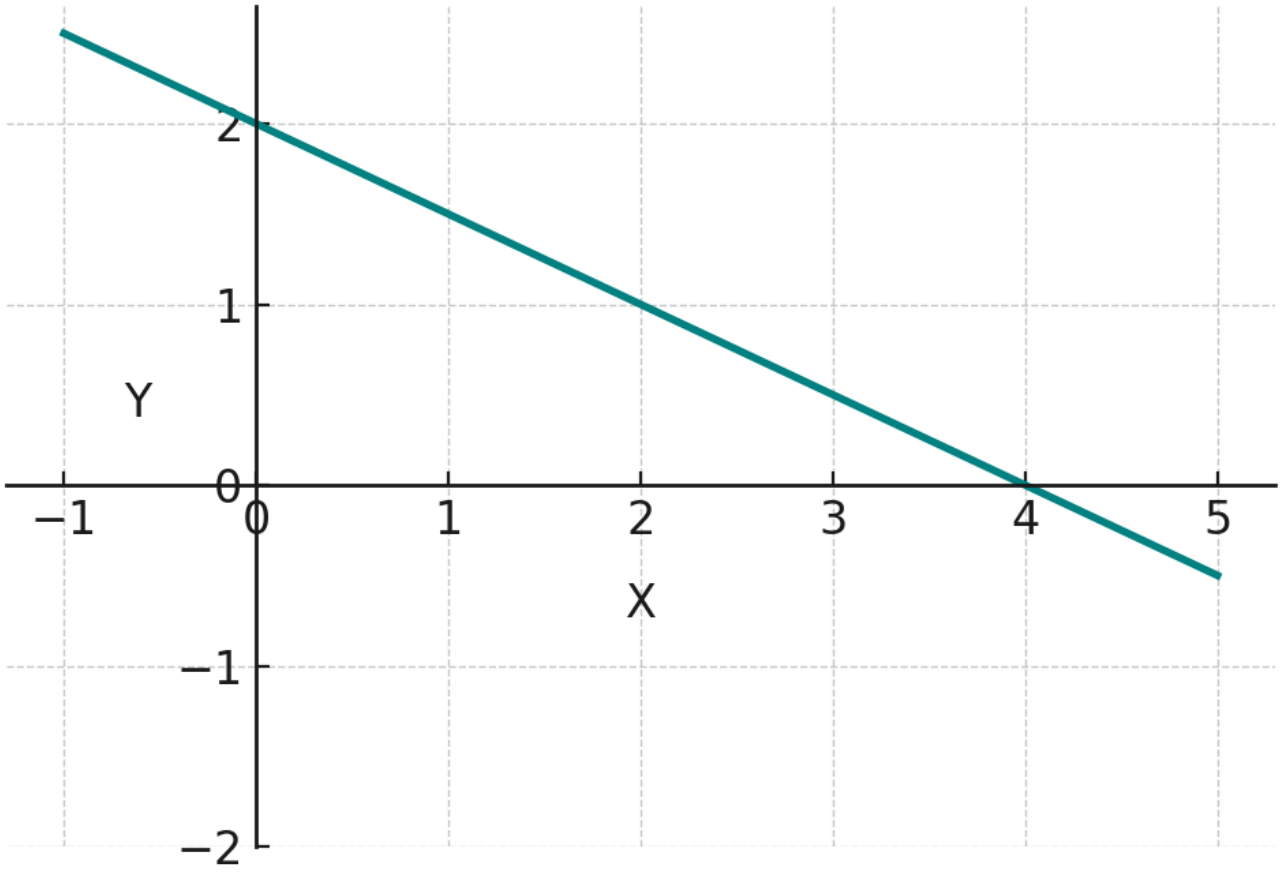
PUBH 526

Topic 1, Part 3: Regression Review

Prof. Nilupa Gunaratna

Department of Public Health

Equation for a Line



Simple Linear Regression

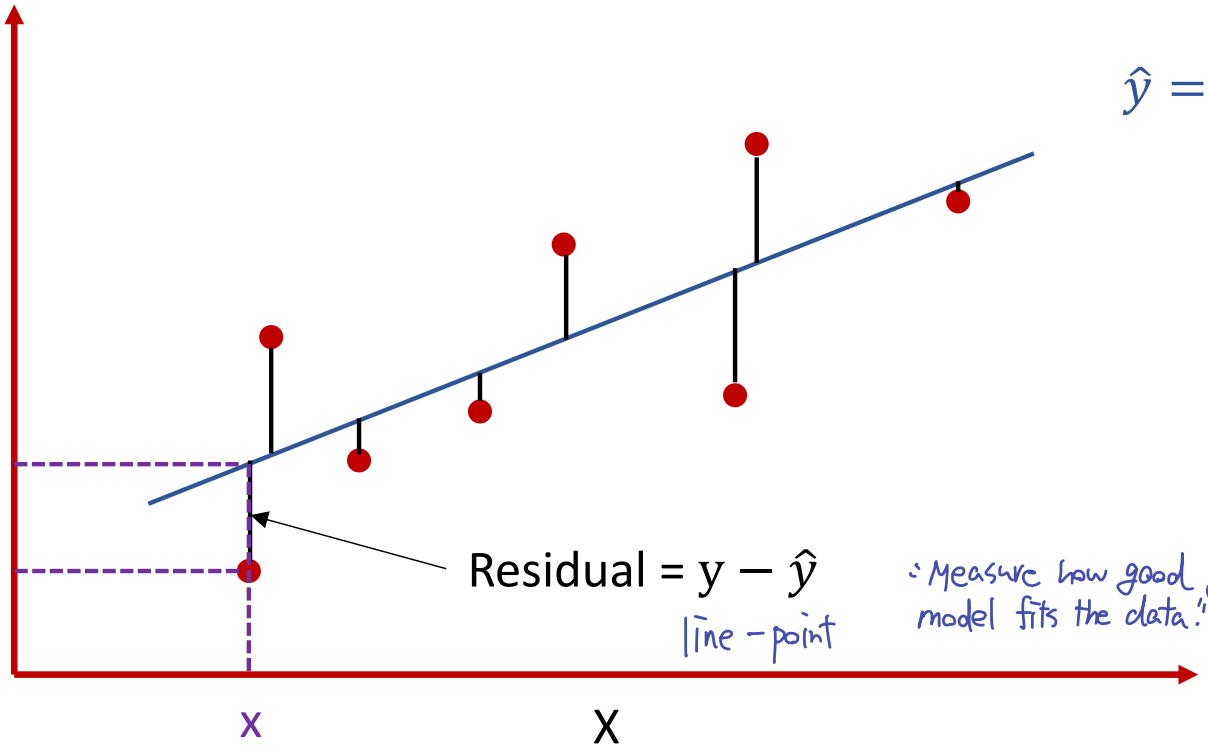
Σ residuals = ?

Σ (residuals)^2 = ?

sum of squares

Why do we care about the latter?

$$\hat{y} = b_0 + b_1 x$$



Y

Dependent variable

Response

Outcome

Target

Label

Output

X

Independent variable

Predictor

Feature

Explanatory variable

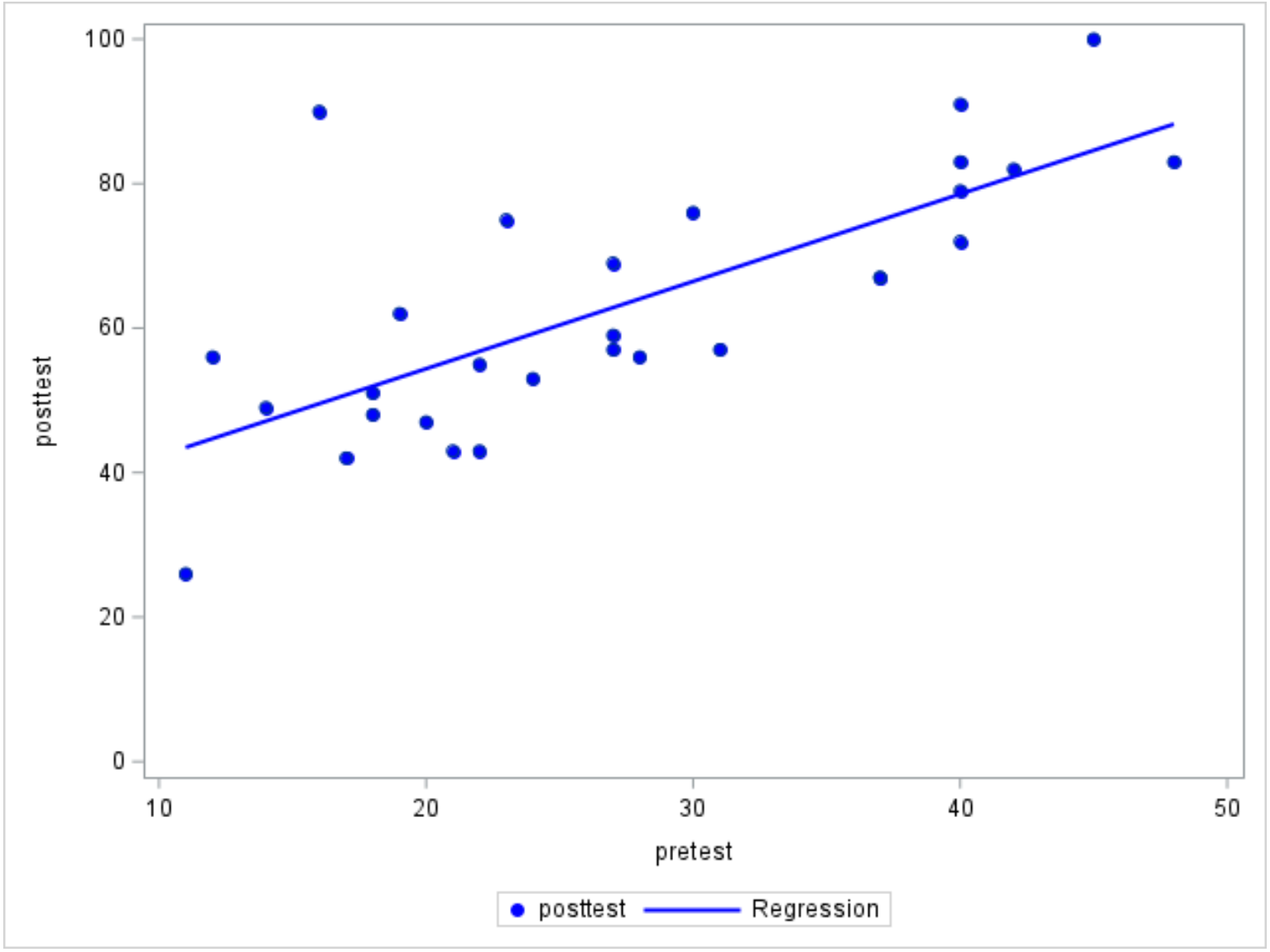
Input

Covariate

Simple Linear Regression

- A local food bank conducted a workshop on healthy meal planning and family nutrition
- All participants took:
  - A pre-test to assess their prior knowledge
  - A post-test to assess their knowledge after attending the workshop
- For each test, the maximum score was 100 points
- The workshop coordinator expected that participants who started out with better scores on the pre-test would do better on the post-test
  - ➢ Linear regression: Y=post-test score, X=pre-test score

Always Visualize Your Data



Regression Results

Source	DF	Sum of Squares	Mean Square	F value	Pr > F
Model	1	SSR	MSR = SSR/1	F_value = MSR/MSE	P(F>F_value,1,n-2)
Error	n-2	SSE	MSE = SSE/(n-2)		
Total	n-1	SSTotal			

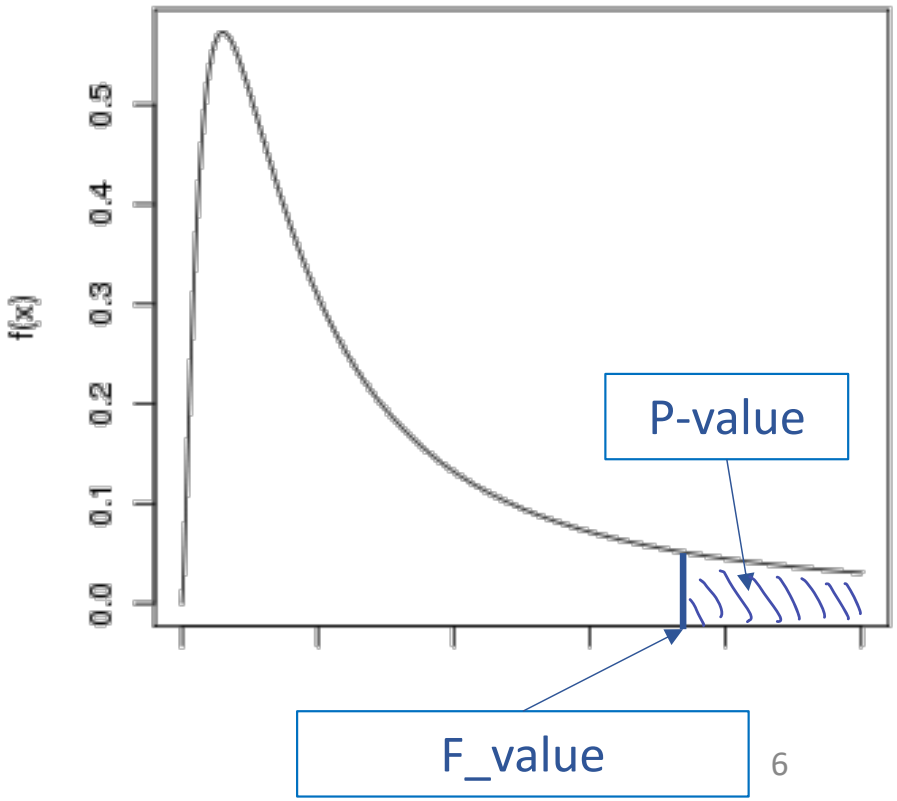
sample size

MS = SS / df

H_0: slope = 0

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4594.25136	4594.25136	31.45	<.0001
Error	27	3944.57622	146.09542		
Corrected Total	28	8538.82759			



Regression Results

- How many participants took the workshop? n=29
- What is the estimated regression model?

Posttest = 30.2 + 1.2*pretest

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4594.25136	4594.25136	31.45	<.0001
Error	27	3944.57622	146.09542		
Corrected Total	28	8538.82759			

Prop. of variation in y explained by the model

Root MSE	12.08699	R-Square	0.5380
Dependent Mean	63.37931	Adj R-Sq	0.5209
Coeff Var	19.07088		

Penalizes the model for adding more predictors: Adjust R^2 based on # of predictors and sample size (n-p-1)

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	30.19458	6.32901	4.77	<.0001
pretest	1	1.20899	0.21559	5.61	<.0001

Conclusion?

- Was the coordinator correct that participants who start out with better scores on the pre-test do better on the post-test? How do you know?
  Yes, $\beta_1 > 0$, $\beta_1 = 1.2$, positive, $p < 0.05$
- Did participants actually learn anything? How much did they learn?
  Yes. $\beta_0 = 30.2$, even when pretest = 0, final test is still higher than original. $\beta_1 = 1.2$
- what does the positive slope actually mean? Do you think the coordinator was happy with this finding?
  Yes. For every 1 point ↑ in pretest, post test ↑ by 1.2 points.
- How could the coordinator use this information?
  - Justify effectiveness of workshop
  - Target future workshops to improve lower-scoring participants.
  - Show stakeholders that ↑ engagement, ↑ baseline knowledge, ↑ gains.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4594.25136	4594.25136	31.45	<.0001
Error	27	3944.57622	146.09542		
Corrected Total	28	8538.82759			

Root MSE	12.08699	R-Square	0.5380
Dependent Mean	63.37931	Adj R-Sq	0.5209
Coeff Var	19.07088		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	30.19458	6.32901	4.77	<.0001
pretest	1	1.20899	0.21559	5.61	<.0001

Assumptions

iid: independent and identically distributed

- Observations (participants' scores) are independent
- In fitting the model:

$$\text{posttest}_i = \beta_0 + \beta_1 * \text{pretest}_i + \varepsilon_i, \quad i=1, 2, \ldots, 29$$ samples

all ε_i follow the same normal distribution with mean 0 and a constant standard deviation, i.e.,

$$\varepsilon_i \sim \text{iid } N(0,\sigma^2) \text{ for all } i$$

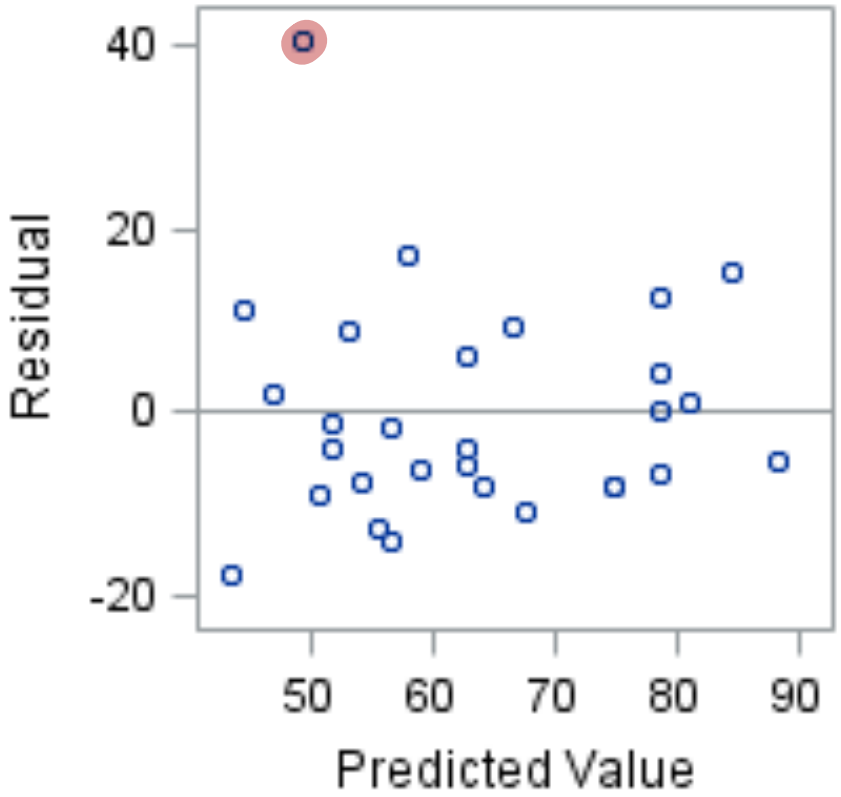
- Normality of error terms, not predictors

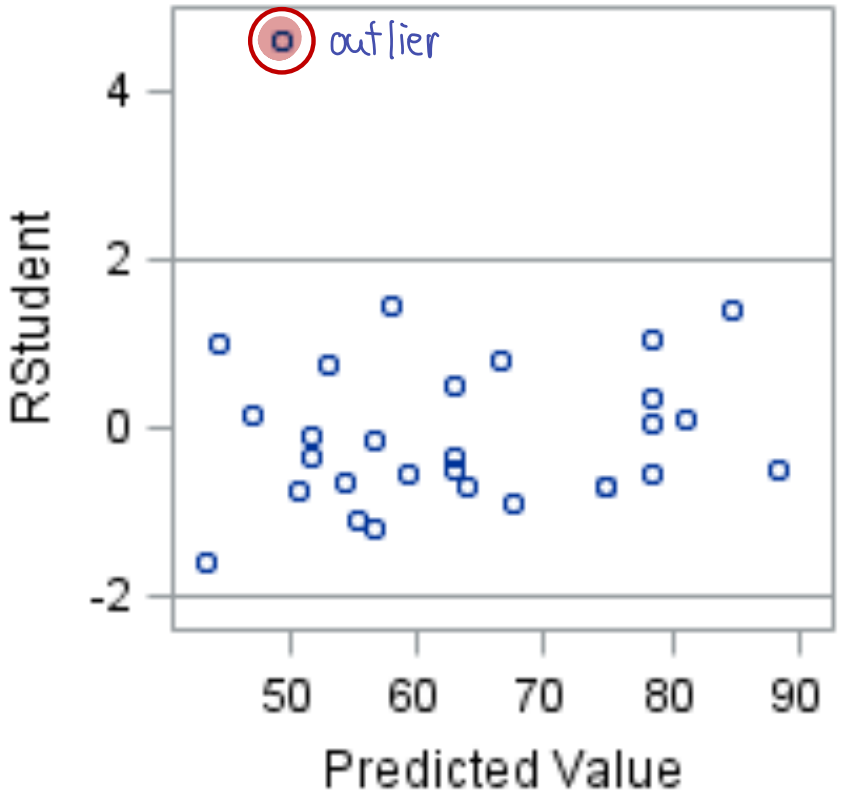
- Should the independent variable be normally distributed? No
- Should the dependent variable be normally distributed? No

only residuals need to be normal

Model Diagnostics

- Residual plots:
  - No particular pattern
  - Linear model probably ok
  - Constant variance

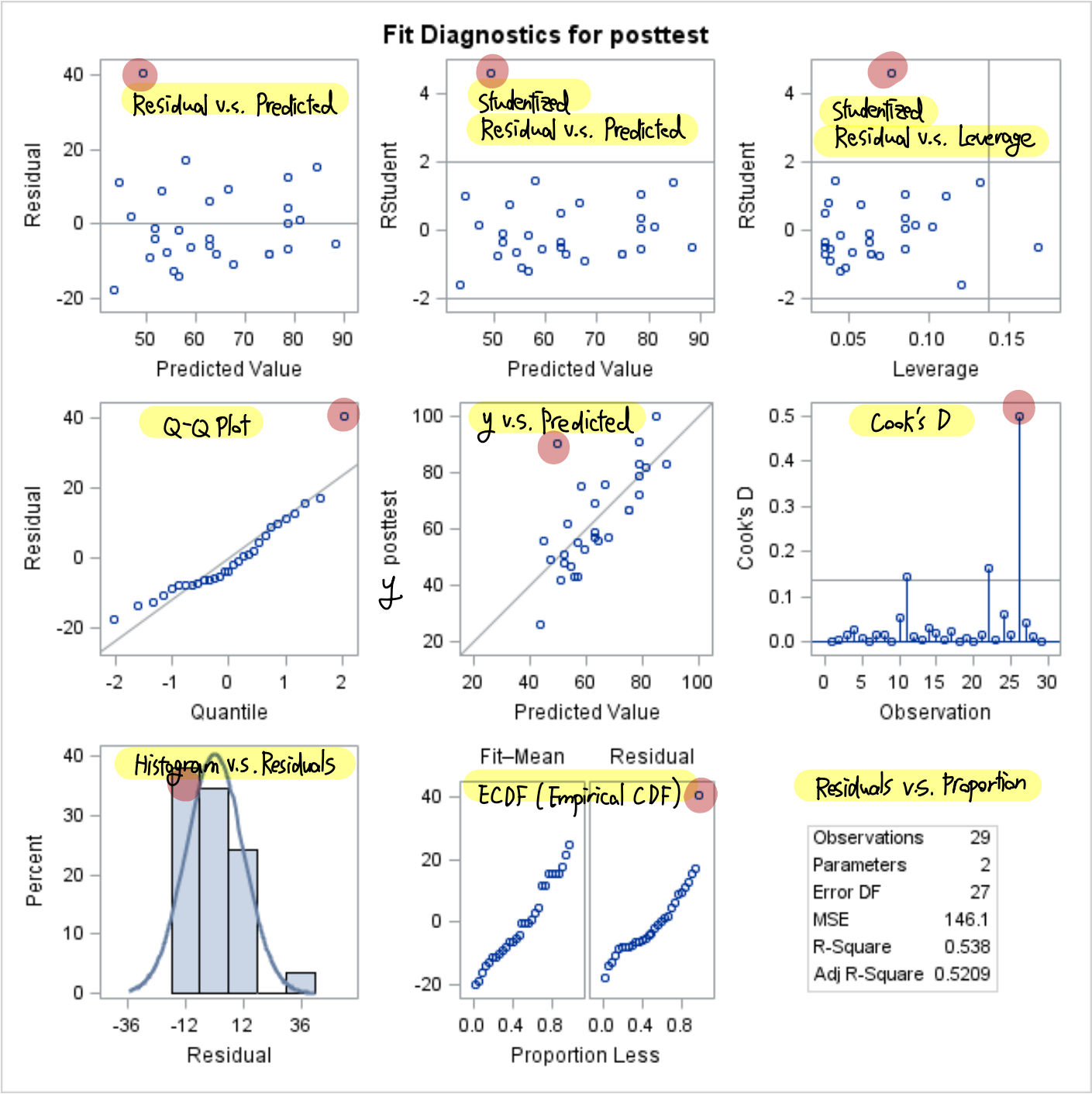




Model Diagnostics

- Potential outlier?
  - Low leverage (not outlying with regard to its X value)
  - High Cook's D (may be influencing our estimated regression line)
  - Return to your plot of the data to explore further

- Residual quantile plot: errors look normal (except perhaps for one outlier)



Hello, SAS…

- “01 regression example.sas”