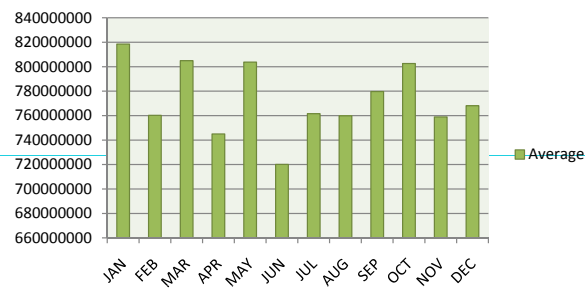


Regression Modelling in R

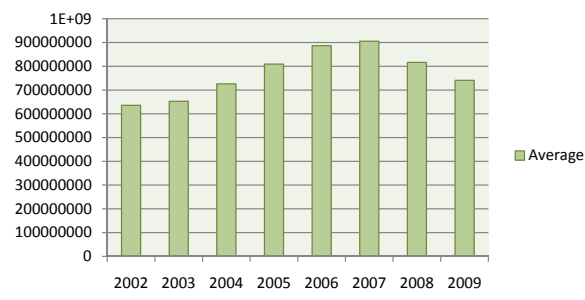
Example 4: The data file gaming_revenue.csv has monthly gaming revenue data for Clark County, Strip, Downtown and Boulder City, for 2002 – 2009. Regression methods are also used for modeling time series data such as gaming revenue. Fit a regression model using dummy variables to account for seasonality (months) and time t (# of month, t = 1, 2, ..., 90. Following table shows the Strip values (in million \$)

642.34	616.36	656.11	647.07	681.19	561.97	627.40	665.12	643.20	654.48	592.87	642.44	635.88
703.26	620.35	684.98	608.89	651.00	656.39	655.90	635.80	683.85	664.50	631.44	635.03	652.61
747.66	731.18	776.29	678.72	748.98	704.33	647.23	726.32	753.27	766.72	720.40	709.89	725.91
793.10	765.80	864.99	723.52	860.38	797.15	765.63	808.36	837.10	884.08	846.69	762.60	809.12
987.70	869.98	908.67	824.20	962.62	757.74	850.26	886.02	807.85	889.91	989.65	908.63	886.93
967.78	901.82	889.67	892.76	968.44	789.66	964.73	838.03	879.14	1001.32	828.73	945.96	905.67
928.65	865.97	871.90	849.97	810.06	806.10	819.68	759.26	853.51	757.51	702.59	771.78	816.41
777.53	710.60	786.46	734.71	747.61	687.55							740.74
818.50	760.26	804.88	744.98	803.78	720.11	761.55	759.84	779.70	802.64	758.91	768.04	0.00

Monthly Average



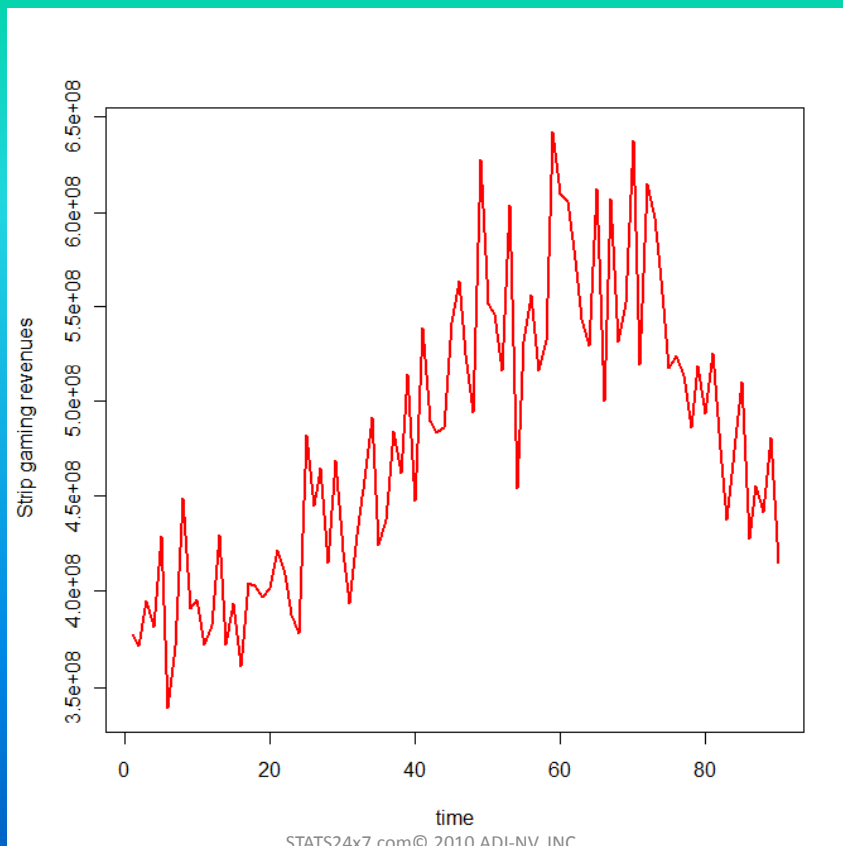
Annual Average



```
plot(Strip)
```

```
layout(1)
```

```
plot(Strip, type = "l", lwd=2, col="red", xlab="time", ylab="Strip gaming revenues")
```



```
x <- read.csv("K:/DataMining/Data/gaming_revenue.csv")

attach(x)
out1 <-
lm(Strip~Month_1+DJan+DFeb+DFeb+DMar+DApr+DMay+DJun+DJul+D
Aug+D Sep+DOct+DNov)
summary(out1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	391304739	18878186	20.728	< 2e-16 ***
T	1870717	232672	8.040	8.51e-12 ***
DJan	42593573	25400646	1.677	0.0976 .
DFeb	-3481394	25392763	-0.137	0.8913
DMar	3516764	25387010	0.139	0.8902
DApr	-22589217	25364451	-0.891	0.3759
DMay	29928816	25361834	1.180	0.2416
DJun	-39821526	25361351	-1.570	0.1205
DJul	3297002	26529926	0.124	0.9014
DAug	7823714	26521147	0.295	0.7688
DSep	14286569	26514406	0.539	0.5916
DOct	26474995	26509706	0.999	0.3211
DNov	-22075879	21655506	-1.019	0.3112

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

Residual standard error: 57160000 on 77 degrees of freedom

Multiple R-squared: 0.5006, Adjusted R-squared: 0.4228

F-statistic: 6.433 on 12 and 77 DF, p-value: 8.984e-08

```

t1 <- (t - mean(t))/sd(t)
t2 <- t1^2
out2 <-
lm(Strip~t1+t2++DJan+DFeb+DFeb+DMar+DApr+DMay+DJun+DJul+D
Aug+DSep+DOct+DNov)

summary(out2)

```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	510066622	12503322	40.794	< 2e-16 ***
t1	48868019	4623458	10.570	< 2e-16 ***
t2	-39600607	5241056	-7.556	7.76e-11 ***
DJan	53176852	19371107	2.745	0.00754 **
DFeb	6869943	19362930	0.355	0.72372
DMar	13752205	19357486	0.710	0.47961
DApr	-12395257	19339985	-0.641	0.52351
DMay	40238971	19339078	2.081	0.04083 *
DJun	-29279130	19340905	-1.514	0.13421
DJul	3388589	20179365	0.168	0.86709
DAug	7683358	20172692	0.381	0.70436
DSep	14030316	20167585	0.696	0.48875
DOct	26218892	20164009	1.300	0.19743
DNov	-21742843	16471808	1.320	0.19080

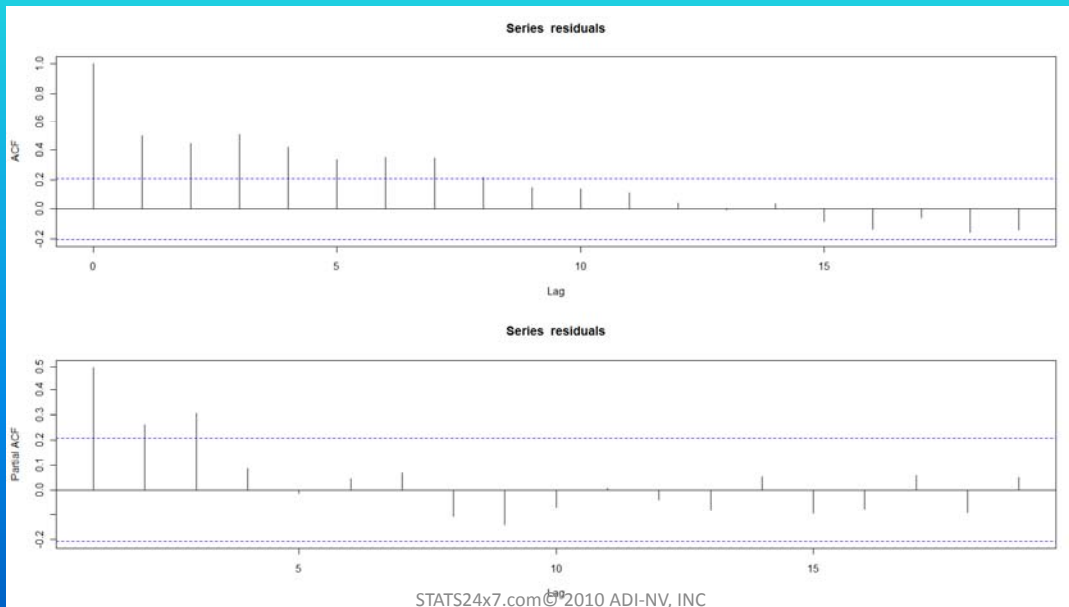
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43480000 on 76 degrees of freedom

Multiple R-squared: 0.7148, Adjusted R-squared: 0.6661

F-statistic: 14.66 on 13 and 76 DF, p-value: 8.716e-16

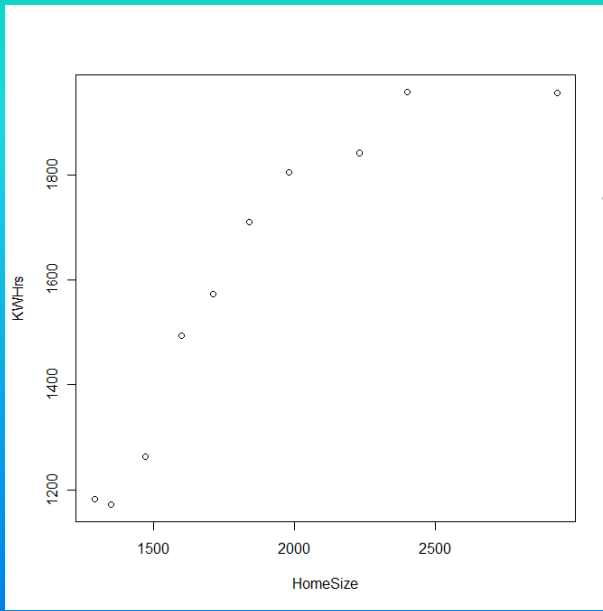
```
layout(1:2)
acf(residuals)
pacf(residuals)
```



Polynomial Regression

Example 5: Following table shows energy used (KWHrs) as a function of HomeSize (square feet). Fit a regression model to KWHrs as a function of HomeSize (Use data file HomeEnergyUse.csv).

HomeSize	KWHrs
1290	1182
1350	1172
1470	1264
1600	1493
1710	1571
1840	1711
1980	1804
2230	1840
2400	1956
2930	1954



Since the graph shows 2 bends, we should try a cubic function:

$$\text{KWHrs} = \beta_0 + \beta_1 \text{HomeSize} + \beta_2 \text{HomeSize}^2 + \beta_3 \text{HomeSize}^3$$

```
xx <- read.csv("K:/DataMining/Data/HomeEnergyUse.csv")
attach(xx)
x <- (HomeSize-mean(HomeSize))/sd(HomeSize)
x2 <- x^2
x3 <- x^3

# install library HH if not installed already
library(HH)

out0 <- lm(KWHrs~HomeSize+HS2+HS3)
summary(out0)
vif(out0)
```

```
> summary(out0)
Call:
lm(formula = KWHrs ~ HomeSize + HS2 + HS3)
```

Residuals:

```
   Min    1Q  Median    3Q   Max
-75.068 -22.133  6.907  30.633  55.594
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.414e+03	1.300e+03	-1.088	0.318
HomeSize	2.707e+00	2.000e+00	1.354	0.225
HS2	-6.033e-04	9.876e-04	-0.611	0.564
HS3	2.436e-08	1.567e-07	0.156	0.882

Residual standard error: 50.45 on 6 degrees of freedom
Multiple R-squared: 0.982, Adjusted R-squared: 0.9729
F-statistic: 108.9 on 3 and 6 DF, p-value: 1.276e-05

```
> vif(out0)
```

HomeSize	HS2	HS3
3788.808	16000.454	4370.732

STATS24x7.com© 2010 ADI-NV, INC

13

One way to take care of multicollinearity is to use
ORTHOGONAL POLYNOMIALS.

Another (easier) way to take care of it is to standardize the X-
variable :

$$X1 = (X - \text{mean}(X)) / \text{sd}(X)$$

and then calculate the square and cubic terms.

```
# standardize x-variables to reduce VIFs
```

```
x <- (HomeSize-mean(HomeSize))/sd(HomeSize)
```

```
x2 <- x^2
```

```
x3 <- x^3
```

```
out1 <- lm(KWHrs~x+x2+x3)
```

```
summary(out1)
```

```
vif(out1)
```

```
summary(out1)
```

Call:

```
lm(formula = KWHrs ~ x + x2 + x3)
```

Residuals:

Min	1Q	Median	3Q	Max
-75.068	-22.133	6.907	30.633	55.594

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1704.872	24.553	69.436	6.00e-10 ***
x	360.832	38.349	9.409	8.19e-05 ***
x2	-124.828	32.205	-3.876	0.0082 **
x3	3.379	21.727	0.156	0.8815

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

Residual standard error: 50.45 on 6 degrees of freedom

Multiple R-squared: 0.982, *Adjusted R-squared:* 0.9729

F-statistic: 108.9 on 3 and 6 DF, *p-value:* 1.276e-05

```
> vif(out1)
```

x	x2	x3
5.200281	5.435543	13.034283

```
> out2 <- lm(KWHrs~x+x2)
> summary(out2)
Call:
lm(formula = KWHrs ~ x + x2)
```

Residuals:

```
   Min    1Q  Median    3Q   Max
-73.792 -22.426  5.886  31.689  52.436
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 1703.22    20.54  82.914 9.77e-12 ***
x           365.84    19.27  18.985 2.80e-07 ***
x2          -120.58    15.83  -7.618 0.000124 ***
```

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

```
Residual standard error: 46.8 on 7 degrees of freedom
Multiple R-squared:  0.9819,    Adjusted R-squared:  0.9767
F-statistic: 189.7 on 2 and 7 DF, p-value: 8e-07
```

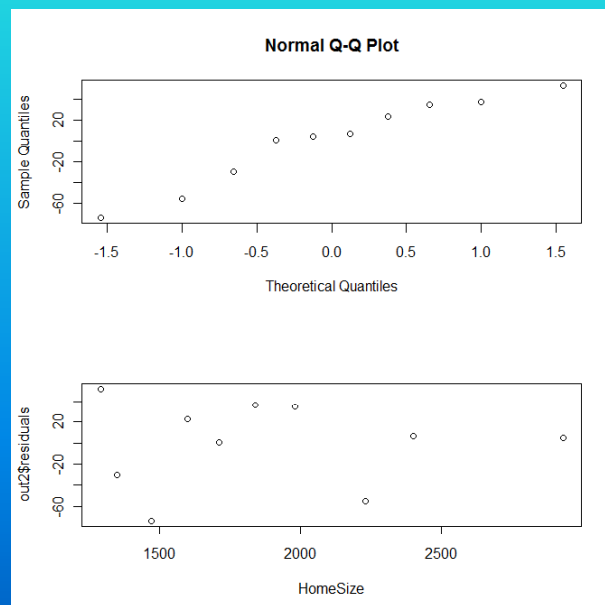
```
> vif(out2)
```

```
      x      x2
1.525748 1.525748
```

```
>
```

RESIDUAL DIAGNOSTIC PLOTS

```
layout(1:2)
qqnorm(out2$residuals)
plot(out2$residuals~HomeSize)
```



Exercises on Regression:

1. Fit $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ to data in the file anyregression.csv.

Does the fitted model provide good fit to the data?

2. Repeat for $Y_1 = \ln(Y)$ and $Y_2 = \sqrt{Y}$.

Call:

```
lm(formula = Y ~ X1 + X2 + X3)
```

Residuals:

```
   Min     1Q  Median     3Q    Max
-47.283 -14.245  1.012  13.757  59.142
```

Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 97.8002    2.3164  42.222 < 2e-16 ***
X1           1.3774    0.3280   4.199 4.64e-05 ***
X2           0.7369    0.3518   2.095 0.03791 *
X3           0.9420    0.3560   2.646 0.00903 **
```

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

Residual standard error: 20.06 on 146 degrees of freedom

Multiple R-squared: 0.9676, Adjusted R-squared: 0.967

F-statistic: 1455 on 3 and 146 DF, p-value: < 2.2e-16

Call:
lm(formula = lnY ~ X1 + X2 + X3)

Residuals:
Min 1Q Median 3Q Max
-0.428783 -0.079935 0.005036 0.090634 0.388577

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.703160 0.016245 289.506 < 2e-16 ***
X1 0.007665 0.002300 3.332 0.00109 **
X2 0.003177 0.002467 1.288 0.19989
X3 0.002955 0.002497 1.184 0.23846

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

Residual standard error: 0.1407 on 146 degrees of freedom
Multiple R-squared: 0.9254, Adjusted R-squared: 0.9239
F-statistic: 604.1 on 3 and 146 DF, p-value: < 2.2e-16

Call:
lm(formula = sqrtY ~ X1 + X2 + X3)

Residuals:
Min 1Q Median 3Q Max
-2.22719 -0.53358 0.02841 0.52334 2.38650

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.27832 0.09346 109.981 < 2e-16 ***
X1 0.05064 0.01323 3.827 0.000192 ***
X2 0.02401 0.01419 1.692 0.092790 .
X3 0.02672 0.01436 1.861 0.064811 .

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

Residual standard error: 0.8093 on 146 degrees of freedom
Multiple R-squared: 0.9529, Adjusted R-squared: 0.9519
F-statistic: 984.3 on 3 and 146 DF, p-value: < 2.2e-16

Question : Why are all these different models providing very good fit to this data set (anyregression.csv)?