KORELAČNÍ A REGRESNÍ ANALÝZA

PŘÍKLAD: Procentní ztráta amoniaku při výrobě kyseliny dusičné.
V aparatuře pro výrobu kyseliny dusičné byl po dobu 21 dní sledován provoz oxidace amoniaku na kyselinu dusičnou. Závisle proměnná Y je relativní ztráta amoniaku v procentech, nezávisle proměnné (regresory) jsou koncentrace kyseliny dusičné X_1, průtok vzduchu X_2 a teplota chladící vody X_3.

Měření veličin X_1, X_2, X_3 byla uložena v softwarovém prostředí NCSS do proměnných *Acid*, *Air*, *Water*, hodnoty veličiny Y do proměnné *Amoniac*. Uvedeme výstupy ze jmenovaného statistického softwaru.

VÝBĚROVÉ KORELAČNÍ KOEFICIENTY

	Acid	**Air**	**Water**	**Amoniac**
Acid	1.0000	0.5395	0.3909	0.3998
Air		1.0000	0.7636	**0.9079**
Water			1.0000	**0.8755**
Amoniac				1.0000

Poslední sloupec: proměnná *Amoniac* je vysoce korelována s proměnnými *Air* a *Water*, které jsou též korelované mezi sebou.

Závislost proměnné *Amoniac* pouze na proměnné *Air* a pouze na proměnné *Water* je možné popsat regresní přímkou.

První řádek: proměnná *Acid* nemá výrazné korelace s ostatními.

Zjistili jsme, že **procentní ztráta amoniaku Y je vysoce korelována s veličinami průtok vzduchu X_2 a teplota chladící vody X_3.**

LINEÁRNÍ REGRESNÍ MODEL

Do modelu pro Y zařadíme všechny regresory X_1, X_2, X_3.

Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$, $i = 1, \ldots, n = 21$.

Estimated Model:
Amoniac =
-37.6769 - 0.2167*Acid + 0 .7336*Air + 1.3883*Water

Variable	**Coefficient**	**T-Value**	**p-value**
Intercept	-37.6769	**-3.137**	**0.0060**
Acid	-0.2167	-1.343	0.1968
Air	0.7336	**5.285**	**0.0001**
Water	1.3883	**3.894**	**0.0012**

2. sloupec: odhady b_0, b_1, b_2, b_3 parametrů β_0, β_1, β_2, β_3.
3. sloupec: hodnoty testové statistiky T_j pro test hypotéz $H_0^{(j)}$: $\beta_j = 0$.
4. sloupec: p-hodnoty pro test hypotéz $H_0^{(j)}$: $\beta_j = 0$, $j = 0, 1, 2, 3$.

Kritický obor testu hypotézy $H_0^{(j)}$: $\beta_j = 0$, $j = 0, 1, 2, 3$:

$$T_j = |b_j| / (s^2 v_{jj})^{1/2} > t_{n-k-1}(1-\alpha/2) \Rightarrow \text{zamítáme } H_0^{(j)} \text{ na hladině } \alpha,$$

s^2 je odhad rozptylu σ^2 náhodných chyb ε_i, $i = 1, \ldots, n = 21$,
který počítáme jako $s^2 = (\boldsymbol{Y}^T \boldsymbol{Y} - b^T \boldsymbol{X}^T \boldsymbol{Y}) / (n - k - 1)$,
v_{jj} je diagonální prvek matice $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$,
sloupcový vektor $\boldsymbol{Y}$ obsahuje pozorování veličin Y_i, $i = 1, \ldots, n = 21$,
matice $\boldsymbol{X}$ má v nultém sloupci jedničky, v j-tém sloupci 21 pozorování veličin X_j, $j = 1, 2, 3$,
$t_{n-k-1}(1-\alpha/2)$ je kvantil t-rozdělení o n-k-1 stupních volnosti.

V našem případě n-k-1 = 21-3-1 = 17 a $t_{17}(0.975) = 2{,}1098$.

Nezamítneme-li $H_0^{(j)}$: $\beta_j = 0 \Rightarrow$
Y nezávisí v rámci uvažovaného modelu na regresoru X_j.

Tučně vyznačené hodnoty statistik T_2, T_3 překročí v absolutní hodnotě kvantil $t_{17}(0.975) = 2{,}1098$, u stejných regresorů X_2, X_3 tučně vyznačené p-hodnoty nepřekročí zvolenou hladinu $\alpha = 0{,}05$ $\Rightarrow$ **procentní ztráta amoniaku *Y* závisí v uvažovaném modelu na regresorech průtok vzduchu X_2 a teplota chladící vody X_3. Závislost na koncentraci kyseliny dusičné X_1 se v datech neprokázala.** Statisticky významný je i parametr β_0 (*Intercept*).

Závěr je v souladu s hodnotami výběrových korelačních koeficientů.

Koeficient determinace v modelu se 3 nezávisle proměnnými:
$R^2 = 0.9132$ $\Rightarrow$ model s regresory X_1, X_2, X_3 vysvětlí více než 91 % variability procentní ztráty amoniaku *Y*.

Podívejme se, jak se změní koeficient determinace, když vyloučíme z modelu proměnnou *Acid*, tedy regresor X_1, který test nulovosti koeficientu β_1 indikuje jako nadbytečný.

Odhadneme model s nezávisle proměnnými *Air* a *Water*.

Estimated Model:
Amoniac = -52.1635 + 0.6580*Air + 1.4068*Water

Variable	Coefficient	T-Value	p-value
Intercept	-52.1635	**-9.652**	**0.0000**
Air	0.6580	**5.074**	**0.0001**
Water	1.4068	**3.864**	**0.0011**

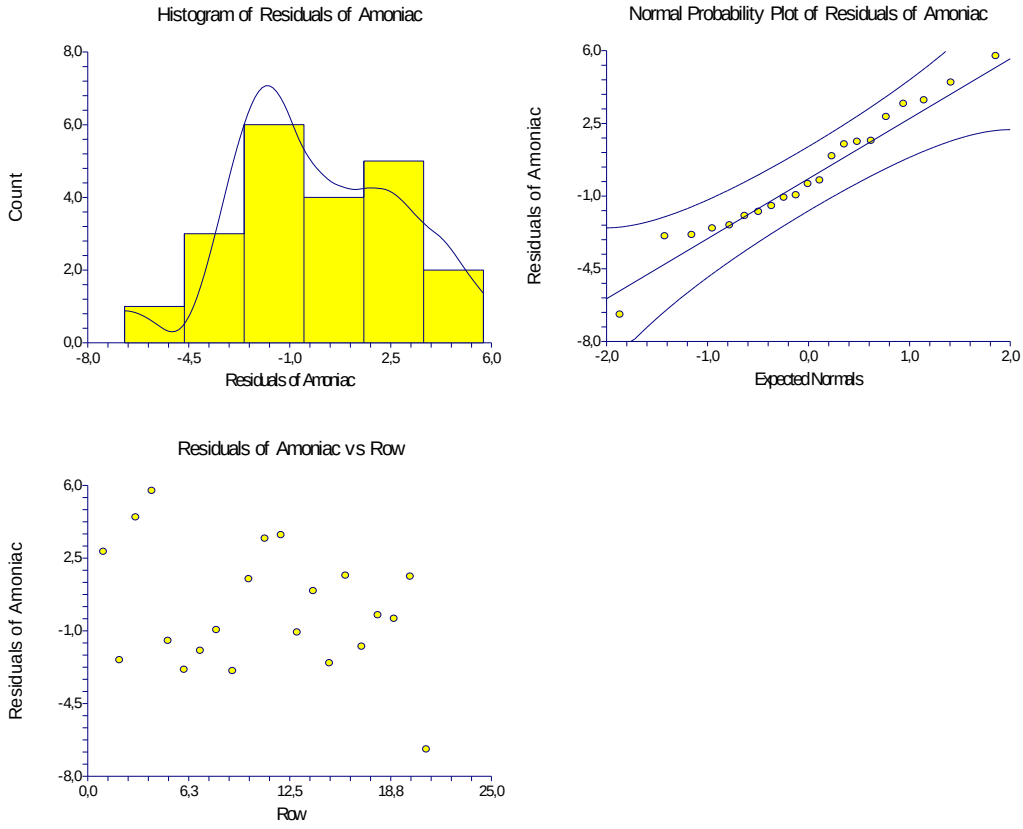
Tučně vyznačené hodnoty statistik T_2, T_3 překročí v absolutní hodnotě kvantil $t_{18}(0.975) = 2{,}1009$, u stejných regresorů X_2, X_3 tučně vyznačené p-hodnoty nepřekročí zvolenou hladinu $\alpha = 0{,}05$. Rovněž parametr β_0 je statisticky významný.

Koeficient determinace v modelu se 2 nezávisle proměnnými:
$R^2 = 0.9040$ $\Rightarrow$ model s regresory X_2, X_3 vysvětlí více než 90 % variability procentní ztráty amoniaku *Y* $\Rightarrow$ vynechání regresoru X_1 nesnížilo výrazně vysvětlenou variabilitu závisle proměnné *Y*.

ANALÝZA REZIDUÍ

Na závěr zpětně ověříme předpoklady modelu, normalitu a nezávislost náhodných chyb ε_i, $i = 1, \ldots, 21$. V modelu se 3 regresory byla spočítána rezidua $e_i = Y_i - b_0 - b_1 X_{i1} - b_2 X_{i2} - b_3 X_{i3}$, $i = 1, \ldots, 21$.

Na rezidua bylo aplikováno 5 testů normality s p-hodnotami: 0.87797, 0.630909, 0.972679, 0.847933, 0.981207.



Vzhledem k p-hodnotám výrazně větším než $\alpha = 0{,}05$ **žádný z 5 testů nezamítá hypotézu normality náhodných chyb** ε_i, $i = 1, \ldots, 21$.
Histogram reziduí neposkytuje stran normality jasnou odpověď, normální diagram (Q-Q graf) indikuje normalitu tím, že se body shlukují kolem diagonální přímky.
Bodový graf reziduí ukazuje rezidua nepravidelně roztroušená kolem nulové úrovně, což naznačuje vzájemnou **nezávislost a nulovou střední hodnotu náhodných chyb.**
Stejný závěr poskytla anylýza reziduí v modelu se 2 regresory.