Ekonometria I (Wykład nr 10)

1 Zmienne jakościowe w modelach ekonometrycznych.

Czynniki jakościowe często mają postać informacji binarnych: osoba jest kobietą lub mężczyzną; osoba jest właścicielem komputera osobistego lub jenie posiada komputera; firma oferuje pewien rodzaj pracowniczego programu emerytalnego lub nie; państwo oferuje powszechne ubezpieczenie zdrowotne lub tego nie robi. We wszystkich tych przykładach odpowiednie informacje można uchwycić, definiując zmienną binarną czyli zmienną zero-jedynkową. W ekonometrii najczęściej występują zmienne binarne nazywane zmienne fikcyjne (dummy variables), chociaż ta nazwa nie jest szczególnie opisowa. Definiując zmienną fikcyjną, musimy zdecydować, które zdarzenie ma przypisaną wartość jeden a któremu przypisano wartość zero. Na przykład w badaniu indywidualnych płac możemy zdefiniować zminną *kobieta* jako zmienną binarną przyjmującą wartość jeden dla kobiet i wartość zero dla mężczyzn. Nazwa w tym przypadku wskazuje zdarzenie, któremu przypisujemy wartość jeden. Ta sama informacja jest zanotowana przez zdefiniowanie zmiennej *mężczyzna* jako jednego jeśli osoba jest mężczyzną i zero, jeśli osoba jest kobietą. Każda z nich jest lepsza niż używaje zmiennej *płeć*, ponieważ ta nazwa nie wyjaśni, kiedy zmienna fikcyjna jest równa jeden. Czy *płeć=1* odpowiada mężczyźnie czy kobiecie? To, jak nazywamy zmienne nie jest nieistotne dla wyników regresji, ale zawsze pomaga wybór nazw, które wyjaśniają znaczenie przyjmowanych przez zmienne modelu wartści. Załóżmy w modelu opisującym płace, że wybraliśmy nazwę zmiennej opisujacej płeć osoby *kobieta*. Ponadto definiujemy zmienną binarną *w-zwazku* równą jeden, jeśli dana osoba jest w związku małżeńskim i zero, jeśli jest inaczej.
Dlaczego używamy wartości zero i jeden do opisu informacji jakościowych? W sensie, wartości te są dowolne: wystarczyłyby dowolne dwie różne wartości. Prawdziwa korzyść z przechwytywania informacji jakościowych z wykorzystaniem zmiennych zero-jedynkowych wynika z faktu, że prowadzą one do modeli regresji, w których parametry mają bardzo naturalne interpretacje co teraz zilustrujemy.

1.1 Pojedyncza zero-jedynkowa zmienna opisowa

Jak uwzględnić informacje binarne w modelu regresji? W najprostszym przypadku gdy należy uwzględnić tylko jedną zmienną objaśniającą tego typu, dodajemy ją jako niezależną zmienną w równaniu. Rozważmy na przykład następujący prosty model płacy godzinowej:

$$placa = \beta_0 + \delta_0\, kobieta + \beta_1\, edu + u. \tag{1}$$

Używamy δ_0 jako parametru dla zmiennej *kobieta* w celu podkreślenia interpretacji parametru. Używamy notacji, która jest najbardziej wygodna do interpretacji. W modelu (1) tylko dwa zaobserwowane czynniki wpływają na wynagrodzenie: płeć i wykształcenie. Ponieważ *kobieta=1*, gdy osoba jest kobietą, i *kobieta= 0*, gdy osoba jest mężczyzną, parametr δ_0 ma następującą interpretację: δ_0 jest różnicą w stawkach godzinowych dla kobiet i dla mężczyźn, biorąc pod uwagę taki sam poziom wykształcenia (i ten sam błęd u). Więc współczynnik δ_0 określa, czy istnieje dyskryminacja kobiet: jeśli $\delta_0 < 0$, to, przy takim samym poziomie innych czynników kobiety zarabiają średnio mniej niż mężczyźni. W terminach wartości oczekiwanych, jeśli przyjmiemy zerową wartość warunkowe wartości oczekiwanej

$$\mathrm{E}(u \mid kobieta, edu) = 0,$$

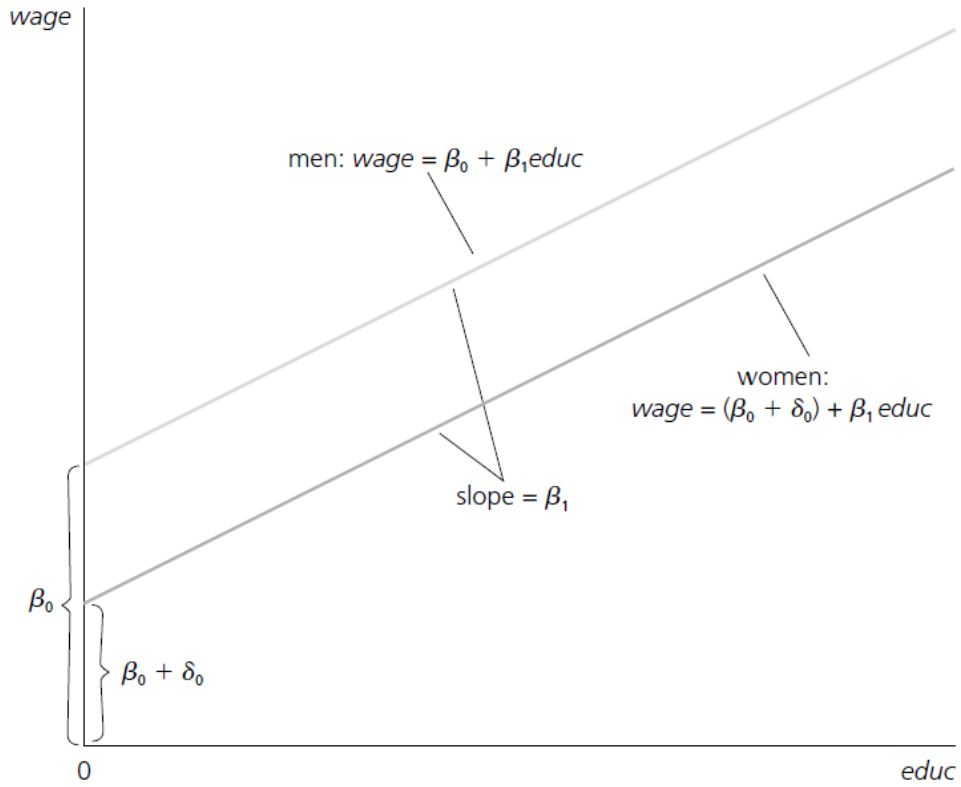
wtedy

$$\delta_0 = \mathrm{E}(placa \,|\, kobieta = 1,\, edu) - \mathrm{E}(placa \,|\, kobieta = 0,\, edu).$$

Ponieważ *kobieta= 1* odpowiada kobietom, a *kobieta=0* mężczyznom, możemy napisz to prościej jako

$$\delta_0 = \mathrm{E}(placa \,|\, kobieta,\, edu) - \mathrm{E}(placa \,|\, mezczyzna,\, edu). \tag{2}$$

Tutaj istotne jest to, że poziom wykształcenia jest taki sam dla obu wartości oczekiwaniach i różnica, δ_0, wynika wyłącznie z płci. Sytuację można przedstawić graficznie jako przesunięcie wyrazu wolnego równania prostej regresji dla mężczyz i dla kobiet. Na rysunku 1 pokazano przypadek $\delta_0 < 0$, dzięki czemu mężczyźni zarabiają ustaloną kwotę więcej na godzinę niż kobiety.



Rysunek 1: Wykres zależności $placa = \beta_0 + \delta_0\, kobieta + \beta_1\, edu$ dla $\delta_0 < 0$.

Różnica nie zależy od poziomu wykształcenia oraz wyjaśnia to, dlaczego profile opisujące zależnoć płacy od wykształcenia dla kobiet i mężczyzn są równoległe. W tym momencie możemy się zastanawiać, dlaczego nie uwzględniamy również w (1) zmiennej binarnej, powiedzmy *mezczyzna*, czyli jeden dla mężczyzn i zero dla kobiet. To byłoby to jednak zbędne. W (1) wyraz wolny prostej dla mężczyzn wynosi β_0, a dla kobiet wynosi $\beta_0 + \delta_0$. Ponieważ są tylko dwie grupy, potrzebujemy tylko dwóch różnych wyrazów wolnych. Oznacza to, że oprócz β_0 musimy użyć tylko jednej zmiennej binarnej. Postanowiliśmy uwzględnić zmienną *kobieta*. Zastosowanie dwóch zmiennych wprowadziłoby idealną kolinearność, ponieważ

$$kobieta + mezczyzna = 1,$$

co oznacza, że zmienna *mezczyzna* jest funkcją liniową zmiennej *kobieta*. Włącznie zmiennych binarnych dla obu płci jest najprostszym przykładem tak zwanej pułapki zmiennych fikcyjnych, która powstaje, gdy zbyt wiele zmiennych opisuje ustaloną liczbę grup. Omówimy ten

problem później. W (1) wybraliśmy mężczyzn jako grupę bazową lub grupę porównawczą, tj. grupę z którą dokonuje się porównań pozostałych grup. W naszym przypadku grupy kobiet. Właśnie dlatego β_0 jest wyrazem wolnym dla mężczyzn, a δ_0 to różnica pomiędzy wyrazem wolnym dla kobietami i wyrazem wolnym dla mężczyzn. Moglibyśmy wybrać kobiety jako grupę podstawową, pisząc model jako

$$placa = \alpha_0 + \gamma_0\, mezczyzna + \beta_1\, edu + u,$$

gdzie wyraz wolny dla kobiet wynosi α_0, a wyraz wolny dla mężczyzn wynosi $\alpha_0 + \gamma + 0$. To implikuje że $\alpha_0 = \beta_0 + \delta_0$ i $\alpha_0 + \gamma_0 = \beta_0$. W konstrukcji modelu nie ma znaczenia, jak wybierzemy grupę podstawowa, ale ważne jest, aby śledzić, która grupa jest grupą podstawową. Niektórzy badacze wolą nie uwzględniać w modelu ogólnego wyrazu wolnego i dołączyć zmienne dla każdej grupy. Równaniem byłoby wówczas

$$placa = \beta_0\, mezczyzna + \alpha_0\, kobieta + \beta_1\, edu + u,$$

gdzie wyraz wolny dla mężczyzn wynosi β_0, a wyraz wolny dla kobiet wynosi α_0. Zauważmy, że w tym przypadku nie ma pułapki zmiennej fikcyjnej, ponieważ nie mamy ogólnego wyrazu wolnego. Jednak, tak skonstruowany model okazuje się trudniejszy do analizy. Dlatego zawsze będziemy uwzględniać ogólny wyraz wolny dla grupy podstawowej. Nic się nie zmienia, gdy w grę wchodzi więcej zmiennych objaśniających. Biorąc mężczyzn jako grupa podstawowa, model, który kontroluje na przykad doświadczenie i staż pracy oprócz edukacji ma postać

$$placa = \beta_0 + \delta_0\, kobieta + \beta_1\, edu + \beta_2\, dosw + \beta_3\, staz + u. \tag{3}$$

Jeśli edukacja, doświadczenie i staż są istotnymi cechami wydajności, hipoteza zerowa mówiąca, że nie ma żadnej różnicy między mężczyznami i kobietami ma postać

$$H_0 :\ \delta_0 = 0.$$

Podobnie hipoteza, że jest dyskryminacja kobiet ma postać

$$H_1 :\ \delta_0 < 0.$$

Jak możemy faktycznie przetestować płacę pod kątem dyskryminacji płacowej? Odpowiedź jest prosta: wystarczy oszacować model przy użyciu metody najmniejszych kwadratów, dokładnie tak jak poprzednio, i użyć zwykłej statystyki $t-$Studenta. Nic się nie zmienia w metodzie najmniejszych kwadratów ani w jej teorii statystycznej, gdy niektóre ze zmiennych niezależnych są zdefiniowane jako zmienne binarne. Jedyną różnicą w stosunku do tego, co zrobiliśmy do tej pory, jest w interpretacji współczynnika dla zmienne binarnej.

Przykład 1. Korzystając z danych w WAGE1 estymujemy model (3). Używamy płacy jako zmiennej zależnej:

$$\widehat{placa} = -1,57 - 1,81\, kobieta + .572\, edu + 0,025\, dosw + 0,141\, staz. \tag{4}$$

$$n = 526, \quad R^2 = 0,364.$$

Ujemne wartość wyrazu wolnego - w tym przypadku wyraz wolny dla mężczyzn - nie ma zbyt dużego znaczenia, ponieważ nikt w próbie nie ma jednocześnie zerowych wartości dla wszystkich zmiennych: wykształcenie, doświadczenie i staż pracy. Współczynnik dla zmiennej *kobieta* jest interesujący, ponieważ mierzy średnią różnicę w stawce godzinowej między a kobietami i mężczyznami, mającymi takie same poziomy wykształcenia, doświadczenia i stażu pracy. Jeśli

weźmiemy kobietę i mężczyzna o tym samym poziomie wykształcenia, doświadczenia i stażu pracy, to z modelu wynika, że jak kobieta zarabia średnio 1,81 USD mniej na godzinę niż mężczyzna. (Zanim zaczniemy wyciągać wnioski dotyczące obecnej sytuacji w Polsce, proszę zwrócić uwagę, że są to płace w USA z 1976 r.) Należy pamiętać, że w modelu kontrolowaliśmy edukację, doświadczenie i staż pracy więc różnicy w płacy na poziomie 1,81 USD nie można tłumaczyć różnicą w średnim poziomie wykształcenia, doświadczenia lub stażu pracy mężczyzn i kobiet. Możemy wyciągnąć wniosek, że różnica w wysokości 1,81 USD wynika z płci lub czynników związanych z płcią. nie kontrolowaliśmy w regresji. Przydatne jest porównanie współczynnika dla kobiety w równaniu (4) z oszacowaniem, gdy wszystkie inne zmienne objaśniające zostaną usunięte z równania:

$$\widehat{placa} = 7,10 - 2,51\, kobieta \tag{5}$$
$$n = 526, \quad R^2 = 0,116.$$

Współczynniki w (5) mają prostą interpretację. Wyraz wolny to średnia płaca mężczyźn w próbie (niech *kobieta= 0*), więc mężczyźni zarabiają średnio 7,10 USD za godzinę. Współczynnik na kobietach jest różnica w średniej płacy między kobietami i mężczyznami. Tak więc średnia płaca dla kobiet w próbie wynosi $7,10 - 2,51 = 4,59$ lub 4,59 USD za godzinę. (Nawiasem mówiąc, próba obejmuje 274 mężczyzn i 252 kobiety). Równanie (5) zapewnia prosty sposób przeprowadzenia testu porównania średnich między dwiema grupami, którymi w tym przypadku są mężczyźni i kobiety. Szacunkowa różnica 2,51 ma wartość t statystyki na poziomie 8,37, która jest bardzo istotna statystycznie (i oczywiście 2,51 USD jest ekonomicznie również duże). Aby zwykły test t mógł być użyty, musimy założyć, że obowiązuje założenie homoskedastyczności. Oznacza to, że zakładamy iż wariancje płacy dla mężczyzn i dla kobiet są takie same. Szacowana różnica płac między mężczyznami i kobietami jest większa w (5) niż w (4) ponieważ (5) nie kontroluje różnic w wykształceniu, doświadczeniu i kadencji, i one są średnio niższa dla kobiet niż dla mężczyzn w tej próbie. Równanie (4) daje bardziej wiarygodne oszacowanie różnicy w wynagrodzeniach kobiet i mężczyzn i nadal wskazuje na bardzo dużą różnicę.

W wielu przypadkach fikcyjne zmienne niezależne odzwierciedlają wybory jednostek lub inne czynniki ekonomiczne jednostki (w przeciwieństwie do czegoś z góry określonego, takiego jak płeć). W takich sytuacjach kwestia związku przyczynowego jest ponownie centralną kwestią. W poniższym przykładzie chcielibyśmy wiedzieć, czy posiadanie komputera osobistego powoduje wyższą średnią na świadectwie.

Przykład 2. Aby określić wpływ posiadania komputera na średnią ocen, oszacujemy model postaci

$$srednia = \beta_0 + \delta_0\, PC + \beta_1\, sredniaP + \beta_2\, TP + u,$$

gdzie zmienna zmienna PC równa się jeden, jeśli uczeń posiada komputer osobisty, a zero w przeciwnym razie. Istnieje wiele powodów, dla których posiadanie komputera może mieć wpływ na *srednia*. Praca ucznia może być bardziej wydajna, jeśli odbywa się to z użyciem własnego komputerze, a czas można zaoszczędzić, nie musząc tego robić czekaj na komputer do momentu aż przestaną go używać rodzice (Microsoft Teams) lub rodzeństwo (Google Clasroom). Oczywiście uczeń może być bardziej skłonny do grania w gry komputerowe lub surfować po Internecie, jeśli on lub ona ma komputer, więc nie jest oczywiste, że δ_0 jest dodatnia. Zmienne *sredniaP* to średnia z szkoły podstawowej PT (wynik testu osiągnięć) są używane jako zmienne kontrole. Może być tak że silniejsi uczniowie, mierzeni wynikami ze szkoły podstawowej i gimnazium, są bardziej skłonni do posiadania komputerów. Kontrolujemy te czynniki, ponieważ chcielibyśmy poznać średni wpływ na wynik w szkole średniej, jeśli uczeń jest wybierany

losowo i otrzymuje komputer osobisty. Korzystając z danych w GPA1, otrzymujemy

$$\widehat{srednia} = 1,26 + 0,157\,PC + 0,447\,sredniaP + 0,0087\,TP. \tag{6}$$

$$n = 141, \quad R^2 = 0,219.$$

To równanie sugeruje, że uczń będący właścicielem komputera ma przewidywanś średnią około 0,16 punktów wyższą niż porównywalny uczeń bez komputera (pamiętaj, że zarówno *srednia,* jak i *sredniaP* są włączone na czteropunktowej skali). Efekt jest również bardzo istotny statystycznie, z $t_PC = 0,157/0,057 = 2,75$. Co się stanie, jeśli usuniemy *sredniaP* i TP z równania?

Każdy z poprzednich przykładów można uznać za mający znaczenie dla analizy polityki. W pierwszym przykładzie byliśmy zainteresowani dyskryminacją płciową wśród siły roboczej. W Drugi przykład dotyczył wpływu własności komputera na politykę edukacyjną. Szczególnym przypadkiem analizy politycznej jest ewaluacja programu, tam gdzie chcemy znać wpływ programów gospodarczych lub społecznych na obywateli, firmy, dzielnice, miasta i tak dalej. W najprostszym przypadku mamy dwie grupy. Grupa kontrolna nie uczestniczy w programie. Grupa eksperymentalna lub grupa terapeutyczna bierze udział w programie. Te nazwy pochodzą z literatury nauk eksperymentalnych i nie powinny być brane zbyt dosłownie. Z wyjątkowo rzadkich przypadków wybór grupy kontrolnej i terapeutycznej nie jest losowy. Jednak w niektórych przypadkach do kontroli innych czynników można zastosować wielowymiarową regresję aby oszacować przyczynowo-skutkowy wpływ programu.

Przykład 3. Korzystając z danych z 1988 r. dla firm produkcyjnychw JTRAIN, otrzymujemy następujace równanie

$$\widehat{godztrn} = 46,67 + 26,25\,dotacja - 0,98\,\log(sprzedaz) - 6,07\,\log(zatrdienie) \tag{7}$$

$$n = 105, \quad R^2 = 0,237.$$

Zmienna zależna to liczba godzin szkolenia przypadająca na pracownika na poziomie firmy. Zmienna dotacja jest zmienną fikcyjną równą jeden, jeśli firma otrzymała dotacje na szkolenie i zero w innym przypadku. Zmienne sprzedaż i zatrudnienie reprezentują odpowiednio roczną sprzedaż i liczbę pracowników. Nie możemy wprowadzić *godztrn* w postaci logarytmicznej, ponieważ *godztrn* wynosi zero dla 29 z 105 firm wykorzystanych przy estymacji modelu regresji. Zmienna *dotacja* jest bardzo istotna statystycznie, z $t_dotacja = 4,70$. Kontrolujac sprzedaż i zatrudnienie możemy stwierdzić, że firmy, które otrzymały dotację, przeszkoliły każdego pracownika średnio o 26,25 godziny dłużej. Ponieważ średnia liczba godzin szkolenia na pracownika w próbie wynosi około 17, z maksymalna wartość wynosząca 164, zgodnie z oczekiwaniami, dotacja ma duży wpływ na szkolenie. Współczynnik dla $\log(sprzedaz)$ jest niewielki i nieistotny. Współczynnik dla $\log(zatrudnienie)$ oznacza, że jeśli firma jest o 10% większa, szkoli swoich pracowników o około 0.61 godzin mniej. Dla tego współczynnika wartość t statystyki to 1,56, co jest jedynie w niewielkim stopniu istotne statystycznie.

Jak w przypadku każdej innej zmiennej niezależnej, powinniśmy zapytać, czy zmierzony efekt zmienna jakościowej ma wpływ na analizowaną zmienną objaśnianą . W równaniu (7) jest różnica w szkoleniu między firmami które otrzymują dotacje i tymi, które nie korzystały z dotacją. Moze jednak otrzymanie dotacji jest jedynie wskaźnikiem czegoś innego? Możliwe, że firmy otrzymujące dotacje szkoliły by swoich pracowników średnio dłużej nawet przy braku dotacji. Nic w tej analizie nie mówi nam czy oszacowaliśmy efekt przyczynowy; musimy wiedzieć, w jaki sposób firmy otrzymujące dotacje zostały wybrane. Możemy mieć tylko nadzieję, że udało nam się kontrolować tak wiele czynników, jak to możliwe aby można było powiązać fakt czy firma otrzymała dofinansowanie z poziomem szkolenia.

1.2 Interpretacja zmiennych jakościowych gdy zmienna objaśniana jest postaci $\log(y)$.

Często w zastosowanich zmienna objaśniana pojawia się w postaci logarytmicznej, a jedna lub więcej zmiennymi jakościowymi jako zmienne objaśniające. W jaki sposób w tym przypadku interpretujemy współczynniki dla tych zmiennych?

Przykład 4. *(Regresja cen domów).*
Korzystając z danych w HPRICE1, otrzymujemy równanie

$$\widehat{\log(cena)} = -1,35 + 0,168\log(dziaka) + 0,707\log(metraz) + 0,027\,lpokoi + 0,054\,styl \quad (8)$$

$$n = 88, \quad R^2 = 0,649.$$

Wszystkie zmienne są oczywiste, z wyjątkiem styl, która jest zmienną binarną równą zero jeśli dom jest z czasów PRL. Co oznacza współczynnik styl? Dla ustalonego poziomy powierzchi działki, metrażu i liczby pokoii, różnica w $\log(cena)$ między domem w innym stylu i stylu z PRL jest 0,054. Oznacza to, że przewiduje się cena domu w stylu innym niż PRL jest o 5,4% większa, przy ustalonych innych czynnikach.

Ten przykład pokazuje, że gdy $\log(y)$ jest zmienną zależną w modelu, współczynnik dla zmiennej jakościowej, pomnożonej przez 100, jest interpretowana jako różnica procentowa w y, przy zachowaniu wszystkich innych czynników.

Przykład 5. *(Logarytm godzinowej stawki płacowej).*
Ponownie przeprowadźmy porównanie płac z pierwszego przykładu. Tym razem używając logarytmu wynagrodzenia jako zmiennej zależnej i kwadratów doświadczenia i kwadratu stażu pracy jako zmiennych objaśniających:

$$\widehat{placa} = -0,417 - 0,297\,kobieta + 0,080\,edu + 0,029\,dosw - 0,0058\,dosw^2 + 0,032\,staz - 0,0059\,staz^2. \quad (9)$$

$$n = 526, \quad R^2 = 0,441.$$

Stosując to samo przybliżenie jak w poprzednim przykładzie, współczynnik dla zmiennej *kobieta* sugeruje, że dla na tym samym poziomie wykształcenia, doświadczenia i stażu pracy kobiety zarabiają około $100 \times 0,297 = 29,7\%$ mniej niż mężczyźni. Możemy to zrobić lepiej, obliczając dokładną różnicę procentową w przewidywanych wynagrodzenach. Chcemy porównać różnicę w wynagrodzeniach kobiet i mężczyzn, utrzymując inne czynniki stałe:

$$(\widehat{placa_K} - \widehat{placa_M})/\widehat{placa_M}.$$

To, co otrzymalismy na podstawie (9), to

$$\widehat{\log(placa_K)} - \widehat{\log(placa_M)} = -0,297.$$

Nakładając funkcję wykładniczą i odejmując otrzymamy

$$(\widehat{placa_K} - \widehat{placa_M})/\widehat{placa_M}\exp(-0,297) - 1 = -0.257.$$

To dokładniejsze oszacowanie sugeruje, że wynagrodzenie kobiety jest średnio o 25,7% niższe niż porównywalne wynagrodzenie mężczyzny.

Gdybyśmy dokonali tej samej korekty w pierwszym przykładzie, uzyskalibyśmy $\exp(0,054) - 1 = 0,0555$ lub około 5,6%. Korekta ma mniejszy wpływ w przykładzie ceny domów niż w

w obecnym przykładzie, ponieważ wielkość współczynnika zmiennej jakościowej jest znacznie mniejszy w (8) niż w (9). Ogólnie jeśli, $\hat{\beta}_1$ jest współczynnikiem zmiennej jakościowej, powiedzmy x_1, gdy $\log(y)$ jest zmienną zależną, dokładna różnica procentowa w przewidywanej wartości y, gdy $x_1 = 1$ względem tej wartości gdy $x_1 = 0$ jest równa

$$100[\exp(\hat{\beta}_1) - 1]. \tag{10}$$

Oszacowanie $\hat{\beta}_1$ może być dodatnia lub ujemna i ważne jest, aby zachować jego znak gdy obliczamy (10).

1.3 Używanie zmiennych jakościowych z wieloma kategoriami

Możemy użyć kilku niezależnych zmiennych jakościowych w tym samym równaniu. Na przykład można dodać zmienną jakościową *w-zwiazku* do równania (9). Współczynnik dla zmiennej *w-zwiazku* daje przybliżoną proporcjonalną różnicę płac między tymi, którzy są i tymi, którzy nie są w związku małżeńskim przy tych samych wartościach pozostałych zmiennych niezależnych. Gdy oceniamy ten model, współczynnik przy zmiennej *w-zwiazku* (ze standardowym błędem w nawiasach) wynosi 0,053 (0,041), a współczynnik przy zmiennej *kobieta* jest równy 0,290 (0,036). Zatem szacuje się, że „premia za małżeńsko" wynosi około 5,3%, ale nie różni się statystycznie od zera (t statystyka na poziomie 1,29). Możemy mieć wątpliwości co do przewidywań opartych na tym modelu gdyż zakładamy w nim , że premia za małżeństwo jest taka sama dla mężczyzn i kobiet. Dlatego powinniśmy raczej rozważyć trochę ogólniejszy model.

Przykład 6. Oszacujmy model, który pozwala na zbadanie różnicy płac między czterema grupami: żonaci mężczyźni, zamężne kobiety, samotni mężczyźni i samotne kobiety. Aby to zrobić, musimy wybrać grupę podstawową; my wybierzemy samotnych mężczyzn. Następnie musimy zdefiniować zmienne jakościowe dla każdej z pozostałych grup. Nazwijmy te zmienne mM, mK i sK. Umieszczenie tych trzech zmiennych w (9) (i, oczywiście, opuszczając zmienną *kobieta*, ponieważ jest ona teraz zbędna) daje następujacy model

$$\widehat{\log(place)} = 0,321 + 0,213\,mM - 0,198\,mK \tag{11}$$

$$- \; 0,110\,sK + 0,079\,eduk + 0,027\,dosw - 0,00054\,dosw^2 \tag{12}$$

$$+0,029\,staz - 0,00053\,staz^2$$

$$n = 526, \quad R^2 = 0,461.$$

Wszystkie współczynniki, z wyjątkiem sK, mają t statystyki znacznie powyżej co do modułu dwa. Statystyka t dla sF wynosi około 1,96, co jest znaczące przy poziomie istotności 95% w przypadku dwustronnej alternatywy. Aby zinterpretować współczynniki zmiennych jakościowych, musimy pamiętać, że grupa podstawowa to samotni mężczyźni. Zatem oszacowania trzech zmiennych jakościowych mierzą proporcjonalną różnicę w wynagrodzeniu w stosunku do wynagrodzenia samotnych mężczyzn. Na przykład szacuje się, że mężczyźni w związku małżeńskim zarabiają o 21,3% więcej niż samotni mężczyźni, posiadający ustalony poziom wykształcenia, doświadczenia i stażu pracy. Z drugiej strony zamężna kobieta zarabia prognozowane 19,8% mniej niż jeden samotny mężczyzna z tymi samymi poziomami innych zmiennych. Ponieważ grupa podstawowa jest reprezentowana przez wyraz wolny w (11), dołączyliśmy zmienne tylko dla trzech z czterech grup. Gdybyśmy dodali zmienną dla samotnych mężczyzn do modelu (11), wpadlibyśmy w pułapkę zmiennych , wprowadzając idealną kolinearność. Niektóre pakiety regresji automatycznie naprawią ten błąd, podczas gdy inne powiedzą, że istnieje

idealna kolinearność. Najlepiej jest dokładnie określić zmienne jakościowe ponieważ wtedy jesteśmy zmuszeni poprawnie interpretować ostateczny model. Mimo że samotni mężczyźni są grupą podstawową w (11), możemy użyć tego równania do uzyskania oszacowania różnicy między dowolnymi dwiema grupami. Ponieważ ogólne wyraz wolny jest współy dla wszystkich grup, możemy zignorować to przy znajdowaniu różnic. Tak więc oszacowanie proporcjonalnej różnicy między samotnymi i zamężnymi kobietami jest równe $-0,110 - (-0,198) = 0,088$, co oznacza, że samotne kobiety zarabiają około 8,8% więcej niż zamężne kobiety. Niestety nie możemy użyć równanie (11) do testowania, czy oszacowana różnica między samotnymi i zamężnymi kobiety są statystycznie istotne. Znajomość standardowych błędów mK i sK nie jest wystarczająca do przeprowadzenia testu. Najłatwiej jest wybrać jedeną z tych grupy jako grupę podstawową i odtworzyć równanie. Nic merytorycznego się nie zmiania, ale bezpośrednio otrzymujemy potrzebne oszacowanie i jego standardowy błąd. Kiedy używamy grupy zamężnych kobiet jako grupę podstawową, otrzymujemy równanie

$$\widehat{\log(place)} = 0,123 + 0,411\,mM + 0,198\,sM + 0,088\,sK \cdots,$$

gdzie oczywiście nie zmienił się żaden z pozostałych współczynników ani ich standardowe błędy. Oszacowanie dla sK wynosi, zgodnie z oczekiwaniami, 0,088. Teraz mamy do bezposrednio do czynienia ze standardowym błędem. Statystyka t dla hipotezy zerowej, mówiącej, że nie ma różnicy w populacji zamężnych i samotnych kobiet jest $t_{sK} = 0,088/0,052 = 1,69$. To niezbyt silny dowód przeciwko hipotezie zerowej. Widzimy również, że szacowana różnica między żonatymi mężczyznami i żonatymi kobiety są wyraźnie istotne statystycznie ($t_{mM} = 7,34$).

Poprzedni przykład ilustruje ogólną zasadę włączania zmiennych jakościowych do wskazania różnych grup. Jśli model regresji ma mieć różne wyrazy wolne dla, powiedzmy g grup lub kategorie, musimy dołączyć do modelu $g-1$ zmiennych jakośiowych wraz ogólnym wyrazem wolnym. Wyraz wolny dla grupy bazowej to ogólny wyraz wolny w modelu oraz współczynnik dla określonej grupy reprezentuje szacunkową różnicę w wyrazach wolnych między tą grupą a grupą bazową. Włącznie g zmiennych wraz z ogólnym wyrazem wolnym doprowadzi do pułapki zmiennych. Alternatywą jest włączenie g zmiennych jakościowych i wykluczenie ogólnego wyrazu wolnego. Nie jest to jednak zalecane, ponieważ testowanie różnic w stosunku do grupa bazowej staje się trudna. Zwykle pakiety statystyczne stosowane do analizy modeli regresji zmieniają sposób obliczania R^2 gdy regresja nie zawiera wyrazu wolnego.

1.4 Włączanie do modelu informacji opisanych przez zmienne porządkowe.

Załóżmy, że chcielibyśmy oszacować wpływ ratingów kredytowych miasta na wysokość oprocentowania emitowanych obligacji (MBR). Kilka firm finansowych, takich jak Moody's Investors Service oraz Standard and Poor's, ocenia jakość długu dla lokalnych państw. Oceny zależą od takich czynników, jak prawdopodobieństwo niewypłacalności. (Państwa preferują niższe stopy procentowe w celu zmniejszenia kosztów pożyczek.) Dla uproszczenia załóżmy, że zakres rankingów jest od zera do czterech, przy czym zero to najgorszy rating kredytowy, a cztery to najlepszy. To jest przykład zmiennej porządkowej. Nazwijmy tą zmienną CR dla uproszczenia. Pytanie na jakie musimy odpowiedzieć przy konstrukcji modeluto: jak włączyć zmienną CR do modelu w celu wyjaśnienia MBR? Jedną z możliwości jest włączenie CR, tak jak w przypadku każdego innego czynnika objaśniającego zmienną zależną.

$$MBR = \beta_0 + \beta_1\,CR + \quad inne\ \ czynniki,$$

gdzie nie specyfikujemy wyraźnie, jakie inne czynniki występują w modelu. Zatem β_1 to zmiana punktu procentowego w MBR, gdy CR wzrośnie o jedną jednostkę, uwzględniając inne

czynniki na tym samym poziomie. Niestety trudno jest zinterpretować wzrost CR o jedną jednostkę. Wiemy jakie znaczenie ilościowe możemy przypisać kolejnemu rokowi nauki lub innej złotówce wydanej na jednego ucznia, ale rzeczy takie jak ratingi kredytowe mają zazwyczaj jedynie porządkowe znaczenie. Wiemy, że CR wynoszący cztery jest lepszy niż CR wynoszący trzy, ale czy różnica między czterema a trzema jest taka sama jak różnica między jednym a zero? Jeśli nie, to może nie mieć sensu zakładać, że wzrost CR o jedną jednostkę ma stały wpływ na MBR. Lepsze podejście, które możemy wdrożyć, ponieważ CR przyjmuje stosunkowo niewiele wartości, jest zdefiniowanie zmiennych zastępczych dla każdej wartości CR. Zatem niech $CR_1 = 1$ jeśli $CR = 1$, i $CR_1 = 0$ w przeciwnym przypadku, podobnie $CR_2 = 1$, jeżeli $CR = 2$, i $CR_2 = 0$ w innym przypadku i tak dalej. Efektywnie, bierzemy pojedynczy rating kredytowy i dzielimy go na pięć kategorii. Następnie możemy dokonać estymacji modelu postaci:

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + \quad inne\ \ czynniki. \tag{13}$$

Zgodnie z naszą zasadą włączania zmiennych jakościowych do modelu, uwzględniamy cztery zmienne, ponieważ mamy pięć kategorii. Pominiętą kategorią jest tutaj zdolność kredytowa na poziomie zero, a więc jest to grupa bazowa. (Dlatego nie musimy definiować zmiennej zastępczej dla tej kategorii). Współczynniki są łatwe interpretować: 1 to różnica w MBR (inne czynniki ustalone) między państwem o zdolności kredytowej jednego i państwem o zdolności kredytowej zero; 2 to różnica w MBR między państwem o ratingu kredytowym dwa i państwem o zdolności kredytowej zero; i tak dalej. Zmiany pomiędzy poszczególnymi ratingami kredytowymi mogą mieć różny efekt, więc używanie (13) jest znacznie bardziej elastyczne niż zwykłe umieszczanie CR jako pojedynczej zmiennej. Po zdefiniowaniu zmiennych zastępczych oszacowanie (13) jest stosunkowo proste. Równanie (13) zawiera model ze stałym efektem częściowym jako przypadek szczególny. Jednym ze sposobów napisania trzech ograniczeń, które implikują stały efekt częściowy, jest $\delta_2 = 2\delta_1$, $\delta_3 = 3\delta_1$ i $\delta_4 = 4\delta_1$. Kiedy wstawimy te patametry do równania (13), otrzymamy

$$MBR = \beta_0 + \delta_1(CR_1 + 2CR_2 + 3CR_3 + 4CR_4) + \quad inne\ \ czynniki.$$

Teraz czynnik, przez który mnożymy δ_1 to po prostu pierwotna zmienna ratingu kredytowego, CR.

Przykład 7. *(Wpływ atrakcyjności fizycznej na wynagrodzenie)* Hamermesh i Biddle (1994) zastosowali miary atrakcyjności fizycznej w równaniu płac. (Plik BEAUTY zawiera mniej zmiennych, ale więcej obserwacji niż używane przez Hamermesh i Biddle.) Każda osoba w próbie została oceniona przez ankietera pod względem atrakcyjności fizycznej, przy użyciu pięciu kategorii (niewyszukana, dość przeciętna, średnia, dobrze wyglądająca i uderzająco piękna lub przystojny). Ponieważ jest tak mało ludzi w obu krańcowych kategoriach, autorzy umieść ludzi w jednej z trzech grup do analizy regresji: średnia, poniżej średniej i powyżej średniej, gdzie grupą podstawowa jest średnia. Wykorzystując danych dotyczących zatrudnienia Hamermesh i Biddle dokonali estymacji następującego modelu dla mężczyzn:

$$\widehat{\log(placa)} = \hat{\beta}_0 - 0.164\ pnsred + 0,016\, posred + \quad inne\ \ czynniki$$

$$n = 700, \quad \bar{R}^2 = 0,403$$

i modelu dla kobiet:

$$\widehat{\log(placa)} = \hat{\beta}_0 - 0.124\ pnsred + 0,035\, posred + \quad inne\ \ czynniki$$

$$n = 409, \quad \bar{R}^2 = 0,330.$$

Inne czynniki kontrolowane przez regresory obejmują wykształcenie, doświadczenie, staż pracy, stan cywilny i rasę. Szacuje się, że w przypadku mężczyzn o wyglądzie poniżej średniej zarobki są o 16,4% niższe niż w przypadku mężczyzn o wyglądzie ocenianym jako powyżej średniej, który są tacy sami pod innymi względami (w tym wykształcenie, doświadczenie, stanowiska, stanu cywilnego i rasa). Efekt jest statystycznie różny od zera, przy t na poziomie -3,57. Podobnie mężczyźni o ponadprzeciętnym wyglądzie zarabiają około 1,6% więcej, chociaż efekt nie jest statystycznie istotny ($t < 0,5$). Kobieta z wyglądem poniżej średniej zarabia około 12,4% mniej niż porównywalnym przypadku przeciętnie wyglądająca kobieta, z $t = 1,88$. Podobnie jak w przypadku mężczyzn, porównanie z przecientnej urodą kobietami nie różni się statystycznie od zera. W niektórych przypadkach zmienna porządkowa przyjmuje zbyt wiele wartości, więc zmienna zastępcza nie można uwzględnić zmiennej zastępczej dla każdej wartości. Możemy je wtedy podzielić na kategorie co zmniejsza ich liczbę.

1.5 Interakcje między zmiennymi jakościowymi.

Podobnie jak zmienne o znaczeniu ilościowym mogą wchodzić w interakcje w modelach regresji, tak też może się dziać pomiędzy zmiennymi jakościowymi. Jako ilustracja może posłużyć rozszerzenie modelu z przykładu 6 opisanego równaniem (11), w którym zdefiniowano cztery kategorie na podstawie stanu cywilnego i płci. W rzeczywistości możemy ten model przekształcić poprzez dodanie terminu interakcji między zmiennymi *kobiet* i *w-zwiazku* do modelu gdzie *kobieta* i *w-zwiazku* pojawiają się osobno. Pozwala to na uzależnienie premii za małżeńskwo od płeci. W celów porównania rozważmy oszacowanie model uwzględniającego interakcję płeć-małżeństwo

$$\widehat{\log}(place) = 0,321 - 0,110\, kobieta + 0,213\, w-zwiazku - 0,301\ kobieta \cdot w-zwiazku + \cdots, \tag{14}$$

gdzie reszta czynników jest identyczna z (11). Oszacowanie (14) wyraźnie pokazuje, że istnieje statystycznie istotna interakcja między płcią a stanem cywilnym. Model ten pozwala nam również uzyskać szacunkową różnicę płac między wszystkimi czterema grupy, ale tutaj musimy uważać, aby poprawnie zinterpretować kombinację zer i jedynek. Kombinacja $kobiet = 0$ i $w-zwiazku = 0$ odpowiada grupie samotnych mężczyzn, czyli grupie podstawowej, ponieważ eliminuje to *kobieta*, *w-zwiazku* i $kobieta \cdot w-zwiazku$. Możemy wyznaczyć wyraz wolny dla żonatych mężczyzn przez ustalenie $kobieta = 0$ i $w-zwiazku = 1$ w (14), co daje wyraz wolny równy $0.321+0,213 = 0,534$ i podobnie dla pozostałych grup. Równanie (14) to po prostu inny sposób na znalezienie różnic w wynagrodzeniu we wszystkich przypadkach kombinacje płci i stanu cywilnego. Pozwala nam łatwo przetestować hipotezę zerową, że różnica wynikająca z płci nie zależy od stanu cywilnego (równoważnie, że różnica wynikająca z faktu małżeńska nie zależy od płci). Równanie (11) jest wygodniejsze do testowania różnic płac między dowolną grupą a grupą podstawową samotnych mężczyzn.

Przykład 9. *(Wpływ użytkowania komputera na płace)* W pracy z roku 1993 Krueger ocenił wpływ użytkowania komputera na płace. Definiuje zmienną jakościową, *compwork*, równą jedne, jeśli dana osoba korzysta z komputera w pracy. Inna zmienna jakościowa, *comphome,* jest równa jednej, jeśli osoba korzysta z komputera w domu. Na podstawie 13379 elementowej próby estymuje model

$$\widehat{\log}(place) = \hat{\beta}_0 + 0,177\, compwork + 0,070\, comphome + 0,017\, compwork \cdot comphome + other\ factors. \tag{15}$$

(Inne przyjete przez Krugera czynniki są standardowe dla regresji płac, w tym wykształcenie, doświadczenie, płeć, stan cywilny itp.) Krueger nie wyszczególnił wartości estymatora

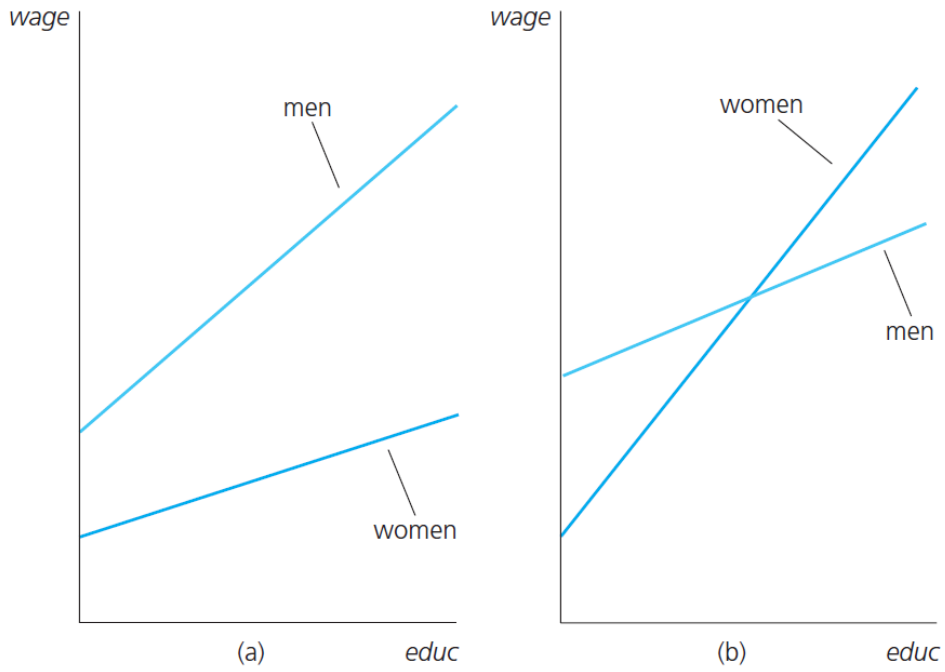
wyrazu wolnego w modelu, ponieważ nie ma to żadnego znaczenia; wszystko, co musimy wiedzieć, to że grupa podstawowa składa się z osób, które nie korzystają z komputera w domu lub w pracy. Warto zauważenie, że szacunkowa zmiana wynikająca z korzystania z komputera w pracy (ale nie w domu) jest na poziomie 17,7%. Podobnie osoby korzystające z komputerów w w domu, ale w pracy nie ma premię płacową na poziomie około 7% w stosunku do tych, którzy nie korzystają z komputera w ogóle. Różnica między tymi, którzy używają komputera w obu miejscach, w stosunku do tych, którzy nie używają komputera w żadnym miejscu, to około 26,4% (uzyskana przez dodanie wszystkich trzech współczynników i pomnożenie przez 100), lub bardziej precyzyje oszacowanie 30,2% uzyskane przy użyciu równania (10). Czynnik interakcji w modelu nie jest istotny statystycznie, jak również nie jest bardzo istotny z punktu widzenia ekonomiczniego.

1.6 Wpływ interakcji między zmiennymi ilosciowymi i jakosciowymi w modelu ekonometrycznym.

Kontynuując przykład płac, załóżmy, że chcemy przetestować czy wpływ wykształcenia na płace jest taki sam dla mężczyzn i kobiet, przy założeniu różnic w wynagrodzeniach kobiet i mężczyzn (różnica, dla której już znaleźliśmy przekowujące dowody). Dla uproszczenia uwzględniamy w modelu tylko wykształcenie i płeć. Jaki model powinniśmy uwzględić? Rozważmy model

$$\log(place) = \beta_0 + \delta_0\, kobieta) + (\beta_1 + \delta_1 kobieta)\, edukacja + u. \tag{16}$$

Jeśli w równaniu (16) przyjmiemy $kobieta = 0$, to okaże się, że punkt przecięcia dla mężczyzn równy β_0, a nachylenie edukacji dla mężczyzn równy β_1. W przypadku kobiet kładąc w równaniu (16) $kobieta = 1$, otrzymamy punkt przecięcia dla kobiet równy $\beta_0 + \delta_0$, a nachylenie równe $\beta_1 + \delta_1$. Dlatego, δ_0 mierzy różnicę w wyrazie wolnym między kobietami i mężczyznami, a δ_1 mierzy różnicę w płacach, wynikającą z wykształcenia, między kobietami i mężczyznami.



Rysunek 2: Wykres zależności (16) (a) dla $\delta_0 < 0, \quad \delta_1 < 0$; (b) dla $\delta_0 < 0, \quad \delta_1 > 0$.

Dwa, z czterech czterech mozliwych wyborów znaków dla δ_0 i δ_1 przedstawiono na rysunku 2. Wykres (a) pokazuje przypadek, w którym punkt przecięcia dla kobiet jest niższy niż

dla mężczyzn, oraz nachylenie linii jest mniejsze dla kobiet niż dla mężczyzn. Oznacza to, że kobiety zarabiają mniej niż mężczyźni na wszystkich poziomach edukacji, a różnica rośnie wraz ze wzrostem wykształcenia. W wykres (b), punkt przecięcia dla kobiet jest niższy niż dla mężczyzn, ale nachylenie edukacji jest większy dla kobiet. Oznacza to, że kobiety zarabiają mniej niż mężczyźni na niskim poziomie wykształcenia, ale różnica zmniejsza się wraz ze wzrostem wykształcenia. W pewnym momencie kobieta zarabia więcej niż mężczyzna, mający ten sam poziom wykształcenia (i ten punkt można łatwo znaleźć, biorąc pod uwagę oszacowane równanie). Jak możemy estymować rozpatrywany model? Aby zastosować, metodę najmniejszych kwadratów wmusimy rozpatrzeć model z interakcją między zmiennymi*kobieta* i zmienną *edukacja*:

$$\log(placa) = \beta_0 + \delta_0\, kobieta + \beta_1\, educja + \delta_1\, kobieta \cdot\ edukacja + u. \tag{17}$$

Parametry można teraz oszacować na podstawie regresji $\log(placa)$ wzgledem zmiennych *kobieta*, *educja* i $kobieta \cdot edukacja$. Estymacja wartosci interakcji mozna w łatwy sposób przeprowadzić w każdym pakiecie statystycznym. Zwróćmy uwagę na interpretację zmiennej $edukaca \cdot kobiea$, która wynosi zero dla każdego mężczyzny w próbie i jest równa poziomowi wykształcenia każdej kobiety w próbie. Ważną hipotezą jest hipoteza mówiąca, że wpływ wykształcenia na płace jest taki sam dla kobiet i mężczyźni. W odniesieniu do modelu (17) jest ona postaci

$$H_0 :\ \delta_1 = 0,$$

co oznacza, że nachylenie $\log(placa)$ w odniesieniu do wykształcenia jest takie samo dla mężczyzn i kobiet. Zauważ, że ta hipoteza nie nakłada żadnych ograniczeń na δ_0. Różnica w wynagrodzeniach mężczyzn i kobiet jest dozwolona, ale musi być taka sa na wszystkich poziomów edukacji. Sytuację tę opisuje rysunek 1.

Interesuje nas również hipoteza, że średnie płace są identyczne dla mężczyzn i kobiet o tym samym poziomie wykształcenia. Oznacza to, że przy założeniu prawdziwości hipotezy zerowej zarówno δ_0 jak i δ_1 muszą być równe zero. Użyjemy testu F do przetestowania

$$H_0 :\ \delta_0 = 0,\ \delta_1 = 0.$$

W modelu z różnicą wyrazów wolnych dla kobiet i mężczyzn odrzucamy tę hipotezę, ponieważ $H_0 :\ \delta_0 = 0$, jest odrzucona na korzyść alternatywy $H_1 :\ \delta_0 < 0$.