Homework 11

Regression Diagnostics and Model Refinement

David Josephs

Southern Methodist University
Masters in Data Science
November 11, 2018

Contents

List of Codes

List of Figures

1 Testing Kleiber's Law

1.1 Examining the Raw Data

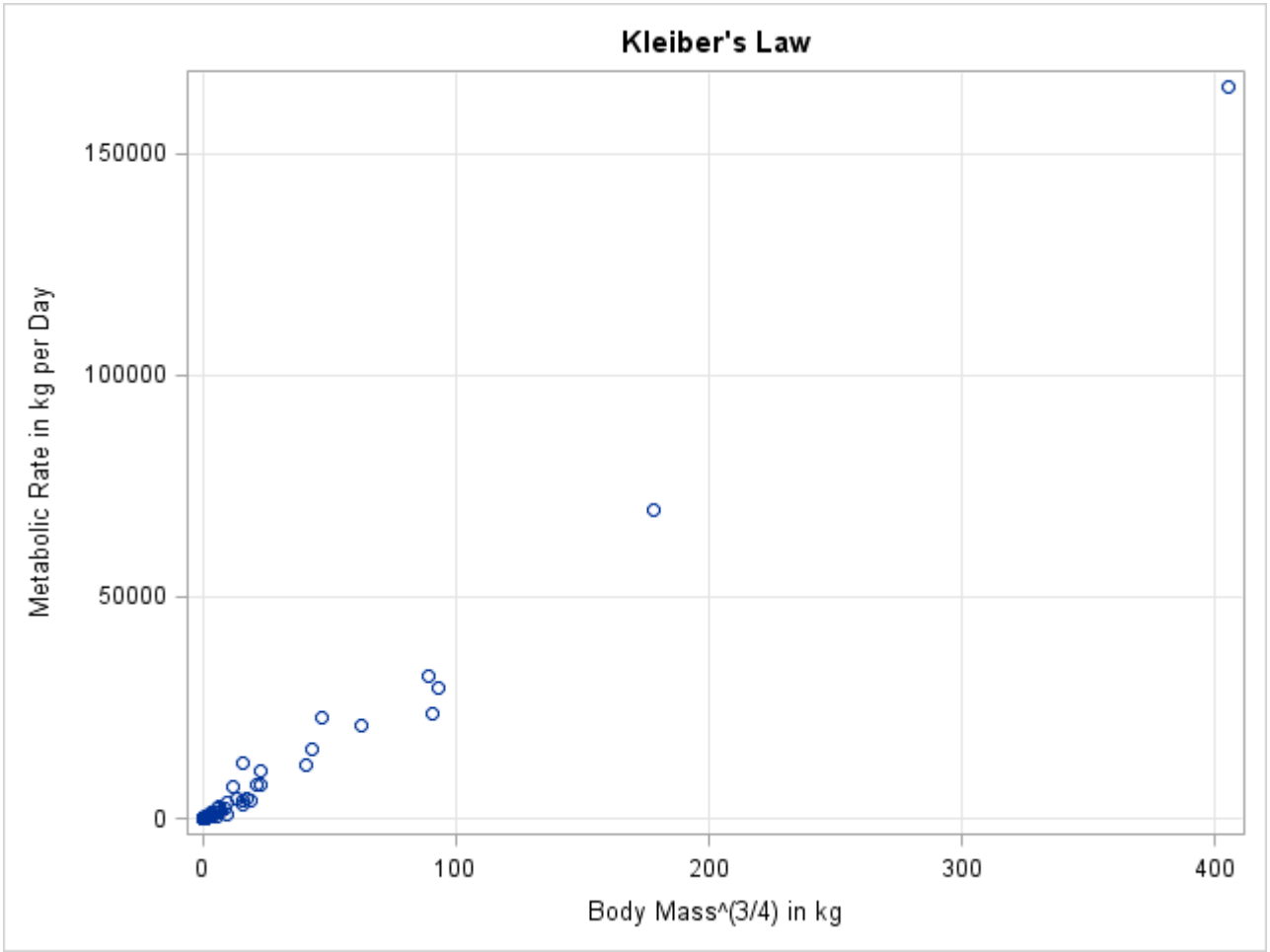
First, we will plot the raw data as a scatterplot, to see if $MASS^{\frac{3}{4}}$ is a reasonable for the data. To do this, we will use the following SAS code:

Code 1. Plotting the Raw Data in SAS

```
data powermetabolism;
set metabolism;
powerMass=Mass**(3/4);
Metab=Metab;
run;
proc sgplot data=powermetabolism noautolegend;
title "Kleiber's Law";
scatter x=powerMass y=Metab;
xaxis label="Body Mass^(3/4) in kg" grid;
yaxis label="Metabolic Rate in kg per Day" grid;
run;
```

Lets see what that graph looks like:

Figure 1. Scatter Plot of The Raw Data



From this graph, it does appear to be linear but it needs to be transformed a bit. a non transformed model will not fit, but in general it should fit.

1.2 Assumption Testing

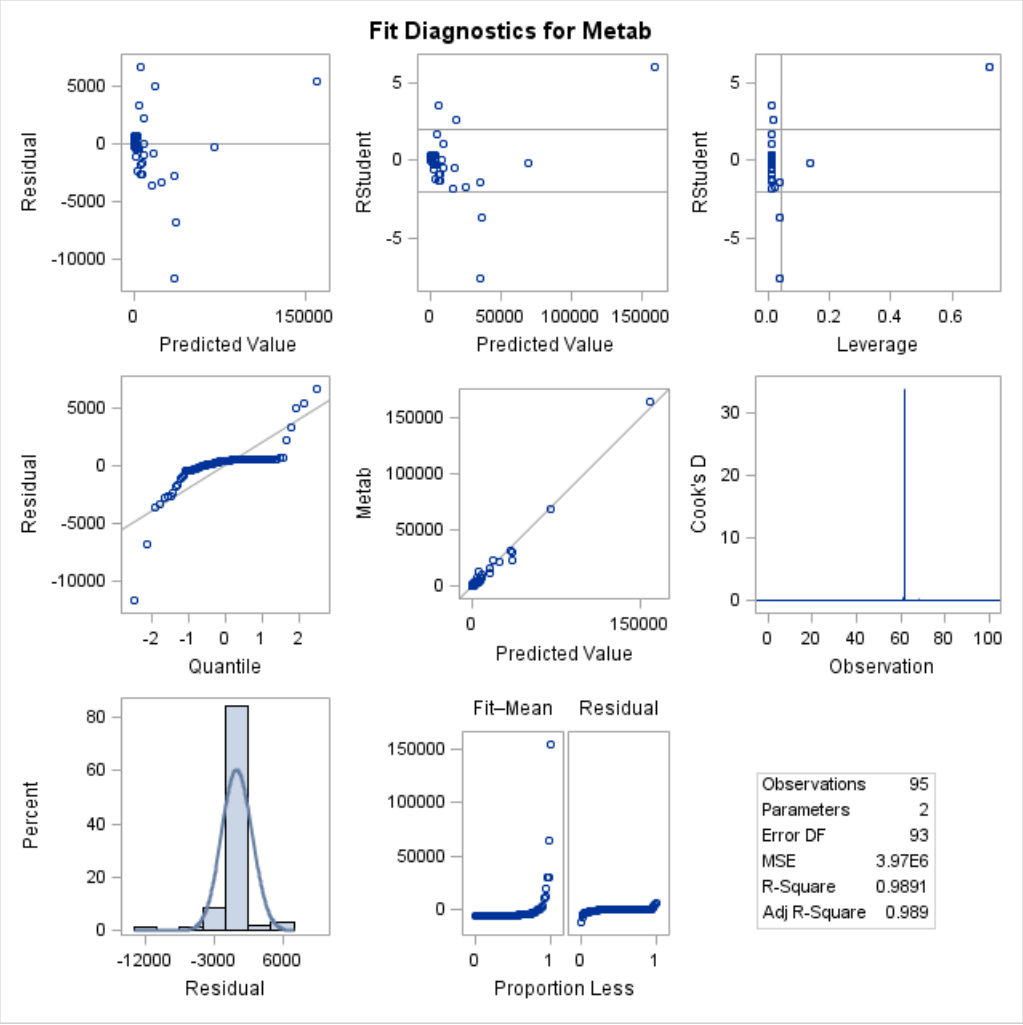
To test the assumptions of the regression of the *raw* data, the following SAS code was used:

Code 2. Regression Diagnostics with SAS

```
proc glm data=powermetabolism plots=all alpha=.05;
model Metab=powerMass / CLPARM;
run;
```

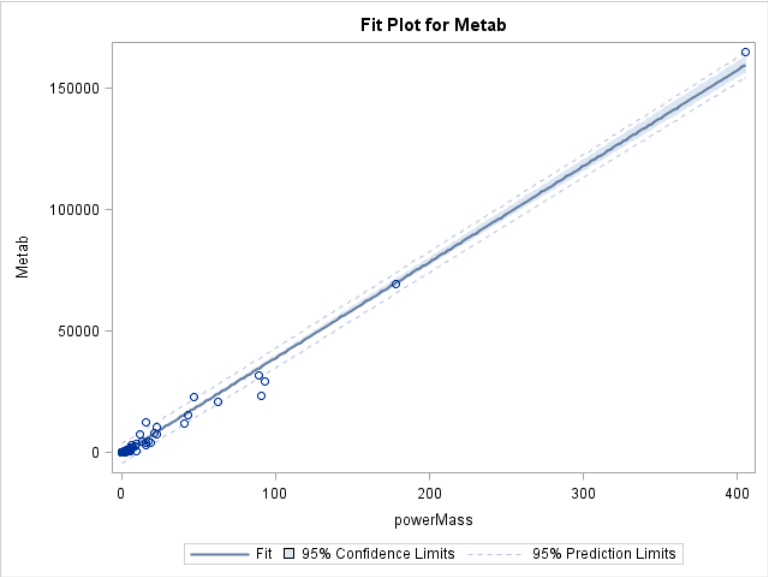
Lets check to see what this looks like in the following figures:

Figure 2. Diagnostic Plots on the Raw Metabolism Data



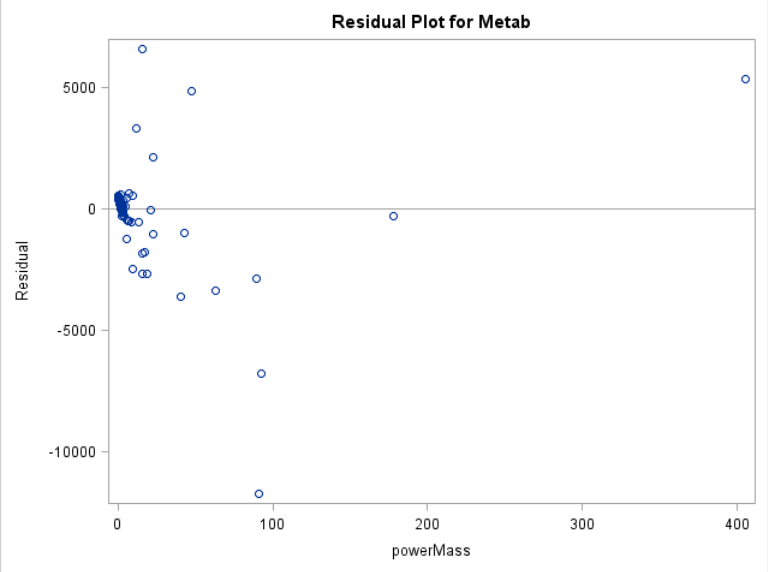
The first two plots,the residuals vs predicted values and the studentized residuals vs predicted values, show us that a lot of our residuals are extreme values. The Q-Q plot and histogram tells us are residuals are not normally distributed at all, which is pretty evident.

Figure 3. Fit Plot



This scatterplot of our data with a linear fit as well as confidence and prediction intervals tells us our data is not very linear, as it is not a random cloud about the fit line.

Figure 4. Residual vs Independent Variable Plot



It is increasingly evident that our data is not linear, as the residuals do not follow a linear pattern at all, not clouding about the x axis. It is also evident, from the distance of the residuals to the x axis, that they do not have a constant variance/spread at all, which is not a good sign for the no-log model

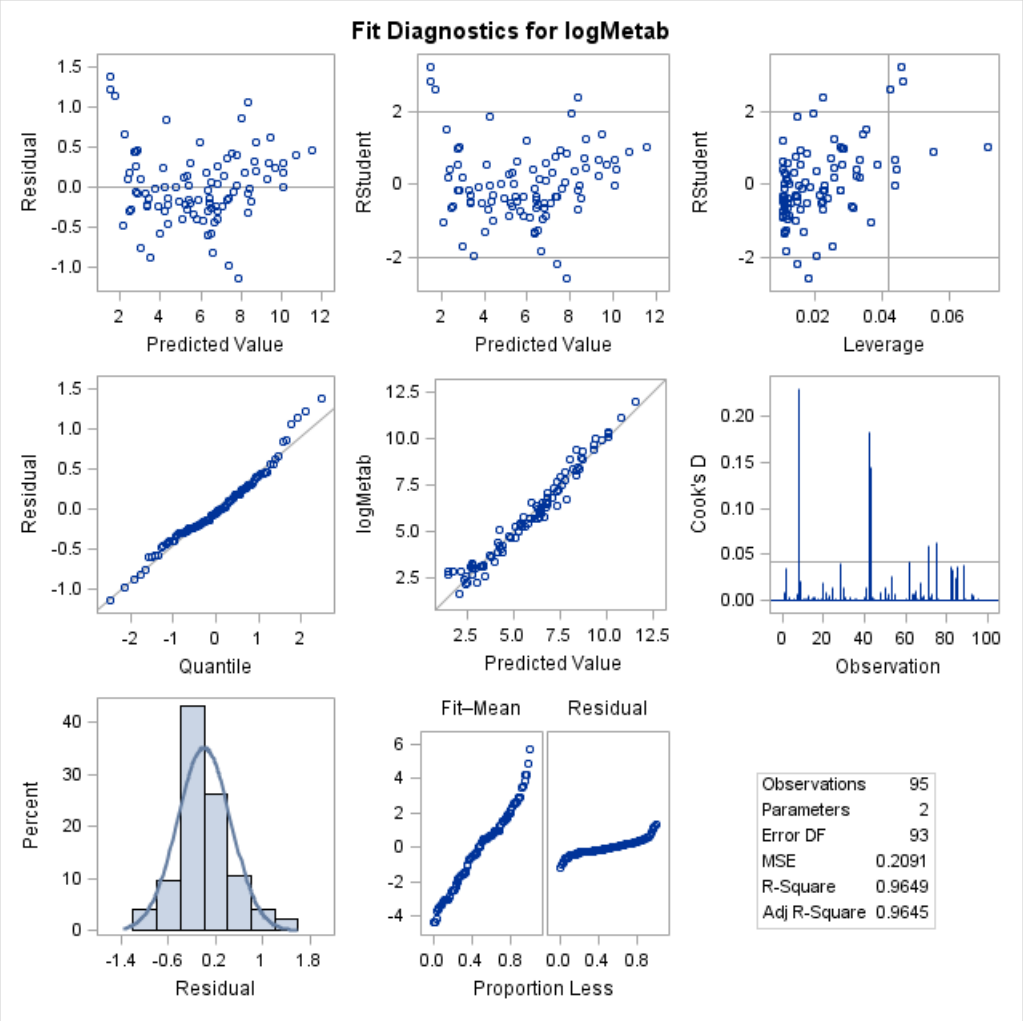
It is clear from Figures 2, 4, and 3 that the linear (no-log) model does not work. It is not linear, does not have constant variance/equal spread, and it is in no way normal. We will try a log-log transformation on the data to see if we can improve things! To do this we will use the following SAS code:

Code 3. Log-Log Transformation in SAS

```
data logmetabolism;
set powermetabolism;
logMass=log(powerMass);
logMetab=log(Metab);
run;
proc glm data=logmetabolism plots=all;
model logMetab=logMass /CLPARM;
run;
```

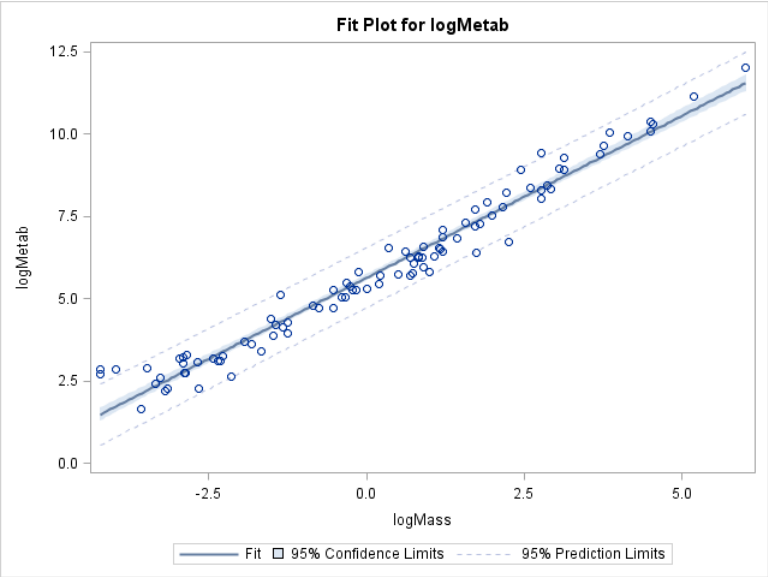
To test our assumptions, we will examine the following figures, produced by Code 3:

Figure 5. Diagnostic Plots of the Log-Log Transformed Data



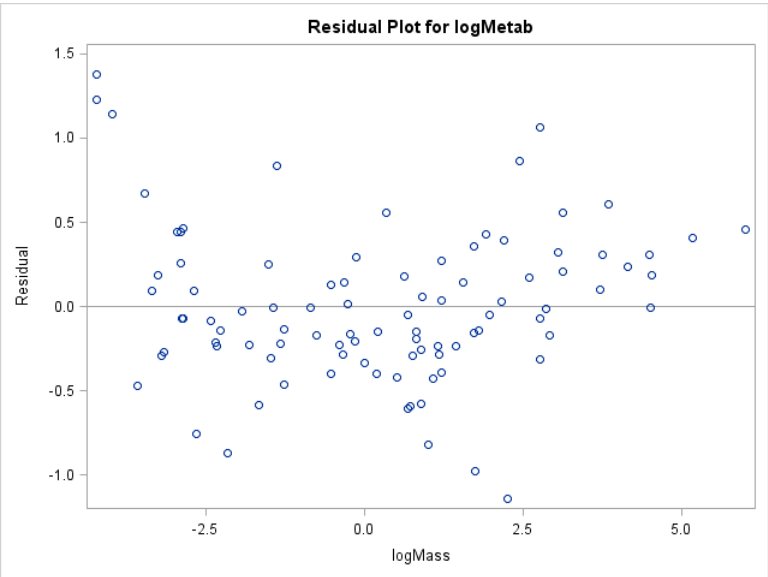
The first two plots,the residuals vs predicted values and the studentized residuals vs predicted values, show us that we do not have a lot of extreme values. The Q-Q plot and histogram tells us are residuals are pretty normally distributed.

Figure 6. Fit Plot of the Log-Log Transformed Data



This scatterplot of our data with a linear fit as well as confidence and prediction intervals tells us our data is pretty linear, a random cloud about the line within reasonable intervals.

Figure 7. Residual vs Independent Variable Plot of the Log-Log Transformed Data



It is increasingly evident that our data is linear, as the residuals cloud about the x axis. It is also evident, from the distance of upper residuals to the lower ones, that they have an equal spread/constant variance.

It is clear from Figures 5, 7, and 6 that the log-log model meets the assumptions of constant variance, normality, and linearity. We will assume independence.

1.3 Assessment of the Model

To assess our model, we will look at the p values and t statistics of the regression parameters. Those were produced in Code 3, and are displayed in the following table:

Table 1. p-Values and T-statistics of Regression Parameters

	Parameter	Estimate	Standard Error	t-Statistic	p-Value	Lower CI	Upper CI
1	Intercept	5.638330664	0.04709325	119.73	<.0001	5.544812798	5.731848530
2	logMass	0.984991519	0.01949275	50.53	<.0001	0.946282773	1.023700265

From table 1, it is clear to see that both our intercept and slope are statistically significant!

1.4 Regression Equation

The regression equation is

$$\ln\left(\widehat{\text{metabolism}}\right) = 5.638330664 + 0.984991519 \ln\left(\widehat{\text{mass}}^{\frac{3}{4}}\right)$$

By unlogging (or raising e to everything), we can untransform the data. That is:

$$e^{\ln\left(\widehat{\text{metabolism}}\right)} = e^{5.638330664 + 0.984991519 \ln\left(\widehat{\text{mass}}^{\frac{3}{4}}\right)} = e^{5.638330664} e^{\left[\ln\left(\widehat{\text{mass}}^{\frac{3}{4}}\right)\right]^{0.984991519}}$$

Or

$$\widehat{\text{metabolism}} = e^{5.638330664} \left(\widehat{\text{mass}}^{\frac{3}{4}}\right)^{0.984991519}$$

I am still not sure why we are testing it this way... Why do we not just use the normal mass and see if the exponent is .75 or not?

1.5 Interpretation

The model states that the natural log of metabolism is equal to 5.638330664 plus 0.984991519 times the natural log of mass to three quarters. That says that if the natural log of mass to three over four were to increase by 1, the natural log of metabolism would increase by almost 1. If we unlog, we can say that a 1% increase in mass to the three quarters leads to a .985% increase in metabolism. A 95% confidence interval for this is $[0.946282773, 1.023700265]$, meaning that a 1% increase in mass to the three quarters leads to a percent increase in the *median* of metabolism between those two values. The problem did not ask for a confidence interval on the intercept but you can see it in Table 1. Because one is contained in the confidence interval, we can say that mass to the three quarters is a good estimate of the metabolism.

1.6 Proportions

Table 2. Proportion explained by the model

	R-Square	Coeff Var	Root MSE	logMetab Mean
1	0.964858	7.819557	0.457235	5.847322

R^2 is 0.965, which means 96.5% of the variation is explained by the model. A good fit! Well done mr. Kleiber!

2 Autism Study

2.1 Assumption Checking

To check the assumptions of the *raw* data, the following R Code was used:

Code 4. Plotting the Raw Data in R

```
data<-read.csv("Data/Autism.csv")
data<-data%>%mutate(logPrevalence=log(Prevalence))
model1 <-lm(Prevalence~Year,data=data)
pint<-predict(model1,interval="prediction",level=.95)
new_df<-cbind(data,pint)
g<-ggplot(new_df, aes(x=Year, y=Prevalence))+geom_point() +
geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
geom_line(aes(y=upr), color = "red", linetype = "dashed")+
geom_smooth(method=lm, se=TRUE,level=0.95)+
theme_classic()
resp<-ggplot(model1, aes(.fitted, .resid))+
geom_point(aes(color = .resid)) +
scale_color_gradient2(low = "blue",
mid = "gray",
high = "red") +
guides(color = FALSE)
resp<-resp+
stat_smooth(method="lm",se=F)+
geom_hline(yintercept=0,
col="red",
linetype="dashed")
resp<-resp+
xlab("Fitted values")+
ylab("Residuals")
resp<-resp+theme_classic()+geom_segment(aes(y=0,x=.fitted,xend=.fitted, yend=.resid,
alpha = 2*abs(.resid)))+guides(alpha=FALSE)
resx<-ggplot(model1, aes(Year, .resid))+
geom_point(aes(color = .resid)) +
scale_color_gradient2(low = "blue",
mid = "gray",
high = "red") +
guides(color = FALSE)
resx<-resx+
stat_smooth(method="lm",se=F)+
geom_hline(yintercept=0,
col="red",
linetype="dashed")
resx<-resx+
xlab("Year")+
ylab("Residuals")
resx<-resx+theme_classic()+geom_segment(aes(y=0,x=Year,xend=Year, yend=.resid,
alpha = 2*abs(.resid)))+guides(alpha=FALSE)
extraVal<-fortify(model1)
ghist<-ggplot(model1,aes(x=.resid))+
geom_histogram(bins=4,
fill='gray22',
color='ghostwhite')+
theme_classic()
ghist<-ghist+stat_function(fun=dnorm,
color="forestgreen",args=list(mean=mean(extraVal$.resid),
sd=sd(extraVal$.resid)),size=1)
```

This results in the following figures:

Figure 8. Plot of the Raw Fit

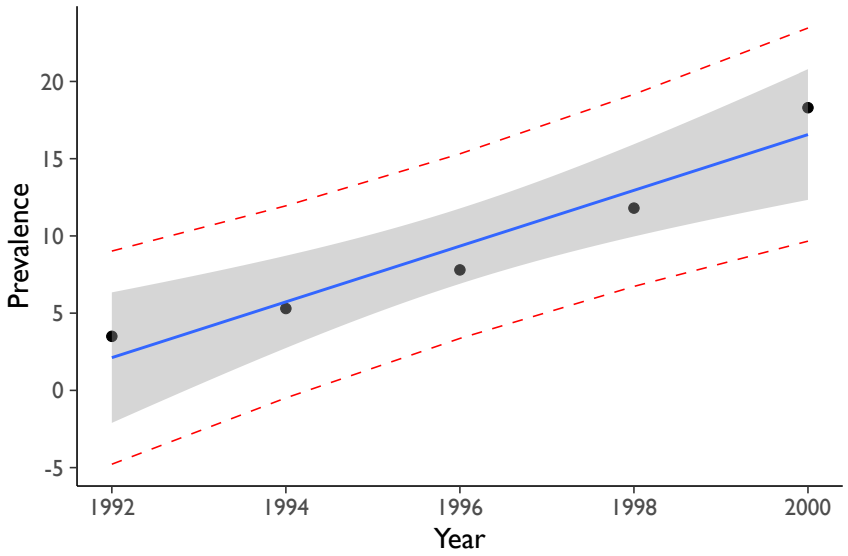


Figure 9. Plot of the Raw Residuals vs Fitted Values

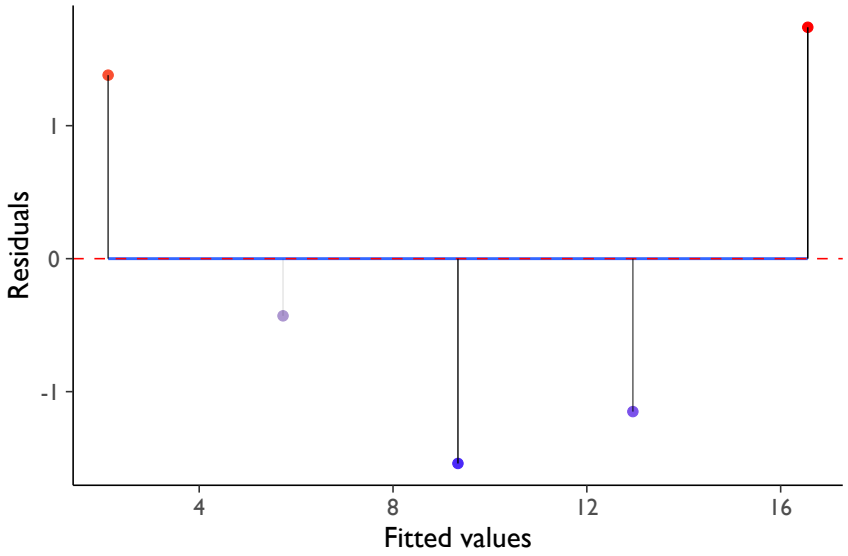


Figure 10. Plot of the Raw Residuals vs Year

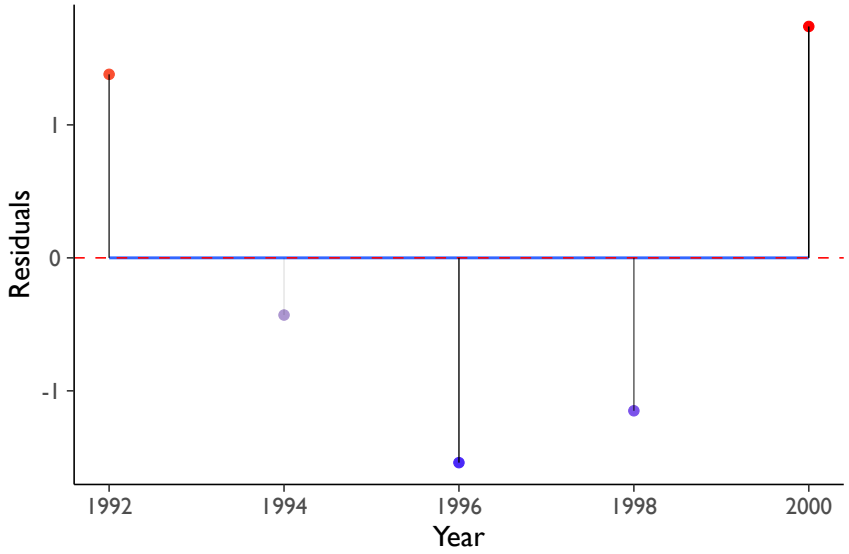
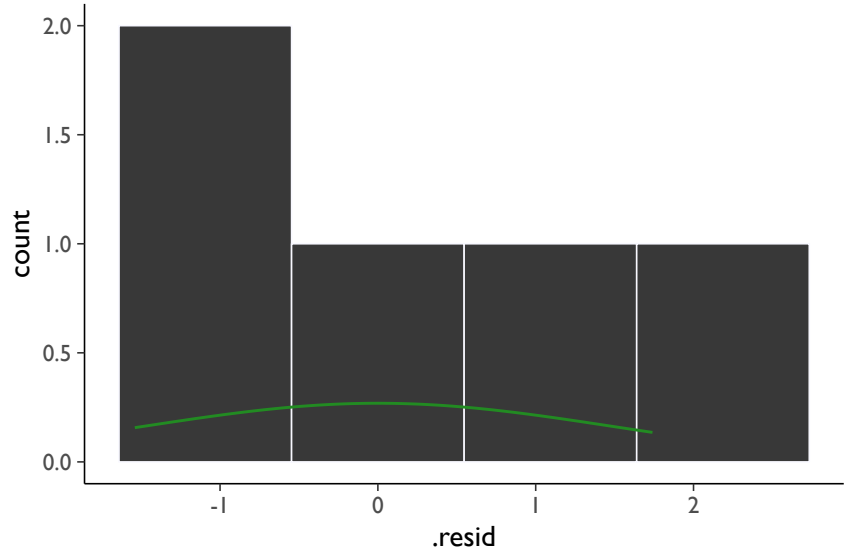


Figure 11. Histogram of the Raw Residuals



As we can see, this data needs to be transformed! Lets log transform the Y axis! We do this with the following Code:

Code 5. Plotting the Log-Linear Data in R

```
model2 <-lm(logPrevalence~Year,data=data)
pint2<-predict(model2,interval="prediction",level=.95)
new_df<-cbind(data,pint2)
g2<-ggplot(new_df, aes(x=Year, y=logPrevalence))+geom_point() +
geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
geom_line(aes(y=upr), color = "red", linetype = "dashed")+
geom_smooth(method=lm, se=TRUE,level=0.95)+
theme_classic()
resp2<-ggplot(model2, aes(.fitted, .resid))+
geom_point(aes(color = .resid)) +
scale_color_gradient2(low = "blue",
mid = "gray",
high = "red") +
guides(color = FALSE)
resp2<-resp2+
stat_smooth(method="lm",se=F)+
geom_hline(yintercept=0,
col="red",
linetype="dashed")
resp2<-resp2+
xlab("Fitted values")+
ylab("Residuals")
resp2<-resp2+theme_classic()+geom_segment(aes(y=0,x=.fitted,xend=.fitted, yend=.resid,
alpha = 2*abs(.resid)))+guides(alpha=FALSE)
resx2<-ggplot(model2, aes(Year, .resid))+
geom_point(aes(color = .resid)) +
scale_color_gradient2(low = "blue",
mid = "gray",
high = "red") +
guides(color = FALSE)
resx2<-resx2+
stat_smooth(method="lm",se=F)+
geom_hline(yintercept=0,
col="red",
linetype="dashed")
resx2<-resx2+
xlab("Year")+
ylab("Residuals")
resx2<-resx2+theme_classic()+geom_segment(aes(y=0,x=Year,xend=Year, yend=.resid,
alpha = 2*abs(.resid)))+guides(alpha=FALSE)
extraVal2<-fortify(model2)
ghist2<-ggplot(model2,aes(x=.resid))+
geom_histogram(bins=4,
fill='gray22',
color='ghostwhite')+
theme_classic()
ghist2<-ghist2+stat_function(fun=dnorm,
color="forestgreen",args=list(mean=mean(extraVal2$.resid),
sd=sd(extraVal2$.resid)),size=1)
```

Figure 12. Plot of the Logged Fit

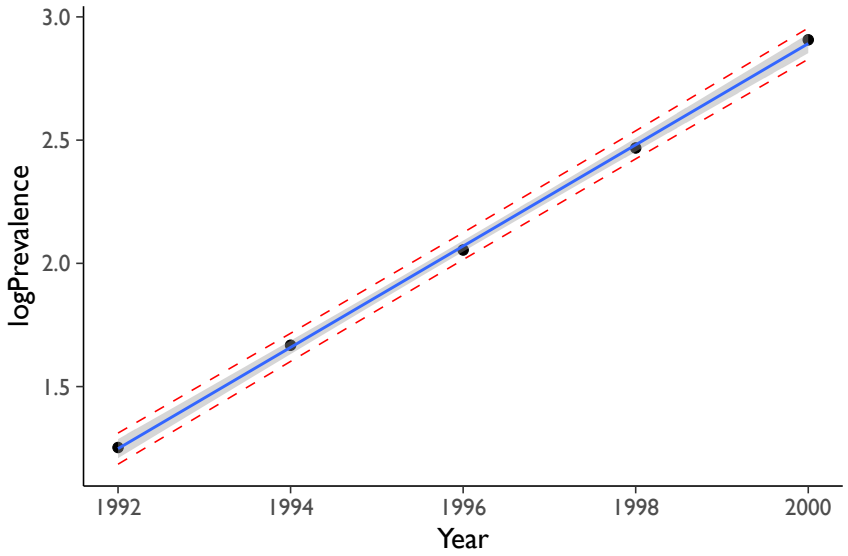


Figure 13. Plot of the Logged Residuals vs Fitted Values

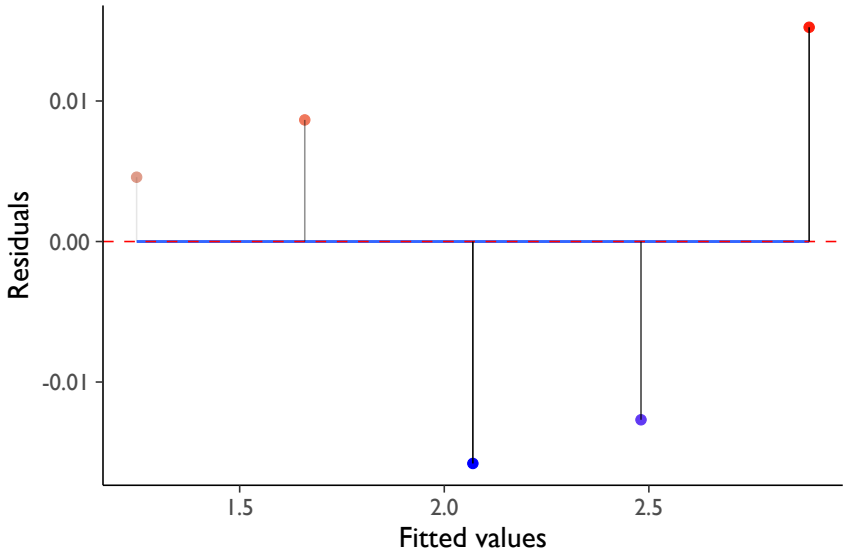


Figure 14. Plot of the Logged Residuals vs Year

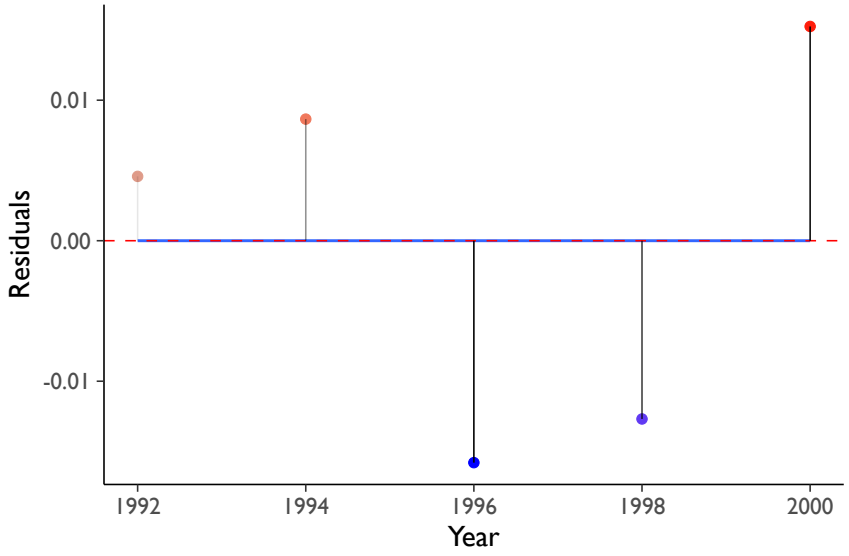
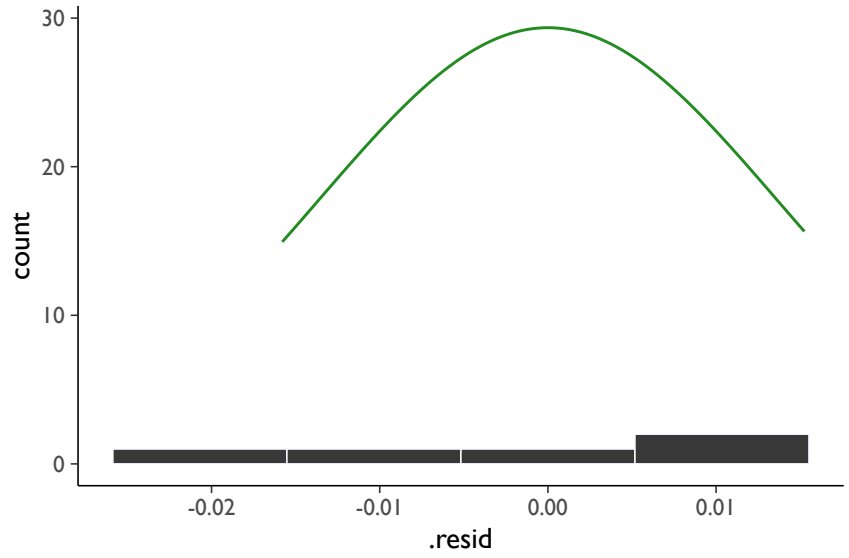


Figure 15. Histogram of the Logged Residuals



This is a way better fit!

2.2 Model Assessment

Lets look at the model parameters!

Table 3. Hypothesis Test on Regression Parameters

	term	estimate	std.error	statistic	p.value	conf.low	conf.high
1	(Intercept)	-407.98	4.95	-82.38	0.00	-423.74	-392.21
2	Year	0.21	0.00	82.79	0.00	0.20	0.21

2.3 Regression Equation

The regression equation is:

$$\ln\left(\widehat{\text{Prevalence}}\right) = -408 + .21\left(\widehat{\text{Year}}\right)$$

2.4 Interpretation

A one unit increase in Year is associated with a multiplicative change of $e^{\cdot}21$ in the median of Prevalence. A confidence interval for this is shown in Table 3, it is very narrow.

2.5 R squared

```
rsquare<-summary(model2)$r.squared
```

$R^2 = 0.9995625$