Midterm Coursework 2025-2026
MATH48011/68011 Linear Models with Nonparametric Regression

Student ID 14199981

November 12, 2025

Introduction

This study analyses the effect of Semaglutide on walking capacity, focusing on the ratio of maximum walking distance after treatment to before treatment (MWDR) across different demographic groups. Using data from 280 participants who received weekly treatment for 59 weeks, a statistical model incorporating quadratic age effects and gender interaction was developed.

1. Exploratory Data Analysis

1.1 Relationship Between Age and MWDR

Figure 1.1 displays the relationship between age and maximum walking distance ratio for all participants. The quadratic trend suggests that treatment efficacy varies with age in a non-linear fashion. The ratio tends to increase initially with age, reaching a peak in middle age (around 40-45 years), and then gradually decreases for older participants.

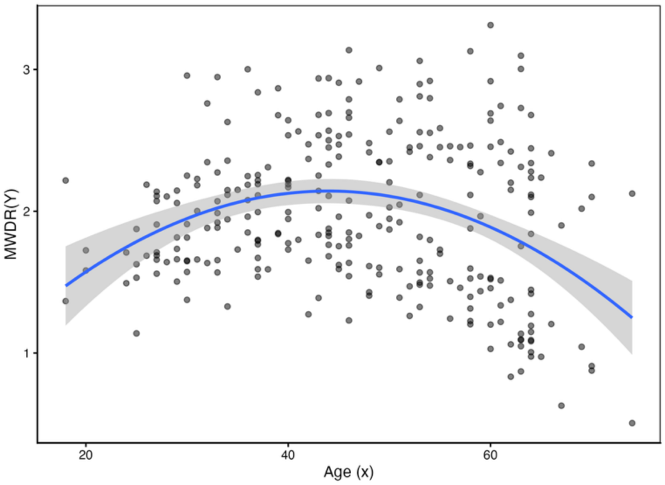

Figure 1.1: The relationship between Age and MWDR.

1.2 Distribution of MWDR by Sex

Figure 1.2 shows clear differences in treatment response between male and female participants. Males exhibit a higher median MWDR, approximately 2.4, compared to females, approximately 1.6, suggesting that on average, male participants experienced a greater improvement in walking capacity after Semaglutide treatment.

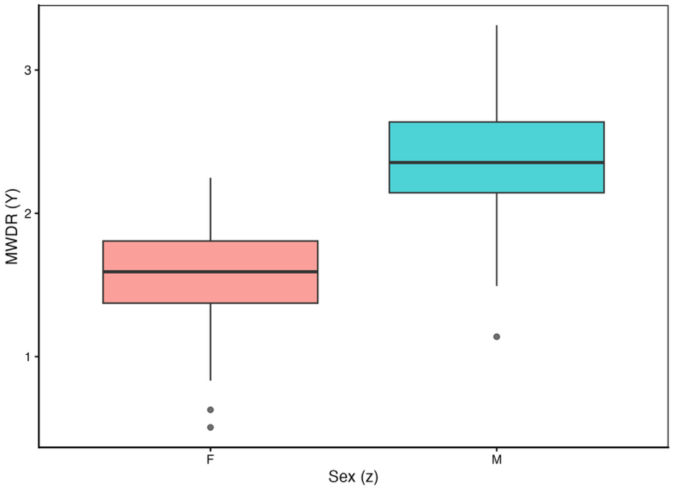

Figure 1.2: The distribution of maximum walking distance ratio by sex.

1.3 Age-Sex Interaction

Figure 1.3 reveals that both males and females show quadratic age trends. Males maintain a higher MWDR across most age groups, and their curve peaks at a slightly older age compared to females. The female trend line shows a steeper decline in older age groups, suggesting that the treatment efficacy may diminish more rapidly for older female participants.

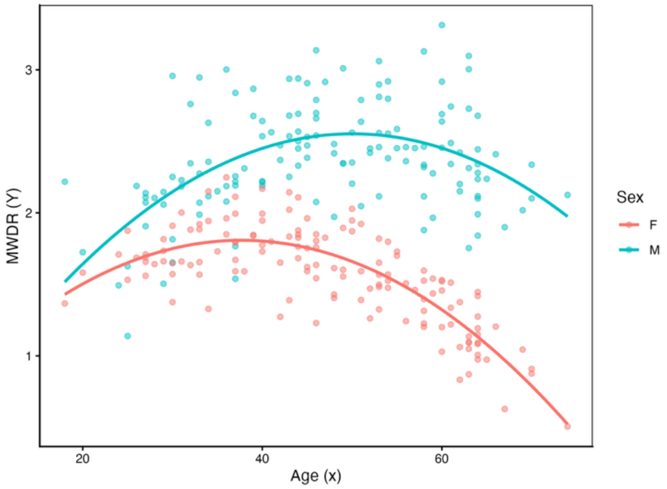

Figure 1.3: MWDR vs. age with respective quadratic trend lines for each sex.

2. Model Fitting

The model is specified as:

$$E[Y] = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 w + \theta_4 xw$$

where Y MWDR, x is age, and w is the dummy variable for sex (0 for male, 1 for female).

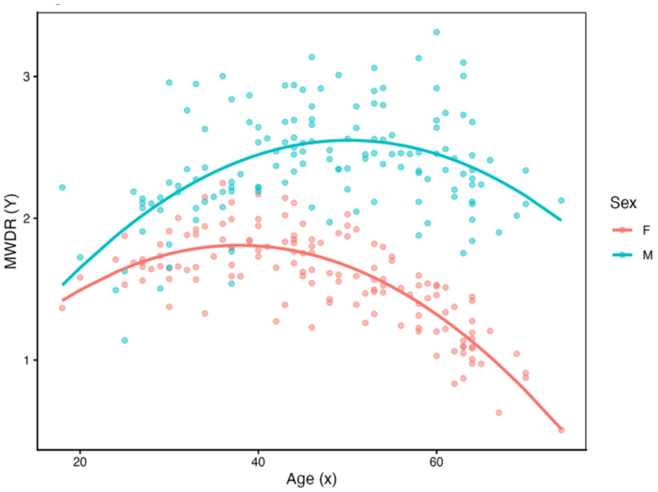

Figure 2.1: The relationship between age and MWDR for males and females.

Table 1: Estimated Coefficients for Model

Parameter	Estimate	Standard Error	t value	p value
Intercept	0.0599	0.2120	0.2830	7.77×10^{-1}
age	0.0994	0.0092	10.8000	5.90×10^{-23}
age^2	-0.0010	0.0001	-10.2000	5.96×10^{-21}
w	0.3300	0.1230	2.69--	7.61×10^{-3}
age:w	-0.0243	0.0025	-9.7000	2.51×10^{-19}

3. Parameter Interpretation

3.1 $\boldsymbol{\theta_0}$

θ_0, estimated as 0.0599, represents the regression intercept for males, corresponding to the expected MWDR at age 0.

3.2 $\boldsymbol{\theta_0 + \theta_3}$

$\theta_0 + \theta_3$, estimated as 0.3899, represents the regression intercept for females. The negative estimate of θ_3 indicates that, at baseline, females exhibit a lower MWDR compared to males, implying a comparatively reduced initial response to Semaglutide treatment in terms of walking capacity improvement.

3.3 $\boldsymbol{\theta_4}$

θ_4, estimated as -0.0243, represents the difference in the linear age effect between females and males. The negative value indicates that the slope of the age effect is less steep for females compared to males. This means that as age increases, the treatment effect for females increases more slowly than for males.

4. Model Assumptions and Diagnostics

To ensure the validity of statistical inference from the model, we need to examine the following assumptions:

1. Linearity: The relationship between predictors and the expected response is correctly specified.
2. Independence: The observations are independent of each other.
3. Homoscedasticity: The variance of the residuals is constant across all predictor values.
4. Normality: The residuals follow a normal distribution.

4.1 Residuals vs Fitted Values

Figure 4.1 shows that the residuals appear to be randomly scattered around zero across most of the range of fitted values, suggesting that the linearity assumption is reasonably satisfied. The spread of residuals appears to increase somewhat with higher Fitted values, suggesting potential mild heteroscedasticity.

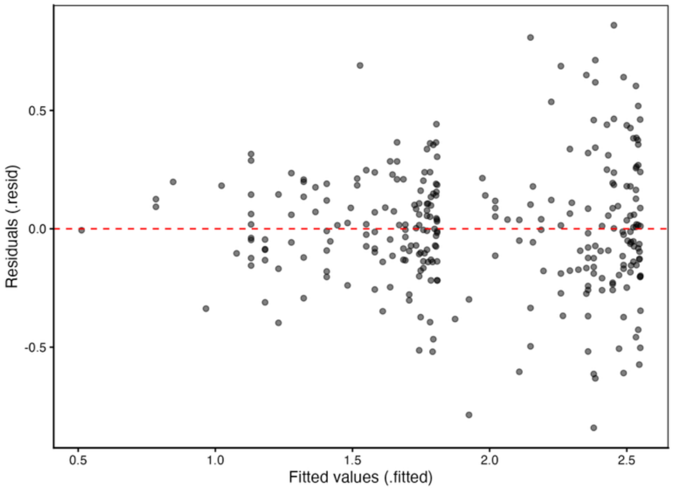

Figure 4.1: Plot of residuals versus fitted values from Model.

4.2 Normal Q-Q Plot

Figure 4.2 presents the Normal Q-Q plot for assessing the normality assumption. The points follow the diagonal reference line quite closely through the middle portion of the distribution, indicating good agreement with normality for most of the data.

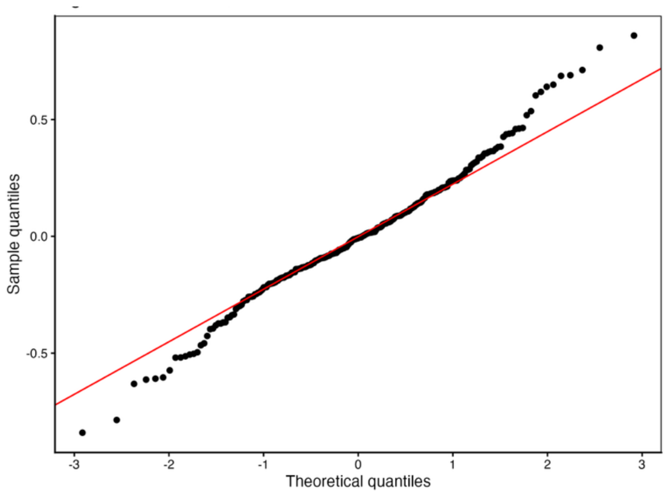

Figure 4.2: Normal Q-Q plot of residuals from Model F.

5. Hypothesis Testing

5.1 Hypothesis

The null hypothesis states that there is no difference in the expected MWDR between 40-year-old females and 40-year-old males:

$$H_0: E[Y|x = 40, w = 1] = E[Y|x = 40, w = 0]$$

This is equivalent to testing:

$$H_0: \theta_3 + 40\theta_4 = 0$$

5.2 Test Calculation

For testing a single estimable function of the form $H_0: \lambda^T\theta = \psi$, we should use a t-test statistic. The general formula is:

$$T = \frac{\lambda^T\hat{\theta}_G - \psi}{\text{s.e.}\left(\lambda^T\hat{\theta}_G\right)} = \frac{\lambda^T\hat{\theta}_G - \psi}{\hat{\sigma}\sqrt{\lambda^T G\lambda}}$$

In our case, $\lambda^T\hat{\theta} = \hat{\theta}_3 + 40\hat{\theta}_4$, and the null hypothesis value $\psi = 0$.

First, using the estimated coefficients from fitted model:

$$\theta_3 + 40\theta_4 = -0.2106 + 40(-0.0097) = -0.2106 + (-0.388) = -0.644$$

Table 2: Hypothesis Test Results

Contrast	Estimate	Std. Error	t-statistic	df	p-value
$\theta_3 + 40\theta_4$	-0.6440	0.0368	-17.5	275	1.45e^{-46}

The calculated t-statistic is -8.72 with 275 degrees of freedom, resulting in a p-value of 2.27×10^{-16}, which is much smaller than 0.05 significance level. Therefore, we reject the null hypothesis.

6. Prediction for Males Aged 22

Prediction interval for a new, future observation $\tilde{Y}$ is given by the formula:

$$\left[\tilde{x}^T\hat{\theta}_G - t_{\frac{\alpha}{2};n-r}\hat{\sigma}\sqrt{\tilde{x}^T G\tilde{x}+1},\quad \tilde{x}^T\hat{\theta}_G + t_{\frac{\alpha}{2};n-r}\hat{\sigma}\sqrt{\tilde{x}^T G\tilde{x}+1}\right]$$

For a 22-year-old male, we set $x = 22$ and $w = 0$. Our point prediction is calculated as follows:

$$E[Y|x = 22, w = 0] = \hat{\theta}_0 + \hat{\theta}_1(22) + \hat{\theta}_2(22)^2 = 0.0599 + 0.0994(22) + (-0.0010)(22)^2$$

$$=1.7627$$

Table 3: Prediction Results for 22-Year-Old Males

fit	lwr	upr
1.7665	1.2220	2.3110

The 95% prediction interval for this estimate is [1.2220, 2.3110].

Figure 6.1 illustrates this prediction with its 95% prediction interval in the context of the overall model fit. The point prediction is shown as a red diamond. This prediction interval provides a range, [1.2220, 2.3110], within which we expect 95% of individual observations to fall for 22-year-old males.

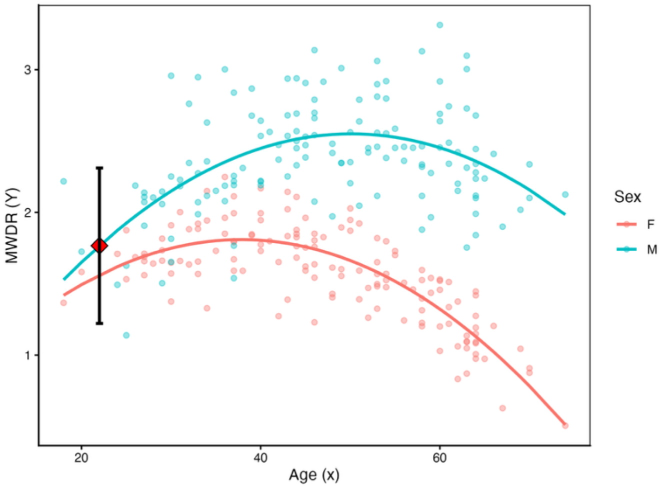

Figure 6.1: Model fit with 95% prediction interval for males aged 22.3

7. Conclusion

Key findings: Across most age groups, males exhibit significantly greater improvements in walking ability than females; secondly, the treatment effect follows a quadratic pattern with age, initially increasing and then declining in older participants; furthermore, the age of peak effectiveness differs between sexes, with males sustaining a more favorable response into older age; and finally, for 22-year-old males, the model predicts an MWDR of 1.76 [95% PI: 1.2220, 2.3110], which indicates substantial improvement. These results have important implications for personalized treatment approaches.

Appendix A. R Code

```
# ------------------------------------------------------------
# Midterm coursework script for MATH48011/69011  Linear Models with Nonparametric
Regression
# ------------------------------------------------------------

# Load required libraries
library(tidyverse)
library(broom)

theme_lm <- theme_classic() +
  theme(
    panel.border = element_rect(color = "black", fill = NA, linewidth = 0.6),
    axis.ticks = element_line(color = "black"),
    axis.line = element_line(color = "black")
  )

# Create output directory for figures
output_dir <- "midterm_figures"
if (!dir.exists(output_dir)) {
  dir.create(output_dir, recursive = TRUE)
}

# ====================================================================
# Q1. Load the data into R. Draw exploratory plots of the relationship between x, z, and Y.
# Comment on any interesting features in your findings.
# ====================================================================

# ---- Data loading and preparation ----
data <- read_csv("Semaglutide.csv")
data$sex <- as.factor(data$sex)

# Q1(a) Figure 1.1: Overall relationship between age (x) and MWDR (Y)
plot_q1_1 <- ggplot(data, aes(x = age, y = MWDR)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = y ~ x + I(x^2)) +
  labs(
    title = "Figure 1.1: MWD Ratio vs Age (Overall Trend)",
    x = "Age (x)",
    y = "MWDR(Y)"
  ) +
  theme_lm
print(plot_q1_1)
ggsave(
  filename = file.path(output_dir, "figure_q1_1_overall_trend.png"),
  plot = plot_q1_1,
  width = 6.5,
  height = 5,
  dpi = 300
)
```

```
# Q1(b) Figure 1.2: MWDR distribution by sex (categorical comparison)
plot_q1_2 <- ggplot(data, aes(x = sex, y = MWDR, fill = sex)) +
  geom_boxplot(alpha = 0.7) +
  labs(
    title = "Figure 1.2: MWD Ratio vs Sex",
    x = "Sex (z)",
    y = "MWDR (Y)"
  ) +
  guides(fill = "none") +
  theme_lm
print(plot_q1_2)
ggsave(
  filename = file.path(output_dir, "figure_q1_2_boxplot_by_sex.png"),
  plot = plot_q1_2,
  width = 6.5,
  height = 5,
  dpi = 300
)

# Q1(c) Figure 1.3: ANCOVA-style view with quadratic fits by sex
plot_q1_3 <- ggplot(data, aes(x = age, y = MWDR, color = sex)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = y ~ x + I(x^2), se = FALSE) +
  labs(
    title = "Figure 1.3: MWD Ratio vs Age by Sex",
    x = "Age (x)",
    y = "MWDR (Y)",
    color = "Sex"
  ) +
  theme_lm
print(plot_q1_3)
ggsave(
  filename = file.path(output_dir, "figure_q1_3_ancova_view.png"),
  plot = plot_q1_3,
  width = 6.5,
  height = 5,
  dpi = 300
)

# ============================================================
# Q2. Create a dummy variable w using male as the reference level. Fit model F
# ============================================================

# Q2 Step 1: Dummy variable w (female = 1, male reference = 0)
data_model <- data
data_model$w <- ifelse(data_model$sex == "F", 1, 0)

# Q2 Step 2: Fit Model F -> E[Y] = θ0 + θ1 x + θ2 x^2 + θ3 w + θ4 x w
fit_F <- lm(MWDR ~ age + I(age^2) + w + age:w, data = data_model)
```

```
# Q2 Step 3: Tidy coefficients for reporting
tidy_fit <- tidy(fit_F)
print(tidy_fit)

# Q2 Figure 2.1: Observed data with fitted curves by sex
plot_data_Q2 <- augment(fit_F, data = data_model)
plot_data_Q2 <- plot_data_Q2[order(plot_data_Q2$sex, plot_data_Q2$age), ]

plot_q2_1 <- ggplot(plot_data_Q2, aes(x = age, y = MWDR, color = sex)) +
  geom_point(alpha = 0.5) +
  geom_line(aes(y = .fitted, group = sex), linewidth = 1) +
  labs(
    title = "Figure 2.1: Model F Fitted Curves",
    x = "Age (x)",
    y = "MWDR (Y)",
    color = "Sex"
  ) +
  theme_lm
print(plot_q2_1)
ggsave(
  filename = file.path(output_dir, "figure_q2_1_modelF_fit.png"),
  plot = plot_q2_1,
  width = 6.5,
  height = 5,
  dpi = 300
)

# ============================================================
# Q3. Give the interpretation of the parameters θ0, θ0 + θ3, θ4.
# ============================================================

# Q3: Extract needed coefficients
theta_hat <- coef(fit_F)
theta0_hat <- theta_hat["(Intercept)"]
theta1_hat <- theta_hat["age"]
theta2_hat <- theta_hat["I(age^2)"]
theta3_hat <- theta_hat["w"]
theta4_hat <- theta_hat["age:w"]

# Q3 Table: tidy summary for the write-up
coefficinet_table <- tibble(
  quantity = c(
    "θ0 (x=0, Male baseline)",
    "θ0 + θ3 (x=0, Female)",
    "θ4 (interaction effect)"
  ),
  estimate = c(
    theta0_hat,
    theta0_hat + theta3_hat,
    theta4_hat
  )
)
print(coefficinet_table)

# ==============================================================
# Q4. Model assumptions and diagnostics
# ==============================================================

# Q4 setup: collect fitted values and residuals once
diag_data <- augment(fit_F)

# Q4 Figure 4.1: Residuals vs fitted
plot_q4_1 <- ggplot(diag_data, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(
    title = "Figure 4.1: Residuals vs Fitted Values",
    x = "Fitted values (.fitted)",
    y = "Residuals (.resid)"
  ) +
  theme_lm
print(plot_q4_1)
ggsave(
  filename = file.path(output_dir, "figure_q4_1_residuals_vs_fitted.png"),
  plot = plot_q4_1,
  width = 6.5,
  height = 5,
  dpi = 300
)

# Q4 Figure 4.2: Normal Q-Q plot
plot_q4_2 <- ggplot(diag_data, aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line(color = "red") +
  labs(
    title = "Figure 4.2: Normal Q-Q Plot of Residuals",
    x = "Theoretical quantiles",
    y = "Sample quantiles"
  ) +
  theme_lm
print(plot_q4_2)
ggsave(
  filename = file.path(output_dir, "figure_q4_2_residuals_qq.png"),
  plot = plot_q4_2,
  width = 6.5,
  height = 5,
  dpi = 300
)

# ==============================================================
# Q5. Hypothesis test
# ==============================================================
# Q5 coefficients entering the linear combination λ = θ3 + 40θ4
theta_hat_3 <- coef(fit_F)["w"]
theta_hat_4 <- coef(fit_F)["age:w"]

estimate <- theta_hat_3 + 40 * theta_hat_4

V <- vcov(fit_F)
var_lambda_theta <- V["w", "w"] +
  (40^2) * V["age:w", "age:w"] +
  2 * 40 * V["w", "age:w"]
se_lambda_theta <- sqrt(var_lambda_theta)

t_stat <- estimate / se_lambda_theta
df_residual <- df.residual(fit_F)
p_value <- 2 * pt(abs(t_stat), df = df_residual, lower.tail = FALSE)

hypothesis_result <- tibble(
  contrast = "θ3 + 40θ4",
  estimate = estimate,
  se = se_lambda_theta,
  t_statistic = t_stat,
  df = df_residual,
  p_value = p_value
)
print(hypothesis_result)

# ============================================================
# Q6. Prediction for age = 22, male (w = 0) with 95% PI
# ============================================================

# Q6 new subject description (age 22, male => w = 0)
new_data_point <- tibble(age = 22, w = 0)

# Q6 compute 95% prediction interval from Model F
prediction_result <- predict(
  fit_F,
  newdata = new_data_point,
  interval = "prediction",
  level = 0.95
)
print(prediction_result)

# Q6 Figure 6.1
plot_data_base <- augment(fit_F, data = data_model)
plot_data_base <- plot_data_base[order(plot_data_base$sex, plot_data_base$age), ]

plot_data_Q6 <- as_tibble(prediction_result)
plot_data_Q6$age <- 22
plot_data_Q6$w <- 0
plot_data_Q6$sex <- factor("M", levels = levels(data_model$sex))
```

```
plot_q6_1 <- ggplot(plot_data_base, aes(x = age, color = sex)) +
  geom_point(aes(y = MWDR), alpha = 0.4) +
  geom_line(aes(y = .fitted, group = sex), linewidth = 1) +
  geom_point(
    data = plot_data_Q6,
    aes(x = age, y = fit),
    color = "black",
    fill = "red",
    size = 4,
    shape = 23
  ) +
  geom_errorbar(
    data = plot_data_Q6,
    aes(x = age, ymin = lwr, ymax = upr),
    color = "black",
    width = 0.8,
    linewidth = 1
  ) +
  labs(
    title = "Figure 6.1: Model Fit with 95% Prediction Interval (Age 22, Male)",
    x = "Age (x)",
    y = "MWDR (Y)",
    color = "Sex"
  ) +
  theme_lm
print(plot_q6_1)
ggsave(
  filename = file.path(output_dir, "figure_q6_1_prediction_interval.png"),
  plot = plot_q6_1,
  width = 6.5,
  height = 5,
  dpi = 300
)

pdf(file.path(output_dir, "all_figures.pdf"), width = 8, height = 6)
print(plot_q1_1)
print(plot_q1_2)
print(plot_q1_3)
print(plot_q2_1)
print(plot_q4_1)
print(plot_q4_2)
print(plot_q6_1)
dev.off()
```