Module 2 Assessment

SOLUTIONS

Logistics

- Due Feb 21 at 11:59pm
- Please submit both a pdf and .Rmd file to canvas

Salaries data

We will be using the "Salaries" dataset from "An R Companion to Applied Regression" by Fox and Weisberg (2019). It can be loaded into `R` using the `carData` package as shown below. The dataset has the 2008-2009 salaries from 397 faculty members at a college in the US.

```
# Install the package carData if you don't already have it using the code below
# install.packages("carData")
```

```
# load the package
library("carData")

# load the dataset. Running this should result in a variable named Salaries
# being created. Note that R is case sensitive
data(Salaries)

head(Salaries)
```

```
##        rank discipline yrs.since.phd yrs.service  sex salary
## 1      Prof          B            19          18 Male 139750
## 2      Prof          B            20          16 Male 173200
## 3  AsstProf          B             4           3 Male  79750
## 4      Prof          B            45          39 Male 115000
## 5      Prof          B            40          41 Male 141500
## 6 AssocProf          B             6           6 Male  97000
```

```
# The ? command opens the documentation for the dataset
?Salaries
```

In the following questions, fit the model which most directly answers the question of interest; i.e., don't take transformations or add additional covariates unless they are required. In many cases, the model you fit will clearly violate some of the assumptions we've discussed in class. For now, just ignore those violations and simply report the estimates.

Question 1 (2pts)

Suppose I am interested in knowing the expected difference in salary between two individuals who differ by 1 year in years since phd. Using the `lm` function, write out the code for the model that directly estimates that quantity. Write one sentence to interpret your finding to a collaborator.

Answer

```
mod1 <- lm(salary ~ yrs.since.phd, data=Salaries)
summary(mod1)
```

```
##
## Call:
## lm(formula = salary ~ yrs.since.phd, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -84171 -19432  -2858  16086 102383
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    91718.7     2765.8  33.162   <2e-16 ***
## yrs.since.phd    985.3      107.4   9.177   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27530 on 395 degrees of freedom
## Multiple R-squared:  0.1758, Adjusted R-squared:  0.1737
## F-statistic: 84.23 on 1 and 395 DF,  p-value: < 2.2e-16
```

Based on a simple linear model, if we compare two individuals who differ in 1 year since PhD, we would expect the individual with 1 more year since PhD to have a salary that is 985.34 higher.

Question 2 (2pts)

Suppose I am interested in knowing the expected difference in salary between two individuals who differ by 1 year in years since phd, but are the same rank and discipline. Using the `lm` function, write out the code for the model that directly estimates that quantity. Write one sentence to interpret your finding to a collaborator.

Answer

```
mod2 <- lm(salary ~ yrs.since.phd+rank+discipline, data=Salaries)
summary(mod2)
```

```
##
## Call:
## lm(formula = salary ~ yrs.since.phd + rank + discipline, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67395 -13480  -1536  10416  97166
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   71405.40    3278.32  21.781  < 2e-16 ***
## yrs.since.phd    71.92     126.68   0.568   0.5706
## rankAssocProf 13030.16    4168.17   3.126   0.0019 **
## rankProf      46211.57    4238.52  10.903  < 2e-16 ***
## disciplineB   14028.68    2345.90   5.980 5.03e-09 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22670 on 392 degrees of freedom
## Multiple R-squared:  0.4454, Adjusted R-squared:  0.4398
## F-statistic: 78.72 on 4 and 392 DF,  p-value: < 2.2e-16
```

If we compare two individuals who differ in 1 year since PhD, we would expect the individual with 1 more year since Phd to have a salary that is 71.92 higher, provided discipline and rank are the same between the two individuals.

Question 3 (2pts)

Suppose I am interested in knowing the % difference in expected salary between two individuals who differ by 1 year in years since phd. Using the `lm` function, write out the code for the model that directly estimates that quantity. Write one sentence to interpret your finding to a collaborator.

Answer

Since we are looking for salary difference in percentages, we take the log transform of salary.

```
mod3 <- lm(log(salary) ~ yrs.since.phd, data=Salaries)
summary(mod3)
```

```
##
## Call:
## lm(formula = log(salary) ~ yrs.since.phd, data = Salaries)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88900 -0.16833  0.00347  0.16163  0.61047
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.142e+01  2.368e-02 482.055   <2e-16 ***
## yrs.since.phd 8.591e-03  9.193e-04   9.345   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2357 on 395 degrees of freedom
## Multiple R-squared:  0.1811, Adjusted R-squared:  0.179
## F-statistic: 87.33 on 1 and 395 DF,  p-value: < 2.2e-16
```

If we compare two individuals who differ in 1 year since PhD, we would expect the individual with 1 more year since Phd to have a salary that is

$$100(e^{8.591e-03} - 1) = 0.863\%$$

higher

Question 4 (2pts)

Suppose I am interested in knowing the % difference in expected salary between two individuals who differ by 1 year in years since phd, but are in the same discipline. Using the `lm` function, write out the code for the model that directly estimates that quantity. Write one sentence to interpret your finding to a collaborator.

Answer

```
mod4 <- lm(log(salary) ~ yrs.since.phd + discipline, data=Salaries)
summary(mod4)
```

```
##
## Call:
## lm(formula = log(salary) ~ yrs.since.phd + discipline, data = Salaries)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84485 -0.14962 -0.01235  0.14973  0.63243
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.308533   0.028292 399.712  < 2e-16 ***
## yrs.since.phd  0.009825   0.000899  10.928  < 2e-16 ***
## disciplineB    0.146202   0.023233   6.293 8.31e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.225 on 394 degrees of freedom
## Multiple R-squared:  0.2559, Adjusted R-squared:  0.2521
## F-statistic: 67.73 on 2 and 394 DF,  p-value: < 2.2e-16
```

If we compare two individuals who differ in 1 year since PhD, but are in the same discipline, we would expect the individual with 1 more year since Phd to have a salary that is

$$100(e^{0.009825} - 1) = 0.987\%$$

higher

Question 5 (2pts)

Write a short explanation which interprets the results of model below to a collaborator.

```
summary(lm(salary ~ relevel(rank, ref = "AssocProf"), data = Salaries))
```

```
##
## Call:
## lm(formula = salary ~ relevel(rank, ref = "AssocProf"), data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -68972 -16376  -1580  11755 104773
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               93876       2954  31.777  < 2e-16
## relevel(rank, ref = "AssocProf")AsstProf  -13100       4131  -3.171  0.00164
## relevel(rank, ref = "AssocProf")Prof       32896       3290   9.997  < 2e-16
##
## (Intercept)                               ***
## relevel(rank, ref = "AssocProf")AsstProf **
## relevel(rank, ref = "AssocProf")Prof      ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23630 on 394 degrees of freedom
## Multiple R-squared:  0.3943, Adjusted R-squared:  0.3912
## F-statistic: 128.2 on 2 and 394 DF,  p-value: < 2.2e-16
```

The expected salary for an associate professor is $93,876. The expected salary for an assistant professor is $13,100 lower than an associate professor. The expected salary for a full professor is $32,896 higher than an associate professor.

Answer

Question 6 (3pts)

I think the association of years since phd and salary is different in discipline A when compared to discipline B. What is the estimate of the expected difference in salary between two individuals who differ by 1 year in years since phd and are both in discipline A? What is the estimate of the expected difference in salary between two individuals who differ by 1 year in years since phd and are both in discipline B?

Answer

Since we expect the relationship between salary and years since PhD will be different between the two disciplines, we will fit the model using an interaction between the two covariates:

```r
mod7 <- lm(salary ~ yrs.since.phd * discipline, data=Salaries)
summary(mod7)
```

```
##
## Call:
## lm(formula = salary ~ yrs.since.phd * discipline, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -84580 -16974  -3620  15733  92072
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 84845.4     4283.9  19.806  < 2e-16 ***
## yrs.since.phd                 933.9      150.0   6.225 1.24e-09 ***
## disciplineB                  7530.0     5492.2   1.371   0.1711
## yrs.since.phd:disciplineB     365.3      211.0   1.731   0.0842 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26400 on 393 degrees of freedom
## Multiple R-squared:  0.2458, Adjusted R-squared:  0.2401
## F-statistic:  42.7 on 3 and 393 DF,  p-value: < 2.2e-16
```

The estimate of the expected difference in salary between two individuals who differ by 1 year in years since phd and are both in discipline A is 933.8790005. The estimate of the expected difference in salary between two individuals who differ by 1 year in years since phd and are both in discipline B is 933.88 + 365.32 = 1299.2.

Question 7 (8pts)

The following data records data across 55 mammals. It is available in the R package `mice`, but I've taken out some species with missing data and only have a subset of the original variables. In particular, we have:

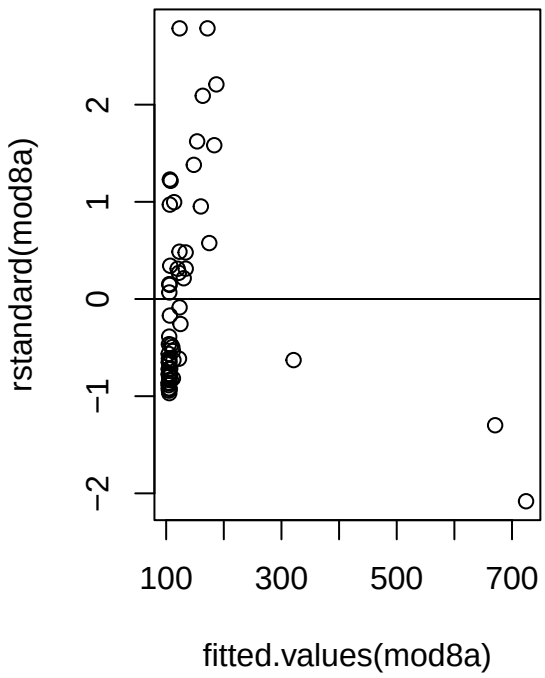
- species: Species of animal
- gt: gestational time
- bw: body weight (kg)
- brw: brain weight (g)

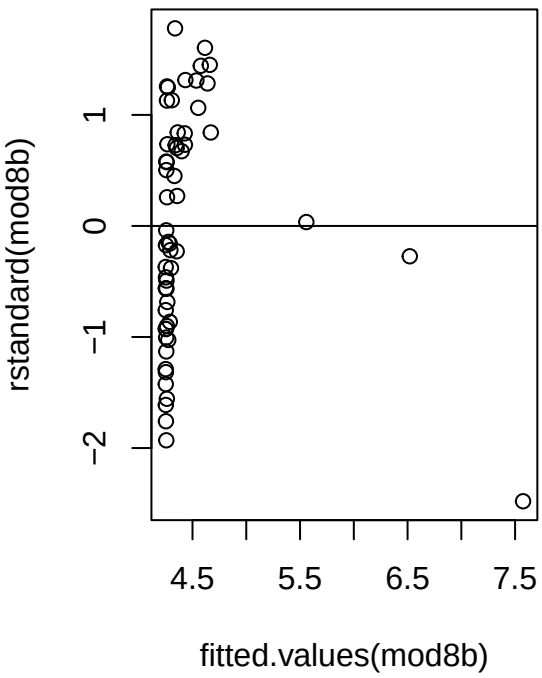
Suppose we want to fit a model which explores the relationship between the dependent variable gestational time (gt) and the covariates body weight (bw) and brain weight (brw). Choose an appropriate model to fit by considering potential transformations of the data (it doesn't have to be perfect, but it shouldn't be terrible either). Using various plots, explain why the model you choose is appropriate and discuss whether the linear model assumptions hold. For simplicity, don't consider models with interactions.

```
fileName <- "https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lab3/mammal_data.csv"
mammals <- read.csv(fileName)
```

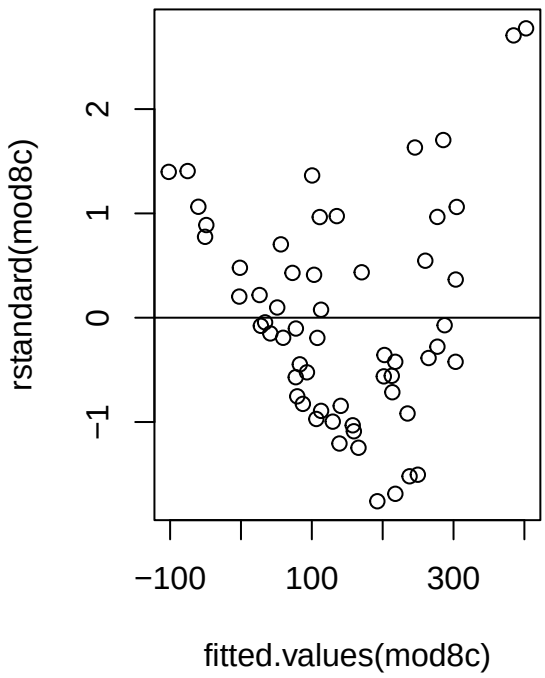
Answer

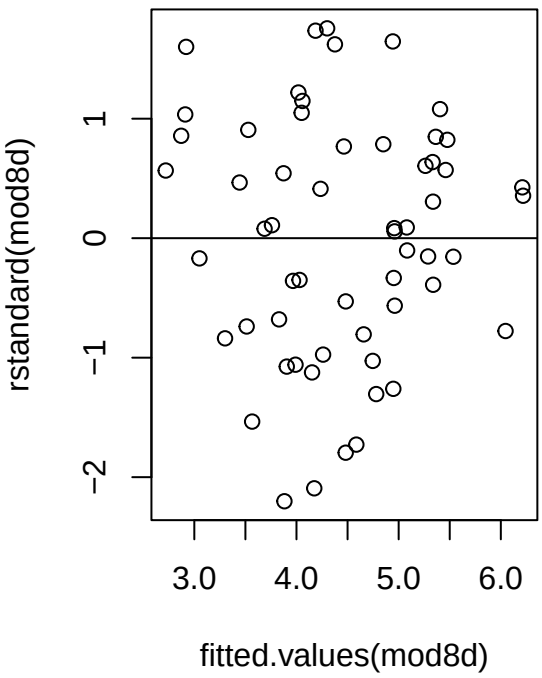
```
#Fit models
mod8a <- lm(gt~bw +brw, data = mammals) #No transformation
mod8b <- lm(log(gt)~bw +brw, data = mammals) # Response transformed
mod8c <- lm(gt ~ log(bw) + log(brw),data = mammals) # covariates transformed
mod8d <- lm(log(gt)~log(bw) + log(brw),data = mammals) # Both response and covariates transformed

# Residual plots
par(mfrow = c(1,2))
plot(fitted.values(mod8a),rstandard(mod8a))
abline(h=0)
plot(fitted.values(mod8b),rstandard(mod8b))
abline(h=0)
```



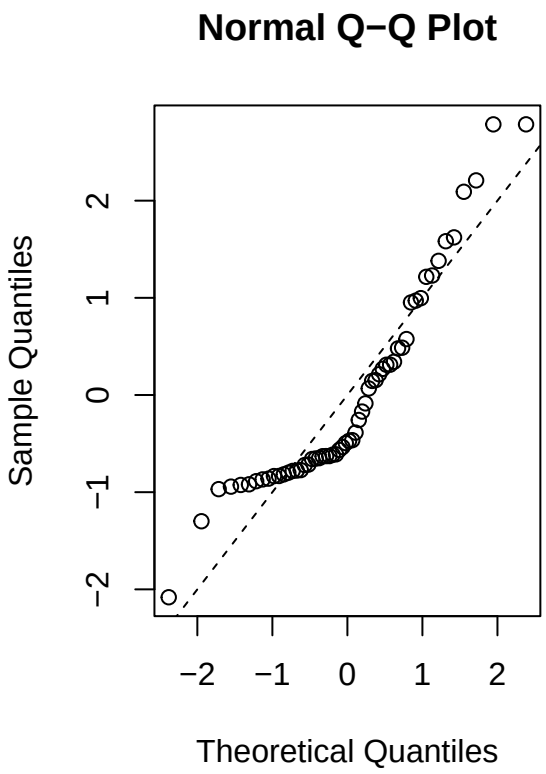


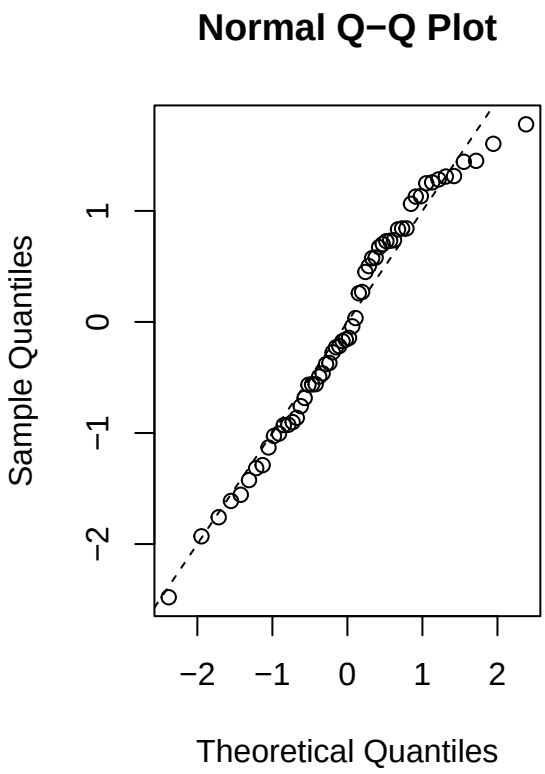

```
plot(fitted.values(mod8c),rstandard(mod8c))
abline(h =0)
plot(fitted.values(mod8d),rstandard(mod8d))
abline(h=0)
```



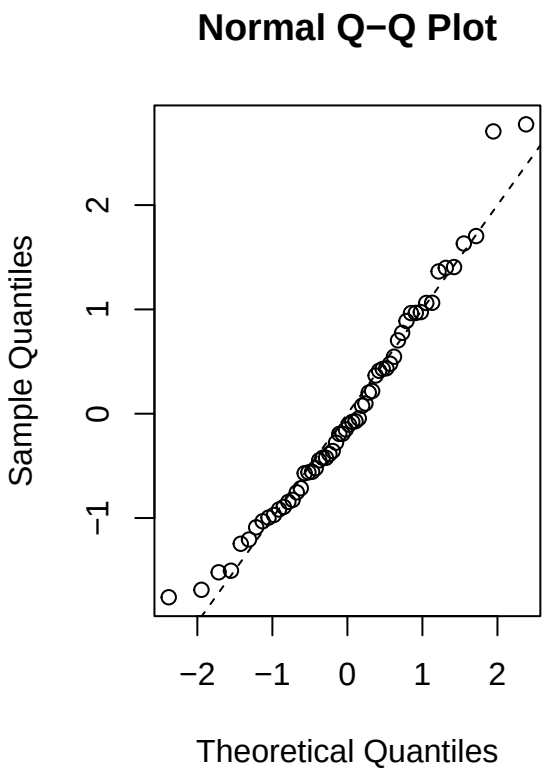


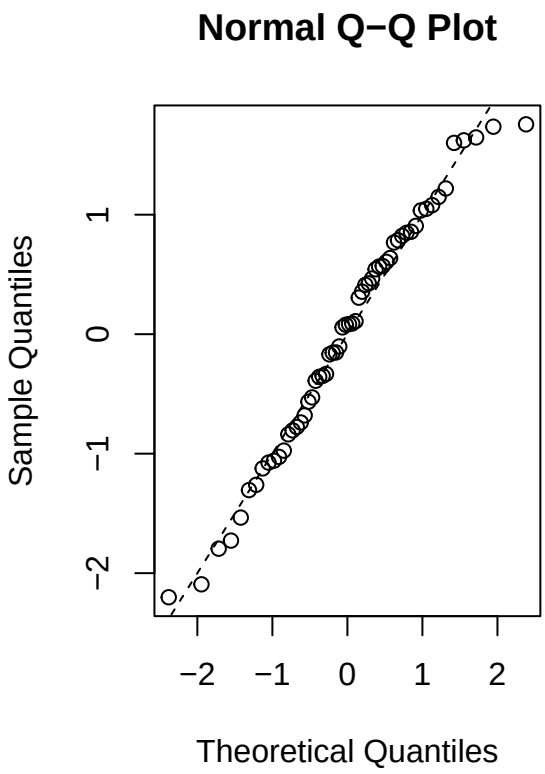

```
# Normal Q-Q plot
qqnorm(rstandard(mod8a))
abline(0,1,lty=2)
qqnorm(rstandard(mod8b))
abline(0,1,lty =2)
```






```
qqnorm(rstandard(mod8c))
abline(0,1,lty=2)
qqnorm(rstandard(mod8d))
abline(0,1,lty=2)
```





Based on the above diagnostic plots I would choose the fourth model where we perform a log transformation on the response variable and both covariates. The plot of the standardized residuals with the fitted values for model 4 shows no clear patterns. Only two points are outside of [-2,2] which is fine as we expect around 5% of the observations to be outside of this range by chance. There are no clear signs of changing variance. There is a reasonable fit to a straight line in the Normal Q-Q plot which indicates the normality assumption is fine.

Question 8 (4pts)

Give an interpretation of the estimated coefficients for covariates corresponding to body weight and brain weight.

Answer

We will use the model with both the explanatory covariates and the response variable transformed:

```
summary(mod8d)
```

```
##
## Call:
## lm(formula = log(gt) ~ log(bw) + log(brw), data = mammals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.39694 -0.46559  0.05181  0.49266  1.11590
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  3.05429    0.28201  10.830 2.97e-15 ***
## log(bw)     -0.11623    0.09705  -1.198 0.236182
## log(brw)     0.48313    0.12365   3.907 0.000258 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6467 on 55 degrees of freedom
## Multiple R-squared:  0.6411, Adjusted R-squared:  0.628
## F-statistic: 49.12 on 2 and 55 DF,  p-value: 5.792e-13
```

Comparing two mammals which differ in 1 unit of logarithm of the body weight but have the same brain weight, we would expect the logarithm of the gestation time for the larger mammal to be 0.1162322 units smaller . Converted into percentages, we would say: comparing two mammals which differ in body weight by 1/% but have the same brain weight, we would expect the gestation time for the larger mammal to have $100(1.01^{\hat{b_1}} - 1)$ = -0.116% difference.

Comparing two mammals which differ in 1 unit of logarithm of the brain weight but have the same body weight, we would expect the logarithm of the gestation time for the larger mammal to be -0.4831337 units smaller . Converted into percentages, we would say: comparing two mammals which differ in brain weight by 1/% but have the same body weight, we would expect the gestation time for the larger mammal to have $100(1.01^{\hat{b_1}} - 1)$ = 0.482% difference.