````
---
title: "Homework 1"
author: "Your Name"
format: pdf
  html:
    self-contained: true
    embed-resources: true
editor: visual
execute:
  echo: TRUE
  include: TRUE
---

For the problems in which calculations are needed, please include your R code with your answers, otherwise you will not be given full credit. Please upload your assignment by Wednesday, September 10, 6 pm in a html/pdf file to Sakai.

1.  In a simple linear regression problem where $n=30$, we obtain $$\sum_{i=1}^n x_i = 60, \sum_{i=1}^n y_i = 90, \sum_{i=1}^n x_i^2 = 240, \sum_{i=1}^n y_i^2 = 540, \sum_{i=1}^n x_iy_i = 234.$$

-   Calculate $\hat{\beta}_0$ and $\hat{\beta}_1$.

```{r}
n <- 30
sum_x <- 60
sum_y <- 90
sum_x2 <- 240
sum_y2<- 540
sum_xy <- 234
xbar <- sum_x/n
ybar <- sum_y/n
beta1 <- (sum_xy - n*xbar*ybar) / (sum_x2 - n*xbar^2)
beta0 <- ybar - beta1*xbar

beta1
beta0

```

-   $\hat{\beta_1}$ = 0.45 & $\hat{\beta_0}$ = 2.1
-   Calculate $\hat{\sigma}^2$ under both maximum likelihood and least squares methods.

```{r}
#Residual sum of squares
RSS<- sum_y2 +n*beta0^2 +beta1^2*sum_x2 - 2*beta0*sum_y -2*beta1*sum_xy +
2*beta0*beta1*sum_x
#maximum likelihood
sigma2_mle <- RSS / n
sigma2_mle
#least squares (aka unbiased variance)
sigma2_ls <- RSS / (n-2)
sigma2_ls
```

-   $\hat{o^2} MLE = 8.190$
-   $\hat{o^2} LSE = 8.775$
-   Calculate a test statistic $t$ and p-value for a hypothesis test for a significant linear relationship between $X$ and $Y$.
-   $t = 1.664$ & $p = 0.1072$

```{r}
#standard error of beta1
SS_x <- sum_x2 - n*xbar^2
````
````

```
SE_B1 <- sqrt(sigma2_ls/ SS_x)
#t-statistic
t_stat <- beta1 / SE_B1
t_stat
#pvalue (two sided)
p_value <- 2* pt(-abs(t_stat),n-2)
p_value
#we cannot reject the null hypothesis
```

2. Consider the dataset lifeexp in the Data folder attached to this assignment. Suppose we wish to predict life expectancy using different covariates (independent variables) in the dataset.

- Draw a scatterplot of life expectancy by income composition of resources (ICOR). Comment on the scatterplot.
- The scatterplot appears to show a relatively linear relationship between ICOR and Life Expectancy with a minimum Life Expectancy of 50 years.

```{r}
library(tidyverse)
lifeexp <- read.csv("~/stat408/Homework1/Data/lifeexp.csv")
ggplot(data = lifeexp, aes(x=Income.composition.of.resources,y=Life.expectancy)) +
geom_point()
```

- Using R, find the equation of the least squares regression line predicting life expectancy by ICOR.
- Our equation for least squares regression line is
- $\hat{Life.expectancy} = 39.264 + 46.906_{ICOR}$

```{r}
#fit for the regression model
mod <- lm(Life.expectancy ~ Income.composition.of.resources,data = lifeexp)
#view coefficients
summary(mod)$coefficients
```

- What do we expect the life expectancy to be for a country with a 0.7 ICOR? 0.3 ICOR?
- We expect the life expectancy for a country with a 0.7 ICOR to be 72.10 years
- We expect the life expectancy for a country with a 0.3 ICOR to be 53.34 years

```{r}
#data frame with values we want to check
newdata <- data.frame(Income.composition.of.resources = c(0.7,0.3))
#rename to ICOR bc its too long to type out
names(newdata)[names(newdata) == "ICOR"] <- "Income.composition.of.resources"
#predict life expectancy
pred <- predict(mod,newdata)
#print
pred
```

- Calculate and interpret a 95% confidence interval for the slope of the least squares regression line.
- We are 95% confident that the slope of the least squares regression line is between 43.563 and 50.250

```{r}
confint(mod, level = 0.95)
```

- Perform a hypothesis test for significance of a linear relationship between ICOR and life expectancy. Be sure to include all pieces of information needed to conduct a formal

hypothesis test.
- $H_0: {\beta_1} = 0$
- $H_a: {\beta_1} \neq 0$
- $p_value = 1.097e-64 < \alpha = 0.05$
- therefore we reject the null hypothesis and conclude there is a significant linear relationship between ICOR and life expectancy

```{r}
summary(mod)$coefficients
```

- Calculate and interpret a 90% prediction interval for the life expectancy of a randomly selected country with an ICOR of 0.79.
- We are 90% confident that the average life expectancy of any country with an Income Composition of Resources of 0.79 is between 70.743 and 81.898 years

```{r}
newdata <- data.frame(Income.composition.of.resources = 0.79)
prediction_int <- predict(mod,newdata,interval = "prediction", level = 0.90)
prediction_int
```

- Determine if the assumptions of homoscedasticity and normally distributed residuals are violated.
- Breusch-Pagan Test
    - $p-value = 6.245e-4 < \alpha = 0.05$
    - therefore homoscedasticity is violated
- Residual Plot
    - the residuals vs fitted plot shows a funnel growing smaller toward a higher age, therefore normally distributed are violated
- Q-Q Plot
    - the assumption of normality is not violated as there are only a small amount of deviations on each side of the plot.
- We cannot rely on the results as more than one test showed a violation of either homoscedasticity or normally distributed results.

```{r}
#load libraries
library(lmtest)
#check for homoscedasticity
bp_test <- bptest(mod)
p_value <- bptest(mod)$pvalue
bp_test
#check for normally distributed results with Q-Q plot
plot(mod,1)
plot(mod,2)
```