Chimdi Homework 1

Chimdi Chikezie

4/7/2023

```
library(ggplot2)
library(Rmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: plyr
```

```
library(graphics)
library(Sleuth3)
```

```
if(! (packageVersion("ggplot2") >= "2.0.0")) {
 stop("This version of stat_qqline require ggplot2 version 2.0.0 or greater")
}
```

Code drawing qqplot and qqline together in ggplot2, you do not need to understand this code

```
StatQQLine <- ggproto("StatQQLine", Stat,
 compute_group = function(data, scales, distribution = qnorm, dparams = list()) {
 data <- remove_missing(data, na.rm = TRUE, "sample", name = "stat_qqline")
 y <- quantile(data$sample, c(0.25, 0.75))
 x <- do.call(distribution, c(list(p = c(0.25, 0.75)), dparams))
 slope <- diff(y)/diff(x)
 int <- y[1L] - slope * x[1L]
 data.frame(slope = slope, intercept = int)
 },
 required_aes = c("sample")
)

stat_qqline <- function(mapping = NULL, data = NULL, geom = "abline",
 position = "identity", na.rm = FALSE, show.legend = NA,
 distribution = qnorm, dparams = list(),
 inherit.aes = TRUE, ...) {
 layer(
 stat = StatQQLine, data = data, mapping = mapping, geom = geom,
 position = position, show.legend = show.legend, inherit.aes = inherit.aes,
 params = list(na.rm = na.rm, distribution = distribution, dparams = dparams, ...)
 )
}
```

Question 1 Problem 1 (Problem 9.22 in textbook) Mammal Lifespans and Kleiber's Law. Kleiber's law states that the metabolic rate of an animal species, on average, is proportional to its mass raised to the power of 3/4. The Exercise 8.26 data set contains the mass (in kilograms), average basal metabolic rate (in kilojoules per day), and lifespan (in years) for 95 mammal species. The data is from A. T. Atanasov, "The Linear Allo-metric Relationship Between Total Metabolic Energy per Life Span and Body Mass of Mammals," Biosystems 90 (2007): 224-33.

(a) Draw scatterplots of metabolism versus mass, lifespan versus mass, and lifespan versus metabolism. Alternatively, you can draw a scatterplot matrix too

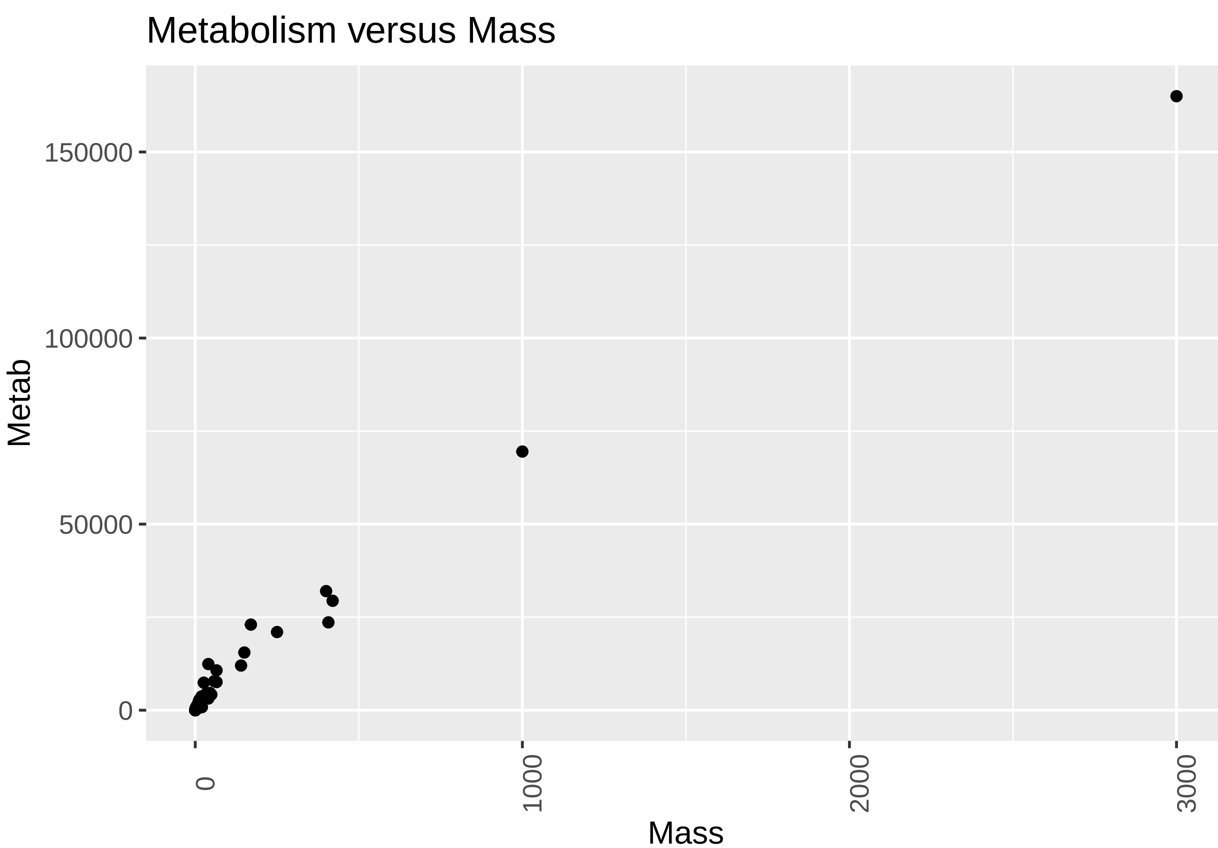
Loading Data

```
#?ex0826
HW1Data <- ex0826
#View(HW1Data)
```

Plotting the Scatterplot metabolism versus mass

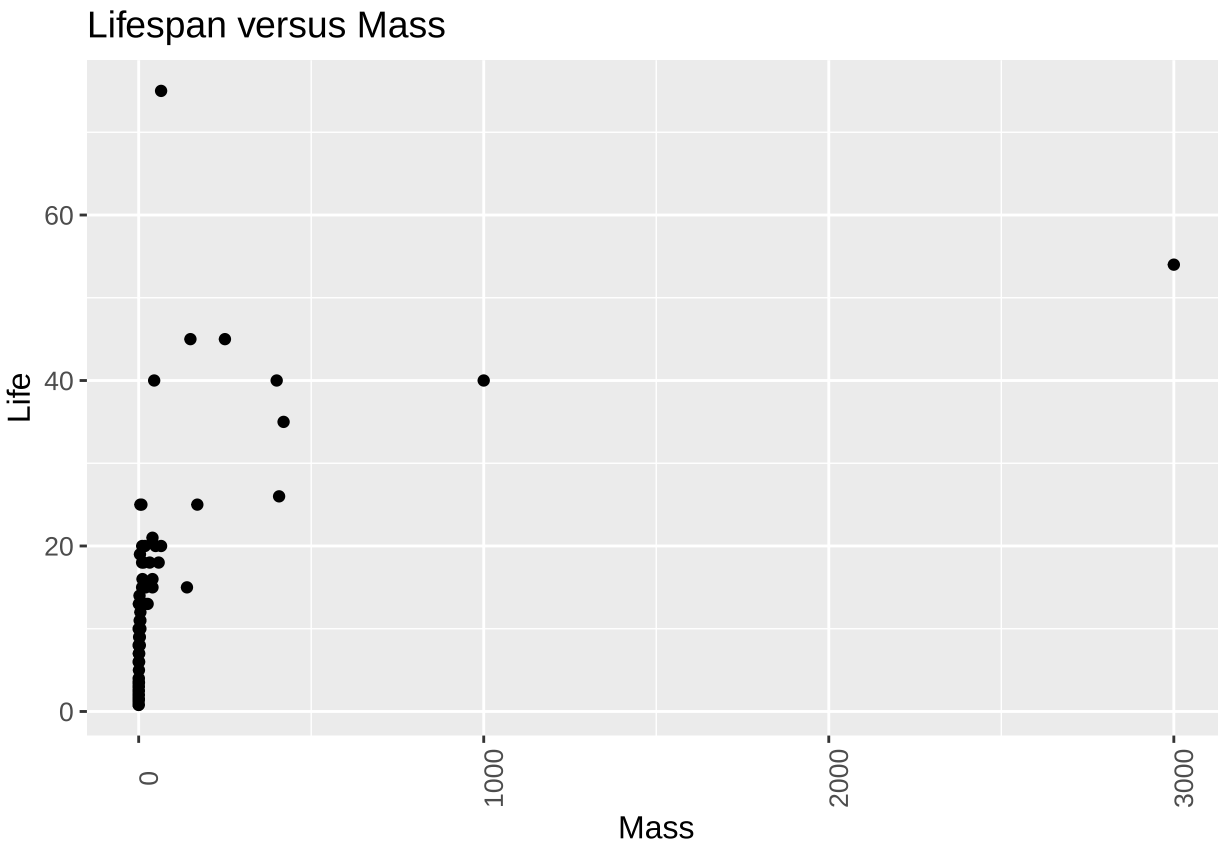
```
qplot(Mass, Metab, data = HW1Data) +  ggtitle('Metabolism versus Mass') +
  theme(text = element_text(size=12),
        axis.text.x = element_text(angle=90, vjust=1))
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Plot the Scatterplot lifespan versus mass

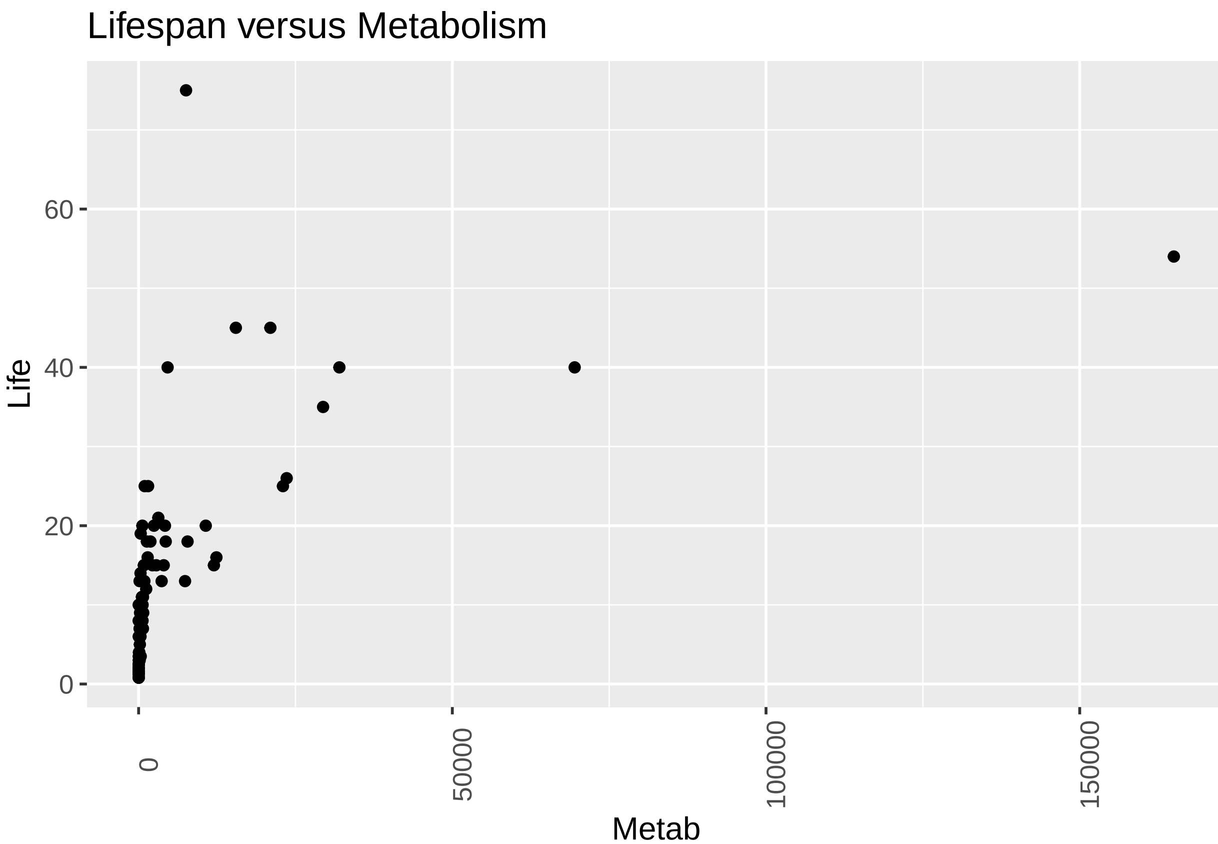
```r
qplot(Mass, Life, data = HW1Data) +  ggtitle('Lifespan versus Mass') +
  theme(text = element_text(size=12),
        axis.text.x = element_text(angle=90, vjust=1))
```




```r
#+ geom_smooth(method='lm', se = FALSE)
```

Plot the Scatterplot lifespan versus metabolism

```r
qplot(Metab, Life, data = HW1Data) +  ggtitle('Lifespan versus Metabolism') +
  theme(text = element_text(size=12),
        axis.text.x = element_text(angle=90, vjust=1))
```



(b) Obtain the least squares fit to the linear regression of metabolism on mass, metabolism on mass to the power of 3/4, and lifespan on mass, separately as simple linear regression.

```
sprintf("Least squares fit to the linear regression of metabolism on mass")
```

```
## [1] "Least squares fit to the linear regression of metabolism on mass"
```

```
lin_mod1 = lm(Metab ~ Mass, data = HW1Data)
summary(lin_mod1)
```

```
##
## Call:
## lm(formula = Metab ~ Mass, data = HW1Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7283.5 -1179.0 -1034.3  -370.2 12102.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1203.2019   291.2699   4.131 7.89e-05 ***
## Mass          57.0268     0.8695  65.586  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2777 on 93 degrees of freedom
## Multiple R-squared:  0.9788, Adjusted R-squared:  0.9786
## F-statistic:  4302 on 1 and 93 DF,  p-value: < 2.2e-16
```

```r
sprintf("Least squares fit to the linear regression of metabolism on mass to the power of 3/4")
```

```
## [1] "Least squares fit to the linear regression of metabolism on mass to the power of 3/4"
```

```r
mass <- (HW1Data$Mass)^(3/4)
lin_mod2 = lm(Metab ~ mass, data = HW1Data)
summary(lin_mod2)
```

```
##
## Call:
## lm(formula = Metab ~ mass, data = HW1Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11712.7   -117.4    368.5    474.3   6598.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -481.346    213.467  -2.255   0.0265 *
## mass         395.016      4.299  91.895   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1992 on 93 degrees of freedom
## Multiple R-squared:  0.9891, Adjusted R-squared:  0.989
## F-statistic:  8445 on 1 and 93 DF,  p-value: < 2.2e-16
```

```r
sprintf("Least squares fit to the linear regression of lifespan on mass")
```

```
## [1] "Least squares fit to the linear regression of lifespan on mass"
```

```r
lin_mod3 = lm(Life ~ Mass, data = HW1Data)
summary(lin_mod3)
```

```
##
## Call:
## lm(formula = Life ~ Mass, data = HW1Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.464  -7.431  -2.929   3.856  62.791
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.919177   1.163746   9.383 4.14e-15 ***
## Mass         0.019848   0.003474   5.713 1.32e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.1 on 93 degrees of freedom
## Multiple R-squared:  0.2598, Adjusted R-squared:  0.2518
## F-statistic: 32.64 on 1 and 93 DF,  p-value: 1.323e-07
```

(c) (3 points) Plot the residuals versus the fitted values for each regression. Is there evidence that the variance of the residuals varies with the fitted values or that there are any outliers?

Adding the lm results to the data set

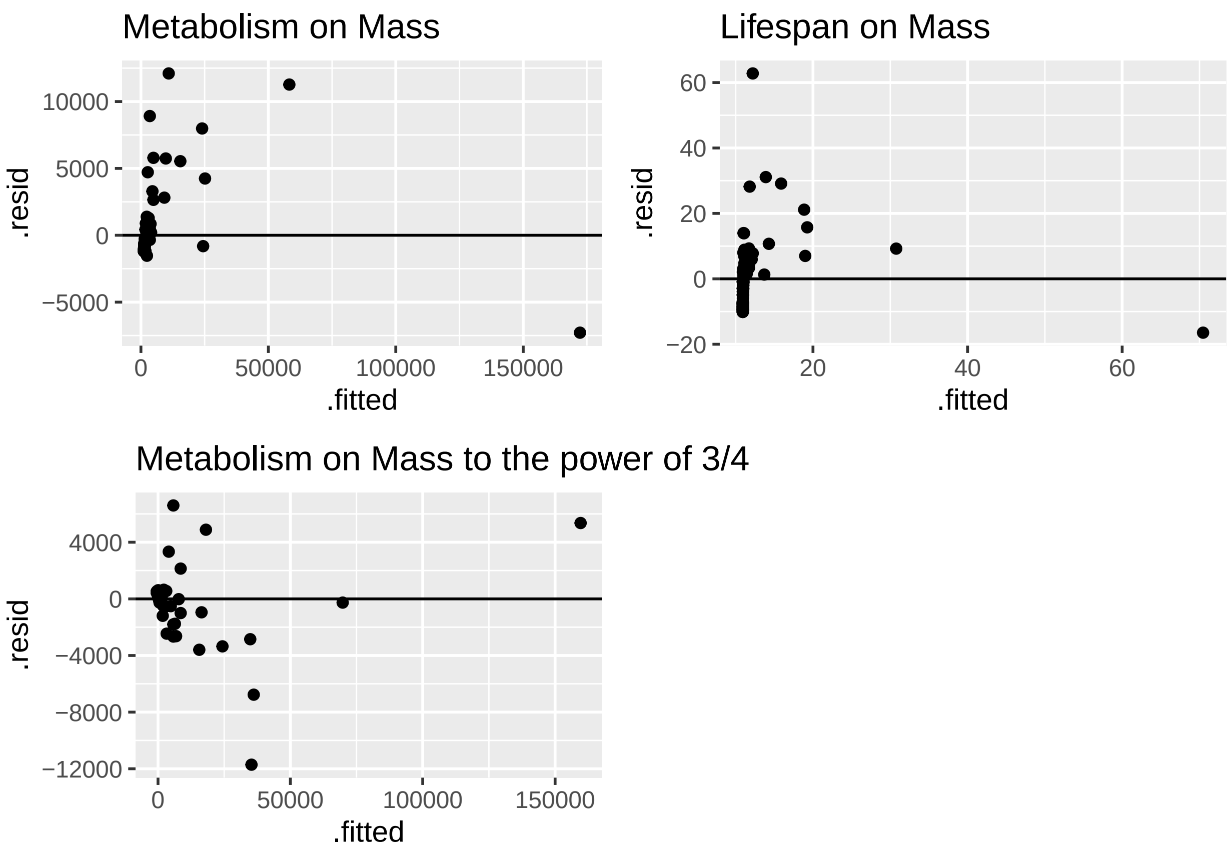
```
sprintf("Least squares fit to the linear regression of metabolism on mass")
```

```
## [1] "Least squares fit to the linear regression of metabolism on mass"
```

```
lin_mod1Data <- fortify(lin_mod1, HW1Data)
lin_mod2Data <- fortify(lin_mod2, HW1Data)
lin_mod3Data <- fortify(lin_mod3, HW1Data)
#View(lin_mod1Data)

# Fitted vs Residual Plot
plot1 <- qplot(.fitted, .resid, data = lin_mod1Data) + geom_hline(aes(yintercept = 0)) + ggtitle('Metabo
plot2 <- qplot(.fitted, .resid, data = lin_mod2Data) + geom_hline(aes(yintercept = 0)) + ggtitle('Metabo
plot3 <- qplot(.fitted, .resid, data = lin_mod3Data) + geom_hline(aes(yintercept = 0)) + ggtitle('Lifesp

multiplot(plot1, plot2, plot3, cols=2)
```



-

There is evidence that there are outliers.For Metabolism on Mass and Metabolism on Mass to the power of 3/4, we can see one value on each plot around 150000 on x axis. While we see a value above 60 on x axis in the Lifespan on Mass plot.

(d) (3 points) Report a summary of each least squares fit in one sentence each, using the t-test p-values and R2 values.

-Answer

-Metabolism on Mass

The t-value of 65.586 is a statistic that measures the significance of the estimated slope coefficient (Mass) and has a p-value of 2e-16. Since this p-value is less that $\alpha = 0.05$, we reject the null hypothesis that the slope is zero.

The F-statistic = 4302 is measuring the overall significance of the regression model. Since the p value(2.2e-16) associated to it is less that $\alpha = 0.05$, we reject the null hypothesis that there is no relationship between Metabolism and Mass.

The R2 = 0.978 value shows that the model fits the data well and there is a strong relationship between Metabolism and Mass.

-Metabolism on Mass to the power of 3/4 The t-value of 91.895 is a statistic that measures the significance of the estimated slope coefficient (Mass to the power of 3/4) and has a p-value of 2e-16. Since this p-value is less that $\alpha = 0.05$, we reject the null hypothesis that the slope is zero.

The F-statistic = 8445 is measuring the overall significance of the regression model. Since the p value(2.2e-16) associated to it is less that $\alpha = 0.05$, we reject the null hypothesis that there is no relationship between Metabolism and Mass.

The R2 = 0.989 value shows that the model fits the data well and there is a strong relationship between Metabolism and Mass to the power of 3/4.

-Lifespan on mass The t-value of 5.713 is a statistic that measures the significance of the estimated slope coefficient (Mass) and has a p-value of 1.32e-07. Since this p-value is less that $\alpha = 0.05$, we reject the null hypothesis that the slope is zero.

The F-statistic = 32.64 is measuring the overall significance of the regression model. Since the p value(1.323-07) associated to it is less that $\alpha = 0.05$, we reject the null hypothesis that there is no relationship between Lifespan and Mass.

The R2 = 0.256 value shows that the model does not fit the data well and there is a weak relationship between Lifespan and Mass.

#Question 2 (Problem 9.19 in textbook)

Depression and Education. Has homework got you depressed? It could be worse. Depression, like other illnesses, is more prevalent among adults with less education than you have. R. A. Miech and M. J. Shanahan investigated the association of depression with age and education, based on a 1990 nationwide (U.S.) telephone survey of 2,031 adults aged 18 to 90. Of particular interest was their finding that the association of depression with education strengthens with increasing age - a phenomenon they called the "divergence hypothesis." They constructed a depression score from responses to several related questions. Education was categorized as (i) college degree, (ii) high school degree plus some college, or (iii) high school degree only. (See "Socioeconomic Status and Depression over the Life Course," Journal of Health and Social Behaviour 41(2) (June, 2000): 162-74.) Note: This question does not involve any data analysis.

(a) (4 points) Construct a multiple linear regression model in which the mean depression score changes linearly with age in all three education categories, with possibly unequal slopes and intercepts. Identify the change in difference of mean depression score between categories (iii) and (i) with one unit change in age.

- Answer Multiple linear regression model Where:

(i) College degree = degree_1
(ii) high school degree plus some college = degree_2
(iii) high school degree = degree_3

$\mu\{depression|age, degree\} = \beta_0 + \beta_1(age) + \beta_2(degree_2) + \beta_3(degree_3) + \beta_4(age * degree_2) + \beta_5(age * degree_3)$

(i) mean depression score: $\beta_0 + \beta_1(age)$ (degree_2 and degree_3 are 0 in this case)

(ii) mean depression score iii $\beta_0 + \beta_1(age) + \beta_3 + \beta_5(age)$ $\beta_0 + \beta_3 + (\beta_1 + \beta_5)age$

difference of mean depression score between categories (iii) and (i) = β_5

(b) (4 points) Modify the model to specify that the slopes of the regression lines with age are equal in categories (i) and (ii) but possibly different in category (iii). Again identify the change in difference of mean depression score between categories (iii) and (i) with one unit change in age.

-Answer

Model: $\beta_0 + \beta_1(age) + \beta_2(degree_2) + \beta_3(degree_3) + \beta_4(age * degree_3)$

(i) mean depression score iv : $\beta_0 + \beta_1(age)$

(ii) mean depression score ii: $\beta_0 + \beta_1(age) + \beta_3 + \beta_4(age) = \beta_0 + \beta_3 + (\beta_1 + \beta_4)age$

difference of mean depression score between categories (iii) and (i) $= \beta_4$

(c) (4 points) This and other studies found evidence that the mean depression is high in the late teens, declines toward middle age, and then increases towards old age. Construct a multiple linear regression model in which these type of association can be captured by converting the continuous variable age into a categorical variable with three categories (late teen, middle age, and old age). Include the three education categories in the model too. Define the model in such a way that the association between mean depression score and age variable can be different for different education categories.

-Answer

College degree = degree_0 (reference) high school degree plus some college = degree_1 high school degree = degree_2 late teen = teen (reference) middle age = middle old age = old

$\mu\{depression|age, degree\} = \beta_0 + \beta_1(middle) + \beta_2(old) + \beta_3(degree_1) + \beta_4(degree_2) + \beta_5(middle * degree_1) + \beta_6(old * degree_1) + \beta_7(middle * degree_2) + \beta_8(old * degree_2)$