# Zadanie 1

Plik **truffles.gdt** zawiera dane o ilości trufli sprzedanych na rynku w uncjach (*q*), ceny trufli wyrażonej w $ za uncję (*p*), ceny trufli sztucznych w $ za uncję (*ps*), miesięcznym dochodzie rozporządzalnym per capita w tys. $ (*di*) oraz godzinowej cenie wynajmu świni truflowej (*pf*).

1. Oszacuj parametry równania:
2. Zbadaj istotność modelu oraz każdej ze zmiennych osobno.
3. Czy na podstawie danych możemy odrzucić hipotezę, że każde 1000 dolarów miesięcznych dochodów rozporządzalnych prowadzą do spadku podaży trufli na rynku o 0,589 uncji?
4. Czy wszystkie zmienne istotne zostały uwzględnione przy budowie modelu z podpunktu *a*? Podaj specyfikację najlepszego modelu według jego łącznej istotności i współczynnika .

# Zadanie 1 rozwiązanie

1. W celu oszacowania parametrów równania wymienionym w podpunkcie a zbudujemy model regresji liniowej za pomocą klasycznej metody najmniejszych kwadratów. Wchodzimy do zakładki Model i wybieramy pierwszy model z listy rozwijanej, czyli KMNK. Jaką zmienną objaśnianą (zależną) ustawiamy ilość trufli sprzedanych, czyli ку. Natomiast jako zmienne objaśniające (niezależne) ustawiamy miesięczny dochód (di) oraz cena na trufli (p). Otrzymujemy następujący model…
2. W modelu oszacowany jest dwa estymatory di oraz p a także mamy konstant, żeby zbadać istotność modelu trzeba popatrzyć na wyniki testu F. p-value danego testu jest równa 0,011 co oznacza, że model nie jest istotny przy współczynniku istotności 0,01 bo p-value jest większe od tego współczynnika, natomiast gdy zakładamy, że nasz współczynnik będzie równy 0,05 odrzucamy hipotezę o nie istotności modelu na rzecz, że nasz model jest istotny. Tę gwiazdki pokazują przy zmiennych pokazują stopień parametrów, a także patrzymy na wyniki testu t-studenta i także porównuje wartości p z wartościami krytycznymi czyli 0,01; 0,05; 0,1. Okazuje się, że zmienna di nie jest istotna przy żadnej z zadanych wartości krytycznych, a zmienna p jest istotna kiedy współczynnik istotności jest równy 0,05 lub więcej.
3. Żeby prawidłowo wykonać podpunkt c musimy najpierw popatrzeć na parametr stojący przy zmiennej *di*. Na pierwszy rzut wydaje się że rzeczywiście spadek dochodów rozporządzalnych o 1000 dolarów prowadzi do spadku podaży trufli, skoro wartość parametru jest ujemna. Jednakże musimy nie zapomnieć o tym, że zmienna musi być istotna. W tym modelu wartość p, zarówno jak i statystyka testowa t wskazują na to, że zmiana dochodu rozporządzalnego nie wpływa na zmianę podaży trufli w sposób statystycznie istotny. Czy odpowiedź na to pytanie jest nie, bo zmienna nie jest istotna.
4. Przechodząc do podpunktu *d*, żeby sprawdzić czy wszystkie istotne zmienne zostały uwzględnione, musimy wykonać test dodanych zmiennych. W tym celu wchodzimy w zakładkę *Testy* naszego modelu i wybieramy odpowiedni test. W okienku dodajemy jako zmienne objaśniające pozostałe zmienne ze zbioru, czyli cenę trufli sztucznych za uncję *ps* i cenę wynajmu świni truflowej *pf.* W wyniku mamy rozszerzony model, który otworzy się w nowym okienku. W celu ustalenia czy naprawdę ten model jest lepszy musimy popatrzeć na wartość p przy statystyce testowej F i na skorygowany współczynnik R kwadrat. No i już wyraźnie widać że ten model jest znaaacznie lepszy od poprzedniego.

Jednakże nadal nie wiemy czy ten model jest najlepszym. W takim celu musimy jeszcze dokonać testów pominiętych zmiennych wyrzucając po kolei zmienne nieistotne. Wchodzimy znowu do zakładki testy i dodajemy zmienne nieistotne, czyli di i ps. I patrzymy na p-value dla łącznej statystyki F zarówno jak i na skorygowany R kwadrat. I widzimy że w tym przypadku r kwadrat i p-value się wzrosły więc możemy stwierdzić że ten model jest lepszy w stosunku do modelu poprzedniego względem wymienionych w zadaniu kryterium.

# Zadanie 2

Plik **vacation.gdt** zawiera dane o ilości mil przejechanych rocznie (*mile*), przychodach rocznych wyrażonych w tys. dolarów (*income*), przeciętnemu wieku osoby dorosłej (*age*) oraz liczbie dzieci (*kids*) w gospodarstwie domowym w Stanach Zjednoczonych.

1. Zbuduj model, w którym *miles* zależy od zmiennych *income* i *age*. Porównaj specyfikację liniową względem specyfikacji wielomianowej 2 stopnia. Przeprowadź test RESET.
2. Powtórz czynności z punktu a dla modelu, w którym zmienną objaśniającą *income* jest zlogarytmizowana. Zinterpretuj oszacowania parametrów strukturalnych.
3. Wprowadź zmienną interakcyjną *kids* do modelu z punktu *b*. Jak się różni ilość mil przejechanych rocznie dla osób z trzema dziećmi w porównaniu do gospodarstw domowych bez dzieci.
4. Popraw specyfikację z podpunktu *c* w ten sposób żeby wszystkie kategorię zmiennej *kids* były statystycznie istotne.

# Zadanie 2 rozwiązanie

1. Najpierw budujemy model regresji gdzie za zmienną zależną ustawiamy *miles*, a w komórce regresory ustawiamy zmienne *income* oraz *age*. W wyniku otrzymujemy liniową zależność miedzy współczynnikami. Dla sprawdzenia poprawności specyfikacji korzystamy z testu RESET. Żeby jego przeprowadzić wchodzimy w zakładkę *Testy*, i wybieramy punkt *test specyfikacji Ramsey RESET*. Dalszym krokiem jest wybór możliwości przeprowadzenia testu specyfikacji, gdzie mamy 4 opcji do wyboru. W zadaniu mamy sprawdzić specyfikacji względem specyfikacji wielomianowej 2 stopnia, co oznacza, że musimy wybrać punkt *Tylko kwadrat zmiennej*. Wybieramy ten punkt i naciskamy OK. Otrzymujemy następującą tablicę w której nas interesuje tylko statystyka testu F. p-value jest równe 0,86, co oznacza, że jest bardzo większa od wartości krytycznych, i nie mamy podstaw do odrzucenia H0 która mówi, że nasz model jest dobre wyspecyfikowany.
2. W następnym punkcie musimy zrobić to samo co i punkcie poprzednim tylko najpierw zlogarytmować zmienną *income*. Dla tego musimy wybrać zmienną *income*, wchodzimy w *Dodawanie zmiennych* która znajduje się w głównym okienku Gretla, i wybieramy pierwszy punkt z listy, czyli *logarytmy dla wybranych zmiennych*.

Można także zrobić to w inny sposób. Naciskamy prawym przyciskiem myszy na zmienną którą chcemy zlogarytmować i wybieramy punkt *Dodaj logarytm zmiennej*. Nie ma różnicy z jakiego podejścia skorzystać, ale kiedy musimy zlogarytmować dużo zmiennych warto korzystać z funkcji w okienku dodawanie zmiennych.

Teraz możemy przystąpić do budowaniu modelu. Wchodzimy do *Model -> KMNK,* ale teraz zamiast regresora *income* wstawiamy naszą nową zmienną *l\_income,* co oznacza logarytm tej zmiennej. Otrzymujemy następujący model. Teraz trzeba poprawnie jego zinterpretować. Interpretacja zmiennej niezależnej age się nie zmieni, bo nadal jest liniowa zależność między *age* a *miles*. Zależność pomiędzy dochodem (*income*) a ilością przejechanych mil (*miles*) możemy zinterpretować następująco: ze wzrostem ilości dochodu o 1%, ilość przejechanych mil wzrośnie o beta/100, czyli o 824,5/100 = 8,24 mil. I taka jest interpretacji liniowo-logarytmiczna (lin-log).

Teraz trzeba sprawdzić poprawność specyfikacji modelu za pomocą Testu RESET. Jak w poprzednim punkcie wchodzimy w zakładkę *Testy* w modelu który już oszacowaliśmy, i wybieramy punkt *test specyfikacji Ramsey RESET*. Sprawdzamy specyfikację względem specyfikacji wielomianowej 2 stopnia, co oznacza, że musimy wybrać punkt *Tylko kwadrat zmienne*j. Naciskamy OK. Znowu patrzymy na wartości p-value Testu F. p-value jest wyższe od minimalnego poziomu istotności, wtedy nie możemy odrzucić H0 o tym, że nasz model jest dobrze wyspecyfikowany.

1. Żeby uwzględnić zmienną *kids* jaką zmienną interakcyjną musimy najpierw dać gretlowi zrozumieć ze chcemy żeby uznał *kids* za taką zmienną. Dlatego zanim zbudujemy model, przechodzimy do zakładki *Dodawanie zmiennych* na wejściowym ekranie. W liście rozwijanej wybierzemy *zmienne zero-jedynkowe dla wybranych dyskretnych zmiennych* z opcją przekodowania **wszYstkich** wartości zmiennej dyskretnej. Mówiąc po ludzku, po prostu rozdzielimy tą zmienną na 5 odrębnych zmiennych, gdyś każda będzie odpowiadała za konkretną wartość dzieci w rodzinie. Dla przykładu zmienna *Dkids\_cztery* będzie równa się 1 dla każdej obserwacji gdzie w rodzinie było 4 dziecka i zero dla każdej obserwacji z inną liczbą dzieci. Teraz musimy ustalić która z tych sztucznych zmiennych będzie punktem odniesienia przy interpretacji modelu. Nie możemy dodać je wszystkich bo inaczej będziemy mieli do czynienie z dokładną współliniowością czego baardzo nie chcemy. Patrząc na opis podpunktu c, musimy porównać *Dkids\_trzy* z *Dkids\_zero*, wiec wybieramy *Dkids* zero jako punkt odniesienia.

Jesteśmy gotowi do wprowadzenia zmiennej interakcyjnej więc wybieramy znów model KMNK i w okienku dodajemy zmienne od *Dkids\_jeden* do *Dkids\_4*. I teraz po prostu musimy zinterpretować wartość parametru strukturalnego. Wychodzi że liczba mil przejechanych rocznie przez rodzinę z 3 dziećmi spada o 201,65 w porównaniu do rodzin bez dzieci.

1. W poprzednim punkcie mieliśmy za małą statystykę testową t dla zmiennej *Dkids\_1*. Żeby poprawić istotność każdej z kategorii zmiennej *kids* możemy zmniejszyć odpowiednio liczbę tych kategorii. Skoro jedno dziecko nie wpłynęło na różnicę dochodów to możemy zgrupować rodziny z jednym dzieckiem i rodziny bez dzieci. Natomiast drugą grupą będą rodziny z dwoma plus dziećmi. Stworzymy w tym celu dodatkową zmienną *kids\_bin*, która będzie przyjmowała wartości zero dla rodzin z 0 lub 1 dzieckiem i 1 dla rodzin z chociażby dwoma dziećmi. W tym celu wchodzimy do zakładki dodawanie zmiennych i przechodzimy na samy dół, gdzie wybieramy opcję *Definiowanie nowej zmiennej*. Tutaj dopóki wpiszemy tylko samą nazwę i zamkniemy okienko edycji wartości. Dalej przechodzimy do zakładki *Zmienna* i wybieramy tutaj *edycję atrybutów*. I w nowym okienku zmieniamy opis na formułę. Skoro chcemy żeby ta wartość przyjmowała wartości jeden dla rodzin z 2+ dziećmi, wpiszemy tutaj polecenie *kids* >= 2. Klikamy okej i widzimy ze pojawiła się nowa zmienna *kids\_bin*, możemy nawet wejść i sprawdzić czy na pewno wartości są okej. No i teraz pozostaje tylko dodać tą zmienną do naszego modelu. Więc kolejny raz wchodzimy do *Model -> KMNK* i dodajemy tutaj zmienną *kids\_bin*. I jak możemy zauważyć, statystyka t jest bardzo wysoka. Co więcej poprawiła się wartość skorygowanego współczynnika determinacji R kwadrat jak i łączna istotność modelu F.