Exercise 3

Tobias Raidl, 11717659

2023-10-30

Contents

1

a

```
df = read.csv("schooldata.csv")
train = df[1:55,]
test = df[56:70,]

model = lm(cbind(reading, mathematics,
↪  selfesteem)~education+occupation+visit+counseling+teacher, data=train)
```

b

Having the p-values for each response variable seperatly does not help us in identifying the most significant variables for our multivariate case. lm() does not fit a multivariate model.

```
summary(model)
```

```
## Response reading :
##
## Call:
```

```
## lm(formula = reading ~ education + occupation + visit + counseling +
##     teacher, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0530  -2.1761  -0.3764   1.9319  11.1642
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.04835    1.73593   1.180 0.243709
## education    0.19819    0.07821   2.534 0.014515 *
## occupation   3.85251    1.05364   3.656 0.000624 ***
## visit        0.15602    0.30223   0.516 0.608002
## counseling  -0.61218    0.29448  -2.079 0.042890 *
## teacher     -0.35742    0.29115  -1.228 0.225461
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.752 on 49 degrees of freedom
## Multiple R-squared:  0.8917, Adjusted R-squared:  0.8806
## F-statistic: 80.66 on 5 and 49 DF,  p-value: < 2.2e-16
##
##
## Response mathematics :
##
## Call:
## lm(formula = mathematics ~ education + occupation + visit + counseling +
##     teacher, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.026  -2.932  -1.056   2.946  19.045
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.03920    2.20920   1.376 0.175170
## education    0.04791    0.09953   0.481 0.632398
## occupation   5.39390    1.34089   4.023 0.000199 ***
## visit       -0.12549    0.38462  -0.326 0.745609
## counseling  -0.47832    0.37477  -1.276 0.207866
## teacher     -0.43418    0.37053  -1.172 0.246945
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.048 on 49 degrees of freedom
## Multiple R-squared:  0.8657, Adjusted R-squared:  0.852
## F-statistic: 63.17 on 5 and 49 DF,  p-value: < 2.2e-16
##
##
## Response selfesteem :
##
## Call:
## lm(formula = selfesteem ~ education + occupation + visit + counseling +
##     teacher, data = train)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3731 -0.7791 -0.1575  0.9080  3.1324
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.09948    0.46471   0.214  0.83138
## education   -0.03501    0.02094  -1.672  0.10086
## occupation   2.20222    0.28206   7.808 3.76e-10 ***
## visit        0.27701    0.08091   3.424  0.00126 **
## counseling  -0.13581    0.07883  -1.723  0.09124 .
## teacher     -0.06543    0.07794  -0.840  0.40524
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.272 on 49 degrees of freedom
## Multiple R-squared:  0.9886, Adjusted R-squared:  0.9874
## F-statistic: 850.7 on 5 and 49 DF,  p-value: < 2.2e-16
```

c

With manova() we receive a multivariate model. The p values are expressive in comparison to the linear regression model before. Say critical vale $\alpha = 0.05$, the variables education, occupation and visit are significant.

```
multivariate_model = manova(cbind(reading, mathematics,
↪  selfesteem)~education+occupation+visit+counseling+teacher, data=train)
summary(multivariate_model)
```

```
##            Df  Pillai approx F num Df den Df    Pr(>F)
## education   1 0.97968   755.16      3     47 < 2.2e-16 ***
## occupation  1 0.97738   676.96      3     47 < 2.2e-16 ***
## visit       1 0.28992     6.40      3     47  0.001004 **
## counseling  1 0.10501     1.84      3     47  0.153155
## teacher     1 0.03230     0.52      3     47  0.668656
## Residuals  49
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2

The p value is too high to be significant in the case of $alpha = 0.05$. Therefore the model is fine without the two variables counseling and teacher.

```
reduced_multivariate_model = manova(cbind(reading, mathematics,
↪  selfesteem)~education+occupation+visit, data=train)
anova(multivariate_model, reduced_multivariate_model)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: cbind(reading, mathematics, selfesteem) ~ education + occupation +
##     visit + counseling + teacher
## Model 2: cbind(reading, mathematics, selfesteem) ~ education + occupation +
##     visit
##   Res.Df Df Gen.var.  Pillai approx F num Df den Df Pr(>F)
## 1     49      6.3311
## 2     51  2   6.3770 0.13341   1.1436      6     96 0.3432
```

#3 ## a This command conducts k-fold cross validation. The default k is 5, so the dataset is split into 5 folds. Each of these folds will be used as test set once. The error is averaged. This process is repeated 100 times, as we set R=100

```
library(cvTools)
```

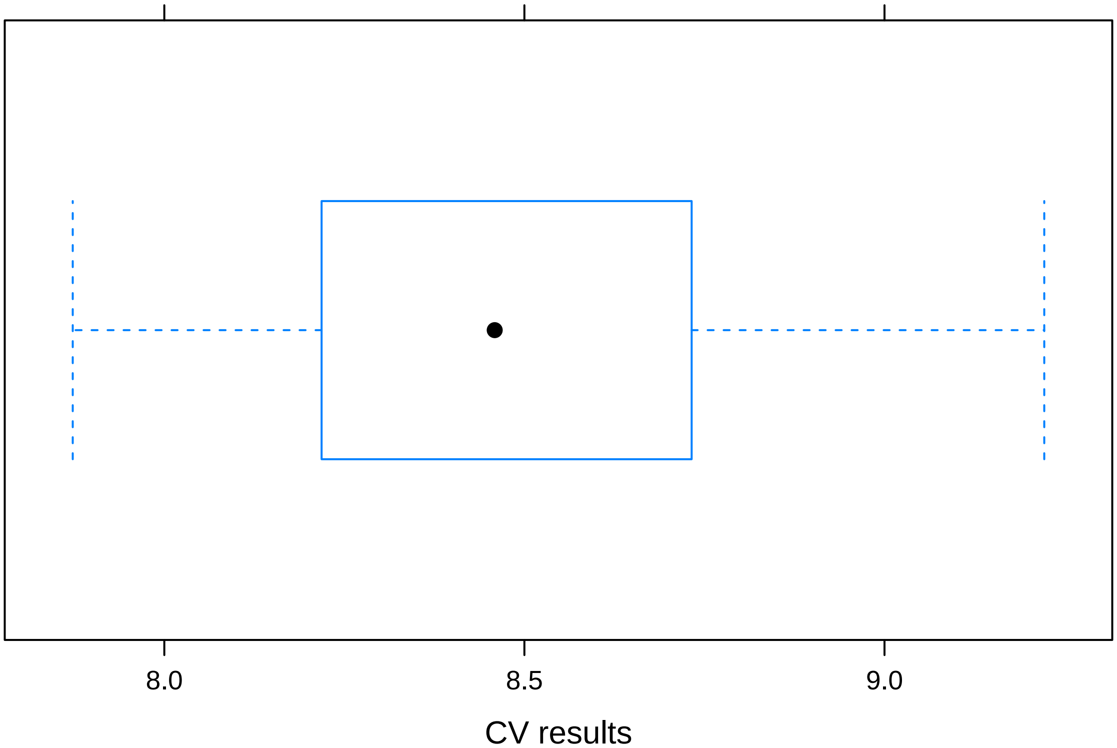
```
## Warning: Paket 'cvTools' wurde unter R Version 4.1.3 erstellt
```

```
## Lade nötiges Paket: lattice
```

```
## Lade nötiges Paket: robustbase
```

```
## Warning: Paket 'robustbase' wurde unter R Version 4.1.3 erstellt
```

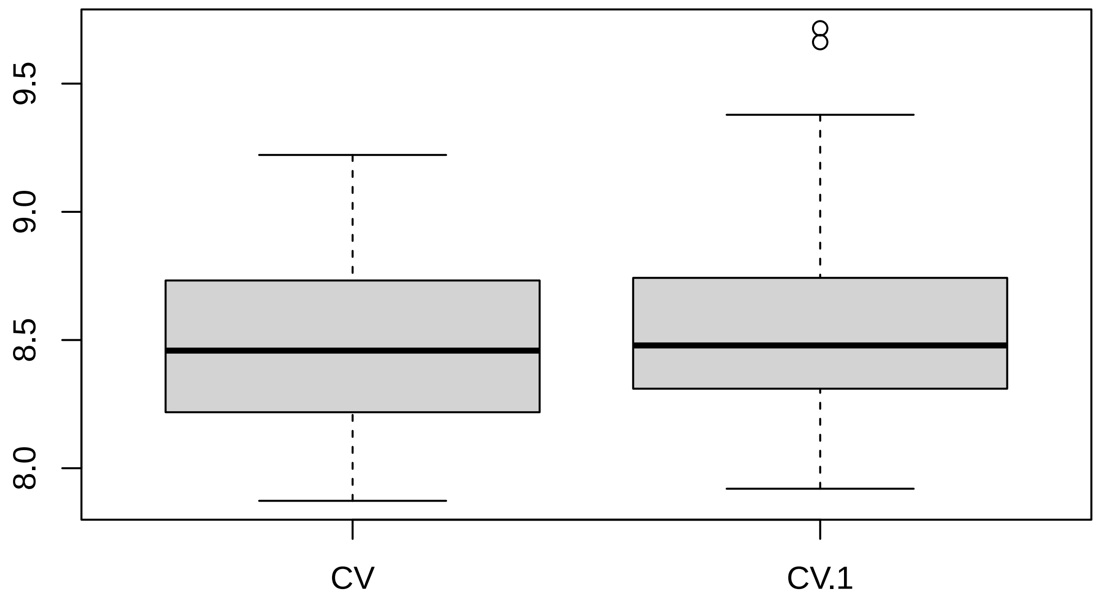
```
multivariate_model.cv =
↪  cvFit(multivariate_model,data=train,y=cbind(train$reading,train$mathematics,train$selfesteem),R=100)

plot(multivariate_model.cv)
```



b

The reduced multivariate model seems to perform slightly better. Also it is less complex due to a lower number of variables.

```
reduced_multivariate_model.cv =
↪  cvFit(reduced_multivariate_model,data=train,y=cbind(train$reading,train$mathematics,train$selfesteem

data = data.frame(A = unlist(multivariate_model.cv$reps),
                  B = unlist(reduced_multivariate_model.cv$reps))
boxplot(data)
```



4

Here I plot the residuals to compare the predicted to the ground truth values for each response variable. I conclude that selfesteem is the variable that is most accuratlly predicted using this reduced multivariate model. All in all though, all three response variables are predicted “well”.

```r
library(dplyr)
```

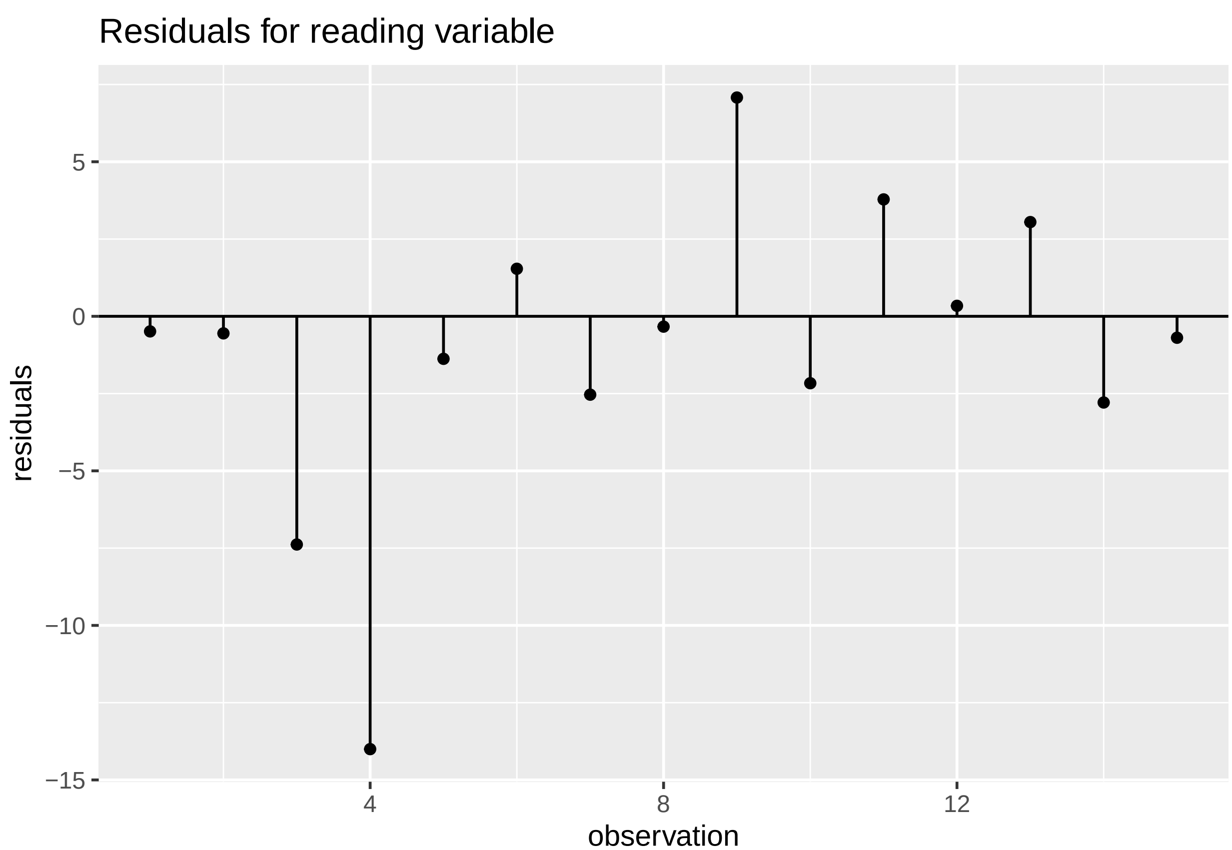
```
## Warning: Paket 'dplyr' wurde unter R Version 4.1.3 erstellt
```

```
##
## Attache Paket: 'dplyr'
```

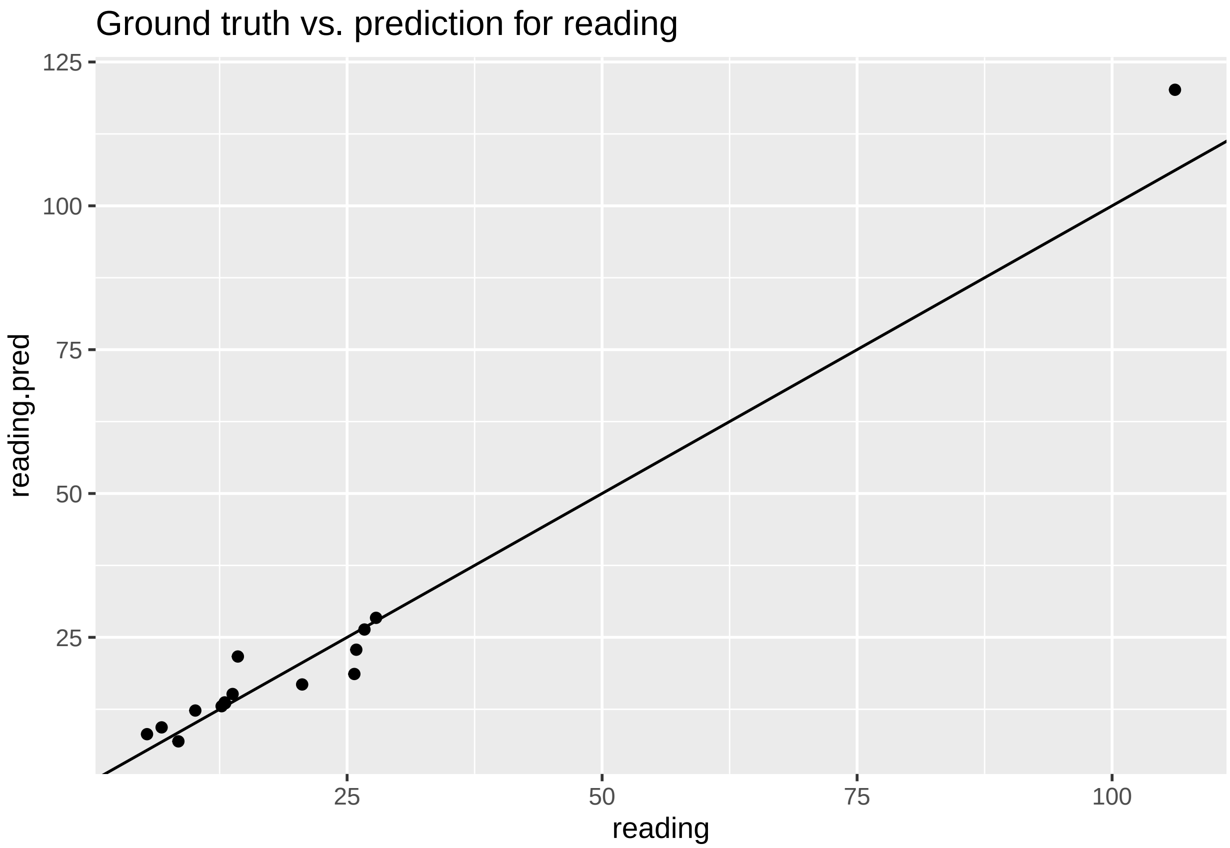
```
## Die folgenden Objekte sind maskiert von 'package:stats':
##
##     filter, lag
```

```
## Die folgenden Objekte sind maskiert von 'package:base':
##
##     intersect, setdiff, setequal, union
```

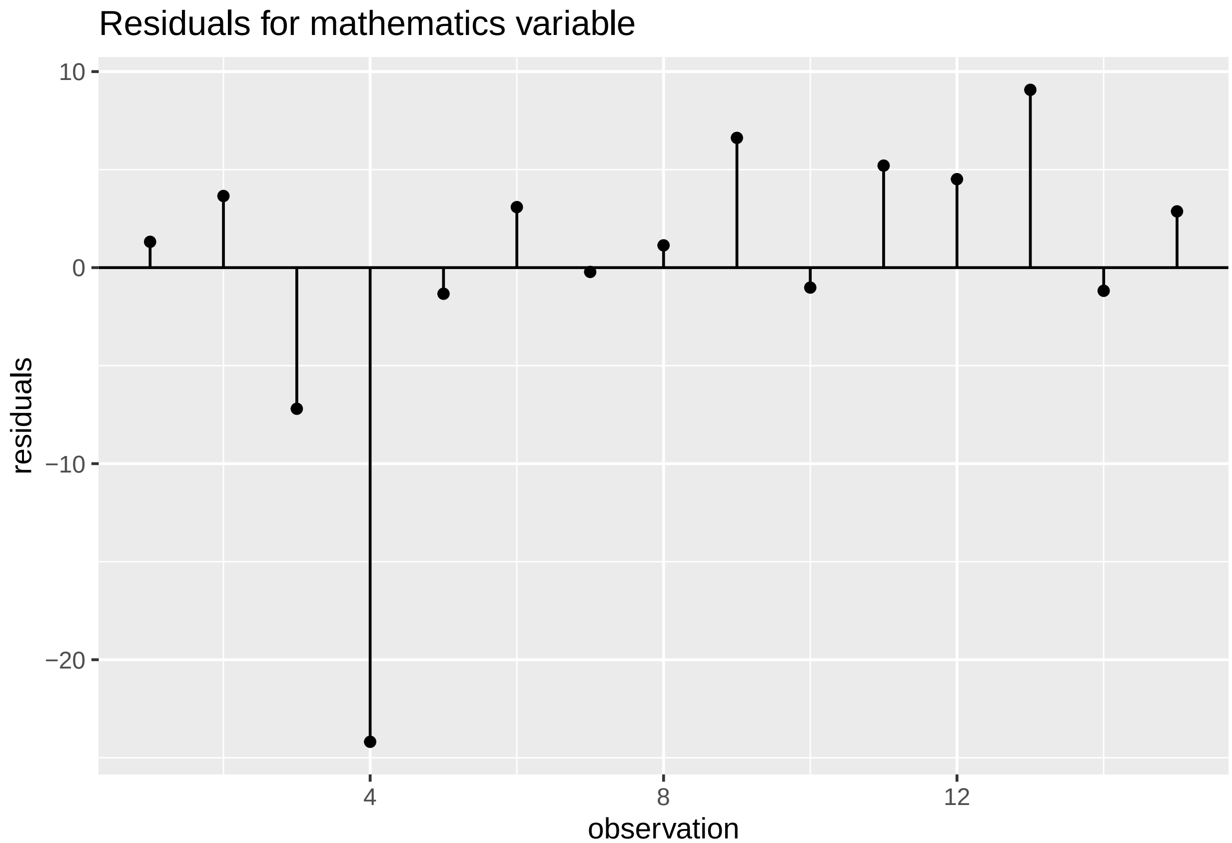
```r
predicted = data.frame(predict(reduced_multivariate_model, select(test, education,
↪   occupation, visit)))
gt = select(test, reading, mathematics, selfesteem)
data = data.frame(cbind(predicted, gt))
data = data %>%
  rename(
       reading.pred=reading.1,
       mathematics.pred=mathematics.1,
       selfesteem.pred=selfesteem.1)

library(ggplot2)
ggplot(data, aes(x=1:nrow(data), y=reading-reading.pred)) +
  geom_point() +
  geom_segment(aes(xend=1:nrow(data)), yend=0) +
  expand_limits(y=0) +
  geom_hline(yintercept=0) +
  ggtitle("Residuals for reading variable") +
  xlab("observation") + ylab("residuals")
```



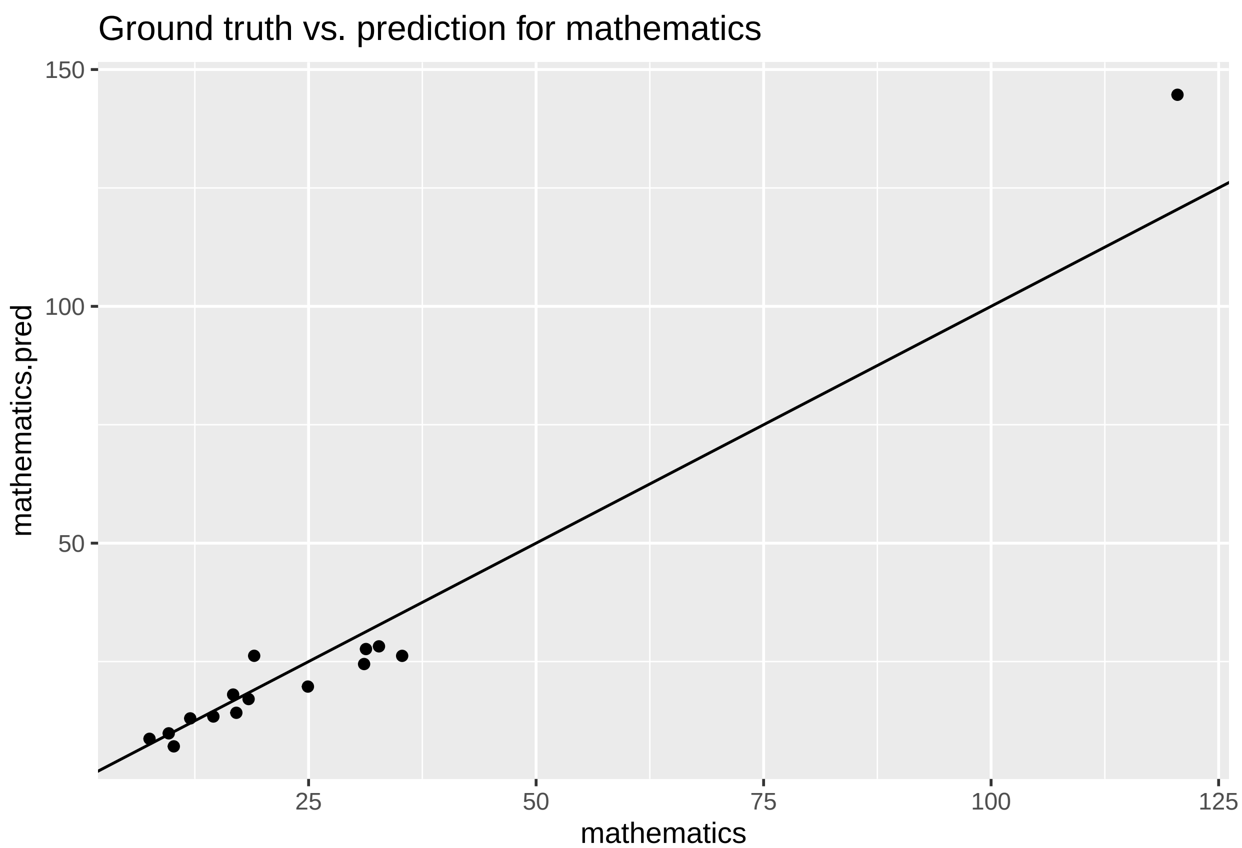

```r
ggplot(data, aes(x=reading, y=reading.pred)) +
  geom_point() +
  ggtitle("Ground truth vs. prediction for reading") +
  geom_abline(intercept=0, slope=1)
```



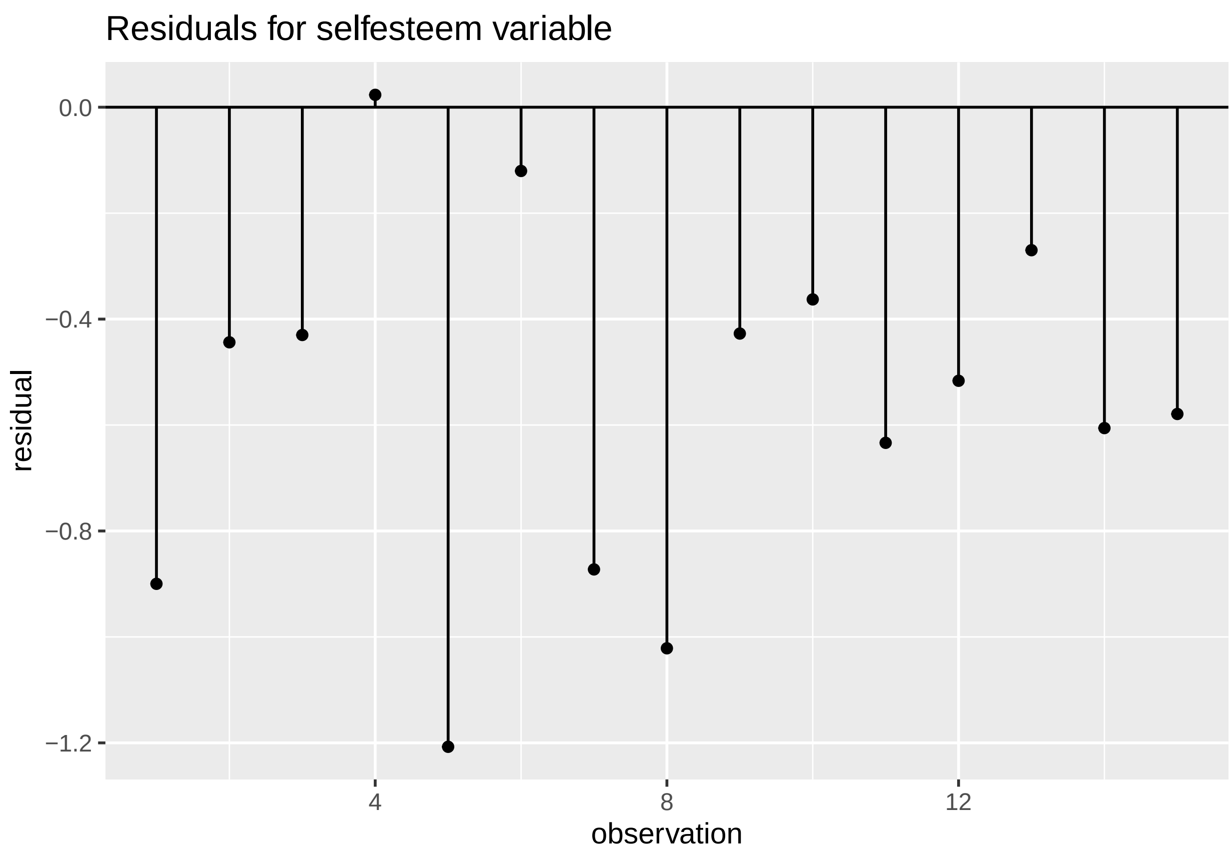

```r
ggplot(data, aes(x=1:nrow(data), y=mathematics-mathematics.pred)) +
  geom_point() +
  geom_segment(aes(xend=1:nrow(data)), yend=0) +
  expand_limits(y=0) +
  geom_hline(yintercept=0) +
  ggtitle("Residuals for mathematics variable") +
  xlab("observation") + ylab("residuals")
```



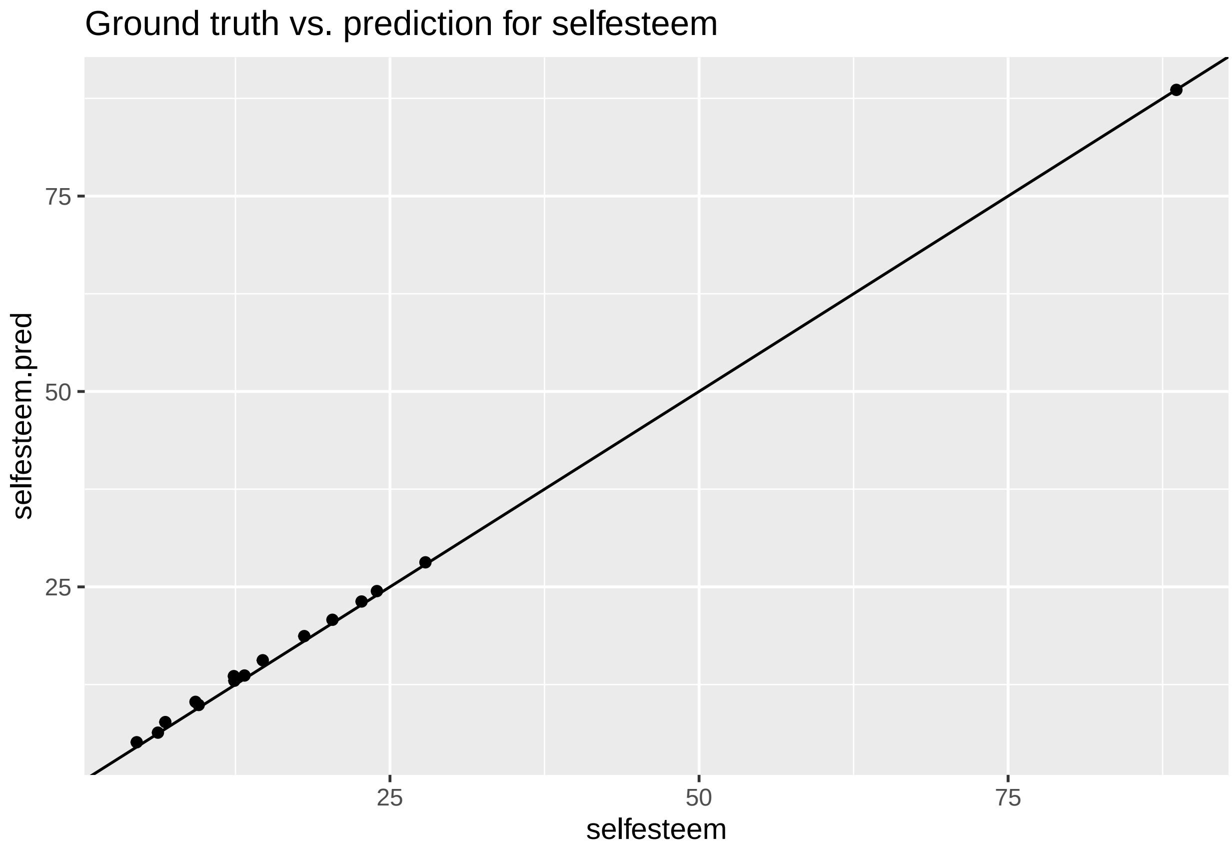

```
ggplot(data, aes(x=mathematics, y=mathematics.pred)) +
  geom_point() +
  ggtitle("Ground truth vs. prediction for mathematics") +
  geom_abline(intercept=0, slope=1)
```




```
ggplot(data, aes(x=1:nrow(data), y=selfesteem-selfesteem.pred)) +
  geom_point() +
  geom_segment(aes(xend=1:nrow(data)), yend=0) +
  expand_limits(y=0) +
  geom_hline(yintercept=0) +
  ggtitle("Residuals for selfesteem variable") +
  xlab("observation") + ylab("residual")
```




```r
ggplot(data, aes(x=selfesteem, y=selfesteem.pred)) +
  geom_point() +
  ggtitle("Ground truth vs. prediction for selfesteem") +
  geom_abline(intercept=0, slope=1)
```


Ground truth vs. prediction for selfesteem
75
50
25
selfesteem.pred
25
50
75
selfesteem