final

jason grahn

11/30/2018

library(tidyverse) #tidy tools  
library(psych)   
library(modelr) #modeling tools  
library(moderndive) #other modeling tools  
library(infer) #tidy inference statistics tools  
#library(corrgram)   
library(rms) #automate p-hacking / back/forward regression  
library(cowplot) #combining ggplots into groups  
library(broom) #modeling utilities

# The R Parts Section 1

## 1

### Download run10.csv (or run10.txt) and read corresponding data into R.

library(readr)  
R10 <- read\_csv("run10.csv", na = "NA")  
#glimpse(R10)

## 2

### Calculate the population mean and standard deviation for divTot. Apply rep function and for loop to collect 50 simple random samples of size 100 from divTot, and then use these stored statistics to calculate 50 confidence intervals of 95% confidence level for population mean. Include both the code/command and the output/graph.

population <- R10$divTot  
  
mean(population, na.rm = TRUE)

## [1] 1440.327

SD(population, na.rm = TRUE)

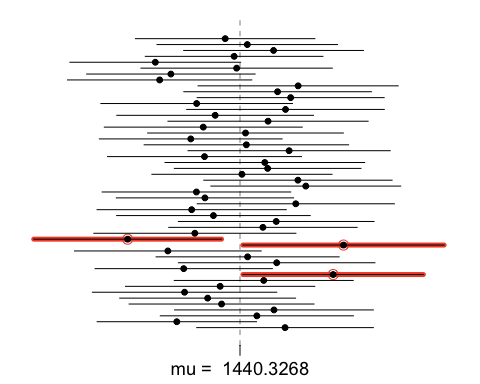
## [1] 811.1689

set.seed(40)  
  
ci.95 <- qnorm(1 - 0.05/2)  
  
samp\_mean <- rep(NA, 50)  
samp\_sd <- rep(NA, 50)  
  
n <- 100  
  
for(i in 1:50){  
 samp <- sample(population, n)   
 samp\_mean[i] <- mean(samp)   
 samp\_sd[i] <- sd(samp)   
}  
  
lower\_vector <- samp\_mean - ci.95 \* samp\_sd / sqrt(n)   
upper\_vector <- samp\_mean + ci.95 \* samp\_sd / sqrt(n)

## 3

### Apply function plot\_ci to display all the 50 confidence intervals in question (2). What proportion of your confidence intervals include the true population mean? Is this proportion consistent with the confidence level? Why or why not? Include both the code/command and the output/graph.

plot\_ci(lower\_vector, upper\_vector, mean(population, na.rm = TRUE))



p <- 1-(3/50)  
p

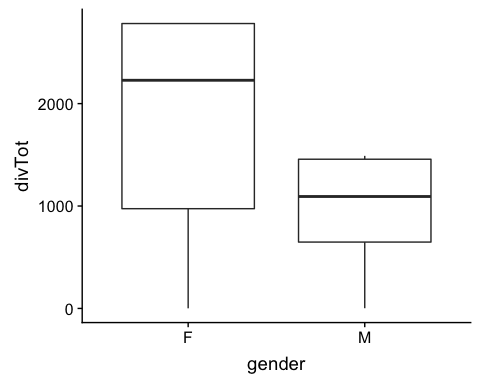
## [1] 0.94

Only 3 CIs do not include the population mean. Therefore, 94 percent of the CIs include the true population mean. This is fairly close to the selected confidence interval but not exact. This is due to the non-exact nature of confidence intervals. They predict a range that is *likely* to contain the true value (as likely as the CI level), but not always.

## 4

### Make a side-by-side boxplot of gender and divTot. What does the plot highlight about the relationship between these two variables? Include both the code/command and the output/graph.

R10 %>%   
 filter(is.na(divTot) == FALSE) %>%   
 ggplot() +  
 geom\_boxplot(aes(gender, divTot))



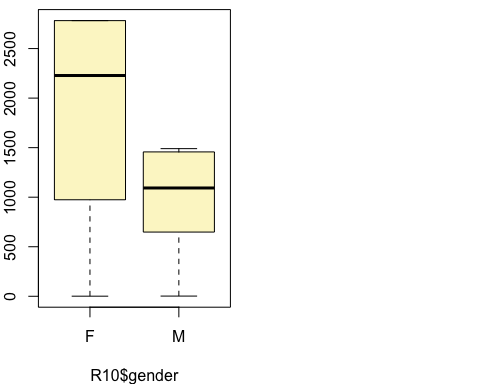
The boxplots showcase the differences in symmetry of divToT between Females and Males. Females have a much higher median than for Males as well as an increased distribution and range. We also see that neither category has any outliers.

## 5

### Calculate and a 95% confidence intervals for the difference between the mean of male divTot and the mean of female divTot, and interpret. Include both the code/command and the output/graph.

inference(y = R10$divTot,  
 x = R10$gender,   
 est="mean",   
 type="ci",  
 null=0,   
 alternative = "twosided",  
 method="theoretical")

## Response variable: numerical, Explanatory variable: categorical  
## Difference between two means  
## Summary statistics:  
## n\_F = 9732, mean\_F = 1747.493, sd\_F = 879.1942  
## n\_M = 7191, mean\_M = 1024.621, sd\_M = 449.3532



## Observed difference between means (F-M) = 722.8711  
## Standard error = 10.3685   
## 95 % Confidence interval = ( 702.5492 , 743.1931 )

We are 95% confident the difference that the mean of divTot between males and females is between 702.5492 and 743.1931.

## 6

### Conduct a hypothesis test at 95% significant level evaluating whether the mean of male divTot is different from the mean of female divTot? Make sure you indicate the hypotheses, the test rest, and the conclusion(s) clearly. Include both the code/command and the output/graph.

First we check for appropriate size for inferrence.

#how many values available from n to determine if   
r.rows <- nrow(R10)  
  
#test conditions  
#This also requires a sample size of np >= 10 and n(1-p)>=10 to determine if n is appropriate size  
R10\_condition\_1 <- if\_else((r.rows\*0.05) >= 10, "Sample size is large and passes condition 1", "Sample size is not large and does not pass condition 1")   
   
R10\_condition\_2 <- if\_else((r.rows\*0.95) >= 10, "Sample size is large and passes condition 2", "Sample size is not large and does not pass condition 2")  
  
R10\_condition\_1

## [1] "Sample size is large and passes condition 1"

R10\_condition\_2

## [1] "Sample size is large and passes condition 2"

• H0 = The average divTot of males and females the same. • HA = The average divTot of males and females is not the same.

males <- R10 %>%   
 filter(gender == "M") %>%   
 select(divTot)  
  
females <- R10 %>%   
 filter(gender == "F") %>%   
 select(divTot)  
  
t.test(females$divTot, males$divTot)

##   
## Welch Two Sample t-test  
##   
## data: females$divTot and males$divTot  
## t = 69.718, df = 15248, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 702.5476 743.1947  
## sample estimates:  
## mean of x mean of y   
## 1747.493 1024.621

With a studentized T-test score of 69.718 and p-value less than 2.2e-16, we reject the null hypothesis and can safely conclude that there is a significant difference between the male and female means of divTot.

# The R Parts Section 2

## 1

### Download smoking.csv (or smoking.txt) and read corresponding data into R. Example command in R: MyData<- read.csv(file=“path/TheDataIWantToReadIn.csv”,header=TRUE, sep=“,”). Note: use forward slash “/” instead of backward slash“” in the path. Make sure to include the code/command.

smoking <- read\_csv("smoking.csv")  
glimpse(smoking)

## Observations: 1,691  
## Variables: 12  
## $ gender <chr> "Male", "Female", "Male", "Female", "Fema...  
## $ age <int> 38, 42, 40, 40, 39, 37, 53, 44, 40, 41, 7...  
## $ maritalStatus <chr> "Divorced", "Single", "Married", "Married...  
## $ highestQualification <chr> "No Qualification", "No Qualification", "...  
## $ nationality <chr> "British", "British", "English", "English...  
## $ ethnicity <chr> "White", "White", "White", "White", "Whit...  
## $ grossIncome <chr> "2,600 to 5,200", "Under 2,600", "28,600 ...  
## $ region <chr> "The North", "The North", "The North", "T...  
## $ smoke <chr> "No", "Yes", "No", "No", "No", "No", "Yes...  
## $ amtWeekends <int> NA, 12, NA, NA, NA, NA, 6, NA, 8, 15, NA,...  
## $ amtWeekdays <int> NA, 12, NA, NA, NA, NA, 6, NA, 8, 12, NA,...  
## $ type <chr> NA, "Packets", NA, NA, NA, NA, "Packets",...

## 2

### Create a new data frame that contains only the rows in smoking.csv associated with male smokers (Male for gender and Yes for smoke).

male\_smokers <- subset(smoking,   
 smoking$gender == "Male" &   
 smoking$smoke == "Yes")  
  
male.smokers <- smoking %>%   
 filter(gender == "Male",  
 smoke == "Yes")

Then, calculate the proportion of Divorced in maritalStatus. Include both the code/command and the output/graph.

tot.smokers <- nrow(male.smokers)  
  
status.males <- male.smokers %>%   
 group\_by(maritalStatus) %>%   
 summarise(count = n(),  
 prop = count/tot.smokers)  
  
divorced.males <- status.males$prop[1]

The proportion of divorced maritial status within males is 0.0855615, or 8.55615%.

## 3

### Check if the conditions for inference are reasonable. If so, apply inference function to calculate the standard error and construct a 95% confidence interval for the proportion of Divorced in maritalStatus. Include both the code/command and the output/graph.

#how many values available from n to determine if   
m.rows <- nrow(male.smokers)  
  
#test conditions  
#This also requires a sample size of np >= 10 and n(1-p)>=10 to determine if n is appropriate size  
msmokers\_condition\_1 <- if\_else((m.rows \* divorced.males) >= 10,   
 "Sample size is large and passes condition 1",   
 "Sample size is not large and does not pass condition 1")   
   
msmokers\_condition\_2 <- if\_else((m.rows \* divorced.males) >= 10,   
 "Sample size is large and passes condition 2",   
 "Sample size is not large and does not pass condition 2")  
  
msmokers\_condition\_1

## [1] "Sample size is large and passes condition 1"

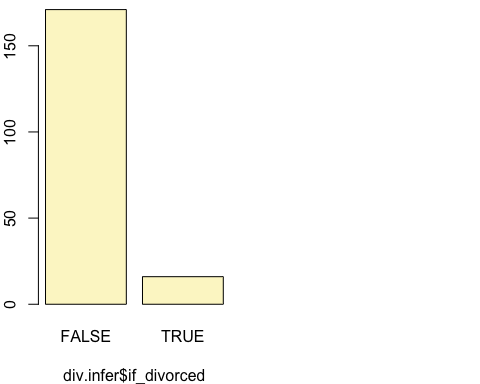
msmokers\_condition\_2

## [1] "Sample size is large and passes condition 2"

The volume of male smokers enough to pass both the first and second conditions for inference.

div.infer <- male.smokers %>%   
 mutate(if\_divorced = if\_else(maritalStatus == "Divorced", TRUE, FALSE))   
  
inference(div.infer$if\_divorced,  
 est = "proportion",  
 type = "ci",  
 method = "theoretical",  
 success = "TRUE")

## Single proportion -- success: TRUE   
## Summary statistics:



## p\_hat = 0.0856 ; n = 187   
## Check conditions: number of successes = 16 ; number of failures = 171   
## Standard error = 0.0205   
## 95 % Confidence interval = ( 0.0455 , 0.1257 )

The 95% confidence interval for the proportion of males that smoke that are also divorced is between 0.0455 and 0.1257. We can also state this as:

We are 95% confident that the true population proportion for males that smoke and are divorced is between 0.0455 and 0.1257.

## 4

### Based on the R output in question (3), what is the standard error and the margin of error for the estimate of the proportion of Divorced in maritalStatus?

#The standard error is part of the output in question 3.  
standard.error <- 0.0205  
  
#one way is to use the SE from above and manufacture the Margin of error  
margin.error <- ci.95 \* standard.error  
margin.error

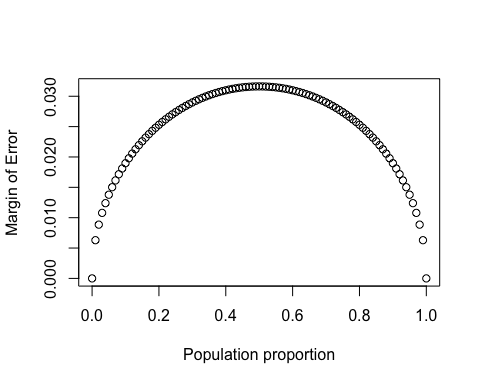
## [1] 0.04017926

The standard error is 0.0205. The margin of error is 0.0402

## 5

### Use simulation to show how the proportion affect the margin of error. Describe the relationship between the proportion and the margin of error. Include both the code/command and the output/graph.

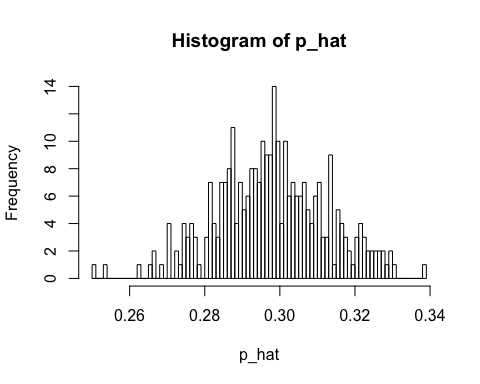
p <- seq(0,1,0.01)  
n <- 1000  
me <- 2\*sqrt(p\*(1-p)/n)  
  
plot(me~p, ylab="Margin of Error", xlab="Population proportion")

 This is nearly the exact same question we had in lab6. The relationship between proportion p and margin of error me is a non-linear arc. We see the proportion of 0.5 is the proportion with the peak margin of error, following the arc from 0 ME at 0 proportion to 0 ME at 1.0 proportion (100%).

## 6

### Apply for loop to simulate the process of drawing 300 samples of size 1000 from a population with a true Divorced proportion of 0.3. For each of the 3000 samples, compute p-hat and then plot a histogram to visualize the distribution. Describe the sampling distribution of the sample proportions. Be sure to note the center, spread, and shape clearly. Include both the code/command and the output/graph.

p<-0.3  
n<-1000  
p\_hat<-rep(0,300)  
  
set.seed(99)  
for(i in 1:300){  
 samp <- sample(c("divorced", "non-divorced"),  
 n,  
 replace = TRUE,  
 prob=c(p,1-p))  
 p\_hat[i] <- sum(samp=="divorced")/n}  
  
hist(p\_hat,breaks=100)



psych::describe(p\_hat)

## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 300 0.3 0.01 0.3 0.3 0.01 0.25 0.34 0.09 -0.05 0.03  
## se  
## X1 0

The distribution of the p-hat samples are approximately normal. The median and mean are 0.3, which is our population proportion of divorcees. The distribution has a standard deviation of 0.01. The range is 0.09 with a very slight left skew at -0.05.

# The R Parts Section 3

## 1

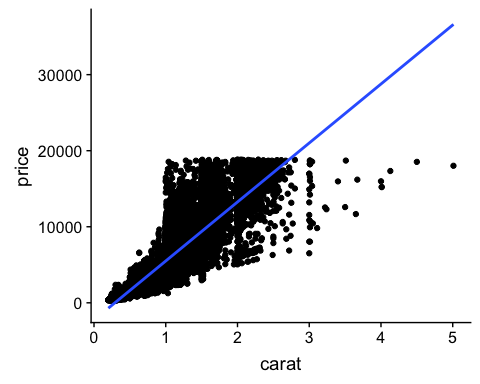
### Download diamonds.csv (or diamonds.txt) and read corresponding data into R. Make sure to include the code/command.

diamonds2 <- read\_csv("diamonds.csv")

## 2

### What type of plot is the most appropriate one to display the relationship between price and carat? Plot this relationship using the variable carat as the predictor. Does the relationship look linear? Include both the code/command and the output/graph.

diamonds2 %>%   
 ggplot(aes(carat, price)) +  
 geom\_point() +  
 geom\_smooth(method = lm)



The scatterplot is the most appropriate type of chart to show the relationship between two continuous variables. The relationship does *not* look linear as the price points ramps up rapidly from 1 carat onward. As we draw the linear line, we see that after 3 carats there’s no way a linear model would be able to predict the value of larger carat diamonds.

## 3

### If the relationship looks linear in question (2), what is the correlation coefficient? Interpret this result. Include both the code/command and the output/graph

The relationship does not appear linear as previously stated, but let’s run through these questions anyway.

cor(diamonds2$price, diamonds2$carat)

## [1] 0.9215913

The correlation coefficient between price and carat is approximately 0.922, which is a strong positive relationship. This is likely due to the clustering of data around the lower carat levels.

## 4

### Fit a simple linear model (use carat to predict price).

dm1 <- lm(price ~ carat,  
 data = diamonds2)  
  
summary(dm1)

##   
## Call:  
## lm(formula = price ~ carat, data = diamonds2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18585.3 -804.8 -18.9 537.4 12731.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2256.36 13.06 -172.8 <2e-16 \*\*\*  
## carat 7756.43 14.07 551.4 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1549 on 53938 degrees of freedom  
## Multiple R-squared: 0.8493, Adjusted R-squared: 0.8493   
## F-statistic: 3.041e+05 on 1 and 53938 DF, p-value: < 2.2e-16

### Using the estimates from the R output, write down the regression equation.

The regression equation is Y = -2256 + 7756\*X.

### What are the y-intercept and the slope?

The y-intercept is -2256.36 slope is 7756.43.

### Interpret the regression line.

With every positive unit change of 1 carat, price increases by 7756.43 units, minus the intercept value of -2256.36.

### Is variable carat significant? Why or why not?

Yes, the variable carat is significant because it has a p-value of <2e-16.

What is the coefficient of determination? Interpret the results. Include both the code/command and the output/graph. The coefficient of determination is r-squared at 0.8493, which states that approximately 84.93% of the variation in price is explained by the variable carat.

## 5

### Fit a multiple linear model with carat, depth, and table as independent variables, and variable price as dependent variable.

dm2 <- lm(price ~ carat + depth + table,   
 data = diamonds2)  
summary(dm2)

##   
## Call:  
## lm(formula = price ~ carat + depth + table, data = diamonds2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18288.0 -785.9 -33.2 527.2 12486.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13003.441 390.918 33.26 <2e-16 \*\*\*  
## carat 7858.771 14.151 555.36 <2e-16 \*\*\*  
## depth -151.236 4.820 -31.38 <2e-16 \*\*\*  
## table -104.473 3.141 -33.26 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1526 on 53936 degrees of freedom  
## Multiple R-squared: 0.8537, Adjusted R-squared: 0.8537   
## F-statistic: 1.049e+05 on 3 and 53936 DF, p-value: < 2.2e-16

### Using the estimates from the R output, write down the regression equation. What are the y-intercept and the slopes? Interpret the regression line.

The regression equation: Y= 13003.441 + 7858.771carat - 151.236depth - 104.473table.

The y-intercept is 13003.441, there are 3 slope factors:

\* carat: 7858.771  
\* depth: - 151.236  
\* table: - 104.473

When carat increases by 1, price increases by 7858.771; when depth increases by 1, price decreases by 151.236, and when table increase by 1, price decreases by 104.473.

### Are all the independent variables in the model significant? Why or why not?

Yes, it appears that all variables in the current model are significant as they each have a p-value less than <2e-16 and are valid predictors of price.

What is the coefficient of determination? Interpret the result, and compare it with question (4). The coefficient of determination is 0.8537 which states approximately 85.37% of the variation in price is explained by the variables in this model; carat, depth, table. This model slightly improves on the single-variable model developed in question (4) by about 0.44%.

## 6

### Using backward-selection and p-value as the selection criterion, determine the best model. You do not need to show all steps in your answer, just the output for the final model. Also, write down the final regression equation and interpret it. Include both the code/command and the output/graph.

#using the rms library to automate this  
library(rms)

#manually  
p.hack.backstep <- lm(price ~ carat + cut + color + clarity + depth + table + x,   
 data = diamonds2)  
summary(p.hack.backstep)

##   
## Call:  
## lm(formula = price ~ carat + cut + color + clarity + depth +   
## table + x, data = diamonds2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21385.0 -592.4 -183.7 376.5 10694.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2366.086 390.351 6.061 1.36e-09 \*\*\*  
## carat 11256.968 48.600 231.626 < 2e-16 \*\*\*  
## cutGood 580.240 33.572 17.283 < 2e-16 \*\*\*  
## cutIdeal 833.260 33.396 24.951 < 2e-16 \*\*\*  
## cutPremium 762.759 32.225 23.670 < 2e-16 \*\*\*  
## cutVery Good 726.820 32.212 22.564 < 2e-16 \*\*\*  
## colorE -209.237 17.893 -11.694 < 2e-16 \*\*\*  
## colorF -272.834 18.093 -15.080 < 2e-16 \*\*\*  
## colorG -481.943 17.716 -27.204 < 2e-16 \*\*\*  
## colorH -980.122 18.836 -52.035 < 2e-16 \*\*\*  
## colorI -1466.182 21.162 -69.283 < 2e-16 \*\*\*  
## colorJ -2369.504 26.131 -90.678 < 2e-16 \*\*\*  
## clarityIF 5344.338 51.015 104.761 < 2e-16 \*\*\*  
## claritySI1 3664.905 43.627 84.005 < 2e-16 \*\*\*  
## claritySI2 2702.077 43.812 61.674 < 2e-16 \*\*\*  
## clarityVS1 4577.589 44.535 102.786 < 2e-16 \*\*\*  
## clarityVS2 4266.612 43.847 97.308 < 2e-16 \*\*\*  
## clarityVVS1 5007.061 47.152 106.190 < 2e-16 \*\*\*  
## clarityVVS2 4950.168 45.847 107.972 < 2e-16 \*\*\*  
## depth -66.769 4.091 -16.322 < 2e-16 \*\*\*  
## table -26.457 2.911 -9.089 < 2e-16 \*\*\*  
## x -1029.478 20.549 -50.098 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1130 on 53918 degrees of freedom  
## Multiple R-squared: 0.9198, Adjusted R-squared: 0.9198   
## F-statistic: 2.944e+04 on 21 and 53918 DF, p-value: < 2.2e-16

#or this, the better way, using RMS  
kitchen.sink <- ols(price ~ ., data = diamonds2)  
rms.phack.ols <- fastbw(fit = kitchen.sink,   
 rule = "p",   
 sls = 0.05)  
rms.phack.ols

##   
## Deleted Chi-Sq d.f. P Residual d.f. P AIC R2   
## y 0.25 1 0.6192 0.25 1 0.6192 -1.75 0.92  
## z 2.08 1 0.1488 2.33 2 0.3117 -1.67 0.92  
##   
## Approximate Estimates after Deleting Factors  
##   
## Coef S.E. Wald Z P  
## Intercept 2366.09 390.350 6.061 1.349e-09  
## carat 11256.97 48.600 231.627 0.000e+00  
## cut=Good 580.24 33.572 17.283 0.000e+00  
## cut=Ideal 833.26 33.396 24.951 0.000e+00  
## cut=Premium 762.76 32.225 23.670 0.000e+00  
## cut=Very Good 726.82 32.211 22.564 0.000e+00  
## color=E -209.24 17.893 -11.694 0.000e+00  
## color=F -272.83 18.093 -15.080 0.000e+00  
## color=G -481.94 17.716 -27.204 0.000e+00  
## color=H -980.12 18.836 -52.036 0.000e+00  
## color=I -1466.18 21.162 -69.283 0.000e+00  
## color=J -2369.50 26.131 -90.679 0.000e+00  
## clarity=IF 5344.34 51.015 104.761 0.000e+00  
## clarity=SI1 3664.91 43.627 84.005 0.000e+00  
## clarity=SI2 2702.08 43.812 61.675 0.000e+00  
## clarity=VS1 4577.59 44.535 102.786 0.000e+00  
## clarity=VS2 4266.61 43.846 97.308 0.000e+00  
## clarity=VVS1 5007.06 47.152 106.190 0.000e+00  
## clarity=VVS2 4950.17 45.847 107.972 0.000e+00  
## depth -66.77 4.091 -16.322 0.000e+00  
## table -26.46 2.911 -9.089 0.000e+00  
## x -1029.48 20.549 -50.098 0.000e+00  
##   
## Factors in Final Model  
##   
## [1] carat cut color clarity depth table x

The final regression model is:

Price = 2366.09 + (carat)11256.97 + (cut=Good)580.24 + (cut=Ideal)833.26 + (cut=Premium)762.76 + (cut=VeryGood)726.82 - (color=E)209.24 - (color=F)272.83 - (color=G)481.94 -(color=H)980.12 -(color=I)1466.18 -(color=J)2369.5 + (clarity=IF)5344.34 + (clarity=SI1)3664.91 + (clarity=SI2)2702.08 + (clarity=VS1)4577.59 + (clarity=VS2)4266.61 + (clarity=VVS1)5007.06 + (clarity=VVS2)4950.17 -(depth)66.77 - (table)26.46 -(x)1029.48

The intercept of the final model is 2366.09.

Cut, color, and clarity are all categorical variables, which is why the model splits these out into separate variables depending on the categorical level.

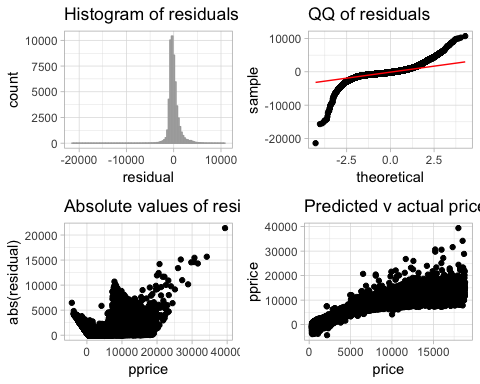
In this ordinary linear regression model, each slope next to a variable denotes the number of units that price will change changes when each variable changes by one unit. Because of the length of the model, I’ll summarize by example: when carat increases by one, price increases by 11256.97; when cut=Good increases by one, price increases by 580.24; if the color is colorJ, the price decreases by 2469.504; and so on.

## 7

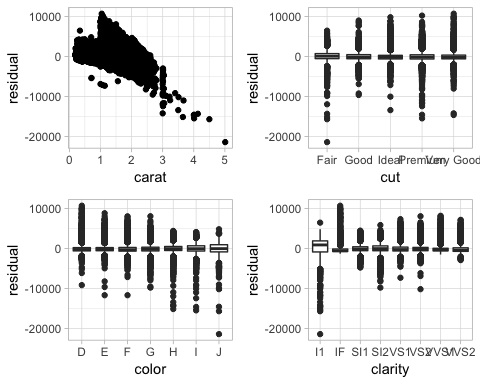
### What are the conditions to validate the model? Apply appropriate plots to check these conditions. Are all conditions passed? Is the model valid? Why or why not? Include both the code/command and the output/graph.

#need our regression points  
p\_hack\_regression\_points <- get\_regression\_points(p.hack.backstep)  
  
diamonds2 <- diamonds2 %>%   
 add\_residuals(p.hack.backstep, "residual") %>%   
 add\_predictions(p.hack.backstep, "pprice")

#histogram of resids  
histo <- diamonds2 %>%   
 ggplot() +   
 theme\_light() +  
 geom\_histogram(aes(residual), fill = "lightgrey", color = "darkgrey",   
 bins = 100) +  
 labs(title = "Histogram of residuals")   
  
#qqplot of resids  
qqplot <- diamonds2 %>%   
 ggplot(aes(sample = residual)) +  
 theme\_light() +  
 stat\_qq() +  
 stat\_qq\_line(color = "red") +  
 labs(title = "QQ of residuals")  
  
absolute <- diamonds2 %>%   
 ggplot() +   
 theme\_light() +  
 geom\_point(aes(pprice, abs(residual))) +   
 labs(title = "Absolute values of residuals against predicted values")  
  
#predictions vs actuals  
predict <- diamonds2 %>%   
 ggplot() +   
 theme\_light() +  
 geom\_point(aes(price, pprice)) +  
 labs(title = "Predicted v actual price")  
  
plot\_grid(histo, qqplot, absolute, predict)



#scatter distribtuion of residuals  
resid <- diamonds2 %>%   
 ggplot() +  
 theme\_light() +   
 geom\_point(aes(carat, residual))  
  
#boxplots of categorical variables vs residuals  
bcut <- ggplot(diamonds2, aes(cut, residual)) +   
 theme\_light() +  
 geom\_boxplot()   
  
bcolor <- ggplot(diamonds2, aes(color, residual)) +   
 theme\_light() +  
 geom\_boxplot()   
  
bclarity <- ggplot(diamonds2, aes(clarity, residual)) +   
 theme\_light() +  
 geom\_boxplot()   
  
plot\_grid(resid, bcut, bcolor, bclarity)



* The residuals of the model are nearly normal in the histogram with a slightly positive right skew (skew factor of 0.58)
* The QQ plot leads me to believe this model is not trustworthy for predicting values on the upper and lower quadrants.
* We assume independence based on study method.
* Boxplots show linearity related to categorical variables, but there certainly are a lot of outliers.
* Carat vs residuals is certainly not evenly distributed, but it is the strongest predictor we have for the values that it DOES serve.
* The Predicted v actual price plot shows us we probably could have done a log transform on some object for better predictions.

We have likely overfit the model. Overall, I think this model is acceptable for only a small, centered, bundle of the data. it does not accurately predict final price for the outer ends of prices.

## 8

### Based on your final model, what are the most important factors that influence the selling price of a diamond? Why?

No code is required to answer. Based on the final model, the most *important* factors for the selling price of a diamond are the carat of the diamond (the weight), the clarity, the cut, and the color; in that order. Carat has the largest point estimate change of all the supplied variables, clarity adds value to price in thousands and only increases value, cut also only adds to price but less-so than clarity; and finally color.