Übung 05 - Datenanalyse und Statistik WS 2008/2009

Aufgabe 1: Laden Sie die Ihnen hoffentlich bereits bekannte Datei "Grube.txt"von der Website herunter, von wo Sie auch die Übungen beziehen und lesen Sie den Datensatz in R ein!
Um nach Möglichkeit Messungen einzusparen, ist es nun das Ziel, Zusammenhänge zwischen den verschiedenen Belastungswerten des Wassers zu finden und diese zu quantifizieren.

(1) Welche statistische Graphik würde sich eignen, um einen ersten Eindruck von dem Zusammenhang zweier Verschmutzungsgrade zu bekommen?

---

(2) Beschreiben Sie den Zusammenhang zwischen der Bleikonzentration an der Messstelle A und der an der Messstelle B mittels einer entsprechenden Graphik!

---

(3) Die Grubenleitung möchte unter Zuhilfenahme einer linearen Regression die Bleikonzentration an der Messstelle B durch die Bleikonzentration an der Stelle A beschreiben. Formulieren Sie (ohne Berechnung der Koeffizienten) die Gleichung der dazugehörigen Regressionsgeraden!

---

(4) Spricht laut Streudiagramm etwas gegen die Anwendung einer linearen Regression?

---

(5) Führen Sie die unten angegebene Regression durch! Stellen Sie die entsprechende Regressionsgleichung auf!

```
reg <- lm( B.Pb ~ A.Pb, data = grube) # Lineare Regression
reg
```

---

(6) Was besagt der hier bestimmte Wert für R^2?

```
Y <- predict(reg) + resid(reg)  # Schätzung + Residuen
print( "R^2 = ")
print(  ( var(Y) - var(resid(reg)) ) / var(Y)  )
```

---

(7) Der p-Wert des nachfolgenden Tests wird unter Pr(>F) angegeben. Was genau wird hier getestet, und wie lautet das Ergebnis?

```
anova(reg) # Varianzanalyse (analysis of variance)
```

---

(8) Geben die hier aufgeführten diagnostischen Graphiken (Abb. 2-4) Anlass zur Beunruhigung (in Bezug auf die Regression)? Wenn ja, inwiefern? Begründen Sie in jedem Fall!

```
par( mfrow = c(2,2) )
plot( grube$A.Pb, grube$B.Pb )
abline(reg)      # Regressionsgerade

plot( predict(reg), resid(reg) , ylab="Residuen" )
plot( predict(reg), influence(reg)$hat  ,ylab="Hebelwirkung"
     ,ylim=c(0,1) )
plot( predict(reg), cooks.distance(reg) ,ylab="Cook distance" )
```

---

(9) Geben Sie für jede der in der vorherigen Aufgabe angegebenen diagnostischen Graphiken (Abb. 2-4) kurz an, was mit ihr überprüft wird!

---

(10) An einem beliebigem Tag betrage die Verunreinigung mit Blei an der Stelle A exakt 12 ppm pro 10 ml Grubenwasser. Geben Sie einen Bereich an, innerhalb dessen die Bleikonzentration an der Stelle B mit einer Wahrscheinlichkeit von 0.95 liegen wird! Wie nennt man diesen Bereich?

Tip: Wird das Konfidenzintervall der Regressionsgeraden schmaler oder breiter als dieser Bereich sein?

```
# Berechnungen für Konfidenzintervall und Vorhersageintervall
alpha <- 0.05
reg <- lm( B.Pb ~ A.Pb, data=grube, x=TRUE, y=TRUE)
a <- coef(reg)[1]
b <- coef(reg)[2]
xr <- range(reg$x[,2])
x <- seq( xr[1], xr[2], length.out=200)
v <- vcov(reg)
y <- a + b * x
sigmaHat2 <- summary(reg)$sigma^2
tq <- qt(1-alpha/2,df=reg$df.resid)
sqyVor  <- v[1,1]+2*v[2,1]*x+v[2,2]*x*x+sigmaHat2
sqyKonf <- v[1,1]+2*v[2,1]*x+v[2,2]*x*x

# Graphik
plot( grube$A.Pb, grube$B.Pb )
abline(reg)                      # Regressionsgerade
# Vorhersageintervall für die Punkte
lines(x, y + sqrt(sqyVor)*tq, col="red")
lines(x, y - sqrt(sqyVor)*tq, col="red")
# Konfidenzintervall für die Regressionsgerade
lines(x, y + sqrt(sqyKonf)*tq )
lines(x, y - sqrt(sqyKonf)*tq )
title(main="Vorhersage- und Konfidenzbereiche")
```

---

(11) Geben Sie eine Vorhersage für die an der Messstelle B auftretende Bleikonzentration an, wenn an der Messstelle A eine Verunreinigung von 18 ppm pro 10 ml Grubenwasser festgestellt wurde!

---

(12) Betrachten Sie mit nachfolgendem Befehl die Streudiagramme zu den Variablen des Datensatzes - für welche Kombination(en) von Variablen würde(n) sich lineare Regression(en) eignen?

```
pairs(grube)   # Streudiagramme aller Variablenkombinationen
```
