Exam 2 - Stat 431

Sebastian Nuxoll

14 Oct 2024

1. Assume that you want to model the relationship between weight (as the response variable), age, height, and nationality for individuals from US, Canada, and Mexico. Assume that there is an interaction between height and nationality.

- Write down the correct R formula to run this model.

ANSWER:

```
model <- lm(weight ~ age + height + nationality + height:nationality, dat)
```

- How many parameters would be estimated in this model?

ANSWER:

There should be 6 parameters that need to be estimated: an intercept, one slope for age, another for height, two slopes for nationality, since it is categorical, and 2 interactions between height and nationality.

- Giving the estimated slopes names such as β_{US} for the effect of living in the US or $\beta_{\text{Age}\times\text{Height}}$ for an interaction slope, and indicate variable values by the variable name, write down the 3 equations for the 3 estimated lines.

ANSWER:

Where A is age and H is height:

$$W = \beta_0 + \beta_A A + \beta_H H + \beta_{US} + \beta_{H\times US} H$$

$$W = \beta_0 + \beta_A A + \beta_H H + \beta_{Canada} + \beta_{H\times Canada} H$$

$$W = \beta_0 + \beta_A A + \beta_H H + \beta_{Mexico} + \beta_{H\times Mexico} H$$

- Assume that you have the following observation: Weight=80kg, Age=18yo, Height=188cm, Nationality=Mexico. Write down the row in the *model matrix* that would correspond to this observation. Label the "column" according to the subscripts you put on the slopes in the previous part. For example, if you were looking at a age and height interaction, the "column" for this would look like "Age x Height" = 3384. You should have the same number of "columns" as you do parameters in your model.

ANSWER:

Intercept	Age	Height	Canada	Mexico	HeightXCanada	HeightXMexico
1	18	188	0	0	0	188

---

2. You get the following output from R:

```
anova(m2,m1)

## Analysis of Variance Table
##
## Model 1: PlantMercuryConc ~ SoilMercuryConc/Crop
## Model 2: PlantMercuryConc ~ Crop * SoilMercuryConc
##   Res.Df   RSS Df.  SumofSq      F  Pr(>F)
## 1     86 14948
## 2     84 14925  2    23.345 0.0657 0.9365
```

- Which model should you favor? Why?

ANSWER:

We can see that we have an F value of 0.0657, which is tiny, and a p value of 94%, which is almost 1, meaning the second model make minimal improvements, and we should stick to the first model.

- What does the `Df = 2` mean? (How is it calculated?)

ANSWER:

`Df` is the number of extra parameters the second model has over the first model. In this case, the second model has two extra parameters.

- Using the numbers given in the table, how do you calculate the F value given in the table? (i.e. actually write down the equation)

ANSWER:

$$F = \frac{\frac{SumofSq}{Df}}{\frac{RSS_2}{Res.Df_2}} = \frac{\frac{23.345}{2}}{\frac{14925}{84}} = \frac{11.673}{177.679} = 0.0657$$

- Write the R command that would yield the $p = 0.9365$ displayed in the table.

ANSWER:

```
anova(m2,m1)[6][2]
```

---

3. Read in the files `ModelMatrix.rds` and `ResponseVar.rds` as `X` and `y`, respectively. Using those variables answer the following. DO NOT USE LM TO ANSWER ANY PART OF THIS QUESTION. You must show your R code for each calculation.

ANSWER:

```
X <- readRDS("ModelMatrix.rds")
y <- readRDS("ResponseVar.rds")
```

- Using $\mathbb{X}$ find the hat matrix.

ANSWER:

```
H <- X %*% solve(t(X) %*% X) %*% t(X)
```

- Using the hat matrix, find the predicted values of y for this model.

ANSWER:

```
predy <- H %*% y
```

- Using $\mathbb{X}$ find the parameter estimate vector.

ANSWER:

```
betahat <- solve(t(X) %*% X) %*% t(X) %*% y
```

- Calculate the residual, regression, and total sums of squares using the *linear algebra* operations. (i.e. use vectors and their transposition to get the sums of squares)

ANSWER:

```
RSS <- t(y - predy) %*% (y - predy)
RegSS <- t(predy) %*% predy
TSS <- t(y) %*% y
```

RSS: 782.3206284

Regression SS: 1.2755514×10^6

TSS: 1.2763338×10^6

- Calculate the residual standard error σ_ϵ.

ANSWER:

```
df <- nrow(X)-ncol(X)
RSE <- sqrt(RSS/df)
```

RSE: 2.8848847

- Using $(\mathbb{X}^T\mathbb{X})^{-1}$, find the standard error of the parameter estimates.

ANSWER:

```
covmat <- solve(t(X) %*% X)
paramse <- sqrt(diag(covmat)) * RSE
```

```
## Warning in sqrt(diag(covmat)) * RSE: Recycling array of length 1 in vector-array arithmetic is depre
##   Use c() or as.vector() instead.
```

```
kable(t(paramse))
```

Intercept	x1	x2	x3	x4	x5
1.694813	0.5022856	0.4175067	0.0422061	0.0361321	0.0381531

- Find the t-values for the hypothesis $H_0 : \beta_i = 0, H_a : \beta_i \neq 0$.

ANSWER:

```
tvals <- betahat / paramse
kable(t(tvals))
```

Intercept	x1	x2	x3	x4	x5
0.4308854	0.3171648	-1.415963	70.32332	139.9767	-261.5233

- Find the p-values for the t-values.

ANSWER:

```
pvals = 2 * pt(-abs(tvals), df=df)
kable(format(t(pvals)))
```

Intercept	x1	x2	x3	x4	x5
6.675382e-01	7.518219e-01	1.600915e-01	4.400980e-83	6.688306e-111	2.385164e-136

- What is R^2?

ANSWER:

```
R2 <- RegSS/TSS
```

R^2: 0.9993871

- What is the F-value for the hypothesis that all of the slopes are equivalent to 0?

ANSWER:

```
f <- (TSS-RSS)/(ncol(X)-1)/(RSS/df)
```

F value: 3.0652863×10^4

- What is the p-value for the F-value you just calculated?

ANSWER:

```
p <- pf(f, df1 = ncol(X), df2 = df, lower.tail = F)
```

p value: $2.287201 \times 10^{-152}$

- What are the leverage values for each observation? Which of the leverage values, if any, are of concern?

ANSWER:

```
kable(t(diag(X)))
```

1	1.749811	7.162172	18.67876	16.42092	1.76952

- What are Cook's D values for each observation? What of the Cook's D values, if any, are of concern?

ANSWER:

```
cooksd <- ((y - predy) ^ 2 / (p * RSE^2)) * (diag(X) / (1 - diag(X))^2)
```

```
## Error in (y - predy)^2/(p * RSE^2): non-conformable arrays
```

```
kable(t(cooksd))
```

```
## Error in t(cooksd): object 'cooksd' not found
```

---

4. Using the data in Exam2.rds do the following. *The response variable is labelled 'c' in these data.*

- Using forward selection based on the χ^2 statistic, add variables sequentially to find the best model. The largest model to consider would be one that has all terms and two-interactions between the terms. Use a cutoff value of $p = 0.15$ for a variable to enter the model. Start with a model that is simply the

intercept. At each step of the selection process record/calculate: R^2, adjusted R^2, PRESS, Mallow's Cp, AIC, and BIC. Put those values in a table.

ANSWER:

```
dat <- readRDS("Exam2.rds")

form <- c ~ d + T1 +   T2 +   s +   pr +   ne +   ct +   bw +   n +   pt +
              d:T1 + d:T2 + d:s + d:pr + d:ne + d:ct + d:bw + d:n + d:pt +
                    T1:T2 +T1:s +T1:pr +T1:ne +T1:ct +T1:bw +T1:n +T1:pt +
                           T2:s +T2:pr +T2:ne +T2:ct +T2:bw +T2:n +T2:pt +
                                  s:pr + s:ne + s:ct + s:bw + s:n + s:pt +
                                        pr:ne +pr:ct +pr:bw +pr:n +pr:pt +
                                               ne:ct +ne:bw +ne:n +ne:pt +
                                                      ct:bw +ct:n +ct:pt +
                                                            +bw:n +bw:pt +
                                                                  + n:pt
model <- lm(c~1, dat)
table <- data.frame()
repeat {
  table = rbind(table, c(summary(model)$r.squared,
                         summary(model)$adj.r.squared,
                         sum((residuals(model) / (1 - hatvalues(model)))^2),
                         sum(residuals(model)^2) / summary(model)$sigma^2 - length(coefficients(model)),
                         AIC(model),
                         BIC(model)))
  perms <- add1(model, form, test = "Chisq")
  if(min(perms$'Pr(>Chi)', na.rm = T) > 0.1) break
  model <- update(model, as.formula(paste(". ~ . + ", rownames(perms)[which.min(perms$'Pr(>Chi)')])))
}
```

```
## Warning: attempting model selection on an essentially perfect fit is nonsense
```

```
## Warning in min(perms$"Pr(>Chi)", na.rm = T): no non-missing arguments to min;
## returning Inf
```

```
names(table) <- c("$R^2$", "Adj. $R^2$", "PRESS", "Mallow's Cp", "AIC", "BIC")
kable(table)
```

R^2	Adj. R^2	PRESS	Mallow's Cp	AIC	BIC
0.0000000	0.0000000	955987.98	30	422.53262	425.46409
0.3726544	0.3517429	625603.09	28	409.61237	414.00958
0.5841319	0.5554513	451165.64	26	398.45623	404.31918
0.6910419	0.6579392	338292.70	24	390.94704	398.27572
0.7518682	0.7151080	294231.78	22	385.93117	394.72558
0.7733022	0.7297064	304073.67	20	385.04023	395.30038
0.8017144	0.7541258	273801.43	18	382.75513	394.48101
0.8188402	0.7660019	278967.00	16	381.86459	395.05622
0.8393129	0.7834218	300990.02	14	380.02713	394.68449
0.8577762	0.7995937	288321.21	12	378.12131	394.24441
0.8766952	0.8179786	296594.05	10	375.55356	393.14239
0.8976810	0.8414055	206991.34	8	371.58351	390.63808
0.9100956	0.8533139	201221.80	6	369.44434	389.96465
0.9198293	0.8619282	199284.36	4	367.77751	389.76355

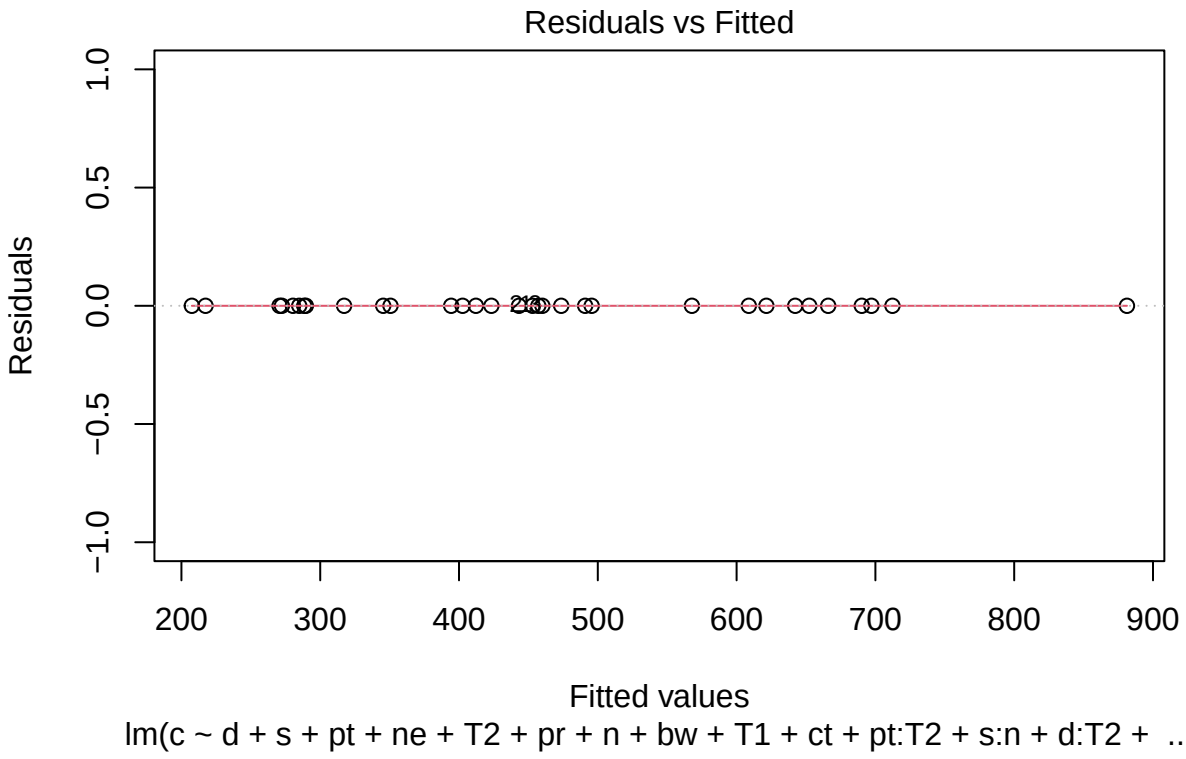
R^2	Adj. R^2	PRESS	Mallow's Cp	AIC	BIC
0.9370465	0.8852025	189327.73	2	362.04112	385.49289
0.9496173	0.9023836	205640.52	0	356.91317	381.83068
0.9581953	0.9136036	227831.65	-2	352.94074	379.32399
0.9690512	0.9314705	148874.85	-4	345.31916	373.16814
0.9759066	0.9425466	121606.22	-6	339.30642	368.62114
0.9790966	0.9459995	166205.76	-8	336.76163	367.54209
0.9899750	0.9717479	144184.78	-10	315.24692	347.49311
0.9953441	0.9855666	75137.40	-12	292.70502	326.41695
0.9963960	0.9875862	118730.89	-14	286.50987	321.68753
0.9972282	0.9892591	107816.56	-16	280.10873	316.75213
0.9976141	0.9894341	154065.80	-18	277.31024	315.41937
0.9980437	0.9898923	157316.56	-20	272.95883	312.53370
0.9986394	0.9915644	515505.05	-22	263.33760	304.37820
0.9989681	0.9920027	473171.57	-24	256.48956	298.99590
0.9996847	0.9967418	304326.95	-26	220.54979	264.52187
0.9999920	0.9998765	149689.91	-28	104.85868	150.29649
1.0000000	0.9999991	30916.39	-30	-71.82788	-24.92434
1.0000000	NaN	NaN	NaN	-Inf	-Inf

- Plot the default diagnostic plots of your best model. Comment on each one of the plots and what you see in them with respect to meeting model assumptions.

ANSWER:

```
plot(model)
```

```
## Warning: not plotting observations with leverage one:
##   1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27,
```




```
## Warning in min(x): no non-missing arguments to min; returning Inf
## Warning in max(x): no non-missing arguments to max; returning -Inf
## Error in qqnorm.default(rs, main = main, ylab = ylab23, ylim = ylim, ...): y is empty or has only NAs
```

I'm more than a little confused by the model I've created. It seems to be perfect (or at least drastically overfit). It seems that there is no noise in the data. The plots reflect this, by having no residuals.

- Calculate the VIF for each term in your model and make a table of the results. Which term shows the highest collinearity with the others?

ANSWER:

I don't even need to calculate the VIF, as it'll be 0. Once again, either I did something wrong or there is no noise in the data. The model shouldn't be perfect like this.

- Find the 70% prediction interval *at the mean value* of all the predictors in your best model.

ANSWER:

The prediction interval is just a point, since there is no error in the model. I don't know what to say.

---

5. What is the model underlying mulitple regression? First write the model in terms of $\mu_i = f(x_i)$. Then specify how y_i relates to μ_i. From that relationship, find the formula for the residual ϵ_i and give the distribution for all residuals. Using matrix algebra notation is not required (but nice).

ANSWER:

$$\mu_i = \beta_0 + \beta_1 + x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$
$$\epsilon_i = y_i - \mu_i = y - \beta_0 - \beta_1 - x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}$$
$$\epsilon_i \sim N(0, \sigma^2)$$

- How many parameters are we estimating in a multiple regression model? How can we determine this number from the model matrix?

ANSWER:

Generally, there is one parameter for the intercept, then one per predictor. This is the width of the model matrix.

- Finally, state what the assumptions are from the model you have outlined above.

ANSWER:

We assume that the relationships are linear, the observations are independent, and that the residuals are normally distributed with a constant variance.

---

6. Assuming we have a model where want to estimate an observer effect; the observers in the data set are Eddie, Claire, Bob, and Sloane. Assume R's default factor encoding to answer the following.
- How many columns will be added to the model matrix if we wish to estimate this effect?

ANSWER:

Since there are 4 observers, we'll need to add 3 columns.

- Which observer will be the base/default level? What does this mean in terms of interpreting model parameters?

ANSWER:

Bob will be the base, since he comes first alphabetically. This means any effect inherent to Bob will be reflected in the intercept.

- Write down a matrix with rows for each observer and the correct number of columns; fill in the dummy encoding given by R.

ANSWER:

Observer	Claire	Eddie	Sloane
Bob	0	0	0
Claire	1	0	0
Eddie	0	1	0
Sloane	0	0	1

- If we compare a model with the observer effect versus without, what value of F, at $\alpha = 0.05$, would be required to add the observer effect to the model? Assume there are 30 observations and 10 parameters in the model with the effect.

ANSWER:

```
qf(0.95, df1 = 3, df2 = 20)
```

F value: 3.0983912
