ST553 HW 4

Nick Sun

April 29, 2019

Question 1

Y is our response vector, **X** is our design matrix, and **H** is our hat matrix defined as $X(X^TX)^{-1}X^T$.

a. Give an expression for R, the residual vector as a function of Y and H

Given that $\hat{Y} = HY = X(X^TX)^{-1}X^TY$:

$$R = Y - \hat{Y} = Y - X(X^TX)^{-1}X^TY = Y(I - H)$$

b. Calculate the variance-covariance matrix of R

Here we use the fact that the variance-covariance matrix of **Y** is $\sigma^2 I$

$$\begin{aligned} Var(R) &= Var(Y(I-H)) \\ &= (I-H)Var(Y)(I-H)^T \\ &= (I-H)\sigma^2 I(I-H)^T \\ &= \sigma^2(I-H) \end{aligned}$$

We get the last equality from the fact that $(I-H)$ is symmetric so $(I-H) = (I-H)^T$ and idempotent so $(I-H)^2 = (I-H)$.

c. For the balanced CRD with one treatment factor, what is the kth diagonal element of the variance-covariance matrix?

Since we are assuming constant variance of the residuals, the diagonal elements will just be $\sigma^2\left(1 - \frac{1}{n}\right)$.

An heuristic argument using $g = 3, n = 2$:

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]  0.5 -0.5  0.0  0.0  0.0  0.0
## [2,] -0.5  0.5  0.0  0.0  0.0  0.0
## [3,]  0.0  0.0  0.5 -0.5  0.0  0.0
## [4,]  0.0  0.0 -0.5  0.5  0.0  0.0
## [5,]  0.0  0.0  0.0  0.0  0.5 -0.5
## [6,]  0.0  0.0  0.0  0.0 -0.5  0.5
```

We can see that the diagonal entries are $\frac{1}{2}$ which is equivalent to $(1 - \frac{1}{n})$ since our n = 2.

Question 2

Let's derive this cool sample size formula for $H_0 : \mu_1 = \mu_2$

$$n \geq 2(Z_{\alpha/2} + Z_\beta)^2 \frac{\sigma^2}{\delta^2}$$

where $\delta = \mu_1 - \mu_2$ and Z_α is the standard normal quantile with α area in the *upper* tail.

a. Write down the rejection region for an α level test of H_0

Assuming the two populations are normal with equal variance σ^2 with sample size of n from each population, the following is true:

$$\frac{\bar{X}_1 - \bar{X}_2 - \delta}{\sqrt{2\sigma^2/n}} \sim N(0,1)$$

Therefore, the rejection region for a two sided test is

$$|\frac{\bar{X}_1 - \bar{X}_2 - \delta}{\sqrt{2\sigma^2/n}}| > Z_{\alpha/2}$$

b. Show some powuh

If $\delta > 0$, show that the power of the test is approximately equal to

$$P\left(Z > Z_{\alpha/2} - \frac{\delta}{\sqrt{2\sigma^2/n}}\right)$$

Power is defined as the probability of rejection given that the alternative hypothesis is true. Our in hypothesis test, we always begin under the assumption that the null hypothesis $\delta = 0$ is true. In that case, our test statistic is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{2\sigma^2/n}} \sim N(0,1) \text{ under } H_0$$

We reject H_0 when $Z > Z_{\alpha/2}$. However, if $\delta > 0$, to find the power of our test we need to find the probability that Z falls in the rejection region of our test given that the true distribution is actually centered around δ. Really what we have is that $Z \sim N(\delta, 1)$.

Assuming that the difference between our null distribution and true distribution is just a location shift from 0 to δ, the quantile $Z_{\alpha/2}$ in the null distribution is the same as $Z_{\alpha/2} - \frac{\delta}{\sqrt{2\sigma^2/n}}$ in the true distribution.

c.

β is our Type II error i.e. the probability we *do not* reject, given that the alternative hypothesis is true.

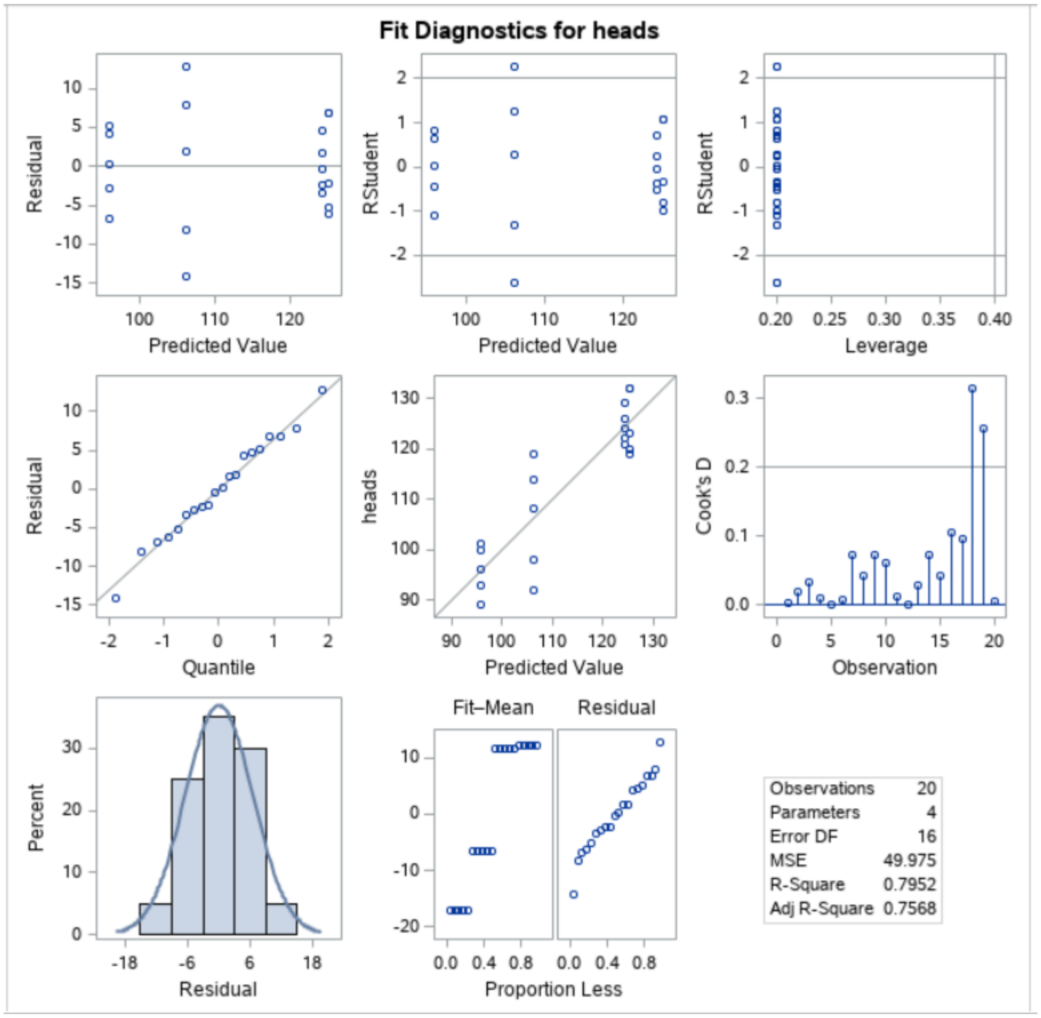
In part b., we found that we would reject if our test statistic was greater than $Z_{\alpha/2} - \frac{\delta}{\sqrt{2\sigma^2/n}}$, so getting a statistic less than this quantile would lead to a type II error. Therefore, $Z_{\alpha/2} - \frac{\delta}{\sqrt{2\sigma^2/n}} = -Z_\beta$

$$
\begin{aligned}
Z_{\alpha/2} - \frac{\delta}{\sqrt{2\sigma^2/n}} &= -Z_\beta \\
Z_{\alpha/2} + Z_\beta &= \frac{\delta\sqrt{n}}{\sqrt{2\sigma^2}} \\
\left(Z_{\alpha/2} + Z_\beta\right)\frac{\sqrt{2\sigma^2}}{\delta} &= \sqrt{n} \\
n = 2(Z_{\alpha/2} - Z_\beta)^2 \frac{\sigma^2}{\delta^2}
\end{aligned}
$$

To complete the formula, we recognize that this n is just a lower bound so

$$n \geq 2(Z_{\alpha/2} + Z_\beta)^2 \frac{\sigma^2}{\delta^2}$$

Question 3



```
Observations       20
Parameters          4
Error DF           16
MSE            49.975
R-Square       0.7952
Adj R-Square   0.7568
```

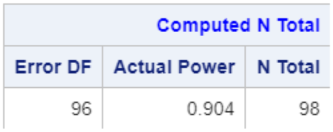
The normality of the residuals actually looks pretty good - the QQ plot and histogram both look approximately normal. The RF plot doesn't look bad either, we see that the spread of the residuals is approximately equal to the spread of the fitted values. A bad RF plot would be if the spread of the residuals was significantly larger than the spread of the fitted values. The only potential problem is constant variance - the top two plots suggest that one of the treatment groups has a much larger spread than the other groups. This nonconstant variance might lead to a poor variance estimate from the residuals. Since this is only one group though, we should be fine.

Question 4

Here we are using `PROC GLMPOWER` to do a power calculation for an experiment. The setup for this is that we have an estimate of the error variance (.218), we want power of .90, $\alpha = .05$, and the alternative is that soy treatment raises the estradiol concentration by 25% or .22 log units.

```
data dat2; /* Create example data set. */
   input group estradiol;
   datalines;
      1 1
      2 1.22
   ;
run;
proc glmpower data=dat2;
   class group;
   model estradiol=group;
   power
        stddev=0.33
        alpha=0.05
        ntotal=.
        power=0.90;
run;
```

The output of this command is:



Computed N Total		
Error DF	Actual Power	N Total
96	0.904	98

So N = 98 is the sample size we need to detect a difference of .22 log units with power = .90