Stat 104: Quantitative Methods for Economists

Class 38: Wrapping Things Up

Dummy Variables-Recoding

- Suppose I have dummy variables for seasons of the year [fall,winter,spring,summer]
- I want to model hockey stick sales so I fit the model
- Sales = 200+50(winter)-20(spring)-40(summer)
- How do I interpret this model?

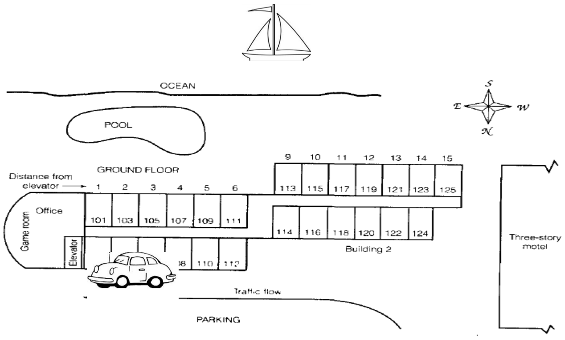
Interpret the Model

- Sales = 200+50(winter)-20(spring)-40(summer)
- This model says average sales are:

Example: Condo Prices

- This case study analyzes the factors affecting the prices of condominium units (in Florida) that were sold at a public auction.
- The condominium complex was completed in 2005.
- Due to a recession, sales were slow, and the developer was forced to sell most of the units at auction approximately 18 months after opening.

Layout of Complex



Some Details of the Complex

- ❑ The complex has eight floors (each identical).
- ❑ The units are all approximately the same size (around 1500 square feet each).
- ❑ The units on the south, called *ocean-view, face* the beach and ocean. The units on the first floor open out to the beach, ocean, and pool.
- ❑ The units on the north, called *bay-view, face the* parking lot and an area of land that ultimately borders a bay (but the bay is distant and barely visible). The view from the upper floors of these units is primarily of wooded, sandy terrain.
- ❑ The only elevator in the complex is located at the east end of Building 1, as are the office and the game room.
- ❑ Some units, called *end units, have their view* partially blocked. These units are numbered _11 and _14.
- ❑ Some units were furnished and rented by the developer prior to the auction.

Collected Data

- The following variables were collected:
  - ❑ Price : Sales price of the unit
  - ❑ Floor : floor of the unit
  - ❑ DISTELEV : distance of the unit from the elevator
  - ❑ OCEAN : 1 if an ocean view, 0 otherwise
  - ❑ ENDUNIT : 1 if end unit, 0 otherwise
  - ❑ FURN : 1 if furnished, 0 otherwise

Purpose of the Data

- The goal is to identify the factors that influence sales price of the condo units and to quantify the effects of the different factors.
- This information can be used by the owner in order to help determine sales prices and rental rates in the future.

Anticipated Effects

- Before doing any analysis whatsoever, what can we say about the anticipated effects of the explanatory variables on PRICE?
  - ❑ FLOOR Lower floors may have a positive effect due to easier access to the pool and ocean (which also suggests a possible interaction effect with OCEAN). Higher floors may have a positive effect due to a better view (again, more so for ocean-side units). The units in the middle would probably have the lowest prices.
  - ❑ DISTELEV On the one hand, units at a greater distance from the elevator are less convenient. On the other hand, there is less foot traffic for these units. Presumably , the first effect is dominant, meaning that DISTELEV should have a negative effect on PRICE.
  - ❑ OCEAN A positive effect is expected.
  - ❑ ENDUNIT A negative effect is expected.
  - ❑ FURN A positive effect is expected

Initial R Work

```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/condosales.csv")

> install.packages("car")  ## beware-this can take a few minutes

> library(car) ## need this package for ncvtest() and vif()
```

Fit the model and look at vif's

- What are we looking for when we examine the vif values?

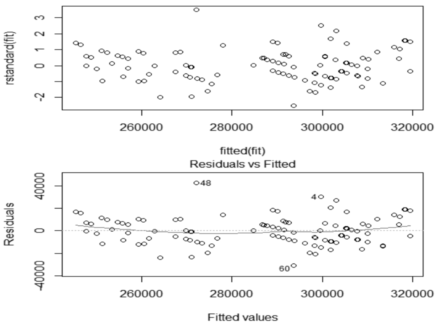
```
> fit=lm(price~floor+distelev+ocean+endunit+furn,data=mydata)

> vif(fit)
   floor distelev    ocean  endunit     furn 
1.179000 1.055194 1.216485 1.058251 1.115909
```

The Naïve Regression Model

```
> fit=lm(price~floor+distelev+ocean+endunit+furn,data=mydata)
> summary(fit)

Call:
lm(formula = price ~ floor + distelev + ocean + endunit + furn)

Residuals:
   Min     1Q Median     3Q    Max 
-31228  -8280   -217   7235  42850 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 267792.4     5265.0  50.863  < 2e-16 ***
floor        -1203.9      670.3  -1.796 0.075511 .  
distelev     -1164.0      310.0  -3.755 0.000293 ***
ocean        44799.1     2835.0  15.802  < 2e-16 ***
endunit     -25196.0     5031.6  -5.008 2.40e-06 ***
furn         12745.1     2612.1   4.879 4.08e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Examine the residuals

```
> plot(fitted(fit),rstandard(fit))
> plot(fit,which=1)
```



R is kind to point out extreme values for us

Examine the Floor Variable

- The failure of floor height to reveal itself as a useful contributor to the model goes against our intuition. One might argue that units on the higher floors posses a better view and hence should command a higher sale price.
- Or, you might argue that units on the lower floors have greater accessibility to the pool and ocean, and consequently should be in greater demand.
- Why then is the t-test not significant ?
- The answer is that both of the preceding arguments are correct, one for the oceanside and one for the bayside. Ocean-view units on the lower floors sell at higher prices than ocean-view units on the higher floors.
- In contrast, bay-view units on the higher floors command higher prices than bay-view units on the lower floors.
- These two contrasting effects tend to cancel and thereby give the false impression that floor height is not an important variable for predicting sale price.

Include an ocean*floor interaction

```
> fit=lm(price~floor+distelev+ocean+endunit+furn+ocean*floor,data=mydata)
> summary(fit)

Call:
lm(formula = price ~ floor + distelev + ocean + endunit + furn + 
    ocean * floor, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max 
-29857  -7302   -640   5113  43258 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 231321.6     9729.9  23.774  < 2e-16 ***
floor         4432.6     1442.5   3.073 0.002745 ** 
distelev     -1102.5      285.9  -3.856 0.000206 ***
ocean        85988.7     9878.1   8.705 7.77e-14 ***
endunit     -26429.4     4643.2  -5.692 1.31e-07 ***
furn         14927.5     2458.3   6.072 2.40e-08 ***
floor:ocean  -6754.1     1562.1  -4.324 3.69e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 11520 on 98 degrees of freedom
Multiple R-squared:  0.7933,    Adjusted R-squared:  0.7806 
F-statistic: 62.67 on 6 and 98 DF,  p-value: < 2.2e-16
```

How do We Interpret?

- The Model

  Price = 231321 + 4432 Floor - 1102 DistElev + 85988 Ocean
  - 26429 EndUnit+ 14927 Furn - 6754 oce*flr

- No Ocean View (ocean=0)

  Price = 231321 + 4432 Floor - 1102 DistElev - 26429 EndUnit+ 14927 Furn

- Ocean View (ocean=1)

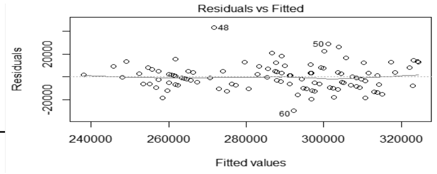
  Price = 317309 - 2322 Floor - 1102 DistElev - 26429 EndUnit+ 14927 Furn

From this model we see that ocean view condos on the lower floors sell for more than ocean view condos on the higher floors. In contrast, bay view condos on the higher floors sell for more than bay view condos on the lower floors.

Always do Diagnostics

- What is this test doing?

```
> ncvTest(fit)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.650776    Df = 1     p = 0.1988533
```



Always do Diagnostics

- What is this test doing?

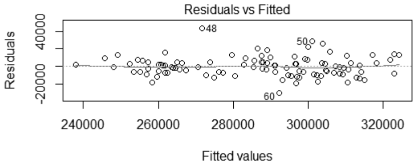
```
> shapiro.test(residuals(fit))

        Shapiro-Wilk normality test

data:  residuals(fit)
W = 0.97229, p-value = 0.02673
```

Stop: HW Help

- The shapiro.test() function only works up to 5000 rows but on the hw we have you work with the homer dataset which has 6057 rows.

```
> install.packages("nortest")

> library(nortest)
> ad.test(residuals(fit))

        Anderson-Darling normality test

data:  residuals(fit)
A = 10.328, p-value < 2.2e-16
```

Deleting an Observation



Hmm….maybe delete that one extreme point and fit again.

Deleting an Observation

- This is the easy way to delete an observation

```
fit=lm(price~floor+distelev+ocean+endunit+furn+ocean*floor,
data=mydata,subset=-48)
```

- If you wanted to delete two observations it would be

```
fit=lm(price~floor+distelev+ocean+endunit+furn
  +ocean*floor,data=mydata,subset=c(-48,-50))
```

Assume we delete row 48

```
> fit=lm(price~floor+distelev+ocean+endunit+furn+ocean*floor,data=mydata,subset=-48)

> ncvTest(fit)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 6.325374    Df = 1     p = 0.01190224


> shapiro.test(residuals(fit))

        Shapiro-Wilk normality test

data:  residuals(fit)
W = 0.99316, p-value = 0.8842
```

Going to log(y) to fix the hetero

```
> fit=lm(log(price)~floor+distelev+ocean+endunit+furn+ocean*floor,data=mydata,subset=-48)
> summary(fit)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.3573270  0.0305970 403.874  < 2e-16 ***
floor        0.0161801  0.0045332   3.569 0.000559 ***
distelev    -0.0033105  0.0009043  -3.661 0.000409 ***
ocean        0.3078659  0.0310571   9.913  < 2e-16 ***
endunit     -0.0896335  0.0145924  -6.142 1.79e-08 ***
furn         0.0490533  0.0077556   6.325 7.79e-09 ***
floor:ocean -0.0238676  0.0049095  -4.862 4.49e-06 ***
---

> ncvTest(fit)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 2.939528    Df = 1     p = 0.08643599
> shapiro.test(residuals(fit))

        Shapiro-Wilk normality test

data:  residuals(fit)
W = 0.9942, p-value = 0.941
```

Back to Model Building Issues

- There are two stepwise commands we have quickly seen
- `step()` does backward regression minimizing AIC
- `model.select()` does “regular” backward regression

Setup Information

- `install.packages("car")`
- `source("http://people.fas.harvard.edu/~mparzen/stat100/model_select.txt")`

Example

- Data on ER visits per year per 1,000 inhabitants, health insurance coverage (%), poverty rate (%), unemployment rate (%), non-citizenship status (%), hospital expenses per inpatient day, teen birth rate per 1,000 for ages 15-19, adults who are overweight/obese (%) and who smoke (%) were collected for each state.

Load the data in

```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/ervisits.csv")
> names(mydata)
[1] "ervisits"    "uninsured"   "poverty"     "unemploy"    "noncitizen"
[6] "expense"     "teenbirth"   "overweight"  "smokers"
```

Fit the full model

- The following commands are equivalent

```
>
fit=lm(ervisits~uninsured+poverty+unemploy+noncitizen+expense+teenbirth+overweight+sm
okers,data=mydata)
```

```
> fit=lm(ervisits~.,data=mydata)
```

Check for heteroskedasticity

- If hetero is present try to log the response variable before going any further (refit full model).

```
> ncvTest(fit)
Non-constant Variance Score Test
Variance formula: ~
fitted.values
Chisquare = 0.04185953    Df = 1
p = 0.8378878
```

Time for Stepwise: Step Command

```
fit1=step(fit)
> summary(fit1)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -169.75074  194.51212  -0.873   0.3877
poverty       16.18174    3.00331   5.388 2.82e-06 ***
noncitizen   -14.59537    3.11116  -4.691 2.76e-05 ***
expense        0.06011    0.02358   2.549   0.0145 *
teenbirth     -3.97088    0.84181  -4.717 2.54e-05 ***
overweight     4.81511    3.26969   1.473   0.1481
smokers        8.03895    3.20234   2.510   0.0159 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46.36 on 43 degrees of freedom
Multiple R-squared:  0.7187,    Adjusted R-squared:  0.6794
F-statistic: 18.31 on 6 and 43 DF,  p-value: 2.037e-10
```

Time for Stepwise: model.select

```
> fit1=model.select(fit)
> summary(fit1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.80617   62.71665   1.623  0.11168
poverty      16.60414    3.02902   5.482 1.94e-06 ***
noncitizen  -14.81738    3.14851  -4.706 2.53e-05 ***
expense       0.05047    0.02295   2.199  0.03321 *
teenbirth    -3.92411    0.85231  -4.604 3.52e-05 ***
smokers      10.15379    2.90004   3.501  0.00107 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46.97 on 44 degrees of freedom
Multiple R-squared:  0.7045,    Adjusted R-squared:  0.6709
F-statistic: 20.98 on 5 and 44 DF,  p-value: 1.156e-10
```

Compare and Contrast

- How do the models agree?
- How do they differ?
- Which one is better? How can you tell?

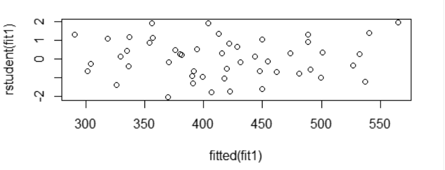
Diagnostic Plot

- We always want to check the residuals versus the fitted values.
- If we are being very careful we want to plot residuals versus each x variable in the model.

Diagnostic Plot

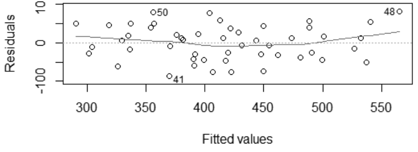
- Residuals versus Fitted Values

```
plot(fitted(fit1),rstudent(fit1))
```



Lazy Diagnostic Plot

```
plot(fit1,which=1)
```



R prints our row numbers of observations you may want to investigate.

Example: Extreme residuals

```
cbind(1:50,rstudent(fit1))

38   38 -0.5802768
39   39  0.8151473
40   40 -0.7279558
41   41 -2.0510622
42   42  0.8768419
43   43  0.4123925
44   44  1.2829659
45   45 -0.6704279
46   46  0.8673237
47   47 -0.6584675
48   48  1.9427279
49   49 -1.6362080
50   50  1.8820404
```

Subsetting Dara

- Say we want to get rid of rows with extreme residuals (for example purposes > 1.8).

```
> newdata=subset(mydata,abs(rstudent(fit1))<1.8)
> dim(newdata)
[1] 46  9
> dim(mydata)
[1] 50  9
```

Refit your final stepwise model

- You could go back and do stepwise again on the reduced data set, but better to just refit final model:

```
fit2=update(fit1,.~.,data=newdata)
```

- This command says fit the same model as before but to a different data set.

Transform Variables

- Suppose you want to try 1/poverty instead of poverty in the model:

```
> fit2=update(fit1,.~.-poverty+I(1/poverty),data=mydata)
```

The Output

```
> fit2=update(fit1,.~.-poverty+I(1/poverty),data=mydata)
> summary(fit2)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.639e+02  1.214e+02   4.644 3.09e-05 ***
noncitizen   -1.479e+01  3.633e+00  -4.070 0.000192 ***
expense       5.001e-02  2.585e-02   1.934 0.059519 .  
teenbirth    -3.036e+00  9.181e-01  -3.307 0.001885 ** 
smokers       9.989e+00  3.322e+00   3.007 0.004350 ** 
I(1/poverty) -3.498e+03  9.166e+02  -3.816 0.000420 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 52.82 on 44 degrees of freedom
Multiple R-squared:  0.6263,    Adjusted R-squared:  0.5838 
F-statistic: 14.75 on 5 and 44 DF,  p-value: 1.711e-08
```

Example: Wine Data

- Data on white wine produced in a particular area of Portugal.
- Data are collected on 12 different properties of the wines one of which is Quality, based on sensory data, and the rest are on chemical properties of the wines including density, acidity, alcohol content etc.
- All chemical properties of wines are continuous variables.
- Quality is an ordinal variable with possible ranking from 1 (worst) to 10 (best). Each variety of wine is tasted by three independent tasters and the final rank assigned is the median rank given by the tasters.

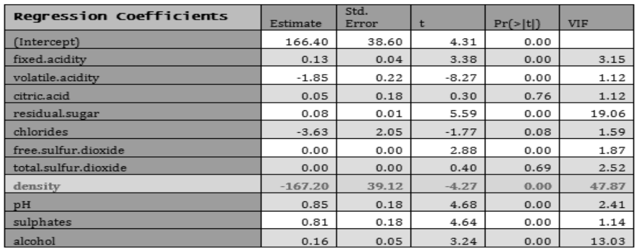
Example

Note lm(quality~., data=mydata)

```
> fit=lm(quality~.,data=mydata)
> summary(fit)

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           1.664e+02  3.860e+01   4.311 1.70e-05 ***
fixed.acidity         1.289e-01  3.813e-02   3.381 0.000736 ***
volatile.acidity     -1.848e+00  2.235e-01  -8.269 2.41e-16 ***
citric.acid           5.470e-02  1.814e-01   0.302 0.763035    
residual.sugar        8.112e-02  1.452e-02   5.588 2.60e-08 ***
chlorides            -3.629e+00  2.047e+00  -1.773 0.076420 .  
free.sulfur.dioxide   4.187e-03  1.456e-03   2.875 0.004086 ** 
total.sulfur.dioxide  2.471e-04  6.208e-04   0.398 0.690678    
density              -1.672e+02  3.912e+01  -4.273 2.01e-05 ***
pH                    8.481e-01  1.813e-01   4.678 3.09e-06 ***
sulphates             8.124e-01  1.752e-01   4.636 3.78e-06 ***
alcohol               1.550e-01  4.778e-02   3.243 0.001201 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7303 on 2025 degrees of freedom
Multiple R-squared:  0.2588,    Adjusted R-squared:  0.2548 
F-statistic: 64.29 on 11 and 2025 DF,  p-value: < 2.2e-16
```

Get the vif's

- Any issues?

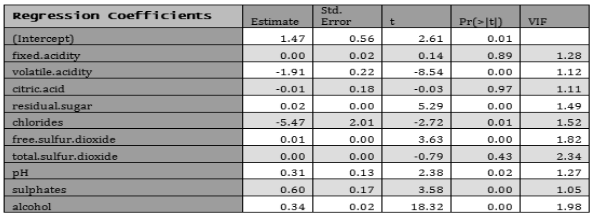


Regression Coefficients	Estimate	Std. Error	t	Pr(>|t|)	VIF
(Intercept)	166.40	38.60	4.31	0.00	
fixed.acidity	0.13	0.04	3.38	0.00	3.15
volatile.acidity	-1.85	0.22	-8.27	0.00	1.12
citric.acid	0.05	0.18	0.30	0.76	1.12
residual.sugar	0.08	0.01	5.59	0.00	19.06
chlorides	-3.63	2.05	-1.77	0.08	1.59
free.sulfur.dioxide	0.00	0.00	2.88	0.00	1.87
total.sulfur.dioxide	0.00	0.00	0.40	0.69	2.52
density	-167.20	39.12	-4.27	0.00	47.87
pH	0.85	0.18	4.68	0.00	2.41
sulphates	0.81	0.18	4.64	0.00	1.14
alcohol	0.16	0.05	3.24	0.00	13.03

Remove density and refit

```
> fit=lm(quality~.-density,data=mydata)
> vif(fit)
```

- The VIF values are much better



Regression Coefficients	Estimate	Std. Error	t	Pr(>|t|)	VIF
(Intercept)	1.47	0.56	2.61	0.01	
fixed.acidity	0.00	0.02	0.14	0.89	1.28
volatile.acidity	-1.91	0.22	-8.54	0.00	1.12
citric.acid	-0.01	0.18	-0.03	0.97	1.11
residual.sugar	0.02	0.00	5.29	0.00	1.49
chlorides	-5.47	2.01	-2.72	0.01	1.52
free.sulfur.dioxide	0.01	0.00	3.63	0.00	1.82
total.sulfur.dioxide	0.00	0.00	-0.79	0.43	2.34
pH	0.31	0.13	2.38	0.02	1.27
sulphates	0.60	0.17	3.58	0.00	1.05
alcohol	0.34	0.02	18.32	0.00	1.98

Full Model without Density

```
> summary(fit)

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           1.4698740  0.5642289   2.605 0.009252 ** 
fixed.acidity         0.0035206  0.0244523   0.144 0.885531    
volatile.acidity     -1.9114802  0.2239115  -8.537  < 2e-16 ***
citric.acid          -0.0062076  0.1816262  -0.034 0.972738    
residual.sugar        0.0215600  0.0040737   5.293 1.34e-07 ***
chlorides            -5.4665194  2.0100052  -2.720 0.006591 ** 
free.sulfur.dioxide   0.0052361  0.0014416   3.632 0.000288 ***
total.sulfur.dioxide -0.0004756  0.0005998  -0.793 0.427915    
pH                    0.3140129  0.1318996   2.381 0.017372 *  
sulphates             0.6044193  0.1690694   3.575 0.000358 ***
alcohol               0.3429519  0.0187239  18.316  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7334 on 2026 degrees of freedom
Multiple R-squared:  0.2521,    Adjusted R-squared:  0.2485 
F-statistic: 68.31 on 10 and 2026 DF,  p-value: < 2.2e-16
```

Any Evidence of Heteroskedasticity?

```
> fit=lm(quality~.-density,data=mydata)
> ncvTest(fit)
Non-constant Variance Score Test 
Variance formula: ~ fitted.values 
Chisquare = 16.35592    Df = 1     p = 5.2492e-05
```

Use Log to fix

```
> fit1=lm(log(quality)~.-density,data=mydata)
> ncvTest(fit1)
Non-constant Variance Score Test 
Variance formula: ~ fitted.values 
Chisquare = 0.8949963    Df = 1     p = 0.344127
```

The Full Model

```
 summary(fit1)

Call:
lm(formula = log(quality) ~ . - density, data = mydata)

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           1.055e+00  9.612e-02  10.977  < 2e-16 ***
fixed.acidity        -2.541e-04  4.165e-03  -0.061 0.951362    
volatile.acidity     -3.473e-01  3.814e-02  -9.105  < 2e-16 ***
citric.acid          -3.084e-03  3.094e-02  -0.100 0.920620    
residual.sugar        3.559e-03  6.940e-04   5.128  3.2e-07 ***
chlorides            -9.548e-01  3.424e-01  -2.788 0.005347 ** 
free.sulfur.dioxide   9.150e-04  2.456e-04   3.726 0.000200 ***
total.sulfur.dioxide -5.733e-05  1.022e-04  -0.561 0.574823    
pH                    4.478e-02  2.247e-02   1.993 0.046386 *  
sulphates             1.107e-01  2.880e-02   3.843 0.000125 ***
alcohol               5.737e-02  3.190e-03  17.987  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1249 on 2026 degrees of freedom
Multiple R-squared:  0.2475,    Adjusted R-squared:  0.2438 
F-statistic: 66.65 on 10 and 2026 DF,  p-value: < 2.2e-16
```

Do a backward stepwise regression

49

```
fit2=step(fit1)

> summary(fit2)

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          1.0466198  0.0775847  13.490  < 2e-16 ***
volatile.acidity    -0.3507931  0.0367358  -9.549  < 2e-16 ***
residual.sugar       0.0035128  0.0006885   5.102 3.67e-07 ***
chlorides           -0.9924030  0.3362986  -2.951 0.003204 ** 
free.sulfur.dioxide  0.0008321  0.0001960   4.246 2.28e-05 ***
pH                   0.0448842  0.0204734   2.192 0.028470 *  
sulphates            0.1082149  0.0284751   3.800 0.000149 ***
alcohol              0.0577924  0.0030748  18.795  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1248 on 2029 degrees of freedom
Multiple R-squared:  0.2474,    Adjusted R-squared:  0.2448 
F-statistic: 95.29 on 7 and 2029 DF,  p-value: < 2.2e-16
```

Let's Start Testing Things

50

- Note-I do things in weird, non-efficient ways.
- In no way is the following the correct way to do things. Just an illustration of what can be done.

Check for hetero and normality

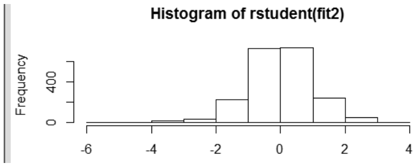
51

```
> ncvTest(fit2)
Non-constant Variance Score Test 
Variance formula: ~ fitted.values 
Chisquare = 0.8620769    Df = 1     p = 0.3531581 


> shapiro.test(residuals(fit2))

        Shapiro-Wilk normality test

data:  residuals(fit2) 
W = 0.98411, p-value = 2.81e-14
```

Histogram of the Standardized Residuals

52

```
hist(rstudent(fit2))
```



Naïve-Delete all "bad" residuals

53

```
> newdata=subset(mydata,abs(rstudent(fit1))<2)
> dim(mydata)
[1] 2037   12
> dim(newdata)
[1] 1933   12
> fit3=update(fit2,.~.,data=newdata)
```

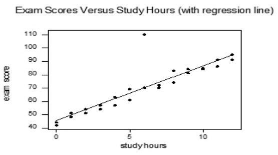
New Test Results-still not normal

54

```
> ncvTest(fit3)
Non-constant Variance Score Test 
Variance formula: ~ fitted.values 
Chisquare = 1.076325    Df = 1     p = 0.2995212 

> shapiro.test(residuals(fit3))

        Shapiro-Wilk normality test

data:  residuals(fit3) 
W = 0.99283, p-value = 4.214e-08
```

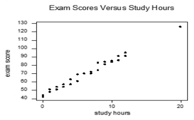
More advanced residual analysis

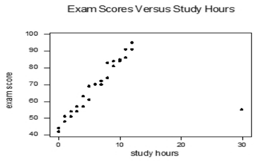
- Lets quickly review what we know about residuals and introduce a useful diagnostic called Cook’s Distance.
- **Cook’s Distance** measures how influential each point is



Recall: Not all outliers are bad

Example of an outlier that doesn't have a large residual:

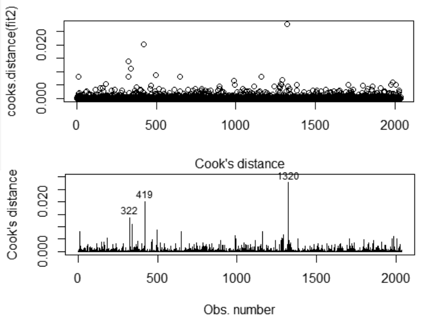




Influential observations are the worst.

Using R

```
plot(cooks.distance(fit2))
```



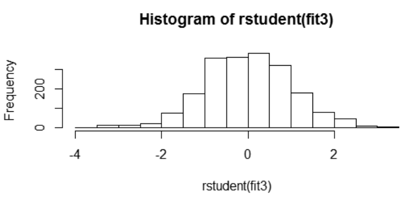
Cook’s distance values larger than usual are cases we want to examine more.
Some people say larger than 4/n are troublesome.

```
plot(fit2,which=4)
```

A little arbitrary what to drop

```
> newdata=mydata[-c(322,419,1320),]
> fit3=update(fit2,.~.,data=newdata)
> ncvTest(fit3)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.0691454    Df = 1     p = 0.7925852
> shapiro.test(residuals(fit3))

        Shapiro-Wilk normality test

data:  residuals(fit3)
W = 0.99438, p-value = 5.652e-07
```

Check Residuals



Where to from here?

- Diagnostic plots
- Maybe transform some x variables
- Remove more observations

Review Time

1) A dummy variable can be assigned up to three values.
   a) True
   b) False

2) Transformations may be used when nonlinear relationships exist between the response and explanatory variable when performing regression.
   a) True
   b) False

3) The value of the coefficient of determination can never decrease when more variables are added to the model.
   a) True
   b) False

Review Time

4) For statistical tests of significance about the regression coefficients, the null hypothesis is that the slope is 1.
   a. True
   b. False

5) If the assumptions of regression have been met, residuals plotted against the independent variable(s) will typically show patterns.
   a) True
   b) False

6) The noise in a regression model is assumed to have zero variance.
   a. True
   b. False

Review Time

11) If the equation of the least squares regression line was computed to be y=45.7+3.1x, then the correlation cannot be less than 0.
   a. True
   b. False

12) If the equation of the regression line that relates percent blood alcohol (x) to reaction time in milliseconds (y) is y=36 - 1.3x, then the slope tells us that for every percent increase in blood alcohol, we can expect reaction time to go down by 1.3 milliseconds
   a. True
   b. False

13) A researcher found the correlation between age of death and number of cigarettes smoked per day to be -0.95. Based just on this information, the researcher can justly conclude that smoking causes early death.
   a. True
   b. False

Review Time

18) A least-squares regression line is not just any line drawn through the points of a scatterplot. What is special about a least-squares regression line?

   a) It passes through all the points.
   b) It minimizes the squared values of the data.
   c) It has slope equal to the correlation between the two variables.
   d) It minimizes the sum of the squared vertical distances of the data points from the line.

Review Time

20) Suppose that the least-squares regression line for predicting y from x is y = 100 + 1.3x. Which of the following is a possible value for the correlation between x and y?

   a) 1.3
   b) –1.3
   c) 0
   d) –0.5
   e) 0.5

Review Time

25) Which of the following is NOT an assumption of the Binomial distribution?

   a) All trials must be identical.
   b) All trials must be independent.
   c) Each trial must be classified as a success or a failure
   d) The number of successes in the trials is counted.
   e) The probability of success is equal to .5 in all trials.

Review Time

34) The weight of a gum drop (piece of candy) in ounces is normally distributed with mean 2 and standard deviation 0.25. A bag contains 10 independent gum drops. The probability that the total weight of the gum drops in the bag exceeds 20 ounces is

   a) 0.25
   b) 0.5
   c) 0.33
   d) 0.75
   e) 0.35

Review Time

36) The purpose of hypothesis testing is to help the researcher reach a conclusion about _______________ by examining the data contained in _______________.

a) a population, a sample
b) an experiment, a computer printout
c) a population, an event
d) a sample, a population

Review Time

37) If the coefficient of determination (R^2) is 0.80, then which of the following is true regarding the slope of the regression line?

a) All we can tell is that it must be positive.
b) It must be 0.80
c) It must be 0.89.
d) Cannot tell the sign or the value.
e) The slope must be significant.

Review Time

39) A multiple regression model with two independent variables exhibits a highly significant F-ratio, but each variable's individual t-statistic is insignificant. The most likely cause of such a situation is

a) Heteroskedasticity
b) Homoskedasticity
c) Multicollinearity
d) Non-normality of residuals

Review Time

41) What is the meaning of the term "heteroscedasticity"?

a) The variance of the errors is not constant
b) The variance of the dependent variable is not constant
c) The errors are not linearly independent of one another
d) The errors have non-zero mean

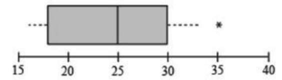
Review Time

61) Suppose we obtain the following regression model for baseball bat sales (Y) when regressed against seasonal indicator variables; $\hat{y} = 100 - 40Spring + 20Wtr - 15Fall$. If we decide to make the baseline season Fall, what would then be the resulting coefficient for Winter (Wtr)?

a) 25
b) -40
c) 30
d) 15
e) None of the above

Review Time

65) Season's Pizza delivers food items to homes in their local area. The following box-and-whisker plot describes the distribution for delivery times in minutes.



Based on this plot, which one of the following statements is correct?

A) The average delivery time is 25 minutes.
B) There are no outliers in this data set.
C) The 75th percentile in this data set is 30 minutes.
D) The second quartile is approximately 18 minutes.
E) None of the above

Review Time

43) Which of the following can NOT be answered from a regression equation?

a) Predict the value of y at a particular value of x.
b) Estimate the slope between y and x.
c) Estimate whether the linear association is positive or negative.
d) Estimate whether the association is linear or non-linear

Review Time

42) Suppose you have estimated wage = 5 + 3education + 2gender – edu*gender, where gender is one for male and zero for female. Suppose instead that gender had been one for female and zero for male. Under this coding what would be the the sum of the coefficients for the gender and interaction variables? (that is we want $b_{gender} + b_{edu*gender}$)

a) -3
b) -1
c) 0
d) 1
e) 2