Summative 1

2024-10-23

R Markdown

```
#Removing all objects from the current environment
rm(list = ls())

#Installing packages
if(!require(ggplot2)) install.packages("ggplot2")
```

```
## Loading required package: ggplot2
```

```
#Loading in the packages
library(ggplot2)
library(car)
```

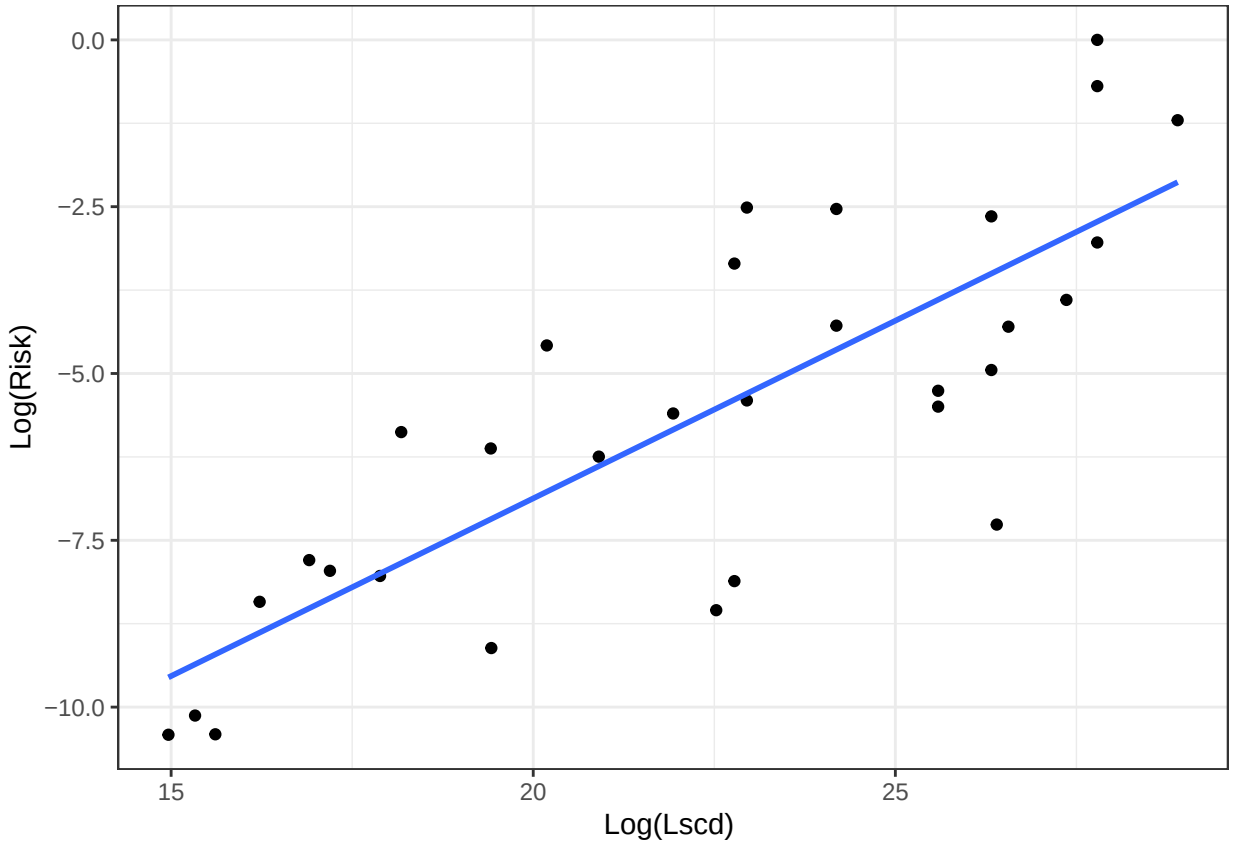
```
## Loading required package: carData
```

```
#Reading in the datset
tumour_data <- read.csv("/Users/tarunrajan/Library/Mobile Documents/com~apple~CloudDocs/ST300/Summative

#Fitting a Simple Linear Regression model of Log(Risk) against Log(Lscd)
x_1 <- log(c(tumour_data$Lscd))
y_1 <- log(c(tumour_data$Risk))

lm1 <- lm(y_1~x_1)

summary(lm1)
```

```
##
## Call:
## lm(formula = y_1 ~ x_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8013 -1.0721  0.1434  0.9945  2.7873
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.52379    1.66436  -10.53 2.02e-11 ***
## x_1           0.53260    0.07316    7.28 5.12e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.725 on 29 degrees of freedom
## Multiple R-squared:  0.6463, Adjusted R-squared:  0.6341
## F-statistic:    53 on 1 and 29 DF,  p-value: 5.117e-08
```

```r
#Plotting the estimated regression line
plot_data <- data.frame(x_1,y_1)

ggplot(data = plot_data, aes(x = x_1,y = y_1)) +
geom_point() +
theme_bw() +
geom_smooth(method = "lm", se = FALSE) +
labs(x = "Log(Lscd) ", y = "Log(Risk)")
```

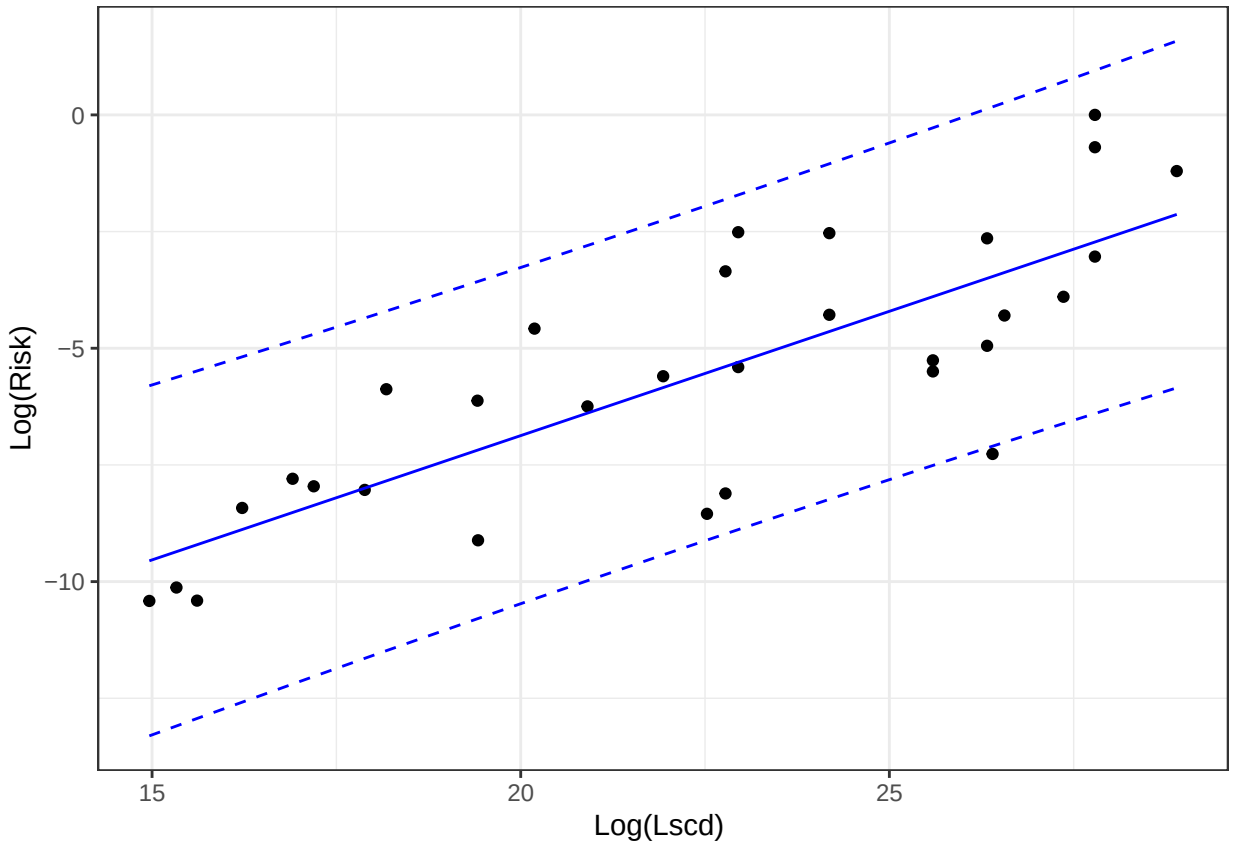
```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
#Adding 95% prediction bands
prediction_intervals <- predict(lm1, interval = "prediction", level = 0.95)
```

```
## Warning in predict.lm(lm1, interval = "prediction", level = 0.95): predictions on current data refer
```

```r
tumour_data_new <- cbind(plot_data,prediction_intervals)

ggplot(tumour_data_new, aes(x = x_1)) +
  geom_point(aes(y = y_1)) +
  geom_line(aes(y = fit), col = "blue") +

  geom_line(aes(y= upr), col = "blue", linetype = "dashed") +

  geom_line(aes(y= lwr), col = "blue",linetype = "dashed") +

  theme_bw() +
  labs(x = "Log(Lscd) ", y = "Log(Risk)")
```




```
#Interpretation at a Lscd 10^10
Lscd_val <- log(10^10)
new_data <- data.frame(x_1 = Lscd_val)
predict_value <- predict(lm1, newdata = new_data, interval = "prediction", level = 0.95)
print(predict_value)
```

```
##         fit       lwr       upr
## 1 -5.260253 -8.845771 -1.674736
```

```
#The prediction interval is [-8.845771, -1.674736]
# If we were to randomly sample a new tissue type with Lscd =10^10 then  would expect the log(Risk) for


# Hypothesis test

#Null Hypothesis : The slope of regression line = 0
#Alternative Hypothesis : The slope of regression line != 0

p_value = summary(lm1)$coefficients[2,4]
alpha = 0.05

if (p_value < alpha) {
  print("There IS enough evidence to reject the Null Hypothesis")

} else {
  print("There is NOT enough evidence to reject the Null Hypothesis")
}
```

```
## [1] "There IS enough evidence to reject the Null Hypothesis"
```

```
#Conclusion
"p_value = 5.117112e-08"
```

```
## [1] "p_value = 5.117112e-08"
```

```
# 95% confidence interval for slope
confint(lm1, 'x_1' , level = 0.95)
```

```
##         2.5 %    97.5 %
## x_1 0.3829708 0.6822268
```

```
confidence_interval <-c(0.3829708, 0.6822268) #This is the 95% confidence interval for the slope of the

#Interpretation of the confidence interval

"For a 1 unit increase in X (log of the lifetime stem cell divisions), we are 95% confident that the tr
```

```
## [1] "For a 1 unit increase in X (log of the lifetime stem cell divisions), we are 95% confident that
```

```
# Seperating into classification of tumour
D_tumours <- tumour_data[(tumour_data$Cluster  == "Deterministic"),]
R_tumours <- tumour_data[(tumour_data$Cluster  == "Replicative"),]

#Fitting linear models for D - tumours
x_d <- log(D_tumours$Lscd)
y_d <- log(D_tumours$Risk)
lm_d <- lm(y_d ~ x_d, D_tumours)

summary(lm_d)
```

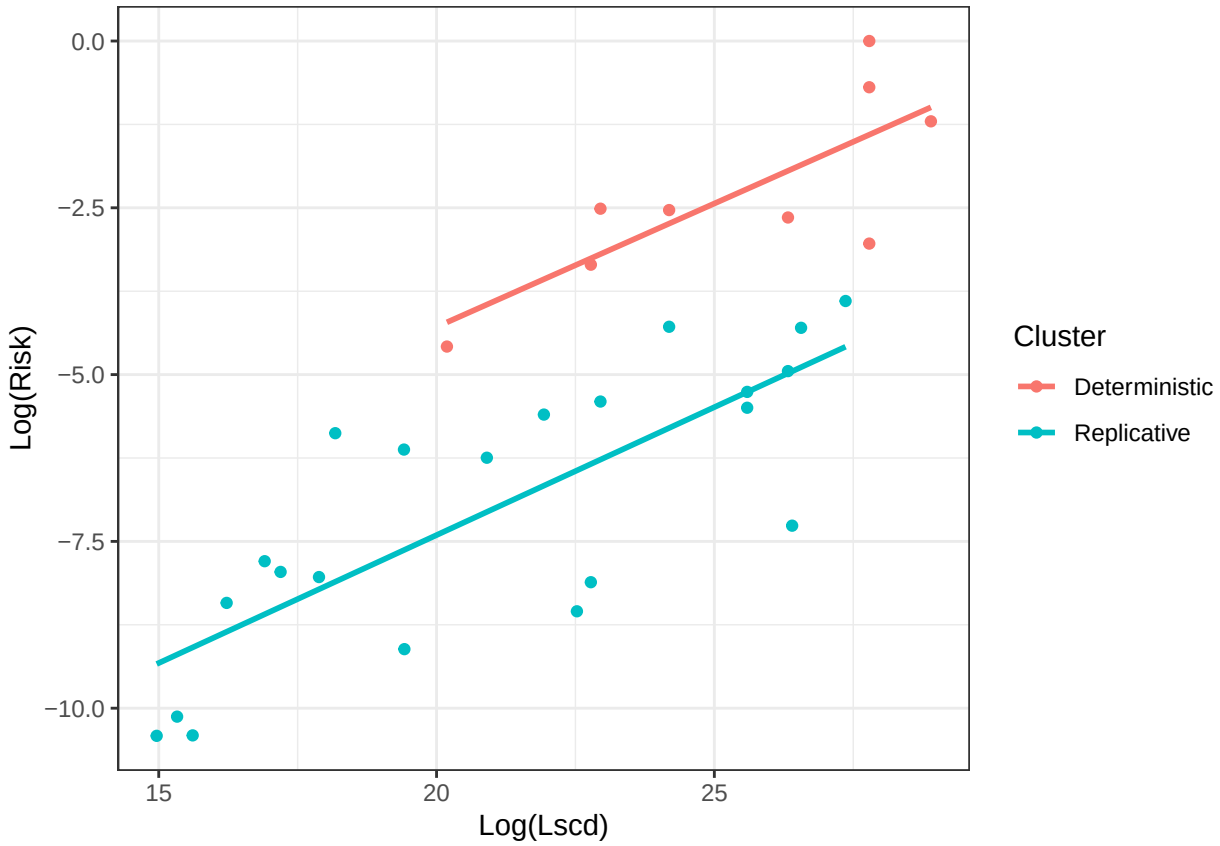
```
##
## Call:
## lm(formula = y_d ~ x_d, data = D_tumours)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63211 -0.36336 -0.09444  0.68079  1.40445
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.6837     2.8687  -4.073  0.00473 **
## x_d           0.3699     0.1122   3.297  0.01318 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9503 on 7 degrees of freedom
## Multiple R-squared:  0.6083, Adjusted R-squared:  0.5523
## F-statistic: 10.87 on 1 and 7 DF,  p-value: 0.01318
```

```
#Fitting linear models for R - tumours
x_r <- log(R_tumours$Lscd)
y_r <- log(R_tumours$Risk)
lm_r <- lm(y_r ~ x_r, R_tumours)

summary(lm_r)
```

```
##
## Call:
## lm(formula = y_r ~ x_r, data = R_tumours)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3133 -1.0401  0.3074  0.8080  2.2266
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.07606    1.44401 -10.440 1.53e-09 ***
## x_r           0.38352    0.06719   5.708 1.38e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.285 on 20 degrees of freedom
## Multiple R-squared:  0.6196, Adjusted R-squared:  0.6006
## F-statistic: 32.58 on 1 and 20 DF,  p-value: 1.382e-05
```

```r
#Combined Plot
combined_data <- rbind(
  data.frame(x = x_d, y = y_d, Cluster = "Deterministic"),
  data.frame(x = x_r, y = y_r, Cluster = "Replicative")
)

ggplot(combined_data, aes(x = x, y = y, color = Cluster)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw() +
  labs(x = "Log(Lscd)", y = "Log(Risk)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#F - test
combined_data$Cluster <- factor(combined_data$Cluster, levels = c("Deterministic", "Replicative")) #Con
reduced_model <- lm(y_1 ~ x_1, data = plot_data)
full_model <- lm(y_1 ~ x_1 + Cluster, data = combined_data)
anova_result <- anova(reduced_model, full_model)
print(anova_result)
```

```
## Analysis of Variance Table
##
## Model 1: y_1 ~ x_1
## Model 2: y_1 ~ x_1 + Cluster
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     29 86.275
## 2     28 85.563  1   0.71171 0.2329 0.6331
```

```
# p-value = 0.63316
```