Homework 6 Solutions
401 Fall 2001
Brian Junker and Kimberly Sellers

Thursday November 15, 2001

1. RwG problem 16 chapter 5 page 180.

   The first test in this problem is to find decent initial values. Recall that the negative exponential model has the form $Y_i = \alpha(1 - e^{-\beta x_i}) + \epsilon_i$. Thus, initial values for both α and β are needed. Recall from the book (p164) that α represents the asymptotic limit. This can also be seen by letting $x \to \infty$. From the data set, the largest value for Y is 14, so a good initial value for α is 14. For β it is a bit trickier. Here, I used the Taylor Series expansion of $e^x \approx 1 + x$. Thus, $Y_i \approx \alpha(1 - (1 - \beta x_i)) = \alpha(1 - (1 - \beta x_i)) = \alpha(\beta x_i)$. Next, solve this for β to get $\beta \approx \frac{Y_i}{\alpha x_i}$. Next plug in the estimate for α which is 14. Then take some sample (Y_i, x_i) pairs to get an idea for the intial estimate of β. For example $\beta \approx 23/(14 * 10) = 0.031$, or $\beta \approx 12.5/(14 * 37) = 0.024$, or $\beta \approx 8/(14 * 70) = 0.008$.

   Next, these initial estimates are used with the `nls()` command to obtain some parameter estimates. I first used the beginning values of $\alpha = 14$ and $\beta = 0.024$. I did this because 0.024 appeared to be about the middle estimate of the guesses I reviewed above. Playing with the initial value of β indicates that it is relatively insensitive to the intial value. That is, β values of 0.008, 0.03, 1, and 3 all provided the same parameter estimates.

   Similarily, setting β to 0.024, and letting α vary indicates that α is relative insensitive to different parameter estimates also. I let α vary between 1, 5, 20, and 50. All of these initial values provided the same parameter estimates.

```
>eggs.dat_read.table("../HW6/eggs.dat", header=T)
> neg.exp.egg_nls(ratio~alpha*(1-exp(-beta*density)), data=eggs.dat,
  start=list(alpha=14, beta=0.024), trace=T)
117.269 : 14 0.024
111.261 : 10.8257 0.0359856
76.6683 : 10.3768 0.0532978
73.1699 : 10.7843 0.0552609
73.1694 : 10.7741 0.0554083
> summary(neg.exp.egg)

Formula: ratio ~ alpha * (1 - exp( - beta * density))

Parameters:
           Value Std. Error  t value
alpha 10.7741000  0.8511820 12.65780
```

```
 beta  0.0554083  0.0115094  4.81417

Residual standard error: 1.74606 on 24 degrees of freedom

Correlation of Parameter Estimates:
      alpha
beta -0.814
```

From the above read-out, the estimate for the α parameter is 10.774 and the estimate for the β parameter is 0.055. From this, we can calculate an R^2 value of $1-((1.746)^2*24)/\text{sum}((\text{ratio-mean(ratio)})^2) = 0.7710$ This is an acceptible fit!.

Next, make some plots of the fit to evaluate the model. Figure 2 contains the data along with the fitted line. From this graph we can see that there is a bit of variance in the original data itself. The fitted line appears to fall in the middle of the data.

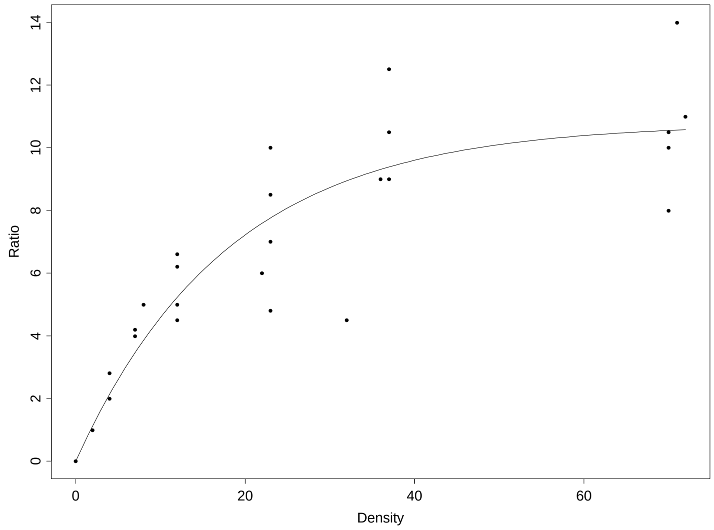

Figure 1: Simple preditor/Prey System Data vs Fitted Line

The next plot in figure 2 contains the residuals vs. the fitted values. From this figure we see that there is quite a bit of non-constant variance. Looking again at figure 2, it is also clear that there is non constant variance. While this line appears to give a good fit, it does not meet the regression assumption of non-constant variance.

Finally, check the normal quantile plot in figure 3. Here, we can see that the normality assumption appear to be violated. This is a bit odd for a textbook problem!

I imagine that they were looking for a problem to put in that was resistant to intial values for the parameters. In doing so, they provided a data set which does not meet the regression assumptions!

2. RwG Problems 5 and 6 chapter 6 page 213.

   - Perform OLS and robust regression of distance covered (Y) on relative dose (X). Graph the results.

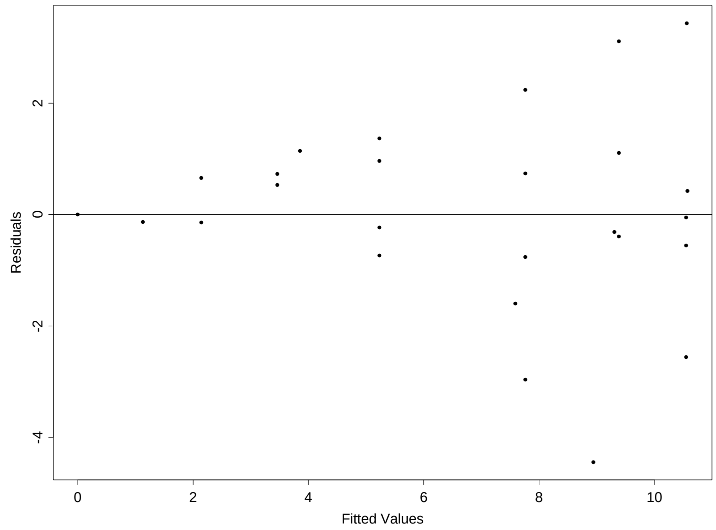

Figure 2: Simple preditor/Prey System Data – Residuals vs. Fitted

First, I created a scatterplot of the data to get an idea of its structure. See figure 4. From this scatterplot, it may be a bit difficult to discern a pattern. If the large distance value of 3000 is not taken into account, there appears to be a slight decreasing trend in the data. Taking that large data point into account, there may be a slight positive trend, or more of a flat line. Next we use OLS and robust regression to analyze this problem.

The results of the OLS regression are below.

```
> summary(kob.ols)

Call: lm(formula = distance ~ dosage, data = kob.dat)
Residuals:
    Min      1Q Median   3Q  Max
 -672.3 -425.1 -231.4 29.7 2063

Coefficients:
                  Value  Std. Error     t value    Pr(>|t|)
(Intercept)   1597.6664   1144.3368      1.3962      0.2053
     dosage -28244.0909  51292.3748     -0.5506      0.5990

Residual standard error: 886.1 on 7 degrees of freedom
Multiple R-Squared: 0.04152
F-statistic: 0.3032 on 1 and 7 degrees of freedom, the p-value is 0.599

Correlation of Coefficients:
       (Intercept)
dosage -0.9661
```

From this, we can see that it appears that neigher of the regression coefficients are significant. This is again reflected in the large p-value of the F-test. That is, it appears that knowledge of the

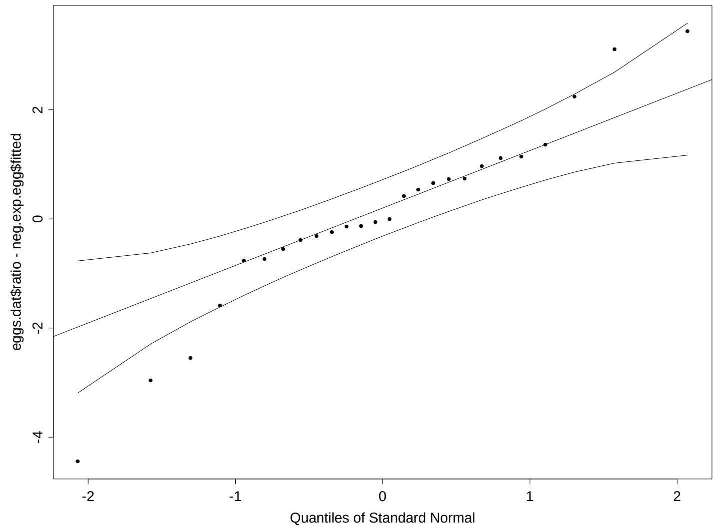

Figure 3: Simple preditor/Prey System Data – Quantile Plot

dart dosage has no predictive effect on the distance the kob ran before dropping.

Next we check the residuals in Figure 5. The fitted vs residuals plot indicates that there is one very large residual. This residual probably reflects the large 3000 value noted earlier. This value is seen as an outlier in all the graphs, creating problems with all the residuals assumptions. We can also see that the distribution of the residuals does not appear normal. It has heavy tails which suggest that a robust regression may be appropriate.

As can be seen in the Cook's distance plot in Figure 5, the cooks distance is very large for two data points. Figure 6 is the residual vs fitted values with the magnitude of the cooks distance proportional to the size of the circle for each data point. We can see that data points 4 and 5 both have large cooks distance.

Because of the poor residual analysis, it appears that the regression statistics are not valid. (i.e. the t and F statistics are not reliable).

Often, it is good to report OLS results with and without the "outliers". For now, (because of our hint in problem 6-6), I'll consider data point 4 to be an outlier. The OLS without the fourth data point is below. Without the "outlier", we can see that both the intercept and the dosage are significant in predicting the distance the kob has run. The p-value on the whole regression is now 0.06, much larger than 0.559 when including the outlier.

```
> kob.ols2_lm(kob.dat$distance[-4]~kob.dat$dosage[-4])
> summary(kob.ols2)

Call: lm(formula = kob.dat$distance[-4] ~ kob.dat$dosage[-4])
Residuals:
    Min      1Q Median    3Q Max
 -428.3 -151.8 -35.71 108.6 567

Coefficients:
                             Value  Std. Error     t value    Pr(>|t|)
```

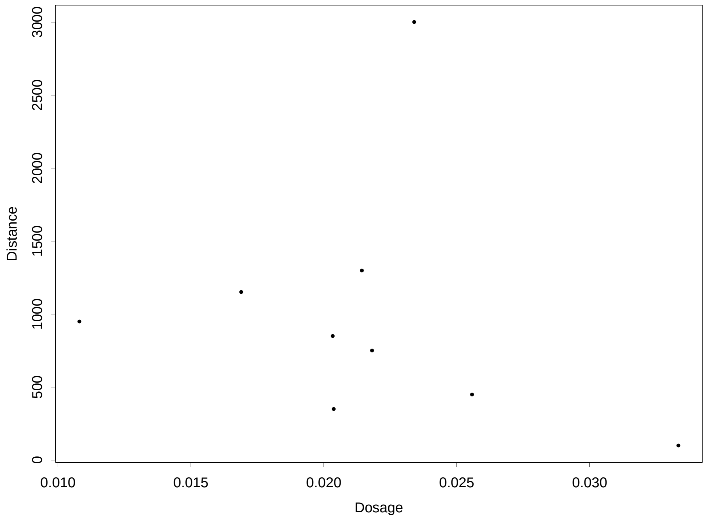

Figure 4: Western Kob Data – Scatter Plot

```
       (Intercept)   1650.6248    423.0347      3.9019      0.0080
kob.dat$dosage[-4] -42823.6941  19081.8293     -2.2442      0.0660

Residual standard error: 327.5 on 6 degrees of freedom
Multiple R-Squared: 0.4563
F-statistic: 5.036 on 1 and 6 degrees of freedom, the p-value is 0.06596

Correlation of Coefficients:
                   (Intercept)
kob.dat$dosage[-4] -0.9618
```

The summary plots are in figure 7. From these summary plots, we can see that there are now three additional points with high influence via Cooks Distance. The normal probability plot still indicates some heavy tails, though it is not too bad. It appears that our best bet is to go on to the robust regression.

```
My Splus code is below.

> kob.dat_read.table("../HW6/kob.dat", header=T)
> library(MASS)
> plot(kob.dat$dosage, kob.dat$distance, xlab="Dosage", ylab="Distance")
> kob.ols_lm(distance~dosage, data=kob.dat)
> summary(kob.ols)
> par(mfrow=c(2,3))
> plot(kob.ols)
> CookPlot(kob.ols, id=T)
```

Next I evaluate this using Robust Regression. The regression summary is below.

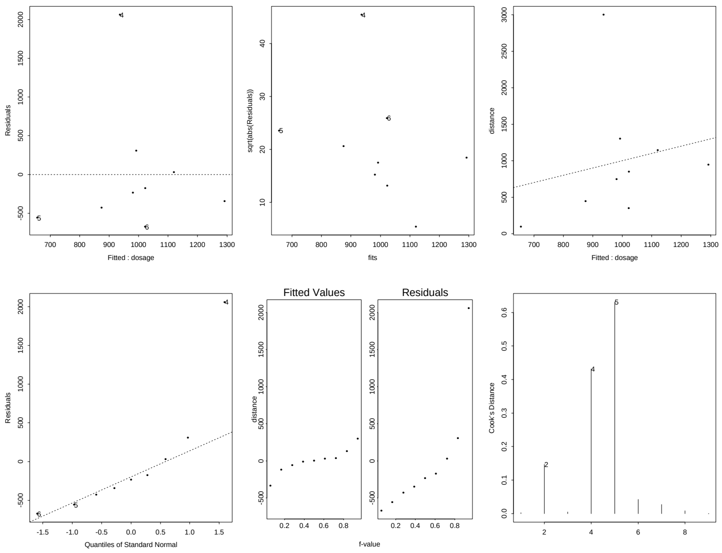

Figure 5: Western Kob Data – OLS Residual Plots

```
> kob.rob_rlm(distance~dosage, data=kob.dat)
> summary(kob.rob)

Call: rlm(formula = distance ~ dosage, data = kob.dat)
Residuals:
    Min     1Q Median  3Q  Max
 -466.8 -192.9 -8.273 193 2306

Coefficients:
                  Value  Std. Error     t value
(Intercept)   1640.0848    634.9771      2.5829
     dosage -40416.3178  28461.4492     -1.4200

Residual standard error: 286.4 on 7 degrees of freedom
Multiple R-Squared: NA
F-statistic: on and degrees of freedom, the p-value is NA

Correlation of Coefficients:
       (Intercept)
dosage -0.9661
```

The residual diagnostic plots are in figure 8. In the residual vs fitted values we can see that there is one large residual, probably denoting the fo0urth data point. The normal probability plot looks fairly normal, except for the one large upper quantile. This again probably reflects the fourth data points. We can see that the one large residual received a much lower weight than the rest of the data points. Most of the data points have a weight of 0.75 or above, however the one outlier residual has a weight under 0.2. Thus, we can see that the one data point does not have the same relative effect on the regression line as the other data points.

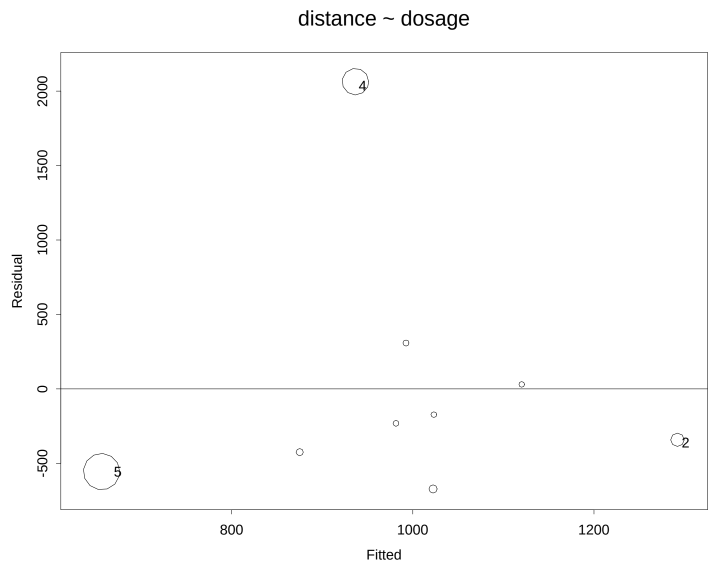

Figure 6: Western Kob Data – OLS Cooks Distance

- With kob #4, the tranquilize r dart failed to administer the intended dosage; the animal actually received only about half the relative dose shown in the data. Hence the unusual Y value for this case reflects an experimtenatl error. How do your OLS and robust analyses differ in their treatment of kob #4?

  The comparison of the three methods are summarized in figure 9. From this, we can see that the OLS regression line has a slope closest to zero. This is because of the one influential point which is pulling the line closer to zero. When the outlier is removed from the OLS regression, we can see that the line has a much more negative slope. The robust regression line is in between those two regression lines. This reflects the fact that the OLS regression line gives the inflential point a weight of 1, the OLS w/o the outlier regression line gives the influential point a weight of 0, and the robust regression line gives that point a weight in between zero and one.

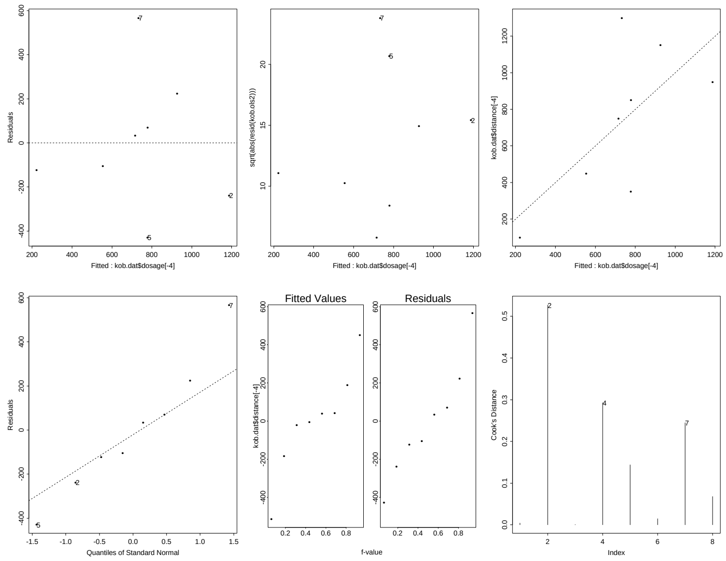

Figure 7: Western Kob Data – OLS Residual Plots w/o Outliers

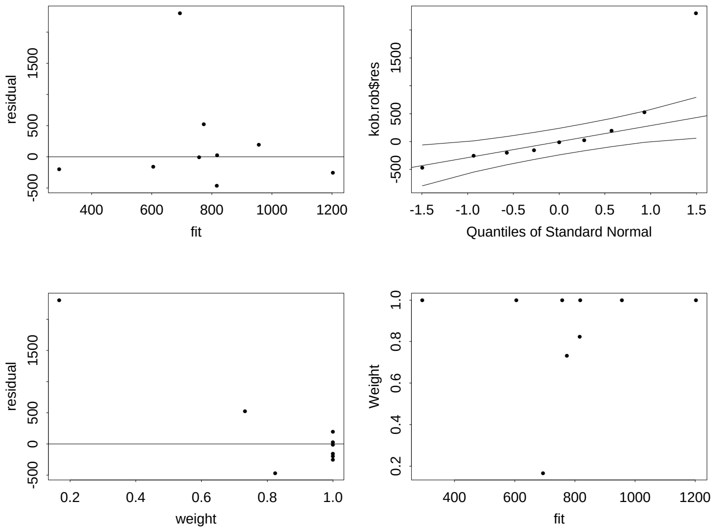

Figure 8: Western Kob Data – Robust Residual Diagnostics

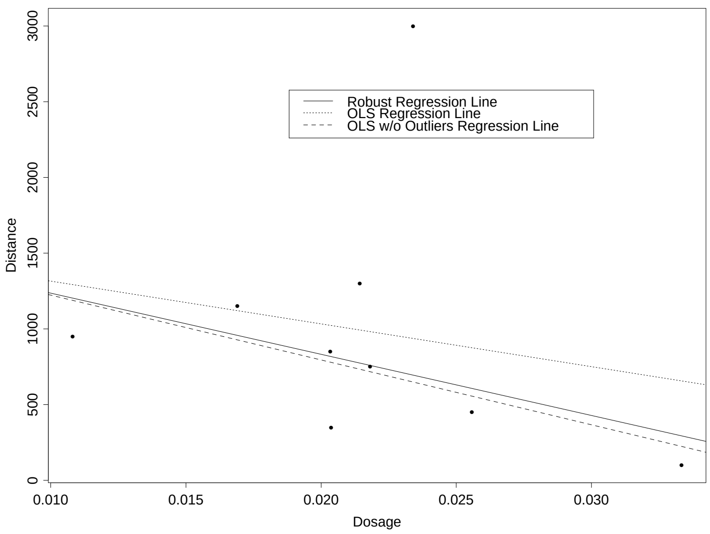

Figure 9: Western Kob Data – Summary