Table of Contents

[Unit 7: Conclusion 2](#_Toc89892814)

[Simple Linear Regression 2](#_Toc89892815)

[Building a Model & Assessing Conditions 3](#_Toc89892816)

[Transformations 3](#_Toc89892817)

[Outliers & Influential Points 3](#_Toc89892818)

[Inference for Regression Slope/Correlation 3](#_Toc89892819)

[ANOVA for simple linear regression 3](#_Toc89892820)

[Confidence and Prediction Intervals 3](#_Toc89892821)

[Multiple Regression 3](#_Toc89892822)

[Building a Model & Assessing Conditions 3](#_Toc89892823)

[Methods for Choosing Predictors 3](#_Toc89892824)

[Interaction Terms 3](#_Toc89892825)

[Polynomial Models 3](#_Toc89892826)

[Second Order Models 3](#_Toc89892827)

[Multicollinearity & VIF 3](#_Toc89892828)

[Nested F-Test 3](#_Toc89892829)

[Cross Validation Correlation 3](#_Toc89892830)

[Logistic Regression 3](#_Toc89892831)

[Building a Model with a single predictors 3](#_Toc89892832)

[Building a Model with a single predictors 3](#_Toc89892833)

[Building a Model with multiple predictors 3](#_Toc89892834)

[Tests of significance for a model (single or multiple predictors) 3](#_Toc89892835)

[Test of significance for nested models 3](#_Toc89892836)

[Methods for Choosing Predictors 3](#_Toc89892837)

[Assessing linearity of the logit model 3](#_Toc89892838)

[Confidence interval for Odds Ratio 3](#_Toc89892839)

[Interpreting Odds Ratio 3](#_Toc89892840)

[Predicting Probabilities 3](#_Toc89892841)

[ANOVA for Means 3](#_Toc89892842)

[One Way ANOVA 3](#_Toc89892843)

[Two Way ANOVA 3](#_Toc89892844)

[Two Way ANOVA with Interaction 3](#_Toc89892845)

[Tukey HSD 3](#_Toc89892846)

[Levene’s Test 3](#_Toc89892847)

# Unit 7: Conclusion

## Simple Linear Regression

### Building a Model & Assessing Conditions

Building a Model and Assessing Conditions

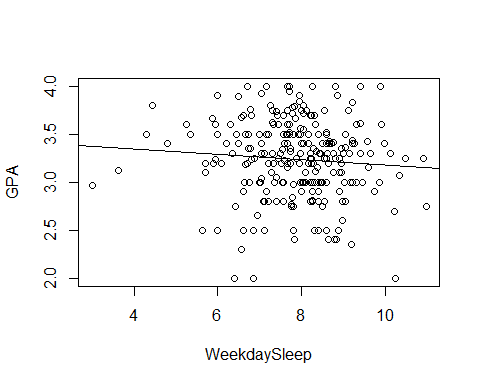
R Markdown

library(readr)  
  
Sleep <- read.csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

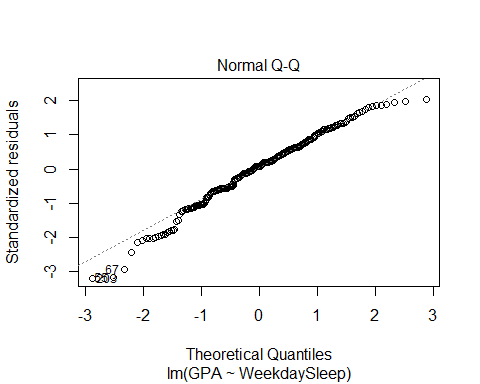
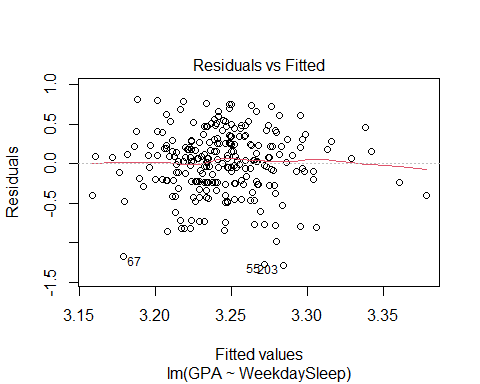
1. Build a linear model predicting GPA based on average hours of sleep during weekdays. Comment on how well your data appear to fit the conditions for a linear model.

Looking at original plot of the data with the linear model, the data does not seem to follow a linear pattern as described by the model line. The residuals vs. fitted plot shows a relatively linear pattern (looking at the plotted line as compared to the origin line) and the variance seems relatively constant across the residual data, with possible slight tapering in towards the right-hand side. The normal Q-Q line shows some skew on the left hand side of the plot, showing possible deviation from a normal distribution pattern in the residuals. Left-skew is also observed in the histogram for the residual data, also implying a deviation from normal distribution.

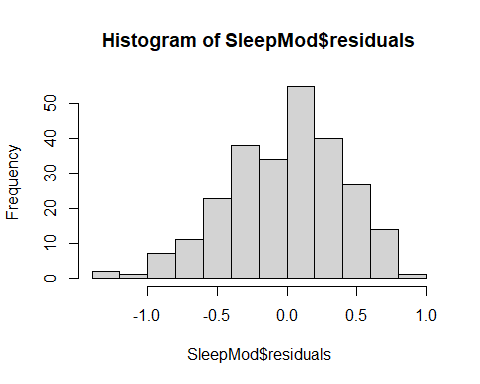
SleepMod = lm(GPA~WeekdaySleep, data=Sleep)  
plot(GPA~WeekdaySleep, data=Sleep)  
abline(SleepMod)



plot(SleepMod, 1:2)



hist(SleepMod$residuals)



### Transformations

R Notebook   
library(readr)   
  
sleep <- read\_csv("https://raw.githubusercontent.com/JA-  
McLean/STOR455/master/data/SleepStudy.csv")   
## Rows: 253 Columns: 27   
## -- Column specification --------------------------------------------------  
------   
## Delimiter: ","   
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse   
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA,   
ClassesMissed, ...   
##   
## i Use `spec()` to retrieve the full column specification for this data.   
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this   
message.   
#Transformations   
#Building a Model with a single predictors   
head(sleep)   
## # A tibble: 6 x 27   
## Gender ClassYear LarkOwl NumEarlyClass EarlyClass GPA ClassesMissed   
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>   
## 1 0 4 Neither 0 0 3.6 0   
## 2 0 4 Neither 2 1 3.24 0   
## 3 0 4 Owl 0 0 2.97 12   
## 4 0 1 Lark 5 1 3.76 0   
## 5 0 4 Owl 0 0 3.2 4   
## 6 1 4 Neither 0 0 3.5 0   
## # ... with 20 more variables: CognitionZscore <dbl>, PoorSleepQuality   
<dbl>,   
## # DepressionScore <dbl>, AnxietyScore <dbl>, StressScore <dbl>,   
## # DepressionStatus <chr>, AnxietyStatus <chr>, Stress <chr>, DASScore   
<dbl>,   
## # Happiness <dbl>, AlcoholUse <chr>, Drinks <dbl>, WeekdayBed <dbl>,   
## # WeekdayRise <dbl>, WeekdaySleep <dbl>, WeekendBed <dbl>, WeekendRise   
<dbl>,   
## # WeekendSleep <dbl>, AverageSleep <dbl>, AllNighter <dbl>   
1. Using CognitionZscore as a predictor for GPA, calculate the least squares regression   
line that best fits your data and plot this.

mod1 = lm(GPA ~ CognitionZscore, data = sleep)   
plot(GPA ~ CognitionZscore, data = sleep)   
abline(mod1)   
  
plot(mod1)

2. Experiment with some transformation(s) to attempt to find one that seems to do a   
better job of satisfying the linear model conditions. Plot this as well.   
mod2 = lm(log(GPA) ~ log(CognitionZscore), data = sleep)

### Outliers & Influential Points

Group 2 Questions

library(readr)  
sleep = read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

1. Start by building a model that predicts GPA based on the number of classes a student has missed, ClassesMissed

mod = lm(GPA~ClassesMissed, data = sleep)

1. Find and display the top 10 standardized residuals (be sure to look at absolute values when sorting). Of these, which data points may be considered influential?

Any values greater than 2 / less than -2 are considered to be influential. In our case, the top 9 data points have high standardized residuals.

head(sort(abs(rstandard(mod)), decreasing = TRUE), 10)

## 55 67 203 126 89 224 34 72   
## 3.198291 3.198291 3.140456 2.273283 2.250347 2.077591 2.037753 2.021602   
## 169 50   
## 2.021602 1.998586

1. Find and display the top 10 studentized residuals (be sure to look at absolute values when sorting). Of these, which data points may be considered influential?

Values we consider to be high: Any values greater than 2 / less than -2 In our case, all of the top 10 data points have high standardized residuals.

head(sort(abs(rstudent(mod)), decreasing = TRUE), 10)

## 55 67 203 126 89 224 34 72   
## 3.259012 3.259012 3.197646 2.292473 2.268865 2.091510 2.050724 2.034199   
## 169 50   
## 2.034199 2.010663

1. Based on only the information from the previous two questions, do we have any indications about data points that may exert significant influence on the model?

For question 4, I think the answer should be we have no evidence that the data points may have significant influence because standardized residuals and studentized residuals only tell us how far away each point deviates from the mean, they don’t tell us how influential each point is.

1. Briefly explain the difference between influence and leverage.

Influence is the effect of a single data point on the regression line, and represents how well the data point matches the of the rest of the points. Leverage represents how much influence a data point has on the regression line. It explains how far away the data point is from the other points (explains how much of an outlier it is)

For question 5, a better explanation of leverage is the “potential” to exert influence.

1. Briefly explain the difference between standardized and studentized residuals.

Standardized: Residuals of the data are standardized If a rstandard value is greater than 2 (or less than -2) it is a mild outlier. If the value is greater than 3 (or less than -3), it is a strong case of being an outlier. Studentized: An outlier is removed from the data, then standardized.

For question 6, the studentized residual is calculated by removing the data point whose studentized value you want to calculate, not removing an outlier.

1. Find and display the top 10 points with highest leverage. Do any points have significantly large leverage?

None of these data points appear to have high leverage.

2/12 # Average leverage (2/n)

## [1] 0.1666667

2 \* 2/12 # Points we might need to worry about

## [1] 0.3333333

3 \* 2/12 # Points we might need to worry about

## [1] 0.5

head(sort(hatvalues(mod), decreasing=TRUE), 10)

## 20 26 85 163 65 204 148   
## 0.12357263 0.12357263 0.06578311 0.06578311 0.05649286 0.05649286 0.04795851   
## 3 27 66   
## 0.04018004 0.02689076 0.02689076

1. Calculate the Cook’s Distance for the points in our data set. How many of these points may be considered influential outliers?

From class, we said to study any case with Di > 0.5; worry if Di > 1.0. In our data set, none of the cook’s distances exceed 0.5, so we have 0 influential outliers in our dataset.

head(sort(cooks.distance(mod), decreasing=TRUE),10)

## 65 26 55 67 163 245 34   
## 0.09610117 0.02662156 0.02314754 0.02314754 0.02079628 0.02033074 0.01966539   
## 203 89 126   
## 0.01965083 0.01476538 0.01341317

### Inference for Regression Slope/Correlation

STOR 455 Homework #9 Part 1

Inference for Regression Slope/Correlation

library(readr)  
SleepStudy <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

1. Create a model to predict GPA from the average amount of sleep and use the test for correlation and test for slope to test the strength of the linear relationship between your variables.

test for corelation: H0: p = 0 Hα: p != 0

test for slope: H0:β1 = 0 Hα:β1 != 0 According to the test for slope, our p-value is 0.337 while the p-value of the correlation test is 0.3368; Both of which do not give us enough evidence to reject the null hypothesis, which means that there isnt a strong relationship between GPA and AverageSleep.

mod1 = lm(GPA~AverageSleep, data = SleepStudy)  
cor.test(SleepStudy$GPA,SleepStudy$AverageSleep)

##   
## Pearson's product-moment correlation  
##   
## data: SleepStudy$GPA and SleepStudy$AverageSleep  
## t = -0.96238, df = 251, p-value = 0.3368  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.18259574 0.06316715  
## sample estimates:  
## cor   
## -0.06063317

summary(mod1)

##   
## Call:  
## lm(formula = GPA ~ AverageSleep, data = SleepStudy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.29018 -0.24395 0.03294 0.26317 0.80407   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.44618 0.21183 16.269 <2e-16 \*\*\*  
## AverageSleep -0.02541 0.02640 -0.962 0.337   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4043 on 251 degrees of freedom  
## Multiple R-squared: 0.003676, Adjusted R-squared: -0.000293   
## F-statistic: 0.9262 on 1 and 251 DF, p-value: 0.3368

### ANOVA for simple linear regression

Group\_Review

library(readr)  
sleep = read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

head(sleep)

## # A tibble: 6 x 27  
## Gender ClassYear LarkOwl NumEarlyClass EarlyClass GPA ClassesMissed  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 0 4 Neither 0 0 3.6 0  
## 2 0 4 Neither 2 1 3.24 0  
## 3 0 4 Owl 0 0 2.97 12  
## 4 0 1 Lark 5 1 3.76 0  
## 5 0 4 Owl 0 0 3.2 4  
## 6 1 4 Neither 0 0 3.5 0  
## # ... with 20 more variables: CognitionZscore <dbl>, PoorSleepQuality <dbl>,  
## # DepressionScore <dbl>, AnxietyScore <dbl>, StressScore <dbl>,  
## # DepressionStatus <chr>, AnxietyStatus <chr>, Stress <chr>, DASScore <dbl>,  
## # Happiness <dbl>, AlcoholUse <chr>, Drinks <dbl>, WeekdayBed <dbl>,  
## # WeekdayRise <dbl>, WeekdaySleep <dbl>, WeekendBed <dbl>, WeekendRise <dbl>,  
## # WeekendSleep <dbl>, AverageSleep <dbl>, AllNighter <dbl>

#Topic 4: ANOVA for simple linear regression  
#Question:Test the strength of the linear relationship between \_\_GPA\_\_ and \_\_AverageSleep\_\_ using ANOVA for regression. Include hypotheses and your conclusions in the context of the problem.

#Solution:  
mod1 = lm(GPA~AverageSleep, data = sleep)  
anova(mod1)

## Analysis of Variance Table  
##   
## Response: GPA  
## Df Sum Sq Mean Sq F value Pr(>F)  
## AverageSleep 1 0.151 0.15142 0.9262 0.3368  
## Residuals 251 41.037 0.16349

#Null Hypothesis: The coefficient for AverageSleep is 0.  
#Alternative Hypothesis: The coefficient for AverageSleep is not 0.  
#Conclusion: According to the table, p-value is 0.3368, which is a really large number. Therefore, we fail to reject the null hypothesis and come to the conclusion that the coefficient for AverageSleep is 0.

#Topic 17: Test of significance for nested models

#Question\_1:Construct a logistic model to predict if the student has an early class using Stress, AnxietyStatus, and the interaction between them as the predictor variables.

#Solution\_1:  
mod2 = glm(EarlyClass~Stress+AnxietyStatus+Stress\*AnxietyStatus,family = binomial, data = sleep)  
summary(mod2)

##   
## Call:  
## glm(formula = EarlyClass ~ Stress + AnxietyStatus + Stress \*   
## AnxietyStatus, family = binomial, data = sleep)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6651 -1.4944 0.8906 0.8906 1.7941   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.51083 0.36515 1.399 0.1618   
## Stressnormal 0.58779 0.59628 0.986 0.3243   
## AnxietyStatusnormal -0.04082 0.67700 -0.060 0.9519   
## AnxietyStatussevere 0.47000 0.76920 0.611 0.5412   
## Stressnormal:AnxietyStatusnormal -0.33774 0.84118 -0.402 0.6881   
## Stressnormal:AnxietyStatussevere -2.95491 1.43662 -2.057 0.0397 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 322.99 on 252 degrees of freedom  
## Residual deviance: 317.01 on 247 degrees of freedom  
## AIC: 329.01  
##   
## Number of Fisher Scoring iterations: 4

#Question2:Conduct a drop in deviance hypothesis test to determine the effectiveness of the interaction terms in the model constructed in the previous question. Cite your hypotheses, p-value, and conclusion in context.

#Solution2:  
reduced = glm(EarlyClass~Stress+AnxietyStatus,family = binomial, data = sleep)  
anova(mod2, reduced,test = "Chisq")

## Analysis of Deviance Table  
##   
## Model 1: EarlyClass ~ Stress + AnxietyStatus + Stress \* AnxietyStatus  
## Model 2: EarlyClass ~ Stress + AnxietyStatus  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 247 317.01   
## 2 249 322.16 -2 -5.1531 0.07604 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Null Hypothesis:The coefficient for the interaction terms is 0.  
#Alternative Hypothesis:The coefficient for the interaction terms is not 0.  
#Conclusion: According to the table, p-value is 0.07604, which is greater than 0.05. Therefore, we fail to reject the null hypothesis and come to the conclusion that the coefficient for interaction terms is 0.

### Confidence and Prediction Intervals

Homework 9 - Topic 5

Rhea Bhagia, Meghna Sharma, Olivia Voss, Connor Hammond

11/28/2021

SleepStudy <- read.csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

*Question*: Construct 90% confidence and prediction intervals to assess the interval for average hours of sleep for all days for a first year and fourth year student. Interpret your findings in context.

SleepStudy\_mod = lm(AverageSleep ~ factor(ClassYear), data=SleepStudy)  
  
first\_year = data.frame(ClassYear = 1)  
fourth\_year = data.frame(ClassYear = 4)  
   
predict.lm(SleepStudy\_mod, first\_year, interval = "confidence", level = 0.90 )

## fit lwr upr  
## 1 7.925319 7.691953 8.158685

predict.lm(SleepStudy\_mod, first\_year, interval = "prediction", level = 0.90 )

## fit lwr upr  
## 1 7.925319 6.30851 9.542129

predict.lm(SleepStudy\_mod, fourth\_year, interval = "confidence", level = 0.90 )

## fit lwr upr  
## 1 7.95 7.738091 8.161909

predict.lm(SleepStudy\_mod, fourth\_year, interval = "prediction", level = 0.90 )

## fit lwr upr  
## 1 7.95 6.336148 9.563852

We are 90% confident that a first year student will sleep on average 7.691953 to 8.158685 hours every night.

We can predict 90% of the time that a first year student will sleep on average between 6.30851 and 9.542129 hours a night.

We are 90% confident that a fourth year student will sleep on average 7.738091 to 8.161909 hours every night.

We can predict 90% of the time that a first year student will sleep on average between 6.336148 and 9.563852 hours a night.

## Multiple Regression

### Building a Model & Assessing Conditions

STOR 455 Homework #9

Due 11/29

library(readr)  
library(Stat2Data)  
"SleepStudy" <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

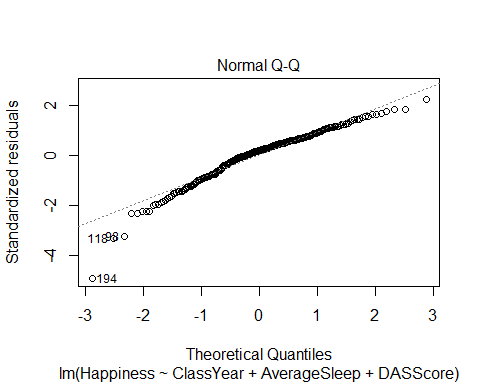
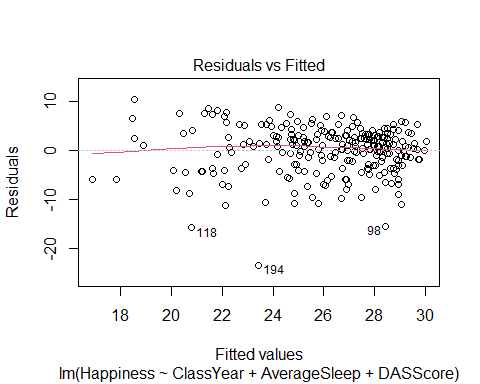
1. Multiple Regression: Building a Model & Assessing Conditions
2. Construct a linear model with *Happiness* as the response and *ClassYear* , *AverageSleep* , *DASScore* as the predictors. Include the output for the summary of the model.

mod1 = lm(Happiness~ClassYear + AverageSleep + DASScore, data=SleepStudy)  
summary(mod1)

##   
## Call:  
## lm(formula = Happiness ~ ClassYear + AverageSleep + DASScore,   
## data = SleepStudy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.4348 -2.8943 0.8034 3.0193 10.4424   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 25.34113 2.68833 9.426 < 2e-16 \*\*\*  
## ClassYear 0.39920 0.29229 1.366 0.173   
## AverageSleep 0.37165 0.31511 1.179 0.239   
## DASScore -0.15870 0.01839 -8.630 7.37e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.81 on 249 degrees of freedom  
## Multiple R-squared: 0.2412, Adjusted R-squared: 0.2321   
## F-statistic: 26.39 on 3 and 249 DF, p-value: 7.432e-15

1. Asses the conditions for linearity of the model you constructed in question 1

plot(mod1, 1:2)

 Looking at our residuals vs. fitted plot, we can see that there seem to be some issues with constant variance, as the data is more compact towards the larger fitted values. Again looking at the residuals vs. fitted plot, there are not issues with linearity as there is not a clear, defined non-linear pattern. Looking at the normal quantile plot, there seem to be issues of normality with the tails of the data, especially the lower tail, as it does not seem to quite fit a normal distribution.

### Methods for Choosing Predictors

R Notebook Group Project 3 - Topic 7: Methods for Choosing Predictors

library(readr)  
  
SleepData = read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

head(SleepData)

## # A tibble: 6 x 27  
## Gender ClassYear LarkOwl NumEarlyClass EarlyClass GPA ClassesMissed  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 0 4 Neither 0 0 3.6 0  
## 2 0 4 Neither 2 1 3.24 0  
## 3 0 4 Owl 0 0 2.97 12  
## 4 0 1 Lark 5 1 3.76 0  
## 5 0 4 Owl 0 0 3.2 4  
## 6 1 4 Neither 0 0 3.5 0  
## # ... with 20 more variables: CognitionZscore <dbl>, PoorSleepQuality <dbl>,  
## # DepressionScore <dbl>, AnxietyScore <dbl>, StressScore <dbl>,  
## # DepressionStatus <chr>, AnxietyStatus <chr>, Stress <chr>, DASScore <dbl>,  
## # Happiness <dbl>, AlcoholUse <chr>, Drinks <dbl>, WeekdayBed <dbl>,  
## # WeekdayRise <dbl>, WeekdaySleep <dbl>, WeekendBed <dbl>, WeekendRise <dbl>,  
## # WeekendSleep <dbl>, AverageSleep <dbl>, AllNighter <dbl>

library(leaps)  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/ShowSubsets.R")

Find the best multiple regression model predicting GPA using the following four method selections: all subsets, forward regression, backward regression, and stepwise regression.

1. All Subsets Method

all = regsubsets(GPA~., data=SleepData)

## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =  
## force.in, : 1 linear dependencies found

## Reordering variables and trying again:

ShowSubsets(all)

## Gender ClassYear LarkOwlNeither LarkOwlOwl NumEarlyClass EarlyClass  
## 1 ( 1 )   
## 2 ( 1 )   
## 3 ( 1 )   
## 4 ( 1 )   
## 5 ( 1 ) \*   
## 6 ( 1 ) \*   
## 7 ( 1 ) \* \*   
## 8 ( 1 ) \* \*   
## 9 ( 1 ) \* \*   
## ClassesMissed CognitionZscore PoorSleepQuality DepressionScore  
## 1 ( 1 )   
## 2 ( 1 ) \*   
## 3 ( 1 ) \*   
## 4 ( 1 ) \*   
## 5 ( 1 ) \*   
## 6 ( 1 ) \*   
## 7 ( 1 ) \*   
## 8 ( 1 ) \*   
## 9 ( 1 ) \* \*  
## AnxietyScore StressScore DepressionStatusnormal DepressionStatussevere  
## 1 ( 1 )   
## 2 ( 1 )   
## 3 ( 1 ) \*   
## 4 ( 1 ) \*   
## 5 ( 1 ) \*   
## 6 ( 1 ) \*   
## 7 ( 1 ) \*   
## 8 ( 1 ) \*   
## 9 ( 1 ) \*   
## AnxietyStatusnormal AnxietyStatussevere Stressnormal DASScore  
## 1 ( 1 )   
## 2 ( 1 )   
## 3 ( 1 )   
## 4 ( 1 ) \*  
## 5 ( 1 ) \*  
## 6 ( 1 ) \*  
## 7 ( 1 ) \*  
## 8 ( 1 ) \*  
## 9 ( 1 ) \*  
## Happiness AlcoholUseHeavy AlcoholUseLight AlcoholUseModerate Drinks  
## 1 ( 1 ) \*  
## 2 ( 1 ) \*  
## 3 ( 1 ) \*  
## 4 ( 1 ) \*  
## 5 ( 1 ) \*  
## 6 ( 1 ) \*  
## 7 ( 1 ) \*  
## 8 ( 1 ) \*  
## 9 ( 1 ) \*  
## WeekdayBed WeekdayRise WeekdaySleep WeekendBed WeekendRise  
## 1 ( 1 )   
## 2 ( 1 )   
## 3 ( 1 )   
## 4 ( 1 )   
## 5 ( 1 )   
## 6 ( 1 ) \*   
## 7 ( 1 ) \*   
## 8 ( 1 ) \* \*   
## 9 ( 1 ) \* \*  
## WeekendSleep AverageSleep AllNighter Rsq adjRsq Cp  
## 1 ( 1 ) 7.25 6.88 36.44  
## 2 ( 1 ) 12.43 11.73 22.49  
## 3 ( 1 ) 16.16 15.15 13.03  
## 4 ( 1 ) 19.71 18.41 4.11  
## 5 ( 1 ) 23.07 21.52 -4.25  
## 6 ( 1 ) 23.92 22.06 -4.85  
## 7 ( 1 ) 24.41 22.25 -4.37  
## 8 ( 1 ) 24.78 22.31 -3.50  
## 9 ( 1 ) 25.18 22.41 -2.73

ShowSubsetsModel = lm(GPA~Drinks+StressScore, data=SleepData)  
summary(ShowSubsetsModel)

##   
## Call:  
## lm(formula = GPA ~ Drinks + StressScore, data = SleepData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.29843 -0.24307 0.03535 0.23565 0.89868   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.298430 0.052105 63.30 < 2e-16 \*\*\*  
## Drinks -0.024851 0.005946 -4.18 4.04e-05 \*\*\*  
## StressScore 0.008848 0.003062 2.89 0.00419 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3845 on 250 degrees of freedom  
## Multiple R-squared: 0.1025, Adjusted R-squared: 0.09533   
## F-statistic: 14.28 on 2 and 250 DF, p-value: 1.346e-06

1. Forwards Selection Method

full = lm(GPA~., data=SleepData)  
none = lm(GPA~1, data=SleepData)  
MSE = (summary(full)$sigma)^2  
  
step(none, scope=list(upper=full), scale=MSE, direction="forward", trace=FALSE)

##   
## Call:  
## lm(formula = GPA ~ Drinks + CognitionZscore + StressScore + DASScore +   
## ClassYear + WeekdayRise, data = SleepData)  
##   
## Coefficients:  
## (Intercept) Drinks CognitionZscore StressScore   
## 3.79938 -0.01700 0.13410 0.03478   
## DASScore ClassYear WeekdayRise   
## -0.01288 -0.07212 -0.04114

ForwardModel = lm(GPA ~ Drinks + CognitionZscore + StressScore + DASScore +   
 ClassYear + WeekdayRise, data = SleepData)  
summary(ForwardModel)

##   
## Call:  
## lm(formula = GPA ~ Drinks + CognitionZscore + StressScore + DASScore +   
## ClassYear + WeekdayRise, data = SleepData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.21505 -0.19065 0.01833 0.24037 0.84845   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.799377 0.223741 16.981 < 2e-16 \*\*\*  
## Drinks -0.017005 0.005692 -2.987 0.003098 \*\*   
## CognitionZscore 0.134102 0.032478 4.129 4.99e-05 \*\*\*  
## StressScore 0.034780 0.007691 4.522 9.53e-06 \*\*\*  
## DASScore -0.012883 0.003648 -3.531 0.000494 \*\*\*  
## ClassYear -0.072116 0.021940 -3.287 0.001160 \*\*   
## WeekdayRise -0.041143 0.024927 -1.651 0.100113   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3569 on 246 degrees of freedom  
## Multiple R-squared: 0.2392, Adjusted R-squared: 0.2206   
## F-statistic: 12.89 on 6 and 246 DF, p-value: 1.17e-12

1. Backwards Regression Method

step(full, scale=MSE, trace=FALSE)

##   
## Call:  
## lm(formula = GPA ~ Gender + ClassYear + CognitionZscore + DepressionScore +   
## AnxietyScore + StressScore + Drinks + WeekdaySleep, data = SleepData)  
##   
## Coefficients:  
## (Intercept) Gender ClassYear CognitionZscore   
## 3.712435 -0.086172 -0.074506 0.128105   
## DepressionScore AnxietyScore StressScore Drinks   
## -0.009911 -0.018460 0.022651 -0.014895   
## WeekdaySleep   
## -0.029240

BackwardsModel = lm(GPA ~ Gender + ClassYear + CognitionZscore + DepressionScore +   
 AnxietyScore + StressScore + Drinks + WeekdaySleep, data = SleepData)  
summary(BackwardsModel)

##   
## Call:  
## lm(formula = GPA ~ Gender + ClassYear + CognitionZscore + DepressionScore +   
## AnxietyScore + StressScore + Drinks + WeekdaySleep, data = SleepData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.20958 -0.18108 0.02274 0.23235 0.79364   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.712435 0.175763 21.122 < 2e-16 \*\*\*  
## Gender -0.086172 0.052490 -1.642 0.101944   
## ClassYear -0.074506 0.022046 -3.380 0.000845 \*\*\*  
## CognitionZscore 0.128105 0.032871 3.897 0.000126 \*\*\*  
## DepressionScore -0.009911 0.004927 -2.012 0.045354 \*   
## AnxietyScore -0.018460 0.006133 -3.010 0.002890 \*\*   
## StressScore 0.022651 0.004556 4.972 1.25e-06 \*\*\*  
## Drinks -0.014895 0.006041 -2.466 0.014364 \*   
## WeekdaySleep -0.029240 0.019654 -1.488 0.138103   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3568 on 244 degrees of freedom  
## Multiple R-squared: 0.2457, Adjusted R-squared: 0.221   
## F-statistic: 9.937 on 8 and 244 DF, p-value: 5.801e-12

1. Stepwise Regression Method

step(none, scope=list(upper=full), scale=MSE, trace=FALSE)

##   
## Call:  
## lm(formula = GPA ~ Drinks + CognitionZscore + StressScore + DASScore +   
## ClassYear + WeekdayRise, data = SleepData)  
##   
## Coefficients:  
## (Intercept) Drinks CognitionZscore StressScore   
## 3.79938 -0.01700 0.13410 0.03478   
## DASScore ClassYear WeekdayRise   
## -0.01288 -0.07212 -0.04114

StepwiseModel = lm(GPA ~ Drinks + CognitionZscore + StressScore + DASScore +   
 ClassYear + WeekdayRise, data = SleepData)  
summary(StepwiseModel)

##   
## Call:  
## lm(formula = GPA ~ Drinks + CognitionZscore + StressScore + DASScore +   
## ClassYear + WeekdayRise, data = SleepData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.21505 -0.19065 0.01833 0.24037 0.84845   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.799377 0.223741 16.981 < 2e-16 \*\*\*  
## Drinks -0.017005 0.005692 -2.987 0.003098 \*\*   
## CognitionZscore 0.134102 0.032478 4.129 4.99e-05 \*\*\*  
## StressScore 0.034780 0.007691 4.522 9.53e-06 \*\*\*  
## DASScore -0.012883 0.003648 -3.531 0.000494 \*\*\*  
## ClassYear -0.072116 0.021940 -3.287 0.001160 \*\*   
## WeekdayRise -0.041143 0.024927 -1.651 0.100113   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3569 on 246 degrees of freedom  
## Multiple R-squared: 0.2392, Adjusted R-squared: 0.2206   
## F-statistic: 12.89 on 6 and 246 DF, p-value: 1.17e-12

### Interaction Terms

R Notebook

Group Members: Katherine Bacon Kay Youngstrom Ali Floyd Bethany Newcomb

library(readr)  
sleep <- read.csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

1. Interaction Terms

Create a model that predicts GPA based on Average Sleep.

mod1 = lm(GPA~ AverageSleep, data = sleep)  
summary(mod1)

##   
## Call:  
## lm(formula = GPA ~ AverageSleep, data = sleep)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.29018 -0.24395 0.03294 0.26317 0.80407   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.44618 0.21183 16.269 <2e-16 \*\*\*  
## AverageSleep -0.02541 0.02640 -0.962 0.337   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4043 on 251 degrees of freedom  
## Multiple R-squared: 0.003676, Adjusted R-squared: -0.000293   
## F-statistic: 0.9262 on 1 and 251 DF, p-value: 0.3368

Now create a model precting GPA based on Average Sleep, Classes Missed, and an interaction between these two variables.

mod2 = lm(GPA~ AverageSleep + ClassesMissed + AverageSleep\*ClassesMissed, data = sleep)  
summary(mod2)

##   
## Call:  
## lm(formula = GPA ~ AverageSleep + ClassesMissed + AverageSleep \*   
## ClassesMissed, data = sleep)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.33450 -0.23587 0.04836 0.23665 0.77052   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.5748832 0.2608322 13.706 <2e-16 \*\*\*  
## AverageSleep -0.0350662 0.0323790 -1.083 0.280   
## ClassesMissed -0.0310116 0.0454508 -0.682 0.496   
## AverageSleep:ClassesMissed 0.0009665 0.0057040 0.169 0.866   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3987 on 249 degrees of freedom  
## Multiple R-squared: 0.0388, Adjusted R-squared: 0.02722   
## F-statistic: 3.35 on 3 and 249 DF, p-value: 0.01968

Does the ClassesMissed term improve our model? Use Anova or another test we have learned in class. No need for a formal hypothesis.

anova(mod1, mod2)

## Analysis of Variance Table  
##   
## Model 1: GPA ~ AverageSleep  
## Model 2: GPA ~ AverageSleep + ClassesMissed + AverageSleep \* ClassesMissed  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 251 41.037   
## 2 249 39.590 2 1.4466 4.5492 0.01147 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The P-value is smaller than .05 which shows that it is significant. This means that at least one of the coefficients of ClassesMissed is non-zero and therefore makes our model better.

### Polynomial Models

STOR 455: Homework 9 Topic 9

1. Fit a quadratic model using *Happiness* to predict GPA and construct a scatterplot of the data with the quadratic fit.

library(readr)  
Sleep = read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

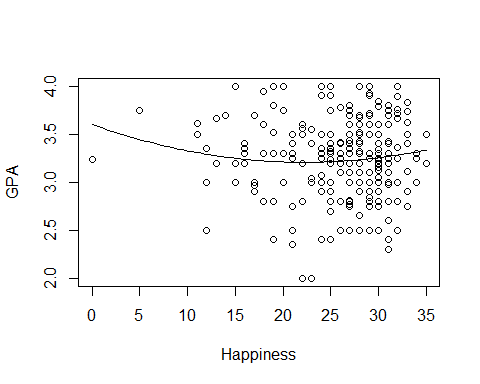
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

mod1 = lm(GPA~Happiness+I(Happiness^2), data=Sleep)  
summary(mod1)

##   
## Call:  
## lm(formula = GPA ~ Happiness + I(Happiness^2), data = Sleep)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.20924 -0.24386 0.03966 0.27400 0.78894   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.6035793 0.2724150 13.228 <2e-16 \*\*\*  
## Happiness -0.0353191 0.0240513 -1.468 0.143   
## I(Happiness^2) 0.0007902 0.0005248 1.506 0.133   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4041 on 250 degrees of freedom  
## Multiple R-squared: 0.008996, Adjusted R-squared: 0.001068   
## F-statistic: 1.135 on 2 and 250 DF, p-value: 0.3232

plot(GPA~Happiness, data=Sleep)  
curve(summary(mod1)$coef[3,1]\*x^2 + summary(mod1)$coef[2,1]\*x + summary(mod1)$coef[1,1], add=TRUE)



1. Determine if the fit would improve if a cubic term was included.

mod2 = lm(GPA~Happiness+I(Happiness^2)+I(Happiness^3), data=Sleep)  
anova(mod1, mod2)

## Analysis of Variance Table  
##   
## Model 1: GPA ~ Happiness + I(Happiness^2)  
## Model 2: GPA ~ Happiness + I(Happiness^2) + I(Happiness^3)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 250 40.818   
## 2 249 40.721 1 0.096534 0.5903 0.443

Doing a nested F-test, we have a p-value greater than .05 so adding the cubic term would not significantly improve the model.

### Second Order Models

R Notebook

This is the “Iris” dataset. Originally published at UCI Machine Learning Repository: Iris Data Set, this small dataset from 1936 is often used for testing out machine learning algorithms and visualizations (for example, Scatter Plot). Each row of the table represents an iris flower, including its species and dimensions of its botanical parts, sepal and petal, in centimeters.

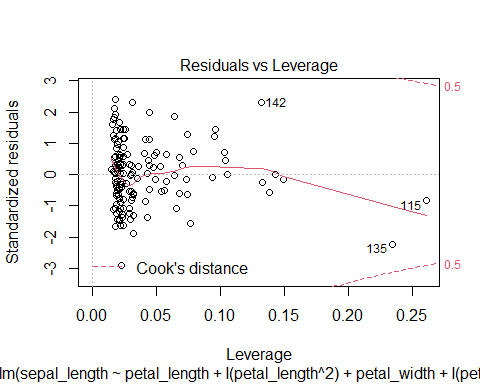
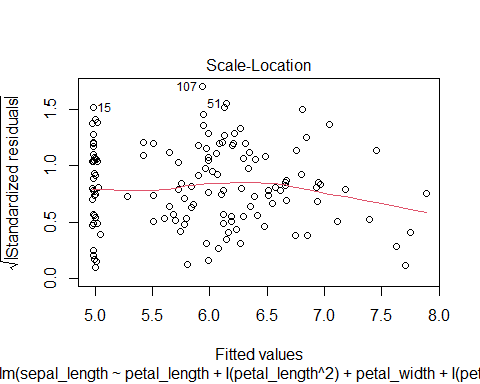
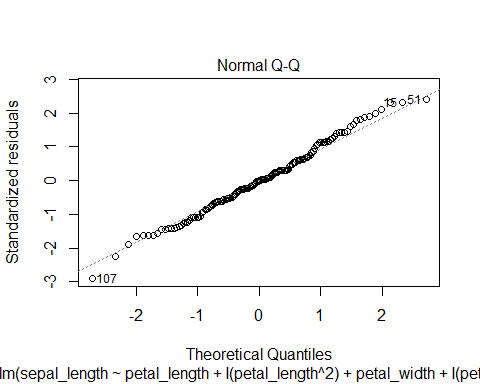
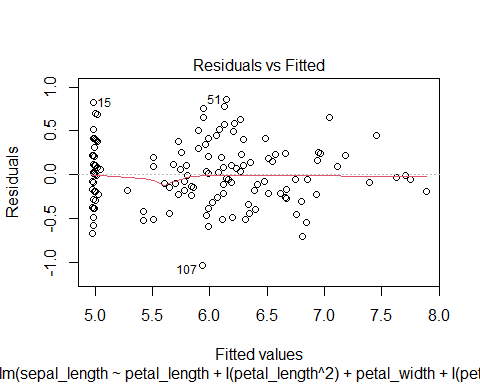
Iris = readr::read\_csv(  
 "https://gist.githubusercontent.com/curran/a08a1080b88344b0c8a7/raw/0e7a9b0a5d22642a06d3d5b9bcbad9890c8ee534/iris.csv", show\_col\_types = FALSE)

Q10: Fit a complete second order model for predicting a sepal\_length based on petal\_length and petal\_width and examine the residuals. Examine the conditions for your model. Discuss any issues that you see in the conditions.

Iris10 = lm(sepal\_length~petal\_length + I(petal\_length^2) + petal\_width + I(petal\_width^2)   
 + petal\_length\*petal\_width, data = Iris)  
summary(Iris10)

##   
## Call:  
## lm(formula = sepal\_length ~ petal\_length + I(petal\_length^2) +   
## petal\_width + I(petal\_width^2) + petal\_length \* petal\_width,   
## data = Iris)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.03541 -0.22444 -0.00044 0.21695 0.85811   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.06985 0.23422 21.646 <2e-16 \*\*\*  
## petal\_length -0.17961 0.30746 -0.584 0.560   
## I(petal\_length^2) 0.07061 0.09779 0.722 0.471   
## petal\_width 0.00520 0.68842 0.008 0.994   
## I(petal\_width^2) -0.27946 0.51092 -0.547 0.585   
## petal\_length:petal\_width 0.13631 0.44288 0.308 0.759   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3613 on 144 degrees of freedom  
## Multiple R-squared: 0.816, Adjusted R-squared: 0.8096   
## F-statistic: 127.7 on 5 and 144 DF, p-value: < 2.2e-16

plot(Iris10)



Overall, the residuals are constantly distributed with no obvious pattern. When fitted value is equal to 5.0, the residuals seem to be more compact.

Linear condition is satisfied since the fitted line is generally linear.

According to the qqplot, the points are generally lying on the fitted line, indicating satisfied normality condition.

### Multicollinearity & VIF

R Notebook

library(readr)  
Sleep = read.csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv  
")  
library(Stat2Data)  
data("CountyHealth") ##from Stat2Data

1. Create a linear model using GPA as the response and AlcoholUse, DepressionStatus, Happiness and Drinks as predictors.

mod1 = lm(GPA~AlcoholUse+Drinks+DepressionStatus+Happiness, data=Sleep)  
summary(mod1)

##   
## Call:  
## lm(formula = GPA ~ AlcoholUse + Drinks + DepressionStatus + Happiness,   
## data = Sleep)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.32735 -0.23308 0.03205 0.23060 0.85236   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.442522 0.148267 23.218 < 2e-16 \*\*\*  
## AlcoholUseHeavy 0.227005 0.152703 1.487 0.138   
## AlcoholUseLight 0.114326 0.087170 1.312 0.191   
## AlcoholUseModerate 0.148325 0.097688 1.518 0.130   
## Drinks -0.035619 0.008522 -4.180 4.07e-05 \*\*\*  
## DepressionStatusnormal -0.060470 0.076218 -0.793 0.428   
## DepressionStatussevere -0.194575 0.144806 -1.344 0.180   
## Happiness -0.002487 0.005029 -0.494 0.621   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.391 on 245 degrees of freedom  
## Multiple R-squared: 0.09052, Adjusted R-squared: 0.06454   
## F-statistic: 3.484 on 7 and 245 DF, p-value: 0.001405

1. Compute the VIF values for the model

library(car)

## Loading required package: carData

vif(mod1)

## GVIF Df GVIF^(1/(2\*Df))  
## AlcoholUse 2.029196 3 1.125176  
## Drinks 2.007391 1 1.416824  
## DepressionStatus 1.279483 2 1.063552  
## Happiness 1.255598 1 1.120535

Since the vif values are less than 5, this suggests there is not multicollinearity between the predictors, where multicollinearity is when one or more of the predictors is strongly correlated with some combination of the other predictors in the set.

##Using CountyHealth dataset 3. Construct a linear model using the number of Doctors in a county, MDs, as the response and number of hospitals in the county, Hospitals, and the number of beds in the county, Beds, as predictors

mod2 = lm(MDs~Hospitals+Beds, data=CountyHealth)  
summary(mod2)

##   
## Call:  
## lm(formula = MDs ~ Hospitals + Beds, data = CountyHealth)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1238.57 -291.40 79.93 200.92 1511.16   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -472.0761 135.3677 -3.487 0.00103 \*\*   
## Hospitals 117.3967 55.1110 2.130 0.03810 \*   
## Beds 1.2600 0.1435 8.778 1.07e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 494.8 on 50 degrees of freedom  
## Multiple R-squared: 0.9304, Adjusted R-squared: 0.9276   
## F-statistic: 334.1 on 2 and 50 DF, p-value: < 2.2e-16

1. Create a correlation matrix to look for multicollinearity between the Hospitals and Bed predictors

cor(CountyHealth[,3:4])

## Hospitals Beds  
## Hospitals 1.0000000 0.9094098  
## Beds 0.9094098 1.0000000

The correlation is 0.9 which is close to 1 so they are highly correlated and therefore explain the same kind of variability i.e. show multicollinearity.

1. Compute the VIF values for the model

library(car)  
vif(mod2)

## Hospitals Beds   
## 5.781221 5.781221

Since the vif values are greater than 5, this suggests multicollinearity. So you should consider removing one of the variables from the model.

### Nested F-Test

None

### Cross Validation Correlation

STOR 455 Cross Validation Correlation Practice

The data in FirstYearGPA contains information on 219 college students.

Variable Description HSGPA | High school GPA SATV | Verbal/critical reading SAT score SATM | Math SAT score Male | 1 for male, 0 for female HU | Number of credit hours earned in humanities courses in high school SS | Number of credit hours earned in social science courses in high school FirstGen | 1 if the student is the first in her or his family to attend college White | 1 for white students, 0 for others CollegeBound | 1 if attended a high school where of students intend to go on to college

library(Stat2Data)  
data("FirstYearGPA")

#Creating training sample  
GPATrain = data.frame(FirstYearGPA[c(0:150),])  
  
#Creating holdout sample  
GPAHoldout = data.frame(FirstYearGPA[c(150:219),])

1. Use the training sample to fit a multiple regression to predict GPA using HSGPA, HU, and White. Compute the predicted GPA for each case in the holdout sample using this model, then compute the residuals for each of the holdout cases.

GPAModel = lm(GPA~HSGPA + HU + factor(White), data=GPATrain)  
fitGPA = predict(GPAModel, newdata=GPAHoldout)  
GPAResid = GPAHoldout$GPA-fitGPA  
GPAResid

## 150 151 152 153 154   
## -0.6289464013 0.1774394377 0.7093971450 0.7199727824 -0.5703145067   
## 155 156 157 158 159   
## -0.4140782175 -0.0382432643 -0.2047372186 -0.6550454987 0.3798684276   
## 160 161 162 163 164   
## 0.4580618768 0.1985923190 0.2664629939 -0.2747119553 0.2351734504   
## 165 166 167 168 169   
## -0.7751471876 -0.2536553431 -0.0018953044 -0.3524895565 -0.3600627205   
## 170 171 172 173 174   
## 0.0141157790 0.7683507724 -0.7868886116 0.2257948302 0.6026659164   
## 175 176 177 178 179   
## -0.0881825223 -0.6084714179 0.4681877699 0.2216631324 -0.0818348277   
## 180 181 182 183 184   
## 0.1975249313 -0.3461618179 -0.6202984124 0.2034320264 -0.4424330626   
## 185 186 187 188 189   
## 0.1059798830 -0.4148954015 -0.4579081385 -0.3659294161 -0.3708417600   
## 190 191 192 193 194   
## -0.0751703326 0.1907910433 -0.5770129223 -0.1300272176 0.0181296694   
## 195 196 197 198 199   
## 0.4429532055 -0.5626864554 0.3596999810 0.2423307665 -0.2842874743   
## 200 201 202 203 204   
## -0.1221447143 0.1743707542 -0.6322232584 -0.5838442333 0.7272796483   
## 205 206 207 208 209   
## -0.7945570625 0.1053929468 -0.1978247748 -0.4516368696 0.4624243698   
## 210 211 212 213 214   
## 0.4690425865 0.4176008337 -0.2202564063 0.0842287668 0.0529403002   
## 215 216 217 218 219   
## 0.0408244100 -0.2425121782 -0.2950094662 -0.0006948518 -0.1901644095

1. Compute the mean and standard deviation for the residuals. Is the mean reasonably close to zero? Is the standard deviation reasonably close to the standard deviation of the error term from the fit to the training sample?

mean(GPAResid)

## [1] -0.06760761

sd(GPAResid)

## [1] 0.4092978

summary(GPAModel)

##   
## Call:  
## lm(formula = GPA ~ HSGPA + HU + factor(White), data = GPATrain)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.09844 -0.23079 0.03517 0.23600 0.82933   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.147478 0.311524 3.683 0.000323 \*\*\*  
## HSGPA 0.466053 0.088393 5.273 4.75e-07 \*\*\*  
## HU 0.015328 0.004091 3.747 0.000257 \*\*\*  
## factor(White)1 0.199174 0.076152 2.615 0.009846 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3773 on 146 degrees of freedom  
## Multiple R-squared: 0.2842, Adjusted R-squared: 0.2695   
## F-statistic: 19.32 on 3 and 146 DF, p-value: 1.319e-10

The mean of -0.06760761 is reasonably close to 0. The standard deviation of the residuals of 0.4092978 is quite close to the standard deviation from our model of 0.3773. This points to our model being a good predictor of GPA in general, not just for our specific training sample.

1. Compute the cross-validation correlation between the actual and predicted GPA values for the cases in the holdout sample.

crosscor = cor(GPAHoldout$GPA, fitGPA)  
crosscor

## [1] 0.5795404

1. Square the cross-validation correlation and subtract from for the training sample to compute the shrinkage. Does it look like the training model works reasonably well for the holdout sample or has there been a considerable drop in the amount of variability explained?

crosscor^2

## [1] 0.3358671

summary(GPAModel)$r.squared - crosscor^2

## [1] -0.05166974

The shrinkage here is actually a negative value, which means that my model created from the training data actually predicts GPA from the holdout model better than the original training data. The values in the holdout model are predicted around 5% better than the training model values, which is not a considerable problem as it is less than 10% and does not point to a significant difference in the effectiveness of model prediction between two datasets.

## Logistic Regression

### Building a Model with a single predictors

Logistic Regression

### Building a Model with a single predictors

library(readr)  
sleep = read.csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

1. Construct a model predicting *AllNighter* by *ClassesMissed*. Include a summary of this model. Use this model to perform a hypothesis test to determine if there is significant evidence of a relationship between *AllNighter* and *ClassesMissed*.

allnighter = glm(AllNighter ~ ClassesMissed, family = binomial, data = sleep)  
summary(allnighter)

##   
## Call:  
## glm(formula = AllNighter ~ ClassesMissed, family = binomial,   
## data = sleep)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.1200 -0.5224 -0.4986 -0.4758 2.1140   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.12138 0.23121 -9.175 <2e-16 \*\*\*  
## ClassesMissed 0.09924 0.04587 2.164 0.0305 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 199.69 on 252 degrees of freedom  
## Residual deviance: 195.45 on 251 degrees of freedom  
## AIC: 199.45  
##   
## Number of Fisher Scoring iterations: 4

Hypothesis test

anova(allnighter, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: AllNighter  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 252 199.69   
## ClassesMissed 1 4.244 251 195.44 0.03939 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Null hypothesis: The slope to the relationship between *AllNighter* and *ClassesMissed* is zero.

Alternative hypothesis: The slope to the relationship between *AllNighter* and *ClassesMissed* is nonzero.

1. For a student that has missed 9 classes, what is the probability the model predicts that the student has pulled an all nighter?

sleepmod = glm(AllNighter~ClassesMissed, data=sleep, family=binomial)  
predict(sleepmod, data.frame(ClassesMissed = 9), type="response")

## 1   
## 0.2264939

### Building a Model with multiple predictors

Group Project #3

library(readr)  
SleepStudy <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

head(SleepStudy)

## # A tibble: 6 x 27  
## Gender ClassYear LarkOwl NumEarlyClass EarlyClass GPA ClassesMissed  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 0 4 Neither 0 0 3.6 0  
## 2 0 4 Neither 2 1 3.24 0  
## 3 0 4 Owl 0 0 2.97 12  
## 4 0 1 Lark 5 1 3.76 0  
## 5 0 4 Owl 0 0 3.2 4  
## 6 1 4 Neither 0 0 3.5 0  
## # ... with 20 more variables: CognitionZscore <dbl>, PoorSleepQuality <dbl>,  
## # DepressionScore <dbl>, AnxietyScore <dbl>, StressScore <dbl>,  
## # DepressionStatus <chr>, AnxietyStatus <chr>, Stress <chr>, DASScore <dbl>,  
## # Happiness <dbl>, AlcoholUse <chr>, Drinks <dbl>, WeekdayBed <dbl>,  
## # WeekdayRise <dbl>, WeekdaySleep <dbl>, WeekendBed <dbl>, WeekendRise <dbl>,  
## # WeekendSleep <dbl>, AverageSleep <dbl>, AllNighter <dbl>

1. Construct a logistic model to predict the *Stress* of a student using *GPA*, *AverageSleep*, and *ClassYear* as the predictor variable.

mod1 = glm(factor(Stress) ~ GPA + AverageSleep + ClassYear, data = SleepStudy, family = binomial)  
summary(mod1)

##   
## Call:  
## glm(formula = factor(Stress) ~ GPA + AverageSleep + ClassYear,   
## family = binomial, data = SleepStudy)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1843 0.4386 0.6356 0.7592 1.0841   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 5.64974 2.10371 2.686 0.00724 \*\*  
## GPA -1.15547 0.42717 -2.705 0.00683 \*\*  
## AverageSleep -0.01023 0.16400 -0.062 0.95026   
## ClassYear -0.20183 0.14690 -1.374 0.16946   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 267.47 on 252 degrees of freedom  
## Residual deviance: 258.64 on 249 degrees of freedom  
## AIC: 266.64  
##   
## Number of Fisher Scoring iterations: 4

1. For a sophomore student that has a 3.50 GPA, gets an average of 8 hours of sleep a night, what does your model predict is their stress level? Interpret.

'The student has a 0.75 chance of having a high level of stress.'

## [1] "The student has a 0.75 chance of having a high level of stress."

student = data.frame(GPA = 3.50, AverageSleep = 8, ClassYear = 2)  
predict(mod1, student, type="response")

## 1   
## 0.7540053

1. Construct a second logistic model to predict the *Stress* of a student using *GPA*, *AverageSleep*, and *ClassYear* as well as the interaction between *GPA* and *ClassYear* as the predictor variable.

mod2 = glm(factor(Stress) ~ GPA + AverageSleep + ClassYear + GPA\*ClassYear, data = SleepStudy, family = binomial)  
summary(mod2)

##   
## Call:  
## glm(formula = factor(Stress) ~ GPA + AverageSleep + ClassYear +   
## GPA \* ClassYear, family = binomial, data = SleepStudy)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2103 0.4257 0.6382 0.7649 1.0378   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.72521 3.89996 1.724 0.0846 .  
## GPA -1.47051 1.05151 -1.398 0.1620   
## AverageSleep -0.00903 0.16423 -0.055 0.9562   
## ClassYear -0.63892 1.33515 -0.479 0.6323   
## GPA:ClassYear 0.12824 0.38933 0.329 0.7419   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 267.47 on 252 degrees of freedom  
## Residual deviance: 258.53 on 248 degrees of freedom  
## AIC: 268.53  
##   
## Number of Fisher Scoring iterations: 4

### Tests of significance for a model (single or multiple predictors)

STOR 455 Homework #9 Part 1 #16

library(readr)  
Sleep <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Tests of significance for a model (single or multiple predictors)

1. Construct a logistic model to predict *Stress* using *WeekdaySleep* and *WeekendSleep* as predictor variables. Perform a hypothesis test to determine if there is significant evidence of a relationship between *Stress*, *WeekdaySleep*, and *WeekendSleep*. Cite your hypotheses, p-value, and conclusion in context.

H0: B1 = 0 HA: B1 != 0 The p-value estimated p-value for WeekdaySleep is 0.362 and the p-value for WeekendSleep is 0.186. This concludes that both tests are statistically insignificant at any level less than or equal to 0.1, because their p-value are greater than 0.1, and you should fail to reject the null.

mod2 = glm(factor(Stress)~WeekdaySleep+WeekendSleep, data = Sleep, family = binomial)  
summary(mod2)

##   
## Call:  
## glm(formula = factor(Stress) ~ WeekdaySleep + WeekendSleep, family = binomial,   
## data = Sleep)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0400 0.5892 0.6761 0.7319 1.0052   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 1.5883 1.3020 1.220 0.223  
## WeekdaySleep 0.1173 0.1288 0.911 0.362  
## WeekendSleep -0.1507 0.1139 -1.324 0.186  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 267.47 on 252 degrees of freedom  
## Residual deviance: 265.14 on 250 degrees of freedom  
## AIC: 271.14  
##   
## Number of Fisher Scoring iterations: 4

I think that unless specifically stated, we tend to use the 95 percent confidence interval which means that alpha, or the p-value to determine whether or not the null should be rejected would be .05. This would mean that the test is not statistically significant. I think that for the interpretation it could be put into context, like the slope would be zero or non-zero. I also think that the conclusion would need to be in context by saying that the slope for the logit model predicting stress by weekend and weekday sleep is zero, and that there is not statistically significant evidence from the model to say that there is a non-zero slope for stress with those variables. - Group 11: Haley Hawkins, Rafel Al Ghrary, Alyssa Warnock, Gloria Su.

### Test of significance for nested models

Group\_Review

library(readr)  
sleep = read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

head(sleep)

## # A tibble: 6 x 27  
## Gender ClassYear LarkOwl NumEarlyClass EarlyClass GPA ClassesMissed  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 0 4 Neither 0 0 3.6 0  
## 2 0 4 Neither 2 1 3.24 0  
## 3 0 4 Owl 0 0 2.97 12  
## 4 0 1 Lark 5 1 3.76 0  
## 5 0 4 Owl 0 0 3.2 4  
## 6 1 4 Neither 0 0 3.5 0  
## # ... with 20 more variables: CognitionZscore <dbl>, PoorSleepQuality <dbl>,  
## # DepressionScore <dbl>, AnxietyScore <dbl>, StressScore <dbl>,  
## # DepressionStatus <chr>, AnxietyStatus <chr>, Stress <chr>, DASScore <dbl>,  
## # Happiness <dbl>, AlcoholUse <chr>, Drinks <dbl>, WeekdayBed <dbl>,  
## # WeekdayRise <dbl>, WeekdaySleep <dbl>, WeekendBed <dbl>, WeekendRise <dbl>,  
## # WeekendSleep <dbl>, AverageSleep <dbl>, AllNighter <dbl>

#Topic 4: ANOVA for simple linear regression  
#Question:Test the strength of the linear relationship between \_\_GPA\_\_ and \_\_AverageSleep\_\_ using ANOVA for regression. Include hypotheses and your conclusions in the context of the problem.

#Solution:  
mod1 = lm(GPA~AverageSleep, data = sleep)  
anova(mod1)

## Analysis of Variance Table  
##   
## Response: GPA  
## Df Sum Sq Mean Sq F value Pr(>F)  
## AverageSleep 1 0.151 0.15142 0.9262 0.3368  
## Residuals 251 41.037 0.16349

#Null Hypothesis: The coefficient for AverageSleep is 0.  
#Alternative Hypothesis: The coefficient for AverageSleep is not 0.  
#Conclusion: According to the table, p-value is 0.3368, which is a really large number. Therefore, we fail to reject the null hypothesis and come to the conclusion that the coefficient for AverageSleep is 0.

#Topic 17: Test of significance for nested models

#Question\_1:Construct a logistic model to predict if the student has an early class using Stress, AnxietyStatus, and the interaction between them as the predictor variables.

#Solution\_1:  
mod2 = glm(EarlyClass~Stress+AnxietyStatus+Stress\*AnxietyStatus,family = binomial, data = sleep)  
summary(mod2)

##   
## Call:  
## glm(formula = EarlyClass ~ Stress + AnxietyStatus + Stress \*   
## AnxietyStatus, family = binomial, data = sleep)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6651 -1.4944 0.8906 0.8906 1.7941   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.51083 0.36515 1.399 0.1618   
## Stressnormal 0.58779 0.59628 0.986 0.3243   
## AnxietyStatusnormal -0.04082 0.67700 -0.060 0.9519   
## AnxietyStatussevere 0.47000 0.76920 0.611 0.5412   
## Stressnormal:AnxietyStatusnormal -0.33774 0.84118 -0.402 0.6881   
## Stressnormal:AnxietyStatussevere -2.95491 1.43662 -2.057 0.0397 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 322.99 on 252 degrees of freedom  
## Residual deviance: 317.01 on 247 degrees of freedom  
## AIC: 329.01  
##   
## Number of Fisher Scoring iterations: 4

#Question2:Conduct a drop in deviance hypothesis test to determine the effectiveness of the interaction terms in the model constructed in the previous question. Cite your hypotheses, p-value, and conclusion in context.

#Solution2:  
reduced = glm(EarlyClass~Stress+AnxietyStatus,family = binomial, data = sleep)  
anova(mod2, reduced,test = "Chisq")

## Analysis of Deviance Table  
##   
## Model 1: EarlyClass ~ Stress + AnxietyStatus + Stress \* AnxietyStatus  
## Model 2: EarlyClass ~ Stress + AnxietyStatus  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 247 317.01   
## 2 249 322.16 -2 -5.1531 0.07604 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Null Hypothesis:The coefficient for the interaction terms is 0.  
#Alternative Hypothesis:The coefficient for the interaction terms is not 0.  
#Conclusion: According to the table, p-value is 0.07604, which is greater than 0.05. Therefore, we fail to reject the null hypothesis and come to the conclusion that the coefficient for interaction terms is 0.

## Methods for Choosing Predictors

### Assessing linearity of the logit model

Assessing linearity of the logit model

1). Build a logistic model that predicts whether a student has pulled a “AllNighter”, based on their “WeekdaySleep”. Comment on the linearity of the model using a emplogitplot.

library(readr)  
library(Stat2Data)  
Sleep <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

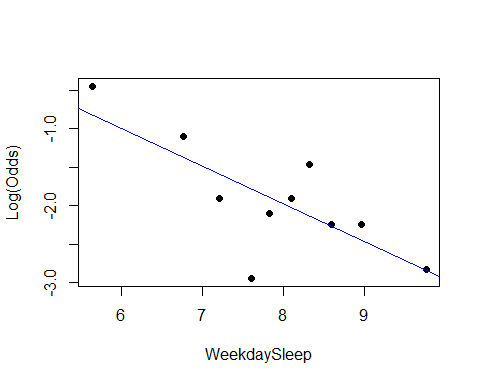
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

All\_Nighter = glm(AllNighter~WeekdaySleep, family=binomial, data=Sleep)  
summary(All\_Nighter)

##   
## Call:  
## glm(formula = AllNighter ~ WeekdaySleep, family = binomial, data = Sleep)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5137 -0.5438 -0.4543 -0.3618 2.5528   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.4767 1.1789 2.101 0.035656 \*   
## WeekdaySleep -0.5713 0.1579 -3.617 0.000298 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 199.69 on 252 degrees of freedom  
## Residual deviance: 185.78 on 251 degrees of freedom  
## AIC: 189.78  
##   
## Number of Fisher Scoring iterations: 5

emplogitplot1(AllNighter~WeekdaySleep, data=Sleep, ngroups=10)

 As we can see in our summary output, our WeekdaySleep term looks significant, and looking at the plot it seems as if the points follow a linear pattern, as there is no clear curve.

### Confidence interval for Odds Ratio

Stor455-Hw9

library(readr)  
library(Stat2Data)

data("Election16")  
head(Election16)

## State Abr Income HS BA Adv Dem.Rep TrumpWin  
## 1 Alabama AL 43623 84.3 23.5 8.7 -17 1  
## 2 Alaska AK 72515 92.1 28.0 10.1 -17 1  
## 3 Arizona AZ 50255 86.0 27.5 10.2 -1 1  
## 4 Arkansas AR 41371 84.8 21.1 7.5 -7 1  
## 5 California CA 61818 81.8 31.4 11.6 16 0  
## 6 Colorado CO 60629 90.7 38.1 14.0 -1 0

1. Run a logistic regression model to predict *TrumpWin* for each state using the per capita *Income* of the state. Print a summary of the model.
2. Find a 95% confidence interval for the odds ratio using the model constructed in question 1. (Explain if you wanna)

Answers

1. Run a logistic regression model to predict *TrumpWin* for each state using the per capita *Income* of the state. Print a summary of the model.

Election\_logitmod = glm(TrumpWin~Income,   
 data = Election16, family = binomial)  
summary(Election\_logitmod)

##   
## Call:  
## glm(formula = TrumpWin ~ Income, family = binomial, data = Election16)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2049 -0.7510 0.4074 0.6566 2.5000   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.118e+01 3.076e+00 3.635 0.000277 \*\*\*  
## Income -1.967e-04 5.582e-05 -3.523 0.000426 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 67.301 on 49 degrees of freedom  
## Residual deviance: 45.923 on 48 degrees of freedom  
## AIC: 49.923  
##   
## Number of Fisher Scoring iterations: 5

1. Find a 95% confidence interval for the odds ratio using the model constructed in question 1. (Explain if you wanna)

exp(confint.default(Election\_logitmod))

## 2.5 % 97.5 %  
## (Intercept) 173.032045 2.980697e+07  
## Income 0.999694 9.999127e-01

We are 95% confident that for every increase of one dollar, the person’s odds that Trump Wins changes by a factor between 0.999694 and 9.999127e-01.

### Interpreting Odds Ratio

R Notebook

Group Members: Katherine Bacon Kay Youngstrom Ali Floyd Bethany Newcomb

library(readr)  
sleep <- read.csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

Create a logistic model that predicts gender based on GPA.

mod1 = glm(Gender~GPA, data=sleep)

how does your model predict the odds of being male or female will change based on GPA?

summary(mod1)

##   
## Call:  
## glm(formula = Gender ~ GPA, data = sleep)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.7730 -0.3865 -0.2675 0.5243 0.8217   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.36768 0.24322 5.623 4.98e-08 \*\*\*  
## GPA -0.29734 0.07441 -3.996 8.47e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.2280315)  
##   
## Null deviance: 60.877 on 252 degrees of freedom  
## Residual deviance: 57.236 on 251 degrees of freedom  
## AIC: 347.97  
##   
## Number of Fisher Scoring iterations: 2

exp(summary(mod1)$coef[2])

## [1] 0.7427898

For every 1 point increase in GPA, the odds of being male increase by a factor of .74.

### Predicting Probabilities

Homework 9

Alyssa Warnock, Helen Johnston, Ashley Wade, Hunter Barbee

Sleep<-read.csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")  
head(Sleep)

## Gender ClassYear LarkOwl NumEarlyClass EarlyClass GPA ClassesMissed  
## 1 0 4 Neither 0 0 3.60 0  
## 2 0 4 Neither 2 1 3.24 0  
## 3 0 4 Owl 0 0 2.97 12  
## 4 0 1 Lark 5 1 3.76 0  
## 5 0 4 Owl 0 0 3.20 4  
## 6 1 4 Neither 0 0 3.50 0  
## CognitionZscore PoorSleepQuality DepressionScore AnxietyScore StressScore  
## 1 -0.26 4 4 3 8  
## 2 1.39 6 1 0 3  
## 3 0.38 18 18 18 9  
## 4 1.39 9 1 4 6  
## 5 1.22 9 7 25 14  
## 6 -0.04 6 14 8 28  
## DepressionStatus AnxietyStatus Stress DASScore Happiness AlcoholUse Drinks  
## 1 normal normal normal 15 28 Moderate 10  
## 2 normal normal normal 4 25 Moderate 6  
## 3 moderate severe normal 45 17 Light 3  
## 4 normal normal normal 11 32 Light 2  
## 5 normal severe normal 46 15 Moderate 4  
## 6 moderate moderate high 50 22 Abstain 0  
## WeekdayBed WeekdayRise WeekdaySleep WeekendBed WeekendRise WeekendSleep  
## 1 25.75 8.70 7.70 25.75 9.50 5.88  
## 2 25.70 8.20 6.80 26.00 10.00 7.25  
## 3 27.44 6.55 3.00 28.00 12.59 10.09  
## 4 23.50 7.17 6.77 27.00 8.00 7.25  
## 5 25.90 8.67 6.09 23.75 9.50 7.00  
## 6 23.80 8.95 9.05 26.00 10.75 9.00  
## AverageSleep AllNighter  
## 1 7.18 0  
## 2 6.93 0  
## 3 5.02 0  
## 4 6.90 0  
## 5 6.35 0  
## 6 9.04 0

1. For an female who has a stress score of 15, what does your model predict is the probability they have pulled an all nighter?

mod1=glm(AllNighter~Gender+AnxietyScore,family="binomial", data = Sleep)  
summary(mod1)

##   
## Call:  
## glm(formula = AllNighter ~ Gender + AnxietyScore, family = "binomial",   
## data = Sleep)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.9740 -0.6274 -0.4097 -0.3440 2.4140   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.85800 0.40034 -7.139 9.4e-13 \*\*\*  
## Gender 1.27236 0.39675 3.207 0.00134 \*\*   
## AnxietyScore 0.06036 0.03428 1.761 0.07824 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 199.69 on 252 degrees of freedom  
## Residual deviance: 187.23 on 250 degrees of freedom  
## AIC: 193.23  
##   
## Number of Fisher Scoring iterations: 5

newx=data.frame(Gender=0,AnxietyScore=15)  
predict(mod1,newx,type="response")

## 1   
## 0.1242689

For a female with an anxiety score of 15, the model predicts the probability of having had an all-nighter this semester is 0.1242689

## ANOVA for Means

### One Way ANOVA

R Notebook

This is the “Iris” dataset. Originally published at UCI Machine Learning Repository: Iris Data Set, this small dataset from 1936 is often used for testing out machine learning algorithms and visualizations (for example, Scatter Plot). Each row of the table represents an iris flower, including its species and dimensions of its botanical parts, sepal and petal, in centimeters.

Iris = readr::read\_csv(  
 "https://gist.githubusercontent.com/curran/a08a1080b88344b0c8a7/raw/0e7a9b0a5d22642a06d3d5b9bcbad9890c8ee534/iris.csv", show\_col\_types = FALSE)

Q23: Construct a one-way ANOVA model for the sepal length for each species of flower, using the species type as the predictor. Include the output showing the ANOVA table. Comment on what this output tells you about the sepal length across different spieces.

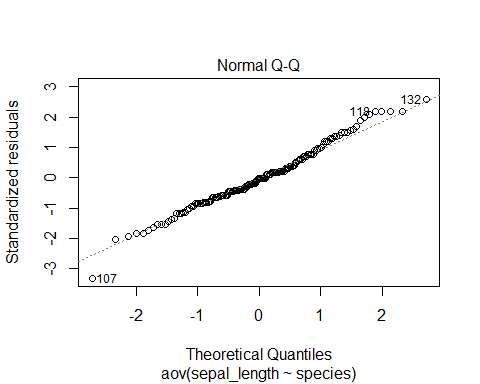
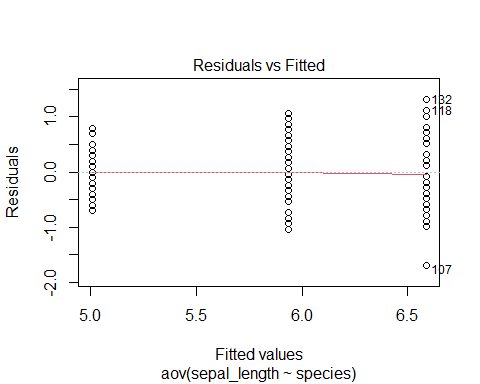
Iris23 = aov(sepal\_length~species, data = Iris)  
summary(Iris23)

## Df Sum Sq Mean Sq F value Pr(>F)   
## species 2 63.21 31.606 119.3 <2e-16 \*\*\*  
## Residuals 147 38.96 0.265   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

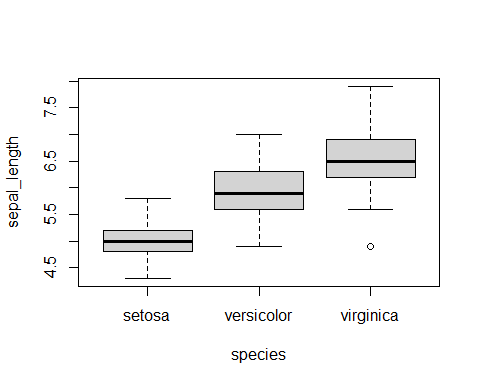
round(tapply(Iris$sepal\_length,Iris$species,sd),2)

## setosa versicolor virginica   
## 0.35 0.52 0.64

plot(Iris23, 1:2)



boxplot(sepal\_length~species, data = Iris)

 Since p-value is less than 0.05, we conclude that species is a significant predictor for sepal length. The spread of residuals is slightly increasing at each specie, however, there’s no specie that has a standard deviation of more than double of the other. From the qqplot, we can see that the points are generally following the fitted line as well. Therefore, there’s no big issue with constant variance and normalityï¼Œ which means species is an useful predictor in predicting sepal length.

### Two Way ANOVA

R Notebook

library(readr)  
  
Sleepstudy = read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Sleepstudy

## # A tibble: 253 x 27  
## Gender ClassYear LarkOwl NumEarlyClass EarlyClass GPA ClassesMissed  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 0 4 Neither 0 0 3.6 0  
## 2 0 4 Neither 2 1 3.24 0  
## 3 0 4 Owl 0 0 2.97 12  
## 4 0 1 Lark 5 1 3.76 0  
## 5 0 4 Owl 0 0 3.2 4  
## 6 1 4 Neither 0 0 3.5 0  
## 7 1 2 Lark 2 1 3.35 2  
## 8 0 2 Lark 0 0 3 0  
## 9 0 1 Neither 2 1 4 0  
## 10 0 4 Neither 2 1 2.9 0  
## # ... with 243 more rows, and 20 more variables: CognitionZscore <dbl>,  
## # PoorSleepQuality <dbl>, DepressionScore <dbl>, AnxietyScore <dbl>,  
## # StressScore <dbl>, DepressionStatus <chr>, AnxietyStatus <chr>,  
## # Stress <chr>, DASScore <dbl>, Happiness <dbl>, AlcoholUse <chr>,  
## # Drinks <dbl>, WeekdayBed <dbl>, WeekdayRise <dbl>, WeekdaySleep <dbl>,  
## # WeekendBed <dbl>, WeekendRise <dbl>, WeekendSleep <dbl>,  
## # AverageSleep <dbl>, AllNighter <dbl>

summary(Sleepstudy)

## Gender ClassYear LarkOwl NumEarlyClass   
## Min. :0.0000 Min. :1.000 Length:253 Min. :0.000   
## 1st Qu.:0.0000 1st Qu.:2.000 Class :character 1st Qu.:0.000   
## Median :0.0000 Median :2.000 Mode :character Median :2.000   
## Mean :0.4032 Mean :2.478 Mean :1.735   
## 3rd Qu.:1.0000 3rd Qu.:3.000 3rd Qu.:3.000   
## Max. :1.0000 Max. :4.000 Max. :5.000   
## EarlyClass GPA ClassesMissed CognitionZscore   
## Min. :0.000 Min. :2.000 Min. : 0.000 Min. :-1.62e+00   
## 1st Qu.:0.000 1st Qu.:3.000 1st Qu.: 0.000 1st Qu.:-4.80e-01   
## Median :1.000 Median :3.300 Median : 1.000 Median :-1.00e-02   
## Mean :0.664 Mean :3.244 Mean : 2.209 Mean :-3.95e-05   
## 3rd Qu.:1.000 3rd Qu.:3.500 3rd Qu.: 3.000 3rd Qu.: 4.40e-01   
## Max. :1.000 Max. :4.000 Max. :20.000 Max. : 1.96e+00   
## PoorSleepQuality DepressionScore AnxietyScore StressScore   
## Min. : 1.000 Min. : 0.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 4.000 1st Qu.: 1.000 1st Qu.: 1.000 1st Qu.: 3.000   
## Median : 6.000 Median : 3.000 Median : 4.000 Median : 8.000   
## Mean : 6.257 Mean : 5.202 Mean : 5.372 Mean : 9.466   
## 3rd Qu.: 8.000 3rd Qu.: 7.000 3rd Qu.: 8.000 3rd Qu.:14.000   
## Max. :18.000 Max. :35.000 Max. :26.000 Max. :37.000   
## DepressionStatus AnxietyStatus Stress DASScore   
## Length:253 Length:253 Length:253 Min. : 0.00   
## Class :character Class :character Class :character 1st Qu.: 7.00   
## Mode :character Mode :character Mode :character Median :16.00   
## Mean :20.04   
## 3rd Qu.:28.00   
## Max. :82.00   
## Happiness AlcoholUse Drinks WeekdayBed   
## Min. : 0.00 Length:253 Min. : 0.000 Min. :21.80   
## 1st Qu.:24.00 Class :character 1st Qu.: 3.000 1st Qu.:24.20   
## Median :28.00 Mode :character Median : 5.000 Median :24.80   
## Mean :26.11 Mean : 5.569 Mean :24.85   
## 3rd Qu.:30.00 3rd Qu.: 8.000 3rd Qu.:25.50   
## Max. :35.00 Max. :24.000 Max. :29.10   
## WeekdayRise WeekdaySleep WeekendBed WeekendRise   
## Min. : 5.500 Min. : 3.000 Min. :21.50 Min. : 5.25   
## 1st Qu.: 8.000 1st Qu.: 7.200 1st Qu.:24.88 1st Qu.: 9.25   
## Median : 8.500 Median : 7.950 Median :25.50 Median :10.25   
## Mean : 8.586 Mean : 7.866 Mean :25.58 Mean :10.20   
## 3rd Qu.: 9.150 3rd Qu.: 8.600 3rd Qu.:26.25 3rd Qu.:11.00   
## Max. :12.020 Max. :10.970 Max. :30.25 Max. :15.00   
## WeekendSleep AverageSleep AllNighter   
## Min. : 4.000 Min. : 4.950 Min. :0.0000   
## 1st Qu.: 7.250 1st Qu.: 7.430 1st Qu.:0.0000   
## Median : 8.250 Median : 8.000 Median :0.0000   
## Mean : 8.217 Mean : 7.966 Mean :0.1344   
## 3rd Qu.: 9.250 3rd Qu.: 8.590 3rd Qu.:0.0000   
## Max. :12.750 Max. :10.620 Max. :1.0000

1. Construct a two-way ANOVA model for the mean GPA in the dataset, *GPA*, using the *LarkOwl* and *Stress* as the predictors. Include the output showing the ANOVA table. Comment on what this output tells you about the GPA across college stduents. Provide formal hypotheses, p-values, and conclusions.

amodA = aov(GPA~LarkOwl+Stress, data=Sleepstudy)  
summary(amodA)

## Df Sum Sq Mean Sq F value Pr(>F)   
## LarkOwl 2 0.43 0.2150 1.347 0.262   
## Stress 1 1.02 1.0215 6.401 0.012 \*  
## Residuals 249 39.74 0.1596   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Hypotheses:

H0: μ1 = μ2 = μ3 (1= Lark, 2 = Owl, 3= Neither) (Mean GPA across LarkOwl status are all same) Ha: Some μi != μk (There is at least one mean GPA of one LarkOwl status that is different from the mean GPA of another LarkOwl status)

P-value: 0.262

Conclusion: Because p-value is not small enough (0.262), we fail to reject the null hypothesis. Mean GPA across LarkOwl statuses are all same.

Hypotheses:

H0: α1 = α2 = 0 (1 = Normal, 2 = High) (The affect for both stress scores are zeros) Ha: One of αi != 0 (There is at least one stress score that does not have zero for the affect)

p-value: 0.012

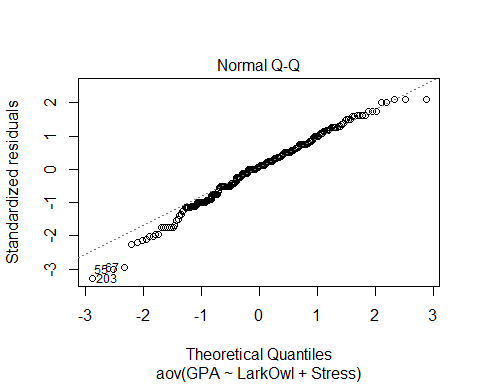
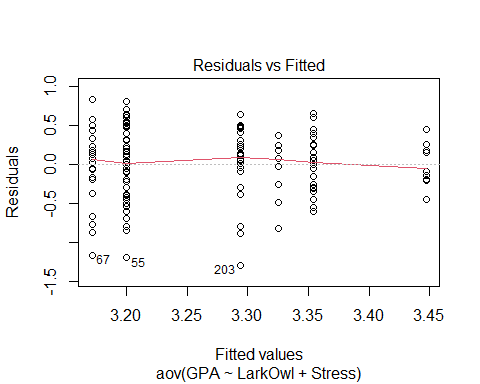
Conclusion: Because our p-value is small enough (0.012), we reject the null hypothesis. There is at least one stress score that does not have zero for the affect.

1. Construct residual plots and comment on the conditions of equality of variances and normality of residuals for the model that you created.

amodA

## Call:  
## aov(formula = GPA ~ LarkOwl + Stress, data = Sleepstudy)  
##   
## Terms:  
## LarkOwl Stress Residuals  
## Sum of Squares 0.43001 1.02147 39.73688  
## Deg. of Freedom 2 1 249  
##   
## Residual standard error: 0.399482  
## Estimated effects may be unbalanced

plot(amodA, 1:2)



tapply(Sleepstudy$GPA, Sleepstudy$LarkOwl, sd)

## Lark Neither Owl   
## 0.3972581 0.3935868 0.4413586

tapply(Sleepstudy$GPA, Sleepstudy$Stress, sd)

## high normal   
## 0.3513277 0.4123180

Variances: In general, we have constant variances, but as fitted values increase, the variances (vertical distances) are decreasing. In addition, the standard deviations for different combinations of LarkOwl and Stress predictors are roughly similar.

Normality: The normal Q-Q line shows some skewness on the left tail.

### Two Way ANOVA with Interaction

### Tukey HSD

library(readr)  
data=read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

1. Construct an ANOVA model for the mean AverageSleep value using LarkOwl and DepressionStatus as factors along with the interaction term and show the summary output. Is there evidence for a significant difference in mean AverageSleep values for any of the predictors?

amod1= aov(GPA~factor(ClassYear)+DepressionStatus+factor(ClassYear)\*DepressionStatus,data=data)  
summary(amod1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(ClassYear) 3 5.13 1.7109 11.606 3.93e-07 \*\*\*  
## DepressionStatus 2 0.22 0.1105 0.750 0.474   
## factor(ClassYear):DepressionStatus 6 0.31 0.0514 0.349 0.910   
## Residuals 241 35.53 0.1474   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

There is evidence for a significant difference in the mean AverageSleep value between class years but not for depression status or the interaction term.

1. Perform pairwise comparisons using Tukey HSD methods to show where any significant differences may occur among factors that showed a significant difference in part a).

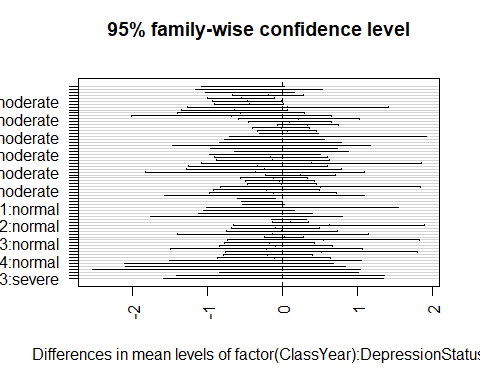
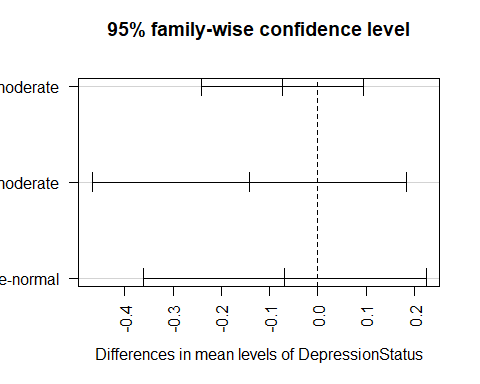
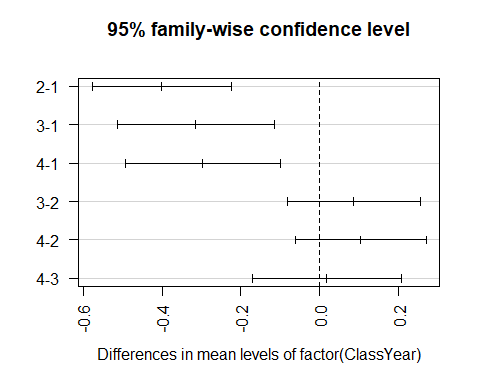
hsd = TukeyHSD(amod1)  
hsd$`factor(ClassYear)`[,4][hsd$`factor(ClassYear)`[,4]<.05] #significant differences in means between class groups

## 2-1 3-1 4-1   
## 9.727889e-08 3.285752e-04 6.721431e-04

There appear to be significant differences in mean AverageSleep values between class years 1:2, 1:3, and 1:4.

1. Produce 95% family-wise confidence level plots for the ANOVA model created in part a).

hsd=TukeyHSD(amod1)  
 plot(hsd,las=2)



### Levene’s Test

- Levene test tries to see if difference in group means when comparing things

- Useful, but may have some issues with

- asses how likely we get these diff by chance if we assume the two groups are similar

- TUkey looks at the group means to see wehre the differences are

- Leven only looks at evidence

- Only looked at levene in a one way ANOVA

- Tests to see if htere are difference in the variances,

More noets on Levene's Test

- Basically does an anova of differences rather than anova of abs values

- main point: see if looking bt the categories of teh variables are the abs deviances

- just testing to see if the variances are difference

- Tests for constant variance

- Use to check conditions