ST505/697R: Fall 2012. Homework 5 Solution.

1. **a and b**. Here are means, standard deviations and sample sizes associated with each of the levels of virus density. There is a clear decline in proportion surviving and the data suggests that the standard deviation among observation is smaller when the mean is smaller. This is not unexpected. Each observation was the proportion of gypsy moth in a bag that survived. If the bag had m individuals and observations across individuals in a bag were independent (a bit assumption) and the probability of surviving were π then the variance of the proportion surviving would be $\pi(1-\pi)/m$ (from binomial results) which has a maximum at $\pi = .5$ and declines as π goes to 0 or 1. While we don't want to necessarily accept that variance model is does suggest in part why the sd's go down with the mean. Later we will do a weighted analysis that allows the variances to change but without modeling it.

```
R:
data<-read.table("e:/s505/data/virus.dat")  #no names
attach(data)
tree<-V1; vden<-V2; totno<-V3; inf<-V4
group<-factor(vden)
x<-1/vden
psurv = 1- (inf/totno)
 gmean<-tapply(psurv,group,mean)
 gsd<-tapply(psurv,group,sd)
 n<-tapply(psurv,group,length)
 groupsum<-cbind(mean=gmean,st.dev=gsd,samplesize=n)
 groupsum
       mean    st.dev samplesize
5  0.6902083 0.2335024          8
10 0.4382847 0.2548244          8
25 0.1746658 0.1761442          8
50 0.1552835 0.1112793          8
70 0.1293478 0.1338735          8

SAS
data a;
infile 'e:/s505/data/virus.dat';
input tree vden totno inf;
psurv=1-(inf/totno);
x = 1/vden; run;
proc means;
class vden; var psurv; run;
                          Analysis Variable : psurv
                N
         vden  Obs   N          Mean       Std Dev       Minimum       Maximum
            5    8   8     0.6902083     0.2335024     0.3200000     0.9583333
           10    8   8     0.4382847     0.2548244     0.1600000     0.7916667
           25    8   8     0.1746658     0.1761442             0     0.4400000
           50    8   8     0.1552835     0.1112793             0     0.3200000
           70    8   8     0.1293478     0.1338735             0     0.3500000
```

c: The F-statistic i the anova table (F = 13.03 with a small p-value) is testing the null hypothesis that $\theta_1 = \theta_2 = \ldots = \theta_5$ where θ_j is the population mean associated with the jth level of vden; i.e. $\theta_j = E(Y)$ for a Y observed at the jth level of vden. This makes no assumption about how this expected value relates to vden. We reject the null since the p-value is $< .0001$.

```
R
```

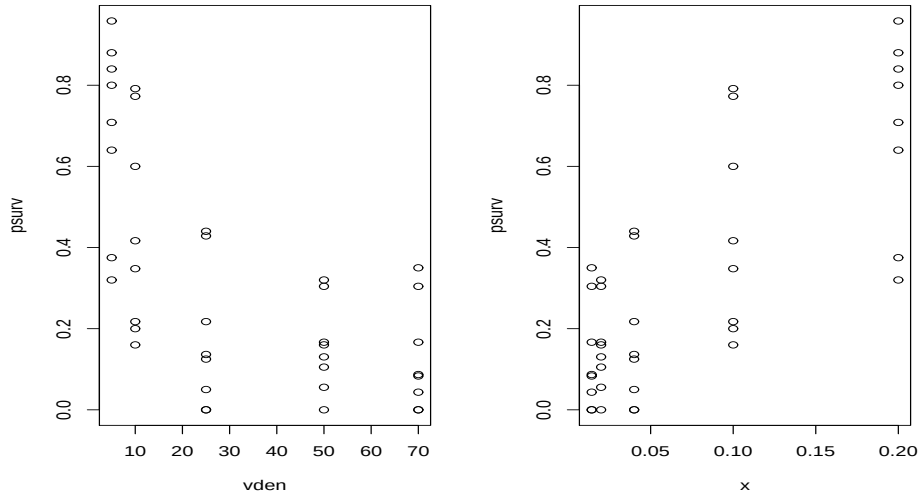

Figure 1: Plot psurv versus vden and x = 1/vden.

```
oneway<-lm(psurv~group)
anova(oneway)

Analysis of Variance Table
Response: psurv
          Df Sum Sq Mean Sq F value    Pr(>F)
group      4 1.8849 0.47123  13.033 1.342e-06 ***
Residuals 35 1.2655 0.03616

proc anova;
class vden;
model psurv=vden;
means vden/hovtest=levene; /* gives result for part d*/
run;
                              The ANOVA Procedure
Dependent Variable: pinf
                                        Sum of
 Source                       DF        Squares    Mean Square   F Value   Pr > F
 Model                         4     1.88493987     0.47123497     13.03   <.0001
 Error                        35     1.26553574     0.03615816
 Corrected Total              39     3.15047561
```

Part d: In the model that just allows different means at each level of the virus density, test the hypothesis that the variance is constant across the vden groups.

In SAS, the hovtest=levene in previous part gives the test based on the squared residuals (output below). This has a P-value of .0425, leading to rejection at $\alpha = .05$. In R you can use Bartlett's test which is automatic, or run a one-way analysis on the squared or absolute residuals, which gives two versions of Levene's test. You'll see that the test using squared residuals is equivalent to what SAS runs as Levene's test. The Levene's test based on absolute residuals has a p-value of .067, but Bartlett's test, with a P-value of .1875 leads to a different conclusion in this case than. It is best here to try and accommodate unequal variances.

```
                The ANOVA Procedure
                Levene's Test for Homogeneity of psurv Variance
                 ANOVA of Squared Deviations from Group Means
                               Sum of        Mean
          Source        DF     Squares      Square    F Value    Pr > F
          vden           4      0.0128     0.00320       2.77    0.0425
          Error         35      0.0405     0.00116
```

```
R.
# Can use bartlett.test which is automatic
# or can use residuals from fit with group means and run
# one-way analysis on the absolute residuals.
residg<-residuals(oneway)
abresidg<-abs(residg)
anova(lm(abresidg~group))  #Levene's test for equal variance in group mean
                           # mean model based on absolute residuals.
resid2<-residg^2
anova(lm(resid2~group))   # Levene's test for equal variance using squared residuals
bartlett.test(psurv,group)   # Bartlett's test for equal variance in model

Analysis of Variance Table
Response: abresidg
          Df  Sum Sq   Mean Sq F value  Pr(>F)
group      4 0.09068 0.0226692  2.4185 0.06691 .
Residuals 35 0.32806 0.0093731

Analysis of Variance Table
Response: resid2
          Df   Sum Sq   Mean Sq F value  Pr(>F)
group      4 0.012799 0.0031996  2.7669 0.04247 *
Residuals 35 0.040474 0.0011564

        Bartlett test of homogeneity of variances
data:  psurv and group
Bartlett's K-squared = 6.1599, df = 4, p-value = 0.1875
```

Part e: Use a plot of psurv versus vden) AND a test for lack of fit to assess whether a simple linear regression model of psurv on vden is adequate here.

The model does not look good from the plot. For testing lack of fit, from the anova of the regression model

SSE = 1.896384 with 38 dof. Also, SSPE = 1.26553574 (with 35 degrees of freedom)

$c = 5$ and $n = 40$, $SSLF = SSE - SSPE$, $MSLF = SSLF/(c-2)$. This leads to
$F_{lof} = MSLF/MSPE = 5.81984$ with 3 and 35 degrees of freedom.
The P-value (obtained via probf in SAS or pf in R) is .00246. [Using tables in the book, we have $F(.995, 3, 30) = 5.24$ and $F(.995, 3, 60) = 4.73$). $F(.995, 3, 35)$ will be somewhere in between, so we know that 5.81984 is more than $F(.99, 3, 5)$ so the p-value is less than .005]

$H_0 : \theta_j = \beta_0 + \beta_1 X_j$ is rejected, the linear regression model in terms of vden is not an adequate fit. This is also seen from the plot.

SAS

```
proc reg;
model psurv=vden;
plot psurv*vden;
run;
                              The REG Procedure
                            Dependent Variable: pinf
                                       Sum of           Mean
 Source                    DF        Squares         Square    F Value    Pr > F
 Model                      1        1.25363        1.25363      25.11    <.0001
  Error                     38        1.89684        0.04992
 Corrected Total           39        3.15048

COMPUTING THE LACK OF FIT TEST IN SAS.
 data lof;
 sspe= 1.26553574;  sse=1.89684;
 n=40;   c=5;
 mse=sse/(n-2);   mspe=sspe/(n-c);
 sslf = sse-sspe;   mslf=sslf/(c-2);
 f=mslf/mspe;
 fpvalue= 1 - probf(f,c-2,n-c);
 proc print;

 Obs   sspe    sse    n c    mse     mspe     sslf    mslf     f       fpvalue
  1  1.26554 1.89684 40 5 0.049917 0.036158 0.63130 0.21043 5.81984 .002456857

COMPUTING THE LACK OF FIT TEST IN R
sspe<-deviance(oneway)
dfpe<-df.residual(oneway)
mspe<-sspe/dfpe
regout<-lm(psurv~vden)
anova(regout)
sse<-deviance(regout)
dfe<-df.residual(regout)   # dof for SSPE in linear regression = n - 2
sslf<-sse-sspe
dflf<-dfe-dfpe     # dof for sslf =  n-2(n-c) = c-2
mslf<-sslf/dflf
Flof<-mslf/mspe    # the f statistic for testing lack of fit
pvalue<- 1 - pf(Flof,dflf,dfpe)  # gets p-value for lack of fit test
                                 # = area to the right of Flof for
                                 # t with dflf = c-2 and dfpe = n -c
                                 # degrees of freedom.
cat("lack of fit test using vden", "sspe =", sspe, "dfpe = ", dfpe, "sse =", sse, "dfe = ", dfe,
"Flof = ", Flof, "P-value = ", pvalue, "\n")

Response: psurv
          Df Sum Sq Mean Sq F value    Pr(>F)
vden       1 1.2536 1.25363  25.114 1.281e-05 ***
Residuals 38 1.8968 0.04992

lack of fit test using vden sspe = 1.265536 dfpe =  35 sse = 1.896841 dfe =  38 Flof =  5.81985 P-value =  0.00245683
```

Part f. Repeat the previous problem but now considering regressing psurv on x = 1/vden.

The plot of psurv on x = 1/vden, suggests a simple linear regression model is a reasonable fit. The regression of psurv on x yields SSE = 1.29224. The lack of fit test has F = .2462 with 3 and 35 degrees

of freedom, with a p-value of .8635. Do not reject the model that has $E(Y)$ linear in $x = 1/vden$.

```
                      Analysis of Variance
                                      Sum of           Mean
 Source                    DF        Squares         Square    F Value    Pr > F
 Model                      1        1.85823        1.85823      54.64    <.0001
 Error                     38        1.29224        0.03401
 Corrected Total           39        3.15048

           lack of fit for model with x=1/vden
 Obs   sspe    sse    n c    mse     mspe     sslf      mslf      f    fpvalue
  1  1.26554 1.29224 40 5 0.034006 0.036158 0.026704 .00890142 0.24618 0.86348

USING R

lack of fit test using 1/vden sspe = 1.265536 dfpe =  35 sse = 1.292244 dfe =  38 Flof =  0.2462177 P-value =  0.8634
```

g Assuming the linear model for psurv on x = 1/vden is good, test for constant variance, by running Levene's test. There are different versions of Levene's test that get used here as seen in the class example. We first fit the regression model and save the residuals. There are then three ways to go 1.run a one-way analysis comparing means but with the response being the absolute residual; 2. like 1 but using squared residuals. 3.In SAS you can just run a one-way anova with the residual (not squared or absolute value) and use the hovtest=levene result. These are all testing the hypothesis of equal variances of the errors with the regression framework with grouped data. (Another option is to get medians and do Brown-Forsythe). There is no best test here. In this problem, there are conflicting answers from the strict testing perspective as the p-values are .1173,.0810 and .0425 for 1 to 3, respectively. Note that 1 is border-line significant if one use $\alpha = .10$ which many people do in screening assumptions like this. As earlier, we are best served by trying to accommodate unequal variances.

```
SAS
proc reg data=a;
model psurv=x;
plot psurv*x;
output out=result r=resid;
run;
proc anova data=result;
class x;
model resid=x;
means x/hovtest=levene;
run;
data b;
set result;
absr = abs(resid);
r2 = resid**2;
run;
proc anova data=b;
class x;
model r2=x;
run;

     The ANOVA Procedure
```

```
        Levene's Test for Homogeneity of resid Variance
         ANOVA of Squared Deviations from Group Means
                        Sum of        Mean
Source        DF     Squares      Square    F Value    Pr > F
x              4      0.0128     0.00320       2.77    0.0425
Error         35      0.0405     0.00116
```

Using R

```
 resid<-residuals(lm(psurv~x))   # residuals from regression fit
 r2<-resid^2
 ar<-abs(resid)
 anova(lm(r2~group))   # Levene's test using squared residuals

Analysis of Variance Table
Response: r2
          Df   Sum Sq   Mean Sq F value  Pr(>F)
group      4 0.013510 0.0033775   2.273 0.08099 .
Residuals 35 0.052006 0.0014859

 anova(lm(ar~group))   # Levene's test using absolute residuals

Analysis of Variance Table
Response: ar
          Df  Sum Sq  Mean Sq F value Pr(>F)
group      4 0.08275 0.020687  1.9917 0.1173
Residuals 35 0.36352 0.010386
```

2. Using the residuals from the fit of psurv on x = 1/den, plot them versus tree number. Does it look like the model should account for tree effects in some manner?

   There is some suggestion that we should allow for tree effects; see trees 1, 5 and 8 in particular. If the trees are random this can be done in a way that allow random tree effects to be part of the error. (This comes under the heading of repeated measures/mixed model regression, which we don't have time to do much with in this course). An alternative, or what we'd need to do if the trees were fixed by design, is look at alternatives to simple linear regression that accommodate tree effects in some manner. We will do this in the context of multiple regression.

3. The full model is our usual regression model with SSE(F) = our usual SSE with n-2 degrees of freedom. Under H_0 the SEE(R) (under the null model) is $SSE(R) = \sum_i (Y_i - X_i)^2$ with $n - 0 = n$ dof since there are no unknown parameters in the reduced model for $E(R)$. So, we would use $F = (SSE - SSE(R))/(n-2-n)/MSE$, which under H_0 will follow an F with 2 and n-2 degrees of freedom. We reject H_0 if $F_{obs} > F(1-\alpha, 2, n-2)$ or if $P(F > F_{obs}) < \alpha$ where F_{obs} is the observed value of the F-statistic and F in the probability is distributed $F(2, n-2)$.

   There are some general ways to test linear hypotheses (of which the above is a special case) in both SAS and R, that we will explore a little later once we have a matrix representation of multiple regression under our belts.

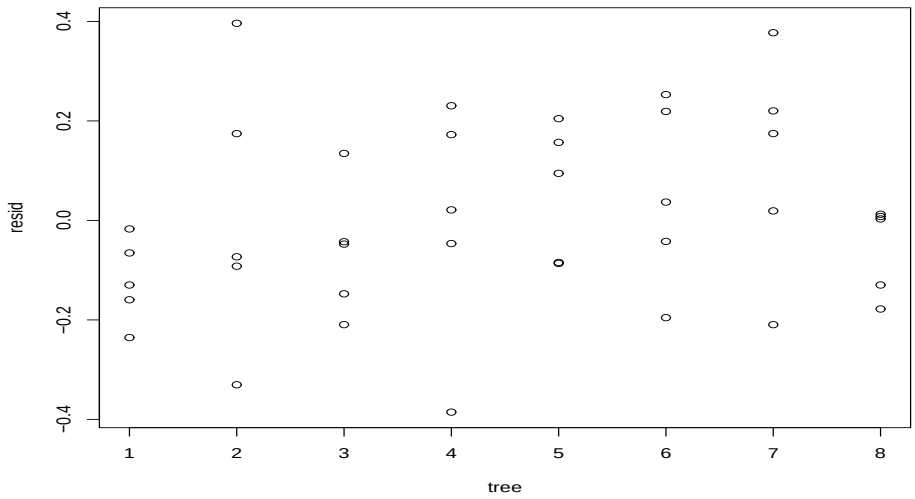

Figure 2: Plot of residual versus tree id