week5Homework

M. Onimus

10/8/2020

Contents

Read in Data

```
week5 <- read_sav("assignments/GLM_homework_phmc.sav")
```

Question 1

David is hoping to come up with a linear model that can be used to predict a person's BMI using a subset of the PHMC community health survey database.

We first wish to examine the relationship between BMI and age.

Question/Answer 1a

What is the correlation between age and BMI?

```
lm <- lm(BMI ~ RESPAGE, data = week5)




#lm
```

The correlation between BMI and age is 0.0042962.

Question/Answer 1b

Write out the regression equation for the prediction of BMI using subject age.

$BMI = \beta_0 + \beta_1 * age$

Question/Answer 1c

What are the null and alternative hypotheses for the test for an association between age and BMI, using your model?

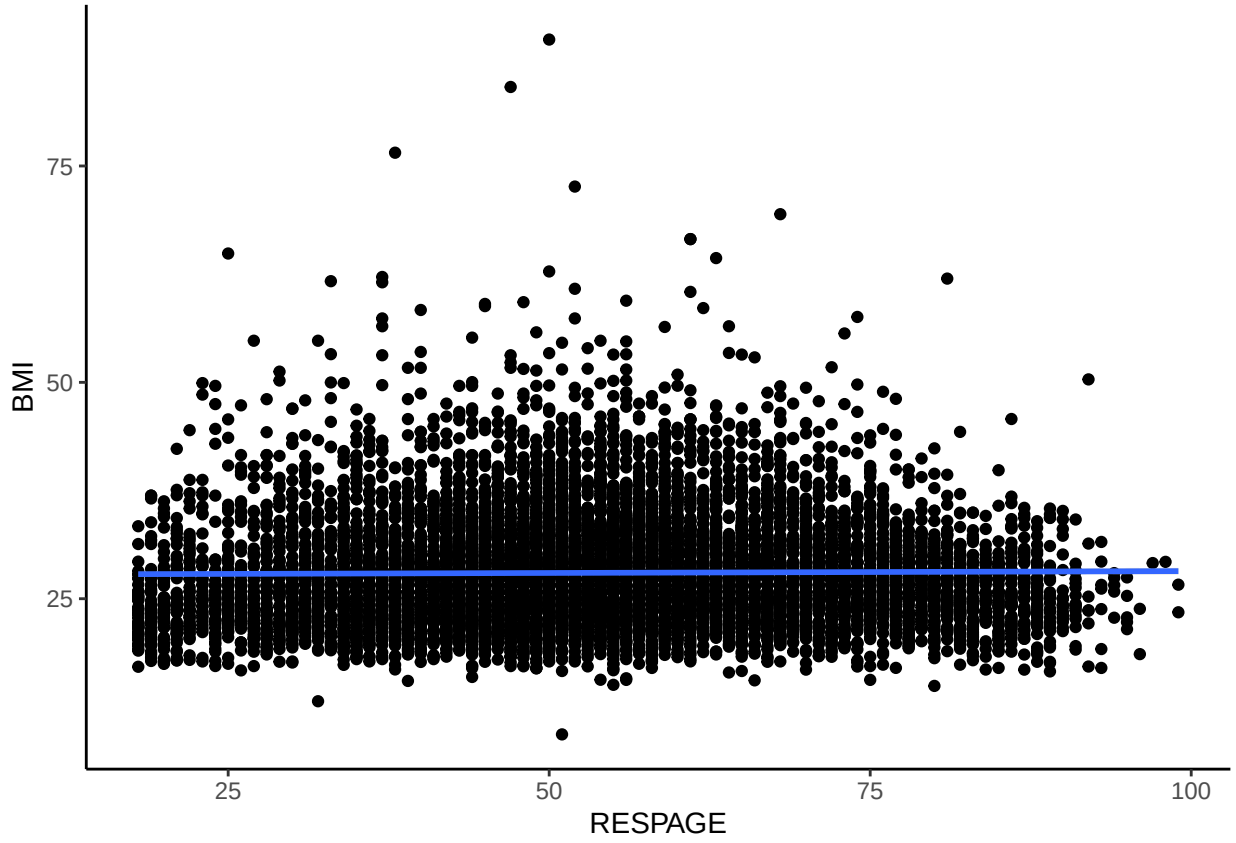
$H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$

Question/Answer 1d

What is the estimated regression equation? Also, provide a plot of it against a scatter plot of the data.

$BMI = 27.7601 + 0.0043 * age$

```
ggplot(week5, aes(x = RESPAGE, y = BMI)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  theme_classic()
```



Question/Answer 1e

What are the results of inference about the association (parameter estimate table)?

```
summary(lm)
```

```
##
## Call:
## lm(formula = BMI ~ RESPAGE, data = week5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.654  -4.314  -1.136   3.027  61.631
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.760111   0.223472 124.222   <2e-16 ***
```

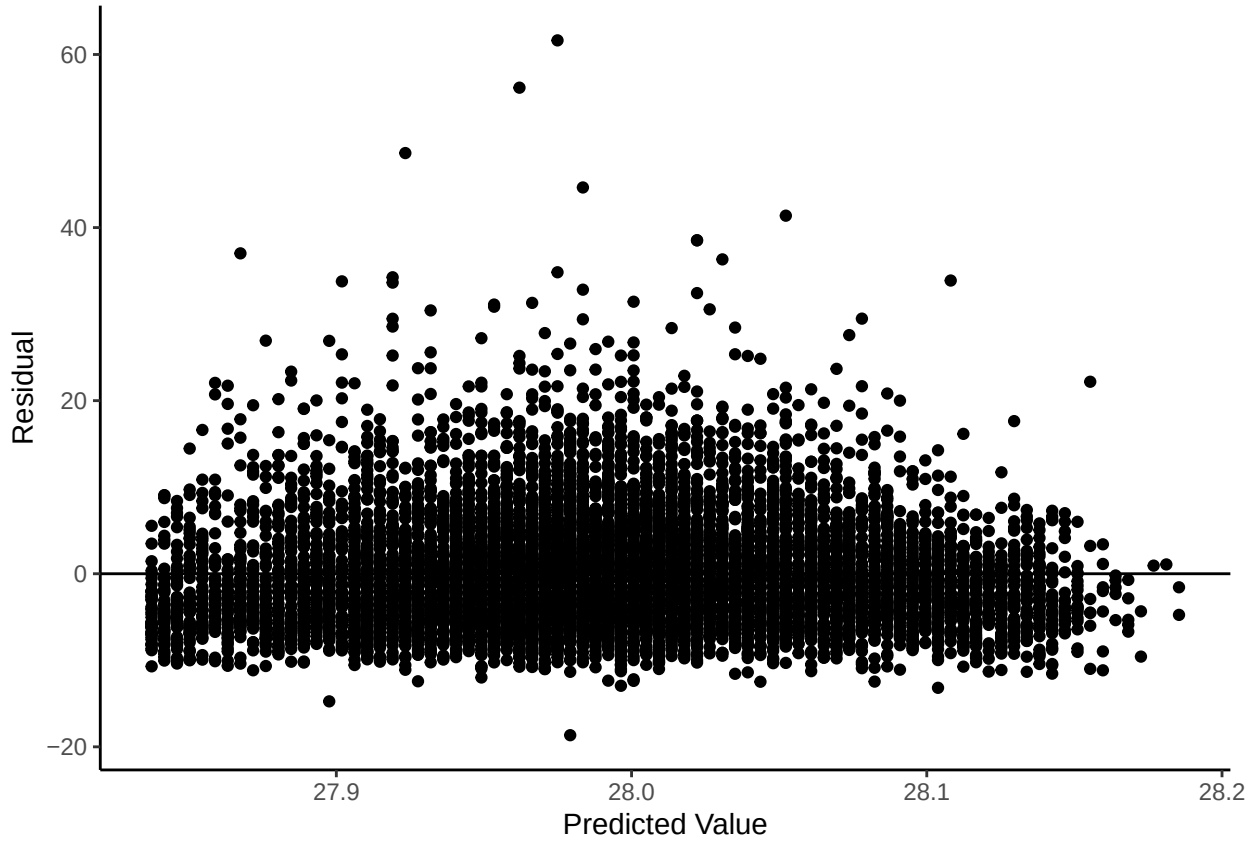
```
## RESPAGE      0.004296   0.003996   1.075    0.282
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.275 on 9817 degrees of freedom
##   (229 observations deleted due to missingness)
## Multiple R-squared:  0.0001177,  Adjusted R-squared:  1.59e-05
## F-statistic: 1.156 on 1 and 9817 DF,  p-value: 0.2823
```

The results of interference indicate that the two parameters are not signifcantly associated with each other ($t_1 = 1.075, p = 0.282$) at an $\alpha = 0.05$ level.

Question/Answer 1f

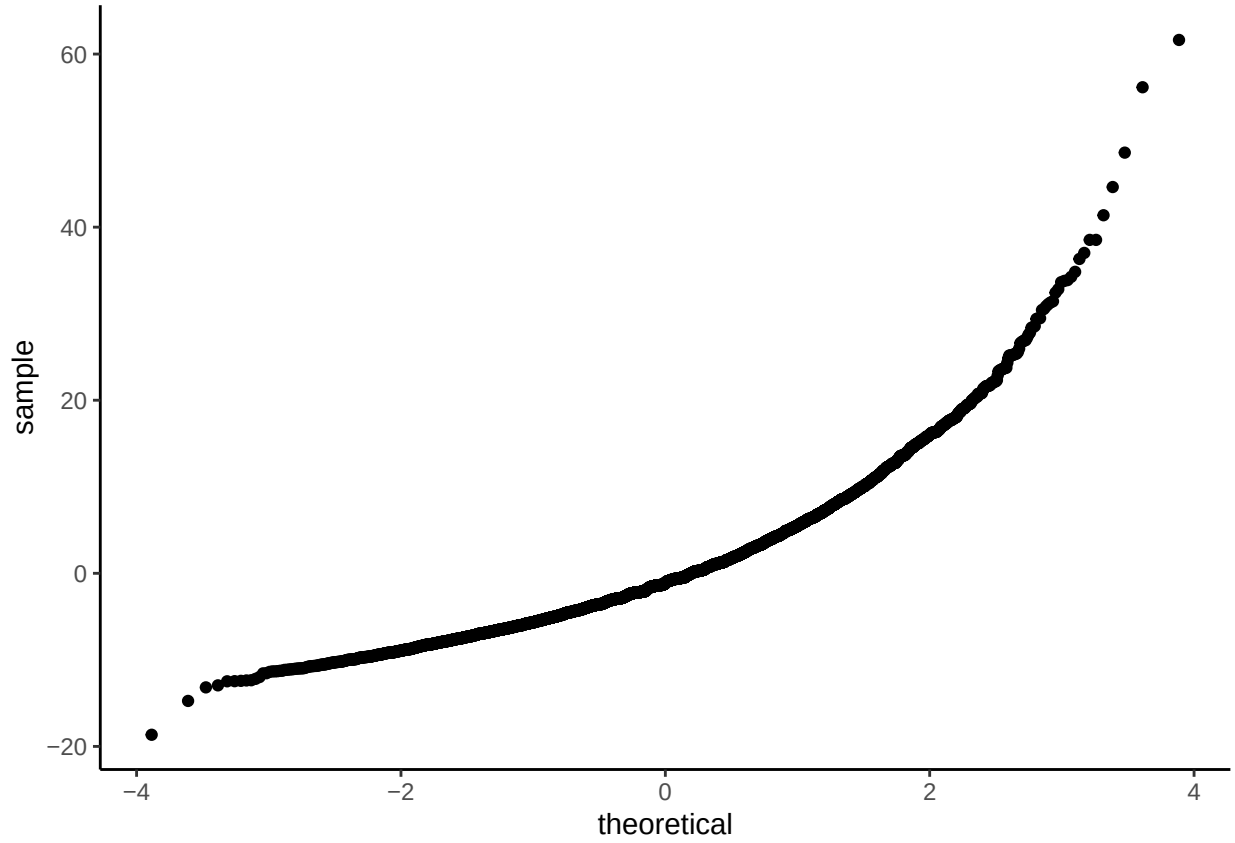
How do the assumptions of linear regression hold up in this case? Use the appropriate plots to support your discussion.

```
ggplot(data = NULL, aes(x = lm$fitted.values, y = lm$residuals)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs(
    x = "Predicted Value",
    y = "Residual"
  ) +
  theme_classic()
```

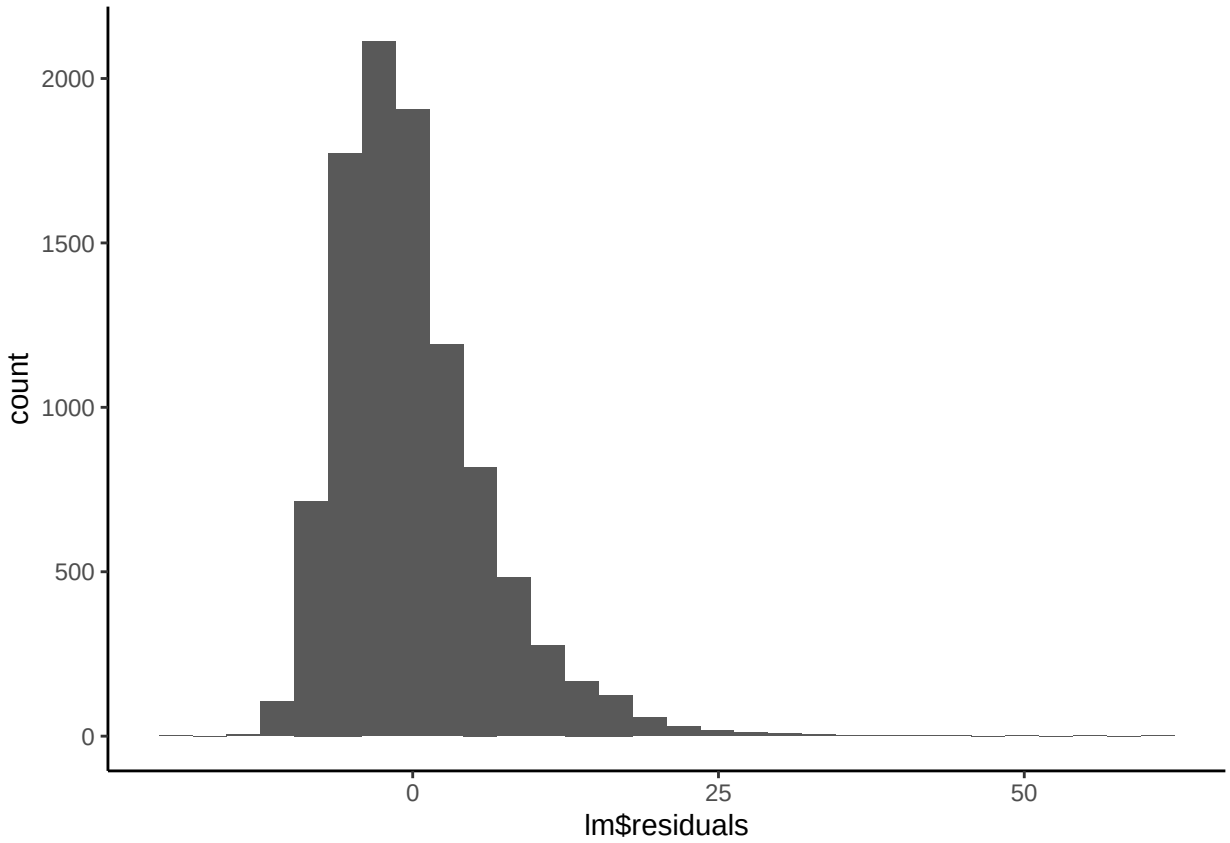


The residual/predicted plot above shows a pretty equal distrubution of points above and below 0 although there may be less points below 0.

```
ggplot(data = NULL, aes(sample = lm$residuals)) +
  geom_qq() +
  theme_classic()
```




```
ggplot(data = NULL, aes(x = lm$residuals)) +
  geom_histogram() +
  theme_classic()
```



The qqplot and the histogram of residuals both show distrubtions which seem to not be normal. The qqplot looks closer to an exponential plot as opposed to the expected straightline; the histrogram shows tailing.

Question/Answer 1g

Write a few sentences describing the results.

A correlation between BMI and age was investigated using a linear regression. With a correlation coefficent of 0.0043 and a p-value of 0.282, there is no significant correlation between the two variables. Additionally, the data failed to meet qqplot and histogram normality assumptions.