Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Probe-Test
SA.01

Donnerstag, 20. 12. 2018

Dauer: 90 Minuten

Name, Vorname: _____

IDS Nr.: (Auf Rückseite HSLU-Karte) L _____

Unterschrift: _____

Aufgabe	1	2	3	4	Total
max. Punkte	5	10	10	10	35
erreichte Punktzahl					

Wichtige Hinweise

- Zugelassene Hilfsmittel:
  - Taschenrechner
  - Zusammenfassung: 10 einseitige A4-Seiten
  - Beliebige käufliche Formelsammlung
  - Papier und Schreibzeug
- Der Lösungsweg muss vollständig schriftlich festgehalten werden.
- **Alle Antworten müssen begründet werden. Ohne Lösungsweg gibt es 0 Punkte!**

Viel Erfolg!
Peter Büchel

Aufgabe 1: ..... (5 Punkte)

Nach einem internationalen Sportwettkampf werden Dopingtests durchgeführt. Dabei wird ein neues Schnelltestverfahren eingesetzt. Falls ein Sportler gedopt ist, dann ist der Schnelltest mit einer Wahrscheinlichkeit von 90 % positiv. Falls ein Sportler in Tat und Wahrheit nicht gedopt ist, zeigt der Schnelltest dennoch in 5 % der Fälle ein positives Ergebnis. Aus Erfahrung wissen die Dopingkontrolleure, dass 10 % der Sportler gedopt sind. Wir betrachten die folgenden Ereignisse:

D : Der Sportler ist gedopt	$\overline{D}$: Der Sportler ist nicht gedopt
T^+ : Der Schnelltest ist positiv	T^- : Der Schnelltest ist negativ

Wie gross ist die Wahrscheinlichkeit, dass ein Sportler gedopt ist, wenn der Test negativ ist?

Lösung 1:

Die bedingte Wahrscheinlichkeit, dass ein Sportler gedopt ist, wenn der Test negativ ist, ist gegeben durch

$$\begin{aligned}
P(D|T^-) &= \frac{P(D \cap T^-)}{P(T^-)} \\
&= \frac{P(T^-|D)P(D)}{P(T^-)} \\
&= \frac{P(T^-|D)P(D)}{P(T^-|D)P(D) + P(T^-|\overline{D})P(\overline{D})} \\
&= \frac{(1 - P(T^+|D))\, P(D)}{(1 - P(T^+|D))\, P(D) + (1 - P(T^+|\overline{D}))P(\overline{D})} \\
&= \frac{(1 - 0.9) \cdot 0.1}{(1 - 0.9) \cdot 0.1 + (1 - 0.05) \cdot 0.9} \\
&= 0.01
\end{aligned}$$

Aufgabe 2: ..... (10 Punkte)

Man vermutet, dass Cholesterinwerte bei Männern mit zunehmenden Alter höher werden. Die Ergebnisse einer Untersuchung der Cholesterinwerte von je 11 Männern in den Altersgruppen 20-30 und 40-50 liegen im Datensatz `cholesterin` vor. Gibt es einen statistisch signifikanten Unterschied zwischen den beiden Altersgruppen?

```r
cholesterin <- read.csv("cholesterin.csv")
```

a) Handelt es sich um einen gepaarten oder ungepaarten Test? Begründen Sie Ihre Antwort

b) Welche Art von Test würden Sie hier verwenden? Begründen Sie Ihre Antwort.

c) Führen Sie auf Signifikanznivneau von 5 % diesen Test durch.
Formulieren Sie Null- und Alternativhypothese und fällen Sie den Testentscheid.

Lösung 2:

a) Die Messungen sind hier sicher *ungepaart*, weil es keine Zuordnung des Cholesterinspiegel eines Mannes der einen Gruppe auf den Cholesterinspiegel eines Mannes der anderen Gruppe gibt. Wir könnten auch einfach verschieden grosse Gruppen wählen.

Bemerkung: Etwas anderes wäre es, wenn wir eine Gruppe aus 11 Männern zwischen 20 und 30 wählen, diese auf den Cholesterinspiegel untersuchen und die Messung bei denselben Männern 20 Jahre später nochmals machen würden. Dann hätten wir einen gepaarten Test, da zu jedem Mann genau 2 Testergebnisse gehören.

b) Da wir aus der Aufgabenstellung keinen Anhaltspunkt über die Art der Verteilung haben, wählen wir einen ungepaarten Wilcoxon-Test.

Weiter wählen wir einen *einseitigen* Test, da wir vermuten, dass die Cholesterinwerte der jüngeren Männer tiefer sind, als die der älteren.

Bemerkung: Dabei wird hier noch stillschweigend die Annahme gemacht, dass die Verteilung symmetrisch ist. Diese Annahme ist allerdings oft erfüllt. Sind wir ganz vorsichtig, so müssten wir einen Vorzeichentest machen.

c) Wir bezeichnen mit X die Zufallsvariable für die Cholesterinwerte der jüngeren Männer und mit Y die Zufallsvariable für die Cholesterinwerte der älteren Männer.

Die Nullhypothese H_0 ist

$$H_0: \quad \mu_X = \mu_Y$$

und die Alternativalternative H_A

$$H_A: \quad \mu_X < \mu_Y$$

Denn Test führen wir mit R durch

```r
setwd("/home/euler/Dropbox/Statistics/Statistics_W_master/Testat/HS18")
cholesterin <- read.csv("cholesterin.csv")
head(cholesterin)

##   mann.jung mann.alt
## 1       135      294
## 2       222      311
## 3       251      286
## 4       260      264
## 5       269      277
## 6       235      336

x <- cholesterin[,"mann.jung"]

y <- cholesterin[,"mann.alt"]

wilcox.test(x, y, paired=F, alternativ="less")

##
##  Wilcoxon rank sum test
##
## data:  x and y
## W = 42, p-value = 0.1213
## alternative hypothesis: true location shift is less than 0
```

Der p-Wert ist 0.1213, also grösser als 0.05 und damit wird die Nullhypothese nicht verworfen. Das heisst, aufgrund dieser Testgruppen kann *kein* statistisch signifikanter Unterschied der Cholesterinwerte zwischen jüngeren und älteren Männern festgestellt werden.

Aufgabe 3: ..... (10 Punkte)

Im Datensatz **creativity** ist das Alter der höchsten Kreativität von 36 Künstlern aufgeführt. Diese werden in 3 *Kategorien* aufgeteilt: Novelisten, Poeten und Filmregisseure (directors). Zusätzlich wird zwischen zwei *Typen* unterschieden: Conceptualists (finders) / experimentalists (seekers).

Das Kriterium für das Alter der grössten Kreativität:

- Poeten: Medianalter für ihre 5 der am meisten gedruckten Gedichten
- Novelisten: Medianalter der Publikation der 3 wichtigsten Romane
- Regisseure: Medianalter für die Filme mit den höchsten Filmbewertungen

```
df <- read.csv("creativity.csv")
head(df)
```

a) Erstellen Sie je ein Boxplot, der das Alter der höchsten Kreativität mit der Kategorie bzw. Typ zeigt. Interpretieren Sie diese beiden Boxplots.

```
boxplot(...,
        data=df,
        col=c("lightblue", "forestgreen", "orange"))
```

b) Erstellen Sie einen Interaktionsplot mit Typ, Kategorie und Alter der höchsten Kreativität. Interpretieren Sie diesen Plot.

```
interaction.plot(...,
                 type="b",
                 col=c("blue","forestgreen","orange"),
                 pch=c(19,17,15))
```

c) Gibt es auf Signifikanznivneau von 5 % statistisch signifikante Unterschiede (ohne Interaktion) zwischen dem Alter der höchsten Kreativität und Kategorie und Typ?
Stellen Sie Null- und Alternativhypothese auf und fällen Sie den Testentscheid.

d) Gibt es auf Signifikanznivneau von 5 % statistisch signifikante Interaktion zwischen Kategorie und Typ?
Stellen Sie Null- und Alternativhypothese auf und fällen Sie den Testentscheid.

Lösung 3:

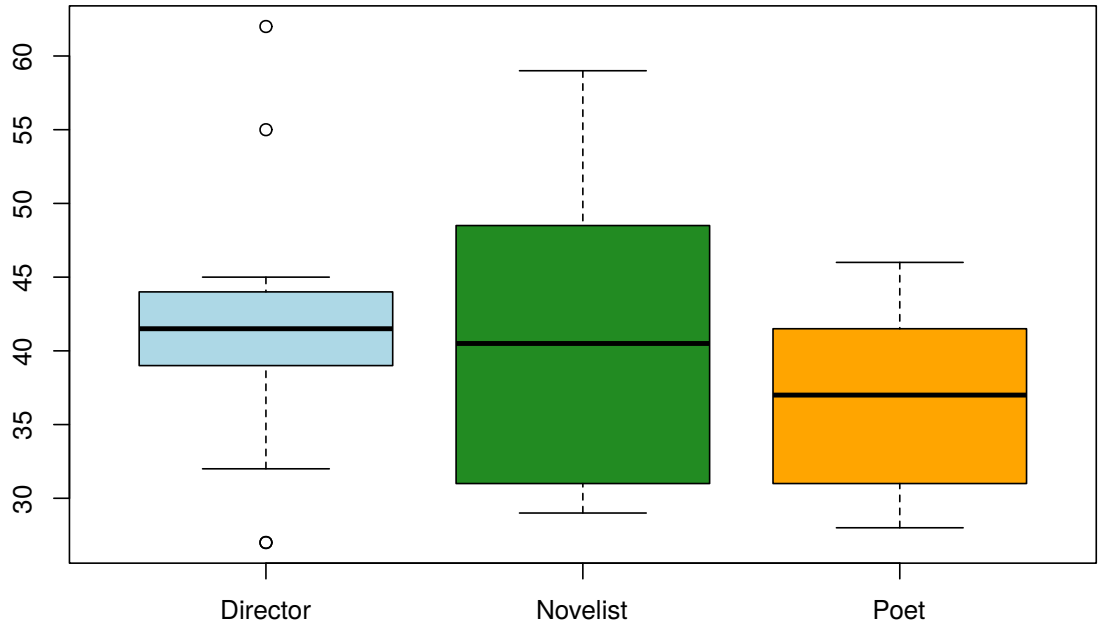
Einlesen der Datei:

```
df <- read.csv("creativity.csv")
head(df)

##    creator crCateg     crType agePeak
## 1    Eliot    Poet    concept      32
## 2 Cummings    Poet    concept      37
## 3    Plath    Poet    concept      30
## 4    Pound    Poet    concept      28
## 5   Wilbur    Poet    concept      29
## 6 Williams    Poet experiment      40
```

a) Boxplot:

```
boxplot(agePeak~crCateg,
        data=df,
        col=c("lightblue","forestgreen","orange"))
```



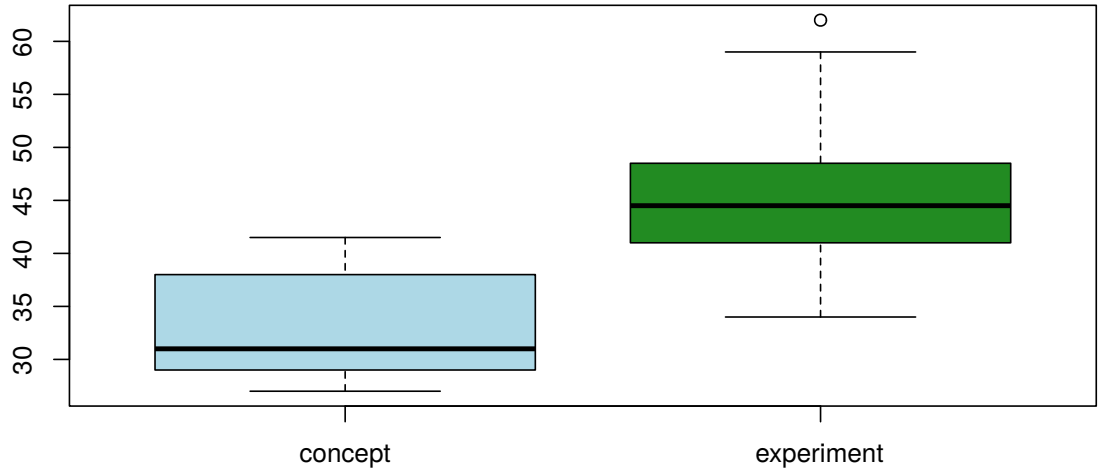
Bei den Regisseuren ist der Median des Alters der höchsten Kreativität bei etwa 42 Jahren, die Streuung ist nicht sehr gross.

Bei den Novelisten ist der Median etwas kleiner, aber die Streuung ist doch sehr gross.

Bei den Poeten ist der Median des Alters der höchsten Kreativität „deutlich" geringer als bei den Regisseuren und Novelisten. Ist der Unterschied aber statistisch signifikant? Dies muss mit einem Hypothesentest untersucht werden (siehe unten).

Boxplot:

```
boxplot(agePeak~crType,
        data=df,
        col=c("lightblue", "forestgreen", "orange"))
```

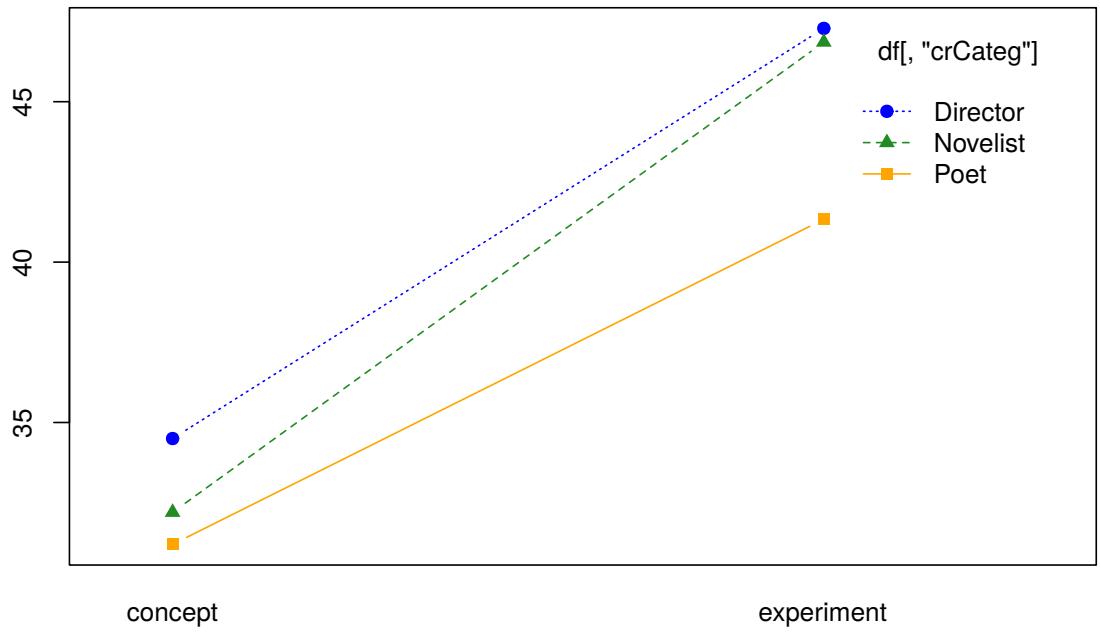


Hier zeigt sich ein deutlicher Unterschied in den Medianen. Bei den Konzeptialisten ist der das Alter der höchsten Kreativität bei gut 30 Jahren und bei

den Existentialisten bei 45 Jahren. Aber auch hier gilt: Ob dieser Unterschied statistisch signifikant ist oder nicht, muss ein Hypothesentest entscheiden.

b) Interaktionplot:

```
interaction.plot(df[, "crType"],
                 df[, "crCateg"],
                 df[, "agePeak"],
                 type="b",
                 col=c("blue","forestgreen","orange"),
                 pch=c(19,17,15))
```



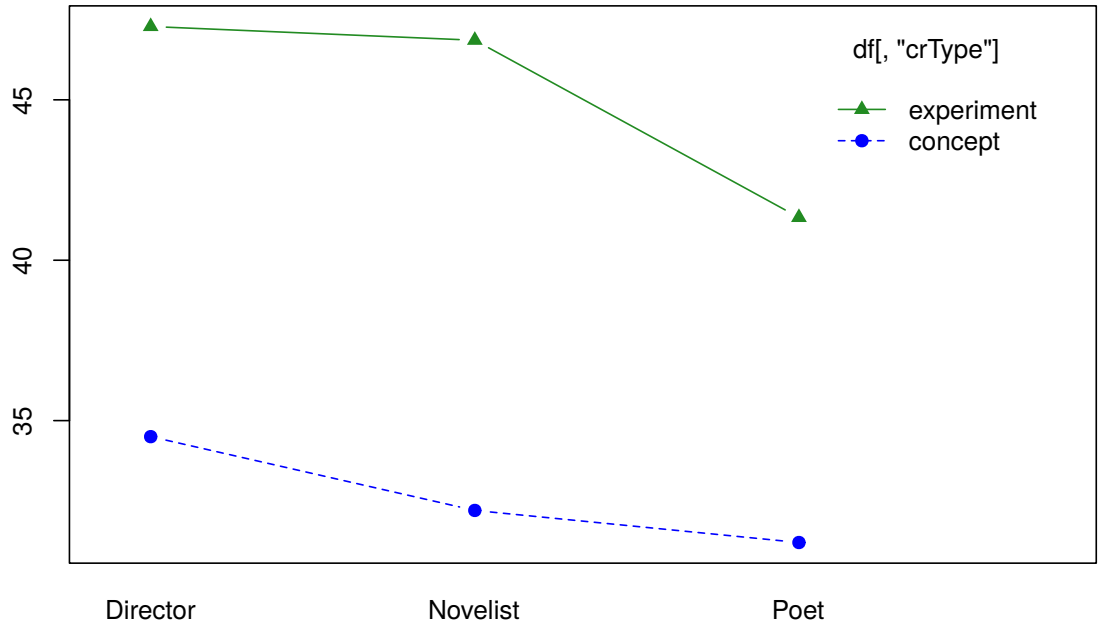
Im zweiten Boxplot von a) haben wir gesehen, dass die Konzeptialisten früher ihr Alter der grössten Kreativität erreichen, als die Existentialisten. Im Interaktionplot oben sehen wir, dass dies für alle Kategorien gilt (steigende Geraden). Diese Abhängigkeit ist also *unabhängig* von der Kategorie. Dies hat zur Folge, dass sich die Linien nicht überschneiden.

Das heisst, wir können hier Interaktion ausschliessen. Aber auch dies müssen wir erst mit einem Hypothesentest bestätigen.

Des weiteren liegen die Linien nicht auf einer Linie. Der Unterschied von Regisseuren und Novelisten ist relativ gering, aber die Poeten haben ein tieferes Alter der höchsten Kreativität. Die bestätigt die Aussagen zun ersten Boxplot in a).

Interaktionplot:

```
interaction.plot(df[,"crCateg"],
                 df[,"crType"],
                 df[,"agePeak"],
                 type="b", col=c("blue","forestgreen","orange"),
                 pch=c(19,17,15))
```



Die Argumentation geht analog zum Interaktionsplot vorher. Die Linien liegen weit auseinander (zweiter Boxplot a)) und es zeigt sich kein grosser Unterschied zwischen Regisseuren und Novelisten. Die Novelisten sind wieder jünger in ihrem Alter der höchsten Kreativität (erster Boxplot a)).

c) Wir haben zwei Nullhypothesen:

- Nullhypothese H_{01}: Das Alter der höchsten Kreativität ist bei allen drei Kategorien gleich.
  Alternativhypothese: H_{A1}: Das Alter der höchsten Kreativität ist mindestens einer Kategorie wesentlich anders als bei den anderen Kategorien.
- Nullhypothese H_{02}: Das Alter der höchsten Kreativität ist bei beiden Typen gleich.
  Alternativhypothese: H_{A2}: Das Alter der höchsten Kreativität der Typen ist verschieden.

```
summary(aov(df[, "agePeak"] ~ df[, "crType"] + df[, "crCateg"]))
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## df[, "crType"]   1 1411.2  1411.2  36.766 9.03e-07 ***
## df[, "crCateg"]  2  141.8    70.9   1.847    0.174
## Residuals       32 1228.3    38.4
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Der p-Wert für die Nullhypothese für Typ ist $9 \cdot 10^{-9}$, also wesentlich kleiner als das Signifikanzniveau von 5 %. Damit wird die Nullhypothese H_{02} deutlich verworfen. Es gibt also einen Unterschied des Alters der höchsten Kreativität

zwischen den Konzeptialisten und Existentialisten. Dies bestätigt unsere Beobachtungen von a) und b).

Der p-Wert der für die Nullhypothese für Typ ist 0.174, also einiges grösser als das Signifikanzniveau von 5 %. Damit wird die Nullhypothese H_{01} *nicht* verworfen. Es gibt also keinen Unterschied des Alters der höchsten Kreativität zwischen den Regisseuren, Novelisten und Poeten. Der tiefere Alter der Poeten ist also nicht statistisch signifikant.

d) Wird die Interaktion mitberücksichtigt, so kommt noch eine 3. Nullhypothese dazu:

Nullhypothese H_{03}: Es gibt keine Interaktion zwischen Typ und Kategorie.

Alternativhypothese: H_{A3}: Es gibt eine Interaktion zwischen Typ und Kategorie.

```
summary(aov(df[, "agePeak"] ~ df[, "crType"] * df[, "crCateg"]))

##                                 Df Sum Sq Mean Sq
## df[, "crType"]                   1 1411.2  1411.2
## df[, "crCateg"]                  2  141.8    70.9
## df[, "crType"]:df[, "crCateg"]   2   29.0    14.5
## Residuals                       30 1199.2    40.0
##                                 F value   Pr(>F)
## df[, "crType"]                   35.303 1.64e-06 ***
## df[, "crCateg"]                   1.773    0.187
## df[, "crType"]:df[, "crCateg"]    0.363    0.698
## Residuals
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Die p-Werte für Typ und Kategorie werden genau gleich interpretiert wie unter c).

Der p-Wert für die Interaktion ist 0.698 und damit weit grösser als 0.05. Die Nullhypothese H_{03} wird also nicht verworfen, es gibt also keine statistisch signifikante Interaktion.

Aufgabe 4: ..... (10 Punkte)

Forscher waren an den Faktoren interessiert, die den Wasserverbrauch (consumption) in 48 US-Staaten beeinflussen. Für jeden Staat wurde der Wasserverbrauch pro Person (in Gallonen pro Tag), das Einkommen pro Kopf (in $1000), der jährliche durchschnittliche Niederschlag (in inches) und die mittleren Kosten des Wassers von 1000 Gallonen (in Dollar) gemessen. Die Daten sind im Datensatz **water** aufgeführt

```
df <- read.csv("water.csv", header = T)
```

a) Stellen Sie formelmässig ein multiples Regressionsmodell mit dem Wasserverbrauch als Zielvariablen und den restlichen Variablen als Prädiktoren auf.

b) Bestimmen Sie die Parameter dieses Modelles und interpretieren Sie diese.

c) Welcher Anteil der Variation wird durch das Regressionsmodell erklärt?

d) Interpretieren Sie den p-Wert des zugehörigen F-Wertes.

e) Wir betrachten nun die einzelnen Regressionskoeffizienten. Deuten die Resultate an, dass wir irgendwelche Variablen vom Modell entfernen können? Begründen Sie Ihre Antwort.

Lösung 4:

Datei einlesen:

```
df <- read.csv("water.csv", header = T)
head(df)

##   STATE INCOME CONSUMPTION  RAIN COST
## 1    ME   9.04         115 43.52 1.30
## 2    NH  11.66          99 63.23 1.56
## 3    VT   9.62         100 33.69 2.50
## 4    MA  12.51          78 43.81 1.01
## 5    RI  10.89          66 45.32 1.56
## 6    CT  14.09          66 44.39 2.42
```

a) Modell:

$$\texttt{CONSUMTION} = \beta_0 + \texttt{INCOME} \cdot \beta_1 + \texttt{RAIN} \cdot \beta_2 + \texttt{COST} \cdot \beta_3$$

b) Parameter:

```r
fit <- lm(CONSUMPTION ~ INCOME + RAIN + COST, data=df)
summary(fit)

##
## Call:
## lm(formula = CONSUMPTION ~ INCOME + RAIN + COST, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.259 -20.499  -0.687  17.350 112.284
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 250.9477    41.8117   6.002 3.35e-07 ***
## INCOME       -4.9158     3.7692  -1.304    0.199
## RAIN         -1.9342     0.3683  -5.252 4.18e-06 ***
## COST        -17.7517    10.8043  -1.643    0.108
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.78 on 44 degrees of freedom
## Multiple R-squared:  0.4436,Adjusted R-squared:  0.4057
## F-statistic: 11.69 on 3 and 44 DF,  p-value: 9.21e-06
```

Der geschätzte Koeffizient für β_0 ist 251. Dieser Wert ist ein bisschen schwierig zu interpretieren. Das wäre der Wasserkonsum in Gallonen pro Tag, wenn man kein Einkommen hat, es im Jahr nichts regnet und die Wasser gratis ist.

Der geschätzte Koeffizient für β_1 ist -4.91. Verdient man 1000\$ mehr pro Jahr so sinkt der Wasserverbrauch um fast 5 Gallonen pro Tag. Dieser Zusammenhang ist nicht wirklich zu erklären, aber der zugehörige p-Wert ist mit 0.2 über dem Signifikanzniveau von 0.05. Somit ist der Einfluss des Einkommens auf den Wasserverbrauch nicht signifikant.

Der geschätzte Koeffizient für β_2 ist -1.93. Regnet es einen Inch pro Jahr mehr, so sinkt der Wasserverbrauch um fast 2 Gallonen pro Tag. Oder anders gesagt, nimmt der Wasserverbrauch mit weniger Regen zu. Es ist heisser, man duscht mehr, muss Rasen (oder in den USA Golfplätze) tränken etc. Der p-Wert ist auch statistisch signifikant.

Der geschätzte Koeffizient für β_3 ist -17.75. Nehmen die Kosten für 1000 Gallonen um einen Dollar zu, so sinkt der Wasserkonsum um knapp 18 Gallonen. Dies mag zwar einleuchtend sein, aber der p-Wert liegt über dem Signifikanzniveau von 5 %. Somit sind die Kosten des Wassers nicht statistisch signifikant für den Wasserverbrauch. Dies mag überraschen, aber kann vielleicht damit erklärt werden, dass das Wasser so billig ist, dass die Kosten auch für arme Haushalte nicht ins Gewicht fallen.

c) Der Anteil der Variation des Regressionsmodelles wird durch R^2 beschrieben. Dieser Wert ist 0.4436, also nicht sonderlich nahe bei 1. Es werden also nur etwa 44 % der Variation durch das Modell erklärt.

d) Die Nullhypothese für den F-Wert ist

$$\beta_1 = \beta_2 = \beta_3 = 0$$

Der p-Wert ist $9 \cdot 10^{-6}$ und damit weit unter dem Signifikanzniveau von 0.05. Somit ist ein β von 0 verschieden. Wenn man die p-Werte der einzelnen Koeffizienten betrachtet, dürfte dies β_2 sein (Regen).

e) Aus den Beobachtungen oben ist an sich nur die jährliche Niederschlagsmenge für den Wasserkonsum verantwortlich. Die beiden erklärenden Variablen können wir (vermutlich) ignorieren.