Regression: Advanced Topics in Stats III

2022-02-24

Today we will finally discuss Simple and Multiple Linear Regression.

# Analysis Framework

1. Load the Data  
  
2. Understand the data (Distribution, Variability, Outliers)  
 \* Summary Statistics  
 \* Data Visualization  
 \* Outliers  
  
3. Check Assumptions  
 \* Normality of Residuals  
 \* Homoskedasticity  
 \* Independent Observations  
 \* Linear Relationship  
 \* Continuous variables  
  
4. Conduct the Statistical Test  
 \* Run Linear Regression `lm()`  
   
5. Conduct Post-Hoc Analyses as needed

# Multiple Regression

## 1. Load the Data

# Loading the Data  
dat <- pl462::happyparent %>%   
 select(sex, str\_slf1, negmood1, health)# Loads the dataset called 'happyparent' from the pl462 package and assigns it the name "dat"  
?happyparent # Loads the help menu giving some descriptive information for the dataset  
glimpse(dat) # Gives a quick view of the structure of the dataset.

## Rows: 302  
## Columns: 4  
## $ sex <fct> female, female, female, female, female, female, male, female,…  
## $ str\_slf1 <dbl> 2, 0, NA, 0, NA, 0, 2, 0, 2, 1, 0, 1, 0, 3, 1, 1, 0, 0, 2, 2,…  
## $ negmood1 <dbl> 4.166667, 4.666667, NA, 4.833333, NA, 4.166667, 4.500000, 5.0…  
## $ health <dbl> 4, 4, 5, 5, 3, 5, 4, 4, 5, 4, 3, 4, 3, 2, 3, 1, 4, 4, 3, 4, 4…

Research Question: How are stress regarding self, self-rated physical health, and sex related to negative mood?

What is our DV? >

What are our IVs? >

## 2. Understand the data (Distribution, Variability, Outliers)

* Summary Statistics

dat %>%   
 skimr::skim()

Data summary

|  |  |
| --- | --- |
| Name | Piped data |
| Number of rows | 302 |
| Number of columns | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 1 |
| numeric | 3 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| sex | 0 | 1 | FALSE | 2 | mal: 178, fem: 124 |

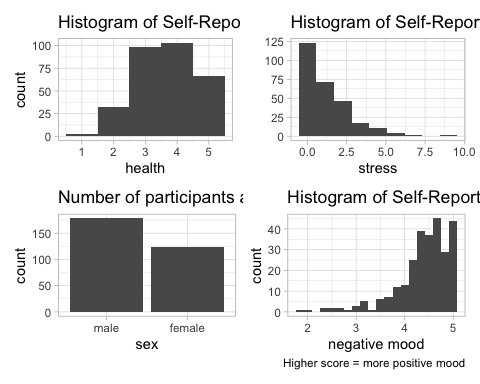
**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| str\_slf1 | 27 | 0.91 | 1.10 | 1.42 | 0.00 | 0.00 | 1.0 | 2.00 | 9 | ▇▂▁▁▁ |
| negmood1 | 27 | 0.91 | 4.38 | 0.58 | 1.83 | 4.17 | 4.5 | 4.83 | 5 | ▁▁▁▅▇ |
| health | 0 | 1.00 | 3.65 | 0.97 | 1.00 | 3.00 | 4.0 | 4.00 | 5 | ▁▂▇▇▅ |

What are your observations given the descriptive statistics?

* Data Visualization

p1 <- dat %>%  
 ggplot() +  
 aes(x = health) +  
 geom\_histogram(bins = 5) +  
 labs(title = "Histogram of Self-Reported Health")  
p2 <- dat %>%  
 ggplot() +  
 aes(x = str\_slf1) +  
 geom\_histogram(bins = 9) +  
 labs(title = "Histogram of Self-Reported Stress",  
 x = "stress",  
 y = "")  
p3 <- dat %>%  
 ggplot() +  
 aes(x = sex) +  
 geom\_histogram(stat = "count", bins = 2) +  
 labs(title = "Number of participants according to Sex")  
p4 <- dat %>%  
 ggplot() +  
 aes(x = negmood1) +  
 geom\_histogram(bins = 20) +  
 labs(title = "Histogram of Self-Reported Negative Mood",  
 x = "negative mood",  
 caption = "Higher score = more positive mood")  
  
(p1 + p2) / (p3 + p4)



* Correlations

dat %>%  
 select(-sex) %>%  
 cor\_mat() %>%  
 cor\_mark\_significant(cutpoints = c(0, .001, .01, .05, 1),  
 symbols = c("\*\*\*", "\*\*", "\*", ""))

## rowname str\_slf1 negmood1 health  
## 1 str\_slf1   
## 2 negmood1 -0.69\*\*\*   
## 3 health -0.15\* 0.14\*

What are your observations (distribution and correlations)?

* Outliers

dat %>%   
 rstatix::identify\_outliers(negmood1) # Produces a table of outliers and extreme cases.

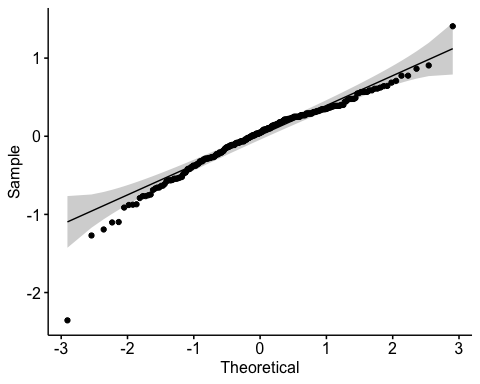
## # A tibble: 17 × 6  
## sex str\_slf1 negmood1 health is.outlier is.extreme  
## <fct> <dbl> <dbl> <dbl> <lgl> <lgl>   
## 1 female 2 3.167 4 TRUE FALSE   
## 2 female 3 3.167 5 TRUE FALSE   
## 3 female 5 3 3 TRUE FALSE   
## 4 female 6 2.667 3 TRUE FALSE   
## 5 male 3 3 3 TRUE FALSE   
## 6 female 1 2 4 TRUE TRUE   
## 7 female 4 2.333 4 TRUE FALSE   
## 8 female 6 1.833 2 TRUE TRUE   
## 9 male 4 2.333 3 TRUE FALSE   
## 10 male 7 2.5 3 TRUE FALSE   
## 11 male 5 3 4 TRUE FALSE   
## 12 female 2 3.167 2 TRUE FALSE   
## 13 male 4 2.5 3 TRUE FALSE   
## 14 male 4 2.833 4 TRUE FALSE   
## 15 female 6 3.167 3 TRUE FALSE   
## 16 male 3 3.167 5 TRUE FALSE   
## 17 male 9 2.667 3 TRUE FALSE

Are there any outliers in these data?

## 3. Check Assumptions

* Normality of Residuals

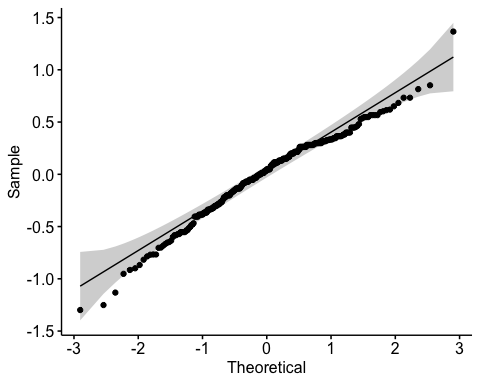
# Assumption testing- Normality of Residuals- QQ Plot  
model <- lm(negmood1 ~ sex + str\_slf1 + health,   
 data = dat)  
ggqqplot(residuals(model)) # Graphically depicts the correlation of these data and a normal distribution.



# Shapiro-Wilkes Test  
shapiro\_test(residuals(model)) # Statistically determines normality of these data.

## # A tibble: 1 × 3  
## variable statistic p.value  
## <chr> <dbl> <dbl>  
## 1 residuals(model) 0.9533 0.0000001134

# With outliers removed  
model2 <- lm(negmood1 ~ sex + str\_slf1 + health,   
 data = dat %>%   
 filter(!is\_extreme(negmood1)))  
ggqqplot(residuals(model2))



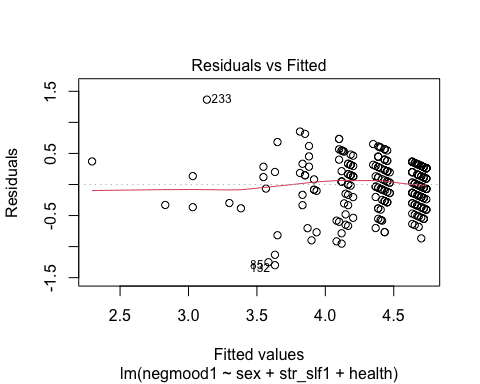
shapiro\_test(residuals(model2))

## # A tibble: 1 × 3  
## variable statistic p.value  
## <chr> <dbl> <dbl>  
## 1 residuals(model2) 0.9792 0.0005415

Are these residuals normally distributed?

* Homoskedasticity

# Residuals vs Fitted Plot  
plot(model2, 1)# Graphically compares the residuals to fitted values

 \*If the data has homoskedasticity, the points should be equally dispursed away from the line as you move from one side to the other. If the points resemble more of a sideways “V”, you have heteroskedasticity and you would need to find a way to correct for that (Box-Cox transformation)

Do these data meet the assumption of homoskedasticity?

## 4. Conduct the Statistical Test

* Run linear regression

# We've already run the linear regression above. Now to see the results:   
summary(model2)

##   
## Call:  
## lm(formula = negmood1 ~ sex + str\_slf1 + health, data = dat %>%   
## filter(!is\_extreme(negmood1)))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.29861 -0.22879 0.04761 0.28044 1.36570   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.64569 0.09941 46.732 <2e-16 \*\*\*  
## sexfemale -0.06593 0.04886 -1.349 0.178   
## str\_slf1 -0.26729 0.01726 -15.486 <2e-16 \*\*\*  
## health 0.01847 0.02510 0.736 0.463   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3918 on 267 degrees of freedom  
## (2 observations deleted due to missingness)  
## Multiple R-squared: 0.485, Adjusted R-squared: 0.4792   
## F-statistic: 83.81 on 3 and 267 DF, p-value: < 2.2e-16

Interpret the output?

What is the predicted mood score for a woman with 5 stressors and a self-rated health score of 3?

Ŷ = Intercept - b(stress-self) + b(health) - b(gender)

* Find standardized regression coefficients for reporting

summary(lm.beta(model2), standardized = TRUE)

##   
## Call:  
## lm(formula = negmood1 ~ sex + str\_slf1 + health, data = dat %>%   
## filter(!is\_extreme(negmood1)))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.29861 -0.22879 0.04761 0.28044 1.36570   
##   
## Coefficients:  
## Estimate Standardized Std. Error t value Pr(>|t|)   
## (Intercept) 4.64569 0.00000 0.09941 46.732 <2e-16 \*\*\*  
## sexfemale -0.06593 -0.05936 0.04886 -1.349 0.178   
## str\_slf1 -0.26729 -0.68701 0.01726 -15.486 <2e-16 \*\*\*  
## health 0.01847 0.03265 0.02510 0.736 0.463   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3918 on 267 degrees of freedom  
## (2 observations deleted due to missingness)  
## Multiple R-squared: 0.485, Adjusted R-squared: 0.4792   
## F-statistic: 83.81 on 3 and 267 DF, p-value: < 2.2e-16