#9 Logistic Regression: First-Year GPA

Troy Edwards

2023-03-26

a)

$$H_0 : \mu_0 = \mu_1$$

$$H_a : \mu_0 \neq \mu_1$$

$$\alpha = 0.05$$

Test: 2-sample difference in means t-test

Random: assume data was collected using simple random sample
Normal: $n_0 = 194 \geq 30$, $n_1 = 25 \ngeq 30$, CLT fails, proceeding with caution
Independent: reasonable to assume that $n \leq 0.1N$

```
t.test(GPA ~ FirstGen, data = gpa)
```

```
##
##  Welch Two Sample t-test
##
## data:  GPA by FirstGen
## t = 2.4474, df = 31.406, p-value = 0.02017
## alternative hypothesis: true difference in means between group 0 and group 1 is not e
## 95 percent confidence interval:
##  0.03820979 0.41912630
## sample estimates:
## mean in group 0 mean in group 1
##        3.122268        2.893600
```

$$t = \frac{(\bar{x}_0 - \bar{x}_1) - (\mu_0 - \mu_1)}{\sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}}}$$

$$t = \frac{(3.122268 - 2.8936) - 0}{\sqrt{\frac{0.4637245^2}{194} + \frac{0.4365001^2}{25}}}$$

$$t = 2.4474$$

$$p = P(t \geq 2.4474) = 0.02017 \text{ on } 31.406 \text{ df}$$

Since our p-value (0.02017) is less than α (0.05), we reject our null hypothesis that the true mean GPA for first-generation college students is equal to that of non-first-generation college students. Therefore, we have enough evidence to conclude that the true mean GPAs for these two groups are not equal.

b)

```r
regres01 <- lm(GPA ~ FirstGen, data = gpa)
summary(regres01)
```

```
##
## Call:
## lm(formula = GPA ~ FirstGen, data = gpa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.19227 -0.31293  0.04773  0.37773  1.02773
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.12227    0.03308  94.377   <2e-16 ***
## FirstGen1   -0.22867    0.09792  -2.335   0.0204 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4608 on 217 degrees of freedom
## Multiple R-squared:  0.02452,    Adjusted R-squared:  0.02002
## F-statistic: 5.454 on 1 and 217 DF,  p-value: 0.02044
```

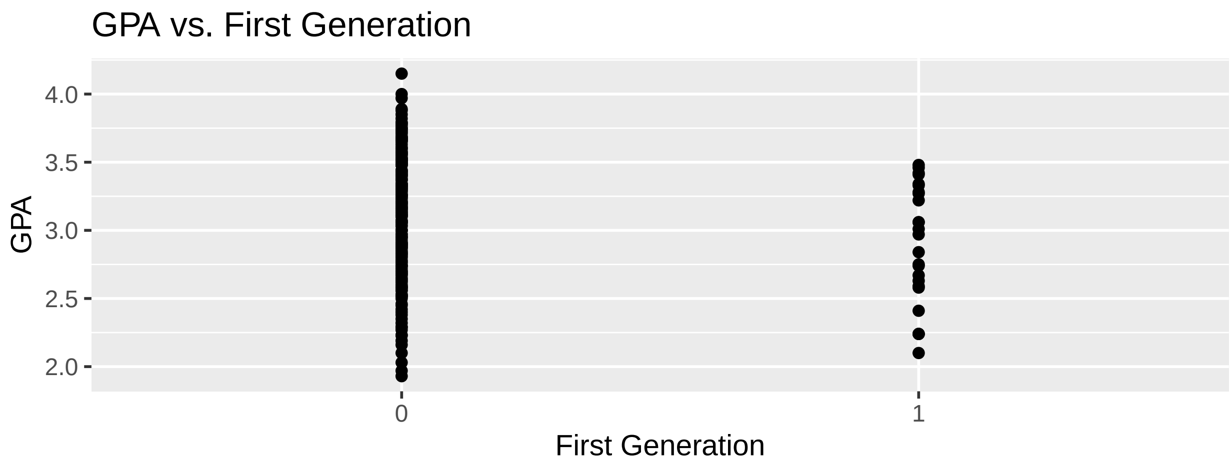
$$H_0 : \beta = 0$$

$$H_a : \beta \neq 0$$

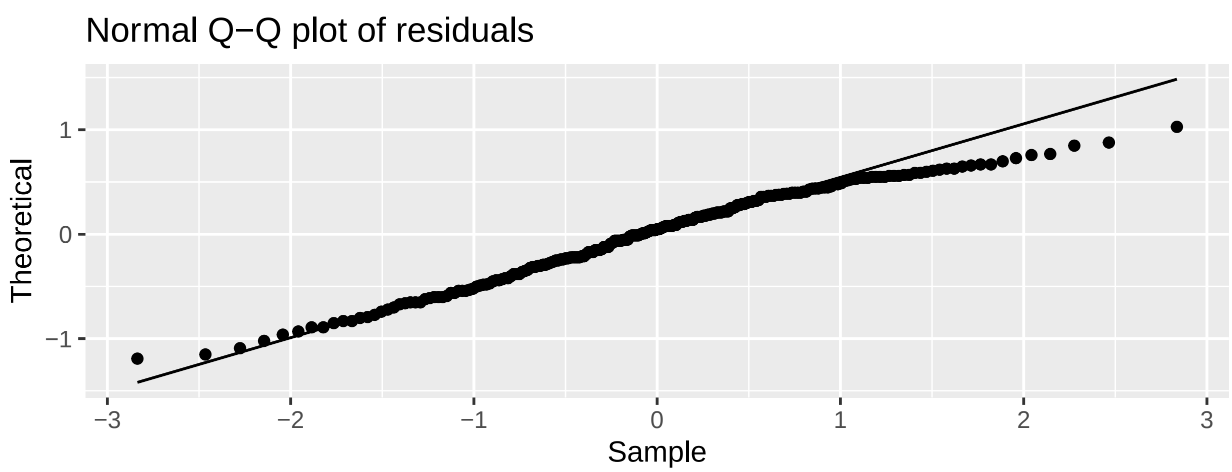
$$\alpha = 0.05$$

Test: Slope model utility t-test

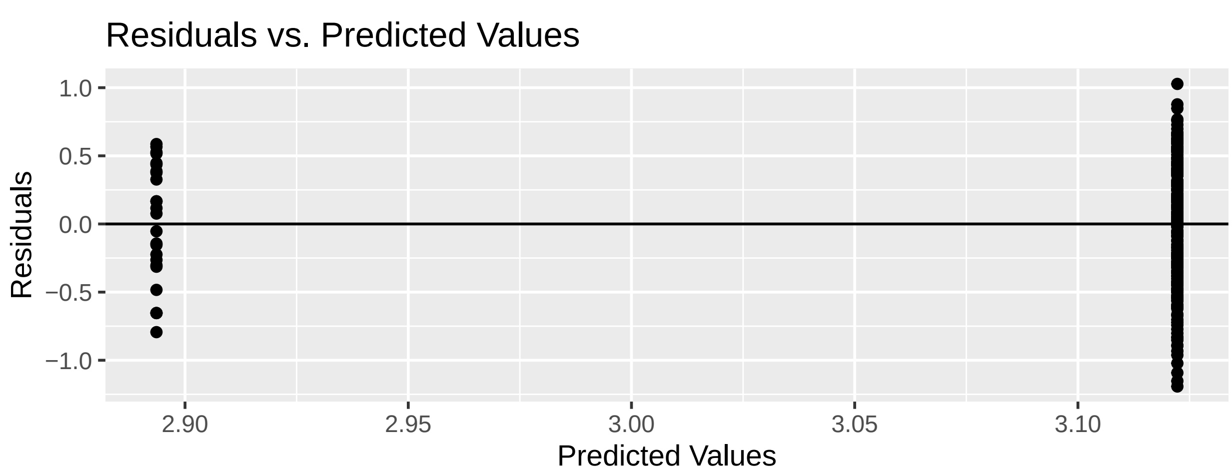
Linear: relationship between GPA and FirstGen does not appear linear, proceeding with caution



Normal: residuals appear normally distributed



Equal Variance: residuals appear randomly scattered about the x-axis



$$t = \frac{b-\beta}{S_b} = \frac{b-\beta}{\frac{\sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}}{\sqrt{\sum(x-\bar{x})^2}}}$$
$$t = \frac{-0.22867 - 0}{0.09792}$$
$$t = -2.335$$
$$p = P(|t| \geq |-2.335|) = 0.02044 \text{ on } 217 \text{ df}$$

Since our p-value (0.02044) is less than α (0.05), we reject our null hypothesis that the true slope equals 0. Therefore, we have enough evidence to support the alternative hypothesis that the true slope does not equal 0.

c)

```r
regres02 <- glm(FirstGen ~ GPA, data = gpa, family = binomial("logit"))
summary(regres02)
```

```
##
## Call:
## glm(formula = FirstGen ~ GPA, family = binomial("logit"), data = gpa)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8161  -0.5295  -0.4374  -0.3637   2.2863
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0751     1.3527   0.795    0.427
## GPA          -1.0381     0.4567  -2.273    0.023 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 155.54  on 218  degrees of freedom
## Residual deviance: 150.28  on 217  degrees of freedom
## AIC: 154.28
##
## Number of Fisher Scoring iterations: 5
```

Let $\alpha = 0.05$

Since our p-value (0.023) is less than α (0.05), we reject our null hypothesis that the true slope of the relationship between the logit transformation of the predicted probability of being a first-generation student and the GPA of the student is 0. Therefore, we have enough evidence to conclude that the true slope is not 0.

d)

Part 1

$$P(FirstGen|GPA = 4.0) = \frac{e^{1.0751-1.0381(4.0)}}{1 + e^{1.0751-1.0381(4.0)}} = 0.04405638$$

Part 2

$$ln\left(\frac{0.5}{1-0.5}\right) = 1.0751 - 1.0381(GPA)$$
$$1.0381(GPA) = 1.0751 - ln\left(\frac{0.5}{1-0.5}\right)$$
$$GPA = \frac{1.0751 - ln\left(\frac{0.5}{1-0.5}\right)}{1.0381}$$
$$GPA = \frac{1.0751 - 0}{1.0381}$$
$$GPA = 1.035642$$

Part 3

$$ln(2) = 1.0751 - 1.0381(GPA)$$
$$1.0381(GPA) = 1.0751 - ln(2)$$
$$GPA = \frac{1.0751 - ln(2)}{1.0381}$$
$$GPA = 0.3679345$$

e)

Model	$R^2(adj)$	S_e	F	p
$\widehat{GPA} = 1.17985 + 0.55501(HSGPA)$	0.196	0.4174	54.15	3.783×10^{-12}
$\widehat{GPA} = 1.160826 + 0.569504(HSGPA) - 0.284592(FirstGen) + 0.003677(HSGPA)(FirstGen)$	0.2235	0.4102	21.91	2.018×10^{-12}
$\widehat{GPA} = 2.0684133 + 0.0016986(SATV)$	0.08842	0.4444	22.15	4.499×10^{-6}
$\widehat{GPA} = 0.2459 + 0.7593(HSGPA) + 1.3010(White) - 0.2923(HSGPA)(White)$	0.2684	0.3981	27.65	2.647×10^{-15}
$\widehat{GPA} = 0.99579 + 0.54004(HSGPA) + 0.29842(White)$	0.2613	0.4001	39.56	2.303×10^{-15}

Note: I did test more models with more predictors but they were too big to put in the table.

The best model I found was:

$$\widehat{GPA} = 0.99579 + 0.54004(HSGPA) + 0.29842(White)$$

I chose this model because it has a good F-statistic and a good adjusted R^2 value.