Statistics for the Sciences

Simple Linear Regression Models

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- Bivariate numerical data
- Linear correlation
  - Pearson linear correlation
  - Spearman's rank correlation
- Simple Linear Regression
  - SLR model
  - LS estimation
  - Statistical inferences
  - Coefficient of Determination and ANOVA
- Lab

Bivariate numerical data

- When two variables are measured (not always but usually on a single experimental unit), the resulting data are called bivariate data (or Paired data).
- When both of the variables (X, Y) are quantitative/numerical, call the variable X - the *independent variable* or *explanatory variable*, and Y - the *dependent variable* or *response*.
  - A random sample is of the form

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n).$$

    Scatter plot can be used to check the relationship between X and Y.

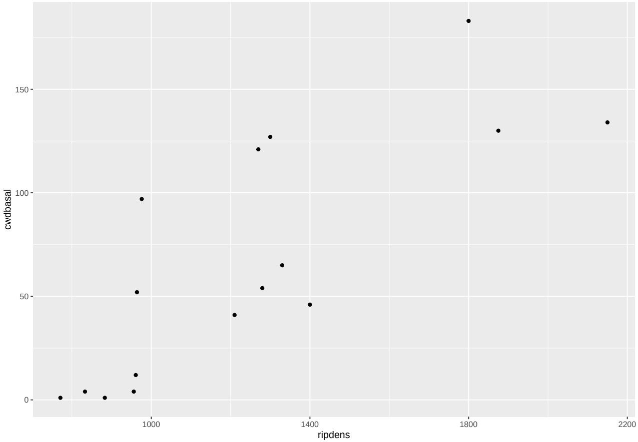
Bivariate numerical data

- An example

```
##    cwdbasal ripdens
## 1       121    1270
## 2        41    1210
## 3       183    1800
## 4       130    1875
## 5       127    1300
## 6       134    2150
## 7        65    1330
## 8        52     964
## 9        12     961
## 10       46    1400
## 11       54    1280
## 12       97     976
## 13        1     771
## 14        4     833
## 15        1     883
## 16        4     956
```

Bivariate numerical data

- We visualize bivariate data with **scatter plot**
  - Scatter plot is used to check the relationship between X and Y.

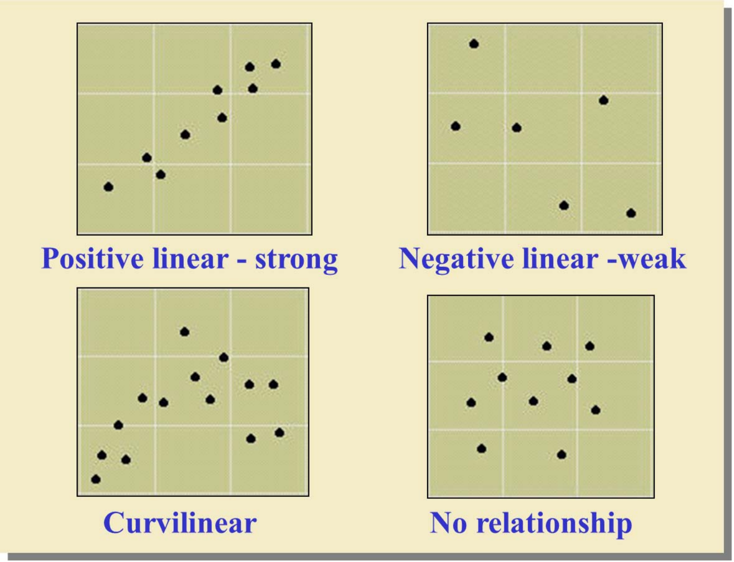


Linear Correlation

When analyzing a scatterplot, you should look for:

- **Type of Association**
  - *Straight line*?
  - Curve?
  - No pattern at all?
- **Direction of Association**
  - Positive or negative?
- **Strength of Association**
  - Strong or weak?

Linear Correlation



Linear Correlation

- Assume that the two random variables X and Y exhibit a **linear pattern** or form. How do we measure the strength and direction of the linear pattern?
- Parameter: Denote the population **correlation coefficient** by ρ_{XY} which measures the **linear relationship** between X and Y.

Linear Correlation

- Let X and Y be two numerical random variables. Recall that $\mu_X = \sum_x xP_X(x)$(discrete) or $= \int_{-\infty}^{\infty} xf_X(x)dx$(continuous).
- And $\sigma_X^2 = Var(X) = E\left[(X-\mu_X)^2\right] =$

$$\begin{cases} \sum_x (x-\mu_X)^2 P_X(x) & (discrete) \\ \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)dx & (continuous) \end{cases}$$

- **Definition.** The **covariance** of two random variables X and Y is defined by

$$\sigma_{XY}^2 = Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right].$$

- **Definition.** The **correlation coefficient** of two random variables X and Y is defined by

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

ρ_{XY} measures the **linear relationship** between X and Y.

Linear Correlation

- Assume that the two random variables X and Y exhibit a **linear pattern** or form.

Suppose n paired measurements, $(x_i,\ y_i),\ i = 1, \ldots, n$ are taken on the variables X and Y (for example, the heights and armspans of n individuals).

We can summarize the location (or center) of each variable by the **sample means**:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

We can summarize the spread of each variable by the **sample standard deviations**:

$$S_x = \sqrt{S_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \text{ and } S_y = \sqrt{S_y^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Linear Correlation

- However, none of these summary measures says anything about the relationship between the two variables.
- One measure of the relationship between X and Y is the **Pearson correlation** which is an estimate of ρ_{XY}.
- **Statistic** The Pearson correlation between X and Y computed from these data is

$$r = \frac{1}{n-1}\sum_{i=1}^{n} x_i' y_i' = \frac{1}{n-1}\frac{S_{xy}}{S_x S_y},$$

where

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}), \text{ and}$$

$$x_i' = \frac{x_i - \overline{x}}{S_x} \text{ and } y_i' = \frac{y_i - \overline{y}}{S_y}$$

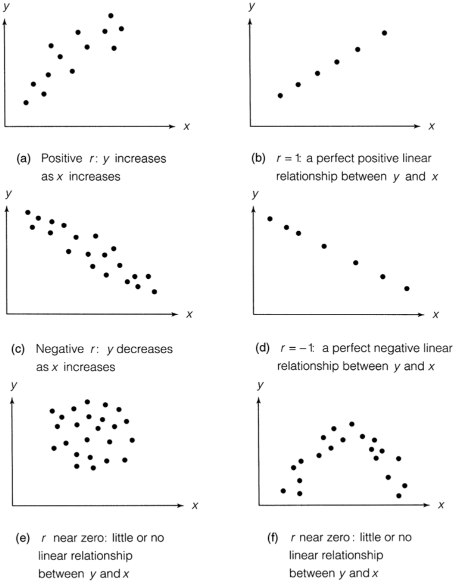
are the standardized data.

Properties of Pearson Correlation r

- Correlation between X and Y is the same as the correlation between Y and X.
- $-1 \leq r \leq 1$. Sign of r indicates direction of the linear relationship
- If $r \approx 0$, Weak linear relationship
- If $r \approx 1$ or -1, Strong relationship; either positive or negative
- If $r = 1$ or 1, All points fall exactly on a straight line
- If $S_x = 0$ and/or $S_y = 0$, we define r to be 0. This is because
  - The formula doesn't work in this case (division of 0 by 0)
  - The standard deviation equals 0 if and only if all data values are equal, so
  - There can be no association since there is no variation.

Properties of Pearson Correlation r

The following figures illustrate what Pearson correlation measures.



(a) Positive r: y increases as x increases

(b) $r = 1$: a perfect positive linear relationship between y and x

(c) Negative r: y decreases as x increases

(d) $r = -1$: a perfect negative linear relationship between y and x

(e) r near zero: little or no linear relationship between y and x

(f) r near zero: little or no linear relationship between y and x

Linear correlation - statistical inference

- The Pearson correlation coefficient, r, tells us about the strength of the linear relationship between X and Y in the sample data.
- To make inference about the population correlation coefficient ρ, we perform a hypothesis test of the significance of the correlation coefficient to decide whether the evidence in the sample data is strong enough to indicate a significant linear correlation at the population level.
- Generally, we test the null hypothesis $H_0 : \rho = 0$ against a two-sided alternative $H_a : \rho \neq 0$. That is, we check if the population correlation coefficient ρ is significantly different from 0.
- To this end, we must assume that the random vector (X, Y) has a bivariate-normal distribution.

Linear correlation - statistical inference

- Example: Test of

$$H_0 : \rho = 0$$

```
##
##  Pearson's product-moment correlation
##
## data:  ripdens and cwdbasal
## t = 4.9298, df = 14, p-value = 0.0002216
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4971417 0.9264444
## sample estimates:
##       cor
## 0.7965489
```

Spearman's rank correlation

- The Spearman's rank correlation test is a non-parametric test that uses **ranks** of sample data consisting of matched pairs. It is used to test for the **strength** of a linear or nonlinear association between two variables.
  - The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the **rank** variables.

An example

- Test of

$$H_0 : \rho = 0$$

```
##
##  Spearman's rank correlation rho
##
## data:  ripdens and cwdbasal
## S = 89.13, p-value = 1.252e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8689258
```

Linear Regression - Introduction

- Assume that visual examination of the scatter plot confirms that the points approximate a straight-line pattern

$$y = \beta_0 + \beta_1 x.$$

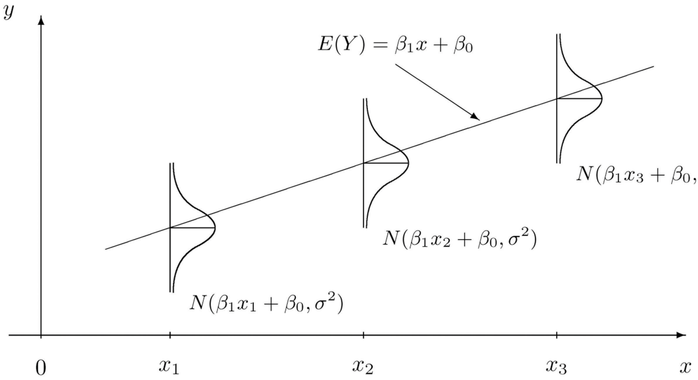
- However, the bivariate measurements that we observe do not generally fall exactly on a straight line, we choose to use a **probabilistic** model

$$Y|_{X=x} = \beta_0 + \beta_1 x + \varepsilon,$$

  where ε is a random variable possessing a specified probability distribution with mean 0. For example, assume that ε's are independent normal random variables with mean 0 and common variance σ^2.
  - $\beta_0 + \beta_1 x = E(Y|X = x)$ is the average value of Y for any given value of x.

Linear Regression - Introduction

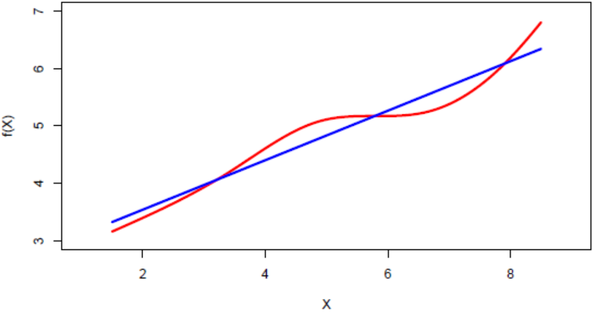
Simple Linear Regression Model Introductory Statistics (Shafer and Zhang), UC Davis Stat Wiki



- We estimate the population **parameters** β_0 and β_1 using sample information.

Linear Regression - Introduction

- Linear regression assumes that the dependence of Y on the predictors $X_1, X_2, \ldots, X_p$ is linear.
- True regression functions are never linear!



- Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

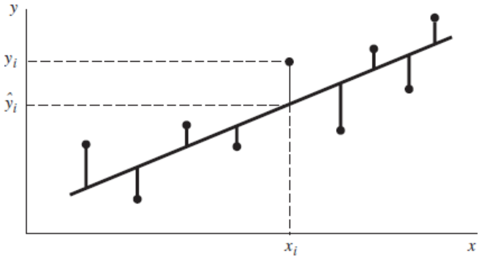
SLR Point Estimations

- We choose our estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to estimate β_0 and β_1 so that the vertical distances of the points y_i from the line, are minimized. That is, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are chosen to minimize the sum of squares of deviations

$$SSE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}[y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)]^2.$$

- The estimates are call LS (Least Square) estimators.

Fitting a straight line through a set of data points



SLR Point Estimations

Least-Squares Estimators for the Simple Linear Regression Model.

1. $\widehat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$ where $S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ and $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$.
2. $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$.

SLR Hypothesis testing

- Both $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are random variables, so they have their sampling distributions (we skip the details).
- Let $S = \sqrt{MSE}$. Then under $H_0 : \beta_i = \beta_{i0}, i = 0, 1$,

$$T = \frac{\widehat{\beta}_i - \beta_{i0}}{S\sqrt{c_{ii}}}, \quad i = 0, 1$$

  possess a Student's t distribution with $n - 2$ df, where $c_{00} = \sum x_i^2/(nS_{xx})$ and $c_{11} = 1/S_{xx}$.
- Note that the testing of β_1 is actually to test

$$H_0 : \text{There is no relationship between } X \text{ and } Y \text{ versus}$$
$$H_a : \text{There is some relationship between } X \text{ and } Y$$

SLR Hypothesis testing

- Test of Hypothesis for β_i:

$$H_0 : \beta_i = \beta_{i0}$$

$$H_a : \begin{cases} \beta_i > \beta_{i0}, & \text{(upper-tail alternative);} \\ \beta_i < \beta_{i0}, & \text{(lower-tail alternative);} \\ \beta_i \neq \beta_{i0}, & \text{(two-tailed alternative).} \end{cases}$$

$$\text{Test statistic:}\, T = \frac{\widehat{\beta}_i - \beta_{i0}}{S\sqrt{c_{ii}}}$$

$$\text{Rejection region:} \begin{cases} t > t_\alpha, & \text{(upper-tail rejection region);} \\ t < -t_\alpha, & \text{(lower-tail rejection region);} \\ |t| > t_{\alpha/2}, & \text{(two-tailed rejection region).} \end{cases}$$

where

$$c_{00} = \sum x_i^2/(nS_{xx}), c_{11} = 1/S_{xx}.$$

Notice that the t-distribution is based on (n-2) df.

SLR Interval Estimations

- A $100(1-\alpha)\%$ Confidence Interval for β_i

$$\boxed{\begin{array}{l} \hfill \widehat{\beta}_i \pm t_{\alpha/2,n-2} S \sqrt{c_{ii}} \hfill \\ \text{where} \\ \qquad c_{00} = \sum x_i^2/(nS_{xx}),\, c_{11} = 1/S_{xx}. \end{array}}$$

- One useful application of the hypothesis-testing and confidence interval techniques just presented is to the problem of estimating $E(Y)$, the mean value of Y , for a fixed value of the independent variable $x = x^*$.

$$\boxed{\begin{array}{c} \text{A } 100(1-\alpha)\% \text{ Confidence Interval for } E(Y) = \beta_0 + \beta_1 x^* : \\ \widehat{\beta}_0 + \widehat{\beta}_1 x^* \pm t_{\alpha/2,n-2} S \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}} \end{array}}$$

SLR Interval Estimations

- Predicting a particular value of Y:

Let $x = x^*$ be a fixed value of the independent variable. Instead of estimating the mean Y value at $x = x^*$, we wish to predict the particular (individual) response Y that we will observe if the experiment is run at some time in the future (such as next Monday), denoted by Y^*. Then

$$Y^* = \beta_0 + \beta_1 x^* + \varepsilon.$$

It is natural to estimate Y^* by $\widehat{Y^*} = \widehat{\beta}_0 + \widehat{\beta}_1 x^*$.

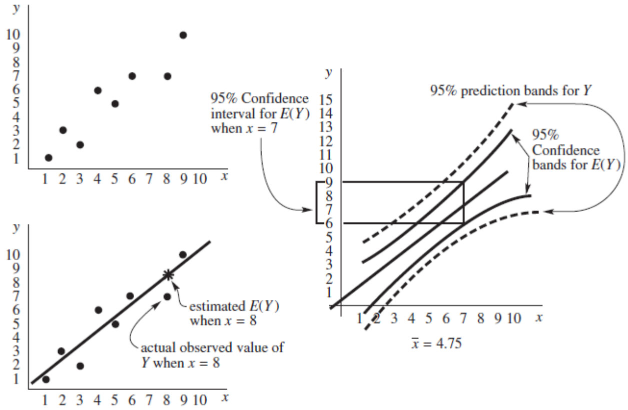
- A $100(1-\alpha)\%$ Prediction Confidence Interval for Y when $x = x^*$

$$\boxed{\widehat{\beta}_0 + \widehat{\beta}_1 x^* \pm t_{\alpha/2,n-2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}.}$$

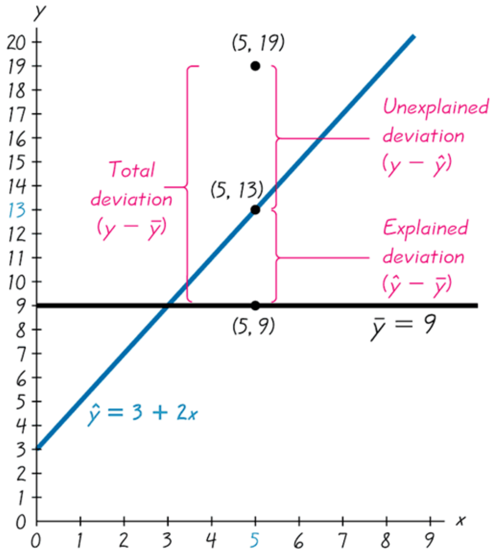
SLR Interval Estimations

Remark. Prediction intervals for the actual value of Y are longer than confidence intervals for $E(Y)$ if both confidence levels are the same and both are determined for the same value of $x = x^*$.

Some hypothetical data and associated confidence and prediction bands



Overall Accuracy of the Model: Coefficient of Determination



Example:

- There is sufficient evidence of a linear correlation.
- The equation of the line is

$$\hat{y} = 3 + 2x$$

- The sample mean of the y-values is 9.
- One of the pairs of sample data is $x = 5$ and $y = 19$.
- The point **(5,13)** is on the fitted regression line.

Overall Accuracy of the Model: Coefficient of Determination

- It can be shown that

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

  That is,

$$SS_{total} = SSR + SSE.$$

- The coefficient of determination, denoted by R^2, is the proportion of the variation in y that is explained by the regression line

$$R^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{SSR}{SS_{total}} = 1 - \frac{SSE}{SS_{total}}.$$

  That is, it is a measure of: How much of the variation in the response is "explained" by the regression (the linear relationship between X and Y).

Analysis of Variance

The procedure of Analysis of Variance for SLR models can be summarized in the following table.

Source	df	SS	MS	F
Regression	1	SSR	$MSR = SSR/1$	MSR/MSE
Error	n-2	SSE	$MSE = SSE/(n-2)$	
Total	n-1	SS_{total}		

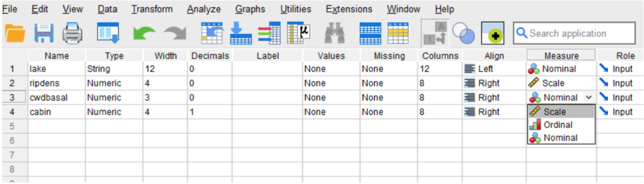
Note. The F-test for $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ is exactly equivalent to the t-test, with

$$t^2 = F.$$

And the F-test statistic has an F distribution under H_0 with $df_1 = 1, df_2 = n-2$.

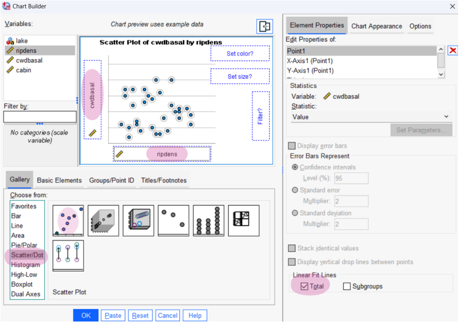
Lab

- Consider data `christ.csv`: Christensen et al. (1996) studied the relationships between coarse woody debris (CWD) and shoreline vegetation and lake development in a sample of 16 lakes in North America. The main variables of interest are the density of cabins(no.km^{-1}), density of riparian trees (trees km^{-1}), the basal area of riparian trees (m^{2}km^{-1}), density of coarse woody debris (no. km^{-1}), basal area of coarse woody debris (m^{2}km^{-1}).
- Import the data and then
- In the `Variable View`, change measure of the response variable `cwdbasal` from `Nominal` to `Scale`



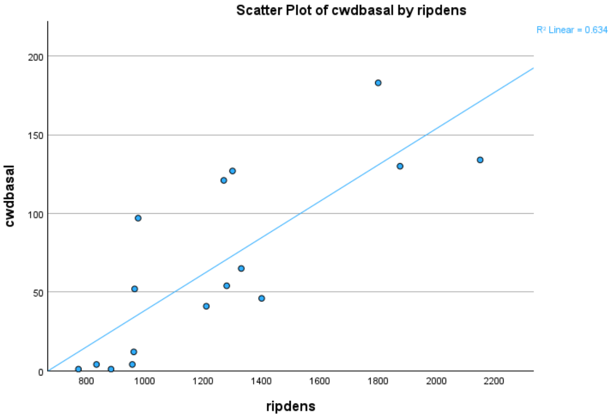
Lab

- Build a scatter plot: Click on `Graphs` in the top menu → Select `Chart Builder` then
  - In the Chart Builder dialog box, drag the `Scatter/Dot` icon from the Gallery tab into the Chart Preview area.
  - Drag the variable you want on the X-axis (independent variable) to the `X-Axis?` box.
  - Drag the variable you want on the Y-axis (dependent variable) to the `Y-Axis?` box.



Lab

- Scatter plot:

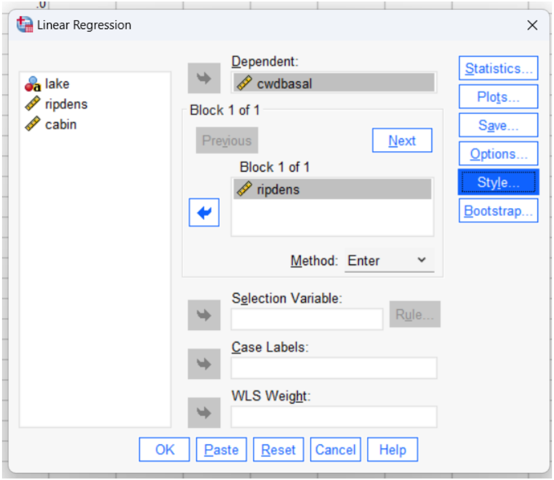


Lab

- Or you can get a scatter plot by clicking on `Graphs` $\rightarrow$ `Regression Variable Plots`
- Or click on `Graphs` $\rightarrow$ `Scatter/Dot ...` $\rightarrow$ `Simple Scatter`

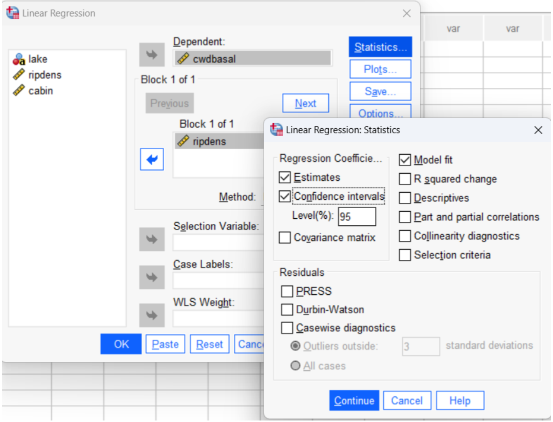
Lab

- To fit a simple linear regression model, Click on `Analyze` → `Regression` → `Linear ...`



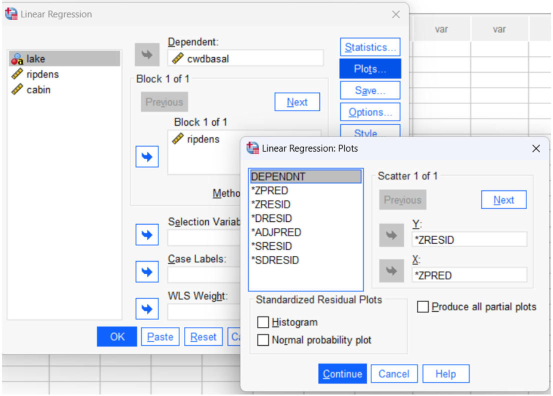
Lab

- To fit the model with confidence intervals



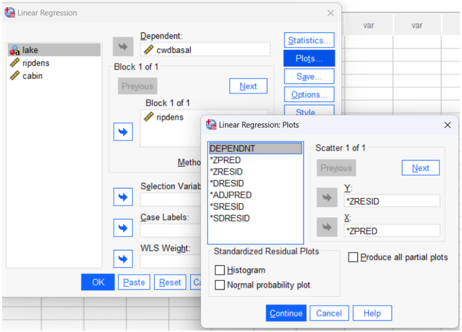
Lab

- To get residual plot of the standardized residuals against the standardized predicted values



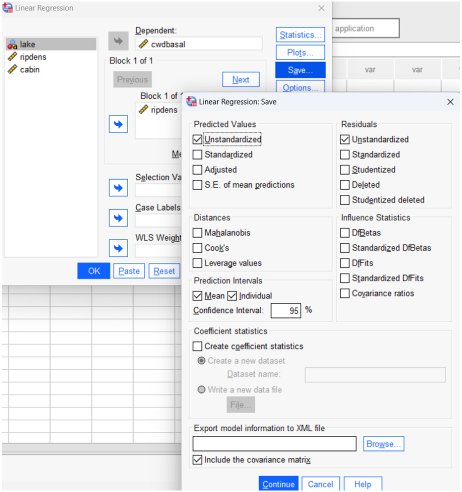
Lab

- To get residual plot of the standardized residuals against the standardized predicted values



Lab

- Save the residuals and predicted values
- Save confidence interval and prediction confidence interval for each observation



Lab

- So six more columns will be added to the data

lake	ripdens	cwdbasal	cabin	PRE_1	RES_1	LMCI_1	UMCI_1	LICI_1	UICI_1
Bay	1270	121	.0	69.60633	51.39367	50.09996	89.11270	-10.69248	149.90514
Bergner	1210	41	.0	62.67537	-21.67537	43.11130	82.23943	-17.63748	142.98821
Crampton	1800	183	.0	130.82985	52.17015	96.91224	164.74745	45.87218	215.78751
Long	1875	130	.0	139.49355	-9.49355	102.42657	176.56053	53.23023	225.75688
Roach	1300	127	.0	73.07181	53.92819	53.42007	92.72355	-7.26244	153.40606
Tenderfoot	2150	134	.6	171.26047	-37.26047	121.89673	220.62422	79.04236	263.47858
Palmer	1330	65	1.9	76.53730	-11.53730	56.62675	96.44784	-3.86066	156.93525
Street	964	52	3.6	34.25841	17.74159	10.13110	58.38573	-47.28623	115.80305
Laura	961	12	4.1	33.91187	-21.91187	9.69523	58.12851	-47.65925	115.48298
Annabelle	1400	46	4.8	84.62342	-38.62342	63.69494	105.55190	3.96735	165.27949
Joyce	1280	54	6.0	70.76149	-16.76149	51.21947	90.30352	-9.54599	151.06897
Lake_hills	976	97	6.7	35.64461	61.35539	11.86837	59.42084	-45.79685	117.08606
Towanda	771	1	11.8	11.96381	-10.96381	-18.89962	42.82724	-71.82134	95.74897
Black oak	833	4	12.3	19.12581	-15.12581	-9.38805	47.63967	-63.82262	102.07423
Johnson	883	1	17.0	24.90161	-23.90161	-1.83182	51.63504	-57.45175	107.25498
Arrowhead	956	4	24.6	33.33429	-29.33429	8.96743	57.70114	-48.28155	114.95012

Lab

- Model fit results

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.797[a]	.634	.608	36.318

a. Predictors: (Constant), ripdens

b. Dependent Variable: cwdbasal

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	32054.439	1	32054.439	24.303	<.001[b]
	Residual	18465.561	14	1318.969		
	Total	50520.000	15			

a. Dependent Variable: cwdbasal

b. Predictors: (Constant), ripdens

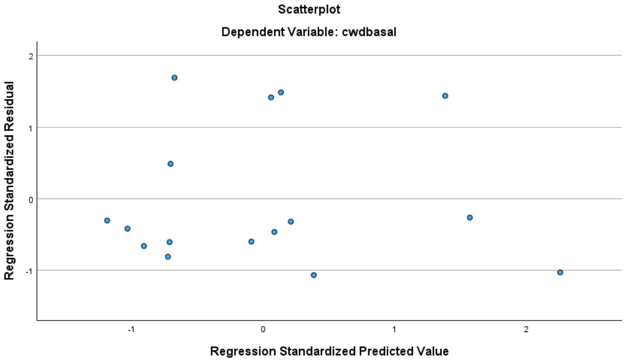
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-77.099	30.608		-2.519	.025	-142.747	-11.451
	ripdens	.116	.023	.797	4.930	<.001	.065	.166

a. Dependent Variable: cwdbasal

Lab

- Standardized residual plot to check two model assumptions: independence and constant variance



License