Data 605 - W12 HW

Avery Davidowitz

2022-11-13

Libraries

```
library(tidyverse)
```

Data

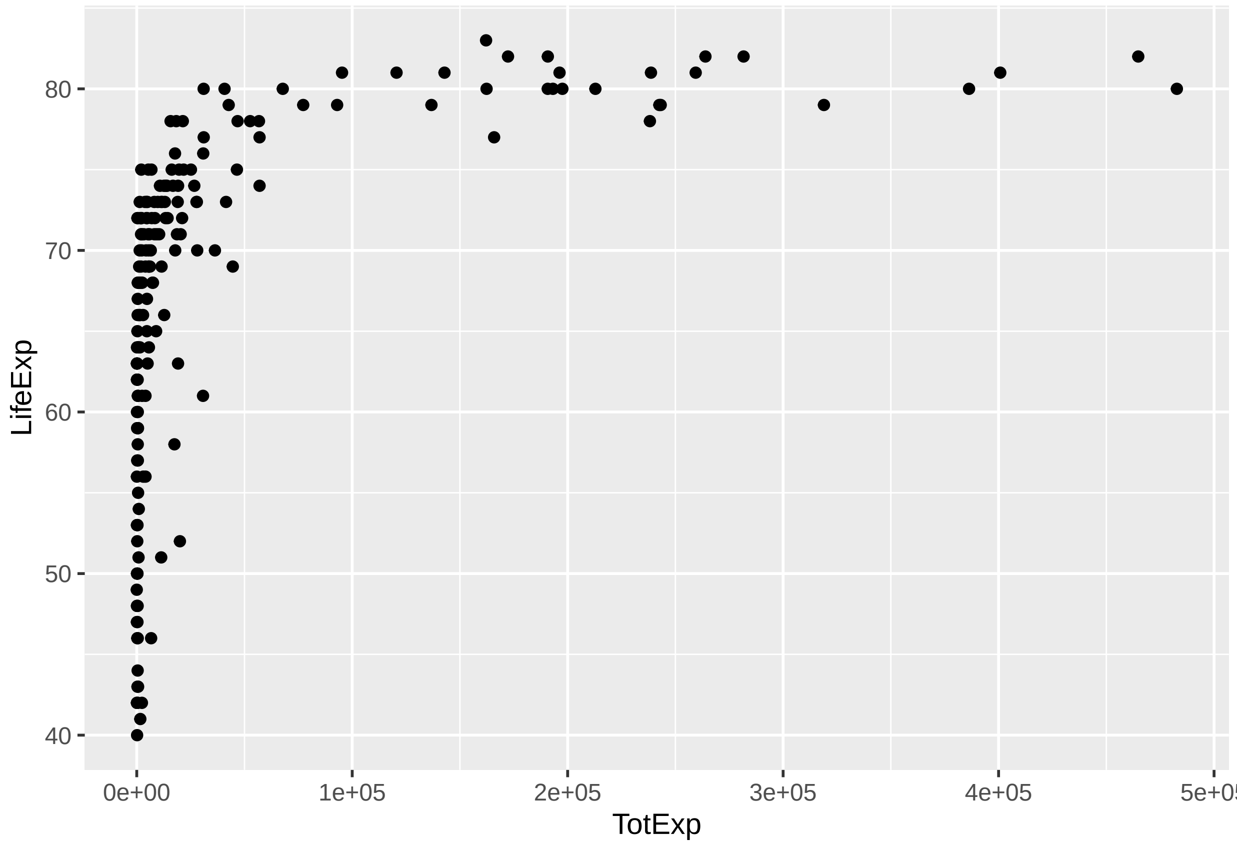
```
df <- readr::read_csv(paste0(getwd(),"/who.csv"))
head(df)
```

```
## # A tibble: 6 x 10
##   Country  LifeExp Infan~1 Under~2 TBFree  PropMD  PropRN PersExp GovtExp TotExp
##   <chr>      <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>
## 1 Afghani~      42   0.835   0.743  0.998 2.29e-4 5.72e-4      20      92    112
## 2 Albania       71   0.985   0.983  1.00  1.14e-3 4.61e-3     169    3128   3297
## 3 Algeria       71   0.967   0.962  0.999 1.06e-3 2.09e-3     108    5184   5292
## 4 Andorra       82   0.997   0.996  1.00  3.30e-3 3.5 e-3    2589  169725 172314
## 5 Angola        41   0.846   0.74   0.997 7.04e-5 1.15e-3      36    1620   1656
## 6 Antigua~      73   0.99    0.989  1.00  1.43e-4 2.77e-3     503   12543  13046
## # ... with abbreviated variable names 1: InfantSurvival, 2: Under5Survival
```

1

Provide a scatterplot of LifeExp~TotExp, and run simple linear regression. Do not transform the variables. Provide and interpret the F statistics, R^2, standard error,and p-values only. Discuss whether the assumptions of simple linear regression met.

```
ggplot(data=df, aes(x=TotExp, y=LifeExp)) + geom_point()
```

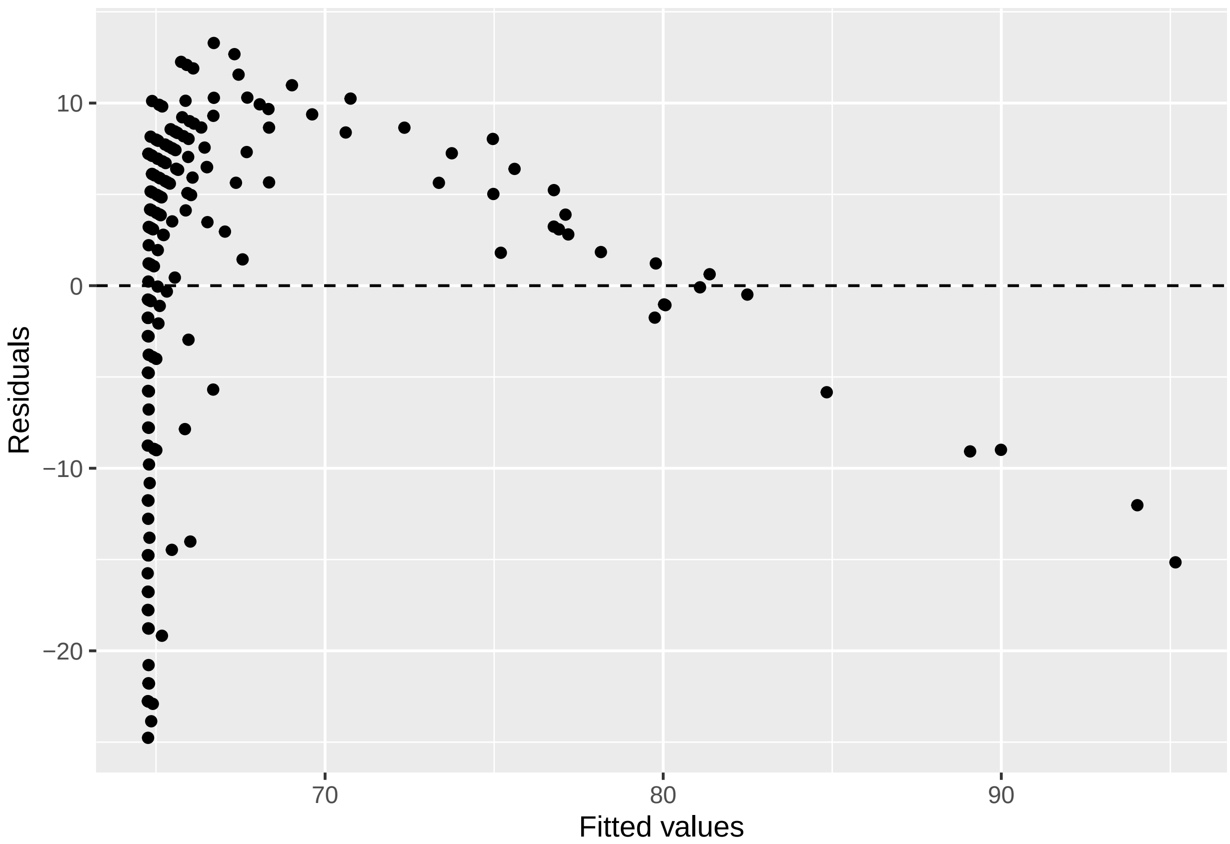



```r
life_exp_lm <- lm(LifeExp ~ TotExp, data = df)
summary(life_exp_lm)
```

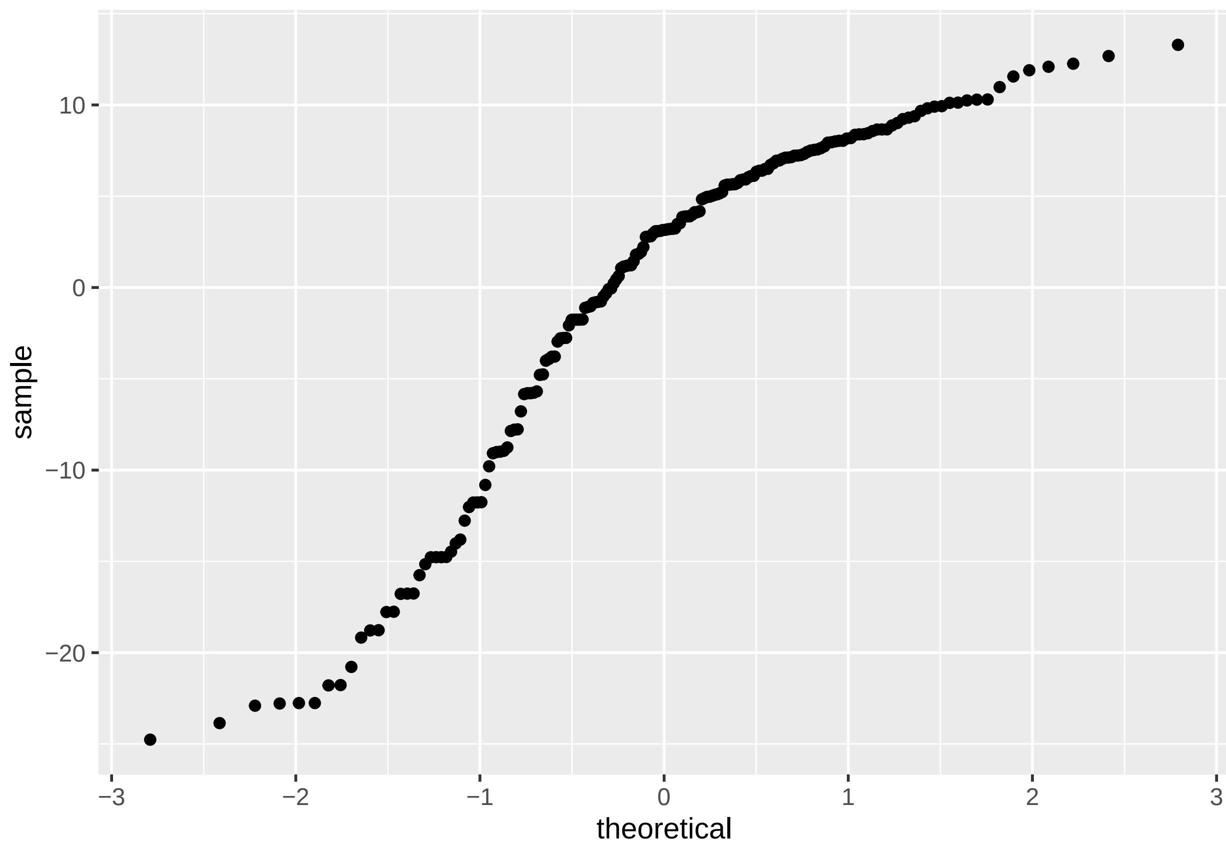
```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.764  -4.778   3.154   7.116  13.292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.475e+01  7.535e-01  85.933  < 2e-16 ***
## TotExp      6.297e-05  7.795e-06   8.079 7.71e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.371 on 188 degrees of freedom
## Multiple R-squared:  0.2577, Adjusted R-squared:  0.2537
## F-statistic: 65.26 on 1 and 188 DF,  p-value: 7.714e-14
```

```r
ggplot(data = life_exp_lm, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
```

```
  xlab("Fitted values") +
  ylab("Residuals")
```




```
ggplot(data = life_exp_lm, aes(sample = .resid)) +
  stat_qq()
```



The R squared value of .25 is very low and indicates that the model does not explain the dependent variable well. From the residuals plots we can see that we are violating assumptions for linear models regarding the normal distribution of residuals and constant variability.

2

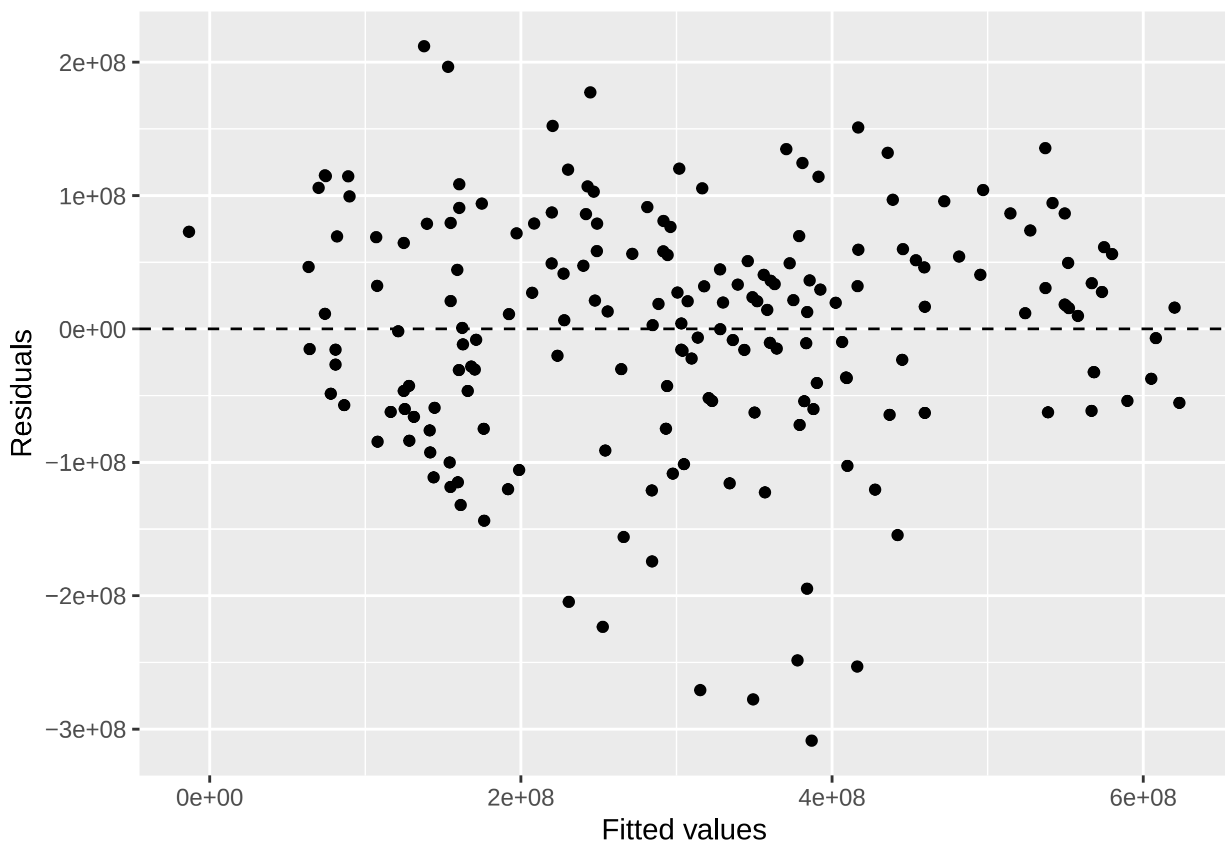
Raise life expectancy to the 4.6 power (i.e., LifeExp^4.6). Raise total expenditures to the 0.06 power (nearly a log transform, TotExp^.06). Plot LifeExp^4.6 as a function of TotExp^.06, and r re-run the simple regression model using the transformed variables. Provide and interpret the F statistics, R^2, standard error, and p-values. Which model is "better?"

```
df2 <- df |> dplyr::mutate(LifeExp = LifeExp^4.6) |> dplyr::mutate(TotExp = TotExp^.06)
life_exp_lm2 <- lm(LifeExp ~ TotExp, data = df2)
summary(life_exp_lm2)
```

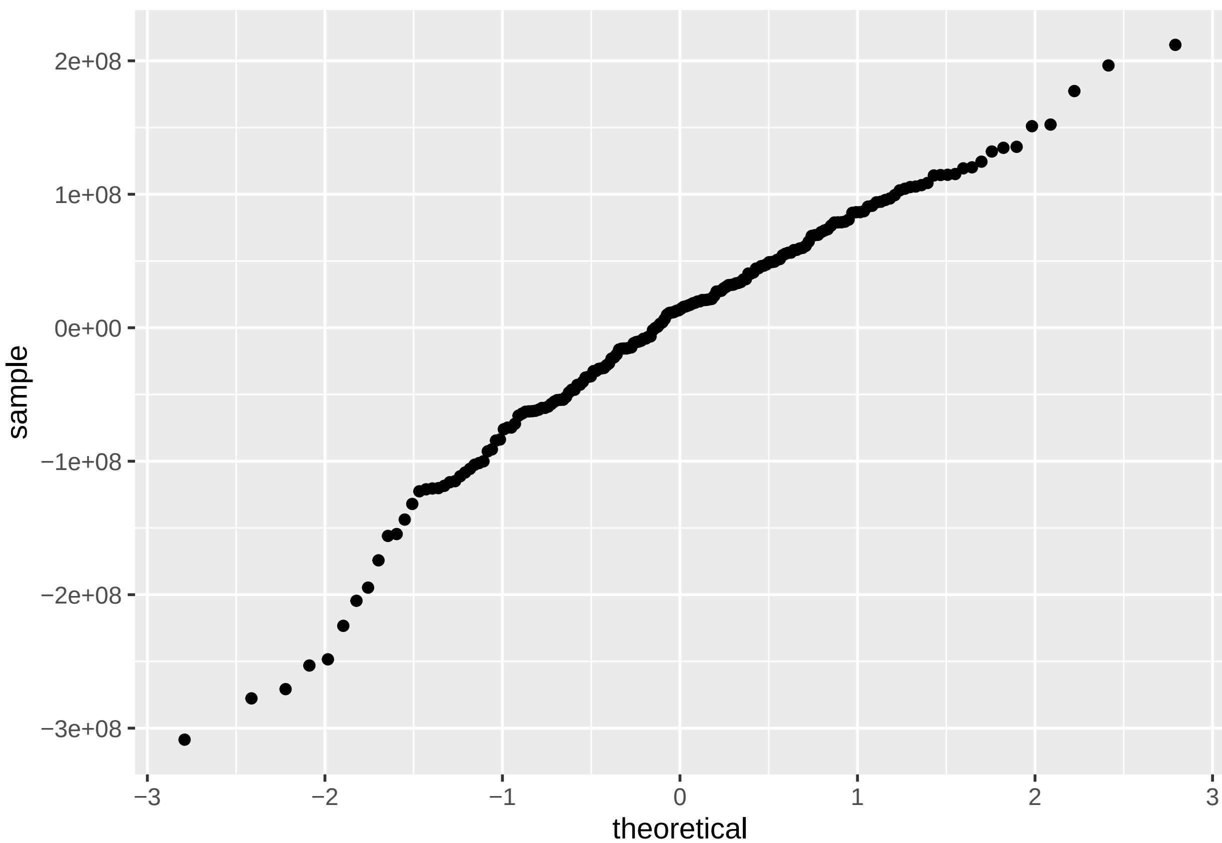
```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = df2)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -308616089  -53978977   13697187   59139231  211951764
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -736527910   46817945  -15.73   <2e-16 ***
## TotExp       620060216   27518940   22.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 90490000 on 188 degrees of freedom
## Multiple R-squared:  0.7298, Adjusted R-squared:  0.7283
## F-statistic: 507.7 on 1 and 188 DF,  p-value: < 2.2e-16
```

```
ggplot(data = life_exp_lm2, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```




```
ggplot(data = life_exp_lm2, aes(sample = .resid)) +
  stat_qq()
```



The assumptions are much more closely met after transformation with both the normality and constant variance were satisfied unlike before. The power scaled model also performs significantly better across all metrics. The R squared explains an additional 50% of the variance in the dependent variable. The p-value representing the probability that the relationship between variables was due to chance also decreased by a factor of 10. The ration of residual standard error/mean estimate(Intercept) also decreased by 2% so while the error of 90490000 in the power model looks enormous it is a strict improvement.

3

Forecast life expectancy when TotExp^.06 =1.5. Then forecast life expectancy when TotExp^.06=2.5.

```
new <- data.frame(TotExp = c(1.5,2.5))
predict.lm(life_exp_lm2, new)
```

```
##         1         2
## 193562414 813622630
```

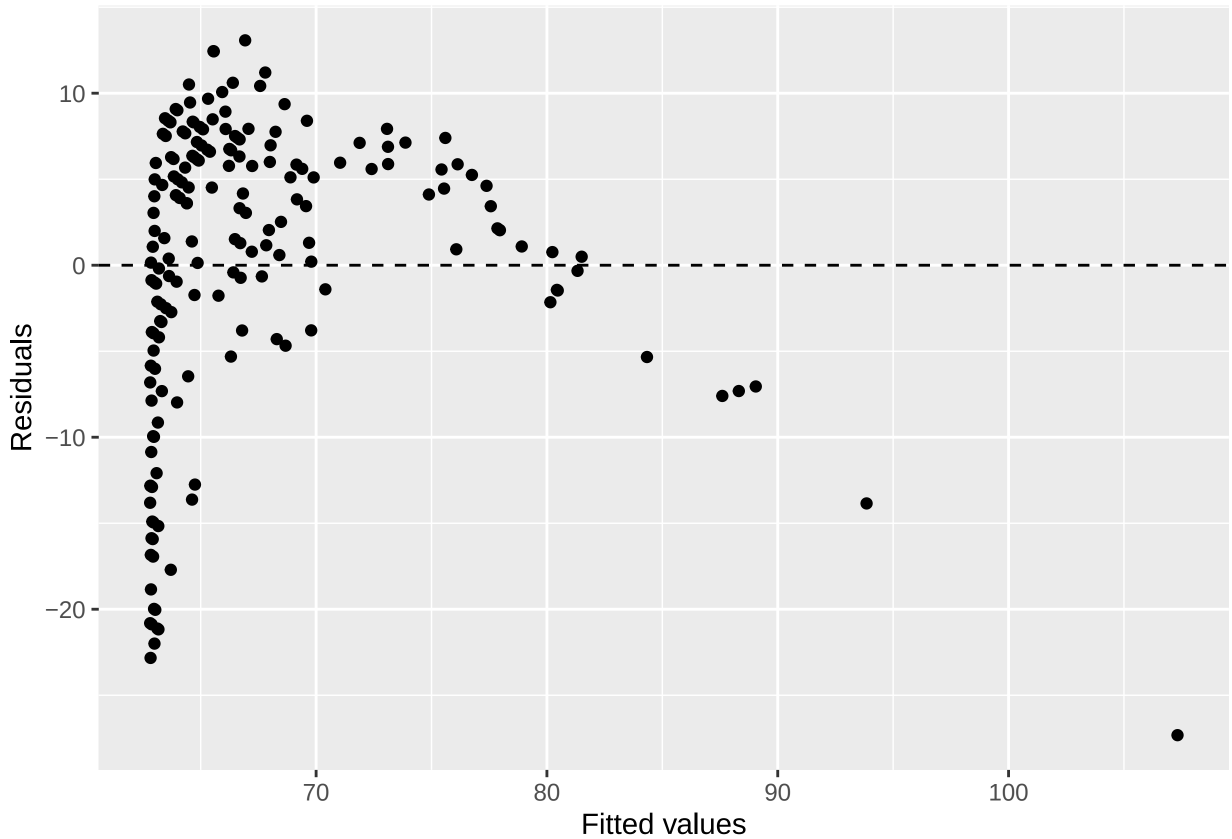
4

Build the following multiple regression model and interpret the F Statistics, R^2, standard error, and p-values. How good is the model? LifeExp = b0+b1 x PropMd + b2 x TotExp +b3 x PropMD x TotExp

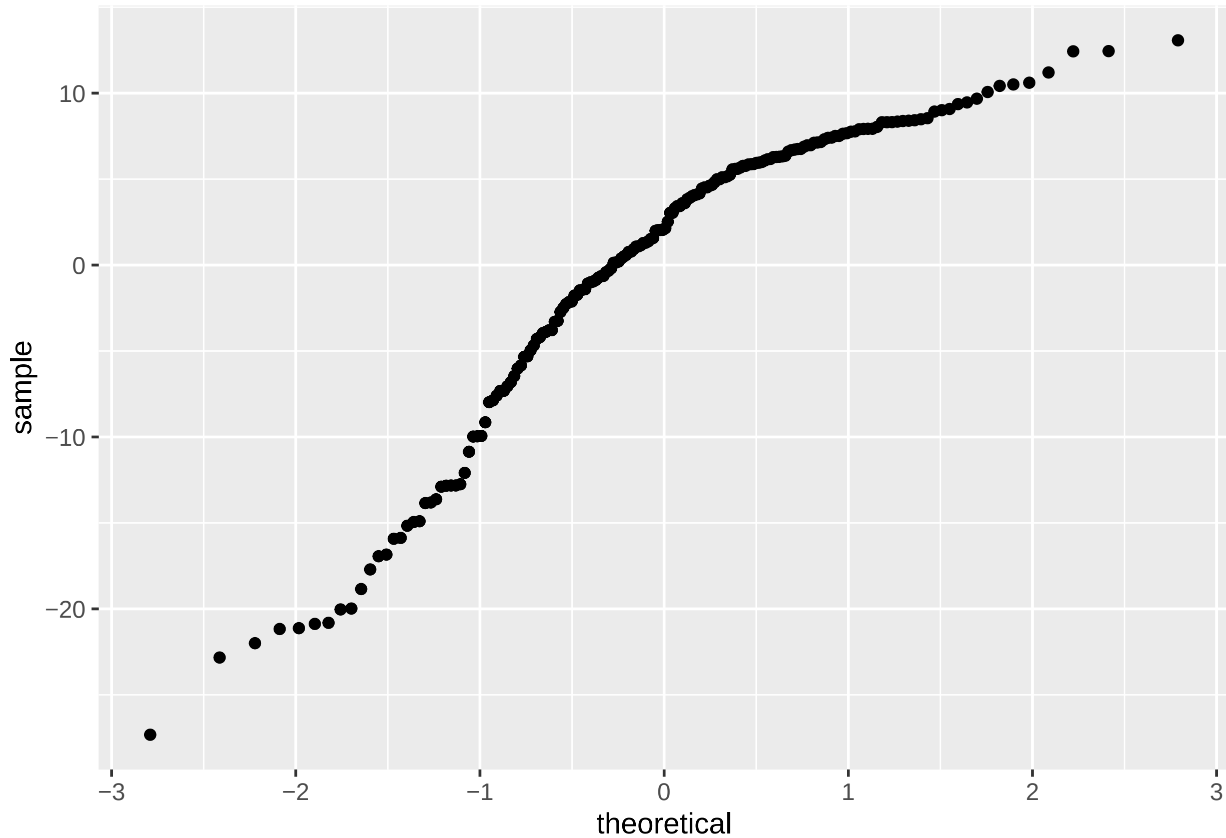
```r
df3 <- df |> dplyr::mutate(MDExp = PropMD * TotExp)
life_exp_lm3 <- lm(LifeExp ~ PropMD + TotExp + MDExp, data = df3)
summary(life_exp_lm3)
```

```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + MDExp, data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.320  -4.132   2.098   6.540  13.074
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.277e+01  7.956e-01  78.899  < 2e-16 ***
## PropMD       1.497e+03  2.788e+02   5.371 2.32e-07 ***
## TotExp       7.233e-05  8.982e-06   8.053 9.39e-14 ***
## MDExp       -6.026e-03  1.472e-03  -4.093 6.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.765 on 186 degrees of freedom
## Multiple R-squared:  0.3574, Adjusted R-squared:  0.3471
## F-statistic: 34.49 on 3 and 186 DF,  p-value: < 2.2e-16
```

```r
ggplot(data = life_exp_lm3, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```




```
ggplot(data = life_exp_lm3, aes(sample = .resid)) +
  stat_qq()
```



This model has a very poor performance with an adjusted R squared showing that it only explains 34% of the variance in the dependent variable. The model is "accurate" in that the standard error and p-values are low. However, it still isn't very useful. It seems to still violate assumptions for linear models for normal distribution of residuals and constant variability.

5

Forecast LifeExp when PropMD=.03 and TotExp = 14. Does this forecast seem realistic? Why or why not?

```
new <- data.frame(TotExp = 14,PropMD = .03, MDExp = 14*.03)
predict.lm(life_exp_lm3, new)
```

```
##       1
## 107.696
```

This prediction doesn't seem remotely accurate because that life expectancy value is 20 over the max from the data set.