Stat500(Section002): Homework #9

Due on Dec.06, 2021 at 3:00pm

Instructor:Naisyin Wang

Tiejin Chen **tiejin@umich.edu**

Problem 1

Part b

We using the following code to prepare the data ,get the model and test for a difference between breeds.

```
library(faraway)
data(butterfat)
butterfat = butterfat[which(butterfat$Age== "Mature"),]
lmod = lm(Butterfat~Breed, butterfat)
lmnull = lm(Butterfat ~ 1, butterfat)
anova(lmnull,lmod)
```

We can get the result:

```
Analysis of Variance Table

Model 1: Butterfat ~ 1
Model 2: Butterfat ~ Breed
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     49 25.4546
2     45  5.8306  4    19.624 37.864 7.284e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

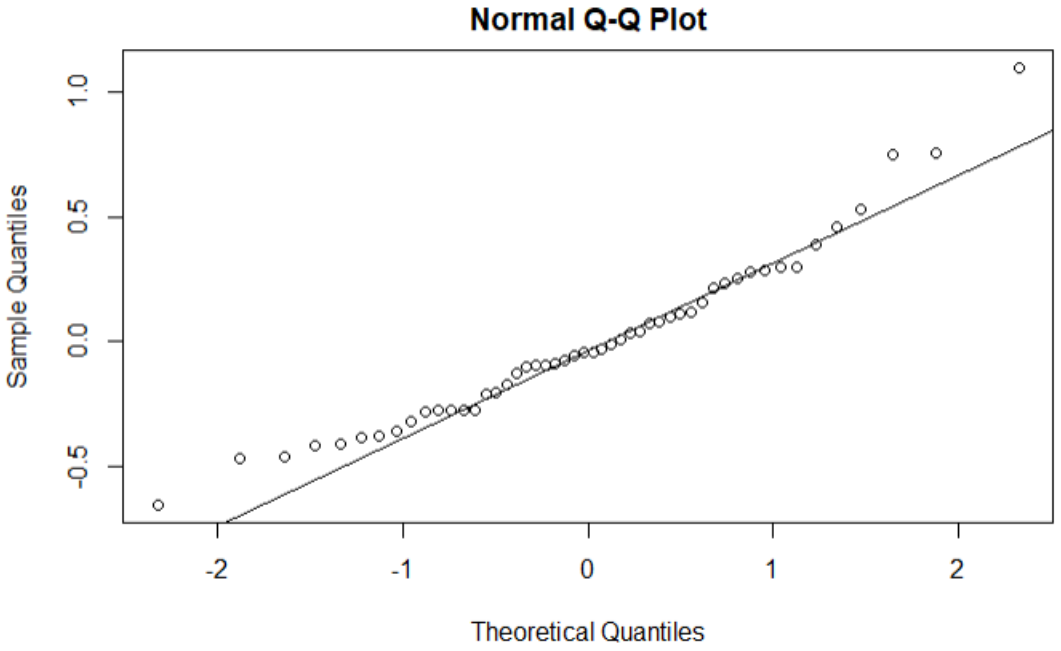
The p value here is smaller than 0.05. Hence there is a difference between breeds indeed.

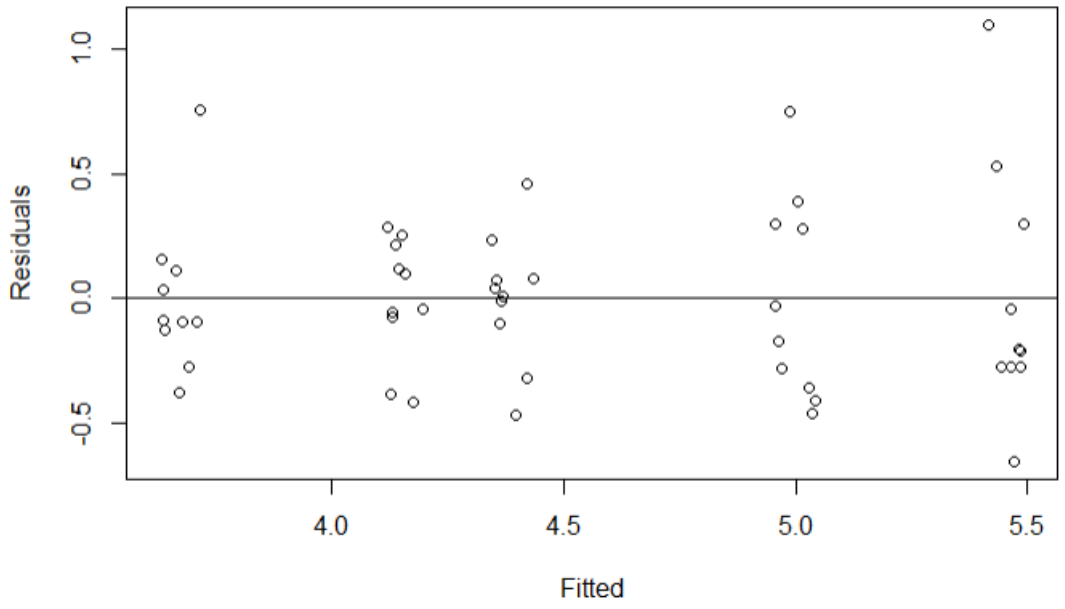
Part c

Using the following code to check the diagnostics.

```
qqnorm(residuals(lmod))
qqline(residuals(lmod))
plot(jitter(fitted(lmod)),residuals(lmod),xlab="Fitted",
ylab="Residuals")
abline(h=0)
```

We can get the result:





We can see the residuals center at 0, Hence there is no non-linear trend. And from the QQ-plot. on the whole,we can see it is a normal distribution.

Part d

```
TukeyHSD(aov(Butterfat~Breed, butterfat))
plot(TukeyHSD(aov(Butterfat~Breed, butterfat)))
```

We can get the results:


```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Butterfat ~ Breed, data = butterfat)

$Breed
                              diff          lwr         upr     p adj
Canadian-Ayrshire            0.235 -0.222410494  0.69241049 0.5932057
Guernsey-Ayrshire            0.847  0.389589506  1.30441049 0.0000366
Holstein-Fresian-Ayrshire   -0.478 -0.935410494 -0.02058951 0.0365566
Jersey-Ayrshire              1.298  0.840589506  1.75541049 0.0000000
Guernsey-Canadian            0.612  0.154589506  1.06941049 0.0037519
Holstein-Fresian-Canadian   -0.713 -1.170410494 -0.25558951 0.0005502
Jersey-Canadian              1.063  0.605589506  1.52041049 0.0000004
Holstein-Fresian-Guernsey   -1.325 -1.782410494 -0.86758951 0.0000000
Jersey-Guernsey              0.451 -0.006410494  0.90841049 0.0549908
Jersey-Holstein-Fresian      1.776  1.318589506  2.23341049 0.0000000
```

From the plot and the result of TukeyHSD. We can say that the different between Canadian-Ayrshire pair is not significant since 0 is in their 95% confidence interval. Another insignificant pair might be Jersey-Guernsey with the same reason, though their p value only greater than 0.05 a little bit and their lower bound is just slightly smaller than 0.

Problem 2

Part a

With the output of R code, we can write the model: Let $p = Pr(test = 1)$, we can know:

$$log(\frac{p}{1-p}) = -9.778884 + 0.300602 \times pregnhigh + 0.042540 \times glucose + 0.085366 \times bmi$$
$$+ 0.936986 \times diabetes - 0.001317 \times pregnhigh \times glucose.$$

Part b

According to the requirement, we know:

$$H_0 : \beta_{bmi} \leq 0, H_A : \beta bmi > 0$$

```
confint(test.out)
```

We can get:

```
Waiting for profiling to be done...
                              2.5 %        97.5 %
(Intercept)       -11.632065434 -8.0610203527
pregnant            0.041357327  0.5567367215
glucose             0.031513080  0.0543252163
bmi                 0.056970749  0.1149063238
diabetes            0.350942720  1.5358685515
pregnant:glucose   -0.003310256  0.0007120542
```

We can see that the lower bound of 95% confidence interval of bmi is greater than 0. Hence, we reject H_0, and we will think that a higher level of bmi would lead to a higher chance of showing signs of diabetes.

Part c

Since there only two level of pregn in test2, what we want to consdier the is the $\beta pregn : glucose$.

```
Call:
glm(formula = test ~ pregn + glucose + bmi + diabetes + pregn:glucose,
    family = binomial, data = pima.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8719  -0.7328  -0.4344   0.7428   2.4432

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -8.71418    0.69166 -12.599  < 2e-16 ***
pregnhigh          3.29173    1.48525   2.216  0.02667 *
glucose            0.03971    0.00368  10.790  < 2e-16 ***
bmi                0.07943    0.01437   5.526 3.28e-08 ***
diabetes           0.89498    0.29917   2.992  0.00278 **
pregnhigh:glucose -0.02072    0.01175  -1.763  0.07785 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

We can see the p-value of $\beta pregn : glucose$ is greater than 0.05, which shows that it is not significant. Hence the difference bewteen two level of pregnant associated with glucose level is not significant. Hence we get a conclusion that the association between glucose level and the chance of showing signs of diabetes does not changes with two levels of times being pregnant.