

Data Transformations and Variance

David M. Rocke

March 2, 2026

Basic Assumptions of Linear Regression

- Linearity:** The mean value of y (the response) conditional on the values of the predictors is a linear function of the predictors. Predictors themselves can be non-linear combinations.
- Independence:** The error terms of different data points are statistically independent of each other (which can be encouraged by randomization).
- Constant Variance:** The error terms of the data points all have the same variance.
- Normal Errors:** The distribution of errors is normal (not particularly important).

Nonconstant Variance

A particularly common “violation” of the assumptions is heteroscedasticity or non-constant variance and one very common version is when the variance is a function of the mean.

$$y \sim f(\mu, h(\mu))$$

Distribution	Pars	Mean	Variance
Lognormal	(μ, σ)	$E(y) = \eta = \exp(\mu + \sigma^2/2)$	$V(y) = [\exp(\sigma^2) - 1] \exp(2\mu + \sigma^2)$
Poisson	λ	$E(y) = \mu = \lambda$	$V(y) = \lambda = \mu$
Exponential	$1/\lambda = \mu$	$E(y) = \mu$	$V(y) = \mu^2$
Binomial	(n, p)	$E(y) = \mu = np$	$V(y) = np(1 - p) = \mu(n - \mu)/n$

For the binomial, the variance is a function of the mean that also depends on n , which is known. For the lognormal, the square of the coefficient of variation is the ratio of the variance to the square of the mean

$$\frac{V(y)}{\eta^2} = \frac{[\exp(\sigma^2) - 1] \exp(2\mu + \sigma^2)}{\exp(2\mu + \sigma^2)} = [\exp(\sigma^2) - 1]$$

Thus, for the lognormal, the variance is a constant multiple of the square of the mean. Also note that for the exponential distribution, the variance is exactly the square of the mean.

The Delta Method

The delta method is a way of approximating the behavior of a function of a random variable. Suppose Y is a random variable with mean $E(Y) = \mu$ and variance $V(Y) = \sigma^2$. And suppose $g(\cdot)$ is a smooth transformation function. Then the Taylor series expansion of $g(Y)$ around the mean is

$$W = g(Y) \approx g(\mu) + g'(\mu)(Y - \mu) + \frac{1}{2}g''(\mu)(Y - \mu)^2 + \dots$$

Or to the first order

$$W = g(Y) \approx g(\mu) + g'(\mu)(Y - \mu)$$

Then, to the first order, $E(W) \approx g(\mu)$ and $V(W) \approx [g'(\mu)]^2 \sigma^2$. This approximation is called the (first-order) delta method.

Nonconstant Variance

Now suppose that $E(Y) = \mu$ and $V(Y) = h(\mu)$, so that the variance depends on the mean. We call $h(\mu)$ the variance function. Can we find a transformation function $g(\cdot)$ such that (to the first order) $W = g(Y)$ has constant variance? By the delta method,

$$\begin{aligned}V(W) &\approx [g'(\mu)]^2 h(\mu) = C^2 \\g'(\mu) \sqrt{h(\mu)} &= C \\ \frac{dg(\mu)}{d\mu} &= \frac{C}{\sqrt{h(\mu)}} \\ g(\mu) &= \int \frac{Cd\mu}{\sqrt{h(\mu)}}\end{aligned}$$

with the constant being irrelevant.

Variance a Power of the Mean

If

$$V(Y) = h(\mu) = c\mu^2,$$

then

$$g(\mu) \propto \int \mu^{-1} d\mu = \ln(\mu).$$

If

$$V(Y) = h(\mu) = c\mu^{2p}, \quad p \neq 1$$

then

$$g(\mu) \propto \int \mu^{-p} d\mu \propto \mu^{1-p}.$$

So, if the CV is constant, take logs. If the standard deviation is proportional to a different power $p \neq 1$ of the mean, then use a power transformation $g(y) = y^{1-p}$.

For example, if the variance function $h(\mu) = \mu$, as in the Poisson distribution, then $p = 1/2$, $c = 1$, and $g(\mu) = \mu^{1/2} = \sqrt{\mu}$. If Y is Poisson with parameter λ , then the mean and variance of $W = g(Y)$ are

$$\begin{aligned} E(W) &\approx g(\lambda) = \sqrt{\lambda} = \lambda^{1/2} \\ V(W) &\approx g'(\mu)^2 \sigma^2 \\ &= g'(\lambda)^2 \lambda \\ &= \left[\frac{1}{2\sqrt{\lambda}} \right]^2 \lambda \\ &= \frac{1}{4\lambda} \lambda = 1/4 \end{aligned}$$

So the square-root of a Poisson random variable with parameter λ has approximate mean $\sqrt{\lambda}$ and approximate variance $1/4$.

If we let $\beta = 1 - p$ and define $g(\mu) = (y^\beta - 1)/\beta$, then this is a linear transformation of the function $g(\cdot)$ derived above when $p \neq 1$. When $p = 1$,

$$\lim_{\beta \rightarrow 0} (y^\beta - 1)/\beta = \ln(y)$$

by L'Hôpital's rule. Thus the transformation we want is

$$g(y; \beta) = \begin{cases} (y^\beta - 1)/\beta & \text{if } p \neq 1 \\ \ln(y) & \text{if } p = 1 \end{cases}$$

and this is continuous in y and β .

The Binomial Distribution

We'll confirm here the variance stabilizing transformation for the binomial proportion derived by Fisher, which is $g(\hat{p}) = 2 \arcsin(\sqrt{\hat{p}})$. First let's find the derivative of the $y = \arcsin(x)$ function wrt x :

$$y = \sin^{-1}(x)$$

$$\sin(y) = x$$

$$y' \cos(y) = 1$$

$$y' = \frac{1}{\cos(y)} = \frac{1}{\sqrt{1 - \sin^2(y)}} = \frac{1}{\sqrt{1 - x^2}}$$

$$\frac{d}{dx} 2 \sin^{-1}(\sqrt{x}) = \frac{x^{-1/2}}{\sqrt{1 - x}}$$

$$[g'(p)]^2 \sigma^2 = \frac{p^{-1}}{(1-p)} \frac{p(1-p)}{n} = n^{-1}$$

Zinc Data

The zinc data consist of test runs of EPA method 1638 ICPMS from spiked samples with 11 different concentrations from 0 to 25,000 $\mu\text{g}/\text{L}$.

```
> summary(zinc)
```

Concentration	Peak.Area
Min. : 0	Min. : 93
1st Qu.: 100	1st Qu.: 1187
Median : 500	Median : 4200
Mean : 4387	Mean : 31725
3rd Qu.: 5000	3rd Qu.: 34942
Max. : 25000	Max. : 189657

```
> dim(zinc)
```

```
[1] 91 2
```

```
> concs <- sort(unique(zinc$Concentration))
```

```
[1] 0 10 20 100 200 500 1000 2000 5000 10000 25000
```

```
> counts <- table(zinc$Concentration)
```

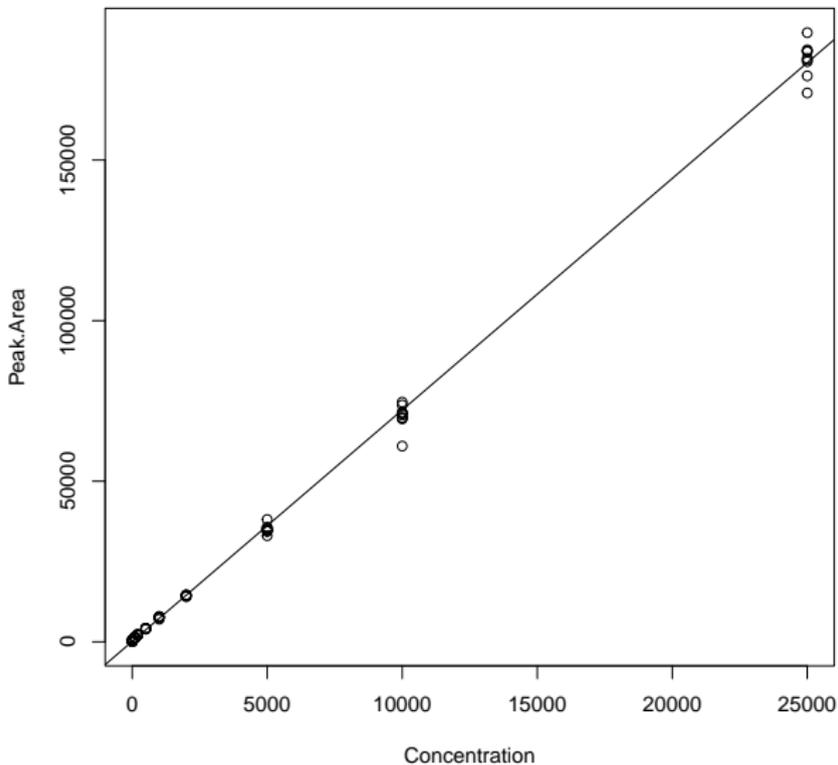
0	10	20	100	200	500	1000	2000	5000	10000	25000
8	7	7	11	7	7	9	7	9	10	9

Zinc Data

```
> mns <- tapply(zinc$Peak.Area,zinc$Concentration,mean)
      0      10      20      100      200      500
264.7500 316.8571 692.4286 1291.3636 2187.8571 4109.2857
 1000  2000  5000 10000 25000
7589.4444 14388.5714 35072.3333 70267.2000 181354.4444

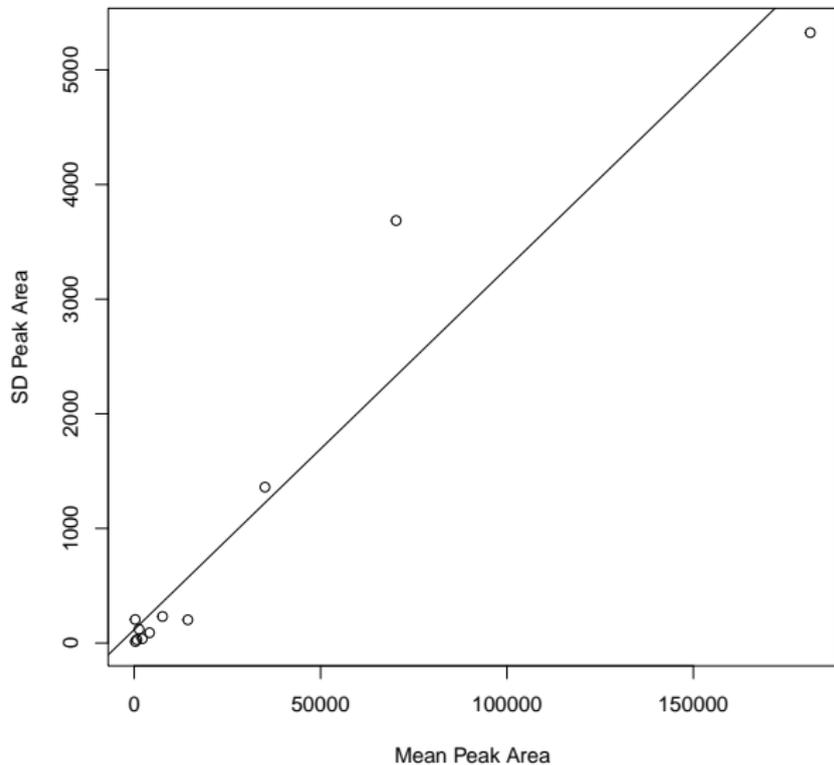
> vars <- tapply(zinc$Peak.Area,zinc$Concentration,var)
      0      10      20      100      200      500
4.203907e+04 1.714762e+02 7.869524e+02 1.438745e+04 1.431810e+03 8.074238e+03
 1000  2000  5000 10000 25000
5.365478e+04 4.098095e+04 1.849742e+06 1.358443e+07 2.835566e+07
```

EPA Zinc Data

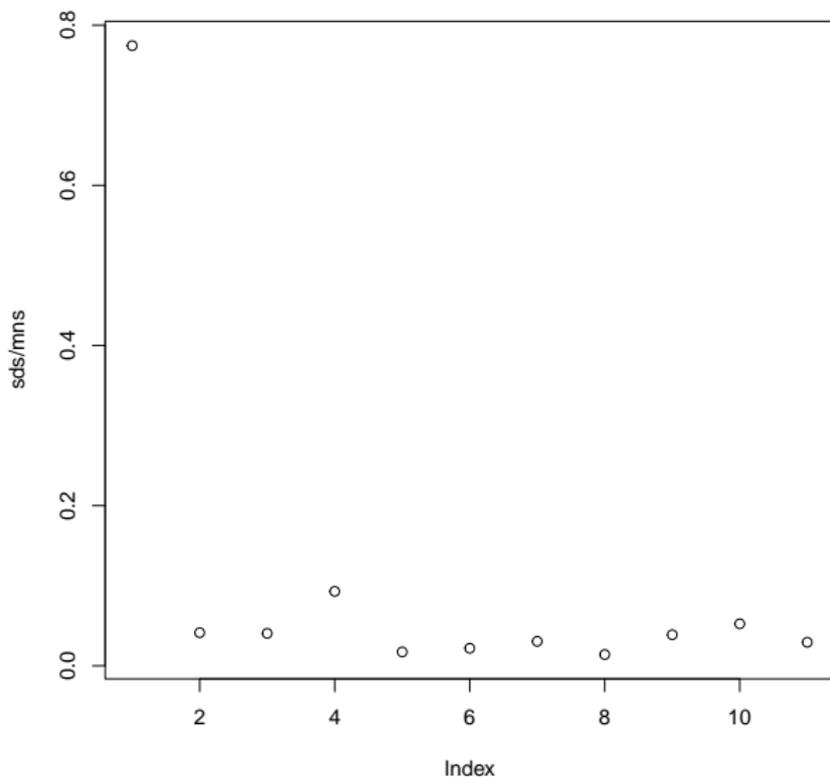


Plot of peak area vs. concentration from ICPMS along with linear calibration curve. Variance appears to increase with the mean. If the CV is nearly constant, we can take logs to stabilize the variance..

EPA Zinc Data



This is roughly linear.
Standard deviations
from only 5–10 points
are quite variable.



This is the ratio of the standard deviation to the mean for the 11 concentrations. The mean of the last 10 is 0.038 which is the average CV of those groups. For zero concentration, the ratio is higher as would be expected.

For the non-zero concentrations, the CV is roughly constant, so the log transform should work.

The standard deviation of the zero-concentration samples is 205 and the CV of the remainder is 0.0379, so the overall variance of the peak areas is

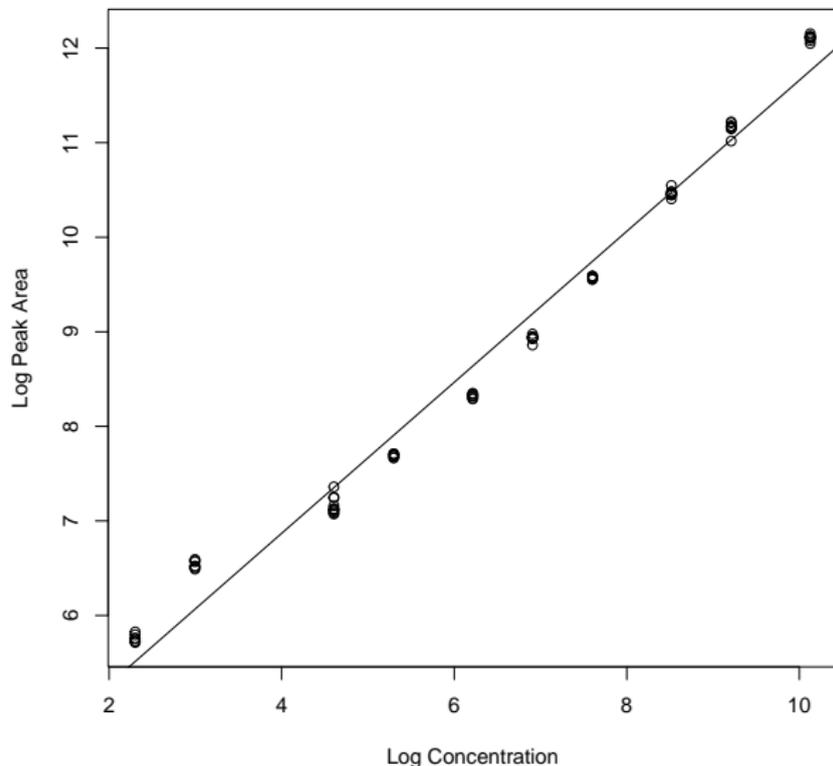
$$V(Y) = h(\mu) = 205^2 + (0.0379)^2 \mu^2.$$

It can be shown that if $h(\mu) = a^2 + b^2 \mu^2$, then a variance-stabilizing transformation is

$$g(y) = \ln(y^2 + \sqrt{y^2 + a^2/b^2})$$

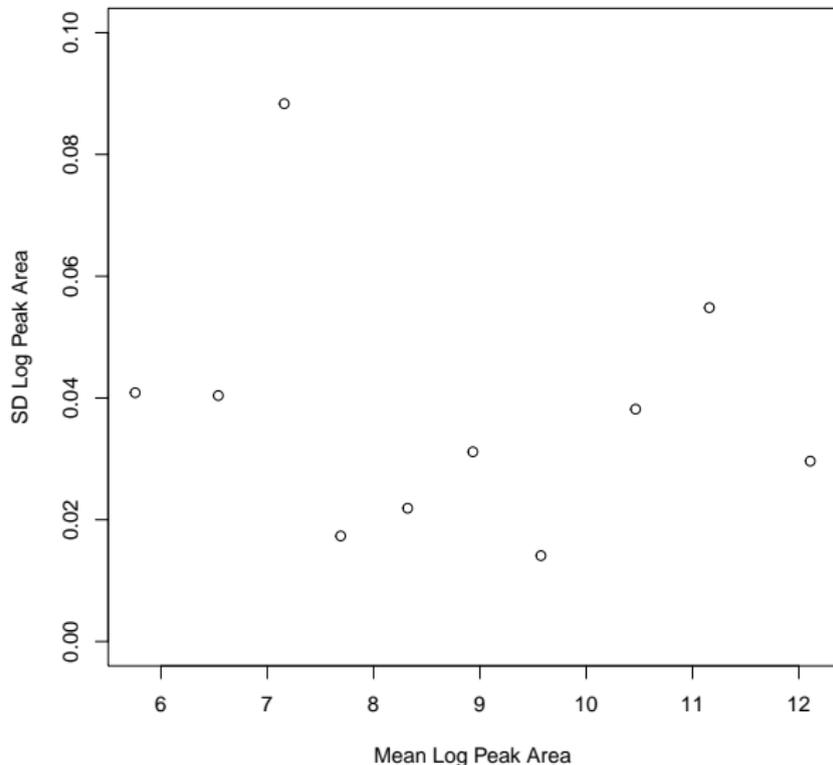
but this might disturb the linearity of the calibration curve, so we will just use the log transform on the non-zero data.

EPA Zinc Data Log Transformed



The variance no longer seems to rise with the mean. We can confirm this with a plot of the variance vs. the mean of each non-zero concentration.

EPA Zinc Data Log Log Transformed



The standard deviation is on the average close to the previous average CV of 0.038.

Poison Data

These are data from a toxicology study on three poisons and four possible treatments with the response being the survival time of the animal in minutes (unit change from the text).

Goals are to determine if the treatments differ in prolonging life, if the poisons differ in toxicity, and if different treatments are more effective against one poison than another.

	A	B	C	D
I	186	492	258	270
	270	660	270	426
	276	528	378	396
	258	432	456	372
II	216	552	264	336
	174	366	210	612
	240	294	186	426
	138	744	240	228
III	132	180	138	180
	126	222	150	216
	108	228	144	186
	138	174	132	198

A quick look shows that some cells seem to have larger survival times than others; for example I/B and II/B. But we need a more systematic analysis via `lm()` and the analysis of variance.

```
> summary(lm(survTime ~ Poison*Treatment,data=poison))
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-195.00	-29.25	3.00	25.88	255.00

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	247.50	44.74	5.532	2.94e-06	***
PoisonII	-55.50	63.27	-0.877	0.3862	
PoisonIII	-121.50	63.27	-1.920	0.0628	.
TreatmentB	280.50	63.27	4.433	8.37e-05	*** Prolongs survival
TreatmentC	93.00	63.27	1.470	0.1503	
TreatmentD	118.50	63.27	1.873	0.0692	.
PoisonII:TreatmentB	16.50	89.48	0.184	0.8547	
PoisonIII:TreatmentB	-205.50	89.48	-2.297	0.0276	* except with poison III
PoisonII:TreatmentC	-60.00	89.48	-0.671	0.5068	
PoisonIII:TreatmentC	-78.00	89.48	-0.872	0.3892	
PoisonII:TreatmentD	90.00	89.48	1.006	0.3212	
PoisonIII:TreatmentD	-49.50	89.48	-0.553	0.5836	

```
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 89.48 on 36 degrees of freedom
```

```
Multiple R-squared:  0.7335,    Adjusted R-squared:  0.6521
```

```
F-statistic:  9.01 on 11 and 36 DF,  p-value: 1.986e-07
```

```
> drop1(lm(survTime ~ Poison*Treatment,data=poison),test="F")
```

Single term deletions

Model:

```
survTime ~ Poison * Treatment
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			288261	441.62		
Poison:Treatment	6	90050	378310	442.67	1.8743	0.1123

```
> drop1(lm(survTime ~ Poison+Treatment,data=poison),test="F")
```

Single term deletions

Model:

```
survTime ~ Poison + Treatment
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			378310	442.67		
Poison	2	371885	750195	471.53	20.643	5.704e-07 ***
Treatment	3	331634	709945	466.88	12.273	6.697e-06 ***

Interactions as a whole are not “significant” but both main effects are strongly significant.

```
> summary(lm(survTime ~ Poison+Treatment,data=poison))
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-151.000	-57.750	-8.937	37.063	299.000

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	271.38	33.55	8.088	4.22e-10	***
PoisonII	-43.87	33.55	-1.308	0.19813	
PoisonIII	-204.75	33.55	-6.102	2.83e-07	***
TreatmentB	217.50	38.75	5.614	1.43e-06	***
TreatmentC	47.00	38.75	1.213	0.23189	
TreatmentD	132.00	38.75	3.407	0.00146	**

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 94.91 on 42 degrees of freedom
```

```
Multiple R-squared:  0.6503,    Adjusted R-squared:  0.6087
```

```
F-statistic: 15.62 on 5 and 42 DF,  p-value: 1.123e-08
```

In the main-effects model, Poison III is more lethal and treatments D and especially B are better than treatment A.

```
mns <- with(poison, tapply(survTime,list(Poison,Treatment),mean))
```

	A	B	C	D
I	247.5	528	340.5	366.0
II	192.0	489	225.0	400.5
III	126.0	201	141.0	195.0

```
vars <-with(poison, tapply(survTime,list(Poison,Treatment),var))
```

	A	B	C	D
I	1737	9312	8841	4584
II	2040	40716	1164	26433
III	168	780	60	252

```
sds <-with(poison, tapply(survTime,list(Poison,Treatment),sd))
```

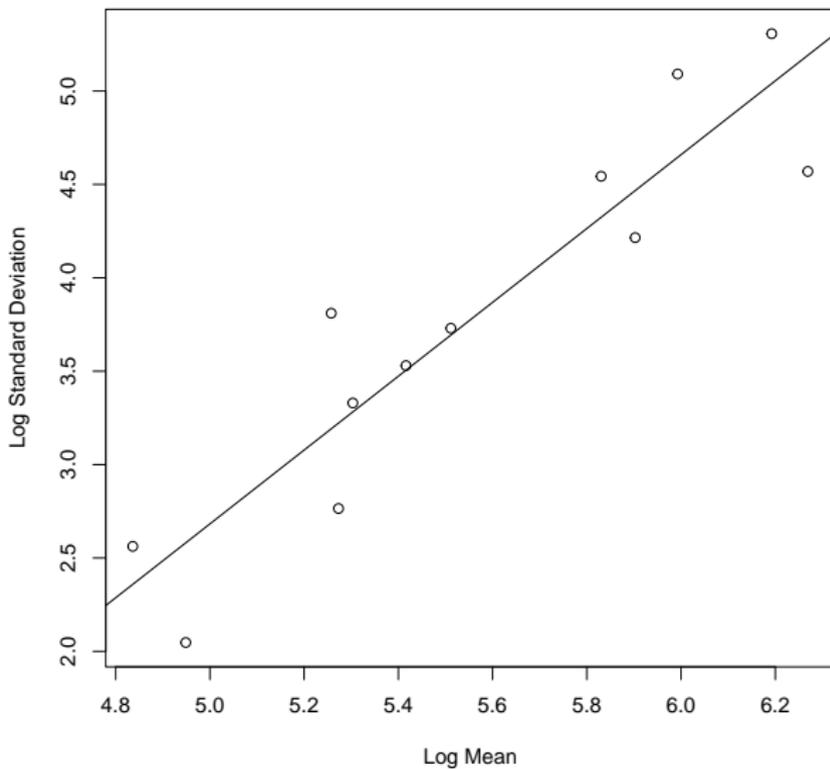
	A	B	C	D
I	41.67733	96.49870	94.026592	67.70524
II	45.16636	201.78206	34.117444	162.58229
III	12.96148	27.92848	7.745967	15.87451

So, let's examine the assumption of equality of variances. These vary from 40,716 down to 60, which suggests that perhaps there is heteroscedasticity. Let's see if the variance is a function of the mean.

Finding a Transformation

We have seen that if the standard deviation is proportional to a power of the mean, then that suggests a transformation. If the CV is constant, take logs. If the standard deviation is proportional to a different power $p \neq 1$ of the mean, then use a power transformation $g(y) = y^{1-p}$. If $\sigma \propto \mu^p$, then $\ln(\sigma) \approx p \ln(\mu)$, so let's compare the log standard deviation to the log mean and see what the slope is.

Log Standard Deviation vs. Log Mean



```
> summary(lm(log(sdvec) ~ log(mvec)))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-7.2029	1.4689	-4.904	0.00062	***
log(mvec)	1.9770	0.2633	7.509	2.04e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4113 on 10 degrees of freedom

Multiple R-squared: 0.8494, Adjusted R-squared: 0.8343

F-statistic: 56.39 on 1 and 10 DF, p-value: 2.041e-05

The slope of the log/log plot is $p = 1.977$, which suggests that a good transformation would be near $1 - p = -0.997$. We will use $g(y) = y^{-1} = 1/y$. I will multiply this by 10,000 to make the numbers easier to read

```

invsurv <- 10000/poison$survTime

> mnst <- with(poison, tapply(invsurv,list(Poison,Treatment),mean))
      A      B      C      D
I    41.44801 19.39107 31.04539 28.16137
II   54.47450 23.22320 45.23199 28.35890
III  80.04475 50.48288 71.08311 51.53009

> varst <-with(poison, tapply(invsurv,list(Poison,Treatment),var))
      A      B      C      D
I    68.52052 11.05536 66.52508 36.94870
II   187.83932 85.00486 48.42160 136.85308
III  77.92049 49.33560 15.31568 16.54518

> sdst <-with(poison, tapply(invsurv,list(Poison,Treatment),sd))
      A      B      C      D
I    8.277712 3.324960 8.156291 6.078545
II   13.705449 9.219808 6.958563 11.698422
III  8.827258 7.023930 3.913525 4.067576

```

The ratio of the largest variance to the smallest is $187.84/11.06 = 16.98$, while the ratio in the untransformed data is $40716/60 = 678.6$.

```
> summary(lm(invsurv~Poison*Treatment,data=poison))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	41.448	4.083	10.151	4.16e-12	***
PoisonII	13.026	5.775	2.256	0.030252	*
PoisonIII	38.597	5.775	6.684	8.56e-08	***
TreatmentB	-22.057	5.775	-3.820	0.000508	***
TreatmentC	-10.403	5.775	-1.801	0.080010	.
TreatmentD	-13.287	5.775	-2.301	0.027297	*
PoisonII:TreatmentB	-9.194	8.166	-1.126	0.267669	
PoisonIII:TreatmentB	-7.505	8.166	-0.919	0.364213	
PoisonII:TreatmentC	1.160	8.166	0.142	0.887826	
PoisonIII:TreatmentC	1.441	8.166	0.176	0.860928	
PoisonII:TreatmentD	-12.829	8.166	-1.571	0.124946	
PoisonIII:TreatmentD	-15.228	8.166	-1.865	0.070391	.

Residual standard error: 8.166 on 36 degrees of freedom

Multiple R-squared: 0.8681, Adjusted R-squared: 0.8277

F-statistic: 21.53 on 11 and 36 DF, p-value: 1.289e-12

The $III \times B$ interaction has disappeared, so the main effects model looks good.

```
> drop1(lm(survTime~Poison*Treatment,data=poison),test="F")
```

Single term deletions

Model:

```
survTime ~ Poison * Treatment
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			288261	441.62		
Poison:Treatment	6	90050	378310	442.67	1.8743	0.1123

```
> drop1(lm(invsurv~Poison*Treatment,data=poison),test="F")
```

Single term deletions

Model:

```
invsurv ~ Poison * Treatment
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			2400.9	211.79		
Poison:Treatment	6	436.33	2837.2	207.81	1.0904	0.3867

The interaction effect has the F-statistic dropping from 1.87 to 1.09 (with 1.00 being the center of the no evidence of effect range).

This, along with the coefficients, means that any evidence of an interaction effect has been removed by the transformation.

```
> drop1(lm(survTime~Poison+Treatment,data=poison),test="F")
Single term deletions
```

Model:

```
survTime ~ Poison + Treatment
      Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                378310 442.67
Poison   2    371885 750195 471.53  20.643 5.704e-07 ***
Treatment 3    331634 709945 466.88  12.273 6.697e-06 ***
```

```
> drop1(lm(invsurv~Poison+Treatment,data=poison),test="F")
Single term deletions
```

Model:

```
invsurv ~ Poison + Treatment
      Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                2837.2 207.81
Poison   2    9688.1 12525.3 275.09  71.708 2.865e-14 ***
Treatment 3    5670.6  8507.8 254.52  27.982 4.192e-10 ***
```

The F-statistic for poisons has risen from 20.6 to 71.7 with a large change in the p-value. The F-statistic for treatments has risen from 12.3 to 28.0 with a large change in the p-value. The transformation has increased the evidence of differences.

```
> summary(poison.lm)
```

```
Call:
```

```
lm(formula = invsurv ~ Poison + Treatment, data = poison)
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	44.961	2.906	15.473	< 2e-16	***
PoisonII	7.811	2.906	2.688	0.01026	*
PoisonIII	33.274	2.906	11.451	1.69e-14	***
TreatmentB	-27.623	3.355	-8.233	2.66e-10	***
TreatmentC	-9.536	3.355	-2.842	0.00689	**
TreatmentD	-22.639	3.355	-6.747	3.35e-08	***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 8.219 on 42 degrees of freedom
```

```
Multiple R-squared:  0.8441,    Adjusted R-squared:  0.8255
```

```
F-statistic: 45.47 on 5 and 42 DF,  p-value: 6.974e-16
```

All the coefficient tests are statistically significant, so Poison II and Poison III are each more toxic than Poison I (if inverse survival time is higher, then survival time is lower). Each of treatments B, C, and D are better than A.

Multiple Comparisons

The coefficients in the linear models for the poison data set are tests of the hypothesis that the given poison/treatment differs from the default (first) level. So we have tests of the difference between Poison II and Poison I and between Poison III and Poison I, but not between Poison III and Poison II. Similarly, we have tests of differences between each of Treatments B, C, and D and Treatment A, but no tests of the three differences within the the three non-default treatment levels. From the F-statistic on each factor, we know differences exist, but not the full list of significant differences.

The Treatment factor has four levels and therefore six distinct pairwise comparisons. Three of them are given in the output to `lm()` but we can compute the other three, as well as the one comparison among poisons that is not already given.

```

cmat <- matrix(c(0,0,0,0,-1,0,0,0,1,0,0,0,0,-1,-1,0,0,1,0,-1,0,0,1,1),ncol=6)
      [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0   -1    1    0    0    0      Compare Poison II to Poison III
[2,]    0    0    0   -1    1    0      Compare Treatments B and C
[3,]    0    0    0   -1    0    1      Compare Treatments B and D
[4,]    0    0    0    0   -1    1      Compare Treatments C and D
xp <- cmat %*% coefp
      [,1]
[1,] 25.463061
[2,] 18.087782
[3,]  4.984402
[4,] -13.103381

varx <- diag(cmat %*% vcvp %*% t(cmat))
> varx
[1]  8.443994 11.258659 11.258659 11.258659
sdx <- sqrt(varx)
[1] 2.905855 3.355393 3.355393 3.355393
> t(xp)/sdx
      [,1]      [,2]      [,3]      [,4]
[1,] 8.762674 5.390661 1.48549 -3.905171

```

The last computation is the t-scores on 42 df for the four computed comparisons.

```

> coef(summary(poison.t.lm))
              Estimate Std. Error  t value    Pr(>|t|)
(Intercept)  44.960943   2.905855  15.472534 5.936991e-19
PoisonII     7.810688    2.905855   2.687914 1.026221e-02
PoisonIII    33.273749    2.905855  11.450587 1.690903e-14
TreatmentB  -27.623373    3.355393  -8.232531 2.655965e-10
TreatmentC   -9.535591    3.355393  -2.841870 6.892762e-03
TreatmentD  -22.638971    3.355393  -6.747041 3.347340e-08

> t(xp)/sdx
           [,1]      [,2]      [,3]      [,4]
[1,]  8.762674  5.390661  1.48549 -3.905171

```

In addition to the previous conclusions, Poison III is more toxic than Poison II. Treatment B is better than Treatment C, Treatment D is better than Treatment C, but Treatments B and D are not significantly different. Poisons more toxic to less toxic: III, II, I. Treatments better to worse: (B, D), C, A.