OSF Microsurgery Study: Final Report
COSC 6325, Spring 2018
Instructor: Dr. Pavlidis
T.A: George Panagopoulos

Group 10
Zhuo Ai
Michael Bremner
Greg Brumbaugh

1 Introduction

The following is a summary of the statistical findings of *Group 10* with regard to the OSF Microsurgery Study Dataset. Our analysis includes an overview of relevant data trends followed by a process of regression modeling to arrive at an appropriate model. Conclusions are drawn in both statistical and real-world terms.

Our primary motivation is to investigate the relationship between stress & learning ability in microsurgical tasks. We wish to answer the question: Does stress inhibit learning? Findings with regard to the learning vs. stress relationship are useful for assessing the viability of candidates for dangerous or high stress jobs which require great focus & attention to detail. such as a pilot or graduate C.S. student. We should be careful however; when generalizing beyond the scope of the experiment, which is learning in a microsurgery environment.

In addition we wish to discover / demonstrate additional significant aspects of the learning process that exist within the data. Did learning take place? What is the meaning of the learning effect?

As the study is primarily concerned with stress and learning, we must have solid quantitative analogies for those attributes. For stress we consider Perinasal Perspiration Signal data. More heat/sweat on the nose corresponds to higher levels of stress. We consider two attributes to assess learning. (1) Accuracy Score (2) Completion Time (scaled). We determine that learning has taken place when accuracy scores improve without a loss in speed, or if completion time improves without a loss in accuracy score.

In general, our analysis reflects mindfulness of two primary considerations:

(1)Does stress significantly impact learning?

(2)How is learning exhibited in the data?

2 Summary Plots

The summarization plots herein are crucial to the unfolding of our analysis. The linear models we develop are motivated by the patterns in the data.

2.1 Normality

It is crucial to check the continuous variables for normality. By understanding the normal behavior of the data, we are better able to understand the behavior and limitations of our models, where they perform well, and why they break down.

In general, for the data in this study, the Shapiro test for normality is not effective due to sensitivity to tail behavior and the tail behavior of our data. We are left to assess the normality of the data by qualitative inspection in both Raw, and Log transformed form. We may select the data which adheres more closely to and makes reliable errors relative to the normal distribution.

2.1.1 Perinasal Perspiration

The log data appears to be a marginal improvement to the raw signal. The irregular tail behavior is noted for the errors it may cause in our models.

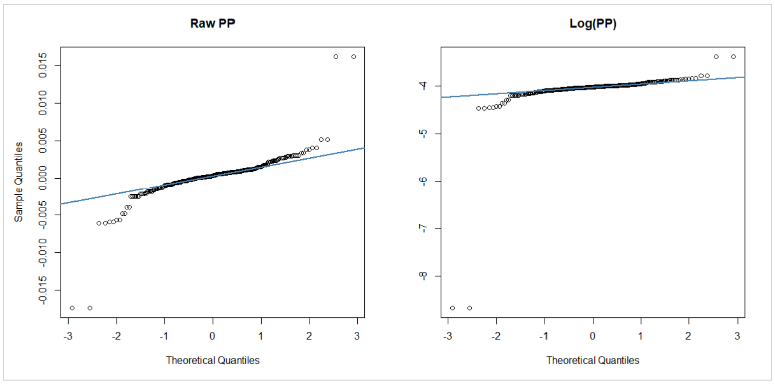


Figure 1: Qualitative analysis of signal PP for normality

2.1.2 Accuracy Score

The tail behavior is not as strong as the other data, but is still a problem. The log transform only makes matters worse. We conclude that the accuracy is sufficiently normal in its raw form.

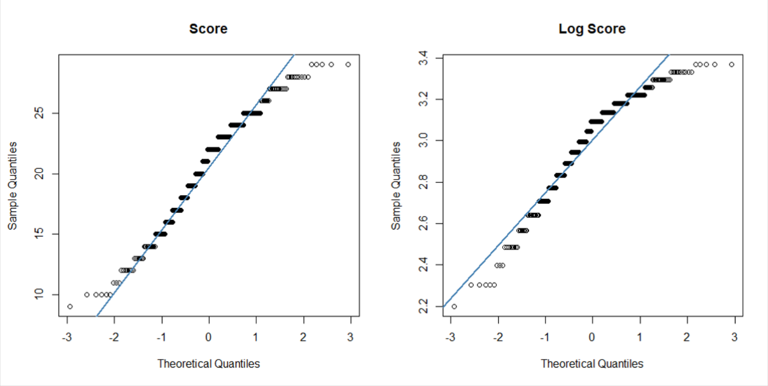


Figure 2: Qualitative analysis of response variable: Accuracy Score, for normality

2.1.3 Raw Completion Time

The raw completion time is highly problematic as a response variable. The distribution is highly not normal & the data is poorly mixed by task, with cut times being lower, and suture times higher. The log transform makes no improvement. We conclude that in order for time to be useful as a response variable, it must be transformed.

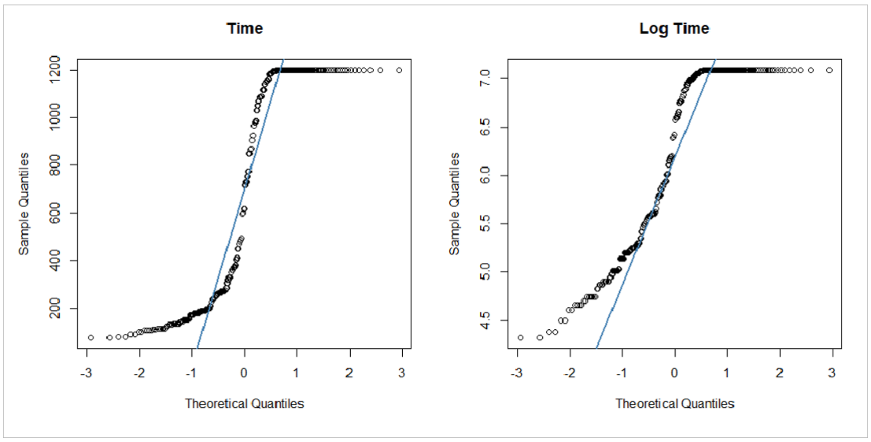


Figure 3: Qualitative analysis of response variable: Completion Time, for normality

2.1.4 Scaled Completion Time

A new response variable, "Scaled Time" is achieved by dividing the suture times by the number of sutures completed, allowing for an one-to-one comparison. Additionally the data are separately scaled to improve the comparison.

The QQ plots for scaled time show a large improvement over the previous time metric. The data is not perfect; however, when we apply the log transform, we have strong normal behavior within +- 1 sd. We conclude by inspection, that the new performance metric, Scaled Time, is sufficiently normal to apply to our linear models.

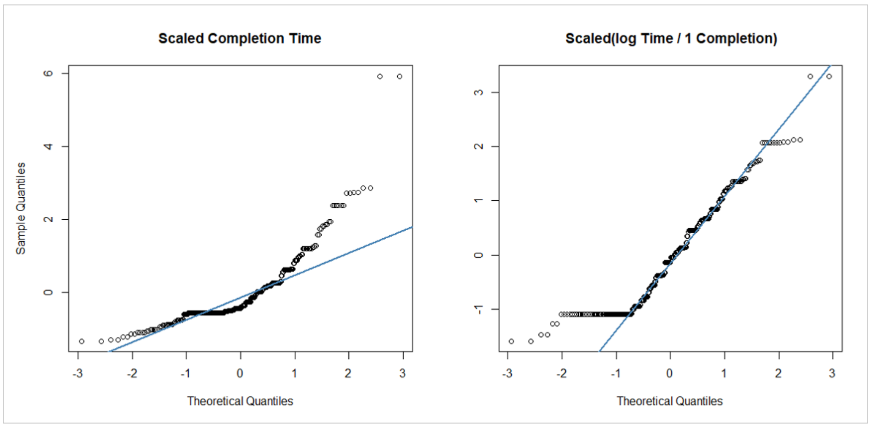


Figure 4: Qualitative analysis of response variable: Scaled Time, for normality

2.2 Factor-Level Summaries

The data is summarized by factor in a side by side comparative fashion. This way potential stress and learning relationships may be qualitatively observed for later verification. These figures are highly informative of the statistical analysis conducted in Section 3.

2.2.1 State Psychometrics

Observing the summation of the Trait Psychometric data for the cutting task, we see that in general each subject had the sentiment of the task getting easier over the repetition of the sessions.

When compared to the previous results, the psychometric data has different behavior for the suturing task for the summation of all subjects. It would appear that the relative stasis of the values for each response would suggest that the subjects felt they weren't improving or felt the task was not getting easier.

When analyzing significant learning trends in the data, and making real-world conclusions, we may keep in mind the general perception of the subjects on the difficulty of the tasks.

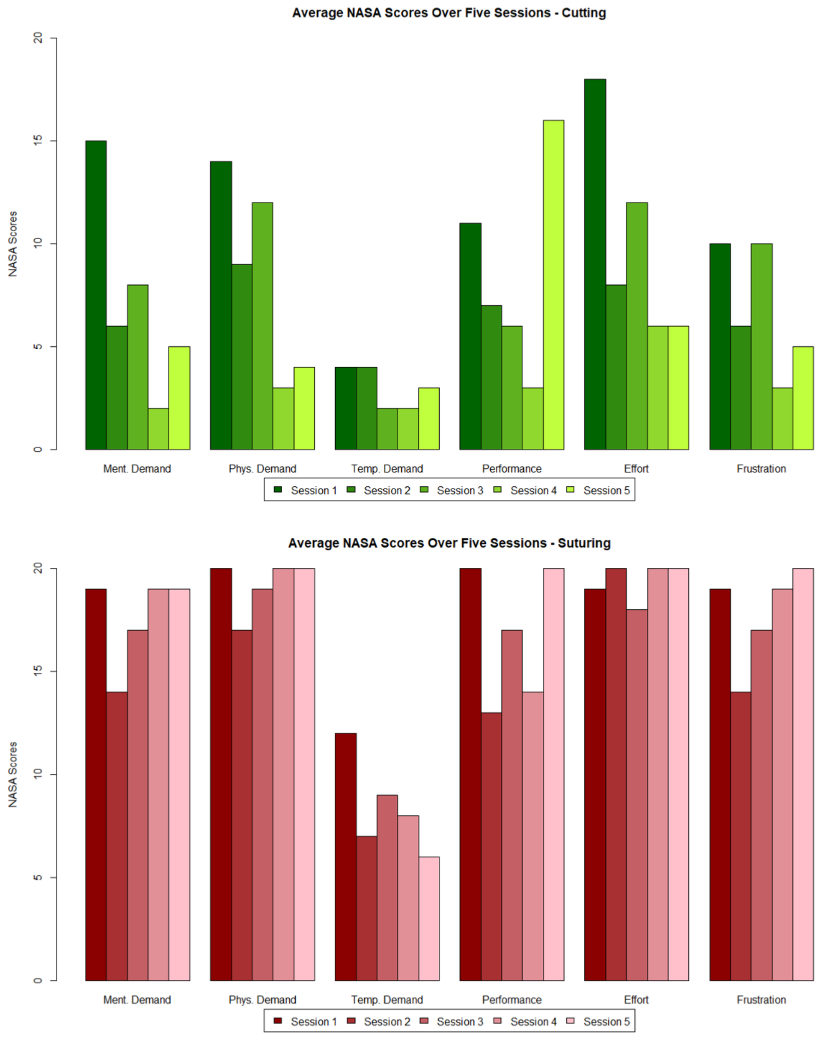


Figure 5: State Psychometrics Summary: Cut (top) Suture (bot)

2.2.2 Perinasal Perspiration

The Raw PP signal is observed in its' translated state, right before the LN transform is made. We do not make hypothesis about the stress signal; however, qualitatively it appears to be fairly uniform across the 10 primary factorial combinations. The lone exception is some highly abberant behavior on the 5th session of suturing. It seems there are very mixed feelings once session # 5 is reached, ranging from relief (low stress) to very high stress.

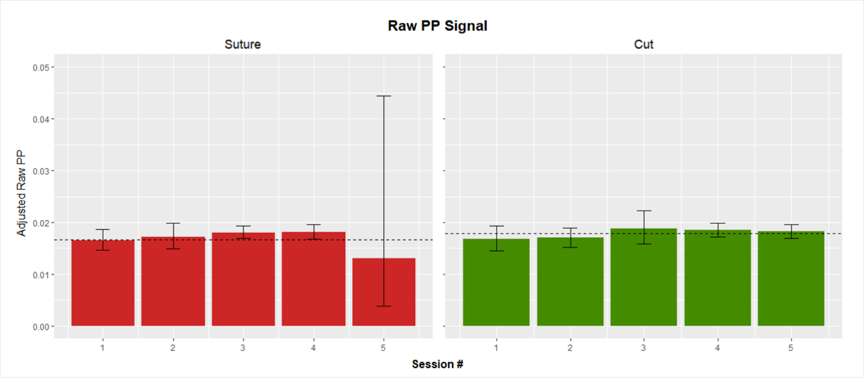


Figure 6: Summary of Raw PP Signal by Factors

2.2.3 Accuracy Score

Here we observe a learning trend across sessions. The response variable Accuracy Score continues to improve with each session. We note, qualitatively, that the mean score for Cut & Suture appear to differ. We will asses if the differences between tasks are significant.

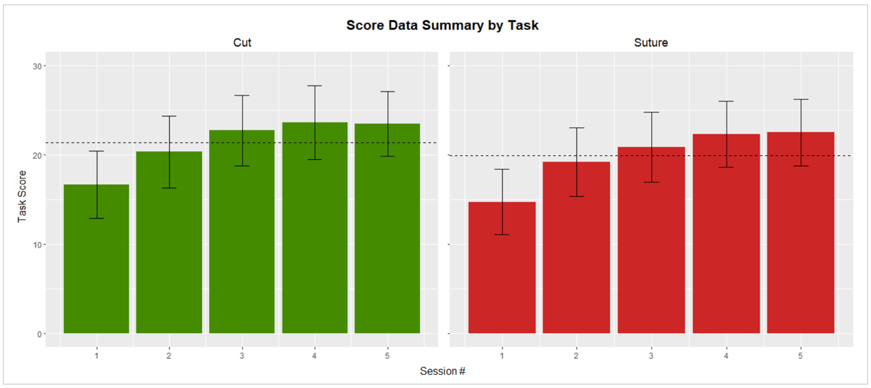


Figure 7: Summary of Accuracy Score by Factors

2.2.4 Raw Completion Time

This figure is useful for a discussion on the difficulty of using raw completion time as a response variable. The task-wise distributions do not describe the same information. The experiment could have been designed better if they controlled for inconsistencies in the task constrains. Cutting and Suturing have their own set of constraints for time limit and # of tasks one can complete. The figure is a reminder that time, in its raw form, is not able to properly represent the learning process. response variable.

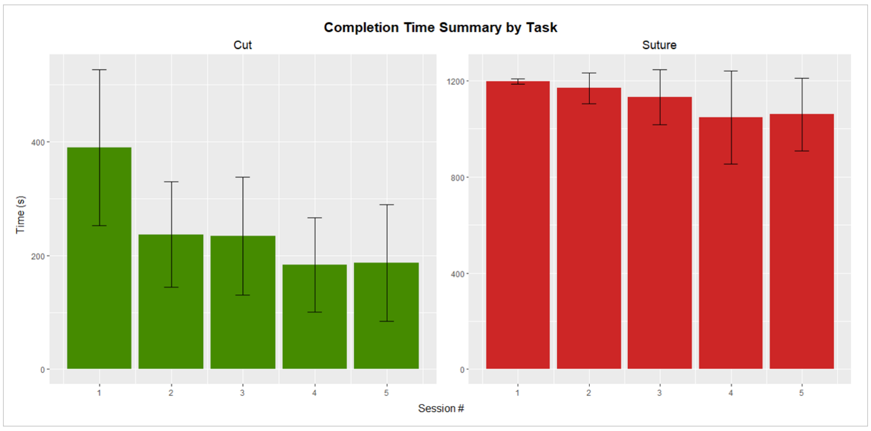


Figure 8: Summary of Raw Completion Time by Factors

2.2.5 # of Suture Considerations

We observe & incorporate No. of Sutures Completed with the suture time data. The goal is to create a new feature called ”per one” which considers the time to complete a single non-repeated task. Cutting is limited to one cut per session. Sutures are allowed to go up to 6. The suture time data is divided by the # of sutures completed to form a more meaningful representation of completion time as a a learning indicator. Observe that the clear learning trend in the left figure. Given a fixed time constraint of 20 minutes, the subjects are able to complete more and more sutures by session. Completion time also shows a trend much more favorable and comparable to the cutting times.

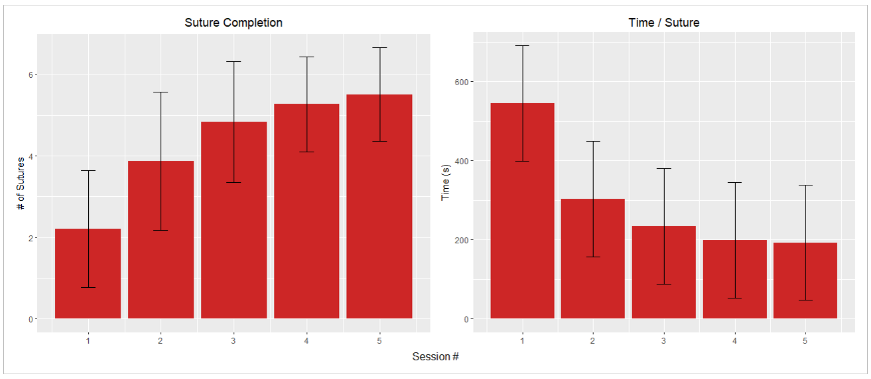


Figure 9: Suture Learning, Left: Improved quantity, Right: Improved time per suture

2.2.6 Scaled Completion Time

The adjusted time data is centered by scaling to improve the analogy between tasks. We now observe highly comparable learning curves between tasks. The session learning curve is very clear and highly pronounced. It seems there is a possibility for significant differences between tasks for this performance metric but it is not entirely clear by observation.

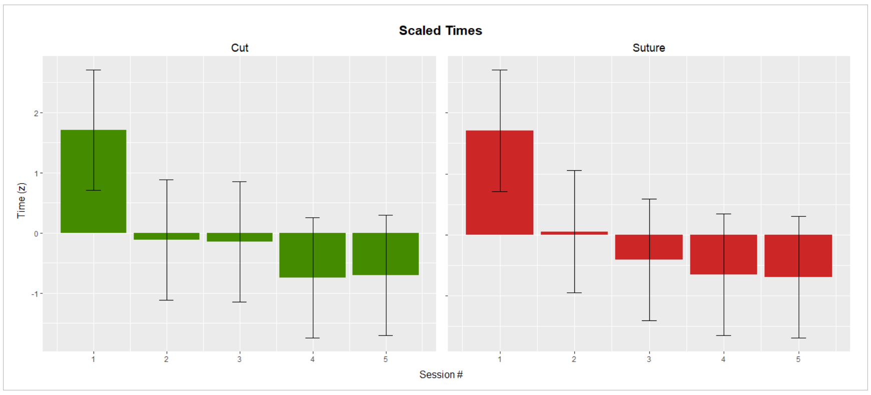


Figure 10: Comparison of scaled time learning curves between tasks

3 Statistical Inference

We developed the following set of hypothese to address the primary concerns of the study:

H_{01}: No main effects on all performance attributes for all factorial combinations
H_{02}: No main effect on all performance attributes for PP Signal
H_{03}: No blocking effect on all performance attributes for Gender or Tai
H_{04}: No Random effect on all performance attributes for subject
H_{05}: No Interaction effects on performance between pairwise combinations of Session, Task, & PP

We consider two distinct response variables, Accuracy Score, and Scale Time, for each mixed and fixed effects, noting the differences between models. The models presented are the highest quality models achieved.

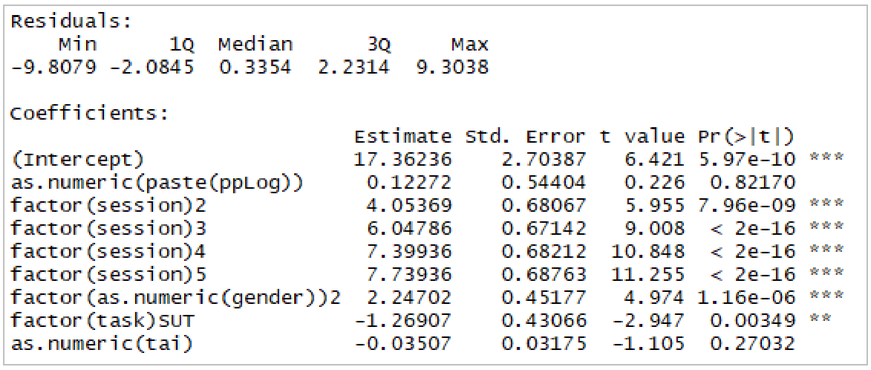
Effects are tested by trial and error, sometimes informed by the summary data. The models shown here are, in general, the highest quality model we developed for each model type. Trial and error is manageable because there are very few factorial combinations worth considering for interaction effects.

3.1 Response Variable: Accuracy Score

1. Fixed Effects Model

$$\text{Accuracy} \sim \ln(\text{PP}) + \text{session} + \text{task} + \text{tai} + \text{gender}$$

From the linear model on Accuracy Score using fixed effects, We can see stress does not appear to have any significant effect on Accuracy, while session, gender, and task do have significant effects. Task notably has a negative slope,which suggests that the suturing task may be more difficult (lower scoring) compared to the cutting task. Tai score does not have any signifcance, so we can remove it for subsequent models.



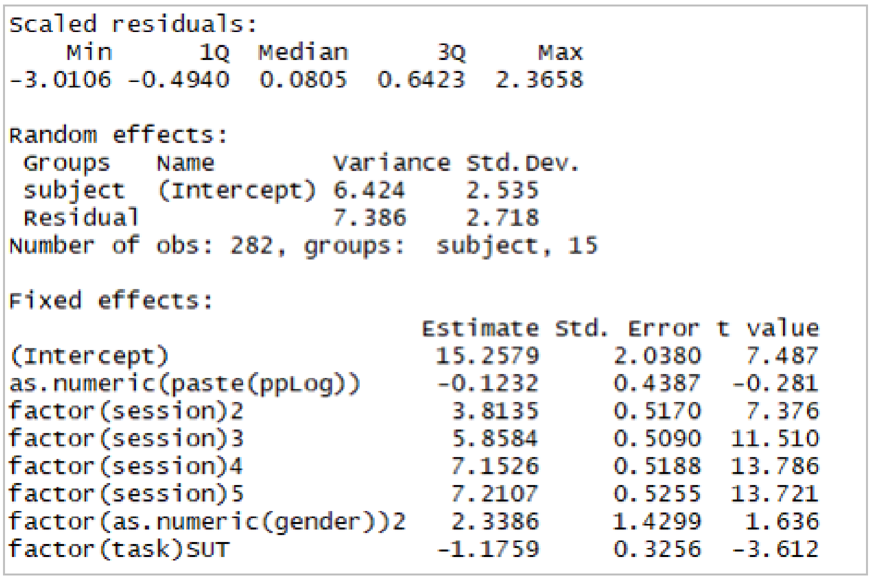
```
Residuals:
    Min      1Q  Median      3Q     Max 
-9.8079 -2.0845  0.3354  2.2314  9.3038 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                17.36236    2.70387   6.421 5.97e-10 ***
as.numeric(paste(ppLog))    0.12272    0.54404   0.226  0.82170    
factor(session)2            4.05369    0.68067   5.955 7.96e-09 ***
factor(session)3            6.04786    0.67142   9.008  < 2e-16 ***
factor(session)4            7.39936    0.68212  10.848  < 2e-16 ***
factor(session)5            7.73936    0.68763  11.255  < 2e-16 ***
factor(as.numeric(gender))2  2.24702   0.45177   4.974 1.16e-06 ***
factor(task)SUT            -1.26907    0.43066  -2.947  0.00349 ** 
as.numeric(tai)            -0.03507    0.03175  -1.105  0.27032    
```

Figure 11: Fixed Effects Model: Response = Accuracy,

2. Mixed Effects Model

$$\text{Score} \sim \ln(\text{PP}) + \text{session} + \text{task} + \text{tai} + \text{gender} + (\ 1 \mid \text{subject})$$

When adding mixed effects to the Score response, we can see that gender suddenly loses significance while the other variables remain the same. Significance we assumed was due to gender in fact belonged to the random effects between subjects.



```
Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0106 -0.4940  0.0805  0.6423  2.3658

Random effects:
 Groups   Name        Variance Std.Dev.
 subject  (Intercept) 6.424    2.535
 Residual             7.386    2.718
Number of obs: 282, groups:  subject, 15

Fixed effects:
                            Estimate Std. Error t value
(Intercept)                  15.2579     2.0380   7.487
as.numeric(paste(ppLog))     -0.1232     0.4387  -0.281
factor(session)2              3.8135     0.5170   7.376
factor(session)3              5.8584     0.5090  11.510
factor(session)4              7.1526     0.5188  13.786
factor(session)5              7.2107     0.5255  13.721
factor(as.numeric(gender))2   2.3386     1.4299   1.636
factor(task)SUT              -1.1759     0.3256  -3.612
```

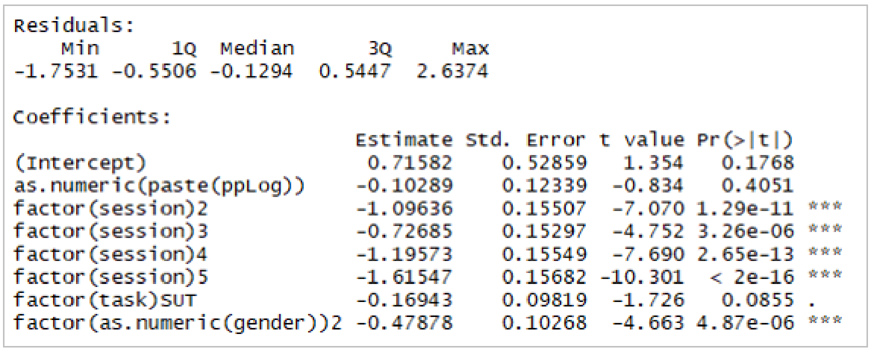
Figure 12: Mixed Effects Model: Response = Accuracy t.thresh ≈ 1.95

3.2 Response Variable: Scaled Time

1. Fixed Effects Model

   From the linear model on Scaled Time data, Stress response does not appear to have any signifance on the Scaled Time. Session has a negative relationship, so each session the subjects are completing their time per task faster. Task has no significance while gender does appear to play a roll in Scaled Time measures (female subjects are able to complete in tasks in times faster than the male subjects)

$$\mathit{Scaled\ Time} \sim \mathit{ln(PP)} + \mathit{session} + \mathit{task} + \mathit{gender}$$

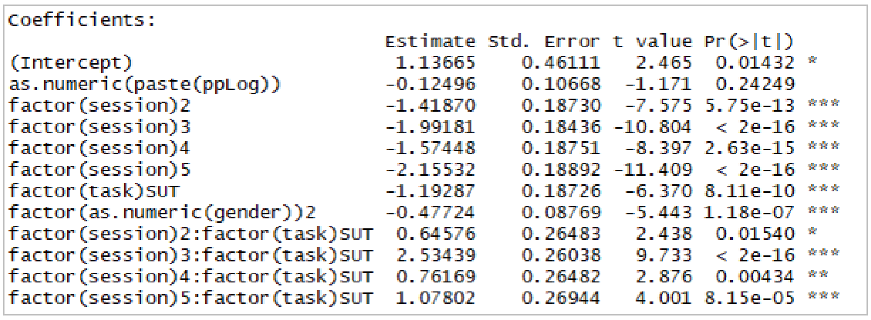


```
Residuals:
    Min      1Q  Median      3Q     Max
-1.7531 -0.5506 -0.1294  0.5447  2.6374

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.71582    0.52859   1.354   0.1768
as.numeric(paste(ppLog))    -0.10289    0.12339  -0.834   0.4051
factor(session)2            -1.09636    0.15507  -7.070 1.29e-11 ***
factor(session)3            -0.72685    0.15297  -4.752 3.26e-06 ***
factor(session)4            -1.19573    0.15549  -7.690 2.65e-13 ***
factor(session)5            -1.61547    0.15682 -10.301  < 2e-16 ***
factor(task)SUT             -0.16943    0.09819  -1.726   0.0855 .
factor(as.numeric(gender))2 -0.47878    0.10268  -4.663 4.87e-06 ***
```

Figure 13: Fixed Effects Model: Response = Scaled Time

When adding an interaction effect for session * task, we find that these interactions are significant. Task is once more significant. It appears the interaction effect completely robbed task of significance. Including this interaction term is crucial if wish for our model to be representative.

$$Scaled\ Time \sim ln(PP) + session + task + gender + task{*}session$$



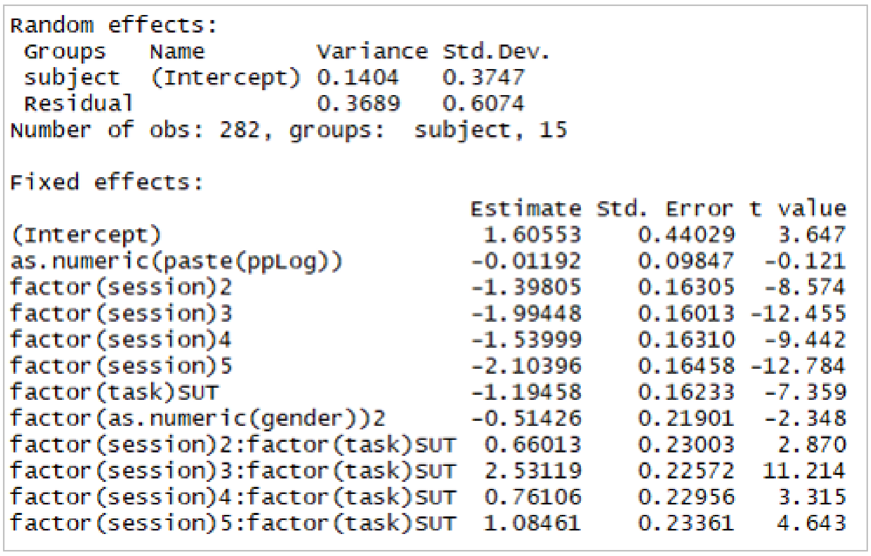
```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    1.13665    0.46111   2.465  0.01432 *  
as.numeric(paste(ppLog))      -0.12496    0.10668  -1.171  0.24249    
factor(session)2              -1.41870    0.18730  -7.575 5.75e-13 ***
factor(session)3              -1.99181    0.18436 -10.804  < 2e-16 ***
factor(session)4              -1.57448    0.18751  -8.397 2.63e-15 ***
factor(session)5              -2.15532    0.18892 -11.409  < 2e-16 ***
factor(task)SUT               -1.19287    0.18726  -6.370 8.11e-10 ***
factor(as.numeric(gender))2   -0.47724    0.08769  -5.443 1.18e-07 ***
factor(session)2:factor(task)SUT  0.64576    0.26483   2.438  0.01540 *  
factor(session)3:factor(task)SUT  2.53439    0.26038   9.733  < 2e-16 ***
factor(session)4:factor(task)SUT  0.76169    0.26482   2.876  0.00434 ** 
factor(session)5:factor(task)SUT  1.07802    0.26944   4.001 8.15e-05 ***
```

Figure 14: Fixed Effects Model w/ interactions: Response = Scaled Time

2. Mixed Effects Model

   This model incorporates random effects as well as the fixed effects from the previous model. Its important to note that gender does not lose significance in this inclusion of random effects. By ANOVA (not shown) , the Random Effect model is significantly different than the Fixed model. In all cases, random effects were found to be significant when comparing to the Fixed model. In this case, the random effect is significant, and must be included, although the inclusion does not alter the significance of the other predictors in a meaningful way.

$$Scaled\ Time \sim ln(PP) + session + task + gender + task{*}session + (\ 1 \mid subject)$$



```
Random effects:
 Groups   Name        Variance Std.Dev.
 subject  (Intercept) 0.1404   0.3747  
 Residual             0.3689   0.6074  
Number of obs: 282, groups:  subject, 15

Fixed effects:
                                 Estimate Std. Error t value
(Intercept)                       1.60553    0.44029   3.647
as.numeric(paste(ppLog))         -0.01192    0.09847  -0.121
factor(session)2                 -1.39805    0.16305  -8.574
factor(session)3                 -1.99448    0.16013 -12.455
factor(session)4                 -1.53999    0.16310  -9.442
factor(session)5                 -2.10396    0.16458 -12.784
factor(task)SUT                  -1.19458    0.16233  -7.359
factor(as.numeric(gender))2      -0.51426    0.21901  -2.348
factor(session)2:factor(task)SUT  0.66013    0.23003   2.870
factor(session)3:factor(task)SUT  2.53119    0.22572  11.214
factor(session)4:factor(task)SUT  0.76106    0.22956   3.315
factor(session)5:factor(task)SUT  1.08461    0.23361   4.643
```

Figure 15: Mixed Effects Model: Response = Scaled Time, t.thresh ≈ 1.95

4 Discussion / Conclusions

Stress Effects

- No main effects for all stress indicators on each performance measure. (Correlation coefficients are zero)
- Null Hypothesis for stress fails to reject on all levels
- No Significant Relationship between Stress & Performance

Learning Effects

- Main Effects
  - Session
  - Task
- Blocking Effects
  - Gender
    - No Blocking Effect on Score
    - Significant blocking effect on Scaled Time
  - TAI - No significant blocking effects
- Interaction Effects: Significant Interaction on Scaled time for Task * Session. The interaction turns out to be crucial to the significance of the main effect on Task
- Random Effects Significant Random effects on performance for subject null hypothesis for Random Effects rejected on all levels

In lay terms, stress (given by PP) is not a significant predictor of microsurgical performance for accuracy or completion time. Stressed individuals are equally competent as unstressed ones. There should not be significant concern about the competency of stressed individuals in a microsurgery setting.

The learning process is exhibited by the by improvements in the two selected performance metrics. Performance is shown to significantly improve over the course of 5 sessions, peaking on session 5. Suturing is shown to be a significantly more difficult than cutting. There is a significant expectation that subjects score lower and take longer to complete suturing tasks.

The broad presence of significant random effects is unsurprising. Microsurgery is a complex set of tasks, and humans are even more complex with any number of uncontrolled factors that may influence the learning process. Accounting for their presence gives strength and credibility to the significant effects that remain once random effects are accounted for.