

Multiple Regression

Part 1

Using RStudio for Ag Data

STAT 441/541 Statistical Methods II

Ag Data Excel File

First row contains variable names. Recall that R is case sensitive.

Note that Column 1 is not an independent variable. It only identifies each observation.



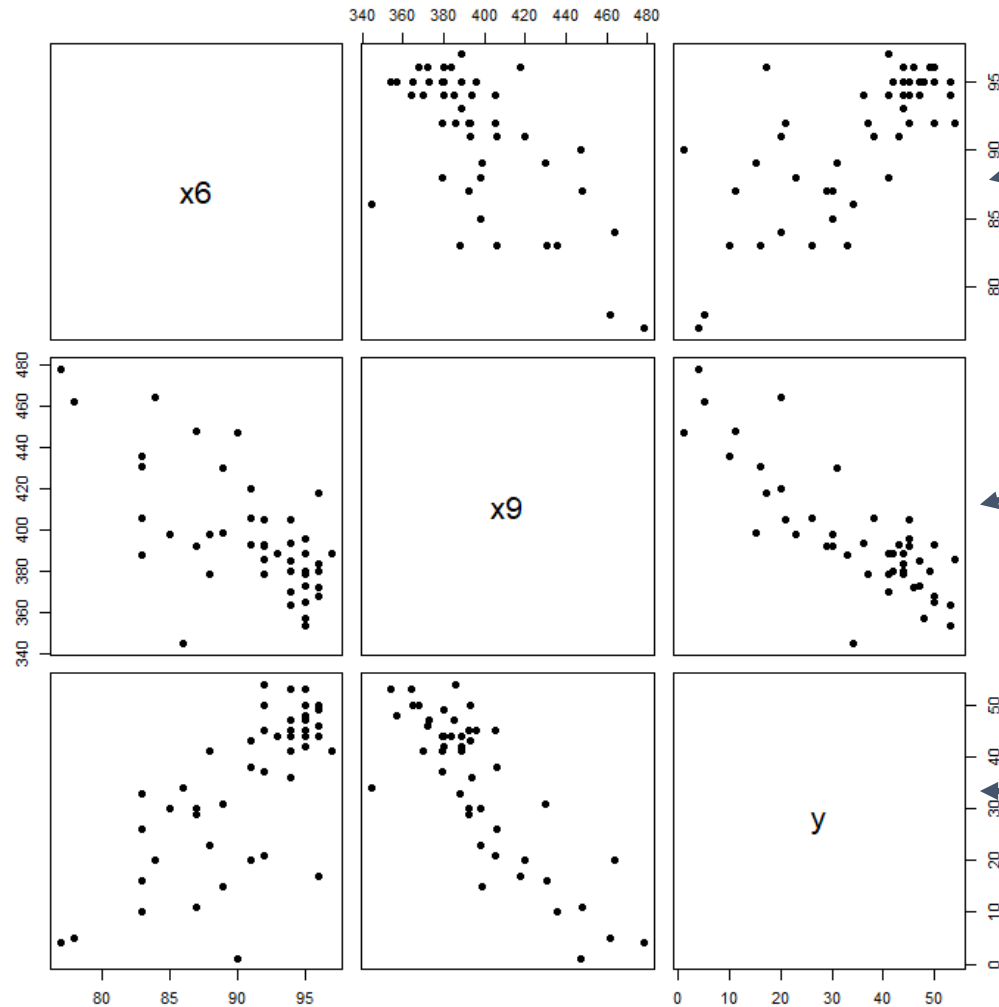
	A	B	C	D	E	F	G	H	I	J	K	L
1	Obs	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
2	1	65	84	95	40	59	85	147	151	398	273	30
3	2	65	84	94	28	61	86	149	159	345	140	34
4	3	66	84	94	41	64	83	142	152	388	318	33
5	4	67	79	94	50	65	83	147	158	406	282	26
6	5	68	81	93	46	69	88	167	180	379	311	41
7	6	66	74	96	73	67	77	131	147	478	446	4
8	7	66	73	96	72	69	78	131	159	462	294	5
9	8	67	75	95	70	68	84	134	159	464	313	20
10	9	68	84	95	63	71	89	161	195	430	455	31
11	10	72	86	93	56	76	91	169	206	406	604	38
12	11	73	88	94	55	76	91	178	208	393	610	43

This is a partial listing. There are n=46 observations in this data file.

Always look at the actual data file to verify it is a valid statistical dataset, with columns as variables and rows as observations. For example, column 7 is variable x6.

Scatterplot Matrix

```
# Scatterplot Matrix (note that columns 7,10,12 correspond to x6,x9,y)  
pairs(dataobj[,c(7,10,12)],pch=19)
```



For this row, x6 is on the vertical axis with variables x9 and y on the horizontal axes.

For this row, x9 is on the vertical axis with variables x6 and y on the horizontal axes.

For this row, y is on the vertical axis with independent variables x6 and x9 on the horizontal axes. Shows relationship between dependent and independent variables.

Correlation Coefficients

Correlation test for each pair of variables

```
cor.test(dataobj$'y',dataobj$'x6')
```

```
cor.test(dataobj$'y',dataobj$'x9')
```

Examine correlation between dependent and each independent variable.

Pearson's product-moment correlation

```
data: dataobj$y and dataobj$x6
t = 6.9447, df = 44, p-value = 1.378e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5478454 0.8376013
sample estimates:
      cor
0.7231368
```

Test Statistic

p-value

Sample Correlation Coefficient

The sample correlation coefficient is 0.7231 so there is positive correlation between y and x6. The p-value is 0.00000001378, which is less than 0.05, so this is statistically significant at the 0.05 level. The scatterplot for y and x6 shows this is a linear relationship.

Pearson's product-moment correlation

```
data: dataobj$y and dataobj$x9
t = -9.7024, df = 44, p-value = 1.682e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9001065 -0.7039041
sample estimates:
      cor
-0.8255148
```

The sample correlation coefficient is -0.8255 so there is negative correlation between y and x9. The p-value is 0.000000000001682, which is less than 0.05, so this is statistically significant at the 0.05 level. The scatterplot for y and x9 shows this is a linear relationship.

Correlation Coefficients (continued)

```
cor.test(dataobj$'x6', dataobj$'x9')
```

Examine correlation between each pair of independent variables. Since there are only two independent variables, there is only one pair which is x6 and x9.

Pearson's product-moment correlation

data: dataobj\$x6 and dataobj\$x9

t = -5.7571, df = 44, p-value = 7.668e-07

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.7945750 -0.4509865

sample estimates:

cor

-0.6554701

Test Statistic

p-value

Sample Correlation Coefficient

The sample correlation coefficient is -0.6555 so there is negative correlation between x6 and x9. The p-value=0.0000007668, which is less than 0.05, so this is statistically significant at the 0.05 level. The scatterplot for x6 and x9 shows this is a linear relationship.

Multiple Regression Model

The proposed model is: $y = \beta_0 + \beta_1 x_6 + \beta_2 x_9 + \varepsilon$

Where:

y is the daily amount of evaporation from the soil

x_6 is average air temperature

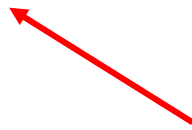
x_9 is average relative humidity

β_0 is the y intercept

β_1 is the slope parameter for average air temperature

β_2 is the slope parameter for average relative humidity

ε is random error



Model equations
always include the
random error term

Estimated Multiple Regression Model

```
# Fit the proposed model
```


```
model <- lm(y ~ x6 + x9, data=dataobj)
```

```
summary(model)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	72.45876	43.16227	1.679	0.10045	
x6	0.92221	0.29775	3.097	0.00343	**
x9	-0.30603	0.05117	-5.981	3.9e-07	***

Estimated
regression models
NEVER include a
random error term



The estimated regression model is:

$$\hat{y} = 72.45876 + 0.92221x_6 - 0.30603x_9$$

Interpretation of estimated slope parameters:

$\hat{\beta}_0 = 72.45876$ is the estimated y intercept when $x_6=0$ and $x_9=0$. Since these are outside the range of observed values of the independent variables, it does not make sense to interpret the y intercept.

$\hat{\beta}_1 = 0.92221$ is the estimated slope parameter for average air temperature. The average daily amount of evaporation from the soil will increase by 0.92221 for a one unit increase in average air temperature, holding average relative humidity constant.

$\hat{\beta}_2 = -0.30603$ is the estimated slope parameter for average relative humidity. The average daily amount of evaporation from the soil will decrease by 0.30603 for a one unit increase in average relative humidity, holding average air temperature constant.

Estimated Model Standard Deviation

```
# Fit the proposed model  
model <- lm(y ~ x6 + x9, data=dataobj)  
summary(model)
```

Residual standard error: 7.642 on 43 degrees of freedom
Multiple R-squared: 0.7396, Adjusted R-squared: 0.7275
F-statistic: 61.06 on 2 and 43 DF, p-value: 2.737e-13

The estimate of the model standard deviation is

$$s_{\varepsilon} = 7.642$$

Overall F Test

Residual standard error: 7.642 on 43 degrees of freedom

Multiple R-squared: 0.7396, Adjusted R-squared: 0.7275

F-statistic: 61.06 on 2 and 43 DF, p-value: 2.737e-13

Hypotheses

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_a: \text{At least one } \beta_j \neq 0$$

Test Statistic: $F = 61.06$ with 2 numerator df and 43 denominator df (df = degrees of freedom)

$$p\text{-value} \approx 0 \quad (2.737e-13 = 0.00000000000002737)$$

Decision about the null hypothesis: Reject the null hypothesis since p-value is less than $\alpha = 0.05$

Conclusion: At least one of the slope parameters for average air temperature and average relative humidity is not equal to zero, there is good evidence of some degree of predictive value among the two independent variables

Hypothesis Test for $\beta_1 = 0$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	72.45876	43.16227	1.679	0.10045	
x6	0.92221	0.29775	3.097	0.00343	**
x9	-0.30603	0.05117	-5.981	3.9e-07	***

Hypotheses

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

Test Statistic: $t = 3.097$

p -value = 0.003430

Decision about the null hypothesis: Reject the null hypothesis since p -value is less than $\alpha = 0.05$

Conclusion: The slope parameter for average air temperature is not equal to zero, there is good evidence of some degree of predictive value for average air temperature

Hypothesis Test for $\beta_2 = 0$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	72.45876	43.16227	1.679	0.10045	
x6	0.92221	0.29775	3.097	0.00343	**
x9	-0.30603	0.05117	-5.981	3.9e-07	***

Hypotheses

$$H_0: \beta_2 = 0$$

$$H_a: \beta_2 \neq 0$$

Test Statistic: $t = -5.981$

p -value = 0.00000039

Decision about the null hypothesis: Reject the null hypothesis since p -value is less than $\alpha = 0.05$

Conclusion: The slope parameter for average relative humidity is not equal to zero, there is good evidence of some degree of predictive value for average relative humidity