Regression Analysis

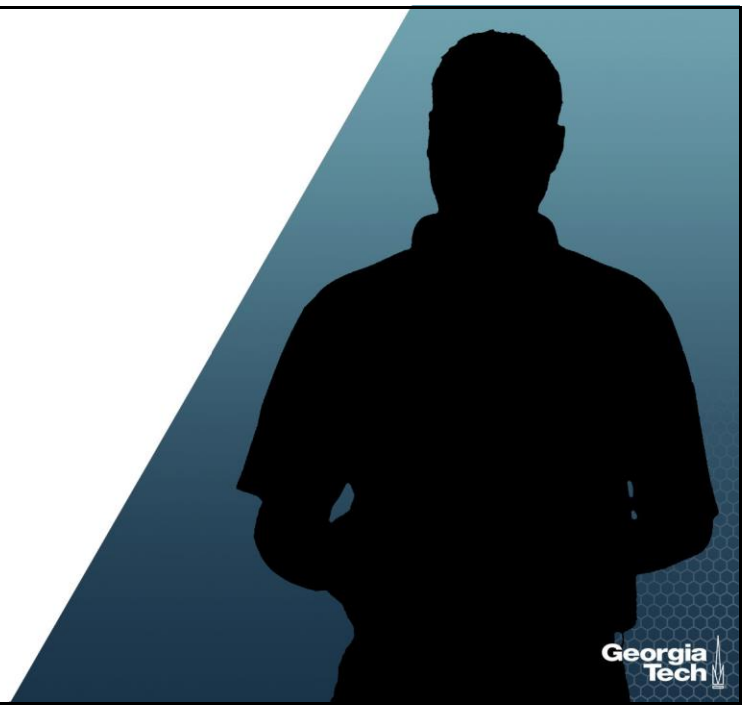
Simple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Example 2: 2000 Presidential Elections in Florida

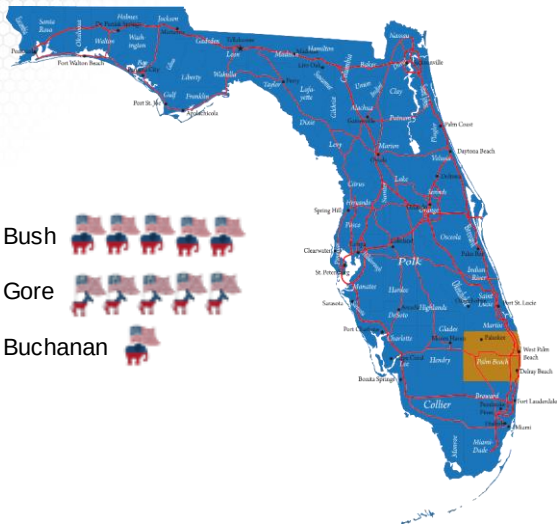
1

About This Lesson



2

Elections in 2000: Florida



3

Data Example in R

Read data with read.table R command which is used for reading ASCII files

elections = read.table("elections.txt",header=TRUE)

Check the data content elections[1:4,]

	co	lat	lon	npop	whit	blac	hisp	o65	hsed	coll	inco	bush	gore	brow
1	1	29.7	82.4	198326	74.4	21.8	4.7	9.4	82.7	34.6	19412	34124	47365	658
2	2	30.3	82.3	20761	82.4	16.8	1.5	7.7	64.1	5.7	14859	5610	2392	17
3	3	30.2	85.6	146223	84.2	12.4	2.4	11.9	74.7	15.7	17838	38637	18850	171
4	4	29.9	82.2	24646	76.1	22.9	2.6	11.8	65.0	8.1	13681	5414	3075	28

	nade	harr	hage	buch	mcre	phil	moor
1	3226	6	42	263	4	20	21
2	53	0	3	73	0	3	3
3	828	5	18	248	3	18	27
4	84	0	2	65	0	2	3

The data file includes many other variables characterizing the counties. We will focus only on the number of votes in this analysis.

4

Exploratory Data Analysis in R

Extract number of votes for each candidates

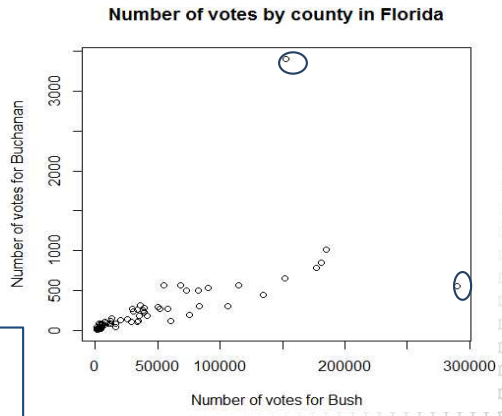
```
buch = elections$buch
bush = elections$bush
```

Visualize the relationship between number of votes between Buchanan and Bush

```
plot(bush,buch,xlab="Number of votes for Bush",ylab="Number of votes for Buchanan", main="Number of votes by county in Florida")
cor(buch,bush)
```

Linearity Assumption:

- The scatterplot shows a strong positive relationship between the number of votes for the two candidates except for two outliers, one corresponding to the Palm Beach county. The correlation is high also (0.625).
- Curvature in the relationship – consider transformations

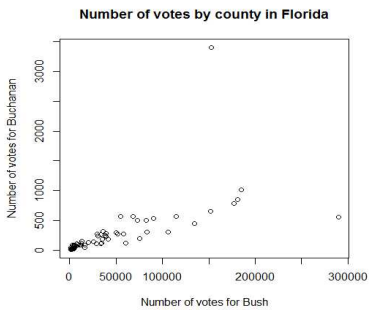


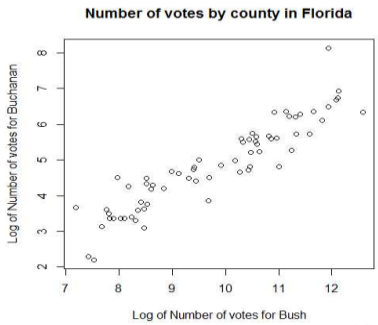
5

Linearity using Transformation

Transform both variables using the log-transformation

```
plot(log(bush),log(buch),xlab="Log of Number of votes for Bush",ylab="Log of Number of votes for Buchanan", main="Number of votes by county in Florida")
cor(log(bush),log(buch))
```





6

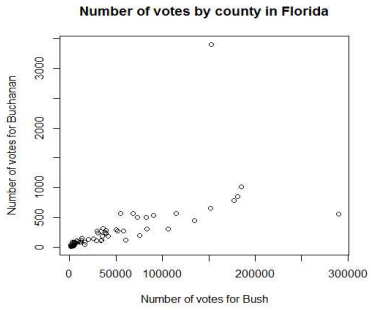
Linearity using Transformation

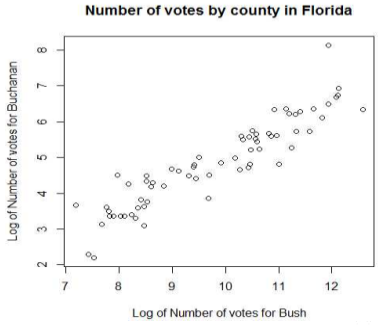
Transform both variables using the log-transformation

```
plot(log(bush),log(buch),xlab="Log of Number of votes for Bush",ylab="Log of Number of votes for Buchanan",
main="Number of votes by county in Florida")
cor(log(bush),log(buch))
```

Linearity Assumption:

- **The linear relationship has improved with the transformations**
- **The correlation has increased from 0.625 to 0.922**
- **We will perform the regression analysis using the transformed data**





7

Linear Regression Analysis

model = lm(log(buch)~log(bush))

```
Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) -2.55079    0.38903   -6.557    1.04e-08 ***
log(bush)    0.75620    0.03934   19.222    < 2e-16 ***
---
```

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

$\hat{\beta}_0 = -2.55$, $se(\hat{\beta}_0) = 0.389$
$\hat{\beta}_1 = 0.756$, $se(\hat{\beta}_1) = 0.039$
Test for statistical significance:
$\hat{\beta}_0$: t-value= -6.557, p-value ≈ 0
$\hat{\beta}_1$: t-value= 19.22, p-value ≈ 0

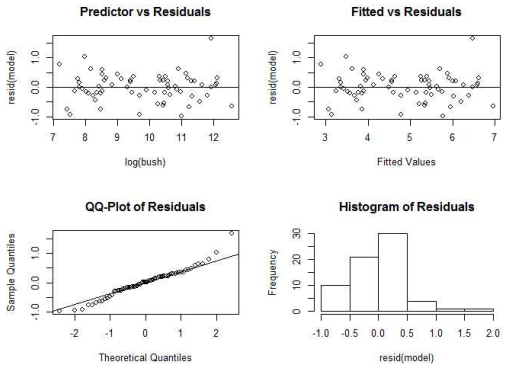
Residual standard error: 0.4672 on 65 degrees of freedom
Multiple R-squared: 0.8504, Adjusted R-squared: 0.8481
F-statistic: 369.5 on 1 and 65 DF, p-value: < 2.2e-16

$\hat{\sigma}$= 0.4672, n-2 = 65
R^2 ~ 85% variability explained

8

Residual Analysis

```
## Perform Residual Analysis
par(mfrow=c(2,2))
plot(log(bush),resid(model), main="Predictor vs
Residuals")
abline(0,0)
plot(fitted(model),resid(model),main="Fitted vs
Residuals",
   xlab="Fitted Values")
abline(0,0)
qqnorm(resid(model),main="QQ-Plot of Residuals")
qqline(resid(model))
hist(resid(model),main="Histogram of Residuals")
```



9

Model Interpretation

```
## Estimated Regression Coefficients
   betas = coef(model)
   Betas
     (Intercept)    log(bush)
    -2.5507857    0.7561963
## Confidence intervals for the coefficients
   confint(model)
                      2.5 %          97.5 %
(Intercept) -3.3277351    -1.7738363
log(bush)    0.6776289     0.8347638
```

Interpretation:

- As number of log-votes for Bush increase by 1% the expected % increase of log-votes for Buchanan is 0.756.
- The maximum % increase is 0.677 and the minimum % increase is 0.834

10

Is Palm Beach an Outlier?

```
## Omit Palm Beach
    model.red = lm(log(buch[-50])~log(bush[-50]))
    summary(model.red)
    Coefficients:
                    Estimate   Std. Error  t value  Pr(>|t|)
    (Intercept)     -2.31657   0.35470    -6.531   1.23e-08 ***
    log(bush[-50]) 0.72960    0.03599   20.271   < 2e-16 ***
## Obtain the predicted vote count for Palm Beach given the fitted model without
    new = data.frame(bush = bush[50])
    ## The difference between predicted on the original scale and the observed vote count
    buch[50]-exp(predict(model.red,new))
    [1] 2809
## Prediction Confidence Interval for log(vote count)
    predict(model.red,new,interval='prediction',level=.95)
## Prediction Confidence Interval on the original scale
    exp(predict(model.red,new,interval='prediction',level=.95))
      fit        lwr        upr
    597.5019 252.738 1412.564
## Is the observed vote count in the prediction interval?
    buch[50]
    [1] 3407
```

11

Is Palm Beach an Outlier?

```
## Omit Palm Beach
    model.red = lm(log(buch[-50])~log(bush[-50]))
    summary(model.red)
    Coefficients:
                    Estimate   Std. Error  t value  Pr(>|t|)
    (Intercept)     -2.31657   0.35470    -6.531   1.23e-08 ***
    log(bush[-50]) 0.72960    0.03599   20.271   < 2e-16 ***
## Obtain the predicted vote count for Palm Beach given the fitted model without
    new = data.frame(bush = bush[50])
    ## The difference between predicted on the original scale and the observed vote count
    buch[50]-exp(predict(model.red,new))
    [1] 2809
## Prediction Confidence Interval for log(vote count)
    predict(model.red,new,interval='prediction',level=.95)
## Prediction Confidence Interval on the original scale
    exp(predict(model.red,new,interval='prediction',level=.95))
      fit        lwr        upr
    597.5019 252.738 1412.564
## Is the observed vote count in the prediction interval?
    buch[50]
    [1] 3407
```

Interpretation:

- The difference between predicted and observed vote count for Bush in the Palm Beach county is 2809.
- The upper bound of the prediction confidence interval for the vote count is 1412 which is much lower than the observed vote count, 3407.
- While a difference of 2809 votes is not large given the total U.S. votes, this was particularly decisive for the 2000 elections.
- Recall that George W. Bush won Florida by a margin of 537 votes.

12

Summary

13