STA303 Summer 2018

Midterm

July 25th, 2018

First Name:_____

Last Name:_____

Student Number:_____

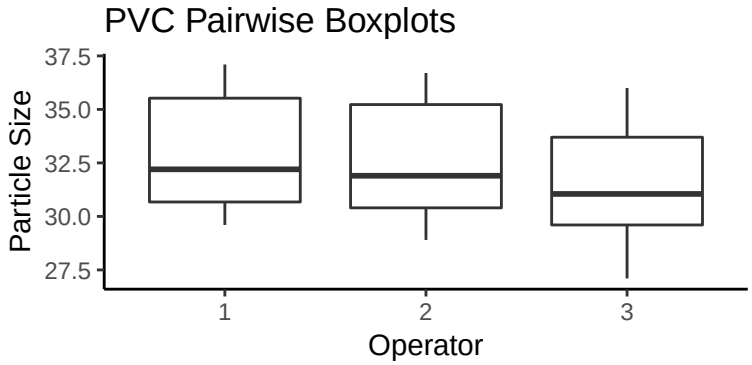
This exam booklet contains 15 pages, and 40 multiple choice questions. Write all answers on the attached scantron sheet, in pencil. Nothing written in the exam booklet will be marked. You must hand in your exam booklet. Aids permitted: non-programmable calculator.

1. Which of the following statements regarding Analysis of Variance (ANOVA) is correct?
   a. ANOVA analyzes differences in variances between several normally distributed populations
   b. ANOVA analyzes differences in means between several normally distributed populations
   c. ANOVA analyzes differences in variances between several arbitrarily distributed populations
   d. ANOVA analyzes differences in means between several arbitrarily distributed populations
2. Which of the following statements regarding Analysis of Variance (ANOVA) is correct?
   a. ANOVA is a linear model $y = X\beta + \epsilon$, where the columns of X can be anything
   b. ANOVA is a non-parametric regression model $y = X\beta + \epsilon$, where the columns of X can be anything
   c. ANOVA is a linear model $y = X\beta + \epsilon$, where the columns of X are all continuous
   d. ANOVA is a linear model $y = X\beta + \epsilon$, where the columns of X are all indicator variables
3. Which of the following statements regarding Analysis of Variance (ANOVA) is correct?
   a. ANOVA assumes that the population variances of all the groups are equal
   b. ANOVA assumes that the population means of all the groups are equal
   c. ANOVA assumes that the sample variances of all the groups are equal
   d. ANOVA assumes that the sample means of all the groups are equal
4. Which of the following is NOT a model assumption when performing an ANOVA, assuming all the groups have an equal number of observations?
   a. The population variances of all the groups are equal
   b. The observations are mutually independent within and across groups
   c. The observations are normally distributed
   d. The total sum of squares equals the sum of the within-group sum of squares (error, SSE) and the between-group sum of squares (model SSM), $SST = SSM + SSE$
5. Recall the cell means coding method of writing a linear model with one discrete predictor variable: $y_{ij} = \alpha_i + \epsilon_{ij}$, $i = 1 \dots m, j = 1 \dots n$. Writing this as a linear model in matrix form $\mathbf{y} = X\beta + \boldsymbol{\epsilon}$ with $\beta = (\alpha_1, ...\alpha_K)$, what is the cross-product matrix $X^T X$? I_m is the m-dimensional identity matrix.
   a. $X^T X = nI_m$
   b. $X^T X = (1/n)I_m$
   c. $X^T X = I_m$
   d. It is not possible to say
6. Recall effects coding: $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, with $\alpha_1 = 0$. The two codings give the same estimated means; they differ in the manner in which their regression coefficients are interpreted in the context of the problem. Which coding would be preferable if we are designing a study to compare a new treatment to a control group, with the aim of making a statement about the difference in means between the treatment and control group?
   a. Effects coding
   b. Cell means coding
   c. Not possible to say

7. Which coding is used by the linear model underlying the standard ANOVA table?
    a. Effects coding
    b. Cell means coding
    c. Not possible to say

Recall the **pvc** data discussed in lecture: Data from an experiment to study factors affecting the production of the plastic PVC, 3 operators used 8 different devices called resin railcars to produce PVC, that hard rubber-like stuff that sewer pipes are made of (look under your kitchen sink when you get home). For each of the 24 combinations, two samples were produced. Here is a glimpse of the data:

```
## Observations: 48
## Variables: 3
## $ psize    <dbl> 36.2, 36.3, 35.3, 35.0, 30.8, 30.6, 29.8, 29.6, 32.0,...
## $ operator <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2,...
## $ resin    <fct> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 1, 1,...
```



8. For the **pvc** data above, which of the following **ggplot** commands generates the pairwise boxplots shown in the figure entitled "PVC Pairwise Boxplots"? (note: code shown does not include the code that sets the title, axis labels, or theme of the plot).
    a. `pvc %>% ggplot(aes(x = operator,y = psize)) + geom_boxplot()`
    b. `pvc %>% ggplot(aes(x = resin,y = psize)) + geom_boxplot()`
    c. `pvc %>% ggplot(aes(y = psize)) + geom_boxplot(x = operator)`
    d. `pvc %>% ggplot(aes(y = psize,group = operator)) + geom_boxplot()`

9. Consider the `pvc` data and the "PVC Pairwise Boxplots" plot. We wish to run an ANOVA to assess differences in particle size for levels of operator. Does the plot contain enough information to assess the assumption of equality of variances across groups, and why?
    a. Yes- the only way to assess this assumption is by looking at this type of plot
    b. No- it is impossible to assess this assumption without the aid of a statistical test
    c. Yes- it is reasonable to assess this assumption by looking at this type of plot
    d. No- this type of plot does not contain information about the variances of the groups
10. Consider the `pvc` data and the "PVC Pairwise Boxplots" plot. We wish to run an ANOVA to assess differences in particle size for levels of operator. Based on this plot, are you comfortable with the assumption of normality of the data, and why?
    a. Yes- we can tell that the distribution of each group is symmetric, and therefore normal, by looking at this plot
    b. Yes- the groups all look similar, which indicates normality
    c. No- the range of the y-axis is not -2 to 2, so the data is not normally distributed
    d. No- while we can sort of tell the distribution of the data looks normal based on this plot, there exist better diagnostic methods for assessing normality that are readily available, so we should use those.

We fit an ANOVA to these data, obtaining the below ANOVA table:

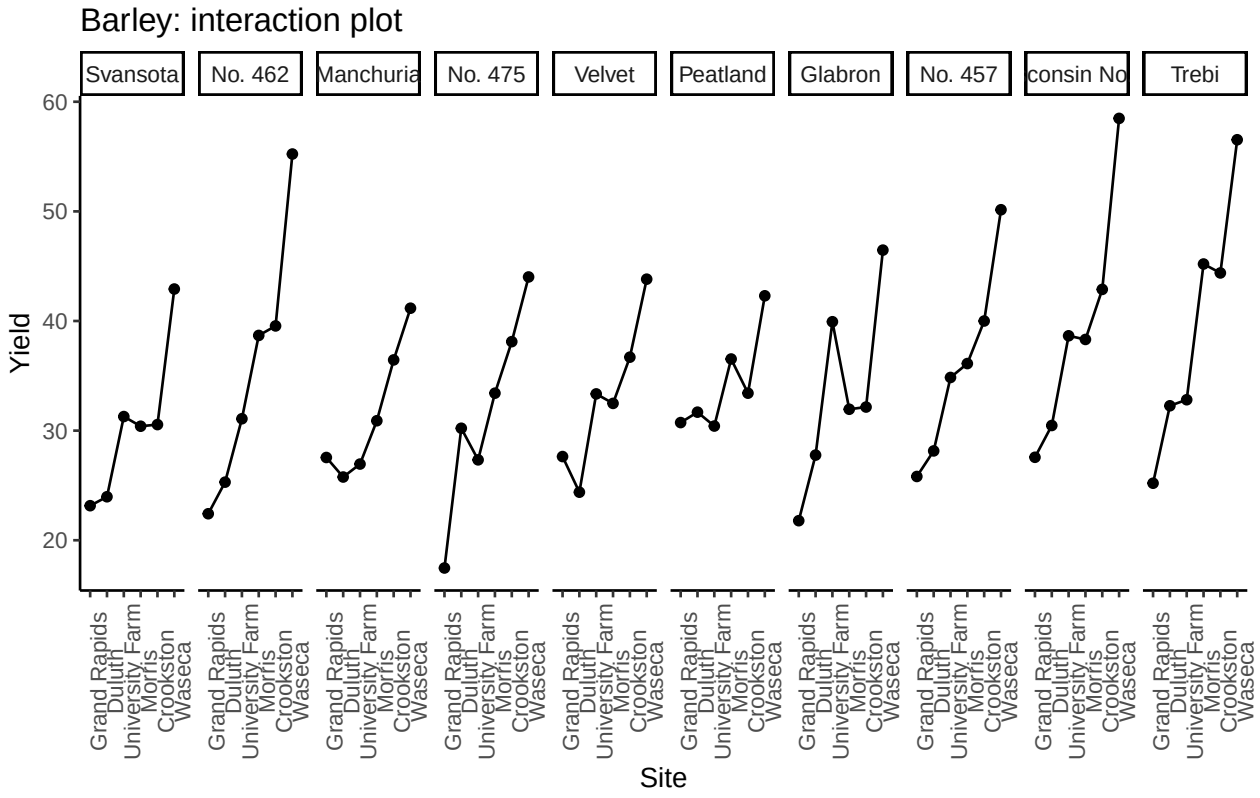
	Df	Sum of Squares	Mean Square	F	p
operator	(1)	20.7	(2)	(3)	.258
Residuals	(4)	333.8	(5)		

11. What is the total sum of squares, $SST = \sum_{i=1}^{3} \sum_{j=1}^{16} (y_{ij} - \bar{y}_{..})^2$
    a. 354.5
    b. 333.8
    c. 20.7
    d. 313.1
12. What value should go in the cell labelled by (1)?
    a. 1
    b. 2
    c. 45
    d. 48
13. What value should go in the cell labelled by (2)?
    a. 10.35
    b. 20.7
    c. 0.062
    d. 0.058

14. What value should go in the cell labelled by (3)?
    a. 1.4
    b. -1.4
    c. 16.13
    d. 0.742
15. What value should go in the cell labelled by (4)?
    a. 1
    b. 2
    c. 45
    d. 48
16. What value should go in the cell labelled by (5)?
    a. 7.418
    b. 6.954
    c. 16.13
    d. 47.00
17. Denote the population group means μ_1, μ_2, μ_3 and the population group standard deviations $\sigma_1, \sigma_2, \sigma_3$. What null hypothesis is the F statistic testing?
    a. $H_0 : \mu_1 = \mu_2 = \mu_3$
    b. $H_0 : \mu_1 = \mu_2 = \mu_3 = 0$
    c. $H_0 : \sigma_1 = \sigma_2 = \sigma_3$
    d. $H_0 : \mu_1 = \mu_2 = \mu_3$ and $\sigma_1 = \sigma_2 = \sigma_3$
18. Using the usual 0.05 significance level, what do you conclude?
    a. These data provide evidence to suggest that the population means of the groups are equal.
    b. These data provide evidence to suggest that the population means of the groups are not equal.
    c. These data fail to provide evidence to suggest that the population means of the groups are equal.
    d. These data fail to provide evidence to suggest that the population means of the groups are not equal.

Consider the following experiment: yield of barley per acre was measured for 10 varieties at 6 sites. We wish to assess whether the yield differs across sites, whether the yield differs across varieties, and whether the relationship between mean yield and site is the same for each variety.

```
## Observations: 120
## Variables: 4
## $ yield   <dbl> 27.00000, 48.86667, 27.43334, 39.93333, 32.96667, 28.9...
## $ variety <fct> Manchuria, Manchuria, Manchuria, Manchuria, Manchuria,...
## $ year    <fct> 1931, 1931, 1931, 1931, 1931, 1931, 1931, 1931, 1931, ...
## $ site    <fct> University Farm, Waseca, Morris, Crookston, Grand Rapi...
```



19. Consider the plot entitled "Barley: interaction plot". Which of the following aspects of these data cannot be reasonably assessed by this plot?
    a. Normality of the data
    b. The marginal mean of **`yield`** across values of **`site`**
    c. The marginal mean of **`yield`** across values of **`variety`**
    d. How the relationship between mean of **`yield`** and **`site`** changes across levels of **`variety`**
20. Which is a reasonable claim to make about the mean of **`yield`** for **`Grand Rapids`**?
    a. The mean **`yield`** for **`Grand Rapids`** is lower than all the other sites, for every variety of barley
    b. The mean **`yield`** for **`Grand Rapids`** is lower than all the other sites, averaged across varieties of barley
    c. The mean **`yield`** for **`Grand Rapids`** is higher than all the other sites, for every variety of barley
    d. The mean **`yield`** for **`Grand Rapids`** is higher than all the other sites, averaged across varieties of barley

21. Which of the following four statements is the most reasonable claim to make about the mean **yield** across sites, based on this plot?
    a. The mean **yield** appears to differ across sites; a formal significance test of the existence of the interaction of **yield** and **site** would be an appropriate procedure to help decide whether this difference is due to random chance
    b. The relationship between the mean of **yield** and **site** appears to be different for different varieties of barley; a formal significance test of the existence of the interaction between **variety** and **site** would be an appropriate procedure to help decide whether this difference is due to random chance
    c. The mean **yield** does not appear to differ across sites; a formal significance test is not necessary or desirable in this context
    d. This plot does not give information about the manner in which mean **yield** differs across sites

Consider the below linear model. For what follows, you can assume any necessary model assumptions are satisfied.

```
##
## Call:
## lm(formula = yield ~ site + variety, data = barley)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7439  -5.3051  -0.0758   5.1164  16.7011
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                20.889      2.445   8.542 1.12e-13 ***
## siteDuluth                  3.065      2.187   1.401 0.164082
## siteUniversity Farm         7.735      2.187   3.536 0.000605 ***
## siteMorris                 10.468      2.187   4.786 5.59e-06 ***
## siteCrookston              12.488      2.187   5.709 1.06e-07 ***
## siteWaseca                 23.177      2.187  10.596  < 2e-16 ***
## varietyNo. 462              5.000      2.824   1.771 0.079518 .
## varietyManchuria            1.086      2.824   0.385 0.701291
## varietyNo. 475              1.383      2.824   0.490 0.625238
## varietyVelvet               2.683      2.824   0.950 0.344164
## varietyPeatland             3.803      2.824   1.347 0.180980
## varietyGlabron              2.964      2.824   1.050 0.296307
## varietyNo. 457              5.469      2.824   1.937 0.055443 .
## varietyWisconsin No. 38     9.017      2.824   3.193 0.001858 **
## varietyTrebi                9.022      2.824   3.195 0.001846 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.917 on 105 degrees of freedom
## Multiple R-squared:  0.6048, Adjusted R-squared:  0.5521
## F-statistic: 11.48 on 14 and 105 DF,  p-value: 1.542e-15
```

22. What null hypothesis is the F statistic testing?
    a. The population mean **`yield`** is the same across sites
    b. The population mean **`yield`** is the same across varieties
    c. The population mean **`yield`** is the same across sites and varieties
    d. There is no significant interaction between **`site`** and **`variety`**
23. Using the usual arbitrary 0.05 significance level, do these data support the hypothesis that the population mean **`yield`** differs across sites?
    a. Yes, some of the **`site`** p-values are less than the significance level
    b. No, not all of the **`site`** p-values are less than the significance level
    c. Yes, the intercept p-value is less than the significance level
    d. The p-values presented in this output do not provide sufficient information to answer this question

We think there might be an interaction between **`site`** and **`variety`**. We want to fit another model incorporating this interaction term, and see whether the improvement in fit to the data is enough to justify the increase in model complexity.

24. What **`lm`** call will generate the desired model?
    a. **`lm(yield ~ site + variety, data = barley)`**
    b. **`lm(yield ~ site * variety, data = barley)`**
    c. **`lm(yield ~ site : variety, data = barley)`**
    d. **`lm(yield ~ site ! variety, data = barley)`**

25. What would be the most reasonable inferential procedure for assessing whether a statistically significant interaction exists between these variables?
    a. Just look at the interaction plot; clearly there is none (this is why we make these plots in the first place). We don't need to do a significance test with an arbitrary level every time we want to make a conclusion.
    b. Fit the model with only main effects, and the model with both main effects and the interaction to the data, and look at which one has a higher adjusted R^2. Adjusted R^2 is important to use over regular R^2, because adding terms to a model will always increase R^2
    c. Fit the model with only the main-effects, then fit the model with only the interaction, and compare their fit to the data. The model with main effects and interaction is saturated and will fit the data perfectly, so should not be considered.
    d. Fit the model with only main effects, and the model with both main effects and the interaction to the data, and compare them using an F test.

Recall the discussion from lecture of the Wisconsin Breast Cancer data: 681 cases of potentially cancerous tumours. The response variable is whether the tumor is benign, with $y = 0$ corresponding to malignant and $y = 1$ corresponding to benign. We fit a binary regression model with two predictors: **`Adhes`** (marginal adhesion) and **`BNucl`** (bare nuclei) as follows:

```
##
## Call:
## glm(formula = Class ~ Adhes + BNucl, family = binomial, data = wbca)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6635  -0.0365   0.2418   0.2418   3.4547
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.95908    0.35760  13.868  < 2e-16 ***
## Adhes       -0.69749    0.09616  -7.253 4.07e-13 ***
## BNucl       -0.74367    0.07760  -9.584  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 881.39  on 680  degrees of freedom
## Residual deviance: 243.65  on 678  degrees of freedom
## AIC: 249.65
```

```
##
## Number of Fisher Scoring iterations: 7
```

26. For a binary response, the binomial residual deviance reduces to $D = -2\sum_{i=1}^{n}(\hat{p}_i \log \hat{p}_i + (1 - \hat{p}_i)\log(1 - \hat{p}_i))$. Would you recommend comparing binary regression models using residual deviance, and why or why not?
    a. Yes, binary regression is just a special case of binomial regression, and residual deviance is an appropriate statistic to consider when comparing the fits of binomial models
    b. Yes, when the data is binary the approximate distribution of the residual deviance is χ^2_1, which has a mean of 1, and hence it will reject the null hypothesis of equal fit more often (higher power)
    c. No, when the data is binary the approximate distribution of the residual deviance is χ^2_1, which has a mean of 1, and hence it will reject the null hypothesis of equal fit less often (lower power)
    d. No, the point of looking at residual deviance is to compare the model fitted values to the observed responses, but the above expression depends only on the model fitted values
27. Recall the form of the binary regression model: $g(p_i) = \eta_i = \mathbf{x}_i^T\beta$, with $g(p) = \log\left(\frac{p}{1-p}\right)$ According to the above model, what is the value of the linear predictor η_i for an observation with **`Adhes`** $= 1$ and **`BNucl`** $= 1$?
    a. 3.51
    b. -3.67
    c. 4.959
    d. -1.44
28. What is the predicted probability of the tumor being malignant for an observation with a linear predictor $\hat{\eta} = 1$?
    a. 0.731
    b. 0.269
    c. 0.5
    d. Not enough information; we also need to know $\mathbf{x}$ and $\hat{\beta}$
29. Suppose observation y_1 has **`BNucl`** $= 1$, and observation y_2 has **`BNucl`** $= 2$, and that they both have the same value of **`Adhes`**. What is the ratio of the odds of y_1 having a benign tumor to the odds of y_2 having a benign tumour?
    a. 0.744
    b. 0.475
    c. 2.103
    d. Impossible to say without knowing their common value of **`Adhes`**.

We think there might be an interaction between these two variables, so we fit a model:

```
##
## Call:
## glm(formula = Class ~ Adhes * BNucl, family = binomial, data = wbca)
```

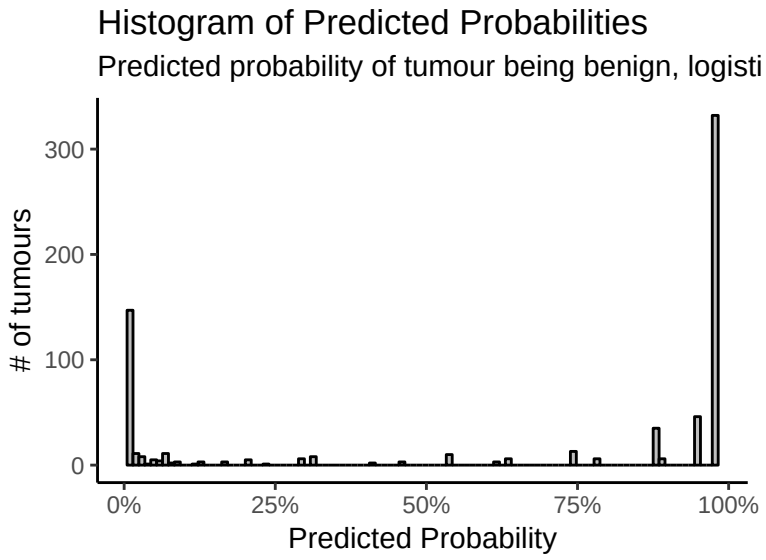
```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7837  -0.1525   0.2049   0.2049   2.9977
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.80758    0.48774  11.907  < 2e-16 ***
## Adhes       -1.02800    0.15401  -6.675 2.48e-11 ***
## BNucl       -1.02960    0.12277  -8.387  < 2e-16 ***
## Adhes:BNucl  0.10345    0.02631   3.933 8.40e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 881.39  on 680  degrees of freedom
## Residual deviance: 233.08  on 677  degrees of freedom
## AIC: 241.08
##
## Number of Fisher Scoring iterations: 7
```

30. Under this model, what is the effect of a unit increase in **Adhes**, holding **BNucl** constant?
    a. Decrease the log-odds of the tumor being benign by -1.028
    b. Change the log-odds of the tumor being benign by $-1.028 + 0.103\times$ **BNucl** (we can't say if it will be an increase or decrease without knowing the value of **BNucl**)
    c. Decrease the probability of the tumor being benign by a factor of -1.028
    d. Multiply the probability of the tumor being benign by $-1.028 + 0.103\times$ **BNucl** (we can't say if it will be an increase or decrease without knowing the value of **BNucl**)
31. Both main effects are negative, but their interaction is positive. What of the following explanations for this sounds the most reasonable?
    a. The relationship between **Adhes** and **BNucl** is increasing: larger **Adhes** implies larger **BNucl**
    b. The model coefficients are unstable (have high variance)
    c. The intercept is huge, so the difference in sign between the interaction term and the main effects won't affect the model predictions much, and hence their estimates could be anything
    d. The relationship between **Adhes** and the log-odds of the tumour being benign is different for different values of **BNucl**.

32. How would you interpret the relationship between the predictor variables and the log-odds of a tumour being benign?
    a. A larger `BNucl` decreases the amount by which increasing `Adhes` decreases the log-odds of the tumor being benign
    b. A larger `Adhes` decreases the amount by which increasing `BNucl` decreases the log-odds of the tumor being benign
    c. Both A and B are correct
    d. Neither A nor B are correct

We wish to assess the ability of our model to classify tumours as benign or malignant- the predictive accuracy of our model. To operationalize our model for this purpose, we need to pick a cutoff $0 < h < 1$ such that we classify a tumour as benign if $\hat{p} > h$ and malignant otherwise, where $\hat{p}$ is the prediction from our model. How to choose this cutoff? We make a histogram of the predicted probabilities for each individual in the dataset, entitled "Histogram of Predicted Probabilities".

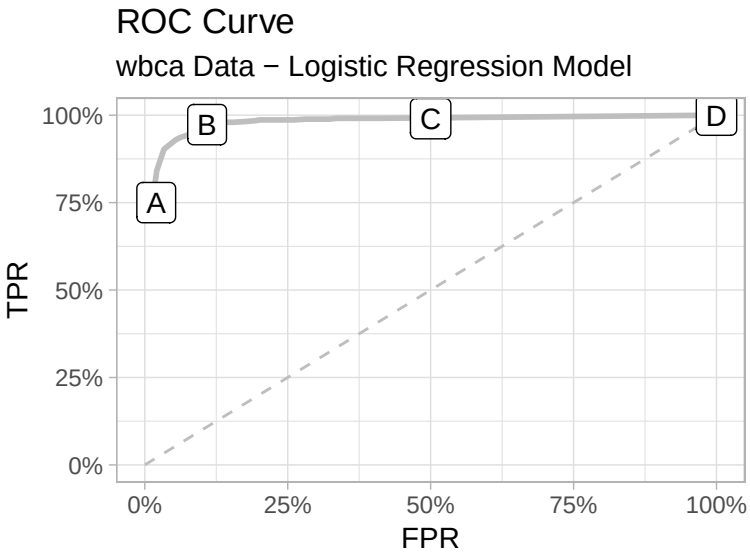
33. How would you use this chart to pick a cutoff? Select the most reasonable response.
    a. Choose $h = 0.5$ because this bisects the range of possible predicted probabilities.
    b. Choose $h = 0.9$ because it appears that most of the predicted probabilities are close to 1
    c. We cannot tell from this plot alone; we need more information about the cost associated with each type of error- e.g. is it worse to predict a benign tumour as malignant, or a malignant tumour as benign?
    d. We cannot tell from this plot alone; we also need to look at a plot of the residuals vs fitted values



Suppose we choose a cutoff of $h = 0.5$, leading to the following table of predicted vs actual (predicted is columns, actual is rows, i.e. there are 26 observations with actual = 0 and predicted = 1):

```
##
##        0   1
##   0 212  26
##   1  12 431
```

34. What is the false positive rate for this cutoff?
    a. 10.9%
    b. 12.3%
    c. 5.4%
    d. Cannot tell from this table alone
35. What is the true positive rate for this cutoff?
    a. 94.3%
    b. 97.3%
    c. 63.3%
    d. Cannot tell from this table alone

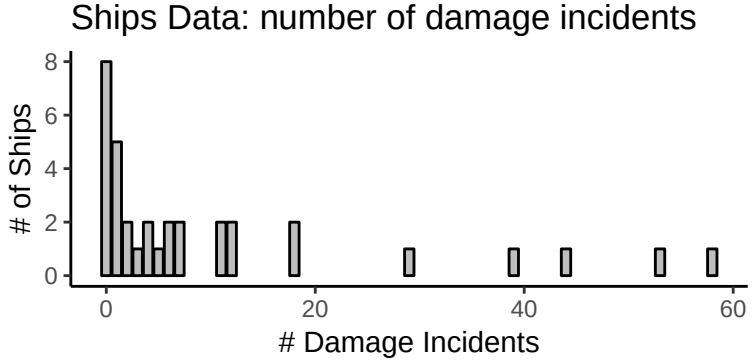


36. Consider the ROC curve shown in the plot "WBCA Model - ROC Curve". There are four letters on the curve. Which one corresponds to the above contingency table?
    a. A
    b. B
    c. C
    d. D

The **ships** data gives the number of damage **incidents** by number of months of **service**, **year** of construction, **period** of operation, and **type** of ship. We wish to build a model of **incidents**.

```
## Observations: 34
```

```
## Variables: 5
## $ type      <fct> A, A, A, A, A, A, A, B, B, B, B, B, B, B, C, C, C, C...
## $ year      <int> 60, 60, 65, 65, 70, 70, 75, 60, 60, 65, 65, 70, 70, ...
## $ period    <int> 60, 75, 60, 75, 60, 75, 75, 60, 75, 60, 75, 60, 75, ...
## $ service   <int> 127, 63, 1095, 1095, 1512, 3353, 2244, 44882, 17176,...
## $ incidents <int> 0, 0, 3, 4, 6, 18, 11, 39, 29, 58, 53, 12, 44, 18, 1...
```



37. Which of the following is assumed about the mean and variance of the response Y when fitting a Poisson regression?
    a. $Var(Y) = \sigma^2$, a constant
    b. $Var(Y) > E(Y)$
    c. $Var(Y) = E(Y)$
    d. There is no assumption; we know the mean-variance relationship because we know the data is Poisson distributed
38. How do you check this assumption?
    a. Plot the residuals vs the fitted values
    b. Plot the squared residuals vs the fitted values
    c. Plot the standardized residuals vs the fitted values
    d. Use a normal QQ-plot

39. I fit a model `incidents ~ service + year + period`. Then I fit another model: `incidents ~ log(service) + year + period`. Why do you think I tried a log transformation of `service`?
    a. `service` is a count, so a log transformation is appropriate
    b. `service` represents the aggregate lifetime in months of the ship. We would think that damage `incident`s would increase approximately linearly with `service`. Since the poisson regression model models the log of the mean number of incidents as a linear function of the predictors, including `log(service)` models this perceived linear relationship.
    c. `service` represents the aggregate lifetime in months of the ship. We would think that damage `incident`s would increase approximately multiplicatively with `service`. Since the poisson regression model models the mean number of incidents as a linear function of the predictors, including `log(service)` models this perceived multiplicative relationship.
    d. `service` is strictly positive, which violates the normality assumption of the model
40. Suppose we think there might be overdispersion in these data. We estimate the dispersion parameter and refit the model. What will change?
    a. Regression coefficients remain the same; their standard errors will change
    b. Regression coefficients remain the same; their standard errors remain the same
    c. Regression coefficients will change; their standard errors will change
    d. Regression coefficients will change; their standard errors remain the same