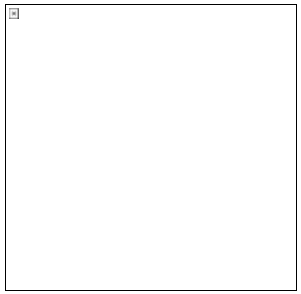
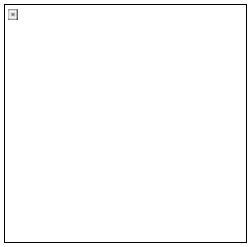
Part 2: Analysis of Relationship Between Two Variables



- Linear Regression
- Linear correlation
- Significance Tests
- Multiple regression

Linear Regression



$$\mathbf{Y} = a\,\mathbf{X} + b$$

Dependent Variable — **Independent Variable**

- To find the relationship between Y and X which yields values of Y with the least error.

Predictor and Predictand

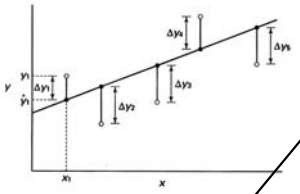
- In meteorology, we want to use a variable ***x*** to predict another variable ***y***. In this case, the independent variable ***x*** is called the "predictor". The dependent variable y is called the "predictand"

$$Y = a + b\,X$$

the dependent variable
the predictand

the independent variable
the predictor

Linear Regression



- We have ***N*** paired data point (x_i, y_i) that we want to approximate their relationship with a linear regression:

$$\hat{y} = a_o + a_1 \cdot x$$

- The errors produced by this linear approximation can be estimated as:

$$Q = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (\hat{y} - y_i)^2$$

- The least square linear fit chooses coefficients a and b to produce a minimum value of the error Q.

a_0 = intercept
a_1 = slope *(b)*

Least Square Fit

- Coefficients a and b are chosen such that the error Q is minimum:

$$\frac{\partial Q}{\partial a_o} = 0; \quad \frac{\partial Q}{\partial a_1} = 0$$

- This leads to:

$$\frac{\partial Q}{\partial a_o} = 2a_o N + 2a_1 \sum x_i - 2\sum y_i = 0$$

$$\frac{\partial Q}{\partial a_1} = 2a_o \sum x_i + 2a_1 \sum x_i^2 - 2\sum x_i y_i = 0$$

covariance between *x* and *y*

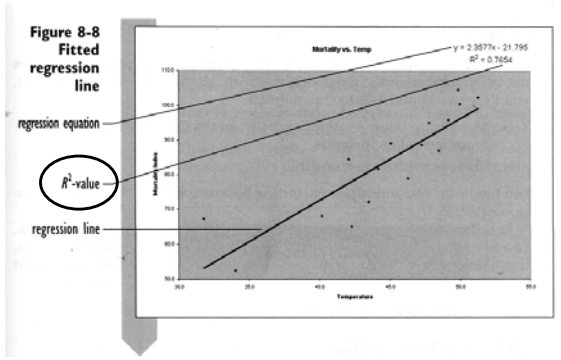
- Solve the above equations, we get the linear regression coefficients:

$$b = a_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\overline{x'y'}}{\overline{x'^2}}; \quad \text{where } (\,)' = (\,) - \overline{(\,)} \quad \text{where} \quad \overline{x'y'} \equiv \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

$$a_o = \bar{y} - a_1 \bar{x}$$

variance of *x*

Example



R^2-value

- R^2-value measures the percentage of variation in the values of the dependent variable that can be explained by the variation in the independent variable.
- R^2-value varies from 0 to 1.
- A value of 0.7654 means that 76.54% of the variance in y can be explained by the changes in X. The remaining 23.46% of the variation in y is presumed to be due to random variability.

Significance of the Regression Coefficients

- There are many ways to test the significance of the regression coefficient.
- Some use t-test to test the hypothesis that b=0.
- The most useful way for the test the significance of the regression is use the "analysis of variance" which separates the total variance of the dependent variable into two independent parts: variance accounted for by the linear regression and the error variance.

How Good Is the Fit?

$$
\begin{aligned}
S_y^2 &= \Sigma(y_i-\bar{y})^2 \\
&= \Sigma(y_i-\hat{y}+\hat{y}-\bar{y})^2 \\
&= \Sigma(\varepsilon+a+bx_i-\bar{y})^2 \\
&= \Sigma(\varepsilon+\bar{y}-b\bar{x}+bx_i-\bar{y})^2 \\
&= \Sigma(\varepsilon+b(x_i-\bar{x}))^2 \\
&= \Sigma\varepsilon^2+b^2\Sigma(x_i-\bar{x})^2+2b\times\underbrace{\Sigma(\varepsilon(x_i-\bar{x}))}_{0} \\
&= \Sigma(y_i-\hat{y})^2+b^2\Sigma(x_i-\bar{x})^2 \\
&= S_\varepsilon^2+b^2S_x^2
\end{aligned}
$$

- The quality of the linear regression can be analyzed using the "***Analysis of Variance***".
- The analysis separates the total variance of ***y*** (S_y^2) into the part that can be accounted for by the linear regression ($b^2S_x^2$) and the part that can not be accounted for by the regression (S_ε^2):

$$S_y^2 = b^2S_x^2 + S_\varepsilon^2$$

ESS210B
Prof. Jin-Yi Yu

Variance Analysis

source	d.f.	sums of squares	mean squares
Regression	1	$SSR=\sum(\hat{y}_i-\bar{y})^2$	$MSR=SSR/1$
Error	$(N-2)$	$SSE=\sum(y_i-\hat{y}_i)^2$	$MSE=SSE/(N-2)$
Total	$(N-1)$	$SST=\sum(y_i-\bar{y})^2$	

Table 5.1 Analysis of Variance for Linear Regression

- To calculate the total variance, we need to know the "mean" ➔ DOF=N-1
- If we know the mean and the regression slope (B), then the regression line is set ➔ The DOF of the regressed variance is only 1 (the slope).
- The error variance is determined from the difference between the total variance (with DOF = N-1) and the regressed variance (DOF=1) ➔ The DOF of the error variance = (N-1)-1=N-2.

ESS210B
Prof. Jin-Yi Yu

Analysis of Variance (ANOVA)

- We then use ***F***-statistics to test the ratio of the variance explained by the regression and the variance not explained by the regression:

$$F = (b^2S_x^2/1) \,/\, (S_\varepsilon^2/(N-2))$$

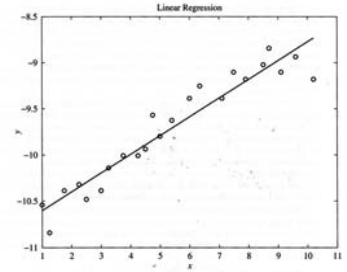
- Select a X% confidence level
- H_0: $\beta = 0$
  (i.e., variation in y is not explained by the linear regression but rather by chance or fluctuations)
  H_1: $\beta \neq 0$

regression slope in population

- Reject the null hypothesis at the α significance level if $F>F_\alpha(1, N-2)$

ESS210B
Prof. Jin-Yi Yu

Example



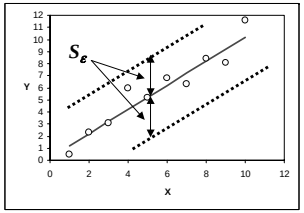

```
>>[a,yest,sst,rr] = xyplot(asmax2(:,1),asmax2(:,2),
Analysis of Variance Table.
Source      d.f.      Sums of Squares      Mean Squares
Regression   1            7.18                7.18
Residuals   21            0.66                0.03
Total       22            7.83

F = 229.069
Correlation coeff: r = 0.957091
Equation: y = -10.8073 + 0.203254 x
```

Figure 5.2 Regression of maximum size vs distance down stream, for Arroyo Seco gravels.

ESS210B
Prof. Jin-Yi Yu

Scattering



- One way to estimate the "badness of fit" is to calculate the scatter:

scatter $S_{scatter} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N-2}} = \sqrt{\frac{S_y^2 - b^2 S_x^2}{N-2}}$

- The relation between the scatter to the line of regression in the analysis of two variables is like the relation between the standard deviation to the mean in the analysis of one variable.
- If lines are drawn parallel to the line of regression at distances equal to $\pm (S_{scatter})^{0.5}$ above and below the line, measured in the ***y*** direction, about 68% of the observation should fall between the two lines.

Correlation and Regression

- Linear Regression: Y = a + bX

  A *dimensional measurement* of the linear relationship between X and Y.

➔ How does Y change with one unit of X?

- Linear Correlation

  A *non-dimensional measurement* of the linear relationship between X and Y.

➔ How does Y change (in standard deviation) with one standard deviation of X?

Linear Correlation

- The linear regression coefficient (***b***) depends on the unit of measurement.
- If we want to have a non-dimensional measurement of the association between two variables, we use the linear correlation coefficient (***r***):

$$r = \frac{\overline{x'y'}}{\sigma_x \sigma_y} = \text{the correlation coefficient; } -1 < r < 1$$

$$\overline{x'y'} \equiv \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \qquad \sigma_x = \sqrt{\frac{1}{N}\Sigma(x_i - \bar{x})^2} \text{ and } \sigma_y = \sqrt{\frac{1}{N}\Sigma(y_i - \bar{y})^2}$$

Correlation and Regression

- Recall in the linear regression, we show that:

$$S_y^2 = S_\varepsilon^2 + b^2 S_x^2$$

the fraction of the variance of *y* explained by linear regression

$$1 = \frac{S_\varepsilon^2}{S_y^2} + \boxed{\frac{b^2 S_x^2}{S_y^2}}$$

- We also know:

$$b = \frac{\overline{x'y'}}{\sigma_x^2},\ S_y = \sigma_y^2, \text{ and } S_x = \sigma_x^2$$

- It turns out that

$$\frac{b^2 S_x^2}{S_y^2} = \frac{(\overline{x'y'})^2}{\sigma_x^2 \sigma_y^2} = \boxed{r^2}$$

The square of the correlation coefficient is equal to the fraction of variance explained by a linear least-squares fit between two variables.

An Example

- Suppose that the correlation coefficient between sunspots and five-year mean global temperature is 0.5 (r = 0.5).
- The fraction of the variance of 5-year mean global temperature that is "explained" by sunspots is $\boldsymbol{r^2}$ = 0.25.
- The fraction of unexplained variance is 0.75.

Significance Test of Correlation Coefficient

- **When the true correlation coefficient is zero** (H_0: ρ=0 and H_1: $\rho\neq$0)
  Use Student-t to test the significance of ***r***

  $t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$ and $\boldsymbol{\nu = N-2}$ degree of freedom

- **When the true correlation coefficient *is not* expected to be zero**
  We can not use a symmetric normal distribution for the test.
  We must use Fisher's Z transformation to convert the distribution of r to a normal distribution:

$$Z = \frac{1}{2}\ln\left\{\frac{1+r}{1-r}\right\};\ \mu_Z = \frac{1}{2}\ln\left\{\frac{1+\rho_o}{1-\rho_o}\right\};\ \sigma_Z = \frac{1}{\sqrt{N-3}}$$

mean of Z **std of Z**

An Example

- Suppose N = 21 and ***r*** = 0.8. Find the 95% confidence limits on ***r***.

Answer:

(1) Use Fisher's Z transformation:

$$Z = \frac{1}{2}\ln\left\{\frac{1+0.8}{1-0.8}\right\} = 1.0986$$

(2) Find the 95% significance limits

$$Z - 1.96\,\sigma_Z < \mu_Z < Z + 1.96\,\sigma_Z$$
$$0.6366 < \mu_Z < 1.5606$$

(3) Convert Z back to ***r***

$$\mu_z = 0.6366 = \frac{1}{2}ln\left\{\frac{1+\rho}{1-\rho}\right\} \Rightarrow \rho = 0.56$$

$$\rho = \frac{\left(e^{2\mu_z} - 1\right)}{\left(e^{2\mu_z} + 1\right)}$$

a handy way to convert Z back to *r*

(4) The 95% significance limits are: 0.56 < $\boldsymbol{\rho}$ < 0.92

Another Example

- In a study of the correlation between the amount of rainfall and the quality of air pollution removed, 9 observations were made. The sample correlation coefficient is –0.9786. Test the null hypothesis that there is no linear correlation between the variables. Use 0.05 level of significance.

Answer:

1. Ho: ρ = 0; H1: ρ≠ 0
2. α = 0.05
3. Use Fisher's Z

$$Z\text{-}statistic = \frac{[Z-\mu_Z]}{\sigma_Z} = \frac{\left[\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)\right]}{\sqrt{\frac{N-3}{3}}} = \frac{\sqrt{9-3}}{2}\ln\left(\frac{1+(-0.9786)}{1-(-0.9786)}\right) = -5.55$$

4. $Z < Z_{0.025}$ (= -1.96) ➔ Reject the null hypothesis

Test of the Difference Between Two Non-Zero Coefficients

- We first convert r to Fisher's Z statistics:

$$Z_1 = \frac{1}{2}\ln\left\{\frac{1+r_1}{1-r_1}\right\}; \quad Z_2 = \frac{1}{2}\ln\left\{\frac{1+r_2}{1-r_2}\right\}$$

- We then assume a normal distribution for Z_1-Z_2 and use the z-statistic (not Fisher's Z):

$$z = \frac{Z_1 - Z_2 - \Delta_{z_1-z_2}}{\sigma_{z_1-z_2}}; \text{ where } \Delta_{z_1-z_2} = \mu_{z_1} - \mu_{z_2}$$

$$\text{and } \sigma_{z_1-z_2} = \sqrt{\sigma_{z_1}^2 + \sigma_{z_2}^2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}.$$

Multiple Regression

- If we want to regress y with more than one variables ($x_1, x_2, x_3, \ldots x_n$):

$$y = a_o + a_1 x_1 + a_2 x_2 + ... + a_n x_n$$

- After perform the least-square fit and remove means from all variables:

$$a_1\overline{x_1^2} + a_2\overline{x_1x_2} + a_3\overline{x_1x_3} + ... + a_n\overline{x_1x_n} = \overline{x_1y}$$

$$a_1\overline{x_1x_2} + a_2\overline{x_2^2} + a_3\overline{x_2x_3} + ... + a_n\overline{x_2x_n} = \overline{x_2y}$$

$$a_1\overline{x_1x_3} + a_2\overline{x_2x_3} + a_3\overline{x_3^2} + ... + a_n\overline{x_3x_n} = \overline{x_3y}$$

- Solve the following matrix to obtain the regression coefficients: $a_1, a_2, a_3, a_4, \ldots, a_n$:

$$\begin{bmatrix} \overline{x_1^2} & \overline{x_1x_2} & \overline{x_1x_3} & .. \\ \overline{x_2x_1} & \overline{x_2^2} & \overline{x_2x_3} & .. \\ \overline{x_3x_1} & \overline{x_3x_2} & \overline{x_3^2} & ... \\ ... & ... & ... & ... \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ ... \end{bmatrix} = \begin{bmatrix} \overline{x_1y} \\ \overline{x_2y} \\ \overline{x_3y} \\ ... \end{bmatrix}$$

Fourier Transform

- Fourier transform is an example of multiple regression. In this case, the independent (predictor) variables are:

$$x_1 = \sin\frac{2\pi z}{L}; \quad x_2 = \cos\frac{2\pi z}{L}; \quad x_3 = \sin\frac{4\pi z}{L}; \quad x_4 = \cos\frac{4\pi z}{L}; \ldots\ldots$$

- These independent variables are orthogonal to each other. That means:

$$(f_n, f_m) = \int_0^L f_n(x) f_m(x)\,dx = \begin{cases} 0 & if\ m \neq n \\ 1 & if\ m = n \end{cases}$$

Therefore, all the off-diagonal terms are zero in the following matrix:

$$\begin{bmatrix} \overline{x_1^2} & \overline{x_1x_2} & \overline{x_1x_3} & .. \\ \overline{x_2x_1} & \overline{x_2^2} & \overline{x_2x_3} & .. \\ \overline{x_3x_1} & \overline{x_3x_2} & \overline{x_3^2} & ... \\ ... & ... & ... & ... \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ ... \end{bmatrix} = \begin{bmatrix} \overline{x_1y} \\ \overline{x_2y} \\ \overline{x_3y} \\ ... \end{bmatrix}$$

This demonstrates Fourier analysis is optimal in least square sense.

- We can easily get: $a_j = \frac{2}{N}\sum_{i=1}^{N}\{y_i \cdot x_j(z_i)\}$

How Many Predictors Are Needed?

- Very often, one predictor is a function of the other predictors.
- It becomes an important question: How many predictors do we need in order to make a good regression (or prediction)?
- Does increasing the number of the predictor improve the regression (or prediction)?
- If too many predictors are used, some large coefficients may be assigned to variables that are not really highly correlated to the predictant (y). These coefficients are generated to help the regression relation to fit y.
- To answer this question, we have to figure out how fast (or slow) the "fraction of explained variance" increase with additional number of predictors.

Explained Variance for Multiple Regression

- As an example, we discuss the case of two predictors for the multiple regression.
- We can repeat the derivation we perform for the simple linear regression to find that the fraction of variance explained by the 2-predictors regression (***R***) is:

$$R^2 = \frac{r_{1y}^2 + r_{2y}^2 - 2 r_{1y} r_{2y} r_{12}}{1 - r_{12}^2}$$ here ***r*** is the correlation coefficient

- We can show that if r_{2y} is smaller than or equal to a "***minimum useful correlation***" value, it is not useful to include the second predictor in the regression.
- The ***minimum useful correlation*** = $r_{1y} * r_{12}$ We want $r_{2y} > r_{1y} * r_{12}$
- This is the minimum correlation of $\mathbf{x}_2$ with **y** that is required to improve the R^2 given that x_2 is correlated with x_1.

An Example

- For a 2-predictor case: $r_{1y} = r_{2y} = r_{12} = 0.50$
  If only include one predictor (x_1) ($r_{2y} = r_{12} = 0$) ➔ R^2=0.25
  By adding x_2 in the regression ($r_{2y} = r_{12} = 0.50$) ➔ R^2=0.33
  In this case, the 2nd predictor improve the regression.
- For a 2-predictor case: $r_{1y} = r_{12} = 0.50$ but $r_{2y} = 0.25$
  If with only x_1 ➔ R^2=025
  Adding x_2 ➔ R^2=025 (still the same!!)
  In this case, the 2nd predictor is not useful. It is because
  $r_{2y} \leq r_{1y} * r_{12} = 0.50*0.50 = 0.25$

Independent Predictors

- Based on the previous analysis, we wish to use predictors that are independent of each other
  ➔ $r_{12} = 0$
  ➔ minimum useful correlation = 0.
- The worst predictors are $r_{12} = 1.0$
- The desire for independent predictors is part of the motivation for Empirical Orthogonal Function (EOF) analysis.
- EOF attempts to find a relatively small number of independent quantities which convey as much of the original information as possible without redundancy.