Midterm Yash Hathi

2025-03-08

Contents

PREAMBLE defines variables

```
Dataset<-read.csv(file="C:\\Users\\yashi\\Downloads\\colorectal_fixed.csv")
DataName<-"colorectal_cancer"
x1name<-"Cancer_Stage"
x2name<-"Physical_Activity"
x3name <- "Alcohol_Consumption"
x4name <- "Smoking_History"
x5name <- "Urban_or_Rural"
x6name <- "Age"
y1name<-"Tumor_Size_mm"
```

```
## Getting data from variable names (this should only depend upon preamble)
x1<-as.factor(Dataset[,x1name])
x2<-as.factor(Dataset[,x2name])
x3<- as.factor(Dataset[,x3name])
x4 <-as.factor(Dataset[,x4name])
x5 <-as.factor(Dataset[,x5name])
x6 <- Dataset[,x6name]
y1<-Dataset[,y1name]
```

Part 1: ANOVA Test

The first thing we are going to do is make a factorial ANOVA model to test if Smoking_History, Physical_Activity, or a interaction of the two have a significant effect on Tumor_Size_mm. We define our null hypothesis as followed:

H0: Smoking_History has no significant effect on Tumor_Size_mm

H0: Physical_Activity has no significant effect on Tumor_Size_mm

H0: There is no significant interaction between Smoking_History and Physical_Activity that affects Tumor_Size_mm.

For the purposes of this test we will define our significance level as 0.05.

This is our test and its results:

```
anova_y1 <- aov(y1 ~ x4 * x2, data = Dataset)
summary(anova_y1)
```

```
##                 Df   Sum Sq Mean Sq  F value Pr(>F)
## x4               1   160297  160297 1012.872 <2e-16 ***
## x2               2    44588   22294  140.870 <2e-16 ***
## x4:x2            2      306     153    0.965  0.381
## Residuals   167491 26507170     158
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We visualize the results as followed:

```
par(mfrow=c(1,2))
boxplot(split(y1,x4), col="green", xlab = "Smoking History", ylab="Tumor Size(mm)",
main="Smoking History vs Tumor Size")
boxplot(split(y1,x2), col = "red", xlab = "Physical Activity", ylab="Tumor Size(mm)",
main="Physical Activity vs Tumor Size")
```

I have used box plots to visualize both variables. We see that for Smoking_History that skew the results to the right. We also see that the difference in the medians of Tumor_Size_mm between our three groups is slightly different. These two factors could indicate there is a significant difference in the means between both groups. Meanwhile for Physical_Activity the box plots appears to have medians that look similar to each other, however once again we see that there outliers skewing the data. This could cause a statistical significant difference in means between the two groups. This is backed up by our ANOVA test. Based on P values from the ANOVA test, we reject the null hypothesis for Smoking_History and Physical_Activity, since they are less than 0.05. We have evidence to suggest that on their own, both variables have a significant effect on Tumor_Size_mm. However, when looking at the interaction between both variables, we fail to reject the null hypothesis. The p-value is greater than 0.05, meaning there is not a significant effect on Tumor_Size_mm from an interaction of Smoking_History and Physical_Activity.

Part 2: Predictive Model

Next we will create a predictive model were we will try to predict Tumor_Size_mm using demographic data such as Cancer_Stage, Age. To do this we will create a multiple linear regression model. But first lets create two scatter plots, one for each of our variables to visualize them separately.

```
par(mfrow=c(1,2))
plot(x1,y1,pch=16, col="blue", names=c("Local", "Meta", "Regional"), xlab="Cancer Stage", ylab="Tumor Si
plot(x6,y1,pch=13, xlab="Age", ylab="Tumor Size(mm)", main="Tumor Size vs Age", col="orange")
```

Some interpretation we can get from our first plot is that the median of regional Tumor_Size_mm is bigger than the metastatic or local Cancer_Stage. This is contrary to science as a metastatic Cancer_Stage

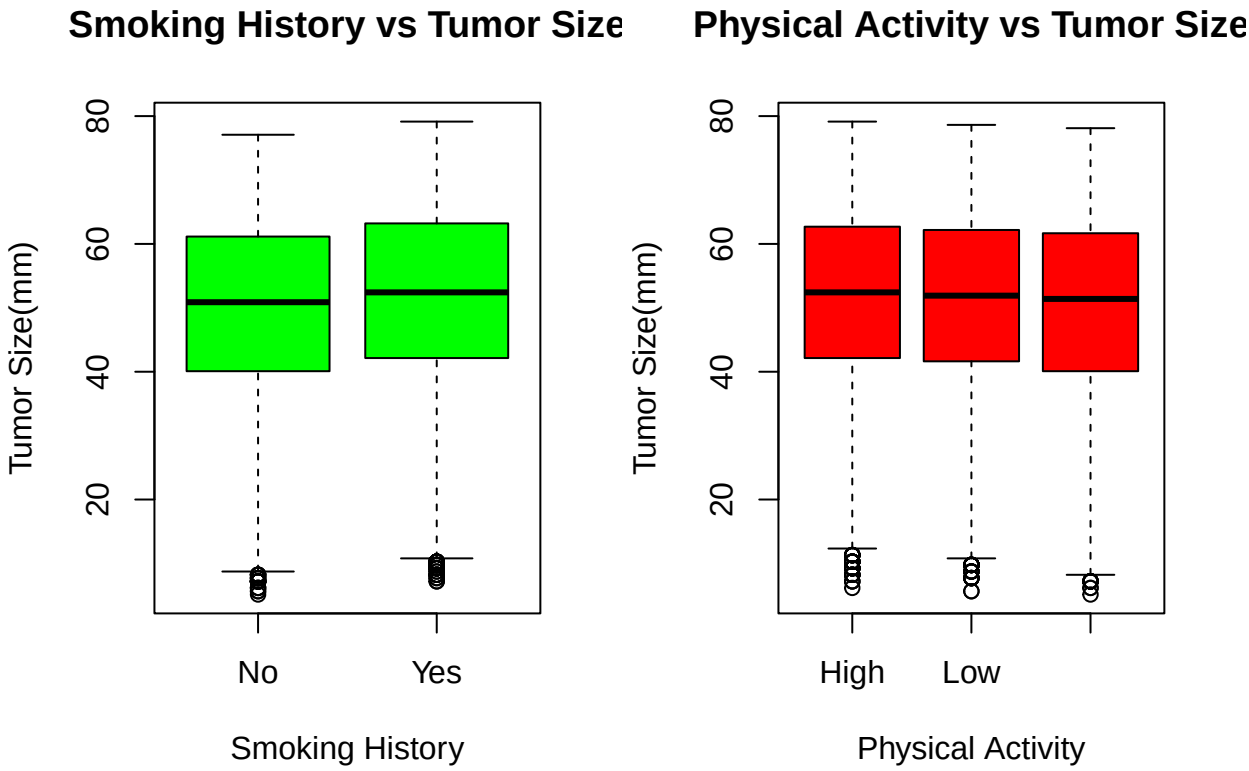

Figure 1: Left: Boxplot comparing Tumor Size and Smoking History, Right: Scatterplot comparing Tumor Size and Age

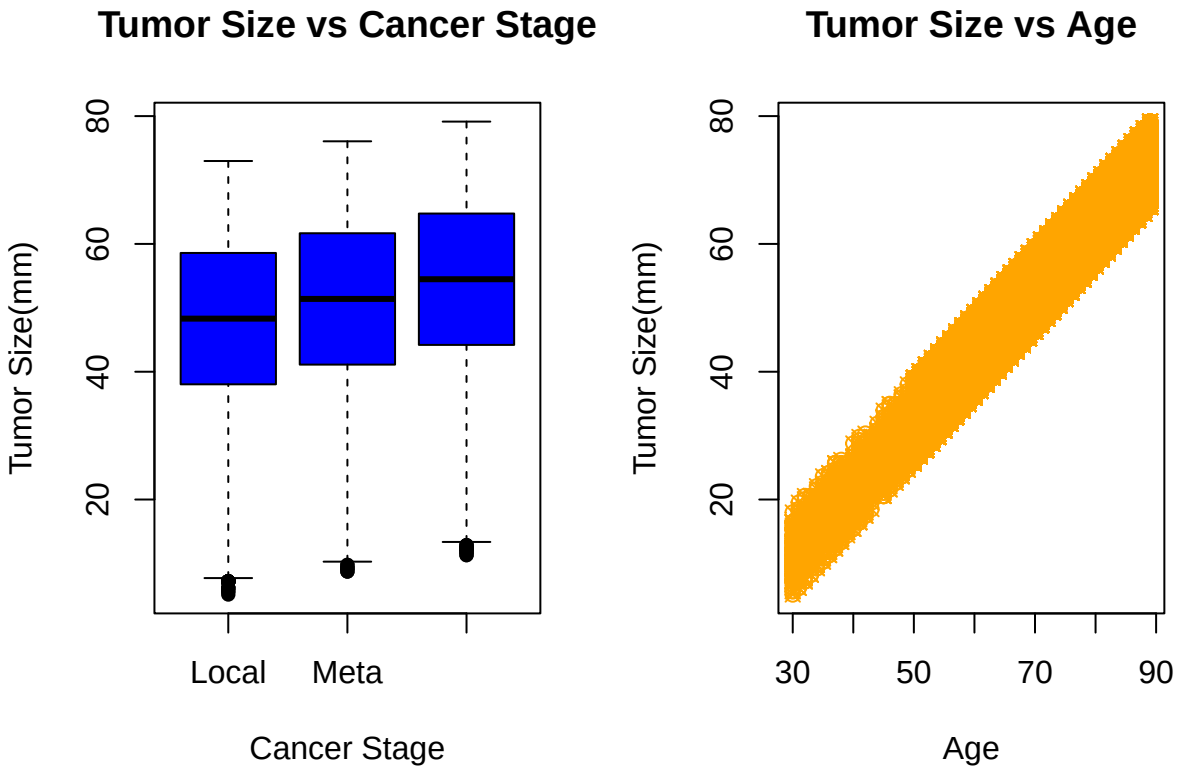

Figure 2: Left: Boxplot comparing Tumor Size and Cancer Stage, Right: Scatterplot comparing Tumor Size and Age

should have the largest Tumor_Size_mm, not regional. But what we do see are some outliers are causing Tumor_Size_mmdata to skew left for all three. So there is a possibility they have a larger affect for the regional group. For our second plot, we see that Age and Tumor_Size_mm are strongly correlated, as it is almost a perfect linear line. These two metrics will help us build our multiple linear regression to predict Tumor_Size_mm. Now we can build our multiple linear regression model as follows:

```
model <- lm(y1 ~x1+x6, data = Dataset)
summary(model)
```

```
##
## Call:
## lm(formula = y1 ~ x1 + x6, data = Dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9448 -0.8823 -0.3546  1.1874  4.2791
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.279e+01  2.610e-02  -873.0   <2e-16 ***
## x1Metastatic  3.097e+00  1.181e-02   262.3   <2e-16 ***
## x1Regional    6.176e+00  9.664e-03   639.1   <2e-16 ***
## x6            1.028e+00  3.637e-04  2826.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.767 on 167493 degrees of freedom
## Multiple R-squared:  0.9804, Adjusted R-squared:  0.9804
## F-statistic: 2.795e+06 on 3 and 167493 DF,  p-value: < 2.2e-16
```

If we look at our multiple r^2 value, we see value of around 0.98. This means that our model takes into account 98% percent of the variance in Tumor_Size_mm. This means that our model is pretty accurate in predicting Tumor_Size_mm. This means that when we take into account Cancer_Stage and Physical_Activity, these two indicators combined are very good predictors for Tumor_Size_mm.

Part 3: The Joe Question

Our next problem involves a scenario about a 40 year old man named Joe. Joe smokes, drinks, lives in a urban area, and does not exercise. As such he is at risk for colorectal_cancer. We want to figure out which variable can we change to reduce Tumor_Size_mm. To do this we can create a linear regression model of these variables.

```
Joe <- lm(y1 ~ x2+x3+x4+x5+x6, data = Dataset)
summary(Joe)
```

```
##
## Call:
## lm(formula = y1 ~ x2 + x3 + x4 + x5 + x6, data = Dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -3.15425 -3.07668 -0.00313  3.07025  3.13562
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.951e+01  4.336e-02 -449.93   <2e-16 ***
## x2Low       -4.923e-01  1.739e-02  -28.31   <2e-16 ***
## x2Moderate  -1.019e+00  1.628e-02  -62.60   <2e-16 ***
## x3Yes        2.093e+00  1.347e-02  155.41   <2e-16 ***
## x4Yes        2.056e+00  1.373e-02  149.76   <2e-16 ***
## x5Urban     -2.041e+00  1.470e-02 -138.77   <2e-16 ***
## x6           1.027e+00  5.671e-04 1810.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.756 on 167490 degrees of freedom
## Multiple R-squared:  0.9524, Adjusted R-squared:  0.9524
## F-statistic: 5.584e+05 on 6 and 167490 DF,  p-value: < 2.2e-16
```

Let's visualize this:

```
par(mfrow=c(1,2))
coefficients <- coef(Joe)
coefficients <- coefficients[names(coefficients) != "(Intercept)"]
coefficients <- coefficients[names(coefficients) != "x2Moderate"]
barplot(coefficients, main = "Linear Regression Coefficients", ylab= "Coefficient Estimate", las=2, yli
plot(x3,y1,pch=5, xlab="Alcohol Consumption", ylab="Tumor Size(mm)", main="Tumor Size vs Age", col="ora
```

Our model shows that Alcohol_Consumption has the slightly has the largest impact on Tumor_Size_mm. We found this by looking at the coefficients of the linear regression model. This means that if Joe wants to change exactly one factor to reduce his Tumor_Size_mm, he should give up Alcohol_Consumption.This is backed up by our bar plot, which shows a peak at Alcohol_Consumption. That being said, Smoking_History and Urban_or_Rural do still have a significant impact on Tumor_Size_mm. The coefficients are slightly lower than Alcohol_Consumption, so in reality he should those into account. However the single biggest factor would be Alcohol_Consumption. We can also see that our bar plot shows that the median Tumor_Size_mm is higher when Alcohol_Consumption is yes, and with outliers skewing the data, that could be even a larger gap. In conclusion, Joe should quit Alcohol_Consumption if he has the choice to change one habit.

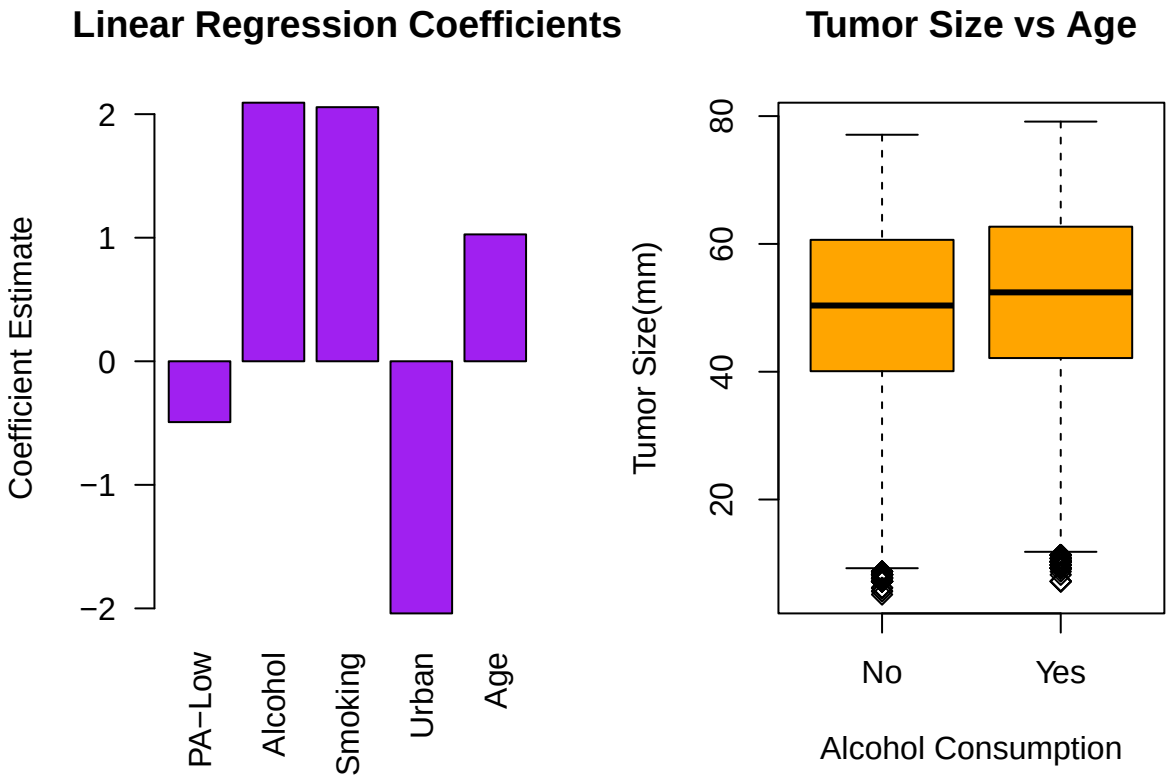

Figure 3: Right: Barplot of Linear Regression Coefficients of each factor for Low Physical Activity, Alchol Consumption, Smoking, Urban living, and Age, Right: Boxplot of Alcohol Consumption vs Tumor Size