Aufgabe 3:

Betrachten Sie den Datensatz „wage2“, dieser enthält Daten zum logarithmierten Lohn, zum gemessenen IQ und zu den gearbeiteten Stunden. Betrachten Sie den folgenden R-Output.

```
> summary(lm(log(wage)~IQ,data=wage2))

Call:
lm(formula = log(wage) ~ IQ, data = wage2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.09324 -0.25547  0.02261  0.27544  1.21486

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.8869943  0.0890206   66.13   <2e-16 ***
IQ          0.0088072  0.0008694   10.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3999 on 933 degrees of freedom
Multiple R-squared:  0.09909,   Adjusted R-squared:  0.09813
F-statistic: 102.6 on 1 and 933 DF,  p-value: < 2.2e-16
```

a) Interpretieren Sie die Eingabe. Was wurde hier gemacht?

b) Interpretieren Sie die geschätzten Koeffizienten.

c) Sind die Koeffizienten signifikant? Was bedeutet Signifikanz?

d) Was bedeuten die Angaben „Residual standard error“, „Multiple R-squared“, „Adjusted R-Squared“ und „F-statistic“? Berechnen Sie das SSR.

e) Lässt sich aus (c) bereits ein kausaler Zusammenhang zwischen dem Log-Lohn und der Intelligenz folgern?

f) Betrachten Sie den nächsten Output. Was wurde hier gemacht? Wieso hat das Vorteile, auch wenn Sie sich nur für den Zusammenhang zwischen dem Log-Lohn und der Intelligenz interessieren?

g) Warum kann man jetzt annehmen, dass bereits im ersten Modell unverzerrt geschätzt wurde? Aus welchen zwei Gründen könnte dies passiert sein? Welcher trifft eher zu?

h) Wie erklären Sie sich inhaltlich den signifikant negativen Regressionskoeffizienten von „hours“?

```
> summary(lm(log(wage)~IQ+hours,data=wage2))

Call:
lm(formula = log(wage) ~ IQ + hours, data = wage2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.06915 -0.24533  0.02365  0.27646  1.28406

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.0536074  0.1150416  52.621   <2e-16 ***
IQ           0.0089535  0.0008698  10.293   <2e-16 ***
hours       -0.0041302  0.0018124  -2.279   0.0229 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3991 on 932 degrees of freedom
Multiple R-squared:  0.1041,    Adjusted R-squared:  0.1022
F-statistic: 54.14 on 2 and 932 DF,  p-value: < 2.2e-16
```

Lösung:

f) Man bezieht die Anzahl an gearbeiteten Stunden in die erste Regression ein.
Man benutzt 'hours' als Maß für den Fleiß um den Effekt des omitted-variable bias entgegen zu wirken.

g) Da 0.0089535 sehr nah bei 0,0088072 liegt, könnte die erste Schätzung bereits unverzerrt sein (es sei denn es liegen Faktoren außer Intelligenz und Fleiß vor, die den Lohn beeinflussen).
Dies kann sein, wenn $\boldsymbol{\gamma} = \mathbf{0}$ oder $\mathbf{X}^T\mathbf{Z} = \mathbf{0}$.
Da aber hours signifikant zu 5% ist (also $\boldsymbol{\gamma}$ wohl nicht $\mathbf{0}$) wird $\mathbf{X}^T\mathbf{Z} = \mathbf{0}$, also IQ nicht mit Fleiß korreliert sein.

h) Weil Personen (zum Beispiel Theaterschauspieler*innen) die trotz gleichen IQs (im Vergleich zu zum Beispiel Professor*innen) weniger pro Stunde verdienen, gezwungen sein könnten mehr zu arbeiten, aber am Ende doch nicht an das höhere Monatsgehalt ran kommen. Also ergibt der negative Koeffizient Sinn.