Golas, Ekal

UNIVERSITY OF TEXAS AT DALLAS | 800 W CaMPBELL RD RICHARDSON TX 75080

MINI PROJECT 6

Statistical methods for data science



Table of Contents

[Answers 2](#_Toc436917485)

[R Code 6](#_Toc436917486)

# Answers

The answers to the questions stated in (Choudhary, 2015) are as follows:-

1. **Consider the prostate cancer dataset from Mini Project 5. Choose one more quantitative or qualitative variable to add to the final model built in that project and justify by our choice. Fit a linear model using the two variables and carry out regression diagnostics. The analysis should include an assessment of the degree to which the key regression assumptions are satisfied. If an assumption is not met, attempt to remedy the situation. Comment on the fit of the final model using appropriate tests and statistics.**

Models were constructed for each variable added to the existing model. The anova table was then constructed to compare these models. As per the results of anova table, Seminal vesical invasion seems to be of most value. Hence, Seminal vesical invasion is chosen as one more variable to be added to the final model. The model now is:-

Call:

lm(formula = log(cancer.data$psa) ~ log(cancer.data$weight) + (cancer.data$vesinv), data = cancer.data)

Residuals:

Min 1Q Median 3Q Max

-2.08878 -0.49949 0.05155 0.56001 1.76260

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

(Intercept) -0.3454 0.6804 -0.508 0.612921

log(cancer.data$weight) 0.6849 0.1856 3.691 0.000375 \*\*\*

cancer.data$vesinv 1.4879 0.2227 6.680 1.67e-09 \*\*\*

---

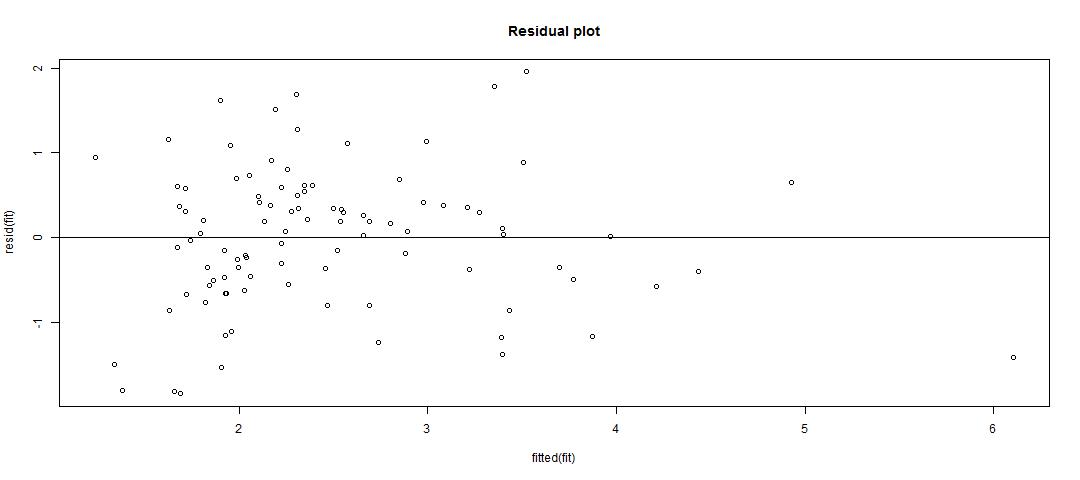
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

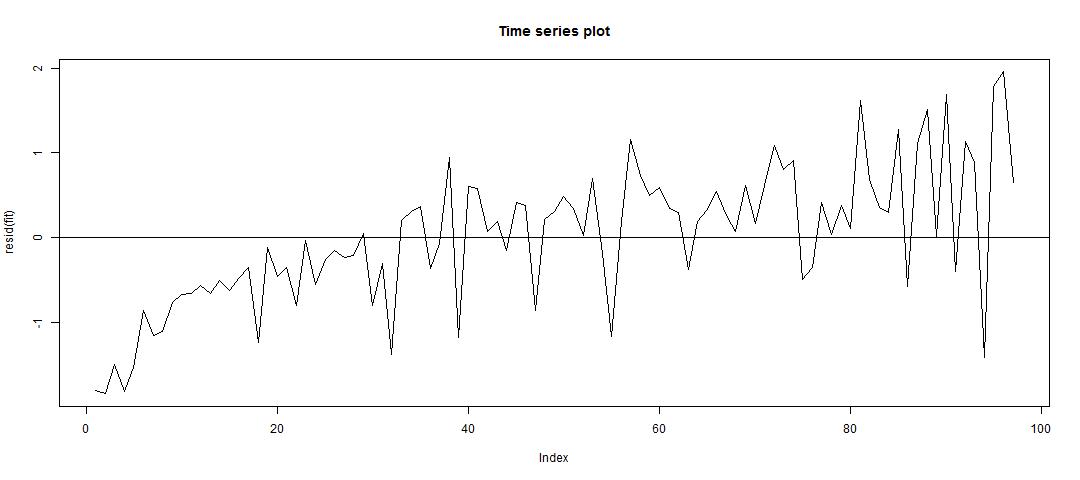
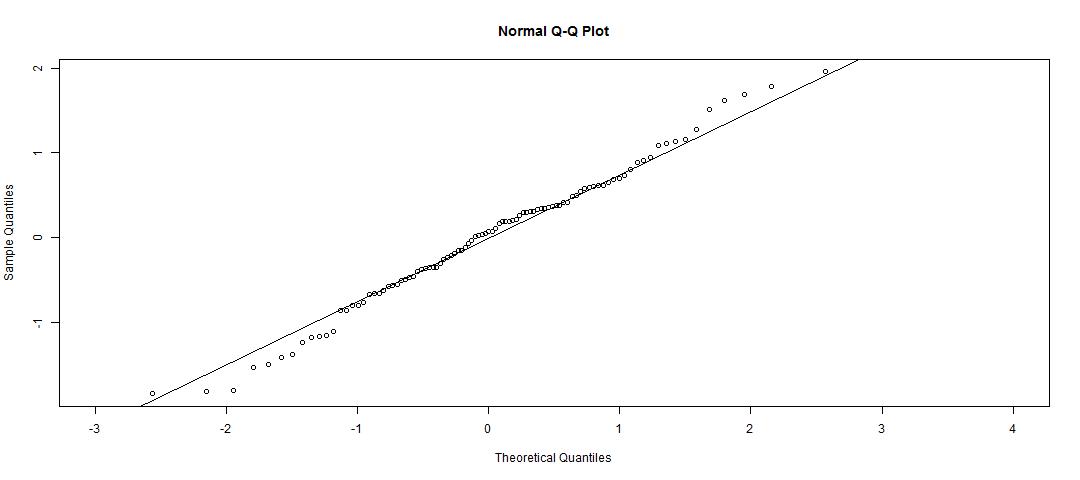
Residual standard error: 0.898 on 94 degrees of freedom

Multiple R-squared: 0.4067, Adjusted R-squared: 0.3941

F-statistic: 32.22 on 2 and 94 DF, p-value: 2.201e-11

The plots for the above model are as follows:-





From the above plots we can see that the key regression assumptions are slightly violated. Hence log transform of the new variable is applied and the final model is computed as:-

Call:

lm(formula = log(cancer.data$psa) ~ log(cancer.data$weight) + log(cancer.data$cancervol), data = cancer.data)

Residuals:

Min 1Q Median 3Q Max

-1.6235 -0.5043 -0.0219 0.5202 1.9060

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

(Intercept) -0.30097 0.56903 -0.529 0.5981

log(cancer.data$weight) 0.51096 0.15723 3.250 0.0016 \*\*

log(cancer.data$cancervol) 0.67662 0.06628 10.208 <2e-16 \*\*\*

---

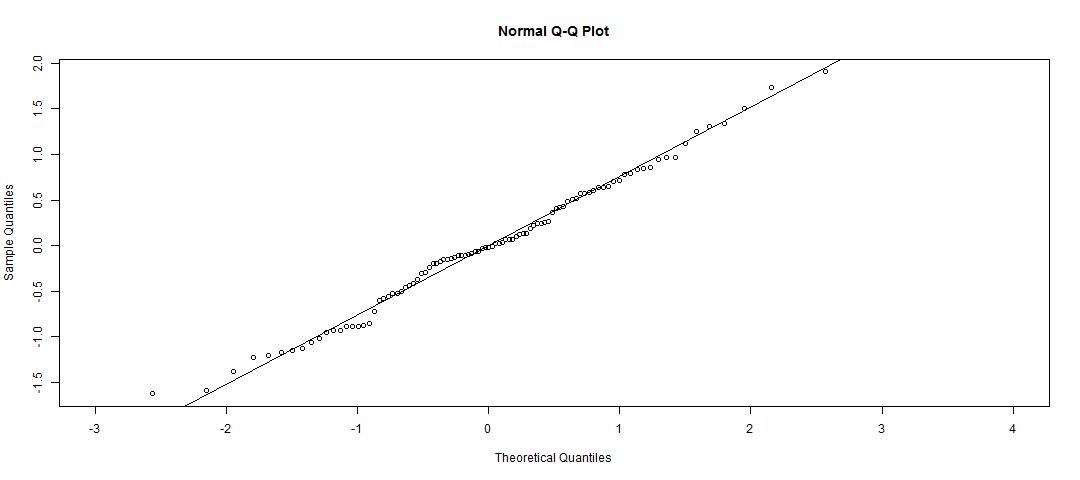
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

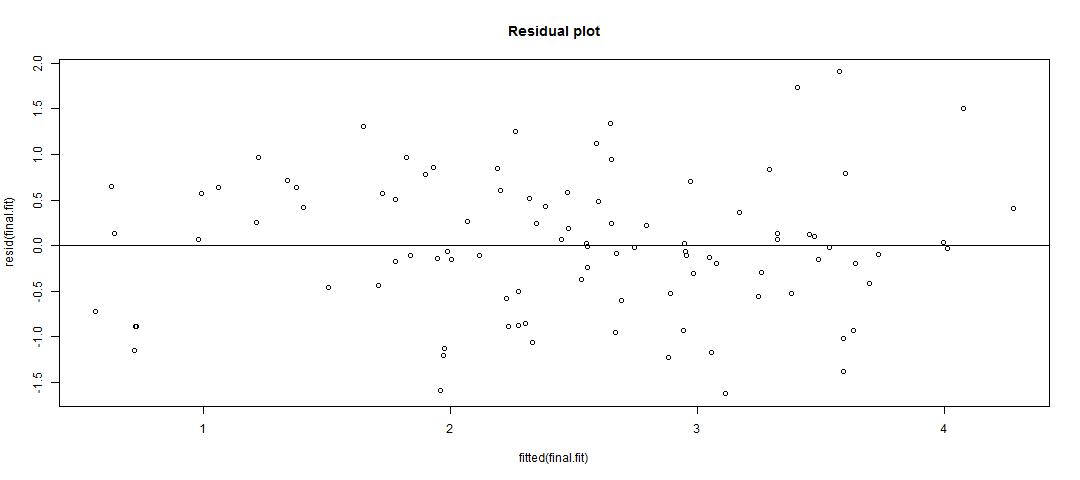
Residual standard error: 0.751 on 94 degrees of freedom

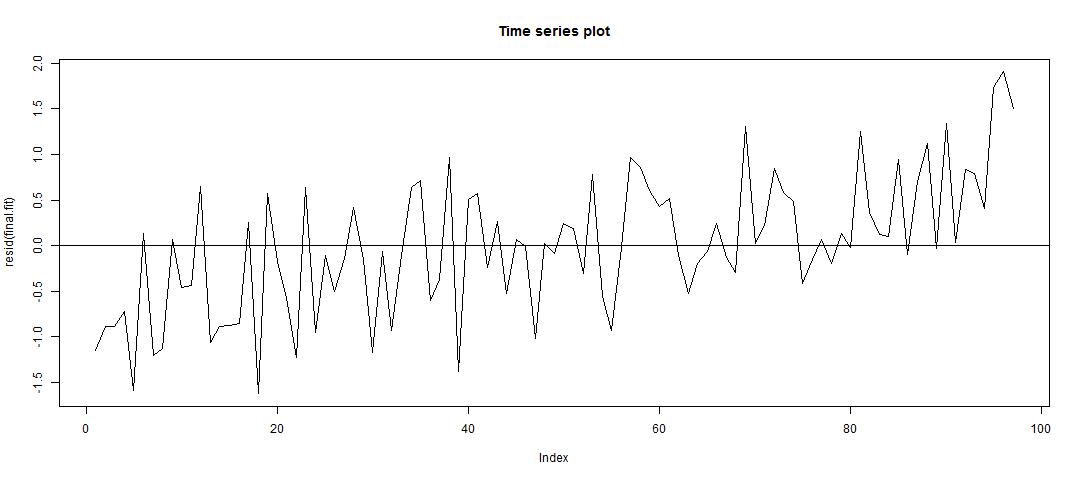
Multiple R-squared: 0.5851, Adjusted R-squared: 0.5763

F-statistic: 66.28 on 2 and 94 DF, p-value: < 2.2e-16

The plots for the final model are as follows:-







From the above plots, we can see that the key regression assumptions are more closely satisfied than the initial model.

1. **Use the final model to predict the PSA level for a patient whose predictor variables are at the sample medians of the variables.**

The median of the Weight variable is 37.338 and the median for seminal vesical invasion variable is 0.

The PSA level for the patient at the median value of predicted variables is 12.71679

# R Code

# Read the data

cancer.data = read.table("prostate\_cancer.csv", sep = ",", header = TRUE)

# Fit the model with two variables

fit1 = lm(log(cancer.data$psa) ~ log(cancer.data$weight) + (cancer.data$cancervol), cancer.data)

fit2 = lm(log(cancer.data$psa) ~ log(cancer.data$weight) + (cancer.data$age), cancer.data)

fit3 = lm(log(cancer.data$psa) ~ log(cancer.data$weight) + (cancer.data$benpros), cancer.data)

fit4 = lm(log(cancer.data$psa) ~ log(cancer.data$weight) + (cancer.data$vesinv), cancer.data)

fit5 = lm(log(cancer.data$psa) ~ log(cancer.data$weight) + (cancer.data$capspen), cancer.data)

fit6 = lm(log(cancer.data$psa) ~ log(cancer.data$weight) + factor(cancer.data$gleason), cancer.data)

anova(fit1, fit2, fit3, fit4, fit5, fit6)

summary(fit4)

# Residual plot

jpeg("Residual plot.jpg", width = 1080, height = 480)

plot(fitted(fit4), resid(fit4), main = "Residual plot")

abline(h=0)

dev.off()

# QQ plot

jpeg("QQ Plot.jpg", width = 1080, height = 480)

qqnorm(resid(fit4), xlim = c(-3, 4))

qqline(resid(fit4))

dev.off()

# Time series plot of residuals

jpeg("Time series plot.jpg", width = 1080, height = 480)

plot(resid(fit4), type="l", main = "Time series plot")

abline(h=0)

dev.off()

# Log transform and fit the model with two variables

final.fit = lm(log(cancer.data$psa) ~ log(cancer.data$weight) + log(cancer.data$vesinv), cancer.data)

summary(final.fit)

# Residual plot

jpeg("Final - Residual plot.jpg", width = 1080, height = 480)

plot(fitted(final.fit), resid(final.fit), main = "Residual plot")

abline(h=0)

dev.off()

# QQ plot

jpeg("Final - QQ Plot.jpg", width = 1080, height = 480)

qqnorm(resid(final.fit), xlim = c(-3, 4))

qqline(resid(final.fit))

dev.off()

# Time series plot of residuals

jpeg("Final - Time series plot.jpg")

plot(resid(final.fit), type="l", main = "Time series plot")

abline(h=0)

dev.off()

# Predict for median

weight.median = median(cancer.data$weight)

ves.median = median(cancer.data$vesinv)

index = cancer.data[which(cancer.data$weight == weight.median & cancer.data$vesinv == ves.median), ]

new.data = data.frame(x = c(weight.median, ves.median))

prediction = predict(fit, newdata = new.data)

exp(prediction[index$subject])