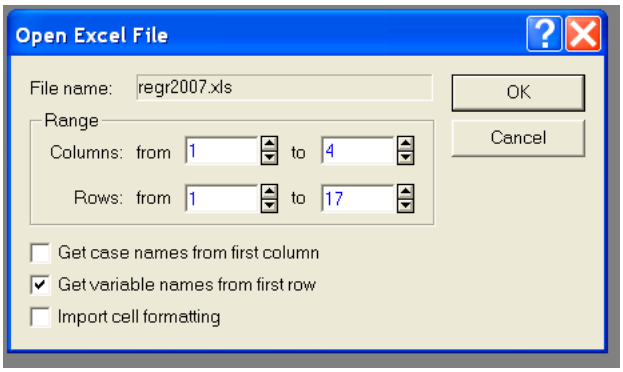
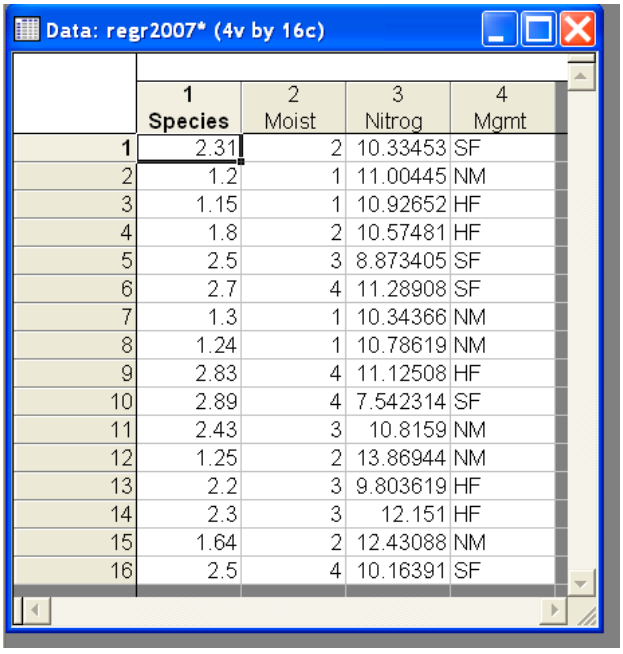
# Klasická lineární regrese

Pracujeme se souborem *regr2007.xls* (na síti učebny je ve folderu CANOCO[1] nebo ke stažení jako *http://regent.jcu.cz/regr2007.xls*). Soubor ma dva listy, my budeme pracovat s prvnim (*Regr*), tak si ho naimportujte do programu Statistica.

Nezapomeňte vybrat správný list a zaškrtnout, že v prvém řádku jsou názvy proměnných:




Výsledná naimportovaná data by měla vypadat zhruba takto:

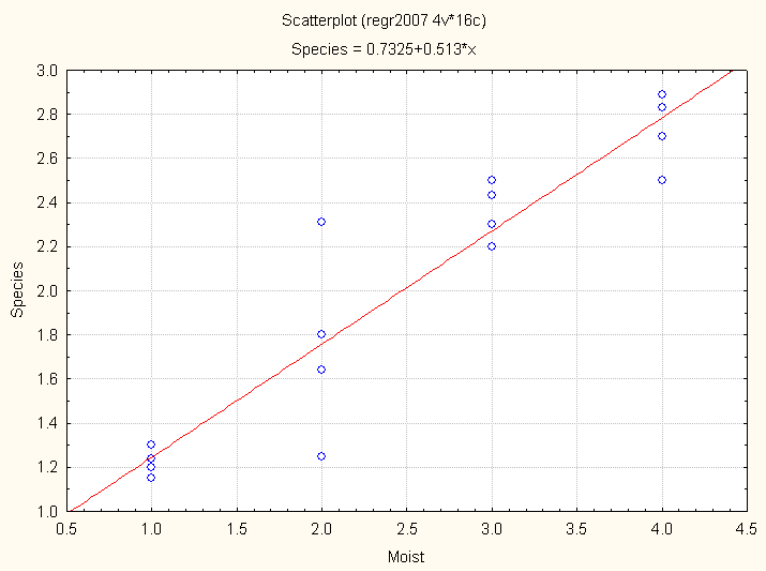



| | 1 Species | 2 Moist | 3 Nitrog | 4 Mgmt |
|---|---|---|---|---|
| 1 | 2.31 | 2 | 10.33453 | SF |
| 2 | 1.2 | 1 | 11.00445 | NM |
| 3 | 1.15 | 1 | 10.92652 | HF |
| 4 | 1.8 | 2 | 10.57481 | HF |
| 5 | 2.5 | 3 | 8.873405 | SF |
| 6 | 2.7 | 4 | 11.28908 | SF |
| 7 | 1.3 | 1 | 10.34366 | NM |
| 8 | 1.24 | 1 | 10.78619 | NM |
| 9 | 2.83 | 4 | 11.12508 | HF |
| 10 | 2.89 | 4 | 7.542314 | SF |
| 11 | 2.43 | 3 | 10.8159 | NM |
| 12 | 1.25 | 2 | 13.86944 | NM |
| 13 | 2.2 | 3 | 9.803619 | HF |
| 14 | 2.3 | 3 | 12.151 | HF |
| 15 | 1.64 | 2 | 12.43088 | NM |
| 16 | 2.5 | 4 | 10.16391 | SF |

Vysvětlujeme pokryvnost určitého rostlinného druhu (*Species*) na jednotlivých lokalitách pomocí vlhkosti půdy (*Moist*) a obsahu dusíku v půdě (*Nitrog*). Navíc známe způsob obhospodařování jednotlivých lokalit (proměnná *Mgmt* s hladinami *SF*, *NM* a *HF*). Vlhkost je vyjádřena nepřesně ("semi-kvantitativně"), způsob obhospodařování je faktor se třemi hladinami (*levels*).

Podíváme se nejprve, jak proměnná *Moist* ovlivňuje pokryvnost druhu. Z menu *Graphs* zvolíme *Scatterplots* a vlhkost vyneseme na vodorovnou a druh na svislou osu. Dostáváme obrázek podobný tomuto:

[1] Tento folder je nejlepší si vymapovat jako disk ve Windows Exploreru (Tools / Map Network Drive v menu), jeho síťová adresa je *\\prfsw01.prf.jcu.cz\canoco*



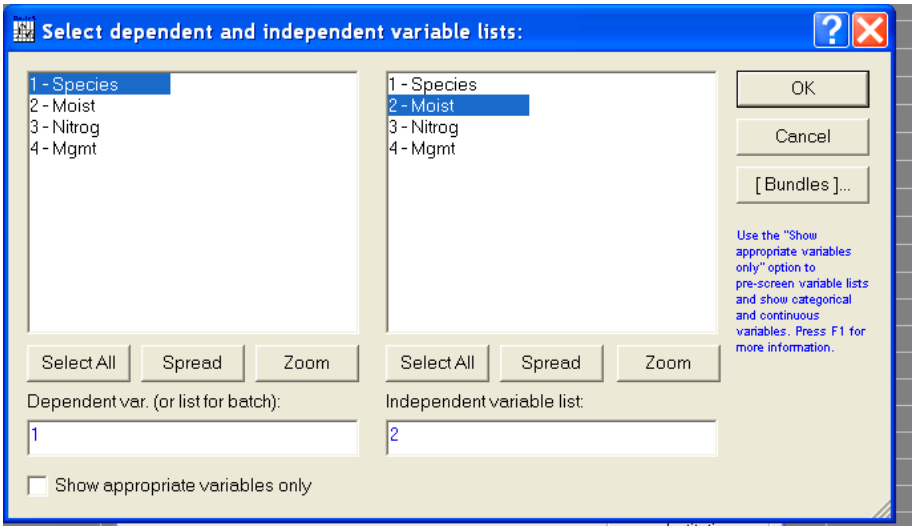

Statistica nám krom obrázku proložila body i regresní přímku, která potvrzuje zjevnou tendenci druhu mít větší pokryvnost na vlhčích lokalitách. Přímka je popsána regresní rovnicí v horní části obrázku. Ta nám říká, že průměrná pokryvnost druhu bude pro dvě lokality, lišící se o jednotku ve své vlhkosti, vyšší o 0.513 na té vlhčí. **Regresní koeficient** 0.513 je sklon přímky a jeho znaménko (plus) nám ukazuje pozitivní závislost pokryvnosti na vlhkosti. Tento koeficient je představován symbolem $b_1$ v následující rovnici:

$$Y = b_0 + b_1*X + e$$

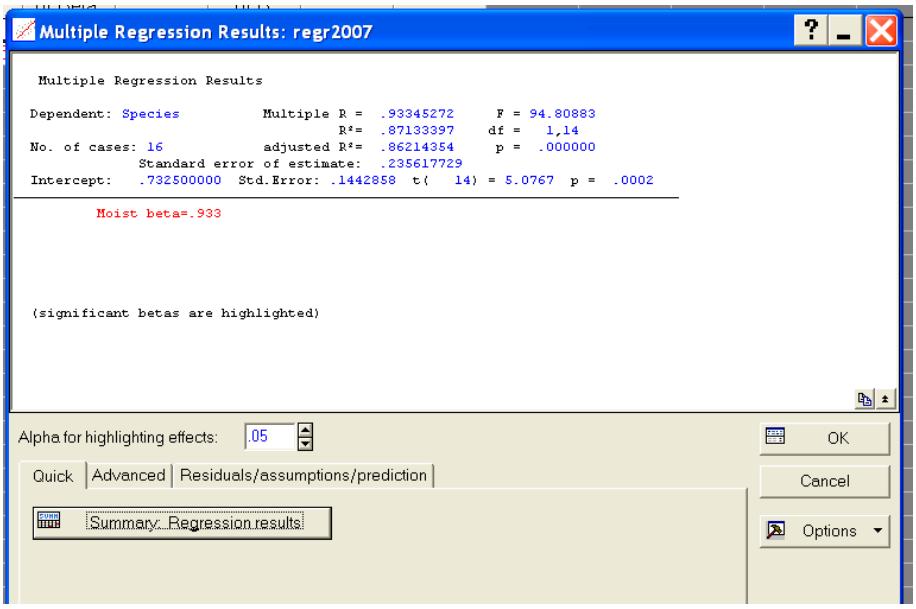
Hodnota druhého přítomného regresního koeficientu - průsečíku přímky ($b_0$ – *intercept*) je pro naše data rovna 0.7325 a nic tak zajímavého nám neříká (průměrná hodnota pokryvnosti by měla být 0.7325 pro vlhkost rovnou 0 – to je ale asi nereálná hodnota pro vlhkost, nejmenší v našich datech je jednička). Z toho, že modrá kolečka (představující kombinace pozorovaných hodnot vlhkosti a pokryvnosti) neleží přímo na nafitované červené přímce[2], lze usuzovat, že předpovědi se od reality odchylují a tyto odchylky jsou v rovnici vyjádřeny členem e. Ten nám představuje **regresní residuály** – velmi důležitou věc v regresní analýze. Jsou to vlastně rozdíly mezi skutečnými (pozorovanými, *observed*) a předpovídanými (regresním modelem fitovanými, *predicted* či *fitted values*) hodnotami vysvětlované proměnné (v rovnici Y, v angličtině jí můžeme říkat třeba *response variable*, i když Statistica upřednostňuje termín *dependent variable*). Fitovaná hodnota Y (tj. hodnota $b_0 + b_1*X$ spočtená ze dvou odhadnutých koeficientů a pozorované hodnoty X) se obvykle označuje jako Y se stříškou.

Podíváme se teď na výsledky regrese důkladněji. V programu Statistica zvolíme *Statistics* / *Multiple Regression* a v dialogu zvolíme *Species* jako Dependent variable a *Moist* jako Independent variable:

[2] Na tom není obecně nic špatného, ani to nezpochybňuje zvolený přímkový model. To, že by na každé lokalitě se shodnou půdní vlhkostí musel mít nějaký druh vždy přesně stejnou pokryvnost, neodpovídá nám známe realitě. Mimo jiné i proto, že na druh působí i mnohé další faktory.




Po spočtení se objeví přehledový dialog:




V dialogu jsou vidět různé zajímavé informace, ale většinu z nich bude lepší předvést ve více detailních výstupech. Zde si jen všimněte statistiky $R^2$ (ve třetím řádku), s hodnotou asi 0.8713. Jde o takzvaný **koeficient determinace** a vyjadřuje, jak velkou část z celkové variability hodnot vysvětlované proměnné (pokryvnosti druhu) se podařilo naším modelem vysvětlit (tedy něco přes 87% v tomto případě). Vzhledem k tomu, že v našem modelu je jedinou vysvětlující proměnnou vlhkost (*Moist*), představuje hodnota 0.8713 schopnost této proměnné vysvětlovat pokryvnost druhu *Species*[3]. Než se podíváme na údaje o jednotlivých parametrech modelu (regresních koeficientech), zůstaneme ještě chvíli u kvality modelu, kterou koeficient determinace představuje. V přehledovém dialogu klikněte na záložku *Advanced* a zvolte tlačítko *ANOVA (Overall goodness of fit)*. Zobrazí se vám tato tabulka:

[3] přesněji "schopnost vysvětlovat" při užití zvoleného typu modelu. Většina našich závěrů, které o vlastnostech fitovaného modelu děláme, závisí na naší často a priorní volbě složitosti modelu (mohli jsme místo přímky zvolit např. polynom) a transformace hodnot vysvětlované a vysvětlujících proměnných (zde žádné).

| Analysis of Variance; DV: Species (regr2007) | | | | | |
|---|---|---|---|---|---|
| Effect | Sums of Squares | df | Mean Squares | F | p-level |
| **Regress.** | 5.263380 | 1 | 5.263380 | 94.80883 | 0.000000 |
| Residual | 0.777220 | 14 | 0.055516 | | |
| Total | 6.040600 | | | | |

Celková variabilita hodnot vysvětlované proměnné *Species* (v následujících vzorečcích ji ale budu označovat jednodušeji jako Y) je zde vyjádřena celkovou sumou čtverců (TSS, *total sum of squares*), a ta je v posledním řádku, s hodnotou 6.0406. Vypočtu ji sečtením druhých mocnin rozdílů mezi hodnotami Y a celkovým aritmetickým průměrem Y (značí se jako Y s pruhem):

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Součet je přes všech *n* pozorování (je jich 16).

Veškerou variabilitu Y (t.j. hodnotu TSS) se mi ale pomocí regresního modelu obvykle nepodaří vysvětlit (to by pak v našem grafu výše všechny body ležely na přímce). Mírou vysvětlené variability (modelová suma čtverců, MSS, *model sum of squares*) je opět variabilita kolem celkového průměru Y, tentokrát ale variabilita fitovaných hodnot (tj. těch, které pro dané X spočteme pomocí regresních koeficientů):

$$MSS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

Tuto hodnotu nám Statistica zobrazuje v prvém řádku tabulky (Regress.) – je to tedy 5.26338.

Modelem nevysvětlenou variabilitu – residuální sumu čtverců (RSS, *residual sum of squares*) můžeme dopočíst lehko jako rozdíl TSS – MSS (0.77722 v našem příkladu), ale ukážeme si, jak ji spočíst i jinak:

$$RSS = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$$

Rozdíly mezi fitovanými a pozorovanými hodnotami vysvětlované proměnné Y jsou nám již známe regresní residuály *e*, takže RSS je rovna součtu druhých mocnin regresních residuálů[4].

V druhém sloupečku tabulky analýzy variance regresního modelu jsou čísla 1 a 14. Představují tzv. stupně volnosti (DF, *degrees of freedom*), odpovídající jednotlivým složkám variability vysvětlované proměnné (tj. části vysvětlené modelem a části nevysvětlené). V posledním řádku (pro TSS) tento údaj chybí, ale odpovídá mu číslo 15. To je počet pozorování zmenšený o jednotku. Než jsme začali model fitovat, měli jsme k dispozici 16 nezávislých pozorování (stupňů volnosti). Museli jsme ale z těchto údajů nejprve spočítat celkový průměr proměnné Y, a tím jsme jeden stupeň volnosti "spotřebovali". Další stupeň volnosti spotřebujeme pro nafitování regresního modelu a proto je v prvním řádku sloupečku *df* hodnota 1[5]. Residuální

[4] Mimochodem, regresní přímku fitujeme tzv. metodou nejmenších čtverců, která volí hodnoty koeficientů b tak, aby právě suma RSS byla co nejmenší.

[5] Proč jen jeden DF, když má náš regresní model dva parametry, které je třeba vypočíst ($b_0$ a $b_1$)? Představme si jednodušší model, který by neobsahoval vysvětlující proměnnou X a tedy ani regresní koefient $b_1$: Y = $b_0$ + e. V tomto modelu by jeho jediný parametr představoval aritmetický průměr proměnné Y, a jeho hodnotu jsme již spočetli. Jinými slovy – pokud již známe hodnotu průměru Y, stačí nám znalost koeficientu $b_1$ k dopočtení hodnoty $b_0$. Velmi praktickým důsledkem tohoto vztahu (v kontextu ordinačních metod) je to, že pokud od hodnot proměnných X a Y odečteme jejich odpovídající aritmetické průměry, bude hodnota koeficientu $b_0$ vždy rovna 0.

stupně volnosti jsou opět rozdílem mezi celkovými stupni volnosti a modelovými stupni volnosti (15-1 = 14).

Hodnota TSS dělená jí odpovídajícím počtem DF (tj. 15) je, jak si jistě pamatujete z Biostatistiky☺, odhadem variance. Pokud odpovídajícími stupni volnosti vydělíme MSS nebo RSS, dostaneme tzv. průměrné sumy čtverců (MS, *mean squares*) – modelovou resp. residuální. V našem případě dělíme modelovou sumu čtverců hodnotou 1, proto je $MS_{mod}$ shodná s MSS. Naproti tomu u RSS dělíme 14, takže residuální MS má hodnotu 0.055516. Mimochodem, tato průměrná suma čtverců se obvykle nazývá *error mean square* a značí se pak $MS_{err}$.

Průměrné sumy čtverců můžeme využít k testu signifikance regresního modelu. V našem jednoduchém modelu – pouze s jednou vysvětlující proměnnou – odpovídá tento test testu signifikance koeficientu $b_1$, tedy testu nulové hypotézy H0: $b_1 = 0$[6]

Pokud nulová hypotéza platí, měl by být poměr $MS_{mod}$ / $MS_{err}$ hodnotou "poměrně blízkou" hodnotě 1 (tj. vysvětlená a nevysvětlená variabilita by měly být podobně velké), přesněji by měly pocházet z F distribuce s parametry 1, 14 (pro náš konkrétní model). Pravděpodobnost, že skutečná hodnota tohoto poměru, tzv. F statistiky (t.j. 94.80883) z této F distribuce pochází je menší než 0.000001, jak je vidět z hodnoty ve sloupečku *p-level*, a proto H0 zamítáme s touto pravděpodobností chyby I. typu (na této hladině signifikance).

Nyní se vraťte do přehledového dialogu a zvolte tlačítko *Summary: Regression results*. Program Statistica nám zobrazí tuto tabulku:

Regression Summary for Dependent Variable: Species (regr2007)
R= .93345272 R²= .87133397 Adjusted R²= .86214354
F(1,14)=94.809 p<.00000 Std.Error of estimate: .23562

| N=16 | Beta | Std.Err. of Beta | B | Std.Err. of B | t(14) | p-level |
|---|---|---|---|---|---|---|
| Intercept | | | 0.732500 | 0.144286 | 5.076730 | 0.000169 |
| Moist | 0.933453 | 0.095867 | 0.513000 | 0.052686 | 9.736983 | 0.000000 |

Všimněte si nejprve, že v horním políčku jsou zopakovány údaje o koeficientu determinace a také o F testu rozkladu variability vysvětlované proměnné. Hodnoty regresních koeficientů našeho výše popsaného přímkového modelu jsou v této tabulce zobrazeny až ve sloupečku *B*. První sloupeček tabulky, s označením *Beta* sice představuje také regresní koeficienty, ale odlišného modelu. V něm byly hodnoty vysvětlované i vysvětlující proměnné změněny tak, aby po této změně měly nulový průměr a jednotkovou varianci[7]. Praktickou výhodu takového postupu uvidíme až v okamžiku, kdy bude náš regresní model obsahovat více než jednu vysvětlující proměnnou. Hodnoty "hrubých" regresních koeficientů (B v tabulce programu Statistica) se totiž mezi různými proměnnými špatně porovnávají, jejich absolutní velikost závisí na škále, na které jsou vyjádřeny hodnoty vysvětlujících proměnných.

Ještě zmíním poslední dva sloupečky tabulky – testovací statistiku *t* a odpovídající signifikanci (*p-level*). Pro náš jednoduchý regresní model je tento T test v případě regresního koeficientu $b_1$

[6] Tady jsem potrestán za svoji "nematematickou nepřesnost", nerozlišuji totiž pečlivě odhady regresních koeficientů (b) a jejich skutečné (neznáme a nepoznatelné) hodnoty v základní populaci, ze které vybíráme, které by měly být značeny písmenkem β. Hypotéza se samozřejmě týká veličiny $\beta_1$ – o $b_1$ totiž s určitostí víme, že není rovno nule (je rovno 0.513).

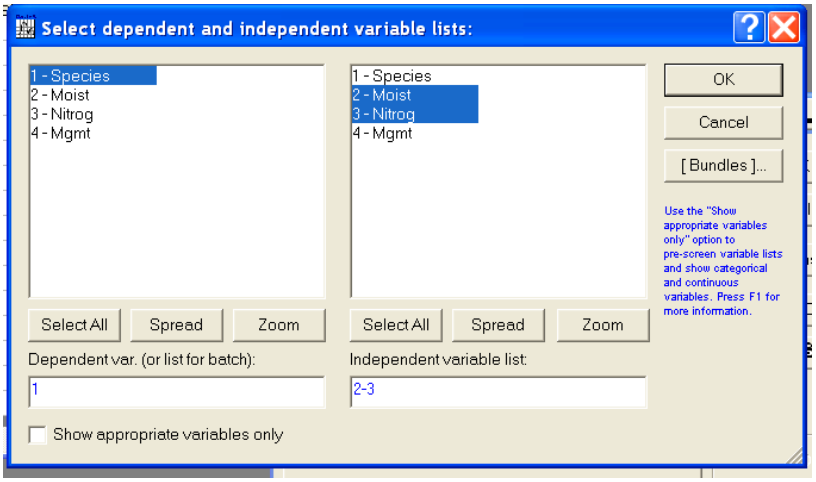
[7] Postupu vedoucímu k nulovému průměru se říká centrování (*centering*), a spočívá v odečtení průměru proměnné od každé její hodnoty, jak jsem již zmínil v dřívější poznámce pod čarou. Postup vedoucí k jednotkové varianci se nazývá standardizace (*standardization*) a spočívá ve vydělení každé hodnoty směrodatnou odchylkou dané proměnné. Standardizace se obvykle provádí po centrování.

(řádek *Moist*) shodný s celkovým F testem[8]. Ve složitějších modelech (se dvěma či více vysvětlujícími proměnnými) tomu ale tak často nebývá: to, že je model jako celek průkazný (tedy schopný vysvětlit významnou část variability hodnot Y), obvykle neznamená, že všechny použité vysvětlující proměnné jsou v modelu potřebné nebo že na tom dokonce mají stejný podíl.

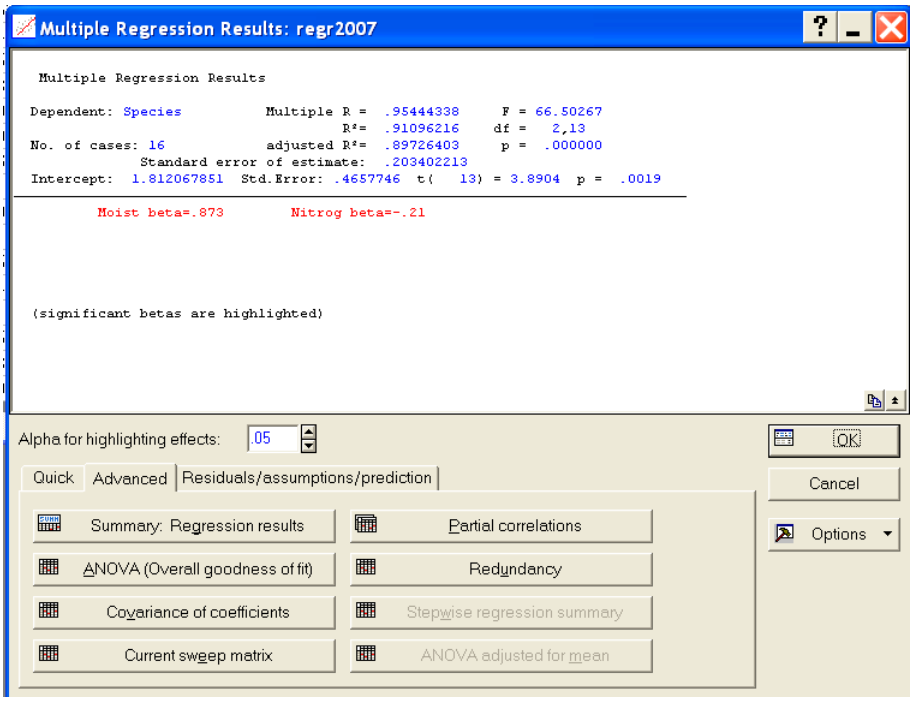
Pro zopakování postupů se nyní podívejte, zda proměnná *Nitrog* má průkazný vliv na pokryvnost druhu *Species* a jakou část variability jeho hodnot vysvětlí.

Proměnná *Nitrog* by měla vysvětlovat asi 21.4% z variability v hodnotách proměnné *Species*, ale její vliv není průkazný na hladině $\alpha = 0.05$ ($F=3.8083$, $p = 0.0713$). Přes tento "neúspěch" se podíváme, jak by vypadaly výsledky z modelu mnohonásobné regrese (*multiple regression*), tedy lineárního modelu, ve kterém máme více než jednu vysvětlující proměnnou.

Při výběru proměnných zadáme tentokrát jak *Moist*, tak *Nitrog* jako vysvětlující proměnné:




Po provedení analýzy nás čeká překvapení již v přehledovém dialogu:




[8] dokonce i hodnoty testových statistik souvisí – *t* statistika je zde odmocninou F statistiky, tedy 9.736983 * 9.736983 = 94.809

Obě vysvětlující proměnné mají průkazný vliv ve společném modelu, jak je vidět také v tabulce s regresními koeficienty:

- Regression Summary for Dependent Variable: Species (regr2007)

Regression Summary for Dependent Variable: Species (regr2007)
R= .95444338 R²= .91096216 Adjusted R²= .89726403
F(2,13)=66.503 p<.00000 Std.Error of estimate: .20340

| N=16 | Beta | Std.Err. of Beta | B | Std.Err. of B | t(13) | p-level |
|---|---|---|---|---|---|---|
| **Intercept** | | | 1.812068 | 0.465775 | 3.89044 | 0.001860 |
| Moist | 0.872813 | 0.086514 | 0.479674 | 0.047546 | 10.08873 | 0.000000 |
| Nitrog | -0.208100 | 0.086514 | -0.092656 | 0.038520 | -2.40540 | 0.031763 |

Právě v případě modelu násobné regrese se nám dobře hodí standardizované koeficienty (ve sloupečku *Beta*), při porovnání absolutních hodnot (znaménko udává směr závislosti, nikoliv její sílu) vidíme, že standardizovaná změna hodnoty proměnné *Moist* vyvolá zhruba čtyřnásobně větší změnu vysvětlované proměnné (pokryvnosti druhu) než standardizovaná změna v nabídce dusíku.

Ovšem případ, kdy se vliv proměnné, která sama o sobě neměla vliv na vysvětlovanou proměnnou, stane průkazný, pokud ji v modelu skombinujeme s proměnnou jinou, je v praxi méně častý než případ, kdy samostatný vliv proměnné vymizí, pokud je kombinována s jinou (jinými) vysvětlujícími proměnnými. To se stává proto, že často říkají různé vysvětlující proměnné, které máme k dispozici, o vysvětlované proměnné "podobné věci" (např. změřená půdní vlhkost a C/N poměr v půdě), část své vysvětlující schopnosti sdílí. Ve statistických analýzách – při tvorbě regresních či ordinačních modelů – proto musíme rozlišovat dvě různé role určité vysvětlující proměnné. Za prvé nás může zajímat její nezávislý efekt (*independent effect*, často též *marginal effect*), a v takovém případě předstíráme, že žádné jiné vysvětlující proměnné neexistují, resp. nebereme jejich existenci v daném modelu na zřetel. Nebo nás může zajímat tzv. podmíněný efekt (*conditional effect*) dané proměnné, podmíněný hodnotami ostatních vysvětlujících proměnných. Jinými slovy, je to to, co nám o vysvětlované proměnné daná proměnná říká navíc k informaci, kterou nám podávají proměnné jiné.

Na závěr se ještě podíváme na vliv způsobu obhospodařování na pokryvnost druhu *Species*. V programu Statistica provádím takovou analýzu obvykle pomocí modelu analýzy variance (ANOVA, *analysis of variance*), my si ale později ukážeme, že ANOVA není tak moc odlišná od mnohonásobné regrese a vliv obhospodařování spočteme i pomocí metody pro Multiple Regression. Nejprve ale klasický postup. Z menu zvolíme *Statistics* / *ANOVA* a pak v dialogu *One-way ANOVA* (máme totiž jen jednu vysvětlující proměnnou, faktor *Mgmt*).

Pomocí tlačítka *Variables* zadáme *Species* do *Dependent variable list* a *Mgmt* do druhého seznamu (*Categorical predictor (factor)*). Po odsouhlasení tlačítkem *OK* zvolíme na záložce *Quick* tlačítko *All effects*.

- Univariate Tests of Significance for Species (regr2007)

Univariate Tests of Significance for Species (regr2007)
Sigma-restricted parameterization
Effective hypothesis decomposition

| Effect | SS | Degr. of Freedom | MS | F | p |
|---|---|---|---|---|---|
| **Intercept** | 66.65879 | 1 | 66.65879 | 298.2065 | 0.000000 |
| Mgmt | 3.13468 | 2 | 1.56734 | 7.0117 | 0.008597 |
| Error | 2.90592 | 13 | 0.22353 | | |

Jak si asi pamatujete z Biostatistky, prvý řádek zobrazené tabulky (*Intercept*) ignorujeme. Celkovou variabilitu v hodnotách proměnné *Species* (6.0406, stejně jako v dříve počítané

regresi) je možné rozdělit na variabilitu, vysvětlitelnou odlišnými průměry pokryvnosti ve třech skupinách definovaných hladinami faktoru *Mgmt* (MSS = 3.13468), a variabilitu uvnitř skupin (RSS), která má velikost 2.9059. Obdobným způsobem jako dříve u regresního modelu spočteme průměrné sumy čtverců a také F statistiku (zde s hodnotou 7.0117) a na ní pak může být založen test významnosti vlivu obhospodařování na pokryvnost studovaného druhu (p=0.0086).

Postup je tedy velmi podobný regresi, ale jsou tu určité rozdíly. Graficky nelze nafitovaný model vyjádřit pomocí přímky, hladiny faktoru představují samostatné skupiny s předpovídanou konstantní hodnotou (skupinový průměr vysvětlované proměnné).

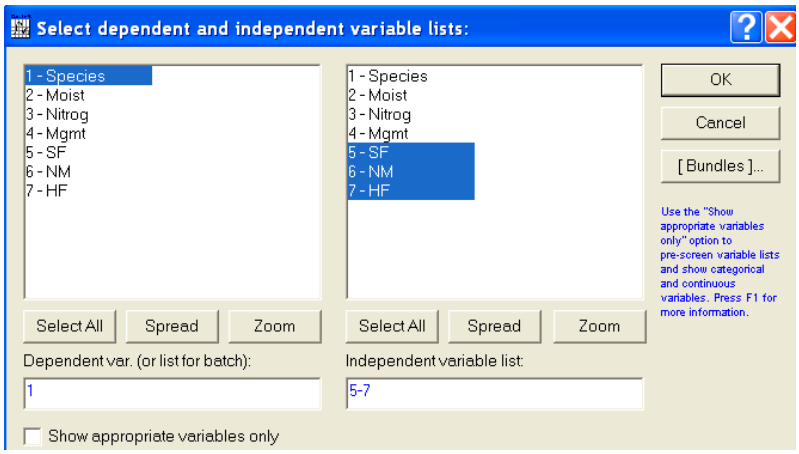
Zajímavá je skutečnost, že jedna vysvětlující proměnná (faktor *Mgmt*) nám v modelu "spotřebovala" ne jeden, ale dva stupně volnosti. Je to proto, že jde o faktor se třemi hladinami, obecně faktoru s *p* hladinami odpovídá *p-1* stupňů volnosti. Důvod pochopíme, až si ukážeme zadání analýzy variance pomocí mnohonásobné regrese.

Abychom takovou analýzu mohli provést, musíme nejprve zadat naši vysvětlující proměnnou jiným způsobem. Jeden sloupeček obsahující v každém řádku jméno jedné ze tří hladin faktoru (SF, HF nebo NM) nahradíme třemi proměnnými (sloupečky), které se mohou jmenovat např. podle jednotlivých hladin (na pojmenování ovšem úspěch analýzy nezávisí☺). Důležité je, že v každém řádku bude právě jedna hodnota 1, pro další dvě proměnné v něm budou nuly. Jednička bude, pochopitelně, v proměnné, která odpovídá hladině faktoru, kterou dané pozorování mělo. Výsledná tabulka dat, rozšířená o tyto tři proměnné (nazývají se *dummy variables* nebo také 0/1 proměnné), bude vypadat takto:

Data: regr2007* (7v by 16c)

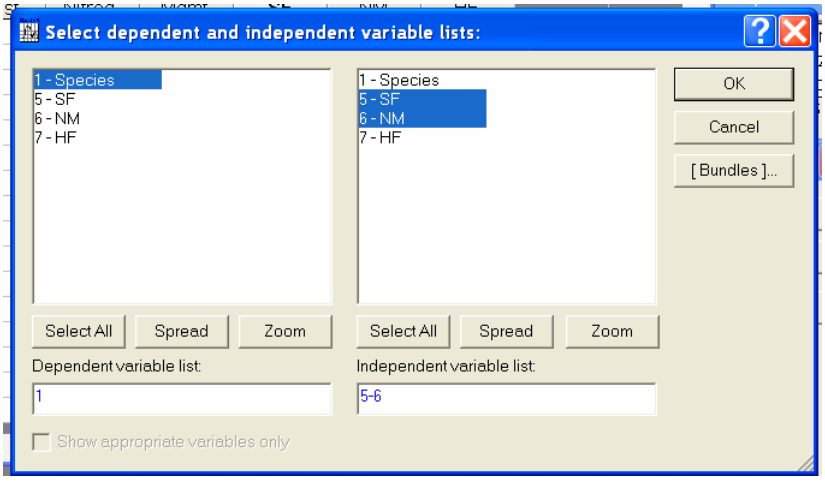
| | 1 Species | 2 Moist | 3 Nitrog | 4 Mgmt | 5 SF | 6 NM | 7 HF |
|---|---|---|---|---|---|---|---|
| 1 | 2.31 | 2 | 10.33453 | SF | 1 | 0 | 0 |
| 2 | 1.2 | 1 | 11.00445 | NM | 0 | 1 | 0 |
| 3 | 1.15 | 1 | 10.92652 | HF | 0 | 0 | 1 |
| 4 | 1.8 | 2 | 10.57481 | HF | 0 | 0 | 1 |
| 5 | 2.5 | 3 | 8.873405 | SF | 1 | 0 | 0 |
| 6 | 2.7 | 4 | 11.28908 | SF | 1 | 0 | 0 |
| 7 | 1.3 | 1 | 10.34366 | NM | 0 | 1 | 0 |
| 8 | 1.24 | 1 | 10.78619 | NM | 0 | 1 | 0 |
| 9 | 2.83 | 4 | 11.12508 | HF | 0 | 0 | 1 |
| 10 | 2.89 | 4 | 7.542314 | SF | 1 | 0 | 0 |
| 11 | 2.43 | 3 | 10.8159 | NM | 0 | 1 | 0 |
| 12 | 1.25 | 2 | 13.86944 | NM | 0 | 1 | 0 |
| 13 | 2.2 | 3 | 9.803619 | HF | 0 | 0 | 1 |
| 14 | 2.3 | 3 | 12.151 | HF | 0 | 0 | 1 |
| 15 | 1.64 | 2 | 12.43088 | NM | 0 | 1 | 0 |
| 16 | 2.5 | 4 | 10.16391 | SF | 1 | 0 | 0 |

Zkusme nyní zadat analýzu variance pomocí modelu mnohonásobné regrese, ve kterém proměnné SF, NM a HF nahradí faktor *Mgmt*:

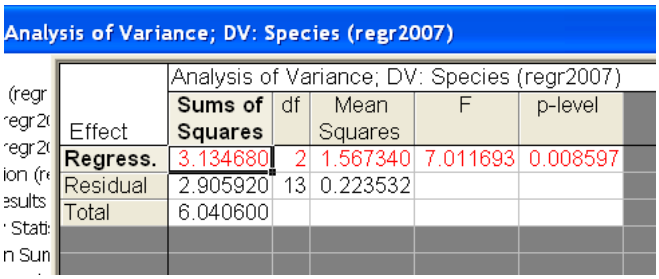


Bohužel, Statistica odmítne tento model spočítat:


Pomoc je, naštěstí, poměrně snadná – jednu z vysvětlujících proměnných vynecháme. Nezáleží (alespoň z hlediska hlavního výsledku ANOVA modelu, tj. testu signifikance pro daný faktor) na tom, která ze tří proměnných to bude, může to být například poslední:




Pak již můžeme model nafitovat a tlačítko *ANOVA (Overall goodness of fit)* na záložce *Advanced* nám dá tento výsledek, shodný s dřívějším výsledkem získaným "klasickým" postupem:




Proč jsme ale museli jednu z proměnných z modelu odstranit? Důvodem je její "nadbytečnost": způsob kódování faktoru (hodnota 1 jen pro jednu z hladin, ostatní nuly) nezávisí na konkrétních hodnotách daného faktoru, je to naše apriorní rozhodnutí. Jestliže tedy známe *p-1* z *p* proměnných kódujících faktor s *p* hladinami, je poslední z proměnných nadbytečná – její hodnotu snadno dopočítáme tak, že součet hodnot ostatních proměnných odečteme od hodnoty 1. Tomu také odpovídá již dříve zmíněný fakt, že našemu faktoru se třemi hladinami odpovídaly v klasickém ANOVA modelu jen dva stupně volnosti.

Alternativní vysvětlení je, že hodnota regresního koeficientu pro každou ze tří 0/1 proměnných představu průměrnou hodnotu pokryvnosti v odpovídající skupině pozorování, ale v našem modelu máme navíc ještě koeficient $b_0$, který odpovídá (i když zde není roven) celkovému průměru. Ale když známe celkový průměr, jsme schopni vypočíst průměr jedné ze tří skupin z něj a z hodnot dvou zbývajících skupinových průměrů (pokud známe velikosti skupin). Pokud jsme jednu z 0/1 proměnných v regresi vynechali, bude odhad $b_0$ představovat průměrnou hodnotu odpovídající skupiny (zde HF) a zbylé dva regresní koeficienty ($b_1$ a $b_2$) udávají, o kolik jsou průměry zbylých dvou skupin větší nebo menší než průměr "referenční" skupiny.