**Lecture 4 Correlation and Simple Linear Regression**

**4.1 Covariance and Correlation between two random variables**

Two random variables X and Y are independent if and only if

P(X ≤ x, Y ≤ y) = P(X ≤ x)P(Y ≤ y),

where P(X ≤ x, Y ≤ y) is the joint distribution of X and Y, while P(X ≤ x) and P(Y ≤ y) are the marginal distributions of X and Y, respectively. (Note that this definition can be extended to more than two random variables.)

If two random variables X and Y are not independent, we use the ***covariance***, denoted by *Cov*, and ***correlation coefficient***, denoted by *Corr*,or , to measure their relationship.

Cov(X, Y) = E([X − ] [Y − ])

Corr(X, Y) =  = 

Note that

* Cov(X, X) = V(X) and Corr(X, X) = 1
* -1 ≤ Corr(X, Y) ≤ 1.
* if X and Y are independent, then Cov(X, Y) = Corr(X, Y) = 0.
* If Corr(X, Y) = 0, we say that X and Y are uncorrelated.
* If X and Y are both normal random variables, then Corr(X, Y) = 0 implies independence.
* Cov(cX, dY) = cdCov(X, Y) for any constants c and d. Hence the covariance is not scale invariant, while |Corr(cX, dY)| = |Corr(X, Y)|.
* Var(X ± Y) = Var(X) + Var(Y) ± 2Cov(X, Y).
* Correlation is a measure of the ***linear*** association between two random variables. Two random variables may be strongly related in a curvilinear way and have a low level of correlation.

Suppose that X and Y are two jointly distributed random variables and {(,), (,), …, (,)} is an independent sample of size *n*. The covariance of X and Y is estimated by the sample covariance ,

 = .

Note that  is the sample variance of X, denoted by , and  is the sample variance of Y, denoted by .

The correlation coefficient of X and Y is estimated by the sample correlation coefficient ,

 = .

Note that  =  = 1.

**Example 4.1** Re Figure 1.7, it seems that there is strong correlation between the two random variables, the daily maximum temperature and minimum temperature. We may test the hypothesis.

|  |
| --- |
| attach(Temperature)  cor.test(Max.Temperature,Min.Temperature)  detach(Temperature) |
| Results of Hypothesis Test  --------------------------  Null Hypothesis: correlation = 0  Alternative Hypothesis: True correlation is not equal to 0  Test Name: Pearson's product-moment correlation  Estimated Parameter(s): cor = 0.7498076  Data: Max.Temperature and Min.Temperature  Test Statistic: t = 21.59091  Test Statistic Parameter: df = 363  P-value: 0  95% Confidence Interval: LCL = 0.7011211  UCL = 0.7915352 |

We thus conclude that there is a significant linear association between the two variables.

Note: The covariance and correlation can be obtained by using the commands *cov* and *cor*, respectively. █

**4.2 Simple Linear Regression**

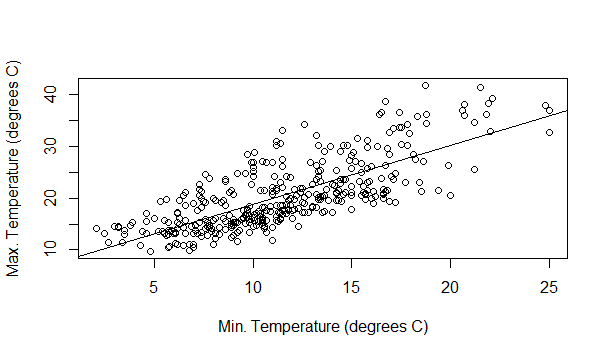
The statistical model for two variables X and Y is given by

Y =  + *X* + ε

where ε is a random error and ε ~ *N*(0, ). We call Y the ***response variable*** and X the ***predictor variable***.

**Example 4.2** Re Example 4.1, let Y be the daily maximum temperature and X the daily minimum temperature. The estimates of the coefficients,  and , and related hypothesis tests are obtained as follows.

|  |
| --- |
| attach(Temperature)  temp.fit<- lm(Max.Temperature~Min.Temperature)  summary(temp.fit)  plot(Min.Temperature, Max.Temperature)  abline(temp.fit)  detach(Temperature) |
| Call:  lm(formula = Max.Temperature ~ Min.Temperature)  Residuals:  Min 1Q Median 3Q Max  -9.564 -3.234 -1.080 3.003 13.115  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 7.4082 0.6484 11.43 <2e-16 \*\*\*  x 1.1378 0.0527 21.59 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 4.361 on 363 degrees of freedom  Multiple R-squared: 0.5622, Adjusted R-squared: 0.561  F-statistic: 466.2 on 1 and 363 DF, p-value: < 2.2e-16 |



Thus, the linear regression equation is given by

Fitted *Max.Temperature* = 7.41 + 1.14 *Min.Temperature*.

We conclude that both coefficients are highly significant (i.e., significantly different from zero), based on the two *t* tests. Note that the *F* test of ***overall significance*** is the same as the *t* test for the slope in simple linear regression (Specifics of the overall significance test will be discussed in Lecture 5). The value of , the ***coefficient of determination***, is not high, which indicates that only about 56% of the variation in the response (*Max.Temperature*) is explained by the variation in the predictor (*Min.Temperature*).

Suppose that we want to forecast the maximum temperature of a certain day, given that the minimum temperature of the day is 22. The confidence interval and prediction interval are shown below.

|  |
| --- |
| predict(temp.fit,data.frame(Min.Temperature=22), interval="confidence")  fit lwr upr  1 32.43944 31.26402 33.61486 |
| predict(temp.fit,data.frame(Min.Temperature=22), interval="predict")  fit lwr upr  1 32.43944 23.78353 41.09535 |

█

**4.3 Regression diagnostics**

Define the *i*th residual, between the *i*th observed value and *i*th fitted value of the response, as  = ** - . The residuals may be perceived as the observed values of the error ε and they should be independent and identically distributed (i.i.d.) random variables, having a normal distribution with mean 0 and variance . Several statistical and graphical procedures are available for checking the assumptions:

* Plot of residuals vs. fitted values of *Y*: If the linear model is appropriate, then the residuals should be bouncing around 0 and there should not be any kind of marked pattern to the plot, since the variance of the errors is supposed to be constant.
* Plot of square-root of absolute value of residuals vs. fitted values: This plot is similar to the first, but may be more helpful in detecting deviations from the assumption of a constant variance in the errors. If the error variance is constant, there should not be a marked pattern; if the error variance increases or decreases with the fitted values of *Y*, then you will see a pattern that looks a bit like a megaphone.
* Normal Q-Q plot of residuals: This plot is used to determine the validity of the assumption that the errors come from a normal distribution.
* Plot of Cook’s Distances: Cook’s distance is a measure of how the fitted value or slope coefficient changes if you leave one of the observations out of the fit. Observations that do not greatly affect the fitted model have small Cook’s distances, while observations that greatly affect the fitted model have large Cook’s distances. Observations with large Cook’s distances are often called ***high-leverage points*** and they are usually associated with values of the predictor variable that are “far away” from the average value of the predictor variable. They are not necessarily “bad” observations (outliers), but they do exert a large influence on the fitted model.

**Example 4.3** Re Example 4.2, the diagnostics for the regression model are show below.

|  |
| --- |
| plot(temp.fit) |
|  |

The red lines indicate the drifts away from the assumed trends. We make a few remarks as follows, based on the graphs:

* There is slight a tendency in the residual-vs-fitted-value plots to imply that the error variance is not constant;
* The normality assumption may not be quite satisfied;
* The Cook’s distance plot suggests several high-leverage points. Are they potential “outliers”? It is subject to further examination. It is important to bear in mind that “outliers” may infer that the model is not able to effectively explain the variation in the response (Recall that  = 56%, Example 4.2). Additional predictors should be sought, which lead to the multiple linear regression models.

█

**Exercises**

* 1. Re-do the Examples in this talk.
  2. Re the data *Ozone*, use *airoz*  as the response variable to establish simple linear regression models against predictors *solar*, *wind* and *temp*, respectively, and perform model diagnostics for each of the models.

**References**

* Mendenhall, W., Scheaffer, R. L. and Wackerly, D. D., (1990), Mathematical Statistics with Applications, PWS Pub Co.
* Millard, S.P. and Neerchal, N. K. (2000), *Environmental Statistics with S-PLUS*, Chapman & Hall.