Lecture 4: Simple Linear Regression

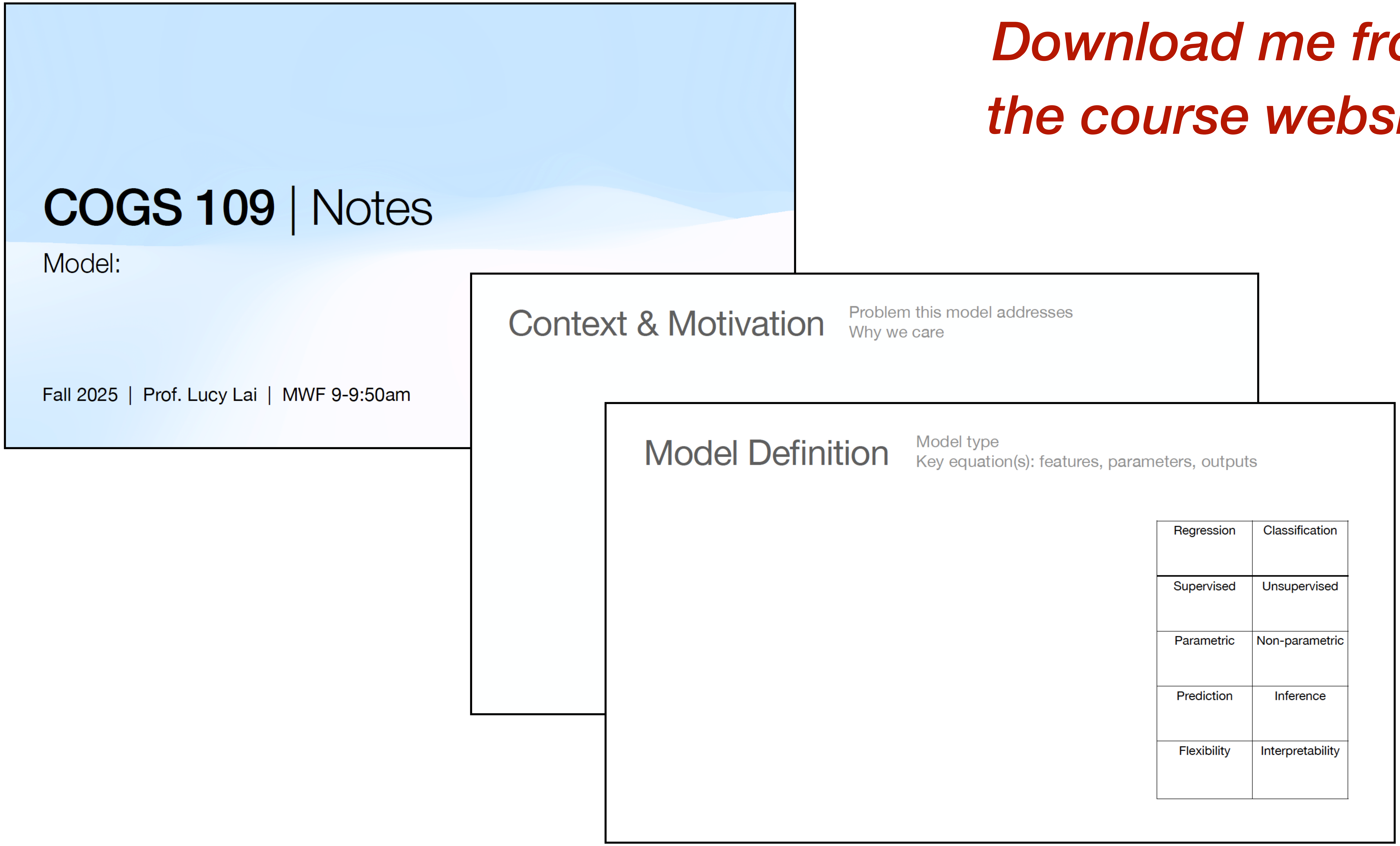
Monday, Oct 6

Going forward

As we discuss each model, lectures will roughly follow this general format:

1. Context & motivation (why we care / how it's useful)
2. Model definition (key equations, features, parameters, outputs)
3. Core assumptions (when to use the model)
4. Model fitting (how to fit model to data)
5. Interpretation & intuition (how to communicate what it means)
6. Model assessment (how good your model is)
7. Applications & examples
8. Strengths & limitations

Lecture notes template

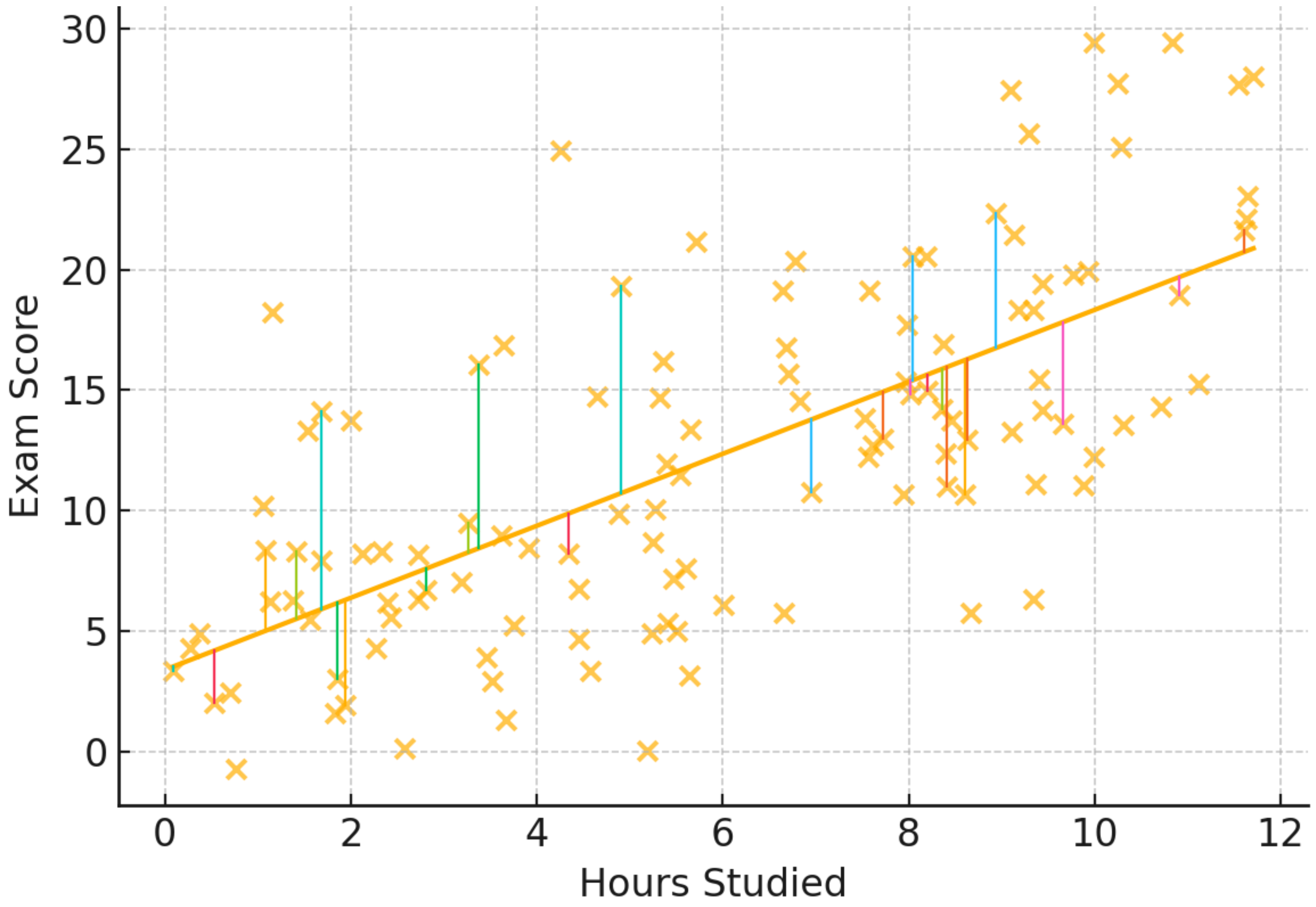


Download me from the course website!

Context & Motivation

Problem this model addresses
Why we care

- **Simple linear regression:** understanding the relationship between two variables
- “Basic” but teaches ***important*** core ideas and sets us up for more complex models (e.g., multiple and polynomial regression, GAMs, etc.)



$$Y = mx + b$$

Model Definition

Model type
Key equation(s): features, parameters, outputs

$$Y = f(X) + \epsilon$$

Model Definition

Model type
Key equation(s): features, parameters, outputs

unknown function we are trying to estimate

$$Y = f(X) + \epsilon$$

output, response

predictors, features

irreducible error

Model Definition

Model type
Key equation(s): features, parameters, outputs

unknown *parameters* we are trying to estimate

$$Y = \beta_0 + \beta_1 X + \epsilon$$

output, response

predictors, features

irreducible error

β_0: "slope" or expected change in Y for a 1-unit increase in X

β_1: "y-intercept," or expected Y when X = 0

Model specs

Regression	Classification
Supervised	Unsupervised
Parametric	Non-parametric
Prediction	Inference
Flexibility: LOW	Interpretability: HIGH

Model Definition

Model type
Key equation(s): features, parameters, outputs

unknown *parameters* we are trying to estimate

$$Y = \beta_0 + \beta_1 X + \epsilon$$

output, response

predictors, features

irreducible error

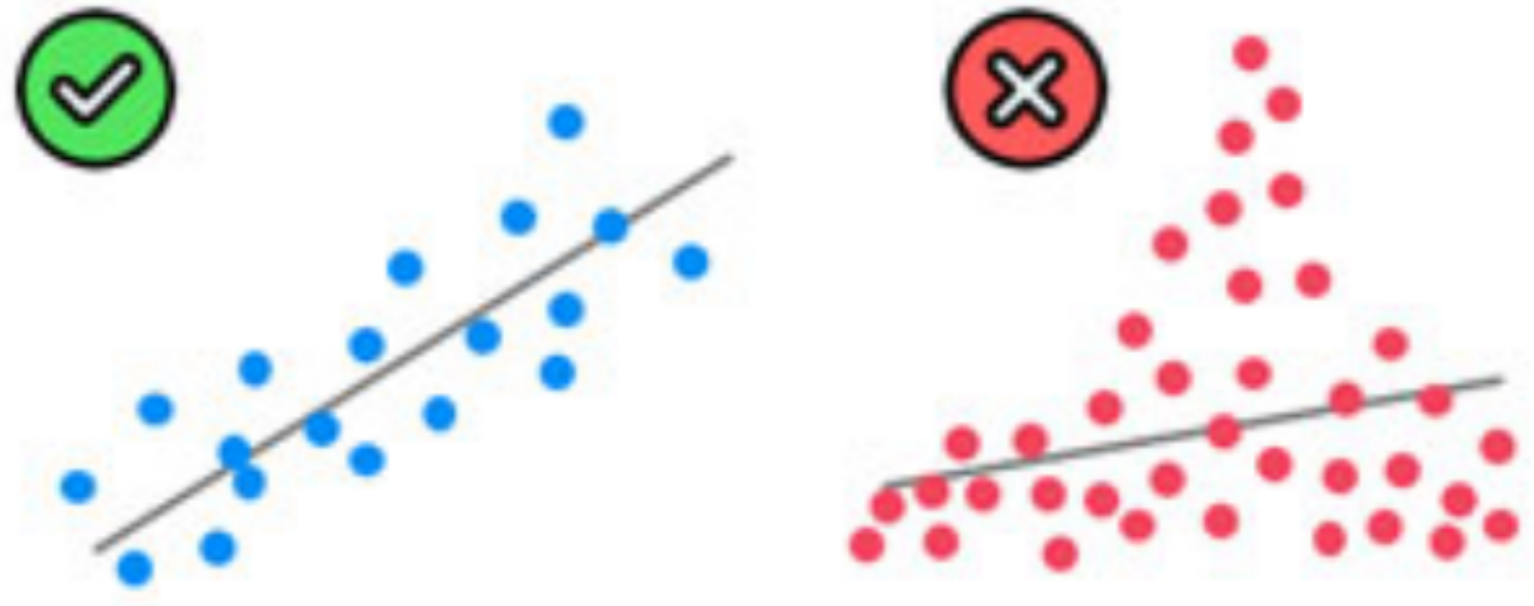
H_0: null hypothesis is that $\beta_1 = 0$; there is *no* relationship between X and Y

H_A: alternative hypothesis is that $\beta_1 \neq 0$; there is a *significant relationship* between X and Y

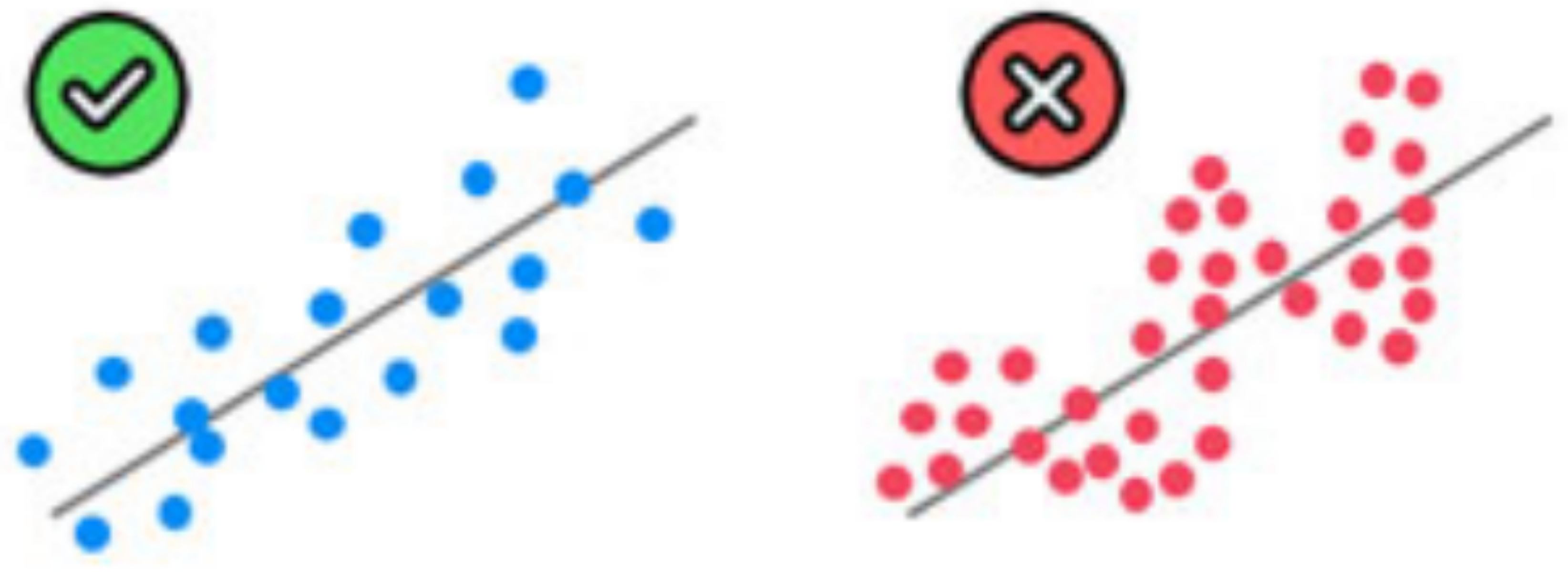
Assumptions

Core assumptions
When to use the model

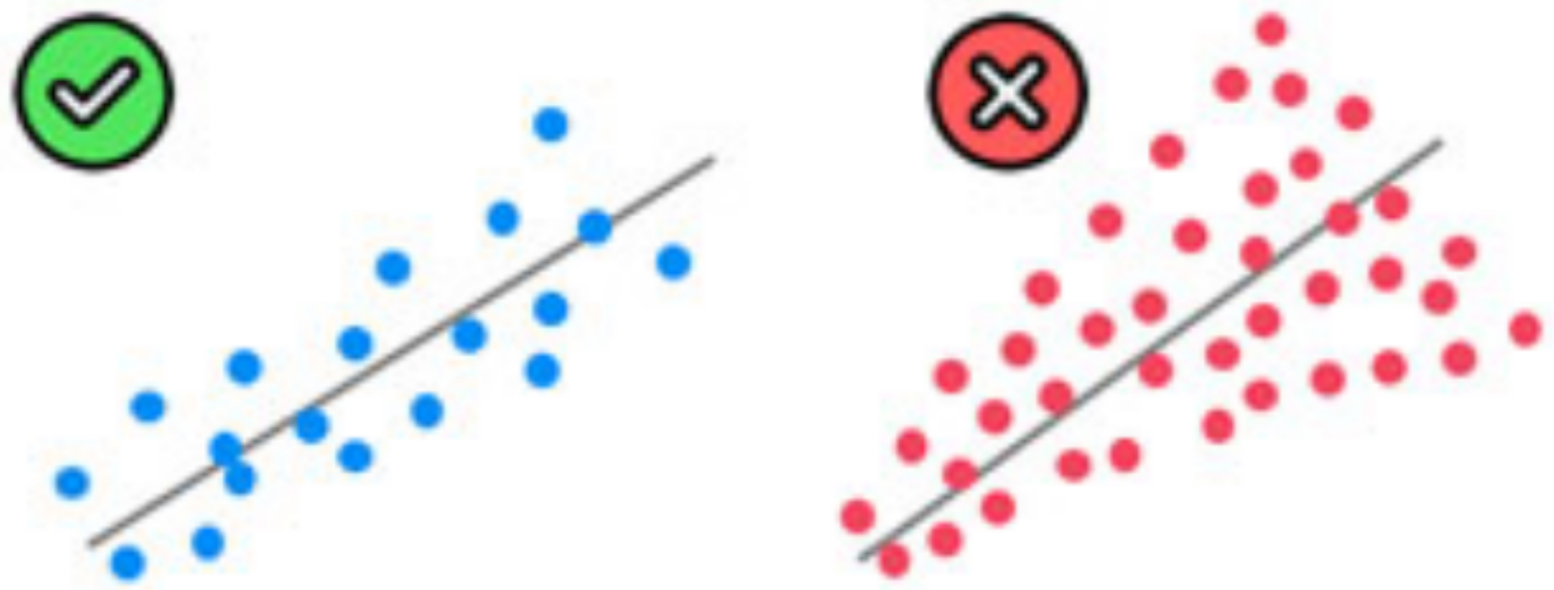
Linearity:
linear relationship between X and Y



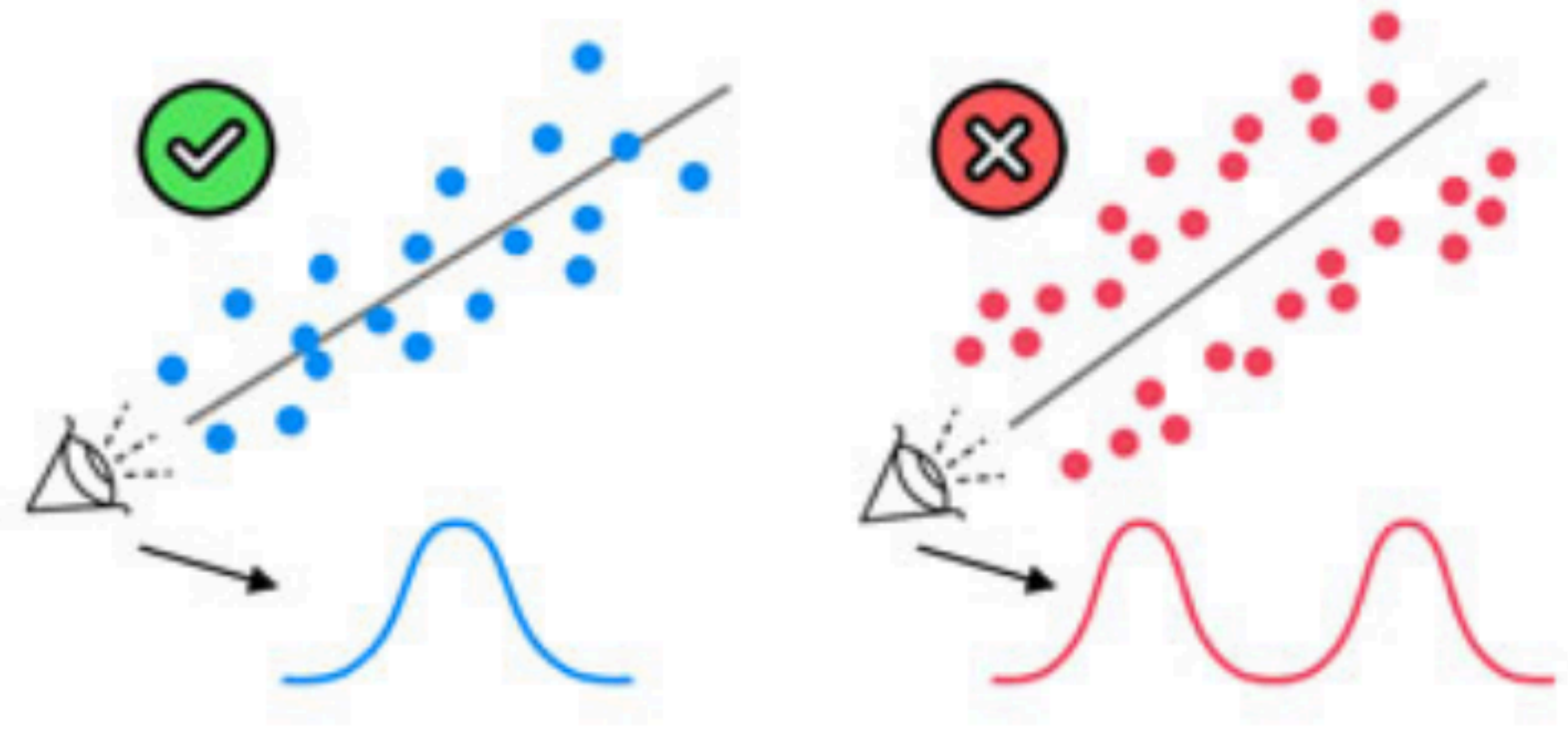
Independent errors:
errors are uncorrelated



Homoscedasticity:
errors have constant variance



Multivariate normality:
errors are normally distributed



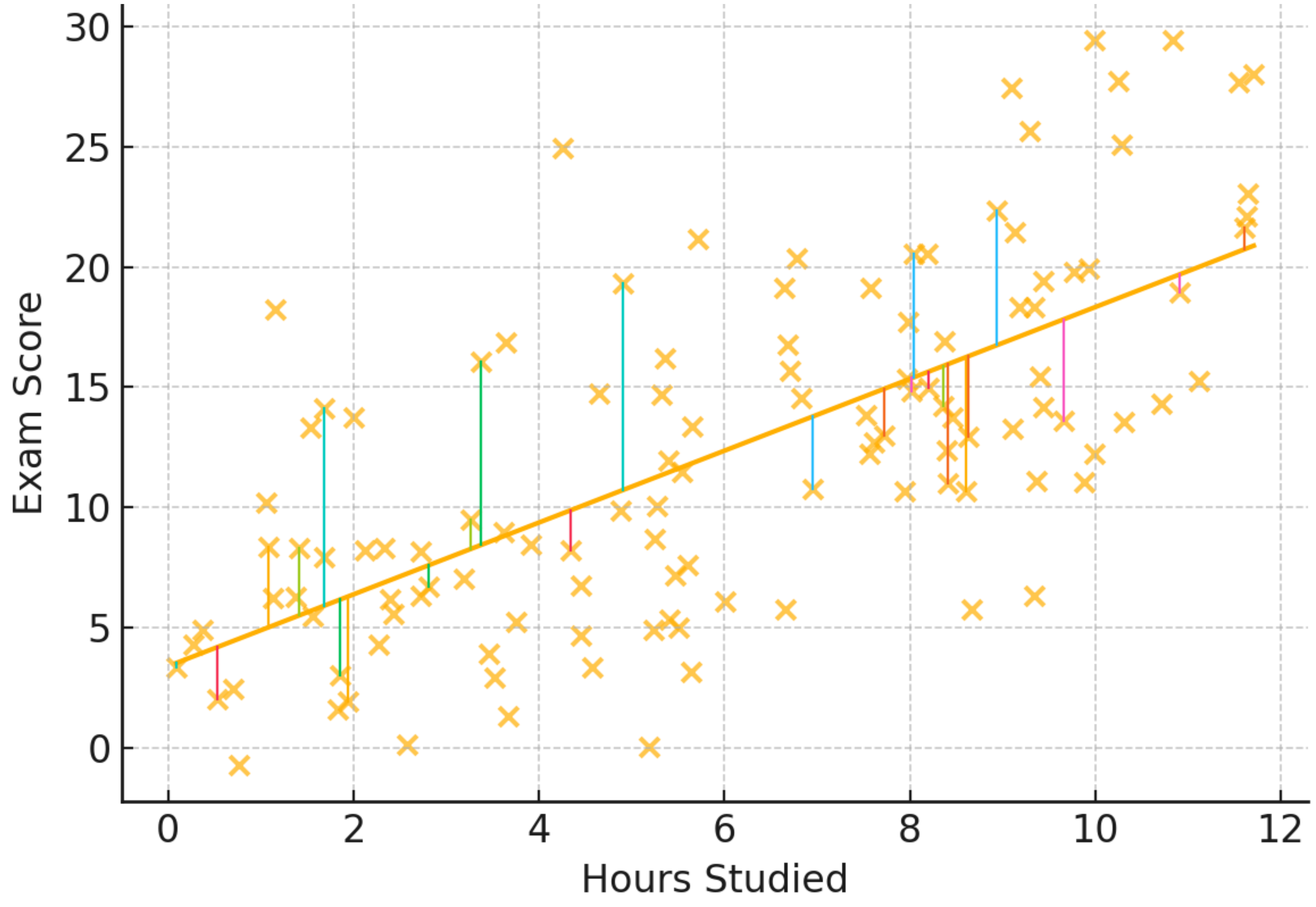
Model Fitting

How to fit the model to data
Parameter estimation
Tools / packages

Choose $\hat{\beta}_0, \hat{\beta}_1$ that minimizes the ***residual sum of squares (RSS):***

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

"Ordinary least squares" (OLS) method



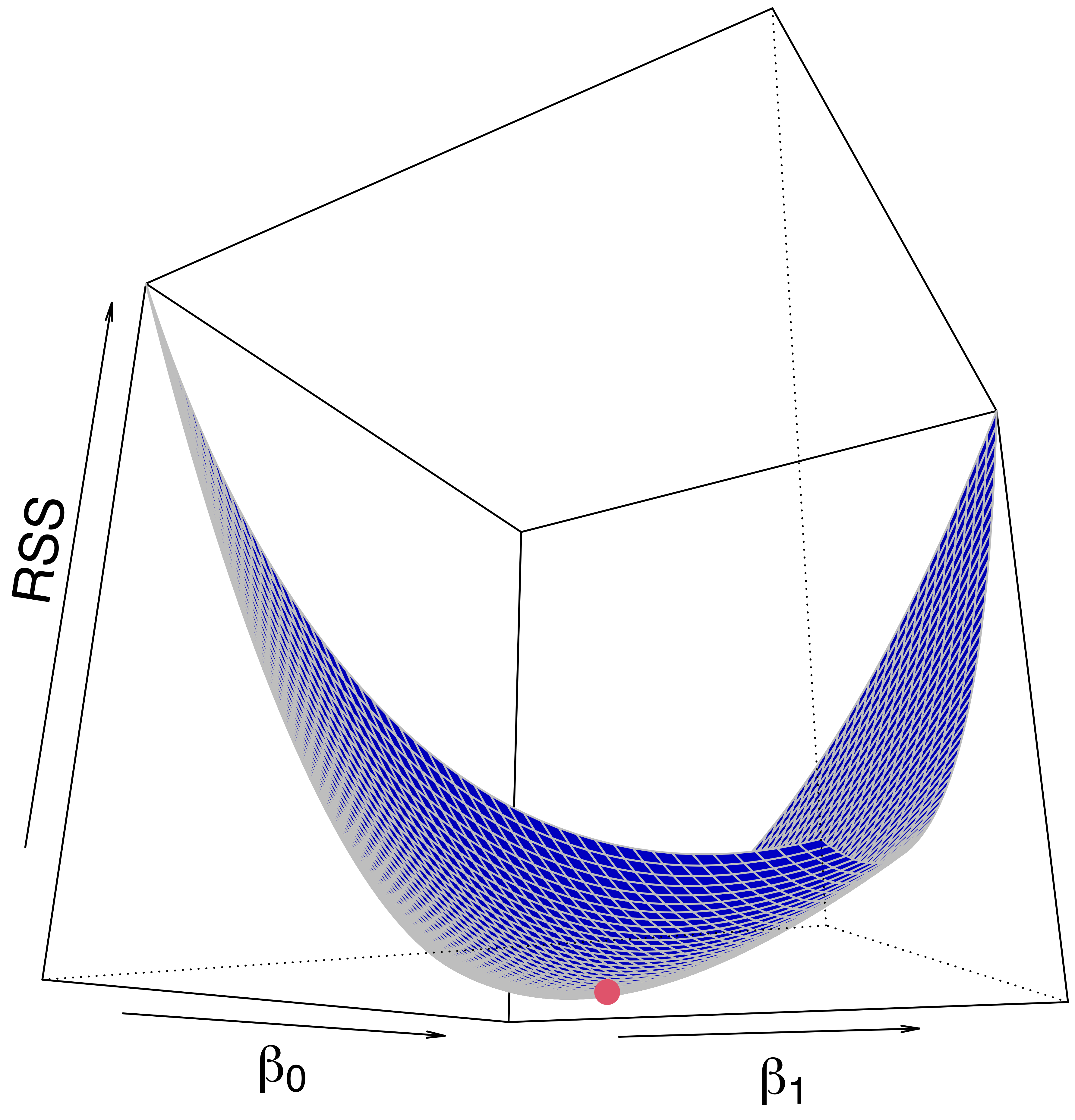
Model Fitting

How to fit the model to data
Parameter estimation
Tools / packages

Choose $\hat{\beta}_0, \hat{\beta}_1$ that minimizes the ***residual sum of squares (RSS):***

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

Does this equation look somewhat familiar?



Model Fitting

How to fit the model to data
Parameter estimation
Tools / packages

Choose $\hat{\beta}_0, \hat{\beta}_1$ that minimizes the ***residual sum of squares (RSS):***

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

Closed-form solutions (you can derive it!):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Model Fitting

How to fit the model to data
Parameter estimation
Tools / packages

Python: use statsmodels to estimate parameters

```
import statsmodels.api as sm
X = sm.add_constant(x)    # adds intercept
model = sm.OLS(y, X).fit()
model.summary()           # β-hats, SEs, t, p
```

...and scikit-learn to predict

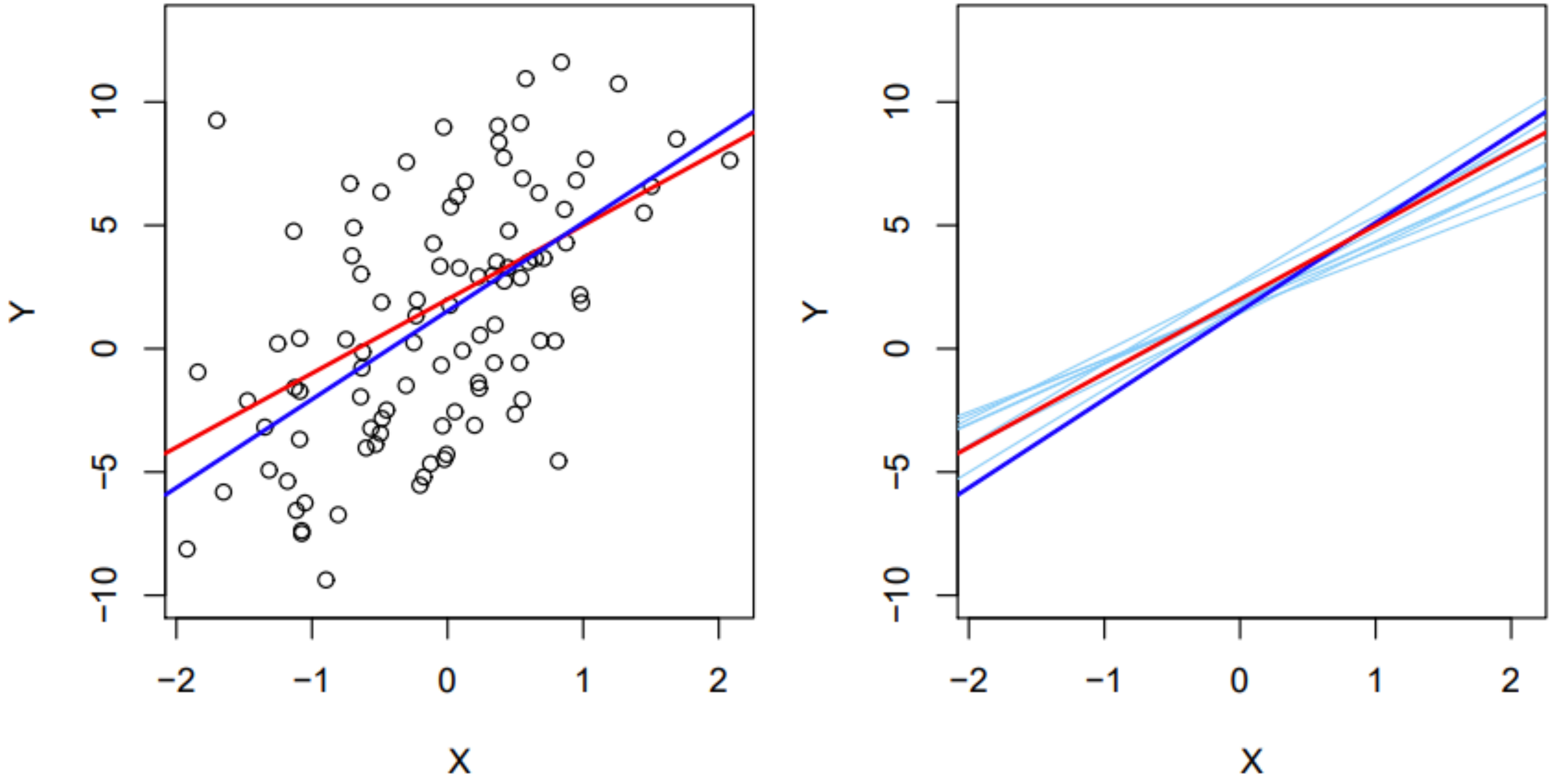
```
from sklearn.linear_model import LinearRegression
lr = LinearRegression().fit(x.reshape(-1,1), y)
lr.coef_, lr.intercept_   # no SE/p-values
```

Interpretation & Intuition

Interpreting the model
Understanding the parameters
Communicating findings

Population line (unknown): $Y = \beta_0 + \beta_1 X + \epsilon$

Least-squares line (estimated from a *sample*): $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \epsilon$



The average of many different least-squares lines ⇒ population line

Interpretation & Intuition

Interpreting the model
Understanding the parameters
Communicating findings

Standard errors: quantify the sampling variability of a parameter; the average amount that $\hat{\beta}_0$ and $\hat{\beta}_1$ differs from the true β_0 and β_1

$$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \qquad \mathrm{SE}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

Confidence intervals (CIs): a range of values (i.e., $\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$) that will contain the true unknown value of the parameter with 95% probability.

"There is approximately a 95% chance that the interval $[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)]$ will contain the true value of β_1."

Model Assessment

Goodness-of-fit and performance metrics
Model diagnostics

Residual standard error (RSE): $\text{RSE} = \sqrt{\dfrac{\text{RSS}}{n-2}}$ where $\text{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$

⇒ Predictions tend to deviate from the regression line by RSE (in units of Y)

⇒ An **absolute measure** of the lack of fit of the model

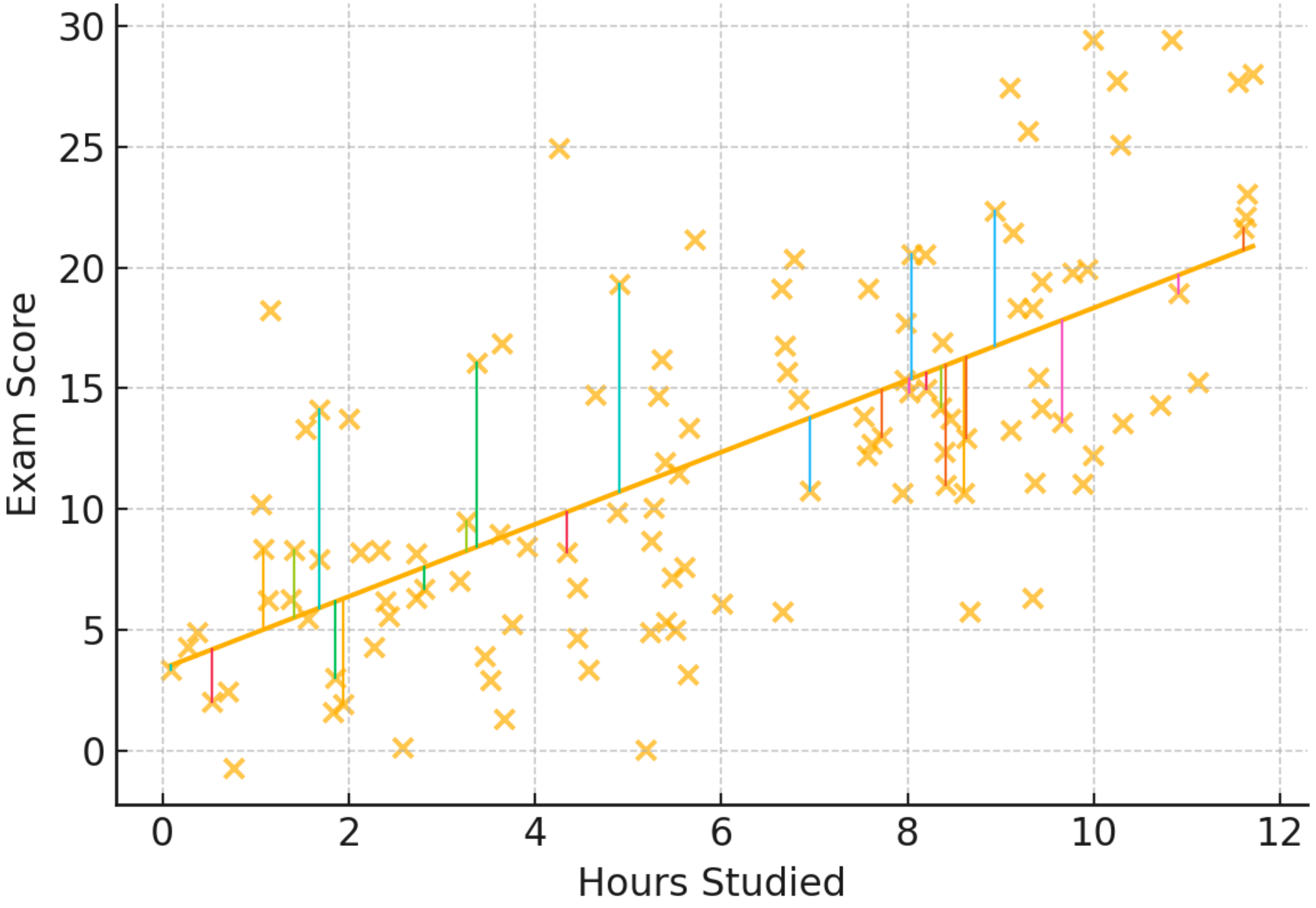
$$R^2 = 1 - \frac{RSS}{TSS} \text{ where TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

⇒ The **proportion** of variance in Y explained by X $\in [0,1]$

⇒ For simple linear regression, $R^2 = r^2$ (Pearson correlation squared)

Applications & Examples

$$ExamScore = \beta_1 + \beta_0 \cdot HoursStudied$$



Groups of 2-3 (6 mins):

1. What is H_0 and H_A?
2. You fit the model and find that $\hat{\beta}_0 = 3.4$ and $\hat{\beta}_1 = 1.5$. How would you interpret each of the coefficients?
3. Suppose the 95% CIs for $\beta_0 = [1.8{,}5.0]$ and for $\beta_0 = [1.1{,}1{,}9]$. Write one sentence of interpretation for each CI.
4. You compute RSE=5.2 and R^2=0.62. Interpret each metric in the context of the problem.

The person who is older will share!

Applications & Examples

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.316
Model:                            OLS   Adj. R-squared:                  0.315
Method:                 Least Squares   F-statistic:                     358.4
Date:                Mon, 06 Oct 2025   Prob (F-statistic):           5.46e-66
Time:                        05:20:26   Log-Likelihood:                -7403.3
No. Observations:                 777   AIC:                         1.481e+04
Df Residuals:                     775   BIC:                         1.482e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       6906.4586    221.614     31.164      0.000    6471.424    7341.493
x1           128.2437      6.774     18.931      0.000     114.946     141.541
```

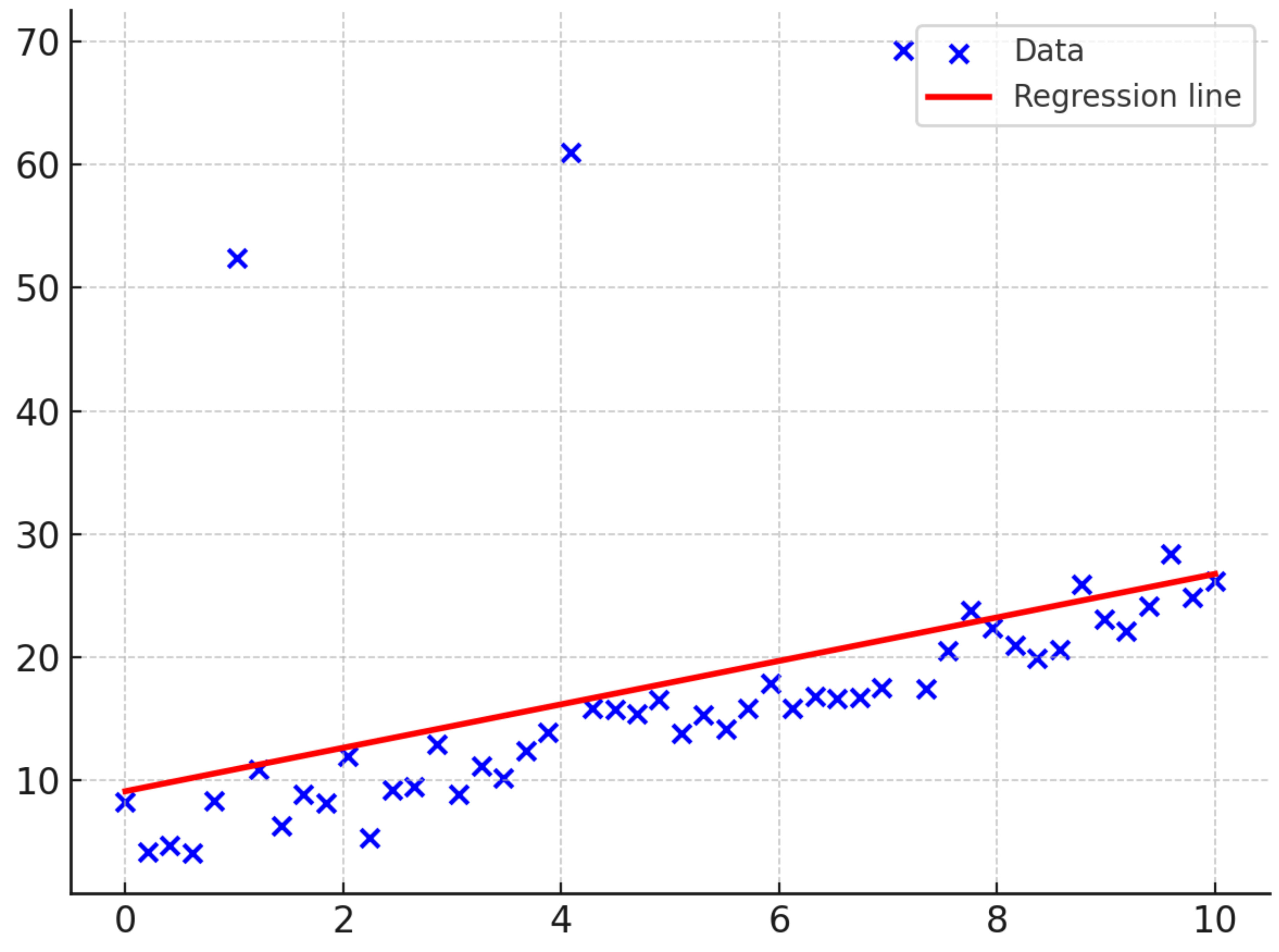
95% CI for coefficients

Strengths & Limitations

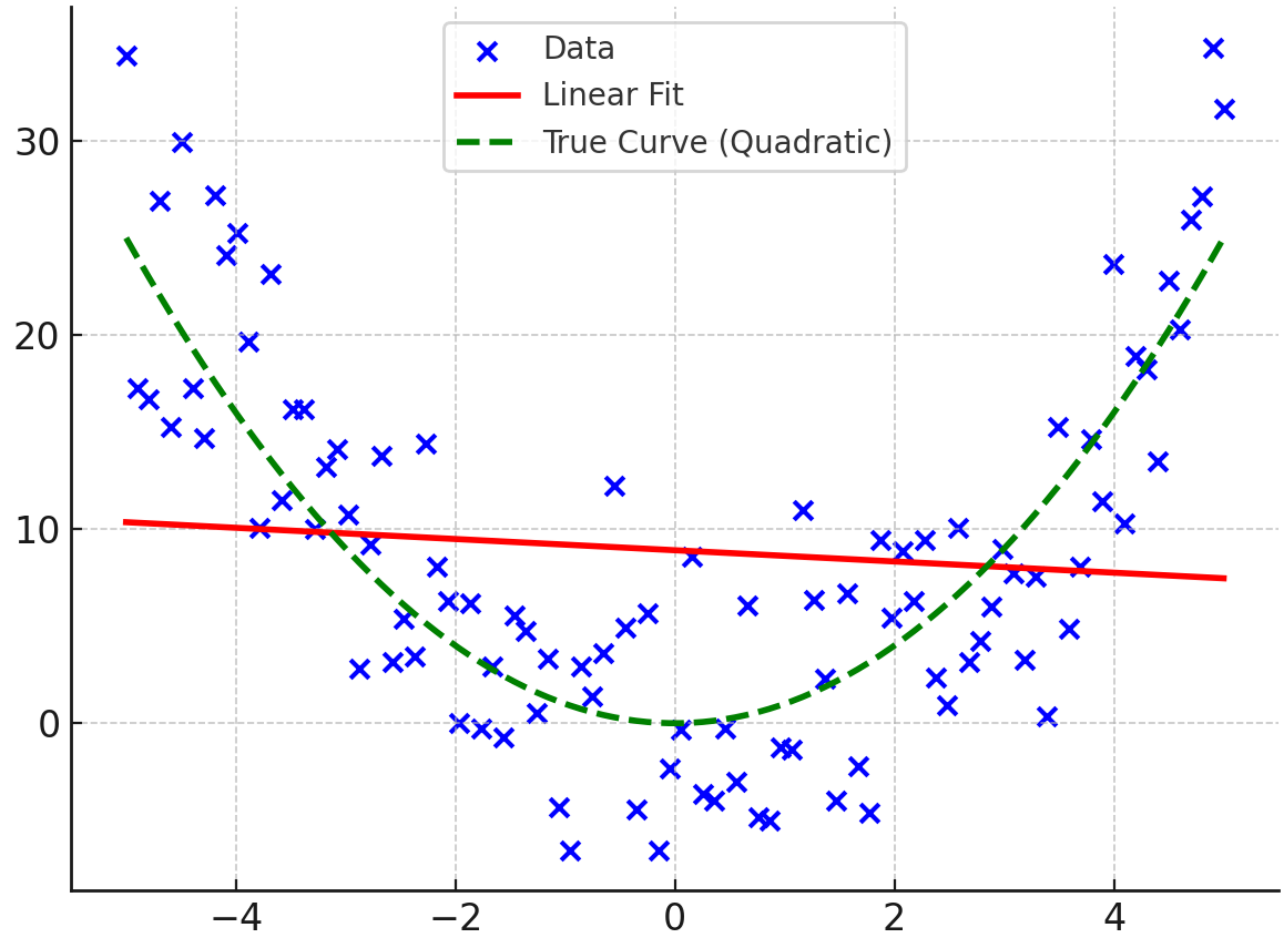
Strengths: simple, fast, interpretable, can use it for inference

Limitations: assumes linearity; sensitive to outliers; limited flexibility

Extreme outliers



Misspecified model or underfitting on training set



Upcoming + Reminders

Updates:

- We are currently sorting groups and will release them by Wednesday
- HW1 feedback will be done by the end of today
- Quiz 1 feedback done by end of Wednesday

Assignments:

- Quiz 1 (DUE: TODAY @ 11:59pm)

Wednesday's topic: ***Multiple Linear Regression***

- **Read:** ISLP Ch. 3.2

Questions?