Arbeitsblatt 6

Multiple lineare Regression (Interaktionen)

Aufgabe 1: Hausisolierung

Wir analysieren den Effekt einer Isolierungssanierung in einem Haus. Der Datensatz heisst `whiteside` im R-Package `MASS` (siehe Unterricht Woche 1). Der Datensatz enthält Daten über drei Variablen:

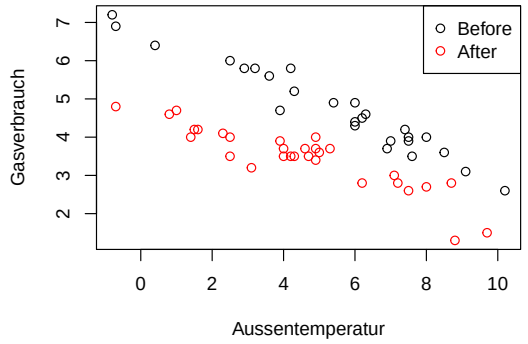
- Temp: Mittlere Aussentemperatur über eine Woche
- Gas: Wöchentlicher Gasverbrauch im Haus (in 1000 Feet^3)
- Insul (before/after): vor oder nach Einführung der Wärmedämmungsmassnahmen

```
library(MASS)
data(whiteside)
```

(a)

Zuerst stellen wir die Daten in einem geeigneten Streudiagramm dar. Tragen Sie `Gas` auf der y-Achse und `Temp` auf der x-Achse auf. Zusätzlich färben Sie die Datenpunkte bezüglich der Variable `Insul` ein. Beschreiben Sie den Zusammenhang.

```
plot(whiteside$Temp, whiteside$Gas, col = whiteside$Insul,
     xlab = "Aussentemperatur", ylab = "Gasverbrauch")
legend("topright", levels(whiteside$Insul), col = 1:2, pch = 1)
```

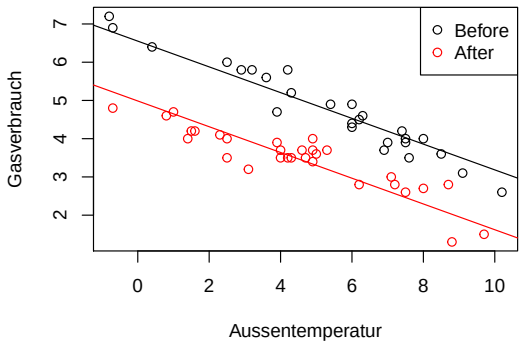


Die Grafik zeigt: Je höher die Temperatur desto geringer ist der Gasverbrauch. Man kann einen klaren Unterschied im Verbrauch zwischen isolierten (rote Punkte) und nicht-isolierten Häusern (schwarze Punkte) sehen.

(b)

Führen Sie eine multiple Regression mit erklärenden Variablen `Insul` und `Temp` durch, um den Gasverbrauch zu modellieren. Zeichnen Sie das angepasste Modell in das Streudiagramm ein.

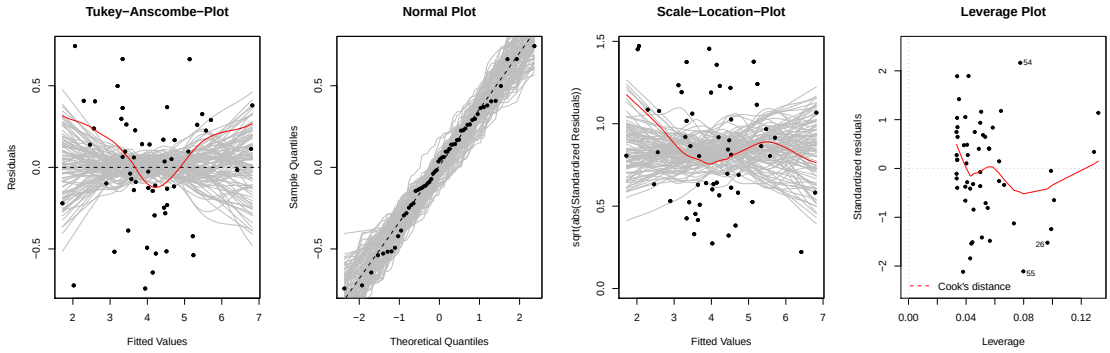
```
fit1 <- lm(Gas ~ Temp + Insul, data = whiteside)
plot(whiteside$Temp, whiteside$Gas, col = whiteside$Insul,
     xlab = "Aussentemperatur", ylab = "Gasverbrauch")
legend("topright", levels(whiteside$Insul), col = 1:2, pch = 1)
abline(a = coef(fit1)[1], b = coef(fit1)[2])
abline(a = coef(fit1)[1] + coef(fit1)[3], b = coef(fit1)[2], col = "red")
```



(c)

Prüfen Sie die Gültigkeit des Regressionsmodells mittels Residuenanalyse. Erzeugen Sie noch zwei zusätzliche Grafiken, einmal für die isolierten und einmal für die nicht isolierten Häuser, mit den Residuen auf der y-Achse und der `Temperatur` auf der x-Achse. Wie sollten die Punkte in dieser Zusatzgrafik verteilt sein, falls die Modellannahmen erfüllt sind? Ist dies der Fall?

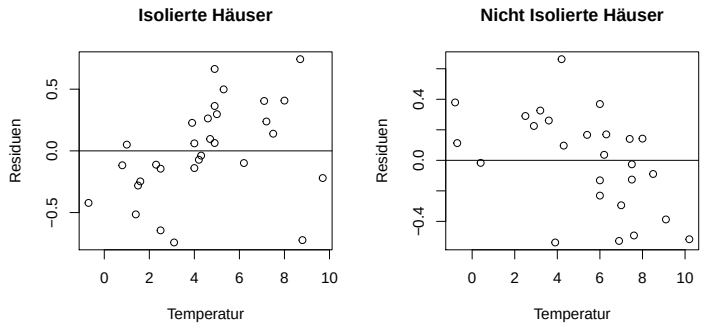
```
par(mfrow = c(1,4))
resplot(fit1)
```



Der Tukey-Anscombe Plot zeigt eine systematische Abweichung von der Horizontalen. Der Zusammenhang ist nicht korrekt modelliert.

Zusatzgrafik: Falls die Modellvoraussetzungen erfüllt sind, dann müssten die Datenpunkte zufällig um die horizontale Nulllinie verteilt sein.

```
par(mfrow = c(1,2))
plot(whiteside$Temp[whiteside$Insul == "After"],
     resid(fit1)[whiteside$Insul == "After"],
     main = "Isolierte Häuser", xlab = "Temperatur",
     ylab = "Residuen")
abline(h = 0)
plot(whiteside$Temp[whiteside$Insul == "Before"],
     resid(fit1)[whiteside$Insul == "Before"],
     main = "Nicht Isolierte Häuser", xlab = "Temperatur",
     ylab = "Residuen")
abline(h = 0)
```



Die Datenpunkte zeigen ein Muster. Bei den isolierten Häuser sind bei niedrigen Temperaturen die meisten Residuen negativ und bei hohen Temperaturen hat die Mehrheit der Residuen ein positives Vorzeichen. Bei nicht isolierten Häusern zeigt sich das umgekehrte Bild. Dies deutet auf eine fehlende Interaktion zwischen `Temp` und `Insul` hin.

(d)

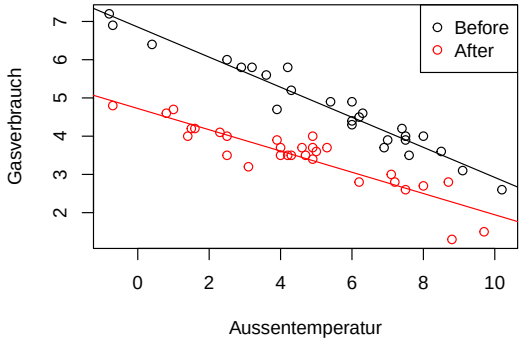
Fügen Sie dem Modell zusätzlich eine Interaktion zwischen `Temp` und `Insul` hinzu. Zeichnen sie das Modell in ein Streudiagramm auf. Sind die Modellannahmen nun besser erfüllt?

Anpassung der Regression mit Interaktionsterm:

```
fit2 <- lm(Gas ~ Temp + Insul + Temp:Insul, data = whiteside)
```

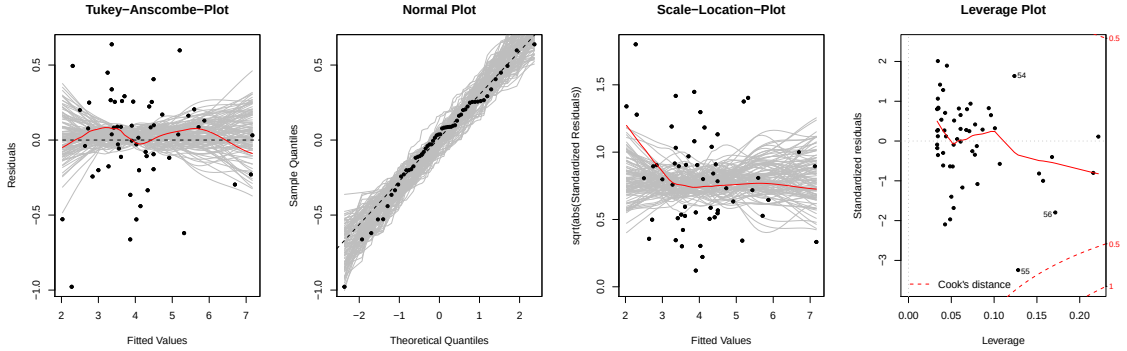
Das Modell entspricht nun 2 Geraden mit unterschiedlichem Achsenabschnitt und Steigung:

```
plot(whiteside$Temp, whiteside$Gas, col = whiteside$Insul,
     xlab = "Aussentemperatur", ylab = "Gasverbrauch")
legend("topright", levels(whiteside$Insul), col = 1:2, pch = 1)
abline(a = coef(fit2)[1], b = coef(fit2)[2])
abline(a = coef(fit2)[1] + coef(fit2)[3], b = coef(fit2)[2] + coef(fit2)[4], col = "red")
```



Die Residuenanalyse zeigt nun ein besseres, aber noch nicht ganz optimales Bild:

```
par(mfrow = c(1,4))
resplot(fit2)
```



(e)

Prüfen Sie die Signifikanz des Effekts einer Isolierungssanierung mit einem geeigneten Test auf dem 5% Niveau.

```
summary(fit2)
```

```
Call:
lm(formula = Gas ~ Temp + Insul + Temp:Insul, data = whiteside)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97802 -0.18011  0.03757  0.20930  0.63803

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.85383    0.13596  50.409  < 2e-16 ***
Temp         -0.39324    0.02249 -17.487  < 2e-16 ***
InsulAfter   -2.12998    0.18009 -11.827 2.32e-16 ***
```

```
Temp:InsulAfter  0.11530    0.03211   3.591 0.000731 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.323 on 52 degrees of freedom
Multiple R-squared:  0.9277,    Adjusted R-squared:  0.9235
F-statistic: 222.3 on 3 and 52 DF,  p-value: < 2.2e-16
```

Der Effekt der Isolierung lässt sich mittels t-Test auf Signifikanz prüfen. Er ist signifikant und hängt nachweisbar mit der Aussentemperatur zusammen. Je kälter es ist, desto grösser ist der Unterschied im Gasverbrauch.

(f)

Wie hoch schätzen Sie, dass der Gasverbrauch in einem isolierten und einem nicht isoliertem Haus sein wird, wenn die Aussentemperatur über eine Woche um 0 Grad beträgt? Geben Sie einen Bereich an, in welchem der Gasverbrauch mit 95% Wahrscheinlichkeit enthalten sein wird?

```
dat <- data.frame(Temp = c(0,0), Insul = c("Before","After"))
predict(fit2, newdata = dat, interval = "prediction")
```

```
       fit      lwr      upr
1 6.853828 6.150591 7.557065
2 4.723850 4.033731 5.413968
```

Der erwartete wöchentliche Gasverbrauch bei einem nicht isoliertem Haus liegt bei 6854 Feet^3. Mit 95% liegt der Gasverbrauch zwischen 6151 Feet^3 und 7557 Feet^3. Bei einem isoliertem Haus liegt der Gasverbrauch mit 4724 Feet^3 deutlich tiefer. Mit 95% liegt der Verbrauch zwischen 4034 Feet^3 und 5414 Feet^3.

Aufgabe 2: Berufsansehen

Der Datensatz `Prestige` stammt aus einer kanadischen Studie über das Ansehen von 98 Berufsgruppen in Kanada (Statistics Canada, 1971). Folgende Variablen stehen zur Verfügung:

`census`	ID für die Berufgruppe
`prestige`	mittleres Ansehen der Berufsgruppen
`education`	durchschittliche Ausbildungszeit (in Jahren)
`income`	mittleres Einkommen (in kanadischen Dollar)
`women`	mittlerer Anteil an Frauen (in Prozent)
`type`	Art der Beschäftigung: bc = Arbeiter; wc = Angestellte; prof = Selbstständige, Manager und Techniker

Die Daten stehen ihnen in `Prestige` im R-Package `car` zur Verfügung.

```
library(car, quietly = TRUE)
data(Prestige)
```

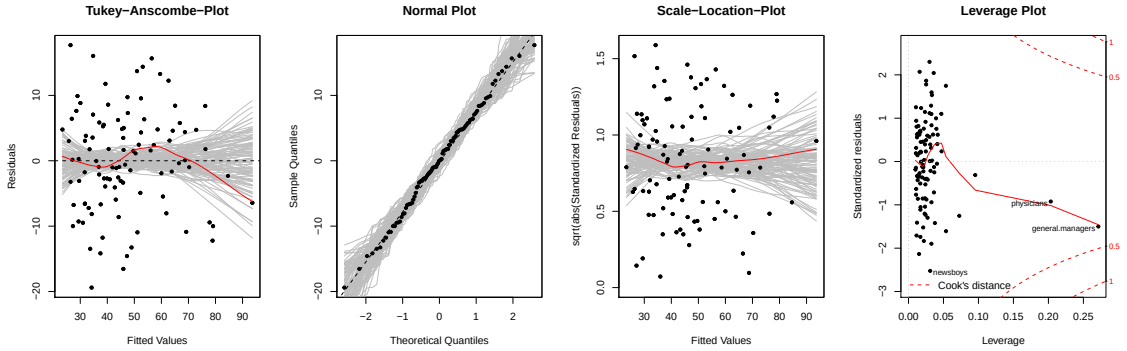
(a)

Passen Sie zuerst das folgende Modell an:

```
fit.prestige1 <- lm(prestige ~ income + education, data = Prestige)
```

und prüfen Sie mittels grafischer Residuenanalyse, ob die Modellvoraussetzungen erfüllt sind.

```
par(mfrow = c(1,4))
resplot(fit.prestige1)
```



- **Tukey-Anscombe-Plot**: Glätter verläuft wellenförmig.
- **Normalplot**: Die Punkte verlaufen auf der Winkelhalbierenden, Normalverteilung ok.
- **Scale-Location-Diagramm**: Der Glätter hat eine leichte Krümmung, ist aber mit den grauen simulierten Glättungskurven vereinbar.
- **Leverage-Plot**: Es gibt zwei Punkte mit recht grossem Hebelwirkung, deren Einfluss ist aber nach Distanz von Cook nicht gefährlich für die Anpassung.

FAZIT: Die Voraussetzungen sind nicht erfüllt.

(b)

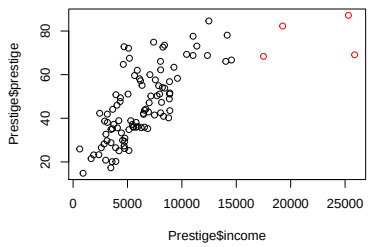
Bestimmen Sie alle Berufsgruppen mit grosser Hebelwirkung. Wenn Sie die Daten deskriptiv betrachten, können Sie sich erklären, warum diese Berufsgruppen eine grosse Hebelwirkung haben?

```
s <- hatvalues(fit.prestige1) > 2*3/nrow(Prestige) # Faustregel: 2*(p+1)/n
rownames(Prestige)[s]
```

```
[1] "general.managers"         "lawyers"                  "physicians"
[4] "osteopaths.chiropractors"
```

Die 4 Berufsgruppen fallen durch ihr sehr hohes Einkommen auf. Ihr Ansehen ist im Vergleich zum Verlauf der anderen Punkte etwas zu klein.

```
plot(Prestige$income, Prestige$prestige, col = s+1)
```



(c)

Passen Sie das Regressionsmodell mit und ohne die Punkte mit grosser Hebelwirkung an. Vergleichen Sie die Koeffizienten.

```
fit.prestige1b <- lm(prestige ~ income + education, data = Prestige[!s,])
coef(fit.prestige1)
```

```
 (Intercept)       income    education
-6.847778720  0.001361166  4.137444384
```

```
coef(fit.prestige1b)
```

```
(Intercept)      income   education
-7.70955035  0.00210334  3.81262803
```

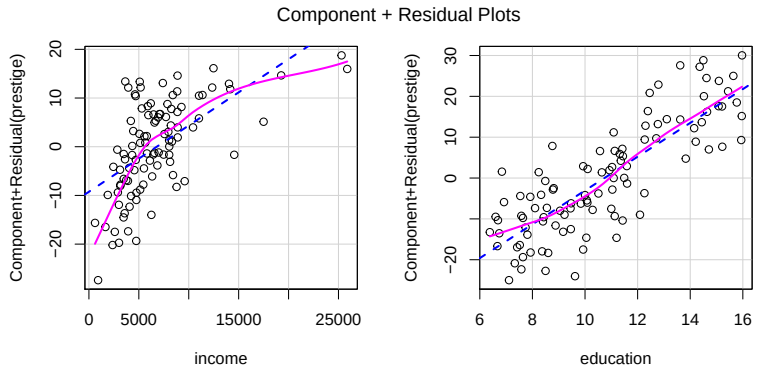
Die Änderungen liegen innerhalb der Standardabweichung der geschätzten Koeffizienten. Daher werden die Punkte von der Distanz von Cook auch nicht als einflussreich kategorisiert.

(d)

Studieren Sie die Modelldefizite aus (a) anhand der partiellen Residuenplots genauer und zeichnen Sie die Residuen auch gegen die 2 (potentiellen) erklärenden Variablen, die noch nicht im Modell aufgenommen sind, auf.

Die partiellen Residuenplots lassen sich mit der Funktion `crPlots` aus dem R-Package `car` erstellen:

```
crPlots(fit.prestige1)
```

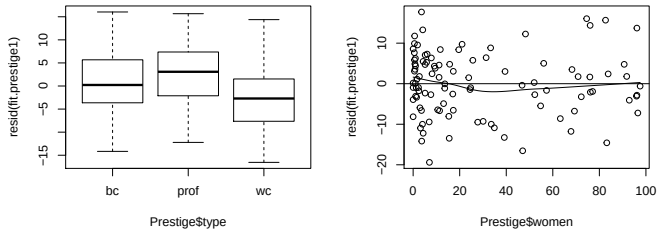


Der partielle Zusammenhang zwischen Income und Prestige ist eindeutig nichtlinear. Es braucht hier eine Logarithmus-Transformation.

Für die Variable Education ist der Glätter einigermassen linear. Eventuell könnte man noch eine kleine Verbesserung des Modells erzielen, wenn man zusätzlich noch eine Variable education kleiner und grösser 12 Jahre einfügt.

Nun betrachten wir noch die 2 (potentiellen) erklärenden Variablen:

```
par(mfrow = c(1,2))
plot(resid(fit.prestige1) ~ Prestige$type)
scatter.smooth(resid(fit.prestige1) ~ Prestige$women)
abline(h = 0)
```



Die Art der Beschäftigung (`type`) zeigt ein Muster in den Residuen. Diese sollte also in die Modellierung aufgenommen werden. Beim Frauenanteil (`women`) ist kein zu deutliches Muster erkennbar. Hier muss aber sicher auch der Einfluss der anderen Variablen beachtet werden.

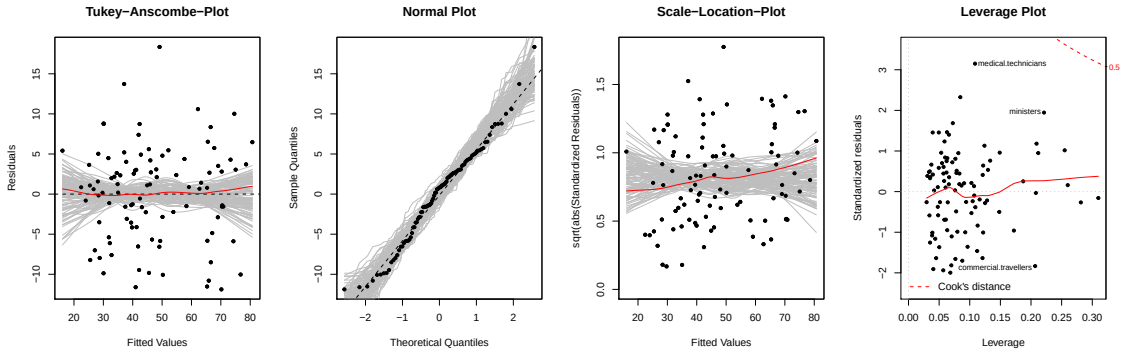
(e)

Passen Sie das folgende Regressionsmodell an:

```
Prestige$log_Income <- log(Prestige$income)
fit.prestige2 <- lm(prestige ~ log_Income + education + women
                    + type + log_Income:type + log_Income:women,
          data = Prestige)
```

und überprüfen Sie wider die Modelleignung mittels Residuenanalyse.

```
par(mfrow = c(1,4))
resplot(fit.prestige2)
```



- **Tukey-Anscombe-Plot**: Glätter verläuft horizontal innerhalb der stochastischen Fluktuation - ok.
- **Normalplot**: Die Punkte verlaufen auf der Winkelhalbierenden, Normalverteilung ok.
- **Scale-Location-Diagramm**: Der Glätter hat eine leichte Steigung, ist aber innerhalb der grauen simulierten Glättungskurven.
- **Leverage-Plot**: Es gibt keine einflussreichen Punkte.

FAZIT: Die Voraussetzungen sind erfüllt. Das Modell ist somit gültig.

(f)

Wie ist die Interaktion zwischen `Income` und `type` zu interpretieren? Ist diese auf dem 5% Niveau signifikant?

Die Signifikanz der Interaktion lässt sich mit einem F-Test prüfen:

```
drop1(fit.prestige2, test = "F")
```

```
Single term deletions

Model:
prestige ~ log_Income + education + women + type + log_Income:type +
    log_Income:women
                Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                       3387.3 365.20
education        1   1066.86 4454.1 390.03 28.0318 8.524e-07 ***
log_Income:type  2    419.72 3807.0 372.64  5.5140  0.005526 **
log_Income:women 1    153.58 3540.8 367.54  4.0353  0.047587 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Der entsprechende p-Wert (= 0.005526) ist kleiner als 0.05. Damit ist die Interaktion signifikant.

Zur Interpretation betrachten wir die Koeffizienten:

```
dummy.coef(fit.prestige2)
```

```
Full coefficients are

(Intercept):          -179.0867
log_Income:            21.89802
education:             3.094687
women:                 1.259971
type:                        bc        prof          wc
                       0.0000000 116.0252289   0.6817113
log_Income:type:             bc        prof          wc
                       0.0000000 -12.4257224  -0.5043697
log_Income:women:     -0.1396265
```

Je nach Art der Beschäftigung, hat das Einkommen einen unterschiedlich starken Einfluss auf das Berufsansehen. Vergleicht man Arbeiter (`bc`) mit Selbstständigen, Managern und Technikern (`prof`), so steigt das Ansehen mit wachsendem Einkommen etwa ums Doppelte. Für Arbeiter (`bc`) und Angestellte (`wc`) kann der Einfluss von Einkommen aufs Ansehen als identisch angenommen werden, da der t-Test des zugehörigen Koeffizienten nicht signifikanter ist (siehe summary-Output).

(g)

Für die Berufsgruppe der Statistiker gibt es keine Messung des Ansehens. Wir nehmen folgendes an:

```
  log_Income women education type
1   10.59663    20        15 prof
```

Welches Ansehen geniessen die Statistiker? Geben Sie den Bereich an, in welchem das Ansehen mit 90% Wahrscheinlichkeit liegen wird. Ist das Ansehen höher als jenes der Mediziner (physicians)?

Und wie ist das Ansehen im Vergleich zu jenem von Buchhaltern (accountants)?

```
predict(fit.prestige2, newdata = x0, interval = "prediction", level=0.9)
```

```
       fit      lwr      upr
1 79.34128 66.37433 92.30823
```

```
Prestige[c("physicians", "accountants"),]
```

```
            education income women prestige census type log_Income
physicians      15.96  25308 10.56     87.2   3111 prof  10.138876
accountants     12.77   9271 15.70     63.4   1171 prof   9.134647
```

Die Mediziner sind mit einem Ansehen von 87.2 Punkten innerhalb des 90%-Prognoseintervall der Statistiker und die Buchhalter mit einem Ansehen von 63.4 Punkten liegen ausserhalb des 90%-Prognoseintervalls. Daher gilt nach dem Modell: das Ansehen der Statistiker ist höher als jenes der Buchhalter, jedoch gleich hoch wie jenes der Mediziner.