[STAT 4400] HW-1

Michael Ghattas

1/17/2022

# Question-1

### (a)

### (b)

### (c)

### (d)

~ N(0, )

y = c(1 , 2 , 3)  
x = c(2 , 0 , 4)  
alpha = 0.5  
n = length(x)  
  
beta.0.hat = 2  
beta.1.hat = 0  
sigma2.hat = 2  
std.t = dt(x , (n - 2))  
t.star = qt((1 - alpha / 2) , df = (n - 2))  
  
y.hat = beta.0.hat + (beta.1.hat \* x)  
eps.hat = y - y.hat  
SSE = sum(eps.hat \*\* 2)  
sigma2.hat = SSE / (n - 2)  
tau.hat = sqrt(sigma2.hat / sum((x - mean(x)) \*\* 2))  
t.hat = beta.1.hat / tau.hat  
  
t.star

## [1] 1

t.hat

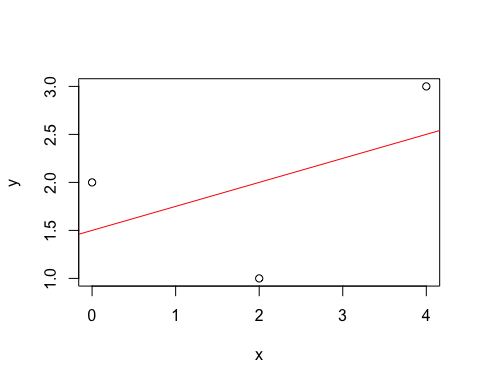
## [1] 0

print("(|t| < t\*) Thus we fail to reject the the null-hypothesis")

## [1] "(|t| < t\*) Thus we fail to reject the the null-hypothesis"

### (e)

lmod = lm(y ~ x)  
plot(x, y)  
abline(lmod, col = "red")



# Question-2

### (a)

### (b)

### (c)

##### I would recommend the solution from part (a), as it has an increasing function that exhibits a positive correlation.

# Question-3

### (a)

set.seed(123)  
  
var1 = rnorm(1000, mean = 0, sd = 1)  
var2 = rnorm(1000, mean = 0, sd = 1)  
  
lmod1 = lm(var1 ~ var2)  
summary(lmod1)

##   
## Call:  
## lm(formula = var1 ~ var2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.7168 -0.6290 -0.0060 0.6451 3.2383   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.01252 0.03129 0.400 0.68909   
## var2 0.08494 0.03097 2.742 0.00621 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9885 on 998 degrees of freedom  
## Multiple R-squared: 0.007479, Adjusted R-squared: 0.006484   
## F-statistic: 7.52 on 1 and 998 DF, p-value: 0.006211

print("Yes, based on the p-value generated we can conclude that the slope coefficient var2 is statistically significant.")

## [1] "Yes, based on the p-value generated we can conclude that the slope coefficient var2 is statistically significant."

### (b)

set.seed(321)  
  
z.scores <- rep (NA, 100)  
  
for (k in 1:100)   
{  
 var1 <- rnorm (1000 ,0 ,1)   
 var2 <- rnorm (1000 ,0 ,1)  
 fit <- lm (var2 ~ var1)  
 z.scores[k] <- coef(fit )[2] / summary(fit)$coef[2,"Std. Error"]  
}  
  
alpha = .05  
cutoffn = qnorm((1 - alpha) / 2, lower.tail=TRUE)  
sum(abs(z.scores) > cutoffn)

## [1] 100

for (k in 1:100)   
{  
 result = sum(abs(z.scores) < 1.96)  
}  
  
result

## [1] 95

print("95 estimated slope coefficients are statistically significant at the α = .05 level of significance.")

## [1] "95 estimated slope coefficients are statistically significant at the α = .05 level of significance."

# Question-4

### (a)

library(haven)  
data <- read\_dta("/Users/Home/Documents/Michael\_Ghattas/School/CU\_Boulder/2022/Spring 2022/STAT - 4400/HW/1/child.iq.dta")  
head(data)

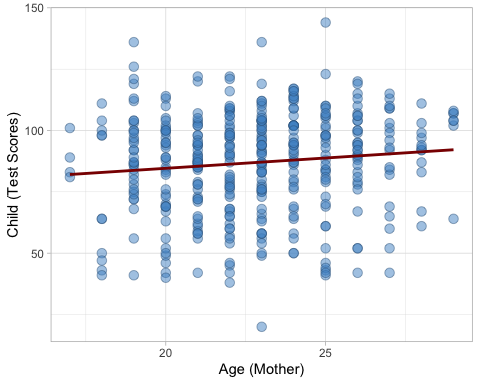
## # A tibble: 6 × 3  
## ppvt educ\_cat momage  
## <dbl> <dbl> <dbl>  
## 1 120 2 21  
## 2 89 1 17  
## 3 78 2 19  
## 4 42 1 20  
## 5 115 4 26  
## 6 97 1 20

lmod = lm(ppvt ~ momage, data = data)  
summary(lmod)

##   
## Call:  
## lm(formula = ppvt ~ momage, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -67.109 -11.798 2.971 14.860 55.210   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 67.7827 8.6880 7.802 5.42e-14 \*\*\*  
## momage 0.8403 0.3786 2.219 0.027 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 20.34 on 398 degrees of freedom  
## Multiple R-squared: 0.01223, Adjusted R-squared: 0.009743   
## F-statistic: 4.926 on 1 and 398 DF, p-value: 0.02702

library(ggplot2)  
ggplot(data, aes(momage, ppvt)) +  
geom\_point(shape = 21, color="steelblue4", fill="steelblue3", size = 3,  
alpha=0.5,show.legend = FALSE) +  
theme\_light() + xlab("Age (Mother)") + ylab("Child (Test Scores)") +  
geom\_smooth(method = lm, color="darkred", se=FALSE)

## `geom\_smooth()` using formula 'y ~ x'



print("Based on our summary and plots, it seems that the mother's age is a significant predictor, though not enough on its own to find a direct correlation. That being said, the plot clearly shows that the majority of the observations with a higher test score belong to the mothers in their late 20's. Thus mothers should give birth in their late 20s.")

## [1] "Based on our summary and plots, it seems that the mother's age is a significant predictor, though not enough on its own to find a direct correlation. That being said, the plot clearly shows that the majority of the observations with a higher test score belong to the mothers in their late 20's. Thus mothers should give birth in their late 20s."

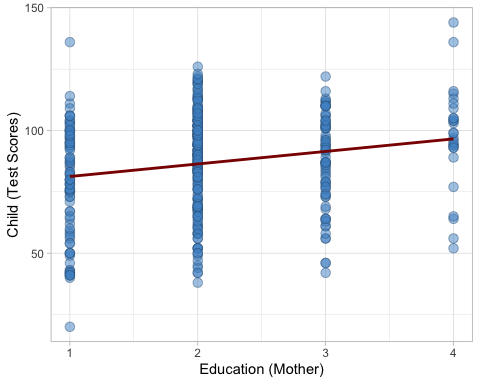
### (b)

lmod = lm(ppvt ~ momage + educ\_cat, data = data)  
summary(lmod)

##   
## Call:  
## lm(formula = ppvt ~ momage + educ\_cat, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61.763 -13.130 2.495 14.620 55.610   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 69.1554 8.5706 8.069 8.51e-15 \*\*\*  
## momage 0.3433 0.3981 0.862 0.389003   
## educ\_cat 4.7114 1.3165 3.579 0.000388 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 20.05 on 397 degrees of freedom  
## Multiple R-squared: 0.04309, Adjusted R-squared: 0.03827   
## F-statistic: 8.939 on 2 and 397 DF, p-value: 0.0001594

library(ggplot2)  
ggplot(data, aes(educ\_cat, ppvt)) +  
geom\_point(shape = 21, color="steelblue4", fill="steelblue3", size = 3,  
alpha=0.5,show.legend = FALSE) +  
theme\_light() + xlab("Education (Mother)") + ylab("Child (Test Scores)") +  
geom\_smooth(method = lm, color="darkred", se=FALSE)

## `geom\_smooth()` using formula 'y ~ x'



print("Based on our summary and plots, it seems that the mother's education is a strong and significant predictor. That being said, the plot clearly shows that the majority of the observations with a higher test score belong to the mothers having completed a high-school education.")

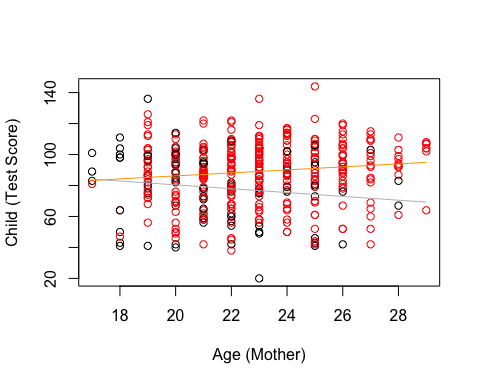
## [1] "Based on our summary and plots, it seems that the mother's education is a strong and significant predictor. That being said, the plot clearly shows that the majority of the observations with a higher test score belong to the mothers having completed a high-school education."

### (c)

data$mom.hs <- ifelse(data$educ\_cat >= 2, 1, 0)  
  
lmod <- lm(ppvt ~ (mom.hs \* momage), data = data)  
summary(lmod)

##   
## Call:  
## lm(formula = ppvt ~ (mom.hs \* momage), data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -56.696 -12.407 2.022 14.804 54.343   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 105.2202 17.6454 5.963 5.49e-09 \*\*\*  
## mom.hs -38.4088 20.2815 -1.894 0.0590 .   
## momage -1.2402 0.8113 -1.529 0.1271   
## mom.hs:momage 2.2097 0.9181 2.407 0.0165 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 19.85 on 396 degrees of freedom  
## Multiple R-squared: 0.06417, Adjusted R-squared: 0.05708   
## F-statistic: 9.051 on 3 and 396 DF, p-value: 8.276e-06

spread <- ifelse(data$mom.hs == 1, "red", "black")  
plot(data$momage, data$ppvt, xlab = "Age (Mother)", ylab = "Child (Test Score)", col = spread, pch = 1)  
curve(cbind(1, 1, x, 1 \* x) %\*% coef(lmod), add = TRUE, col = "orange") # Mother finished hs  
curve(cbind(1, 0, x, 0 \* x) %\*% coef(lmod), add = TRUE, col = "grey") # Mother did not finish hs

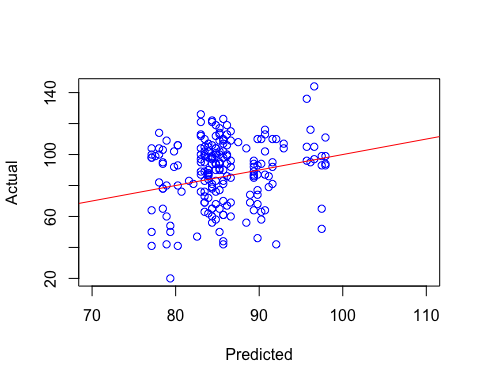


### (d)

lmod <- lm(ppvt ~ momage + educ\_cat, data = data[1:200, ])  
summary(lmod)

##   
## Call:  
## lm(formula = ppvt ~ momage + educ\_cat, data = data[1:200, ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -46.358 -12.967 2.866 14.435 58.428   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 63.6295 11.8202 5.383 2.07e-07 \*\*\*  
## momage 0.4473 0.5516 0.811 0.41836   
## educ\_cat 5.4434 1.8228 2.986 0.00318 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 19.58 on 197 degrees of freedom  
## Multiple R-squared: 0.06199, Adjusted R-squared: 0.05246   
## F-statistic: 6.509 on 2 and 197 DF, p-value: 0.001831

pmod <- predict(lmod, data[201:400, ])  
plot(pmod, data$ppvt[201:400], xlim = c(70, 110), xlab = "Predicted",   
 ylab = "Actual", col = "blue")  
abline(a = 0, b = 1, col = "red")



# Question-5

### (a)

library(faraway)  
data(prostate)  
head(prostate)

## lcavol lweight age lbph svi lcp gleason pgg45 lpsa  
## 1 -0.5798185 2.7695 50 -1.386294 0 -1.38629 6 0 -0.43078  
## 2 -0.9942523 3.3196 58 -1.386294 0 -1.38629 6 0 -0.16252  
## 3 -0.5108256 2.6912 74 -1.386294 0 -1.38629 7 20 -0.16252  
## 4 -1.2039728 3.2828 58 -1.386294 0 -1.38629 6 0 -0.16252  
## 5 0.7514161 3.4324 62 -1.386294 0 -1.38629 6 0 0.37156  
## 6 -1.0498221 3.2288 50 -1.386294 0 -1.38629 6 0 0.76547

lmod = lm(log(lpsa) ~ log(lcavol) + ., data = prostate)

## Warning in log(lpsa): NaNs produced

## Warning in log(lcavol): NaNs produced

summary(lmod)

##   
## Call:  
## lm(formula = log(lpsa) ~ log(lcavol) + ., data = prostate)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.42557 -0.12542 0.02419 0.19314 0.51973   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.228166 0.665962 -0.343 0.73286   
## log(lcavol) -0.073779 0.082178 -0.898 0.37221   
## lcavol 0.298653 0.093642 3.189 0.00209 \*\*  
## lweight 0.132774 0.086027 1.543 0.12700   
## age -0.008582 0.005988 -1.433 0.15596   
## lbph 0.068535 0.030029 2.282 0.02535 \*   
## svi 0.247254 0.116662 2.119 0.03741 \*   
## lcp -0.049417 0.044109 -1.120 0.26619   
## gleason 0.086846 0.075419 1.152 0.25323   
## pgg45 0.001981 0.002088 0.948 0.34605   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3323 on 74 degrees of freedom  
## (13 observations deleted due to missingness)  
## Multiple R-squared: 0.4948, Adjusted R-squared: 0.4334   
## F-statistic: 8.053 on 9 and 74 DF, p-value: 3.254e-08

confint(lmod, 'log(lcavol)', level = 0.95)

## 2.5 % 97.5 %  
## log(lcavol) -0.2375227 0.08996445

### (b)

lmod = lm(log(lpsa) ~ log(lcavol) + lcavol + lbph + svi, data = prostate)

## Warning in log(lpsa): NaNs produced

## Warning in log(lcavol): NaNs produced

summary(lmod)

##   
## Call:  
## lm(formula = log(lpsa) ~ log(lcavol) + lcavol + lbph + svi, data = prostate)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.48196 -0.12939 0.05611 0.22173 0.48638   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.35496 0.12566 2.825 0.005987 \*\*   
## log(lcavol) -0.07497 0.08041 -0.932 0.354038   
## lcavol 0.29829 0.08713 3.423 0.000983 \*\*\*  
## lbph 0.07834 0.02645 2.961 0.004043 \*\*   
## svi 0.24564 0.10428 2.356 0.020976 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3384 on 79 degrees of freedom  
## (13 observations deleted due to missingness)  
## Multiple R-squared: 0.4407, Adjusted R-squared: 0.4124   
## F-statistic: 15.56 on 4 and 79 DF, p-value: 1.98e-09

confint(lmod, 'log(lcavol)', level = 0.95)

## 2.5 % 97.5 %  
## log(lcavol) -0.2350253 0.08509177

### (c)

##### The model from part (a) has a slightly better fit, as the R^2 value is slightly higher, indicating a better fitted model.