Analýza a modelovanie kvality kávy

Dáta

V tejto práci sa venujem dátam o kvalite kávy z rôznych fariem – Coffee Quality Data (CQI May-2023). Tento dataset (dáta pochádzajú z Coffee Quality Institute) uvádza pre kávy dopestované na rôznych farmách okrem iného ich spôsob spracovania, odrodu kávovníka, nadmorskú výšku, v ktorej sú pestované, a tiež metriky kvality ako acidita, aróma, balans chutí a iné. Výstupným hodnotením kvality je `TotalCupPoints`, ktorý spája 10 rôznych kategórií, z ktorých tri boli spomenuté.

Tieto dáta boli vyčistené, mierne upravené (pridané kontinenty pre štáty a interval nadmorských výšok som previedol na jeho stred) a spojené Pythonovským programom, ktorý k práci neprikladám.

Dôvod, prečo som si na analýzu vybral práve takéto dáta, je čisto z osobného záujmu o túto tematiku. Už pár rokov ma fascinuje svet tzv. *specialty coffee* (dataset sa pravdepodobne venuje iba práve takýmto kávam, čiže kávam s jasným pôvodom a najvyššou kvalitou) a škála procesu, ktorý je za tým, aby si ktokoľvek mohol dať šálku kávy. Odvtedy som subjektívne odpozoroval nejaké charakteristiky rôznych káv a chcel by som zistiť, či sa niektoré moje pohľady zhodujú aj s informáciou zachytenou v tomto datasete.

Testovanie strednej hodnoty kvality

V rôznych knihách sa uvádzajú rôzne typické charakteristiky chuťového profilu káv pre regióny. Moje osobné preferencie sú u káv z Afriky, preto chcem otestovať, či je ich priemerná kvalita (teda `TotalCupPoints`) vyššia ako z iných kontinentov. Budem teda používať dvojsúborový jednostranný t-test na hladine významnosti 5%, kde alternatívna hypotéza tvrdí, že káva z Afriky má vyššiu strednú kvalitu, ako káva z iného kontinentu (Ázia, Južná Amerika, Severná Amerika):

$$H_0 : \mu_{\text{Afrika}} \leq \mu_{\text{iný}} \quad \text{vs.} \quad H_1 : \mu_{\text{Afrika}} > \mu_{\text{iný}}$$

Kontinent	$\bar{X}$	p-hodnota t-testu
Afrika	84.626	–
Ázia	84.024	0.0229
Severná Amerika	83.068	0.0008
Južná Amerika	83.277	$3.093 \cdot 10^{-5}$

Vidíme, že vo všetkých troch prípadoch zamietame nulovú hypotézu, že očakávaná kvalita Africkej kávy je nižšia alebo rovná kvalite kávy z iného kontinentu. Teda to, že priemerná kvalita vyššia o 0.6 bodu od Ázie, resp. o viac ako bod od Amerík, je natoľko signifikantné, že môžeme (voľne) tvrdiť, že Africké kávy sú kvalitnejšie.

Je známe, že kávovníky pestované vo vyšších nadmorských vyškach dozrievajú pomalšie a teda zrná plodov majú viac času „naťahať“ chuťový profil dužiny. Toto by malo teda mať za následok chuťovo zaujímavejšiu a tým pádom aj kvalitnejšiu kávu. Otestujeme teda rovnakou formou t-testu ako vyššie, či kávy pestované v nadmorskej výške viac ako 1500 m.n.m. majú vyššie očakávané `TotalCupPoints` oproti kávam pestovaným nižšie:

$$H_0 : \mu_{>1500} \leq \mu_{\leq 1500} \quad \text{vs.} \quad H_1 : \mu_{>1500} > \mu_{\leq 1500}$$

Nadmorská výška	$\bar{X}$	p-hodnota t-testu
> 1500 m.n.m.	84.340	–
≤ 1500 m.n.m.	83.436	$6 \cdot 10^{-6}$

Vidíme, že teória je potvrdená v dátach – hovoria, že *specialty* kávy majú vyššiu priemernú kvalitu, ak sú pestované vo výške nad 1500 m.n.m.

Posledný zo série t-testov (viem, že som ich nemal robiť veľa, ale príliš ma to zaujímalo na to, aby som ich nespravil) sa zaoberá spracovaním kávy. Dva najčastejšie spôsoby sú *natural* (kávové čerešne sa sušia na slnečnom svetle celé) a *washed* (káva sa suší po tom, ako je dužina odstránená napríklad vodou). Ja subjektívne neviem určiť, že chuťový profil ktorého procesu mi vyhovuje viac, preto chcem otestovať, či je v dátach dostatočná informácia o tom, že ich priemerné kvality sú rozdielne. Hypotézy dvojsúborového t-testu budú teda nasledovné:

$$H_0 : \mu_{\text{natural}} = \mu_{\text{washed}} \quad \text{vs.} \quad H_1 : \mu_{\text{natural}} \neq \mu_{\text{washed}}$$

Spracovanie	$\bar{X}$	p-hodnota t-testu
natural	83.679	–
washed	83.633	0.8917

Vysoká p-hodnota hovorí, že nemôžeme zamietnuť, že stredné hodnoty kvality pre obe spracovania sú rovnaké. Toto naznačuje, že aj keď proces spracovania sa odráža na charakteristikách, môže to mať za následok aj kvalitnejšie, aj menej kvalitné hodnotenie.

Podiel pestovaných kávovníkov odrody *Gesha*

V posledných rokoch narástla popularita kávy z kávovníka odrody *Gesha* pre jej výnimočné charakteristiky a konzistentne vysoké hodnotenia v rôznych rebríčkoch. Je však náročné dohľadať podiel jednotlivých pestovaných odrôd kávovníka v celkovej produkcii arabiky (voľne povedané arabika je „naddruh“ väčšiny pestovaných odrôd). Rád by som teda na základe dát z *Coffee Quality Institute* odvodil 95% interval spoľahlivosti pre podiel odrody *Gesha* vo svete. Budeme ho konštruovať nasledovne

$$(L,\ U) = (\hat{p} - S \cdot t_{n-1}(0.025)),\ \hat{p} + S \cdot t_{n-1}(0.025)$$

Kde $\hat{p} = \frac{\#\,\text{Gesha}}{n}$ je odhad tohto podielu a $S^2 = \frac{\hat{p}(1-\hat{p})}{n}$ je odhad disperzie $\hat{p}$. Výsledky sú nasledovné:

$$\hat{p} = 0.1337$$
$$S^2 = 0.000573$$
$$(L,\ U) = (0.0865,\ 0.1809)$$

Dáta naznačujú, že skutočný podiel odrody *Gesha* v pestovaní kávovníka druhu arabika je na 95% medzi 8.65% a 18.09%. Tento fakt však nie je v úplnej zhode s teoretickými znalosťami, že *Gesha* je pomerne vzácna, aj keď presnú kvantifikáciu tohto faktu som nenašiel. Môže sa teda jednať o výchylku spôsobenú napríklad zberom dát, ak jedným zo zámerou bolo aj komplexnejšie vyhodnocovanie kvality tejto odrody kvôli nárastu jej popularity.

Môžeme sa ešte pozrieť na to, či sa zastúpenie tejto odrody v produkcii líši medzi kontinentami. Táto odroda pochádza z Etiópie, no aktuálne je známa svojou kvalitou *Gesha* vypestovaná v Paname. Pozrieme sa teda na to, či môžeme tvrdiť, že podiel, ktorý v produkcii predstavujú, je rozdielny pre Afriku ($\hat{p}_1$) a Severnú Ameriku ($\hat{p}_2$):

$$H_0 : p_1 = p_2 \quad \text{vs.} \quad H_1 : p_1 \neq p_2$$

$$\begin{aligned}
\hat{p}_1 &= \frac{1}{23} = 0.043 \\
\hat{p}_2 &= \frac{6}{64} = 0.094 \\
T &= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = 0.8977548 \\
p\text{-value} &= 2(1 - \mathrm{pt}_{(23+64-2)}(|T|)) = 0.3719
\end{aligned}$$

Označením $\mathrm{pt}_n(\dots)$ značíme kumulatívnu distribučnú funkciu t-rozdelenia s n stupňami voľnosti. Vidíme, že p-hodnota $0.3719 > 0.05$, čiže dáta nenesú v sebe informáciu o tom, že by produkcie odrody *Gesha* boli v týchto dvoch kontinentoch rozdielne.

Predikovanie kvality lineárnou regresiou

Ako bolo spomenuté `TotalCupPoints` je skóre vypočítané z 10 rôznych metrík (aróma, chuť, dochuť, kyslosť, telo, balans, uniformita, čistota, sladkosť a celkový dojem), ktoré sú evaluované profesionálnymi „ochutnávačmi“. Samozrejme, moje amatérske zručnosti nie sú schopné spraviť takýto komplexný chuťový obraz. Bežne je však k dispozícii informácia o tom, v akej nadmorskej výške je káva pestovaná a akej odrody je. Čo sa týka samotných metrík, kyslosť a aróma sú dve, ktoré z môjho pohľadu sú prístupnejšie na zhodnotenie aj ľuďom, ako ja. Preto budeme modelovať kvalitu kávy `TotalCupPoints` na základe štyroch premenných – nadmorská výška, aróma, kyslosť a indikátor, či daná káva pochádza z odrody *Gesha*:

$$\mathrm{TotalCupPoints}_i = \beta_0 + \beta_1 \,\mathrm{Altitude}_i + \beta_2 \,\mathrm{Aroma}_i + \beta_3 \,\mathrm{Acidity}_i + \beta_4 \,\mathrm{Gesha}_i + \varepsilon_i$$

Na nájdenie optimálnych parametrov použijeme metódu najmenších štvorcov. Optimálne hodnoty parametrov sú nasledovné:

$$\hat{\beta} = (33.7, 0.0000138, 2.76, 3.72, 0.146)$$

Vidíme, že nárast v ľubovoľnej premennej spôsobí nárast predikovanej hodnoty. Pre účely porovnávania signifikancie vplyvov, preškálujeme všetky premenné okrem `Gesha` na nulový priemer a jednotkovú varianciu. Potom optimálne parametre vyzerajú takto:

$$\hat{\beta}_{\text{scaled}} = (83.68, 0.0068, 0.802, 0.972, 0.146)$$

Z výsledkov vidíme, že nárast kyslosti má najväčší vplyv na predikovanú hodnotu, spolu s arómou. Tento výsledok nie je prekvapujúci, keďže aj z týchto dvoch hodnôt sa skutočná hodnota počíta. Ďalej budeme v práci používať iba koeficienty pôvodného (nenaškálovaného) modelu.

Poznamenajme, že Shapirov-Wilkov test normality na rezíduách modelu vyšiel s p-hodnotou 0.1162, čiže nezamietol ich normalitu, takže môžeme používať nasledovné metódy.

Otestujme, či skutočná hodnota β_1, teda koeficient pri nadmorskej výške, sa nerovná nule:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

Použijeme t-test pre kontrastový vektor a ($n = 202$ je počet pozorovaní, $k = 5$ je počet parametrov v našom modeli):

$$\begin{aligned} \mathrm{SSE} &= \sum_{i=1}^{n} (y_i - x_i^T \hat{\beta})^2 \\ S &= \frac{\mathrm{SSE}}{n-k} \\ T &= \frac{a^T \hat{\beta}}{S\sqrt{a^T (X^T X)^{-1} a}} \\ p\text{-value} &= 2(1 - \mathrm{pt}_{(n-k)}(|T|)) \end{aligned}$$

Pre kontrastový vektor $a = (0, 1, 0, 0, 0)$ zodpovedajúci hypotéze H_0 dostávame p-hodnotu $0.8566 > 0.05$, čiže nemôžeme zamietnuť, že skutočný vplyv nadmorskej výšky je nulový. Rovnako môžeme otestovať parameter β_4 pre premennú `Gesha`. Pre vektor $a = (0, 0, 0, 0, 1)$ je p-hodnota $0.2173 > 0.05$, čiže rovnako nemôžeme tvrdiť, že skutočný vply odrody *Gesha* na kvalitu kávy je nenulový.

Rovnakým spôsobom môžeme otestovať, či to, že $\hat{\beta}_2 = 2.76$ a $\hat{\beta}_3 = 3.72$ je signifikantný rozdiel, alebo skutočné β_2 a β_3 môžu byť zhodné.

$$H_0 : \beta_2 = \beta_3 \quad \text{vs.} \quad H_1 : \beta_2 \neq \beta_3$$

Použijeme kontrastový vektor $a = (0, 0, 1, -1, 0)$, p-hodnota je $0.0083 < 0.05$. Tento výsledok teda hovorí, že skutočný vplyv arómy a acidity na výslednú kvalitu kávy v našom modeli nie je rovnaký.

Ďalej sa môžeme pozrieť, či náš model funguje lepšie, ako submodel, ktorý predikuje iba konštantným členom:

$$\mathrm{TotalCupPoints}_i = \beta_0 + \varepsilon_i$$

Chceme zistiť, či náš model má sumu kvadratických chýb nižšiu, ako takýto jednoduchý model (túto druhú sumu označíme SSE_{submodel}):

$$H_0 : SSE_{\text{submodel}} \leq SSE \quad \text{vs.} \quad H_1 : SSE_{\text{submodel}} > SSE$$

Použijeme prispôsobený F-test:

$$\begin{aligned} F &= \frac{\frac{SSE_{\text{submodel}} - SSE}{k-1}}{\frac{SSE}{n-k}} \\ p\text{-value} &= (1 - \mathrm{pf}_{(k-1),(n-k)}(F)) \end{aligned}$$

Funkcia $\mathrm{pf}_{(m),(n)}(\dots)$ je kumulatívna distribučná funkcia F rozdelenia so stupňami voľnosti (m, n). Pre hodnoty $SSE = 53.26$ a $SSE_{\text{submodel}} = 610.69$ dostávame pre tento test p-hodnotu rovnú 0 (aproximované softvérom R), čiže môžeme tvrdiť, že náš model predikuje kvalitu kávy lepšie, ako konštantný model.

Nakoniec sa pozrieme na to, či má zmysel konštruovať dva rôzne modely pre dva rôzne spôsoby spracovania kávy – *natural* a *washed*. Štruktúra modelu bude rovnaká ako vyššie, jedine odhad β skonštruujeme pomocou dát relevantných pre danú metódu:

$$\hat{\beta}^{(\text{washed})} = (35.22, -0.000054, 2.51, 3.79, 0.28)$$
$$\hat{\beta}^{(\text{natural})} = (33.65, 0.00023, 3.39, 3.06, 0.23)$$

Otestujme, či sú predikované nadroviny paralelné:

$$H_0 : \beta_i^{(\text{washed})} = \beta_i^{(\text{natural})} \quad \text{vs.} \quad H_1 : \beta_i^{(\text{washed})} \neq \beta_i^{(\text{natural})} \quad i \in 1, 2, 3, 4$$

Môžeme na to využiť kontrastový test, ak spojíme tieto dva modely „do jedného“. Vhodným dizajnom kontrastového vektora potom môžeme testovať, či sú dané parametre zhodné:

$$\gamma = (\beta^{(\text{washed})}, \beta^{(\text{natural})})^T$$
$$\tilde{y} = (\mathrm{TotalCupPoints}^{(\text{washed})}, \mathrm{TotalCupPoints}^{(\text{natural})})^T$$
$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}^{(\text{washed})} & \mathbf{0}_{122\times 5} \\ \mathbf{0}_{45\times 5} & \mathbf{X}^{(\text{natural})} \end{bmatrix}$$

Upravený model teda vyzerá $\tilde{y} = \tilde{\mathbf{X}}\gamma + \varepsilon$. Pre hypotézy vyššie skonštruujeme kontrastový vektor $a = (0, 1, 1, 1, 1, 0, -1, -1, -1, -1)$, s ktorým nám kontrastový test pre hodnoty $a, \gamma, \tilde{y}$ a $\tilde{\mathbf{X}}$ vráti p-hodnotu $0.7814 > 0.05$. Nemôžeme teda zamietnuť, že predikčné nadroviny sú paralelné, teda že vplyvy nadmorskej výšky, arómy, acidity a odrody *Gesha* sú rovnaké pre rôzne spôsoby spracovania kávy, jedine konštantný posun môžu mať rozdielny.

Môžeme do tohto testu zahrnúť aj porovnanie *intercept* člena. Kontrastový vektor $a = (1, 1, 1, 1, 1, -1, -1, -1, -1, -1)$ zodpovedá hypotézam:

$$H_0 : \beta^{(\text{washed})} = \beta^{(\text{natural})} \quad \text{vs.} \quad H_1 : \beta^{(\text{washed})} \neq \beta^{(\text{natural})}$$

Z t-testu dostávame p-hodnotu $0.5889 > 0.05$, teda nemôžeme ani zamietnuť, že sa jedná o totožné modely. Ukazuje sa teda, že modelovať kvalitu kávy separátne pre rôzne procesy spracovania kávy nepridáva žiadnu relevantnú informáciu oproti spojenému modelu.