Handout 5: Introduction to Simple Linear Regression

EC 282: Introduction to Econometrics

Spring 2026

Instructions: Run the provided R code and answer the questions. Show your work for calculations.

1 Setup: The Gapminder Data

Does national income buy longer lives? We use the Gapminder dataset—a cross-section of 142 countries in 2007—to explore the relationship between GDP per capita and life expectancy.

Run the code below to load and prepare the data.

```r
library(ggplot2)

# Load the Gapminder data directly from the web
gapminder_full <- read.csv(
  "https://raw.githubusercontent.com/resbaz/r-novice-gapminder-files/
   master/data/gapminder-FiveYearData.csv"
)

# Keep only the year 2007 (a single cross-section)
gap07 <- gapminder_full[gapminder_full$year == 2007, ]

cat("Number of countries:", nrow(gap07), "\n")
cat("Variables:", paste(names(gap07), collapse = ", "), "\n")
summary(gap07[, c("lifeExp", "gdpPercap")])
```

Question 1.1: Create a scatterplot of life expectancy (Y) against GDP per capita (X). What do you notice about the shape of the relationship?

```r
ggplot(gap07, aes(x = gdpPercap, y = lifeExp)) +
  geom_point(color = "steelblue", size = 2, alpha = 0.7) +
  labs(title = "Life Expectancy vs. GDP per Capita (2007)",
       x = "GDP per Capita (USD)",
       y = "Life Expectancy (years)") +
  theme_minimal()
```

Question 1.2: Compute the sample means $\bar{X}$ and $\bar{Y}$, sample standard deviations S_X and S_Y, and the sample correlation r_{XY}.

```r
X <- gap07$gdpPercap
Y <- gap07$lifeExp
n <- length(Y)

```

```r
x_bar <- mean(X)
y_bar <- mean(Y)
s_x   <- sd(X)
s_y   <- sd(Y)
r_xy  <- cor(X, Y)

cat("n:", n, "\n")
cat("X_bar (mean GDP/cap):", round(x_bar, 2), "\n")
cat("Y_bar (mean life exp):", round(y_bar, 2), "\n")
cat("S_X:", round(s_x, 2), "\n")
cat("S_Y:", round(s_y, 2), "\n")
cat("r_XY:", round(r_xy, 4), "\n")
```

What does the sign and magnitude of r_{XY} tell you about the relationship between GDP per capita and life expectancy?

2 Ordinary Least Squares (OLS)

Question 2.1: Write down the population regression model relating life expectancy to GDP per capita. Define each component.

Question 2.2: Using the OLS formulas, compute $\hat{\beta}_1$ and $\hat{\beta}_0$ by hand:

$$\hat{\beta}_1 = \frac{S_{XY}}{S_X^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

```r
# Step-by-step OLS computation
s_xy <- cov(X, Y)          # sample covariance
s_x2 <- var(X)             # sample variance of X

beta1_hat <- s_xy / s_x2
beta0_hat <- y_bar - beta1_hat * x_bar

cat("Sample covariance S_XY:", round(s_xy, 2), "\n")
cat("Sample variance S_X^2:", round(s_x2, 2), "\n\n")
cat("beta1_hat:", round(beta1_hat, 6), "\n")
cat("beta0_hat:", round(beta0_hat, 4), "\n")
```

Now verify with R's built-in regression function:

```r
reg <- lm(lifeExp ~ gdpPercap, data = gap07)
summary(reg)
```

Do your hand calculations match the output from `lm()`?

Question 2.3: Interpret $\hat{\beta}_1$ in a sentence. Is the magnitude of this coefficient large or small? (*Hint:* Think about a realistic change in GDP per capita, e.g., \$1,000 or \$10,000.)

Question 2.4: Interpret $\hat{\beta}_0$. Does the intercept have a meaningful interpretation in this context?

Question 2.5: Predict the life expectancy for a country with GDP per capita of \$5,000 and for a country with GDP per capita of \$40,000.

```
pred_5k  <- beta0_hat + beta1_hat * 5000
pred_40k <- beta0_hat + beta1_hat * 40000

cat("Predicted life exp at GDP/cap = $5,000: ",
    round(pred_5k, 2), "years\n")
cat("Predicted life exp at GDP/cap = $40,000:",
    round(pred_40k, 2), "years\n")
```

Question 2.6: Add the OLS regression line to your scatterplot. Does a straight line seem like a good summary of the data?

```
ggplot(gap07, aes(x = gdpPercap, y = lifeExp)) +
  geom_point(color = "steelblue", size = 2, alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE,
              color = "red", linewidth = 1) +
  labs(title = "OLS Regression: Life Expectancy on GDP per Capita",
       x = "GDP per Capita (USD)",
       y = "Life Expectancy (years)") +
  theme_minimal()
```

3 Predicted Values and Residuals

Question 3.1: Compute the predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ for every observation.

```
gap07$Y_hat <- beta0_hat + beta1_hat * X
gap07$u_hat <- Y - gap07$Y_hat

# Show a few countries
subset(gap07, country %in% c("United States", "China",
                             "Nigeria", "Norway", "Brazil"),
       select = c(country, lifeExp, gdpPercap, Y_hat, u_hat))
```

Pick one country from the table. Explain what its residual means in plain language.

Question 3.2: Which countries have the largest positive and largest negative residuals? What might explain why these countries deviate from the regression line?

```
# Top 5 positive residuals (living longer than predicted)
head(gap07[order(-gap07$u_hat),
     c("country", "lifeExp", "gdpPercap", "u_hat")], 5)

# Top 5 negative residuals (living shorter than predicted)
head(gap07[order(gap07$u_hat),
     c("country", "lifeExp", "gdpPercap", "u_hat")], 5)
```

4 Measures of Fit

Question 4.1: Compute the Total Sum of Squares (TSS), Explained Sum of Squares (ESS), and Residual Sum of Squares (RSS):

$$TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \qquad RSS = \sum_{i=1}^{n}\hat{u}_i^2$$

```
TSS <- sum((Y - y_bar)^2)
ESS <- sum((gap07$Y_hat - y_bar)^2)
RSS <- sum(gap07$u_hat^2)

cat("TSS:", round(TSS, 2), "\n")
cat("ESS:", round(ESS, 2), "\n")
```

```r
cat("RSS:", round(RSS, 2), "\n")
cat("ESS + RSS:", round(ESS + RSS, 2), "\n")
```

Verify that $TSS = ESS + RSS$. Explain in your own words what each quantity measures.

Question 4.2: Compute the coefficient of determination R^2:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

```r
R2 <- ESS / TSS

cat("R-squared:", round(R2, 4), "\n")
cat("R-squared (alternative):", round(1 - RSS/TSS, 4), "\n")
cat("r_XY squared:", round(r_xy^2, 4), "\n")
```

(a) Interpret R^2 in a sentence.

(b) Verify that $R^2 = r_{XY}^2$ in simple regression. Why does this relationship hold?

Question 4.3: Compute the Standard Error of the Regression (SER):

$$SER = \sqrt{\frac{RSS}{n-2}}$$

```r
SER <- sqrt(RSS / (n - 2))
cat("SER:", round(SER, 4), "years\n")
```

What does the SER tell you about the typical size of the prediction errors?

5 Properties of OLS Residuals

OLS residuals satisfy four important algebraic properties *by construction*. Let's verify each one with our data.

Question 5.1: Property 1: The sample mean of the residuals is zero.

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$$

```
cat("Mean of residuals:", mean(gap07$u_hat), "\n")
```

Why must this always be true for OLS? (*Hint:* Think about how $\hat{\beta}_0$ is derived.)

Question 5.2: Property 2: The sample mean of predicted values equals the sample mean of Y.

$$\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \bar{Y}$$

```
cat("Mean of Y_hat:", round(mean(gap07$Y_hat), 4), "\n")
cat("Mean of Y:    ", round(y_bar, 4), "\n")
```

Question 5.3: Property 3: The residuals are uncorrelated with X (and therefore also with $\hat{Y}_i$, since $\hat{Y}_i$ is a linear function of X_i).

$$\sum_{i=1}^{n}\hat{u}_i X_i = 0$$

```
cat("Sum of u_hat * X:", sum(gap07$u_hat * X), "\n")
cat("Correlation(u_hat, X):", cor(gap07$u_hat, X), "\n")
cat("Correlation(u_hat, Y_hat):", cor(gap07$u_hat, gap07$Y_hat), "\n")
```

Why is this property important? What would it mean if the residuals were correlated with X?

Question 5.4: Property 4: $TSS = ESS + RSS$. Verify this using the values you computed in Section 4.

```
cat("TSS:", round(TSS, 2), "\n")
cat("ESS + RSS:", round(ESS + RSS, 2), "\n")
cat("Equal?", all.equal(TSS, ESS + RSS), "\n")
```

Why does this decomposition hold? (*Hint:* Use the fact that $Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + \hat{u}_i$ and Property 3.)

Question 5.5: Visualize the residuals. A residual plot helps us assess whether the linear model is appropriate.

```
ggplot(gap07, aes(x = gdpPercap, y = u_hat)) +
  geom_point(color = "steelblue", size = 2, alpha = 0.7) +
  geom_hline(yintercept = 0, color = "red",
             linetype = "dashed") +
  labs(title = "Residuals vs. GDP per Capita",
       x = "GDP per Capita (USD)",
       y = "Residual (years)") +
  theme_minimal()
```

Do you see any pattern in the residuals? What does this suggest about the linear model?

6 Bringing It Together

Question 6.1: We showed that $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$. This means the OLS regression line always passes through the point $(\bar{X}, \bar{Y})$. Verify this visually:

```
ggplot(gap07, aes(x = gdpPercap, y = lifeExp)) +
  geom_point(color = "steelblue", size = 2, alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE,
              color = "red", linewidth = 1) +
  annotate("point", x = x_bar, y = y_bar,
           color = "red", size = 5, shape = 18) +
  annotate("text", x = x_bar + 3000, y = y_bar - 2,
           label = paste0("(X_bar, Y_bar) = (",
                          round(x_bar, 0), ", ",
                          round(y_bar, 1), ")"),
           color = "red", hjust = 0) +
  labs(title = "The OLS Line Passes Through the Point of Means",
       x = "GDP per Capita (USD)",
       y = "Life Expectancy (years)") +
  theme_minimal()
```

Question 6.2: Suppose a new country enters the dataset with GDP per capita of \$20,000. Using the regression results, predict its life expectancy and construct a range of plausible values using the SER (i.e., $\hat{Y} \pm SER$).

```
pred_20k <- beta0_hat + beta1_hat * 20000
cat("Predicted life expectancy:", round(pred_20k, 2), "years\n")
cat("Plausible range:", round(pred_20k - SER, 2), "to",
    round(pred_20k + SER, 2), "years\n")
```

Question 6.3: Based on everything you've seen in this handout, what are the limitations of fitting a straight line to the GDP–life expectancy relationship? What might we do differently? (We will address this in future handouts.)