ℹ Cover Page

Department of Mathematical Sciences

Examination paper for ST2304 Statistical modelling for biologists and biotechologists

Examination date: 1st June 2022

Examination time (from-to): 09:00 – 13:00

Permitted examination support material: All support material is allowed

Academic contact during examination: Bob O'Hara
Phone: 915 54 416

Technical support during examination: Orakel support services
Phone: 73 59 16 00

OTHER INFORMATION

Do not open Inspera in multiple tabs, or log in on multiple devices, simultaneously. This may lead to errors in saving/submitting your answer.

Get an overview of the question set before you start answering the questions.

Read the questions carefully, make your own assumptions and specify them in your answer. Only contact academic contact if you think there are errors or insufficiencies in the question set.

Cheating/Plagiarism: The exam is an individual, independent work. Examination aids are permitted, but make sure you follow any instructions regarding citations. During the exam it is not permitted to communicate with others about the exam questions or distribute drafts for solutions. Such communication is regarded as cheating. All submitted answers will be subject to plagiarism control. *Read more about cheating and plagiarism here.*

Weighting: Weighting of the questions is given for each question.

ABOUT SUBMISSION

Answering in Inspera: If the question set contains questions that are not upload assignment, you must answer them directly in Inspera. In Inspera, your answers are saved automatically every 15 seconds.

NB! We advise against pasting content from other programs (other than as plain text), as this may cause loss of formatting and/or entire elements (e.g. images, tables).

Automatic submission: Your answer will be submitted automatically when the examination time expires and the test closes, as long as you have answered at least one question. This will happen even if you do not click “Submit and return to dashboard” on the last page of the question set. You can reopen and edit your answer as long as the test is open. If no questions are answered by the time the examination time expires, your answer will not be submitted. This is considered as “did not attend the exam”.

Withdrawing from the exam: If you become ill during the exam or wish to submit a blank answer/withdraw from the exam for another reason, go to the menu in the top right-hand corner and click “Submit blank”. This cannot be undone, even if the test is still open.

Accessing your answer post-submission: You will find your answer in Archive when the examination time has expired.

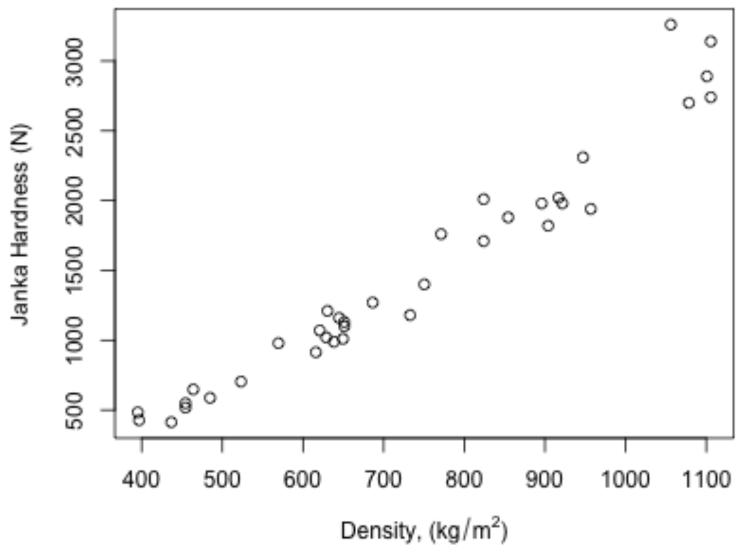
Good luck!

ℹ Hardness Description

Foresters can be interested in all sorts of things. One is how hard their wood is (you don't want to use soft wood as your floor, because it will be dented the first time you sit down on a chair). They can test this by pushing a steel ball into the wood, but would prefer something easier and less destructive. So one thing that has been looked at it wood density: if there is a strong correlated, it could be used as a proxy to predict hardness.

Here we can look at some data on hardness and density, to see what the relationship is, and how best to predict hardness from measures of density. This can be done with regression. Although the positive relationship is obvious, if we want to get good predictions we need to check that the model fits the data well.

This is the data:



The first model was a simple linear regression. It gave the following summary:

```
##
## Call:
## lm(formula = hardness ~ density, data = janka)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -338    -97    -16     93    625
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1160.50     108.58     -11    2e-12 ***
## density         3.59       0.14      25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 180 on 34 degrees of freedom
## Multiple R-squared:  0.95,   Adjusted R-squared:  0.95
## F-statistic: 6.4e+02 on 1 and 34 DF,  p-value: <2e-16
```

1 Simple linear model Slope

What percentage of the variance is explained by the model (to the nearest percent)? [ ]

What is the slope of the model, to 1 decimal place? [ ].

Maximum marks: 2

2 Simple Linear Model Prediction

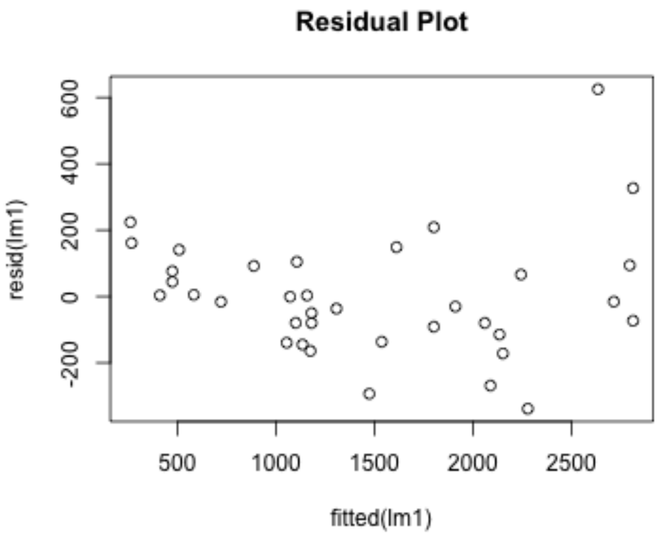
Based on this model, what is the predicted hardness of a piece of wood with a density of 1000 kg/m^2? Answer to the nearest integer.

[ ]

Maximum marks: 1

3 SLM violations

The simple model was checked to see if the assumptions of the model were reasonable



From studying this residual plot, which assumptions do you think are violated?
Select one or more alternatives:

- ☐ Linearity
- ☐ Normality
- ☐ None, it looks OK
- ☐ Outliers
- ☐ Homoscedasticity (i.e. equal variance)

Maximum marks: 2

4 SLM: What to do

Based on this plot, what would you do next?
(this may not be what was actually done, and what comes next may not be the best thing to do)

Fill in your answer here

Format

Words: 0

Maximum marks: 4

It was decided to try a square root transformation of hardness (the response). Fitting the model gave the following summary

```
##
## Call:
## lm(formula = sqrt(hardness) ~ density, data = janka)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -3.5   -1.1   -0.3    1.0    4.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.259      1.094       2     0.05 *
## density        0.047      0.001      33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2 on 34 degrees of freedom
## Multiple R-squared:   1, Adjusted R-squared:   1
## F-statistic: 1e+03 on 1 and 34 DF,  p-value: <2e-16
```

5 Sqrt Prediction

Based on this model, what is the predicted hardness of a piece of wood with a density of 1000 kg/m^2?

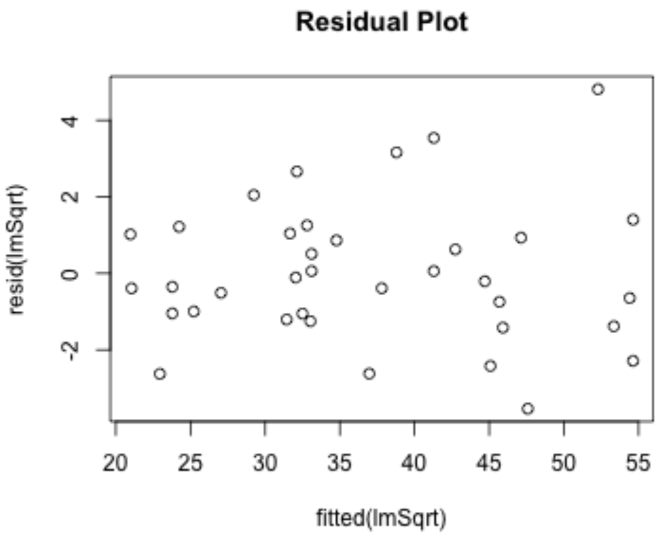
Answer to the nearest integer

[ ] .

Maximum marks: 1

6 Sqrt violations

The model was checked to see if the assumptions of the model were reasonable



From studying this residual plot, which assumptions do you think are violated?
Select one or more alternatives:

- ☐ It looks OK
- ☐ Linearity
- ☐ Outliers
- ☐ Normality
- ☐ Homoscedasticity (i.e. equal variance)

Maximum marks: 2

Next, a log transformation of hardness (the response) was tried. This gave the following summary

```
##
## Call:
## lm(formula = log(hardness) ~ density, data = janka)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.336 -0.087  0.002  0.085  0.233
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.20858    0.08050      65   <2e-16 ***
## density      0.00263    0.00011      25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.14 on 34 degrees of freedom
## Multiple R-squared:  0.95,   Adjusted R-squared:  0.95
## F-statistic: 6.2e+02 on 1 and 34 DF,  p-value: <2e-16
```

7 Log transformation Prediction

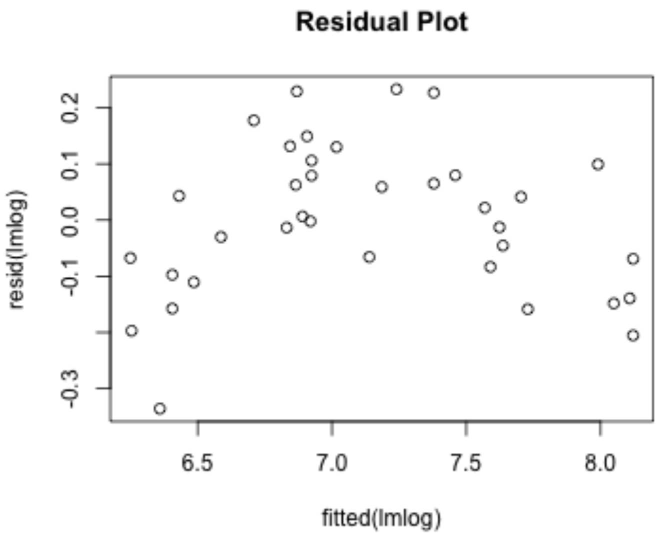
Based on this transformed model, what is the predicted hardness of a piece of wood with a density of 1000 kg/m^2? Answer to the nearest integer.

(note that a natural log transformation was used)

Maximum marks: 1

8 log violations

The model was checked to see if the assumptions of the model were reasonable



From studying this residual plot, which assumptions do you think are violated?
Select one or more alternatives:

- ☐ Homoscedasticity (i.e. equal variance)
- ☐ Normality
- ☐ It looks OK
- ☐ Outliers
- ☐ Linearity

Maximum marks: 2

At the suggestion of another statistician, a generalised linear model was used. This assumed a Gamma distribution, and had a square root link function.

We do not need the details of the gamma distribution: it has to be positive, and the variance increases with the mean. The inverse of the square root link function is the square.

Fitting this model gave the following summary:

```
lmGamma <- glm(hardness ~ density, family=Gamma("sqrt"), data=janka)
print(summary(lmGamma), digits=1)
```

```
##
## Call:
## glm(formula = hardness ~ density, family = Gamma("sqrt"), data = janka)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -0.22   -0.07   -0.02    0.05    0.17
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.867      0.904       2     0.05 *
## density        0.048      0.001      34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.009716208)
##
##     Null deviance: 11.26788  on 35  degrees of freedom
## Residual deviance:  0.32876  on 34  degrees of freedom
## AIC: 453
##
## Number of Fisher Scoring iterations: 4
```

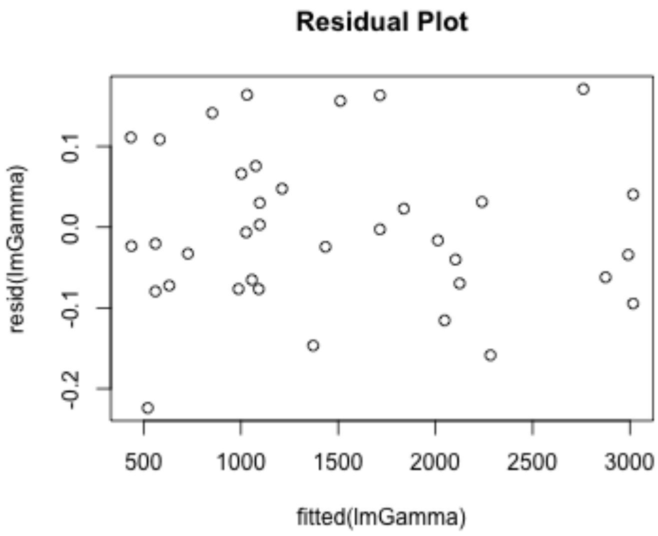
9 Janka Gamma Model Prediction

Based on this model, what is the predicted hardness of a piece of wood with a density of 1000 kg/m^2? Answer to the nearest integer

Maximum marks: 1

10 Janka Gamma violations

The simple model was checked to see if the assumptions of the model were reasonable



From studying this residual plot, which assumptions do you think are violated?
Select one or more alternatives:

- Outliers
- Heteroscedasticity
- It looks OK
- Normality
- Linearity

Maximum marks: 2

Now we have tried several models, we want to compare them, and decide how well they describe the data.

These are the estimated coefficients:

Model	Intercept	Slope
Simple	-1160.5	3.6
Square Root transformation	2.26	0.047
Log transformation	5.21	0.0026
Gamma	1.87	0.048

11 **Why not R2?**

Why can we not compare the R^2 for this model to the value for the untransformed model?

Fill in your answer here

Format

Words: 0

Maximum marks: 4

12 Compare Predictions

How different are the model predictions for a sample of hardness 1000 kg/m^3? (i.e. compare the predictions you have made).

Fill in your answer here

Format

Words: 0

Maximum marks: 4

13 Compare Intercepts

How similar/different are the intercepts? (note: you might have to transform them). How relevant are these for comparing the models, and how well they describe the data?

Fill in your answer here

Format

Σ

Words: 0

Maximum marks: 6

14 Assessment of the Models

On the basis of these analyses explain which model you would prefer, and why. In particular:

- which model looks best?
- how much difference do the different models make?
- are there any other analyses or calculations you would like to have seen?

Fill in your answer here

Format

Words: 0

Maximum marks: 10

You have been hired as a consultant by Project Ekorn to assess their performance. Project Ekorn have been trying to stop a mysterious global group of squirrels called Cyber Squirrel Operations from taking out electric power systems. You have been asked to look at how successful these counter-operations have been. Obviously there are other reasons for power systems to go down (birds, storms, octopus etc.), so you want to look to see if the proportion of attacks by squirrels has changed, and also if this is effective in reducing the numbers of people affected.

You have collected data from around the world on the following variables:
- whether the outage was caused by a squirrel
- the year of the event
- the region (some countries have been combined into a larger region)
- the number of people affected, and
- the duration of the event

We expect the country to have an effect (as some countries do not have many squirrels), but are interested in the other factors.

First we will look at the binary response of whether an attack was caused by a squirrel or not.

(note: this is based on real data. The explanation has been exaggerated a lot)

15 Choose Squirrel Models

If we want to test whether the proportion of squirrel attacks has changed over time (=Year), what models would we compare?

Null model

Select one alternative

- Year + Country
- Country
- Year*Country
- Year

Alternative Model

Select one alternative

- Country
- Year*Country
- Year + Country
- Year

Maximum marks: 2

16 Squirrel Model Comparison type

What approach is being taken to the model selection in this problem (i.e. asking if the proportion of squirrel attacks has changed over time)?

Select one alternative:

- Exploratory, to get a model that predicts the data well
- Comfirmatory, to test a hypothesis
- Exploratory, to get a model that explains the data well

Maximum marks: 1

17 Test Results

When the relevant test was done, we got the following result:

```
## Model 1: REDACTED
## Model 2: REDACTED
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      2524     3134.3
## 2      2523     3104.7  1   29.601 5.307e-08 ***
```

Does this suggest an effect of year?

Select one alternative

- ○ Can't tell
- ○ Yes
- ○ No

What statistics tell you this?

Select one or more alternatives

- ☐ Resid. Df
- ☐ Resid. Dev
- ☐ Df
- ☐ Deviance
- ☐ Pr(>Chi)

Why is this not enough information to assess the effect of year, and what more would you want to be told?

Maximum marks: 2

18 Need More Information

When the relevant test was done, we got the following result:

```
## Model 1: REDACTED
## Model 2: REDACTED
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      2524     3134.3
## 2      2523     3104.7  1   29.601 5.307e-08 ***
```

Why is this not enough information to assess the effect of year on the probability that an attack was by a squirrel, and what more would you want to be told?

Fill in your answer here

Format

Words: 0

Maximum marks: 4

In order to try to get a good model, the following model was fitted:

```
sq1 <- glm(Squirrel ~ Region + YearS + lnAffected + lnDuration,
           data=Squirrels, family = binomial())
```

where YearS is Year - 2000 (so the intercept is at 2000 AD), and lnAffected and lnDuration are the natural logs of the number of people affected and the length of time the power was out (in hours). The reference level is the USA, so the other country effects are contrasts to that. The response, Squirrel, is a binary response: it is 1 if the attack was by a squirrel, 0 if it was not.

This gave the following summary:

```
##
## Call:
## glm(formula = Squirrel ~ Region + YearS + lnAffected + lnDuration,
##     family = binomial(), data = Squirrels)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.29739  -1.19724  -0.00052   1.09060   2.60735
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.501505   0.366771   4.094 4.24e-05 ***
## RegionAsia           -4.017684   1.015025  -3.958 7.55e-05 ***
## RegionCanada         -0.941575   0.344321  -2.735 0.006246 **
## RegionEurope         -1.760425   0.511984  -3.438 0.000585 ***
## RegionSouth America -15.785693 589.340327  -0.027 0.978631
## RegionUK             -1.439707   0.700209  -2.056 0.039772 *
## YearS                -0.086268   0.024656  -3.499 0.000467 ***
## lnAffected           -0.220496   0.077955  -2.829 0.004677 **
## lnDuration            0.009776   0.078283   0.125 0.900613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1089.62  on 785  degrees of freedom
## Residual deviance:  973.11  on 777  degrees of freedom
##   (1791 observations deleted due to missingness)
## AIC: 991.11
##
## Number of Fisher Scoring iterations: 14
```

19 link function

Which link function was used here?
Select one alternative:

- identity
- logit
- cloglog
- square root
- log
- probit

Maximum marks: 1

20 Why year - 2000?

Why was (Year - 2000) used in the model, rather than Year? How does it affect the other parameter estimates?

Fill in your answer here

Format B I U x_2 x^2 I_x Ω Σ

Words: 0

Maximum marks: 6

21 Squirrel Effects Interpretation

Based on this summary, has there been a change in the proportion of squirrel attacks, and if so how much of an effect? Does this mean that actions against squirrels have been effective?

Are there any other interesting effects? (ignore South America for this question: it is doing something weird and obscure)

Fill in your answer here

Format

Words: 0

Maximum marks: 10

In order to try to get a good model, the following model was fitted:

```
sq1 <- glm(Squirrel ~ Region + YearS + lnAffected + lnDuration,
           data=Squirrels, family = binomial())
```

where YearS is Year - 2000 (so the intercept is at 2000 AD), and lnAffected and lnDuration are the natural logs of the number of people affected and the length of time the power was out (in hours). The reference level is the USA, so the other country effects are contrasts to that.

This gave the following summary:

```
##
## Call:
## glm(formula = Squirrel ~ Region + YearS + lnAffected + lnDuration,
##     family = binomial(), data = Squirrels)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##     -2      -1       0       1       3
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)             1.50       0.37     4.1    4e-05 ***
## RegionAsia             -4.02       1.01    -4.0    8e-05 ***
## RegionCanada           -0.94       0.34    -2.7    0.006 **
## RegionEurope           -1.76       0.51    -3.4    6e-04 ***
## RegionSouth America   -15.79     589.34     0.0    0.979
## RegionUK               -1.44       0.70    -2.1    0.040 *
## YearS                  -0.09       0.03    -3.5    5e-04 ***
## lnAffected             -0.22       0.08    -2.8    0.005 **
## lnDuration              0.01       0.08     0.1    0.901
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1089.62  on 785  degrees of freedom
## Residual deviance:  973.11  on 777  degrees of freedom
##   (1791 observations deleted due to missingness)
## AIC: 991.1
##
## Number of Fisher Scoring iterations: 14
```

22 Europe Prediction

What is the probability that a power outage was caused by a squirrel under the following conditions?

- **Region**: Europe
- **Year**: 2020 (note that we use 2000 as the intercept, i.e. YearS = year - 2000)
- **lnAffected**: 1
- **lnDuration**: 0

The log odds (i.e. the value on the linear scale):

The probability:

Maximum marks: 2

As well as trying to find out if the proportion of squirrel attacks have changed, you have also been asked to look at the effects of attacks by different groups, in particular on the numbers that are affected (if squirrels only affect a few people, but birds affect many more, then perhaps we should be more worried bird effects).

So, we looked at several factors to try to understand what influenced the numbers of people affected by each attack. Several models were tried, and their fits to the data summarised in the following table:

Model	AIC	BIC	R^2 (%)
lnDuration	113.0	118.9	1.0
YearS + lnDuration	92.9	104.6	1.9
Countries + lnDuration	73.2	114.2	3.0
Countries + YearS + lnDuration	49.4	96.3	4.0
Operative + lnDuration	61.0	166.4	4.3
Operative + YearS + lnDuration	35.4	146.7	5.3
Operative + Countries + lnDuration	15.0	155.5	6.4
Operative + Countries + YearS + lnDuration	-11.2	135.2	7.4

23 Which Statistic

Which statistic would you use to decide on the best model to explain and understand what is influencing the number of people affected?

Select one alternative:

- Something else
- BIC
- R^2
- AIC

Maximum marks: 2

24 Which model is best?

Which model do you think is best, according to the criterion you chose?
Select one alternative:

- Countries + YearS + lnDuration
- lnDuration
- Operative + lnDuration
- Countries + lnDuration
- YearS + lnDuration
- Operative + Countries + YearS + lnDuration
- Operative + YearS + lnDuration
- Operative + Countries + lnDuration

Maximum marks: 2

This is the summary of the full model (which may not be the best one). The reference levels are squirrels (for Operative) and USA (for Region).

```
##
## Call:
## lm(formula = lnAffected ~ Operative + Region + YearS + lnDuration,
##     data = Squirrels)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -4.2   -0.3    0.2    0.6    2.3
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.28       0.14     1.9    0.057 .
## OperativeAnimal             -0.34       0.28    -1.2    0.219
## OperativeBat                -0.26       0.74    -0.4    0.724
## OperativeBeaver              0.43       0.36     1.2    0.230
## OperativeBird                0.14       0.09     1.6    0.116
## OperativeEagle               0.64       0.59     1.1    0.283
## OperativeMarten              1.49       0.52     2.9    0.004 **
## OperativeMonkey              0.92       1.03     0.9    0.371
## OperativeOther Mammal        0.10       0.31     0.3    0.735
## OperativeOther Vertebrate    0.48       0.17     2.8    0.006 **
## OperativePossum             -0.34       0.40    -0.8    0.402
## OperativeRaccoon             0.41       0.16     2.5    0.013 *
## OperativeRat                -0.11       0.37    -0.3    0.767
## OperativeUnknown             0.24       0.16     1.5    0.131
## RegionAsia                   0.42       0.18     2.3    0.020 *
## RegionCanada                 0.26       0.16     1.6    0.109
## RegionEurope                -0.23       0.24    -1.0    0.322
## RegionSouth America          1.02       0.44     2.3    0.019 *
## RegionUK                    -0.58       0.31    -1.8    0.067 .
## YearS                       -0.03       0.01    -2.9    0.004 **
## lnDuration                   0.10       0.04     2.6    0.009 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1 on 765 degrees of freedom
##   (1791 observations deleted due to missingness)
## Multiple R-squared:  0.07,   Adjusted R-squared:  0.05
## F-statistic:  3 on 20 and 765 DF,  p-value: 8e-06
```

25 Affected Assessment

Based on these results, what are the important effects on the variation in the number of people being affected? Consider both the model you chose, the parameter estimates and other summaries of the model.

Fill in your answer here

Format

Words: 0

Maximum marks: 10