🎤

Linear Regression

Regression analysis is used to:

Predict the value of a dependent variable based on the value of at least one independent variable

Dependent variable: the variable we wish to explain
Independent variable: the variable used to explain the dependent variable

Flexural Strength and Modulus of Elasticity. Values of modulus of elasticity (MOE, the ratio of stress, i.e., force per unit area, to strain, i.e., deformation per unit length, in GPa) and flexural strength (a measure of the ability to resist failure in bending, in MPa) were determined for a sample of concrete beams of a certain type, resulting in the given data (See article “Effects of Aggregates and Microfillers on the Flexural Properties of Concrete,” Magazine of Concrete Research, 1997: 81–98): (Data: ex12.15)

a. Construct a scatterplot of the data Strength depending on MoE.

```
scatterplot(Strength~MoE,data=ex12.15)
```

b. Use the method of least squares to find the coefficients of the linear regression function for predicting strength from modulus of elasticity.

```
fit<-lm(Strength~MoE,data=ex12.15)
```

c. Summarize the p-values of the model coefficients. Give and explain the R^2-value.

```
summary(fit)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.2925 0.6008 5.480 1.08e-05 ***
MoE 0.1075 0.0128 8.399 9.56e-09 ***
Multiple R-squared: 0.7383, Adjusted R-squared: 0.7279
```

d. Predict strength for a beam with a modulus of elasticity = 40.

```
predict(fit,list(MoE=40))
7.591786
```

e. Would you feel comfortable using this least squares line to predict strength when modulus of elasticity is 100? Explain.

f. Calculate the coefficient of correlation.

```
cor(ex12.15$Strength,ex12.15$MoE)
[1] 0.8592721
```

Rainfall and Runoff Volumes. The article “Characterization of Highway Runoff in Austin, Texas, Area” (J. of Envir. Engr., 1998: 131–137) gave a scatter plot, along with the least squares line, of x = rainfall volume (m^3) and y = runoff volume (m^3) for a particular location. (Data: ex12.16)

a. Construct a scatterplot of the data y = f(x). Does the scatter plot of the data support the use of the simple linear regression model?

```
scatterplot(y~x,data=ex12.16) # needs library(car)
```

b. Calculate point estimates of the slope and intercept of the population regression line.

```
fit<-lm(y~x,data=ex12.16)
summary(fit)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.12830 2.36778 -0.477 0.642
x 0.82697 0.03652 22.642 7.9e-12 ***
Multiple R-squared: 0.9753, Adjusted R-squared: 0.9734
```

c. Calculate a point estimate of the true average runoff volume when rainfall volume is 50.

```
predict(fit,list(x=50))
40.22035
```

d. Calculate the coefficient of correlation and R^2.

```
cor.test(ex12.16$x,ex12.16$y)
t = 22.642, df = 13, p-value = 7.896e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.9619174 0.9959699
sample estimates:
cor
0.987557
```

Nitrogen Emissions. The article “An Experimental Correlation of Oxides of Nitrogen Emissions from Power Boilers Based on Field Data” (J. of Engr. for Power, July 1973: 165–170) reports data with X = burner-area liberation rate (MBtu/hr-ft^2) and Y = NOx emission rate (ppm). (Data: ex12.19)
a. Form a scatterplot of the data Y = f(X).
b. Find the least squares regression line.
c. What is the estimate of expected NOx emission rate when burner area liberation rate equals 225?
d. Calculate the coefficient of correlation.

```
plot(Y~X,data=ex12.19)
fit<-lm(Y~X,data=ex12.19)
summary(fit)
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -45.55191 25.46779 -1.789 0.0989 .
X 1.71143 0.09969 17.168 8.23e-10 ***
Multiple R-squared: 0.9609, Adjusted R-squared: 0.9576
abline(fit)
predict(fit,list(X=225))
339.5204
cor.test(ex12.19$X,ex12.19$Y)
t = 17.168, df = 12, p-value = 8.226e-10
cor
0.9802442
```

Bio indicators of Air Pollution. A number of studies have shown lichens (certain plants composed of an alga and a fungus) to be excellent bio indicators of air pollution. The article “The Epiphytic Lichen Hypogymnia Physodes as a Biomonitor of Atmospheric Nitrogen and Sulphur Deposition in Norway” (Envir. Monitoring and Assessment, 1993: 27–47) gives the following data on X = NO3 wet deposition (g N/m2
) and Y = lichen N (% dry weight) (Data ex12.20)
Use simple linear regression to analyse the data.
a. What are the least squares estimates of the intercept and the slope?
b. Predict lichen N for an NO3 deposition value of 0.5.
c. Calculate the coefficient of correlation.
d. Calculate and explain R².

plot(Y~X,data=ex12.20)

```
fit<-lm(Y~X,data=ex12.20)
summary(fit)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.36510 0.09904 3.686 0.003586 **
X 0.96683 0.18292 5.286 0.000258 ***
Multiple R-squared: 0.7175, Adjusted R-squared: 0.6918
abline(fit)
predict(fit,list(X=0.5))
1
0.8485202
cor.test(ex12.20$X,ex12.20$Y)
t = 5.2856, df = 11, p-value = 0.0002581
cor
0.8470501
```

Gas Turbines. Failures in aircraft gas turbine engines due to high cycle fatigue is a pervasive problem. The article “Effect of Crystal Orientation on Fatigue Failure of Single Crystal Nickel Base Turbine Blade Superalloys” (J. of Engineering for Gas Turbines and Power, 2002: 161–176) gave the accompanying data and fit a nonlinear regression model to predict strain amplitude from cycles to failure. (Data: ex13.18)
Fit a log-log model, and predict amplitude when cycles to failure = 5000.

plot(Strampl~Cycfail,data=ex13.18)

> ```
> > plot(Strampl~Cycfail,data=ex13.18)
> > x<-log(ex13.18$Cycfail)
> ```

```
> y<-log(ex13.18$Strampl)
> plot(y~x)
> fit =lm.fit(y~x)
Error in lm.fit(y ~ x) : 'x' must be a matrix
> fit =lm(y~x)
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.37052 -0.23765 -0.01776  0.20089  0.43123 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.73722    0.26946 -13.869 1.07e-10 ***
x           -0.12395    0.03199  -3.874  0.00122 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2731 on 17 degrees of freedom
Multiple R-squared:  0.4689,    Adjusted R-squared:  0.4376 
F-statistic: 15.01 on 1 and 17 DF,  p-value: 0.001218

> abline(fit)
> exp(predict(fit,list(x=log(5000)))
+ )
          1 
0.008288254 
```

Lifetime of Wire. Thermal endurance tests were performed to study the relationship between temperature and lifetime of polyester enameled wire (“Thermal Endurance of Polyester Enameled Wires Using Twisted Wire Specimens,” IEEE Trans. Insulation, 1965: 38–44), resulting in the following data. (Data: ex13.19)
a. Does a scatter plot of the data suggest a linear probabilistic relationship between lifetime and temperature?
b. Plot of the transformed data: log(Lifetime) as a function of Temp.
c. Estimate the parameters of the model suggested in part (b).
d. What lifetime would you predict for a temperature of 220?

plot(Lifetime~Temp,data=ex13.19)

```
x<-ex13.19$Temp
y<-log(ex13.19$Lifetime)
plot(y~x)
fit<-lm(y~x)
summary(fit)
Coefficients:
```

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.018336 0.934966 25.69 1.96e-14 ***
x -0.077951 0.004238 -18.39 3.47e-12 ***
Multiple R-squared: 0.9548, Adjusted R-squared: 0.952
abline(fit)
exp(predict(fit,list(x=220)))
1
962.1202
```