#Linear and Logistic Analysis

#Comparing answers to Q7 rows to whether or not a respondent actually *used* a service

> LogisticReg1 <- glm(Q4N ~ Q7.1N + Q7.2N + Q7.3N, data= X2022FSULibSurveyDataFinalVersion, family = "binomial")
> summary(LogisticReg1)

Call:
glm(formula = Q4N ~ Q7.1N + Q7.2N + Q7.3N, family = "binomial",
        data = CombinedX042122FSULibrariesData)

Deviance Residuals:
        Min   1Q   Median   3Q   Max
-1.10363 -0.76733 -0.45546 -0.08427   2.15288

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.7805   3.7439 -1.811   0.0701 .
Q7.1N   -0.1699   1.0365 -0.164   0.8698
Q7.2N   2.0377   1.1074   1.840   0.0658 .
Q7.3N   -0.7261   0.7021 -1.034   0.3010

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 49.127 on 47 degrees of freedom
Residual deviance: 42.368 on 44 degrees of freedom
  (8 observations deleted due to missingness)
AIC: 50.368

Number of Fisher Scoring iterations: 6

#Inference note for the above output; Q7.2N is a pretty good predictor of whether or not someone will use any data services or not at 90% confidence. The other two questions are not statistically significant for whether or not someone will use a research data service. (The intercept is also statistically significant at 90% confidence, although that's not a predictor variable and doesn't offer any useful insights.)

#Comparing answers to Q7.2, years in college, and major type to whether or not a respondent actually *used* a service

#Here's another logistic regression to see if someone would use a service based on responses to Q7.2, how many years they have been in college, and their major type

```
> LogisticReg2 <- glm(formula = Q4N ~ Q7.2N + Q1N + RankedQ2NV1,
data=X2022FSULibSurveyDataFinalVersion, family = "binomial")
> summary(LogisticReg2)
```

Call:
glm(formula = Q4N ~ Q7.2N + Q1N + RankedQ2NV1, family = "binomial",
        data = X2022FSULibSurveyDataFinalVersion)

Deviance Residuals:
       Min   1Q   Median   3Q   Max
-1.22666 -0.70484 -0.46846 -0.00007   2.00203

Coefficients:
          Estimate Std. Error z value Pr(>|z|)
(Intercept) -26.0197 2091.7522 -0.012   0.990
Q7.2N   1.5785   0.9193   1.717 0.086 .
Q1N   0.3959   0.3919   1.010 0.312
RankedQ2NV11 16.6589 2091.7475   0.008   0.994
RankedQ2NV12 15.6738 2091.7476   0.007   0.994

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

      Null deviance: 44.584 on 43 degrees of freedom
Residual deviance: 35.750 on 39 degrees of freedom
  (12 observations deleted due to missingness)
AIC: 45.75

Number of Fisher Scoring iterations: 17

#How a student responded to Question 7.2 is still statistically significant at the 90% confidence interval for whether a student will use a research data service or not.

#Comparing use of research data services to having previously hearing about research data services AND years in college

#This seems obvious, but I would like to see what the statistical significance is for someone using a service in comparison to how many years they have been in college, as well as whether or not they have heard of a research data service. (My hypothesis is that a student would be more likely to hear about a research data service if they have been in college for longer.)

```
> LogisticReg3 <-glm(formula= Q4N ~ Q1N + RankedQ3N,
data=X2022FSULibSurveyDataFinalVersion, family = "binomial")
> summary(LogisticReg3)
```

Call:
glm(formula = Q4N ~ Q1N + RankedQ3N, family = "binomial", data =
X2022FSULibSurveyDataFinalVersion)

Deviance Residuals:
	Min	1Q	Median	3Q	Max
	-0.84057	-0.76928	-0.70243	-0.00012	1.79017

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-18.8278	1805.0859	-0.010	0.992
Q1N	0.1038	0.2445	0.424	0.671
RankedQ3N	17.4504	1805.0858	0.010	0.992

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 55.486 on 55 degrees of freedom
Residual deviance: 48.720 on 53 degrees of freedom
AIC: 54.72

Number of Fisher Scoring iterations: 17

#There does not appear to be any statistically significant evidence that years in college or previously hearing about a research data service impacts whether or not someone will use a research data service with a logistic model. However, what's interesting is that this is not the case if one uses standard linear regression.

```
> LinearReg1 <-lm(formula= Q4N ~ Q1N + RankedQ3N)
> summary(LinearReg1)
```

Call:
glm(formula = Q4N ~ Q1N + RankedQ3N)

Deviance Residuals:
```
        Min       1Q    Median        3Q       Max
-0.28734  -0.25688  -0.22641   0.02226   0.78882
```

Coefficients:
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03750    0.14100  -0.266   0.7913
Q1N          0.01523    0.03643   0.418   0.6775
RankedQ3N    0.24867    0.12535   1.984   0.0525 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for gaussian family taken to be 0.1539457)

Null deviance: 8.8393 on 55 degrees of freedom
Residual deviance: 8.1591 on 53 degrees of freedom
AIC: 59.053

Number of Fisher Scoring iterations: 2

#For OLS regression, it appears that whether or not someone has heard of a data service is statistically significant in comparison to whether they will proceed to *use* a service, but only with 90% confidence. Years in college isn't statistically significant in this context.

#Comparing use of a data service to previously hearing about a data service ONLY

> LinearReg2 <- lm(formula= Q4N ~ RankedQ3N)
> summary(LinearReg2)

Call:
glm(formula = Q4N ~ RankedQ3N)

Deviance Residuals:
       Min   1Q Median   3Q   Max
-0.2558 -0.2558 -0.2558 0.0000 0.7442

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.967e-16 1.080e-01   0.000   1.0000
RankedQ3N 2.558e-01 1.232e-01 2.076   0.0427 *

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.1515935)

      Null deviance: 8.8393 on 55 degrees of freedom
Residual deviance: 8.1860 on 54 degrees of freedom
AIC: 57.238

Number of Fisher Scoring iterations: 2

#For linear regression, the p-value for *hearing* about a research data service to *using* a research data service *improves* if one leaves out the major type as a variable, with statistical significance at the 95% confidence level.

> LogisticReg4 <- glm(formula = Q4N ~ Q3N, data=X2022FSULibSurveyDataFinalVersion, family = "binomial")
> summary(LogisticReg4)

Call:
glm(formula = Q4N ~ Q3N, family = "binomial", data = X2022FSULibSurveyDataFinalVersion)

Deviance Residuals:
       Min   1Q   Median   3Q   Max
-0.76872 -0.76872 -0.76872 -0.00013 1.65124

Coefficients:

```
          Estimate Std. Error z value Pr(>|z|)
(Intercept)  -18.57    1809.05  -0.01      0.992
Q3N           17.50  1809.05      0.01   0.992
```

(Dispersion parameter for binomial family taken to be 1)

```
        Null deviance: 55.486  on 55  degrees of freedom
Residual deviance: 48.902  on 54  degrees of freedom
AIC: 52.902
```

Number of Fisher Scoring iterations: 17

#For logistic regression, there is no statistical significance between hearing or not hearing about research data services and then proceeding to use it. (Then again, Q3 is also a yes/no (i.e; 0/1) variable, so it might not be as robust in the context of logarithms.)

#**Comparing Q7.X Statements to academic standing in linear regression**
> AnotherLinReg1 <- lm(formula= Q7.1N ~ Q1N)
> summary(AnotherLinReg1)

Call:
lm(formula = Q7.1N ~ Q1N)

Residuals:
```
       Min     1Q  Median    3Q     Max
-3.2620 -0.2125 -0.1488  0.7946  0.8512
```

Coefficients:
```
         Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.29033    0.31038  13.823   <2e-16 ***
Q1N          -0.02832      0.09550  -0.296       0.768
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9762 on 46 degrees of freedom
 (8 observations deleted due to missingness)
Multiple R-squared: 0.001907,   Adjusted R-squared: -0.01979
F-statistic: 0.08791 on 1 and 46 DF, p-value: 0.7682

> AnotherLinReg2 <- lm(formula = Q7.2N ~ Q1N)
> summary(AnotherLinReg2)

Call:

lm(formula = Q7.2N ~ Q1N)

Residuals:
```
       Min    1Q Median    3Q    Max
-3.3188 -0.2745  0.6558  0.7318  0.7825
```

Coefficients:
```
           Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.34417    0.32280  13.458   <2e-16 ***
Q1N         -0.02532    0.09932  -0.255      0.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 1.015 on 46 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared: 0.001411,   Adjusted R-squared: -0.0203
F-statistic: 0.06501 on 1 and 46 DF, p-value: 0.7999

> AnotherLinReg3 <- lm(formula = Q7.3N ~ Q1N)
> summary(AnotherLinReg3)

Call:
lm(formula = Q7.3N ~ Q1N)

Residuals:
```
       Min    1Q Median    3Q    Max
-3.1846 -0.1312 -0.0243  0.8287  1.0291
```

Coefficients:
```
           Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.23809    0.31577   13.42   <2e-16 ***
Q1N         -0.05344    0.09716   -0.55    0.585
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.9931 on 46 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared: 0.006534,   Adjusted R-squared: -0.01506
F-statistic: 0.3025 on 1 and 46 DF, p-value: 0.585

#Comparing Q7.X Statements to major type in linear regression

> MoreLinReg1 <- lm(formula = Q7.1N ~ Q2NV1)
> summary(MoreLinReg1)

Call:
lm(formula = Q7.1N ~ Q2NV1)

Residuals:
     Min   1Q Median   3Q   Max
-3.4099 -0.2511 -0.1717 0.6298 0.9077

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.0923   0.2453 16.684   <2e-16 ***
Q2NV1   0.1588   0.1842   0.862 0.394

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8479 on 42 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared: 0.01738,   Adjusted R-squared: -0.006012
F-statistic: 0.743 on 1 and 42 DF, p-value: 0.3936

> MoreLinReg2 <- lm(formula = Q7.2N ~ Q2NV1)
> summary(MoreLinReg2)

Call:
lm(formula = Q7.2N ~ Q2NV1)

Residuals:
     Min   1Q Median   3Q   Max
-3.5837 -0.3026 0.4163 0.6974 0.9785

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.0215   0.2534 15.873   <2e-16 ***
Q2NV1   0.2811   0.1903   1.477 0.147

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8758 on 42 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared: 0.0494,   Adjusted R-squared: 0.02676
F-statistic: 2.183 on 1 and 42 DF, p-value: 0.147

> MoreLinReg3 <- lm(formula = Q7.3N ~ Q2NV1)
> summary(MoreLinReg3)

Call:
lm(formula = Q7.3N ~ Q2NV1)

Residuals:
      Min   1Q Median   3Q   Max
-3.2661 -0.2661 -0.1159 0.7339 1.0343

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.9657   0.2553 15.535 <2e-16 ***
Q2NV1   0.1502   0.1917 0.783 0.438

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8824 on 42 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared: 0.0144,   Adjusted R-squared: -0.009062
F-statistic: 0.6138 on 1 and 42 DF, p-value: 0.4378

> MoreLinReg4 <- lm(formula = Q7.1N ~ Q2NV2)
> summary(MoreLinReg4)

Call:
lm(formula = Q7.1N ~ Q2NV2)

Residuals:
      Min   1Q Median   3Q   Max
-3.4243 -0.3113 -0.1417 0.6040 0.9148

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.0852   0.2298 17.780 <2e-16 ***
Q2NV2   0.1130 0.1152 0.981 0.332

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8457 on 42 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared: 0.02241,   Adjusted R-squared: -0.0008702
F-statistic: 0.9626 on 1 and 42 DF, p-value: 0.3321

> MoreLinReg5 <- lm(formula = Q7.2N ~ Q2NV2)
> summary(MoreLinReg5)

Call:
lm(formula = Q7.2N ~ Q2NV2)

Residuals:
       Min   1Q Median   3Q   Max
-3.5926 -0.4049 0.4074 0.5951 0.9705

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0295   0.2371 16.995   <2e-16 ***
Q2NV2   0.1877   0.1189   1.579   0.122

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8727 on 42 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared: 0.05602,   Adjusted R-squared: 0.03354
F-statistic: 2.492 on 1 and 42 DF, p-value: 0.1219

> MoreLinReg6 <- lm(formula = Q7.3N ~ Q2NV2)
> summary(MoreLinReg6)

Call:
lm(formula = Q7.3N ~ Q2NV2)

Residuals:
       Min   1Q Median   3Q   Max
-3.2121 -0.2121 -0.0991 0.7879 0.9574

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.04260   0.24085 16.785   <2e-16 ***
Q2NV2   0.05652   0.12077   0.468   0.642

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8865 on 42 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared: 0.005187,   Adjusted R-squared: -0.0185
F-statistic: 0.219 on 1 and 42 DF, p-value: 0.6422

#Comparing the use of a data service to years in college ONLY
#Using years as a categorical variable to get specific details on each year

> LinearReg3 <- lm(formula= Q4N ~ RankedQ1N)
> summary(LinearReg3)

Call:
glm(formula = Q4N ~ RankedQ1N)

Deviance Residuals:
        Min   1Q Median   3Q   Max
-0.3333 -0.2353 -0.2000 0.0000 0.8000

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.33333   0.23600   1.412   0.164
RankedQ1N1 -0.33333   0.27250 -1.223   0.227
RankedQ1N2 -0.13333   0.26908 -0.496   0.622
RankedQ1N3 -0.09804   0.25597 -0.383   0.703
RankedQ1N4 -0.04762   0.28207 -0.169   0.867
RankedQ1N5 -0.13333   0.26908 -0.496   0.622

Residual standard error: 0.4088 on 50 degrees of freedom
Multiple R-squared: 0.05489,   Adjusted R-squared: -0.03962
F-statistic: 0.5808 on 5 and 50 DF, p-value: 0.7144

#While the p-value for freshmen at 0.227 is not statistically significant enough to prompt R to give me significance codes for the traditional 90%, 95%, 99%, and 99.9% confidence levels, it *would* be statistically significant at 75% confidence. Additionally, it sticks out because it's a lot lower than the p-values for all of the other academic rankings with p-values above 0.6, and the dataset itself has absolutely no first-year students that have *used* a research data service. With all of the above in consideration, I suspect that this may be a potential gap in research data services outreach. Unfortunately, we don't have a statistically significant F-statistic, and we also have an R-square value close to zero (and a negative adjusted R-square…)

> LogisticReg5 <- glm(formula= Q4N ~ RankedQ1N,
data=X2022FSULibSurveyDataFinalVersion, family ="binomial")
> summary(LogisticReg5)

Call:
glm(formula = Q4N ~ RankedQ1N, family = "binomial", data =
X2022FSULibSurveyDataFinalVersion)

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-0.90052	-0.73248	-0.66805	-0.00013	1.79412

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.6931	1.2247	-0.566	0.571
RankedQ1N1	-17.8729	2174.2132	-0.008	0.993
RankedQ1N2	-0.6931	1.4577	-0.475	0.634
RankedQ1N3	-0.4855	1.3516	-0.359	0.719
RankedQ1N4	-0.2231	1.4832	-0.150	0.880
RankedQ1N5	-0.6931	1.4577	-0.475	0.634

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.486 on 55 degrees of freedom
Residual deviance: 50.761 on 50 degrees of freedom
AIC: 62.761

Number of Fisher Scoring iterations: 17

#This is has no statistical significance with logistic regression. Ignore the previous hypothesis.

#Using years as a numerical variable in the same logistic model outlined above
> LinearReg4 <- lm(formula = Q4N ~ Q1N)
> summary(LinearReg4)

Call:
lm(formula = Q4N ~ Q1N)

Residuals:

	Min	1Q	Median	3Q	Max
	-0.2511	-0.2009	-0.1758	-0.1508	0.8743

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.12568	0.11759	1.069	0.290
Q1N	0.02508	0.03705	0.677	0.501

Residual standard error: 0.4029 on 54 degrees of freedom
Multiple R-squared: 0.008409,   Adjusted R-squared: -0.009954
F-statistic: 0.4579 on 1 and 54 DF, p-value: 0.5015

> LogisticReg6 <- glm(formula= Q4N ~ Q1N, data=X2022FSULibSurveyDataFinalVersion, family="binomial")
> summary(LogisticReg6)

Call:
glm(formula = Q4N ~ Q1N, family = "binomial", data = X2022FSULibSurveyDataFinalVersion)

Deviance Residuals:
        Min   1Q Median   3Q   Max
-0.7666 -0.6650 -0.6183 -0.5743 2.0109

Coefficients:
        Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.8798   0.7894 -2.381   0.0172 *
Q1N   0.1611 0.2358   0.683   0.4945

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

        Null deviance: 55.486 on 55 degrees of freedom
Residual deviance: 55.011 on 54 degrees of freedom
AIC: 59.011

Number of Fisher Scoring iterations: 4

The intercept is statistically significant at the 95% confidence level. It's mostly a question of *what* it is statistically significant with. It isn't significant with years in college.

#We can conclude that our strongest model for predicting whether someone will *use* a research data service is how someone responds to Question 7.2, as shown with the below code.

> LogisticReg7 <- glm(formula = Q4N ~ Q7.2N, data=X2022FSULibSurveyDataFinalVersion, family = "binomial")
> summary(LogisticReg7)

Call:
glm(formula = Q4N ~ Q7.2N, family = "binomial", data = X2022FSULibSurveyDataFinalVersion)

Deviance Residuals:
        Min   1Q Median   3Q   Max
-0.8879 -0.8879 -0.4779 -0.1076 2.1101

Coefficients:

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.650      3.694  -2.071   0.0383 *
Q7.2N         1.385      0.774   1.789   0.0736 .
---
```

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

```
    Null deviance: 49.127  on 47  degrees of freedom
Residual deviance: 43.712  on 46  degrees of freedom
  (8 observations deleted due to missingness)
AIC: 47.712
```

Number of Fisher Scoring iterations: 6

#With the above model, we can note that intercept is statistically significant at the 95% confidence level, and that how someone responds to Question 7.2 is statistically significant at the 90% confidence level.