**Lecture 5 Multiple Linear Regression and Prediction**

* 1. **Multiple regression models**

Suppose that a response *Y* is ‘linearly’ related to three predictors, ,  and . Then

*Y* =  +  +  +  + ε

where ε is a random error and ε ~ *N*(0, ). Note that the regression model is linear in the regression coefficients (*β*s), i.e., the deterministic element of the model is a linear combination of the coefficients. (Examples of linear and nonlinear models)

Suppose that the data are collected for the model and exhibited below.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Y*** |  |  |  |
| **…** | … | … | … |

Then the model must satisfy the following *n* equations:

**** =  +  +  +  + **

**** =  +  +  +  + **

*……*

**** =  +  +  +  + **

The estimates of the coefficients, *j* = 0, 1, 2, 3, are thus derived from the above equations by the ***ordinary least-squares method***, which requires the errors,,, …, **, to be i.i.d.

**Example 5.1** Re data *Ozone.csv*, use *airoz* as the response variable to establish a multiple linear regression model against predictors *solar*, *wind* and *temp*.

|  |
| --- |
| attach(Ozone)  airoz.fit<- lm(airoz~solar+wind+temp)  summary(airoz.fit)  detach(Ozone) |
| Call:  lm(formula = airoz ~ solar + wind + temp)  Residuals:  Min 1Q Median 3Q Max  -40.485 -14.219 -3.551 10.097 95.619  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -64.34208 23.05472 -2.791 0.00623 \*\*  solar 0.05982 0.02319 2.580 0.01124 \*  wind -3.33359 0.65441 -5.094 1.52e-06 \*\*\*  temp 1.65209 0.25353 6.516 2.42e-09 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 21.18 on 107 degrees of freedom  Multiple R-squared: 0.6059, Adjusted R-squared: 0.5948  F-statistic: 54.83 on 3 and 107 DF, p-value: < 2.2e-16 |

Therefore, the regression equation is given by

Fitted *airoz* =-64.34 + 0.060 *solar* -3.33 *wind* + 1.65 *temp*.

All predictors are significant at the 5% level of significance. At the 1% level, the predictor *solar* is not significant.

The *F* test for the overall model shows that the regression model is highly significant, i.e., it is useful in predicting the response based on the predictors. Note that, however, the significance of the test does not necessarily suggest that all of the predictors are useful in predicting the response.

Suppose that we want to make forecasts for Ozone, if *solar* = 180, *wind* = 10 and *temp* = 60. The confidence intervals are shown below.

|  |
| --- |
| predict(airoz.fit, data.frame(solar=184,wind=12,temp=78),interval='confidence')  fit lwr upr  1 35.52506 30.69638 40.35374 |
| predict(airoz.fit, data.frame(solar=184,wind=12,temp=78),interval='predict')  fit lwr upr  1 35.52506 -6.740045 77.79017 |

If *lwr* has a negative value, we set *lwr* = 0. █

* 1. **Multiple regression with categorical predictor(s)**

***Categorical data*** is the statistical data type consisting of categorical variables or of data that has been converted into that form, such as grouped data. A ***categorical variable*** is a [variable](https://en.wikipedia.org/wiki/Variable_%28research%29) that can take on one of a limited, and usually fixed, number of possible values, thus assigning each individual to a particular group or ***category***. In environmental statistics, it is typical that the data may contain categorical variables. Here is an example where a categorical variable is taken into the regression equation as a ***categorical predictor***.

**Example 5.2** Re Example 5.1.

|  |
| --- |
| attach(Ozone)  air.fit1<- lm(airoz~solar+wind+temp+factor(month))  summary(air.fit1)  detach(Ozone) |
| Call:  lm(formula = airoz ~ solar + wind + temp + factor(month))  Residuals:  Min 1Q Median 3Q Max  -40.344 -13.495 -3.165 10.399 92.689  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -74.23481 26.10184 -2.844 0.00537 \*\*  solar 0.05222 0.02367 2.206 0.02957 \*  wind -3.10872 0.66009 -4.710 7.78e-06 \*\*\*  temp 1.87511 0.34073 5.503 2.74e-07 \*\*\*  factor(month)6 -14.75895 9.12269 -1.618 0.10876  factor(month)7 -8.74861 7.82906 -1.117 0.26640  factor(month)8 -4.19654 8.14693 -0.515 0.60758  factor(month)9 -15.96728 6.65561 -2.399 0.01823 \*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 20.72 on 103 degrees of freedom  Multiple R-squared: 0.6369, Adjusted R-squared: 0.6122  F-statistic: 25.81 on 7 and 103 DF, p-value: < 2.2e-16 |

Thus the regression equation is given by

Fitted *airoz* =

Now we consider making forecasts for *airoz*. Suppose *solar* = 180, *wind* = 10, *temp* = 60 and *month* = 6, the confidence intervals are shown below.

|  |
| --- |
| predict(air.fit1, data.frame(solar=184,wind=12,temp=78,month=6),interval='confidence')  fit lwr upr  1 29.56881 15.86551 43.27211 |
| predict(air.fit1, data.frame(solar=184,wind=12,temp=78,month=6),interval='predict')  fit lwr upr  1 29.56881 -13.75395 72.89157 |

█

* 1. **Model diagnostics**

**Example 5.3** Re Example 5.2, we perform diagnostics for the multiple regression model with a categorical predictor.

|  |
| --- |
| plot(air.fit1) |
|  |

Some remarks 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. In particular, observation #77 seems to be “outstanding”;
* The normality assumption is not severely violated, apart from observation #77;
* The Cook’s distance plot suggests several high-leverage points, namely observations #30, #34 and #77. █
  1. **Variable screening: Backward elimination**

Suppose that a group of potential predictor are available. It is desirable to include the useful predictors in the regression equation and exclude those that are insignificant. The backward elimination method works according to the significance of each of the predictors. The method is illustrated in Example 5.4.

**Example 5.4** Re Example 5.2, suppose that we want to select the “best” group of predictors at the 1% of significance. Notice that two predictors, *solar* and *month*, are with *p*-values more than 1%. In the elimination procedure, only one predictor can be removed from the equation at each step and it is the one most insignificant (i.e., with the highest p-value). After removing predictor *solar*, we have a new regression model:

|  |
| --- |
| attach(Ozone)  air.fit2<- lm(airoz~wind+temp+factor(month))  summary(air.fit2)  detach(Ozone) |
| Call:  lm(formula = airoz ~ wind + temp + factor(month))  Residuals:  Min 1Q Median 3Q Max  -44.82 -13.44 -2.45 11.14 97.17  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -82.5483 26.3045 -3.138 0.00221 \*\*  wind -3.0346 0.6714 -4.520 1.64e-05 \*\*\*  temp 2.1304 0.3264 6.527 2.49e-09 \*\*\*  factor(month)6 -17.6984 9.1912 -1.926 0.05689 .  factor(month)7 -11.1815 7.8939 -1.416 0.15962  factor(month)8 -8.8693 8.0118 -1.107 0.27084  factor(month)9 -19.2489 6.6069 -2.913 0.00438 \*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 21.1 on 104 degrees of freedom  Multiple R-squared: 0.6197, Adjusted R-squared: 0.5978  F-statistic: 28.24 on 6 and 104 DF, p-value: < 2.2e-16 |

Thus the regression equation is given by

Fitted *airoz* =

The forecast intervals for *airoz*, when *wind* = 10, *temp* = 60 and *month* = 6 , are given below.

|  |
| --- |
| predict(air.fit2, data.frame(wind=12,temp=78,month=6),interval='confidence')  fit lwr upr  1 29.51051 15.55637 43.46465 |
| predict(air.fit2, data.frame(wind=12,temp=78,month=6),interval='predict')  fit lwr upr  1 29.51051 -14.60556 73.62658 |

█

**Exercises**

5.1 Repeat the Examples in this talk.

5.2 Re Example 5.2, assume that the response *airoz* follows a lognormal distribution.

1. Use the logarithm of *airoz* as the response variable to attain a linear regression model against the four predictors;
2. Find the confidence interval and prediction interval at the 95% significance level with *solar* = 180, *wind* = 10, *temp* = 60 and *month* = 6. Compare the results with those in Example 5.2;
3. Perform model diagnostics for the regression model. Is the assumption of lognormal distribution valid?
4. Select a “best” group of predictors by the backward elimination method with significance level 5%;
5. Find the confidence and prediction intervals at the 95% significance level with the values of the predictors in (ii), using the model in (iv). Compare the results with those in (ii).

5.2 Re the data frame *swiss* in {datasets},

1. Use *Fertility* as the response variable to establish a linear regression model against the five predictors;
2. Perform model diagnostics for the regression model;
3. Select a “best” group of predictors by the backward elimination method with significance level 5%;
4. Select a “best” group of predictors by the backward elimination method with significance level 2%.

**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.