Paweł Maśluch

[WDEC] Wspomaganie Decyzji

Laboratorium 7A, 7B

Testowanie hipotez statystycznych

Ogólna charakterystyka laboratorium 7A i 7B:
Celem laboratorium jest budowa modeli statystycznych w oparciu o analizę ANOVA (Analysis of Variance).

W modelu ANOVA zmienne zależne są ciągłe,
a zmienne niezależne są zmiennymi kategorycznymi.

Polecenia do wykonania w trakcie laboratorium (część 1, lab 7A):
Pewna firma chce zbadać, jak wybrany typ reklamy produktu wpływa na sprzedaż.
Analiza jest prowadzona w oparciu o tabelę ADS1, zawierającą następujące informacje:

- **Ad** - typ reklamy
- **Area** – obszar kraju
- **Sales** – wielkość sprzedaży

Wstępna analiza danych.

Na początku, przed przystąpieniem do wykonywania poniższych poleceń, wykonuję następujące operacje:

- stworzenie biblioteki Lab7A (która będzie zawierać program realizujący polecenia oraz wynikowe struktury danych): `libname Lab7A'/folders/myfolders/Lab7A';`

- skopiowanie pliku ads1.xlsx do biblioteki Lab7A

- import danych z pliku ads1.xlsx do pliku SAS-owego Lab7A.ads1:

```
proc import
   datafile = '/folders/myfolders/Lab7A/ads1.xlsx'
   DBMS = xlsx REPLACE
   OUT = Lab7A.ads1;
run;
```

1. Wyświetl 10 pierwszych obserwacji plików danych Ads i Ads1.

Kod SAS-owy:

```
options obs=10;
proc print data=Lab7A.ads1;
run;
```

Wynikowa tabela:

Obs.	Ad	Area	Sales
1	paper	1	75
2	radio	1	69
3	people	1	63
4	display	1	52
5	paper	2	57
6	radio	2	51
7	people	2	67
8	display	2	61
9	paper	3	76
10	radio	3	100

2. Oblicz następujące statystyki dla wszystkich danych i dla poszczególnych typów reklamy: średnia, min, max, oraz odchylenie standardowe (PROC MEANS) i zapisz w pliku.

2a) Dla wszystkich danych (zapis do pliku SAS-owego Lab7A.ads1_stats_all_data)

Kod SAS-owy:

```
PROC MEANS DATA = Lab7A.ads1 MEAN MIN MAX STDDEV;
title 'Ads1 - statystyki dla wszystkich danych';
VAR Sales;
output OUT = Lab7A.ads1_stats_all_data (drop=_type_
/*_freq_*/);
RUN;
```

Wynikowa tabela:

Zmienna analizowana: Sales Sales			
Średnia	**Minimum**	**Maksimum**	**Odch. std.**
66.8194444	33.0000000	100.0000000	13.5278282

2b) Dla poszczególnych typów reklamy
(zapis do pliku SAS-owego Lab7A.ads1_stats_kinds_of_ad)

Kod SAS-owy:

```
/* Sortowanie danych po typie reklamy */
proc sort data = Lab7A.ads1;
     by Ad;
run;

/* Właściwy program */
PROC MEANS DATA = Lab7A.ads1 MEAN MIN MAX STDDEV;
     title 'Ads1 - statystyki dla poszczególnych typów
reklamy';
     VAR Sales;
     class /* by */ Ad;
     output OUT = Lab7A.ads1_stats_kinds_of_ad (drop=_type_
/*_freq_*/);
RUN;
```

Wynikowa tabela:

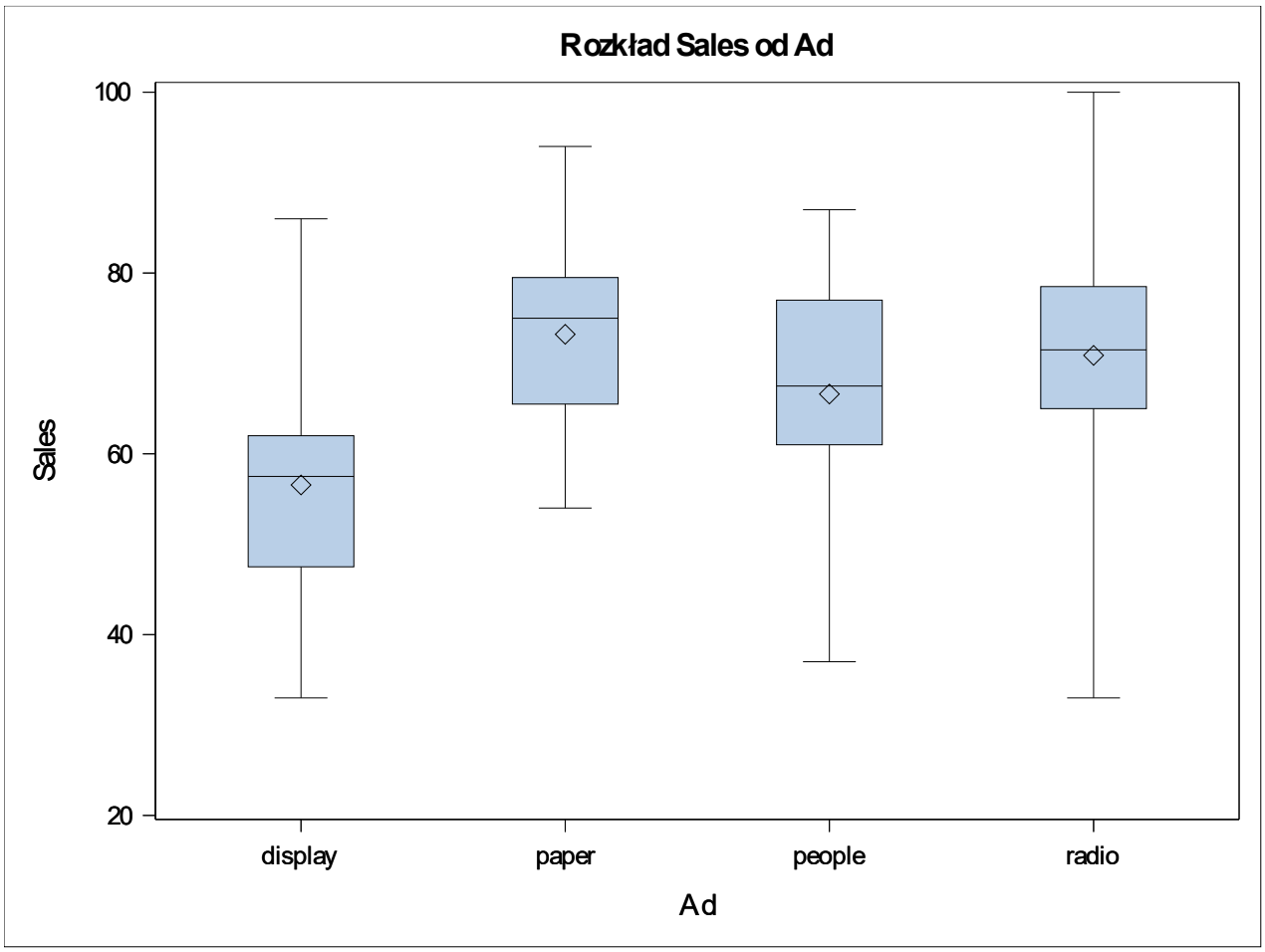
Zmienna analizowana: Sales Sales					
Ad	**N obs.**	**Średnia**	**Minimum**	**Maksimum**	**Odch. std.**
display	36	56.5555556	33.0000000	86.0000000	11.6188134
paper	36	73.2222222	54.0000000	94.0000000	9.7339204
people	36	66.6111111	37.0000000	87.0000000	13.4976776
radio	36	70.8888889	33.0000000	100.0000000	12.9676031

3. Narysuj wykres typu BOX PLOT pokazujący statystyki sprzedaży dla poszczególnych typów reklamy (PROC SGPLOT). Zinterpretuj otrzymany wykres.

Kod SAS-owy:

```
     /* Sortowanie danych względem typu reklamy */
proc sort data = Lab7A.ads1;
     by Ad;
run;

/* Właściwy procstep */
PROC BOXPLOT data=Lab7A.ads1;
     ods graphics on;
    plot Sales*Ad;
RUN;
```

Wykres:



Interpretacja

Dla każdej klasy:

- „wąsy” oznaczają wartości minimalne i maksymalne elementów klasy
- symbol wewnątrz prostokąta oznacza średnią arytmetyczną elementów klasy
- pozioma linia wewnątrz prostokąta to mediana elementów klasy
- wysokość (czyli wymiar liniowy pionowy) prostokąta oznacza zakres między 25. a 75. kwartylem

Dla wszystkich klas (z wyjątkiem klasy paper),
mediana (z dobrym przybliżeniem) jest równa wartości średniej.

Najmniejsza różnica między wartościami maksymalną a minimalną jest dla klasy paper,
a największa dla klasy radio.

Możemy zauważyć, że dla naszych klas, dane są „skupione” wokół pewnych wartości
(wskazują na to stosunkowo nieduże wysokości prostokątów w ramach każdej klasy).
Przy czym największe „skupienie” dostrzegamy dla klasy radio.

Testowanie hipotezy statystycznej .

Hipoteza zerowa:
Ho : S1=S2=S3=S4
H1: S1≠S2 S1≠S2 or S1≠S3 or S1≠S4 or S2≠S3 or S2≠S4 or S3≠S4
S1 – średnia sprzedaż dla reklamy typu: *display*
S2 – średnia sprzedaż dla reklamy typu: *paper*
S3 – średnia sprzedaż dla reklamy typu: *people*
S4 – średnia sprzedaż dla reklamy typu: *radio*

Model podstawowy:
$Mik = \mu + Ti$
Weryfikacja wyjść dla danych rzeczywistych:
$Yik = Mik + \varepsilon ik$ $Yik = \mu + Ti + \varepsilon ik$,
gdzie:
Mik - oznacza k-tą wartość zmiennej wyjściowej dla reklamy typu i (dla modelu)
Yik oznacza k-tą wartość zmiennej wyjściowej dla reklamy typu i
μ jest średnią dla wszystkich obserwacji
Ti jest różnicą między średnią sprzedaży dla wszystkich obserwacji i średnią dla danego typu reklamy i
εik jest różnicą między wartością rzeczywista k-tej obserwacji dla i-tej klasy reklamy oraz wartością tej obserwacji uzyskanej z modelu Mik

Model rozszerzony:
$MR_{jik} = \mu + \alpha_j + T_i$
Weryfikacja wyjść dla danych rzeczywistych:
$Y_{jik} = MR_{jik} + \varepsilon_{jik}$ $Y_{jik} = \mu + \alpha_j + T_i + \varepsilon_{jik}$
gdzie
Y_{jik} oznacza k-tą wartość zmiennej wyjściowej dla reklamy typu i i obszaru j
μ jest średnią dla wszystkich obserwacji
α_j jest różnicą między średnią sprzedaży dla wszystkich obserwacji i średnią dla danego obszaru j
T_i jest różnicą między średnią sprzedaży dla wszystkich obserwacji i średnią dla danego typu reklamy i
ε_{jik} jest błędem dla danej obserwacji

1. Sprawdź, czy hipoteza 0 jest spełniona (PROC GLM). Dokonaj interpretacji wyników procedury GLM.

Model podstawowy

Kod SAS-owy:

```
proc glm
     data=Lab7A.ads1;
     class Ad;
     model Sales = Ad/solution;
run;
```

Uzyskane rezultaty:

Procedura GLM

Informacje o poziomach klasyfikacji		
Klasa	**Poziomy**	**Wartości**
Ad	4	display paper people radio

Liczba obserwacji wczytanych	144
Liczba obserwacji użytych	144

Procedura GLM

Zmienna zależna: Sales Sales

Źródło	DF	Suma kwadratów	Średni kwadrat	Wartość F	Pr. > F
Model	3	5866.08333	1955.36111	13.48	<.0001
Błąd	140	20303.22222	145.02302		
Razem skorygowane	143	26169.30556			

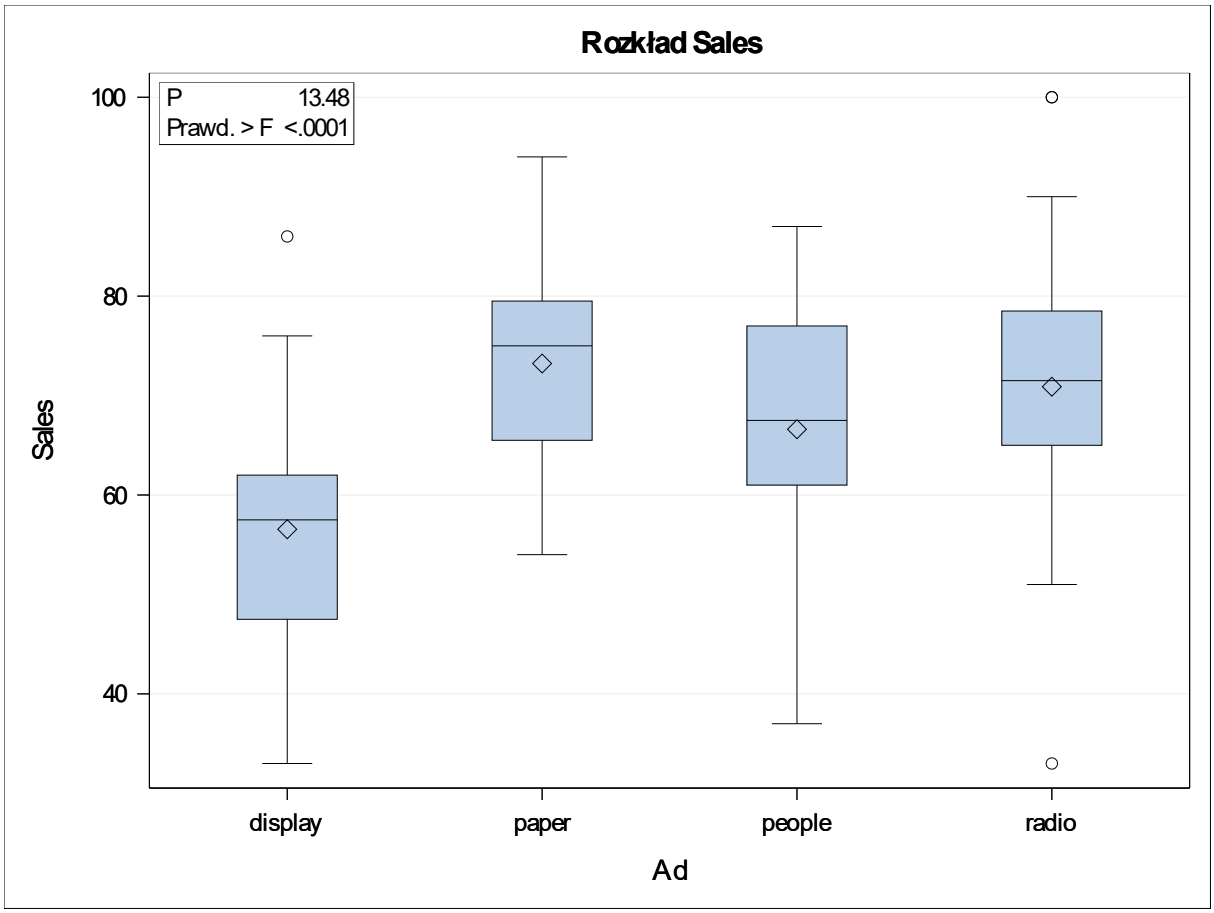
R-kwadrat	Wsp. zmienności	Pierw. z MSE	Średnia Sales
0.224159	18.02252	12.04255	66.81944

Źródło	DF	Suma kwadratów typu Typ I	Średni kwadrat	Wartość F	Pr. > F
Ad	3	5866.083333	1955.361111	13.48	<.0001

Źródło	DF	Suma kwadratów typu Typ III	Średni kwadrat	Wartość F	Pr. > F
Ad	3	5866.083333	1955.361111	13.48	<.0001

Parametr		Ocena		Błąd standardowy	Wartość t	Pr. > |t|
Intercept		70.88888889	B	2.00709170	35.32	<.0001
Ad	**display**	-14.33333333	B	2.83845631	-5.05	<.0001
Ad	**paper**	2.33333333	B	2.83845631	0.82	0.4125
Ad	**people**	-4.27777778	B	2.83845631	-1.51	0.1340
Ad	**radio**	0.00000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.



Ciekawostka. W wyniku wywołania proc glm (nie wyłączyłem opcji wyświetlania tego typu wykresu), otrzymałem kolejny wykres typu BOXPLOT dla naszych danych (względem typu reklamy). Można dostrzec pewne różnice względem poprzedniego takiego wykresu. Otóż, w tym przypadku, niezamalowane kółka oznaczają wartości „odstające" (w ramach danej klasy), które nie były uwzględnione wcześniej. Oczywiście, sam wygląd wykresu również nieco się zmienił (np. inne zakresy „wąsów" lub jeszcze większa różnica między medianą a wartością średnią w klasie paper).

Zauważmy, że w kolumnie Suma kwadratów, podano najpierw SSM, potem SSE, a na koniec SST.

Zauważmy, że miara R-kwadrat (czyli współczynnik determinacji) wynosi ok. 0.224159. Wielkość ta informuje o tym, jaka część zmienności zmiennej objaśnianej została wyjaśniona przez model.

W zależności od wartości R-kwadrat, mamy różne rodzaje dopasowań:
0,0 - 0,5 - dopasowanie niezadowalające
0,5 - 0,6 - dopasowanie słabe
0,6 - 0,8 - dopasowanie zadowalające
0,8 - 0,9 - dopasowanie dobre
0,9 - 1,0 - dopasowanie bardzo dobre

W naszym przypadku, dopasowanie jest niezadowalające.

Odczytujemy (z dowolnej tabeli wynikowej), że wartość F wyniosła dla naszych danych ok. 13,48, natomiast p-value jest mniejsze niż 0.0001 (kolumna Pr. > F).

Korzystając ze strony http://www.socr.ucla.edu/Applets.dir/F_Table.html , możemy odczytać, że dla wartości alfa = 0.05, df1 = 3 (czyli w naszym przypadku to liczba klas – rodzajów reklamy – pomniejszona o 1) oraz df2 = nieskończoność (gdyż nasza liczba obserwacji, pomniejszona o 1, wynosi 144-1=143, i jest większa niż największa wartość z tabeli na wyżej wymienionej stronie), wartość F wynosi ok. 2,6049.

Skoro wartość F = 13.48 (dla naszych danych) jest większa niż wartość F = 2.6049 (uzyskana z tabeli), to wnioskujemy, że hipotezę zerową należy odrzucić (lub inaczej: wartość p-value dla F=13.48 jest mniejsza niż wartość p-value=alfa=0.05 dla F = 2.6049).

Model rozszerzony

Kod SAS-owy:

```
proc glm
        data=Lab7A.ads1;
        class Ad Area;
        model Sales = Ad Area/solution;
run;
```

Uzyskane rezultaty:

Informacje o poziomach klasyfikacji		
Klasa	**Poziomy**	**Wartości**
Ad	4	display paper people radio
Area	18	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Liczba obserwacji wczytanych	144
Liczba obserwacji użytych	144

	DF	Suma kwadratów	Średni kwadrat	Wartość F	Pr. > F
Model	20	15131.38889	756.56944	8.43	<.0001
Błąd	123	11037.91667	89.73916		
Razem skorygowane	143	26169.30556			

R-kwadrat	Wsp. zmienności	Pierw. z MSE	Średnia Sales
0.578211	14.17712	9.473076	66.81944

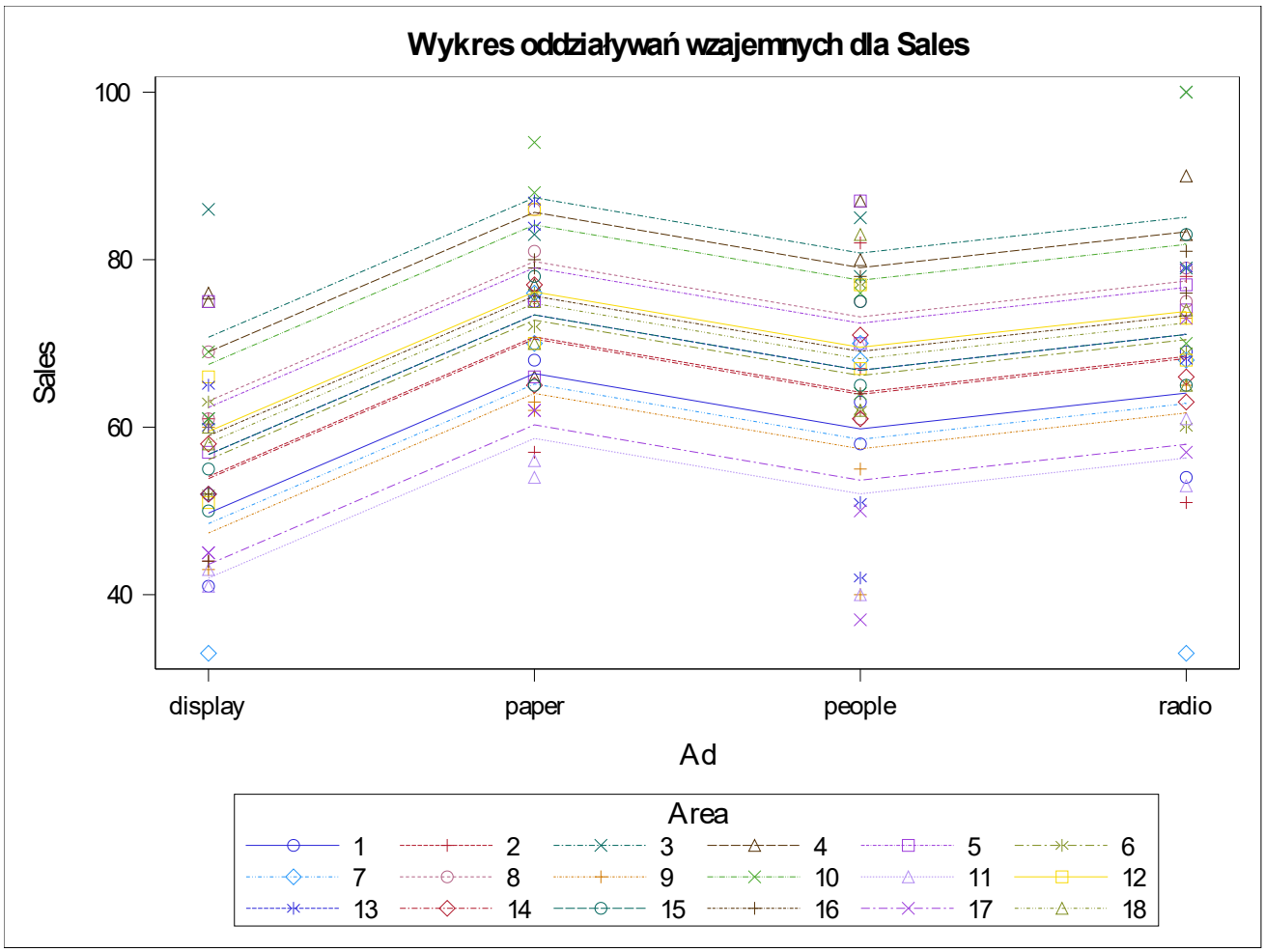
Źródło	DF	Suma kwadratów typu Typ I	Średni kwadrat	Wartość F	Pr. > F
Ad	3	5866.083333	1955.361111	21.79	<.0001
Area	17	9265.305556	545.017974	6.07	<.0001

Źródło	DF	Suma kwadratów typu Typ III	Średni kwadrat	Wartość F	Pr. > F
Ad	3	5866.083333	1955.361111	21.79	<.0001
Area	17	9265.305556	545.017974	6.07	<.0001

Parametr		Ocena		Błąd standardowy	Wartość t	Pr. > |t|
Intercept		72.44444444	B	3.61759047	20.03	<.0001
Ad	**display**	-14.33333333	B	2.23282531	-6.42	<.0001
Ad	**paper**	2.33333333	B	2.23282531	1.05	0.2981
Ad	**people**	-4.27777778	B	2.23282531	-1.92	0.0577
Ad	**radio**	0.00000000	B	.	.	.
Area	**1**	-8.37500000	B	4.73653776	-1.77	0.0795
Area	**2**	-4.00000000	B	4.73653776	-0.84	0.4000
Area	**3**	12.62500000	B	4.73653776	2.67	0.0087
Area	**4**	10.87500000	B	4.73653776	2.30	0.0234
Area	**5**	4.25000000	B	4.73653776	0.90	0.3713
Area	**6**	-2.00000000	B	4.73653776	-0.42	0.6736
Area	**7**	-9.62500000	B	4.73653776	-2.03	0.0443
Area	**8**	5.00000000	B	4.73653776	1.06	0.2932
Area	**9**	-10.75000000	B	4.73653776	-2.27	0.0250

Parametr		Ocena		Błąd standardowy	Wartość t	Pr. > |t|
Area	10	9.37500000	B	4.73653776	1.98	0.0500
Area	11	-16.12500000	B	4.73653776	-3.40	0.0009
Area	12	1.37500000	B	4.73653776	0.29	0.7721
Area	13	-1.37500000	B	4.73653776	-0.29	0.7721
Area	14	-4.25000000	B	4.73653776	-0.90	0.3713
Area	15	-1.37500000	B	4.73653776	-0.29	0.7721
Area	16	0.87500000	B	4.73653776	0.18	0.8537
Area	17	-14.50000000	B	4.73653776	-3.06	0.0027
Area	18	0.00000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms who estimates are followed by the letter 'B' are not uniquely estimable.



Zauważmy, że w kolumnie Suma kwadratów, podano najpierw SSM, potem SSE, a na koniec SST.

W przypadku modelu rozszerzonego, wartość R-kwadrat wyniosła ok. 0.578211.
Zatem osiągnęliśmy dopasowanie słabe.

Odczytujemy (z dowolnej tabeli wynikowej), że wartość F wyniosła dla naszych danych ok. 8.43, natomiast p-value jest mniejsze niż 0.0001 (kolumna Pr. > F).

Korzystając ze strony http://www.socr.ucla.edu/Applets.dir/F_Table.html ,
możemy odczytać, że dla wartości alfa = 0.05, df1 = 20 (w naszym przypadku to liczba klas (rodzajów reklamy 4 + liczba obszarów 18), pomniejszona o 2) oraz df2 = nieskończoność (gdyż nasza liczba obserwacji, pomniejszona o 1, wynosi 144-1=143, i jest większa niż największa wartość z tabeli na wyżej wymienionej stronie), wartość F wynosi ok. 1.5705.

Skoro wartość F = 8.43 (dla naszych danych) jest większa niż wartość F = 1.5705 (uzyskana z tabeli), to wnioskujemy, że hipotezę zerową należy odrzucić
(lub inaczej: wartość p-value dla F=8.43 jest mniejsza niż wartość p-value=alfa=0.05 dla F = 1.5705).

2. Podaj parametry modelu podstawowego i rozszerzonego. Oblicz wyjścia na podstawie modeli.

Średnią wszystkich obserwacji odczytałem z tabeli z zad.2a (wstępna analiza danych) i wynosi mi = 66.8194444.

Średnie dla klas odczytałem z tabeli z zad.2b (wstępna analiza danych) i wynoszą:
mean_display = 56.5555556
mean_paper = 73.2222222
mean_people = 66.6111111
mean_radio = 70.8888889

Średnie względem obszarów uzyskałem, wywołując poniższy fragment kodu SAS-owego:

```
/* Sortowanie danych po typie obszarze */
proc sort data = Lab7A.ads1;
        by Area;
run;

/* Właściwy program */
PROC MEANS DATA = Lab7A.ads1 MEAN;
        title 'Ads1 - statystyki dla poszczególnych obszarów';
```

```
        VAR Sales;
        class /* by */ Area;
        output OUT = Lab7A.ads1_stats_kinds_of_area (drop=_type_ /*_freq_*/);
RUN;
```

Uzyskałem poniższą tabelę:

Zmienna analizowana: Sales Sales		
Area	**N obs.**	**Średnia**
1	8	60.0000000
2	8	64.3750000
3	8	81.0000000
4	8	79.2500000
5	8	72.6250000
6	8	66.3750000
7	8	58.7500000
8	8	73.3750000
9	8	57.6250000
10	8	77.7500000
11	8	52.2500000
12	8	69.7500000
13	8	67.0000000
14	8	64.1250000
15	8	67.0000000
16	8	69.2500000
17	8	53.8750000
18	8	68.3750000

Wartości tau (dla danego rodzaju reklamy – tau_display, tau_paper, tau_people, tau_radio) oraz wartości alfa (dla danego obszaru – alfa1, …, alfa18) obliczyłem za pomocą odpowiedniego programu SAS-owego, a następnie odczytałem z wynikowej struktury danych:

```
/* Model podstawowy i rozszerzony */

data Lab7A.ads1;
    set Lab7A.ads1;
```

```
    /* Model podstawowy i rozszerzony */


      /* Średnia wszystkich obserwacji - wielkość odczytana z tabeli z
zad.2a (wstępna analiza danych) */
    mi = 66.8194444;

      /* Średnie dla klas - wielkości odczytane z tabeli z zad.2b (wstępna
analiza danych) */
      mean_display = 56.5555556;
      mean_paper   = 73.2222222;
      mean_people  = 66.6111111;
      mean_radio   = 70.8888889;

      /* Różnice między średnią sprzedaży dla wszystkich obserwacji i
średnią dla danego typu reklamy */
    if Ad = "display" then tau = mi - mean_display;    tau_display = mi -
mean_display;
    if Ad = "paper"   then tau = mi - mean_paper;    tau_paper = mi -
mean_paper;
    if Ad = "people"  then tau = mi - mean_people;    tau_people = mi -
mean_people;
    if Ad = "radio"   then tau = mi - mean_radio;    tau_radio = mi -
mean_radio;

    /* Wartość zmiennej wyjściowej dla reklamy danego typu (dla modelu
podstawowego) */
    m = mi + tau;

    /* Różnica między wartością rzeczywistą obserwacji oraz wartością m */
    eps = sales - m;

    /* Wartości obserwacji (ta kolumna ma charakter testowy) */
    y = m + eps;

    /* Reszta linii kodu dotyczy tylko modelu rozszerzonego */

      /* Średnie wartości obserwacji względem obszaru - uzyskane z
pomocniczych operacji */
          mean1  = 60.000;
            mean2  = 64.375;
            mean3  = 81.000;
            mean4  = 79.250;
            mean5  = 72.625;
            mean6  = 66.375;
            mean7  = 58.750;
            mean8  = 73.375;
            mean9  = 57.625;
            mean10 = 77.750;
            mean11 = 52.250;
            mean12 = 69.750;
            mean13 = 67.000;
            mean14 = 64.125;
            mean15 = 67.000;
```

```
            mean16 = 69.250;
            mean17 = 53.875;
            mean18 = 68.375;


/* Różnica między średnią sprzedaży dla wszystkich obserwacji i średnią dla
danego obszaru */
    if Area =  1 then   alfa  = mi -  mean1;   alfa1  = mi -  mean1;
    if Area =  2 then   alfa  = mi -  mean2;   alfa2  = mi -  mean2;
    if Area =  3 then   alfa  = mi -  mean3;   alfa3  = mi -  mean3;
    if Area =  4 then   alfa  = mi -  mean4;   alfa4  = mi -  mean4;
    if Area =  5 then   alfa  = mi -  mean5;   alfa5  = mi -  mean5;
    if Area =  6 then   alfa  = mi -  mean6;   alfa6  = mi -  mean6;
    if Area =  7 then   alfa  = mi -  mean7;   alfa7  = mi -  mean7;
    if Area =  8 then   alfa  = mi -  mean8;   alfa8  = mi -  mean8;
    if Area =  9 then   alfa  = mi -  mean9;   alfa9  = mi -  mean9;
    if Area = 10 then   alfa  = mi - mean10; alfa10  = mi -  mean10;
    if Area = 11 then   alfa  = mi - mean11; alfa11  = mi -  mean11;
    if Area = 12 then   alfa  = mi - mean12; alfa12  = mi -  mean12;
    if Area = 13 then   alfa  = mi - mean13; alfa13  = mi -  mean13;
    if Area = 14 then   alfa  = mi - mean14; alfa14  = mi -  mean14;
    if Area = 15 then   alfa  = mi - mean15; alfa15  = mi -  mean15;
    if Area = 16 then   alfa  = mi - mean16; alfa16  = mi -  mean16;
    if Area = 17 then   alfa  = mi - mean17; alfa17  = mi -  mean17;
    if Area = 18 then   alfa  = mi - mean18; alfa18  = mi -  mean18;

    /* Wartość zmiennej wyjściowej dla reklamy danego typu oraz z danego
obszaru (dla modelu rozszerzonego) */
    mr = mi + alfa + tau;

    /* Błędy modelu rozszerzonego */
    epsr = sales - mr;

    /* Wartości obserwacji (ta wartość ma charakter testowy) */
    yr = mr + epsr;
run;
```

Wyjścia modeli m (dla modelu podstawowego) oraz mr (dla modelu rozszerzonego) są obliczone w powyższym programie SAS-owym i zapisywane do wynikowej struktury SAS-owej.

<u>Wartości tau:</u>

tau_display = 10.2638888
tau_paper = -6.4027778
tau_people = 0.2083333
tau_radio = -4.0694445

Wartości alfa:
alfa1 = 6.8194444
alfa2 = 2.4444444
alfa3 = -14.1805556
alfa4 = -12.4305556
alfa5 = -5.8055556
alfa6 = 0.4444444
alfa7 = 8.0694444
alfa8 = -6.5555556
alfa9 = 9.1944444
alfa10 = -10.9305556
alfa11 = 14.5694444
alfa12 = -2.9305556
alfa13 = -0.1805556
alfa14 = 2.6944444
alfa15 = -0.1805556
alfa16 = -2.4305556
alfa17 = 12.9444444
alfa18 = -1.5555556

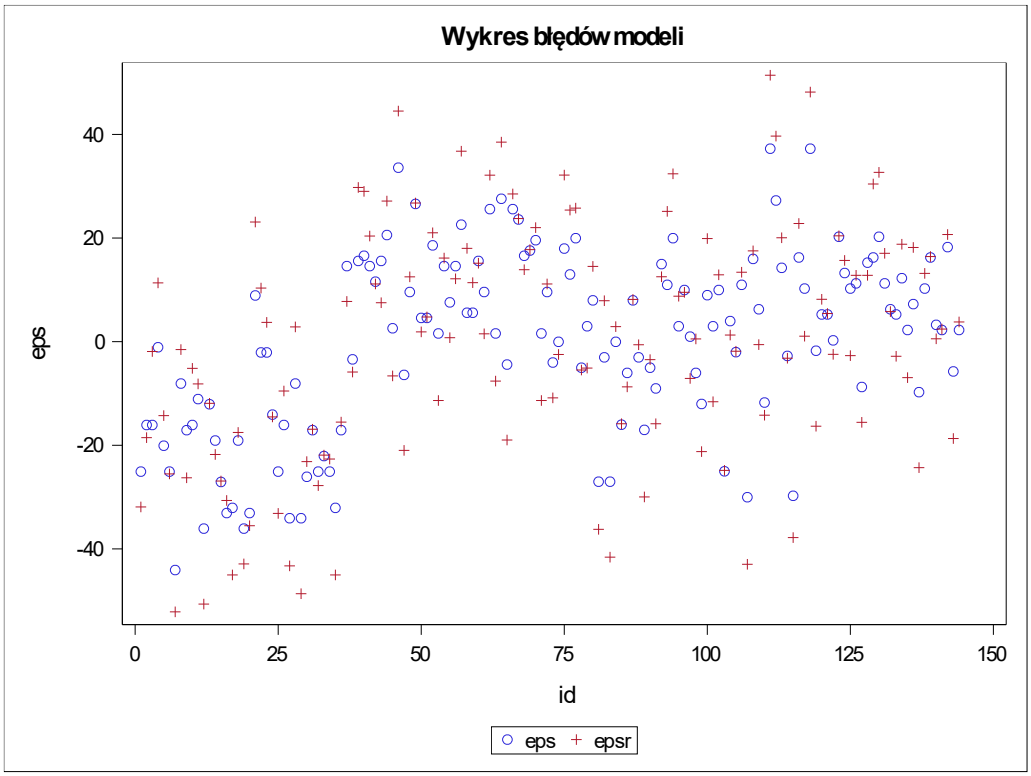
Ze względu na sporą ilość obserwacji, ograniczę się do przedstawienia w postaci wykresów, błędów modeli podstawowego i rozszerzonego (konkretne wartości błędów są w wynikowej strukturze danych, wygenerowanej w ramach poprzedniego fragmentu kodu).

Kody SAS-owe generujące wykresy:

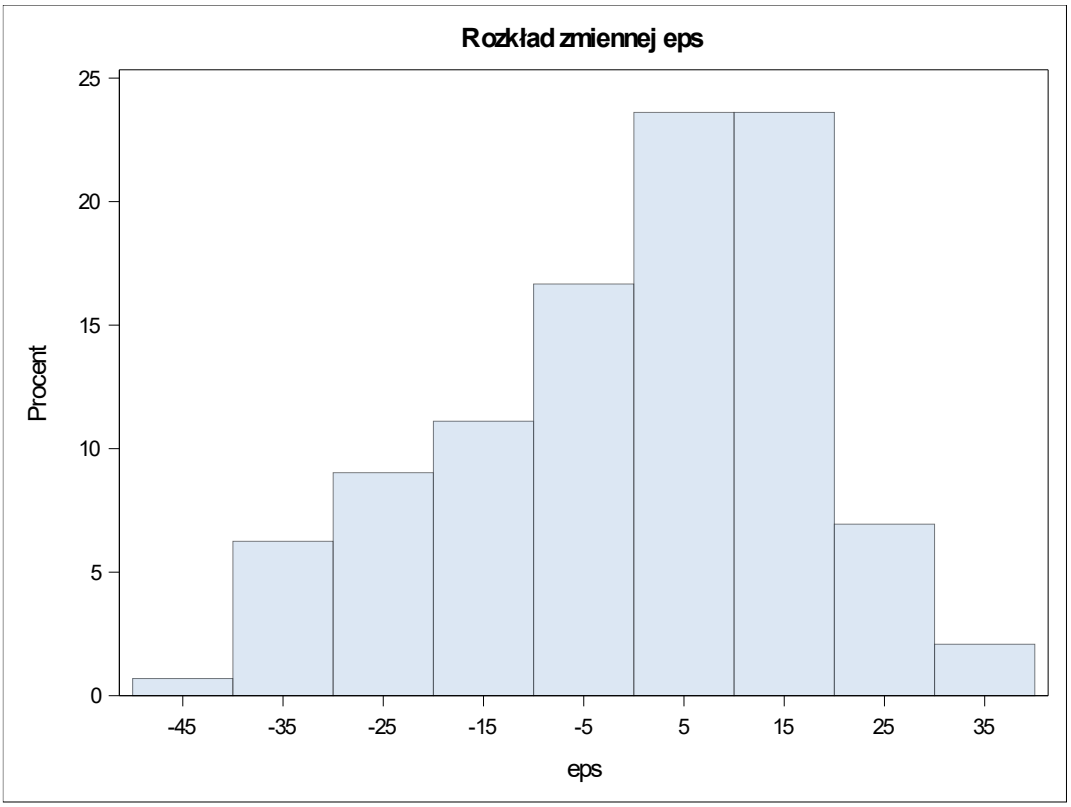
```
/* Wykres błędów modeli */

proc sgplot data = Lab7A.ads1;
        title 'Wykres błędów modeli'; /* tytuł wykresu */
        scatter x = id y = eps; /* wykres punktowy */
        scatter x = id y = epsr; /* wykres punktowy */
run;


/* Histogram - błędy modelu podstawowego */

PROC UNIVARIATE data = Lab7A.ads1; /* miejsce, z którego pobieramy dane do
stworzenia histogramu */

      title 'Histogram - błędy modelu podstawowego';
      VAR eps; /* dla kolumny col1 stworzymy histogram */
      HISTOGRAM eps; /* stworzenie histogramu dla kolumny eps */

run; /* uruchomienie tego fragmentu kodu */
```

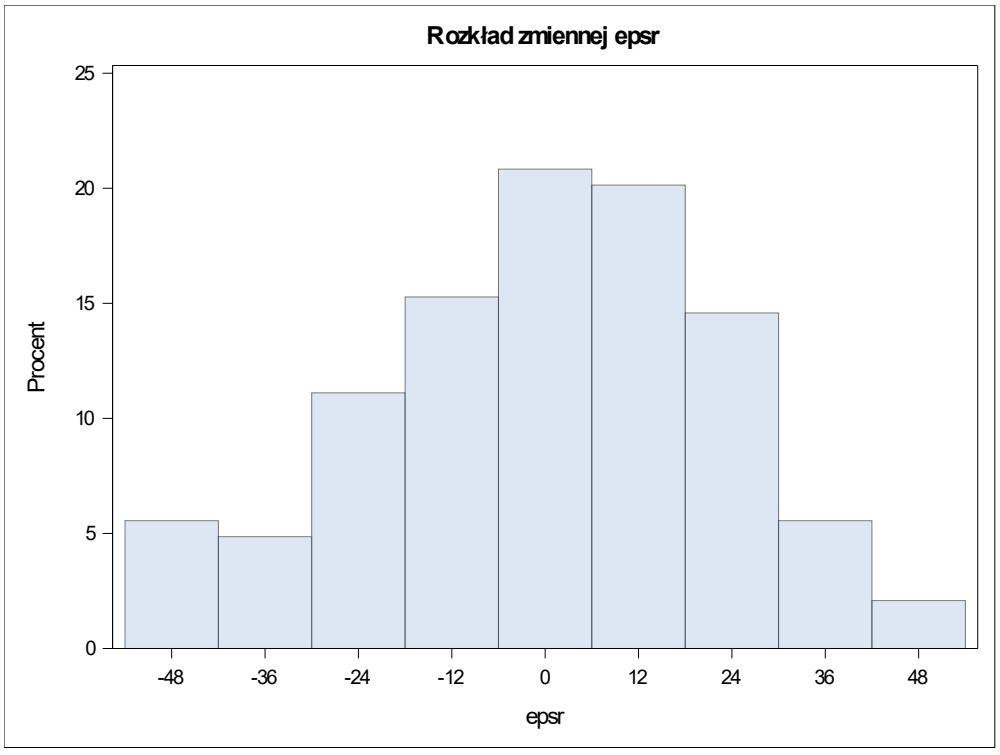
```
/* Histogram - błędy modelu rozszerzonego */

PROC UNIVARIATE data = Lab7A.ads1; /* miejsce, z którego pobieramy dane do
stworzenia histogramu */

	title 'Histogram - błędy modelu rozszerzonego';

	VAR epsr; /* dla kolumny col1 stworzymy histogram */

	HISTOGRAM epsr; /* stworzenie histogramu dla kolumny epsr */

run; /* uruchomienie tego fragmentu kodu */
```

Uzyskane wykresy:



Uwaga. eps – błędy modelu podstawowego, epsr – błędy modelu rozszerzonego.





Obserwacje. Oba histogramy błędów przypominają kształtem rozkład normalny.

Zadania (część 2,lab 7B):
1. Oblicz, korzystając z systemu SAS, wartości (dane w zbiorze Ads1):

- SST= $\sum_i \sum_j \left(y_{ij} - \bar{y}\right)^2$ - całkowita suma odchyleń
- SSM = $\sum_i n_i \cdot (\bar{y_i} - \bar{y})^2$ - odchylenia międzygrupowe
- SSE= $\sum_i \sum_j \left(y_{ij} - \bar{y_i}\right)^2$ - odchylenia w grupie

2. Oblicz:

- F = MSM/MSE
- MSM= SSM/(k-1)
- MSE= SSE/(n-k)

Model podstawowy

Aby wykonać zadanie 1 oraz zadanie 2, użyłem poniższego fragmentu kodu SAS-owego:

```
/* Model podstawowy */

/* Zadanie 1 oraz Zadanie 2 - Obliczenie SST, SSM, SSE, MSM, MSE, F */

data Lab7A.ads1;
    set Lab7A.ads1;

    /* Średnia wszystkich obserwacji - wielkość odczytana z tabeli z zad.1
(wstępna analiza danych) */
    mi = 66.8194444;

        /* Średnie dla klas - wielkości odczytane z tabeli z zad.2 (wstępna
analiza danych) */
        mean_display = 56.5555556;
        mean_paper   = 73.2222222;
        mean_people  = 66.6111111;
        mean_radio   = 70.8888889;



    /* Wkłady do SST */
    SST_basic = (Sales - mi) * (Sales - mi);
```

```
    /* Wkłady do SSE */
        if Ad = "display" then SSE_basic = (Sales - mean_display) * (Sales
- mean_display);
    if Ad = "paper"   then SSE_basic = (Sales - mean_paper) * (Sales -
mean_paper);
    if Ad = "people"  then SSE_basic = (Sales - mean_people) * (Sales -
mean_people);
    if Ad = "radio"   then SSE_basic = (Sales - mean_radio) * (Sales -
mean_radio);

    /* Wkłady do SSM */
        if Ad = "display" then SSM_basic = (mean_display - mi) *
(mean_display - mi);
    if Ad = "paper"   then SSM_basic = (mean_paper - mi) * (mean_paper -
mi);
    if Ad = "people"  then SSM_basic = (mean_people - mi) * (mean_people -
mi);
    if Ad = "radio"   then SSM_basic = (mean_radio - mi) * (mean_radio -
mi);


run;




proc summary data=Lab7A.ads1;
var SST_basic SSM_basic SSE_basic;
output out=Lab7A.ads1_totals_podstawowy sum=;
run;



data Lab7A.ads1_totals_podstawowy;
        set lab7a.ads1_totals_podstawowy;
        n = 144;
        k = 4;
        SSM_basic_test = SST_basic - SSE_basic;
        MSM_basic = SSM_basic / (k-1);
        MSE_basic = SSE_basic / (n-k);
        F_basic = MSM_basic / MSE_basic;
        F90_basic = 2.08380;
        F95_basic = 2.6049;
        F98_basic = 3.1161; /* tak naprawde to F dla alfa = 0.025  */
run;
```

Uzyskałem następujące wyniki (odczytałem je z wynikowej struktury Lab7A.ads1_totals_podstawowy):

n = 144 (liczba obserwacji)
k = 4 (liczba klas – typów reklamy)

SST = 26169.305556
SSE = 20303.222222
SSM = 5866.0833333

MSM = 1955.3611111
MSE = 145.02301587
F = 13.483108866

F90 = 2.08380 (wartość F dla alfa = 0.1, df1 = 4-1 = 3, df2 = nieskończoność)
F95 = 2.6049 (wartość F dla alfa = 0.05, df1 = 4-1 = 3, df2 = nieskończoność)
F98 = 3.1161 (wartość F dla alfa = 0.025, df1 = 4-1 = 3, df2 = nieskończoność)

Model rozszerzony

Aby wykonać zadanie 1 oraz zadanie 2, użyłem poniższego fragmentu kodu SAS-owego:

```
/* Model rozszerzony */


/* Zadanie 1 oraz Zadanie 2 - Obliczenie SST, SSM, SSE, MSM, MSE, F */


data Lab7A.ads1;

    set Lab7A.ads1;


    /* Średnia wszystkich obserwacji - wielkość odczytana z tabeli z zad.2a
(wstępna analiza danych) */

    mi = 66.8194444;
```

```
        /* Średnie dla klas - wielkości odczytane z tabeli z zad.2b (wstępna
analiza danych) */

        mean_display = 56.5555556;

        mean_paper   = 73.2222222;

        mean_people  = 66.6111111;

        mean_radio   = 70.8888889;




        /* Średnie wartości obserwacji względem obszaru - uzyskane z
pomocniczych operacji */

            mean1  = 60.000;

                mean2  = 64.375;

                mean3  = 81.000;

                mean4  = 79.250;

                mean5  = 72.625;

                mean6  = 66.375;

                mean7  = 58.750;

                mean8  = 73.375;

                mean9  = 57.625;

                mean10 = 77.750;

                mean11 = 52.250;

                mean12 = 69.750;

                mean13 = 67.000;

                mean14 = 64.125;

                mean15 = 67.000;

                mean16 = 69.250;

                mean17 = 53.875;

                mean18 = 68.375;
```

```
/* Średnia względem typu reklamy, dla danej obserwacji */
    if Ad = "display" then mean_ad = mean_display;
    if Ad = "paper"   then mean_ad = mean_paper;
    if Ad = "people"  then mean_ad = mean_people;
    if Ad = "radio"   then mean_ad = mean_radio;

    /* Średnia względem obszaru, dla danej obserwacji */
    if Area =  1 then     mean_area = mean1;
    if Area =  2 then     mean_area = mean2;
    if Area =  3 then     mean_area = mean3;
    if Area =  4 then     mean_area = mean4;
    if Area =  5 then     mean_area = mean5;
    if Area =  6 then     mean_area = mean6;
    if Area =  7 then     mean_area = mean7;
    if Area =  8 then     mean_area = mean8;
    if Area =  9 then     mean_area = mean9;
    if Area = 10 then     mean_area = mean10;
    if Area = 11 then     mean_area = mean11;
    if Area = 12 then     mean_area = mean12;
    if Area = 13 then     mean_area = mean13;
    if Area = 14 then     mean_area = mean14;
    if Area = 15 then     mean_area = mean15;
    if Area = 16 then     mean_area = mean16;
    if Area = 17 then     mean_area = mean17;
    if Area = 18 then     mean_area = mean18;
```

```
    /* Wkłady do SST */
    SST_extended = (Sales - mi)**2;


    /* Wkłady do SSE */
        SSE_extended = (Sales - mean_ad - mean_area + mi)**2;


    /* Wkłady do SSM */
        SSM_extended_ad = (mean_ad - mi)**2;
        SSM_extended_area = (mean_area - mi)**2;

run;


proc summary data=Lab7A.ads1;
var   SST_extended   SSM_extended_ad   SSM_extended_area   SSE_extended;
output out=Lab7A.ads1_totals_rozszerzony sum=;
run;


data Lab7A.ads1_totals_rozszerzony;
        set lab7a.ads1_totals_rozszerzony;
        n = 144;
        k = 18+4;
        SSM_extended = SSM_extended_ad + SSM_extended_area;
        MSM_extended = SSM_extended / (k-2);
        MSE_extended = SSE_extended / (n-k+1);
        F_extended = MSM_extended / MSE_extended;
```

```
/*    Biorę wyniki dla df1 = 20 (czyli k-2)
      oraz df2 = INFINITY (bo nie było 144-1=143) */

        F90_extended = 1.4206;

        F95_extended = 1.5705;

        F98_extended = 1.7085; /* tak naprawde to F dla alfa = 0.025  */

run;
```

Uzyskałem następujące wyniki (odczytałem je z wynikowej struktury Lab7A.ads1_totals_rozszerzony):

n = 144 (liczba obserwacji)
k = 22 (liczba typów reklamy 4 + liczba obszarów 18)

SST = 26169.305556
SSE = 11037.916667
SSM = 15131.388849

MSM = 756.56944246
MSE = 89.739159892
F = 8.4307613686

F90 = 1.4206 (wartość F dla alfa = 0.1, df1 = 20, df2 = nieskończoność)
F95 = 1.5705 (wartość F dla alfa = 0.05, df1 = 20, df2 = nieskończoność)
F98 = 1.7085 (wartość F dla alfa = 0.025, df1 = 20, df2 = nieskończoność)

3. Sprawdź hipotezę zerową dla przedziałów ufności: 90, 95, 98

Model podstawowy

Zauważmy, że dla przedziału ufności 90 mamy F90 = 2.0838.
Podobnie, dla przedziału ufności 95 mamy F95 = 2.6049.
Podobnie, dla przedziału ufności 98 mamy F98 = 3.1161.

Dalej, widzimy, że wartość F = 13.483108866 (dla naszych danych) jest większa od wartości F90, F95 oraz F98, co implikuje, że wartość p_value dla naszych danych jest mniejsza od wartości p_value dla przedziałów ufności 90, 95, 98. Ostatni wniosek implikuje, że hipotezę zerową należy odrzucić dla każdego rozważanego przedziału ufności.

Model rozszerzony

Zauważmy, że dla przedziału ufności 90 mamy F90 = 1.4206.
Podobnie, dla przedziału ufności 95 mamy F95 = 1.5705.
Podobnie, dla przedziału ufności 98 mamy F98 = 1.7085.

Dalej, widzimy, że wartość F = 8.4307613686 (dla naszych danych) jest większa od wartości F90, F95 oraz F98, co implikuje, że wartość p_value dla naszych danych jest mniejsza od wartości p_value dla przedziałów ufności 90, 95, 98. Ostatni wniosek implikuje, że hipotezę zerową należy odrzucić dla każdego rozważanego przedziału ufności.

Przy testowaniu hipotez należy skorzystałem z tabel statystycznych:
http://www.socr.ucla.edu/Applets.dir/F_Table.html