Homework 6 - Statistical Data Analysis

Yigit Kasal

2025-06-02

Exercise 1

Load and prepare data

```
stress <- read_delim("StressSymptoms2.txt",
                     delim = "\t",
                     locale = locale(encoding = "UTF-8"),
                     trim_ws = TRUE)
```

```
## Rows: 107 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): participant, gender
## dbl (2): stress, symptoms
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
stress$gender_coded <- ifelse(stress$gender == "man", -0.5,
                              ifelse(stress$gender == "woman", 0.5, NA_real_))
```

Fit regression with interaction

```
mod <- lm(symptoms ~ stress * gender_coded, data = stress)
summary(mod)$coefficients
```

```
##                       Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)          71.0472276  3.2764192 21.684413 3.502579e-40
## stress                0.9116141  0.1315310  6.930794 3.743230e-10
## gender_coded        -16.7660695  6.5528384 -2.558597 1.196385e-02
## stress:gender_coded   0.8497816  0.2630619  3.230348 1.659997e-03
```

Interpretation:
Intercept = 71.047
Slope (stress) = 0.91
Slope (gender_coded) = -16.77
Interaction = 0.85

Manually add interaction term

```
stress$stressXgender <- stress$stress * stress$gender_coded

mod2 <- lm(symptoms ~ stress + gender_coded + stressXgender, data = stress)
summary(mod2)$coefficients
```

```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    71.0472276  3.2764192 21.684413 3.502579e-40
## stress          0.9116141  0.1315310  6.930794 3.743230e-10
## gender_coded  -16.7660695  6.5528384 -2.558597 1.196385e-02
## stressXgender   0.8497816  0.2630619  3.230348 1.659997e-03
```

Observation:
Manually creating the interaction term gives the same results as using *.

Exercise 2

Load and prepare EEG data

```
mydata <- read.csv("EEG_Indiv_RT_Dataset20180706out.csv", sep = ";")
mydata <- mydata[mydata$Accuracy == "1", ]
mydata <- mydata[!is.na(mydata$Latency), ]
mydata <- mydata[!is.na(mydata$AA), ]
mydata$AoA <- as.factor(as.character(mydata$AoA))
contrasts(mydata$AoA) <- c(-0.5, 0.5)
```

Check multicollinearity

```
ex2m <- lmer(Latency ~ AoA + freqfilms2 + (1 | PP.id) + (1 | Item), data = mydata, REML = FALSE)
summary(ex2m)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: Latency ~ AoA + freqfilms2 + (1 | PP.id) + (1 | Item)
##    Data: mydata
##
##      AIC      BIC   logLik deviance df.resid
##  74118.7  74158.4 -37053.4  74106.7     5471
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2244 -0.5110 -0.1862  0.2547 27.0418
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
```

```
##  Item     (Intercept)  3518     59.31
##  PP.id    (Intercept) 12118    110.08
##  Residual             42016    204.98
## Number of obs: 5477, groups:  Item, 99; PP.id, 20
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 817.11799   25.84731  23.62687  31.613  < 2e-16 ***
## AoA1         46.31541   13.19642  97.49809   3.510 0.000681 ***
## freqfilms2   -0.07829    0.24409  97.10986  -0.321 0.749089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##            (Intr) AoA1
## AoA1       -0.014
## freqfilms2 -0.168  0.076
```

```
lm_vif <- lm(Latency ~ AoA + freqfilms2, data = mydata)
car::vif(lm_vif)
```

```
##        AoA freqfilms2
##   1.005744   1.005744
```

Linear model with both predictors

```
lm_full <- lm(Latency ~ freqfilms2 + AoA, data = mydata)
summary(lm_full)$coefficients
```

```
##                 Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 811.07568242   3.876962 209.2039544 0.000000e+00
## freqfilms2   -0.05257349   0.119398  -0.4403213 6.597218e-01
## AoA1         43.30730449   6.473688   6.6897423 2.458454e-11
```

Likelihood ratio test

```
m_freq <- lmer(Latency ~ freqfilms2 + (1 | PP.id) + (1 | Item), data = mydata, REML = FALSE)
m_full <- lmer(Latency ~ freqfilms2 + AoA + (1 | PP.id) + (1 | Item), data = mydata, REML = FALSE)
anova(m_freq, m_full)
```

```
## Data: mydata
## Models:
## m_freq: Latency ~ freqfilms2 + (1 | PP.id) + (1 | Item)
## m_full: Latency ~ freqfilms2 + AoA + (1 | PP.id) + (1 | Item)
##        npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
## m_freq    5 74128 74161 -37059    74118
## m_full    6 74119 74158 -37053    74107 11.616  1  0.0006539 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion:
Even after controlling for frequency, AoA significantly affects naming latency.

Interaction effect

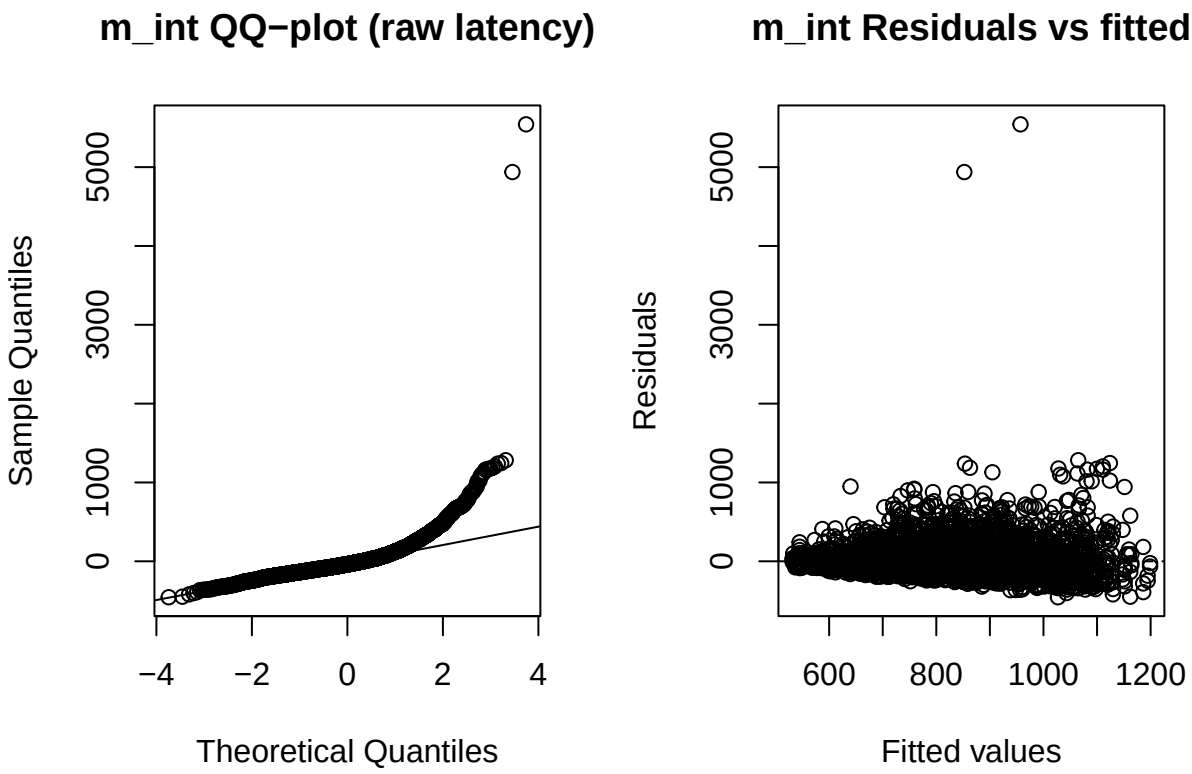
```
mydata$freq_c <- with(mydata, freqfilms2 - mean(freqfilms2, na.rm = TRUE))

m_int <- lmer(Latency ~ AoA * freq_c + (1 | PP.id) + (1 | Item), data = mydata, REML = FALSE)
summary(m_int)$coefficients
```

```
##                Estimate Std. Error       df    t value     Pr(>|t|)
## (Intercept) 815.6560727 25.4839297 22.35198 32.0066835 3.604073e-20
## AoA1         46.3166451 13.1953929 97.49491  3.5100618 6.799622e-04
## freq_c       -0.0754852  0.2455976 96.79888 -0.3073531 7.592349e-01
## AoA1:freq_c  -0.0503354  0.4912061 96.80708 -0.1024731 9.185932e-01
```

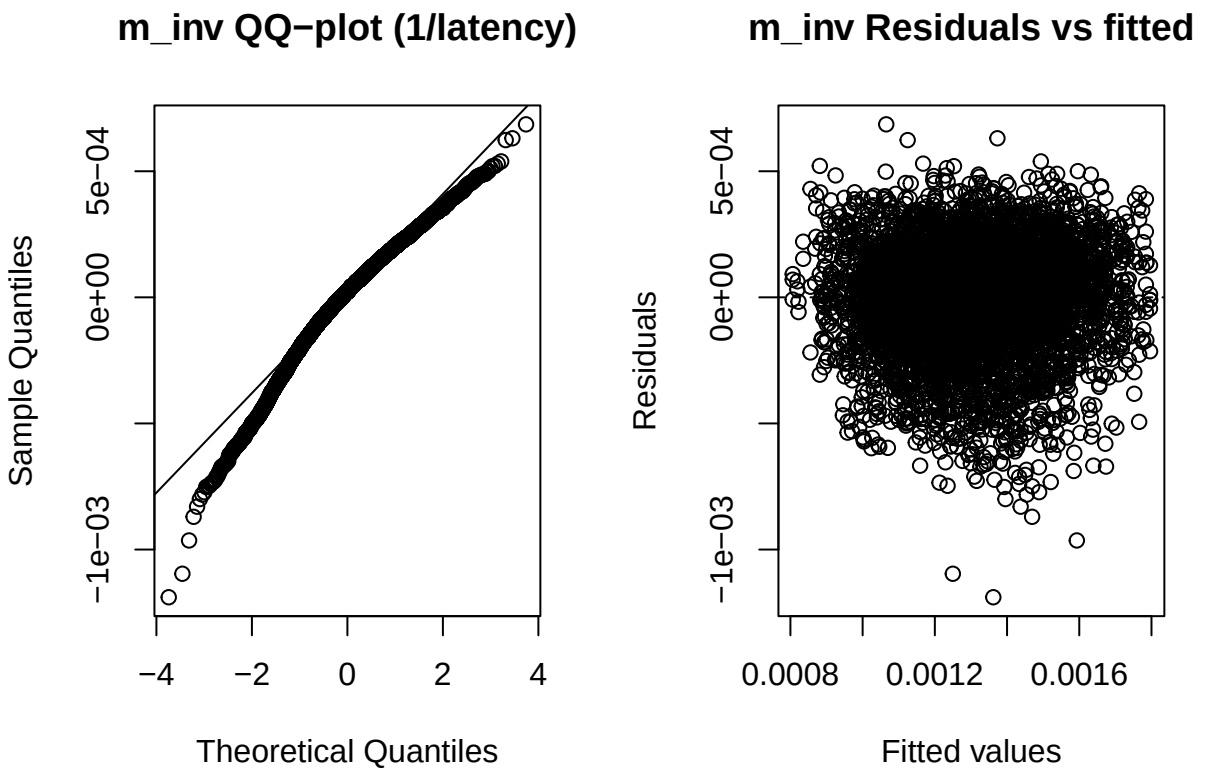
Conclusion:
No significant interaction between AoA and frequency. AoA has a consistent effect.

Assumption checks

```
mydata$invLatency <- 1 / mydata$Latency
m_inv <- lmer(invLatency ~ AoA * freq_c + (1 | PP.id) + (1 | Item), data = mydata, REML = FALSE)

par(mfrow = c(1, 2))
qqnorm(resid(m_int), main = "m_int QQ-plot (raw latency)")
qqline(resid(m_int))

plot(fitted(m_int), resid(m_int), xlab = "Fitted values", ylab = "Residuals",
     main = "m_int Residuals vs fitted")
abline(h = 0, lty = 2)
```




```
par(mfrow = c(1, 2))
qqnorm(resid(m_inv), main = "m_inv QQ-plot (1/latency)")
qqline(resid(m_inv))

plot(fitted(m_inv), resid(m_inv), xlab = "Fitted values", ylab = "Residuals",
     main = "m_inv Residuals vs fitted")
abline(h = 0, lty = 2)
```




```
par(mfrow = c(1, 1))
```

Interpretation:
Reciprocal transformation improves normality and homoscedasticity of residuals.