (t separ.var. est.), a hozzá tartozó szabadsági fokok száma (df) és a meghaladási valószínűség (p 2-sided). Vegyük észre, hogy df nem egész szám. Az eredmények alapján a nullhipotézist (a várhatóértékek egyezését) elutasítjuk. A táblázat többi oszlopa azonos az előző táblázatban található eredményekkel.

1. Hasonlítsuk össze a PULSE1 változónak a két nem szerinti csoportokra (SEX) vonatkozó várható értékeit.

**Kérdés**: Mennyi a nők és a férfiak átlagos pulzusa az első mérésnél? Szignifikáns-e a különbség a nők és a férfiak pulzusának várható értéke között? Miért van a pulzus-adatoknak egyáltalán eloszlása?

1. Hasonlítsuk össze a PULSE2 változónak a helybenfutás szerinti csoportokra (RAN) vonatkozó várható értékeit, és ugyanezt a PULSE1 változóra. (Vegyük észre, hogy egyidejűleg vizsgálhatjuk a PULSE1 és PULSE2 változókat.) Készítsük el a box-plot-ot is. (A *Quick* vagy az *Advanced* fülön bal oldalt *click* Box & wishker plot ).

**Kérdés**: Van-e jelentős különbség a PULSE2 változó várható értékében ill. varian­ciájában aszerint, hogy futottak-e vagy nem?

Vessük össze az ábrákat és a t-próbák eredményeit! A pulse1 és pulse2 változók ábrájának az y tengelyén az ábrázolási határok eltérőek!

**Páros t-próba**

**Példa**

Vizsgáljuk meg, hogy a futás okoz-e változást a pulzusban. természetesen csak azoknál a hallgatóknál van értelme a vizsgálatot elvégezni, akik futottak (Select cases: run=1).

**Kérdés**: Változik-e a pulzus a futás hatására? Mi a szignifikancia-szint? Miért jött ki más p érték, mint a kétmintás t-próbánál?

**Egyváltozós lineáris regresszió**

**Példa: egyváltozós lineáris függvény illesztése**

Illesszünk egyenest az alábbi táblázatban található mérési adatokra. A független változó (x) értéke pontosan beállítható, a függő változó (y) értéke azonban a méréskor elkövetett hiba (ε) miatt különbözik a valódi értéktől ( ).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sorszám** | **x** | **y** | **Sorszám** | **x** | **y** |
| 1 | 1.500 | 2.651 | 14 | 3.200 | 4.920 |
| 2 | 3.000 | 4.682 | 15 | 4.700 | 6.567 |
| 3 | 4.000 | 5.972 | 16 | 6.200 | 9.112 |
| 4 | 5.000 | 7.253 | 17 | 7.700 | 10.766 |
| 5 | 6.000 | 8.509 | 18 | 9.200 | 13.082 |
| 6 | 7.000 | 9.262 | 19 | 10.700 | 14.853 |
| 7 | 8.000 | 11.068 | 20 | 12.200 | 16.190 |
| 8 | 9.000 | 12.101 | 21 | 13.700 | 17.597 |
| 9 | 10.000 | 13.248 | 22 | 9.800 | 13.112 |
| 10 | 11.000 | 14.018 | 23 | 10.100 | 13.236 |
| 11 | 12.000 | 16.009 | 24 | 10.400 | 14.336 |
| 12 | 13.000 | 17.463 | 25 | 10.700 | 13.952 |
| 13 | 1.700 | 3.365 |  |  |  |

A táblázat adatai a *linreg1.sta* file-ban találhatók.

Hívjuk be a STATISTICA program Multiple Regression modulját.

 *(Statistics >Multiple Regression>OK*)

 Megjelenik a Multiple Linear Regression párbeszéd-ablak. Maradjunk a *Quick* fülön! A jobb oldalon található “*Open Data*” gombra kattintva megnyitjuk a linreg1.sta adatállományt, amelyben a fenti táblázat adatai találhatók és kijelöljük a változókat.

 *(Open Data>linreg1.sta>OK>Variables> Dependent: y*

 *Independent: x*

 *OK)*

Megjelenik a "Multiple Regression Results" ablak, melynek fejlécében megtaláljuk a regresszióra vonatkozó legfontosabb információkat (a továbbiakban részletesen megismerkedünk jelentésükkel).

 A *Quick* fülön lévő egyetlen gombra (*Summary: Regression results*) kattintva az illesztett függvény konstansainak összesítő táblázatát kapjuk:

A "B" oszlopban találjuk a tengelymetszet (Intercpt sorában) és az iránytangens (X sorában) becsült értékét, szórásukat pedig az “St. Err. of B” oszlopban. A t(23) oszlopban a 23 szabadsági fokhoz tartozó próbastatisztika értéket találjuk (nullhipotézis: B valódi értéke zérus). A táblázat utolsó oszlopában található *p*-level érték az elsőfajú hiba elkövetésének valószínűsége. Az oszlop első cellájában található *p*=0.000002 annak valószínűsége, hogy a nullhipotézis érvényessége esetén (vagyis ha a tengelymetszet valódi értéke nulla, azaz a valódi egyenes átmegy az origón) 1.017212 legyen a becsült tengelymetszet értéke. A fentiekhez hasonlóan a második sorban található p=0.000000 érték azt jelenti, hogy ha *x* együtthatójának (az iránytangensnek) valódi értéke nulla (ez a nullhipotézis), *p* < 5·10-7 annak valószínűsége, hogy B=1.240713 (vagy annál nagyobb) eredményt kapjunk. Mivel *p* értéke a tengelymetszetre és az iránytangensre egyaránt rendkívül kicsi, a próba szerint mindkettô értéke szignifikánsan eltér nullától.

**Kérdés:** Írja föl az illesztett egyenes egyenletét! Mi a táblázatban lévő t-próbáknál a nullhipotézis?

 A táblázat első oszlopában látható “BETA” az összes változó (*x* és *y*) standard transzformációja  utáni regresszió esetén az együttható értéke. Egyetlen független változó esetén BETA =R (korrelációs koefficiens).

 megjelenik egy másik táblázat is, a neve Summary Statistics, ez az illeszkedés jóságára és a regresszió szignifikanciájára vonatkozó, a Regression Summary ablakban már látott mutatók táblázatos összefoglalása.

Az ebben a táblázatban és a “Regression Summary” ablak fejlécében található R2 érték megadja, hogy az illesztett függvény a teljes négyzetösszeg (az *y* értékek változása) hányadrészére ad magyarázatot. Négyzetgyöke a szintén itt található R korrelációs együttható. Az “Adjusted R2” érték számításánál az adatokra illesztett függvény paraméterei miatti szabadságifok-csökkenést is figyelembe veszik, használata különösen akkor indokolt, ha kevés adathoz sokparaméteres függvény illesztünk. F(1,23) annak a próbastatisztikának a kiszámított értéke, amellyel azt vizsgáljuk, hogy van-e egyáltalán összefüggés a függő és a független változó(k) között, a nullhipotézis az, hogy nincs. A *p* érték ehhez a statisztikai próbához tartozik. Itt *p* kicsi, elutasítjuk tehát a nullhipotézist (hogy nincs kapcsolat). A Std. Err. of Estimate (standard error of estimate) a reziduális szórást, a mérési adatok illesztett függvénytől (itt az egyenestől) való eltérésének szórását jelöli.

Nézzük most meg, milyen további információhoz jutunk, ha nem csak a legegyszerűbb eredményeket kérjük, váltsunk át az Advanced fülre! Az "*ANOVA (Overall goodness of fit*)" gombra kattintva kapjuk az ANOVA táblázatot:

Az ANOVA táblázatban a teljes négyzetösszeget (Sum of Squares a Total sorában) két részre bontják: a Regress. sorában található érték az illesztett függvény által leírt  eltérésekbôl számított, míg a függvény által nem "magyarázott" rész (a maradék) a Residual sorban található . Df a szabadsági fokok száma, Mean Squares a szórásnégyzet, F pedig a próbastatisztika értéke, amely a nullhipotézis érvényessége esetén F-eloszlású. A nullhipotézis az, hogy nincs függvénykapcsolat a függô változó és a független változó(k) között. A táblázat utolsó oszlopában megadott p érték annak valószínűségét adja, hogy ha a nullhipotézis igaz lenne, ilyen eredményt kapnánk.

A táblázatban található eredmény alapján megállapíthatjuk, hogy *y* és *x* között van függvénykapcsolat, de nem tudhatjuk, hogy az illesztett lineáris függvény megfelelô-e.

Az ezen a lapon (Multiple Regression Results, Advanced fül) látható többi gombok a többváltozós regresszióhoz hasznosak.

**Kérdés:** Miért kisebb R2adj az R2-nél? Mindig kisebb? Mitől függ, hogy mennyivel különböznek?

* A regresszió alapfeltételezéseinek teljesülését (az  hibák egymástól függetlenek, varianciájuk konstans, az Y és x közötti "valódi" függvénykapcsolat lineáris) ellenôrizni kell. Erre nyílik lehetôség, ha a "Multiple Regression Results" ablakon a "Residuals/assumptions/predictions" fülre váltunk át. Maga az ablak a szokásos módon a képernyő aljára került ikon-formában, rákattintva kapjuk az eredeti alakjában.
* A megnyíló ablak bal felső részén kattintsunk a "*Perform residual analysis*" gombra! Nagyon sok fül közül választhatunk, vegyük először a Quick feliratút! Ezen belül csak két lehetőségünk van, a reziduumok táblázatának előállítása “Summary: Residuals & predicted” és a "*Normal Plot of resids*" , válasszuk az utóbbit! Ezzel a reziduumok normális eloszlását ellenôrizhetjük. Ha a reziduumok normális eloszlást követnek, a pontok az ábrán piros színnel berajzolt egyenes körül helyezkednek el, nem találunk kiugró pontot vagy szisztematikus eltérést. Esetünkben elfogadhatjuk, hogy a reziduumok normális eloszlást követnek.
* Váltsunk most át az Advanced feliratú fülre, ezen belül is kattintsunk a "Durbin-Watson statistics" gombra. A próba annak ellenôrzésére szolgál, hogy a mérés során elkövetett hibák egymástól függetlenek-e (nincs-e a hibáknak "menete").

A “Serial correlation coefficiens” szintén a hibák függetlenségét méri (a mérés sorrendje és a reziduum között van-e függvénykapcsolat). A táblázatban látható d=1.658811 próbastatisztika-érték nagyobb a függelék táblázatában n=25, α=0.05 és K=1-hez tartozó kritikus értéknél (d>dU=1.34), tehát nem utasítjuk el azt a nullhipotézist, hogy a reziduumok egymástól függetlenek.

* Ha az alkalmazott lineáris modell adekvát (a valódi függvény lineáris), a reziduumok eloszlásának típusa, várható értéke és varianciája megegyezik a mérési hibáéval. Ha a hibák normális eloszlásúak, használhatjuk a normális eloszlás feltételezésén alapuló statisztikai próbákat (t-próba, F-próba). A reziduumok eloszlását grafikusan vizsgálhatjuk. Készítsük el a gyakorisági hisztogramot: (A Residuals fülön *Histograms* *of residuals*). A hisztogram alakja megfelel a normális eloszlásnak.
* A reziduumokat a becsült y értékek függvényében is vizsgálhatjuk (Scatterplots fül: *Predicted vs.residuals*). Ha a mérési hiba varianciája konstans, a reziduumok véletlenszerűen szórnak zérus körül, függetlenül y mért értékétől. Ha pl. a variancia y értékével növekednék, a reziduumok sávja szélesedne a nagyobb y értékek irányában. Ha nem megfelelő függvényt illesztettünk (a lineáris modell nem adekvát), a reziduumoknak menete, görbülete van. Itt az ábrán nem tapasztalunk rendellenességet.
* Készítsük el az y - x ábrát az illesztett egyenessel és a 95%-os konfidencia sávval. Erre nyújt lehetőséget a Scatterplot fülön a "*Bivariate correlation*" gomb, ahol a megnyíló ablakban több változó esetén a változók listájáról kiválaszthatjuk azt a két változót, amelyek közötti korrelációt vizsgálni kívánjuk. Jelenleg egyetlen független változónk van, nincs választási lehetôségünk, független változónak x-et, függônek pedig y-t választjuk. (Hasonló ábrát készíthetünk a főmenü Graph-ból választható 2D Scatterplot feladattal is.)

**Kérdés:** Olvassa le a becsült y értékét x=4.5-nél! Milyen tartományban van az igazi Y ugyanitt?

*x*=10 értékhez adjuk meg azt az intervallumot, amelyben a valódi *Y* érték 95%-os valószínűséggel megtalálható (konfidencia intervallum).

* Menjünk vissza a regresszió elvégzésekor kapott párbeszéd-ablak "*Residuals/assumptions/predictions*" fülére! Alul középen látjuk a Predict dependent variable gombot, alatta a konfidencia-intervallum (Compute confidence limits) ill. a jóslási intervallum (Compute prediction limits) közötti választás lehetősége és az elsőfajú hiba (Alpha) beállítására szolgáló ablak. Válasszuk a *Compute confidence limits* feladatot, *alpha=0.05* legyen, majd *Predict dependent variable.* A program megkérdezi, hogy milyen x értékhez kívánjuk elvégezni a számítást, írjuk be a 10-et!

A megjelenő táblázat utolsó oszlopában az egyenlet iránytangense és tengelymetszete után található a becsült (számított) *y* érték 13.42435, valamint a konfidencia-intervallum alsó - és fölső határa (-95.0%CL ill. +95.0%CL) pedig 13.26791 és 13.58079.

Legyünk kíváncsiak arra, hogy ha az *x*=10 beállításnál újabb mérést végeznénk, a mért y érték 95%-os valószínűséggel mekkora intervallumban fordulhatna elő (jóslás).

* Az előző feladathoz hasonlóan végezzük a számítást, de a Predict dependent variable gomb alatta a *Compute prediction limits* előtti mezőre kattintunk, majd *Predict dependent vaiable.*

A megjelenő táblázatban a számított *y* érték alatt most a jóslási tartomány alsó és fölső határát adja meg a program: 12.72741 és 14.12129. Vegyük észre, hogy a jóslási intervallum szélesebb a konfidencia-intervallumnál.

**Példa: hogyan vesszük észre, ha nem alkalmas a lineáris függvény**

Illesszünk egyenest az alábbi táblázatban található mérési adatokra!

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sorszám** | **x** | **y** | **Sorszám** | **x** | **y** |
| 1 | 1.500 | 6.715 | 14 | 3.200 | 25.470 |
| 2 | 3.000 | 22.793 | 15 | 4.700 | 50.021 |
| 3 | 4.000 | 38.223 | 16 | 6.200 | 87.609 |
| 4 | 5.000 | 57.614 | 17 | 7.700 | 130.190 |
| 5 | 6.000 | 80.889 | 18 | 9.200 | 184.749 |
| 6 | 7.000 | 105.906 | 19 | 10.700 | 245.855 |
| 7 | 8.000 | 139.657 | 20 | 12.200 | 314.008 |
| 8 | 9.000 | 173.927 | 21 | 13.700 | 391.479 |
| 9 | 10.000 | 212.717 | 22 | 9.800 | 205.059 |
| 10 | 11.000 | 253.806 | 23 | 10.100 | 216.246 |
| 11 | 12.000 | 304.392 | 24 | 10.400 | 232.183 |
| 12 | 13.000 | 356.560 | 25 | 10.700 | 241.801 |
| 13 | 1.700 | 10.337 |  |  |  |

A táblázat adatai a *linpol2.sta* file-ban találhatók.

Hívjuk be a STATISTICA program Multiple Linear Regression modulját.

 *(Statistics >Multiple Regression>OK*)

 Megjelenik a Multiple Linear Regression párbeszéd-ablak. Megnyitjuk a linpol2.sta adatállományt, amelyben a fenti táblázat adatai találhatók.

 *(Open Data>linpol2.sta>OK>)*

 Az előbbi példa alapján végezzük el az illesztést, és a reziduumok ábráján figyeljük meg az eredményt. (Ne feledkezzen el a reziduumok grafikus elemzéséről!)

**Kérdés:** Hogyan dönt, adekvát-e a függvény? Mit mond a Gauss-háló ábrája (normal probability plot)?

**Polinom illesztése**

A másodfokú polinom nemlineáris függvény, de az x2 tag beiktatásával többváltozós lineáris regresszióként kezelhető. Hívjuk be a STATISTICA program Fixed Nonlinear Regression modulját.

*(Statistics >Advanced Linear/Nonlinear Models> Fixed Nonlinear Regression >OK*)

* A megnyíló select vars. for... párbeszéd-ablakban kijelöljük azokat a változókat, amelyekre az elemzést végezzük: *Variables>X és YPOL>OK*

A Nonlinear Components Regression ablakba jutunk. A felkínált függvénytípusok közül kiválasztjuk a négyzetes tagot (*click x2.>OK*)

 A “Model Definition” ablakban, a Quick fülön maradva kijelöljük a változókat (vegyük észre, hogy a program mindegyik kijelölt változóra – tehát a függő és független változóra egyaránt – felkínálja a választott transzformációs függvényt):

 *Variables> Independent: X és v2^2* (mivel X a második oszlopban található)

 *Dependent: YPOL*

 *OK>OK*

Megjelenik a "Multiple Regression Results" ablak, melynek fejlécében megtaláljuk a regresszióra vonatkozó legfontosabb információkat. Ezután hajtsuk végre az elemzést, beleértve a reziduumok vizsgálatát. Azt találjuk, hogy a másodfokú polinom alkalmas az adatok leírására.