# Problem Set 1

Thomas Gößl, Dzan Hadzimujic, Leonardo Puehler

Labour Economics

November 4, 2025

# Introduction: Dataset

**Data description.** The dataset `ps1_clean_data.Rda` contains simulated labour-market variables:

- `hours`: number of hours worked per day (dependent variable),
- `motivation`: intrinsic motivation of individuals for a successful career, where higher values denote stronger motivation relative to the average person,
- `education`: years of education,
- `wage`: hourly wage,
- `wage_premium`: dummy variable indicating whether an individual received a randomly assigned 35% wage increase (e.g. through an income-support program).

These variables will be analysed throughout this assignment to study labour-supply behaviour.

# Task 1(a): Descriptive Statistics

**Task:** Generate the log of wages (`ln_wage`). Produce a table with descriptive statistics for `education`, `motivation`, `hours` and `ln_wage`. Also calculate correlations between these variables and plot the density of log wages as well as a histogram for the years of education. Briefly comment on your results.

# Task 1(a): Generate log wages and descriptive statistics

R Code

```
# Load data
load("ps1_clean_data.Rda")

# Generate log wages
df1$ln_wage <- log(df1$wage)

# Summary statistics
df1 %>%
  get_summary_stats(
    education, motivation, hours, ln_wage,
    type = "common"
  )
```

# Task 1(a): Summary Statistics

| variable | n | min | max | median | iqr | mean | sd | se | ci |
|---|---|---|---|---|---|---|---|---|---|
| education | 5000.00 | 10.00 | 24.00 | 12.00 | 4.00 | 12.77 | 2.42 | 0.03 | 0.07 |
| motivation | 5000.00 | -4.56 | 5.14 | -0.00 | 1.88 | 0.01 | 1.39 | 0.02 | 0.04 |
| hours | 5000.00 | 6.35 | 9.70 | 7.95 | 0.67 | 7.95 | 0.49 | 0.01 | 0.01 |
| ln_wage | 5000.00 | 2.37 | 4.12 | 3.35 | 0.32 | 3.36 | 0.23 | 0.00 | 0.01 |

Table: Descriptive Statistics

*Comment:* The average individual in the sample has approximately 12.77 years of education and an average log wage of 3.36. Wages are roughly normally distributed, while education shows slight left skewness. Motivation is centered around zero, suggesting a balanced sample in terms of intrinsic motivation.

# Task 1(a): Correlations – R Code

**Goal:** Calculate Pearson correlations between education, motivation, hours, and log wages.

R Code

```
corrs <- cor(df1[, c("education", "motivation", "hours", "ln_wage")],
             use = "pairwise.complete.obs",
             method = "pearson")
```

# Task 1(a): Correlations – Results

| | education | motivation | hours | ln_wage |
|---|---|---|---|---|
| education | 1.00 | 0.14 | 0.65 | 0.51 |
| motivation | 0.14 | 1.00 | 0.82 | 0.35 |
| hours | 0.65 | 0.82 | 1.00 | 0.60 |
| ln_wage | 0.51 | 0.35 | 0.60 | 1.00 |

Table: Correlation Matrix

*Comment:* There is a strong positive correlation between `hours` and `motivation` ($r = 0.82$), suggesting that individuals with higher intrinsic motivation tend to work longer hours. Education also correlates positively with hours ($r = 0.65$) and log wages ($r = 0.51$), indicating that more educated individuals tend to earn more and work more hours. Motivation is moderately correlated with log wages ($r = 0.35$), implying that motivation contributes positively to wage differences, though less strongly than education.

# Task 1(a): Plot the Density of Log Wages – R Code

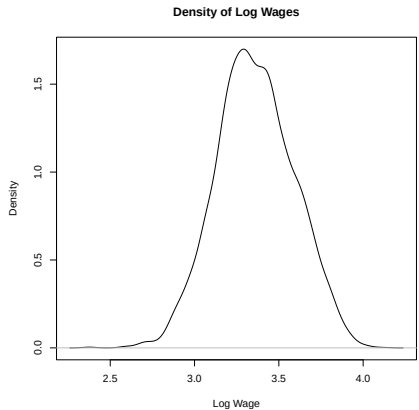
**Goal:** Visualize the distribution of log wages.

R Code

```
# Compute and plot the density of log wages
density_ln_wage <- density(df1$ln_wage)

plot(density_ln_wage,
     main = "Density of Log Wages",
     xlab = "Log Wage",
     ylab = "Density",
     col = "steelblue",
     lwd = 2)
```

# Task 1(a): Density of Log Wages – Results




*Comment:* The distribution of log wages is approximately normal, centered around 3.3, with moderate dispersion. There is only slight skewness, suggesting that most individuals earn wages close to the sample mean.

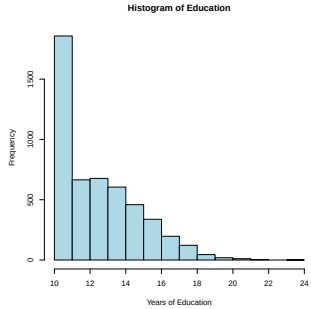
# Task 1(a): Histogram of Education – R Code

**Goal:** Visualize the distribution of years of education.

R Code

```
# Histogram for education (years)
hist(df1$education,
     breaks = 10,
     col = "lightblue",
     main = "Histogram of Education",
     xlab = "Years of Education",
     ylab = "Frequency")
```

# Task 1(a): Histogram of Education – Results




*Comment:* The distribution of education years is slightly right-skewed, with most individuals having between 10 and 14 years of education. A smaller group has completed tertiary education (above 16 years). This pattern suggests a sample where the majority finished secondary schooling, while a smaller portion pursued higher education.

# Task 1(b): Comparing Groups by Motivation and Education

**Task:** Compare descriptive statistics for two groups of individuals:

- those with **motivation** $\geq 0$ versus those with **motivation** $< 0$, and
- those with **education** $\geq 14$ **years** (some college) versus those with **education** $< 14$ **years**.

**Goal:** Investigate whether average hours worked, wages, and motivation levels differ systematically between these subgroups. Generate summary statistics for each group and comment on your findings.

*Hint:* Create dummy variables for the two conditions (`M` and `E`) and use `group_by()` together with `get_summary_stats()` to summarize differences.

# Task 1(b): Group Comparison by Motivation – R Code

**Goal:** Compare descriptive statistics for individuals with `motivation` $\geq 0$ versus those with `motivation < 0`.

## R Code

```
# Create motivation dummy: 1 = motivation >= 0, 0 = motivation < 0
df1$M <- ifelse(df1$motivation >= 0, 1, 0)

# Summary statistics by motivation group
df1 %>%
  group_by(M) %>%
  get_summary_stats(
    education, motivation, hours, ln_wage,
    type = "common"
  )
```

# Task 1(b): Comparison by Motivation – Results and Interpretation

| M | variable | n | min | max | median | iqr | mean | sd | se | ci |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | education | 2501 | 10.00 | 22.00 | 12.00 | 4.00 | 12.48 | 2.30 | 0.05 | 0.09 |
| | motivation | 2501 | -4.56 | -0.00 | -0.92 | 1.14 | -1.10 | 0.83 | 0.02 | 0.03 |
| | hours | 2501 | 6.35 | 9.01 | 7.62 | 0.49 | 7.63 | 0.36 | 0.01 | 0.01 |
| | ln_wage | 2501 | 2.37 | 3.90 | 3.29 | 0.31 | 3.29 | 0.23 | 0.00 | 0.01 |
| 1 | education | 2499 | 10.00 | 24.00 | 13.00 | 4.00 | 13.06 | 2.51 | 0.05 | 0.10 |
| | motivation | 2499 | 0.00 | 5.14 | 0.96 | 1.19 | 1.13 | 0.84 | 0.02 | 0.03 |
| | hours | 2499 | 7.43 | 9.70 | 8.24 | 0.52 | 8.28 | 0.37 | 0.01 | 0.01 |
| | ln_wage | 2499 | 2.83 | 4.12 | 3.42 | 0.30 | 3.42 | 0.21 | 0.00 | 0.01 |

Table: Summary Statistics by Motivation Group

*Comment:* Individuals with higher motivation (`M = 1`) have slightly higher education (mean = 13.1 years vs 12.5 years), work longer hours (8.28 vs 7.63 hours per day), and earn higher log wages (3.42 vs 3.29). This pattern suggests a positive association between motivation and both labor supply and earnings. While the differences are not extreme, wages are on average 13.9% higher, also working 0.65 hours more.

# Task 1(b): Group Comparison by Education – R Code

**Goal:** Compare descriptive statistics for individuals with `education` $\geq 14$ years (some college) versus those with `education` $< 14$.

## R Code

```
# Create education dummy: 1 = education >= 14 years, 0 = below
df1$E <- ifelse(df1$education >= 14, 1, 0)

# Summary statistics by education group
df1 %>%
  group_by(E) %>%
  get_summary_stats(
    education, motivation, hours, ln_wage,
    type = "common"
  )
```

# Task 1(b): Comparison by Education – Results and Interpretation

| E | variable | n | min | max | median | iqr | mean | sd | se | ci |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | education | 3200 | 10.00 | 13.00 | 11.00 | 2.00 | 11.24 | 1.18 | 0.02 | 0.04 |
| | motivation | 3200 | -4.29 | 5.14 | -0.12 | 1.85 | -0.10 | 1.39 | 0.02 | 0.05 |
| | hours | 3200 | 6.35 | 9.19 | 7.75 | 0.56 | 7.76 | 0.41 | 0.01 | 0.01 |
| | ln_wage | 3200 | 2.37 | 3.97 | 3.27 | 0.29 | 3.28 | 0.21 | 0.00 | 0.01 |
| 1 | education | 1800 | 14.00 | 24.00 | 15.00 | 2.00 | 15.48 | 1.53 | 0.04 | 0.07 |
| | motivation | 1800 | -4.56 | 4.17 | 0.20 | 1.85 | 0.22 | 1.37 | 0.03 | 0.06 |
| | hours | 1800 | 7.08 | 9.70 | 8.30 | 0.56 | 8.30 | 0.42 | 0.01 | 0.02 |
| | ln_wage | 1800 | 3.00 | 4.12 | 3.47 | 0.29 | 3.49 | 0.20 | 0.00 | 0.01 |

Table: Summary Statistics by Education Group

*Comment:* Individuals with 14 or more years of education (`E = 1`) have substantially higher average education (15.5 vs 11.2 years), work slightly longer hours (8.30 vs 7.76), and earn higher log wages (3.49 vs 3.28). Motivation is also marginally higher among the more educated group (0.22 vs 0.10). Wages are on average 23.4% higher for highly educated individuals, who also work on average 0.55 hours more. In general, these results indicate that education is positively associated with both the labor supply and the earnings.

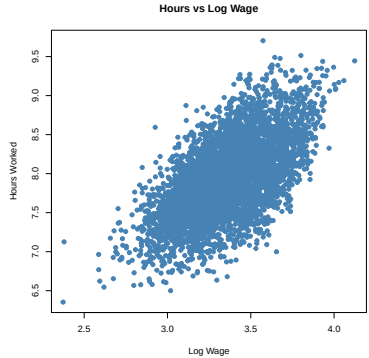
# Task 1(c): Scatterplot of Hours Worked vs Log Wage – R Code

**Goal:** Plot a scatterplot of the number of hours worked against the log of wages.

R Code

```
# Scatterplot: hours worked vs. log wage
plot(df1$ln_wage, df1$hours,
     main = "Scatterplot: Hours Worked vs Log Wage",
     xlab = "Log Wage",
     ylab = "Hours Worked",
     pch = 19, col = rgb(0.2, 0.4, 0.8, 0.4))
```

# Task 1(c): Scatterplot of Hours Worked vs Log Wage – Results




*Comment:* The scatterplot shows a positive relationship between `log(wage)` and `hours worked`. Individuals with higher wages tend to work slightly more hours on average. However, the data points are widely dispersed, suggesting that the relationship is not very strong and that other factors beyond wages may influence the number of hours worked.

# Task 1(d): Simple Regression – R Code

**Goal:** Estimate a simple regression of `hours` on `ln_wage` and add the regression line with its 95% confidence interval to the scatterplot.

**Estimated model:**

$$\text{hours}_i = \beta_0 + \beta_1 \cdot \ln(\text{wage}_i) + \varepsilon_i$$

where $\beta_0$ is the intercept (constant term) and $\beta_1$ measures the marginal effect of log wages on hours worked.

## R Code

```
# Simple OLS regression
simple_OLS <- lm(hours ~ ln_wage, data = df1)
summary(simple_OLS)

# 95% confidence intervals for coefficients
confint(simple_OLS)
```

# Task 1(d): Simple Regression – Results and Interpretation

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.6358 | 0.0807 | 45.05 | 0.0000 |
| ln_wage | 1.2848 | 0.0240 | 53.61 | 0.0000 |

Table: Simple OLS Regression: Hours on Log Wage

*Comment:* The coefficient on `ln_wage` (1.285) indicates a strong and highly significant positive relationship between hourly wages and hours worked. A 1% increase in wages ($\approx 0.01$ in log terms) is associated with an increase of about 0.013 hours per day. This suggests that individuals with higher wages tend to work longer hours, consistent with a positive labor supply elasticity. The high t-value (53.6) and p-value ($< 0.001$) confirm strong statistical significance.

# Task 1(d): Regression Line – R Code

**Goal:** Add the regression line and 95% confidence interval to the scatterplot.

## R Code

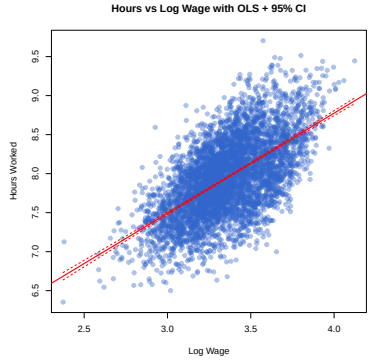
```
# Scatterplot with regression line and confidence band
plot(df1$ln_wage, df1$hours,
     main = "Hours Worked vs Log Wage with Regression Line",
     xlab = "Log Wage", ylab = "Hours Worked",
     pch = 19, col = rgb(0.2, 0.4, 0.8, 0.4))

# Add regression line
abline(simple_OLS, col = "red", lwd = 2)

# 95% confidence band
newdata <- data.frame(ln_wage = seq(min(df1$ln_wage),
                                    max(df1$ln_wage), length = 100))
pred <- predict(simple_OLS, newdata, interval = "confidence")

lines(newdata$ln_wage, pred[, "lwr"], col = "red", lty = 2)
lines(newdata$ln_wage, pred[, "upr"], col = "red", lty = 2)
```

# Task 1(d): Hours Worked vs Log Wage – OLS with 95% Confidence Interval




*Comment:* The scatterplot with the red regression line shows a clear positive relationship between log wages and hours worked. The shaded 95% confidence band indicates that this effect is statistically significant and precisely estimated. Individuals with higher wages tend to work longer hours, although the effect size remains moderate, suggesting that factors beyond wages also influence labor supply.

# Task 1(e): Adding `education` to our regression - R Code

**Goal:** Extend the model by including `education` as an additional regressor.

**Estimated model:**

$$\text{hours}_i = \beta_0 + \beta_1 \cdot \ln(\text{wage}_i) + \beta_2 \cdot \textit{education}_i + \varepsilon_i$$

where $\beta_2$ is the marginal effect of one additional year of education on worked hours.

## R Code

```
# Multiple regression with education
model2 <- lm(hours ~ ln_wage + education, data = df1)
summary(model2)
```

# Task 1(e): Adding `education` to our regression

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.1337 | 0.0714 | 57.90 | 0.0000 |
| ln_wage | 0.7866 | 0.0243 | 32.34 | 0.0000 |
| education | 0.0921 | 0.0023 | 39.89 | 0.0000 |

Table: OLS Regression - Hours on Log Wage and Education

*Comment*: By adding education into the regression, the explanatory power of the log wages is reduced. In the univariate model, a percentual change in wages implies 0.0128 more hours worked, while in the multivariate model, it implies only 0.0078 additional hours. One more year of education implies 0.09 additional hours of work. Both models show significant estimators at 5%, with the second model fitting the data better ($R^2 = 0.365 \rightarrow R^2 = 0.518$).

# Task 1(f): Adding `motivation` to our multivariate regression - R Code

**Goal:** Do the full multivariate regression of `hours` on ln_wage, `education`, and `motivation`.

**Estimated model:**

$$\text{hours}_i = \beta_0 + \beta_1 \cdot \ln(\text{wage}_i) + \beta_2 \cdot \textit{education}_i + \beta_3 \cdot \textit{motivation}_i + \varepsilon_i$$

where $\beta_3$ is the marginal effect of one additional unit of motivation on worked hours.

## R Code

```
# Multiple regression with education and motivation
model3 <- lm(hours ~ ln_wage + education + motivation,data = df1)
summary(model3)
```

# Task 1(f): Adding `motivation` to our multivariate regression

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 5.9681690 | 0.0223144 | 267.46 | 0.0000 |
| ln_wage | 0.2099211 | 0.0075306 | 27.88 | 0.0000 |
| education | 0.0999287 | 0.0006756 | 147.91 | 0.0000 |
| motivation | 0.2499828 | 0.0010803 | 231.39 | 0.0000 |

Table: OLS Regression - Hours on Log Wage, Education, and Motivation

*Comment*: Considering that both `education` and `motivation` are important variables to explain the amount of hours worked, the third model presents the most complete version of the regression. In the univariated version of the model, the regressor associated to `ln_wage` may appear overestimated, or biased, considering the absence of other relevant variables. ($R^2 = 0.95$)