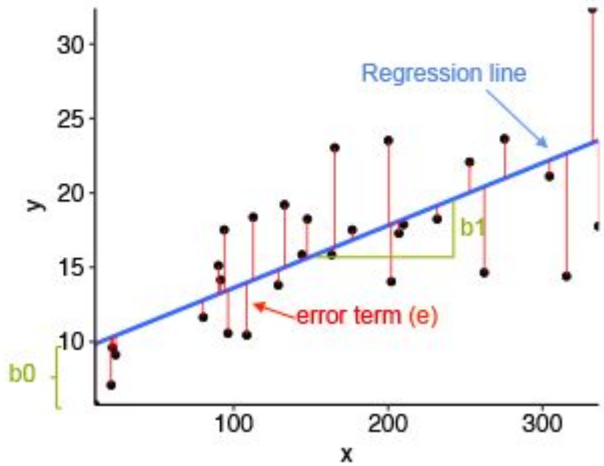
TUT206 Nov 08

Recap: simple linear regression

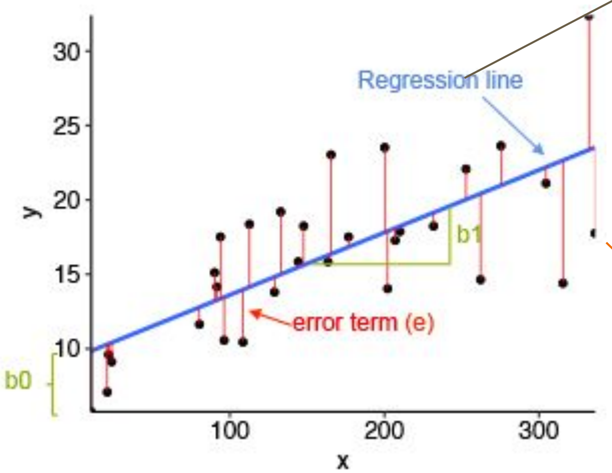


$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

theoretical model

Recap: simple linear regression

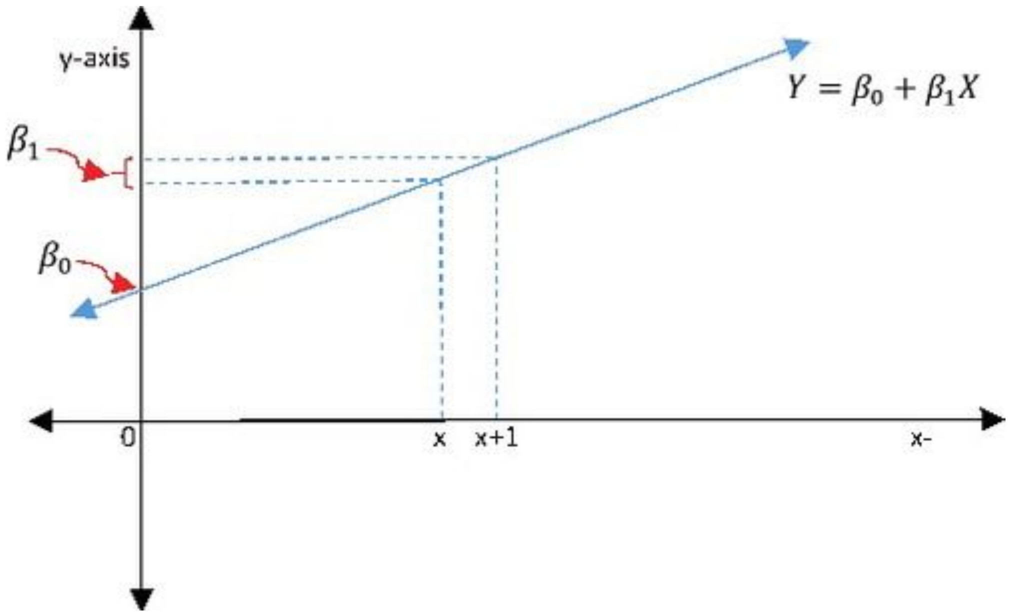
Fitted



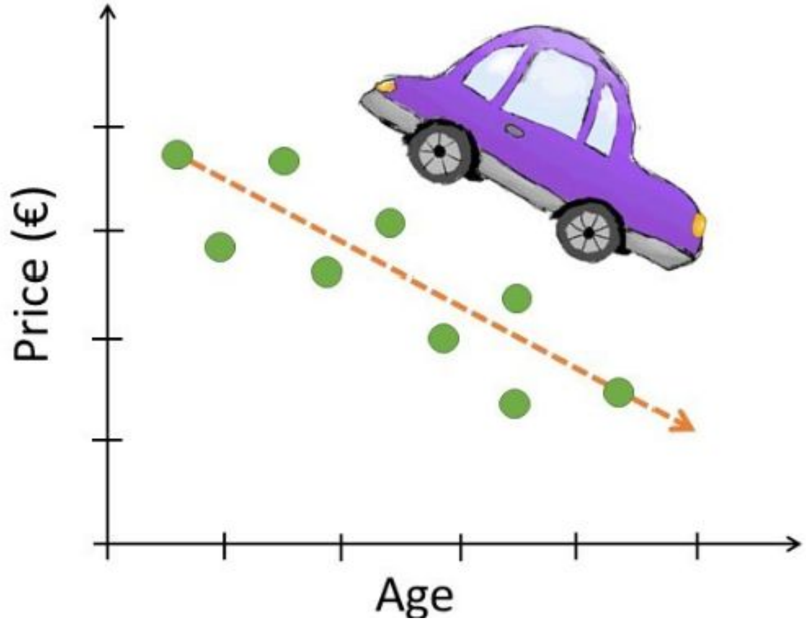
residuals $e_i = \hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ **fitted model**

Recap: simple linear regression



Recap: simple linear regression



$$H0 : \beta_1 = 0$$

$$H1 : \beta_1 \neq 0$$

Project

The project requires students to **analyze real-world data** from the **Canadian Social Connection Survey (CSCS)** and communicate their findings effectively.

1. **Select Three (3) Research Questions**

- Collaboratively choose three distinct research questions to explore using the CSCS data.
- Each research question should focus on:
  - **Key variables** related to social connection, community engagement, or well-being.
  - **Statistical analyses or methodologies** that can answer the question (e.g., hypothesis testing, confidence intervals, regression analysis).

Project

2. **Conduct Data Analysis**

- Perform **data wrangling** and **exploratory data analysis (EDA)** on the CSCS dataset to clean and prepare the data for analysis.
- The analysis should include:
    - **Summary statistics** for key variables.
    - **Visualizations** (e.g., histograms, scatter plots) to help interpret the data.
    - **Statistical tests** or models (e.g., t-tests, linear regression) to answer the research questions.

Project

3. **Create Group Project Slides**

- Prepare a maximum of **23 slides** summarizing your project findings, including:
  - **Title Slide**: Project title, group member names, TUT number, and TA name.
  - **Introduction Slides (1-2 slides)**: Describe the overarching theme of the project and provide context for the research questions.
  - **Data Summary Slides (2-3 slides)**: Include definitions of key variables and descriptions of any data wrangling performed.
  - **Research Question Slides (3-5 slides per question)**:
    - Clearly state each research question.
    - Provide relevant visualizations and set up the analysis methodology.
    - Present the results and interpret them in the context of the research question.
  - **Limitations Slide (1-2 slides)**: Discuss any limitations in the data or analysis methods used.
  - **Conclusion Slides (1-2 slides)**: Summarize the findings from all research questions and suggest next steps or future analyses.
  - **References Slide**: Acknowledge any sources or contributors to the project.

Project

4. **Submit Group Project Slides and Presentation Recording**

- **Slides Submission (Due Mon, Dec 2)**: Submit your finalized slide deck. Ensure that it is well-organized and communicates the findings clearly to a non-technical audience.
- **Presentation Recording**:
    - Record a **4-6 minute video** where all group members present parts of the project.

Project

Grading Breakdown:

- **Individual Proposal**: 2 points (11% of project grade)
- **Practice Presentation (Nov 29)**: 2 points (11% of project grade)
- **Group Project Slides**: 8 points (45% of project grade)
- **Group Presentation Recording**: 2 points (11% of project grade)
- **Individual Q&A Performance (Poster Fair)**: 2 points (11% of project grade)
- **Individual Critiques and Reflections**: 2 points (11% of project grade)

Communication Activity

Which is continuous and which is categorical?

i	study_hours	class_section	exam_score
0	10.9934280	A	86.530831
1	9.7234711	A	84.632809
2	11.2953770	B	87.036506
3	13.0460600	C	97.952866
4	9.5316930	C	79.749848

Communication Activity

Communication Activity

1. How could you use ONLY TWO **binary indicator variables** in combination to represent the ALL THREE levels (A, B, and C) in the example above?

$$I_{[x_i = \text{"B"}]}(x_i) = \begin{cases} 1 & \text{if } x_i = B \\ 0 & \text{o/w} \end{cases}$$

$$I_{[x_i = \text{"C"}]}(x_i) = \begin{cases} 1 & \text{if } x_i = B \\ 0 & \text{o/w} \end{cases}$$

Communication Activity

2. What are the **means** of the different `class_section` groups in terms of the parameters of the following model specification?

$$Y_i = \beta_0 + 1_{[x_i=\text{"B"}]}(x_i)\beta_1 + 1_{[x_i=\text{"C"}]}(x_i)\beta_2 + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

$$Y_i = \beta_0 + I_{[x_i = \text{"B"}]}\beta_1 + I_{[x_i = \text{"C"}]}\beta_2 + \varepsilon_i$$

	$I_{[x_i = \text{"B"}]}$	$I_{[x_i = \text{"C"}]}$
A :	0	0
B :	1	0
C :	0	1

Communication Activity

2. What are the **means** of the different `class_section` groups in terms of the parameters of the following model specification?

$$Y_i = \beta_0 + 1_{[x_i=\text{"B"}]}(x_i)\beta_1 + 1_{[x_i=\text{"C"}]}(x_i)\beta_2 + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

For A: $Y_i = \beta_0 + 0 + 0 + \varepsilon_i$

$E(Y_i) = E(\beta_0) = \beta_0$

For B: $Y_i = \beta_0 + \beta_1 + 0 + \varepsilon_i$

$E(Y_i) = E(\beta_0 + \beta_1) = \beta_0 + \beta_1$

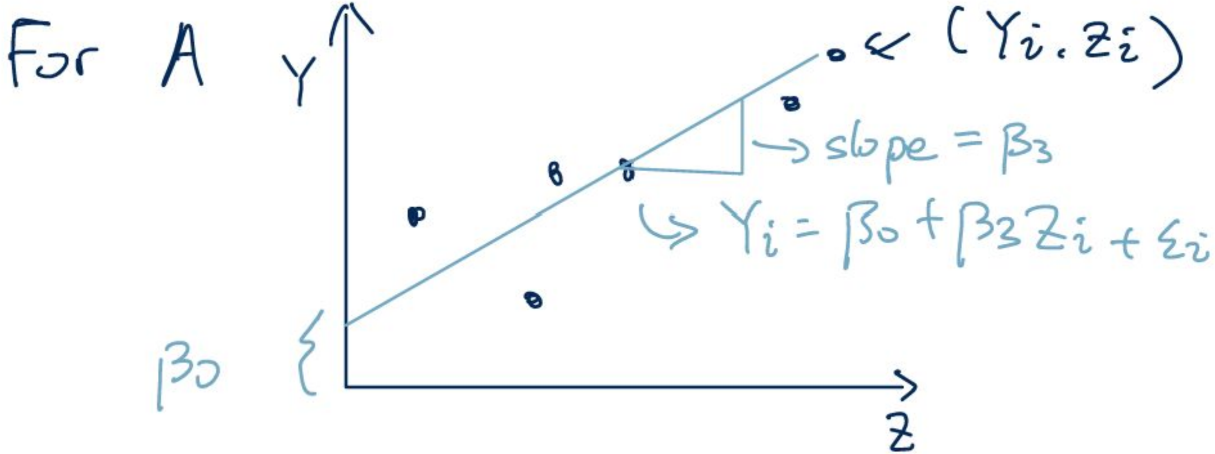
For C: $Y_i = \beta_0 + 0 + \beta_2 + \varepsilon_i$

$E(Y_i) = E(\beta_0 + \beta_2) = \beta_0 + \beta_2$

Communication Activity

3. What is the nature of the data generated under the following model specification if Y_i is the `exam_score` of **observation** i, z_i is the value of `study_hours` for **observation** i, and x_i is as described above?

$$Y_i = \beta_0 + 1_{[x_i=\text{"B"}]}(x_i)\beta_1 + 1_{[x_i=\text{"C"}]}(x_i)\beta_2 + \beta_3 z_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$



Communication Activity

3. What is the nature of the data generated under the following model specification if Y_i is the `exam_score` of **observation** i, z_i is the value of `study_hours` for **observation** i, and x_i is as described above?

$$Y_i = \beta_0 + 1_{[x_i = \text{"B"}]}(x_i)\beta_1 + 1_{[x_i = \text{"C"}]}(x_i)\beta_2 + \beta_3 z_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

- The model describes a **linear relationship** between `exam_score` and `study_hours`, with different intercepts for each `class_section` group:
  - For "A": $Y_i = \beta_0 + \beta_3 z_i + \epsilon_i$
  - For "B": $Y_i = (\beta_0 + \beta_1) + \beta_3 z_i + \epsilon_i$
  - For "C": $Y_i = (\beta_0 + \beta_2) + \beta_3 z_i + \epsilon_i$

Communication Activity

4. What is the practical interpretation of how `exam_score` changes relative to `class_section` according to the model specification of the previous question if β_1 and β_2 are not 0?

$$Y_i = \beta_0 + I_{[x_i = \text{"B"}]}\beta_1 + I_{[x_i = \text{"C"}]}\beta_2 + \varepsilon_i$$

	$I_{[x_i = \text{"B"}]}$	$I_{[x_i = \text{"C"}]}$
A :	0	0
B :	1	0
C :	0	1

Communication Activity

4. What is the practical interpretation of how `exam_score` changes relative to `class_section` according to the model specification of the previous question if β_1 and β_2 are not 0?

- If $\beta_1 \neq 0$ and $\beta_2 \neq 0$:
    - There are **differences in intercepts** between the "A", "B", and "C" groups.
    - This suggests that the **average exam score** differs based on the `class_section`, even after accounting for `study_hours`.
- If $\beta_1 = 0$ and $\beta_2 = 0$:
    - The intercept is the same across all groups, implying no significant difference in `exam_score` across `class_section` groups.

Communication Activity

5. What is the practical interpretation of the behavior of the relationship between `exam_score` and `study_hours` within different `class_section` groups according to the model specification of the previous question?

- The relationship is described by a **single slope** (β_3), implying the **rate of change** of `exam_score` per unit increase in `study_hours` is the same for all `class_section` groups.
- The difference lies only in the intercepts (β_0, $\beta_0 + \beta_1$, $\beta_0 + \beta_2$).

Communication Activity

5. What is the practical interpretation of the behavior of the relationship between `exam_score` and `study_hours` within different `class_section` groups according to the model specification of the previous question?

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	36.3380	12.397	2.931	0.209	-121.186	193.862
is_B	-2.9994	3.968	-0.756	0.588	-53.420	47.421
is_C	-1.1537	3.299	-0.350	0.786	-43.078	40.770
study_hours	4.7540	1.178	4.036	0.155	-10.212	19.720

Communication Activity

6. Is there a different kind of behavior that could be seen for the relationship between `exam_score` and `study_hours` between different `class_section` groups that might be different than what's prescribed by the model specification of the previous question?

   1. Hint 1: what is the meaning of the following model specification?

$$Y_i = \beta_0 + \beta_3 z_i + 1_{[x_i=\text{"B"}]}(x_i)\beta_1 + \beta_4 z_i \times 1_{[x_i=\text{"B"}]}(x_i) + 1_{[x_i=\text{"C"}]}(x_i)\beta_2 + \beta_5 z_i \times 1_{[x_i=\text{"C"}]}(x_i) + \epsilon_i \quad \text{where}$$

$$\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

- The model with interaction terms:
  $$Y_i = \beta_0 + \beta_3 z_i + 1_{[x_i=\text{"B"}]}(x_i)\beta_1 + \beta_4 z_i \times 1_{[x_i=\text{"B"}]}(x_i) + 1_{[x_i=\text{"C"}]}(x_i)\beta_2 + \beta_5 z_i \times 1_{[x_i=\text{"C"}]}(x_i) + \epsilon_i$$
- Here, the slopes ($\beta_3 + \beta_4$ for "B" and $\beta_3 + \beta_5$ for "C") differ between groups.
- This specification allows the **relationship between `exam_score` and `study_hours` to vary** depending on the `class_section` group, capturing potential differences in how `study_hours` impact `exam_score` across groups.

Communication Activity

6. Is there a different kind of behavior that could be seen for the relationship between `exam_score` and `study_hours` between different `class_section` groups that might be different than what's prescribed by the model specification of the previous question?

   1. Hint 1: what is the meaning of the following model specification?

$$Y_i = \beta_0 + \beta_3 z_i + 1_{[x_i=\text{"B"}]}(x_i)\beta_1 + \beta_4 z_i \times 1_{[x_i=\text{"B"}]}(x_i) + 1_{[x_i=\text{"C"}]}(x_i)\beta_2 + \beta_5 z_i \times 1_{[x_i=\text{"C"}]}(x_i) + \epsilon_i \quad \text{where}$$

$$\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

```
# Step 2: Fit the basic model without interaction terms
model_basic = smf.ols('exam_score ~ is_B + is_C + study_hours', data=df).fit()

# Step 3: Fit the model with interaction terms
model_interaction = smf.ols('exam_score ~ is_B * study_hours + is_C * study_hours', data=df).fit()
```