

Stat 242 Quiz – Topics Drawn from Chapters 9 through 12

What's Your Name? _____

We have a data set with the following information about different species of mammals:

- Species: The species of mammal
- Body: Average weight of the body
- Gestation: Average length of pregnancy
- Litter: Average litter size
- Brain: Average weight of the brain

We will use brain size as the response variable and the other variables as explanatory variables. Here is a look at the first few rows of the data, as well as the species in the data set. Note that there appear to be some closely related species in the data set; for example, there are three species of Porcupine, and four species of Deer mouse.

```
head(mammals)
```

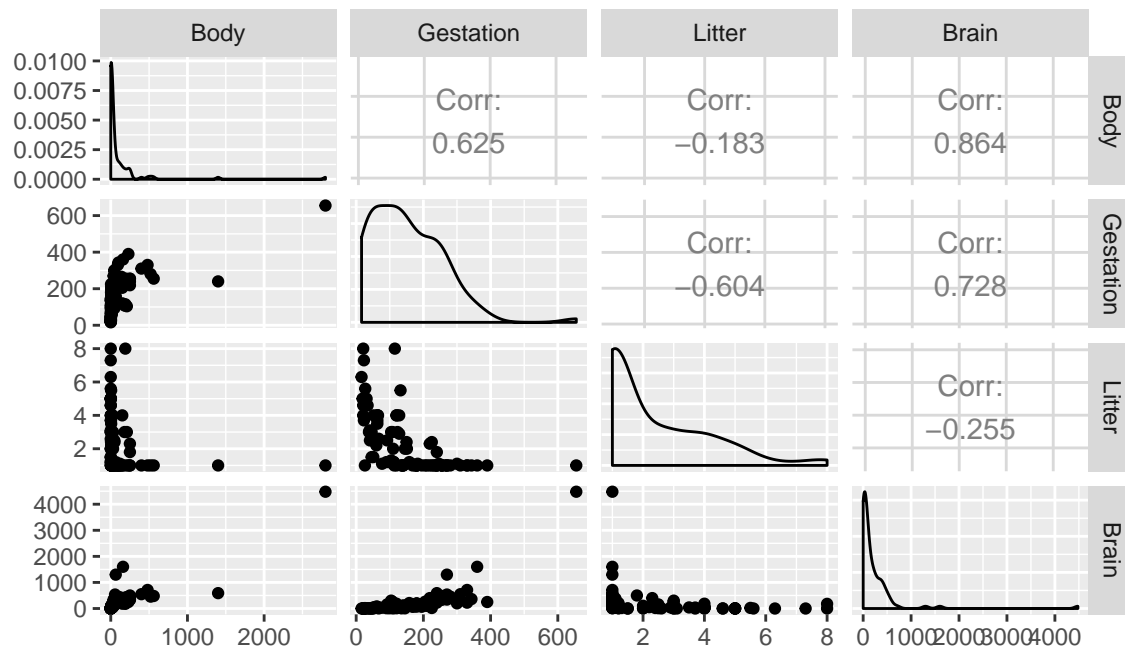
##	Species	Body	Gestation	Litter	Brain
## 1	Aardvark	2.20	31	5.0	9.6
## 2	Acouchis	0.78	98	1.2	9.9
## 3	African elephant	2800.00	655	1.0	4480.0
## 4	Agoutis	2.80	104	1.3	20.3
## 5	Axis deer	89.00	218	1.0	219.0
## 6	Badger	6.00	60	2.2	53.0

```
mammals$Species
```

## [1] Aardvark	Acouchis	African elephant
## [4] Agoutis	Axis deer	Badger
## [7] Barbary sheep	Barking deer	Bat-eared fox
## [10] Beaked whale	Beaver	Black buck antelope
## [13] Bush baby	Canadian beaver	Capybara
## [16] Caribou	Cattle	Chimpanzee
## [19] Chinchilla	Deer mouse I	Deer mouse II
## [22] Deer mouse III	Deer mouse IV	Dog
## [25] Dolphin	Domestic cat	Domestic goat
## [28] Domestic pig	Domestic sheep	Duikers
## [31] Eland	Elephant shrew I	Elephant shrew II
## [34] Elk	Fallow deer	Flying squirrel
## [37] Fur seal	Gentle lemur	Gorilla
## [40] Gray fox	Grizzly bear	Guinea pig
## [43] Hamadryas baboon	Hamster I	Hamster II
## [46] Harp seal	Hedgehog	Hippopotamus
## [49] Hopping mouse	Horse	House mouse
## [52] Howler monkey	Human being	Hyrax
## [55] Jack rabbit	Kinkajou	Leaf monkey
## [58] Lemur	Leopard	Lion
## [61] Llama	Long-nose armadillo	Lynx
## [64] Nutria	Orangutan	Porcupine I
## [67] Porcupine II	Porcupine III	Porpoise
## [70] Pygmy gerbil	Pygmy hippopotamus	Quokka
## [73] Raccoon	Rat I	Rat II
## [76] Red deer	Red fox	Rhesus monkey I
## [79] Rhesus monkey II	Ring-tail monkey	Sambar
## [82] Sea lion	Slow loris	Spider monkey I
## [85] Spider monkey II	Tapir	Tiger
## [88] Tree shrew	Tree squirrel	Vervet guenon
## [91] Vicuna	Weddell seal	Western baboon
## [94] White-handed gibbon	Wild boar	Yak
## 96 Levels: Aardvark Acouchis African elephant Agoutis Axis deer ... Yak		

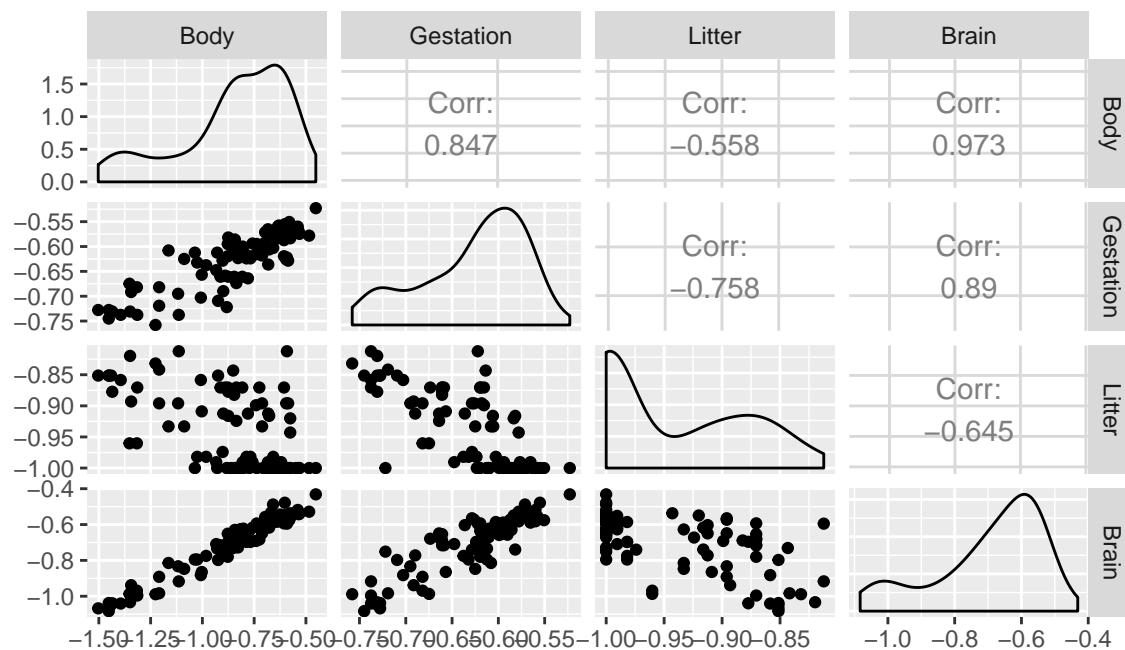
Initial Set Up

```
ggpairs(mammals %>% select(-Species))
```



```
mammals_transformed <- mammals %>%
  mutate(
    Body = -1/(Body^0.1),
    Gestation = -1/(Gestation^0.1),
    Litter = -1/(Litter^0.1),
    Brain = -1/(Brain^0.1)
  )
```

```
ggpairs(mammals_transformed %>% select(-Species))
```



Model 1: All Observations

```
lm_fit <- lm(Brain ~ Body + Gestation + Litter, data = mammals_transformed)
summary(lm_fit)

##
## Call:
## lm(formula = Brain ~ Body + Gestation + Litter, data = mammals_transformed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.069831 -0.020271 -0.002648  0.024110  0.080942
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.22674     0.13609  -1.666  0.09909 .
## Body         0.49035     0.02303  21.295 < 2e-16 ***
## Gestation    0.42641     0.13799   3.090  0.00265 **
## Litter       -0.22420     0.07939  -2.824  0.00581 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03037 on 92 degrees of freedom
## Multiple R-squared:  0.9654, Adjusted R-squared:  0.9643
## F-statistic: 856.5 on 3 and 92 DF,  p-value: < 2.2e-16

confint(lm_fit)

##              2.5 %      97.5 %
## (Intercept) -0.4970256  0.04354491
## Body         0.4446130  0.53607702
## Gestation    0.1523581  0.70046413
## Litter       -0.3818769 -0.06653050
```

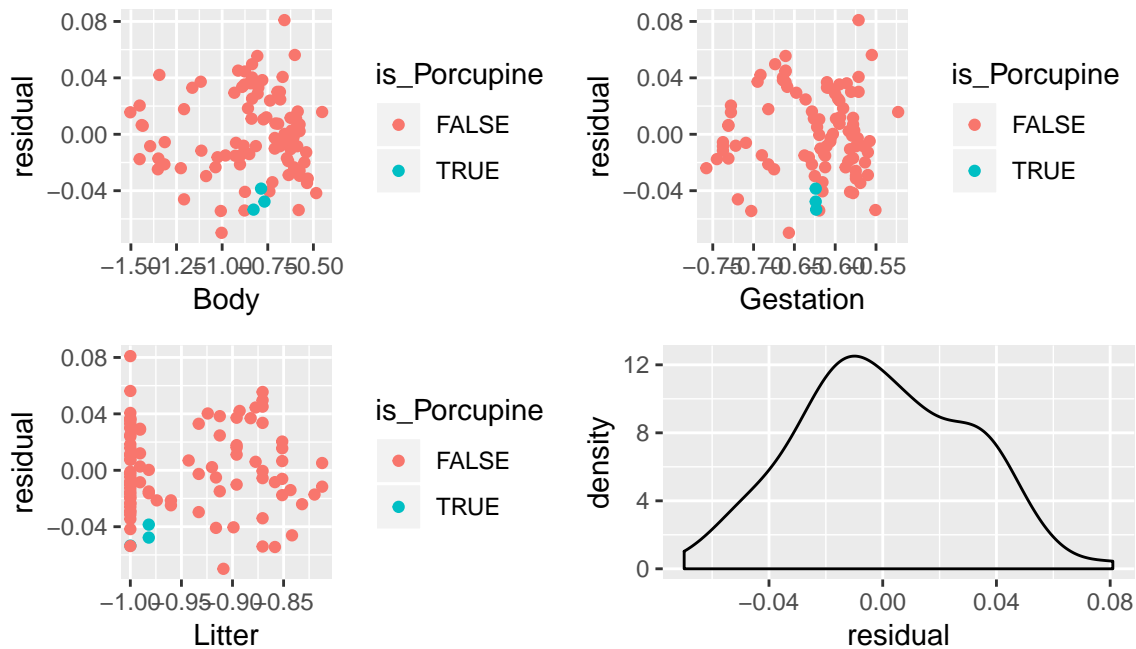
Examining Residuals

Let's look at the residuals plots. Recalling that there were three species of Porcupines in the data set, I have used a different color for the residuals for those species.

```
mammals_transformed <- mammals_transformed %>%
  mutate(
    is_Porcupine = Species %in% c("Porcupine I", "Porcupine II", "Porcupine III"),
    residual = residuals(lm_fit)
  )

p1 <- ggplot(data = mammals_transformed, mapping = aes(x = Body, y = residual, color = is_Porcupine)) +
  geom_point()
p2 <- ggplot(data = mammals_transformed, mapping = aes(x = Gestation, y = residual, color = is_Porcupine)) +
  geom_point()
p3 <- ggplot(data = mammals_transformed, mapping = aes(x = Litter, y = residual, color = is_Porcupine)) +
  geom_point()
p4 <- ggplot(data = mammals_transformed, mapping = aes(x = residual)) +
  geom_density()

grid.arrange(p1, p2, p3, p4, nrow = 2, ncol = 2)
```



Here are the residuals for the three species of porcupines in the data set:

```
mammals_transformed %>%
  filter(is_Porcupine) %>%
  pull(residual)
```

```
## [1] -0.03851209 -0.04770483 -0.05340879
```

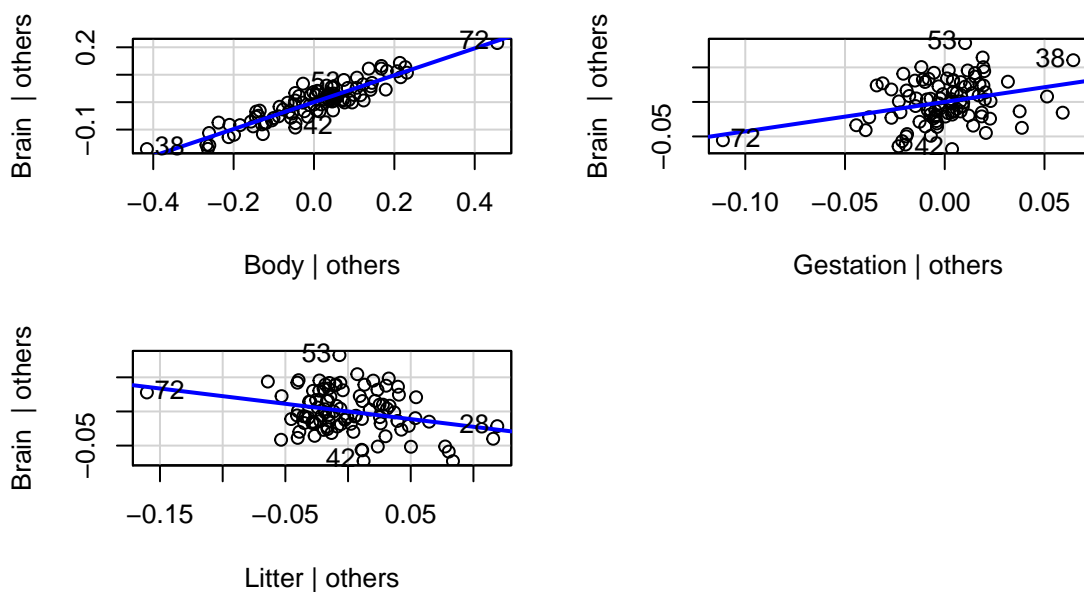
Variance Inflation Factor and added variable plot

```
vif(lm_fit)
```

```
##      Body Gestation  Litter
## 3.754219 6.072703 2.494167
```

```
avPlots(lm_fit)
```

Added-Variable Plots



Model 2: Setting aside some observations

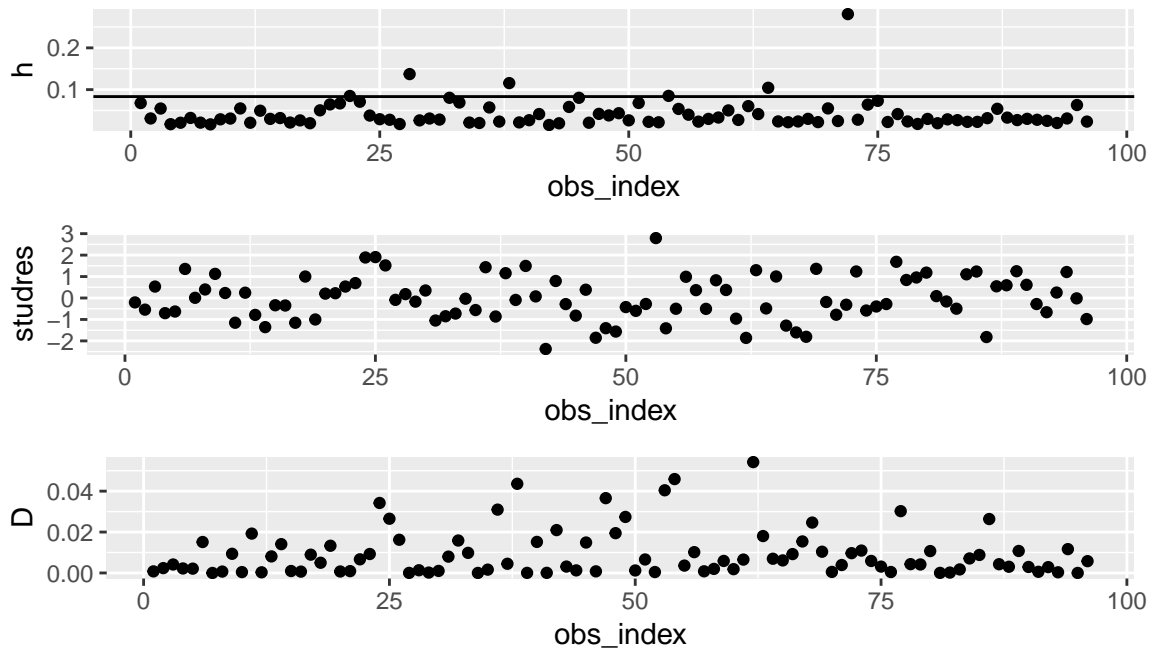
```
mammals_transformed <- mammals_transformed %>%
  mutate(
    obs_index = row_number(),
    h = hatvalues(lm_fit),
    studres = rstudent(lm_fit),
    D = cooks.distance(lm_fit)
  )

p1 <- ggplot(data = mammals_transformed, mapping = aes(x = obs_index, y = h)) +
  geom_hline(yintercept = 2*4/nrow(mammals_transformed)) +
  geom_point()

p2 <- ggplot(data = mammals_transformed, mapping = aes(x = obs_index, y = studres)) +
  geom_point()

p3 <- ggplot(data = mammals_transformed, mapping = aes(x = obs_index, y = D)) +
  geom_point()

grid.arrange(p1, p2, p3, ncol = 1)
```



```
obs_to_investigate <- c(28, 38, 53, 64, 72)

mammals_transformed <- mammals_transformed %>%
  mutate(
    suspicious = row_number() %in% obs_to_investigate
  )
mammals_no_suspicious <- mammals_transformed %>% filter(!suspicious)
```

```
lm_fit_no_suspicious <- lm(Brain ~ Body + Gestation + Litter, data = mammals_no_suspicious)
summary(lm_fit_no_suspicious)
```

```
##
## Call:
## lm(formula = Brain ~ Body + Gestation + Litter, data = mammals_no_suspicious)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-0.068596	-0.019677	-0.000896	0.022801	0.059827

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.31302	0.17954	-1.743	0.0848 .
Body	0.50516	0.02647	19.082	<2e-16 ***
Gestation	0.31709	0.17734	1.788	0.0773 .
Litter	-0.25513	0.10075	-2.532	0.0131 *

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02963 on 87 degrees of freedom
## Multiple R-squared:  0.9679, Adjusted R-squared:  0.9668
## F-statistic: 875.2 on 3 and 87 DF,  p-value: < 2.2e-16
```

Problems

(a) Explain why a transformation was necessary.

(b) Check all model conditions based on the model fit using the transformed data. For any conditions that are not met, suggest a step to take to address the problem.

- (c) Summarize the findings from this analysis about the strength of evidence of an association between body size, gestation length, litter size, and brain weight.
- (d) What is the interpretation of the coefficient estimate for the Body variable in the model fit including all observations (Model 1)?
- (e) What is the interpretation of the confidence interval for the coefficient of the Body variable in the model fit including all observations (Model 1)? Include a description of the meaning of the phrase “95% confident”.
- (f) The variance inflation factor for Body is 3.75. Rounding up to 4 for convenience, what does this value say about the width of a confidence interval for the coefficient of Body in the linear model?

(g) In the added variable plot for the Body variable, what is on the horizontal and vertical axes of the plot? How does the slope of the line in that plot relate to the coefficient estimate in the linear model?

(h) What is a high leverage observation? Why can high leverage observations be problematic?

(You should be able to draw a picture of a simple linear regression setting illustrating why this is a problem.)

(i) What does Cook's distance measure, at an intuitive level?

(j) Define multicollinearity in a sentence or two. Why can multicollinearity be problematic?