Solutions Tutorial 3

Housing prices

The regression model is: `price = 300,000 + 1500 * sqmtr`, with $R^2 = 0.64$.

(a) Interpret the intercept (300,000) and the slope coefficient (1,500) in plain English.

- **Intercept (300,000):** This is the predicted price of a house with 0 square meters of interior surface. In this context, the intercept has no meaningful practical interpretation, as a house cannot have zero square meters. It is a statistical construct that helps position the regression line correctly in the data cloud.
- **Slope (1,500):** For each additional square meter of interior surface, the sale price of a house is predicted to increase by 1,500 euros, holding all other factors constant.

(b) What does the R-squared value of 0.64 tell us about this model?

- An R-squared of 0.64 means that **64% of the total variation in house prices (`price`) is explained by the variation in the interior surface (`sqmtr`)**. The remaining 36% of the variation in price is due to other factors not included in the model (e.g., location, age of the house, number of bedrooms, etc.).

(c) If you were to re-estimate the model with price measured in thousands of euros (e.g., a 250,000 euro house becomes 250), what would the new equation be?

- If we divide the dependent variable `price` by 1,000, we must also divide the entire right-hand side of the equation by 1,000 to maintain the equality. Let `price_k` be the price in thousands of euros.

$$\frac{\text{price}}{1000} = \frac{300,000}{1000} + \frac{1500}{1000} \times \text{sqmtr}$$

- The new equation would be: `price_k = 300 + 1.5 * sqmtr`
- The interpretation changes accordingly: The intercept is now 300 thousand euros, and each additional square meter increases the predicted price by 1.5 measured in the new units (thousands of euros).

Log-Log Model

The regression result is: `log(Sales) = 2.1 - 0.85 * log(Ad_Price)`.

How would you interpret the coefficient -0.85? What is the economic term for this value?

- **Interpretation:** In a log-log model, the coefficient represents an elasticity. A **1% increase in the advertising price (`Ad_Price`) is associated with a 0.85% decrease in product sales (`Sales`)**, on average.
- **Economic Term:** This value is the **price elasticity of demand**. Since the absolute value is less than 1 ($|-0.85| < 1$), we would say that the demand for the product is **inelastic** with respect to the advertising price.

Error Term and Residual

What is the fundamental difference between the population error term (u_i) and the OLS residual (e_i)? Why can we observe one but not the other?

- **Fundamental Difference:**
  - The **population error term (u_i)** is the vertical distance between a data point (y_i) and the *true, unobservable population regression line*. It represents all the unobserved factors that affect y_i besides x_i. $u_i = y_i - (\beta_0 + \beta_1 x_i)$
  - The **OLS residual (e_i or $\hat{u}_i$)** is the vertical distance between a data point (y_i) and the *estimated sample regression line*. It is the prediction error from our estimated model. $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$
- **Why we can't observe u_i:** We cannot observe the population error term u_i because we do not know the true population parameters β_0 and β_1. We can only estimate them using a sample of data, which gives us $\hat{\beta}_0$ and $\hat{\beta}_1$. Because we can calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ from our sample, we can calculate the residual e_i for each observation.

Proving a Fundamental OLS Property

Using the formula for the OLS intercept estimator, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, prove that the regression line passes through the point of sample means, $(\bar{x}, \bar{y})$.

1. The estimated OLS regression line is given by the equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
2. To show that the line passes through the point $(\bar{x}, \bar{y})$, we need to show that when we plug in $x = \bar{x}$, the predicted value $\hat{y}$ is equal to $\bar{y}$.
3. Substitute the formula for the intercept, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, into the regression equation: $\hat{y} = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x$
4. Now, set $x = \bar{x}$: $\hat{y} = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 \bar{x}$
5. The terms $-\hat{\beta}_1 \bar{x}$ and $+\hat{\beta}_1 \bar{x}$ cancel each other out: $\hat{y} = \bar{y}$

This proves that when the input is the sample mean of x, the predicted output is the sample mean of y. Therefore, the OLS regression line always passes through the point of sample means $(\bar{x}, \bar{y})$.

Unbiasedness

(a) Show that the estimator can be rewritten as: $\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

1. Start with the formula for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

   A useful property is $\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum(x_i - \bar{x})y_i$. So,

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}$$

2. Substitute the true population model $y_i = \beta_0 + \beta_1 x_i + u_i$ for y_i:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\sum(x_i - \bar{x})^2}$$

3. Distribute the term $(x_i - \bar{x})$ in the numerator:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})\beta_0 + \sum(x_i - \bar{x})\beta_1 x_i + \sum(x_i - \bar{x})u_i}{\sum(x_i - \bar{x})^2}$$

4. Analyze each term in the numerator:

   - $\sum(x_i - \bar{x})\beta_0 = \beta_0 \sum(x_i - \bar{x}) = \beta_0 \cdot 0 = 0$.
   - $\sum(x_i - \bar{x})\beta_1 x_i = \beta_1 \sum(x_i - \bar{x})x_i$. Using the same property as step 1, $\sum(x_i - \bar{x})x_i = \sum(x_i - \bar{x})(x_i - \bar{x}) = \sum(x_i - \bar{x})^2$. So this term is $\beta_1 \sum(x_i - \bar{x})^2$.
   - $\sum(x_i - \bar{x})u_i$ remains as is.

5. Substitute these back into the expression:

$$\hat{\beta}_1 = \frac{0 + \beta_1 \sum(x_i - \bar{x})^2 + \sum(x_i - \bar{x})u_i}{\sum(x_i - \bar{x})^2}$$

6. Separate the fraction:

$$\hat{\beta}_1 = \frac{\beta_1 \sum(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2} + \frac{\sum(x_i - \bar{x})u_i}{\sum(x_i - \bar{x})^2}$$

7. Simplify to get the final result:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(b) Take the conditional expectation... to prove that $E(\hat{\beta}_1|X) = \beta_1$.

1. Start with the expression from part (a):

$$\hat{\beta}_1 = \beta_1 + \frac{\sum(x_i - \bar{x})u_i}{\sum(x_i - \bar{x})^2}$$

2. Take the expectation of both sides, conditional on $X = \{x_1, x_2, ..., x_n\}$:

$$E(\hat{\beta}_1|X) = E\left(\beta_1 + \frac{\sum(x_i - \bar{x})u_i}{\sum(x_i - \bar{x})^2}\middle| X\right)$$

3. Use the linearity of expectation:

$$E(\hat{\beta}_1|X) = E(\beta_1|X) + E\left(\frac{\sum(x_i - \bar{x})u_i}{\sum(x_i - \bar{x})^2}\middle| X\right)$$

4. Analyze each term:

   - $E(\beta_1|X) = \beta_1$ because β_1 is a constant.
   - For the second term, since we are conditioning on X, all x_i and $\bar{x}$ values are treated as non-random. We can pull them outside the expectation:

$$E\left(\frac{\sum(x_i - \bar{x})u_i}{\sum(x_i - \bar{x})^2}\middle| X\right) = \frac{1}{\sum(x_i - \bar{x})^2} E\left(\sum(x_i - \bar{x})u_i\middle| X\right)$$

$$= \frac{1}{\sum(x_i - \bar{x})^2} \sum(x_i - \bar{x})E(u_i|X)$$

5. Now, use the **Zero Conditional Mean assumption**, $E(u_i|X) = 0$. This means the entire second term becomes zero:

$$\frac{1}{\sum(x_i - \bar{x})^2} \sum(x_i - \bar{x}) \cdot 0 = 0$$

6. Substitute back into the main equation:

$$E(\hat{\beta}_1|X) = \beta_1 + 0$$

$$E(\hat{\beta}_1|X) = \beta_1$$

   This proves that the OLS slope estimator is unbiased.

Omitted Variable Bias

Show that the expected value of the estimator from the incorrect short regression is $E(\hat{\gamma}_1) = \beta_1 + \beta_2 \cdot \delta_1$.

1. The estimated coefficient from the incorrect (short) regression of y on x_1 is:
$$\hat{\gamma}_1 = \frac{\sum(x_{1i} - \bar{x}_1)y_i}{\sum(x_{1i} - \bar{x}_1)^2}$$

2. Substitute the *true* population model for y_i: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$.
$$\hat{\gamma}_1 = \frac{\sum(x_{1i} - \bar{x}_1)(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i)}{\sum(x_{1i} - \bar{x}_1)^2}$$

3. Distribute the numerator and separate the terms, just as in the unbiasedness proof:
$$\hat{\gamma}_1 = \frac{\sum(x_{1i} - \bar{x}_1)\beta_0}{\sum(x_{1i} - \bar{x}_1)^2} + \frac{\sum(x_{1i} - \bar{x}_1)\beta_1 x_{1i}}{\sum(x_{1i} - \bar{x}_1)^2} + \frac{\sum(x_{1i} - \bar{x}_1)\beta_2 x_{2i}}{\sum(x_{1i} - \bar{x}_1)^2} + \frac{\sum(x_{1i} - \bar{x}_1)u_i}{\sum(x_{1i} - \bar{x}_1)^2}$$

4. Simplify each term:

   - The first term is 0. (Covariance with a constant)
   - The second term simplifies to β_1.
   - The third term can be rewritten as $\beta_2 \left(\frac{\sum(x_{1i} - \bar{x}_1)x_{2i}}{\sum(x_{1i} - \bar{x}_1)^2} \right)$.
   - The fourth term remains. So,
   $$\hat{\gamma}_1 = \beta_1 + \beta_2 \left(\frac{\sum(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum(x_{1i} - \bar{x}_1)^2} \right) + \frac{\sum(x_{1i} - \bar{x}_1)u_i}{\sum(x_{1i} - \bar{x}_1)^2}$$
   (Note: We can replace x_{2i} with $(x_{2i} - \bar{x}_2)$ in the numerator because $\sum(x_{1i} - \bar{x}_1)\bar{x}_2 = 0$.)

5. Recognize that the term in the parenthesis is the formula for the OLS slope coefficient from a regression of x_2 on x_1. Let's call this $\hat{\delta}_1$.
$$\hat{\delta}_1 = \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} = \frac{\sum(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum(x_{1i} - \bar{x}_1)^2}$$
   The equation becomes: $\hat{\gamma}_1 = \beta_1 + \beta_2 \hat{\delta}_1 + \text{error term involving u.}$

6. Now, take the expectation. We assume that in the population, the relationship between x_2 and x_1 is $E(x_2|x_1) = \delta_0 + \delta_1 x_1$. So $E(\hat{\delta}_1) = \delta_1$.
$$E(\hat{\gamma}_1) = E(\beta_1) + E(\beta_2 \hat{\delta}_1) + E(\text{error term})$$
$$E(\hat{\gamma}_1) = \beta_1 + \beta_2 E(\hat{\delta}_1) + 0 \quad (\text{since } E(u|X) = 0)$$
$$E(\hat{\gamma}_1) = \beta_1 + \beta_2 \delta_1$$
   The term $\beta_2 \delta_1$ is the **omitted variable bias**.

Marginal Effects

(a) The Quadratic Model: For $y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$, find $\frac{dy}{dx}$.

- To find the marginal effect of x on y, we take the partial derivative of y with respect to x:

$$\frac{\partial y}{\partial x} = \frac{\partial}{\partial x}(\beta_0 + \beta_1 x + \beta_2 x^2 + u)$$

$$\frac{\partial y}{\partial x} = 0 + \beta_1 + 2\beta_2 x + 0$$

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x$$

- This result shows that the marginal effect of a one-unit change in x on y is not constant; it depends on the current level of x. For each value of x, the slope of the relationship is different.

(b) The Level-Log Model: For $y = \beta_0 + \beta_1 \log(x) + u$, show that a 1% change in x leads to an approximate change of $(\beta_1/100)$ units in y.

1. First, find the derivative of y with respect to x:

$$\frac{dy}{dx} = \beta_1 \frac{1}{x}$$

2. Rearrange the equation to find an expression for an infinitesimal change in y, dy:

$$dy = \beta_1 \frac{dx}{x}$$

3. The term $\frac{dx}{x}$ represents the proportional or percentage change in x. For discrete changes, we can write this as an approximation:

$$\Delta y \approx \beta_1 \frac{\Delta x}{x}$$

4. If we consider a 1% change in x, then $\frac{\Delta x}{x} = 0.01$.
5. Substitute this value into the approximation:

$$\Delta y \approx \beta_1 (0.01) = \frac{\beta_1}{100}$$

- Thus, a 1% change in x is associated with an approximate change in y of $(\beta_1/100)$ units.

Perfect Multicollinearity

(a) What does it mean for x_1 and x_2 to have *perfect multicollinearity*?

- Perfect multicollinearity means that one explanatory variable is a perfect linear function of another. For example, $x_1 = c_0 + c_1 x_2$ for some constants c_0 and c_1 where $c_1 \neq 0$. This means there is no independent variation in x_1 that is not associated with x_2. A common example is including a variable in different units (e.g., height in meters and height in centimeters) in the same regression.

(b) Analytically, what happens to the value of R_1^2 under perfect multicollinearity?

- R_1^2 is the R-squared from a regression of x_1 on x_2. If x_1 is a perfect linear function of x_2, then the regression of x_1 on x_2 will explain 100% of the variation in x_1. Therefore, $R_1^2 = 1$.

(c) Explain mathematically why it is impossible to calculate $\hat{\beta}_1$.

- The variance of the OLS estimator $\hat{\beta}_1$ is given by:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_1(1 - R_1^2)}$$

- Under perfect multicollinearity, we established that $R_1^2 = 1$. Substituting this into the denominator:

$$\text{Denominator} = SST_1(1 - 1) = SST_1 \cdot 0 = 0$$

- The variance of the estimator becomes:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{0} \to \infty$$

- Since the variance of the estimator is infinite, the OLS estimator is undefined. The OLS procedure fails because it is mathematically impossible to distinguish the unique effect of x_1 from the effect of x_2 when they are perfectly linearly related.

Zero Conditional Mean

(a) Explain in your own words what this assumption means.

- The Zero Conditional Mean assumption, $E(u|x) = 0$, means that the average value of all unobserved factors (the error term, u) is zero for any given value of the explanatory variable (x). Put simply, it means that the unobserved factors are not systematically related to, or correlated with, the explanatory variable.

(b) Using `wage` on `education`, explain why "innate ability" violates this assumption.

- In a model wage $= \beta_0 + \beta_1$education $+ u$, "innate ability" is an unobserved factor and is therefore part of the error term u.
- It is very likely that innate ability is correlated with both `wage` and `education`.
    1. People with higher ability may earn higher wages regardless of their education level.
    2. People with higher ability may find it easier to succeed in school and are therefore more likely to attain higher levels of education.
- Because `ability` is in `u` and is also correlated with `education`, the average level of `u` is not zero across different levels of `education`. Specifically, $E(u|\text{education})$ will be higher for higher levels of `education`. This violates the Zero Conditional Mean assumption.

(c) In which direction will $\hat{\beta}_1$ be biased?

- The bias will be **positive**. The OLS estimate $\hat{\beta}_1$ will be **overstated**.
- **Reasoning (using the OVB formula):** The bias is $\beta_2 \cdot \delta_1$.
    - β_2 is the effect of the omitted variable (ability) on the outcome (wage). This effect is positive ($\beta_2 > 0$).
    - δ_1 is the correlation between the included variable (education) and the omitted variable (ability). This correlation is also positive ($\delta_1 > 0$).
- Since the bias term is the product of two positive numbers, the bias is positive. The OLS estimate $\hat{\beta}_1$ will mistakenly attribute some of the wage-increasing effect of ability to education, leading to an estimate that is larger than the true causal effect of education on wages.

Variance of the OLS Estimator

What two things could you do to increase the *precision* of your estimate, $\hat{\beta}_1$?

The variance formula is $Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x}$. To increase precision, we need to *decrease* this variance.

1. **Decrease the error variance (σ^2)**: σ^2 is the variance of the unobserved factors, u. In an experimental setting, this means **making the experimental conditions as controlled and uniform as possible**. For example, ensure all crop plots have the same soil type, water access, and sunlight exposure. By minimizing the influence of other factors, you reduce the "noise" in the model, making the relationship between fertilizer and yield clearer.
2. **Increase the Total Sum of Squares of x (SST_x)**: $SST_x = \sum(x_i - \bar{x})^2$. This term measures the total variation in the explanatory variable. In your experiment, this means you should **use a wider range of fertilizer amounts (x) across your different plots**. Intuitively, it is easier to detect a trend line if the points are spread far apart horizontally than if they are all bunched together. More variation in x provides more information to pin down the slope of the regression line.

R-squared

Why is a high R-squared not necessarily the ultimate goal? What is often more important?

- A high R-squared is not the ultimate goal because it only measures **goodness-of-fit**, not **causal validity**. A model can have a very high R-squared but still suffer from severe omitted variable bias, making its coefficients unreliable for policy decisions. For example, a model predicting crime rates using ice cream sales might have a high R-squared in the summer, but the relationship is spurious.
- What is often more important is obtaining an **unbiased and consistent estimate of a specific coefficient** that represents a causal effect of interest. For policy, we need to know the true causal impact of changing a variable (e.g., years of education, police funding, carbon tax). This requires a model specification that is theoretically sound and minimizes biases (like OVB), even if it results in a lower R-squared. **Unbiasedness is usually more important than fit.**

Omitted Variable Bias

What are two or three other variables you would want to include in the `wage` on `education` model? What are the practical challenges?

- **Variables to Include:**
  1. **Cognitive Ability / IQ:** To control for the "innate ability" bias discussed earlier.
  2. **Quality of Institution:** The return to a degree from a top university is likely higher than from a less selective one. This could be measured by university ranking or average test scores of admitted students.
  3. **Parental Background:** Variables like parents' education and income can capture family network effects, financial support, and environmental factors that influence both a child's education and future earnings.
- **Practical Challenges:**
  - **Data Availability:** Most standard economic datasets (like labor force surveys) do not collect information on IQ, school quality, or detailed parental background.
  - **Measurement:** These variables are difficult to measure accurately. "Ability" is a complex construct, and IQ tests are controversial. "School quality" is also multi-faceted and hard to summarize with a single number.
  - **Privacy:** Data on IQ and parental income are highly sensitive and can be difficult to obtain due to privacy concerns.

OLS Minimization

Why do we use the *sum of squared residuals*? Why not absolute values or just the sum?

- **Why not the sum of residuals?** Minimizing $\sum e_i$ is not a useful criterion. An infinite number of lines can make this sum equal to zero (any line passing through the point of means, $(\bar{x}, \bar{y})$), so it does not yield a unique solution.
- **Why we use the sum of *squared* residuals:**

  1. **Treats Positive/Negative Errors Equally:** Squaring makes all errors positive, so large positive errors and large negative errors are treated as equally "bad".
  2. **Penalizes Large Errors More:** Squaring gives much more weight to large errors than to small ones (e.g., an error of 2 becomes 4, but an error of 10 becomes 100). This is often desirable, as it forces the line to fit the bulk of the data well by avoiding large deviations.
  3. **Mathematical Convenience:** The sum of squares is a smooth, differentiable function. Using calculus, we can easily derive a unique, closed-form analytical solution for the estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Minimizing the sum of absolute values (Least Absolute Deviations, or LAD) is computationally more complex and may not have a unique solution.

Polynomials

When might you suspect a quadratic model would be more appropriate? What would a negative coefficient on the `experience²` term imply?

- **When to Suspect Non-linearity:**
  1. **Economic Theory:** Theory might suggest a non-linear relationship. For example, the "law of diminishing marginal returns" is common in economics. The effect of experience on wages is likely positive but decreases as one gets more experienced.
  2. **Visual Inspection:** A scatterplot of the dependent variable against the independent variable might reveal a curved, parabolic shape rather than a straight line.
  3. **Residual Plots:** If you fit a linear model and then plot the residuals against the independent variable, a U-shaped or inverted U-shaped pattern in the residuals suggests that a quadratic term might be missing.
- **Interpretation of a negative `experience²` coefficient:**
  - In the model wage $= \beta_0 + \beta_1 \text{experience} + \beta_2 \text{experience}^2 + u$, if $\beta_1 > 0$ and $\beta_2 < 0$, it implies a **concave, inverted U-shaped relationship** between experience and wage.
  - This means that as a person gains their first few years of experience, their wage increases (β_1 term dominates). However, the rate of this increase slows down over time (the negative β_2 term starts to have more impact). This reflects **diminishing marginal returns to experience**. Eventually, after a certain point, an additional year of experience might even lead to a decrease in predicted wages (if the person becomes less adaptable or their skills become obsolete).