Handout 6: Inference and Binary Regressors

ANSWER KEY

EC 282: Introduction to Econometrics

Spring 2026

1 Setup: Continuing with the Gapminder Data

No questions in this section—just run the setup code from the questions handout to reload the data and re-estimate the regression.

2 Reading the Regression Output

Question 2.1: Identify values from the regression output.

Run `summary(reg)` to obtain:

(a) $\hat{\beta}_0 \approx 59.57$ (intercept), $\hat{\beta}_1 \approx 0.000637$ (slope on GDP per capita)

(b) $SE(\hat{\beta}_1) \approx 0.0000653$

(c) t-statistic for $\hat{\beta}_1$: $t \approx 9.75$

(d) p-value for $\hat{\beta}_1$: $p < 0.0001$ (essentially zero)

(e) $R^2 \approx 0.4636$

The key insight: all of these numbers are reported in the `Coefficients:` table of the summary output. The columns are `Estimate`, `Std. Error`, `t value`, and `Pr(>|t|)`.

Question 2.2: What does the standard error tell you?

The standard error measures the **uncertainty** (or imprecision) in our estimate of $\hat{\beta}_1$.

- A **small** standard error means our estimate is precise—if we drew a new random sample, we'd likely get a similar value of $\hat{\beta}_1$.
- A **large** standard error means our estimate is imprecise—different samples could give quite different estimates.

Think of it this way: $\hat{\beta}_1$ is our best guess for the true slope β_1, and $SE(\hat{\beta}_1)$ tells us roughly how far off that guess might be.

Question 2.3: What affects precision?

Three factors determine the precision of $\hat{\beta}_1$:

1. **Sample size (n):** Increasing n makes $SE(\hat{\beta}_1)$ *smaller* (more precise). More data = more information = better estimates. This is analogous to the standard error of the sample mean shrinking as n grows.
2. **Spread of X ($\mathbf{Var}(X_i)$):** Increasing the spread of X makes $SE(\hat{\beta}_1)$ *smaller*. If all countries had similar GDP per capita, it would be hard to trace out how life expectancy changes with income. You need variation in X to estimate the slope precisely.
3. **Noise in the regression ($\mathbf{Var}(u_i)$):** Increasing the noise makes $SE(\hat{\beta}_1)$ *larger* (less precise). When there are many other factors affecting Y that aren't captured by X, the data points are more scattered around the line, making it harder to pin down the slope.

3 Hypothesis Testing

Question 3.1: Hypotheses.

$$H_0 : \beta_1 = 0 \qquad \text{(GDP per capita has no relationship with life expectancy)}$$
$$H_A : \beta_1 \neq 0 \qquad \text{(GDP per capita is related to life expectancy)}$$

This is a **two-sided** test because we are looking for a relationship in either direction.

Question 3.2: Compute the t-statistic by hand.

Under H_0, $\beta_{1,0} = 0$, so:

$$t\text{-stat} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{0.000637}{0.0000653} \approx 9.75$$

This matches the t-statistic in the `summary()` output. The t-statistic tells us that $\hat{\beta}_1$ is about 9.75 standard errors away from zero. This is very far from zero, suggesting H_0 is unlikely.

Question 3.3: Statistical significance.

(a) At the 5% level: $|t| = 9.75 > 1.96$. **Yes**, the slope is statistically significant at the 5% level. We reject H_0.

(b) At the 1% level: $|t| = 9.75 > 2.576$. **Yes**, the slope is also statistically significant at the 1% level. The evidence against H_0 is overwhelming.

Question 3.4: P-value interpretation.

The p-value is essentially 0 (far less than 0.01). This means: **if there were truly no relationship between GDP per capita and life expectancy ($\beta_1 = 0$), the probability of observing a slope estimate as large as ours (or larger) by pure chance is essentially zero.**

In plain language: the data provide overwhelming evidence that GDP per capita and life expectancy are related. It would be virtually impossible to see this strong a pattern in a sample of 142 countries if the two variables were truly unrelated in the population.

Question 3.5: One-sided vs. two-sided p-value.

The one-sided p-value is exactly **half** the two-sided p-value. This is because the two-sided test looks for extreme values in both tails (both very positive and very negative t-statistics), while the one-sided test only looks in one tail.

Since we have strong evidence that $\hat{\beta}_1 > 0$ (the t-statistic is positive and very large), the one-sided test is even more decisive. But in this case, both p-values are so tiny that the distinction is irrelevant—we reject H_0 either way.

4 Confidence Intervals

Question 4.1: 95% CI by hand.

$$\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1) = 0.000637 \pm 1.96 \times 0.0000653$$
$$= 0.000637 \pm 0.000128 = [0.000509,\ 0.000765]$$

The R output from `confint()` will be very similar (slight differences due to R using the t-distribution rather than the normal approximation, but they are negligible with $n = 142$).

Question 4.2: Interpret the 95% CI.

We are 95% confident that the true effect of a one-dollar increase in GDP per capita on life expectancy lies between 0.000509 and 0.000765 years.

The interval does **not include zero**. This is consistent with rejecting $H_0 : \beta_1 = 0$ at the 5% significance level. If zero were a plausible value for β_1, it would be inside the confidence interval.

In more practical terms: a \$10,000 increase in GDP per capita is associated with an increase in life expectancy of between 5.1 and 7.7 years (multiply the CI bounds by 10,000).

Question 4.3: 99% CI.

$$\hat{\beta}_1 \pm 2.576 \times SE(\hat{\beta}_1) = 0.000637 \pm 2.576 \times 0.0000653$$
$$= 0.000637 \pm 0.000168 = [0.000469,\ 0.000805]$$

The 99% CI is **wider** than the 95% CI. This makes intuitive sense: if we want to be *more confident* that our interval contains the true value, we need to cast a wider net. There is a tradeoff between confidence level and precision.

Question 4.4: Connection between CIs and hypothesis testing.

A 95% confidence interval contains all the values of $\beta_{1,0}$ that we would *fail to reject* in a two-sided test at the 5% level. Conversely, any value *outside* the CI would be rejected.

This works because the CI is constructed as $\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)$, and the hypothesis test rejects when $|\hat{\beta}_1 - \beta_{1,0}|/SE(\hat{\beta}_1) > 1.96$. These are algebraically the same condition.

$$\boxed{\text{Reject } H_0 : \beta_1 = \beta_{1,0} \text{ at } \alpha = 0.05 \quad \Longleftrightarrow \quad \beta_{1,0} \notin 95\% \text{ CI}}$$

Question 4.5: CI for predicted change.

When GDP per capita increases by \$10,000:

Point estimate: $\Delta\hat{Y} = 0.000637 \times 10{,}000 \approx 6.37$ years

95% CI: $[0.000509 \times 10{,}000,\ 0.000765 \times 10{,}000] = [5.09,\ 7.65]$ years

Interpretation: We are 95% confident that a \$10,000 increase in GDP per capita is associated with an increase in life expectancy of between 5.1 and 7.7 years. This is a meaningful range—even the lower bound suggests a substantial association between national income and longevity.

5 Regression with Binary Variables

Question 5.1: Create the binary variable.

Run the code to create the `africa` dummy. You should find 52 African countries and 90 non-African countries in the 2007 cross-section.

Question 5.2: Group means.

- Mean life expectancy in Africa: ≈ 54.81 years
- Mean life expectancy outside Africa: ≈ 74.07 years
- Difference: $\approx 54.81 - 74.07 = -19.26$ years

African countries have, on average, about 19 fewer years of life expectancy than non-African countries. This is a very large gap.

Question 5.3: Binary regression results.

The regression $\widehat{\text{LifeExp}}_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Africa}_i$ yields:

(a) $\hat{\beta}_0 \approx 74.07$ — this is the **average life expectancy of non-African countries** (the group where Africa = 0). The intercept equals the sample mean of the omitted group.

(b) $\hat{\beta}_1 \approx -19.26$ — this is the **difference in average life expectancy** between African and non-African countries (Africa minus non-Africa). It matches exactly the difference in group means computed above.

(c) The t-statistic on $\hat{\beta}_1$ is very large in absolute value (approximately -10.1), and the p-value is essentially zero. **Yes**, the difference is statistically significant at the 5% level (and at any conventional level). We can reject the hypothesis that African and non-African countries have the same average life expectancy.

Question 5.4: Comparison with t-test.

The t-test and the regression produce the same key results:

- The difference in means is identical (≈ -19.26 years)
- The p-values are essentially the same (both ≈ 0)

This demonstrates a fundamental insight: **a two-sample t-test is just a regression with a binary independent variable**. The regression framework generalizes the t-test—it can accommodate continuous regressors, multiple regressors, and more complex specifications, while the t-test is limited to comparing two groups.

Note: The p-values may differ very slightly because the regression uses homoskedastic standard errors by default, while `t.test()` uses the Welch correction (heteroskedastic). This distinction is minor in practice.

Question 5.5: 95% CI for binary coefficient.

The 95% CI for $\hat{\beta}_1$ is approximately $[-23.05, -15.47]$ years.

Interpretation: We are 95% confident that the true difference in average life expectancy between African and non-African countries is between 15.5 and 23.1 years (with Africa being lower). The entire interval is negative, confirming that the gap is statistically significant. Even the most conservative estimate suggests a gap of more than 15 years.

Question 5.6: Visualization.

The boxplot reveals several patterns:

- The **median** life expectancy is much lower in Africa than elsewhere
- There is **some overlap**—a few African countries have life expectancy comparable to non-African countries, and some non-African countries (particularly in Asia) have relatively low life expectancy
- The **spread** within Africa is notable—some countries (like Tunisia, Libya) have much higher life expectancy than others (like Swaziland, Sierra Leone), reflecting the devastating impact of HIV/AIDS in parts of sub-Saharan Africa
- The non-Africa group also has considerable spread, reflecting the diversity of this group (it includes both wealthy European countries and lower-income Asian countries)

6 Bringing It Together

Question 6.1: Omitted variable bias.

Several variables could create omitted variable bias:

1. Healthcare spending / quality of health systems:

- Correlated with GDP per capita: Richer countries spend more on healthcare
- Affects life expectancy: Better healthcare directly improves health outcomes
- Bias direction: Positive (we attribute some of the healthcare effect to GDP)

2. Education levels:

- Correlated with GDP per capita: Richer countries have more educated populations
- Affects life expectancy: Education improves health behaviors and access to care
- Bias direction: Positive (the GDP coefficient picks up education effects too)

Other examples: clean water access, sanitation infrastructure, political stability, inequality, disease environment (e.g., malaria, HIV prevalence).

Because these omitted variables are positively correlated with both GDP per capita and life expectancy, our estimated $\hat{\beta}_1$ likely **overstates** the causal effect of income on longevity. Part of what we attribute to "income" is really driven by these other factors. This is why we need multiple regression (coming next!).

Question 6.2: Comparison table.

	GDP per capita regression	Africa dummy regression
$\hat{\beta}_0$	≈ 59.57	≈ 74.07
$\hat{\beta}_1$	≈ 0.000637	≈ -19.26
$SE(\hat{\beta}_1)$	≈ 0.0000653	≈ 1.91
t-statistic	≈ 9.75	≈ -10.1
R^2	≈ 0.46	≈ 0.42

The GDP per capita regression has a slightly higher R^2 (0.46 vs. 0.42), meaning it explains a bit more of the variation in life expectancy. However, a higher R^2 does **not** automatically mean a "better" model. The two regressions answer different questions:

- The GDP regression asks: "How does life expectancy change with income?"
- The Africa regression asks: "Is there a life expectancy gap between Africa and the rest of the world?"

Both are valid and informative. The "right" model depends on the question you are trying to answer, not just which has the higher R^2.

Question 6.3: What can't we do yet?

The main limitation is that we cannot make **causal claims**. Simple regression gives us *associations*, not *causal effects*.

We cannot say that increasing a country's GDP per capita by \$10,000 would *cause* life expectancy to rise by 6.4 years, because of **omitted variable bias**. Many unobserved factors (healthcare, education, institutions) are correlated with both GDP per capita and life expectancy, so our slope estimate captures the combined effect of all these variables, not just income.

To move toward causal inference, we need:

- **Multiple regression**: Include control variables to reduce omitted variable bias (we will cover this in the next handout)
- **Research design**: Use natural experiments, instrumental variables, or randomized controlled trials to isolate causal effects