STA 210: HW 4

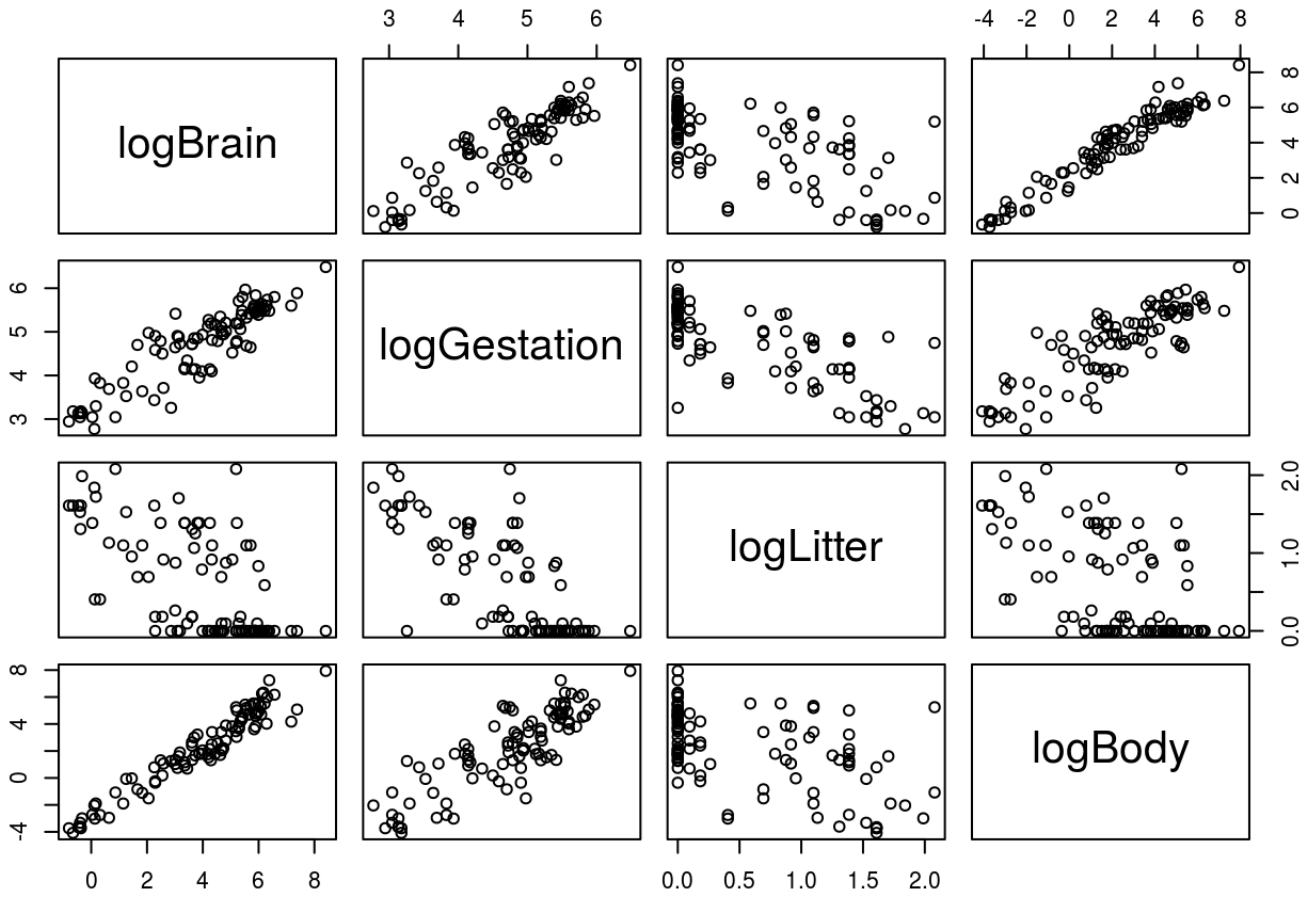
Jackson Hubbard

October 1, 2018

Question 1: Ex. 9.12

a

```
#use case0902 data set
q1 <- as_data_frame(case0902)
q1 <- q1 %>% mutate ( logBrain = log(q1$Brain), logBody = log(q1$Body), logGestation
= log(q1$Gestation), logLitter = log(q1$Litter))


pairs(logBrain ~ logGestation + logLitter + logBody, data = q1)
```



b

```
model1 <- lm(logBrain ~ logBody + logGestation + logLitter, data = q1)
tidy(model1)
```

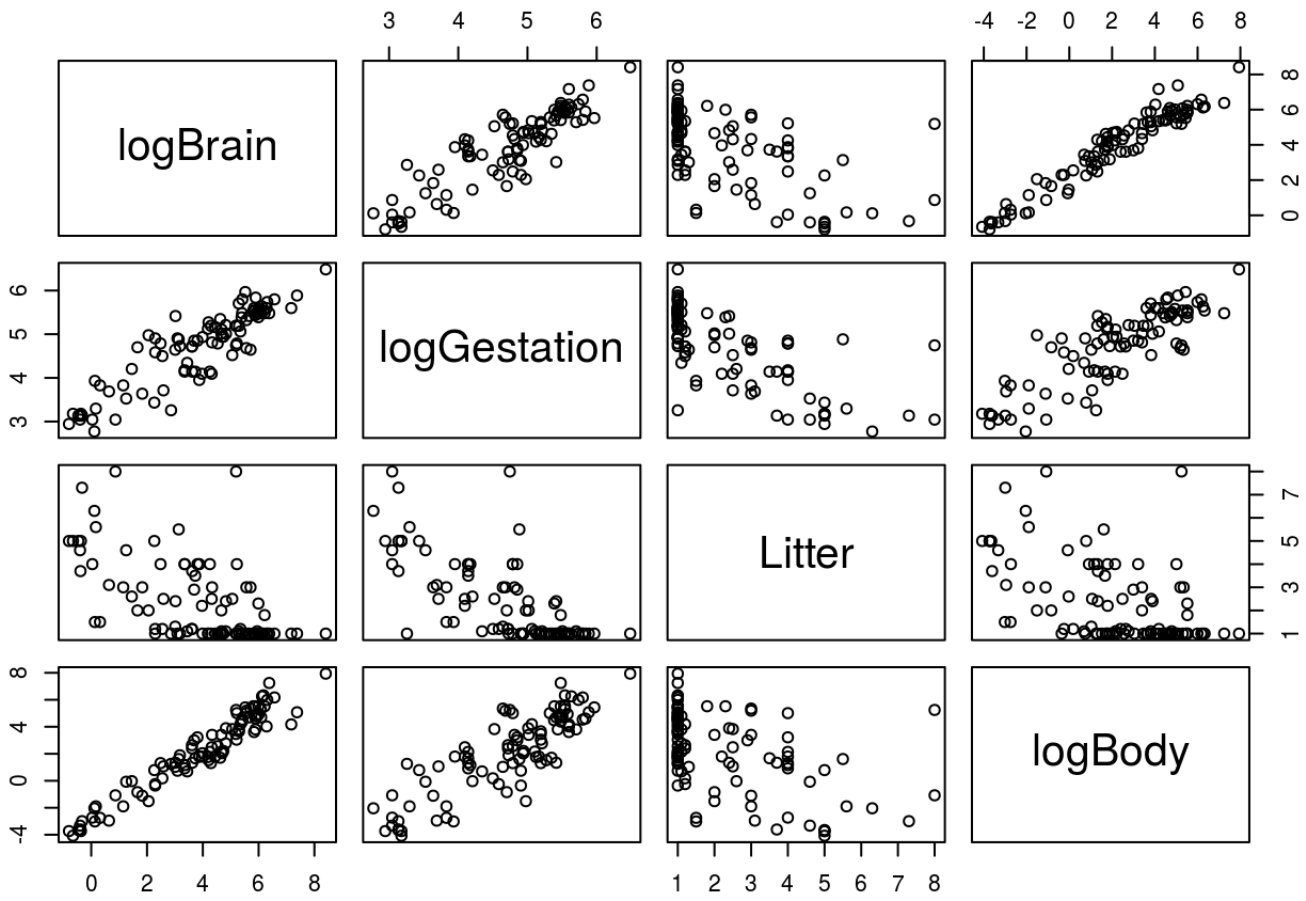
term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	0.8548219	0.66167247	1.291911	1.996239e-01
logBody	0.5750714	0.03258789	17.646784	2.777551e-31
logGestation	0.4179421	0.14078249	2.968708	3.812593e-03
logLitter	-0.3100712	0.11592709	-2.674708	8.852076e-03

4 rows

Looking at the outputttted table, we see that this is the same table as the textbook

C

```
pairs(logBrain ~ logGestation + Litter + logBody, data = q1)
```

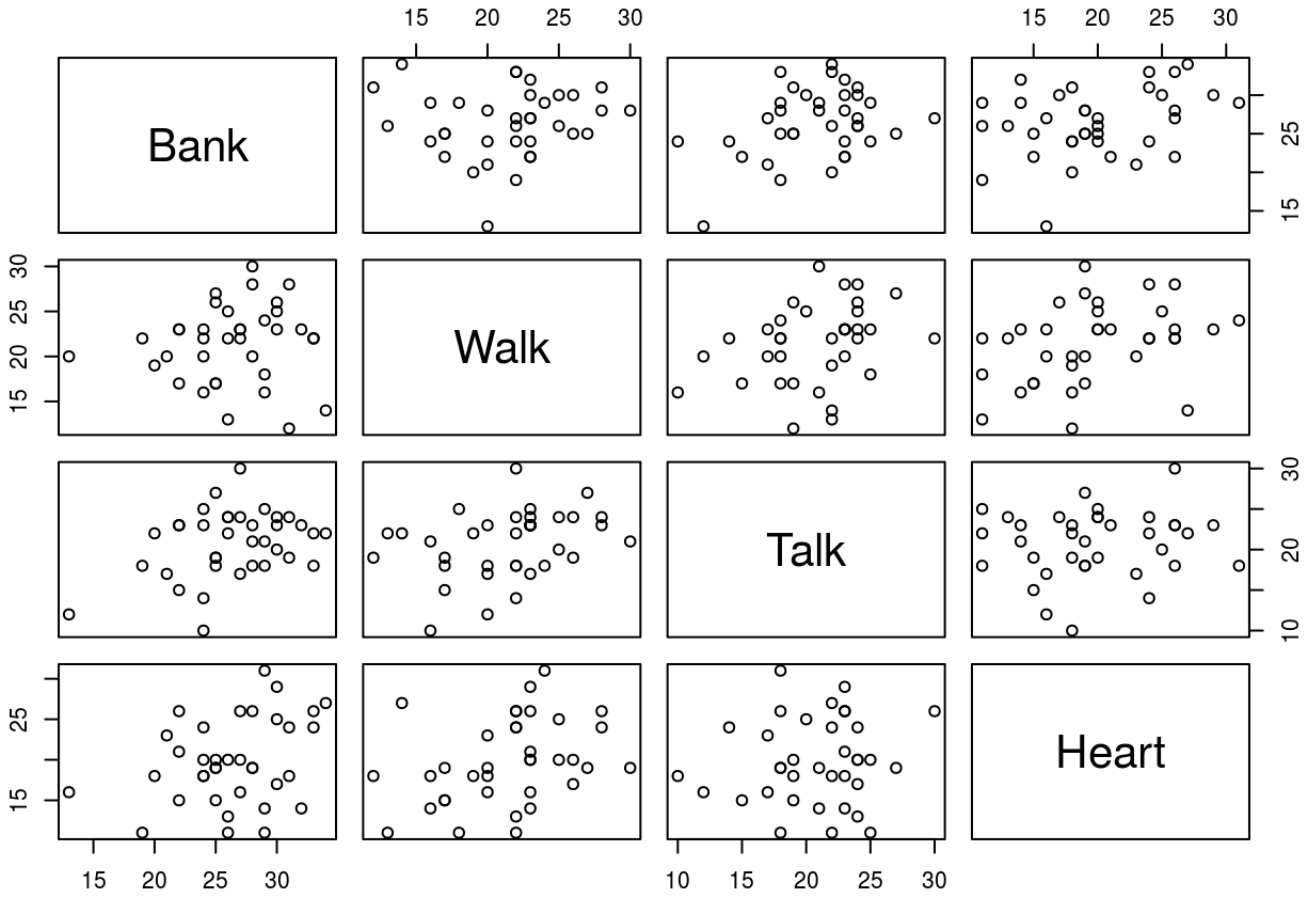


No the relationship between litter size and log(brain weight) does not appear to be any better than the relationship between log(litter size) and log(brain weight). So there is no need to use the logged trasnformed variable for Litter.

Question 2: Ex 9.14

a

```
#use ex0914 data set

q2 <- ex0914

pairs(Bank ~ Walk + Talk + Heart, data = q2)
```



b

```
model2 <- lm(Heart ~ Bank + Walk + Talk, data = q2)
tidy(model2)
```

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	3.1786957	6.3369459	0.5016132	0.61937341
Bank	0.4052170	0.1971021	2.0558738	0.04802911
Walk	0.4516011	0.2008735	2.2481862	0.03158393
Talk	-0.1796096	0.2222154	-0.8082681	0.42490460

4 rows

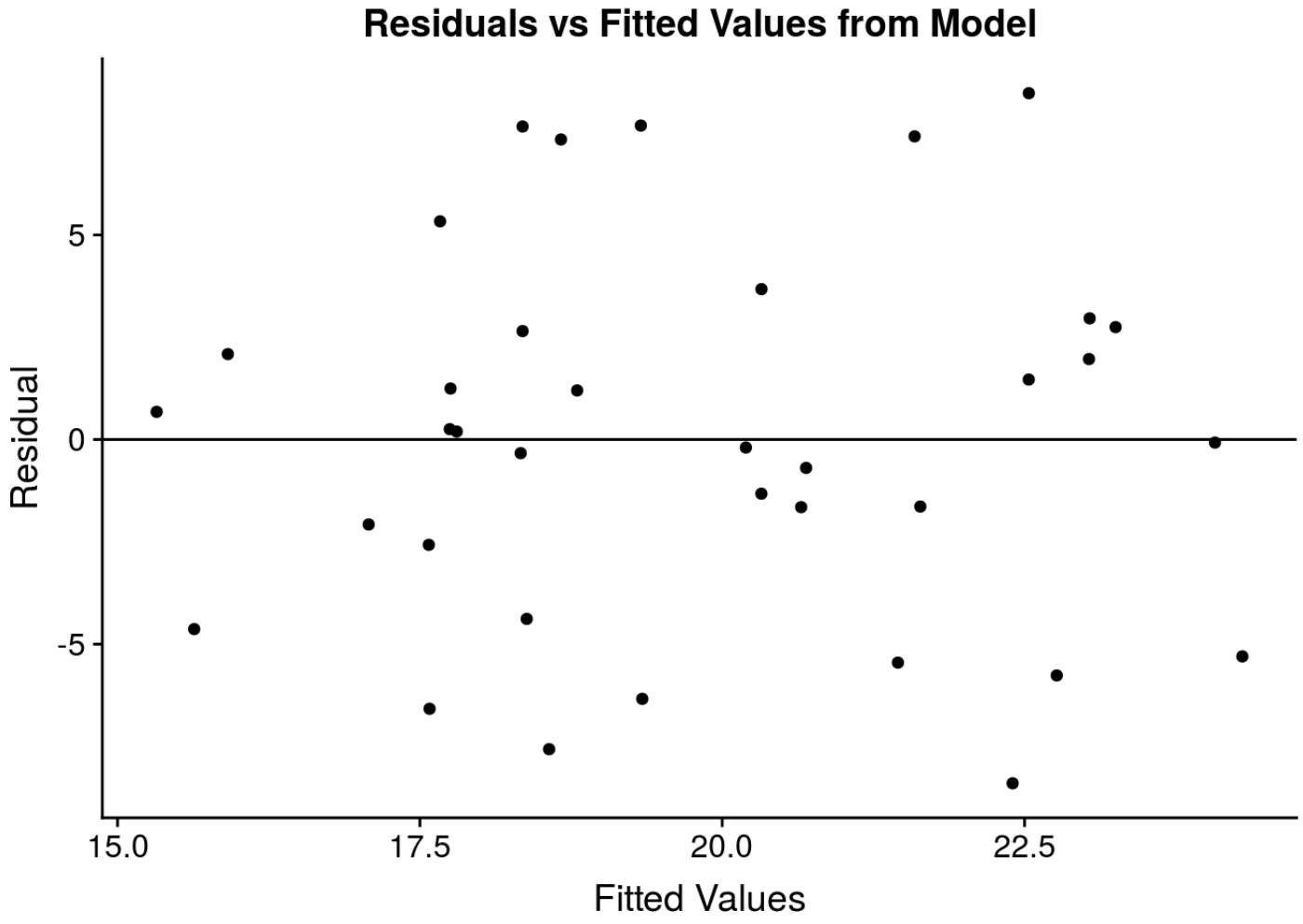
The least squares fit to the linear regression of heart on bank, walk, and talk is:

μ{Heart| Bank, Walk, Talk} = 3.1786957 + 0.4052170 * Bank + 0.4516011 * Walk -0.1796096 * Talk

c

plot the residuals vs the fitted values. Is there evidence that the variance of residuals increases with increasing fitted values or that there are any outliers?

```
q2 <- q2 %>% mutate(residuals = resid(model2), fitted = predict.lm(model2))

ggplot(q2, aes(x = fitted, y= residuals)) +geom_point() + geom_hline(yintercept = 0)
+
  labs(title = "Residuals vs Fitted Values from Model", x = "Fitted Values",
       y= "Residual") + theme(plot.title = element_text(hjust = 0.5, size = 14))
```



looking at the scatterplot of the fitted values vs the residuals, we see that the residuals do not increase with increasing fitted values or that there is any other pattern. this means that the asumption of constant variance no matter what the fitted values are is met. We also see that there are no distinct outliers.

d

Looking at the table from part b, we see that the p values for Bank and for Walk are small (lower than 0.05), which means that they are significant. However, the p value for talk is very high, which means that it is not significant predictor of Heart and should not be included in the model.

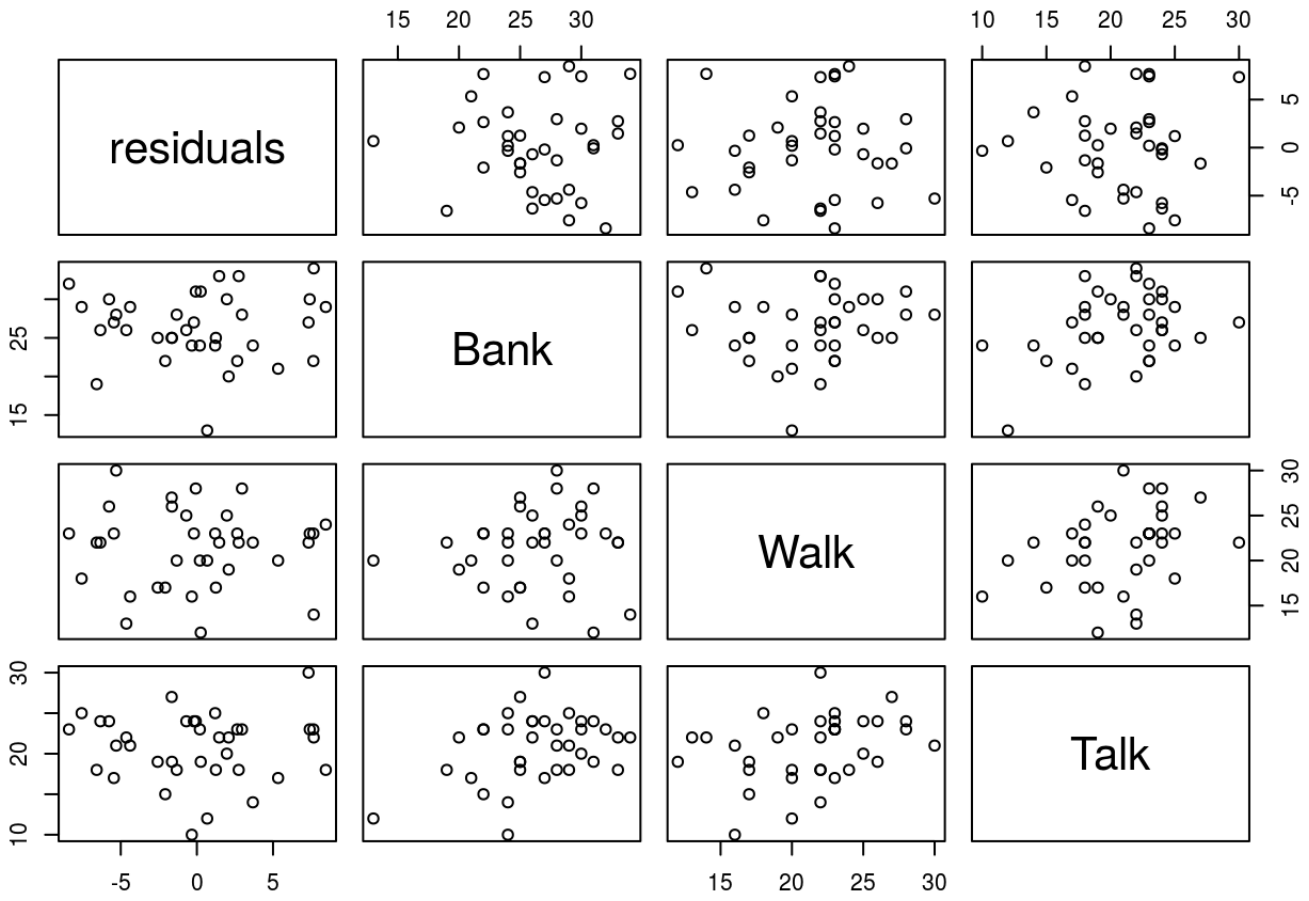
The least squares fit to the linear regression of heart on bank, walk, and talk is:

μ{Heart| Bank, Walk, Talk} = 3.1786957 + 0.4052170 * Bank + 0.4516011 * Walk - 0.1796096 * Talk

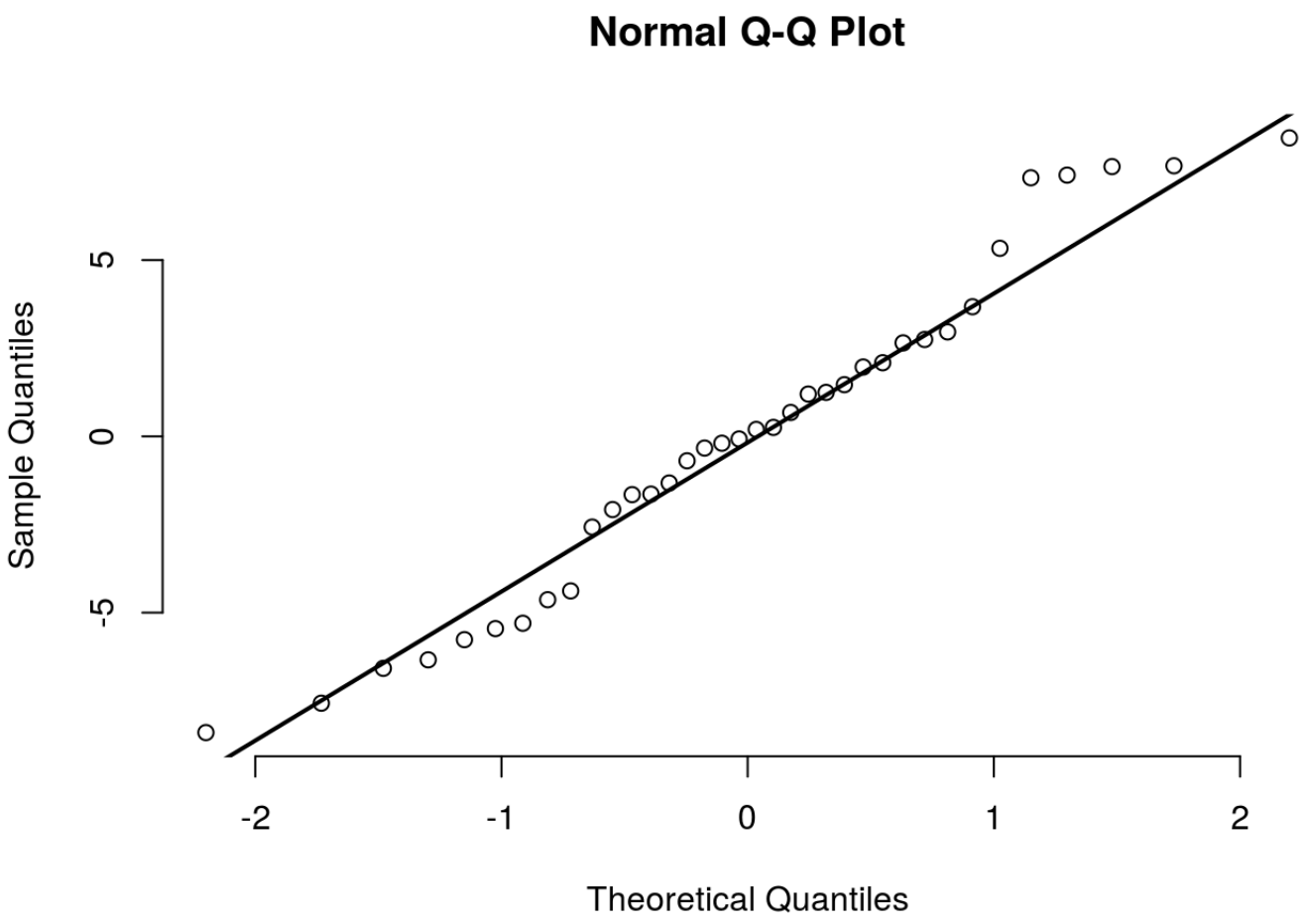
The standard errors are:
Bank- 0.1971021
Walk- 0.2008735
Talk- 0.2222154

Checking assumptions:

```
pairs(residuals ~ Bank + Walk + Talk, data= q2)
```




```
qqnorm(q2$residuals, pch=1, frame= FALSE)
qqline(q2$residuals, lwd=2)
```



From the scatterplot matrix from part a, we see that the linearity assumption is likely not met since there is a very weak (if any) relationship between the response variable (heart) and the explanatory variables. this is especially the case for the Talk variable. The assumption of constant variance is met as there is no pattern for the residuals based on the values of each explanatory variable. Further, the normality assumption is appears to be met as we look at the QQ plot. There are a couple deviations from the line (especially in the top right of the line) but these do not appear to be strong enough departures to say that normality is not met.

Question 3: Ex. 9.20

a

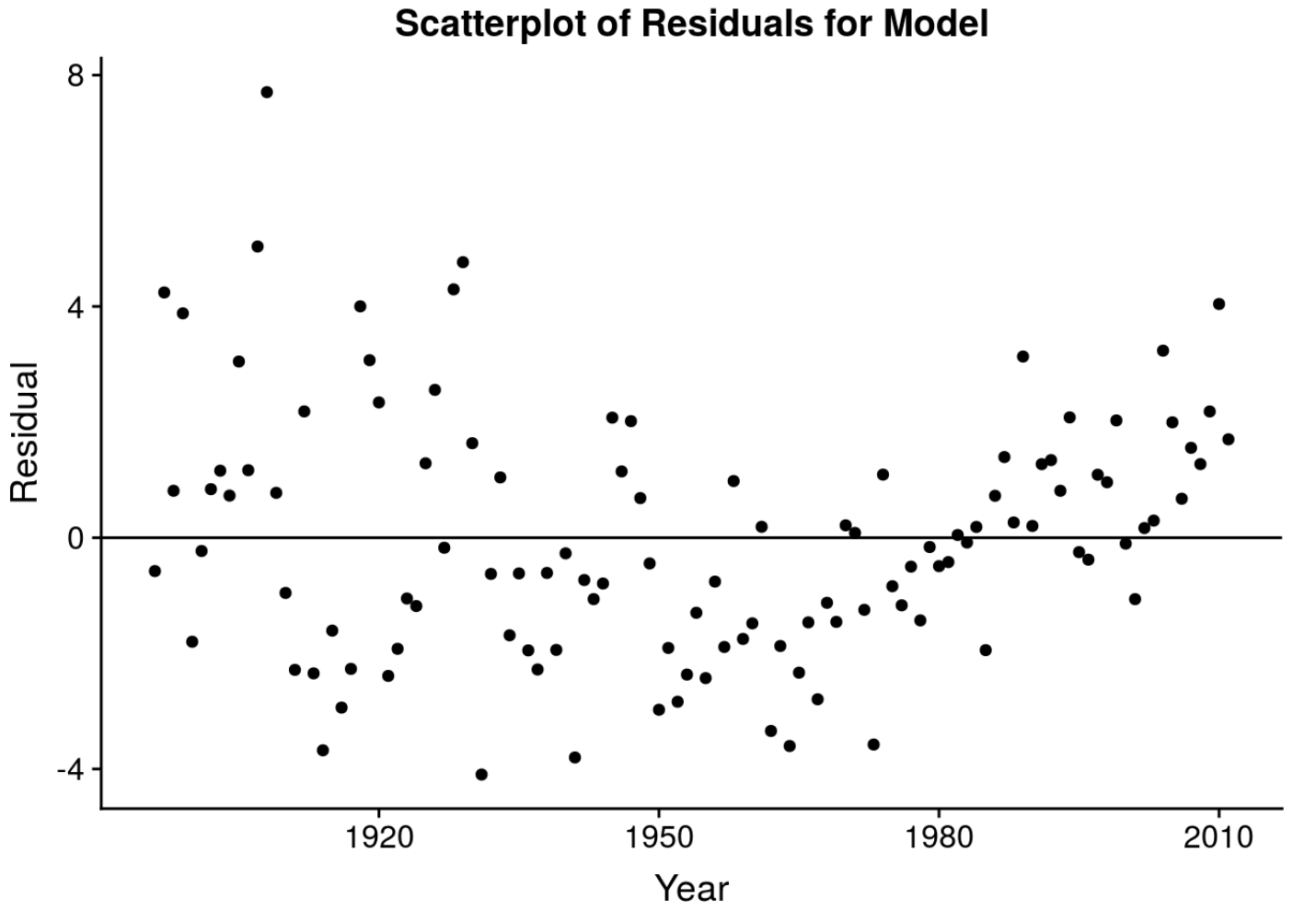
```
#use ex0920 data set

q3 <- ex0920

model_year1 <- lm(Time  ~ Year , data = q3)
tidy(model_year1)
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>

(Intercept)	260.05003571	11.627325519	22.36542	1.681507e-43
Year	-0.06947369	0.005951174	-11.67395	3.424810e-21

2 rows

```
q3 <- q3 %>% mutate(resid0 = resid(model_year1))

q3 %>% ggplot(aes(x= Year, y= resid0)) +geom_point() + labs (title= "Scatterplot of R
esiduals for Model", x= "Year", y= "Residual") + theme(plot.title = element_text(hjus
t = 0.5, size = 14)) + geom_hline(yintercept = 0)
```



Looking at this scaterplot, we see that the residuals have a pattern. From the shape of the graph we can conclude that we likely need a quadratic transformation on the year.

```
q3 <- q3 %>% transform(q3, year2 = Year*Year)
model_year <- lm(Time  ~ Year +year2, data = q3)
tidy(model_year)
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>

(Intercept)	3.989352e+03	6.759480e+02	5.901862	3.835770e-08
Year	-3.888668e+00	6.921802e-01	-5.617999	1.406551e-07
year2	9.775260e-04	1.771589e-04	5.517792	2.208006e-07

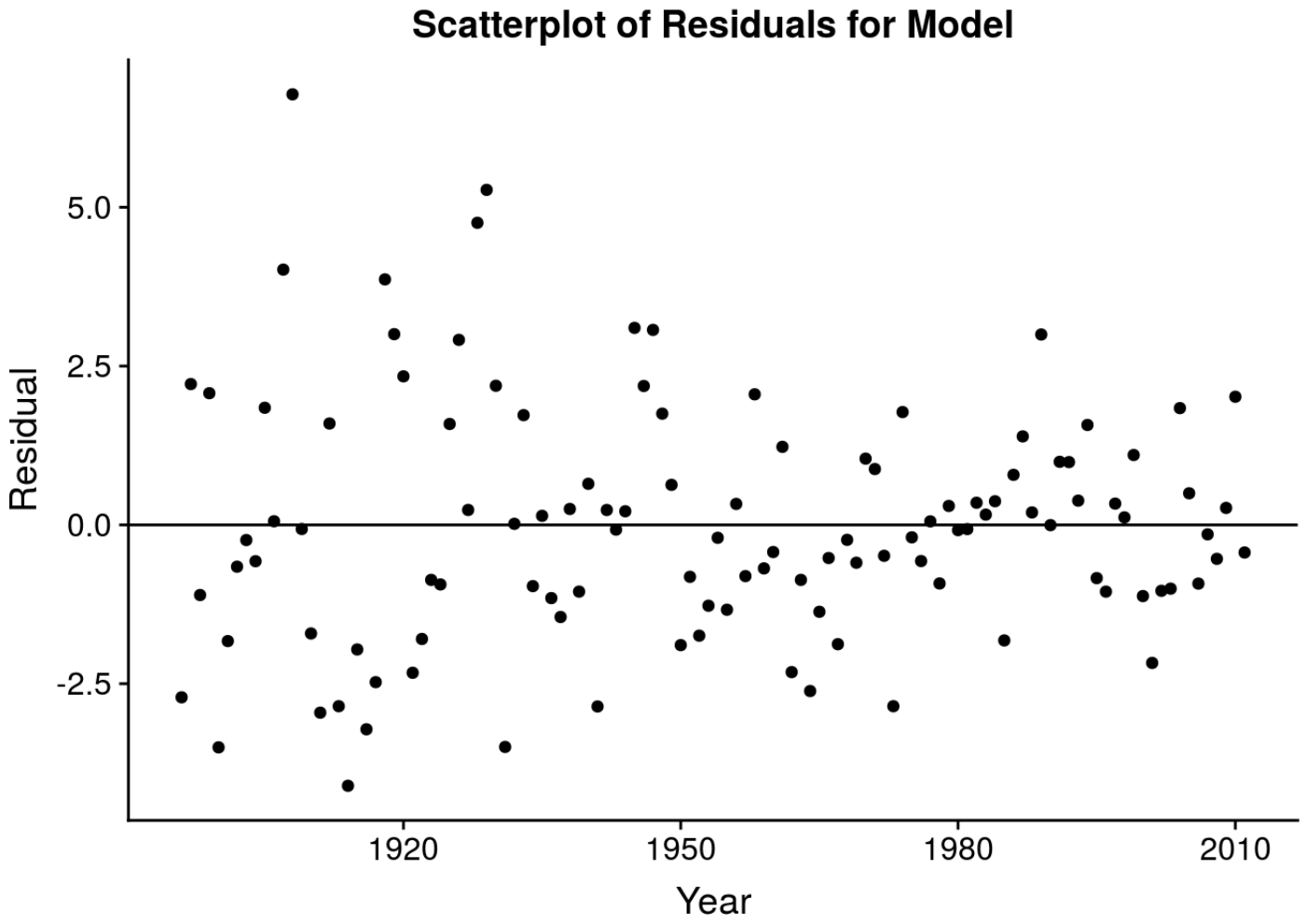
3 rows

```
glance(model_year)
```

r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>	p.value <dbl>	df <int>	logLik <dbl>	AIC <dbl>	
1 0.6411881	0.6348375	1.913335	100.9641	7.079054e-26	3	-238.3435	484.6869	495

1 row | 1-10 of 12 columns

```
q3 <- q3 %>% mutate(resid4 = resid(model_year))

q3 %>% ggplot(aes(x= Year, y= resid4)) +geom_point() + labs (title= "Scatterplot of R
esiduals for Model", x= "Year", y= "Residual") + theme(plot.title = element_text(hjus
t = 0.5, size = 14)) + geom_hline(yintercept = 0)
```



Looking at the new scatterplot, we see that the curvature has improved. However, there still is a funnel effect as the residuals beceome less varied as the year increases. From this we can conclude that the constant variance assumption was likely broken.

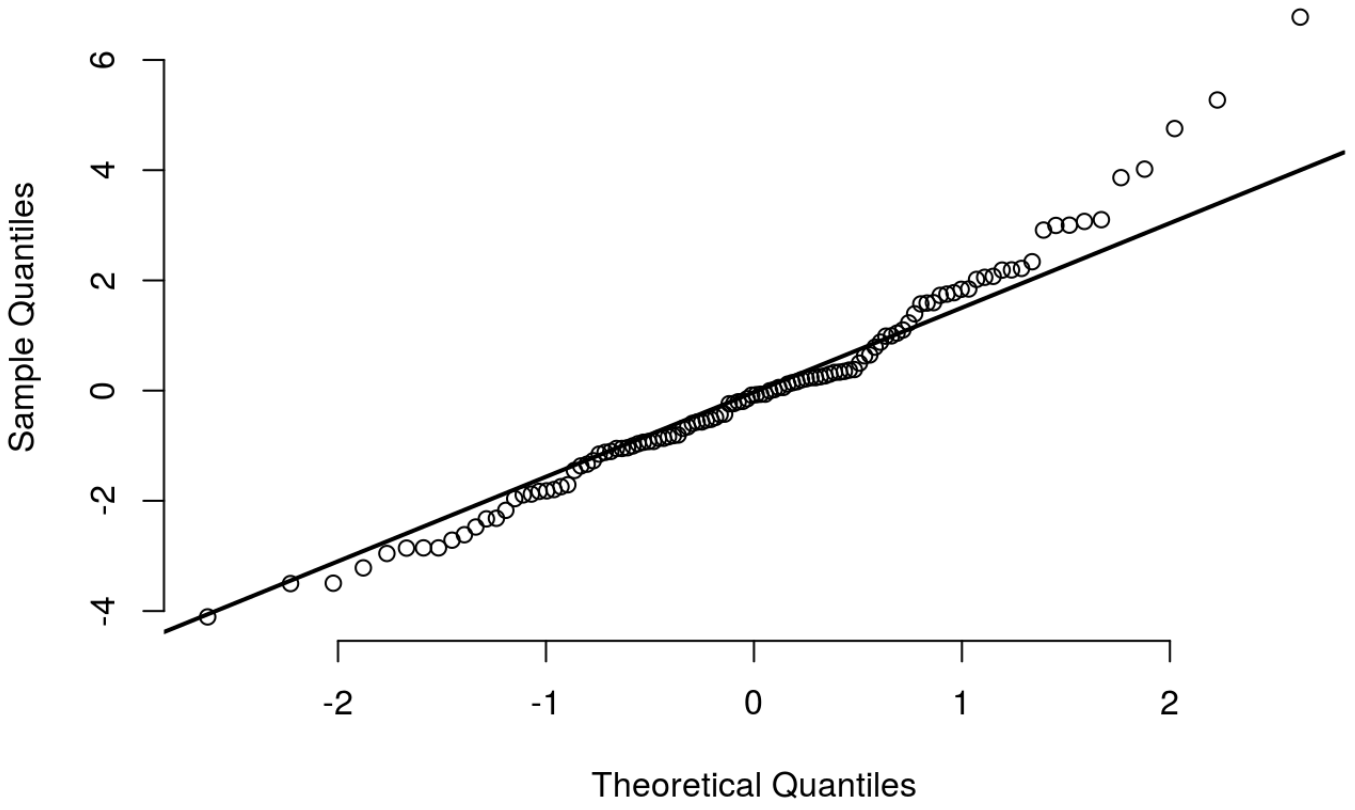
The outputted model is:

μ{Time|Year} = 3989.352 - 3.888668 * Year + 97752.60 * $Year^2$

All of the explanatory variables' p values are low (less than 0.05) which means that they are all significant predictors.

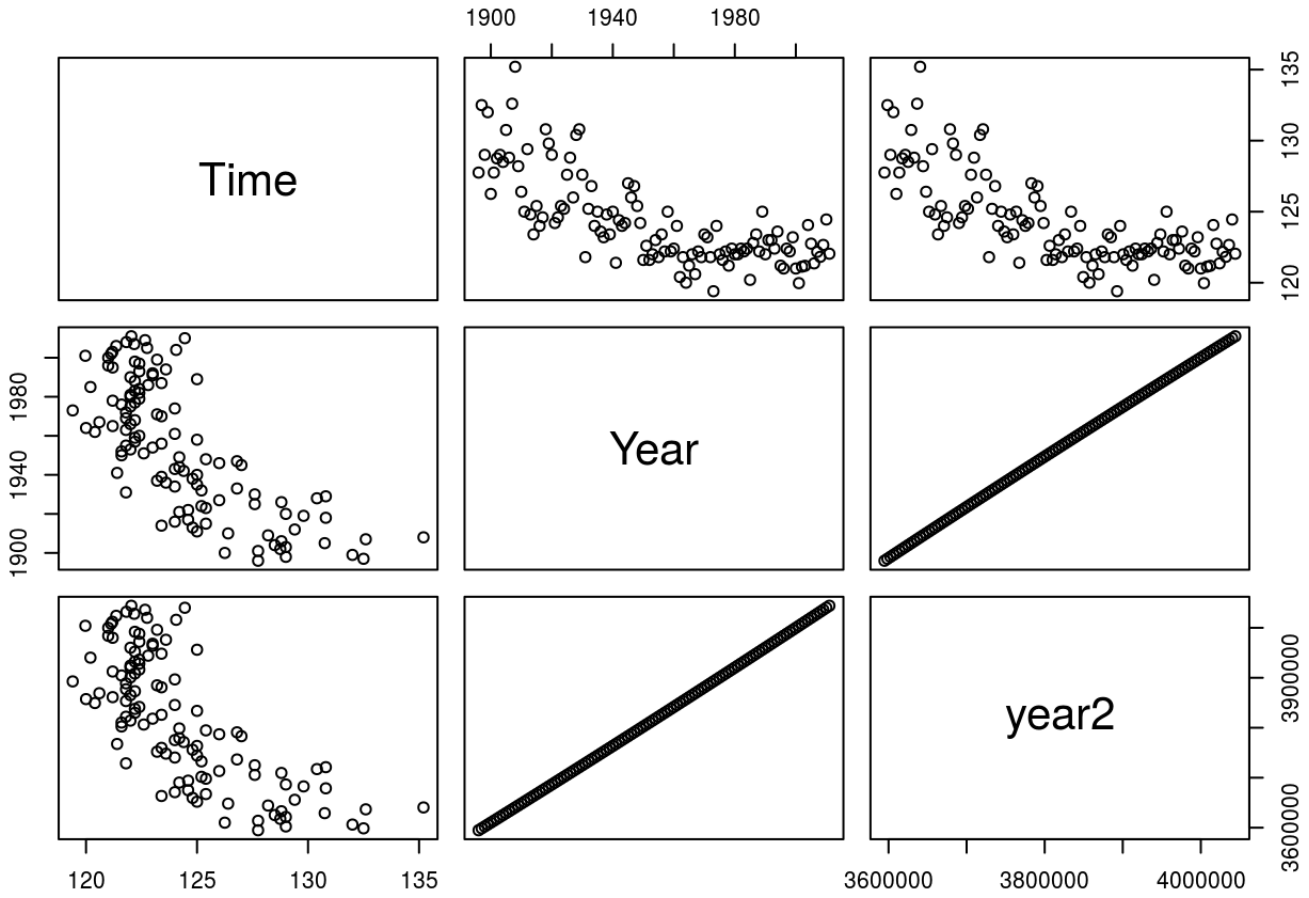
Checking assumptions: Normality:

```
qqnorm(q3$resid4, pch=1, frame= FALSE)
qqline(q3$resid4, lwd=2)
```



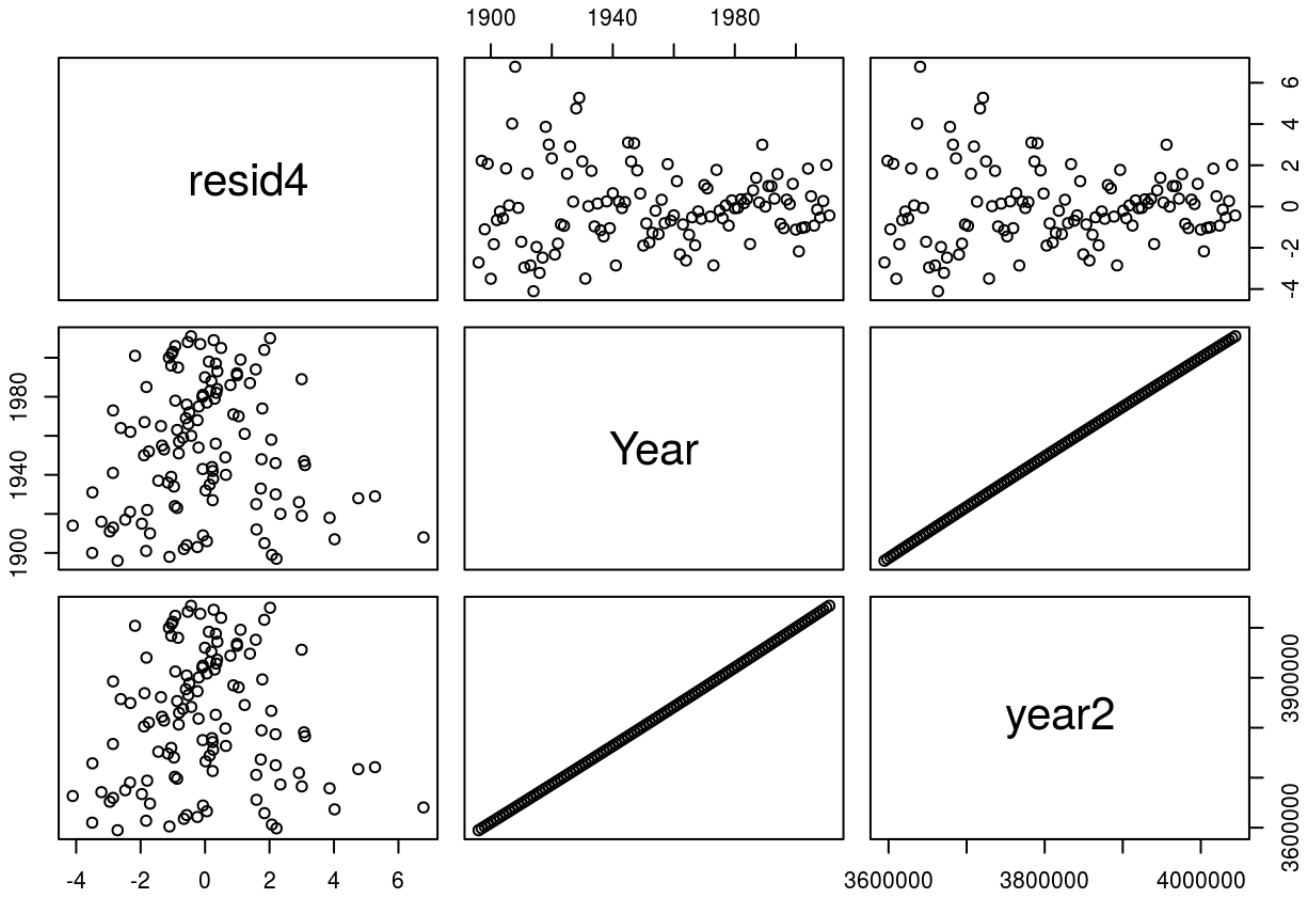
Linearity:

```
pairs(Time ~ Year + year2, data= q3)
```



constant variance:

```
pairs(resid4 ~ Year + year2, data= q3)
```



Looking at the output of these graphs we see that the Normality assumption holds. The top right of the graph with the upper quantiles deviate from the line a little bit but it is likely not enought to say that the normality assumption is not met.interpreting the next graph shows that linearity assumption is still pribably not met and looking at the resdiuals graph we see that there is still not constant variance. Due to the horn shape, we probably need another transformation. For the independence assumption we need to be caredul beacuse the data was collected over time, so a time series model may be needed.

b

Quantify the amount by which the winning time on fast tracks exceeds the winning time on slow tracks

```
q3 <- q3 %>% mutate(conditionSlow = case_when(Conditions == "Fast" ~ 0, Conditions ==
"Slow" ~ 1))

modeltime2cond <- lm(Time ~ Year + year2 + conditionSlow, data = q3 )
tidy(modeltime2cond)
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	3.745853e+03	4.930026e+02	7.598038	9.843159e-12
Year	-3.650410e+00	5.047885e-01	-7.231562	6.284950e-11
year2	9.192196e-04	1.291850e-04	7.115527	1.123464e-10
conditionSlow	3.500250e+00	3.483913e-01	10.046893	2.559317e-17

4 rows

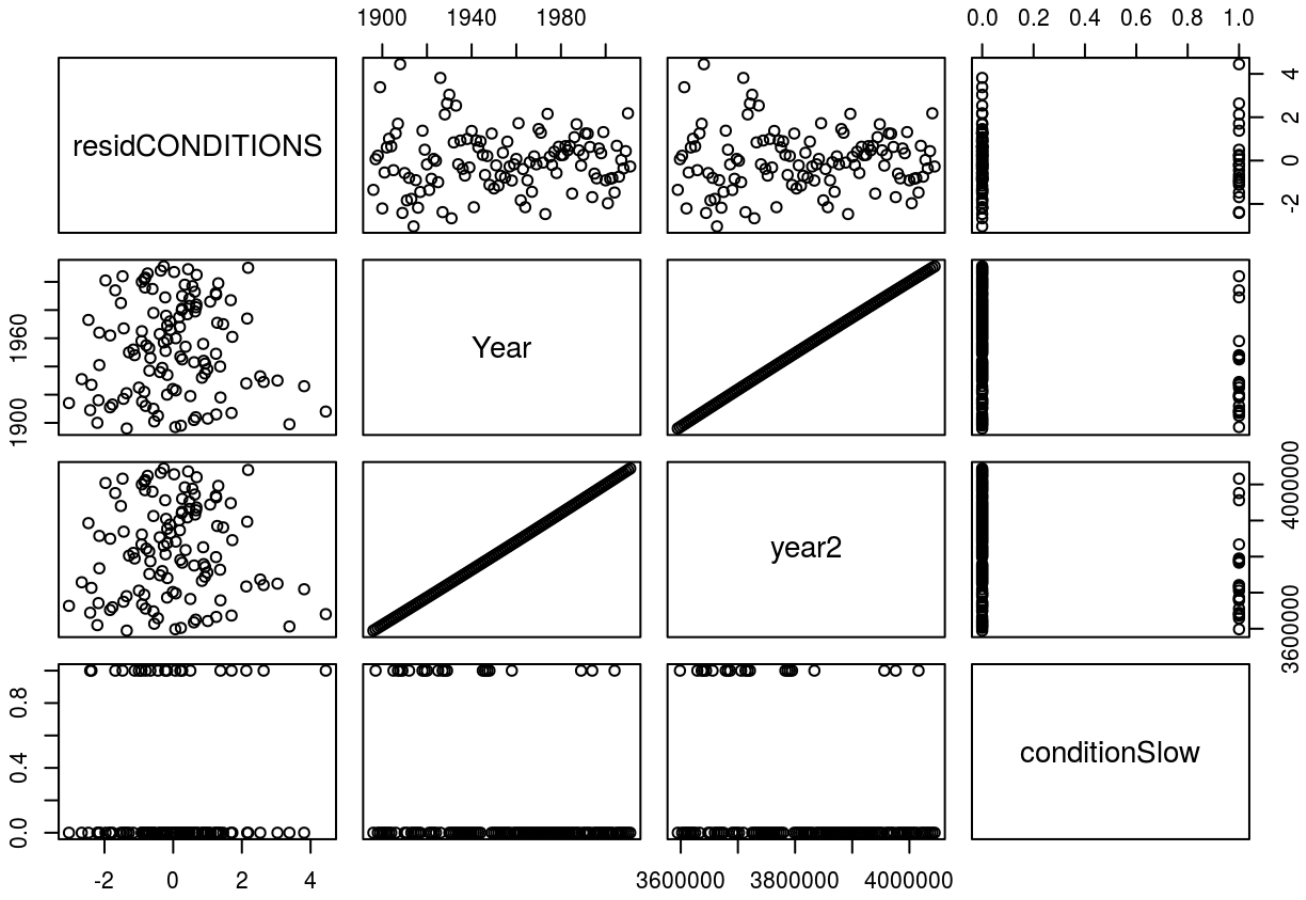
Looking at the outputted table from this new model, we see that the mean time on slow tracks is greater than the mean time on fast tracks by 3.5 seconds after accounting for the year (holding it constant).

The outputted model is:

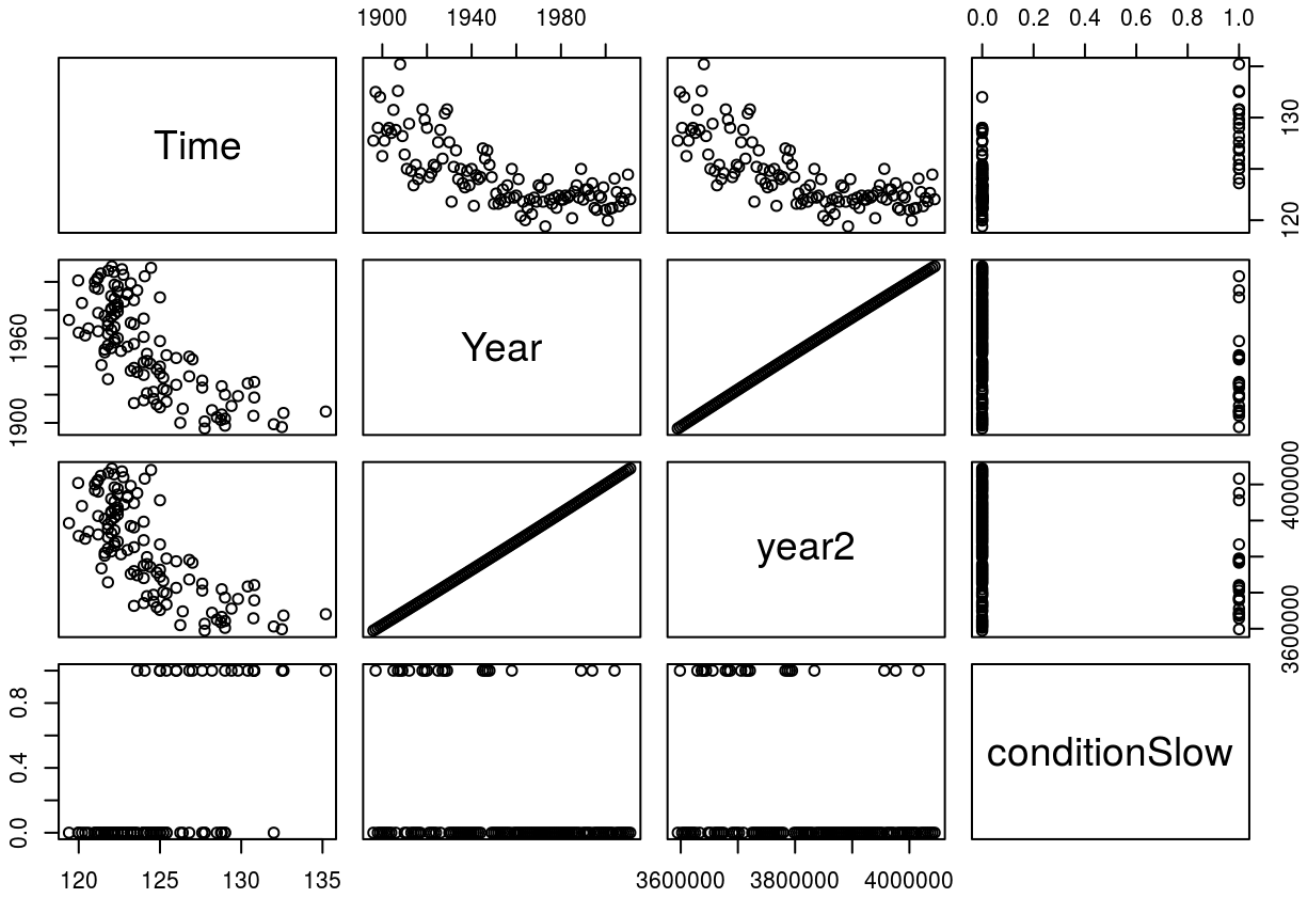
μ{Time|Year, ConditionSlow} = 3745.853 - 3.650410 * Year + 91921.96 * $Year^2$ + 3.500250 * ConditionSlow

We built a new model so we must check assumptions again.

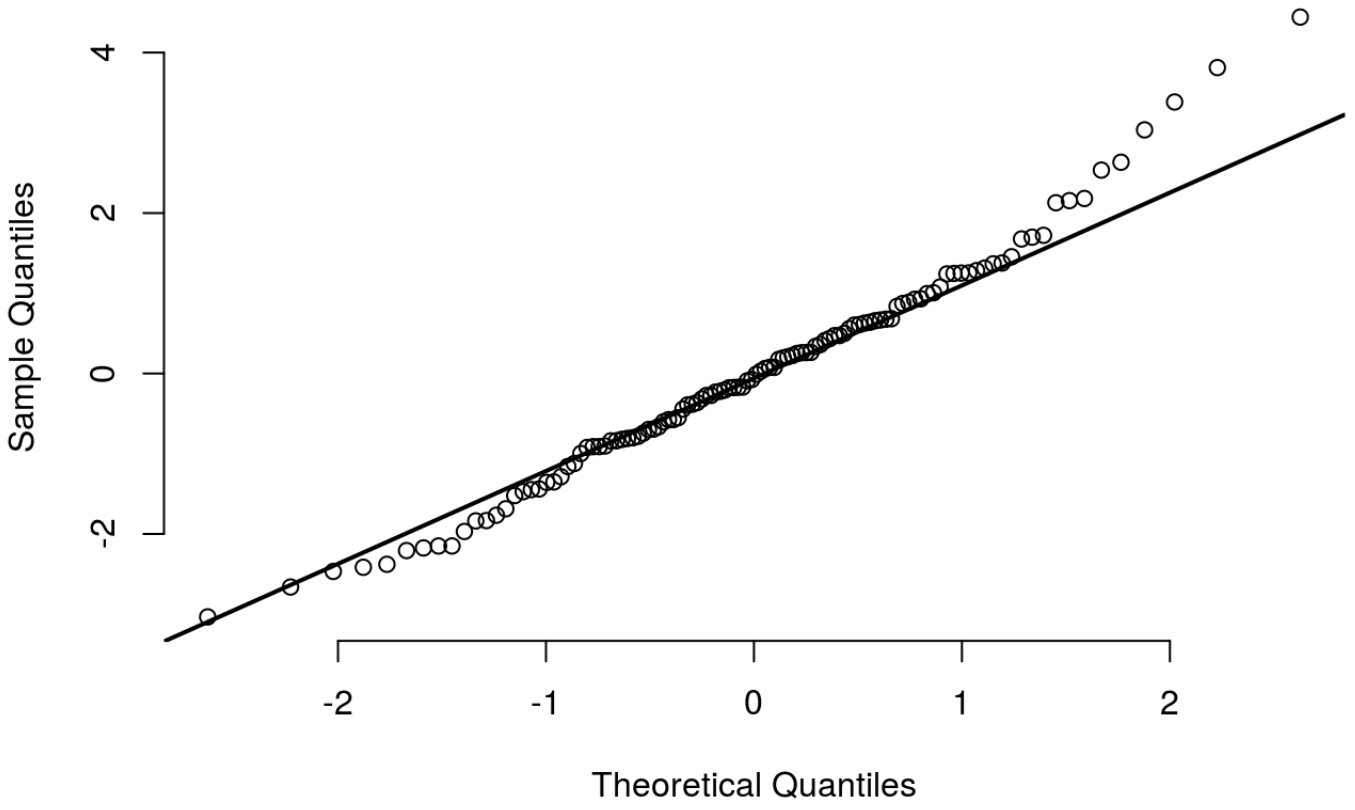
```
q3 <- q3 %>% mutate(residCONDITIONS = resid(modeltime2cond))

pairs(residCONDITIONS ~ Year + year2 + conditionSlow, data=q3)
```



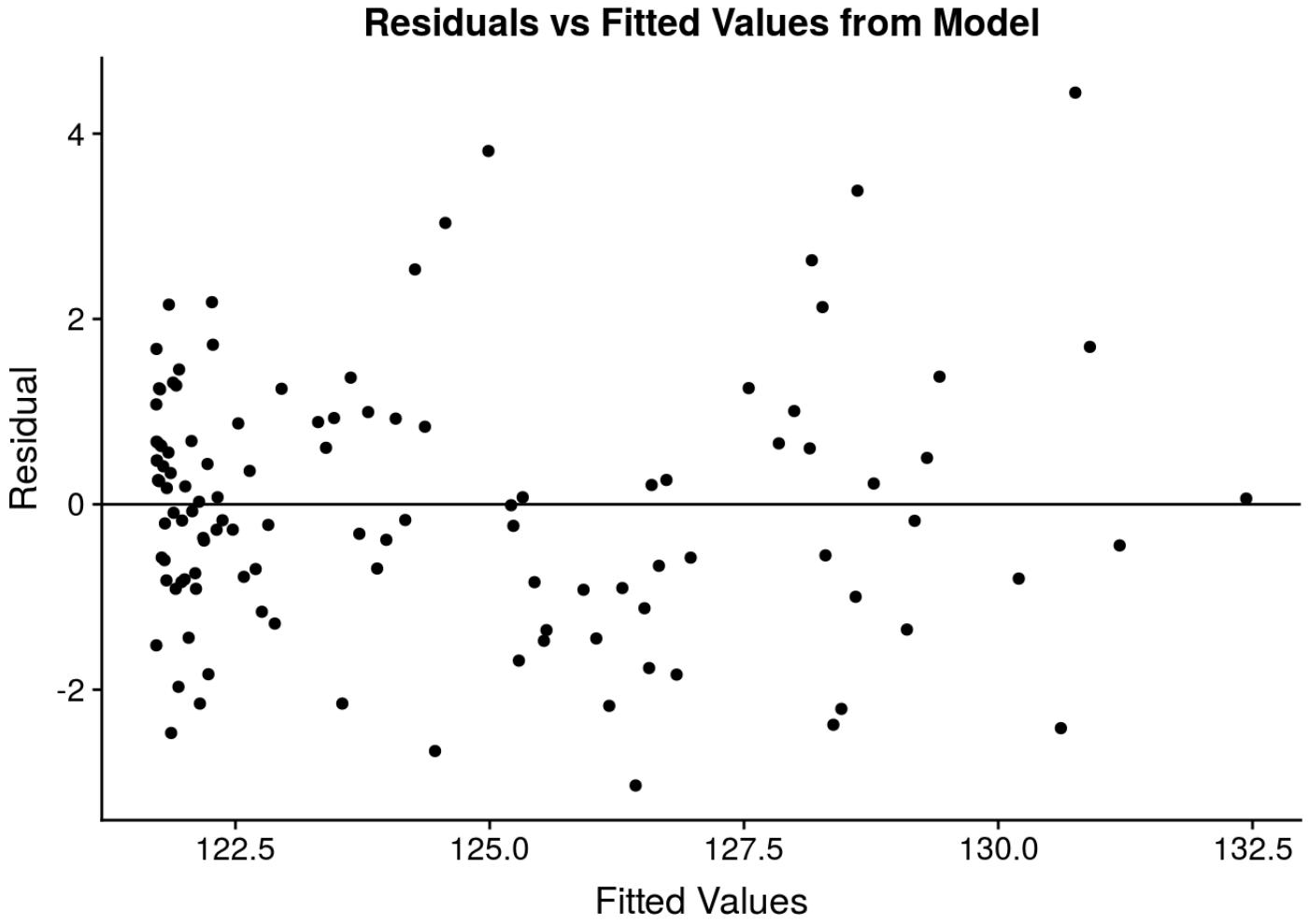

```
pairs(Time ~ Year + year2 + conditionSlow, data= q3)
```




```
qqnorm(q3$residCONDITIONS, pch=1, frame= FALSE)
qqline(q3$residCONDITIONS, lwd=2)
```




```
q3 <- q3 %>% mutate(prediction = predict(modeltime2cond))
ggplot(data= q3, aes(x=prediction, y= residCONDITIONS)) + geom_point() + labs(title =
"Residuals vs Fitted Values from Model", x = "Fitted Values",
       y= "Residual") + theme(plot.title = element_text(hjust = 0.5, size = 14)) + ge
om_hline(yintercept = 0)
```



Looking at these graphs, we see that the Normality assumption is still met (with the same discussion about the upper quantiles). The residual graphs have constant variance and no pattern so the constant variance assumption is met. The linearity assumption is also met. Just as discussed before, the independence assumption is likely met as well as long as we check for the time series. Further, as we look at the residuals vs the fitted values, we can confirm our conclusions about the constant variance and linearity.

c

after accounting for effects of year and track conditons is there any evidence that the mean time depends on the number of horses in race (Starters)?

```
model_numRace <- lm(Time ~ Year + year2 + Starters + conditionSlow, data = q3 )
tidy(model_numRace)
```

term	estimate	std.error	statistic	p.value

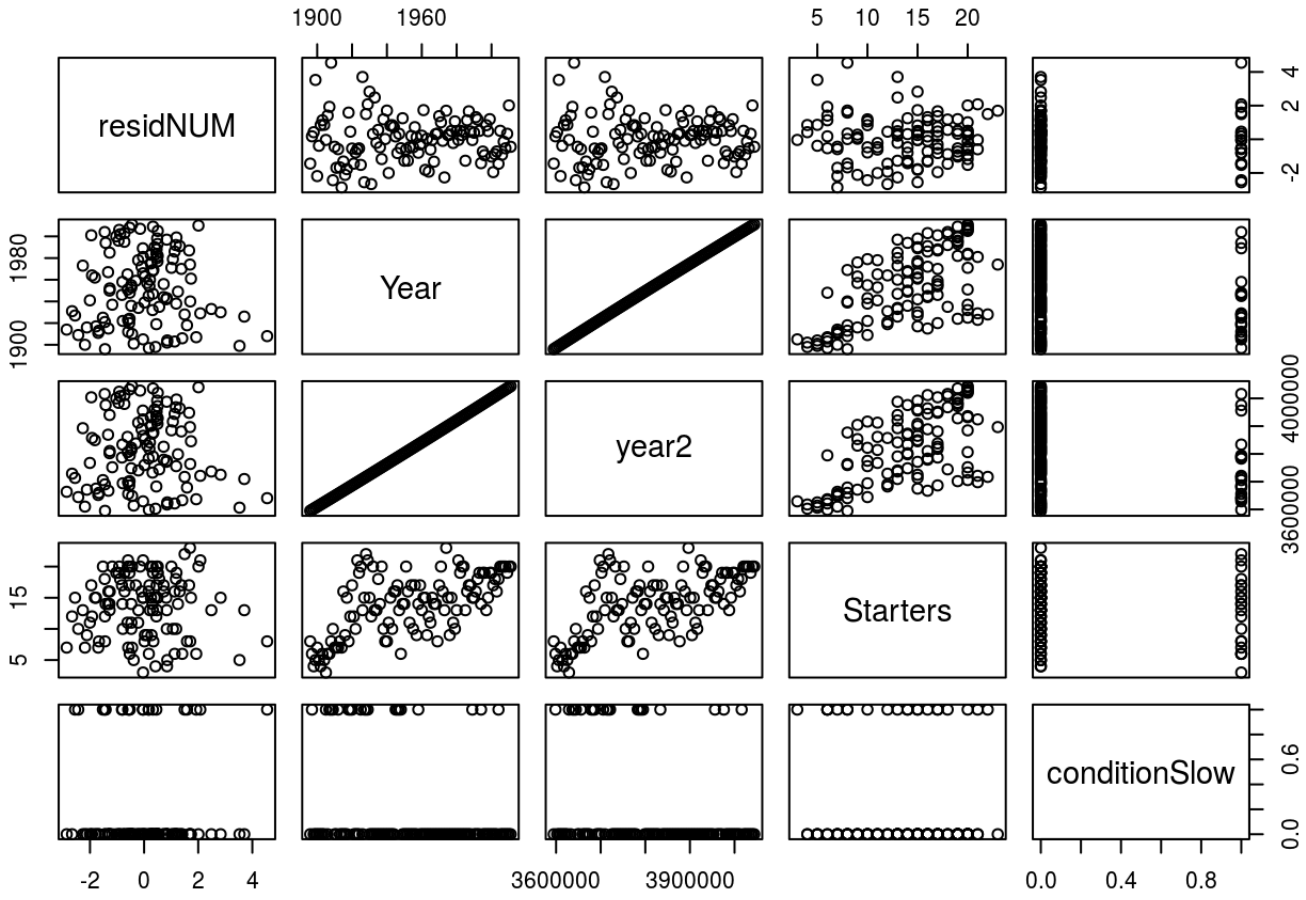
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	3.947904e+03	4.979012e+02	7.929091	1.885979e-12
Year	-3.851999e+00	5.092449e-01	-7.564140	1.216380e-11
year2	9.692148e-04	1.301614e-04	7.446253	2.209374e-11
Starters	6.655570e-02	3.417798e-02	1.947327	5.402151e-02
conditionSlow	3.439134e+00	3.455564e-01	9.952453	4.609892e-17

5 rows

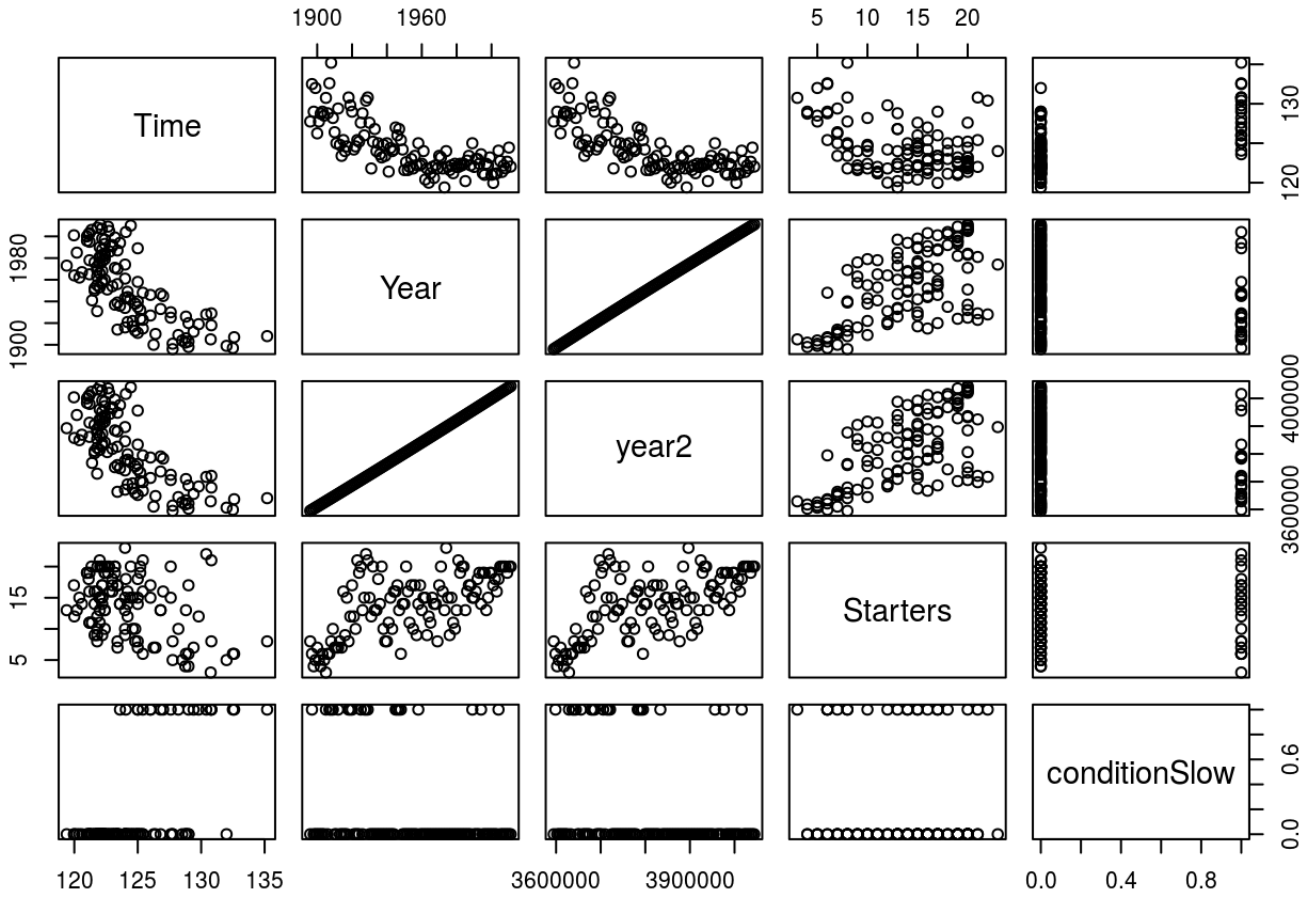
Yes there is evidence that the mean time depends on the number of horses in the race. This is beacuse looking at the output from the model, we see that Starters has an estimated slope of 0.06655570. This slope has a p-value of 0.05402151, which is greater than 0.05 which technically means that it is not significant, but it is so close to 0.05 that it is practically significant so we will keep it in the model. Thus, after accounting for effects of the year and the track conditons, there is a (barely) significant effect of the number of horses in the race on the mean time. For every additional horse that is added to the race, the mean winning time increases by 0.06655570 holding all other variables constant.

Checking assumptions:

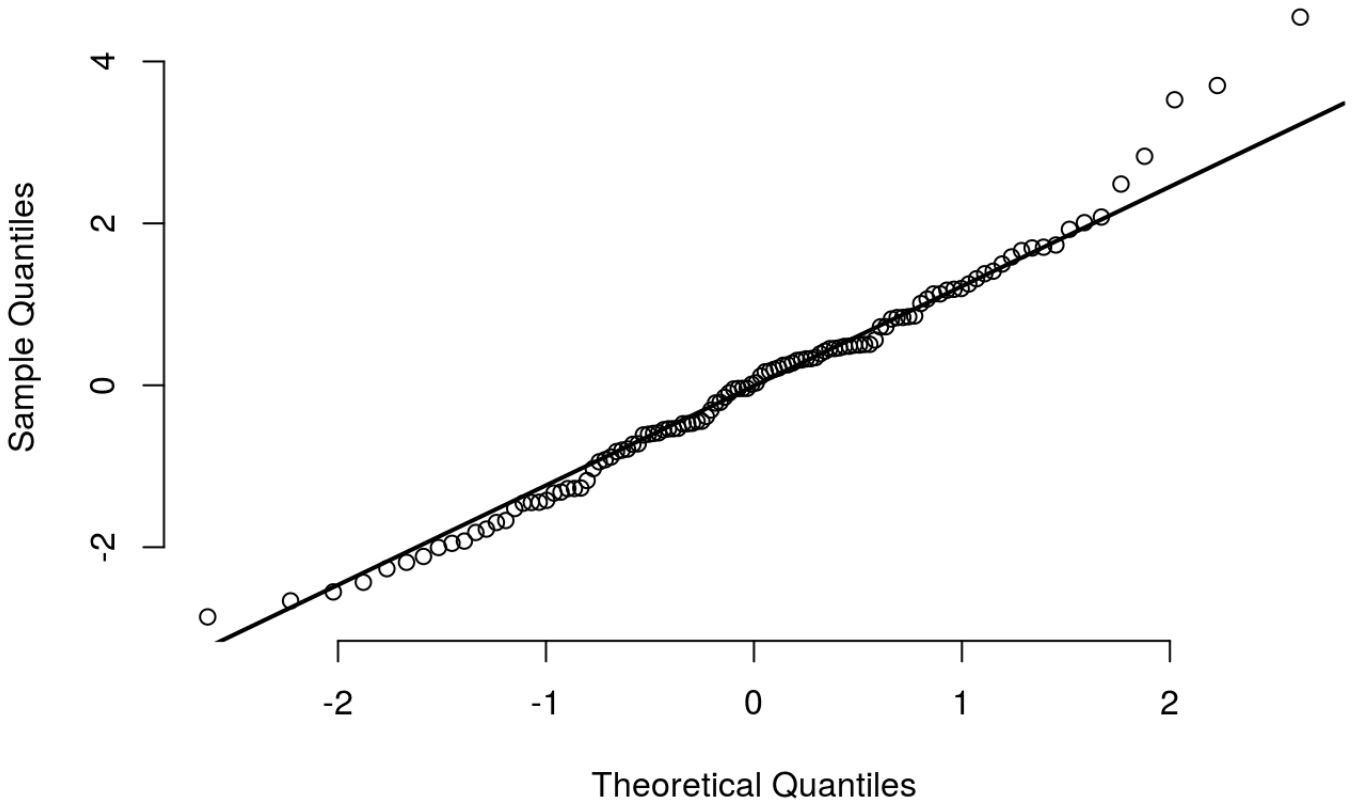
```
q3 <- q3 %>% mutate(residNUM = resid(model_numRace))

pairs(residNUM ~ Year + year2 + Starters + conditionSlow, data=q3)
```



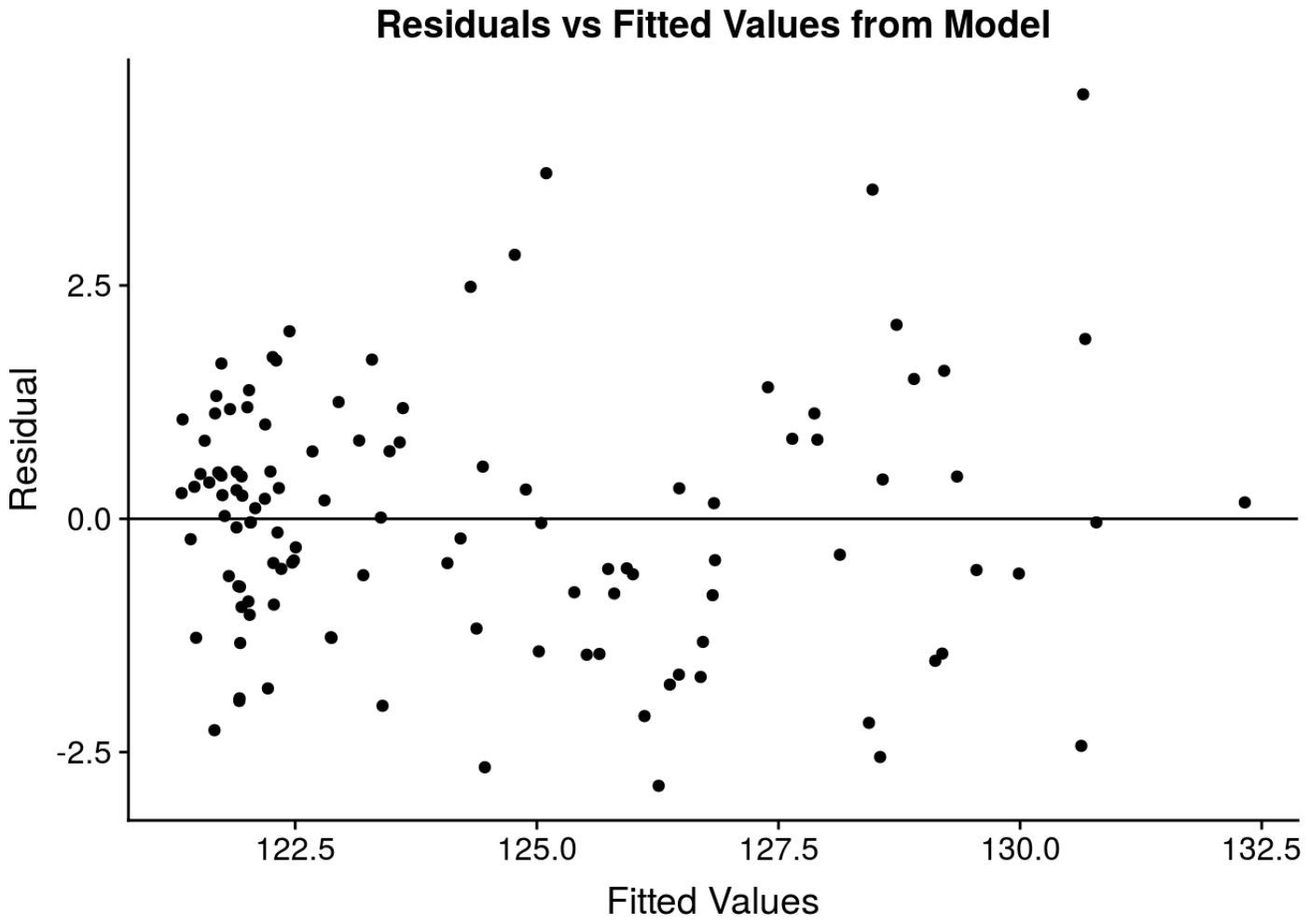

```
pairs(Time ~ Year + year2 + Starters + conditionSlow, data= q3)
```




```
qqnorm(q3$residNUM, pch=1, frame= FALSE)
qqline(q3$residNUM, lwd=2)
```




```
q3 <- q3 %>% mutate(prediction2 = predict(model_numRace))
ggplot(data= q3, aes(x=prediction2, y= residNUM)) + geom_point() + labs(title = "Resi
duals vs Fitted Values from Model", x = "Fitted Values",
       y= "Residual") + theme(plot.title = element_text(hjust = 0.5, size = 14)) + ge
om_hline(yintercept = 0)
```



Looking at these outputted scatterplots, we can see random scatter for the residuals, which means the constant variance assumption is met. The QQ plot looks good (besides the upper quantile) so we can say that the normality assumption is also met. Looking at the grid plot of the reponse variable vs each explanatoty variable, we see that there are linear relationships so this assumption is met. Also, the same conclusion as above for independence is reached.

Is there an interaction effect between Starters and Conditions?

```
q3 <- q3 %>% mutate(horsesXconditions = Starters * conditionSlow)
model_Interaction <- lm(Time ~ Year + year2 + Starters +conditionSlow + horsesXcondit
ions, data = q3)
tidy (model_Interaction)
```

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	3.889532e+03	5.006593e+02	7.768820	4.476243e-12
Year	-3.791596e+00	5.121390e-01	-7.403451	2.845723e-11

year2	9.535282e-04	1.309274e-04	7.282880	5.209122e-11
Starters	8.492145e-02	3.829998e-02	2.217271	2.866096e-02
conditionSlow	4.317279e+00	8.973694e-01	4.811039	4.807266e-06
horsesXconditions	-6.837954e-02	6.449421e-02	-1.060243	2.913565e-01

6 rows

Checking assumptions:

No there is not an interaction effect between Starters and Conditions. Looking at the outputted table, we see that the p-value of the interaction term (horsesXconditions) is very high, which means that the interaction is not significant. Since the p-value is so high, we do not need to check assumptions.

Question 4: Ex. 9.21

Analyze the data to see whether the distribution of ingestion rates depends on the percentage of organic matter in the food after acconting for the species weight

```
#use ex0921 data set
q4 <- as_data_frame(ex0921)

q4 <- q4 %>% mutate(logWeight = log(Weight), logIngestion = log(Ingestion), logOrgani
c = log(Organic))

model_birds <- lm(logIngestion ~ logWeight + logOrganic, data= q4)
tidy(model_birds)
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	3.3950689	0.32446815	10.463489	1.457098e-08
logWeight	0.7708333	0.05552626	13.882319	2.426061e-10
logOrganic	-0.9168148	0.09376788	-9.777493	3.753265e-08

3 rows

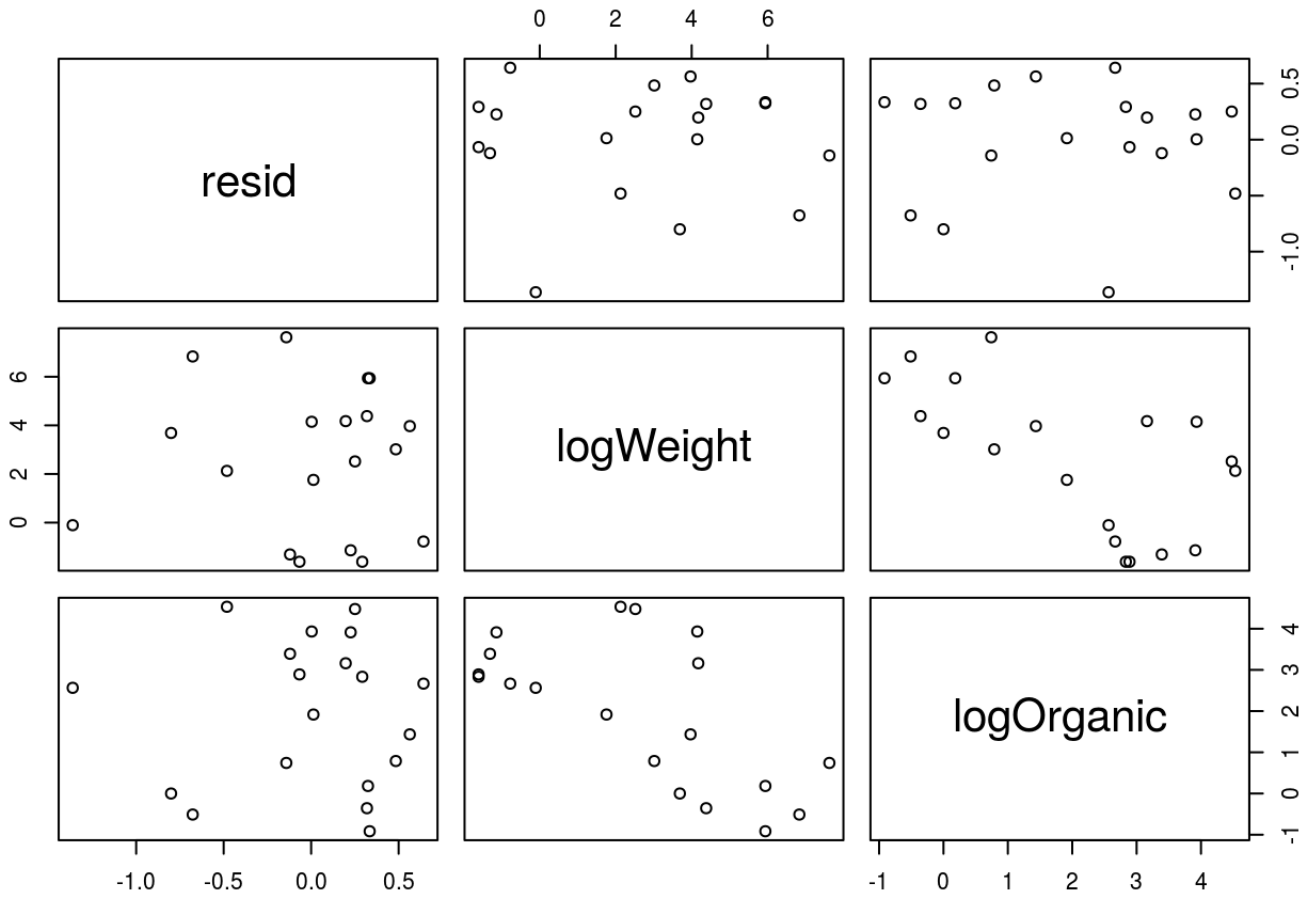
Looking at the outputted table of the model, we see that the estimated slope for logOrganic is -0.9168148 and it has a p-value of 3.753265e-08, which is less than 0.05. This means that it is significant and that the percentage of organic matter in the food does have an effect on the ingestion rate. However, since we logged

the variables, we must interpret them in this context. If we increasing the organic food matter by a factor of C, we expect the median ingestion rate to increase by a multiplicative factor of C^(-0.9168148).

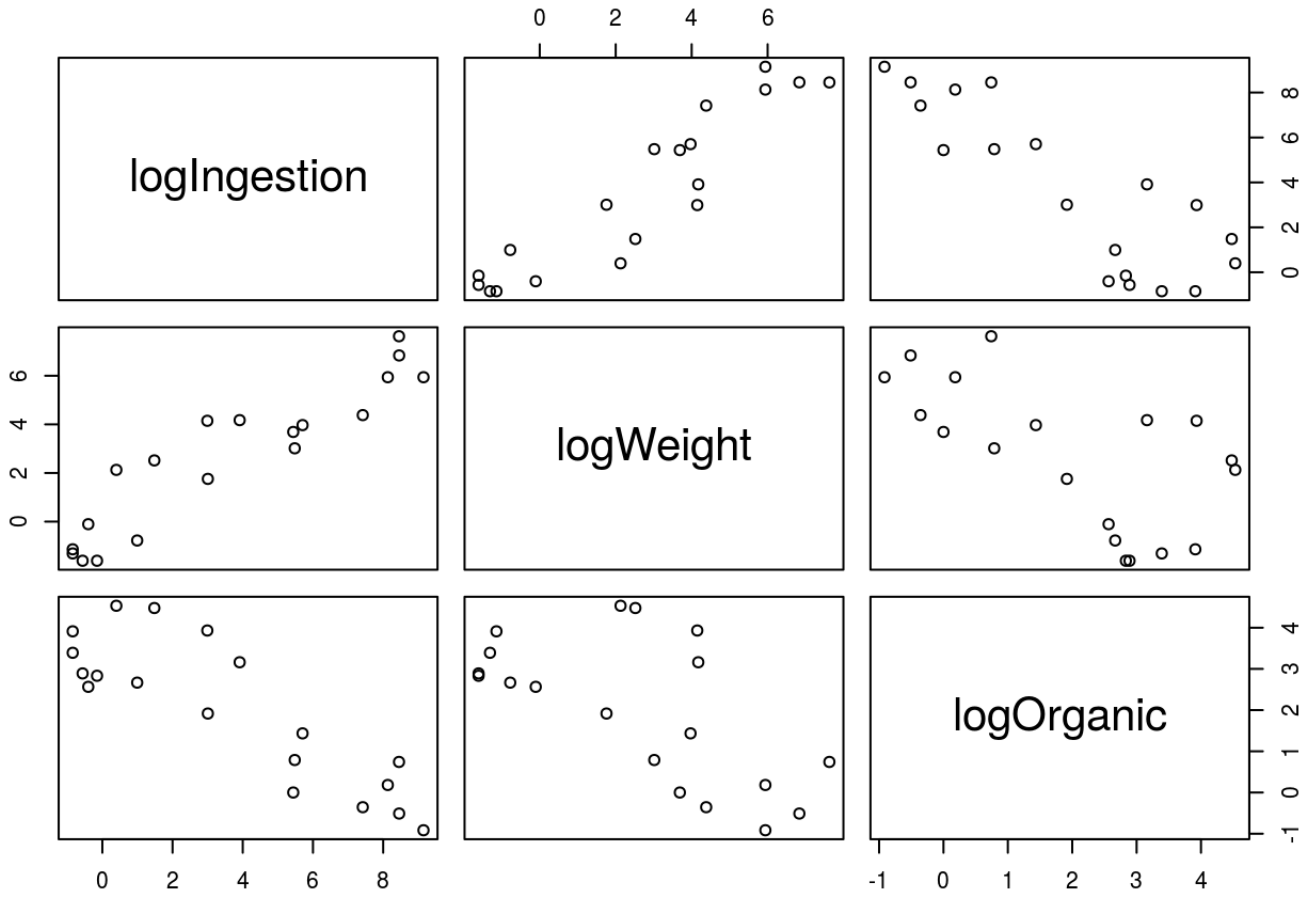
The model is μ{logIngestion |logWeight, logOrganic} = 3.3950689 + 0.7708333 * logWeight - 0.9168148 * logOrganic

To check assumptions:

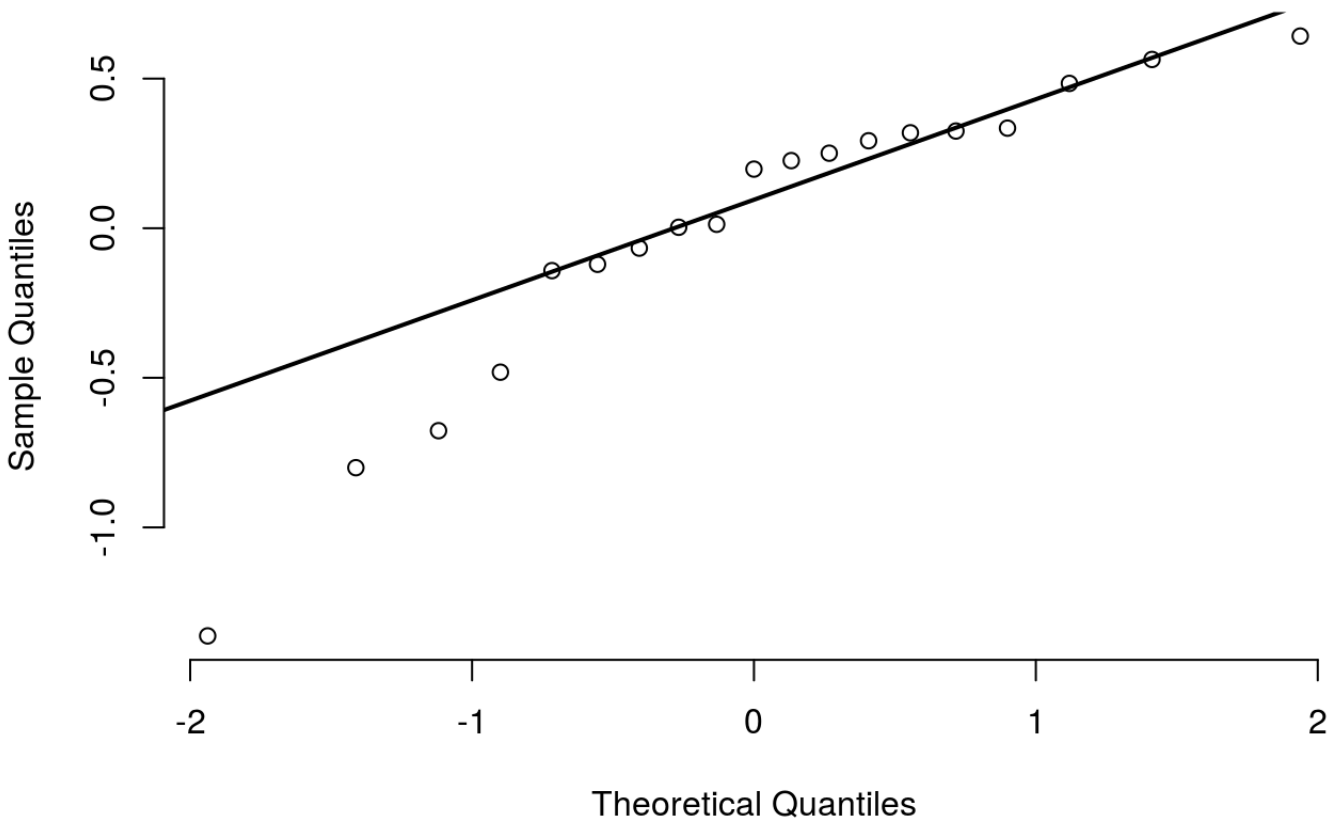
```
q4 <- q4 %>% mutate(resid = resid(model_birds))

pairs(resid ~ logWeight + logOrganic, data=q4)
```



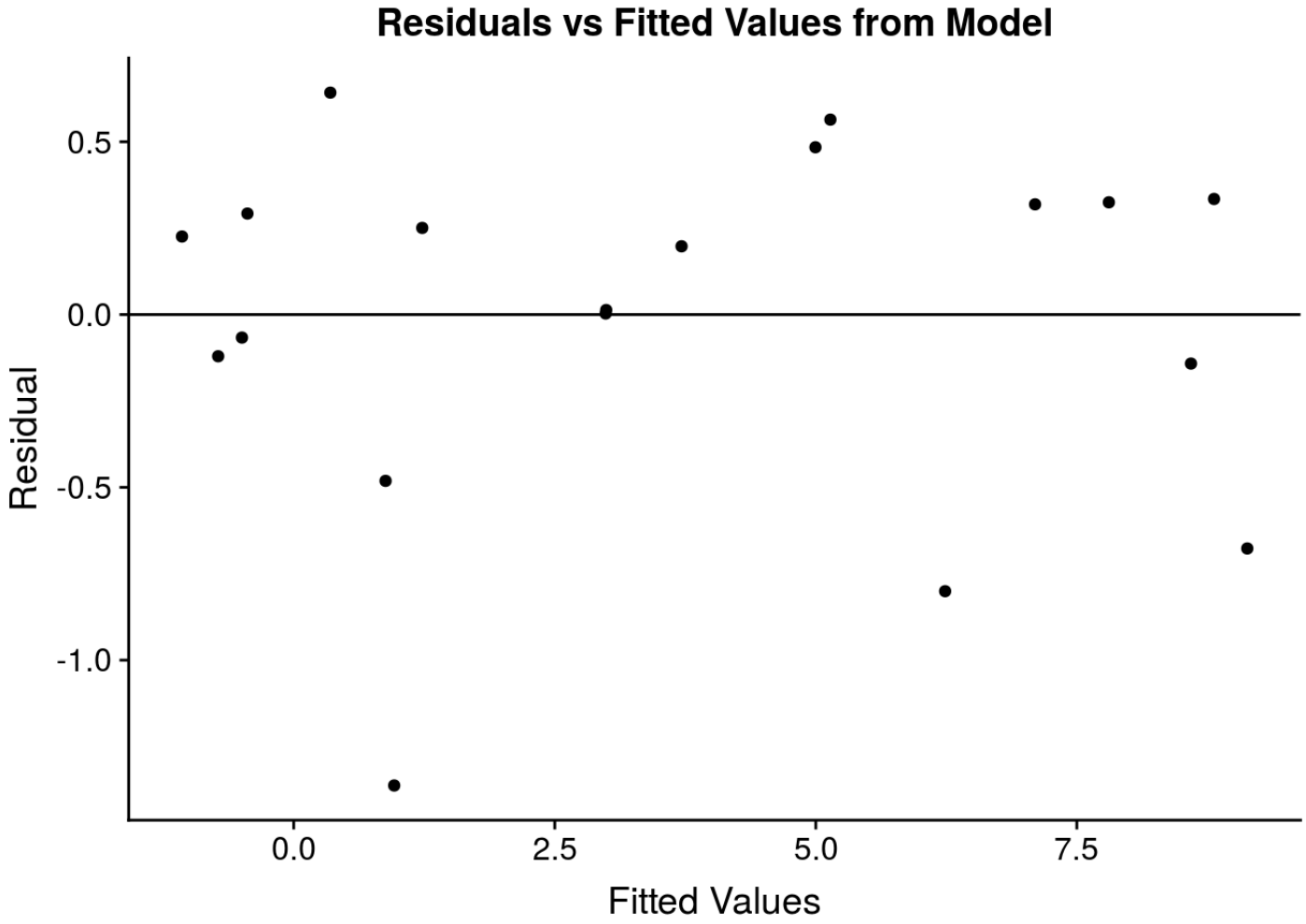

```
pairs(logIngestion ~ logWeight + logOrganic, data= q4)
```




```
qqnorm(q4$resid, pch=1, frame= FALSE)
qqline(q4$resid, lwd=2)
```




```
q4 <- q4 %>% mutate(prediction4 = predict(model_birds))
ggplot(data= q4, aes(x=prediction4, y= resid)) + geom_point() + labs(title = "Residua
ls vs Fitted Values from Model", x = "Fitted Values",
       y= "Residual") + theme(plot.title = element_text(hjust = 0.5, size = 14)) + ge
om_hline(yintercept = 0)
```



Looking at the second grid, we see that there is a linear relationship the response variable and each of the explanatory variables. There is a fairly large departure from the line in the QQ plot, and given the small sample size this is significant so the Normality assumption is not met. The constant variance assumption is met as there is no pattern in the residuals based off the fitted values. There is one outlier with a resiudal of about -1.35, so we need to investigate this data point. The data also meets the independence assumption.

Question 5: Ex. 10.9

a

How many df are there in est of sigma?

n-p = sample size- the number of parameters

38-6 = 32

b

what is the p-value of test for hypo that the

slope in regression of log force on log height is the same for species 2 as species 1?

We can read this off of the outputted table. The reference species is Hemigrapsus nudus (species 1) so we can look at the line lheight x lb (lb is species 2) because this shows the relationship of log force on log height between species 2 and species 1. thus the p value for the hypothesis that the slope in regression of log force on log height is the same for species 2 as species 1 is 0.0014.

c

95% CI for amount by which the slope for species 3 exceeds the slope for species 1?

The formula for a confidence interval is

$$\hat{\beta_1} \pm t * SE(\hat{\beta_1})$$

From the table we plug in:

1.6601 ± 2.036933 x 0.7889

(0.053, 3.267)

```
qt(0.975, 32)
```

```
## [1] 2.036933
```