# Homework 5

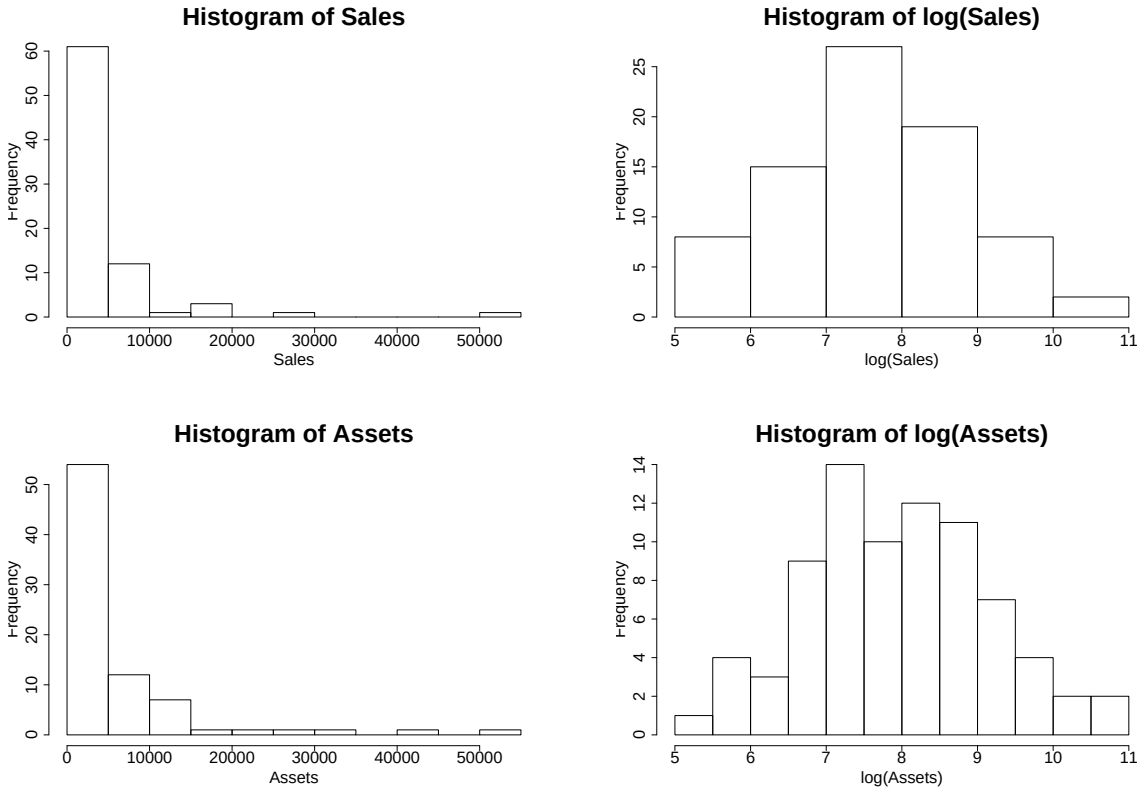
Due November 26, 2007

## Questions

Consider the business data on the Homework 5 website. Here, we have a list of 77 companies, their assets, sales, market value, profits, cash flow, employees, and market sector. (Data is a couple of decades old.)

1. Examine the financial data. Notice that Sales and Assets, in particular, are right-skewed. To satisfy the normality assumptions of the linear modeling approach, apply the log transformation to both these variables before continuing.

```
> d <- read.table("hw5.Rdata", sep="\t", header=T)
> attach(d)
> hist(Assets, cex.axis=2, cex.lab=2, cex.main=3)
> hist(log(Assets), cex.axis=2, cex.lab=2, cex.main=3)
> hist(Sales, cex.axis=2, cex.lab=2, cex.main=3)
> hist(log(Sales), cex.axis=2, cex.lab=2, cex.main=3)
```



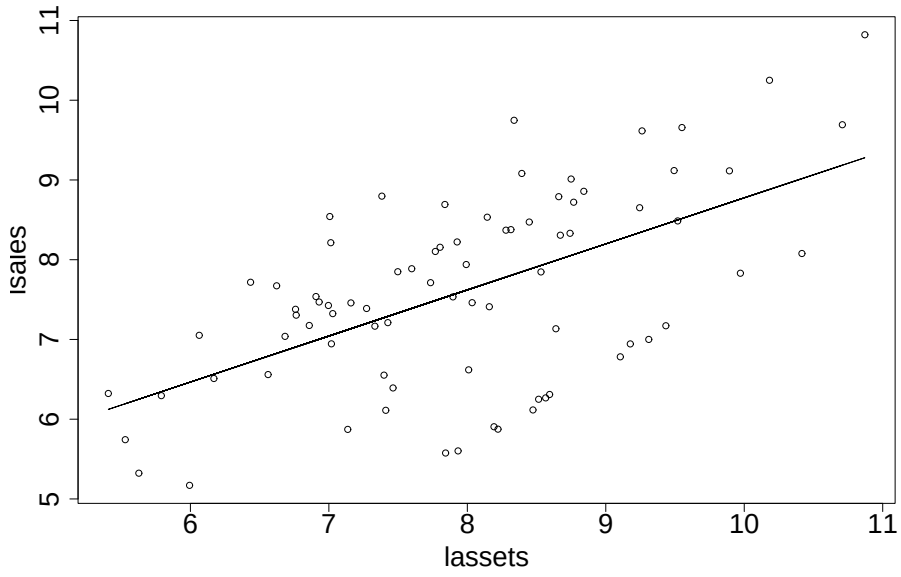

For kicks, let's blindly run one of the tests of normality provided by R.

```
> shapiro.test(Assets)
> shapiro.test(log(Assets))
> shapiro.test(Sales)
> shapiro.test(log(Sales))
```

| Variable | $p$-value |
|---|---|
| Assets | 1.314e-13 |
| log(Assets) | 0.8672 |
| Sales | 1.324e-14 |
| log(Sales) | 0.7976 |

2. Consider a simple linear regression of log(sales) on log(assets). Do assets predict sales?

   R code and analysis below. Yes, clearly there is a strong positive relationship between log assests and log sales, supported by a $p$-value=$1.83 \times 10^{-8}$ for the test of slope $\beta_{\log(\text{Assets})} = 0$. Also notice that the call to anova and the resulting F test were unecessary, as the F test is equivalent to the $t$-test reported in the summary() function on the slope parameter.



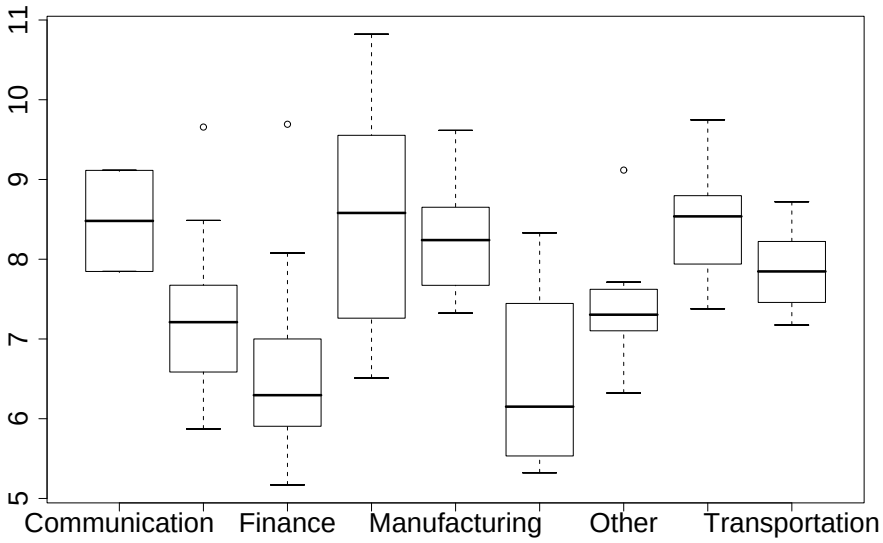

```
> m.a <- lm(lsales ~ lassets)
> m.0 <- lm(lsales ~ 1)
> summary(m.a)
> summary(m.0)
> anova(m.a, m.0)
> plot(lassets, lsales, cex.axis=2, cex.lab=2, cex.main=3)
```

```
> lines(lassets, m.a$fitted.values)

Call:
lm(formula = lsales ~ lassets)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9799 -0.8874  0.2645  0.6253  1.9326

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.00003    0.73942   4.057 0.000118 ***
lassets      0.57758    0.09191   6.284 1.82e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.985 on 77 degrees of freedom
Multiple R-Squared: 0.339,      Adjusted R-squared: 0.3304
F-statistic: 39.49 on 1 and 77 DF,  p-value: 1.817e-08


Call:
lm(formula = lsales ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4238 -0.8939 -0.1247  0.8304  3.2266

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.5943     0.1354   56.07   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.204 on 78 degrees of freedom

Analysis of Variance Table

Model 1: lsales ~ lassets
Model 2: lsales ~ 1
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     77  74.708
2     78 113.026 -1   -38.318 39.493 1.817e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

3. Consider a one-way ANOVA of log(sales) on market sector. Does market sector predict

sales?

R code and analysis below. Yes, market sector substantially affects log sales, with an $F$-test for the classical ANOVA hypothesis of constant means across all sectors of $3.387 \times 10^{-5}$. Again, the call to anova is not necessary, as function lm() reports this particular result (comparing the full model with the intercept-only model) by default. Note, however, that only the coefficients for sectorFinance and sectorMedical are significant, and not by much. Despite no real dramatic sector outliers, the null hypothesis is rejected, indicating some differences in the sector means.

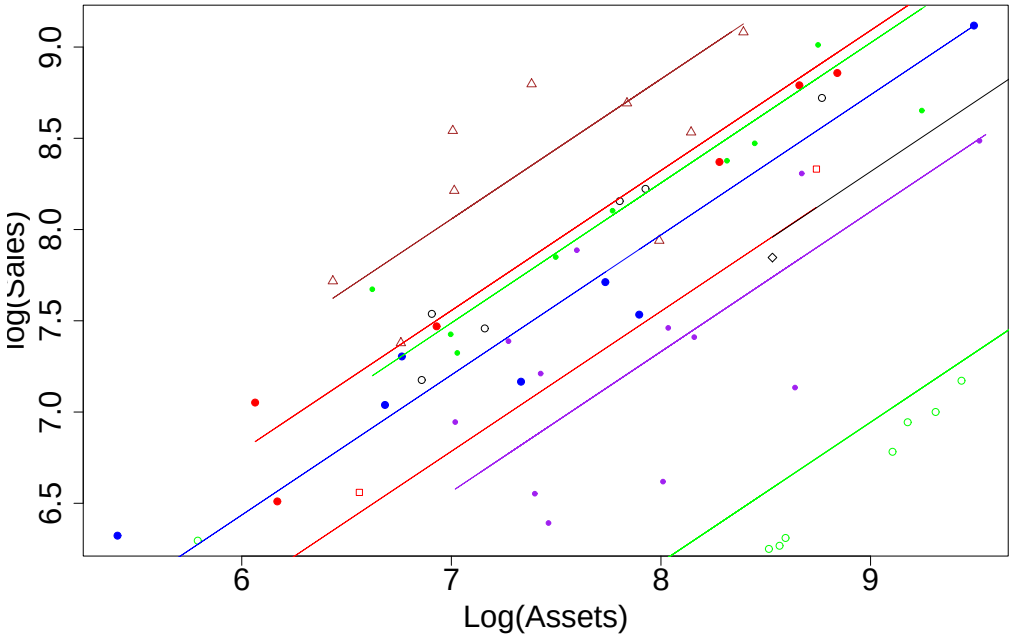



```
> m.s <- lm(lsales ~ sector)
> summary(m.s)
> anova(m.s, m.0)
> plot(sector, lsales, cex.axis=2, cex.lab=2, cex.main=3)

Call:
lm(formula = lsales ~ sector)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0049 -0.6900 -0.1448  0.3527  3.0534

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.48043    0.70938  11.955   <2e-16 ***
sectorEnergy      -1.18482    0.75519  -1.569   0.1212
sectorFinance     -1.84125    0.74995  -2.455   0.0166 *
```

```
sectorHiTech          0.03473    0.79311   0.044   0.9652
sectorManufacturing  -0.23009    0.77709  -0.296   0.7680
sectorMedical        -1.99119    0.86881  -2.292   0.0249 *
sectorOther          -1.02393    0.80436  -1.273   0.2072
sectorRetail         -0.01618    0.77709  -0.021   0.9834
sectorTransportation -0.60184    0.81912  -0.735   0.4649
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.003 on 70 degrees of freedom
Multiple R-Squared: 0.3767,     Adjusted R-squared: 0.3054
F-statistic: 5.288 on 8 and 70 DF,  p-value: 3.387e-05

Analysis of Variance Table

Model 1: lsales ~ sector
Model 2: lsales ~ 1
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     70  70.451
2     78 113.026 -8   -42.575 5.2878 3.387e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

4. Now, combine these two predictors (market sector and log(assets)) in a single ANCOVA model. Are both components of the model important, even in the presence of the other?

   R code and analysis below. The anova output shows that both simpler models, without sector or without log(Assets), are inferior to the model that includes log(Assets) *and* Market Sector, so both are important, even in the presence of the other. The plot shows that while the increase in log(Sales) for a unit increase in log(Assets) may be very similar for all sectors, the sectors differ in the overall level (intercept) of log(Sales). Notice that in this model, more of the market sectors have significant coefficients. Accounting for the covariate log(Asset) allows us to see that more of the sectors have distinct overall means.



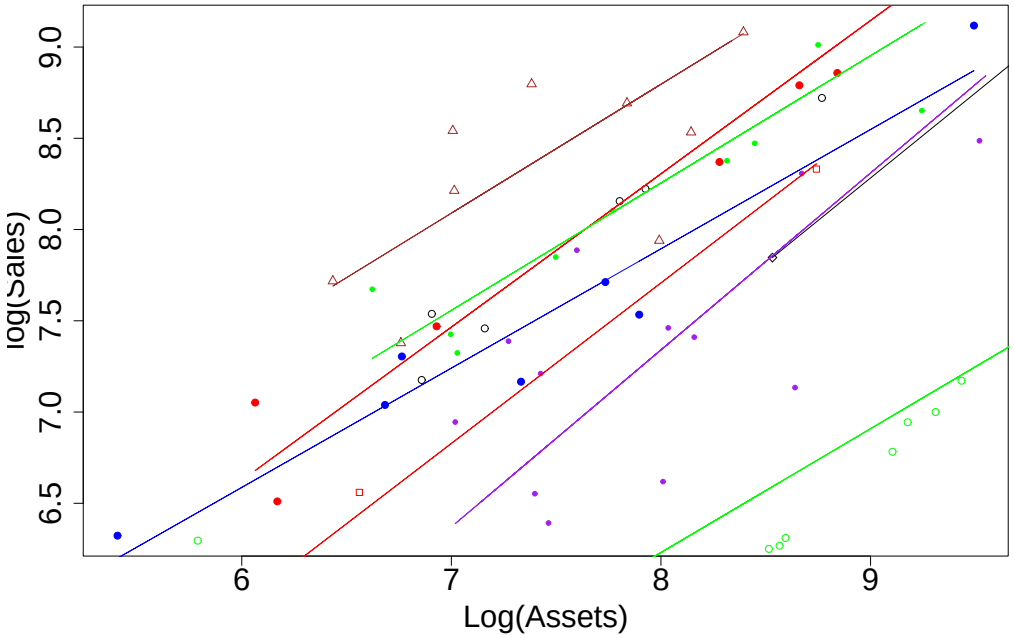

```
> m.sa <- lm(lsales ~ sector + lassets) # Model with sector and log(Assets), no interaction
> summary(m.sa)
> anova(m.sa)
> plot(lassets[sector=="Other"], lsales[sector=="Other"], cex.axis=2, cex.lab=2, cex.main=3, pch=19,
+ col="blue", xlab="Log(Assets)", ylab="log(Sales)")
> points(lassets[sector=="Energy"], lsales[sector=="Energy"], pch=20, col="purple")
> points(lassets[sector=="Finance"], lsales[sector=="Finance"], pch=21, col="green")
> points(lassets[sector=="Medical"], lsales[sector=="Medical"], pch=22, col="red")
> points(lassets[sector=="Communication"], lsales[sector=="Communication"], pch=23, col="black")
> points(lassets[sector=="Retail"], lsales[sector=="Retail"], pch=24, col="brown")
> points(lassets[sector=="HiTech"], lsales[sector=="HiTech"], pch=19, col="red")
> points(lassets[sector=="Manufacturing"], lsales[sector=="Manufacturing"], pch=20, col="green")
> points(lassets[sector=="Transportation"], lsales[sector=="Transportation"], pch=21, col="black")
> lines(lassets[sector=="Other"], m.sa$fitted.values[sector=="Other"], col="blue")
> lines(lassets[sector=="Energy"], m.sa$fitted.values[sector=="Energy"], col="purple")
> lines(lassets[sector=="Finance"], m.sa$fitted.values[sector=="Finance"], col="green")
> lines(lassets[sector=="Medical"], m.sa$fitted.values[sector=="Medical"], col="red")
> lines(lassets[sector=="Communication"], m.sa$fitted.values[sector=="Communication"], col="black")
> lines(lassets[sector=="Retail"], m.sa$fitted.values[sector=="Retail"], col="brown")
> lines(lassets[sector=="HiTech"], m.sa$fitted.values[sector=="HiTech"], col="red")
> lines(lassets[sector=="Manufacturing"], m.sa$fitted.values[sector=="Manufacturing"], col="green")

Call:
lm(formula = lsales ~ sector + lassets)

Residuals:
     Min       1Q   Median       3Q      Max
-0.87909 -0.31926 -0.03989  0.19548  1.81473
```

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.410424   0.588417   2.397 0.019244 *
sectorEnergy         -0.218986   0.376033  -0.582 0.562223
sectorFinance        -1.373386   0.369169  -3.720 0.000402 ***
sectorHiTech          0.773593   0.392150   1.973 0.052538 .
sectorManufacturing   0.705542   0.386281   1.827 0.072101 .
sectorMedical         0.002178   0.446639   0.005 0.996123
sectorOther           0.420587   0.406258   1.035 0.304157
sectorRetail          1.274794   0.390867   3.261 0.001724 **
sectorTransportation  0.658345   0.410566   1.604 0.113391
lassets               0.767448   0.051510  14.899  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4921 on 69 degrees of freedom
Multiple R-Squared: 0.8522,     Adjusted R-squared: 0.8329
F-statistic:  44.2 on 9 and 69 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: lsales
          Df Sum Sq Mean Sq F value    Pr(>F)
sector     8 42.575   5.322  21.981 3.274e-16 *** # Model without sector is significantly worse fit
lassets    1 53.745  53.745 221.981 < 2.2e-16 *** # Model without lassets is significantly worse fit
Residuals 69 16.706   0.242
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

5. Is there evidence of heterogeneity of slopes across market sectors?

   R code and analysis below. Also, see more comments at the end of the R output.

   The anova call shows that there is no significant interaction between log(Assets) and market sector ($p$-value for rejecting the null hypothesis of same slopes is 0.8611), so there is no evidence of distinct slopes across market sectors.




```
> m.sai <- lm(lsales ~ sector * lassets) # model with sector, log(Assets), and interaction
> summary(m.sai)
> anova(m.sa, m.sai)
> plot(lassets[sector=="Other"], lsales[sector=="Other"], cex.axis=2, cex.lab=2, cex.main=3, pch=19, co
> points(lassets[sector=="Energy"], lsales[sector=="Energy"], pch=20, col="purple")
> points(lassets[sector=="Finance"], lsales[sector=="Finance"], pch=21, col="green")
> points(lassets[sector=="Medical"], lsales[sector=="Medical"], pch=22, col="red")
> points(lassets[sector=="Communication"], lsales[sector=="Communication"], pch=23, col="black")
> points(lassets[sector=="Retail"], lsales[sector=="Retail"], pch=24, col="brown")
> points(lassets[sector=="HiTech"], lsales[sector=="HiTech"], pch=19, col="red")
> points(lassets[sector=="Manufacturing"], lsales[sector=="Manufacturing"], pch=20, col="green")
> points(lassets[sector=="Transportation"], lsales[sector=="Transportation"], pch=21, col="black")
> lines(lassets[sector=="Other"], m.sai$fitted.values[sector=="Other"], col="blue")
> lines(lassets[sector=="Energy"], m.sai$fitted.values[sector=="Energy"], col="purple")
> lines(lassets[sector=="Finance"], m.sai$fitted.values[sector=="Finance"], col="green")
> lines(lassets[sector=="Medical"], m.sai$fitted.values[sector=="Medical"], col="red")
> lines(lassets[sector=="Communication"], m.sai$fitted.values[sector=="Communication"], col="black")
> lines(lassets[sector=="Retail"], m.sai$fitted.values[sector=="Retail"], col="brown")
> lines(lassets[sector=="HiTech"], m.sai$fitted.values[sector=="HiTech"], col="red")
> lines(lassets[sector=="Manufacturing"], m.sai$fitted.values[sector=="Manufacturing"], col="green")

Call:
lm(formula = lsales ~ sector * lassets)

Residuals:
     Min       1Q   Median       3Q      Max
-0.85091 -0.29230 -0.02873  0.20833  1.62471
```

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -0.10060    4.87041  -0.021   0.9836
sectorEnergy                 -0.31609    5.04883  -0.063   0.9503
sectorFinance                 0.90325    4.94267   0.183   0.8556
sectorHiTech                  1.68619    4.95483   0.340   0.7348
sectorManufacturing           2.77348    5.07579   0.546   0.5868
sectorMedical                 0.77103    5.04695   0.153   0.8791
sectorOther                   2.76467    5.01981   0.551   0.5838
sectorRetail                  3.24545    5.19956   0.624   0.5348
sectorTransportation          2.17091    5.40056   0.402   0.6891
lassets                       0.93147    0.52725   1.767   0.0823 .
sectorEnergy:lassets          0.03816    0.55289   0.069   0.9452
sectorFinance:lassets        -0.25302    0.53606  -0.472   0.6386
sectorHiTech:lassets         -0.09148    0.53824  -0.170   0.8656
sectorManufacturing:lassets  -0.23369    0.55638  -0.420   0.6759
sectorMedical:lassets        -0.05183    0.56261  -0.092   0.9269
sectorOther:lassets          -0.27767    0.55209  -0.503   0.6168
sectorRetail:lassets         -0.22506    0.57964  -0.388   0.6992
sectorTransportation:lassets -0.16423    0.61013  -0.269   0.7887
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5074 on 61 degrees of freedom
Multiple R-Squared: 0.8611,     Adjusted R-squared: 0.8223
F-statistic: 22.24 on 17 and 61 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Model 1: lsales ~ sector + lassets
Model 2: lsales ~ sector * lassets
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     69 16.706
2     61 15.704  8     1.002 0.4865 0.8611
```

Another way to compare models is using the log likelihood ratio test statistic.

```
> ll.sai <- logLik(m.sai)[1]
> ll.sa <- logLik(m.sa)[1]
> chi <- -2*(ll.sa - ll.sai)  # compute the LRT
> df <- df.residual(m.sa) - df.residual(m.sai)
> p <- 1-pchisq(chi, df = df) # LRT has a chi-square distribution when null is true
> print( paste("Log likelihood ratio test of same slope (chi=", chi, ", df=", df, "): ", p, sep='') )
[1] "Log likelihood ratio test of same slope (chi=4.88658548113283 , df=8): 0.769626440246548"
```

Finally, notice the $R^2$ for the different models.

| | |
|---|---|
| $\log(\text{Sales}) \sim \log(\text{Assets})$ | 0.339 |
| $\log(\text{Sales}) \sim \text{sector}$ | 0.3767 |
| $\log(\text{Sales}) \sim \text{sector} + \log(\text{Assets})$ | 0.8522 |

Clearly, we prefer the last model for predictive ability because the first two models, while significantly better than the null model of a single constant mean, only capture

about a third of the variability in log(Sales). The last model captures 85% of variability in log(Sales), which is not bad for such an unpredictable science.