# Lab01: Data Transformations, Bivariate Regression Analysis, Numerical Integration & Distributions

## Task 1. Univariate Variable Exploration and Transformations [2 points]

Use the **CPS1985** dataset[[1]](#footnote-1) in the library **AER** to explore the distribution of the respondents’ **wage**.

**Task 1.1:** Find the best -value for the Box-Cox transformation (see **summary(car::powerTransform(*varName*))**). Could the *log*-transformation (i.e., ) instead of be used? Justify your answer. [0.5 points]

summary(powerTransform(lm(wage~1, data=CPS1985)))

bcPower Transformation to Normality

Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd

Y1 -0.0658 0 -0.1997 0.068

Likelihood ratio test that transformation parameter is equal to 0

(log transformation)

LRT df pval

LR test, lambda = (0) 0.9245772 1 0.33628

Likelihood ratio test that no transformation is needed

LRT df pval

LR test, lambda = (1) 232.5055 1 < 2.22e-16

, could use log-transformation here due to the p-value are not significant (fail to reject )

**Task 1.2:** For the untransformed (, optimal () and over-adjusted ( Box-Cox transformed **wage** variable repeat the following tasks and ***comparatively interpret*** the results:

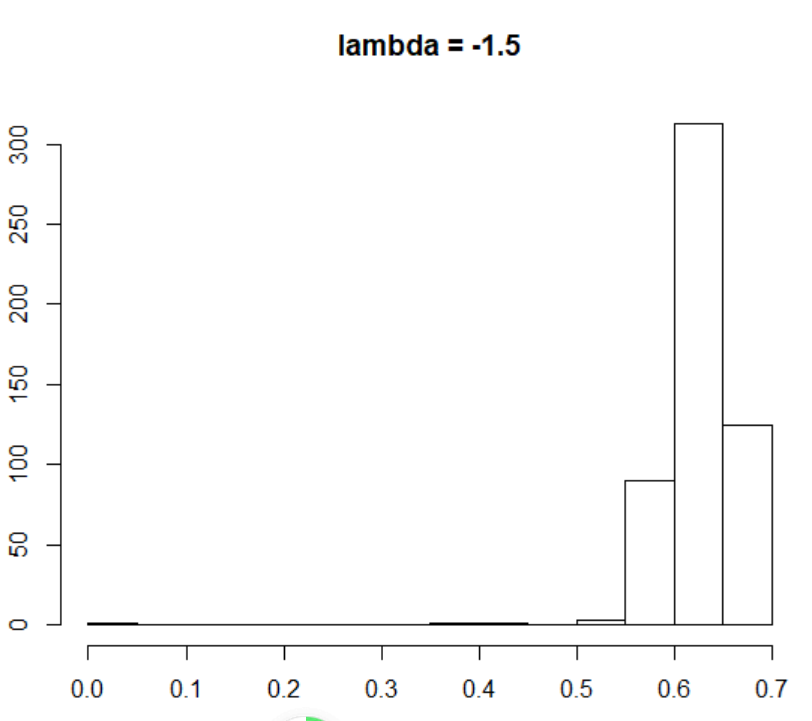
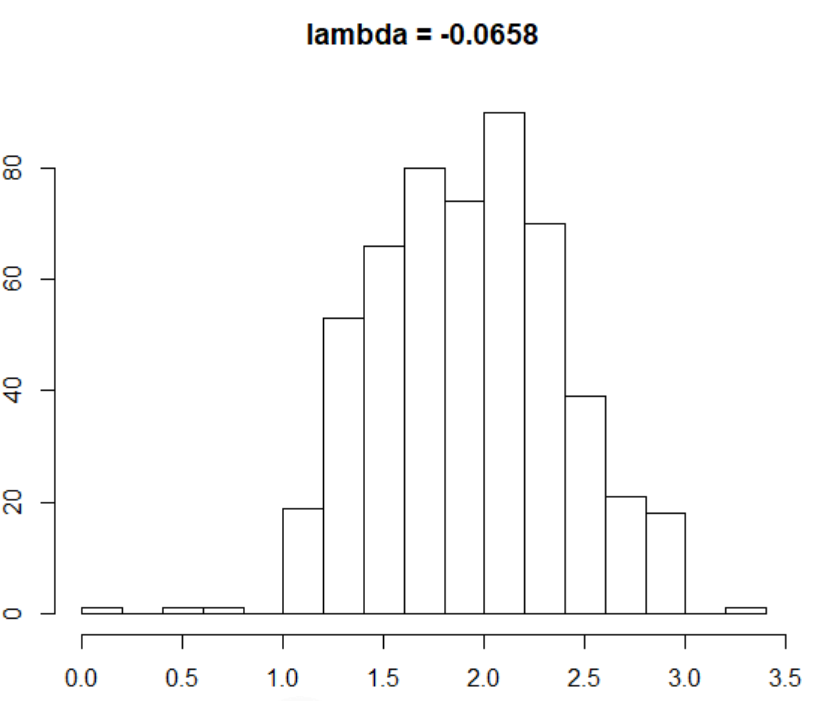
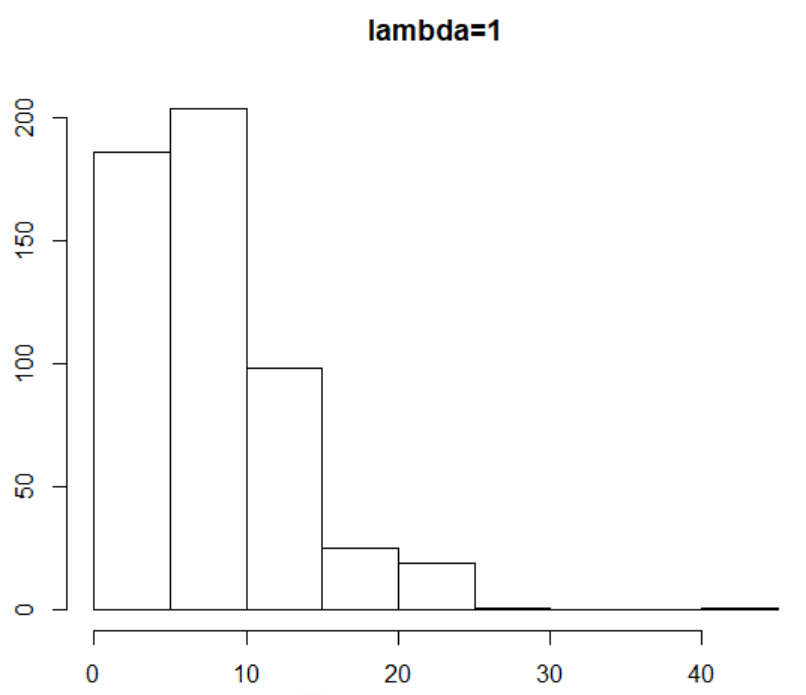
[a] Draw properly constructed histograms of *all three distributions* and discuss their properties,

hist(car::bcPower(CPS1985$wage, lambda=1),breaks = 12,main = 'lambda = 1')

hist(car::bcPower(CPS1985$wage, lambda= -0.0658 ),breaks = 12,main = 'lambda = -0.0658')

hist(car::bcPower(CPS1985$wage, lambda= -1.5 ),breaks = 12,main = 'lambda = -1.5')

are positively skewed, are close to the normal distribution with little positive skewness. are highly negatively skewed.



[b] evaluate their sknewness (see **e1071::skewness( )**), and

e1071::skewness(car::bcPower(CPS1985$wage, lambda=1))

[1] 1.687762

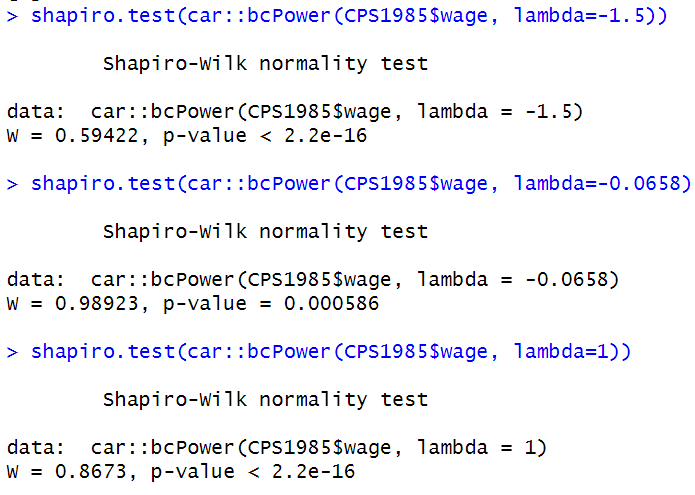
e1071::skewness(car::bcPower(CPS1985$wage, lambda=-0.0658))

[1] 0.001027028

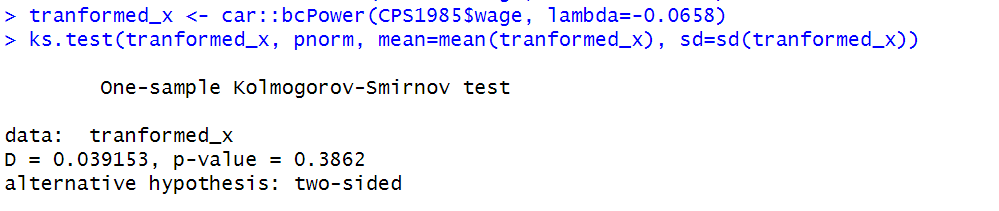
e1071::skewness(car::bcPower(CPS1985$wage, lambda=-1.5))

[1] -7.310307

[c] test whether these variables are approximately normal distributed (see **ks-test( )** and **shapiro.test( )**).



shapiro.test( ) are stricter than ks-test. Even using the best power transformation, we could not get a normal distribution from Shapiro. test, but the ks-test shows our result is good enough.



Address also the questions: Which transformed variable *comes the closest* to the normal distribution? Is the transformation with over-compensating the inherent positive skewness in the **wage** variable? [1.5 points]

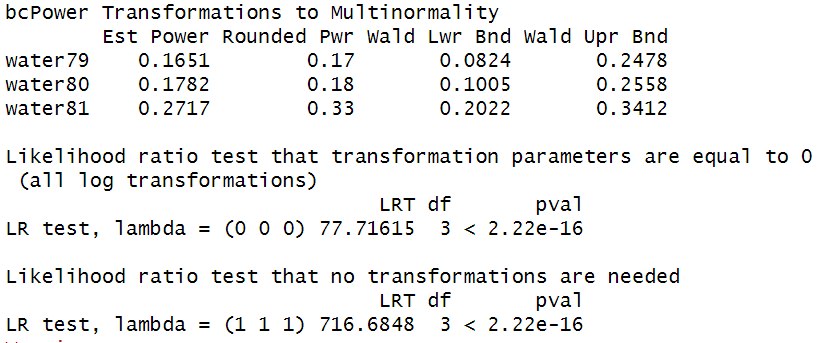
Second one () is the closest to the normal distribution. Yes, it becomes highly negatively skewed.

## Task 2: Explore the function powerTransform to achieve a multivariate normal distribution [1 point]

Read up in ’s online help about the function **car::powerTransfrom( )** in the **car** library. Use the variables **water81**, **water80**, and **water79** from the **Concord1.sav** data file.

**Task 2.1**: ***Simultaneously*** estimate the optimal set of Box-Cox transformation parameters for all four variables so that the set transformed variables becomes approximately multivariate normal distributed. Report your code to do the estimation. [0.5 points]

summary(lambda <- powerTransform(lm(cbind(water79,water80,water81)~1, data=Concord)))

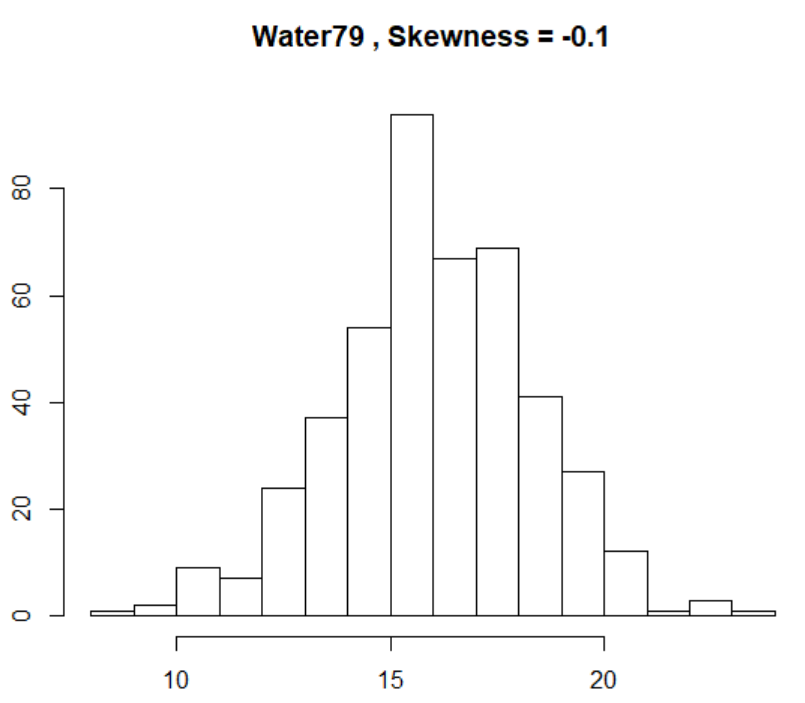
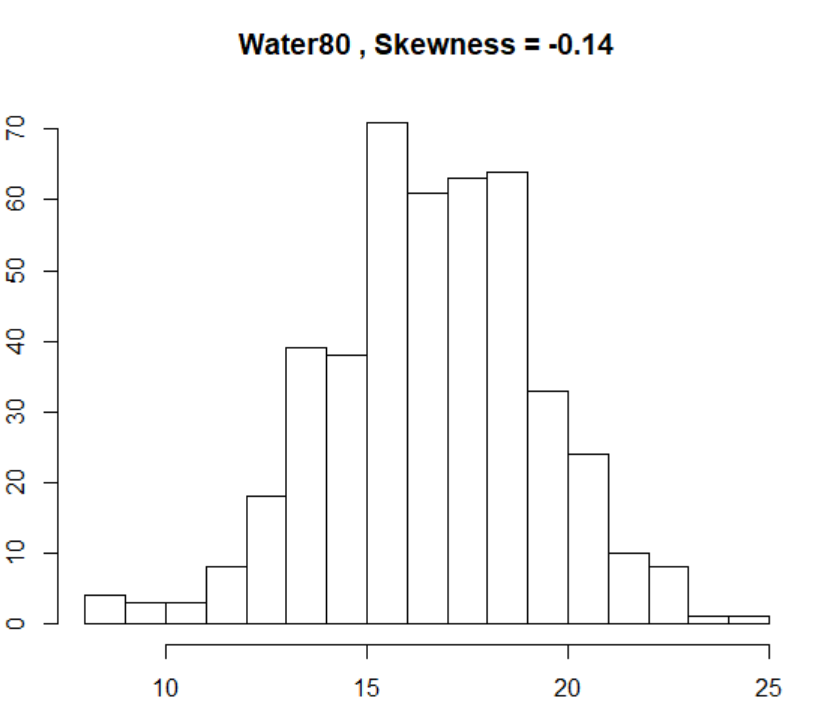
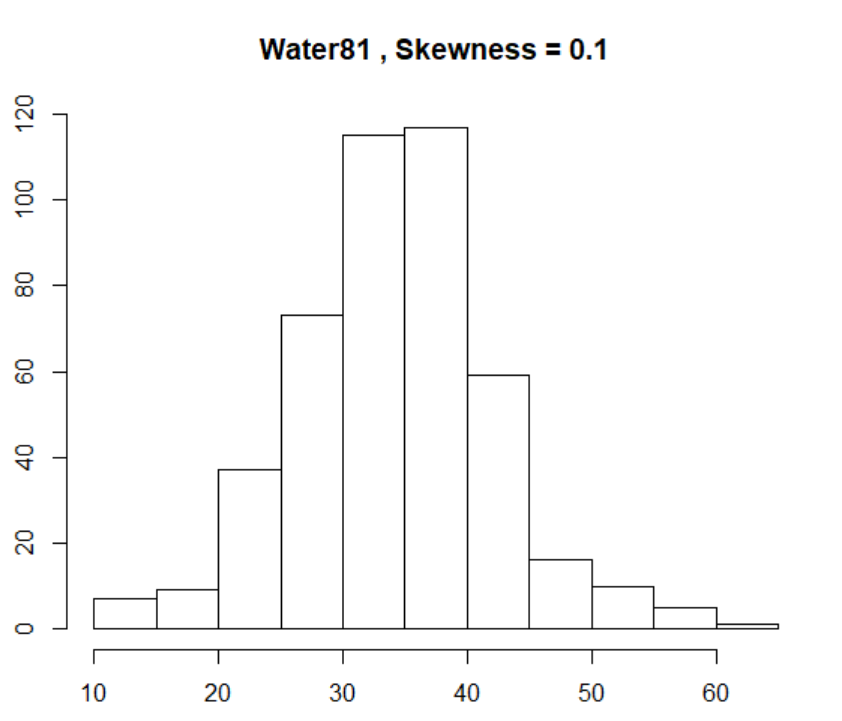


**Task 2.2:** Show the output and interpret the results. [0.5 points]

hist(Concord$Z1.0.17,breaks = 12,main = paste('Water79 , Skewness =',round(e1071::skewness(Concord$Z1.0.17),2)))

hist(Concord$Z2.0.18,breaks = 12,main = paste('Water80 , Skewness =',round(e1071::skewness(Concord$Z2.0.18),2)))

hist(Concord$Z3.0.33,breaks = 12,main = paste('Water81 , Skewness =',round(e1071::skewness(Concord$Z3.0.33),2)))

Although we do the power transformation, their distribution also postively skewed.

## Task 3: Confidence Intervals [2 points]

Continue with the **CPS1985** dataset for this task. To simplify things do not perform variable transformations.

**Task 3.1:** Run a bivariate regression of **wage** (dependent variable) on **education** (independent variable) and interpret the model estimates. [0.5 points]

reg01 <- lm(wage~education, data=CPS1985)

summary(reg01)

Residuals:

Min 1Q Median 3Q Max

-7.911 -3.260 -0.760 2.240 34.740

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

(Intercept) -0.74598 1.04545 -0.714 0.476

education 0.75046 0.07873 9.532 <2e-16 \*\*\*

---

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

Residual standard error: 4.754 on 532 degrees of freedom

Multiple R-squared: 0.1459, Adjusted R-squared: 0.1443

F-statistic: 90.85 on 1 and 532 DF, p-value: < 2.2e-16

Education is positively related to wage, and the regression coefficient is 0.75. R-squared is 0.146, which means education could only explain almost 14% variety in wage. the p-value smaller than 0.05, means we could reject the non-hypothesis, this model is meaningful. Intercept is not meaningful.

**Task 3.2:** Calculate the 99 % confidence intervals around the estimated regression parameters. Can you draw the same conclusion as you did using the *t*-test in the **summary** output of task 3.1? [0.5 points]

cbind("Coef"=coef(reg01), confint(reg01, level=0.99))

Coef 0.5 % 99.5 %

(Intercept) -0.7459797 -3.448585 1.9566260

education 0.7504608 0.546926 0.9539955

Yes

pt(-0.714,df = 532) \* 2 ## for Intercept

pt(9.532,df = 532,lower.tail = F) \*2 ## for regression coefficient

Get the same P-value as above.

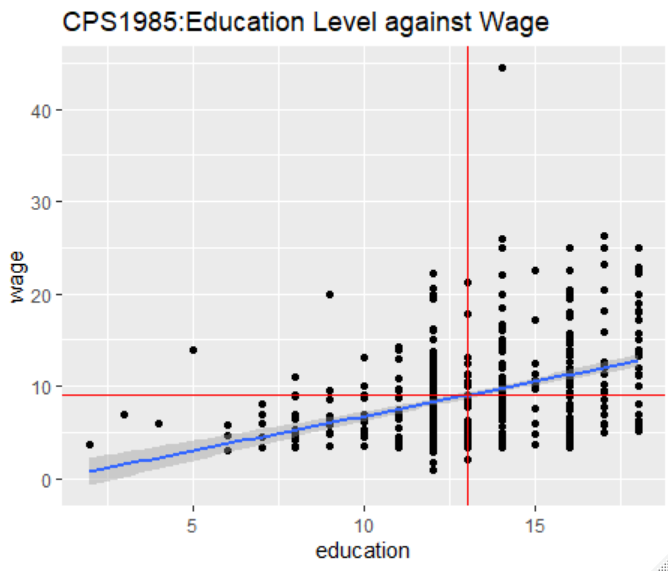
**Task 3.3:** Scatterplot both variables and add the predicted regression line as well as the lower and upper 90% confidence interval lines around the ***point predictions***.(i.e., prediction interval in Hamilton and **interval="prediction"** in the **predict** function).

q <- ggplot(CPS1985,aes(education,wage)) + geom\_point() + geom\_smooth(method=lm, se=TRUE,level =0.9)

q <- q + geom\_vline(xintercept =mean(education),color = "red") +

geom\_hline(yintercept =mean(wage),color = "red")

q + ggtitle('CPS1985:Education Level against Wage')



Task 4: Calibration and Prediction of a Bivariate Regression Model with Skewed Variables [3 points]

The **dBase** file **CampusCrime.dbf** has among other variables the count variables **crime** (number of crimes committed on university campuses) and **police** (size of the campuses police forces).

Note: To import **dBase** files use the  function **foreign::read.dbf( )**.

**Task 4.1:** Plot **police** in dependence of **crime** including their boxplots along the margins. Is a data transformation advisable?

df\_crime <- foreign::read.dbf('CampusCrime.dbf',as.is = T)

boxplot(df\_crime[,c("police","crime")])

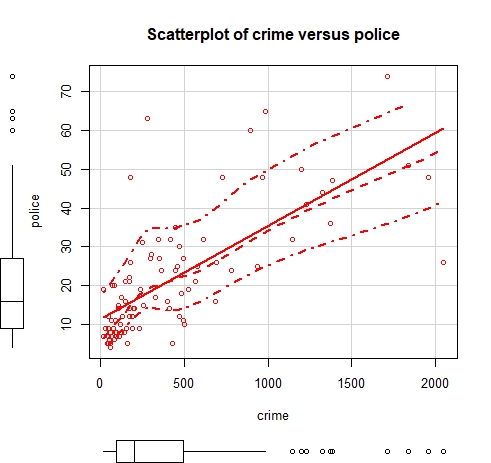
e1071::skewness(df\_crime$police, na.rm=TRUE)

e1071::skewness(df\_crime$crime, na.rm=TRUE)

1.285377

1.781008

Both are positively skewed, could use power transformation with lambda < 1. There is a little difference between median and expectation curve.



**Task 4.2:** Find a proper transformation of both variables in a way that the independent variable **crime** is approximately symmetrically distributed, and that the transformation of the dependent variable **police** leads to approximately symmetrically distributed regression residuals.

summary(x.lambda <- powerTransform(lm(crime~1, data=df\_crime)))

bcPower Transformation to Normality

Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd

Y1 0.0258 0 -0.1444 0.196

Likelihood ratio test that transformation parameter is equal to 0

(log transformation)

LRT df pval

LR test, lambda = (0) 0.08843754 1 0.76617

Likelihood ratio test that no transformation is needed

LRT df pval

LR test, lambda = (1) 107.9414 1 < 2.22e-16

summary(y.lambda <- powerTransform(lm(police~bcPower(df\_crime$crime,lambda = x.lambda$lambda), data=df\_crime))

bcPower Transformation to Normality

Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd

Y1 9e-04 0 -0.241 0.2429

Likelihood ratio test that transformation parameter is equal to 0

(log transformation)

LRT df pval

LR test, lambda = (0) 5.889249e-05 1 0.99388

Likelihood ratio test that no transformation is needed

LRT df pval

LR test, lambda = (1) 62.87245 1 2.2204e-15

**Task 4.3:** Test whether a ***log***-transformation (i.e., ) is appropriate for both, the dependent and the independent, variables.

Yes, according to the result above, is insignificant, so we could use ***log***-transformation here.

bcPower(cbind(df\_crime$police,df\_crime$crime), lambda = c(0,0))

**Task 4.4:** Estimate the model in the transformed system and interpret the estimates. Also ***test*** if the elasticity (i.e., slope parameter) differs significantly from the neutral elasticity of 1, i.e., . This could be done manually by using ’s standard error from the regression output or by using the function **car::linearHypothesis**.

.

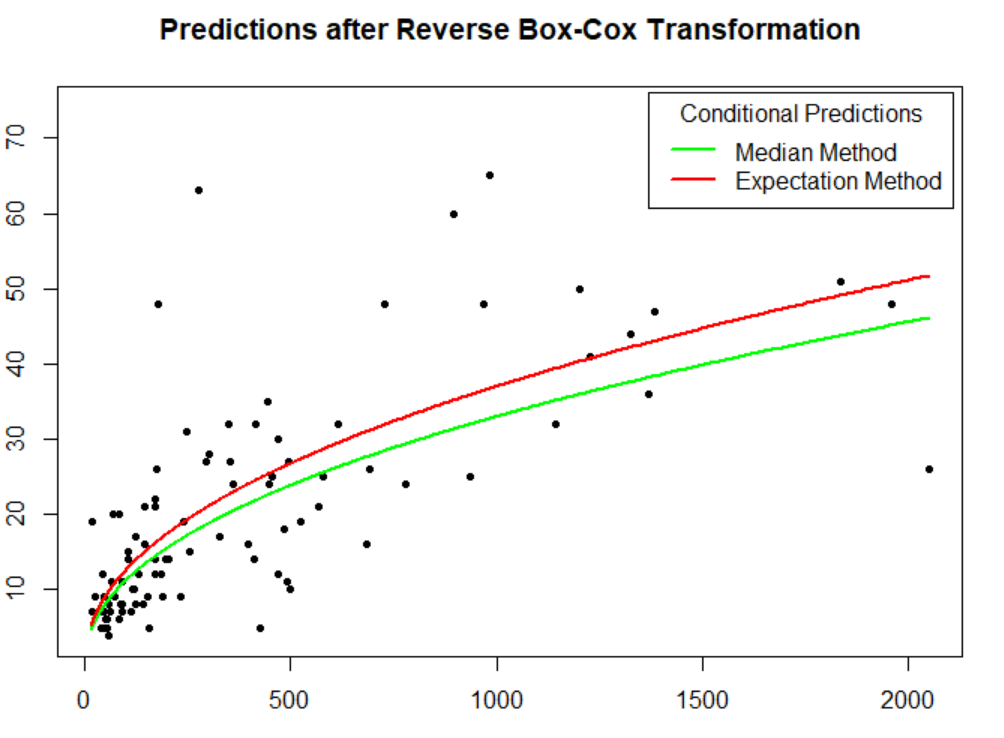
So, = -12.35031 with 91 degree of freedom

pt(t,df =91) \* 2

3.701025e-21

So, we reject the non-hypothesis, which means

**Task 4.5:** Perform a prediction in the original data units and plot the ***median*** and ***expectation*** curves. Interpret the plot both in terms of the median and expectation curves.



The gap between the two lines is smaller when the independent variable is tiny. And it becomes wider when variables go larger.

**Task 4.6:** Provide a ***clean***  script that documents your analysis. Do ***not show*** the provided functions in the sample script.

Already provided above.

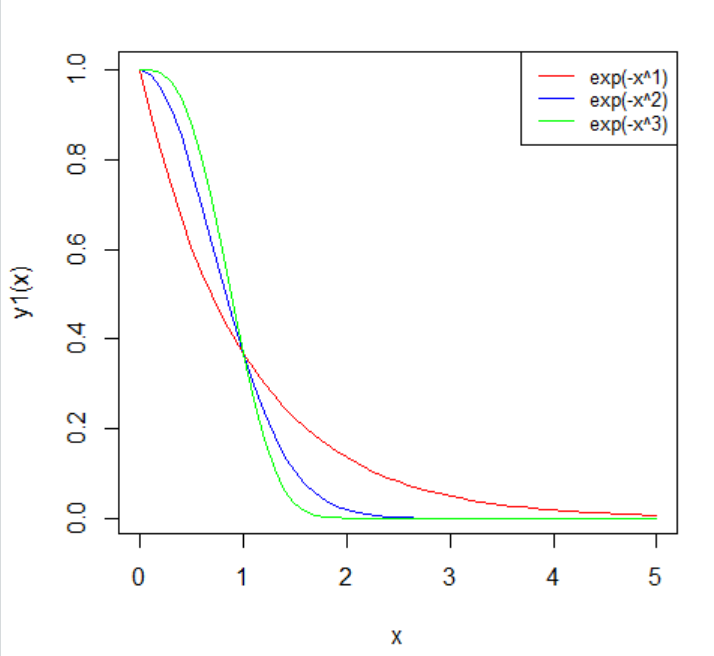
## Task 5: Numerical Integration [2 points]

Evaluate the three distance decay functions along a line of around the central reference point zero.

over their full distance range . Show the ***clean*** code for the tasks below.

**Task 5.1:** Plot these three functions within a reasonable value range of the distance . How does the shape of the curves change with increasing power?

Before x <1, when the power goes up, the decrease speed of the dependent variable (y) goes down. （）After x>1, when the power goes up, the decrease speed of the dependent variable (y) increases.



**Task 5.2:** Evaluate the areas , and underneath the three curves over their full support with ’s **integrate( )**-function.

(a1 <- integrate(y1, 0, Inf))

1 with absolute error < 5.7e-05

(a2 <- integrate(y2, 0, Inf))

0.8862269 with absolute error < 2.2e-06

(a3 <- integrate(y3, 0, Inf))

0.8929795 with absolute error < 3e-05

**Task 5.3:** Calculate the expectation of the distances for all three distance decay functions .

y11 <- function(x){x\*exp(-x^1)/a1$value}

y22 <- function(x){x\*exp(-x^2)/a2$value}

y33 <- function(x){x\*exp(-x^3)/a3$value}

(a11 <- integrate(y11, 0, Inf))

1 with absolute error < 6.4e-06

(a22 <- integrate(y22,0,Inf))

0.5641896 with absolute error < 3e-06

(a33 <- integrate(y33,0,Inf))

0.5054681 with absolute error < 9.1e-05

**Task 5.4:** Calculate the variances of the distances for all three distance decay functions .

y111 <- function(x){(x-a11$value)^2\*exp(-x^1)/a1$value}

y222 <- function(x){(x-a22$value)^2\*exp(-x^2)/a2$value}

y333 <- function(x){(x-a33$value)^2\*exp(-x^2)/a3$value}

integrate(y111, 0, Inf)

1 with absolute error < 5.8e-05

integrate(y222,0,Inf)

0.1816901 with absolute error < 2.8e-06

integrate(y333,0,Inf)

0.1177842 with absolute error < 7.8e-05

## Task 6: Appendix 1 [2 points]

**Task 6.1:** Why is a *t*-distributed random variable with degrees of freedom when it is squared identically to the -distributed random variable with one degree of freedom in the numerator and degrees of freedom for the denominator? Hint: use the definition of both random variables. [0.8 points]

For , has with degrees of freedom

= . Numerator in here is a distribution only contain z, and denominator is the *t*-distribution with degrees of freedom.

**Task 6.2:** Answer the questions **4 a** to **4 d** in Hamilton’s exercises on page 300. [1.2 points]

4a. same to X without error. Expectation of error equal to 0.

4b.

4c. error is not correlated with x. variation between is random. Regression coefficient between equal to 1.

4d. No, it means there is a linear relationship between error and x. So, the regression coefficient between would not equal to 1.

1. See pages 1-3 in Kleiber & Zeileis, 2008. Applied Econometrics with R. Springer Verlag. Available as *e-book* in UTD’s library. To import the data-frame use the statement **data(CPS1985, package=”AER”)** . [↑](#footnote-ref-1)