Handout 6: Inference and Binary Regressors

EC 282: Introduction to Econometrics

Spring 2026

Instructions: Run the provided R code and answer the questions. Show your work for calculations. This handout builds directly on Handout 5—make sure you have that code loaded first.

1 Setup: Continuing with the Gapminder Data

We continue with the same Gapminder 2007 cross-section from Handout 5. Run the code below to reload the data and re-estimate the regression.

```r
library(ggplot2)

options(scipen = 999)

# Reload the Gapminder 2007 data
gapminder_full <- read.csv(
  "https://raw.githubusercontent.com/resbaz/r-novice-gapminder-files/
   master/data/gapminder-FiveYearData.csv"
)
gap07 <- gapminder_full[gapminder_full$year == 2007, ]

# Re-estimate the simple regression from Handout 5
reg <- lm(lifeExp ~ gdpPercap, data = gap07)
```

2 Reading the Regression Output

Question 2.1: Run `summary(reg)` and examine the output carefully.

```r
summary(reg)
```

Identify and write down the following from the output:

(a) The estimated slope $\hat{\beta}_1$ and intercept $\hat{\beta}_0$

(b) The standard error of $\hat{\beta}_1$, denoted $SE(\hat{\beta}_1)$

(c) The t-statistic for $\hat{\beta}_1$

(d) The p-value for $\hat{\beta}_1$

(e) The R^2

Question 2.2: The standard error measures the *precision* of our estimate. In your own words, what does a small standard error tell you? What does a large one tell you?

Question 2.3: What affects the precision of $\hat{\beta}_1$? Three factors determine $SE(\hat{\beta}_1)$: the sample size n, the spread of X (i.e., $\text{Var}(X_i)$), and the amount of noise in the regression (i.e., $\text{Var}(u_i)$). For each factor, explain whether *increasing* it makes $SE(\hat{\beta}_1)$ larger or smaller, and why.

3 Hypothesis Testing

We want to test whether GDP per capita has a *statistically significant* relationship with life expectancy. This means testing whether the true slope β_1 could be zero.

Question 3.1: Write down the null and alternative hypotheses for a two-sided test of whether GDP per capita is related to life expectancy.

Question 3.2: The t-statistic is computed as:

$$t\text{-stat} = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

where $\beta_{1,0}$ is the hypothesized value under H_0. Using the values from the regression output, compute the t-statistic by hand. Verify it matches the output.

```r
# Extract coefficients and standard errors
beta1_hat <- coef(reg)["gdpPercap"]
se_beta1  <- summary(reg)$coefficients["gdpPercap", "Std. Error"]

cat("beta1_hat:", beta1_hat, "\n")
cat("SE(beta1_hat):", se_beta1, "\n")

# Compute t-statistic by hand
t_stat <- beta1_hat / se_beta1
cat("t-statistic (by hand):", t_stat, "\n")
```

Question 3.3: To make a decision, we compare the t-statistic to critical values. For a two-sided test at the 5% significance level, we reject H_0 if $|t| > 1.96$.

(a) Is the slope on GDP per capita statistically significant at the 5% level?

(b) Is it significant at the 1% level? (Critical value: 2.576)

Question 3.4: The p-value gives the probability of observing a t-statistic at least as extreme as ours, *assuming H_0 is true*. Compute the p-value in R and verify it matches the regression output.

```r
# Compute p-value for two-sided test
p_value <- 2 * pnorm(-abs(t_stat))
cat("p-value (by hand):", p_value, "\n")
cat("p-value (from summary):",
    summary(reg)$coefficients["gdpPercap", "Pr(>|t|)"], "\n")
```

Interpret the p-value in a sentence. What does it tell us about the relationship between GDP per capita and life expectancy?

Question 3.5: Suppose instead we wanted to test whether GDP per capita has a *positive* effect on life expectancy (a one-sided test):

$$H_0 : \beta_1 = 0 \qquad H_A : \beta_1 > 0$$

```r
# One-sided p-value (right tail)
p_value_one <- pnorm(-abs(t_stat))
cat("One-sided p-value:", p_value_one, "\n")
```

How does the one-sided p-value compare to the two-sided p-value? Why?

4 Confidence Intervals

A confidence interval provides a range of plausible values for the true population parameter β_1.

Question 4.1: A 95% confidence interval for β_1 is:

$$\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)$$

Compute this by hand using the values from the regression output.

```
# 95% confidence interval by hand
ci_lower <- beta1_hat - 1.96 * se_beta1
ci_upper <- beta1_hat + 1.96 * se_beta1

cat("95% CI: [", round(ci_lower, 7), ",", round(ci_upper, 7), "]\n")
```

Verify with R's built-in function:

```
confint(reg, "gdpPercap", level = 0.95)
```

Question 4.2: Interpret the 95% confidence interval in a sentence. Does the interval include zero? What does this tell you?

Question 4.3: Now construct a 99% confidence interval. Is it wider or narrower than the 95% CI? Why?

```
# 99% confidence interval
ci99_lower <- beta1_hat - 2.576 * se_beta1
ci99_upper <- beta1_hat + 2.576 * se_beta1

cat("99% CI: [", round(ci99_lower, 7), ",", round(ci99_upper, 7), "]\n")
confint(reg, "gdpPercap", level = 0.99)
```

Question 4.4: The connection between confidence intervals and hypothesis testing: if the 95% CI does not include a hypothesized value $\beta_{1,0}$, then we would reject $H_0 : \beta_1 = \beta_{1,0}$ at the 5% level. Explain in your own words why this equivalence holds.

Question 4.5: Suppose we want a confidence interval for the *predicted change* in life expectancy when GDP per capita increases by \$10,000. Compute a 95% CI for $\Delta Y = \beta_1 \times 10{,}000$.

```
# CI for predicted change when X increases by $10,000
delta_X <- 10000
pred_change <- beta1_hat * delta_X
ci_change_lower <- ci_lower * delta_X
ci_change_upper <- ci_upper * delta_X

cat("Predicted change:", round(pred_change, 2), "years\n")
cat("95% CI for change: [", round(ci_change_lower, 2), ",",
    round(ci_change_upper, 2), "] years\n")
```

Interpret this result in a sentence.

5 Regression with Binary Variables

So far, our regressor (GDP per capita) has been continuous. Now we explore what happens when the independent variable is **binary** (taking only values 0 or 1).

Question 5.1: Create a binary variable `africa` that equals 1 if a country is in Africa and 0 otherwise.

```
# Create binary variable
gap07$africa <- ifelse(gap07$continent == "Africa", 1, 0)

cat("Number of African countries:", sum(gap07$africa), "\n")
cat("Number of non-African countries:", sum(1 - gap07$africa), "\n")
```

Question 5.2: Compute the average life expectancy separately for African and non-African countries.

```
# Group means
mean_africa     <- mean(gap07$lifeExp[gap07$africa == 1])
mean_non_africa <- mean(gap07$lifeExp[gap07$africa == 0])

cat("Mean life exp (Africa):", round(mean_africa, 2), "years\n")
cat("Mean life exp (non-Africa):", round(mean_non_africa, 2), "years\n")
cat("Difference:", round(mean_africa - mean_non_africa, 2), "years\n")
```

Question 5.3: Now run a regression of life expectancy on the `africa` dummy:

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

```
reg_binary <- lm(lifeExp ~ africa, data = gap07)
summary(reg_binary)
```

(a) What does $\hat{\beta}_0$ represent? Compare it to one of the group means you computed above.

(b) What does $\hat{\beta}_1$ represent? Compare it to the difference in group means.

(c) Is the difference statistically significant at the 5% level? How do you know?

Question 5.4: This is a key insight: **regression with a binary regressor is the same as a two-sample comparison of means**. Verify this by running a t-test and comparing the results.

```r
# Two-sample t-test
t.test(lifeExp ~ africa, data = gap07, var.equal = FALSE)
```

Compare the difference in means and p-value from the t-test to the regression output. What do you notice?

Question 5.5: Construct a 95% confidence interval for $\hat{\beta}_1$ from the binary regression. What does this interval tell us about the life expectancy gap between African and non-African countries?

```r
confint(reg_binary, "africa", level = 0.95)
```

Question 5.6: Create a visualization comparing the two groups.

```r
ggplot(gap07, aes(x = factor(africa,
                             labels = c("Non-Africa", "Africa")),
                  y = lifeExp)) +
  geom_boxplot(fill = c("steelblue", "coral"), alpha = 0.7) +
  geom_jitter(width = 0.2, alpha = 0.4, size = 1.5) +
  labs(title = "Life Expectancy: Africa vs. Rest of World (2007)",
       x = "", y = "Life Expectancy (years)") +
  theme_minimal()
```

What patterns do you notice in the data? Is there overlap between the two groups?

6 Bringing It Together

Question 6.1: We've seen that the relationship between GDP per capita and life expectancy is not perfectly linear (Handout 5 showed curvature in the residual plot). One concern is **omitted variable bias**. What variables might be in the error term u_i that are both:

1. Correlated with GDP per capita, AND
2. Affect life expectancy?

Name at least two such variables and explain why they create bias.

Question 6.2: Compare the two regressions we've run in this handout. Fill in the table:

	GDP per capita regression	Africa dummy regression
$\hat{\beta}_0$		
$\hat{\beta}_1$		
$SE(\hat{\beta}_1)$		
t-statistic		
R^2		

```r
# Side-by-side comparison
cat("=== GDP per Capita Regression ===\n")
cat("beta0:", round(coef(reg)[1], 4), "\n")
cat("beta1:", round(coef(reg)[2], 7), "\n")
cat("SE(beta1):", round(summary(reg)$coefficients[2, 2], 7), "\n")
cat("t-stat:", round(summary(reg)$coefficients[2, 3], 3), "\n")
cat("R-squared:", round(summary(reg)$r.squared, 4), "\n\n")

cat("=== Africa Dummy Regression ===\n")
cat("beta0:", round(coef(reg_binary)[1], 4), "\n")
cat("beta1:", round(coef(reg_binary)[2], 4), "\n")
cat("SE(beta1):",
    round(summary(reg_binary)$coefficients[2, 2], 4), "\n")
cat("t-stat:",
    round(summary(reg_binary)$coefficients[2, 3], 3), "\n")
cat("R-squared:", round(summary(reg_binary)$r.squared, 4), "\n")
```

Which regression has a higher R^2? Does this mean it is a "better" model? Discuss.

Question 6.3: Reflect on what we've learned across Handouts 5 and 6. We can now:

- Estimate a regression line (Handout 5)
- Assess how well it fits (Handout 5)
- Test whether the relationship is statistically significant (Handout 6)
- Construct confidence intervals for the true effect (Handout 6)
- Use binary variables to compare group means within a regression framework (Handout 6)

What is one important thing we still *cannot* do with simple regression? (Hint: Think about why the coefficient on GDP per capita might not represent a causal effect.)