# Math 150 - Methods in Biostatistics - Homework 5

Ethan Ashby

Due: Friday, March 5, 2021

**Assignment Summary (Goals)**

- fluent use of the multiple logistic model for prediction and for coefficient interpretation
- working with variables that interact and variables which are multicollinear
- practice using `ggplot()` so that visualizations can inform the larger analysis

Note that if you don't know the R code either check my notes or ask me!!! Happy to scaffold, debug, send resources, etc. Don't go down a rabbit hole trying to figure out an R function or syntax.

Also, note that you'll need to get the data from Sakai and use it for this analysis. Look back to your own HW1 file to see the line of code **you** used to import the `games1.csv` dataset. Ask me if it isn't obvious to you after you look at your own HW1. And just like in HW4, you'll need to deal with the missing variables coded as `"*"`.

**Q1. PodQ** Describe one thing you learned from someone in your pod this week (it could be: content, logistical help, background material, R information, etc.) 1-3 sentences.

Annika went on a hike to Thumb Butte last weekend. Annie had her half birthday, and Lian is in Tahoe making snowmen!

**Q2. Chp 7, E1 Bird Nest study** The file `Birdnest` contains data for 99 species of North American passerine birds. Passerine are "perching birds" and include many families of familiar small birds (e.g., sparrows and warblers), as well as some larger species like crows and ravens, but do not include hawks, owls, water fowl, wading birds, and woodpeckers. One hypothesis of interest was about the relationship of body size to type of nest. Body size was measured as average length of the species. Although nests come in a variety of types (see the `Nesttype` variable), in this data set next type was categorized into either closed or open. "Closed" refers to nests with only a small opening to the outside, such as the tree cavity nest of many woodpeckers or the pendant-style nest of an oriole. "Open" nests include the cup-shaped nest of the American robin. (Note: `Closed?` = 1 for closed nests; `Closed?` = 0 for open nests.)

```
birdnest <- read.delim("C7 Birdnest.csv", sep="\t",
                       na="*")
```

(a) Create a logistic regression model using bird length (`Length`) to estimate the probability that a bird species has a closed net type. Interpret the model in terms of the odds ratio.

```
glm(`Closed.` ~ as.numeric(Length), data=birdnest, family="binomial") %>% tidy()
```

```
## # A tibble: 2 x 5
##   term               estimate std.error statistic p.value
##   <chr>                   <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)            0.457     0.753      0.607   0.544
## 2 as.numeric(Length)  -0.0677    0.0425    -1.59    0.112
```

The odds ratio (of a closed nest) associated with a one unit increase in length is `exp(-0.0677)=0.935`.

(b) Use the Wald statistic to create a 95% confidence interval for the odds ratio. (Wald just means normal distribution, use Z. You can do it by hand using the standard output, or you can find the CI for the slope, using `tidy(conf.int = TRUE)`, and exponentiate. You could do it both ways and check to make sure you get the same answer!)

```
glm(`Closed.` ~ as.numeric(Length), data=birdnest, family="binomial") %>% tidy(conf.int=TRUE)
```

```
## # A tibble: 2 x 7
##   term               estimate std.error statistic p.value conf.low conf.high
##   <chr>                 <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept)          0.457     0.753      0.607   0.544   -0.921   2.04
## 2 as.numeric(Length)  -0.0677    0.0425    -1.59    0.112   -0.161   0.00673
```

The confidence interval for the log odds ratio is [-0.161, 0.00673], which implies the 95% confidence interval for the odds ratio is [exp(-0.161), exp(0.00673)]=[0.851, 1.007]!

(c) Test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ using both Wald's test and the likelihood ratio test. State your conclusions based on these tests.

```
glm(`Closed.` ~ as.numeric(Length), data=birdnest, family="binomial") %>% tidy()
```

```
## # A tibble: 2 x 5
##   term               estimate std.error statistic p.value
##   <chr>                 <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)          0.457     0.753      0.607   0.544
## 2 as.numeric(Length)  -0.0677    0.0425    -1.59    0.112
```

```
glm(`Closed.` ~ as.numeric(Length), data=birdnest, family="binomial") %>% glance()
```

```
## # A tibble: 1 x 8
##   null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
##           <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
## 1          120.      94  -58.4  121.  126.     117.          93    95
```

Wald test: The Wald test statistic is calculated as: Z=b_1-0/se(b_1)=(-0.0677-0)/0.0425=-1.59

Our p-value is the $P(|Z| \geq 1.59) = 0.112$ as computed using 'pnorm(1.59, lower.tail=FALSE)*2'. Thus, at the $\alpha = 0.05$ level, we fail to reject the null hypothesis: $H_0 : \beta_1 = 0$.

LRT: The likelihood ratio test statistic is calculated as $G = dev_{\text{null}} - dev_{\text{full}}$. We obtain these deviances from the `glance()` function called above. Thus, $G = 120 - 117 = 3$. Under the null model $G \sim \chi^2_1$ (b/c we are estimating only one additional parameter in our full model). Using 'pchisq(3, df=1, lower.tail=FALSE)*2', I obtained a p-value=0.167, meaning at the $\alpha = 0.05$ level, we fail to reject the null hypothesis: $H_0 : \beta_1 = 0$.

Both of these analyses suggest that the `length` variable does significantly affect closed nest staus.

For the LRT in R, the first thing you'll need to do is pipe the glm into `glance()`. That is: `glm(...) %>% glance()`. The "deviance" value with the MLEs is called `deviance`, the "deviance" value with the null value of $\beta_1 = 0$ is called `null.deviance`. You have to use R as a calculator to subtract the values. Then use `pchisq()` to find the p-value. See page 223 in your book.

skip (d), (e), (f)

**Q3. Chp 7, E9 (no (d)) Donner Party** In 1846, a group of 87 people (called the Donner Party) were heading west from Springfield, Illinois, for California.

The leaders attempted a new route through the Sierra Nevada and were stranded there throughout the winter. The harsh weather conditions and lack of food resulted in the death of many people within the group. Social scientists have used the data to study the theory that females are better able than men to survive harsh conditions.

(a) Create a logistic regression model using `Gender` and `Age` to estimate the probability of survival. Create a plot of the estimated probability of survival using `Age` as the explanatory variable and grouping the data by `Gender`. Use the plot and the model to interpret the coefficients in terms of the odds ratios.

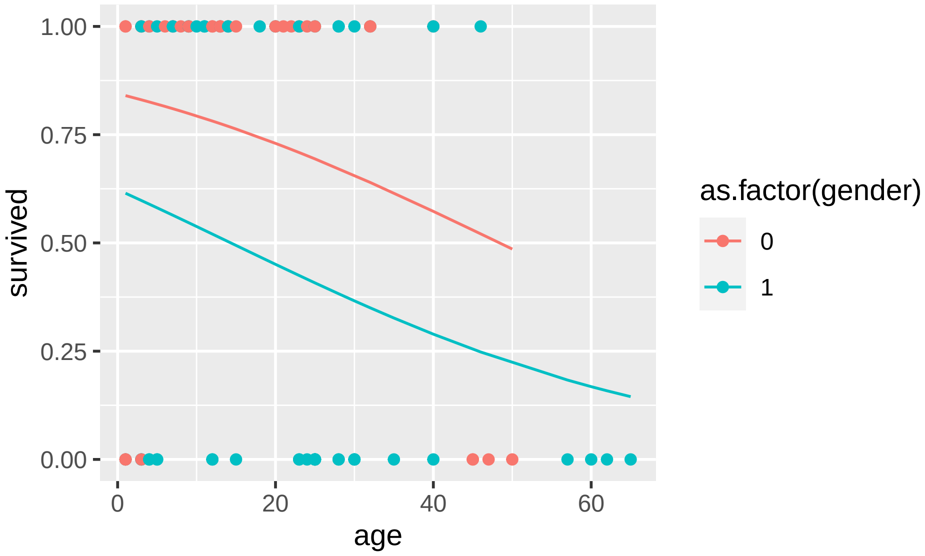
```
donner <- read.delim("C7 Donner.csv", sep="\t",
                     na="*")

names(donner) <- c("name", "gender", "age", "survived", "familysize", "X6", "X7",
                   "X8", "X9", "adultname", "adultgender", "adultage",
                   "adultsurvived", "adultfamilysize")
```

The code will look something like this. Fill in the blanks. And run the code one line at a time so that you know exactly what each line is doing. [Recall: what is the difference between the output of `tidy()` `glance()` and `augment()`? That is, all are data frames. And they give output that is of dimension 1, p (p is the number of variables), and n (n is the number of observations). Which is which?]

```
glm(survived ~ gender + age, data = donner, family="binomial") %>%
  augment()
```

```
## # A tibble: 86 x 10
##    .rownames survived gender   age .fitted .resid .std.resid   .hat .sigma
##    <chr>        <int>  <int> <int>   <dbl>  <dbl>      <dbl>  <dbl>  <dbl>
##  1 1                0      1    23 -0.303  -1.05      -1.06  0.0209   1.13
##  2 2                1      1    13  0.0470  1.16       1.17  0.0239   1.13
##  3 3                1      0     1  1.66    0.590      0.601 0.0350   1.13
##  4 4                1      1     4  0.362   1.03       1.05  0.0362   1.13
##  5 5                1      1    14  0.0119  1.17       1.19  0.0231   1.13
##  6 6                1      0    40  0.294   1.06       1.09  0.0629   1.13
##  7 7                1      1    40 -0.899   1.57       1.61  0.0384   1.12
##  8 8                1      1    11  0.117   1.13       1.14  0.0259   1.13
##  9 9                1      1     7  0.257   1.07       1.09  0.0312   1.13
## 10 10               1      1     9  0.187   1.10       1.12  0.0284   1.13
## # ... with 76 more rows, and 1 more variable: .cooksd <dbl>
```

```
glm(survived ~ gender + age, data = donner, family="binomial") %>%
  augment(type.predict = "response") %>%
  arrange(age) %>%
  ggplot() +
  geom_point(aes(x = age, y = survived, color=as.factor(gender))) +
  geom_line(aes(x = age, y = .fitted, group = gender, color = as.factor(gender)))
```

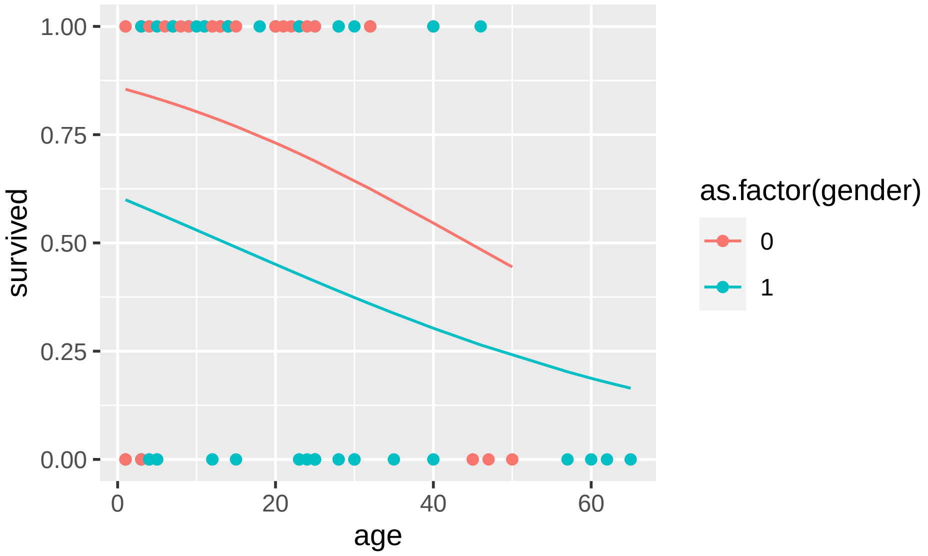


(b) Create and interpret a logistic regression model using `Gender`, `Age`, and `Gender*Age` to estimate the probability of survival. Create a plot of survival. Create a plot of the estimated probability of survival using `Age` as the explanatory variable and grouping the data by `Gender`. [Code from above almost identical.]

```r
glm(survived ~ gender + age + gender*age, data = donner, family="binomial") %>%
  augment()
```

```
## # A tibble: 86 x 10
##    .rownames survived gender   age  .fitted .resid .std.resid   .hat .sigma
##    <chr>        <int>  <int> <int>    <dbl>  <dbl>      <dbl>  <dbl>  <dbl>
##  1 1                0      1    23 -0.294   -1.06      -1.07  0.0209   1.13
##  2 2                1      1    13  0.0234   1.17       1.18  0.0257   1.13
##  3 3                1      0     1  1.77      0.560      0.576 0.0557   1.14
##  4 4                1      1     4  0.309     1.05       1.07  0.0455   1.13
##  5 5                1      1    14 -0.00827   1.18       1.20  0.0244   1.13
##  6 6                1      0    40  0.184     1.10       1.16  0.104    1.13
##  7 7                1      1    40 -0.833     1.55       1.59  0.0503   1.13
##  8 8                1      1    11  0.0869    1.14       1.16  0.0289   1.13
##  9 9                1      1     7  0.214     1.09       1.11  0.0374   1.13
## 10 10               1      1     9  0.150     1.11       1.13  0.0328   1.13
## # ... with 76 more rows, and 1 more variable: .cooksd <dbl>
```

```r
glm(survived ~ gender + age  + gender*age, data = donner, family="binomial") %>%
  augment(type.predict = "response") %>%
  arrange(age) %>%
  ggplot() +
  geom_point(aes(x = age, y = survived, color=as.factor(gender))) +
  geom_line(aes(x = age, y = .fitted, group = gender, color = as.factor(gender)))
```



(c) Explain any key differences between the plots created in parts (a) and (b). Discuss how adding the interaction term `Gender*Age` impacts the model.

The addition of the interacton term allows age to have different effects on survival for each gender. This manifested by a slightly steeper downward slope for the women's curve in the second plot, illustrating that when age and gender interact, the model predicts poorer survivorship for older women.

**Q4. Chp 7, E10 Variable Selection Techniques and Multicollinearity** Wolberg and Mangasarian developed a technique to accurate diagnose breast masses using only visual characteristics of the cells within the tumor (PNAS 1990). A sample is placed on a slide, and characteristics of the cellular nuclei within the tumor, such as size, shape, and texture are examined under a microscope to determine with the cancer cells are benign or malignant. Benign tumors are scar tissue or abnormal growths that do not spread and are typically harmless. Malignant (or invasive) cancer cells are cells that can travel, typically through the bloodstream or lymph nodes, and begin to replace normal cells in other parts of the body. If a tumor is malignant, it is essential to remove or destroy all cancerous cells in order to keep them from spreading. If a tumor is benign, surgery is not needed and the harmless tumor can remain.

(a) Crate a logistic regression model using `Radius`, `Concave`, and `Radius*Radius`, and `Radius*Concave` as explanatory variables to estimate the probability that a mass is malignant. Submit the logistic regression model and the likelihood ratio test results (testing whether any variables at all are significant), including the log-likelihood (or deviance) values and a statement of the null hypothesis. [Note that you need to create the `Radius*Radius` variable before running the `glm`.]

```
cancer <- read.delim("C7 Cancer2.csv", sep="\t",
                     na="*")
cancer <- cancer %>%
  mutate(Radius2 = Radius*Radius)

#log reg results
glm(`Malignant.` ~ Radius+Concave+Radius2+Radius*Concave, data=cancer, family="binomial") %>%
  tidy()
```

```
## # A tibble: 5 x 5
##   term           estimate std.error statistic p.value
##   <chr>             <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)      -9.36      7.78    -1.20    0.229
## 2 Radius            0.346     3.78     0.0915  0.927
## 3 Concave           6.54      3.03     2.16    0.0310
## 4 Radius2           0.351     0.465    0.757   0.449
## 5 Radius:Concave   -0.806     0.749   -1.08    0.282
```

```
#likelihood ratio test p-value
glm(`Malignant.` ~ Radius+Concave+Radius2+Radius*Concave, data=cancer, family="binomial") %>%
  glance() %>% .[,c(1,6)] %>% summarize(G=null.deviance-deviance) %>% .$G %>% pchisq(df=4, lower.tail=FA
```

```
## [1] 3.353182e-113
```

Note that the null deviance is 751 and the residual deviance is 222, leaving a difference in deviance of 529.13. The p-value for obtaining a difference in residuals greater than 529.13 with df=4 is $3.35 \cdot 10^{-113}$. So we can reject the null hypothesis that all the coefficients are 0… indeed, we conclude that at least one coefficient is nonzero and so at least one variable is significant.

(b) Even though in part (a) Wald's test shows the highest p-value for `Radius`, it is typically best to attempt to keep the simplest terms in the model. Generally, keeping simpler terms in the model makes the model easier to interpret. Thus, we suggest as a first attempt keeping `Radius` in the model and eliminating the variable with the next highest p-value. Create a logistic regression model using `Radius`, `Concave`, and `Radius*Concave` as explanatory variables to estimate the probability that a mass is malignant. Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values. Conduct the drop-in-deviance test (aka LRT) to determine if `Radius*Radius` should be included in the model, state the null hypothesis for your test. [Use `glm(...) %>% glance()` on models with and without the squared term.]

```
#log reg results
glm(`Malignant.` ~ Radius+Concave+Radius*Concave, data=cancer, family="binomial") %>%
  tidy()
```

```
## # A tibble: 4 x 5
##   term           estimate std.error statistic       p.value
##   <chr>             <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)     -15.0       2.47      -6.09 0.00000000114
## 2 Radius            3.19      0.605      5.27 0.000000136
## 3 Concave           6.46      3.05       2.12 0.0341
## 4 Radius:Concave   -0.778     0.747     -1.04 0.298
```

```
#resid dev
residual_dev=glm(`Malignant.` ~ Radius+Concave+Radius2+Radius*Concave, data=cancer, family="binomial")
  glance() %>% .$deviance

#null_dev
null_dev=glm(`Malignant.` ~ Radius+Concave+Radius*Concave, data=cancer, family="binomial") %>%
  glance() %>% .$deviance

pchisq(null_dev-residual_dev, df=1, lower.tail=FALSE)
```

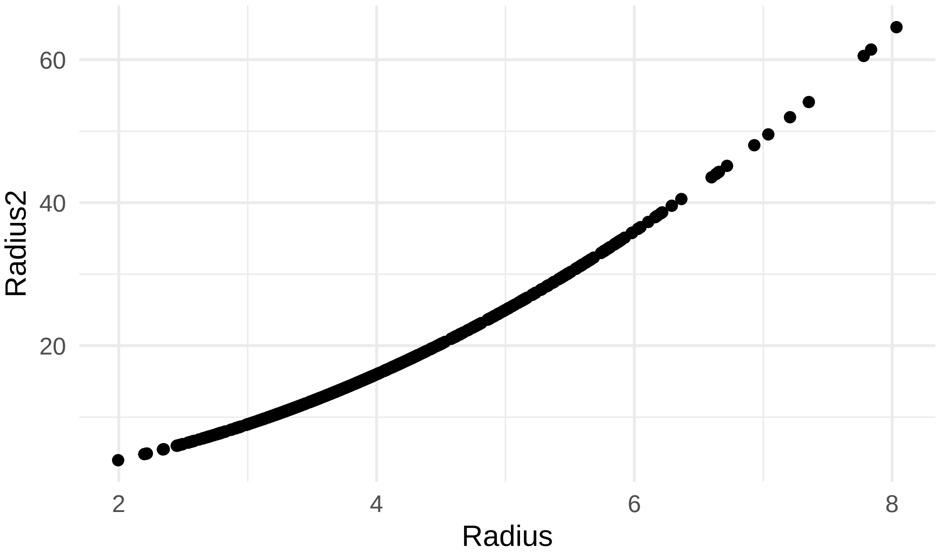
```
## [1] 0.4384992
```

The full model has a deviance of 222.32, while the null model (removed Radius2) has a null deviance of 222.91, producing a drop-in-deviance of drop in deviance of 0.6. With df=1, $\chi^2 = 0.6$ produces a p-value of 0.44. Thus, we fail to reject the null hypothesis that $b$ associated with the Radius2 variable is 0.

Removing Radius2 is a good move! Removing Radius2 cut the number of predictors by 25%, improved model interpretability, and barely effected the model deviance.

(c) Use a scatterplot to compare Radius to Radius*Radius and calculate the correlation between these two terms. Are the two variables highly correlated?

```r
ggplot(aes(x=Radius, y=Radius2), data=cancer)+geom_point()+theme_minimal()
```



Radius and Radiusˆ2 are highly correlated: the pearson correlation between these two variables is 0.99. It is important to note that, Radiusˆ2 is a deterministic function of Radius, however, the relationship is not linear.

(d) Chapter 3 discusses **multicollinearity** (highly correlated explanatory variables). Explain whether you believe Radius is important in the logistic regression model. Why is the p-value for Radius so large in part (a) but very small in part (b)?

Radius is doubtlessly important to the model. In the simpler model, Radius has a relatively large $b$ parameter and a small p-value, suggesting that it is important to include in the model. The p-value for Radius appears large in part (a) because Radius and Radius2 are highly correlated and are competing to provide the same information about the response. In turn, This depresses the $b$ value associated with Radius , rendering the variable nonsignificant. In part (b), when the multicollinearity is removed, Radius becomes highly significant because all the information about the response is allocated to one variable.

(e) Create a logistic regression model using Radius and Concave as explanatory variables to estimate the probability that a mass is malignant. Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values. Conduct the drop-in-deviance test to determine if Radius*Concave should be included in the model, state the null hypothesis for your test.

```
#full dev
residual_dev=glm(`Malignant.` ~ Radius+Concave+Radius*Concave, data=cancer, family="binomial") %>%
  glance() %>% .$deviance

#null dev
null_dev=glm(`Malignant.` ~ Radius+Concave, data=cancer, family="binomial") %>%
  glance() %>% .$deviance

#LRT
pchisq(null_dev-residual_dev, df=1, lower.tail=FALSE)
```

```
## [1] 0.2934193
```

```
#print model
glm(`Malignant.` ~ Radius+Concave, data=cancer, family="binomial") %>%
  tidy()
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   -13.1      1.49      -8.79 1.43e-18
## 2 Radius          2.72     0.366      7.42 1.19e-13
## 3 Concave         3.32     0.355      9.36 7.81e-21
```

The drop-in-deviance test performed above generates a $G = 1.10$. In other words, including the interaction term only marginally increased the drop-in-deviance statistic. With $G = 1.10$ and df=1, the p-value is 0.29; thus, we fail to reject the null hypothesis that the beta associated with the interaction term is 0.

I choose to omit `Radius*Concave` from the model, because it improves model interpretability and doesn't impact the drop in deviance very much.

(f) Create a logistic regression model using only `Concave` as an explanatory variable to estimate the probability that a mass is malignant. Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values. Conduct the drop-in-deviance to test to determine if `Radius` should be included in the model.

```
#full dev
residual_dev=glm(`Malignant.` ~ Radius+Concave, data=cancer, family="binomial") %>%
  glance() %>% .$deviance

#null dev
null_dev=glm(`Malignant.` ~ Concave, data=cancer, family="binomial") %>%
  glance() %>% .$deviance

#LRT
pchisq(null_dev-residual_dev, df=1, lower.tail=FALSE)
```

```
## [1] 2.142426e-23
```

```
#print model
glm(`Malignant.` ~ Concave, data=cancer, family="binomial") %>%
  tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    -2.85     0.236     -12.1 1.75e-33
## 2 Concave         4.71     0.307      15.3 4.39e-53
```

The drop-in-deviance test performed above generates a $G = 99.32$. In other words, leavig out Radius led to a substantial reduction in the drop-in-deviance statistic. With $G = 99.32$ and df=1, the p-value is $2.14 \cdot 10^{-23}$; thus with astronomically high confidence, we can reject the null hypothesis that the betas associated with `Radius` is 0.

I would recommend including `Radius` in the model. Leaving it out leads to a drastic drop in the drop-in-deviance statistic, illustrating that this variable is highly significant. Plus, it is an interpretable variable that doesn't overcomplicate the model.

(g) Submit a final model and provide a justification for choosing that model.

The final model that I would support would be `Malignant~Concave+Radius`. This model is lightweight, interpretable, and shed of all irrelevant and multicollinear variables. Both variables have extremely low p-values, illustrating that their effect on the response is significant!

```
praise()
```

```
## [1] "You are sublime!"
```