Final Project
Stats 510

By
Hilda Ulloa
Shraddha Swaroop and
Emalina Huerta

Introduction

Despite our group being cat owners and not owning dogs as pets, our group shared a collective curiosity in the exploration of dog intelligence. Dogs are known to be affectionate and loyal, as well as having a great sense of smell, taste, and auditory perception. We are curious to find out if some of these attributes can be indicators of their intelligence. Dr. Stanley Coren, a leading dog psychologist, did a study on 199 breeds and found that medium-sized and larger dogs were top performers, they were able to learn commands in less time than other breeds (Business Insider, 2018). We decided to investigate if the size of the dog has any relationship to the dog's intelligence based on a Kaggle dataset that was compiled by Coren on a smaller sample size. The file contained a sample size of 124 different dog breeds and the obedience scale of the dog's intelligence based on their height, weight and how many repetitions it took for the dog to learn a new command (high/ low reps). Based on the repetitions needed for the dogs to follow a command, the dogs were categorized into one of five different rankings of obedience, such as brightest dog, excellent working, above average, average, and fair. This set of information stood out to us since we all have an interest in dogs and who doesn't want to pick the smartest breed?

Questions of Interest

Research Questions

1. Does a dog's obedience rate predict a dog' height?
2. Is there a correlation between a dog's weight and height?
3. What are the strongest predictors in determining a dog's height?

Regression Method

The following is a list of the variables that we used in our research questions including our outcome variable, Average Height, and predictors which included a categorical predictor Obey (split into 5 classifications of Brightest Dogs, Excellent working dogs, Above average, Average and Fair) and other continuous predictors including reps lower, reps upper, weight high, weight low, and average weight.

y = average height

x1 = reps lower

x2 = reps upper

x3 = obey

x4 = weight high (lbs)

x5 = weight low (lbs)

x6 = average weight

Regression Analysis

We first ran our data, through a variable selection process using a multitude of Regressions, Stepwise Regression AIC (see Appendix A), Best Subset Regression (see Appendix B to D), and Mallow's CP (see Appendix E). The results yielded the following model as our best,

Model 5: y = x1+ x2 +x3 +x4 +x5.

Before we could start the analysis process, we made sure that line conditions of our linear regression model were met. . This means checking for independence, normality and equal variances. We ran a residual vs fit plot to measure linearity and equal variances, see Figure 1 Residuals vs Fitted graph.

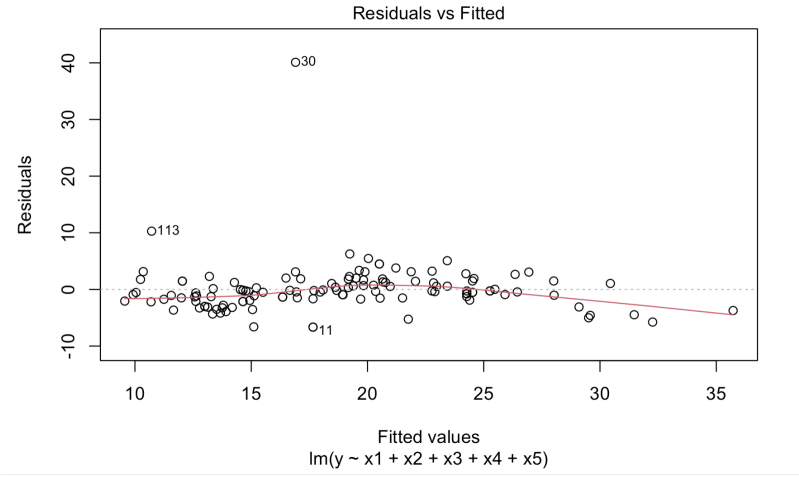


Figure 1 shows that most residuals bounce along the 0-line meeting the equal variances assumption. In addition, residuals fall in a horizontal band showing linearity.

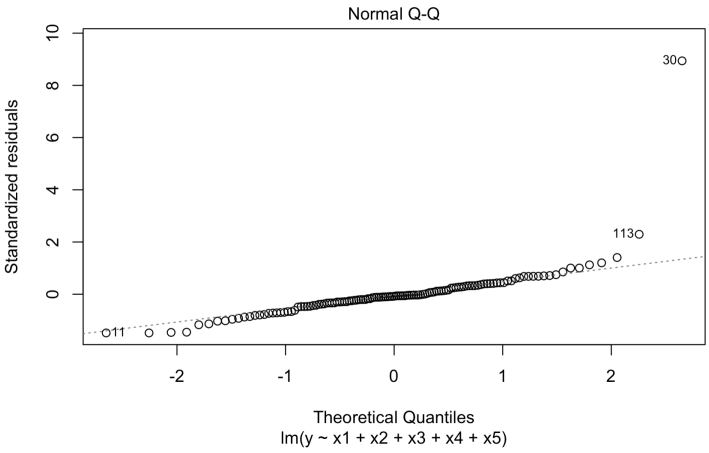
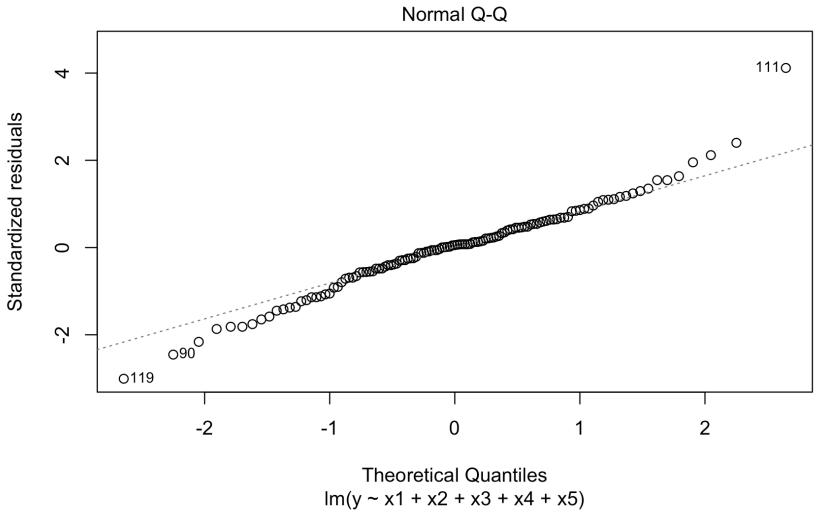
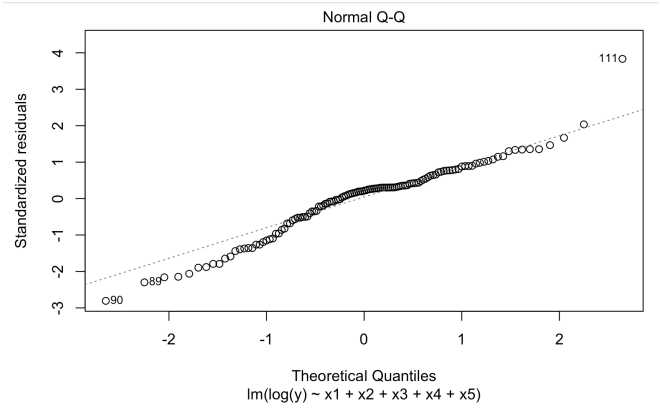


Figure 2, the Normal Q-Q plot of the residuals shows it is approximately linear supporting the condition that the error terms are normally distributed.



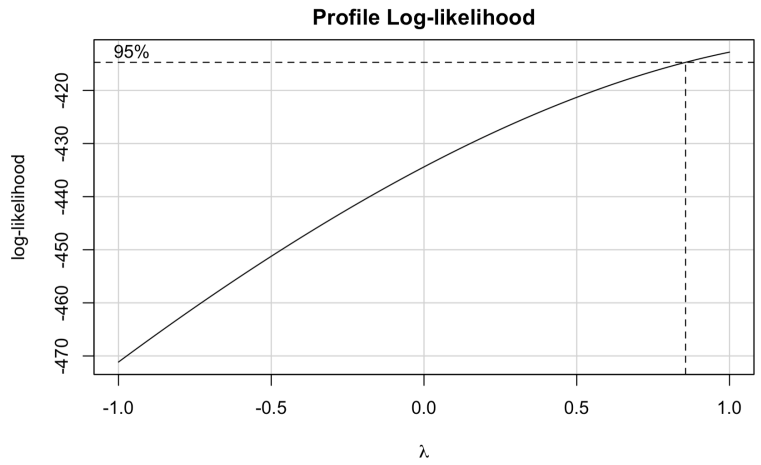
In addition the Shapiro Wilks test shows a p-value of 2.457e-07, which is much smaller than .05 which allows us to fail to reject the null hypothesis which means we cannot assume normality. We removed the outliers of data points 30, 113, 121, 91 but failed to achieve normality as you can see from **Figure 3**. The Shapira Wilks test shows a p-value of 0.000594. This is also smaller than .05 which also leads us to fail to reject the null hypothesis which means we cannot assume normality.

Our next step was to log transform y to address our normality issue, see the result in Figure 4. The Shapira Wilks test shows a p-value of 3.609e-05. This is also smaller than .05, even smaller than our previous attempt, which also leads us to fail to reject the null hypothesis which means we cannot assume normality. We ran a Box Cox transformation which returned a lambda of 1 indicating a transformation



was not necessary. We attempted all the steps that we learned in class to transform variables to adjust for normality and were not successful.

Figure 5: Box-Cox Transformation with a value of 1

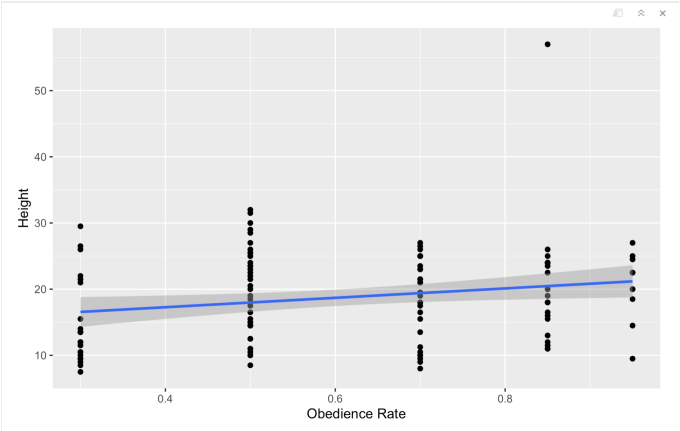


Results and Interpretation

RQ1: Does obedience predict height?

After conducting a multiple linear regression analysis on this question in R Studio, we found that there is a positive linear relationship between obedience and height but this relationship was not significant.

Figure 6: Relationship between dog's height and obedience rate



Multiple linear regression was used to test if dogs obedience categorization significantly predicted height of dog. The fitted regression model was: Height = 34.13 - 0.16(obey) –

12.79(upper reps). The overall regression was not statistically significant (R2 = 0.07, F -test (1, 122) = 2.96, p = < .08768). It was found that obedience rate and commands did not significantly predict dog height (β = -.16, p = .0855).We can be 95% confident that the mean number of commands dogs mastered is between 10.55 and -30.15, regardless of the obedience rate.

The interpretation of this finding is that obedience rate, and upper repetition of commands are not an accurate predictor of a dog's height.

RQ2: Is there a correlation between dogs weight and height?

To answer this question, let's take a look at our response variable; height, and our predictor, average weight. Now if we take a look at our graph we can see that there is strong positive linear correlation between the predictor and the outcome, see Figure 7

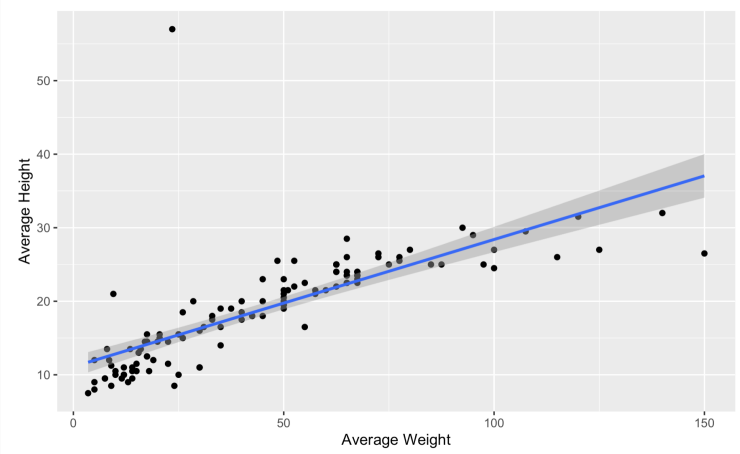


Figure 7: Average Weight vs. Avg Height

The fitted regression model is as follows: Height = 11.10 + 0.17(average weight). The overall regression was statistically significant, with R2 = 0.57, F -test (1, 123) = 162.77, where p <= 2.2e-16). We found that average weight significantly predicts a dog height (β = 0.17, p = .2e-16). Therefore, we can be 95% confident that the mean weight of dogs increases between .1461 and .1998 lbs.

RQ3: What are the strongest predictors in determining a dog's height?

After running multiple regression analysis tests we were able to determine the best model for our data. We had to check for the smallest AIC and the largest adjusted R^2 for each model in our data set. Appendix A shows the smallest AIC and Appendix C shows the largest adjusted R^2. This has concluded the model with y=x1+x2+x3+x4+x5 is the best fit for the data Appendix B also justifies this conclusion. Hence, model 5 is the best model.

Conclusion

Before conducting any sort of analysis we first had to determine which variables would be our response variable, and which others would be our predictors. We initially wanted the obey variable to be our predictor, but after some trial and error we concluded that because this was a categorical variable it could not be used as our predictor, and therefore we used average height as our predictor.

In conclusion we can say there is a positive correlation between our response height and our predictor obedience. Since we were able to conclude there was linearity in our residual vs fit plot. Although this failed the normality test after trying to delete our outliers, running a Box-Cox test, and using the transformation on height it still continued to fail the normality test in our Q-Q plot. Our second research question focused on if a dog's weight would predict height. We ran a residual vs fit plot and we were able to see there was a positive relationship between the response height to the predictor weight. Lastly we wanted to know what would be the best model for our data set therefore we ran a stepwise function and we checked for the largest adjusted R^2 which concluded model 5 was the best.

References

Bray, E. E., Gruen, M. E., Gnanadesikan, G. E., Horschler, D. J., Levy, K. M., Kennedy, B. S., ... & MacLean, E. L. (2021). Dog cognitive development: a longitudinal study across the first 2 years of life. Animal cognition, 24(2), 311-328.

Business Insider. (2018, November 10). *These Are The 'Smartest' Dog Breeds, According to a Canine Psychologist : ScienceAlert*. ScienceAlert. https://www.sciencealert.com/smartest-dog-breeds-canine-psychologist-intelligence-pets

Carballo, F., Cavalli, C. M., Gácsi, M., Miklósi, Á., & Kubinyi, E. (2020). Assistance and therapy dogs are better problem solvers than both trained and untrained family dogs. Frontiers in Veterinary Science, 7, 164.

Duranton, C., Rödel, H. G., Bedossa, T., & Belkhir, S. (2015). Inverse sex effects on performance of domestic dogs (Canis familiaris) in a repeated problem-solving task. Journal of Comparative Psychology, 129(1), 84–87. https://doi.org/10.1037/a0037825

Marshall-Pescini, S., Valsecchi, P., Petak, I., Accorsi, P. A., & Previde, E. P. (2008). Does training make you smarter? The effects of training on dogs' performance (Canis familiaris) in a problem solving task. Behavioural processes, 78(3), 449-454.

Appendices

Appendix A: Stepwise Regression AIC

```
Start:  AIC=492.38
y ~ 1

       Df Sum of Sq    RSS    AIC
+ x5    1    3652.7 2666.3 386.52
+ x6    1    3599.2 2719.8 389.00
+ x4    1    3447.7 2871.3 395.78
+ x2    1     363.3 5955.7 486.97
+ x1    1     302.6 6016.4 488.24
+ x3    1     263.7 6055.2 489.05
<none>              6319.0 492.38

Step:  AIC=386.52
y ~ x5

       Df Sum of Sq    RSS    AIC
+ x2    1      99.9 2566.5 383.75
+ x1    1      86.7 2579.6 384.39
+ x3    1      73.3 2593.0 385.03
<none>              2666.3 386.52
+ x4    1       4.7 2661.6 388.30
+ x6    1       4.7 2661.6 388.30
- x5    1    3652.7 6319.0 492.38

Step:  AIC=383.75
y ~ x5 + x2

       Df Sum of Sq    RSS    AIC
<none>              2566.5 383.75
+ x3    1      10.7 2555.7 385.22
+ x4    1       7.6 2558.8 385.37
+ x6    1       7.6 2558.8 385.37
+ x1    1       2.4 2564.0 385.63
- x2    1      99.9 2666.3 386.52
- x5    1    3389.2 5955.7 486.97
=

 Call:
 lm(formula = y ~ x5 + x2)

 Coefficients:
 (Intercept)           x5           x2
    12.59089      0.20225     -0.03924
```

Appendix B: Best Subsets Regression

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6, data = dogs)

Residuals:
   Min     1Q Median     3Q    Max
-6.658 -1.707 -0.352  1.436 40.094

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.585768  22.987837   1.940  0.05480 .
x1           -0.604330   0.481688  -1.255  0.21208
x2           -0.006466   0.083845  -0.077  0.93866
x3          -33.964969  24.358667  -1.394  0.16580
x4            0.024829   0.041184   0.603  0.54773
x5            0.167676   0.057971   2.892  0.00455 **
x6                  NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.598 on 119 degrees of freedom
Multiple R-squared:  0.6019,    Adjusted R-squared:  0.5851
F-statistic: 35.98 on 5 and 119 DF,  p-value: < 2.2e-16
```

1	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
2	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
3	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
4	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
5	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE

Appendix C: Best Subsets Regression with R^2

```
#best subset regression with r^2

Model 1: y = x5 -----------------[57.80]
Model 2: y = x2+ x5--------------[59.39]
Model 3: y = x1 + x3 + x5--------[60.06]
Model 4: y = x1 + x3+ x4 +x5-----[60.18]
Model 5: y = x1+ x2 +x3 +x4 +x5 -[60.19]
```

```{r}
summary.mod$rsq
```

```
 [1] 0.5780490 0.5938513 0.6006086 0.6018433 0.6018632
```

Appendix D: Best Subsets Regression with R^2, MSE and adjusted R^2

```{r}
n = 124
rss = summary.mod$rss
mses = c(rss[1]/(n-2), rss[2]/(n-3), rss[3]/(n-4), rss[4]/(n-5), rss [5]/(n-6))
mses
```

```
Warning: The working directory was changed to /Users/shraddhaswaroop inside a notebook chunk. The working directory will be reset when the chunk is finished running. Use the knitr root.dir option in the setup chunk to change the working directory for notebook chunks.[1] 21.85498 21.21036 21.03128 21.14245 21.32055
```

```
#best subset regression with r^2 and mse and adr r^2
model >  r^2 > adj r^2 and MSE

Model 1: y = x5 -----------------[57.80]-------- [57.46]-----[21.85]
Model 2: y = x2+ x5--------------[59.39]---------[58.72]-----[21.21]
Model 3: y = x1 + x3 + x5--------[60.06]---------[59.07]-----[21.03]
Model 4: y = x1 + x3+ x4 +x5-----[60.18]---------[58.86]-----[21.14]
Model 5: y = x1+ x2 +x3 +x4 +x5 -[60.19]---------[58.51]-----[21.32]
```

Appendix E: Mallow's CP (prediction bias)

Prediction bias

```{r}
cp = summary.mod$cp
summary.mod$which
```

```
  (Intercept)    x1    x2    x3    x4   x5    x6
1        TRUE FALSE FALSE FALSE FALSE TRUE FALSE
2        TRUE FALSE  TRUE FALSE FALSE TRUE FALSE
3        TRUE  TRUE FALSE  TRUE FALSE TRUE FALSE
4        TRUE  TRUE FALSE  TRUE  TRUE TRUE FALSE
5        TRUE  TRUE  TRUE  TRUE  TRUE TRUE FALSE
```

```{r}
summary.mod$cp
```

```
[1] 4.058080 1.374600 1.371864 3.005898 5.000000
```

Use Model 5 because it is the most unbiased

Model 5: y = x1+ x2 +x3 +x4 +x5 -- 6 & 5 ------------Unbiased

Appendix F: Summary fit for model

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = dogs)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8618 -1.4235  0.1437  1.4560 10.6213

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.52874   13.36136   1.462   0.1465
x1           0.02688    0.28071   0.096   0.9239
x2          -0.11604    0.04903  -2.367   0.0196 *
x3          -8.88429   14.14690  -0.628   0.5312
x4           0.03511    0.02378   1.477   0.1424
x5           0.16370    0.03349   4.888 3.27e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.651 on 117 degrees of freedom
Multiple R-squared:  0.8288,    Adjusted R-squared:  0.8215
F-statistic: 113.3 on 5 and 117 DF,  p-value: < 2.2e-16
```

Appendix G: Summary fit for RQ1

```
#RQ1
Y=B0 + B1x1(Obey) + B2x1(Commands_Upper)
H0: B1 = B2  = 0
H1 = Either B1 or B2 \ne$ 0
```{r}
fit = lm(y ~ x2 +x3, data = dogs)
summary(fit)
```
```

```
Call:
lm(formula = y ~ x2 + x3, data = dogs)

Residuals:
    Min      1Q  Median      3Q     Max
-11.625  -4.729  -0.625   4.275  37.275

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.41205    9.48546   3.206  0.00172 **
x2          -0.15708    0.09124  -1.722  0.08768 .
x3          -9.80051   10.27854  -0.953  0.34223
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.961 on 122 degrees of freedom
Multiple R-squared:  0.06447,   Adjusted R-squared:  0.04913
F-statistic: 4.203 on 2 and 122 DF,  p-value: 0.01716
```

Appendix H: R code for replication and analysis

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library (olsrr)
library (leaps)
library(ggplot2)
library(GGally)
library(car)
library(readr)
```
```{r Auto}
dogs = read.csv ('~/Desktop/Desktop/_Spring 2023 Classes/Stats 510/Group Project/GroupProject/Dataset/dogs.csv', header = T)
summary (dogs)
colnames (dogs)
```
```{r}
view(dogs)
n=nrow(dogs)
summary(dogs)
```
```{r}
y = dogs$average_height
x1 = dogs$reps_lower
x2 = dogs$reps_upper
x3 = dogs$obey
x4 = dogs$weight_high_lbs
x5 = dogs$weight_low_lbs
x6 = dogs$avg_weight
```
```{r}
mod0 = lm (y ~ 1)
mod.upper = lm(y ~ x1+x2+x3+x4+x5+x6)
step (mod0, scope = list (lower = mod0, upper = mod.upper))
```
```{r}
dogslm = lm (y ~x1 + x2 + x3 +x4 + x5+ x6, data = dogs)
summary (dogslm)
```

```
```{r}
library(leaps)
mod = regsubsets(cbind(x1,x2,x3,x4,x5,x6), y)
summary.mod = summary(mod)
summary.mod$which
```
```{r}
summary.mod$rsq
```
```{r}
n = 124
rss = summary.mod$rss
mses = c(rss[1]/(n-2), rss[2]/(n-3), rss[3]/(n-4), rss[4]/(n-5), rss [5]/(n-6))
mses
```
```{r}
summary.mod$adjr2
```
```{r}
cp = summary.mod$cp
summary.mod$which
```
```{r}
summary.mod$cp
```
```{r}
mod1 =lm (y ~ x1+ x2 +x3 +x4 +x5, data = dogs)
summary (mod1)
```
```{r}
plot(mod1, which = 1)
```
```{r}
fit = lm(y ~ x1 + x2 +x3 + x4 + x5 , data = dogs)
summary(fit)
```
```{r}
plot(mod1, which = 2)
```
```

```{r}
shapiro.test(y)
```
```{r}
ggplot(data = dogs, aes(x = x3, y = y)) +
geom_point() +
geom_smooth(method = "lm", formula = y ~ x) +
labs(x = "Obedience Rate", y = "Height")
```
```{r}
ggplot(data = dogs, aes(x = x6, y = y)) +
geom_point() +
geom_smooth(method = "lm", formula = y ~ x) +
labs(x = "Average Weight", y = "Height")
```
```{r}
ggplot(data = dogs, aes(x = x2, y = y)) +
geom_point() +
geom_smooth(method = "lm", formula = y ~ x) +
labs(x = "Upper Reps", y = "Height")
```
```{r}
ggplot(data = dogs, aes(x = x1, y = y)) +
geom_point() +
geom_smooth(method = "lm", formula = y ~ x) +
labs(x = "Lower Reps", y = "Height")
```
```{r}
fit = lm(y ~ x6, data = dogs)
summary(fit)
```
```{r}
mod3 = lm(y ~ x6, data = dogs)
summary(mod3)
```
```{r}
mod.reduced = lm(y ~ 1 )
mod.full = lm(y ~ x6)
anova(mod.reduced, mod.full)
```

```{r}
fit = lm(y ~ x6, data = dogs)
confint(fit)
```

```{r}
dogs_1 <- dogs[-c(30,112),  ]
```
```{r}
View(dogs_1)
```

```{r}
y =  dogs_1$average_height
x1 = dogs_1$reps_lower
x2 = dogs_1$reps_upper
x3 = dogs_1$obey
x4 = dogs_1$weight_high_lbs
x5 = dogs_1$weight_low_lbs
x6 = dogs_1$avg_weight
```

```{r}
mod12 = lm(y ~ x1 + x2 + x3 + x4+ x5, data = dogs_1)
summary(mod12)
```
```{r}
plot(mod12, which = 2)
```
```{r}
dogs_1 <- dogs[-c(113,121,91),  ]
```

```{r}
mod12 = lm(y ~ x1 + x2 + x3 + x4+ x5, data = dogs_1)
summary(mod12)
```
```{r}
plot(mod12, which = 2)
```
```{r}
mod13 = lm(log(y) ~ x1 + x2 + x3 + x4+ x5, data = dogs_1)
summary(mod13)
```
```{r}
plot(mod13, which = 2)
```

```
```
```{r}
shapiro.test (log(y))
```

```{r}
boxcox.trans=boxCox(mod12, lambda = seq(-1, 1, length = 10))
```

```{r}
lam<- boxcox.trans$x[which.max(boxcox.trans$y)]
lam
```