# Data Mining Lab, Exercise 6

Team# 1: Poliakov Valerii, Holovnia Dmytro, Selvaraj Sinju

Dataset: ConcreteData.csv

In the dataset the possible linear relationship is between “Cement” (**X**) and “Concrete compressive strength” (**Y**). The scatter plot shows it very clearly.

> d<-read.csv(file="ConcreteData.csv")

> make.names(names(d))

> x<-d$Cement

> y<-d$Concrete.compressive.strength

plot(x,y)

Chart, scatter chart

Description automatically generated

## Tasks 1

Use regression to estimate Y based on a single predictor X.

1. What is the estimated regression equation (ERE)?

> model<-lm(y ~ x)

> summary(model)

Call:

lm(formula = y ~ x)

Residuals:

Min 1Q Median 3Q Max

-28.295 -5.624 -0.145 5.001 48.260

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 12.066387 1.296877 9.304 <2e-16 \*\*\*

x 0.063101 0.002927 21.556 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.308 on 277 degrees of freedom

Multiple R-squared: 0.6265, Adjusted R-squared: 0.6252

F-statistic: 464.6 on 1 and 277 DF, p-value: < 2.2e-16

Based on the summary(model) output the **ERE** is: **y = 12.066387 + 0.063101 \* x**

1. A scatter plot of “Cement” vs. “Concrete compressive strength” and the line of ERE.

Chart, scatter chart

Description automatically generated

1. What would be a typical prediction error (residual standard error) obtained using the created model to predict Y.

The RSE value is 9.308. The observed values deviate from the predicted values by ~9.308.

1. Does the linear relationship exist between X and Y?

Assuming the regression formula is y = B0 + B1 \* x + E. To answer is linear regression exists, we should test the following hypothesis:

* H0: B1 = 0. No relationship between x and y.
* Ha: B1 != 0. Linear relationship between x and y.

We can test “null hypothesis” because we have p-value less than < 2.2e-16, which is close to 0. If p-value is such small (< 0.05), the hypothesis H0 is rejected.

1. How closely does the model fit the data?

We should use the coefficient of determination. The R2 (Coefficient of determination) value of our model is 0.6265. Not very close 1, but still good enough.

1. For new values of X find the estimates of response Y. Find the 95% confidence interval for the true mean Y and find the 95% prediction interval for a randomly chosen value of Y. Perform the calculations for all new values of Y. What can you observe?

Let’s take new Cement values as 198.6, 412.8 and 616.4.

Determining confidence intervals:

> new <- data.frame((x = c(198.6,412.8, 616.4)))

> pred.conf <- predict(model, new, interval="confidence", level=0.95)

> pred.conf

fit lwr upr

1 24.59823 23.00109 26.19537

2 38.11445 37.01504 39.21386

3 50.96180 49.30114 52.62246

Column “fit” is the predicted values. Column “lwr” is the lower bound of the confidence interval. Column “upr” is the upper bound of the confidence interval.

Prediction intervals:

> pred.pred <- predict(model, new, interval="prediction", level=0.95)

> pred.pred

fit lwr upr

1 24.59823 6.206153 42.99031

2 38.11445 19.758895 56.47000

3 50.96180 32.564095 69.35950

Prediction intervals are wider than confidence intervals.

Let’s evaluate the model on the real data.

> p <- predict(model, new) # estimated values of Y for new values of X

> q <- c(24.89, 47.13, 59) # true values of Y for new values of X

>

> sse <- sum((q - p)^2)

> sst <- sum((y-mean(y))^2)

> pseudo\_r2 <- 1 - sse/sst

> pseudo\_r2

[1] 0.997728

The closer to 1 the better, so we have very good result.

## Task 2

Use multiple regression to estimate Y based on several predictors X.

> x1<-d$Cement

> x2<-d$Blast.Furnace.Slag

> mult\_model <- lm(y ~ x1+x2)

> summary(mult\_model)

Call:

lm(formula = y ~ x1 + x2)

Residuals:

Min 1Q Median 3Q Max

-22.4959 -4.8970 -0.3651 4.9933 31.1790

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.130370 1.764375 -1.207 0.228

x1 0.047632 0.002908 16.379 <2e-16 \*\*\*

x2 0.173764 0.016847 10.314 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.922 on 276 degrees of freedom

Multiple R-squared: 0.7304, Adjusted R-squared: 0.7285

F-statistic: 373.9 on 2 and 276 DF, p-value: < 2.2e-16

1. What is the estimated regression equation?

ERE is: **Y = -2. 130370 + 0.047632 \* x1 + 0.173764 \* x2**

1. Compare R2 values from the multiple regression and the regression done in Task 1.

The R2 for the multiple linear regression is 0.7304, which is better than 0.6265 coefficient of determination from Task 1.

The Residual standard error 7.922 is also better than RSE value 9.308 from Task 1.