Transformations to Remove Heteroscedasticity

BIOS 6611

CU Anschutz

Week 9

1. Homoscedasticity and Transformations

2. Log-Transformation Examples

3. Transformations to Address Non-Linearity

Homoscedasticity and Transformations

Homoscedasticity Assumption

The assumption of homoscedasticity is important, especially if the regression analysis is used for predictions.

Transformations of the response variable (the dependent variable) are often used to remove heteroscedasticity. This type of transformation is called a **variance-stabilization transformation**.

Taking the natural log of the response variable is a particularly useful transformation, especially for removing heteroscedasticity when the residual variance is an increasing function of X.

Other transformations are sometimes used to stabilize the variance, but they may give a model that is more difficult to interpret.

Some Transformations

Relationship of σ^2 **to** $E[Y]$	**Transformation**	**Comment**
$\sigma^2 \propto E[Y]$	$\sqrt{Y}$	Used for Poisson data
$\sigma^2 \propto E[Y](1 - E[Y])$	$\sin^{-1}\sqrt{Y}$	Used for binomial proportions or rates
$\sigma^2 \propto (E[Y])^2$	$\log(Y)$	Also used for non-linearity, non-normality; y>0
$\sigma^2 \propto (E[Y])^3$	$Y^{-1/2}$	
$\sigma^2 \propto (E[Y])^4$	Y^{-1}	

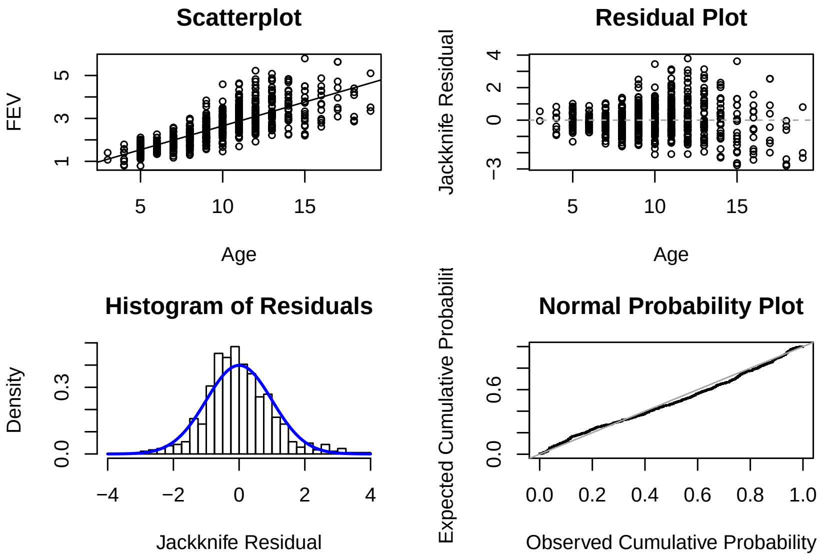
Log-Transformation Examples

FEV Data Set

Let's examine an example using our FEV data set for the regression model to predict FEV status based on age. We can review the four diagnostic plots to see if our assumptions appear to be met:

```r
# Code to generate figures
fev <- read.csv('FEV_rosner.csv')
mod1 <- glm( fev ~ age, data=fev)

par(mfrow=c(2,2), mar=c(4.1,4.1,3.1,2.1))
plot(x=fev$age, y=fev$fev, xlab='Age', ylab='FEV', main='Scatterplot',
     cex=0.7); abline( mod1 )

plot(x=fev$age, y=rstudent(mod1), xlab='Age', ylab='Jackknife Residual',
     main='Residual Plot', cex=0.7); abline(h=0, lty=2, col='gray65')

hist(rstudent(mod1), xlab='Jackknife Residual',
     main='Histogram of Residuals', freq=F, breaks=seq(-4,4,0.25));
  curve( dnorm(x,mean=0,sd=1), lwd=2, col='blue', add=T)

plot( ppoints(length(rstudent(mod1))), sort(pnorm(rstudent(mod1))),
      xlab='Observed Cumulative Probability',
      ylab='Expected Cumulative Probability',
      main='Normal Probability Plot', cex=2, pch='.');
  abline(a=0,b=1, col='gray65', lwd=1)
```

FEV Non-Transformed Diagnostic Plots

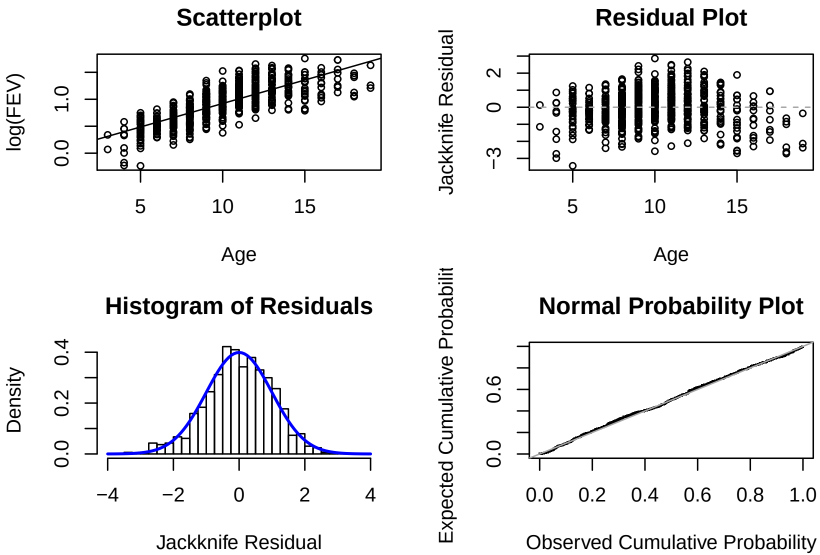


(Natural) Log Transformation

Since we have a potential violation of our assumption of homoscedasticity, let's take a log-transformation of our FEV outcome:

```r
# Code to generate figures
mod2 <- glm( log(fev) ~ age, data=fev)

par(mfrow=c(2,2), mar=c(4.1,4.1,3.1,2.1))
plot(x=fev$age, y=log(fev$fev), xlab='Age', ylab='log(FEV)',
     main='Scatterplot', cex=0.7); abline( mod2 )

plot(x=fev$age, y=rstudent(mod2), xlab='Age', ylab='Jackknife Residual',
     main='Residual Plot', cex=0.7); abline(h=0, lty=2, col='gray65')

hist(rstudent(mod2), xlab='Jackknife Residual',
     main='Histogram of Residuals', freq=F, breaks=seq(-4,4,0.25));
  curve( dnorm(x,mean=0,sd=1), lwd=2, col='blue', add=T)

plot( ppoints(length(rstudent(mod2))), sort(pnorm(rstudent(mod2))),
      xlab='Observed Cumulative Probability',
      ylab='Expected Cumulative Probability',
      main='Normal Probability Plot', cex=2, pch='.');
  abline(a=0,b=1, col='gray65', lwd=1)
```

FEV Non-Transformed Diagnostic Plots



Interpretation with Log-Transformed Response

When a logarithmic transformation of the dependent variable is used, the model is now interpreted on the scale of the transformed outcome. For example:

```
mod2 <- glm(log(fev) ~ age, data=fev)
coef(mod2)
```

```
## (Intercept)         age
##   0.0505960   0.0870833
```

This results in $E[\log(\text{FEV})] = 0.0506 + 0.0871 \times \text{Age}$.

With respect to log(Y), we have our usual interpretations:

- The intercept of 0.0506 is the mean log(FEV) for someone who is age 0.
- The slope term indicates that for each 1 year increase in age, on average, log(FEV) increases by 0.0871.

However, we usually would like to interpret our results on the original scale...

Interpretation with Log-Transformed Response

To get an interpretation back on the original scale, we can transform our beta coefficients. For a log transformation, we exponentiate our β's:

$$E[\log(\text{FEV})] = 0.0506 + 0.0871 \times \text{Age} \to \exp\{0.0506 + 0.0871 \times \text{Age}\}$$

However, our transformed estimates no longer represent the *arithmetic mean* that we are used to. For a log-transformation they represent the *geometric mean*:

$$\exp(E[\log(\text{FEV})]) = \exp\{0.0506\}\exp\{0.0871 \times \text{Age}\} = 1.052 \times (1.091)^{\text{Age}}$$

For the interpretation of our slope, for each year of age, FEV increases, on average, 1.091 times (or ~9% per year).

Interpretation with Log-Transformed Response

From our equation we can also address other questions:

$$\exp(E[\log(\text{FEV})]) = \exp\{0.0506\} \exp\{0.0871 \times \text{Age}\} = 1.052 \times (1.091)^{\text{Age}}$$

The expected geometric mean for a 0 year old:

The expected geometric mean for a 5 year old:

The estimated percent increase in FEV for a difference in 5 years between two individuals:

Confidence Interval Calcuations

```r
summary(mod2)$coefficients
```

```
##               Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept) 0.0505960 0.029104004  1.738455  8.260273e-02
## age         0.0870833 0.002809118 31.000228 2.297876e-130
```

The 95% confidence interval for the slope (or intercept) can also be transformed in the same way.

On our log(FEV) scale we have

$$0.0871 \pm 1.96 \times 0.0028 = (0.082, 0.093)$$

Then we exponentiate to our original FEV scale:

$$(e^{0.082}, e^{0.093}) = (1.085, 1.097)$$

In a brief, but complete, summary we would state:

There is a significant increase in FEV {**decision**} for a one year increase in age ($p<0.001$) {**uncertainty**}. On average, FEV increases by 9.1% {**point estimate**} (95% CI: 8.5% to 9.7%) {**interval estimate**} for every one year increase in age.

log(FEV) Example with a Categorical Predictor

Let's examine the interpretation if we look at smoking status in those 14 or older:

```
mod3 <- glm(log(fev)~smoke, data=fev[which(fev$age>=14),])
summary(mod3)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  1.3266410 0.03336716 39.758888 9.761164e-51
## smokesmoker -0.1302228 0.05240521 -2.484921 1.528284e-02
```

$$E[\log(FEV)] = 1.32664 - 0.13022 \times \text{smoker} \rightarrow E^*(FEV) = 3.768 \times (0.878)^{\text{smoker}}$$

For non-smokers their *geometric* mean FEV is 3.768 liters.

For smokers, their *geometric* mean FEV is $3.768 \times 0.878 = 3.308$ liters.

log(FEV) Example with a Categorical Predictor

We can also summarize the difference between smokers and non-smokers and its accompanying 95% CI:

$$E^*(FEV) = 3.768 \times (0.878)^{\text{smoker}} \rightarrow (1 - 0.878) \times 100 = 12.2\%$$

95% CI on log scale: $-0.13022 \pm 1.96(0.05241) = (-0.2329, -0.0275)$.

Now we exponentiate: $(e^{-0.2329}, e^{-0.0275}) = (0.79, 0.97)$.

In other words, we are 95% confident that FEV is between 3% and 21% lower in smokers compared to non-smokers.

In a brief, but complete, summary:

There is a significant association between smoking status and FEV ($p = 0.0153$). On average, FEV is 0.878 times lower (95% CI: 0.79 to 0.97) in smokers compared to non-smokers. *(Note, we could also have presented it as FEV is 12.2% lower (95% CI: 3% to 21%).)*

Transformations to Address Non-Linearity

Transformations to Address Non-Linearity

Linear regression methods can be used to model curves as long as those curves can be expressed in a linear fashion. The following are examples of curvilinear relationships that can be estimated using linear regression models:

- $Y = e^{\beta_0} e^{\beta_1 X} e^{\epsilon} \rightarrow \log(Y) = \beta_0 + \beta_1 X + \epsilon$
- $Y = \sqrt{\beta_0 + \beta_1 X + \epsilon} \rightarrow Y^2 = \beta_0 + \beta_1 X + \epsilon$
- $Y = \beta_0 + \beta_1 \log(X) + \epsilon$
- $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$
- $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

Transformations of the independent variables are usually performed to address non-linearity or reduce leverage/influence, not non-normality:

- The independent variables need not be normally distributed.
- In fact, we have already seen the use of categorical variables as independent variables, which are far from normally distributed.