Homework 5

Mary Silva, Wyara Silva, Peter Trubey

November 21, 2017

Question 1

```
data("tb.dilute")
tb.dilute$orddose = tb.dilute$logdose %>%
  as.character %>%
  as.numeric %>%
  factor(levels = c(-0.5, 0, 0.5), labels = c('low','med','high'), ordered = TRUE)
q1_model = lm(reaction ~ animal + orddose, data = tb.dilute)
summary(q1_model)
```

```
##
## Call:
## lm(formula = reaction ~ animal + orddose, data = tb.dilute)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4167 -0.3750  0.1667  0.2917  1.7917
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.0000     0.5329  20.640 1.58e-09 ***
## animal2       1.0000     0.7537   1.327 0.214075
## animal3       2.1667     0.7537   2.875 0.016534 *
## animal4       2.1667     0.7537   2.875 0.016534 *
## animal5       3.5833     0.7537   4.754 0.000775 ***
## animal6       4.0833     0.7537   5.418 0.000294 ***
## orddose.L     3.3882     0.3768   8.991 4.18e-06 ***
## orddose.Q    -0.7655     0.3768  -2.031 0.069661 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9231 on 10 degrees of freedom
## Multiple R-squared:  0.9266, Adjusted R-squared:  0.8753
## F-statistic: 18.04 on 7 and 10 DF,  p-value: 6.395e-05
```

This model performs a two-way ANOVA on the tb.dilute data. In the model, we look at the effect of the dosage, as well as the individual animal effect on reaction. Dosage amount has gone through a log-transformation and then been coded as a factor. The model is defined as:

$$y_{i,j} = \mu + \alpha_i + \beta_j + \varepsilon_{i,j} \qquad \varepsilon_{i,j} \sim N(0, \sigma^2)$$

In order to make the model-matrix full rank, R has dropped the first columns from each of the factor variables. Thus, μ now represents the mean value for animal 1, and log dose (+) 0.5.

The R^2 for this model is given as 0.9266, which means 92.66% of the variation in the response variable *reaction* is explained by the predictors.

H_0 : $\alpha_i, \beta_j = 0 \ \forall i, j$ - There is no dosage or animal effect.

H_1 : At least one of the dosage or animal effects is significant.

The overall F-test for the model, which tests whether any of the individual coefficients are significant, is given as 18.04 with 7 and 10 degrees of freedom. This corresponds to an extremely low p-value compared to our standard alpha = 0.05 level of significance, causing us to reject the null hypothesis.

We should check the normality of the residuals as well.

```
shapiro.test(q1_model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  q1_model$residuals
## W = 0.94842, p-value = 0.4008
```

The Shapiro-Wilks test for normality of errors fails to reject the null hypothesis, that the errors are normally distributed around 0. Given this result, we can accept the previous test as valid and sufficient for this analysis.

Question 2

```
data("vitcap2")
vitcap2$group = factor(vitcap2$group)

q2_model = lm(vital.capacity ~ group + age, data = vitcap2)

print(summary(q2_model))
```

```
##
## Call:
## lm(formula = vital.capacity ~ group + age, data = vitcap2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38054 -0.38050  0.01321  0.37909  1.37047
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.927982   0.360863  16.427  < 2e-16 ***
## group2       0.046737   0.224536   0.208    0.836
## group3       0.116935   0.209236   0.559    0.578
## age         -0.039775   0.006322  -6.291 1.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6127 on 80 degrees of freedom
## Multiple R-squared:  0.3696, Adjusted R-squared:  0.3459
## F-statistic: 15.63 on 3 and 80 DF,  p-value: 4.323e-08
```

The model performs an analysis of covariance on the vitacap2 data. In this model we analyse the vital capacity, which is a measure of lung volume, with respect to the age of the workers in the cadmium industry, and by the group divided in three categories of being exposed or not.

Here, the intercept represents the mean value for the group 1.

```
anova(q2_model)
```

```
## Analysis of Variance Table
##
## Response: vital.capacity
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## group      2  2.7473  1.3737  3.6589   0.03016 *
## age        1 14.8589 14.8589 39.5781 1.572e-08 ***
## Residuals 80 30.0347  0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

However, even the 'anova' function given that the two independent variable are significant, we could see in the output of the R, using the 'summary' function that the variable factor 'group2' and 'group3' are not statistically significant, with p-value greater than 0.05.

```
shapiro.test(q2_model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  q2_model$residuals
## W = 0.99227, p-value = 0.9041
```

The Shapiro-Wilks test for normality of errors fails to reject the null hypothesis, that the errors are normally distributed around 0, with p-value greater than 0.05. Thus, we can accept the previous test as valid and sufficient for this analysis.

Analyzing another model, we will have:

```
q2_model2 = lm(vital.capacity ~ age, data = vitcap2)

summary(q2_model2)
```

```
##
## Call:
## lm(formula = vital.capacity ~ age, data = vitcap2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35136 -0.37332  0.02796  0.40735  1.42776
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.033316   0.247487  24.378  < 2e-16 ***
## age         -0.040478   0.005881  -6.883 1.08e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6068 on 82 degrees of freedom
## Multiple R-squared:  0.3662, Adjusted R-squared:  0.3584
## F-statistic: 47.37 on 1 and 82 DF,  p-value: 1.082e-09
```

Now, the model seems to fit better the data. The F-test for the model, which tests whether any of the individual coefficients are significant, has a p-value less than 0.05, which means that the null hyphotesis will be rejected.

```r
shapiro.test(q2_model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  q2_model2$residuals
## W = 0.99204, p-value = 0.8927
```

The Shapiro-Wilks test for normality of errors fails to reject the null hypothesis, that the errors are normally distributed around 0, with p-value greater than 0.05.

However, the model with only the variable 'age' seems not be enough useful to interpret and predict the vital capacity for workers in the cadmium industry.

Question 3

```r
data("malaria")
malaria$subject = factor(malaria$subject)

q3_model = glm(mal ~ age + log(ab), data = malaria, family = binomial(link = 'logit'))
summary(q3_model)
```

```
##
## Call:
## glm(formula = mal ~ age + log(ab), family = binomial(link = "logit"),
##     data = malaria)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8492  -0.7536  -0.4838   0.8809   2.5796
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.57234    0.95184   2.702 0.006883 **
## age         -0.06546    0.06772  -0.967 0.333703
## log(ab)     -0.68235    0.19552  -3.490 0.000483 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 116.652  on 99  degrees of freedom
## Residual deviance:  98.017  on 97  degrees of freedom
## AIC: 104.02
##
## Number of Fisher Scoring iterations: 5
```

This model performs a logistic regression on the malaria data. In the model we analyzed the incidence in children aged 3-15 years with or without symptoms of malaria. It was evaluated whether the age and the log-transformed antibody level have effects in the incidence of symptoms of malaria in children. The model initially is defined as:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 \text{age} + \beta_2 \log(\text{ab})$$

However, we could see in the output of the R, that the variable 'age' are not statistically significant, with p-value greater than 0.05. Thus, we may try to use just the log-transformed antibody level to evalute the data. The Residual Deviance has reduced by 18.635 with a loss of two degrees of freedom.

```
q3_model2 = glm(mal ~ log(ab), data = malaria, family = binomial(link = 'logit'))
summary(q3_model2)
```

```
##
## Call:
## glm(formula = mal ~ log(ab), family = binomial(link = "logit"),
##     data = malaria)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9159  -0.7339  -0.4854   0.8813   2.4722
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.1552     0.8401   2.565 0.010305 *
## log(ab)      -0.7122     0.1932  -3.686 0.000228 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 116.652  on 99  degrees of freedom
## Residual deviance:  98.968  on 98  degrees of freedom
## AIC: 102.97
##
## Number of Fisher Scoring iterations: 4
```

Now the model is defined as:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 \log(\text{ab})$$

With the corresponding values of the parameters: $\beta_0 = 2.1552$ and the $\beta_1 = -0.7122$. The probability of a child having malaria symptoms decreases with the level of antibodides.

The Residual Deviance has reduced by 17.684 with a loss of one degrees of freedom.

Question 4

```
data("graft.vs.host")
graft.vs.host$type = factor(graft.vs.host$type)

model_fitter = function(x){
  model = glm(x, data = graft.vs.host, family = binomial(link = 'logit'))
  aic = model$aic
  # Predicted Values:
  yhats = predict(model, newdata = graft.vs.host, type = 'response')
  # Hosmer Lemeshow GoF test:
  hosm = hoslem.test(graft.vs.host$gvhd, yhats)$p.value
  # Performance of the Model:
```

```
  pred = prediction(yhats, graft.vs.host$gvhd)
  perf = performance(pred, 'auc')@y.values[[1]]
  # CV Error:
  cverror = cv.glm(graft.vs.host, glmfit = model)$delta[1]
  return(c(AIC = aic, AUC = perf, GoF = hosm, CVErr = cverror))
}
```

the *model_fitter* function accepts a model formula, computes a model, and computes the model AIC (Akkaike's Information Criterion, a measure based on likelihood of the model, penalized for the number of predictor variables included in the model).

It also runs a Hosmer-Lemeshow Goodness of Fit test on values predicted by the model, indicating whether this model is a poor fit. The Hosmer Lemeshow Test is interpreted as follows:

$$\begin{aligned} H_0: &\quad \text{The model is a good fit for the data} \\ H_1: &\quad \text{The model is a poor fit for the data} \end{aligned}$$

If the p-value returned by the Hosmer-Lemeshow test is below our alpha value of 0.05, then we would conclude that the model is a poor fit to the data.

Lastly, it calculates an ROC curve, and computes the area under that curve. We can use this metric to compare the performance of competing models.

We have a few models we're interested in, with *preg*, *time*, and various transformations of *index* being used to predict *gvhd*.

```
models = c(
  Raw = (gvhd ~ preg + time + index),
  Log = (gvhd ~ preg + time + log(index)),
  Sqrt = (gvhd ~ preg + time + sqrt(index)),
  OtherLog = (gvhd ~ preg + time + log(index + sqrt(1 + index^2)))
)

sapply(models, model_fitter)
```

```
##              Raw        Log       Sqrt   OtherLog
## AIC   32.3160701 29.4982225 30.8256959 30.4351143
## AUC    0.9294118  0.9441176  0.9441176  0.9470588
## GoF    0.4107372  0.9832021  0.9862216  0.9857342
## CVErr  0.1527924  0.1481630  0.1475760  0.1477514
```

We're interested in the model with the minimum AIC that meets our Goodness of Fit criterion, and has a high model performance as measured by the area under the ROC curve. Given these criteria, it seems that model 2, with the strait log transformation of *index*, has the lowest AIC, and AUC and Cross-validated error not vastly different than the competing models. The goodness of fit test does not find evidence that it is mis-specified either

The ROC curve for the model with variables 'preg', 'time', and 'log(index)' is displayed below: