Statistical Models for Data Science

Assignment 2

Experience sampling data

Submitted by: Haya Salameh

Submitted to: Prof. Philip Tzvi Reiss

Date of Submission: 05/01/2025

Table of Contents

Study Overview

This experience sampling study investigated "mental time travel" - examining how thoughts oriented toward the future vary throughout the day. The data included experience sampling data from 492 participants over three days focused on relationships between time orientation or future-oriented thinking (a binary variable) and:

1. Non-time-varying characteristics (age, sex, personality traits) – used in question 1.

2. Time-varying variables (mood, stress level, social context) – used in question 2.

Question 1 - Logistic Regression of Future-Oriented Thinking

The analysis investigates the relationship between time of day and future-oriented thinking using experience sampling data from 492 participants over three days. This analysis uses logistic regression models to analyze binary outcomes (thinking/not thinking about the future) and then accounts for the hierarchical structure of the data through mixed modeling.

Data structure

Number of participants is 492.

- Subject: Participant identification number

- TIME.S: Time in seconds from midnight (e.g., 9:00 = 32,400 seconds)

- DAY: Study day (1, 2, or 3)

- future: Binary indicator for future-oriented thinking (1 = yes, 0 = no)

Participants were contacted at six random time points each day, resulting in a potential **maximum of 18 observations per person**.

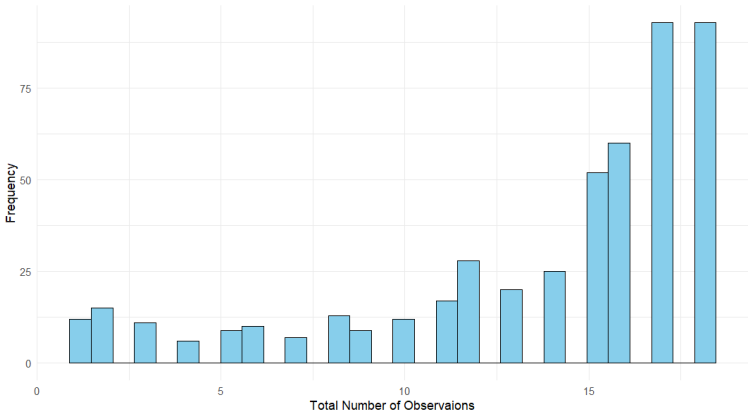
Missing data

Even though the data did not include NA values, if we take into consideration that the complete data should include 18 observations per subject we see that there is missing observations.

Characteristics of missing observations across subjects

- The number of missing observations is 2170 ($number\ of\ subjects\ *\ 18 -\ number\ of\ actual\ observations$), and the proportion of missing observations is 24.5%

- 80.7% of participants have incomplete observations (at least one observation is missing)
- 12 subjects (2.44%) have only one observation
- The median number of observations for participants is 15.5
- The mean number of observations across participants is 13.6



Characteristics of missing observations across days

- There is a progressive increase in missing data rate in days 1-3:
  - Day 1 (18.6%), Day 2 (25.3%), Day 3 (27.3%)
- Using a logistic regression model to model the probability of whether the future variable is missing as a function of the DAY variable suggests that missingness is related to study day:
- Model formula: $logit(P(missing\_future_{ijk} = 1)) = \beta_0 + \beta_1 \cdot DAY2_{ijk} + \beta_2 \cdot DAY3_{ijk}$ where $missing\_future_{ijk}$ is the binary indicator (1 if missing, 0 if observed) for subject i, day j, observation k.
- The reference group is observations on day 1. The intercept term has a value of -1.424. The **odds** of missing data on day one is $e^{\beta_0} = e^{-1.424} \approx 0.240$ $(p < 2e-16)$, thus the **probability** of missing data in day 1 is $p = \frac{e^{\beta_0}}{1+e^{\beta_0}} = 0.194$.
- $\beta_1 = 0.382, p = 1.07e-09$. Meaning that the odds of missing data on day 2 are about $e^{0.382} \approx 1.465$ **times** higher compared to Day 1.
- $\beta_2 = 0.480, p < 1.01e-14$. Meaning that the odds of missing data on day 3 are about $e^{0.48} \approx 1.616$ **times** higher compared to Day 1.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.42355    0.04654 -30.591  < 2e-16 ***
DAY2         0.38194    0.06263   6.098 1.07e-09 ***
DAY3         0.47983    0.06201   7.738 1.01e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **These results show that the likelihood of missing data increases as we move from day 1 to day 2 and day 3. This suggests that the data is not missing completely at random (MCAR).**
- I assume multiple imputation is not feasible because the time values themselves are missing. And the missingness mechanism in the time variable and in the future variable may be related to the **unobserved values of time or other variables.**

Descriptive statistics

6,686 total observations form 492 participants.

Day (DAY)

Distribution of observations across days:

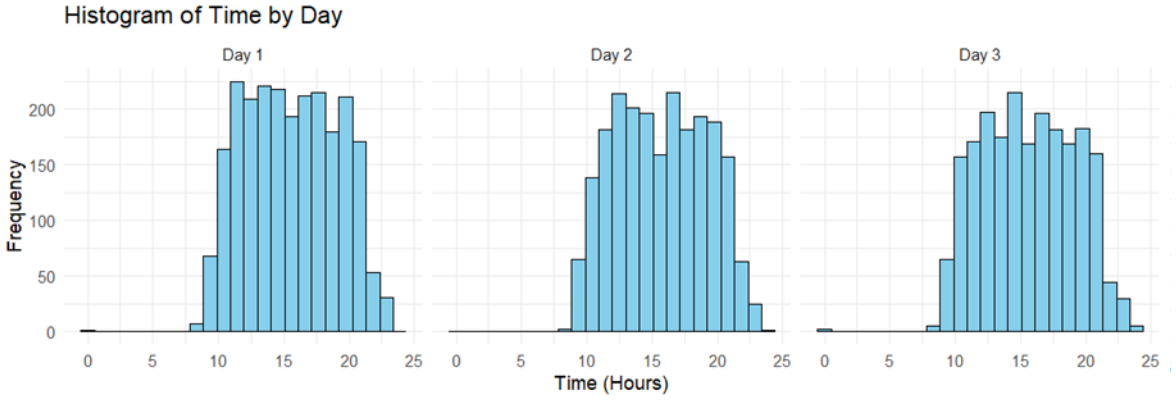
- Day 1: 2,379 observations (35.6%)
- Day 2: 2,182 observations (32.6%)
- Day 3: 2,125 observations (31.8%)

  Shows a progressive decrease in response rates across days (see missing data section).

Time (TIME.S)

Time in seconds from midnight, converted to hours for analysis:

- Range: 0.023 to 23.9 hours
- Mean: 15.6 hours
- Median: 15.6 hours
- Variance: 12.4 hours^2
- Similar distributions across all three days
- Sampling appears to be concentrated during waking hours

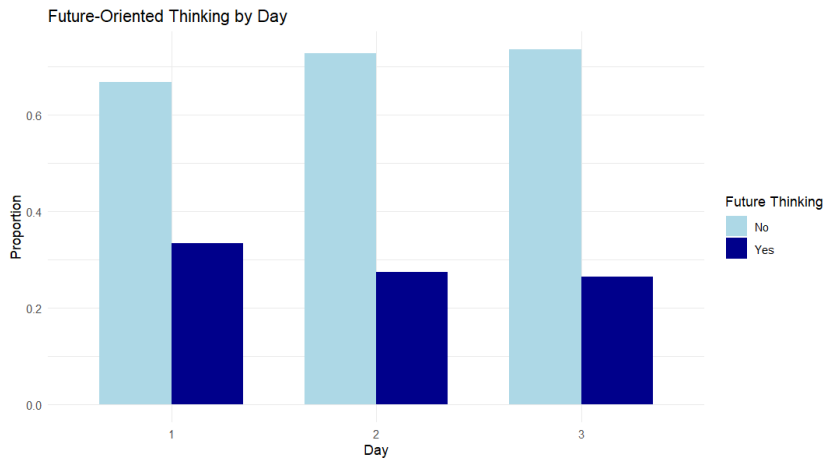


Future-oriented thinking – outcome variable

Binary variable with mean 0.292, meaning that future-oriented thinking occurred in 29.2% of observations.

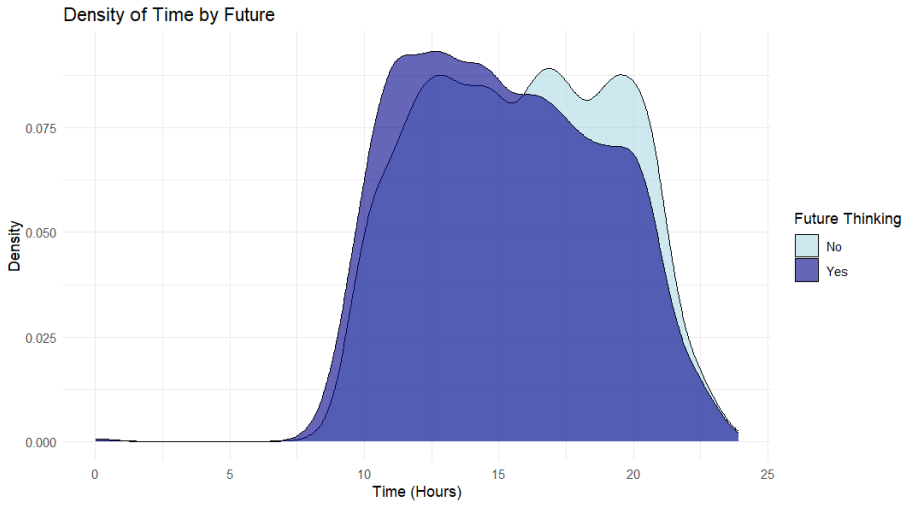
Distribution by Day:

- We see a decrease in future-oriented thinking across days of the study, the percentage of future thinking on day 1 is 33.3%, then on day 2 it is 27.4%, and on day 3 it is 26.4%.
- Complementary increase in non-future thinking across days 66.7%->72.65->73.6%.

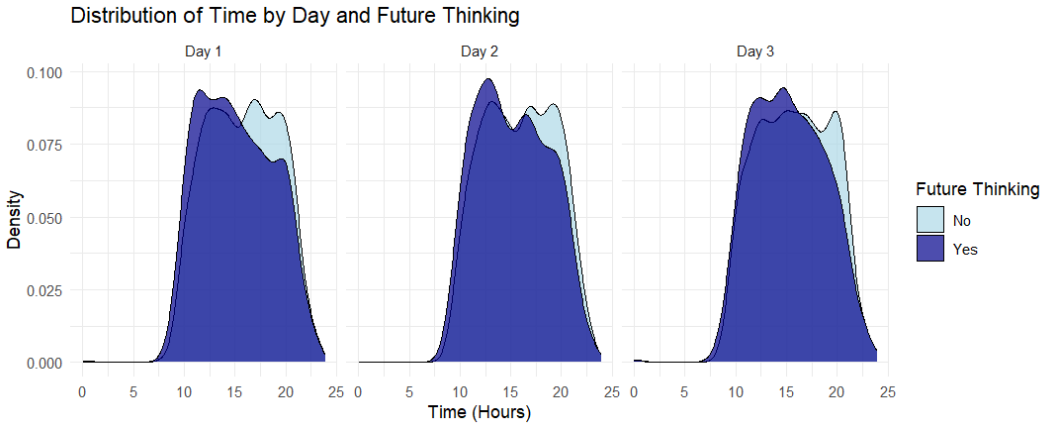


Distribution by Time:

- Qualitatively, by looking at the density plot of time by future, participants are more likely to engage in future thinking during morning hours and gradually shift toward non-future thinking as the day progresses.



- This pattern is consistent time distribution patterns across all three days



Logistic regression of future-oriented thinking

Simple logistic regression

A simple logistic model was fitted (estimated using ML) to predict future with TIME.H. The model ignores the random effect of subject and the hierarchical structure.

Model formula: $logit(P(future_{ijk} = 1)) = \beta_0 + \beta_1 TIME_{ijk}$. Where:

- $future_{ijk}$: Binary outcome of future thinking in observation i from subject j on day k
- $TIME_{ijk}$: Time of day in hours for observation i within day k of participant j

Results of fitting:

- Intercept: $\beta_0 = -0.112, p = 0.355$. A non-significant intercept means that the log odds of future thinking at midnight is 0, which means that the probability of future thinking at midnight is 0.5 ($p = \frac{e^{\beta_0}}{1+e^{\beta_0}} = \frac{1}{2}$).
- Time coefficient is significant and negative, $\beta_1 = -0.050,\ p = 8.8e-11.$ A unit increase in time (an hour) **decreases** the odds of future thinking by $e^{\beta^1} \cong 0.95$

Goodness of fit:

- $R^2 = 1 - \frac{deviance\ of\ model}{deviance\ of null\ model} = 0.0053$ (weak)
- AIC = 8035, BIC=8048
- Deviance: 8031

Mixed-effects logistic regression

A logistic mixed-effects model was fitted (estimated using ML and Nelder-Mead optimizer implemented by *lme4::glmer()*) with time (in hours) as a fixed effect and random intercepts for participant and day within Participant.

Model formula: $logit(P(future_{ijk} = 1)) = \beta_0 + \beta_1 \cdot TIME_{ijk} + u_j + v_{jk}$. Where:

- $future_{ijk}$: Binary outcome of future thinking in observation i from subject j on day k
- $TIME_{ijk}$: Time of day in hours for observation i from subject j on day k
- u_j: Subject-specific random variation
- v_{jk}: Day-specific random variation for day k within participant j

Results of fitting:

- Intercept: $\beta_0 = -0.126, p = 0.35$. A non-significant intercept means that the log odds of future thinking at midnight is 0, which means that the probability of future thinking at midnight is 0.5 ($p = \frac{e^{\beta_0}}{1+e^{\beta_0}} = \frac{1}{2}$).
- Time coefficient is significant and negative, $\beta_1 = -0.056,\ p = 1.8e-11.$ A unit increase in time (an hour) **decreases** the odds of future thinking by $e^{\beta^1} \cong 0.95.$

Random Effects:

- Participant: Variance of 0.533 (SD = 0.730)
- Day within Participant: Variance 0.082 (SD = 0.287)
- The variance at the participant level is larger than the variance at the day-within-participant level. This suggests that differences between participants contribute more to variability in the outcome than differences between days within the same participant.

Goodness of fit:

Note: R^2 for mixed model implemented by *MuMIn* library based on (Nakagawa & Schielzeth, 2013) and on (Nakagawa et al., 2017), it was calculated two ways:

1. Marginal R^2 - variance explained by fixed effects only.
2. Conditional R^2 - variance explained by the fixed **and random effects**.

- Marginal R^2 = 0.0097 (R^2 of fixed effects)
- Conditional R^2 = 0.166 (R^2 of total model) - moderate
- AIC: 7801, BIC = 7828
- Deviance = 7793

Comparison of Simple Logistic and Mixed-Effects Models

The two models show both agreements and important differences:

Agreement:

- Both models show significant negative relationship between time and future thinking, indicatung decreasing probability of future thinking throughout the day
- Similar intercepts: -0.112 (simple) vs -0.126 (mixed)
- Similar time coefficients: -0.050 (simple) vs -0.056 (mixed)

Differences in model structure:

- Simple model treats all observations as independent
- Mixed model accounts for: repeated measures within subjects and nested structure of days within subjects. Thus it accounts for individual variations in baseline future thinking.

Differences in model fit and explanatory power:

- Mixed model shows substantially better fit with AIC (7801 vs 8035) and BIC (8048 vs 7828).
- Using the likelihood ratio test with anova() function, we see that the mixed model has a significantly better fit than the simple model Chisq=237.98, p<2.2e-16
- Mixed model explains more variance:
    - Simple model R^2 = 0.0053 (weak)
    - Mixed model total (conditional) R^2 = 0.166 (moderate)
    - The difference between marginal and conditional R^2 in the mixed model (0.0097 vs 0.166) indicates that most of the explained variance comes from individual differences rather than time effects.

The mixed-effects model is more appropriate for this analysis as it accounts for the repeated measures within participants and the hierarchical structure of the study, resulting in a better overall model fit with more exploratory power.

Question 2 - Poisson Regression of Future-Oriented Thinking

The final dataset consists of 477 subjects (492 before handling missing data) with the following variables:

- Number of observations: Total number of responses per subject (numeric)
- Number of future thoughts: numeric count **outcome variable**.
- Age: Participant's age in years (numeric)
- Sex: Participant's gender (categorical: male/female)
- Personality Traits (all numeric, scale 1-7) with description based on Wikipedia:
    - A – Agreeableness: High scorers tend to be more helpful and empathetic, while low scorers may be more competitive or assertive.
    - C – Conscientiousness: High scorers are typically disciplined and detail-oriented, while low scorers may be more flexible but less structured.
    - E – Extraversion: High scorers are typically outgoing and energized by social interaction, while low scorers (introverts) prefer solitary activities and quieter environments.
    - N – Neuroticism: High scorers may experience more anxiety, mood swings, and emotional instability, while low scorers tend to be more emotionally stable and resilient to stress.
    - O - Openness: High scorers tend to be imaginative and enjoy new experiences, while low scorers prefer routine and familiar situations.

I first verified that trait variables (A, C, E, N, O, age, sex) were constant for each subject. And ranges of variables indicated no impossible values or outliers or data entry errors.

Missing Data

- 14 subjects had missing values for all seven predictor variables (age, sex, A, C, E, N, O). These subjects with complete missingness were removed.
- One subject had missing values for A, C, E, N, and O (all traits) but no missing value in demographic variables was also removed.
- The final sample size is 477 subjects, with 0 missing values.

Descriptive Statistics

Demographics

Age

- Mean: 28.8 years
- Median: 26
- Standard Deviation: 9.61 years
- Range: 18-67 years

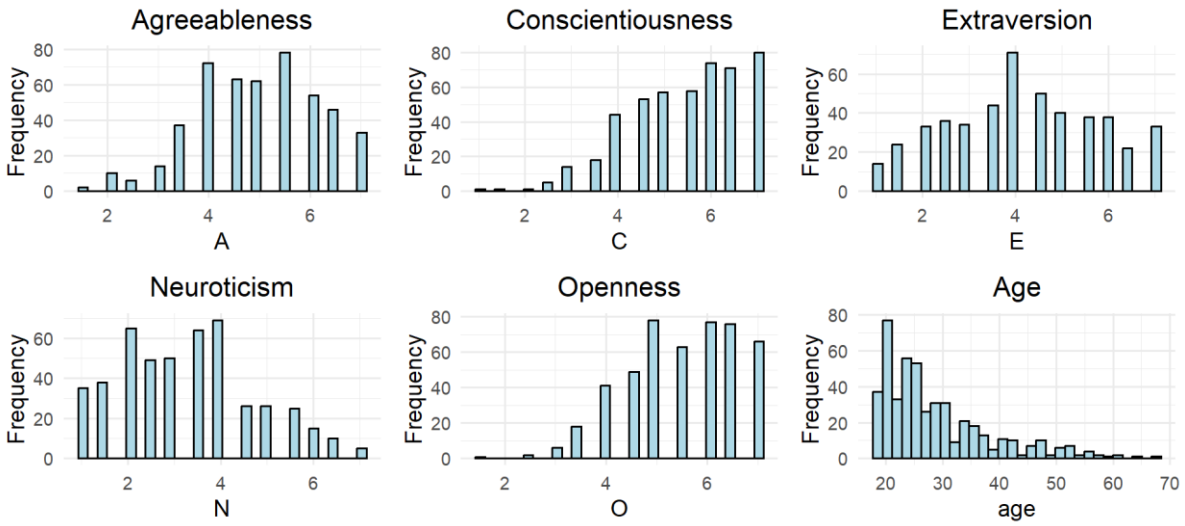
Sex

- Female: 305 (63.8%)
- Male: 173 (36.2%)

Personality traits

Trait	Missing	Min	Max	Mean	SD
Agreeableness (A)	1	1.5	7.0	4.98	1.20
Openness (O)	1	1.5	7.0	5.51	1.07
Conscientiousness (C)	1	1.0	7.0	5.46	1.20
Extraversion (E)	1	1.0	7.0	4.16	1.62
Neuroticism (N)	1	1.0	7.0	3.29	1.45

Histograms of numeric predictor variables

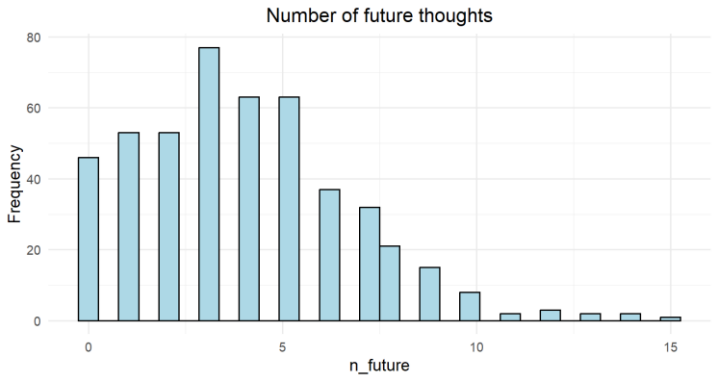


Number of Observations per Subject

- Mean: 13.78
- Median 16
- Standard Deviation: 4.69
- Range: 1-18

Number of Future Thoughts

- Mean: 4.00
- Median: 4
- Skewness: 0.78
- Standard Deviation: 2.81
- Range: 0-15

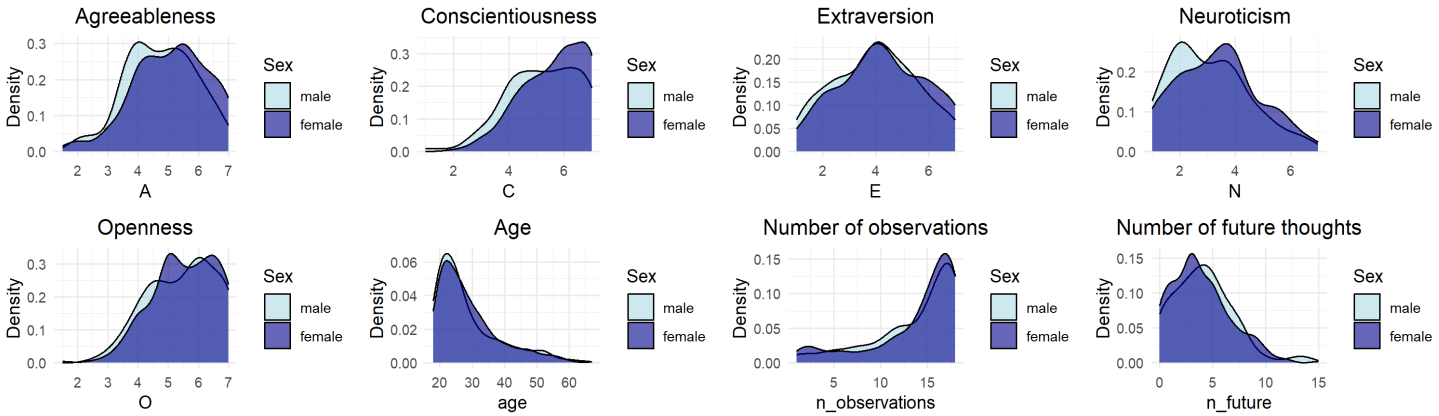


Variable Relationships

Numeric variables and Sex

Qualitatively by examining the distribution of by sex (see figure below) we see:

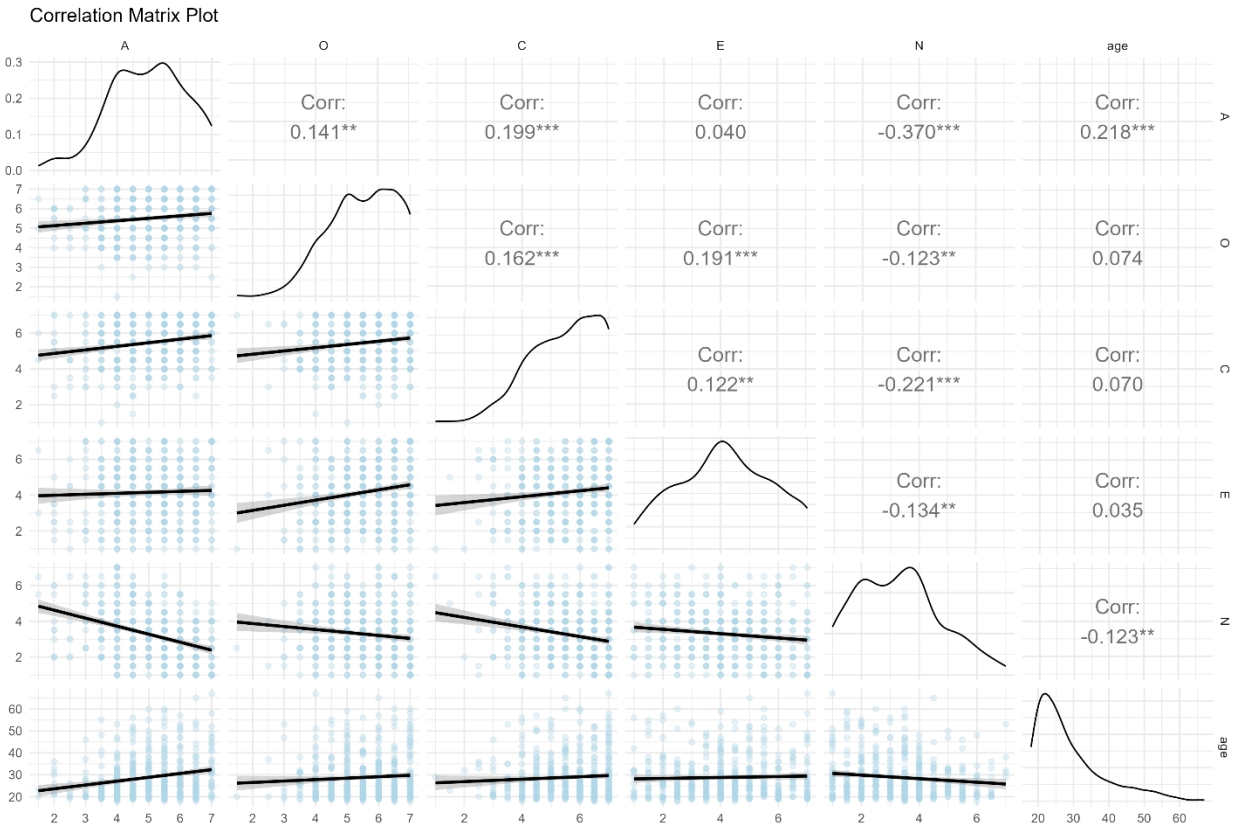
- Neuroticism shows the most pronounced sex difference, with females (peak around 3.5-4) reporting higher values than males (peak around 2.5-3).
- Conscientiousness and agreeableness scores in females show a slight tendency toward higher values,with males showing more spread distribution
- The distributions of other variables (including the number of future thoughts) show minimal sex differences.



Correlations between numeric variables

Pearson correlations between numeric predictor variables showed:

- Age is positively correlated with Agreeableness (r=0.218, p-value=1.47e-06) and negatively correlated with Neuroticism (r = -0.123, p-value = 0.0073).
- Neuroticism is significantly negatively correlated with all the rest of the personality traits.
- The rest of the personality traits are significantly positively correlated with each others. (See pairs correlation plot below).



Individual Poisson regression

Fitting models for predicting the number of future thoughts with individual predictor variables (with an offset term for number of observations) showed that :

- Only Openness shows significance (p=0.008) with a positive coefficient (0.057)
- Sex has largest effect size (coef=-0.058) but not significant (p=0.216)
- Extraversion shows marginal significance (p=0.065, coef=0.026)
- All other variables show weak effects (|coef| < 0.02) and non-significance (p > 0.3)

variable	coefficient	std_error	p_value	abs_coef	significance
sex	-0.058	0.047	0.216	0.058	ns
O	0.057	0.021	0.008	0.057	**
E	0.026	0.014	0.065	0.026	ns
A	0.018	0.019	0.346	0.018	ns
C	-0.013	0.019	0.503	0.013	ns
age	-0.002	0.002	0.470	0.002	ns
N	0.002	0.016	0.883	0.002	ns

Model selection

- Based on the individual regression results, model selection based on the significance of individual models maybe very restrictive as a model selection method as it would result in choosing O and maybe sex and E because of the higher effect size compared to other predictors.
- This is especially true given that potential relationships may exist in combination (as seen in the EDA part).
- I will start with a full interaction model and “prune” the terms using AIC and BIC - based model selection to potentially identify meaningful variable combinations that might be missed by strictly using significance and effect size levels.
- I will do this keeping in mind that probably sex, O, and E are expected to be included in the final model.

Models Compared:

1. Main effects model: All predictors without interactions
2. Full interactions model: All predictors with all possible two-way interactions
3. Stepwise AIC selection model: Starting from full model, using AIC criterion:
   The model after stepAIC(): n_future ~ A + C + E + O + age + sex + A:C + A:O + A:age + A:sex + C:O + C:age + C:sex + E:sex + age:sex + offset(log(n_observations))

4. Stepwise BIC selection model: Starting from full model, using BIC criterion (k=log(n)).
   The model after stepAIC(): n_future ~ A + C + O + age + A:C + C:age + offset(log(n_observations))

Coefficients of best model with BIC step selection

	n future		
Predictors	*Incidence Rate Ratios*	*CI*	*p*
(Intercept)	0.33	0.12 – 0.91	**0.033**
Agreeableness	0.79	0.67 – 0.94	**0.007**
Conscientiousness	0.91	0.76 – 1.09	0.302
Openness to Experience	1.06	1.02 – 1.11	**0.005**
age	1.03	1.00 – 1.05	**0.021**
A:C	1.05	1.02 – 1.08	**0.003**
C:age	0.99	0.99 – 1.00	**0.009**

Coefficients of best model with AIC step selection

	n future		
Predictors	*Incidence Rate Ratios*	*CI*	*p*
(Intercept)	0.29	0.06 – 1.50	0.145
Agreeableness	0.71	0.54 – 0.95	**0.021**
Conscientiousness	1.07	0.83 – 1.38	0.596
Extraversion	0.99	0.95 – 1.04	0.805
Openness to Experience	0.96	0.74 – 1.24	0.758
age	1.05	1.02 – 1.07	**0.001**
sex: female	0.71	0.39 – 1.30	0.267
A:C	1.04	1.01 – 1.08	**0.006**
A:O	1.06	1.02 – 1.10	**0.005**
A:age	1.00	0.99 – 1.00	0.113
A:sexfemale	0.88	0.80 – 0.95	**0.002**
C:O	0.97	0.93 – 1.00	0.082
C:age	0.99	0.99 – 1.00	**0.009**
C:sexfemale	1.08	1.00 – 1.17	0.051
E:sexfemale	1.05	0.99 – 1.11	0.103
age:sexfemale	1.01	1.00 – 1.02	**0.048**

Model Comparison Results (ordered by BIC):

model	AIC	BIC	df
step_bic	2187.981	2217.153	7
main_effects	2198.865	2232.205	8
step_aic	2174.808	2241.488	16
full_interactions	2193.584	2314.442	29

Selected Model: BIC-based Selection

I chose the BIC-selected model (step_bic) as the final model because:

- It has the lowest BIC (2217.153)
- The increase in AIC compared to the AIC-selected model (difference of ~13) is a reasonable trade-off for the substantial reduction in model complexity (**7 vs 16 parameters**)

Final Model Specification

$$\log(y_i) = \beta_0 + \beta_1 A_i + \beta_2 C_i + \beta_3 O_i + \beta_4 age_i + \beta_5 (A \cdot C)_i + \beta_6 (C \cdot age)_i + log\left(n_{observations_i}\right)$$

Equivalent to:

$$\log\left(\frac{y_i}{n_{observations_i}}\right) = \log(rate\ of\ future\ thoghts_i)$$
$$= \beta_0 + \beta_1 A_i + \beta_2 C_i + \beta_3 O_i + \beta_4 age_i + \beta_5 (A \cdot C)_i + \beta_6 (C \cdot age)_i$$

Assuming: $n\_future_i \sim Poisson(\mu_i)$

Where:

y_i is the expected number of future-oriented thoughts for participant i.

$A_i,\ C_i,\ and\ O_i$: Personality trait scores (Agreeableness, Conscientiousness, Openness).

age_i : Participant's age.

$(A \cdot C)_i$: Interaction term between Agreeableness and Conscientiousness.

$(C \cdot age)_i$: Interaction term between Conscientiousness and age.

$log\left(n_{observations_i}\right)$: Offset term accounting for differences in the total number of observations per participant.

(*) See the appendix for the full result of summary() function.

Interpretation of coefficients

Intercept ($\boldsymbol{\beta_0}$)

$\beta_0 = -1.12,\ p = 0.033,\ e^{-1.12} = 0.33$. Thus, the expected rate of future thoughts when all other variables are zero is 0.33 (not interpretable for age).

Agreeableness ($\boldsymbol{\beta_1}$)

$\beta_1 = -0.23,\ p = 0.007,\ e^{-0.23} = 0.79$. Thus, a unit increase in Agreeableness (where all other variables remain the same) is associated with a decrease in the rate of future thoughts by a factor of 0.79. (21% decrease).

Conscientiousness ($\boldsymbol{\beta 2}$)

Not statistically significant ($\beta_2 = -0.094, p = 0.30$).

Openness ($\boldsymbol{\beta_3}$)

$\beta_3 = 0.062,\ p = 0.005,\ e^{0.062} = 1.06$: Thus, a unit increase in Openness (where all other variables remain the same) is associated with an increase in the rate of future thoughts by a factor of 1.06. (6% increase).

Age ($\boldsymbol{\beta_4}$)

$\beta_4 = 0.027,\ p = 0.021,\ e^{0.027} = 1.03$: Thus, a one year increase in age (where all other variables remain the same) is associated with an increase in the rate of future thoughts by a factor of 1.03. (3% increase).

Agreeableness*Conscientiousness ($\boldsymbol{\beta_5}$)

$\beta_5 = 0.046, p = 0.003, e^{0.046}$: Thus, a 1-unit increase in both Agreeableness and Conscientiousness (where all other variables remain the same) is associated with an increase in the rate of future thoughts by a factor of 1.05. (5% increase).

Conscientiousness*Age ($\boldsymbol{\beta_6}$)

$\beta 6 = -0.005, p = 0.009, e^{-0.005} = 0.99$ Thus, an additional year and a 1-unit increase in Conscientiousness (where all other variables remain the same) is associated with a decrease in the rate of future thoughts by a factor of 0.99. (1% decrease).

Overdispersion

Testing for overdispersion with Cameron and Trivedi Test implemented by dispersiontest() function from AER library.

Cameron and Trivedi Test

- Tests if $Var(Yi) = \mu i + \alpha g(\mu i)$ where $g(\mu i)$ is a known function. I chose, g=1, g(μ) = μ).

- H_0: α = 0 (no overdispersion)
- H_1: α > 0 (overdispersion exists)
- Results: $p - value = 5.073e - 06,\ \hat{\alpha} = 0.4745701.$
- Given the significant p-value (5.073e-06), we reject H_0 and conclude there is strong evidence of overdispersion in the data.
- This means the estimated variance is approximately: $Var(Yi) = \mu i + 0.475\mu_i$
- These results suggest that the variance is larger than what would be expected under a Poisson model. This justifies using either a quasi-Poisson or negative binomial model to account for the extra variation in the data.

Refitting the same model with the "quasipoisson" family

Refitted the same model with same offset but with the "quasipoisson" family
The dispersion parameter = 1.536886

Effects on model parameters:

Coefficient estimates:

- Remained exactly the same as the Poisson model, because both models use the same estimating equations.

Standard errors:

- All standard errors increased by a factor of $\sqrt{1.536886} \approx 1.24$
- Example: Intercept SE changed from 0.525998 to 0.652086

Changes in Significance:

- Reduced statistical significance across all coefficients, with 'intercept' and 'age' becoming marginally significant (.) rather than significant (*).

- Original → Quasi-Poisson changes:
  - Intercept: p=0.03325 (*) → p=0.0866 (.)
  - A: p=0.00738 (**) → p=0.0312 (*)
  - C: p=0.30239 → p=0.4059 (remained non-significant)
  - O: p=0.00491 (**) → p=0.0237 (*)
  - age: p=0.02140 (*) → p=0.0641 (.)
  - A:C: p=0.00268 (**) → p=0.0158 (*)

- C:age: p=0.00948 (**) → p=0.0369 (*)

So while the coefficients don't change, our confidence in these estimates (as reflected in standard errors and p-values) is adjusted to account for the overdispersion in the data.

(**) See the appendix for the full result of summary() function.

Refitting with Negative Binomial Regression

Refitted the same model with same offset but with negative binomial regression, implemented by glm.nb() from the MASS library.

Differences between the Poisson model and the Negative Binomial model:

Treatment of Overdispersion

- Poisson: Assumes variance equals mean
- Negative Binomial: Variance = $\mu + \mu^2/\theta$ (estimated $\theta = 7.13$)

2. Changes in Statistical Inference

- Coefficient estimates changed slightly but mostly remain nearly identical. (***) See the appendix for the full result of summary() function.
- Standard errors increase in negative binomial models compared to Poisson
- Several variables show reduced significance levels in the adjusted models:
  - Intercept: Significant → marginally significant
  - age: Significant → marginally significant
  - Other coefficients retain significance but with larger p-values

3. Model Fit

AIC (Akaike Information Criterion)

- Poisson AIC: 2188
- Negative Binomial AIC: 2137.4

BIC (Bayesian Information Criterion)

- Poisson BIC: 2217.153

- Negative Binomial BIC: 2170.704

Both AIC and BIC suggest the negative binomial model provides a better fit, with lower values indicating better fit.

Appendix

(*) Full summary of the final model in question 2 section b

```
glm(formula = n_future ~ A + C + O + age + A:C + C:age + offset(log(n_observations)),
    family = poisson, data = subject_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3711  -0.9213  -0.0807   0.7357   3.7481

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.119905   0.525998  -2.129  0.03325 *
A           -0.230383   0.085984  -2.679  0.00738 **
C           -0.094120   0.091261  -1.031  0.30239
O            0.062017   0.022047   2.813  0.00491 **
age          0.027160   0.011804   2.301  0.02140 *
A:C          0.046111   0.015360   3.002  0.00268 **
C:age       -0.005346   0.002061  -2.594  0.00948 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 820.00  on 476  degrees of freedom
Residual deviance: 796.26  on 470  degrees of freedom
AIC: 2188
```

(**) Full summary of the quasipoisson model in question 2 section c

```
glm(formula = n_future ~ A + C + O + age + A:C + C:age + offset(log(n_obser
vations)),
    family = quasipoisson, data = subject_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3711  -0.9213  -0.0807   0.7357   3.7481

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.119905   0.652086  -1.717   0.0866 .
A           -0.230383   0.106596  -2.161   0.0312 *
C           -0.094120   0.113137  -0.832   0.4059
O            0.062017   0.027332   2.269   0.0237 *
age          0.027160   0.014634   1.856   0.0641 .
A:C          0.046111   0.019042   2.422   0.0158 *
C:age       -0.005346   0.002555  -2.093   0.0369 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 1.536886)

    Null deviance: 820.00  on 476  degrees of freedom
Residual deviance: 796.26  on 470  degrees of freedom
AIC: NA
```

(***) Full summary of the negative binomial model in question 2 section d

```
Call:
MASS::glm.nb(formula = n_future ~ A + C + O + age + A:C + C:age +
    offset(log(n_observations)), data = subject_data, init.theta = 7.132560429,
    link = log)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.91306  -0.75849  -0.08283   0.56334   2.64780

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.116176   0.672563  -1.660   0.0970 .
A           -0.231026   0.111230  -2.077   0.0378 *
C           -0.092713   0.116400  -0.797   0.4257
O            0.060272   0.027803   2.168   0.0302 *
age          0.026666   0.015277   1.746   0.0809 .
A:C          0.046876   0.019831   2.364   0.0181 *
C:age       -0.005349   0.002653  -2.016   0.0438 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(7.1326) family taken to be 1)

    Null deviance: 557.07  on 476  degrees of freedom
Residual deviance: 541.75  on 470  degrees of freedom
AIC: 2137.4

Number of Fisher Scoring iterations: 1


              Theta:  7.13
          Std. Err.:  1.37

 2 x log-likelihood:  -2121.364
```

R code

```r
rm(list = ls())
library(dplyr)
library(ggplot2)
library(tidyr)
library(tidyverse)
library(knitr)
library(Hmisc)
library(lattice)
library(lme4)
library(MASS)
############################## Load Data ##############################

# code from moodle to get data
require(Hmisc)
esm_data = spss.get("Data/ETT_ESM_Study1.sav", use.value.labels=TRUE)

# Plot of binary indicator for thinking of future, versus time of day,
# for first four subjects and all three days
require(lattice)
xyplot(future ~ TIME.S | DAY+Subject, esm_data[esm_data$Subject %in% unique(e
sm_data$Subject)[1:4],],
       xlab = "Time of day", ylab = "Thinking of future?")

# view the structure of the data frame
str(esm_data)

# not really clear, so i will choose the relevant columns then continue
df_q1 <- esm_data %>%
  mutate(
    Subject = as.factor(Subject),
    DAY = as.factor(DAY),
    future = as.numeric(future),
    TIME.S = as.numeric(TIME.S)
  )
# view the structure of the data frame with the choosen cols
str(df_q1)


# num of subs
n_subjects <- df_q1 %>%
  summarise(n_unique_subjects = n_distinct(Subject))
n_subjects

################################## Q1 ##################################
```

```r
################################## Missing data ##################################
##

#### Missing data
colSums(is.na(df_q1))

#  -> no missing data in any variable
# but we cant see this unless we have a "wide format"

num_missing_obs = 492*18 - nrow(df_q1)
num_missing_obs

num_missing_obs_p = (492*18 - nrow(df_q1)) / (492*18)
num_missing_obs_p

# check observations across days for each subject
incomplete_days_subs <- df_q1 %>%
  group_by(Subject, DAY) %>%
  summarise(obs_per_day = n(), .groups = 'drop') %>%
  group_by(Subject) %>%
  summarise(
    incomplete_days = sum(obs_per_day != 6),
    has_incomplete = any(obs_per_day != 6)
  )
head(incomplete_days_subs)
# number of incomplete subjects (who have at least one day with < 6 observati
ons)
incomplete_subs <- incomplete_days_subs %>%
  summarise(
    n_incomplete = sum(has_incomplete),
    percent_incomplete = mean(has_incomplete) * 100
  )
incomplete_subs

# total num of observations per sub
subs_tot_observations <- df_q1 %>%
  group_by(Subject) %>%
  summarise(
    n_observations = n(),
    .groups = 'drop'
  )

# stats of total num of observations per sub
subs_tot_observations_stats  <- subs_tot_observations %>%
  summarise(
    mean_obs = mean(n_observations),
    median_obs = median(n_observations),
    min_obs = min(n_observations),
    max_obs = max(n_observations),
    sd_obs = sd(n_observations)
```

```r
  )
subs_tot_observations_stats
# -> some subs have a very low number of observations (1 obs)

# frequency table of total num of observations per sub
obs_freq <- subs_tot_observations %>%
  count(n_observations) %>%
  mutate(proportion = n/sum(n)*100)
obs_freq
# -> 12 subs have one obs, 15 subs have 2 obs

# histogram of total num of observations per sub
ggplot(subs_tot_observations, aes(x = n_observations)) +
  geom_histogram(fill = "skyblue", color = "black") +
  labs(x = "Total Number of Observaions", y = "Frequency") +
  theme_minimal()

# compare missing observations across days
missing_per_day <- df_q1 %>%
  group_by(DAY) %>%
  summarise(
    total_expected = 487 * 6,  # Using consistent subject count
    actual_obs = n(),
    missing_obs = total_expected - actual_obs,
    missing_rate = (missing_obs/total_expected) * 100
  ) %>%
  arrange(DAY)
missing_per_day

# check missing data patterns

# create a grid for all combinations of day and sub
all_combinations <- expand_grid(
  Subject = unique(df_q1$Subject),
  DAY = unique(df_q1$DAY),
  Obs = 1:6  # Six observations per day
)

# add an observation number to the original data
df_q1 <- df_q1 %>%
  group_by(Subject, DAY) %>%
  mutate(Obs = row_number()) %>%
  ungroup()

# join the original data with the complete grid
expanded_df <- all_combinations %>%
  left_join(df_q1, by = c("Subject", "DAY", "Obs"))

# fill missing in time and future values with NA
expanded_df <- expanded_df %>%
```

```r
  mutate(
    TIME.S = ifelse(is.na(TIME.S), NA, TIME.S),
    future = ifelse(is.na(future), NA, future)
  )

# check if each sub and day has 6 observations
validation <- expanded_df %>%
  group_by(Subject, DAY) %>%
  summarise(n_obs = n(), .groups = "drop") %>%
  filter(n_obs != 6)

if (nrow(validation) == 0) {
  print("All subjects have 6 observations per day.")
}

# check if missing num is correct
colSums(is.na(expanded_df))
# view the expanded dataframe
head(expanded_df)

mar_data <- expanded_df %>%
  dplyr::select(Subject, DAY, future) %>%
  mutate(missing_future = is.na(future))

# test for MAR in long format
mar_test <- glm(missing_future ~ DAY , family = binomial,
                      data = mar_data)

summary(mar_test)

# missingness by day
ggplot(mar_data, aes(x = DAY, fill = missing_future)) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Missing future", y = "Proportion", fill = "Miss
ing future")

################################## Stats ###################################
#### Stats

#### DAY and future
# Frequency table for DAY
day_freq <- df_q1 %>%
  count(DAY) %>%
  mutate(proportion = n/sum(n))
day_freq

# Frequency table for future thinking
future_freq <- df_q1 %>%
  count(future) %>%
  mutate(proportion = n/sum(n))
```

```r
future_freq

future_day_summary <- df_q1 %>%
  group_by(DAY, future) %>%
  summarise(
    count = n(),
    .groups = 'drop'
  ) %>%
  group_by(DAY) %>%
  mutate(
    proportion = count / sum(count),
    percentage = round(proportion * 100, 1)
  ) %>%
  arrange(DAY, future)
future_day_summary

# grouped bar plot
ggplot(future_day_summary, aes(x = DAY, y = proportion, fill = factor(future)
)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7) +
  scale_fill_manual(values = c("lightblue", "darkblue"),
                    name = "Future Thinking",
                    labels = c("No", "Yes")) +
  labs(
    title = "Future-Oriented Thinking by Day",
    x = "Day",
    y = "Proportion"
  ) +
  theme_minimal()

#### Time

time_stats <- df_q1 %>%
  summarise(
    min_hours = min(TIME.S)/3600,
    max_hours = max(TIME.S)/3600,
    mean_hours = mean(TIME.S)/3600,
    median_hours = median(TIME.S)/3600,
    var_hours = var(TIME.S)/(3600^2)
  )
time_stats

# histogram of time in hours
ggplot(df_q1, aes(x = TIME.S/3600)) +
  geom_histogram(bins = 24, fill = "skyblue", color = "black") +
  labs(x = "Time (Hours)", y = "Frequency") +
  theme_minimal()

# time by day
time_by_day_stats <- df_q1 %>%
```

```r
  group_by(DAY) %>%
  summarise(
    min_hours = min(TIME.S)/3600,
    max_hours = max(TIME.S)/3600,
    mean_hours = mean(TIME.S)/3600,
    median_hours = median(TIME.S)/3600,
    var_hours = var(TIME.S)/(3600^2)
  )
time_by_day_stats

ggplot(df_q1, aes(x = TIME.S/3600)) +
  geom_histogram(bins = 24, fill = "skyblue", color = "black") +
  facet_wrap(~DAY, labeller = labeller(DAY = function(x) paste("Day", x))) +
  labs(x = "Time (Hours)", y = "Frequency", title = "Histogram of Time by Day
") +
  theme_minimal()

# time by future
time_by_fut_stats <- df_q1 %>%
  group_by(future) %>%
  summarise(
    min_hours = min(TIME.S)/3600,
    max_hours = max(TIME.S)/3600,
    mean_hours = mean(TIME.S)/3600,
    median_hours = median(TIME.S)/3600,
    var_hours = var(TIME.S)/(3600^2)
  )
time_by_fut_stats

ggplot(df_q1, aes(x = TIME.S/3600, fill = factor(future))) +
  geom_density(color = "black", alpha = 0.6) +
  scale_fill_manual(values = c("lightblue", "darkblue"),
                    name = "Future Thinking",
                    labels = c("No", "Yes")) +
  labs(x = "Time (Hours)", y = "Density", title = "Density of Time by Future"
) +
  theme_minimal()


# time by future and day
time_by_day_by_future_stats <- df_q1 %>%
  group_by(DAY, future) %>%
  summarise(
    min_hours = min(TIME.S)/3600,
    max_hours = max(TIME.S)/3600,
    mean_hours = mean(TIME.S)/3600,
    median_hours = median(TIME.S)/3600,
    var_hours = var(TIME.S)/(3600^2)
  )
time_by_day_by_future_stats
```

```r
ggplot(df_q1, aes(x = TIME.S/3600, fill = factor(future))) +
  geom_density(alpha = 0.7) +
  facet_wrap(~DAY, labeller = labeller(DAY = function(x) paste("Day", x))) +
  scale_fill_manual(values = c("lightblue", "darkblue"),
                    name = "Future Thinking",
                    labels = c("No", "Yes")) +
  labs(x = "Time (Hours)", y = "Density", title = "Distribution of Time by Da
y and Future Thinking") +
  theme_minimal()


################################## Logistic Reg ################################
##

# conver time to hours
df_q1$TIME.H <- df_q1$TIME.S/3600

# fit simple logistic regression
simple_model <- glm(future ~ TIME.H, family = binomial, data = df_q1)

# fit mixed effects model
mixed_model <- glmer(future ~ TIME.H + (1|Subject/DAY),
                     family = binomial, data = df_q1)

# model comparisons
# summary of both models
summary(simple_model)
summary(mixed_model)

anova(mixed_model, simple_model,  test = "Chisq")

# compare AIC values
aic_comparison <- data.frame(
  Model = c("Simple Logistic", "Mixed Effects"),
  AIC = c(AIC(simple_model), AIC(mixed_model))
)
print(aic_comparison)

# 3. compare fixed effects coefs
fixed_effects <- data.frame(
  Model = c("Simple Logistic", "Mixed Effects"),
  Intercept = c(coef(simple_model)[1], fixef(mixed_model)[1]),
  Time_Coefficient = c(coef(simple_model)[2], fixef(mixed_model)[2])
)
print(fixed_effects)

# 4. random effects variance from mixed model
print("Random Effects Variance:")
print(VarCorr(mixed_model))
```

```r
# 5. calculate R-squared
# for simple model
r2_simple <- 1 - simple_model$deviance/simple_model$null.deviance

# for mixed model using MuMIn package
# R² for mixed model calculated two ways:
# 1. Marginal R² - variance explained by fixed effects only:
#    var(fixed)/(var(fixed) + var(random) + var(residual))
# 2. Conditional R² - variance explained by entire model:
#    (var(fixed) + var(random))/(var(fixed) + var(random) + var(residual))
library(MuMIn)
r2_mixed <- r.squaredGLMM(mixed_model)
r2_mixed

r2_comparison <- data.frame(
  Model = c("Simple Logistic", "Mixed Effects (Marginal)", "Mixed Effects (Co
nditional)"),
  R_squared = c(r2_simple, r2_mixed[1], r2_mixed[3])
)
print(r2_comparison)

report::report(simple_model)
report::report(mixed_model)

################################## Q2 ##################################

# make sure traits are constant for subs
cols <- c("A", "O", "C", "E", "N", "age", "sex")
for (col in cols){
  trait_variation <- esm_data %>%
    group_by(Subject) %>%
    summarise(
      unique_values = n_distinct(col, na.rm = TRUE),
      has_variation = unique_values > 1
    )
  cat(paste("variation in col ", col, " "))
  cat(any(trait_variation$has_variation), "\n")
}
## -> all values are ok, they are in fact fixed vars

# create subject-level dataset
subject_data <- esm_data %>%
  group_by(Subject) %>%
  summarise(
    n_observations = n(),
    n_future = sum(future),
    # take the first occurrence of these variables
    # since they're constant per subject
    age = first(age),
```

```r
    sex = first(sex),
    A = first(A),
    C = first(C),
    E = first(E),
    N = first(N),
    O = first(O)
  )
head(subject_data)

################################# Missing data ################################
##
# check for missing data
colSums(is.na(subject_data))

# count missing variables per subject
missing_per_subject <- data.frame(
  Subject = subject_data$Subject,
  num_missing = rowSums(is.na(subject_data[, c("A", "C", "E", "N", "O")]))
) %>%
  arrange(desc(num_missing))
max_missing <- max(missing_per_subject$num_missing)
max_missing
print(table(missing_per_subject$num_missing))

# remove subjects with all 5 trait variables missing
subject_data_clean <- subject_data %>%
  filter(!Subject %in% missing_per_subject$Subject[missing_per_subject$num_mi
ssing == 5])

cat("Removed", sum(missing_per_subject$num_missing == 5),
    "subjects with all variables missing\n")
cat("Remaining subjects:", nrow(subject_data_clean), "\n")

# check the new distribution of missing values
missing_per_subject_new <- data.frame(
  Subject = subject_data_clean$Subject,
  num_missing = rowSums(is.na(subject_data_clean[, c("age", "sex", "A", "C",
"E", "N", "O")]))
)
# print new frequency table
print(table(missing_per_subject_new$num_missing))

subject_data <- subject_data_clean
colSums(is.na(subject_data))

################################# Stats ###################################
numeric_cols <- c("A", "O", "C", "E", "N", "age", "n_observations", "n_future
")
numeric_cols_names <- c(
  "A" = "Agreeableness",
```

```r
  "C" = "Conscientiousness",
  "E" = "Extraversion",
  "N" = "Neuroticism",
  "O" = "Openness",
  "age" = "Age",
  "n_observations" = "Number of observations",
  "n_future" = "Number of future thoughts"
)
# numeric descriptive statistics
for (var in numeric_cols){
  x <- subject_data[[var]]
  # basic stats
  cat("\n")
  stats <- data.frame(
    Variable = var,
    n_obs = length(x),
    n_missing = sum(is.na(x)),
    mean = mean(x, na.rm = TRUE),
    sd = sd(x, na.rm = TRUE),
    median = median(x, na.rm = TRUE),
    min = min(x, na.rm = TRUE),
    max = max(x, na.rm = TRUE)
  )
  print(stats)
  cat("\n")
}

library(moments)
skewness(subject_data$n_future)

sex_freq <- subject_data %>%
  count(sex) %>%
  mutate(proportion = n/sum(n))
sex_freq

## histograms
# create plots list
plots <- list()
for(var in numeric_cols) {
  plots[[var]] <- ggplot(subject_data, aes(x = .data[[var]])) +
    geom_histogram(fill = "lightblue", color = "black", bins = 30) +
    labs(title =  numeric_cols_names[[var]],
         y = "Frequency",
         x = var) +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5))
}

# Print all plots
for(p in plots) {
```

```r
  print(p)
}

library(gridExtra)
combined_plot <- grid.arrange(
  plots$A, plots$C, plots$E,
  plots$N, plots$O,plots$age,
  ncol = 3
)
combined_plot

## Numeric variables by sex histograms
# Create plots list
plots <- list()
for(var in numeric_cols) {
  plots[[var]] <- ggplot(subject_data, aes(x = .data[[var]], fill = sex,)) +
    geom_density(alpha=0.6) +
    labs(title = numeric_cols_names[[var]],
         y = "Density",
         x = var) +
    scale_fill_manual(values = c("lightblue", "darkblue"),
                      name = "Sex",
                      labels = c("male", "female")) +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5))
}

# Print all plots
for(p in plots) {
  print(p)
}

library(gridExtra)
combined_plot <- grid.arrange(
  plots$A, plots$C, plots$E, plots$N,
  plots$O,plots$age, plots$n_observations, plots$n_future,
  ncol = 4
)
combined_plot

## Numeric x Numeric
# correlation plot matrix using GGally
library(GGally)
predictors <- c("A", "O", "C", "E", "N", "age")
pairs_plot <- ggpairs(subject_data[predictors],
                      lower = list(continuous = wrap("smooth",
                                                     alpha = 0.3,
                                                     color = "lightblue")),
                      upper = list(continuous = wrap("cor", size = 5))) +
  theme_minimal() +
```

```r
  labs(title = "Correlation Matrix Plot")
pairs_plot
ggsave("pairs_plot.png", pairs_plot, width = 12, height = 8, dpi = 300)

# Loop over pairs of traits
for (i in 1:(length(predictors)-1)) {
  for (j in (i+1):length(predictors)) {
    test <- cor.test(subject_data[[predictors[i]]],
                     subject_data[[predictors[j]]],
                     use = "pairwise.complete.obs")

    cat("\nCorrelation between", predictors[i], "and", numeric_cols[j], ":\n"
)
    cat("r =", round(test$estimate, 3), "\n")
    cat("p-value =", round(test$p.value, 10), "\n")
  }
}

################################ Poisson Reg ################################
#
# Define variables to model

predictors <- c("A", "C", "E", "N", "O", "age", "sex")

models <- list()
results <- data.frame(
  variable = predictors,
  coefficient = NA,
  std_error = NA,
  p_value = NA
)

# fit models in a loop
for(i in seq_along(predictors)) {
  # Create formula with offset
  formula <- as.formula(paste("n_future ~", predictors[i], "+ offset(log(n_ob
servations))"))

  models[[i]] <- glm(formula, family = poisson(link = "log"), data = subject_
data)

  coef_summary <- summary(models[[i]])$coefficients[2,]
  results$coefficient[i] <- coef_summary[1]
  results$std_error[i] <- coef_summary[2]
  results$p_value[i] <- coef_summary[4]
}

#  absolute coefficient column
results$abs_coef <- abs(results$coefficient)
```

```r
# significance indicators
results$significance <- ifelse(results$p_value < 0.001, "***",
                               ifelse(results$p_value < 0.01, "**",
                                      ifelse(results$p_value < 0.05, "*", "ns
")))

# Create formatted output sorted by p-value
formatted_results <- results %>%
  mutate(
    coefficient = round(coefficient, 3),
    std_error = round(std_error, 3),
    p_value = format.pval(p_value, digits = 3),
    abs_coef = round(abs_coef, 3)
  ) %>%
  arrange(p_value)

# Print results
cat("\nResults sorted by p-value:\n")
print(formatted_results)

# Print results sorted by absolute coefficient value
cat("\nResults sorted by absolute coefficient size:\n")
print(formatted_results[order(formatted_results$abs_coef, decreasing = TRUE),
])

test <- formatted_results[order(formatted_results$abs_coef, decreasing = TRUE
), ]
png("test.png",  width  = 8, height= 5,
    units= "in", res = 1200)
grid.table(test)
dev.off()

############################## Model Selection ######################
n <- nrow(subject_data)

# basic model with all main effects
model_all <- glm(n_future ~ A + C + E + N + O + age + sex +
                   offset(log(n_observations)),
                 family=poisson, data=subject_data)

# full model with all two way interactions
model_all_int <- glm(n_future ~ (A + C + E + N + O + age + sex)^2 +
                       offset(log(n_observations)),
                     family=poisson, data=subject_data)

# stepwise selection with AIC
step_aic <- stepAIC(model_all_int, direction="both", trace=FALSE)

# stepwise selection with BIC
step_bic <- stepAIC(model_all_int, direction="both", k=log(n), trace=FALSE)
```

```r
models <- list(
  main_effects = model_all,
  full_interactions = model_all_int,
  step_aic = step_aic,
  step_bic = step_bic
)

# comparison
metrics <- data.frame(
  model = names(models),
  AIC = sapply(models, AIC),
  BIC = sapply(models, BIC),
  df = sapply(models, function(x) length(coef(x))),
  dispersion = sapply(models, function(x) {
    pearson_chi2 <- sum(residuals(x, type="pearson")^2)
    df <- x$df.residual
    return(pearson_chi2/df)
  })
)

# model comparison
cat("\nModel Comparison (sorted by BIC):\n")
print(metrics[order(metrics$BIC),])

png("BIC.png",  width  = 8, height= 5,
    units= "in", res = 1200)
grid.table(metrics[order(metrics$BIC),])
dev.off()

# Get best model (using BIC as criterion)
best_model_bic <- models[[which.min(metrics$BIC)]]
best_model_aic <- models[[which.min(metrics$AIC)]]

sjPlot::tab_model(best_model_bic)
sjPlot::tab_model(best_model_aic)

best_model <- best_model_bic

print(summary(best_model))

print(anova(model_all, best_model, test="Chisq"))


############ Over-dispersion C + D ##########################
# the Pearson Chi-Square Dispersion
pearson_chi2 <- sum(residuals(best_model, type="pearson")^2)
df <- best_model$df.residual
dispersion <- pearson_chi2/df
```

```r
library(AER)
dispersiontest(best_model)


# Or alternatively using:
qp_model <- glm(formula = n_future ~ A + C + O + age + A:C + C:age + offset(l
og(n_observations)),
                family = quasipoisson, data = subject_data)
summary(qp_model)
summary(qp_model)$dispersion

# fit negative binomial model
nb_model <- MASS::glm.nb(n_future ~ A + C + O + age + A:C + C:age + offset(lo
g(n_observations)), data = subject_data)
summary(nb_model)

poisson_bic <- BIC(best_model)
poisson_bic
nb_bic <- BIC(nb_model)
nb_bic
```