The Chi-square test statistic is: $\chi^2_{STAT} = \sum_{all\ cells} \frac{(f_o - f_e)^2}{f_e}$ $f_e = \frac{\text{row total} \times \text{column total}}{n}$

The test statistic for the McNemar test: $Z_{STAT} = \frac{B - C}{\sqrt{B + C}}$

McNemar test Contingency Table:

	Condition 2		
Condition 1	Yes	No	Totals
Yes	A	B	A+B
No	C	D	C+D
Totals	A+C	B+D	n

Wilcoxon Rank-Sum Test for Large Sample

$$Z_{STAT} = \frac{T_1 - \mu_{T_1}}{\sigma_{T_1}} = \frac{T_1 - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}}$$

The Kruskal-Wallis H-test statistic:

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^{c} \frac{T_j^2}{n_j}\right] - 3(n+1)$$

Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \hat{Y}_i = b_0 + b_1 X_i$$

Total variation is made up of two parts: SST = SSR + SSE

Coefficient of Determination

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression } \textit{sum} \text{ of squares}}{\textit{total} \text{ sum of squares}}$$

Inference about the slope: t test with df = n -2

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}}$$

F test for overall significance with df1 = k, df2 = n – k -1

$$F_{STAT} = \frac{MSR}{MSE}$$, $$MSR = \frac{SSR}{k}$$ $$MSE = \frac{SSE}{n-k-1}$$

Confidence interval estimate about the slope

$$b_1 \pm t_{\alpha/2} S_{b_1}$$

Multiple Regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \, , \; \hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}$$

Adjusted R square

$$r_{adj}^2 = 1 - \left[(1 - r^2) \left(\frac{n-1}{n-k-1} \right) \right]$$

Variance Inflation Factor

$$VIF_j = \frac{1}{1 - R_j^2}$$

Nonlinear Regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$$

Logistic Regression

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1, \pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Regression ANOVA table

Source	**DF**	**Sum of Sq**	**Mean Sq**	**F Value**	**Pr(>F)**
Regression	k	SSR	MSR	F	p-value
Residual	n-k-1	SSE	MSE		
Total	n-1	SST			

R functions for exam 2

Chi-square

```
chisq.test(dataframe[,i:j], correct =FALSE)
```

McNemar test (two-tailed test)

```
x<-matrix(c(A,B,C,D),2,2)
mcnemar.test(x, correct = FALSE)
```

Wilcoxon Signed-Rank Test

```
wilcox.test(Growth_Value$‘Growth‘, mu=5, alternative="greater")
```

Wilcoxon Signed-Rank Test for Matched-Pairs

```
wilcox.test(Growth_Value$'Growth‘, Growth_Value$'Value‘, alternative="two.sided“, paired=TRUE)
```

Wilcoxon Rank-Sum Test for Independent Samples

```
wilcox.test(Undergrad_Salaries$'Computer Science‘, Undergrad_Salaries$'Finance‘,
alternative="two.sided“, paired=FALSE)
```

Kruskal-Wallis Rank Test

```
Stacked2<- melt(KWexample)
colnames(Stacked2)<- c("Major", "Size")
Stacked2
kruskal.test(Stacked2$'Size',Stacked2$'Major')
```

Calculate and interpret the correlation coefficient between debt payments and income.

```
cor(Debt_Payments$'Income', Debt_Payments$'Debt’)
cor(Debt_Payments[2:4], use = "all.obs")
```

Create plot

```
install.packages("tidyverse")
library(tidyverse)
ggplot (data = Debt_Payments) + geom_point (mapping = aes (x = Income, y = Debt))
```

Simple Linear Regression

```
Simple <– lm(Debt~Income, data=Debt_Payments)
summary(Simple)
anova(Simple)
```

Multiple Linear Regression

```
Multiple2 <- lm(pie$'pie sales'~pie$'price ($)'+pie$'advertising ($100s)')
summary(Multiple2)
MR1  <- lm(pie$'pie sales'~ 1)
anova(MR1, Multiple2)
```

Model with dummy variable

```
GNV<- ifelse(GNV_JAX_Jan2022$city == "Gainesville, Florida", 1, 0)
mlr2 <- lm(GNV_JAX_Jan2022$PRICE ~ GNV_JAX_Jan2022$`SQUARE FOOTAGE` + GNV)
summary(mlr2)
```

Model with interaction variable

GNV_Int<- ifelse(GNV_JAX_Jan2022$city == "Gainesville, Florida", GNV_JAX_Jan2022$`SQUARE FOOTAGE`, 0)
mlr2_b <- lm(GNV_JAX_Jan2022$PRICE ~ GNV_JAX_Jan2022$`SQUARE FOOTAGE` + GNV+GNV_Int)
summary(mlr2_b)

Create residual plot

plot(mlr2_b)

Calculate VIF

install.packages("car")
library(car)
lmobject1 <- lm(PRICE ~ BEDS+ SQFT + BEDSANDBATHS+ LOTSIZE, data = Tampa2022)
summary(lmobject1)
vif(lmobject1)

Logistic regression

logmod = glm(Purchase~Age, family = binomial, data = MacysPurchases)
summary(logmod)

Build model using stepwise method

none <-lm(price ~1, data = GainesvilleHomes_Sp2019_Quant)
full <- lm(price ~ beds_baths + square_footage + lot_size+commute + year_built + es_dist + ms_dist + hs_dist, data = GainesvilleHomes_Sp2019_Quant)
MSE <- (summary(full)$sigma)^2
step(none, scope=list(upper= full), scale=MSE) (#by default, it uses stepwise method)