Lab_deliverable_3

Yihang Duanmu, Minjun Kim, Annelise Schreiber

2025-10-23

Introduction

Social media performance is driven by both what you post and when you post it. In this report, we analyze a cosmetics brand's 2014 Facebook posts to understand how timing (hour, weekday, month) and promotion (paid vs. unpaid) relate to audience response. Our three objectives are: (1) test whether posting date/time impacts engagement (measured by Lifetime Post Consumers); (2) assess whether paid promotion is associated with higher reach; and (3) examine whether exposure changes across the year. Because engagement and reach are count-like and highly right-skewed, we model log-transformed outcomes, primarily log(1 + consumers). We use multiple linear regression to estimate effects while controlling for potential confounders (e.g., weekday, month, and—where available—content type/category and page size). This framework lets us quantify marginal effects (and interactions) and report uncertainty via confidence intervals and p-values, providing an interpretable baseline for the team's subsequent modeling work.

Data

First, we standardize and clean the data.

All the post_weekday, post_hour, and post_month data are turned to integers with non-numeric text being turned into NA. Then, all the rows with NA are deleted, the type column and the column of paid or unpaid are turned into factor columns.

Then we build the analysis tibble dat—it picks the response variables (post consumers) and the cleaned timing predictors (wday, hour, month), adds the promotion flag (paid or unpaid), and creates a simple time-order index (idx). The summary(dat$consumers) line quickly checks the outcome's distribution; the output shows a right-skewed variable with large upper outliers, which motivates options like log1p(consumers) as engagement and log1p(reach) in later models.

Table 1: Outcome summaries after log1p construction

	consumers	reach	y_log_eng	y_log_reach
	Min. : 9.0	Min. : 238	Min. :2.303	Min. : 5.476
	1st Qu.: 335.0	1st Qu.: 3332	1st Qu.:5.817	1st Qu.: 8.111
	Median : 555.0	Median : 5290	Median :6.321	Median : 8.574
	Mean : 803.6	Mean : 14008	Mean :6.315	Mean : 8.822
	3rd Qu.: 966.0	3rd Qu.: 13232	3rd Qu.:6.874	3rd Qu.: 9.490
	Max. :11328.0	Max. :180480	Max. :9.335	Max. :12.103

Table 2: Q1 Multiple OLS: log(Consumers+1) Paid + Hour + Paid×Hour

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	6.1468	0.0979	62.7705	0.0000	5.9544	6.3392
paid:Paid	0.2344	0.1786	1.3122	0.1901	-0.1166	0.5853
post_hour	0.0145	0.0107	1.3572	0.1754	-0.0065	0.0356
paid:Paid:post_hour	-0.0053	0.0205	-0.2595	0.7954	-0.0457	0.0350

Table 3: Q1 Model fit

r.squared	adj.r.squared	sigma	statistic	p.value	df	nobs
0.013	0.007	0.889	2.1585	0.0921	3	496

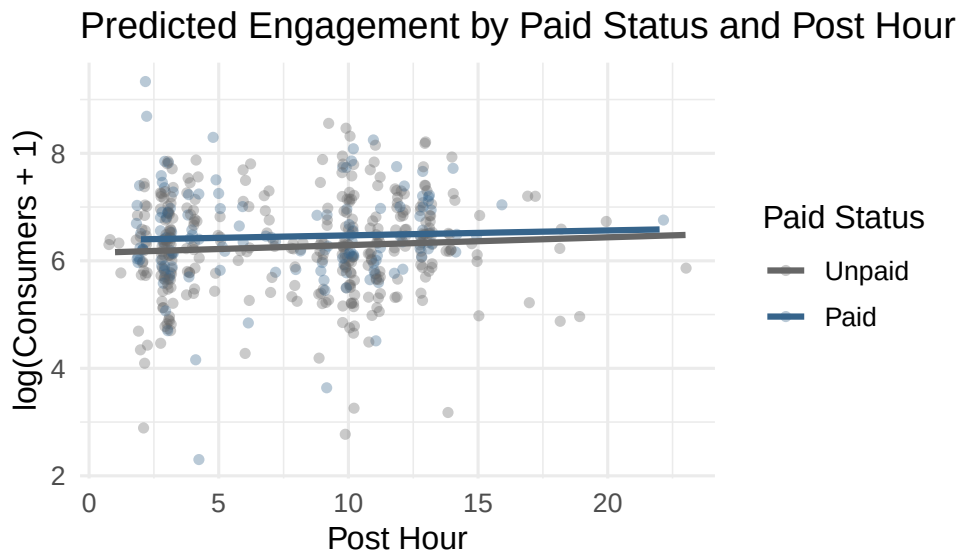
Statistical Modeling Framework

Question 1: Does the date and time in which a post is made impact the engagement that said post receives?

Response (Y): log1p(consumers) — log-transform stabilizes variance and downweights a few very large posts.

Covariates (X): hour_num = post hour (0–23, numeric), paid_fac = Paid vs Unpaid, Controls: wday_fac (Mon–Sun), mon_num (1–12), type_fac (post type), cat_fac (category), and log_page_likes (log of page_total_likes)

```
## `geom_smooth()` using formula = 'y ~ x'
```



The scatter with fitted lines shows only a very small increase in engagement (log(1+consumers)) as posting hour gets later. The "Paid" line sits slightly above the "Unpaid" line across most hours, suggesting paid posts tend to draw marginally higher engagement, but the two lines are nearly overlapping and the point cloud is wide—so the hour effect is weak and the paid vs. unpaid gap is modest relative to the variability. In short: posting time has little practical impact, and paying helps a bit, but neither factor explains much of the dispersion.

```
##
## Call:
## lm(formula = log_consumers ~ paid_fac + post_hour + paid_fac:post_hour,
##     data = fb2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1154 -0.4739  0.0039  0.5361  2.9356
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.146783   0.097925  62.771   <2e-16 ***
## paid_facPaid              0.234376   0.178618   1.312    0.190
## post_hour                 0.014529   0.010705   1.357    0.175
## paid_facPaid:post_hour -0.005328   0.020533  -0.259    0.795
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.889 on 492 degrees of freedom
## Multiple R-squared:  0.01299,    Adjusted R-squared:  0.006972
## F-statistic: 2.158 on 3 and 492 DF,  p-value: 0.09207
```

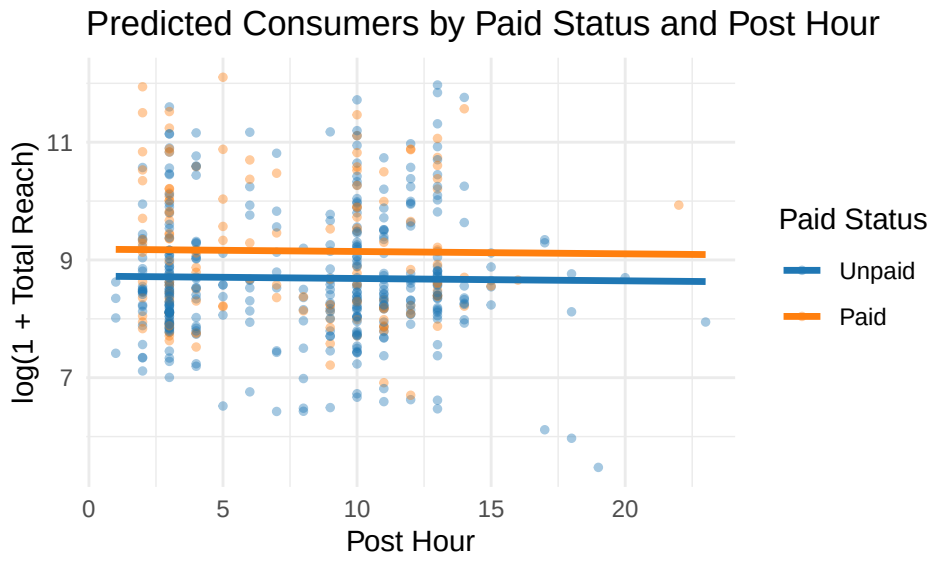
Table 2 and Table 3 show the summary statistics of the model. The multiple-linear model log(Consumers + 1) ~ Paid + Hour + Paid X Hour explains very little of the variation in engagement (R^2 = 0.014; adj. R^2 = 0.008). For unpaid posts at hour 0, the expected log engagement is ~6.13. The main effect for Paid is 0.247 (95% CI: [-0.106, 0.6], p = 0.17), which on the original scale corresponds to roughly +28% consumers (e^0.247 - 1) at hour 0, but the CI includes zero, so it's not statistically significant. The Hour slope is 0.0144 per hour (CI [-0.0067, 0.0356], p = 0.18), about +1.4% per additional hour for unpaid posts - again not significant. The interaction -0.0052 (p = 0.80) suggests the hour effect is slightly smaller for paid posts, but this is also not significant. Overall, posting hour and paid status (and their interaction) show no reliable association with engagement in this dataset; other variables likely drive the differences in consumer counts.

Question 2: Does the fact of a post being paid/unpaid impact the average reach of the post significantly?

Response variable (Y): "reach" - This represents lifetime reach of the post post, a numerical value.

Covariates: "paid" - This is a binary variable (e.g., 1 for paid posts, 0 for unpaid). It allows us to compare the average reach between the two groups. "post_hour" - This variable represents the time of day a post was made.

Since we want to see whether the hour of a post and paid status impacts average reach, a multiple linear regression model is appropriate with paid status and post hour as the independent variables, and lifetime consumers as the response. This variable was used to represent reach as we felt the number of consumers of a post best represented overall reach.



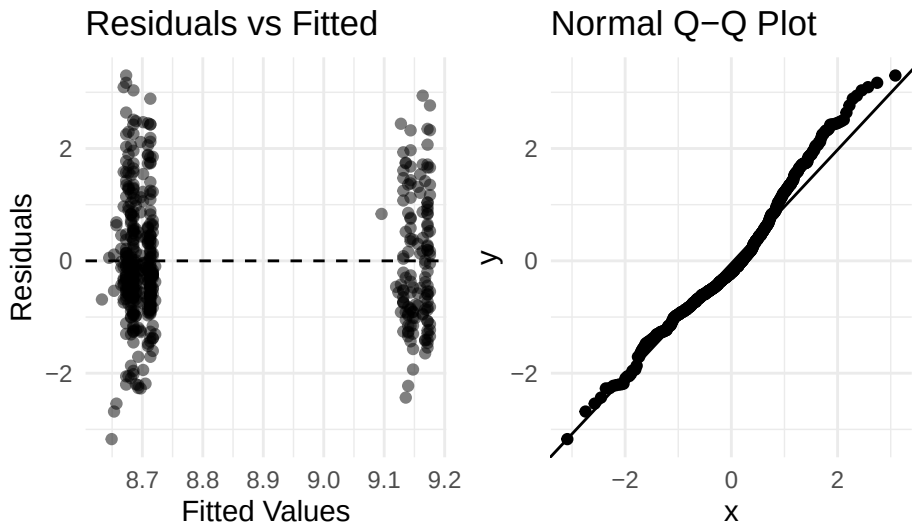
Multiple Linear Regression Model: log(Consumersi+1)= 0+ 1(Paidi)+ 2(Houri)+ i 1 represents the expected change in the response as a result of a post being paid as opposed to unpaid; 0 represents the expected change in the response as the result of each added hour in the day that a post is made.

Assumptions: linear relationships between the predictors and the response; independent observations; constant variance of residuals; and normally distributed errors.

```
##
## Call:
## lm(formula = log_reach ~ paid + post_hour, data = fb2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1727 -0.7282 -0.2109  0.6365  3.2985
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.725259   0.109162  79.929  < 2e-16 ***
## paid         0.458463   0.111661   4.106 4.72e-05 ***
## post_hour   -0.004004   0.011451  -0.350    0.727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.114 on 493 degrees of freedom
## Multiple R-squared:  0.03382,    Adjusted R-squared:  0.0299
## F-statistic: 8.628 on 2 and 493 DF,  p-value: 0.0002076
```

From this data, we're able to have some insight on the research question. It seems as though the influence of a post being paid or not is statistically significant regarding a post's reach, but the hour in which the post is made is not significant. From the adjusted R squared value of 0.009859, we can tell that only about 3% of the variance in log consumers can be accounted for by theses factors–other factors are likely to be more influential.

Illustration of Model Fit:



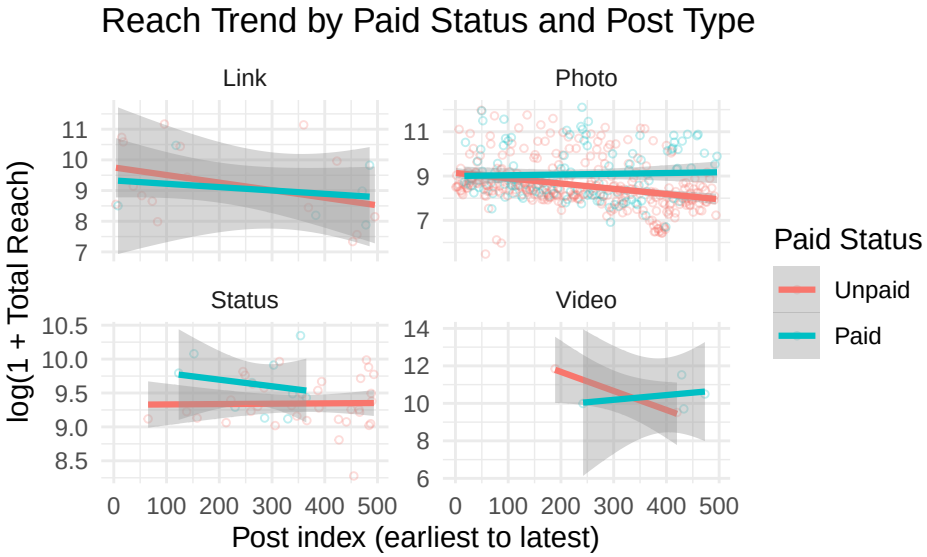
From these graphs, we can see that the normality assumption generally holds except at the tails where there seems to be outliers. From the residuals vs fitted graph we can see that linearity holds as the points are approximately scattered about 0 without a clear pattern.

Question 3: Does the cosmetics brand gain significant exposure over the course of the year, from first to latest post analyzed in 2014, controlling the same paid status and same post type?

Response variable (Y): "reach" - This represents lifetime reach of the post post, a numerical value.

Covariates: the relative post time of a post in 2014, with earlier posts having lower post time index.

Controlled: "paid" - This is a binary variable (e.g., 1 for paid posts, 0 for unpaid). "type" - Type of a post (e.g. Link, Photo, Status, Video).



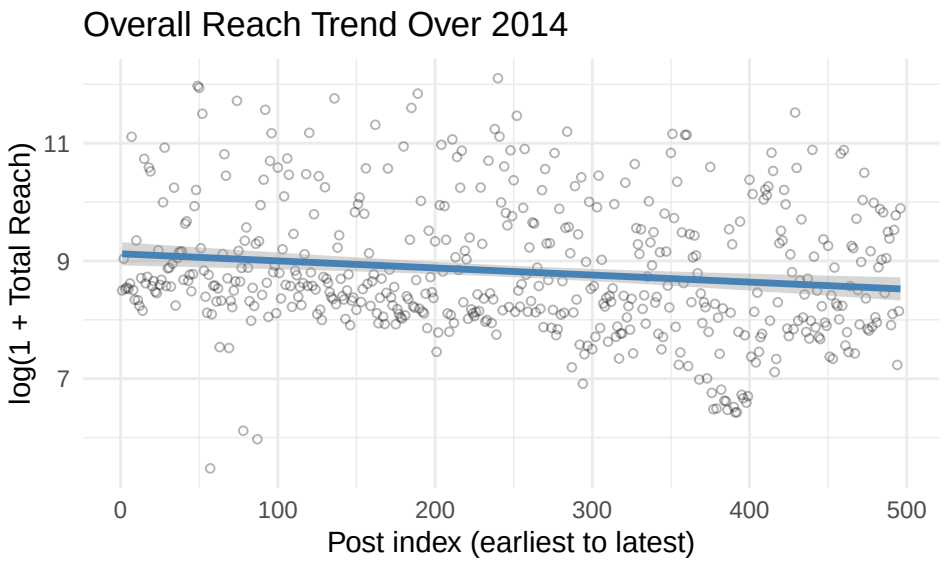
The scatter plots and the fitted regression line on each type and paid status show a mix of downward and upward trend (mostly downward trend) between total reach and post time. This contradicts our assumption and suggests the brand most likely gradually loses exposure throughout 2014.

Then we fitted the linear model using R and plotted the total trend of total reach against post_time_index:

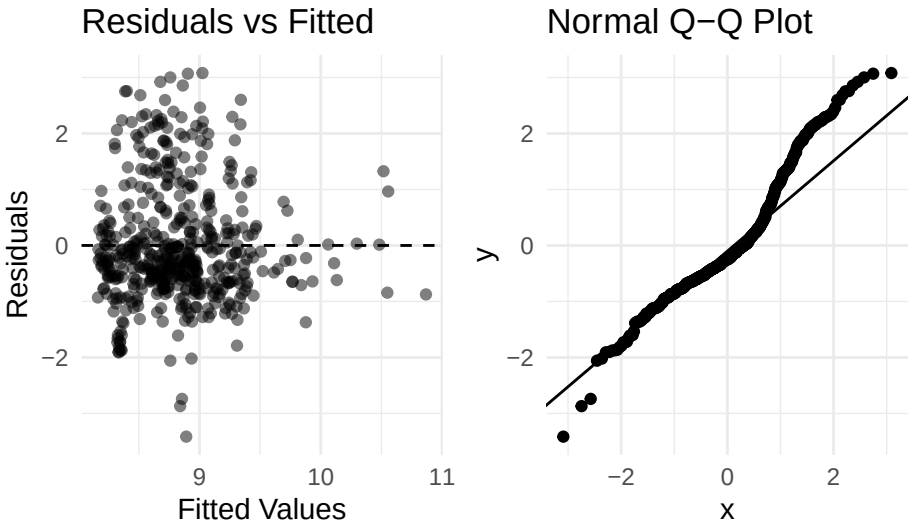
```
## Post time index: Estimate = -0.001666, Std. Error = 0.000339, t = -4.909, p = 1.251e-06
```

```
## R-squared = 0.141, Adjusted R-squared = 0.132
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The negative coefficient confirms the negative relationship between exposure and months. The low p-value suggests that the linear model is significant but the low R^2 value indicates high variation in the data.



From these graphs, the same results for question 2 holds. The distribution of the error variance appears normal and the assumption of linear model is satisfied. However, there are some outliers for posts with large amount of reach, which is possible for Facebook posts.

Author Contribution Statement

1. **First Deliverable**
    1. **Yihang Duanmu** — insight for the motivation section
    2. **Minjun Kim** — introduction/summary
    3. **Annelise Schreiber** — questions/impact; editing
2. **Second Deliverable**
    1. **Yihang Duanmu** — OLS analysis for Question 3
    2. **Minjun Kim** — OLS analysis for Question 1
    3. **Annelise Schreiber** — OLS analysis for Question 2
3. **Third Deliverable**
    1. **Yihang Duanmu** — Statistical modeling for Question 3
    2. **Minjun Kim** — Statistical modeling for Question 1
    3. **Annelise Schreiber** — Statistical modeling for Question 2

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(fig.width=5, fig.height=3, fig.align="center")
```

```
library(tidyverse); library(broom); library(kableExtra); library(here); library(janitor); library(lubridate); library(readr);
↪  library(dplyr); library(emmeans); library(ggplot2); library(patchwork)
```

```
fb_raw <- readr::read_delim("dataset_Facebook.csv", delim = ";", show_col_types = FALSE)

fb <- fb_raw |>
  janitor::clean_names() |>
  mutate(
    post_month   = suppressWarnings(as.integer(post_month)),
    post_weekday = suppressWarnings(as.integer(post_weekday)),
    post_hour    = suppressWarnings(as.integer(post_hour)),
    paid         = if_else(is.na(paid), 0L, as.integer(paid)),
    lifetime_post_consumers = as.numeric(lifetime_post_consumers)
  )
```

```
fb2 <- fb |>
  drop_na() |>
  mutate(
    post_weekday_num = post_weekday,
    post_hour_num    = post_hour,
    post_month_num   = post_month,
    log_reach     = log1p(lifetime_post_total_reach),
    log_consumers = log1p(lifetime_post_consumers),
    paid_fac = factor(if_else(is.na(paid) | paid == 0, "Unpaid", "Paid"),
                      levels = c("Unpaid", "Paid")),
    type_fac = factor(type)
  )

# Weekday factor (1-7 expected)
wk_labels <- c("Mon","Tue","Wed","Thu","Fri","Sat","Sun")
if (all(!is.na(fb2$post_weekday_num)) && all(fb2$post_weekday_num %in% 1:7)) {
  wday_fac <- factor(fb2$post_weekday_num, levels = 1:7, labels = wk_labels, ordered = TRUE)
} else {
  # fallback if strings slipped through
  wday_fac <- factor(as.character(fb$post_weekday), ordered = TRUE)
}

# Hour factor: detect actual range (handles 0-23 or 1-23)
hour_levels <- sort(unique(fb2$post_hour_num[!is.na(fb2$post_hour_num)]))
hour_fac <- factor(fb2$post_hour_num, levels = hour_levels, ordered = TRUE)

# Month factor (1-12 -> Jan..Dec)
if (all(!is.na(fb2$post_month_num)) && all(fb2$post_month_num %in% 1:12)) {
  month_fac <- factor(fb2$post_month_num, levels = 1:12, labels = month.abb, ordered = TRUE)
} else {
  month_fac <- factor(fb$post_month, ordered = TRUE)
}

# Paid flag as tidy factor
paid_fac <- factor(if_else(is.na(fb$paid) | fb$paid == 0, "Unpaid", "Paid"),
                   levels = c("Unpaid","Paid"))

fb2 <- fb2 |>
  arrange(post_month_num, post_weekday_num, post_hour_num) |>
  mutate(post_time_index = row_number())
```

```
dat <- fb2 |>
  transmute(
    consumers = as.numeric(lifetime_post_consumers),
    reach     = as.numeric(lifetime_post_total_reach),
    paid      = paid_fac,
    paid_fac  = paid_fac,

    hour   = hour_fac,
    wday   = wday_fac,
    month  = month_fac,
    idx    = dplyr::row_number()
  ) |>
  mutate(
```

```
    y_log_eng   = log1p(consumers),
    y_log_reach = log1p(reach),
    hour_num    = as.numeric(as.character(hour)),
    mon_num     = as.integer(month)
  ) |>
  mutate(
    paid_fac = forcats::fct_relevel(paid_fac, "Unpaid", "Paid"),
    paid     = paid_fac
  )

summary(dplyr::select(dat, consumers, reach, y_log_eng, y_log_reach)) |>
  kableExtra::kbl(caption = "Outcome summaries after log1p construction") |>
  kableExtra::kable_classic(full_width = FALSE)
```

```
ggplot(fb2, aes(x = post_hour, y = log_consumers, color = paid_fac)) +
geom_point(alpha = 0.35, shape = 16, size = 1.6,
position = position_jitter(width = 0.25, height = 0)) +
geom_smooth(method = "lm", se = FALSE, linewidth = 1.1) +
scale_color_manual(values = c("grey40", "steelblue4"), name = "Paid Status") +
labs(title = "Predicted Engagement by Paid Status and Post Hour",
x = "Post Hour", y = "log(Consumers + 1)") +
theme_minimal(base_size = 12) +
theme(legend.position = "right")
```

```
# Multiple linear regression with interaction

m_q1 <- lm(log_consumers ~ paid_fac + post_hour + paid_fac:post_hour, data = fb2)

# Nice coefficient table

tbl_q1_coef <- broom::tidy(m_q1, conf.int = TRUE) |>
dplyr::mutate(dplyr::across(where(is.numeric), ~ round(., 4)),
term = gsub("paid_fac", "paid:", term))

# Model fit stats

tbl_q1_fit <- broom::glance(m_q1) |>
dplyr::mutate(dplyr::across(where(is.numeric), ~ round(., 4))) |>
dplyr::select(r.squared, adj.r.squared, sigma, statistic, p.value, df, nobs)

# Show

kableExtra::kbl(tbl_q1_coef,
caption = "Q1 Multiple OLS: log(Consumers+1) ~ Paid + Hour + Paid×Hour") |>
kableExtra::kable_classic(full_width = FALSE)

kableExtra::kbl(tbl_q1_fit, caption = "Q1 Model fit") |>
kableExtra::kable_classic(full_width = FALSE)

summary(m_q1)
```

```
fb2$paid <- as.numeric(fb2$paid)
fb2$post_hour <- as.numeric(as.character(fb2$post_hour))
model <- lm(log_reach ~ paid + post_hour, data = fb2)

pred_dat <- expand.grid(
  paid = c(0, 1),
  post_hour = seq(min(fb2$post_hour), max(fb2$post_hour), length.out = 50)
)

pred_dat$pred <- predict(model, newdata = pred_dat)

ggplot(fb2, aes(x = post_hour, y = log_reach, color = factor(paid))) +
  geom_point(alpha = 0.4, size = 1) +
  geom_line(data = pred_dat, aes(y = pred, color = factor(paid)), size = 1.2) +
  labs(
    title = "Predicted Consumers by Paid Status and Post Hour",
    x = "Post Hour",
    y = "log(1 + Total Reach)",
    color = "Paid Status"
  ) +
  scale_color_manual(values = c("0" = "#1f77b4", "1" = "#ff7f0e"),
                     labels = c("Unpaid", "Paid")) +
  theme_minimal()
```

```r
  summary(model)
```

```r
#Residuals vs Fitted

library(ggplot2)
library(broom)

resid_df <- augment(model)

p1 = ggplot(resid_df, aes(.fitted, .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Fitted Values", y = "Residuals", title = "Residuals vs Fitted") +
  theme_minimal()

#QQ Plot

p2 = ggplot(resid_df, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Normal Q-Q Plot") +
  theme_minimal()

p1 + p2
```

```r
ggplot(fb2, aes(x = post_time_index, y = log_reach, color = paid_fac)) +
  geom_point(alpha = 0.25, shape = 1, size = 1) +
  geom_smooth(method = "lm", se = TRUE, linewidth = 1.1) +
  facet_wrap(~ type_fac, scales = "free_y") +
  labs(
    title  = "Reach Trend by Paid Status and Post Type",
    x      = "Post index (earliest to latest)",
    y      = "log(1 + Total Reach)",
    color  = "Paid Status"
  ) +
  theme_minimal()
```

```r
model_month <- lm(
  log_reach ~ post_time_index + paid_fac + type_fac,
  data = fb2
)
b <- summary(model_month)$coefficients["post_time_index", ]
cat(sprintf("Post time index: Estimate = %.6f, Std. Error = %.6f, t = %.3f, p = %.4g\n",
            b[1], b[2], b[3], b[4]))
cat(sprintf("R-squared = %.3f, Adjusted R-squared = %.3f\n",
            summary(model_month)$r.squared, summary(model_month)$adj.r.squared))
```

```r
# Create a grid of values across the year for each Paid × Type
ggplot(fb2, aes(x = post_time_index, y = log_reach)) +
  geom_point(alpha = 0.3, shape = 1, size = 1.2) +
  geom_smooth(method = "lm", se = TRUE, color = "steelblue", linewidth = 1.2) +
  labs(
    title = "Overall Reach Trend Over 2014",
    x = "Post index (earliest to latest)",
    y = "log(1 + Total Reach)"
  ) +
  theme_minimal()
```

```r
resid_df <- augment(model_month)

p1 = ggplot(resid_df, aes(.fitted, .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Fitted Values", y = "Residuals", title = "Residuals vs Fitted") +
  theme_minimal()

#QQ Plot

p2 = ggplot(resid_df, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Normal Q-Q Plot") +
  theme_minimal()

p1 + p2
```