Stat 104: Quantitative Methods for Economists

Class 37: Dummy Variables and More Diagnostics

Example: Brick Houses

- We have data on 128 recent sales in Mid City.
- For each sale, the file shows the neighborhood (1, 2, or 3) in which the house is located, the number of offers made on the house, the square footage, whether the house is made primarily of brick, the number of bathrooms, the number of bedrooms, and the selling price.
- Neighborhoods 1 and 2 are more traditional neighborhoods, whereas neighborhood 3 is a newer, more prestigious neighborhood.

Snapshot of Data

	A	B	C	D	E	F	G	H	I	J	K
1	Home	Nbhd	Offers	Sq Ft	Brick	Bedrooms	Bathrooms	Price	Nbhd1	Nbhd2	Nbhd3
2	1	2	2	1790	0	2	2	114300	0	1	0
3	2	2	3	2030	0	4	2	114200	0	1	0
4	3	2	1	1740	0	3	2	114800	0	1	0
5	4	2	3	1980	0	3	2	94700	0	1	0
6	5	2	3	2130	0	3	3	119800	0	1	0
7	6	1	2	1780	0	3	2	114600	1	0	0
8	7	3	3	1830	1	3	3	151600	0	0	1
9	8	3	2	2160	0	4	2	150700	0	0	1
10	9	2	3	2110	0	4	2	119200	0	1	0
11	10	2	3	1730	0	3	3	104000	0	1	0
12	11	2	3	2030	1	3	2	132500	0	1	0

Is there a brick premium

- All else equal, do buyers pay a premium for a brick house?

```
> fit=lm(Price~Offers+Sq.Ft+Brick+Bedrooms+Bathrooms+Nbhd2+Nbhd3,data=foo)
> summary(fit)

Call:
lm(formula = Price ~ Offers + Sq.Ft + Brick + Bedrooms + Bathrooms +
    Nbhd2 + Nbhd3, data = foo)

Residuals:
     Min       1Q   Median       3Q      Max
-27337.3  -6549.5    -41.7   5803.4  27359.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2159.498   8877.810   0.243  0.80823
Offers      -8267.488   1084.777  -7.621 6.47e-12 ***
Sq.Ft          52.994      5.734   9.242 1.10e-15 ***
Brick       17297.350   1981.616   8.729 1.78e-14 ***
Bedrooms     4246.794   1597.911   2.658  0.00894 **
Bathrooms    7883.278   2117.035   3.724  0.00030 ***
Nbhd2       -1560.579   2396.765  -0.651  0.51621
Nbhd3       20681.037   3148.954   6.568 1.38e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10020 on 120 degrees of freedom
Multiple R-squared:  0.8686,    Adjusted R-squared:  0.861
F-statistic: 113.3 on 7 and 120 DF,  p-value: < 2.2e-16
```


Is there a Neighborhood 3 Premium?

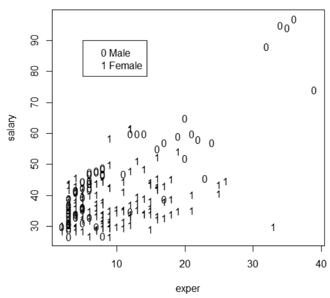

```
> fit=lm(Price~Offers+Sq.Ft+Brick+Bedrooms+Bathrooms+Nbhd2+Nbhd3,data=foo)
> summary(fit)

Call:
lm(formula = Price ~ Offers + Sq.Ft + Brick + Bedrooms + Bathrooms +
    Nbhd2 + Nbhd3, data = foo)

Residuals:
     Min       1Q   Median       3Q      Max
-27337.3  -6549.5    -41.7   5803.4  27359.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2159.498   8877.810   0.243  0.80823
Offers      -8267.488   1084.777  -7.621 6.47e-12 ***
Sq.Ft          52.994      5.734   9.242 1.10e-15 ***
Brick       17297.350   1981.616   8.729 1.78e-14 ***
Bedrooms     4246.794   1597.911   2.658  0.00894 **
Bathrooms    7883.278   2117.035   3.724  0.00030 ***
Nbhd2       -1560.579   2396.765  -0.651  0.51621
Nbhd3       20681.037   3148.954   6.568 1.38e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10020 on 120 degrees of freedom
Multiple R-squared:  0.8686,    Adjusted R-squared:  0.861
F-statistic: 113.3 on 7 and 120 DF,  p-value: < 2.2e-16
```


Interaction Variables

- Another type of variable used in regression models is an interaction variable.
- This is usually formulated as the product of two variables; for example, $x_3 = x_1x_2$
- With this variable in the model, it means the level of x_2 changes how x_1 affects Y

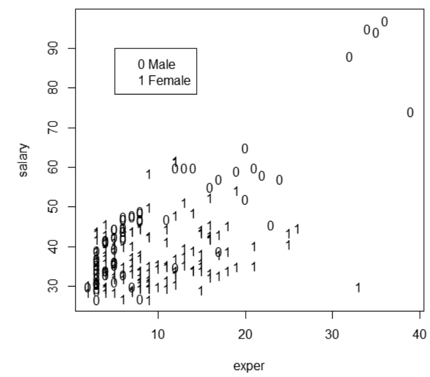
Bank Data Again

- Examine the graph-do you see two lines with different intercepts and slopes?



To model different slopes you need an interaction term.

Salary Versus Years of Experience



1 female
0 male

At all levels of experience, the **male** salaries appear higher.

The Interaction Model

With two *x* variables the model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$$

If we factor out x_1 we get:

$$y = \beta_0 + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2 + e$$

so each value of x_2 yields a different slope in the relationship between *y* and x_1

Interaction Involving an Indicator

If one of the two variables is binary, the interaction produces a model with two different slopes.

When $x_2 = 0$

$$y = \beta_0 + \beta_1 x_1 + e$$

When $x_2 = 1$

$$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + e$$

Example: Discrimination (again)

- In the Bank Case, suppose we suspected that the salary difference by gender changed with different levels of experience
- To investigate this, we created a new variable MEXP = EXPER*MALES and added it to the model.

Regression Output

```
> mexp=exper*male
> fit=lm(salary~exper+male+mexp)
> summary(fit)

Call:
lm(formula = salary ~ exper + male + mexp)

Residuals:
     Min       1Q   Median       3Q      Max
-20.0685  -4.6506  -0.7679   4.4034  23.9122

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.5283     1.1380  30.342  < 2e-16 ***
exper         0.2800     0.1025   2.733  0.00684 **
male         -4.0983     1.6658  -2.460  0.01472 *
mexp          1.2478     0.1367   9.130  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.816 on 204 degrees of freedom
Multiple R-squared:  0.6386,    Adjusted R-squared:  0.6333
F-statistic: 120.2 on 3 and 204 DF,  p-value: < 2.2e-16
```

How do we interpret the equation this time?

A Slope Adjuster

To see the interaction effect, once again evaluate the equation for the two groups.

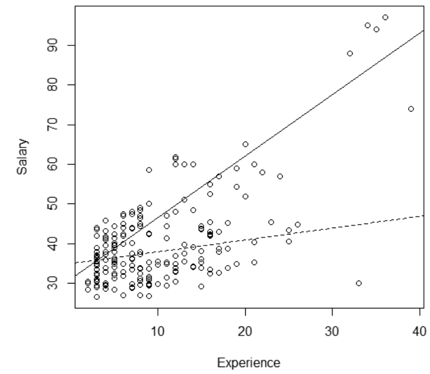
FEMALES (MALES = 0)

SALARY = 35 + 0.3 EXPER - 4 MALES + 1.25 MEXP
= 35 + 0.3 EXPER - 4 (0) + 1.25 (EXPER*0)
= 35 + 0.3 EXPER

MALES (MALES = 1)

SALARY = 35 + 0.3 EXPER - 4 MALES + 1.25 MEXP
= 35 + 0.3 EXPER - 4 (1) + 1.25 (EDUCAT*1)
= 35 + 0.3 EXPER – 4 + 1.25 EXPER
= 31 + 1.55 EXPER

Lines With Two Different Slopes



Women start out at a higher rate, but men make much more money per year of experience.

Are these results significant? What do we examine in the regression output?

What does the following imply?

```
> inter=Brick*Nbhd3
> fit=lm(Price~Offers+Sq.Ft+Brick+Bedrooms+Bathrooms+Nbhd2+Nbhd3+inter)
> summary(fit)

Call:
lm(formula = Price ~ Offers + Sq.Ft + Brick + Bedrooms + Bathrooms +
    Nbhd2 + Nbhd3 + inter)

Residuals:
     Min       1Q   Median       3Q      Max
-26939.1  -5428.7   -213.9   4519.3  26211.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3009.993   8706.264   0.346  0.73016
Offers      -8401.088   1064.370  -7.893 1.62e-12 ***
Sq.Ft          54.065      5.636   9.593  < 2e-16 ***
Brick       13826.465   2405.556   5.748 7.11e-08 ***
Bedrooms     4718.163   1577.613   2.991  0.00338 **
Bathrooms    6463.365   2154.264   3.000  0.00329 **
Nbhd2        -673.028   2376.477  -0.283  0.77751
Nbhd3       17241.413   3391.347   5.084 1.39e-06 ***
inter       10181.577   4165.274   2.444  0.01598 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9817 on 119 degrees of freedom
Multiple R-squared:  0.8749,    Adjusted R-squared:  0.8665
F-statistic:   104 on 8 and 119 DF,  p-value: < 2.2e-16
```


A further look at the noise term

We assume

$$\varepsilon_i \sim N(0, \sigma^2) \quad \textit{independent}$$

This is called homoskedastic noise

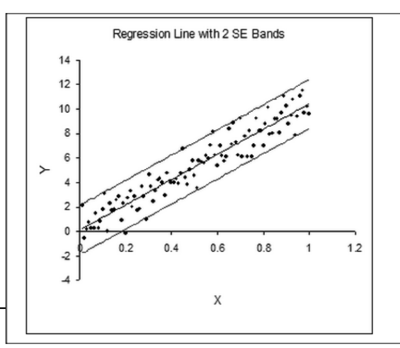
In contrast you could have

$$\varepsilon_i \sim N(0, \sigma_i^2) \quad \textit{independent}$$

This is called heteroskedastic noise

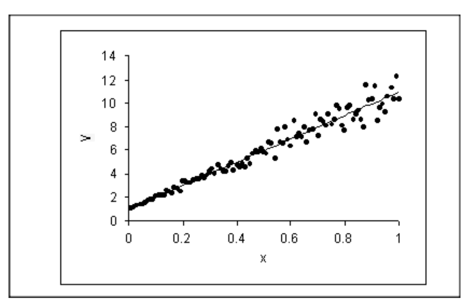
Homoskedasticty Visual

- This is what we assume is happening in regression with our noise:



Heteroskedasticity Visual

- This is (one) example of heteroskedasticity



Many forms of heteroskedasticity

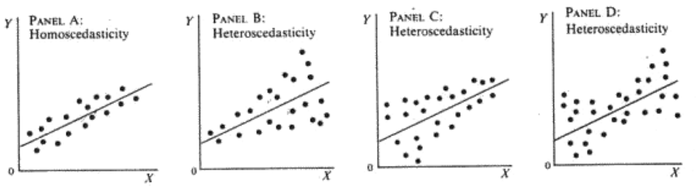


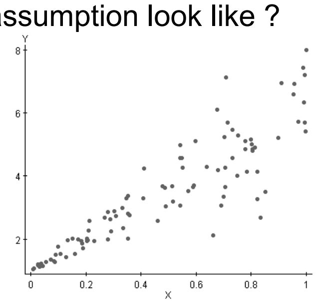
Fig. 9-1

What's the difference?

- If the noise is homoskedastic, there is one term to estimate regarding the noise, σ^2
- If the noise is heteroskedastic, there are *n* things to estimate regarding the noise, σ_i^2
- Would you rather have to estimate 1 thing, or *n* things? That's why we assume we have homoskedastic noise. But we could be wrong.

Non-Constant Variance or Heteroskedasticity

Another of our basic assumptions is that the ε_i all have the same distribution and in particular, the same variance. What does a violation of this assumption look like ?

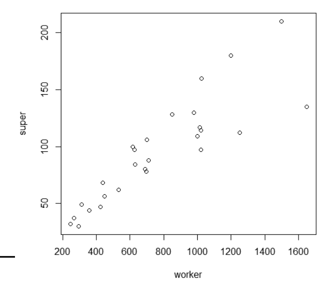


heteroskedasticity means the variance of the errors changes.

our model assumes "homoskedasticity"

Example:

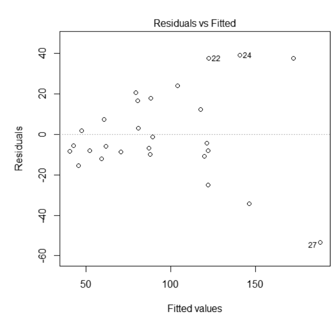
We have data on manufacturing plants for a Fortune 500 company. The data consists of the number of supervisors (Y) and the associated number of supervised workers (X),



Regression Output

```
> fit=lm(super~worker)
> summary(fit)

Call:
lm(formula = super ~ worker)

Residuals:
    Min      1Q  Median      3Q     Max
-53.294  -9.298  -5.579  14.394  39.119

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.44806    9.56201   1.511    0.143
worker       0.10536    0.01133   9.303 1.35e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.73 on 25 degrees of freedom
Multiple R-squared:  0.7759,    Adjusted R-squared:  0.7669
F-statistic: 86.54 on 1 and 25 DF,  p-value: 1.35e-09
```


Residual Diagnostics



If you have heteroskedasaticity, your estimates are ok, but your standard errors are incorrect. That's not good.

Hypothesis testing will be wrong, confidence interval wil be wrong

Good basic solution for heteroskedasticity

- Essentially there is too much variation in the model. That is, there is excess variation in the Y variable.
- An easy way to reduce the variation in Y is to take the log of it.

The log

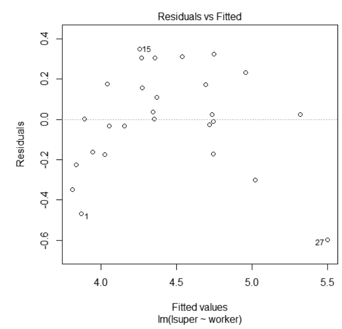
- By “logging” your data, you are transforming it to a different scale.
- The log scale squeezes numbers together, so there is less variation. However, the model becomes slightly different to interpret since the scale of the Y variables changes (instead of dollars we are modeling “log dollars”; what exactly are those ?)

New Regression Output

```
> lsuper=log(super)
> fit1=lm(lsuper~worker)
> summary(fit1)

Call:
lm(formula = lsuper ~ worker)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.59648 -0.16578  0.00244  0.17481  0.34964 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.5150232  0.1110670  31.648  < 2e-16 ***
worker      0.0012041  0.0001316   9.153 1.85e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2524 on 25 degrees of freedom
Multiple R-squared:  0.7702,     Adjusted R-squared:  0.761 
F-statistic: 83.77 on 1 and 25 DF,  p-value: 1.855e-09
```

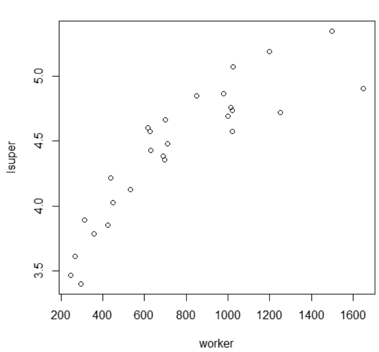
WARNING : now can’t compare s or R-sq
because it’s different units.
earlier was y vs x
now it’s log y vs x

New Residual Plot



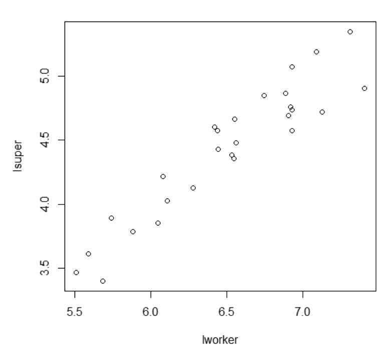
Are we done or is there anything else wrong ?

What transformation should we try ?



Go back to our transformation guide

We will try log(X)

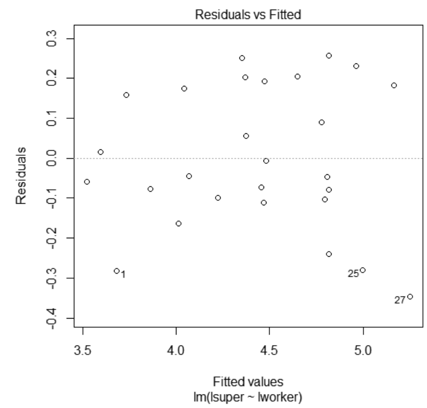


New Output

```
> fit2=lm(lsuper~lworker)
> summary(fit2)

Call:
lm(formula = lsuper ~ lworker)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.3460 -0.1011 -0.0446  0.1783  0.2568 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.48458    0.43544  -3.409  0.00221 ** 
lworker      0.90920    0.06673  13.625 4.51e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.1814 on 25 degrees of freedom
Multiple R-squared:  0.8813,    Adjusted R-squared:  0.8766 
F-statistic: 185.6 on 1 and 25 DF,  p-value: 4.508e-13
```

this we can compare to previous because data is log y in both cases
Se went down
R2 adjusted went up

How does the residual plot look ?



Example: Auto Data

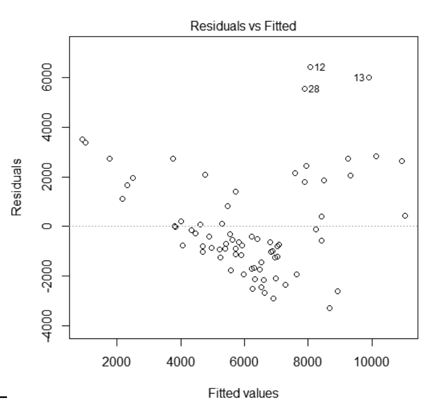
- This (old) data set has information of the price and several explanatory variables of cars from 1978

variable name	storage type	display format	value label	variable label
make	str18	%-18s		Make and Model
price	int	%8.0gc		Price
mpg	int	%8.0g		Mileage (mpg)
rep78	int	%8.0g		Repair Record 1978
headroom	float	%6.1f		Headroom (in.)
trunk	int	%8.0g		Trunk space (cu. ft.)
weight	int	%8.0gc		Weight (lbs.)
length	int	%8.0g		Length (in.)
turn	int	%8.0g		Turn Circle (ft.)
displacement	int	%8.0g		Displacement (cu. in.)
gear_ratio	float	%6.2f		Gear Ratio
foreign	byte	%8.0g	origin	Car type

The Regression Model

```
> fit=lm(price~mpg+weight+trunk+foreign)
> summary(fit)

Call:
lm(formula = price ~ mpg + weight + trunk + foreign)

Residuals:
    Min      1Q  Median      3Q     Max 
-3289.1 -1239.1  -607.1  1346.6  6433.7 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5425.547   3394.248  -1.598    0.115    
mpg            15.394     74.341   0.207    0.837    
weight          3.761      0.685   5.491 6.23e-07 ***
trunk         -87.015     79.047  -1.101    0.275    
foreign      3711.123    683.821   5.427 8.01e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 2128 on 69 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5082,    Adjusted R-squared:  0.4797 
F-statistic: 17.82 on 4 and 69 DF,  p-value: 4.313e-10
```


Diagnostic Plot



Testing for Heteroskedasticity

- This is called the Breusch-Pagan test
- Step 1: Run the full regression
- Step 2: Run the following regression

$$e_i^2 = \alpha_0 + \alpha_1 \hat{Y}_i$$

- Step 3: test if the slope=0 [some finesse to this because of the possible heteroskedasicity]
- This is done with the `ncv.Test()` in the `car` package

Testing for Heteroskedasticity

```
> ncvTest(fit)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 5.938511    Df = 1     p =
0.01481353
```

Ho : It's homoskedastic
Ha: It's heteroskedastic
p is low, Ho must go
there is non constant variation in the noise, we must fix it

One Fix-log the Y variable

```
> lprice=log(price)
> fit1=lm(lprice~mpg+weight+trunk+foreign)
> summary(fit1)

Call:
lm(formula = lprice ~ mpg + weight + trunk + foreign)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.126e+00  4.318e-01  16.502  < 2e-16 ***
mpg         -6.239e-04  9.458e-03  -0.066    0.948
weight       4.756e-04  8.714e-05   5.458 7.08e-07 ***
trunk       -4.928e-03  1.006e-02  -0.490    0.626
foreign      5.370e-01  8.700e-02   6.173 4.05e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2707 on 69 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5496,    Adjusted R-squared:  0.5235
F-statistic: 21.05 on 4 and 69 DF,  p-value: 2.233e-11

> ncvTest(fit1)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.2462385    Df = 1     p = 0.6197362
```


Another Fix-Robust Standard Errors

Heteroscedasticity-consistent standard errors

From Wikipedia, the free encyclopedia

The topic of **heteroscedasticity-consistent (HC) standard errors** arises in statistics and econometrics in the context of linear regression and also time series analysis. The alternative names of **Huber–White standard errors**, **Eicker–White** or **Eicker–Huber–White**[1] are also frequently used in relation to the same ideas.

```
> vcov(fit)
            (Intercept)          mpg       weight        trunk      foreign
(Intercept) 11520922.475 -231194.15857 -1928.3389611 -30744.61425 -976986.8084
mpg          -231194.159    5526.60072    34.6012622    463.88421    8859.3115
weight         -1928.339      34.60126     0.4691758    -21.28866     227.4872
trunk         -30744.614     463.88421   -21.2886589   6248.41732   -2733.2255
foreign      -976986.808    8859.31146   227.4871772  -2733.22550  467610.8021

> hccm(fit)
            (Intercept)          mpg      weight        trunk      foreign
(Intercept) 15895724.483 -307295.78981 -3685.55732 149637.05061 -1514207.9977
mpg          -307295.790    7197.09176    61.29616  -2296.67470   12446.5930
weight         -3685.557      61.29616     1.06737    -64.46125     457.8455
trunk         149637.051   -2296.67470   -64.46125   6679.74624  -20696.8695
foreign     -1514207.998   12446.59295   457.84554 -20696.86951  537292.6440
```


Another Issue: Multicollinearity

- For a regression of Y on k explanatory variables X1,....,Xk, it is hoped that the explanatory variables will be highly correlated with the dependent variable. A relation is sought that will explain a large portion of the variation in Y.
- At the same time, however, it is not desirable for strong relationships to exist **among** the explanatory variables.
- When explanatory variables are correlated with one another, the problem of multicollinearity is said to exist.

The presence of a high degree of multicollinearity among the explanatory variables will result in the following problems:

- The standard deviations of the regression coefficients (s_bi) will be disproportionately large. As a result, the t-ratios will be small. Thus we may think we do not need variables when in fact we do.
- The regression coefficient estimates will be unstable. Because of the high standard errors, reliable estimates are hard to obtain. Signs of the coefficients may be opposite of what is intuitive reasonable. Dropping one variable from the regression will cause large changes in the estimates of the other variables.

Detecting Multicollinearity

- Compare the pairwise correlations between the explanatory variables. One rule of thumb is that multicollinearity may be a serious problem if any pairwise correlation is larger than 0.5

Example:

For the past 12 months, the manager of Pizza Lean-to has been running a series of ads in the local newspaper. The ads are scheduled and paid for in the month before they appear.

Each of the ads contains a two-for-one coupon, which entitles the bearer to two Pizza Lean-to pizzas while only paying for the more expensive pizza.

The manager has collected the data on the following slide and would like to use it to predict pizza sales.

Month	Number of ads appearing X1	Cost of ads appearing X2	Total Pizza Sales Y
May	12	13.9	43.6
June	11	12	38
July	9	9.3	30.1
August	7	9.7	35.3
Septembe	12	12.3	46.4
October	8	11.4	34.2
November	6	9.3	30.2
December	13	14.3	40.7
January	8	10.2	38.5
February	6	8.4	22.6
March	8	11.2	37.6

Simple Model 1

- Regress pi_sales on num_ads

```
> fit1=lm(pisales~numads)
> summary(fit1)

Call:
lm(formula = pisales ~ numads)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8364 -2.7568  0.6804  3.8346  4.8971

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.9369     4.9818   3.400  0.00677 **
numads        2.0832     0.5271   3.952  0.00272 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.206 on 10 degrees of freedom
Multiple R-squared:  0.6097,     Adjusted R-squared:  0.5707
F-statistic: 15.62 on 1 and 10 DF,  p-value: 0.00272
```

looks like we don't need any variable

F-test p-value says Ho must go i.e. you need at least one variable

Simple Model 2

- Regress pi_sales on cost_ads

```
> fit2=lm(pisales~costads)
> summary(fit2)

Call:
lm(formula = pisales ~ costads)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7016 -1.3227 -0.6647  1.7577  6.8957

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.173      7.109   0.587  0.57023
costads        2.873      0.633   4.538  0.00108 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.849 on 10 degrees of freedom
Multiple R-squared:  0.6731,     Adjusted R-squared:  0.6404
F-statistic: 20.59 on 1 and 10 DF,  p-value: 0.001079
```


The Multiple Regression Model

- What happened?

Contradiction! Why?

```
> fit3=lm(pisales~numads+costads)
> summary(fit3)

Call:
lm(formula = pisales ~ numads + costads)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6981 -1.8223 -0.6656  2.4470  6.0123

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5836     8.5422   0.771    0.461
numads        0.6247     1.1203   0.558    0.591
costads       2.1389     1.4701   1.455    0.180

Residual standard error: 3.989 on 9 degrees of freedom
Multiple R-squared:  0.684,     Adjusted R-squared:  0.6138
F-statistic: 9.741 on 2 and 9 DF,  p-value: 0.005604
```


What has happened here ?

In the simple linear regression, each variable is highly significant, and in the multiple regression, they are collectively very significant, but individually not significant.

This apparent contradiction is explained once we notice that the number of ads is highly correlated with the cost of the ads:

```
          pi_sales  num_ads
num_ads     0.781
cost_ads    0.820    0.895
```

Correlation 0.895 is really high. You don't need both. Throw one out. They're giving the same info.

- This is a classic case of multicollinearity. In fact, you might wonder why these two variables are not perfectly correlated. This is because the cost of an ad varies slightly, depending on where it appears in the newspaper.
- Since X1 and X2 are closely related to each other, in effect, they explain the same part of the variability in Y.
- That is why we get R^2=.61 in the first simple regression, R^2=.67 in the second simple regression, but an R^2 of only .68 in the multiple regression.
- Adding the number of ads as a second explanatory variable to the cost of ads explains only about 1 percent more of the variation in total sales.

At this point, it is fair to ask

"Which variable is really explaining the variation in total sales in the multiple regression ?"

The answer is that both are, but we cannot separate out their individual contributions, because they are so highly correlated with each other.

Dealing with Multicollinearity

- Throw out some explanatory variables
- Get more data
- Redefine variables (create an index) (X1+X2)/2
- Step-wise regression

Multicollinearity Again

Variance Inflation Factors (VIF)

- To determine if one X is related to the other X's in the model, we can regress each X on the other X's in the model. That is, let X1,....,Xk be the explanatory variables.
- Perform the regression of Xj on the remaining k-1 explanatory variables and call the coefficient of determination from this model R_j^2.
- We define the variance inflation factor (VIF) for the variable X_j as

$$VIF_j = \frac{1}{1-R_j^2}$$

Interpreting VIF's

- A variance inflation factor can be computed for each X variable in the model. It is a measure of the strength of the relationship between each explanatory variable and all the other explanatory variables in the regression.
- If there is no relationship, R_j^2=0 and VIF_j=1. As R_j^2 increases, VIF_j increases also. For example, if R_j^2=.90, then VIF_j=10.
- A rule of thumb says that if VIF_j>10, then multicollinearity may be a problem with X_j.

Why is it called VIF?

First, consider a multiple regression model with two predictors

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

Let r_{12} denote the correlation between x_1 and x_2 and S_{x_j} denote the standard deviation of x_j. Then it can be shown that

$$\text{Var}(\hat{\beta}_j) = \frac{1}{1-r_{12}^2} \times \frac{\sigma^2}{(n-1)S_{x_j}^2} \qquad j = 1,2$$

Notice how the variance of $\hat{\beta}_j$ gets larger as the absolute value of r_{12} increases. Thus, *correlation amongst the predictors increases the variance of the estimated regression coefficients*. For example, when $r_{12}^2 = 0.99$ the variance of $\hat{\beta}_j$ is $\frac{1}{1-r_{12}^2} = \frac{1}{1-0.99^2} = 50.25$ times larger than it would be if $r_{12}^2 = 0$. The term $\frac{1}{1-r_{12}^2}$ is called a variance inflation factor (VIF).

More on VIF

Let R_j^2 denote the value of R^2 obtained from the regression of x_j on the other x's (i.e., the amount of variability explained by this regression). Then it can be shown that

$$\text{Var}(\hat{\beta}_j) = \frac{1}{1-R_j^2} \times \frac{\sigma^2}{(n-1)S_{x_j}^2} \qquad j = 1,\ldots,p$$

The term $1/(1-R_j^2)$ is called the jth **variance inflation factor (VIF)**.

Example: Auto Data (again)

- Regress Price on Length

```
> fit=lm(price~length)
> summary(fit)

Call:
lm(formula = price ~ length)

Residuals:
    Min      1Q  Median      3Q     Max
-3278.5 -1809.8  -720.8   775.8  8821.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4584.90    2664.44  -1.721 0.089587 .
length         57.20      14.08   4.063 0.000122 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2679 on 72 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1865,    Adjusted R-squared:  0.1752
F-statistic:  16.5 on 1 and 72 DF,  p-value: 0.0001222
```


Example: Auto Data (again)

- Regress Price on Weight

```
> fit=lm(price~weight)
> summary(fit)

Call:
lm(formula = price ~ weight)

Residuals:
    Min      1Q  Median      3Q     Max
-3341.9 -1828.3  -624.1  1232.1  7143.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.7074  1174.4296  -0.006    0.995
weight         2.0441     0.3768   5.424 7.42e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2502 on 72 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.2901,    Adjusted R-squared:  0.2802
F-statistic: 29.42 on 1 and 72 DF,  p-value: 7.416e-07
```


Compare

- What is the sign of the slope coefficient in each model?
- Which model is better?

Full Model

- What is going here with weight and length?

```
> fit=lm(price~mpg+rep78+headroom+trunk+weight+length+turn)
> summary(fit)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16142.479   6487.281   2.488   0.0156 *
mpg          -104.452     80.312  -1.301   0.1983
rep78         723.217    324.880   2.226   0.0297 *
headroom     -655.962    421.397  -1.557   0.1247
trunk          79.229    105.205   0.753   0.4543
weight          5.286      1.133   4.663 1.74e-05 ***
length        -93.325     43.526  -2.144   0.0360 *
turn         -196.632    133.241  -1.476   0.1452
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2215 on 61 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.4811,    Adjusted R-squared:  0.4216
F-statistic:  8.08 on 7 and 61 DF,  p-value: 6.34e-07
```


Check the VIF's

- Aha

```
> vif(fit)
      mpg     rep78  headroom     trunk    weight    length      turn
 3.076441  1.433520  1.791541  2.893446 11.193432 13.586715  4.852827
```