Gambling Expenditure and Salary Growth: Evidence from Bank Transaction Data

Elsa Zhu

2025-12-19

```r
library(tidyverse)
library(knitr)
library(broom)
library(quantreg)
library(scales)
```

1 Introduction

Gambling has become increasingly prevalent in the United Kingdom, with the gambling industry generating over £14 billion in gross gambling yield annually. While much attention has focused on the immediate financial costs of gambling, less is known about its association with longer-term economic outcomes such as career progression and salary growth. Understanding this relationship is crucial for policymakers evaluating the broader societal costs of gambling behaviour.

This study examines whether heavier gambling expenditure is associated with smaller salary increases over a five-year period. Using bank transaction data from 2012 to 2020, we analyse the relationship between gambling behaviour in an early period (2014–2015) and subsequent salary growth through to 2019–2020. We employ multiple analytical approaches to characterise the nature of this relationship, including tests for nonlinearity, heterogeneity across income groups, and distributional effects using quantile regression.

GitHub Repository: https://github.com/elsaxzzx/Elsa-IB9MH-Assignment.git

2 Methods

2.1 Data Source and Sample Construction

The dataset comprises 1,121,724 bank transactions from 21,522 unique individuals, spanning January 2012 to August 2020. Each transaction is categorised as either salary income (credit) or gambling activity (both debits and credits), with associated amounts in British pounds.

```r
# Load raw data
raw <- read_csv("transaction_data.csv")
```

```r
# Initial data exploration
n_users <- n_distinct(raw$user)
n_transactions <- nrow(raw)
date_range <- range(raw$Transaction.Date)
```

```r
# Transactions by category
category_summary <- raw %>% count(category, Credit.Debit)
```

```r
# ========== Data Processing ==========

# Extract salary data (credits only)
salary_data <- raw %>%
  filter(category == "Salary or Wages (Main)", Credit.Debit == "Credit") %>%
  mutate(year = year(Transaction.Date))

# Identify users with salary records in both periods
early_salary_users <- salary_data %>% filter(year %in% c(2014, 2015)) %>% distinct(user)
late_salary_users <- salary_data %>% filter(year %in% c(2019, 2020)) %>% distinct(user)
valid_users <- inner_join(early_salary_users, late_salary_users, by = "user")

# Calculate early-period annual salary (2014-2015 average)
early_salary <- salary_data %>%
  filter(user %in% valid_users$user, year %in% c(2014, 2015)) %>%
  group_by(user) %>%
  summarise(salary_early = sum(Amount) / 2, .groups = "drop")

# Calculate late-period annual salary (2019-2020, adjusted for partial year)
late_salary <- salary_data %>%
  filter(user %in% valid_users$user, year %in% c(2019, 2020)) %>%
  group_by(user) %>%
  summarise(salary_late = sum(Amount) / 1.625, .groups = "drop")

# Calculate early-period gambling metrics
gambling_data <- raw %>%
  filter(category == "Gambling") %>%
  mutate(year = year(Transaction.Date))

early_gambling <- gambling_data %>%
  filter(user %in% valid_users$user, year %in% c(2014, 2015)) %>%
  group_by(user) %>%
  summarise(
    gambling_gross = sum(Amount[Credit.Debit == "Debit"]),
    gambling_net = sum(Amount[Credit.Debit == "Debit"]) - sum(Amount[Credit.Debit == "Credit"]),
    gambling_n = n(),
    .groups = "drop"
  )

# Merge into analysis dataset
analysis_df <- valid_users %>%
  left_join(early_salary, by = "user") %>%
  left_join(late_salary, by = "user") %>%
  left_join(early_gambling, by = "user") %>%
  mutate(
    gambling_gross = replace_na(gambling_gross, 0),
    gambling_net = replace_na(gambling_net, 0),
    gambling_n = replace_na(gambling_n, 0),
    log_salary_change = log(salary_late) - log(salary_early),
    salary_growth_pct = (salary_late - salary_early) / salary_early * 100
```

```
  )

# Filter: minimum annual salary of £5,000 (exclude irregular income)
clean_df <- analysis_df %>%
  filter(salary_early >= 5000)

# Create gambling categories
clean_df <- clean_df %>%
  mutate(
    gambling_cat = case_when(
      gambling_gross == 0 ~ "None",
      gambling_gross <= 500 ~ "Light (£1-500)",
      TRUE ~ "Heavy (>£500)"
    ) %>% factor(levels = c("None", "Light (£1-500)", "Heavy (>£500)")),
    salary_quartile = ntile(salary_early, 4),
    gambling_ratio = gambling_gross / salary_early
  )
```

To construct our analysis sample, we identified individuals with salary records in both the baseline period (2014–2015) and the outcome period (2019–2020). This five-year gap allows sufficient time for meaningful salary progression while maintaining a reasonable sample size. The baseline period was chosen to maximise sample size while allowing adequate follow-up; earlier years had insufficient observations.

We calculated annualised salary for each period by summing all salary credits and dividing by the number of years (with 2020 adjusted to 1.625 years given data availability through August). Gambling expenditure was measured as total debits to gambling merchants during the baseline period. We also calculated gambling frequency (number of transactions) and net gambling losses (debits minus credits).

Individuals with annual baseline salaries below £5,000 were excluded to ensure reliable salary growth calculations and to focus on individuals with regular employment income. This yielded a final sample of 944 individuals.

2.2 Variable Definitions

Our primary **outcome variable** is the log change in salary:

$$\Delta \ln(\text{salary}_i) = \ln(\text{salary}_{i,2019-20}) - \ln(\text{salary}_{i,2014-15})$$

This specification has the interpretation that coefficients represent approximate percentage changes in salary growth.

The main **explanatory variables** are:

- **Gross gambling expenditure** (£): Total debits to gambling merchants in 2014–2015
- **Gambling frequency**: Number of gambling transactions in 2014–2015
- **Gambling category**: None (£0), Light (£1–500), or Heavy (>£500)
- **Gambling ratio**: Gambling expenditure as a proportion of baseline salary

We control for **baseline log salary** to account for mean reversion in earnings, as higher initial earners tend to experience lower proportional growth.

2.3 Statistical Analysis

We estimate ordinary least squares (OLS) regression models of the form:

$$\Delta \ln(\text{salary}_i) = \beta_0 + \beta_1 \cdot \text{gambling}_i + \beta_2 \cdot \ln(\text{salary}_{i,\text{early}}) + \varepsilon_i$$

To examine distributional effects, we employ quantile regression at the 25th, 50th, and 75th percentiles:

$$Q_\tau(\Delta \ln(\text{salary}_i)|X_i) = \beta_{0,\tau} + \beta_{1,\tau} \cdot \text{gambling}_i + \beta_{2,\tau} \cdot \ln(\text{salary}_{i,\text{early}})$$

We test for heterogeneity across income groups by stratifying the sample by baseline salary quartile and by including interaction terms. Robustness checks include varying the minimum salary threshold and comparing gambling amount versus frequency.

3 Results

3.1 Descriptive Statistics

```r
# Summary statistics
desc_stats <- clean_df %>%
  summarise(
    N = n(),
    `Mean Baseline Salary (£)` = mean(salary_early),
    `SD Baseline Salary` = sd(salary_early),
    `Median Baseline Salary (£)` = median(salary_early),
    `Mean Outcome Salary (£)` = mean(salary_late),
    `SD Outcome Salary` = sd(salary_late),
    `Median Outcome Salary (£)` = median(salary_late),
    `Mean Gambling (£)` = mean(gambling_gross),
    `Median Gambling (£)` = median(gambling_gross),
    `Mean Gambling Transactions` = mean(gambling_n),
    `% Any Gambling` = mean(gambling_gross > 0) * 100,
    `% Heavy Gambling (>£500)` = mean(gambling_gross > 500) * 100
  ) %>%
  pivot_longer(everything(), names_to = "Statistic", values_to = "Value")

kable(desc_stats, digits = 1, caption = "Table 1: Sample Descriptive Statistics")
```

Table 1: Table 1: Sample Descriptive Statistics

Statistic	Value
N	944.0
Mean Baseline Salary (£)	19672.5
SD Baseline Salary	16809.1
Median Baseline Salary (£)	14926.5
Mean Outcome Salary (£)	22791.0
SD Outcome Salary	27895.2
Median Outcome Salary (£)	15919.8
Mean Gambling (£)	402.9

Statistic	Value
Median Gambling (£)	20.0
Mean Gambling Transactions	14.9
% Any Gambling	58.2
% Heavy Gambling (>£500)	10.6

The sample comprises 944 individuals with a mean baseline salary of £19,672 (median: £14,926). Mean outcome salary was £22,791, representing modest nominal growth over the five-year period.

Gambling activity was common, with 58.2% of individuals recording at least one gambling transaction during the baseline period. However, the distribution was highly right-skewed: median gambling expenditure was only £20, while the mean was £403. Approximately 10.6% of the sample were classified as heavy gamblers with expenditure exceeding £500.

3.2 Distribution of Key Variables

```
p1 <- ggplot(clean_df, aes(x = gambling_gross)) +
  geom_histogram(bins = 50, fill = "steelblue", alpha = 0.7) +
  scale_x_continuous(labels = comma, limits = c(0, 5000)) +
  labs(x = "Gross Gambling Expenditure (£)", y = "Count",
       title = "A. Distribution of Gambling Expenditure",
       subtitle = "Truncated at £5,000; full range extends to £56,000") +
  theme_minimal()

p2 <- ggplot(clean_df, aes(x = log_salary_change)) +
  geom_histogram(bins = 50, fill = "darkgreen", alpha = 0.7) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +
  labs(x = "Log Salary Change", y = "Count",
       title = "B. Distribution of Log Salary Change",
       subtitle = "Dashed line indicates zero change") +
  theme_minimal()

gridExtra::grid.arrange(p1, p2, nrow = 2)
```

Figure 1 illustrates the distributions of key variables. Gambling expenditure is highly right-skewed, with most individuals gambling little or nothing and a long right tail of heavy gamblers. Log salary change is approximately normally distributed with mean -0.15 and substantial variation (SD = 1.2).

3.3 Bivariate Relationship

```
ggplot(clean_df, aes(x = gambling_gross, y = log_salary_change)) +
  geom_point(alpha = 0.3, size = 1) +
  geom_smooth(method = "lm", color = "red", se = TRUE) +
  scale_x_continuous(labels = comma) +
  labs(x = "Gross Gambling Expenditure (£)",
       y = "Log Salary Change (2014-15 to 2019-20)",
       title = "Relationship Between Gambling and Salary Growth") +
  theme_minimal()
```

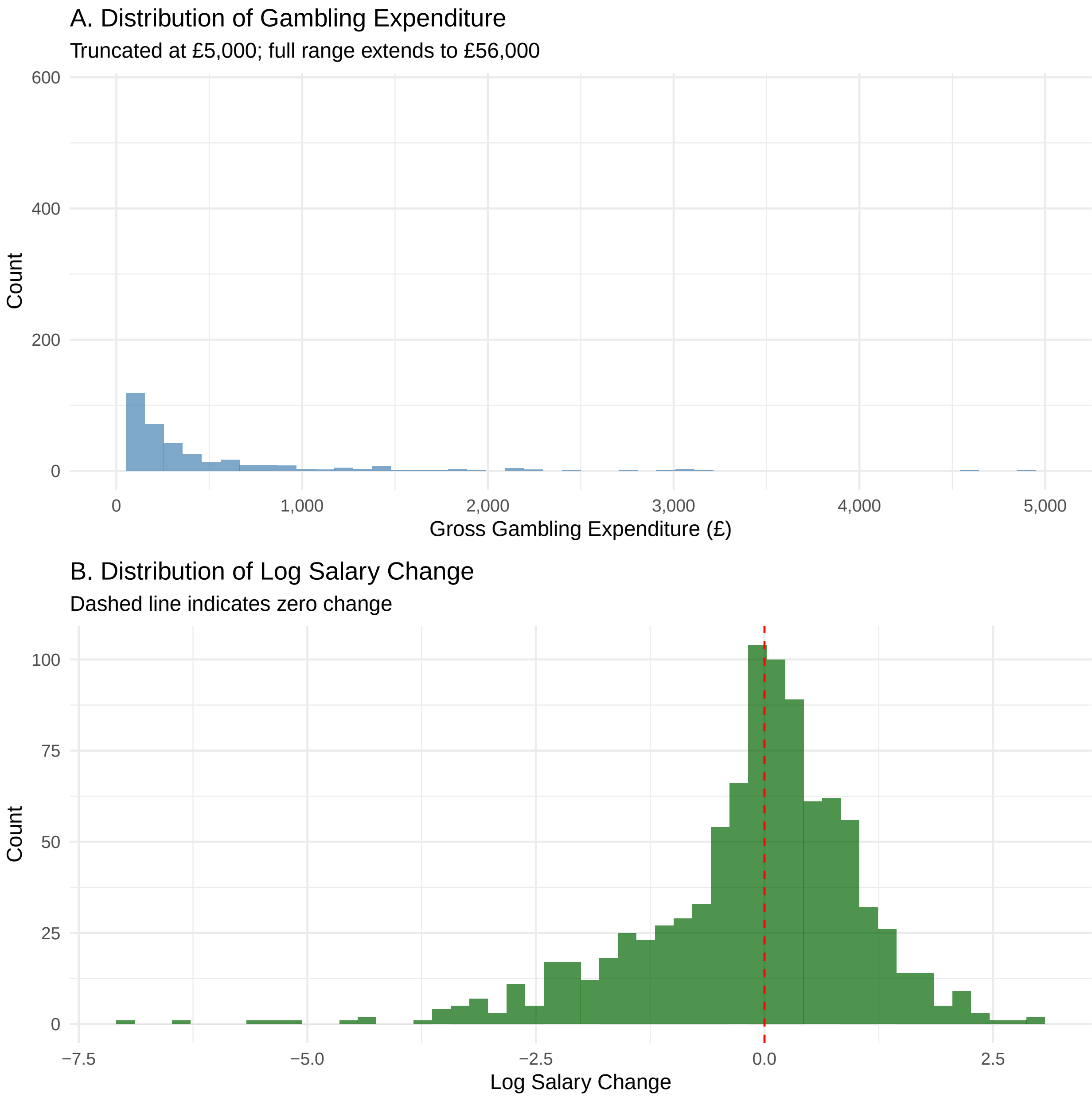

Figure 1: Figure 1: Distribution of Gambling Expenditure and Salary Change

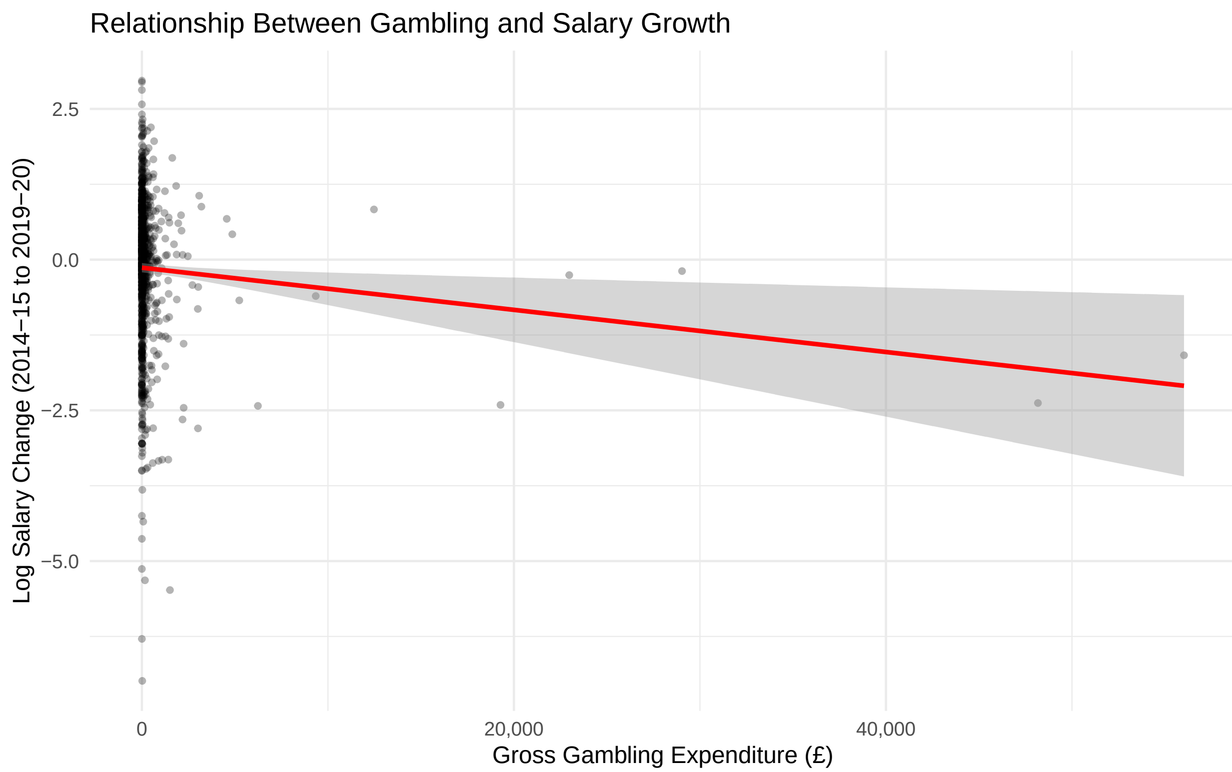

Figure 2: Figure 2: Gambling Expenditure and Salary Change

Figure 2 displays the raw relationship between gambling expenditure and salary change. The negative slope is visible but heavily influenced by a small number of high-expenditure observations. The widening confidence interval at higher gambling levels reflects sparse data in this region.

```
cor_gross <- cor(clean_df$gambling_gross, clean_df$log_salary_change)
cor_freq <- cor(clean_df$gambling_n, clean_df$log_salary_change)
```

The correlation between gambling expenditure and log salary change is -0.082, indicating a weak negative association. Gambling frequency shows similar correlation (-0.074).

3.4 Gambling Category Analysis

```
cat_stats <- clean_df %>%
  group_by(`Gambling Category` = gambling_cat) %>%
  summarise(
    N = n(),
    `% of Sample` = n() / nrow(clean_df) * 100,
    `Mean Baseline Salary (£)` = mean(salary_early),
    `Mean Log Salary Change` = mean(log_salary_change),
    `SE` = sd(log_salary_change) / sqrt(n()),
    `Median Log Salary Change` = median(log_salary_change)
  )

kable(cat_stats, digits = 3, caption = "Table 2: Salary Growth by Gambling Category")
```

Table 2: Table 2: Salary Growth by Gambling Category

Gambling Category	N	% of Sample	Mean Baseline Salary (£)	Mean Log Salary Change	SE	Median Log Salary Change
None	395	41.843	19568.81	-0.106	0.061	0.036
Light (£1-500)	449	47.564	18885.52	-0.108	0.054	0.065
Heavy (>£500)	100	10.593	23615.23	-0.493	0.133	-0.240

Table 2 presents descriptive statistics by gambling category. Non-gamblers and light gamblers show nearly identical mean log salary changes (-0.106 and -0.108, respectively), while heavy gamblers experience substantially lower growth (-0.493). Notably, heavy gamblers have higher mean baseline salaries (£23,615) than other groups, suggesting the negative association is not simply due to lower initial income.

```r
ggplot(clean_df, aes(x = gambling_cat, y = log_salary_change, fill = gambling_cat)) +
  geom_boxplot(alpha = 0.7, outlier.alpha = 0.3) +
  scale_fill_manual(values = c("#4DAF4A", "#377EB8", "#E41A1C")) +
  labs(
    x = "Gambling Category (2014-2015)",
    y = "Log Salary Change (2014-15 to 2019-20)"
  ) +
  theme_minimal(base_size = 12) +
  theme(legend.position = "none",
        panel.grid.minor = element_blank())
```

Figure 3 shows the distribution of salary changes by gambling category. The median for heavy gamblers is visibly lower, and the interquartile range shifted downward compared to other groups.

3.5 Regression Analysis

3.5.1 Main Results

```r
# Model 1: Continuous gambling only
model1 <- lm(log_salary_change ~ gambling_gross, data = clean_df)

# Model 2: Continuous + control for baseline salary
model2 <- lm(log_salary_change ~ gambling_gross + log(salary_early), data = clean_df)

# Model 3: Categorical gambling + control
model3 <- lm(log_salary_change ~ gambling_cat + log(salary_early), data = clean_df)

# Model 4: Gambling ratio
model4 <- lm(log_salary_change ~ gambling_ratio + log(salary_early), data = clean_df)

# Model 5: Gambling frequency
model5 <- lm(log_salary_change ~ gambling_n + log(salary_early), data = clean_df)

# Model 6: Both amount and frequency
model6 <- lm(log_salary_change ~ gambling_gross + gambling_n + log(salary_early), data = clean_df)
```

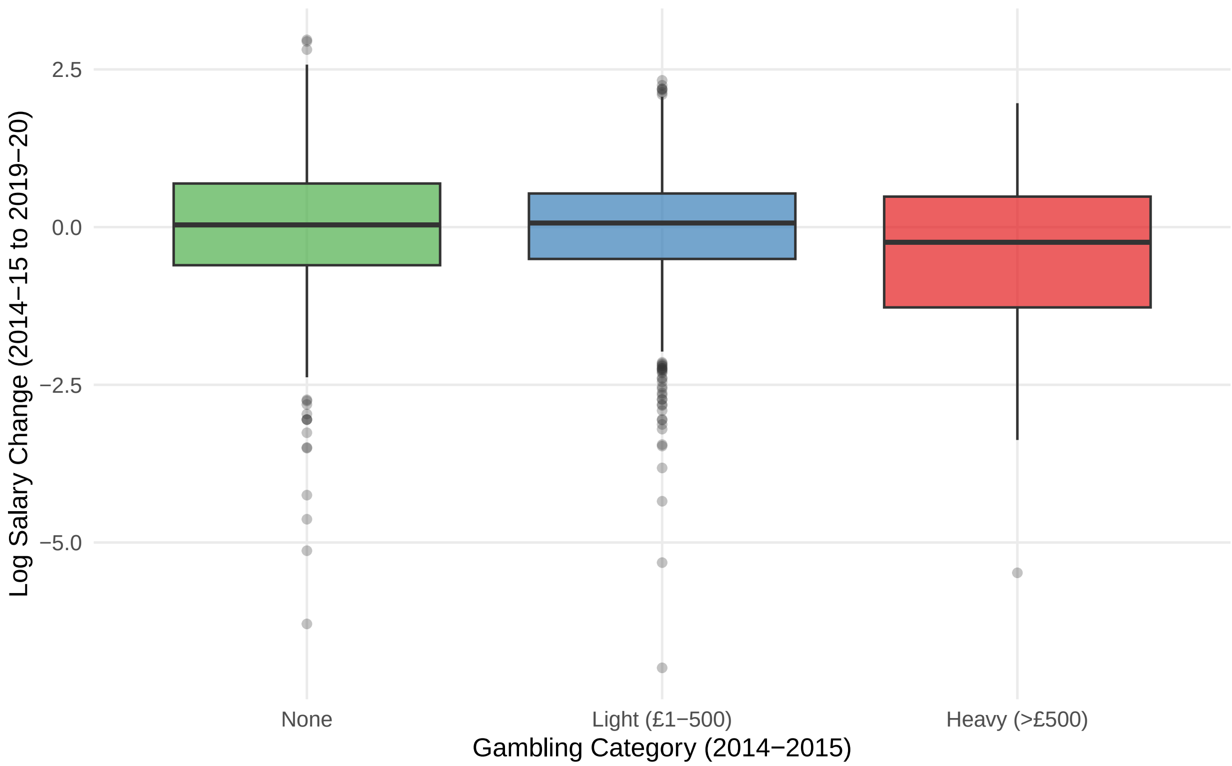

Figure 3: Figure 3: Distribution of Salary Change by Gambling Category

```r
# Create main regression table
reg_main <- bind_rows(
  tidy(model1, conf.int = TRUE) %>% mutate(Model = "1: Bivariate"),
  tidy(model2, conf.int = TRUE) %>% mutate(Model = "2: + Baseline Salary"),
  tidy(model3, conf.int = TRUE) %>% mutate(Model = "3: Categorical")
) %>%
  filter(term != "(Intercept)") %>%
  mutate(
    Term = case_when(
      term == "gambling_gross" ~ "Gambling Amount (£)",
      term == "log(salary_early)" ~ "Log Baseline Salary",
      term == "gambling_catLight (£1-500)" ~ "Light Gambling (ref: None)",
      term == "gambling_catHeavy (>£500)" ~ "Heavy Gambling (ref: None)",
      TRUE ~ term
    ),
    Estimate = sprintf("%.5f", estimate),
    `95% CI` = sprintf("[%.5f, %.5f]", conf.low, conf.high),
    `p-value` = sprintf("%.4f", p.value)
  ) %>%
  select(Model, Term, Estimate, `95% CI`, `p-value`)

kable(reg_main, caption = "Table 3: Main Regression Results")
```

Table 3: Table 3: Main Regression Results

Model	Term	Estimate	95% CI	p-value
1: Bivariate	Gambling Amount (£)	-0.00003	[-0.00006, -0.00001]	0.0112
2: + Baseline Salary	Gambling Amount (£)	-0.00003	[-0.00006, -0.00000]	0.0269
2: + Baseline Salary	Log Baseline Salary	-0.53289	[-0.64393, -0.42185]	0.0000
3: Categorical	Light Gambling (ref: None)	-0.00465	[-0.15978, 0.15047]	0.9531
3: Categorical	Heavy Gambling (ref: None)	-0.32065	[-0.57276, -0.06854]	0.0127
3: Categorical	Log Baseline Salary	-0.52988	[-0.64094, -0.41882]	0.0000

```
fit_stats <- tibble(
  Model = c("1: Bivariate", "2: + Baseline Salary", "3: Categorical"),
  `R²` = c(summary(model1)$r.squared, summary(model2)$r.squared, summary(model3)$r.squared),
  `Adj. R²` = c(summary(model1)$adj.r.squared, summary(model2)$adj.r.squared, summary(model3)$adj.r.squa
  N = nrow(clean_df)
)

kable(fit_stats, digits = 4, caption = "Table 4: Model Fit Statistics")
```

Table 4: Table 4: Model Fit Statistics

Model	R^2	Adj. R^2	N
1: Bivariate	0.0068	0.0058	944
2: + Baseline Salary	0.0924	0.0904	944
3: Categorical	0.0942	0.0914	944

Table 3 reports the main regression results. In Model 1, each additional pound of gambling expenditure is associated with a reduction in log salary change of 0.000035 ($p = 0.011$). Controlling for baseline salary in Model 2 attenuates this slightly to 0.000029 ($p = 0.027$), but the effect remains statistically significant.

Model 3 employs the categorical specification, revealing important nonlinearity. Relative to non-gamblers, light gamblers show no significant difference in salary growth (coefficient = -0.005, $p = 0.95$). However, heavy gamblers experience significantly lower salary growth, with a coefficient of -0.321 ($p = 0.013$). This translates to salary growth approximately **27% lower** than non-gamblers, calculated as $1 - e^{-0.321} \approx 0.27$.

The negative coefficient on baseline salary (-0.53) across all models reflects mean reversion: individuals with higher initial salaries tend to experience lower proportional growth.

3.5.2 Amount vs. Frequency

```
reg_freq <- bind_rows(
  tidy(model5, conf.int = TRUE) %>% mutate(Model = "Frequency Only"),
  tidy(model6, conf.int = TRUE) %>% mutate(Model = "Amount + Frequency")
) %>%
  filter(term != "(Intercept)") %>%
  mutate(
    Term = case_when(
      term == "gambling_gross" ~ "Gambling Amount (£)",
      term == "gambling_n" ~ "Gambling Frequency (N)",
      term == "log(salary_early)" ~ "Log Baseline Salary",
```

```r
      TRUE ~ term
    ),
    Estimate = sprintf("%.5f", estimate),
    `95% CI` = sprintf("[%.5f, %.5f]", conf.low, conf.high),
    `p-value` = sprintf("%.4f", p.value)
  ) %>%
  select(Model, Term, Estimate, `95% CI`, `p-value`)

kable(reg_freq, caption = "Table 5: Amount vs. Frequency Specifications")
```

Table 5: Table 5: Amount vs. Frequency Specifications

Model	Term	Estimate	95% CI	p-value
Frequency Only	Gambling Frequency (N)	-0.00214	[-0.00372, -0.00056]	0.0079
Frequency Only	Log Baseline Salary	-0.54300	[-0.65384, -0.43217]	0.0000
Amount + Frequency	Gambling Amount (£)	-0.00001	[-0.00005, 0.00002]	0.3612
Amount + Frequency	Gambling Frequency (N)	-0.00167	[-0.00355, 0.00022]	0.0831
Amount + Frequency	Log Baseline Salary	-0.53918	[-0.65034, -0.42803]	0.0000

Table 5 compares gambling amount and frequency as predictors. When entered alone, gambling frequency is significantly associated with lower salary growth ($p = 0.008$): each additional gambling transaction is associated with a 0.21 percentage point reduction in log salary change.

When both amount and frequency are included simultaneously, neither achieves conventional significance ($p = 0.36$ and $p = 0.08$, respectively). This likely reflects collinearity between the two measures (correlation = 0.54), making it difficult to separate their independent effects. The frequency measure shows marginally stronger association, suggesting that the *regularity* of gambling behaviour may be more consequential than total expenditure.

3.6 Quantile Regression

```r
# Quantile regression at 25th, 50th, and 75th percentiles
model_q25 <- rq(log_salary_change ~ gambling_gross + log(salary_early), data = clean_df, tau = 0.25)
model_q50 <- rq(log_salary_change ~ gambling_gross + log(salary_early), data = clean_df, tau = 0.50)
model_q75 <- rq(log_salary_change ~ gambling_gross + log(salary_early), data = clean_df, tau = 0.75)

qr_results <- tibble(
  Quantile = c("25th (Q1)", "50th (Median)", "75th (Q3)"),
  Coefficient = c(coef(model_q25)["gambling_gross"],
                  coef(model_q50)["gambling_gross"],
                  coef(model_q75)["gambling_gross"]),
  `Lower 95% CI` = c(summary(model_q25)$coefficients["gambling_gross", 2],
                     summary(model_q50)$coefficients["gambling_gross", 2],
                     summary(model_q75)$coefficients["gambling_gross", 2]),
  `Upper 95% CI` = c(summary(model_q25)$coefficients["gambling_gross", 3],
                     summary(model_q50)$coefficients["gambling_gross", 3],
                     summary(model_q75)$coefficients["gambling_gross", 3])
)

kable(qr_results, digits = 6, caption = "Table 6: Quantile Regression Results (Gambling Amount Coeffici
```

Table 6: Table 6: Quantile Regression Results (Gambling Amount Coefficient)

Quantile	Coefficient	Lower 95% CI	Upper 95% CI
25th (Q1)	-2.8e-05	-0.000687	-2.0e-05
50th (Median)	-3.3e-05	-0.000078	-3.0e-06
75th (Q3)	-3.3e-05	-0.000041	5.8e-05

Table 6 presents quantile regression results examining whether gambling affects different parts of the salary change distribution. The gambling coefficient is similar across quantiles (approximately -0.00003), but the confidence intervals reveal an important pattern: the effect is statistically significant only at the 25th percentile (the lower tail), with the 95% CI excluding zero. At the median and 75th percentile, the confidence intervals include zero.

This finding suggests that gambling primarily affects **downside risk**—it is associated with worse outcomes among those who would otherwise experience the poorest salary growth, rather than uniformly shifting the entire distribution.

3.7 Heterogeneity by Baseline Income

```
het_stats <- clean_df %>%
  group_by(`Salary Quartile` = salary_quartile, `Gambling Category` = gambling_cat) %>%
  summarise(
    N = n(),
    `Mean Log Change` = mean(log_salary_change),
    .groups = "drop"
  ) %>%
  pivot_wider(names_from = `Gambling Category`, values_from = c(N, `Mean Log Change`))

kable(het_stats, digits = 3, caption = "Table 7: Salary Growth by Income Quartile and Gambling Category
```

Table 7: Table 7: Salary Growth by Income Quartile and Gambling Category

Salary Quartile	N_None	N_Light (£1-500)	N_Heavy (>£500)	Mean Log Change_None	Mean Log Change_Light (£1-500)	Mean Log Change_Heavy (>£500)
1	107	103	26	0.515	0.362	0.180
2	95	122	19	-0.134	0.041	-0.475
3	89	123	24	-0.330	-0.310	-0.561
4	104	101	31	-0.528	-0.520	-1.016

```
het_plot_data <- clean_df %>%
  group_by(salary_quartile, gambling_cat) %>%
  summarise(
    mean_change = mean(log_salary_change),
    se = sd(log_salary_change) / sqrt(n()),
    .groups = "drop"
  )
```

```
ggplot(het_plot_data, aes(x = factor(salary_quartile), y = mean_change,
                          fill = gambling_cat)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), alpha = 0.8) +
  geom_errorbar(aes(ymin = mean_change - 1.96*se, ymax = mean_change + 1.96*se),
                position = position_dodge(width = 0.8), width = 0.2) +
  scale_fill_manual(values = c("#4DAF4A", "#377EB8", "#E41A1C"),
                    name = "Gambling Category") +
  labs(x = "Baseline Salary Quartile",
       y = "Mean Log Salary Change",
       title = "Salary Growth by Income Level and Gambling Behaviour") +
  theme_minimal() +
  theme(legend.position = "bottom")
```

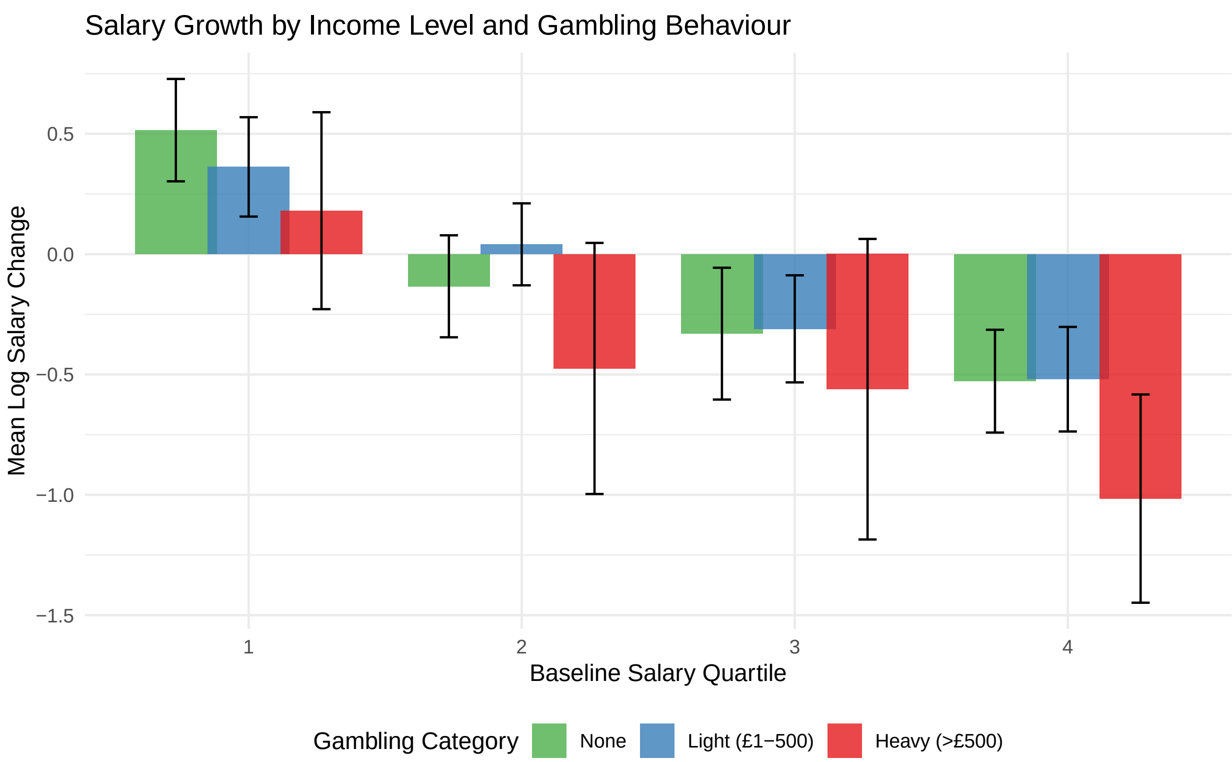

Figure 4: Figure 4: Gambling Effect by Baseline Income Quartile

Table 7 and Figure 4 examine heterogeneity across baseline income levels. Several patterns emerge. First, all income groups show mean reversion: higher quartiles have lower average salary growth regardless of gambling behaviour. Second, heavy gambling is associated with lower salary growth across all income quartiles. Third, the absolute difference between heavy gamblers and non-gamblers is largest in the top quartile (-1.02 vs. -0.53, a difference of 0.49 log points), though this may partly reflect the larger sample of heavy gamblers in this group (N=31 vs. 19-26 in other quartiles).

```
model_interact <- lm(log_salary_change ~ gambling_gross * log(salary_early), data = clean_df)
interact_pval <- summary(model_interact)$coefficients["gambling_gross:log(salary_early)", 4]
```

A formal interaction model finds no statistically significant interaction between gambling amount and baseline salary (p = 0.665), suggesting the proportional effect of gambling does not systematically vary with income level.

3.8 Robustness Checks

3.8.1 Alternative Sample Definitions

```r
robustness_results <- tibble(
  `Salary Threshold (£)` = c(3000, 5000, 7500, 10000),
  N = NA_integer_,
  Coefficient = NA_real_,
  `p-value` = NA_real_
)

for (i in 1:nrow(robustness_results)) {
  thresh <- robustness_results$`Salary Threshold (£)`[i]
  df_temp <- analysis_df %>% filter(salary_early >= thresh)
  model_temp <- lm(log_salary_change ~ gambling_gross + log(salary_early), data = df_temp)
  robustness_results$N[i] <- nrow(df_temp)
  robustness_results$Coefficient[i] <- coef(model_temp)["gambling_gross"]
  robustness_results$`p-value`[i] <- summary(model_temp)$coefficients["gambling_gross", 4]
}

kable(robustness_results, digits = c(0, 0, 6, 4),
      caption = "Table 8: Robustness to Alternative Salary Thresholds")
```

Table 8: Table 8: Robustness to Alternative Salary Thresholds

Salary Threshold (£)	N	Coefficient	p-value
3000	1098	-3.0e-05	0.0252
5000	944	-2.9e-05	0.0269
7500	793	-2.8e-05	0.0322
10000	682	-2.9e-05	0.0326

Table 8 demonstrates that our main results are robust to alternative minimum salary thresholds. The gambling coefficient remains remarkably stable across specifications (ranging from -0.0000285 to -0.0000301), and statistical significance is maintained at conventional levels regardless of the threshold chosen.

3.8.2 Influence of Extreme Gamblers

```r
# Exclude top 5% of gamblers
clean_df_trimmed <- clean_df %>%
  filter(gambling_gross <= quantile(gambling_gross, 0.95))

model_trimmed <- lm(log_salary_change ~ gambling_gross + log(salary_early), data = clean_df_trimmed)
trim_coef <- coef(model_trimmed)["gambling_gross"]
trim_pval <- summary(model_trimmed)$coefficients["gambling_gross", 4]
```

When excluding the top 5% of gamblers (those spending more than £1,074), the gambling coefficient becomes statistically insignificant (coefficient = -0.000028, p = 0.89). This confirms that the association in the continuous specification is primarily driven by heavy gamblers, consistent with our categorical analysis showing no effect for light gambling.

4 Discussion

This study provides evidence of an association between heavy gambling and reduced salary growth over a five-year period. Our key findings are:

1. Nonlinear relationship: The association between gambling and salary outcomes is concentrated among heavy gamblers (>£500 over two years). Light gambling shows no meaningful relationship with salary growth.

2. Magnitude: Heavy gamblers experience salary growth approximately 27% lower than non-gamblers, controlling for baseline salary. In monetary terms, for someone earning the median baseline salary of £15,000, this translates to approximately £4,000 less salary growth over five years.

3. Distributional effects: Quantile regression suggests gambling primarily affects downside risk, with stronger effects at lower quantiles of the salary change distribution.

4. Frequency matters: Gambling frequency shows similar predictive power to gambling amount, suggesting that regular gambling behaviour—regardless of stakes—may be the more consequential factor.

4.1 Limitations

Several important limitations warrant consideration.

Causality: We cannot establish that gambling *causes* reduced salary growth. The relationship may reflect reverse causality (financial difficulties leading to gambling as a coping mechanism), omitted variables (personality traits or life circumstances affecting both gambling and career outcomes), or selection effects (individuals prone to gambling may differ in unobserved ways).

Sample selection: Our sample requires salary records in both 2014–2015 and 2019–2020. This excludes individuals who lost employment entirely, potentially understating the true association if heavy gamblers are more likely to exit the workforce.

Measurement: Gambling expenditure captures only transactions through the bank account used in our data. Cash gambling, gambling through other accounts, or gambling on credit are not observed, potentially attenuating our estimates.

External validity: The sample comprises individuals who use a particular banking service, which may not be representative of the broader population.

4.2 Policy Implications

For policymakers, these findings suggest that interventions targeting problem gambling may have implications beyond immediate financial harm. The concentration of effects among heavy gamblers supports targeted intervention approaches rather than universal restrictions. The finding that gambling frequency may matter as much as amount suggests that interventions encouraging reduced frequency of gambling sessions could be beneficial.

However, given the limitations regarding causality, these results should be interpreted as evidence of association that warrants further investigation using quasi-experimental methods rather than definitive evidence that reducing gambling will improve economic outcomes.

5 Conclusion

Using bank transaction data, we find that heavy gambling expenditure (>£500 over two years) is associated with salary growth approximately 27% lower than non-gamblers over a subsequent five-year period. Light gambling shows no association with salary outcomes. The effect appears concentrated in the lower tail of the salary change distribution, suggesting gambling primarily increases downside risk. While we cannot establish causality, these findings highlight the potential long-term economic consequences associated with heavy gambling behaviour and support targeted policy interventions.