Brandon Maahs

Homework 9

1.

R2=

F=

MSModel= =

MSE= =

F= =

= = 1 +

-1 =

=

F=

2. a.

drug.fit = lm(COST ~ GS + RI + RXPM + COPAY + AGE + F + MM, data=Drug)

> summary(drug.fit)

Call:

lm(formula = COST ~ GS + RI + RXPM + COPAY + AGE + F + MM, data = Drug)

Residuals:

Min 1Q Median 3Q Max

-0.142888 -0.050521 -0.003367 0.047232 0.122523

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.851e+00 7.636e-01 2.424 0.024488 \*

GS -1.137e-02 2.830e-03 -4.018 0.000622 \*\*\*

RI 3.341e-04 2.089e-03 0.160 0.874468

RXPM 2.241e-02 1.100e-02 2.037 0.054483 .

COPAY 1.472e-02 1.870e-02 0.787 0.439791

AGE -3.754e-02 1.491e-02 -2.517 0.020012 \*

F 1.297e-02 9.712e-03 1.335 0.196148

MM 2.908e-08 4.163e-08 0.699 0.492505

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08276 on 21 degrees of freedom

Multiple R-squared: 0.5758, Adjusted R-squared: 0.4344

F-statistic: 4.072 on 7 and 21 DF, p-value: 0.00572

The F-ratio is 4.072 with 7 and 21 degrees of freedom and a p-value of 0.00572. The Multiple R2 is 0.5758.

2.b.

There is strong evidence that the percent generic substitution used by the plan (GS) will reduce drug costs, but there is no evidence that the restrictiveness index (RI) will.

2.c.

Other important variables are RXPM and AGE. We have some evidence that the average member age increases drug costs and weak evidence that the average number of prescriptions per member per year increases drug costs.

2.d.

confint(drug.fit)

2.5 % 97.5 %

(Intercept) 2.626334e-01 3.438775e+00

GS -1.725825e-02 -5.486703e-03

RI -4.010217e-03 4.678382e-03

RXPM -4.721503e-04 4.529921e-02

COPAY -2.416036e-02 5.360987e-02

AGE -6.855516e-02 -6.528358e-03

F -7.231209e-03 3.316418e-02

MM -5.749404e-08 1.156542e-07

If we only consider the important variables (GS, RXMP and AGE) we are 95% confident, after accounting for all other predictors, that there will be a reduction of drug cost by 0.00549 to 0.0173 dollars with the percent generic substitution (GS).

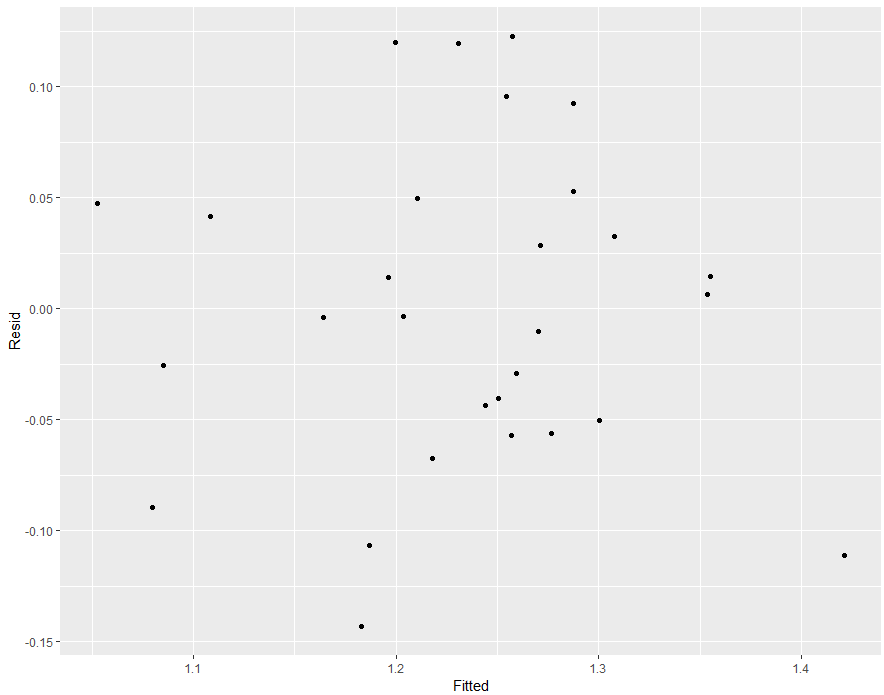
2.e.

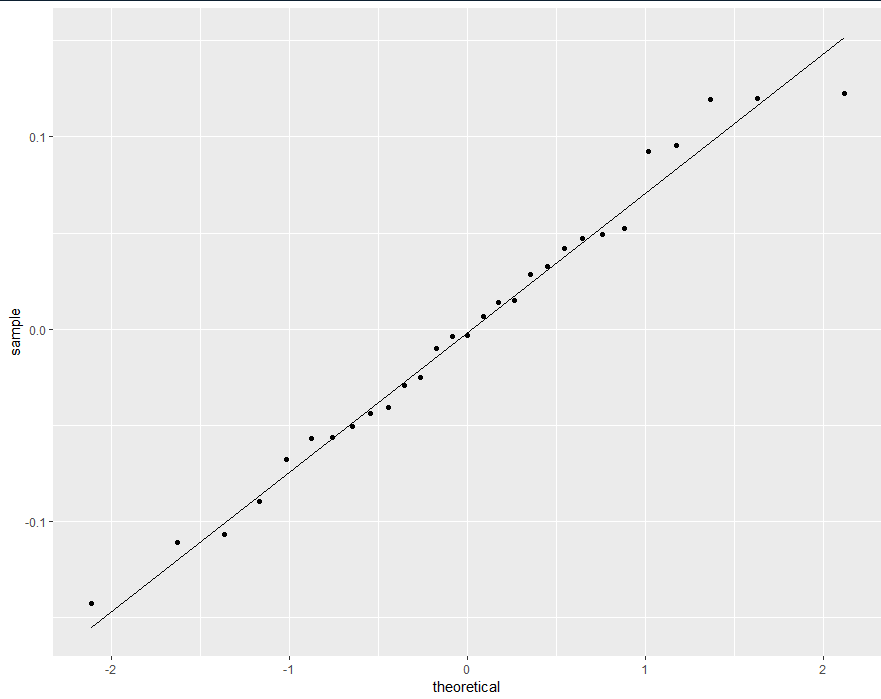
Drug$Resid = resid(drug.fit)

> Drug$Fitted = fitted(drug.fit)

> ggplot(Drug, aes(x=Fitted, y=Resid)) + geom\_point()

> ggplot(Drug, aes(sample=Resid)) + stat\_qq() + stat\_qq\_line()





It doesn’t seem like any assumptions are violated. The points in the residuals vs. fitted plot are equally spread out about zero and the point fall along the line in the QQ-plot.

3.a.

dace.fit = lm(longnosedace ~ acreage + do2 + maxdepth + no3 + so4 + temp, data = dace)

> summary(dace.fit)

Call:

lm(formula = longnosedace ~ acreage + do2 + maxdepth + no3 +

so4 + temp, data = dace)

Residuals:

Min 1Q Median 3Q Max

-57.428 -25.028 -2.215 10.667 170.017

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.276e+02 6.642e+01 -1.921 0.05941 .

acreage 1.962e-03 6.753e-04 2.906 0.00509 \*\*

do2 6.104e+00 5.384e+00 1.134 0.26135

maxdepth 3.542e-01 1.784e-01 1.985 0.05167 .

no3 7.713e+00 2.905e+00 2.655 0.01011 \*

so4 -8.605e-03 7.735e-01 -0.011 0.99116

temp 2.748e+00 1.694e+00 1.622 0.10997

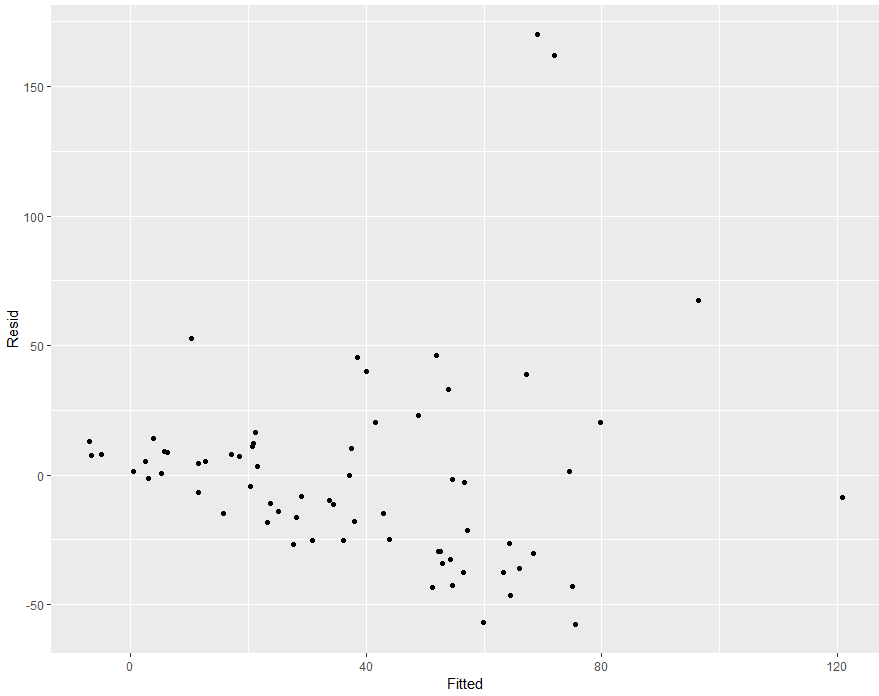
---

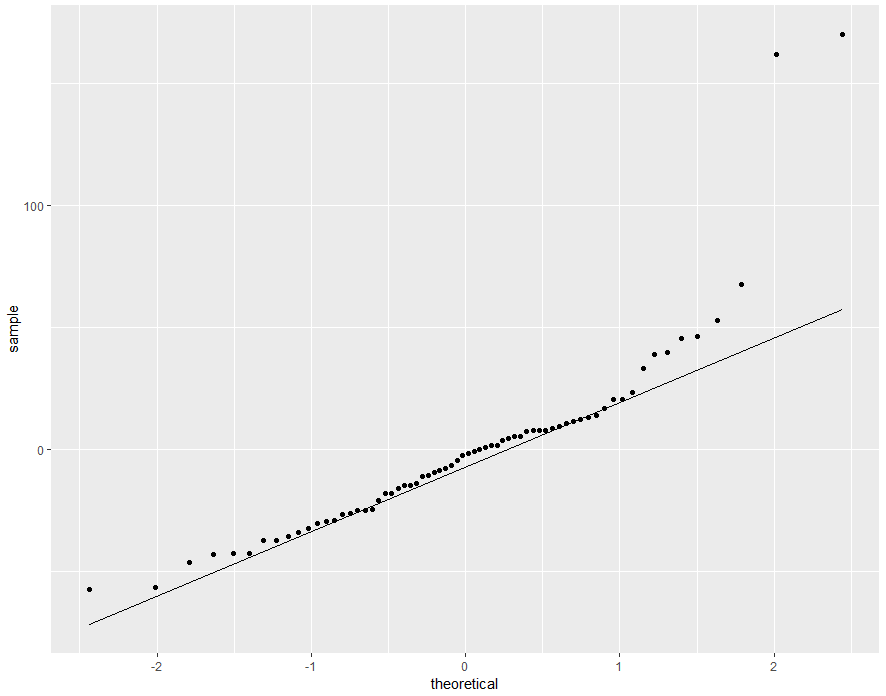
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.05 on 61 degrees of freedom

Multiple R-squared: 0.314, Adjusted R-squared: 0.2465

F-statistic: 4.653 on 6 and 61 DF, p-value: 0.0005905





The assumptions appear to be violated. The points in the residual vs. fitted plot are not evenly around zero, there seems to be a declining trend. Also, the points in the QQ-plot do not fall along the line.

3.b.

dace.fit2 = lm(log(longnosedace) ~ acreage + do2 + maxdepth + no3 + so4 + temp, data = dace)

> summary(dace.fit2)

Call:

lm(formula = log(longnosedace) ~ acreage + do2 + maxdepth + no3 +

so4 + temp, data = dace)

Residuals:

Min 1Q Median 3Q Max

-2.63686 -0.56544 0.05159 0.71044 1.81289

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.758e+00 1.593e+00 -2.359 0.02155 \*

acreage 5.103e-05 1.620e-05 3.150 0.00253 \*\*

do2 3.921e-01 1.291e-01 3.036 0.00352 \*\*

maxdepth 8.997e-03 4.281e-03 2.102 0.03971 \*

no3 2.109e-01 6.970e-02 3.026 0.00363 \*\*

so4 8.863e-03 1.856e-02 0.478 0.63459

temp 8.767e-02 4.065e-02 2.157 0.03497 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9847 on 61 degrees of freedom

Multiple R-squared: 0.4193, Adjusted R-squared: 0.3621

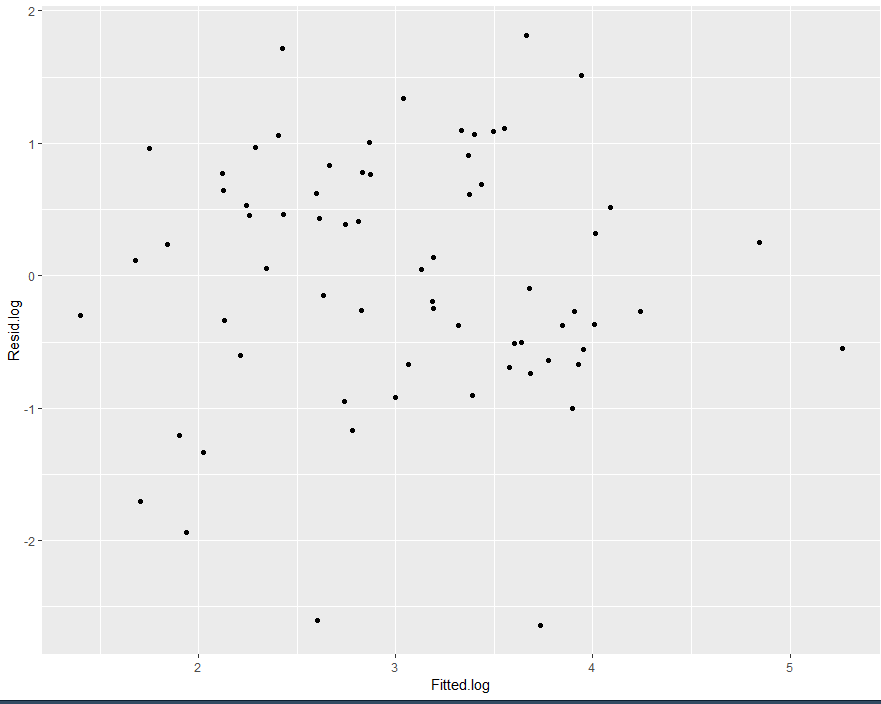
F-statistic: 7.34 on 6 and 61 DF, p-value: 6.217e-06

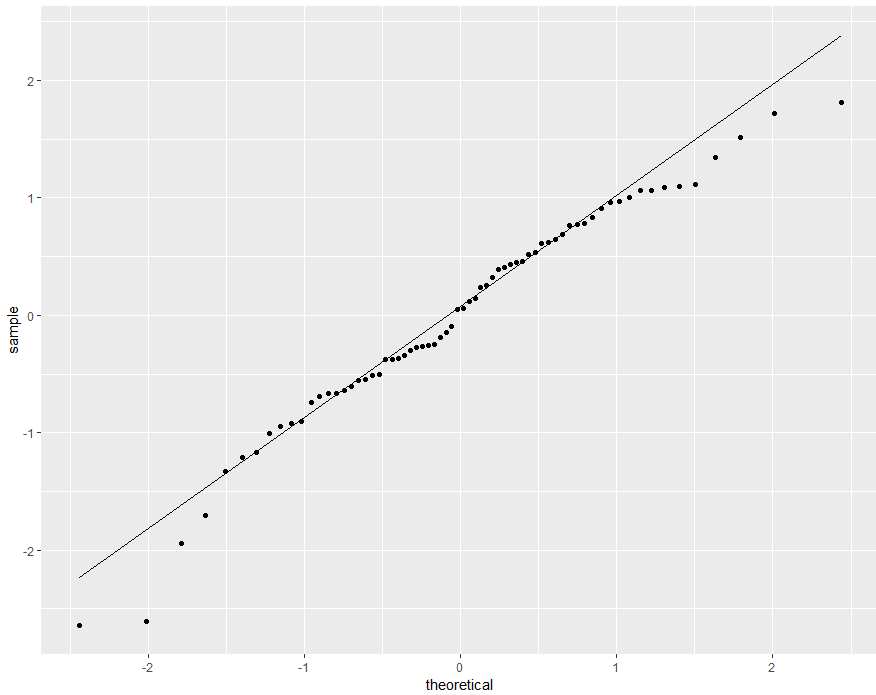
dace$Resid.log = resid(dace.fit2)

> dace$Fitted.log = fitted(dace.fit2)

> ggplot(dace, aes(x=Fitted.log, y=Resid.log)) + geom\_point()

> ggplot(dace, aes(sample=Resid.log)) + stat\_qq() + stat\_qq\_line()





After log-transforming the data the diagnostics plots look better. The points in the residual vs. fitted plot are distributed much more evenly along zero and more points in the QQ-plot fall along the line.

3.c.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.758e+00 1.593e+00 -2.359 0.02155 \*

acreage 5.103e-05 1.620e-05 3.150 0.00253 \*\*

do2 3.921e-01 1.291e-01 3.036 0.00352 \*\*

maxdepth 8.997e-03 4.281e-03 2.102 0.03971 \*

no3 2.109e-01 6.970e-02 3.026 0.00363 \*\*

so4 8.863e-03 1.856e-02 0.478 0.63459

temp 8.767e-02 4.065e-02 2.157 0.03497 \*

The variable that seem to be important are acreage, do2, and no3.

3.d.

confint(dace.fit2)

2.5 % 97.5 %

(Intercept) -6.943987e+00 -5.722102e-01

acreage 1.864086e-05 8.342815e-05

do2 1.338685e-01 6.503553e-01

maxdepth 4.373384e-04 1.755645e-02

no3 7.153662e-02 3.502687e-01

so4 -2.823999e-02 4.596638e-02

temp 6.390792e-03 1.689506e-01

After accounting for all other predictors, we are 95% confident that for every additional increase in acreage there is an increase of 1.864086e-5 to 8.342815e-5 longnose dace in the 75m section of stream.

After accounting for all other predictors, we are 95% confident that for every additional increase of mg/L of dissolved oxygen, longnose dace also increases by 1.338685e-1 to 6.503553e-1 in the 75m section of stream.

After accounting for all other predictors, we are 95% confident that for every additional increase of mg/L of no3, longnose dace also increases by 7.153662e-2 to 3.502687e-1 in the 75m section of stream.

3.e.

The units of this study are the number of longnose dace in a 75m section of the stream with a predictor variable. The units for each predictor variable are: stream is Name, acreage is in acres, do2 is in mg/liter, maxdepth is in centimeters, no3 is in mg/liter, so4 is in mg/liter and temperature is in degrees Celsius. Due to all the predictor variables being natural environmental variables and unchanged by the researcher, this is an observational study.

3.f.

predict(dace.fit2, newdata = data.frame(acreage=6298, do2=9.7, maxdepth= 65, no3= 7.5, so4=44, temp=20), interval = "predict")

fit lwr upr

1 4.676764 2.245116 7.108412

The median abundance longnose dace is 4.68 fish in a 75m stream section.

4.a.

brain.fit = lm(log(Brain) ~ log(Body) + log(Gestation) + log(Litter), data=brain)

> summary(brain.fit)

Call:

lm(formula = log(Brain) ~ log(Body) + log(Gestation) + log(Litter),

data = brain)

Residuals:

Min 1Q Median 3Q Max

-0.95415 -0.29639 -0.03105 0.28111 1.57491

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.85482 0.66167 1.292 0.19962

log(Body) 0.57507 0.03259 17.647 < 2e-16 \*\*\*

log(Gestation) 0.41794 0.14078 2.969 0.00381 \*\*

log(Litter) -0.31007 0.11593 -2.675 0.00885 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4748 on 92 degrees of freedom

Multiple R-squared: 0.9537, Adjusted R-squared: 0.9522

F-statistic: 631.6 on 3 and 92 DF, p-value: < 2.2e-16

The F-statistic is 631.6 with 3 and 92 degrees of freedom. The p-value is <0.0001, the standard error is 0.4748 and the multiple R2 is 0.9537.

4.b.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.85482 0.66167 1.292 0.19962

log(Body) 0.57507 0.03259 17.647 < 2e-16 \*\*\*

log(Gestation) 0.41794 0.14078 2.969 0.00381 \*\*

log(Litter) -0.31007 0.11593 -2.675 0.00885 \*\*

All three predictor variables (Body, Gestation and Litter) are important in predicting brain size.

4.c.

We would use a prediction interval because the interval would be for a specific mammal rather than a general confidence interval for all mammals.

4.d.

predict(brain.fit, newdata = data.frame(Body=log(28.576), Gestation=log(34), Litter=log(2)), interval = "predict")

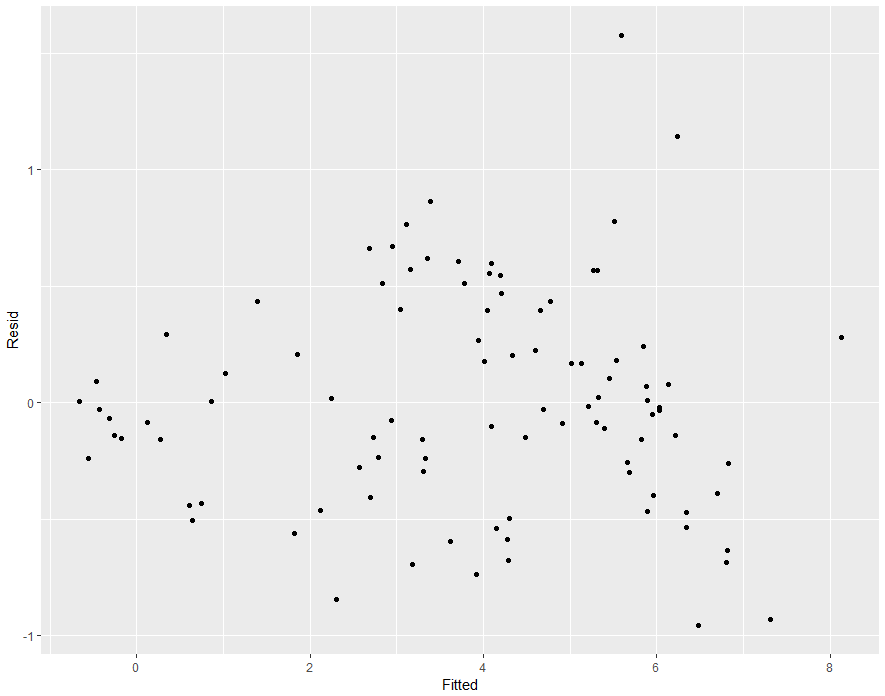
fit lwr upr

1 2.190864 0.7522792 3.62945

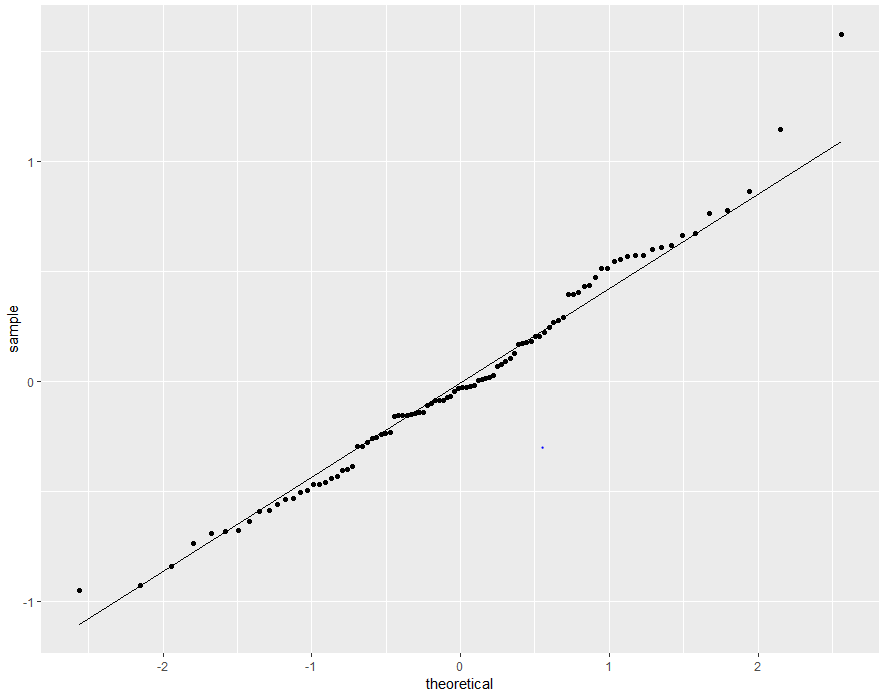
Given the information for red kangaroos, the median brain weight is 2.19 kg. After accounting for all other predictors, we can be 95% confident that the median brain weight for red kangaroos will be between 0.752 and 3.629 kg given the information provided.

4.e.

ggplot(brain, aes(x=Fitted, y=Resid)) + geom\_point()



ggplot(brain, aes(sample=Resid)) + stat\_qq() + stat\_qq\_line()



In the residuals vs. fitted plot there seems to be a disproportionate distribution along zero. The distance from zero seems to be increasing as body weight increases. The QQ-plot shows a similar pattern for larger animals. These plots suggest this model is not suitable in predicting brain weight for large animals.