Week 5 stats

Will Jones

04/03/2021

## Model checking

## clean data

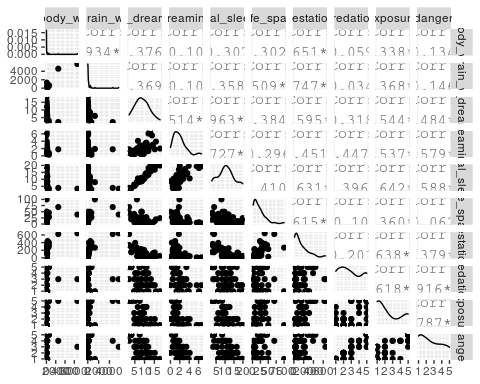
mammals <- read\_csv("data/mammals.csv")

skim(mammals)#missing values  
names(mammals)#all same names  
unique(mammals$species)#names only appear once so all correct   
str(mammals)# 1 character which is species rest are numeric   
head(mammals)# clean data

i have some NAs therefore: 1) Do you have any 2) Are they in the variable you will actually be testing? E.g. if there is an NA but in a different column, you don’t want to drop that whole row if it doesn’t affect your analysis

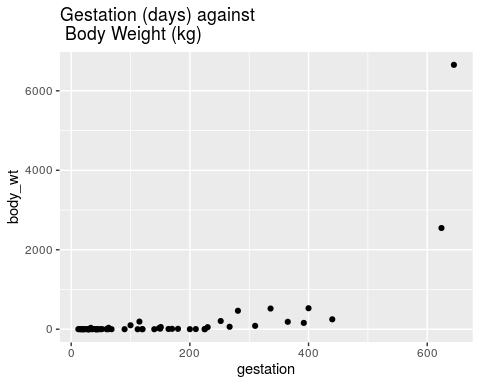
are they in my column= non in body weight but some gestation: therefore:

inspect stats using ggally



Explore a relationship: Gestation and adult body weight We will expore the relationship between the length of gestation and adult bodyweight. With the working hypothesis that the length of gestation is a reliable predictor of adult bodyweight across species.

p <- mammals %>%   
 ggplot(aes(x=gestation, y=body\_wt))+  
 geom\_point()+  
 ggtitle("Gestation (days) against \n Body Weight (kg)")  
  
p



## Questions

Does it look like a linear relationship makes sense?

Do we have any concerns about modeling a linear relationship?

Any outliers?

Any initial thoughts on homoscedasticity (equal variance of the residuals across the model?)

Linear relationship: Maybe, though it is difficult to see because of the very narrow distribution of bodyweights across the mammals

Modeling concerns, definitely - it would be difficult to fit a good least squares line through this data

Clearly two very large outliers on both gestation and body weight at the high end of the scale

Also clearly a funnel shape as the values increase, so major concerns there will be heteroscedasticity (unequal residual variance across the fit of our regression).

## Model it

Once we’ve decided that a linear relationship makes sense, we’ll model it using lm(). Note that we haven’t checked all assumptions yet. That’s because a lot of our assumptions for linear regression are based on model residuals (e.g. normality & homoscedasticity of residuals), which we can’t calculate until after we find the predicted values from the model (residual=y actual−y predicted).

So let’s make our first model (body wt is the dependent variable

gestation is the independent everything right of ~ indicates a predictor or independent variable)

gestation\_model <- lm(body\_wt~gestation, data=mammals)  
summary(gestation\_model)

##   
## Call:  
## lm(formula = body\_wt ~ gestation, data = mammals)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1187.5 -317.0 119.2 251.0 4372.0   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -374.9706 130.6333 -2.87 0.00578 \*\*   
## gestation 4.1193 0.6417 6.42 3.13e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 711.2 on 56 degrees of freedom  
## (4 observations deleted due to missingness)  
## Multiple R-squared: 0.4239, Adjusted R-squared: 0.4136   
## F-statistic: 41.21 on 1 and 56 DF, p-value: 3.132e-08

we found the : slope of the line is 4.12 the y-intercept is -374.97

so extra 4 days of gestation incerases body weght by 1 kg, look this up not too sure

use broom::tidy() to get the model outputs in a nice dataframe format

tidy\_gestation <- broom::tidy(gestation\_model, conf.int=T)  
tidy\_gestation

## # A tibble: 2 x 7  
## term estimate std.error statistic p.value conf.low conf.high  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -375. 131. -2.87 0.00578 -637. -113.   
## 2 gestation 4.12 0.642 6.42 0.0000000313 2.83 5.40

## extract the intercept

tidy\_gestation$estimate[1]

## [1] -374.9706

## extract the slope

tidy\_gestation$estimate[2]

## [1] 4.119331

## model information (degrees of freedom, F-statistic, p-value)

glance\_gestation <- broom::glance(gestation\_model)  
glance\_gestation

## # A tibble: 1 x 12  
## r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.424 0.414 711. 41.2 3.13e-8 1 -462. 930. 937.  
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

To find the predicted values and the residuals for each species adult body weight for their gestation period, we can use broom::augment()

augment\_gestation <- broom::augment(gestation\_model, interval="confidence")  
augment\_gestation

## # A tibble: 58 x 11  
## .rownames body\_wt gestation .fitted .lower .upper .resid .hat .sigma  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 6.65e+3 645 2282. 1609. 2955. 4372. 0.223 260.  
## 2 2 1.00e+0 42 -202. -429. 25.3 203. 0.0254 717.  
## 3 3 3.38e+0 60 -128. -343. 87.1 131. 0.0228 717.  
## 4 4 9.20e-1 25 -272. -512. -31.7 273. 0.0285 717.  
## 5 5 2.55e+3 624 2195. 1549. 2842. 352. 0.206 716.  
## 6 6 1.06e+1 180 367. 173. 560. -356. 0.0184 716.  
## 7 7 2.30e-2 35 -231. -463. 1.67 231. 0.0266 717.  
## 8 8 1.60e+2 392 1240. 868. 1611. -1080. 0.0680 702.  
## 9 9 3.30e+0 63 -115. -329. 97.6 119. 0.0224 717.  
## 10 10 5.22e+1 230 572. 354. 791. -520. 0.0235 714.  
## # … with 48 more rows, and 2 more variables: .cooksd <dbl>, .std.resid <dbl>

### 5. Model assumptions

Let’s use this information from augment\_gestation to manually evaluate some assumptions.

Linearly related variables (CHECK - already looked & thought about this before starting)

Normally distributed residuals

Homoscedasticity (constant residuals variance)

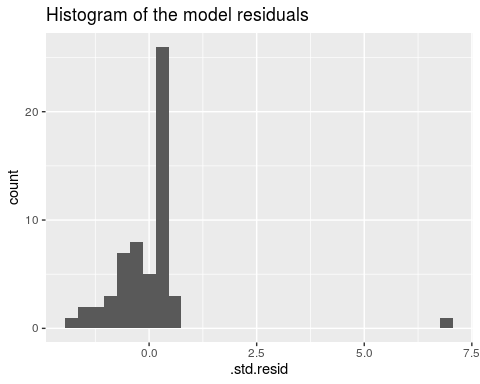
Influential outliers

Residuals distribution The standardised residuals for our model (y actual −y predicted ) are stored in the $.std.resid column from the broom::augment() function.

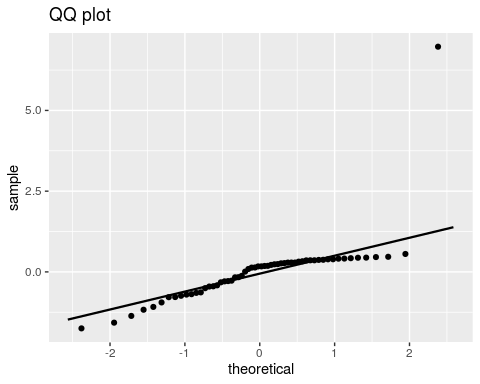
Here we create a histogram eyeball this histogram and see if your residuals look like they follow a normal distribution? yes

augment\_gestation %>%   
ggplot(aes(x = .std.resid)) +  
 geom\_histogram()+  
 ggtitle("Histogram of the model residuals")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 QQ plot of the residuals this shows the same as the histogram but as a different plot

augment\_gestation %>%  
ggplot(aes(sample = .std.resid)) +  
 geom\_qq()+  
 stat\_qq\_line()+  
 ggtitle("QQ plot")

 This looks like an example of an overall left-skew with one extreme outlier

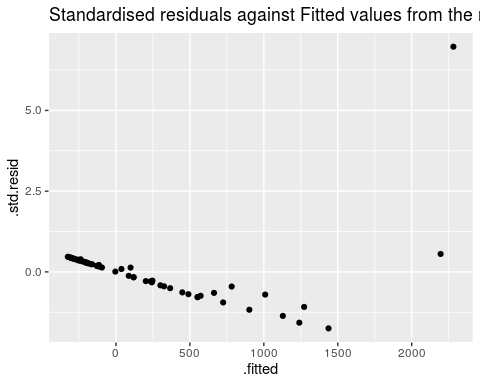
left skew goes from below the line touches it and then goes back down like a c facing down right skew comes from above touches the line then goes back up Under-dispersed data is an s shape Over-dispersed data has a horizontal s shape, =Over-dispersed data has an increased number of outliers

### Homoscedasticity

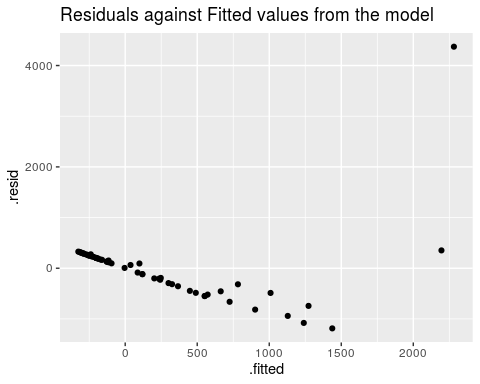
Does it look like the variance (spread) of residuals changes over the span of the model?

Violations of homoscedasticity may lead to wrongly large or small errors (and our confidence intervals) associated with coefficients, but will not affect the model estimates (coefficients) themselves.

augment\_gestation %>%  
ggplot(aes(x=.fitted, y= .std.resid)) +  
 geom\_point()+  
 ggtitle("Standardised residuals against Fitted values from the model")



augment\_gestation %>%  
ggplot(aes(x=.fitted, y= .resid)) +  
 geom\_point()+  
 ggtitle("Residuals against Fitted values from the model")



Note that while the pattern is identical, the scale is very different. this is because your residuals are on the original scale of the dependent variable, whereas your standardised residuals have been fitted onto a distribution of standard deviations

Here we can see there is clearly both an increase in the variance of the residuals as the fitted values increase AND there is a clear trend in the residuals. Not looking good! Remember what we WANT to see is no trends here and an even distribution of the residuals - this would indicate our model has explained most or all of the linear pattern, and has an equal amount of error along the regression.

###Cook’s Distance (influential outliers) If Cook’s D is greater than 4n for any observation, where n is the number of observations used to create the model, then that observation is strongly influential. This does NOT mean you should just remove that observation. In fact, you should plan on leaving all observations in unless you have really good reason not to.

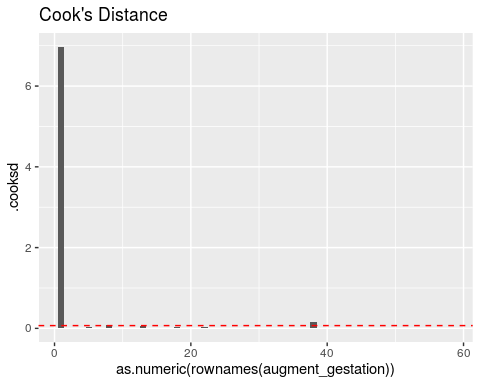
In our example (mammals), n = 62, so the threshold for a second look is 462. Let’s make & store that as a variable here:

cook\_limit <- as.numeric(4 / count(augment\_gestation))  
### I use the augmented dataframe in case any NA values were dropped while fitting the model.   
cook\_limit

## [1] 0.06896552

model:makes teh cook limit for you

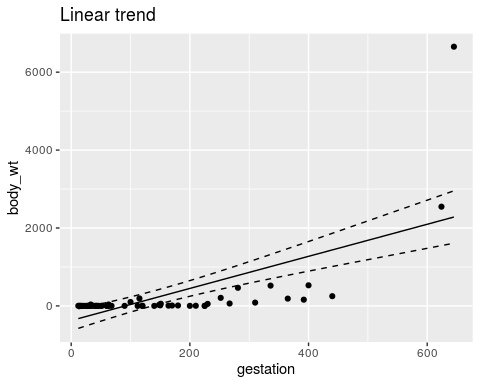
augment\_gestation %>%   
ggplot(aes(x = as.numeric(rownames(augment\_gestation)), y = .cooksd)) +  
 geom\_col() +  
 geom\_hline(yintercept = cook\_limit,  
 color = "red",  
 linetype = "dashed")+  
 ggtitle("Cook's Distance")

 can see one or two points having a huge impact on the model line

### Modelling predictions and residuals

how well our Regression line fits the pattern of our dataset & we can look at the pattern of our residuals against the dependent variable.

augment\_gestation %>%   
 ggplot(aes(x=gestation, y=body\_wt))+  
 geom\_line(aes(x=gestation, y=.fitted))+  
 geom\_line(aes(x=gestation, y=.upper), linetype="dashed")+  
 geom\_line(aes(x=gestation, y=.lower), linetype="dashed")+  
 geom\_point()+  
 ggtitle("Linear trend")



### Refit your model

force our residuals into a more normal distribution transforming data and producing a closer approximation of a linear relationship

suitable transformation: Ok let’s try and fit a new model with a Log10 transformation of our dependent variable

log10\_model <- lm(log10(body\_wt)~gestation, data=mammals)  
summary(log10\_model)

##   
## Call:  
## lm(formula = log10(body\_wt) ~ gestation, data = mammals)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.05037 -0.64966 0.09736 0.57278 1.86558   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.404357 0.162460 -2.489 0.0158 \*   
## gestation 0.007149 0.000798 8.958 2.11e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8845 on 56 degrees of freedom  
## (4 observations deleted due to missingness)  
## Multiple R-squared: 0.589, Adjusted R-squared: 0.5816   
## F-statistic: 80.24 on 1 and 56 DF, p-value: 2.11e-12

from this tranformation we will now see that we have

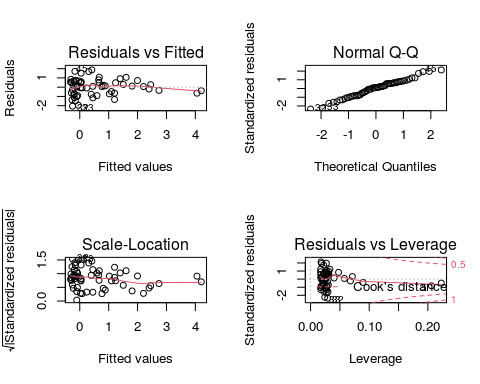
a Normal distribution to our residuals

Removed patterns in our residuals against fitted, and greatly reduced heteroscedasticity (variance along the fit of our regression line)

We still have two influential outliers but on a much reduced scale.

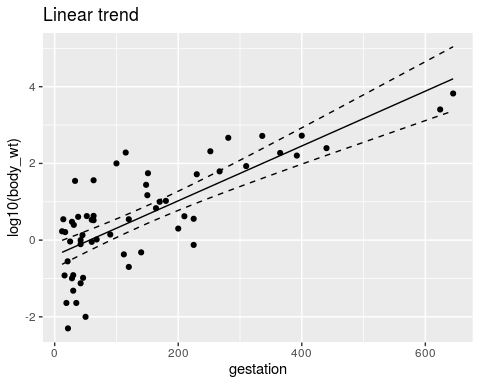
An alternate approach to make those graphs (that you’re more likely to use, and is fine): Just get the diagnostic plots using plot(model\_name):

par(mfrow=c(2,2))  
plot(log10\_model)



Now that we’ve explored the assumptions and have decided that our linear regression is now a valid tool to describe the relationship between gestation and bodyweight, let’s look at the model.

augment\_log10\_model <- broom::augment(log10\_model, interval="confidence")  
plot9 <-   
augment\_log10\_model %>%   
 ggplot(aes(x=gestation, y=`log10(body\_wt)`))+  
 geom\_line(aes(x=gestation, y=.fitted))+  
 geom\_line(aes(x=gestation, y=.upper), linetype="dashed")+  
 geom\_line(aes(x=gestation, y=.lower), linetype="dashed")+  
 geom\_point()+  
 ggtitle("")  
plot9+ggtitle("Linear trend")

 ## Write a brief summary of your best-fit model Things to include:

Mention the data transformation

Describe the slope of the regression

F statistic, degrees of freedom and P-value

The variance explained by the model

broom::tidy(log10\_model, conf.int=T)

## # A tibble: 2 x 7  
## term estimate std.error statistic p.value conf.low conf.high  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -0.404 0.162 -2.49 1.58e- 2 -0.730 -0.0789   
## 2 gestation 0.00715 0.000798 8.96 2.11e-12 0.00555 0.00875

broom::glance(log10\_model)

## # A tibble: 1 x 12  
## r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.589 0.582 0.884 80.2 2.11e-12 1 -74.2 154. 161.  
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

### write up:

A simple linear regression was used to explore the relationship between gestation in days and the average adult body weight across mutliple mammalian species. A log10 transformation was applied to the mammalian bodyweight variable in order to better approximate a linear fit for the regression. A significant regression equation was found β = 0.007(95%CI: 0.006-0.009), F1,56 = 80.24, P < 0.001, with an R^2 of 059.

# stargazer package

this will need to be modififed between html and pdf formats (<https://www.jakeruss.com/cheatsheets/stargazer/>)

stargazer::stargazer(log10\_model, type="html", ci=TRUE)

##   
## <table style="text-align:center"><tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"></td><td><em>Dependent variable:</em></td></tr>  
## <tr><td></td><td colspan="1" style="border-bottom: 1px solid black"></td></tr>  
## <tr><td style="text-align:left"></td><td>log10(body\_wt)</td></tr>  
## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left">gestation</td><td>0.007<sup>\*\*\*</sup></td></tr>  
## <tr><td style="text-align:left"></td><td>(0.006, 0.009)</td></tr>  
## <tr><td style="text-align:left"></td><td></td></tr>  
## <tr><td style="text-align:left">Constant</td><td>-0.404<sup>\*\*</sup></td></tr>  
## <tr><td style="text-align:left"></td><td>(-0.723, -0.086)</td></tr>  
## <tr><td style="text-align:left"></td><td></td></tr>  
## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left">Observations</td><td>58</td></tr>  
## <tr><td style="text-align:left">R<sup>2</sup></td><td>0.589</td></tr>  
## <tr><td style="text-align:left">Adjusted R<sup>2</sup></td><td>0.582</td></tr>  
## <tr><td style="text-align:left">Residual Std. Error</td><td>0.884 (df = 56)</td></tr>  
## <tr><td style="text-align:left">F Statistic</td><td>80.243<sup>\*\*\*</sup> (df = 1; 56)</td></tr>  
## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"><em>Note:</em></td><td style="text-align:right"><sup>\*</sup>p<0.1; <sup>\*\*</sup>p<0.05; <sup>\*\*\*</sup>p<0.01</td></tr>  
## </table>