Simonson Homework 9

Martin A. Simonson

April 5, 2019

1. Multiple Linear Regression Formula

Consider a multiple linear regression model with n observations and m continuous predictor variables. Recall that multiple R2 = SSModel / (SSModel + SSError) and the overall F-ratio is F = MSModel/MSError. Use this to express F in terms of R2 (and m and n).

- **Answer:** Let's start with the F-table as we know it:

Source	DF	SS	MS	F	p-value
Model	.	.	.	$F = \frac{MS_{model}}{\hat{\sigma}^2}$	.
Error	.	.	.	.	.
Total	.	.	.	.	.

$$F = \frac{MS_{model}}{\hat{\sigma}^2}$$

Where

$$MS_{model} = \frac{SS_{model}}{DF_{model}}$$

$F = \frac{R^2}{1-R^2} * \frac{n-m-1}{m}$

2. Drug Costs

Health plans use many tools to try to control the cost of prescription medicines. For older drugs, generic substitutes that are equivalent to name-brand drugs are sometimes available at a lower cost. Another tool that may lower costs is restricting the drugs that physicians may prescribe. For example, if three similar drugs are available for treating the same symptoms, a health plan may require physicians to prescribe only one of them. Since the usage of the chosen drug will be higher, the health plan may be able to negotiate a lower price for that drug.

The data in the file *drugcost.txt*, can be used to explore the effectiveness of these two strategies in controlling drug costs. The response variable is *COST*, the average cost of drugs per prescription per day, and predictors include GS (the extent to which the plan uses generic substitution, a number between zero, no substitution, and 100, always use a generic substitute if available) and RI (a measure of the restrictiveness of the plan, from zero, no restrictions on the physician, to 100, the maximum possible restrictiveness). Other variables that might impact cost were also collected, and are described in Table 1. The data are from the mid-1990s, and are for 29 plans throughout the United States with pharmacies administered by a national insurance company.

The Drug Cost Data

```
## 'data.frame':    29 obs. of  8 variables:
##  $ COST : num  1.34 1.34 1.38 1.22 1.08 1.16 1.25 1.2 1.1 1.04 ...
##  $ RXPM : num  4.2 5.4 7 7.1 3.5 7.2 10.7 7.6 7.2 6.6 ...
##  $ GS   : int  36 37 37 40 40 46 40 43 45 42 ...
##  $ RI   : num  45.6 45.6 45.6 23.6 23.6 22.3 22.3 21.3 20 20 ...
##  $ COPAY: num  10.87 8.66 8.12 5.89 6.05 ...
##  $ AGE  : num  29.7 29.7 29.7 28.7 28.7 29.1 29.1 29.8 32.4 29.8 ...
##  $ F    : num  52.3 52.3 52.3 53.4 53.4 52.2 52.2 51.6 50.8 50 ...
##  $ MM   : int  1158096 1049892 96168 407268 13224 303312 720 73380 513266 1388605 ...
```

A.

Fit a multiple linear regression model of COST on the other predictor variables. Report the overall F-ratio, p-value and the multiple R-squared.

```
##
## Call:
## lm(formula = COST ~ RXPM + GS + RI + COPAY + AGE + F + MM, data = df)
##
## 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 *
## RXPM         2.241e-02  1.100e-02   2.037 0.054483 .
## GS          -1.137e-02  2.830e-03  -4.018 0.000622 ***
## RI           3.341e-04  2.089e-03   0.160 0.874468
## 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
```

- **Answer:** The overall F-statistic is 4.072 on 7 and 21 degrees of freedom, with a p-value of 0.00572. The multiple R-squared is 0.5758.

B.

Summarize your results with regard to the importance of *GS* and *RI*. In particular, can we infer that more use of GS and RI will reduce drug costs?

- **Answer:** The coefficient for *GS* has a t-ratio of -4.018 associated with a p-value of 0.000622, providing very strong evidence that an increase in *GS* will reduce drug costs. On the other hand, *RI* has a low t-ratio of 0.160 associated with a p-value of 0.874468, providing no evidence that an increase in *RI* will affect drug cost.

C.

What are the other important variables and how do they affect the cost?

- **Answer:** There is some evidence (p-value 0.054483) for an effect of *RXPM*; with an increase in *RXPM* corresponding with an increase in drug cost. Also, there is evidence (p-value 0.020012) for an effect of *AGE*, with an increase in *AGE* corresponding with a decrease in drug cost.

D.

Find a 95% confidence interval for the coefficient of GS and interpret it in the given context.

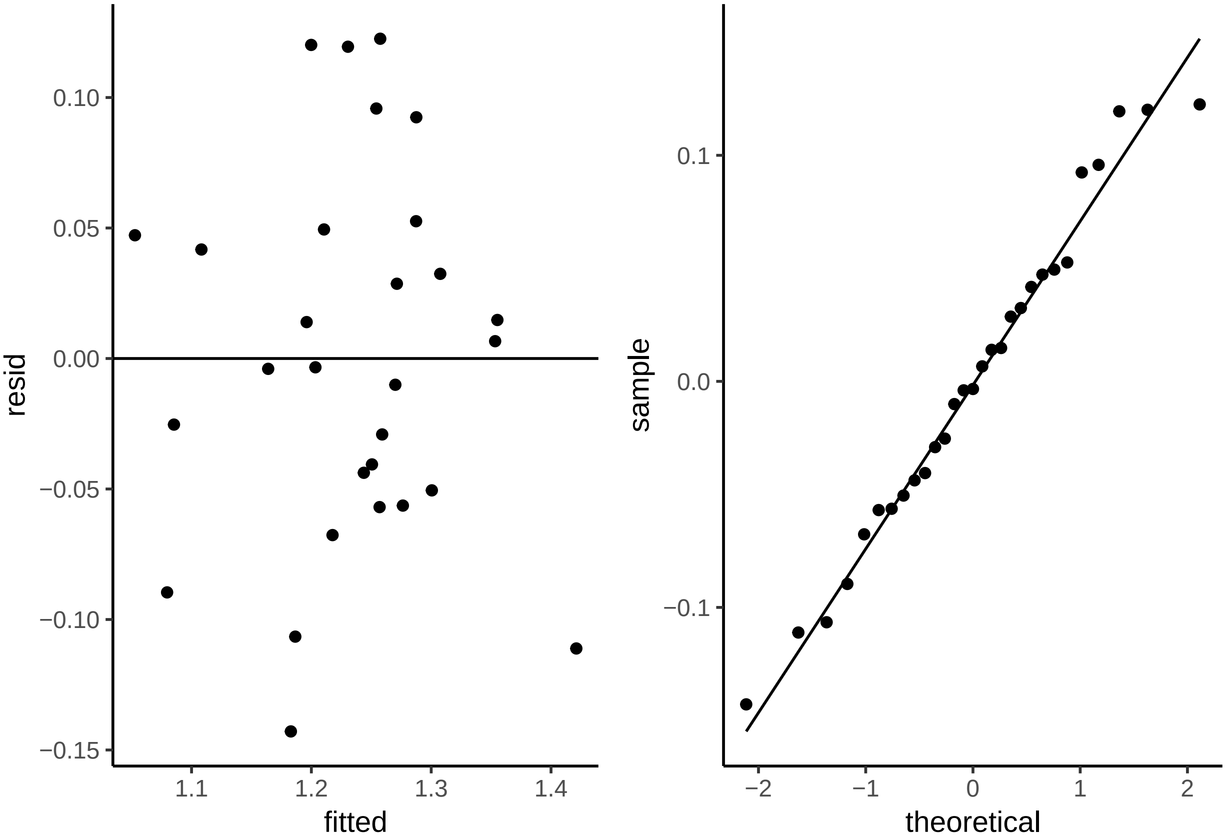
```r
t.star<-2.08 # From t-table with 21 degrees of freedom
se<-2.830 * (10^-3) # SE of GS coefficient from summary output
est<- -1.137 * (10^-2) # GS coefficient estimate from summary output

lower<- est - t.star * se
upper<- est + t.star * se

data.frame(lower,upper)
```

```
##        lower      upper
## 1 -0.0172564 -0.0054836
```

- **Answer:** We are 95% confident that, after accounting for the other variables, the coefficient for *GS* lies between -0.0172564 and -0.0054836. In other words, with all other variables held constant a one unit increase in *GS* would correspond with between 0.0172564 and 0.0054836 reduction in drug cost.

E.

Run model diagnostics and comment



- **Answer:** From both diagnostic plots we see that there may be an issue with the normality assumption (curvature of points at upper end of QQ plot) and that the homoskedasticity assumption is met due to a roughly even scatter of points on the residuals vs. fitted values plot.

3. Longnose Dace

The data in longnose_dace.txt gives the data on the abundance of longnose dace in streams in Maryland. The columns are:

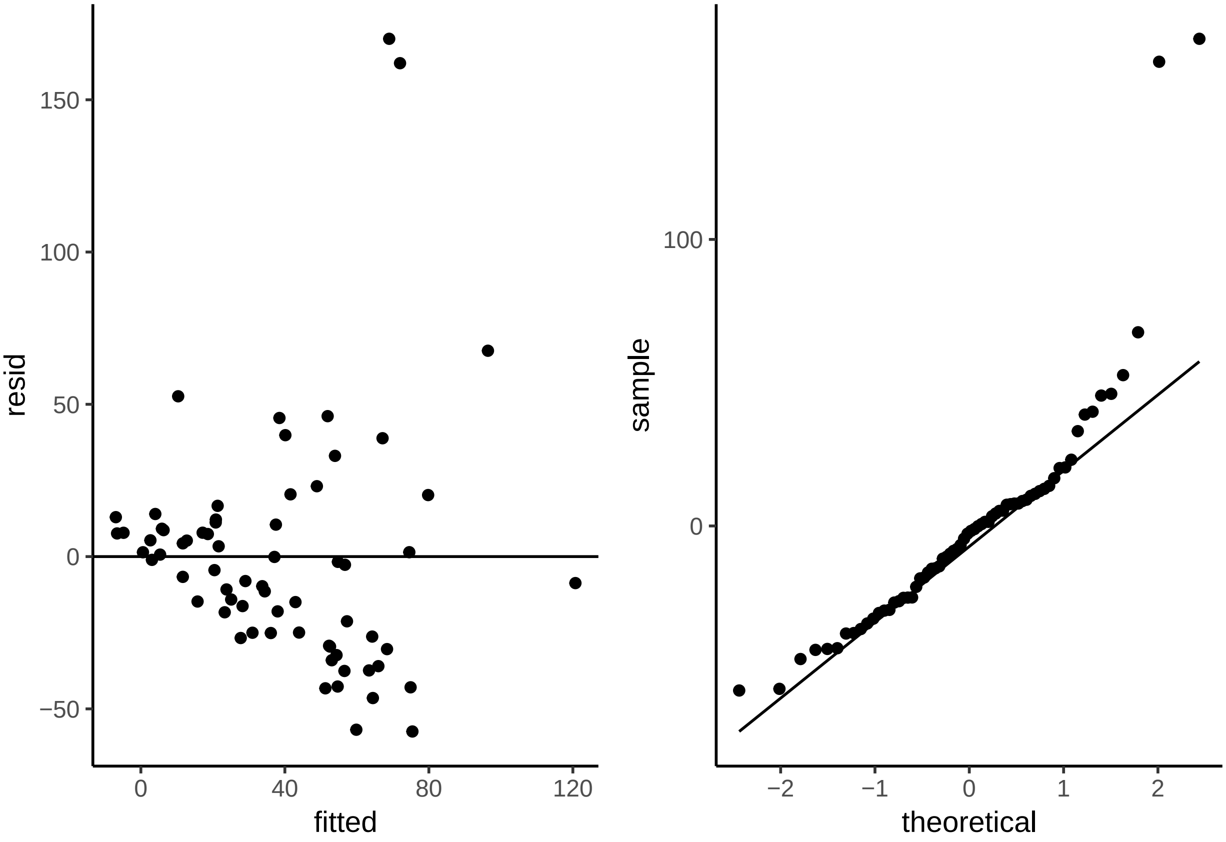
- stream : Name of the stream [ignore this column for fitting model]
- longnosedace : number of longnose in a 75m section of the stream.
- acreage : area (in acres) drained by the stream
- do2 : dissolved oxygen (in mg/litre)
- maxdepth : maximum depth (in cm) of the 75-meter segment of stream
- no3 : nitrate concentration (mg/liter)
- so4 : sulfate concentration (mg/liter)
- temp : water temperature on the sampling date (in oC).

A.

Use multiple linear model with the number of longnose dace as the response variable. Run a model diagnostics and check for model adequacy.

```
##
## Call:
```

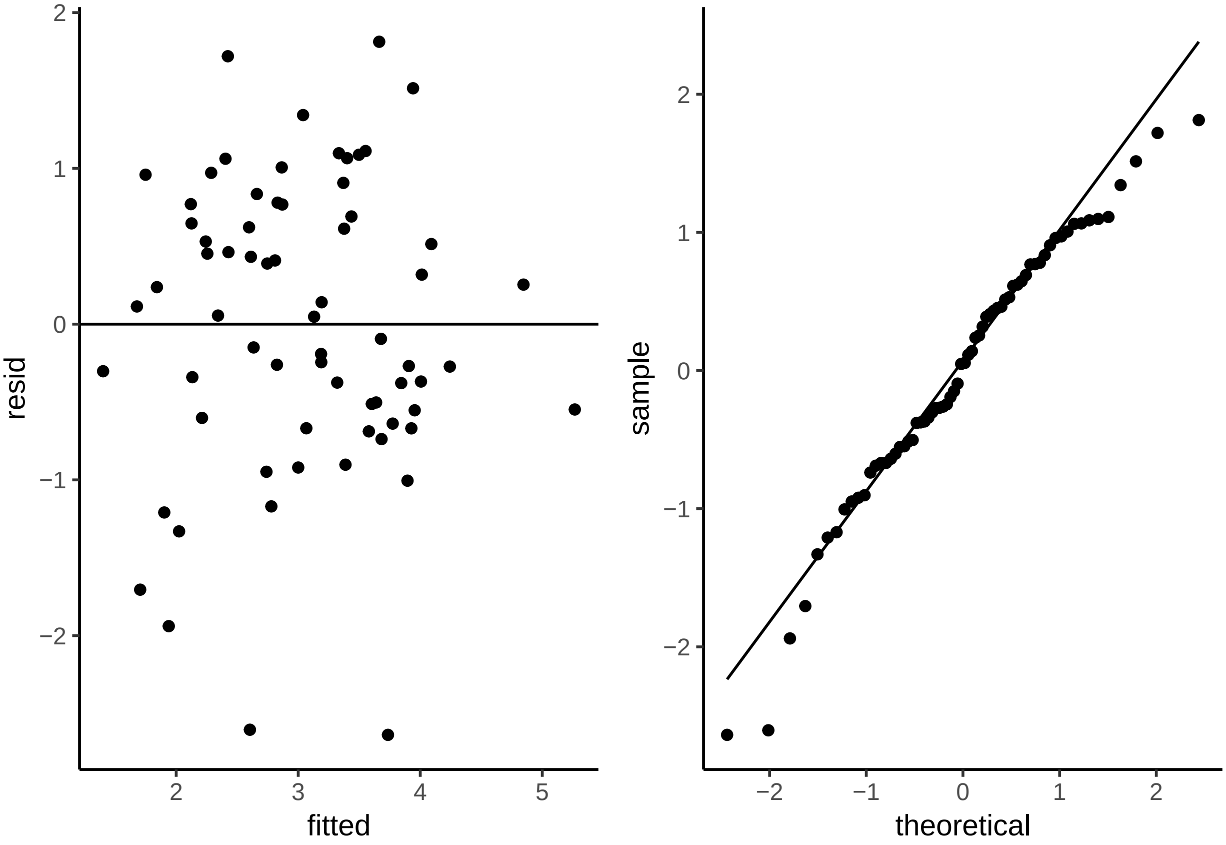
```
## lm(formula = longnosedace ~ acreage + do2 + maxdepth + no3 +
##     so4 + temp, data = dace)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.428 -25.028  -2.215  10.667 170.017
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.276e+02  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
```



- **Answer:** The first longnose dace model had an F-statistic of 4.653 on 6 and 61 degrees of freedom, with an associated p-value of 0.0005905. However, the multiple R-squared value was 0.314, showing that the model only explains about 31% of overall variability. The diagnostic plots also show violations of homoskedascity and non-normal distributions.

B.

Use an appropriate transformation [log] on the response, fit another MLR and run model diagnostics. Do things look better than A.? Irrespective of the answer, use the model in B. to answer all following questions.



- **Answer:** Log-transformation of the response variable yielded diagnostic plots that show homoskedasticity and normality are met.

C.

Check for predictive power of each variable after accounting for the rest [That is, test each regression coefficient]. Report the table of estimates, s.e.'s, T-ratios and p-values. Write a short conclusion [1-2 sentences, include direction of an effect if present].

```
summary(fit)
```

```
##
## 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
```

- **Answer:** The variables that have strongest evidence for an additive effect are agreage, do2, maxdepth, no3, and temperature. Each variable has a positive relationship with the response, log(longnosedace).

D.

Find a 95% confidence intervals for the coefficients of important predictor variables [i.e. the ones you found "significant" in C.].

```
confint(fit)
```

```
##                     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
```

- **Answer:** The only variable in this model with a 95% confidence interval for its beta estimate that overlaps 0 is so4.

E.

What are the units in the study? Is this an experimental sudy or observational?

- **Answer:** The units of this *observational* study are the number of longnose dace in a 75m section of stream.

F.

Use your model to predict the median abundance of longnose dace and a 95% interval for the abundance in a Maryland river where the conditions are as follows:

- Acreage: 6298
- do2: 9.7 mg/l
- maxdepth: 65cm

- NO3: 7.5 mg/l
- SO4: 44 mg/l
- temperature: 20°C

```
LD<-data.frame(predict(fit,
        newdata = data.frame(acreage=6928,do2=9.7,maxdepth=65,no3=7.5,so4=44,temp=20),
        interval="predict"))
exp(LD)
```

```
##        fit      lwr      upr
## 1 110.9318 9.751066 1262.003
```

- **Answer:** The predicted median abundance of longnose dace for 75 km of stream with these characteristics is 111 (rounded to whole number for individual fish).

4. Mammal Brain Weights

The data is given in *brainsize.txt*. Ignore book's questions and answer the following. Use a MLR of log(Brain) on log(Body), log(Gestation) and log(Litter) to answer the following questions:

A.

Report the overall F-test and the p-value associated with it, multiple R-squared and the estimate of error variance (i.e. $\hat{\sigma}$).

```
fit<-lm(log(Brain)~log(Body)+log(Gestation)+log(Litter), data = brain)
summary(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
```

- **Answer:** The F-statistic is 631.6 on 3 and 92 DF, with a p-value < 0.0001. $\hat{\sigma}$ is 0.4748 and the multiple R^2 is 0.9537.

B.

Report the table of estimates for the coefficients. Which predictors seem to be important in predicting brain size?

- **Answer:** [See summary table from A.] All three predictor variables have strong evidence for an additive effect on log-transformed brain size (p-values < 0.001).

C.

Suppose we want to construct a 5% interval for the average brain size of red kangaroos. Should it be a confidence interval or prediction interval?

- **Answer:** We would use a prediction interval because this estimate of the response is based on a single species rather than a group of species.

D.

Obtain a 95% interval for C. Use the following information on red kangaroos:

- Average body weight: 63 lb. [*warning: watch the units*]
- Average gestation period: 34 days
- Average litter size: 2

[Note: youll not be double penalized if you get C. wrong and if your answer in D. is consistent with C.]

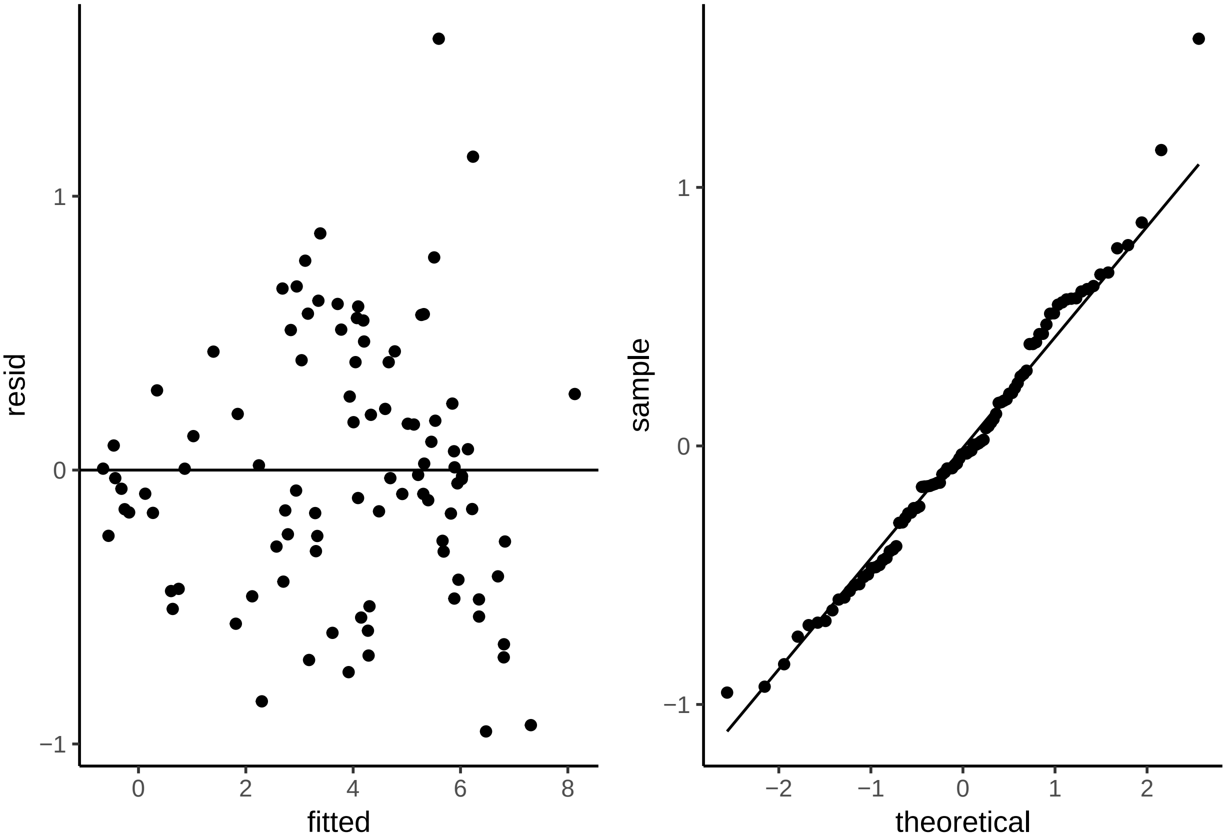
```r
BS<-data.frame(predict(fit,
                       newdata=data.frame(Body=log(63),Gestation=log(34),Litter=log(2)),
                       interval="predict"))
exp(BS)
```

```
##        fit      lwr      upr
## 1 10.10088 2.379071 42.88552
```

- **Answer:** We would expect the median brain weight to be 10.10. We are 95% confident that the median brain weight for red kangaroos will be between 2.38 and 42.89.

E.

Obtain the plot of residual-vs-fitted values. Point out a limitation of your model. [Hint: check the signs of the residuals associated with large body weights.]



- **Answer:** In the residuals vs. fitted values plot there seem to be disproportionate distributions along zero, perhaps overdispersion is occurring. There is also a pattern of non-normal data for larger animals. These plots together suggest the model is not suitable for predicting brain weights in large animals.