Multi Collinearity in R:

**Data Description**  
The datafile ([wagesmicrodata.xls](http://lib.stat.cmu.edu/datasets/CPS_85_Wages)) is downloaded from <http://lib.stat.cmu.edu/datasets/>. It contains 534 observations on 11 variables sampled from the Current Population Survey of 1985. The Current Population Survey (CPS) is used to supplement census information between census years. These data consist of a random sample of 534 persons from the CPS, with information on wages and other characteristics of the workers, including sex, number of years of education, years of work experience, occupational status, region of residence and union membership. We wish to determine whether wages are related to these characteristics.  
In particular, we are seeking for the following model:

$$ wage = \beta\_0 + \beta\_1 occupation + \beta\_2 sector + \beta\_3 union +\beta\_4 education \\+\beta\_5 experience +\beta\_6 age +\beta\_7 sex +\beta\_8 marital\_status \\ +\beta\_0 race +\beta\_10 south + u $$

After estimating the above model and running the post estimation diagnosis in R, it is seen that if we consider log of wages as the dependent variable, the variances seems to be more stabilized. Hence the log-transformed wage is used in the subsequent estimation, that is,  
$$ ln(wage) = \beta\_0 + \beta\_1 occupation + \beta\_2 sector \beta\_3 union +\beta\_4 education \\ +\beta\_5 experience +\beta\_6 age +\beta\_7 sex +\beta\_8 marital\_status \\ +\beta\_0 race +\beta\_10 south + u $$

Data Analysis in R

Import the data, and attach to R allowing you not to load data everytime you run the code below.

library(readxl)

wagesmicrodata <- read\_excel(file.choose(), sheet = "Data", skip = 0)

View(wagesmicrodata)

attach(wagesmicrodata)

**Fitting the Linear Model:**  
Assuming no multicollinearity, the model is being estimated using the following codes:

fit1<- lm(log(WAGE)~OCCUPATION+SECTOR+UNION+EDUCATION+EXPERIENCE+AGE+SEX+MARR+RACE+SOUTH)

To get the model summary:

fit1<- lm(log(WAGE)~OCCUPATION+SECTOR+UNION+EDUCATION+EXPERIENCE+AGE+SEX+MARR+RACE+SOUTH)

summary(fit1)

*Call:*

*lm(formula = log(WAGE) ~ OCCUPATION + SECTOR + UNION + EDUCATION +*

*EXPERIENCE + AGE + SEX + MARR + RACE + SOUTH)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-2.16246 -0.29163 -0.00469 0.29981 1.98248*

*Coefficients:*

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

*(Intercept) 1.078596 0.687514 1.569 0.117291*

*OCCUPATION -0.007417 0.013109 -0.566 0.571761*

*SECTOR 0.091458 0.038736 2.361 0.018589 \**

*UNION 0.200483 0.052475 3.821 0.000149 \*\*\**

*EDUCATION 0.179366 0.110756 1.619 0.105949*

*EXPERIENCE 0.095822 0.110799 0.865 0.387531*

*AGE -0.085444 0.110730 -0.772 0.440671*

*SEX -0.221997 0.039907 -5.563 4.24e-08 \*\*\**

*MARR 0.076611 0.041931 1.827 0.068259 .*

*RACE 0.050406 0.028531 1.767 0.077865 .*

*SOUTH -0.102360 0.042823 -2.390 0.017187 \**

*---*

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

*Residual standard error: 0.4398 on 523 degrees of freedom*

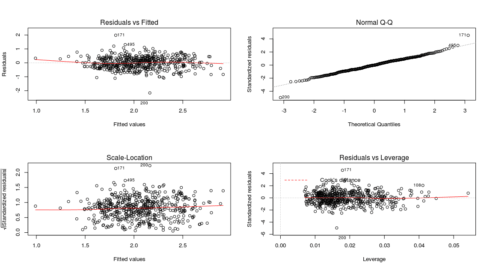
*Multiple R-squared: 0.3185, Adjusted R-squared: 0.3054*

*F-statistic: 24.44 on 10 and 523 DF, p-value: < 2.2e-16*

By looking at the model summary, the R-squared value of 0.31 is not bad for a cross sectional data of 534 observations. The F-value is highly significant implying that all the explanatory variables together significantly explain the log of wages. However, coming to the individual regression coefficients, it is seen that as many as four variables (occupation, education, experience, age) are not statistically significant and two (marital status and south) are significant only at 10 % level of significance.  
Further we can plot the model diagnostic checking for other problems such as normality of error term, heteroscedasticity etc.

par(mfrow=c(2,2))

plot(fit1)

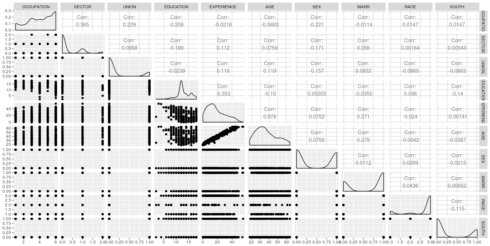
Gives this plot:  
[](https://i1.wp.com/datascienceplus.com/wp-content/uploads/2017/09/plot_zoom_png.png?ssl=1)  
Thus, the diagnostic plot is also look fair. So, possibly the multicollinearity problem is the reason for not getting many insignificant regression coefficients.

For further diagnosis of the problem, let us first look at the pair-wise correlation among the explanatory variables.

X<-wagesmicrodata[,3:12]

library(GGally)

ggpairs(X)

Gives this plot:  
[](https://i1.wp.com/datascienceplus.com/wp-content/uploads/2017/09/plot_zoom_png-1.png?ssl=1)

The correlation matrix shows that the pair-wise correlation among all the explanatory variables are not very high, except for the pair age – experience. The high correlation between age and experience might be the root cause of multicollinearity.  
Again by looking at the partial correlation coefficient matrix among the variables, it is also clear that the partial correlation between experience – education, age – education and age – experience are quite high.

library(corpcor)

cor2pcor(cov(X))

*[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]*

*[1,] 1.000000000 0.314746868 0.212996388 0.029436911 0.04205856 -0.04414029 -0.142750864 -0.018580965 0.057539374 0.008430595*

*[2,] 0.314746868 1.000000000 -0.013531482 -0.021253493 -0.01326166 0.01456575 -0.112146760 0.036495494 0.006412099 -0.021518760*

*[3,] 0.212996388 -0.013531482 1.000000000 -0.007479144 -0.01024445 0.01223890 -0.120087577 0.068918496 -0.107706183 -0.097548621*

*[4,] 0.029436911 -0.021253493 -0.007479144 1.000000000 -0.99756187 0.99726160 0.051510483 -0.040302967 0.017230877 -0.031750193*

*[5,] 0.042058560 -0.013261665 -0.010244447 -0.997561873 1.00000000 0.99987574 0.054977034 -0.040976643 0.010888486 -0.022313605*

*[6,] -0.044140293 0.014565751 0.012238897 0.997261601 0.99987574 1.00000000 -0.053697851 0.045090327 -0.010803310 0.021525073*

*[7,] -0.142750864 -0.112146760 -0.120087577 0.051510483 0.05497703 -0.05369785 1.000000000 0.004163264 0.020017315 -0.030152499*

*[8,] -0.018580965 0.036495494 0.068918496 -0.040302967 -0.04097664 0.04509033 0.004163264 1.000000000 0.055645964 0.030418218*

*[9,] 0.057539374 0.006412099 -0.107706183 0.017230877 0.01088849 -0.01080331 0.020017315 0.055645964 1.000000000 -0.111197596*

*[10,] 0.008430595 -0.021518760 -0.097548621 -0.031750193 -0.02231360 0.02152507 -0.030152499 0.030418218 -0.111197596 1.000000000*

Farrar – Glauber Test

The ‘mctest’ package in R provides the Farrar-Glauber test and other relevant tests for multicollinearity. There are two functions viz. ‘omcdiag’ and ‘imcdiag’ under ‘mctest’ package in R which will provide the overall and individual diagnostic checking for multicollinearity respectively.

library(mctest)

omcdiag(X,WAGE)

*Call:*

*omcdiag(x = X, y = WAGE)*

*Overall Multicollinearity Diagnostics*

*MC Results detection*

*Determinant |X'X|: 0.0001 1*

*Farrar Chi-Square: 4833.5751 1*

*Red Indicator: 0.1983 0*

*Sum of Lambda Inverse: 10068.8439 1*

*Theil's Method: 1.2263 1*

*Condition Number: 739.7337 1*

*1 --> COLLINEARITY is detected*

*0 --> COLLINEARITY in not detected by the test*

*===================================*

*Eigvenvalues with INTERCEPT*

*Intercept OCCUPATION SECTOR UNION EDUCATION EXPERIENCE AGE SEX MARR*

*Eigenvalues: 7.4264 0.9516 0.7635 0.6662 0.4205 0.3504 0.2672 0.0976 0.0462*

*Condition Indeces: 1.0000 2.7936 3.1187 3.3387 4.2027 4.6035 5.2719 8.7221 12.6725*

*RACE SOUTH*

*Eigenvalues: 0.0103 0.0000*

*Condition Indeces: 26.8072 739.7337*

The value of the standardized determinant is found to be 0.0001 which is very small. The calculated value of the Chi-square test statistic is found to be 4833.5751 and it is highly significant thereby implying the presence of multicollinearity in the model specification.  
This induces us to go for the next step of Farrar – Glauber test (F – test) for the location of the multicollinearity.

imcdiag(X,WAGE)

*Call:*

*imcdiag(x = X, y = WAGE)*

*All Individual Multicollinearity Diagnostics Result*

*VIF TOL Wi Fi Leamer CVIF Klein*

*OCCUPATION 1.2982 0.7703 17.3637 19.5715 0.8777 1.3279 0*

*SECTOR 1.1987 0.8343 11.5670 13.0378 0.9134 1.2260 0*

*UNION 1.1209 0.8922 7.0368 7.9315 0.9445 1.1464 0*

*EDUCATION 231.1956 0.0043 13402.4982 15106.5849 0.0658 236.4725 1*

*EXPERIENCE 5184.0939 0.0002 301771.2445 340140.5368 0.0139 5302.4188 1*

*AGE 4645.6650 0.0002 270422.7164 304806.1391 0.0147 4751.7005 1*

*SEX 1.0916 0.9161 5.3351 6.0135 0.9571 1.1165 0*

*MARR 1.0961 0.9123 5.5969 6.3085 0.9551 1.1211 0*

*RACE 1.0371 0.9642 2.1622 2.4372 0.9819 1.0608 0*

*SOUTH 1.0468 0.9553 2.7264 3.0731 0.9774 1.0707 0*

*1 --> COLLINEARITY is detected*

*0 --> COLLINEARITY in not detected by the test*

*OCCUPATION , SECTOR , EDUCATION , EXPERIENCE , AGE , MARR , RACE , SOUTH , coefficient(s) are non-significant may be due to multicollinearity*

*R-square of y on all x: 0.2805*

*\* use method argument to check which regressors may be the reason of collinearity*

*===================================*

The VIF, TOL and Wi columns provide the diagnostic output for variance inflation factor, tolerance and Farrar-Glauber F-test respectively. The F-statistic for the variable ‘experience’ is quite high (5184.0939) followed by the variable ‘age’ (F -value of 4645.6650) and ‘education’ (F-value of 231.1956). The degrees of freedom is \( (k-1 , n-k) \)or (9, 524). For this degrees of freedom at 5% level of significance, the theoretical value of F is 1.89774. Thus, the F test shows that either the variable ‘experience’ or ‘age’ or ‘education’ will be the root cause of multicollinearity. Though the F -value for ‘education’ is also significant, it may happen due to inclusion of highly collinear variables such as ‘age’ and ‘experience’.  
Finally, for examining the pattern of multicollinearity, it is required to conduct t-test for correlation coefficient. As there are ten explanatory variables, there will be six pairs of partial correlation coefficients. In R, there are several packages for getting the partial correlation coefficients along with the t- test for checking their significance level. We’ll the ‘ppcor’ package to compute the partial correlation coefficients along with the t-statistic and corresponding p-values.

library(ppcor)

pcor(X, method = "pearson")

*$estimate*

*OCCUPATION SECTOR UNION EDUCATION EXPERIENCE AGE SEX MARR RACE SOUTH*

*OCCUPATION 1.000000000 0.314746868 0.212996388 0.029436911 0.04205856 -0.04414029 -0.142750864 -0.018580965 0.057539374 0.008430595*

*SECTOR 0.314746868 1.000000000 -0.013531482 -0.021253493 -0.01326166 0.01456575 -0.112146760 0.036495494 0.006412099 -0.021518760*

*UNION 0.212996388 -0.013531482 1.000000000 -0.007479144 -0.01024445 0.01223890 -0.120087577 0.068918496 -0.107706183 -0.097548621*

*EDUCATION 0.029436911 -0.021253493 -0.007479144 1.000000000 -0.99756187 0.99726160 0.051510483 -0.040302967 0.017230877 -0.031750193*

*EXPERIENCE 0.042058560 -0.013261665 -0.010244447 -0.997561873 1.00000000 0.99987574 0.054977034 -0.040976643 0.010888486 -0.022313605*

*AGE -0.044140293 0.014565751 0.012238897 0.997261601 0.99987574 1.00000000 -0.053697851 0.045090327 -0.010803310 0.021525073*

*SEX -0.142750864 -0.112146760 -0.120087577 0.051510483 0.05497703 -0.05369785 1.000000000 0.004163264 0.020017315 -0.030152499*

*MARR -0.018580965 0.036495494 0.068918496 -0.040302967 -0.04097664 0.04509033 0.004163264 1.000000000 0.055645964 0.030418218*

*RACE 0.057539374 0.006412099 -0.107706183 0.017230877 0.01088849 -0.01080331 0.020017315 0.055645964 1.000000000 -0.111197596*

*SOUTH 0.008430595 -0.021518760 -0.097548621 -0.031750193 -0.02231360 0.02152507 -0.030152499 0.030418218 -0.111197596 1.000000000*

*$p.value*

*OCCUPATION SECTOR UNION EDUCATION EXPERIENCE AGE SEX MARR RACE SOUTH*

*OCCUPATION 0.000000e+00 1.467261e-13 8.220095e-07 0.5005235 0.3356824 0.3122902 0.001027137 0.6707116 0.18763758 0.84704000*

*SECTOR 1.467261e-13 0.000000e+00 7.568528e-01 0.6267278 0.7615531 0.7389200 0.010051378 0.4035489 0.88336002 0.62243025*

*UNION 8.220095e-07 7.568528e-01 0.000000e+00 0.8641246 0.8146741 0.7794483 0.005822656 0.1143954 0.01345383 0.02526916*

*EDUCATION 5.005235e-01 6.267278e-01 8.641246e-01 0.0000000 0.0000000 0.0000000 0.238259049 0.3562616 0.69337880 0.46745162*

*EXPERIENCE 3.356824e-01 7.615531e-01 8.146741e-01 0.0000000 0.0000000 0.0000000 0.208090393 0.3482728 0.80325456 0.60962999*

*AGE 3.122902e-01 7.389200e-01 7.794483e-01 0.0000000 0.0000000 0.0000000 0.218884070 0.3019796 0.80476248 0.62232811*

*SEX 1.027137e-03 1.005138e-02 5.822656e-03 0.2382590 0.2080904 0.2188841 0.000000000 0.9241112 0.64692038 0.49016279*

*MARR 6.707116e-01 4.035489e-01 1.143954e-01 0.3562616 0.3482728 0.3019796 0.924111163 0.0000000 0.20260170 0.48634504*

*RACE 1.876376e-01 8.833600e-01 1.345383e-02 0.6933788 0.8032546 0.8047625 0.646920379 0.2026017 0.00000000 0.01070652*

*SOUTH 8.470400e-01 6.224302e-01 2.526916e-02 0.4674516 0.6096300 0.6223281 0.490162786 0.4863450 0.01070652 0.00000000*

*$statistic*

*OCCUPATION SECTOR UNION EDUCATION EXPERIENCE AGE SEX MARR RACE SOUTH*

*OCCUPATION 0.0000000 7.5906763 4.9902208 0.6741338 0.9636171 -1.0114033 -3.30152873 -0.42541117 1.3193223 0.1929920*

*SECTOR 7.5906763 0.0000000 -0.3097781 -0.4866246 -0.3036001 0.3334607 -2.58345399 0.83597695 0.1467827 -0.4927010*

*UNION 4.9902208 -0.3097781 0.0000000 -0.1712102 -0.2345184 0.2801822 -2.76896848 1.58137652 -2.4799336 -2.2436907*

*EDUCATION 0.6741338 -0.4866246 -0.1712102 0.0000000 -327.2105031 308.6803174 1.18069629 -0.92332727 0.3944914 -0.7271618*

*EXPERIENCE 0.9636171 -0.3036001 -0.2345184 -327.2105031 0.0000000 1451.9092015 1.26038801 -0.93878671 0.2492636 -0.5109090*

*AGE -1.0114033 0.3334607 0.2801822 308.6803174 1451.9092015 0.0000000 -1.23097601 1.03321563 -0.2473135 0.4928456*

*SEX -3.3015287 -2.5834540 -2.7689685 1.1806963 1.2603880 -1.2309760 0.00000000 0.09530228 0.4583091 -0.6905362*

*MARR -0.4254112 0.8359769 1.5813765 -0.9233273 -0.9387867 1.0332156 0.09530228 0.00000000 1.2757711 0.6966272*

*RACE 1.3193223 0.1467827 -2.4799336 0.3944914 0.2492636 -0.2473135 0.45830912 1.27577106 0.0000000 -2.5613138*

*SOUTH 0.1929920 -0.4927010 -2.2436907 -0.7271618 -0.5109090 0.4928456 -0.69053623 0.69662719 -2.5613138 0.0000000*

*$n*

*[1] 534*

*$gp*

*[1] 8*

*$method*

*[1] "pearson"*

As expected the high partial correlation between ‘age’ and ‘experience’ is found to be highly statistically significant. Similar is the case for ‘education – experience’ and ‘education – age’ . Not only that even some of the low correlation coefficients are also found to be highyl significant. Thus, the Farrar-Glauber test points out that X1 is the root cause of all multicollinearity problem.

Remedial Measures

There are several remedial measure to deal with the problem of multicollinearity such Prinicipal Component Regression, Ridge Regression, Stepwise Regression etc.  
However, in the present case, I’ll go for the exclusion of the variables for which the VIF values are above 10 and as well as the concerned variable logically seems to be redundant. Age and experience will certainly be correlated. So, why to use both of them? If we use ‘age’ or ‘age-squared’, it will reflect the experience of the respondent also. Thus, we try to build a model by excluding ‘experience’, estimate the model and go for further diagnosis for the presence of multicollinearity.

fit2<- lm(log(WAGE)~OCCUPATION+SECTOR+UNION+EDUCATION+AGE+SEX+MARR+RACE+SOUTH)

summary(fit3)

*Call:*

*lm(formula = log(WAGE) ~ OCCUPATION + SECTOR + UNION + EDUCATION +*

*AGE + SEX + MARR + RACE + SOUTH)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-2.16018 -0.29085 -0.00513 0.29985 1.97932*

*Coefficients:*

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

*(Intercept) 0.501358 0.164794 3.042 0.002465 \*\**

*OCCUPATION -0.006941 0.013095 -0.530 0.596309*

*SECTOR 0.091013 0.038723 2.350 0.019125 \**

*UNION 0.200018 0.052459 3.813 0.000154 \*\*\**

*EDUCATION 0.083815 0.007728 10.846 < 2e-16 \*\*\**

*AGE 0.010305 0.001745 5.905 6.34e-09 \*\*\**

*SEX -0.220100 0.039837 -5.525 5.20e-08 \*\*\**

*MARR 0.075125 0.041886 1.794 0.073458 .*

*RACE 0.050674 0.028523 1.777 0.076210 .*

*SOUTH -0.103186 0.042802 -2.411 0.016261 \**

*---*

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

*Residual standard error: 0.4397 on 524 degrees of freedom*

*Multiple R-squared: 0.3175, Adjusted R-squared: 0.3058*

*F-statistic: 27.09 on 9 and 524 DF, p-value: < 2.2e-16*

Now by looking at the significance level, it is seen that out of nine of regression coefficients, eight are statistically significant. The R-square value is 0.31 and F-value is also very high and significant too.  
Even the VIF values for the explanatory variables have reduced to very lower values.

vif(fit2)

*OCCUPATION SECTOR UNION EDUCATION AGE SEX MARR RACE SOUTH*

*1.295935 1.198460 1.120743 1.125994 1.154496 1.088334 1.094289 1.037015 1.046306*

So, the model is now free from multicollinearity.

# **Should I always transform my variables to make them normal?**

When I first learned data analysis, I always checked normality for each variable and made sure they were normally distributed before running any analyses, such as t-test, ANOVA, or linear regression. I thought normal distribution of variables was the important assumption to proceed to analyses. That’s why stats textbooks show you how to draw histograms and QQ-plots in the beginning of data analysis in the early chapters and see if they’re normally distributed, isn’t it? There I was, drawing histograms, looking at the shape and thinking, *“Oh, no, my data are not normal. I should transform them first or I can’t run any analyses.”*

No, you don’t have to transform your observed variables just because they don’t follow a normal distribution. Linear regression analysis, which includes t-test and ANOVA, does not assume normality for either predictors (IV) or an outcome (DV).

*No way! When I learned regression analysis, I remember my stats professor said we should check normality!*

Yes, you should check normality of errors AFTER modeling. In linear regression, errors are assumed to follow a normal distribution with a mean of zero.

Y = intercept + coefficient \* X + error

Let’s do some simulations and see how normality influences analysis results and see what could be consequences of normality violation.

# Simulation conditions

# sample size = 30, true coefficient = 0.3

# replications = 10000

set.seed(2015) # if you want to get exactly the same results as here

# Case 1: Errors are normally distributed

results1 = data.frame(est=NA, se=NA, t=NA, p=NA)

for(i in 1:10000){

x = scale(rchisq(30, 1)) # non-normal x

error = rnorm(30) # normal error

y = 0 + 0.3\*x + error # y regressed on x and error

m = lm(y ~ x)

results1[i,] = summary(m)$coefficients['x',]

}

# Case 2: Errors are NOT normally distributed

results2 = data.frame(est=NA, se=NA, t=NA, p=NA)

for(i in 1:10000){

x = scale(rchisq(30, 1)) # non-normal x

error = scale(rchisq(30, 1)) # non-normal errors

y = 0 + 0.3\*x + error # y regressed on x and error

m = lm(y ~ x)

results2[i,] = summary(m)$coefficients['x',]

}

If you want to visually assess if the distribution of each variable looks normal:

qqnorm(x); qqline(x)

qqnorm(error); qqline(error)

qqnorm(y); qqline(y)

Tip: Check out another StatLab article, [*Understanding Q-Q Plots*](http://data.library.virginia.edu/understanding-q-q-plots/).

Let’s look at means of the results of 10000 replications.

> colMeans(results1)

est se t p

0.2990674 0.1838041 1.6570579 0.2259360

> colMeans(results2)

est se t p

0.3007691 0.1855503 1.6404420 0.2459877

Wait, didn’t I say the errors should be normally distributed? They are essentially the same! It seems like it’s working totally fine even with non-normal errors.

In fact, linear regression analysis works well, even with non-normal errors. But, the problem is with p-values for hypothesis testing.

After running a linear regression, what researchers would usually like to know is–is the coefficient different from zero? The t-statistics (and its corresponding p-value) answers the question if the estimated coefficient is statistically significantly different from zero.

Let’s look at the distributions of the two results.

# The estimates are normally distributed in Case 1

hist(results1$est, breaks=100, freq=FALSE,

xlim=c(-0.5, 1.1), ylim=c(0, 2.5),

main='Case 1: Normal Errors', xlab='Coefficient Estimation')

curve(dnorm(x, mean=mean(results1$est), sd=sd(results1$est)),

col='red', lwd=3, add=TRUE)

abline(v=0.3, col='red', lwd=3)

# The estimates are NOT normally distributed in Case 2

hist(results2$est, breaks=100, freq=FALSE,

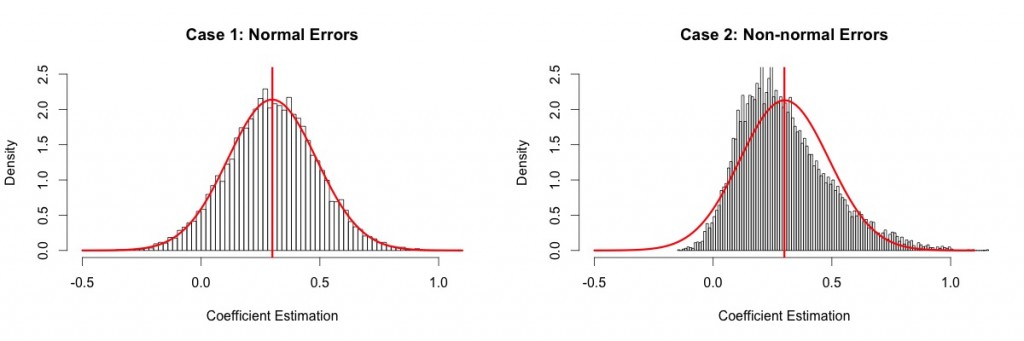
xlim=c(-0.5, 1.1), ylim=c(0, 2.5),

main='Case 2: Non-normal Errors', xlab='Coefficient Estimation')

curve(dnorm(x, mean=mean(results2$est), sd=sd(results2$est)),

col='red', lwd=3, add=TRUE)

abline(v=0.3, col='red', lwd=3)

[](http://data.library.virginia.edu/files/estimations.jpeg)

Now we can see differences. The distribution of estimated coefficients follows a normal distribution in Case 1, but not in Case 2. That means that in Case 2 we cannot apply hypothesis testing, which is based on a normal distribution (or related distributions, such as a t-distribution). When errors are not normally distributed, estimations are not normally distributed and we can no longer use p-values to decide if the coefficient is different from zero. In short, if the normality assumption of the errors is not met, we cannot draw a valid conclusion based on statistical inference in linear regression analysis.

And even then those procedures are actually pretty robust to violations of normality. In our second example above, our simulated sample size was 30 (kind of small) and our errors were drawn from a chi-square distribution with 1 degree of freedom. (You can’t get any more non-normal than that!) And yet the sampling distribution histogram of the coefficient was not as far from normal as you might expect. Now if your sample is small (less than 30) and you detect extremely non-normal errors, you might consider alternatives to constructing standard errors and p-values, such as bootstrapping. But otherwise you can probably rest easy if your errors seem “normal enough”.

*Okay, I understand my variables don’t have to be normal. Why do we even bother checking histogram before analysis then?*

Although your data don’t have to be normal, it’s still a good idea to check data distributions just to understand your data. Do they look reasonable? Your data might not be normal for a reason. Is it count data or reaction time? In such cases, you may want to transform it or use other analysis methods (e.g., generalized linear models or nonparametric methods). The relationship between two variables may also be non-linear (which you might detect with a scatterplot). In that case transforming one or both variables may be necessary.

**Summary:**  
None of your observed variables have to be normal in linear regression analysis, which includes t-test and ANOVA. The errors after modeling, however, should be normal to draw a valid conclusion by hypothesis testing.

**Note:**  
There are other analysis methods that assume multivariate normality for observed variables (e.g., Structural Equation Modeling).

# Outlier detection and treatment with R

* Published on December 9, 2016 at 10:54 am
* Updated on April 29, 2017 at 3:13 pm
* 11kViews

[45](https://www.addtoany.com/share" \l "url=https%3A%2F%2Fdatascienceplus.com%2Foutlier-detection-and-treatment-with-r%2F&title=Outlier%20detection%20and%20treatment%20with%20R%20%7C%20DataScience%2B" \o "Give it a Share)

[Shares](https://www.addtoany.com/share" \l "url=https%3A%2F%2Fdatascienceplus.com%2Foutlier-detection-and-treatment-with-r%2F&title=Outlier%20detection%20and%20treatment%20with%20R%20%7C%20DataScience%2B" \o "Give it a Share)

* [8Comments](https://datascienceplus.com/outlier-detection-and-treatment-with-r/#share-wrapper)

Outliers in data can distort predictions and affect the accuracy, if you don’t detect and handle them appropriately especially in regression models.

## Why outliers detection is important?

Treating or altering the outlier/extreme values in genuine observations is not the standard operating procedure. However, it is essential to understand their impact on your predictive models. It is left to the best judgement of the investigator to decide whether treating outliers is necessary and how to go about it.

So, why identifying the extreme values is important? Because, it can drastically bias/change the fit estimates and predictions. Let me illustrate this using the cars dataset.

To better understand the implications of outliers better, I am going to compare the fit of a simple linear regression model on cars dataset with and without outliers. In order to distinguish the effect clearly, I manually introduce extreme values to the original cars dataset. Then, I predict on both the datasets.

# Inject outliers into data.

cars1 <- cars[1:30, ] # original data

cars\_outliers <- data.frame(speed=c(19,19,20,20,20), dist=c(190, 186, 210, 220, 218)) # introduce outliers.

cars2 <- rbind(cars1, cars\_outliers) # data with outliers.

# Plot of data with outliers.

par(mfrow=c(1, 2))

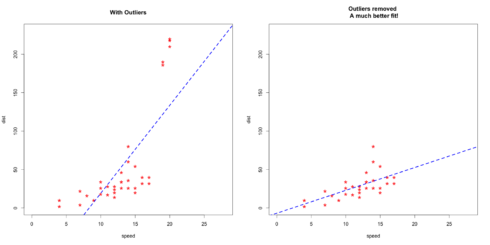
plot(cars2$speed, cars2$dist, xlim=c(0, 28), ylim=c(0, 230), main="With Outliers", xlab="speed", ylab="dist", pch="\*", col="red", cex=2)

abline(lm(dist ~ speed, data=cars2), col="blue", lwd=3, lty=2)

# Plot of original data without outliers. Note the change in slope (angle) of best fit line.

plot(cars1$speed, cars1$dist, xlim=c(0, 28), ylim=c(0, 230), main="Outliers removed \n A much better fit!", xlab="speed", ylab="dist", pch="\*", col="red", cex=2)

abline(lm(dist ~ speed, data=cars1), col="blue", lwd=3, lty=2)

Plot of original data without outliers:  
[](https://datascienceplus.com/wp-content/uploads/2016/12/outliers_effect.png)

Notice the change in slope of the best fit line after removing the outliers. Had we used the outliers to train the model(left chart), our predictions would be exagerated (high error) for larger values of speed because of the larger slope.

## Detect Outliers

**Univariate approach**  
For a given continuous variable, outliers are those observations that lie outside 1.5 \* IQR, where IQR, the ‘Inter Quartile Range’ is the difference between 75th and 25th quartiles. Look at the points outside the whiskers in below box plot.

url <- "http://rstatistics.net/wp-content/uploads/2015/09/ozone.csv"

# alternate source: https://raw.githubusercontent.com/selva86/datasets/master/ozone.csv

inputData <- read.csv(url) # import data

outlier\_values <- boxplot.stats(inputData$pressure\_height)$out # outlier values.

boxplot(inputData$pressure\_height, main="Pressure Height", boxwex=0.1)

mtext(paste("Outliers: ", paste(outlier\_values, collapse=", ")), cex=0.6)

### Bivariate approach

Visualize in box-plot of the X and Y, for categorical X’s

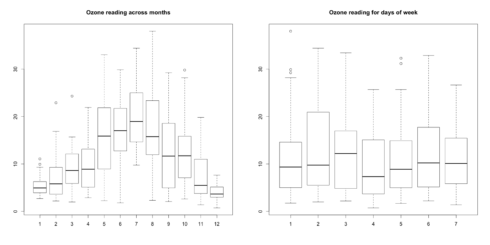
url <- "http://rstatistics.net/wp-content/uploads/2015/09/ozone.csv"

ozone <- read.csv(url)

# For categorical variable

boxplot(ozone\_reading ~ Month, data=ozone, main="Ozone reading across months") # clear pattern is noticeable.

boxplot(ozone\_reading ~ Day\_of\_week, data=ozone, main="Ozone reading for days of week") # this may not be significant, as day of week variable is a subset of the month var.

Box-plot:  
[](https://datascienceplus.com/wp-content/uploads/2016/12/bivariate_boxplot.png)

What is the inference? The change in the level of boxes suggests that Month seem to have an impact in ozone\_reading while Day\_of\_week does not. Any outliers in respective categorical level show up as dots outside the whiskers of the boxplot.

# For continuous variable (convert to categorical if needed.)

boxplot(ozone\_reading ~ pressure\_height, data=ozone, main="Boxplot for Pressure height (continuos var) vs Ozone")

boxplot(ozone\_reading ~ cut(pressure\_height, pretty(inputData$pressure\_height)), data=ozone, main="Boxplot for Pressure height (categorial) vs Ozone", cex.axis=0.5)