Handout 5: Introduction to Simple Linear Regression

ANSWER KEY

EC 282: Introduction to Econometrics

Spring 2026

1 Setup: The Gapminder Data

Question 1.1: Describe the scatterplot.

The scatterplot shows a **positive** relationship between GDP per capita and life expectancy: richer countries tend to have longer-lived populations. However, the relationship is **not linear**—it is concave (curved). Life expectancy rises steeply at low levels of GDP per capita but flattens out at higher income levels. This suggests that the marginal "health benefit" of additional income is larger for poor countries than for rich ones.

There are also notable outliers—some countries have much lower (or higher) life expectancy than their income level would suggest.

Question 1.2: Sample statistics and correlation.

Run the R code to obtain the exact values. You should find:

- $\bar{X}$ (mean GDP per capita) $\approx$ \$11,680
- $\bar{Y}$ (mean life expectancy) $\approx$ 67.0 years
- $S_X \approx$ \$12,860 (large spread—some countries are very poor, others very rich)
- $S_Y \approx$ 12.1 years
- $r_{XY} \approx 0.68$

The correlation is **positive and moderately strong**. Countries with higher GDP per capita tend to have higher life expectancy. However, $r_{XY} = 0.68$ is not close to 1, meaning income alone does not fully explain variation in life expectancy. Other factors (healthcare systems, education, disease burden, inequality) also matter.

2 Ordinary Least Squares (OLS)

Question 2.1: Population regression model.

The population regression model is:

$$\text{LifeExp}_i = \beta_0 + \beta_1 \times \text{GDPperCap}_i + u_i$$

where:

- LifeExp_i = life expectancy in country i (dependent variable, Y)
- GDPperCap_i = GDP per capita in country i (independent variable, X)
- β_0 = intercept (predicted life expectancy when GDP per capita is zero)
- β_1 = slope (change in life expectancy for a one-dollar increase in GDP per capita)
- u_i = error term (all other factors: healthcare quality, education, disease, culture, etc.)

Question 2.2: OLS computation.

Using the formulas:

$$\hat{\beta}_1 = \frac{S_{XY}}{S_X^2} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Run the R code to obtain the exact values. You should find $\hat{\beta}_1 \approx 0.0006371$ and $\hat{\beta}_0 \approx 59.57$.

The values from the hand calculation match the output from `lm()` exactly. This confirms that `lm()` is simply computing the OLS formulas.

Question 2.3: Interpret $\hat{\beta}_1$.

$\hat{\beta}_1 \approx 0.000637$ means: a one-dollar increase in GDP per capita is associated with an increase in life expectancy of about 0.000637 years. This sounds tiny, but a one-dollar change in GDP per capita is a trivially small change.

A more meaningful interpretation: a **$10,000 increase** in GDP per capita is associated with an increase of about $0.000637 \times 10{,}000 \approx 6.4$ years of life expectancy. That is a substantial difference—roughly the gap between a middle-income and high-income country.

Important: This is an *association*, not a causal claim. We cannot say that increasing a country's GDP per capita by $10,000 would *cause* life expectancy to rise by 6.4 years.

Question 2.4: Interpret $\hat{\beta}_0$.

$\hat{\beta}_0 \approx 59.57$ is the predicted life expectancy for a country with GDP per capita of $0. While no country has zero GDP per capita, the intercept serves a mathematical purpose: it anchors the regression line so that it passes through the point of means $(\bar{X}, \bar{Y})$.

In this case, the intercept is not entirely unreasonable—a very poor country might indeed have life expectancy around 60 years. But we should be cautious about interpreting predictions at values of X far outside the range of our data.

Question 2.5: Predictions.

- At GDP/cap = \$5,000: $\hat{Y} = 59.57 + 0.000637 \times 5{,}000 \approx 62.8$ years
- At GDP/cap = \$40,000: $\hat{Y} = 59.57 + 0.000637 \times 40{,}000 \approx 85.0$ years

The prediction at \$5,000 seems plausible. The prediction at \$40,000 may be too high (very few countries have life expectancy above 83). This reflects the non-linear nature of the true relationship: the linear model overpredicts at high income levels.

Question 2.6: Does a straight line fit well?

The straight line captures the general upward trend but misses the curvature in the data:

- It **underpredicts** life expectancy for middle-income countries
- It **overpredicts** for the richest countries
- It **overpredicts** for some of the poorest countries (where life expectancy is very low)

A straight line is a reasonable first approximation, but a non-linear specification (e.g., using the logarithm of GDP per capita) would likely fit better. We will explore this in future handouts.

3 Predicted Values and Residuals

Question 3.1: Interpret a residual.

Each residual $\hat{u}_i = Y_i - \hat{Y}_i$ is the difference between the **actual** life expectancy and the life expectancy **predicted** by our regression.

For example, China has $Y_i = 72.96$ years and $\hat{Y}_i = 62.73$ years, so $\hat{u}_i = 72.96 - 62.73 = 10.23$ years. This means China's life expectancy is about 10.2 years *higher* than what our model predicts based on its GDP per capita alone. Something beyond income—perhaps public health infrastructure, education, or cultural factors—is helping China outperform its income level.

A **positive residual** means the country lives longer than predicted; a **negative residual** means shorter.

Question 3.2: Largest residuals.

Largest positive residuals (living much longer than income predicts): These are typically countries with strong public health systems or cultural factors that promote longevity despite modest income levels (e.g., Cuba, Costa Rica, some Asian countries).

Largest negative residuals (living much shorter than income predicts): These are often countries affected by the HIV/AIDS epidemic (e.g., Swaziland, Zambia, Lesotho, Mozambique) or countries with conflict/instability (e.g., Angola).

The residuals represent the influence of **everything not captured by GDP per capita**—health policy, disease burden, inequality, governance, education, etc. These are all part of the error term u_i.

4 Measures of Fit

Question 4.1: TSS, ESS, RSS.

Run the R code to obtain the values. You should verify that $TSS = ESS + RSS$ (up to rounding).

- **TSS** (Total Sum of Squares): Measures the *total* variation in life expectancy across countries. This is the variation we are trying to explain.
- **ESS** (Explained Sum of Squares): Measures how much of that variation is *explained* by GDP per capita through the regression line.
- **RSS** (Residual Sum of Squares): Measures the *unexplained* variation—what's left over after the regression. This is what OLS minimizes.

The decomposition $TSS = ESS + RSS$ says: total variation = explained variation + unexplained variation.

Question 4.2: R^2.

(a) $R^2 \approx 0.46$. This means that GDP per capita explains about **46% of the variation** in life expectancy across countries. The remaining 54% is due to other factors not included in the model.

An R^2 of 0.46 is moderate. It tells us that income is an important predictor of life expectancy, but far from the only one.

(b) In simple linear regression (one regressor), $R^2 = r_{XY}^2$. This holds because:

$$R^2 = \frac{ESS}{TSS} = r_{XY}^2$$

In simple regression, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ is a linear function of X_i, so the correlation between Y and $\hat{Y}$ reduces to the correlation between Y and X. Verify: $r_{XY} \approx 0.68$, and $0.68^2 \approx 0.46 = R^2$.

Question 4.3: SER.

$SER \approx 8.9$ years. This measures the **typical size of the prediction errors** in our regression. When we use GDP per capita to predict a country's life expectancy, a prediction error of roughly 9 years is typical.

This is a fairly large error relative to the standard deviation of Y ($S_Y \approx 12.1$ years). It confirms that while GDP per capita helps predict life expectancy, there is substantial unexplained variation.

5 Properties of OLS Residuals

Question 5.1: Property 1: Mean of residuals is zero.

The mean of the residuals should be exactly zero (or extremely close, up to machine precision). This is guaranteed by construction: OLS chooses $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, which ensures:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X} = 0$$

This is an **algebraic fact**, not an assumption. It holds for any OLS regression.

Question 5.2: Property 2: Mean of $\hat{Y}$ equals $\bar{Y}$.

Since $Y_i = \hat{Y}_i + \hat{u}_i$ and the residuals average to zero:

$$\bar{Y} = \frac{1}{n}\sum \hat{Y}_i + \frac{1}{n}\sum \hat{u}_i = \overline{\hat{Y}} + 0$$

The R output confirms that the mean of the predicted values equals the mean of the actual values.

Question 5.3: Property 3: Residuals are uncorrelated with X (and $\hat{Y}$).

$\sum \hat{u}_i X_i = 0$ and $\text{Corr}(\hat{u}_i, X_i) = 0$. This follows from the OLS first-order condition (the derivative of the sum of squared residuals with respect to $\hat{\beta}_1$ set to zero).

Since $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ is a linear function of X_i, the residuals are also uncorrelated with $\hat{Y}_i$. The R output confirms both.

Why it matters: If the residuals were correlated with X, it would mean that X still has predictive power that we haven't captured. OLS exhausts all the linear predictive power of X by construction. This is analogous to "squeezing all the juice" out of X.

Question 5.4: Property 4: $TSS = ESS + RSS$.

Write $Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + \hat{u}_i$. Squaring and summing:

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum \hat{u}_i^2 + 2\sum \hat{u}_i(\hat{Y}_i - \bar{Y})$$

The cross-term vanishes because the residuals are uncorrelated with $\hat{Y}_i$ (Property 3):

$$\sum \hat{u}_i(\hat{Y}_i - \bar{Y}) = \sum \hat{u}_i \hat{Y}_i - \bar{Y}\sum \hat{u}_i = 0 - 0 = 0$$

Therefore $TSS = ESS + RSS$. This is verified numerically in Section 4.

Question 5.5: Residual plot.

The residual plot reveals a **clear pattern**:

- For low GDP per capita, residuals are **spread widely** (both very positive and very negative)
- For high GDP per capita, residuals tend to be **negative** (the model overpredicts)
- The residuals are not randomly scattered around zero—they show a curved pattern

This pattern suggests that a linear model is **not the best specification**. The curvature in the residual plot mirrors the concave relationship we saw in the scatterplot. A logarithmic transformation of GDP per capita would likely produce more randomly scattered residuals.

Note: Even though the residuals are uncorrelated with X (Property 3 holds exactly), they can still show a *non-linear* pattern. Zero correlation means no *linear* relationship, but non-linear patterns can remain.

6 Bringing It Together

Question 6.1: The regression line passes through $(\bar{X}, \bar{Y})$.

This follows directly from $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$. Plugging $X = \bar{X}$ into the regression:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1\bar{X} = (\bar{Y} - \hat{\beta}_1\bar{X}) + \hat{\beta}_1\bar{X} = \bar{Y}$$

The red diamond on the plot confirms that the regression line passes through the point of means.

Question 6.2: Prediction for GDP/cap = $20,000.

$$\hat{Y} = 59.57 + 0.000637 \times 20{,}000 \approx 72.3 \text{ years}$$

Using the SER as a rough guide, a plausible range is approximately 72.3 ± 8.9, or roughly 63.4 to 81.2 years.

This is *not* a formal confidence interval (we will learn those later), but it gives a sense of how precise our predictions are. The range is quite wide, reinforcing that GDP per capita alone is an imperfect predictor.

Question 6.3: Limitations and what to do differently.

Key limitations of the linear specification:

1. **Non-linearity**: The true relationship between GDP per capita and life expectancy is concave, not linear. The linear model overpredicts for rich countries and misses the steep gains at low incomes.
2. **Omitted variables**: Many factors affect life expectancy beyond GDP per capita (healthcare spending, education, disease burden, inequality, governance). These are all in u_i.
3. **Causality**: We cannot claim that higher GDP *causes* longer lives. The correlation could be driven by reverse causality (healthier populations are more productive) or confounding variables.

What we might do differently:

- Use log(GDP per capita) instead of GDP per capita—this captures the diminishing returns pattern and often produces a much better fit (we will cover logarithmic transformations later)
- Add more regressors (multiple regression) to reduce omitted variable bias
- Use causal inference techniques to move beyond association