Scenario 1

① Do the residuals indicate any problems with the fit?
② If densities (y) is not reported exactly, how might this affected the fit?
③ What if the blocks were not measured in a random order?

Regression:

① Draw Scatterplot: ⇒ Identify which one is response variable, which one is explanatory variable.
+ Residual Plot — x-axis ↑ y-axis ↑

⇒ Analyze relationship: sign: positive / negative
shape: linear / quadratic / etc.
level of correlation: strong / moderately strong / weak

② Calculate Correlation: Def: strength of the linear association between two variables. $[-1, 1]$

Population mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \cdots + X_n}{n}$

Population variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Population correlation: R (empirical Correlation Coefficient)

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i \cdot Y_i) - (\bar{X}\cdot\bar{Y})}{S_X \cdot S_Y}$$

slope of regression: $\hat{\beta}_1 = \frac{S_y}{S_x} \cdot R$

Interpretation: For each additional % point in explanatory variable (x), we would expect the % point in response variable (y) increase on average $\hat{\beta}_1$ points.

Intercept: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

Intuition: A regression line always passed through $(\bar{X}, \bar{Y})$

Interpretation: With no explanatory variable (x), we expect on average to have $\hat{\beta}_0$ of response variable (y).

③ Fitting:

⇒ Residual: distance between observed and predicted

Goal: We want a line with small residuals

Formula: $\hat{\varepsilon}_i = Y_i - \hat{\beta}_1 X_i - \hat{\beta}_0$

$(\varepsilon_i = Y_i - \hat{Y}_i)$

⇒ least square: Def: = Minimizes the sum of the squares of data point to regression line.

formula = $\text{Min}\sum_{i=1}^{n}(\text{Residuals}_i)^2$

$= \min_{\beta_0 \beta_1}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$

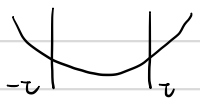
Why Least Square: ① most commonly used
② Easier to compute
③ A residual twice as large as another is usually more than twice as bad.

⇒ least absolute deviation: Def: Minimizes the sum of the data point to regression line.

formula = $\text{Min}\sum_{i=1}^{n}|\text{Residual}_i|$

$= \min_{\beta_0, \beta_1}\sum_{i=1}^{n}|Y_i - \beta_0 - \beta_1 X_i|$

⇒ $\sum_{i=1}^{n} h(\text{residuals}_i)$, $h(\text{residuals}_i) = \begin{cases} |x| & |x| > \tau \\ x^2 & |x| < \tau \end{cases}$



$\tau > 0$

Note: The statements are not causal, unless the study is a randomized controlled experiment.

Scenario 2 :

① adding bands around least square line ⇒ make estimates (Prediction Interval + Confidence Interval)

Note : several measurement of gain was taken for one density

④ Predicting :

* Fitted Value ⟺ Old Value

  New Value ⟺ New Value

* Definition :

  ⇒ Prediction : Give explanatory variable (x) to predict response variable (y) (Plug in x to the model)

  ⇒ Extrapolation : Apply a model estimate to values outside of the realm of the original data ($x=0$) ← unreachable data

* Least Square : $\varepsilon_i \sim N(0, \sigma^2)$ $\hat{\beta_1}$ is MLE

  Least Absolute Value : $\varepsilon_i \sim$ expl) $\hat{\beta_1}$ is MLE

* Requirement :

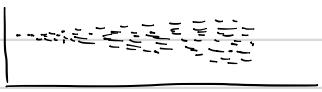
  ① Plot Residuals + Scatterplot, check linearity

  ② Plot QQ-plot + residuals histogram, Check if residuals plot ~ Normal.

  ③ If use scatterplot to plot residuals, should only see scatter plots without clear pattern. (x ⇒ explanatory variable. y ⇒ residuals)

  [homoscedasticity] ⇒ constant variability

  [13.1] Residual Plot

   ← violation ?

* R^2 :

  ⇒ Definition : The strength of the fit of a linear model

  ⇒ formula : $= \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i \cdot y_i) - (\bar{x} \cdot \bar{y})}{s_x \cdot s_y} \right)^2$ ⇒ smaller than R

  ⇒ Interpretation : [what percent of variability in the response variable is explained by the model. The remainder of the variability is explained by variables not included in the model / due to randomness in data]

  $R^2 \times 100\%$ of the variability in the response variable is explained by the model.

  ⇒ Adjusted R^2 : (If we have more than 2 variables ⇒ gives more accurate interpretation)

  * Categorical Data :

    ↳ reference level = (categorical data = 0) = (y = intercept)

    ↳ Use LS to test if X and Y are independent

    H_0 : $\beta_1 = 0$ H_1 : $\beta_1 \neq 0$

Which region (northeast, midwest, west, or south) has the lowest poverty percentage?

$H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$ Hypothesis testing

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

(Intercept) ← 9.50 : $\hat{\beta}_0$

region4midwest, region4west, region4south ← slope based on reference level : $\hat{\beta}_1$

Std. Error of estimator

→ probability of getting non-zero β_1 value

$$SE_{\hat{\beta}} = \frac{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\frac{|\hat{\beta} - \beta|}{SE_{\hat{\beta}}} \sim T_{n-4}$$

$$= \frac{|\text{point estimate} - \text{null}|}{SE}$$

⑤ Outlier:

* Distinguish if that outlier is influential / high leverage

Relationship between biological IQ & foster IQ

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
Multiple R-squared: 0.7779,Adjusted R-squared: 0.769
F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

n-2

Confidence Interval: Estimate ± SE × t_{n-2}