Assignment 2 - Report

Eleni Liarou, Zoë Azra Blei, Frederieke Loth, group 20

2025-03-09

Exercise 1: Titanic

Section a

Section b

Section c

Section d

Section e

Exercise 2: Military Coups

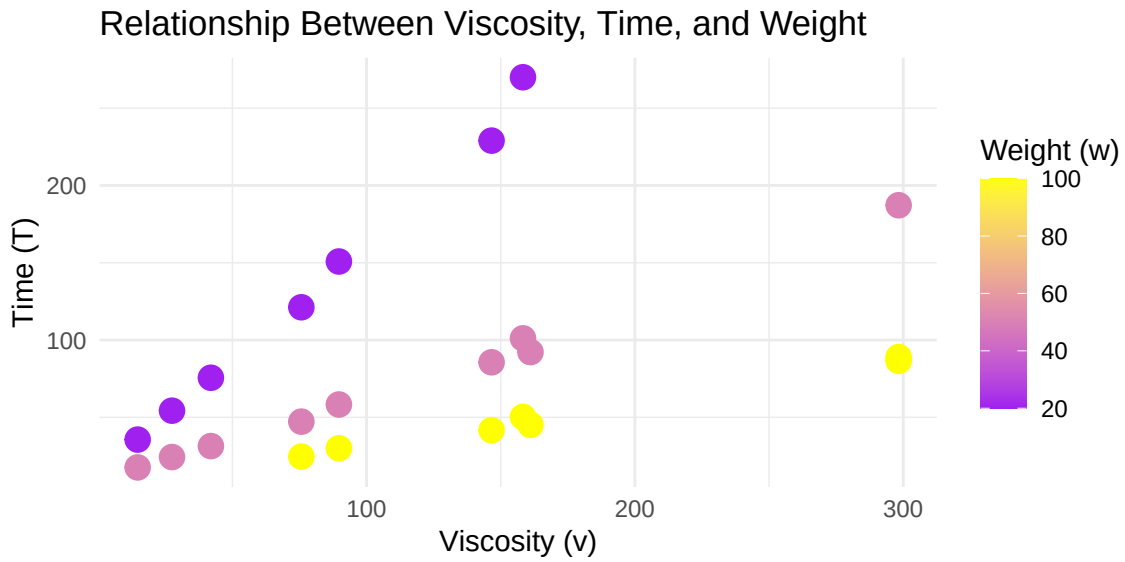
Section a

Section b

Section c

Exercise 3: Stormer viscometer

Section a To understand the relationships between the variables, a scatterplot of the data is plotted below:



The scatterplot visualizes that **higher weight values correspond to lower time values**, demonstrating an inverse relationship between weight and time. Additionally, the pattern of points suggests a **nonlinear relationship** between viscosity and time, supporting the theoretical nonlinear model:

$$T = \frac{\theta_1 v}{w - \theta_2} + e$$

However, in order to estimate the parameters θ_1 and θ_2, since the theoretical model can be rewritten to a linear form, we can first apply linear regression to obtain initial estimates using the following formula:

$$wT = \theta_1 v + \theta_2 T + (w - \theta_2)e$$

As the variance of the error term is not constant, we have to take into account heteroscedasticity. For this, the variance of wT becomes the following:

$$Var(wT) = \sigma^2 (w - \theta_2)^2$$

We use Weighted Least Squares and set the weights in the regression as, using an initial guess for θ_2 to do the linear regression:

$$w_i = \frac{1}{(w_i - \hat{\theta}_2)^2}$$

```
theta2_init = mean(stormer$Wt)
weights_wls = 1 / (stormer$Wt - theta2_init)^2
linear_model_wls = lm(Wt * Time ~ Viscosity + Time, data = stormer, weights = weights_wls)
summary(linear_model_wls)
```

```
##
## Call:
## lm(formula = Wt * Time ~ Viscosity + Time, data = stormer, weights = weights_wls)
##
```

```
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -58.565 -10.645  -2.313   7.604  44.115
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  222.569     79.259   2.808   0.0109 *
## Viscosity     26.499      1.694  15.642 1.11e-12 ***
## Time           5.150      2.774   1.857   0.0781 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.96 on 20 degrees of freedom
## Multiple R-squared:  0.9932, Adjusted R-squared:  0.9926
## F-statistic:  1471 on 2 and 20 DF,  p-value: < 2.2e-16
```

The found estimated values are $\theta_1 = 26.499$ and $\theta_2 = 5.150$, where only the value of θ_1 is statistically significant. Using these estimated values, we can do nonlinear regression:

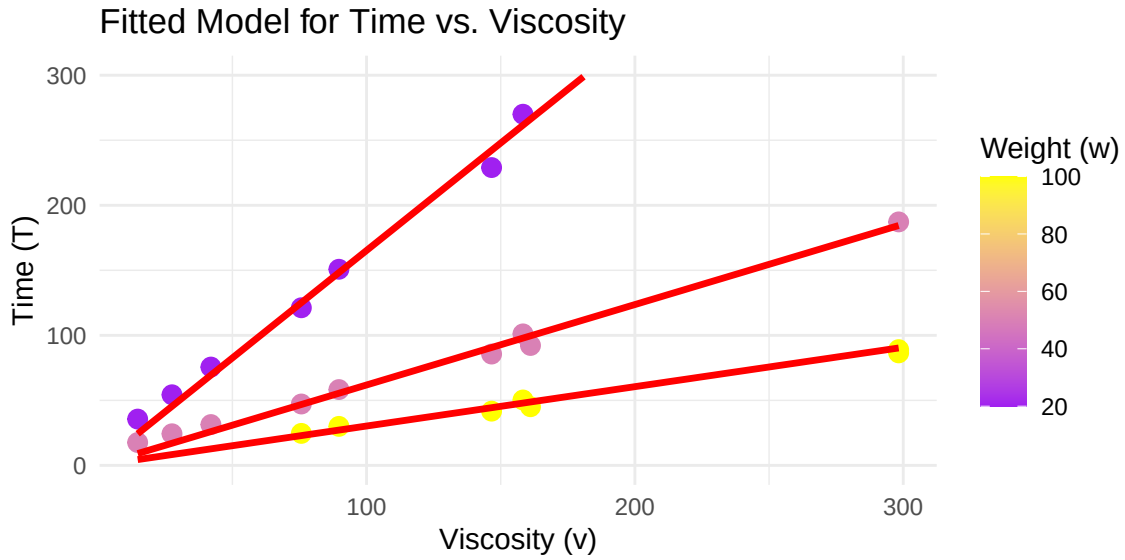
```
theta1_wls <- coef(linear_model_wls)["Viscosity"]
theta2_wls <- coef(linear_model_wls)["Time"]
nls_model_weighted <- nls(Time ~ (theta1 * Viscosity) / (Wt - theta2),
                          data = stormer,
                          start = list(theta1 = theta1_wls, theta2 = theta2_wls),
                          weights = 1 / (Wt - theta2_wls)^2)

summary(nls_model_weighted)
```

```
##
## Formula: Time ~ (theta1 * Viscosity)/(Wt - theta2)
##
## Parameters:
##        Estimate Std. Error t value Pr(>|t|)
## theta1   29.637      2.624  11.295 2.21e-10 ***
## theta2    2.065      1.620   1.275    0.216
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3424 on 21 degrees of freedom
##
## Number of iterations to convergence: 3
## Achieved convergence tolerance: 2.105e-06
```

The final estimated values are $\theta_1 = 29.637$ and $\theta_2 = 2.065$, where only θ_1 seems statistically significant. Additionally, the residual standard error is smaller (0.342) compared to the linear model (22.96) indicating that this model has much less unexplained variation.

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: Removed 41 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



The final plot shows that the nonlinear regression model fits the data well, capturing the expected trend between viscosity and time while accounting for different weight values. The fitted curves align closely with the data points, confirming that the nonlinear model is more appropriate than a linear model.However, the separation of curves suggests some variation in the effect of weight, and θ_2 was not statistically significant, indicating potential refinements in the model.

Section b A two-tailed t-test is conducted, as we have no expectation about whether θ_1 will be greater or smaller than 25; it could differ in either direction. We consider the null hypothesis H0: $\theta_1 = 25$. For the test, we use the estimated θ_1 and its standard error obtained from question a). The test resulted in a t-statistic of 4.81 and a p-value of 9.45e-05, which is far below the typical significance level of 0.05. This means we reject H0 and conclude that θ_1 is significantly different from 25, further supporting the nonlinear model's results.

```
theta1_hat = 29.4013  # Estimated parameter
theta1_se = 0.9155    # Standard error
theta1_h0 = 25        # Hypothesized value under H0
df = 21               # Degrees of freedom from nls summary

t_stat = (theta1_hat - theta1_h0) / theta1_se
p_value = 2 * pt(-abs(t_stat), df)
```

Test Statistic (t): 4.808

P-value: 9.454e-05

Section c For computing the 92% confidence interval for θ_1 and θ_2, we consider the following formula to calculate the z-value:

$$\hat{\theta} \pm z_{\alpha/2} \cdot SE(\theta)$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution. For a 92% confidence level, the significance level is $\alpha = 0.08$, thus:

$$z_{0.04/2} = z_{0.02} \approx 1.75$$

This gave a 92% CI for θ_1 of [27.80, 31.00] and for θ_2 of [1.05, 3.38], meaning we are 92% confident that the true values lie within these intervals. Since the confidence interval for θ_1 does not include 25, it further supports rejecting H0 from question b).

```
theta1_hat <- 29.4013  # Estimated 1
theta1_se <- 0.9155    # Standard error of 1
theta2_hat <- 2.2183   # Estimated 2
theta2_se <- 0.6655    # Standard error of 2

z_value = qnorm(0.96)  # 1.75

theta1_CI = c(theta1_hat - z_value * theta1_se, theta1_hat + z_value * theta1_se)
theta2_CI = c(theta2_hat - z_value * theta2_se, theta2_hat + z_value * theta2_se)
```

92% CI for 1: [27.799 31.004]

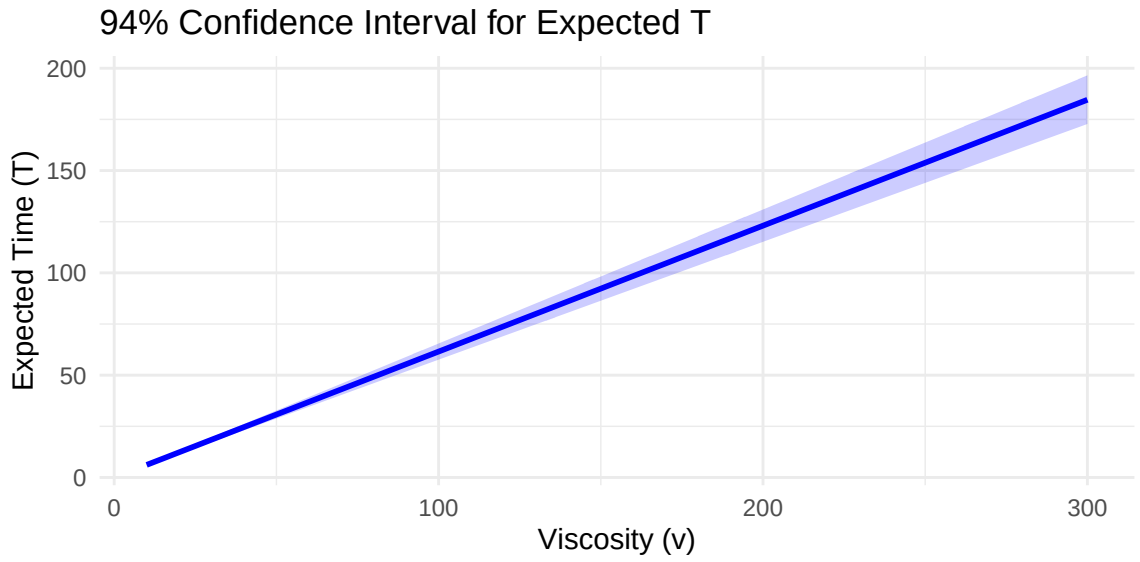
92% CI for 2: [1.053 3.383]

Section d The expected values are computed using the nonlinear model with $w = 50$, and viscosity values ranging from 10 to 300. The 94% confidence intervals were derived using asymptotic normality, where the standard error of T was estimated through error propagation. The confidence bounds were calculated as:

$$T(v) \pm z_{\alpha/2} \cdot SE(T)$$

where $z_{0.03} = 1.88$ is the critical value for a 94% confidence level. The plot shows the expected T along with a shaded confidence band, indicating the uncertainty in our estimates. The confidence interval widens as viscosity increases, reflecting greater uncertainty for larger v. The linear trend suggests a strong relationship between viscosity and time, but further diagnostics are needed to confirm model assumptions. Overall, the plot aligns well with the theoretical nonlinear model, supporting its validity over a simple linear approximation.

```r
theta1_hat = 29.4013  # Estimate of theta1
theta2_hat = 2.2183   # Estimate of theta2
theta1_se = 0.9155    # Standard error of theta1
theta2_se = 0.6655    # Standard error of theta2
w_fixed = 50
v_values = seq(10, 300, length.out = 100)

T_hat = (theta1_hat * v_values) / (w_fixed - theta2_hat)
T_se = sqrt(
  (v_values / (w_fixed - theta2_hat))^2 * theta1_se^2 +
  (theta1_hat * v_values / (w_fixed - theta2_hat)^2)^2 * theta2_se^2
)

z_value = qnorm(0.97)
T_lower = T_hat - z_value * T_se
T_upper = T_hat + z_value * T_se
```



Section e To investigate whether the smaller model with $\theta_1 = 25$ is appropriate, we compare it to the estimated model using hypothesis testing and model evaluation metrics. From question b), the T-test rejected H0: θ_1 =25 with a p-value of 9.45e-05, indicating that setting $\theta_1 = 25$ significantly deviates from the data. Additionally, the 92% confidence interval for θ_1 of [27.80, 31.00] does not contain 25, further supporting that the restriction is inappropriate. A likelihood ratio test could be performed to formally compare the smaller model to the unrestricted model, however, the hypothesis test already suggests a poor fit. Constraining θ_1 may lead to higher residual errors and reduced model flexibility, making the model less accurate. Given these findings, the smaller model does not seem appropriate, as it forces an assumption that contradicts the observed data. Therefore, the unrestricted nonlinear model remains the more valid choice.