**Inference tests on open economical data**

**Variant 1**

1. Use the data about the main economic indicators with foreign investments in Kyrgyzstan.

(A) Is there significant correlation between the inflow of foreign direct investments and the

export?

*> #testing correlation between inflow and export*

*> cor.test(data[5,], data[8,])*

*Pearson's product-moment correlation*

*data: data[5, ] and data[8, ]*

*t = 5.1494, df = 12, p-value = 0.0002412*

*alternative hypothesis: true correlation is not equal to 0*

*95 percent confidence interval:*

*0.5344152 0.9445021*

*sample estimates:*

*cor*

*0.8297261*

ANSWER: YES, p value is small

(\*) between other indicators and export? What is the best export predictor?

*> #other indicators with export*

*> res <- cor.test(data[2,], data[8,])*

*> res$p.value*

*[1] 0.004238976*

*> res$estimate*

*cor*

*0.7125445*

*> res <- cor.test(data[3,], data[8,])*

*> res$p.value*

*[1] 4.128827e-05*

*> res$estimate*

*cor*

*0.8752945*

*> res <- cor.test(data[4,], data[8,])*

*> res$p.value*

*[1] 0.0009136365*

*> res$estimate*

*cor*

*-0.7835888*

*> res <- cor.test(data[5,], data[8,])*

*> res$p.value*

*[1] 0.0002412169*

*> res$estimate*

*cor*

*0.8297261*

*> res <- cor.test(data[6,], data[8,])*

*> res$p.value*

*[1] 0.01757574*

*> res$estimate*

*cor*

*0.6218697*

*> res <- cor.test(data[7,], data[8,])*

*> res$p.value*

*[1] 0.02347006*

*> res$estimate*

*cor*

*0.5994453*

*> res <- cor.test(data[9,], data[8,])*

*> res$p.value*

*[1] 1.435804e-05*

*> res$estimate*

*cor*

*0.89624*

The best predictor is IMPORT

(B) (\*) Check the assumptions behind the simple linear regression model.

> #anova difference in means

> df <- data.frame(

+ indicators = c(data[2,],data[3,],data[4,],data[5,],data[6,],data[7,],data[8,],data[9,])

+ ,levels = gl(8, 14, 8\*14)

+ )

> summary(aov(indicators~levels, data = df))

Df Sum Sq Mean Sq F value Pr(>F)

levels 7 3.956e+10 5.651e+09 55.02 <2e-16 \*\*\*

Residuals 104 1.068e+10 1.027e+08

---

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

There is a difference in means: p-value < 2e-16

> #testing the normality of residuals

> inflow\_predicted <- coefficients(least\_sq\_line)[1] + coefficients(least\_sq\_line)[2]\*inflow

> residuals = inflow - inflow\_predicted

> shapiro.test(residuals)

Shapiro-Wilk normality test

data: residuals

W = 0.84729, p-value = 0.02041

We can say reject normality at 3% significance level

(C) Find the least squares line. Graph this line and scatter diagram.

> #least squares line

> export = data[8,]

> inflow = data[6,]

> least\_sq\_line <- lm (export~inflow) # inflow and import

> least\_sq\_line

Call:

lm(formula = export ~ inflow)

Coefficients:

(Intercept) inflow

2.481e+02 6.066e-03

>

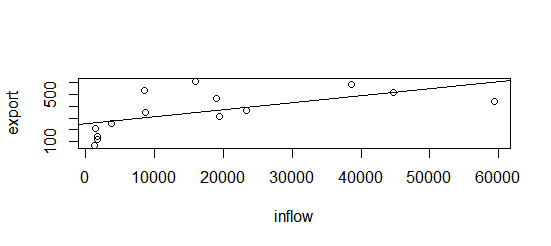
> #graph

> plot(export~inflow,

+ xlab = "inflow",

+ ylab = "export")

> abline(least\_sq\_line)



(D) Find the 96% confidence interval for the average export volume if inflow of foreign direct

investment is 800 million of dollars.

> # Confidence interval

> predict(least\_sq\_line,data.frame(inflow=800),interval="confidence",level=0.96)

fit lwr upr

1 252.9589 **129.034 376.8838**

(\*) Use the predicted production volume for 2015 based

on the analysis of corresponding time series.

*> # Time series (\*)*

*> years <- 1:length(export)*

*> least\_sq\_line2 <- lm(export~years)*

*> inflow\_pr<-coefficients(least\_sq\_line2)[1]+coefficients(least\_sq\_line2)[2]\*(length(export)+1)*

*> inflow\_pr*

*(Intercept)*

*611.5725*

*> # Confidence interval*

*> predict(least\_sq\_line,data.frame(inflow=inflow\_pr),interval="confidence",level=0.96)*

*fit lwr upr*

(Intercept) 251.8159 **127.2285 376.4033**

2. Using the data about the number of recorded crimes in Kyrgyzstan, is there sufficient evidence to conclude that the numbers of recorded crimes are different in different regions of Kyrgyzstan?

*> summary(aov(regions~levels, data=df))*

*Df Sum Sq Mean Sq F value Pr(>F)*

*levels 8 722877664 90359708 293.8 <2e-16 \*\*\**

*Residuals 72 22141136 307516*

*---*

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

ANSWER: YES, p-value is small

(\*)If yes, conduct the multiple comparison tests and give the conclusions.

*> #multiple comparison test*

*> t.test(crimes[1,], crimes[2,]) #batken and jalal-abad*

*Welch Two Sample t-test*

*data: crimes[1, ] and crimes[2, ]*

*t = -14.879, df = 9.177, p-value = 9.81e-08*

*alternative hypothesis: true difference in means is not equal to 0*

*95 percent confidence interval:*

*-2521.339 -1857.550*

*sample estimates:*

*mean of x mean of y*

*813.2222 3002.6667*

*> t.test(crimes[1,], crimes[3,]) #batken and yssyj-kul*

*Welch Two Sample t-test*

*data: crimes[1, ] and crimes[3, ]*

*t = -13.243, df = 9.2432, p-value = 2.548e-07*

*alternative hypothesis: true difference in means is not equal to 0*

*95 percent confidence interval:*

*-2222.343 -1576.102*

*sample estimates:*

*mean of x mean of y*

*813.2222 2712.4444*

*> t.test(crimes[2,], crimes[4,]) #jalal-abad and naryn*

*Welch Two Sample t-test*

*data: crimes[2, ] and crimes[4, ]*

*t = 15.822, df = 8.8411, p-value = 8.654e-08*

*alternative hypothesis: true difference in means is not equal to 0*

*95 percent confidence interval:*

*1974.533 2635.467*

*sample estimates:*

*mean of x mean of y*

*3002.6667 697.6667*

All above regions have different means as can be shown from the result and p-values. And so on for all combination of regions

3. Construct the linear regression estimate for revenue of local budget by choosing the best set of

predictors among the following: average monthly salary; commissioning of houses; consumer

price index; direct foreign investments; employed population; gross output of agriculture; import;

industrial production volume; microcrediting; number of recorded crimes; number of educational

institutions; retail trade turnover; unemployment rate.

(A) Test the significance of the whole model and each individual predictor.

*> budget.lm<-lm(local\_budget~microcrediting+employed\_population+industrial\_volume+output\_agric+fdi+import)*

*> summary(budget.lm)*

*Call:*

*lm(formula = local\_budget ~ microcrediting + employed\_population +*

*industrial\_volume + output\_agric + fdi + import)*

*Residuals:*

*1 2 3 4 5 6 7 8 9*

*2056763 -2067292 -1071570 558288 -1903128 1809310 3264291 -561610 -2085051*

*Coefficients:*

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

*(Intercept) -1.009e+08 2.546e+08 -0.396 0.730*

*microcrediting 2.617e+00 6.537e+00 0.400 0.728*

*employed\_population 6.216e+04 1.135e+05 0.548 0.639*

*industrial\_volume -2.416e-01 7.878e-01 -0.307 0.788*

*output\_agric -4.904e+02 7.035e+02 -0.697 0.558*

*fdi 9.203e+00 2.880e+01 0.320 0.780*

*import 2.753e+00 6.749e+00 0.408 0.723*

*Residual standard error: 4011000 on 2 degrees of freedom*

*Multiple R-squared: 0.8793, Adjusted R-squared: 0.5171*

*F-statistic: 2.428 on 6 and 2 DF,* ***p-value: 0.3202***

(B) Test the presence of multicollinearity.

loc.budg microcred. emp.pop ind.v out.aggr fdi import

loc.budg 1.00 0.83 0.88 0.86 0.81 0.87 0.81

microcred. 0.83 1.00 0.93 0.97 0.99 0.78 0.92

emp.pop 0.88 0.93 1.00 0.94 0.92 0.88 0.84

ind.v 0.86 0.97 0.94 1.00 0.93 0.85 0.85

out.aggr 0.81 0.99 0.92 0.93 1.00 0.77 0.95

fdi 0.87 0.78 0.88 0.85 0.77 1.00 0.76

import 0.81 0.92 0.84 0.85 0.95 0.76 1.00

The component with the highest correlation with the Local Budget is FDI. However, it has multicollinearity with almost all other predictors

(C) Construct the 90% prediction interval for the revenue of local budget using the data of 2013.

Is the actual revenue of local budget in this interval?

> #Prediction interval (10,946,586, 32,650,452.0), real budget=21,236,908.6

> newdata=data.frame(microcrediting=27533900.5, employed\_population=2263.0, industrial\_volume=169520000, output\_agric=171412.4, fdi=2006849.7, import=5987000)

> predict(budget.lm, newdata, interval="predict", level=0.9)

fit lwr upr

1 21798519 **5831562 37765476**

Actual value: 21236908.6

ANSWER: YES, Actual value is in the prediction interval

(D) (\*) Prove the significance of your model using the values of coefficient of determination for

complete and reduced models.

*> #reduced model*

*> budget.lm\_reduced <- lm(local\_budget~fdi)*

*> #testing significance*

*> summary(budget.lm)*

*Call:*

*lm(formula = local\_budget ~ microcrediting + employed\_population +*

*industrial\_volume + output\_agric + fdi + import)*

*Residuals:*

*1 2 3 4 5 6 7 8 9*

*2056763 -2067292 -1071570 558288 -1903128 1809310 3264291 -561610 -2085051*

*Coefficients:*

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

*(Intercept) -1.009e+08 2.546e+08 -0.396 0.730*

*microcrediting 2.617e+00 6.537e+00 0.400 0.728*

*employed\_population 6.216e+04 1.135e+05 0.548 0.639*

*industrial\_volume -2.416e-01 7.878e-01 -0.307 0.788*

*output\_agric -4.904e+02 7.035e+02 -0.697 0.558*

*fdi 9.203e+00 2.880e+01 0.320 0.780*

*import 2.753e+00 6.749e+00 0.408 0.723*

*Residual standard error: 4011000 on 2 degrees of freedom*

*Multiple R-squared: 0.8793, Adjusted R-squared: 0.5171*

*F-statistic: 2.428 on 6 and 2 DF, p-value: 0.3202*

*> summary(budget.lm\_reduced)*

*Call:*

*lm(formula = local\_budget ~ fdi)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-4582799 -1399645 -668789 1748470 5645382*

*Coefficients:*

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

*(Intercept) -5.903e+06 4.813e+06 -1.227 0.25965*

*fdi 1.243e+01 2.720e+00 4.570 0.00258 \*\**

*---*

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

*Residual standard error: 3092000 on 7 degrees of freedom*

*Multiple R-squared: 0.7489, Adjusted R-squared: 0.7131*

*F-statistic: 20.88 on 1 and 7 DF, p-value: 0.002575*

The complete model has higher coefficient of determination, however there are more predictors used in this model.

As looking for the significance of the each predictor, in the complete model independent variables have high p-value for the t-test, which means that they are slightly unreliable.

4. Use the monthly data about the production of main industrial products in Kyrgyzstan in 1994 –

2015.

(A) Decompose the time series for production of bread into seasonal, trend, and irregular

components. Graph this decomposition.

> #Decomposition

>

> products\_ts<-ts(products, start=1994, frequency=12)

> decomposition<-stl(products\_ts, s.window = "periodic")

> decomposition

Call:

stl(x = products\_ts, s.window = "periodic")

Components

seasonal trend remainder

Jan 1994 -422.67431 17125.2489 2195.425434

Feb 1994 -601.08447 16350.8645 952.719925

Mar 1994 -423.73997 15576.4802 -1660.040234

Apr 1994 -379.27785 14817.9394 -490.961563

May 1994 -484.09277 14059.3986 -1295.305850

Jun 1994 -375.02256 13308.7683 61.654230

…………….

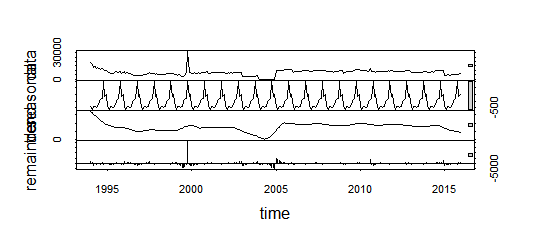
Aug 2015 102.12014 5301.8897 98.390178

Sep 2015 236.03291 5120.8880 312.279067

Oct 2015 1469.52174 4962.6595 -749.181284

Nov 2015 387.77819 4804.4311 661.090735

Dec 2015 525.63688 4662.3762 853.886961



(B) (\*) Find the cyclical component of the above time series.

> #Cyclical components

>

> predicted\_products<-coefficients(least\_sq\_line)[1]+coefficients(least\_sq\_line)[2]\*year

> predicted\_products

[1] 7144.477 7148.173 7151.868 7155.564 7159.259 7162.955 7166.651 7170.346 7174.042

[10] 7177.737 7181.433 7185.129 7188.824 7192.520 7196.215 7199.911 7203.607 7207.302

[19] 7210.998 7214.693 7218.389 7222.085

>

> cyclic<-products/predicted\_products

> cyclic

[1] 2.6451201075 2.3366111525 1.8865979478 1.9492104705 1.7152612057 1.8142512302

…………………..

[253] 0.6355416809 0.6648593604 0.6984869670 0.6975163305 0.6674758544 0.6975919605

[259] 0.7273162359 0.7634479439 0.7861880161 0.7876980692 0.8108873143 0.8365867217

(C) Based on the time series decomposition, give the prediction about the production of bread in

the first quarter of 2016. Compare with the actual data.

> #prediction of the first three months of the 2016

> m <- length(months)

> quarter\_2016 <- c(m+1,m+2,m+3)

> res\_predicted <- (coefficients(least\_sq\_line)[1]+coefficients(least\_sq\_line)[2]\*quarter\_2016[1])\*decomposition$time.series[1]+

+ (coefficients(least\_sq\_line)[1]+coefficients(least\_sq\_line)[2]\*quarter\_2016[2])\*decomposition$time.series[2]+

+ (coefficients(least\_sq\_line)[1]+coefficients(least\_sq\_line)[2]\*quarter\_2016[3])\*decomposition$time.series[3]

> res\_predicted\*(-1)

(Intercept)

11759207

Actual value = 14746.9

Predicted value = 11759.207