MWS Ćwiczenie 5

Regresja liniowa i porównywanie prób

Tomasz Korzeniowski, 265753

18 stycznia 2018

1 Zadanie 1

Mamy do dyspozycji dwie niezależne próby losowe z rozkładu normalnego o tej samej wariancji. Przyjmując, że tymi próbami są $X_1, \ldots, X_n \sim N(\mu_X, \sigma^2)$ oraz $Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma^2)$ możemy wyznaczyć

$$\overline{X} - \overline{Y} \sim N\Big(\mu_X - \mu_Y, \sigma^2(n^{-1} + m^{-1})\Big) \tag{1}$$

Do oszacowana wariancji σ^2 skorzystamy ze wzorów

$$s^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2} \tag{2}$$

gdzie

$$s_X^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

$$s_Y^2 = \frac{\sum_{i=1}^{m}(Y_i - \overline{Y})^2}{m-1}$$

Do wyznaczenia odchylenia standardowego błędu estymowanej różnicy $\mu_X - \mu_Y$ należy obliczyć

$$s = \sqrt{s^2(n^{-1} + m^{-1})} \tag{3}$$

Otrzymane wyniki prezentuje tabela 1:

μ_X	μ_Y	$\mu_X - \mu_Y$	s^2	s
1.0375	1.0754	-0.0379	1.4551	0.8092

Tabela 1: Statystyki otrzymane na podstawie dostępnych danych.

Następnie sprawdzamy hipotezę, czy zachodzi równość między średnimi μ_X oraz μ_Y. Skorzystamy z testu dwustronnego, ponieważ pozwala on stwierdzić ewentualną nierówność średnich. Test jednostronny byłby wstanie wskazać jedynie czy któraś ze średnich jest większa od drugiej. Przyjmiemy

$$H_0\text{: } \mu_X = \mu_Y$$
$$H_1\text{: } \mu_X \neq \mu_Y$$

Statystyką testową będzie

$$T = \frac{\overline{X} - \overline{Y}}{s\sqrt{n^{-1} + m^{-1}}} \tag{4}$$

Do obliczeń wykorzystamy funkcję z pakietu R - *t.test*. Test na poziomie istotności $\alpha = 0.1$ daje następujące wyniki:

```
> t.test(los1,los2,var.equal=TRUE, conf.level = 0.9)

        Two Sample t-test

data:  los1 and los2
t = -0.046837, df = 7, p-value = 0.964
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -1.570972  1.495172
sample estimates:
mean of x mean of y
   1.0375    1.0754
```

Otrzymana p-wartość jest równa 0.964 i jest większa od zadanego poziomu istotności. Wartość statystyki testowej $T = -0.046837$ jest mniejsza niż wartość krytyczna $c = F^{-1}_{t_{n+m-2}}(1 - \frac{\alpha}{2}) = 1.8946$. Z tych powodów nie ma podstaw do odrzucenia hipotezy zerowej ($p > \alpha$ oraz $T < c$).

2 Zadanie 2

Do sprawdzenia tego, że nie ma różnicy między łożyskami wykonanymi z dwóch różnych materiałów wykorzystamy hipotezy

H_0: łożyska są takie same
H_1: łożyska są różne

Jeśli założymy, że czas pracy łożyska do momentu uszkodzenia opisuje się rozkładem normalnym, będziemy mogli skorzystać z testu Studenta.

```
> t.test(lozyska[,1], lozyska[,2])

        Welch Two Sample t-test

data:  lozyska[, 1] and lozyska[, 2]
t = 2.0723, df = 16.665, p-value = 0.05408
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.07752643  7.96352643
sample estimates:
mean of x mean of y
   10.693     6.750
```

Na poziomie istotności $\alpha = 0.05$, p-wartość = 0.05408 świadczy o tym, że nie ma podstaw do odrzucenia hipotezy zerowej.

W sytuacji gdy nie przyjmiemy założenia normalności rozkładów musimy skorzystać z innego testu. Posłużymy się nieparametrycznym testem Wilcoxona (sumy rang). Jako niezależne próbki będziemy uważali łożyska wykonane z różnych materiałów. Pakiet R oferuje funkcję *wilcox.test*

```
> wilcox.test(lozyska[,1], lozyska[,2])

        Wilcoxon rank sum test

data:  lozyska[, 1] and lozyska[, 2]
W = 75, p-value = 0.06301
alternative hypothesis: true location shift is not equal to 0
```

Ponownie dowiadujemy się, że uzyskana p-wartość równa 0.06301 jest większa od przyjętego poziomu istotności. Zatem nie ma podstaw do odrzucenia hipotezy zerowej.

Bardziej odpowiednim testem do zbadania powyższych hipotez jest test Wilcoxona. Jest on zarówno bardziej uniwersalny, jak i nie wymaga założenia normalności rozkładów (nie ma powodów dla których powinniśmy zakładać, że rozkłady opisujące czas zużycia łożyska opisany jest rozkładem normalnym).

Ostatnim elementem zadania jest oszacowanie prawdopodobieństwa, że łożysko wykonane z pierwszego materiału będzie pracowało dłużej niż wykonane z materiału drugiego. Zakładając rozkład normalny, prawdopodobieństwo można opisać jako

$$P = \frac{|\{(x,y) \in X \times Y : x > y\}|}{|X \times Y|} = 0.75 \tag{5}$$

gdzie X i Y oznaczają łożyska wykonane odpowiednio z materiału pierwszego i drugiego.

Bardzo podobny efekt otrzymamy stosując metodę bootstrapu nieparametrycznego. Wykonując 1000 iteracji, w każdej losując 10000 razy ze zwracaniem wartości z odpowiednich typów (rodzajów materiałów) łożysk otrzymamy prawdopodobieństwo = 0.7499826.

3 Zadanie 3

W celu dopasowania zależności liniowej opisującej długość drogi hamowania y w funkcji prędkości v wykorzystamy równania

$$y = a \cdot v \tag{6}$$

$$\sqrt{y} = b \cdot v \tag{7}$$

Do wyliczenia powyższych współczynników można skorzystać z funkcji lm, która służy do dopasowania modelu liniowego do podanych danych. Podając odpowiednie zależności (parametr $y \sim v - 1$ wynika z faktu, że nie chcemy wyznaczać składowej stałej) otrzymujemy następujące wyniki:

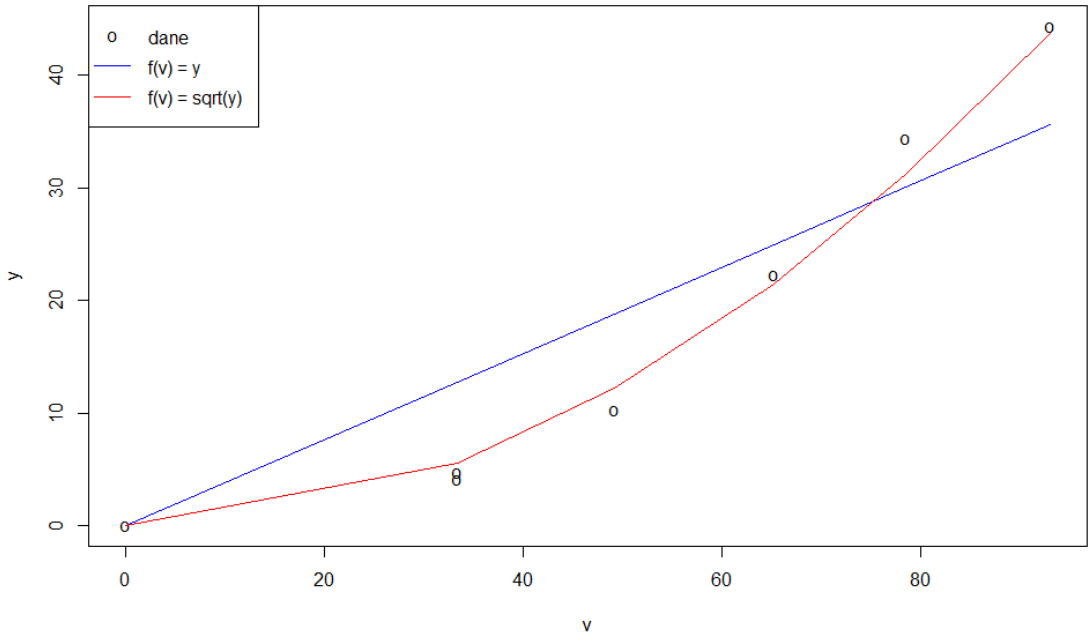
```
> lm(y ~ v - 1)
Call:
lm(formula = y ~ v - 1)

Coefficients:
     v
0.3834

> lm(sqrt(y) ~ v - 1)
Call:
lm(formula = sqrt(y) ~ v - 1)

Coefficients:
      v
0.07113
```

Interpretacją wyników są wartości współczynników $a = 0.3834$ oraz $b = 0.07113$. Graficznie obie wyznaczone zależności przedstawia wykres 1.



Wykres 1: Wyznaczone zależności liniowe.

Do zbadania, która ze znalezionych zależności jest bardziej zgodna z danymi wykorzystamy współczynnik determinacji R^2

$$R^2 = 1 - \frac{s_{\hat{Y}\hat{Y}}}{s_{YY}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - \overline{y})} \in [0, 1] \tag{8}$$

Im większa wartość współczynnika determinacji tym lepsze dopasowanie zależności do dostępnych danych.

	$f(v) = y$	$f(v) = \sqrt{y}$
R^2	0.8181624	0.9901931

Lepszym odwzorowaniem okazuje się zależność $f(v) = \sqrt{y}$. Na podstawie wykresu można zauważyć, że dane znajdują się bliżej czerwonej krzywej oznaczającej tę właśnie zależność.