Paper 2 (Possible Gender Differences) Appendix

Casey Lee

2025-02-16

Overview

Paper 2 appendix provides an analysis of the possible gender differences in overconfidence. We will be creating two regression models (E(Y | intel_theory, attn_to) and E(Y | intel_theory, attn_to, gender):
1. See if the gender variable fits the first model and plot its fitted values.
2. Create plots to see if there is an unexplained variation potentially explained by gender.
3. Calculate in-sample loss for the versions of the first and second models to assess whether the relationship of overplacement to intelligence theory and experimental condition differed for men and women.

We are using the merged file combining the holdout and the original sample to reinforce a paper model with and without a contribution from gender. Then we create two regression models: We find the in-sample loss calculations based on cross-validation and do a hypothesis test only for the holdout sample.

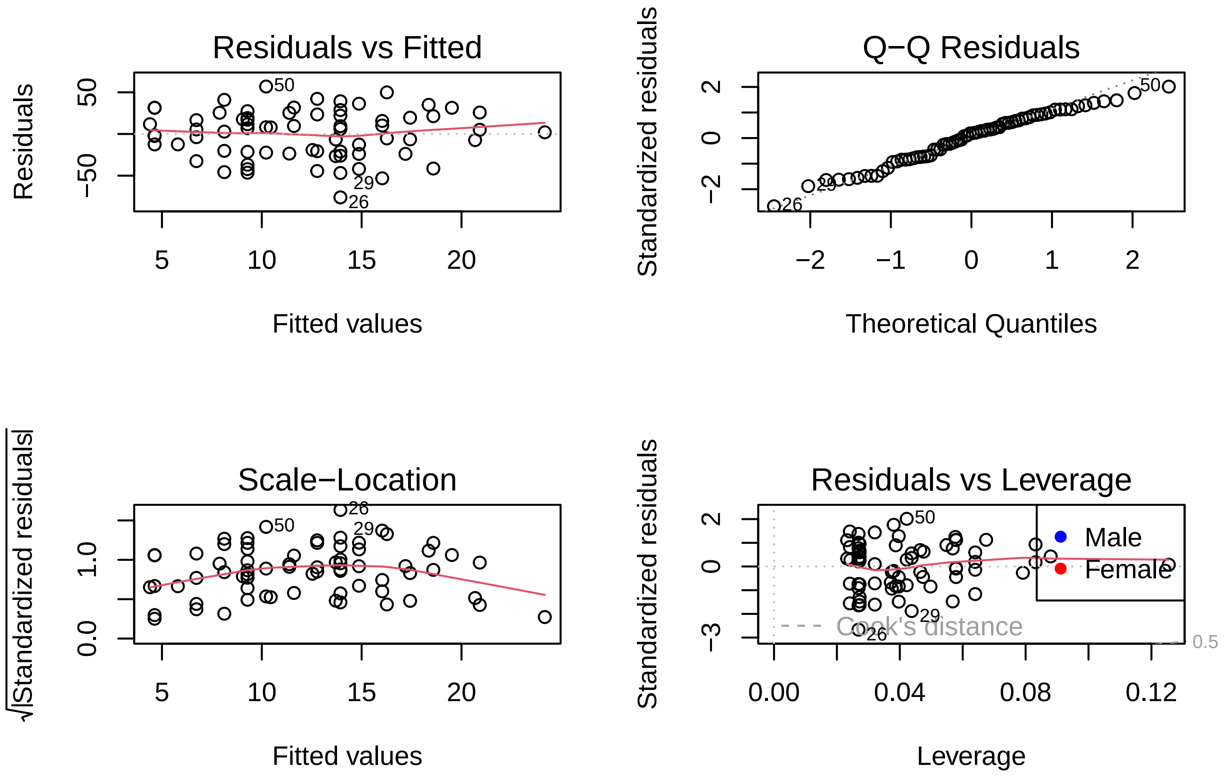
```r
# read data
no_gender <- read.csv("http://dept.stat.lsa.umich.edu/~bbh/s485/data/emdstudy3-small-nogender.csv")
gender <- read.csv("http://dept.stat.lsa.umich.edu/~bbh/s485/data/emdstudy3-small.csv")

# renaming columns
data_no_gender <- c("intel_theory", "attn_to", "ActPerc", "EstPerc") # this is for data w/o gender
names(no_gender) <- data_no_gender

data_col <- c("intel_theory", "gender", "attn_to", "ActPerc", "EstPerc") # this is for data w/ gender
names(gender) <- data_col

# changing all categorical variables in both datasets to factors
no_gender$attn_to <- as.factor(no_gender$attn_to)
gender$attn_to <- as.factor(gender$attn_to)

# fitted model w/o gender
mod0 <- lm(EstPerc - ActPerc ~ ., data = no_gender)
summary(mod0)
```

```
##
## Call:
## lm(formula = EstPerc - ActPerc ~ ., data = no_gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.077 -21.338   5.225  21.850  56.925
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.579     13.105  -0.197    0.845
## intel_theory        4.652      3.559   1.307    0.196
## attn_tohardprobs   -2.093      7.128  -0.294    0.770
##
## Residual standard error: 28.92 on 67 degrees of freedom
## Multiple R-squared:  0.02531,    Adjusted R-squared:  -0.003784
## F-statistic: 0.8699 on 2 and 67 DF,  p-value: 0.4237
```

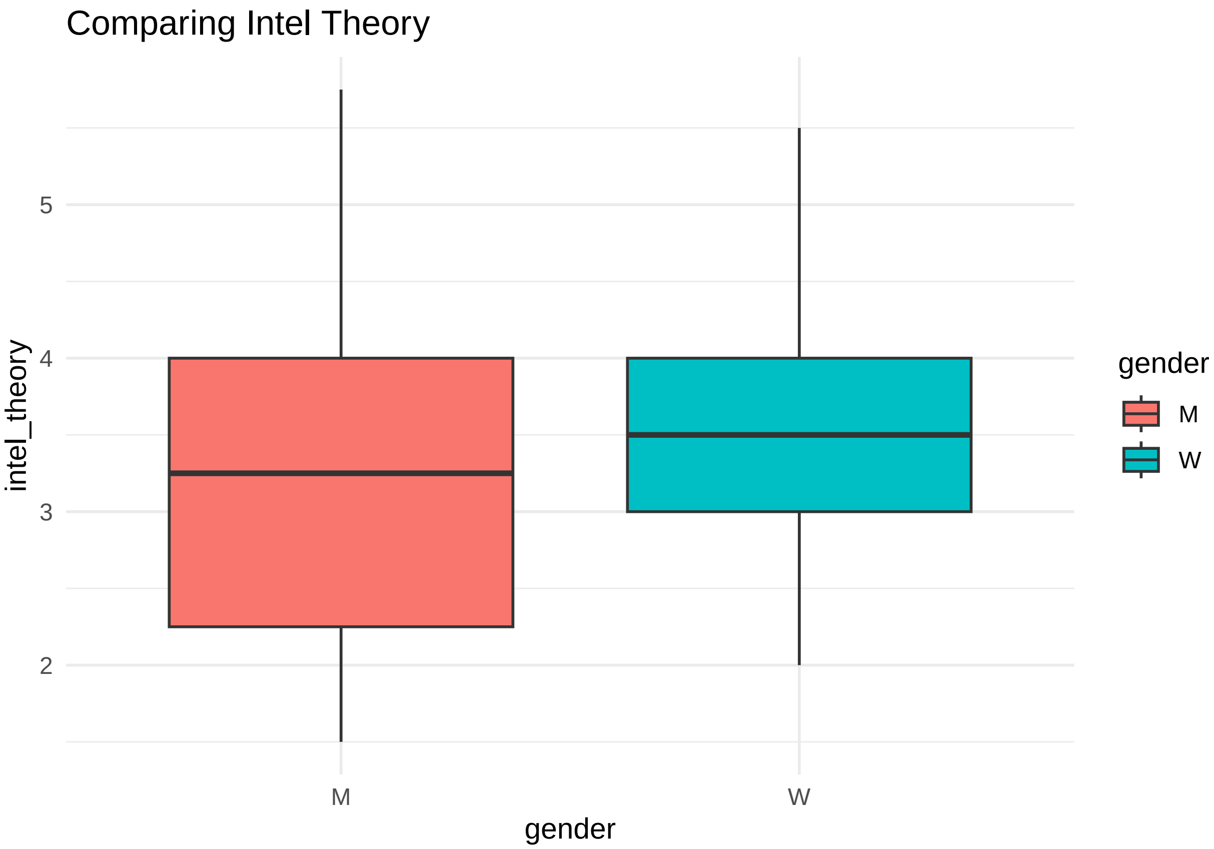
```
mean(mod0$residuals^2)
```

```
## [1] 800.2951
```

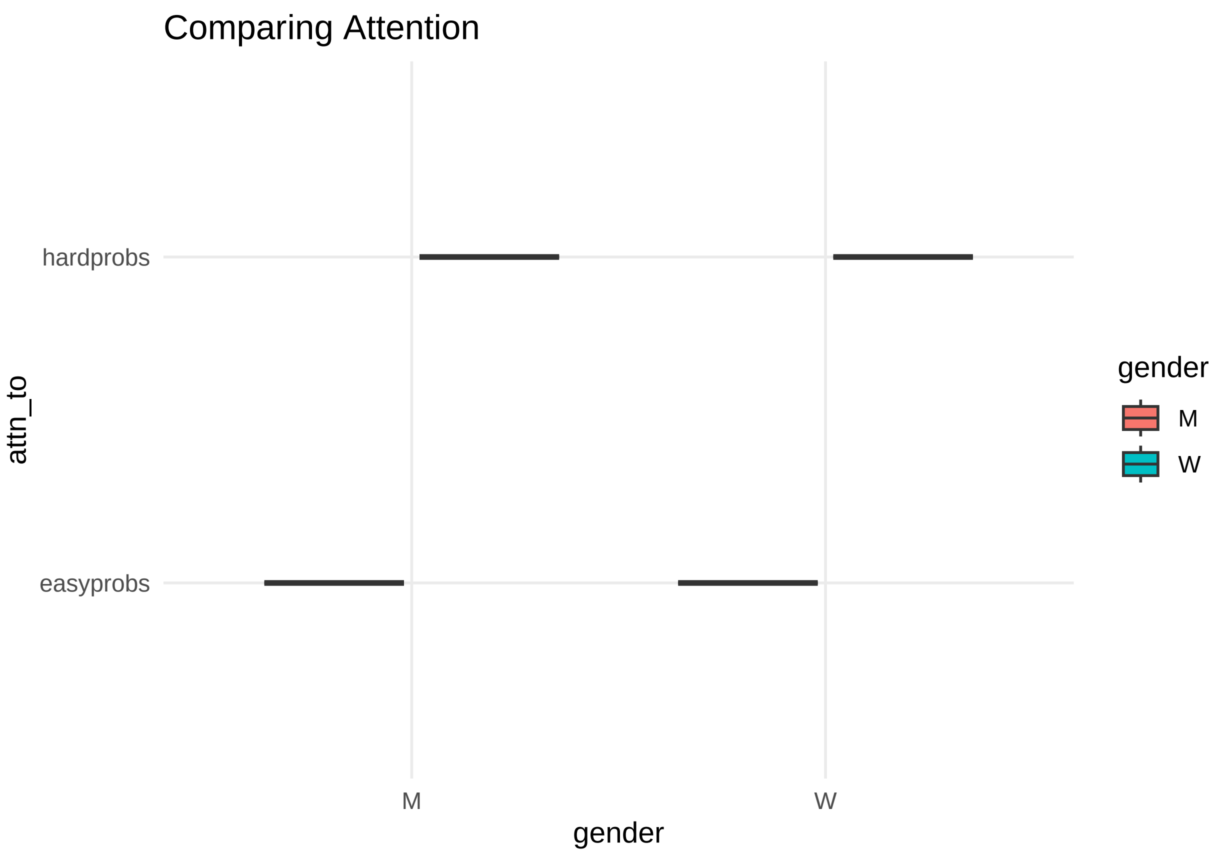
```
par(mfrow=c(2,2)) # checking residuals: Normality, Homoscedasticity, Influential Points
plot(mod0)
points(fitted(mod0), resid(mod0), col = c("blue", "red"), pch = 16)
legend("topright", legend = c("Male", "Female"), col = c("blue", "red"), pch = 16)
```



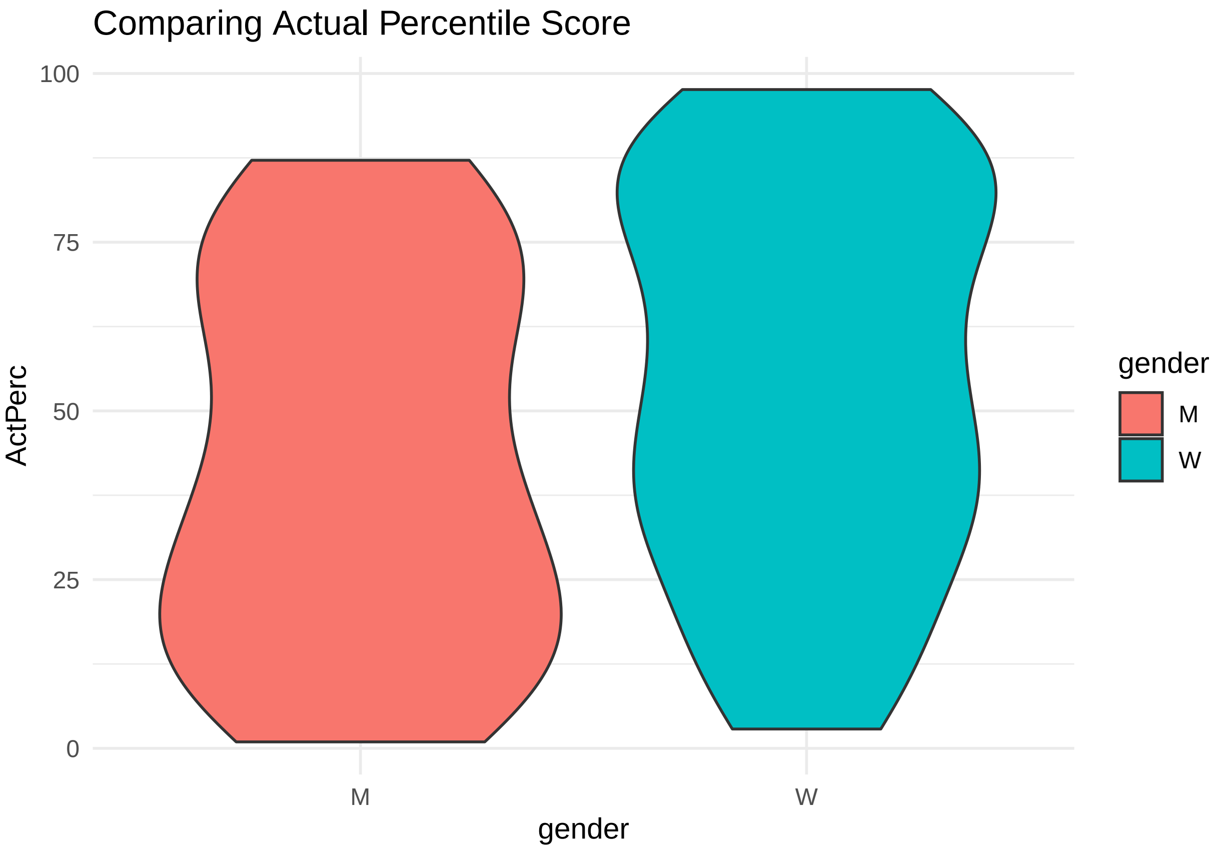
```
# combine genders by independent variables
ggplot(gender, aes(x = gender, y = intel_theory, fill = gender)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Comparing Intel Theory")
```



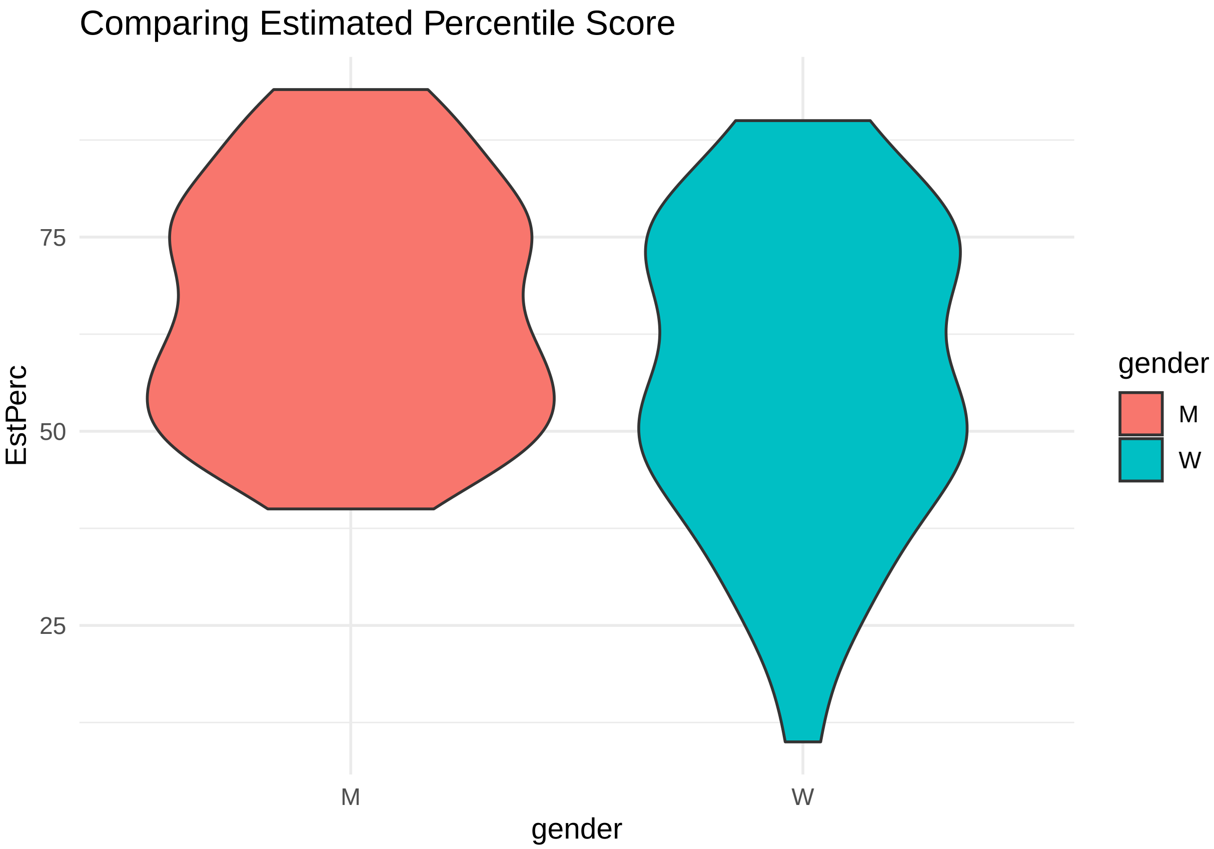
```r
ggplot(gender, aes(x = gender, y = attn_to, fill = gender)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title="Comparing Attention")
```



```r
ggplot(gender, aes(x = gender, y = ActPerc, fill = gender)) +
  geom_violin() +
  theme_minimal() +
  labs(title="Comparing Actual Percentile Score")
```



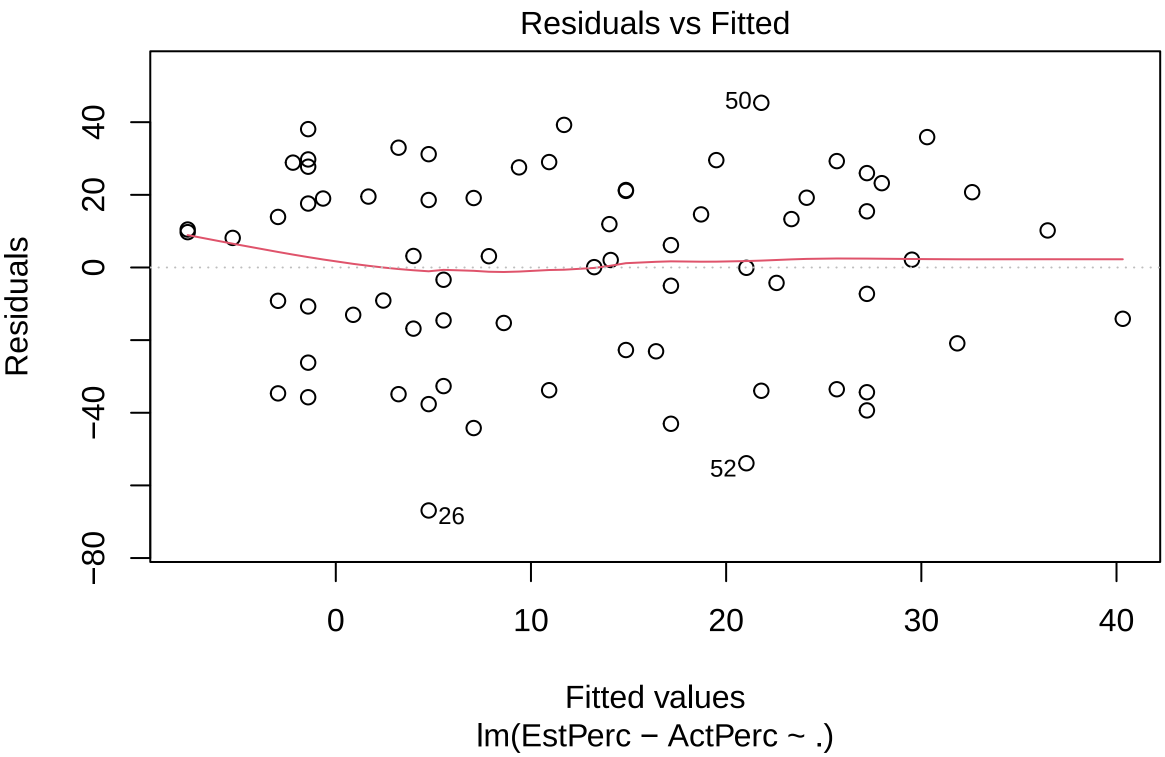
```r
ggplot(gender, aes(x = gender, y = EstPerc, fill = gender)) +
  geom_violin() +
  theme_minimal() +
  labs(title="Comparing Estimated Percentile Score")
```

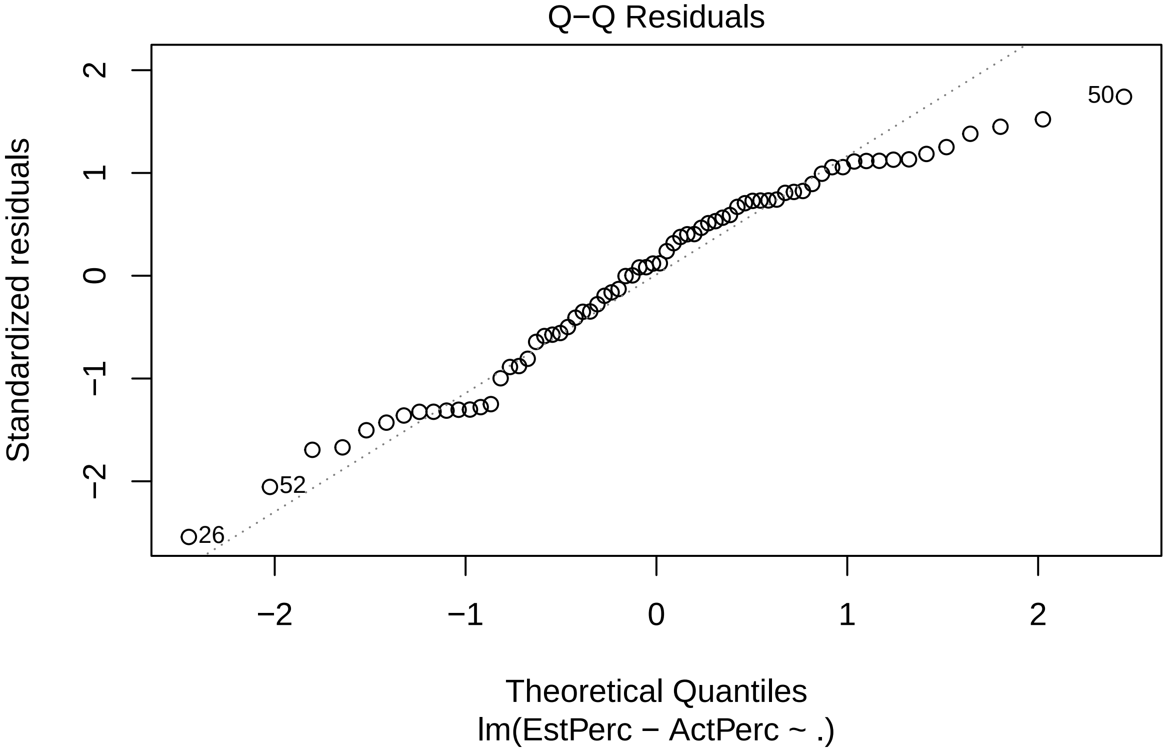


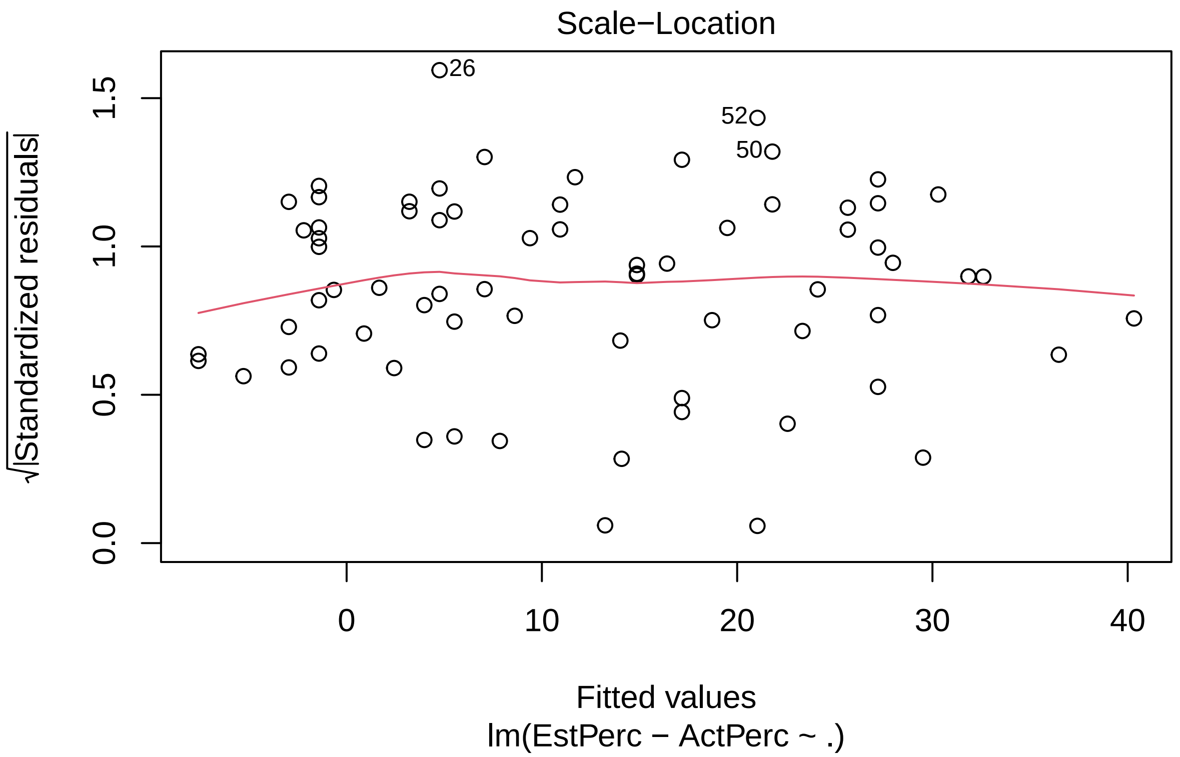
```r
# model w/ gender
mod1<- lm(EstPerc - ActPerc ~ ., data = gender)
summary(mod1)
```

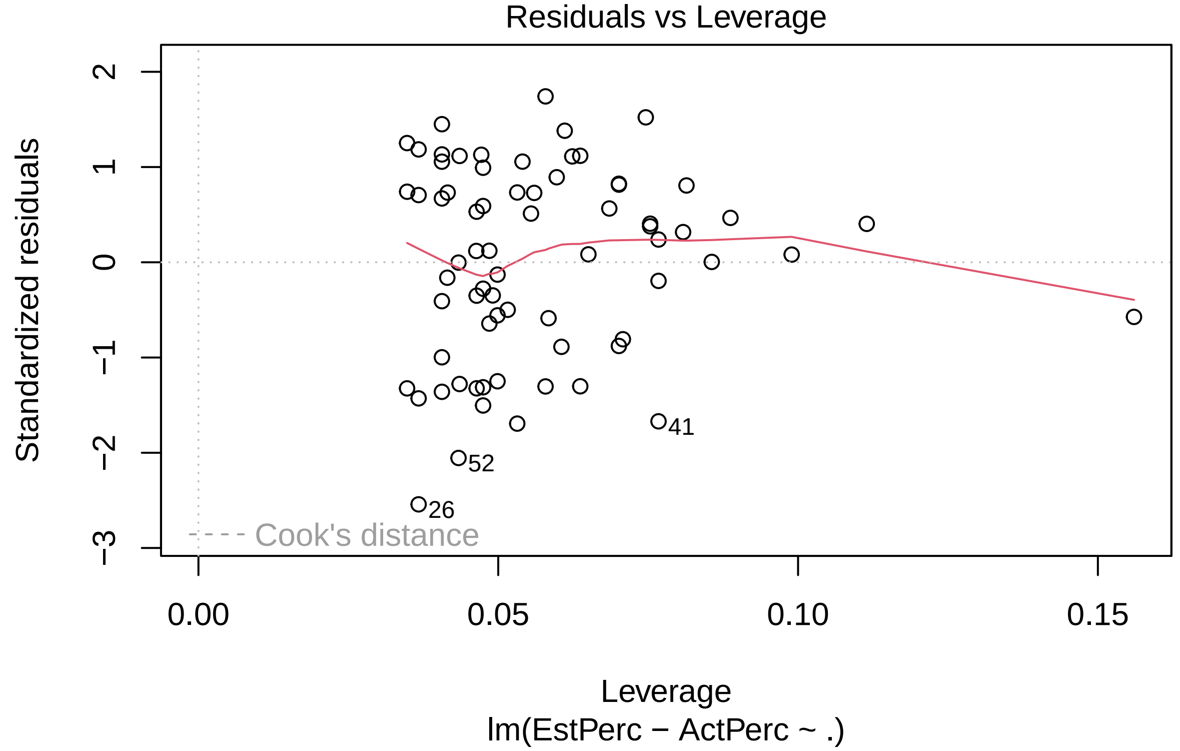
```
##
## Call:
## lm(formula = EstPerc - ActPerc ~ ., data = gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.898 -19.878   3.133  20.424  45.338
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.825     12.342   0.391 0.697104
## intel_theory         6.173      3.330   1.854 0.068271 .
## genderW            -22.453      6.510  -3.449 0.000985 ***
## attn_tohardprobs    -2.307      6.611  -0.349 0.728302
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.82 on 66 degrees of freedom
## Multiple R-squared:  0.1742, Adjusted R-squared:  0.1366
## F-statistic: 4.639 on 3 and 66 DF,  p-value: 0.005277
```

```r
plot(mod1)
```


Residuals vs Fitted
Residuals
−80
−40
0
20
40
0
10
20
30
40
Fitted values
lm(EstPerc − ActPerc ~ .)
50
52
26


Q−Q Residuals
Standardized residuals
−2
−1
0
1
2
−2
−1
0
1
2
Theoretical Quantiles
lm(EstPerc − ActPerc ~ .)
50
52
26





```
mean(mod1$residuals^2)
```

```
## [1] 678.0849
```

```
# compare the 2 models by checking the MSE & in-sample loss
mse_mod0 <- mean(mod0$residuals^2)
mse_mod1 <- mean(mod1$residuals^2)
```

```
MSE_Diff <- mse_mod0 - mse_mod1
MSE_Diff
```

```
## [1] 122.2102
```

```
r_squared <- summary(mod1)$r.squared - summary(mod0)$r.squared
r_squared
```

```
## [1] 0.1488413
```

```
# F test for both models
anova(mod0, mod1)
```

```
## Analysis of Variance Table
##
## Model 1: EstPerc - ActPerc ~ intel_theory + attn_to
## Model 2: EstPerc - ActPerc ~ intel_theory + gender + attn_to
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     67 56021
## 2     66 47466  1    8554.7 11.895 0.0009851 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

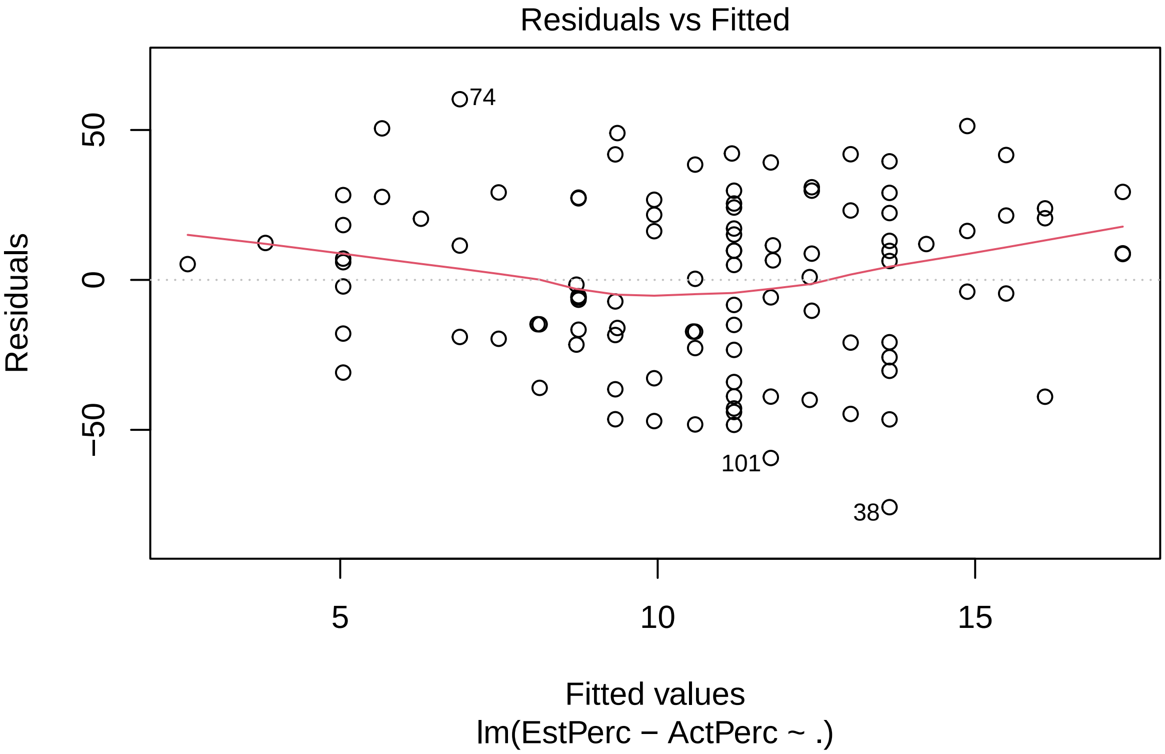
Additional calculations for paper version 2

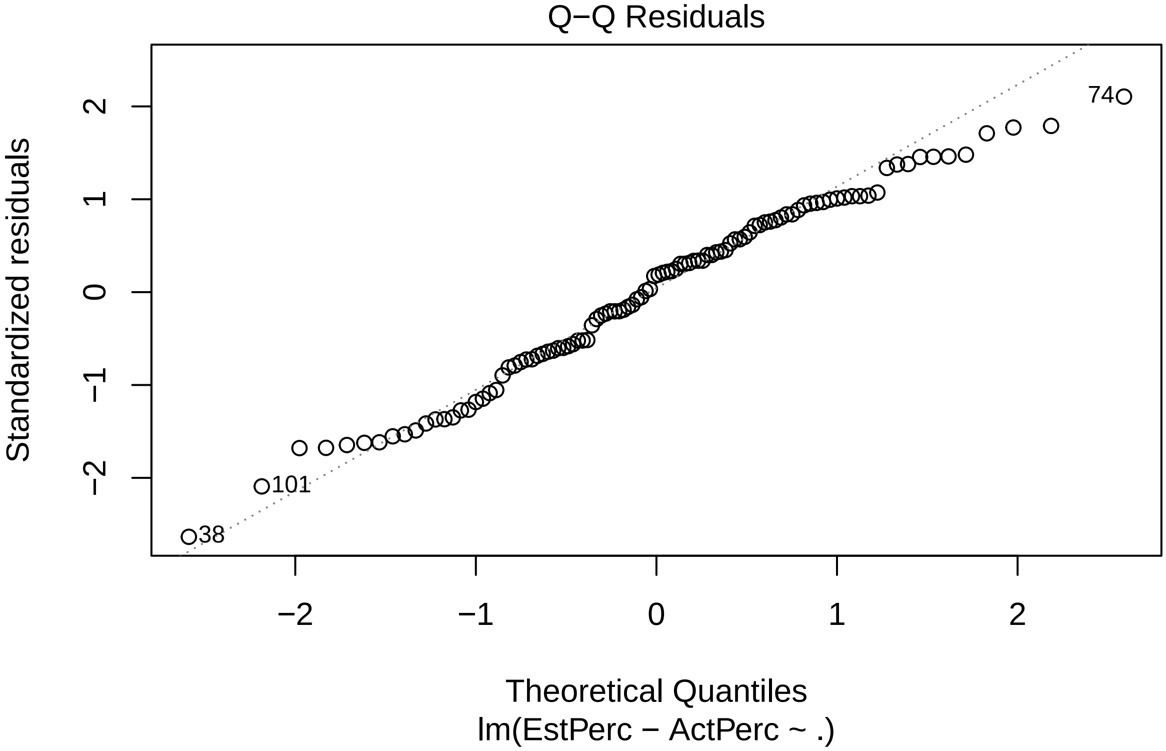
```
holdout_gender <- read.csv(file = "http://dept.stat.lsa.umich.edu/~bbh/s485/data/emdstudy3-all.csv", ski

names(holdout_gender) <- data_col

holdout_no_gender <- read.csv(file = "http://dept.stat.lsa.umich.edu/~bbh/s485/data/emdstudy3-all.csv",

holdout_no_gender <- holdout_no_gender[,-2]

names(holdout_no_gender) <- data_no_gender
```

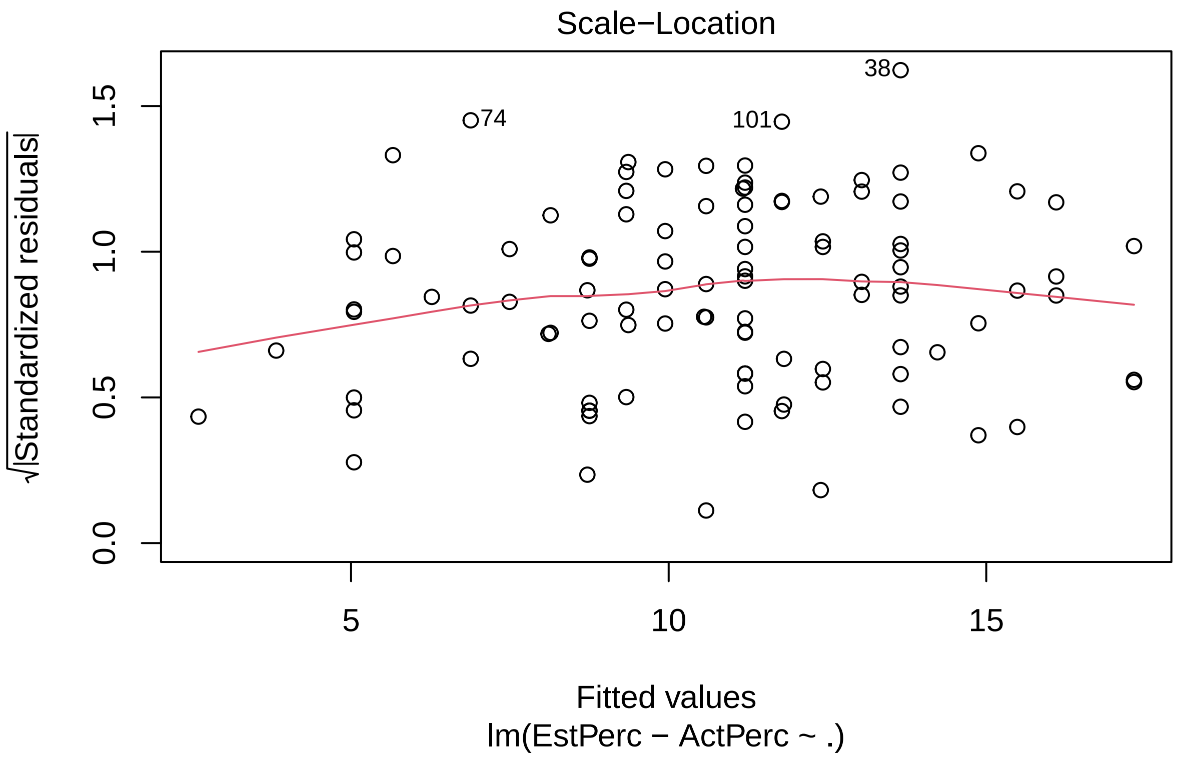
```
m1 <- lm(EstPerc - ActPerc ~ ., data = holdout_no_gender)
summary(m1)
```

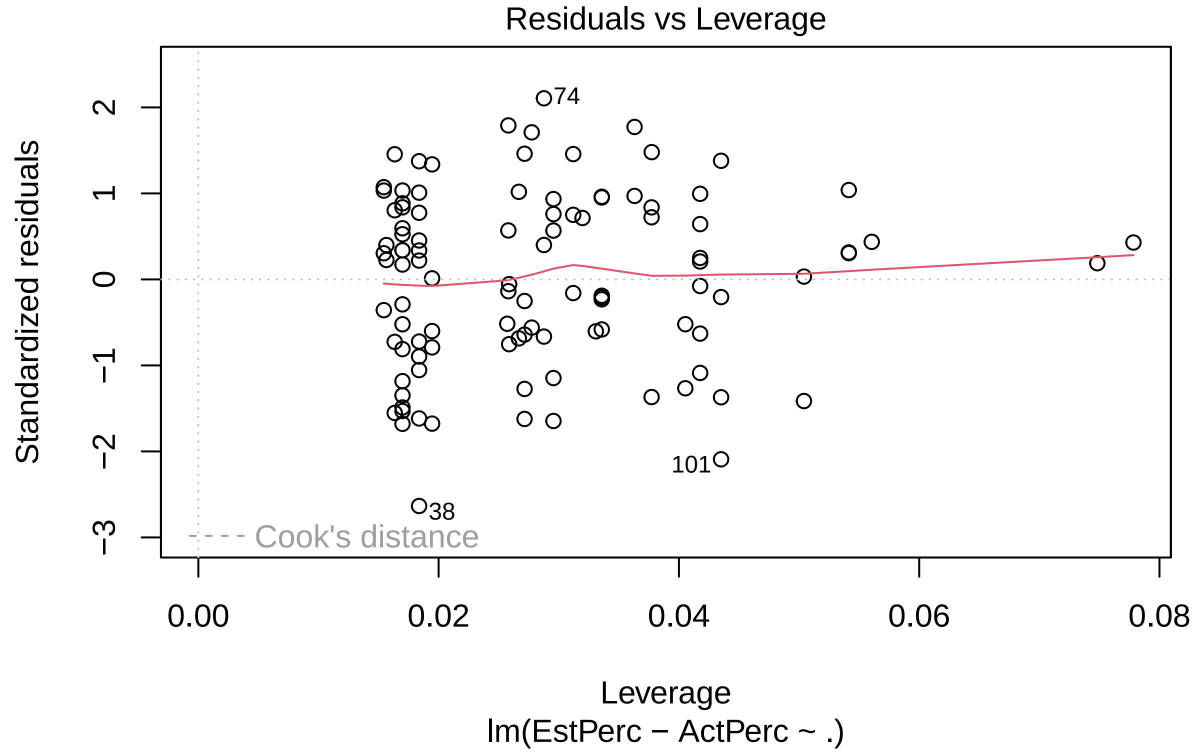
```
##
## Call:
## lm(formula = EstPerc - ActPerc ~ ., data = holdout_no_gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.791 -19.925   5.125  22.512  60.257
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.1482    10.3010   0.014    0.989
## intel_theory       2.4491     2.7523   0.890    0.376
## attn_tohardprobs   3.7067     5.8857   0.630    0.530
##
## Residual standard error: 29.04 on 101 degrees of freedom
## Multiple R-squared:  0.01207,    Adjusted R-squared:  -0.007493
## F-statistic: 0.617 on 2 and 101 DF,  p-value: 0.5416
```

```r
plot(m1)
```


Residuals vs Fitted
Residuals
50
0
−50
74
101
38
5
10
15
Fitted values
lm(EstPerc − ActPerc ~ .)


Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
74
101
38
−2
−1
0
1
2
Theoretical Quantiles
lm(EstPerc − ActPerc ~ .)





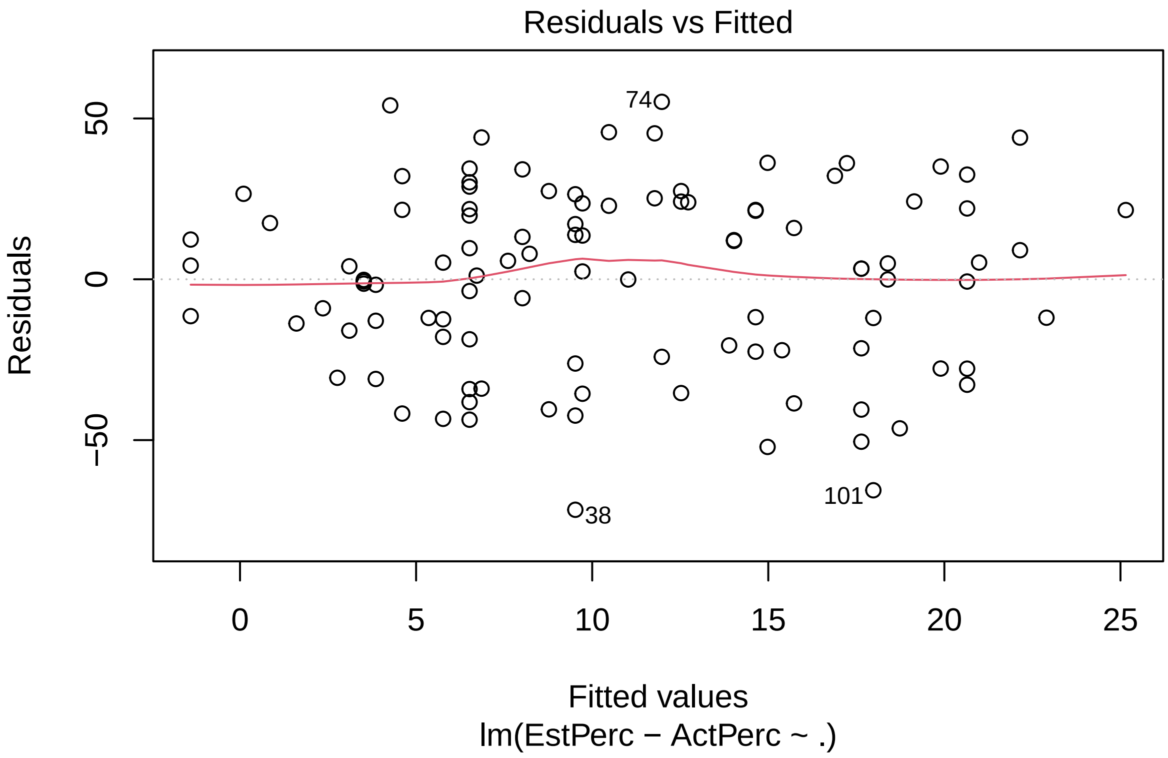
```
mean(m1$residuals^2)
```

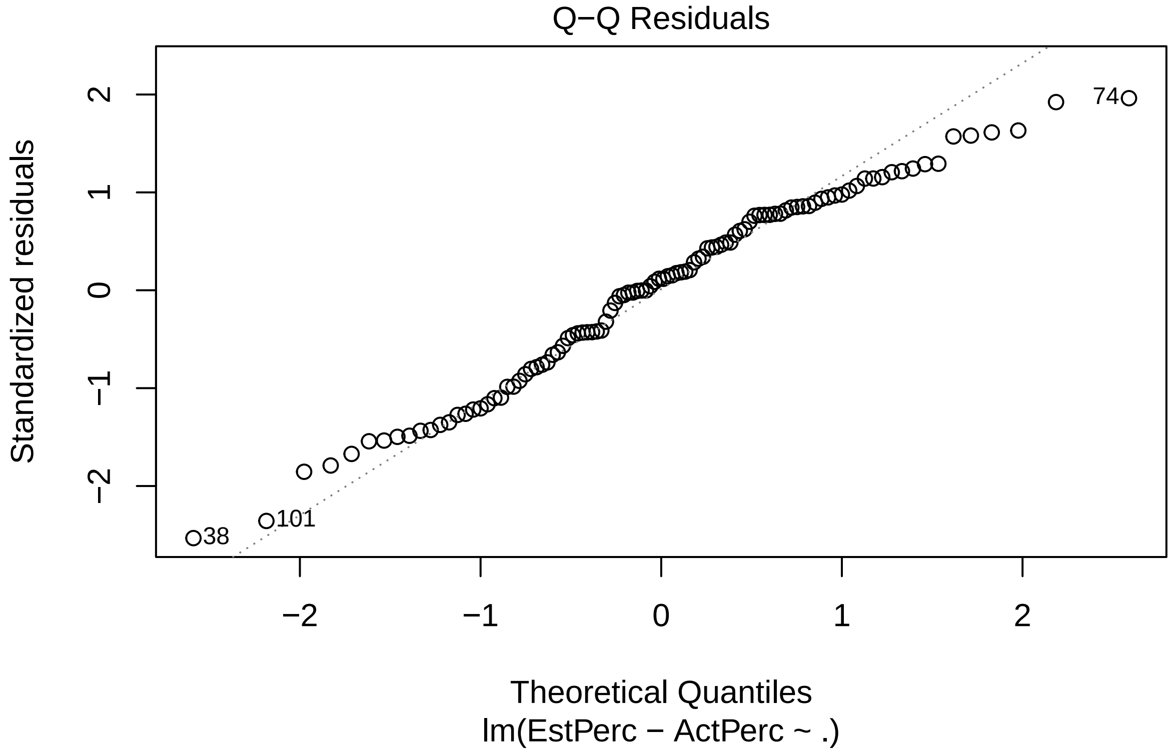
```
## [1] 818.7944
```

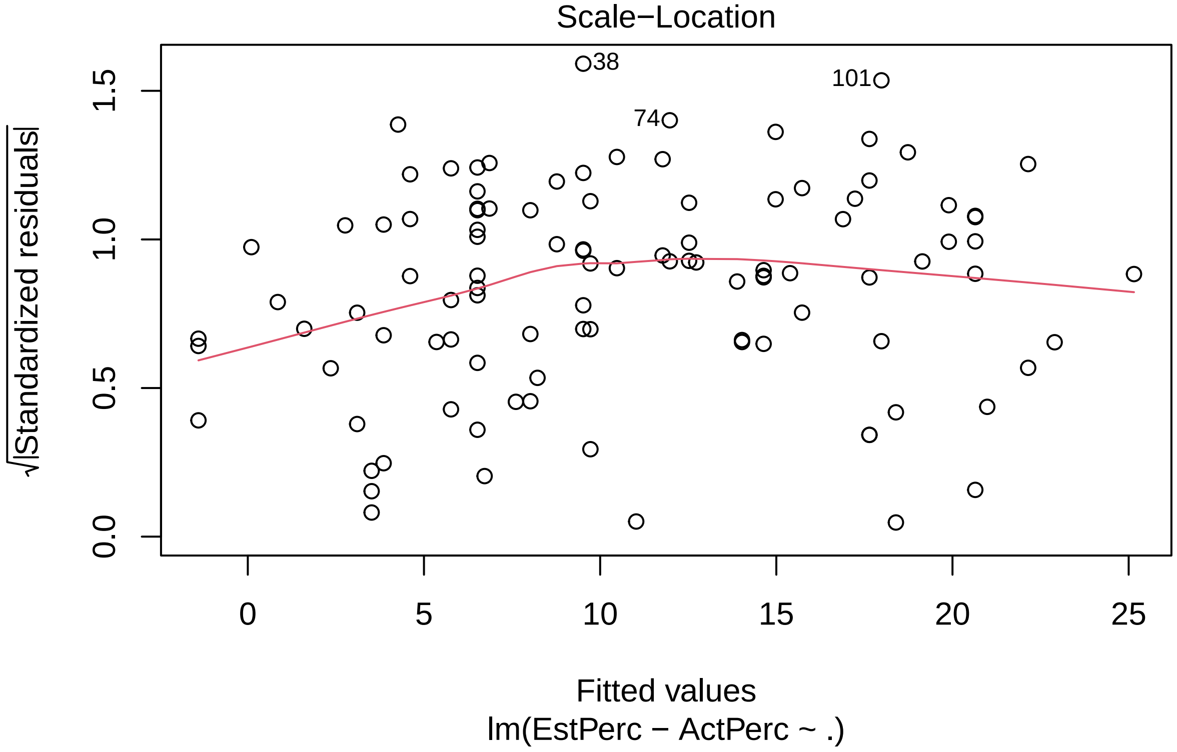
```
m2 <- lm(EstPerc - ActPerc ~ ., data = holdout_gender)
summary(m2)
```

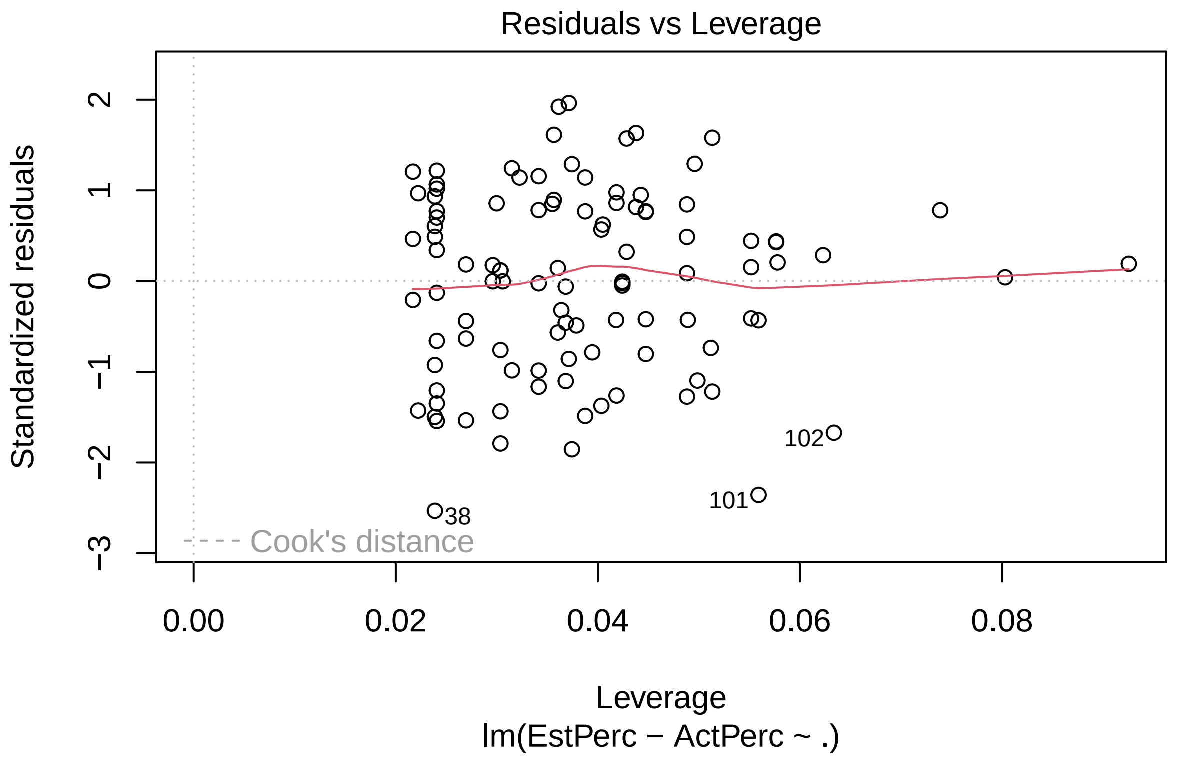
```
##
## Call:
## lm(formula = EstPerc - ActPerc ~ ., data = holdout_gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.661 -21.605   3.307  22.232  55.164
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.717     10.326   0.360   0.7196
## intel_theory        3.003      2.730   1.100   0.2739
## genderW           -11.125      5.721  -1.945   0.0546 .
## attn_tohardprobs    4.916      5.840   0.842   0.4019
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.64 on 100 degrees of freedom
## Multiple R-squared:  0.04806,    Adjusted R-squared:  0.01951
## F-statistic: 1.683 on 3 and 100 DF,  p-value: 0.1755
```

```r
plot(m2)
```




Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical Quantiles
lm(EstPerc − ActPerc ~ .)
74
38
101


Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
0
5
10
15
20
25
Fitted values
lm(EstPerc − ActPerc ~ .)
38
101
74



```r
mean(m2$residuals^2)
```

```
## [1] 788.9624
```

```r
MSE_Diff = mean(m1$residuals^2) - mean(m2$residuals^2)
MSE_Diff
```

```
## [1] 29.83194
```

```r
r_squared_diff = 0.04806 - 0.01207
r_squared_diff
```

```
## [1] 0.03599
```

```r
holdout_gender <- read.csv(file = "http://dept.stat.lsa.umich.edu/~bbh/s485/data/emdstudy3-all.csv",
                           skip = 1, header = FALSE, sep = ",")

# column names
data_col <- c("intel_theory", "gender", "attn_to", "ActPerc", "EstPerc")
colnames(holdout_gender) <- data_col

# w/o gender
holdout_no_gender <- holdout_gender[, !(colnames(holdout_gender) %in% "gender")]

# holdout data set only
holdout <- read.csv("http://dept.stat.lsa.umich.edu/~bbh/s485/data/emdstudy3-holdout.csv")

# Cross-validation
cv.lm <- function(data, formulae, nfolds = 5) {
  data <- na.omit(data)
  formulae <- sapply(formulae, as.formula)
  n <- nrow(data)
  fold.labels <- sample(rep(1:nfolds, length.out = n))
  mses <- matrix(NA, nrow = nfolds, ncol = length(formulae))
  colnames(mses) <- as.character(formulae)

  for (fold in 1:nfolds) {
    test.rows <- which(fold.labels == fold)
    train <- data[-test.rows, ]
    test <- data[test.rows, ]

    for (form in 1:length(formulae)) {
      current.model <- lm(formula = formulae[[form]], data = train)
      predictions <- predict(current.model, newdata = test)
      test.responses <- eval(formulae[[form]][[2]], envir = test)
      test.errors <- test.responses - predictions
      mses[fold, form] <- mean(test.errors^2)
    }
  }
  return(colMeans(mses))
}

# models
model_w_gender <- "EstPerc - ActPerc ~ gender + intel_theory + attn_to"
model_wo_gender <- "EstPerc - ActPerc ~ intel_theory + attn_to"

cv_results <- cv.lm(holdout_gender, c(model_w_gender, model_wo_gender))
print(cv_results)
```

```
## EstPerc - ActPerc ~ gender + intel_theory + attn_to
##                                            855.0002
##          EstPerc - ActPerc ~ intel_theory + attn_to
##                                            860.9368
```

```r
# model fitting
m1 <- lm(EstPerc - ActPerc ~ ., data = holdout_no_gender)  # w/o gender
m2 <- lm(EstPerc - ActPerc ~ ., data = holdout_gender)  # w/ gender

# MSE
mse_w_gender <- mean((holdout$EstPerc - predict(m2, holdout_gender))^2)
```

```
## Warning in holdout$EstPerc - predict(m2, holdout_gender): longer object length
## is not a multiple of shorter object length
```

```r
mse_wo_gender <- mean((holdout$EstPerc - predict(m1, holdout_no_gender))^2)
```

```
## Warning in holdout$EstPerc - predict(m1, holdout_no_gender): longer object
## length is not a multiple of shorter object length
```

```r
mse_w_gender
```

```
## [1] 2789.614
```

```r
mse_wo_gender
```

```
## [1] 2744.426
```

```r
# fitted models for holdout sample
holdout_mod0 <- lm(EstPerc - ActPerc ~ intel_theory + attn_to, data = holdout_no_gender)
```

```
holdout_mod1 <- lm(EstPerc - ActPerc ~ gender + intel_theory + attn_to, data = holdout_gender)

anova(holdout_mod0, holdout_mod1)
```

```
## Analysis of Variance Table
##
## Model 1: EstPerc - ActPerc ~ intel_theory + attn_to
## Model 2: EstPerc - ActPerc ~ gender + intel_theory + attn_to
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    101 85155
## 2    100 82052  1    3102.5 3.7812 0.05464 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```