Data Analytics

Yong Zheng

Illinois Institute of Technology

Chicago, IL, 60616, USA

School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

Assignment #3

- Two-sample hypothesis testing
  - If they are independent ➔ z or t test with paired=F
  - If they are dependent
    - Option 1: diff = μ1 – μ2 ➔ one sample z or t test
    - Option 2: z or t test with paired=T
    - In class, we showed Option 1, and the z.test function in the package “BSDA” has no options on “paired”, https://www.rdocumentation.org/packages/BSDA/versions/1.2.0/topics/z.test
    - For option2, you can use z.test function in the package “PASWR2”, https://www.rdocumentation.org/packages/PASWR2/versions/1.0.2/topics/z.test

Multiple Linear Regression

- General Workflow
- Advanced Topics
  - Multicollinearity Problems
  - Dummy Variables (When X is a qualitative variable)
  - Higher-Order Multiple Linear Regressions
  - Interaction Terms
  - Influential Points
- Final Note: Predictions

Multiple Linear Regression

- General Workflow
- Advanced Topics
  - Multicollinearity Problems
  - Dummy Variables (When X is a qualitative variable)
  - Higher-Order Multiple Linear Regressions
  - Interaction Terms
  - Influential Points
- Final Note: Predictions

Multiple Linear Regression (Hold-out Eval)

Important Steps in Multiple Linear Regression

- Data Splits – build a model based on train set, and evaluate it based on the test set
- Determine x and y, examine their linear relationships
- Build a multiple linear regression model by parameter estimates → build diff models by using feature selection
- Goodness of fit test
- Residual analysis – the last step to tell your model is qualified
- Interpret the performance of the training process
- Evaluations and predictions – evaluate it based on test set

Data Splits for Evaluations

1). Hold-out Evaluation → If your data is large enough

Color	Weight (lbs)	Stripes	Tiger?
Orange	300	no	no
White	50	yes	no
Orange	490	yes	yes
White	510	yes	yes
Orange	490	no	no
White	450	no	no
Orange	40	no	no
Orange	200	yes	no
White	500	yes	yes
Green	560	yes	no
Orange	500	yes	?
White	50	yes	?

Training Data Set

Validation Data Set

Unseen data set

Example

```
mydata=read.table("clerical.txt",header=T)
mydata=mydata[sample(nrow(mydata)),]
select.data = sample (1:nrow(mydata), 0.8*nrow(mydata))
train.data = mydata[select.data,]
test.data = mydata[-select.data,]
```

Do not forget to shuffle the data

We use hold-out evaluation
For example. 80% as training

Multiple Linear Regression (N-folds Eval)

Important Steps in Multiple Linear Regression

- ~~Data Splits – build a model based on train set, and evaluate it based on the test set~~
- Determine linear relationship between y and x variables
- Build a multiple linear regression model by parameter estimates
- Goodness of fit test
- Residual analysis – the last step to tell your model is qualified
- Interpret the performance of the training process
- Evaluations and predictions – evaluate it based on test set

N-fold Cross validation

Data Splits for Evaluations

2). N-folds Cross Evaluation → If your data is relatively small

Validation Set

Training Set

Usually we choose N as 5 or 10

Round 1 Round 2 Round 3 Round 10

Validation Accuracy:	93%	90%	91%	95%

Final Accuracy = Average(Round 1, Round 2, ...)

Example

Run 5-fold cross validation

cv.glm() in the package boot

```
> m3=glm(hours~cert+acc+change+check)
> m4=glm(hours~cert+acc+change+check+misc)
> m5=glm(hours~acc+check)
>
> mse3=cv.glm(mydata,m3,K=5)$delta
> mse4=cv.glm(mydata,m4,K=5)$delta
> mse5=cv.glm(mydata,m5,K=5)$delta
>
> mse3
[1] 137.1981 134.5955
> mse4
[1] 132.9957 129.9830
> mse5
[1] 168.4418 166.0293
```

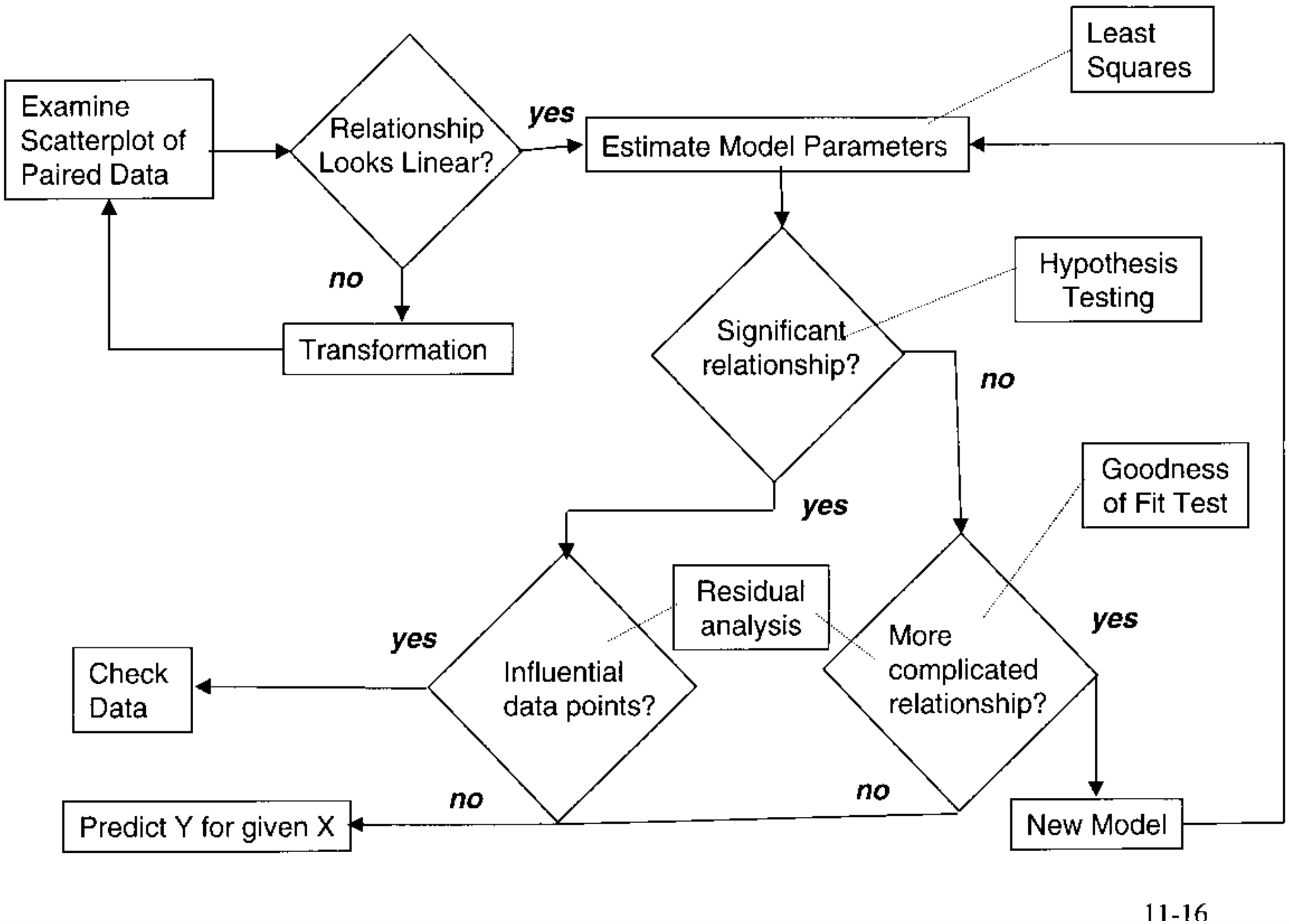
You should build models based on glm() function

Raw MSE value

Adjusted MSE value

MSE = $RMSE^2$

Multiple Linear Regression



Multiple Linear Regression

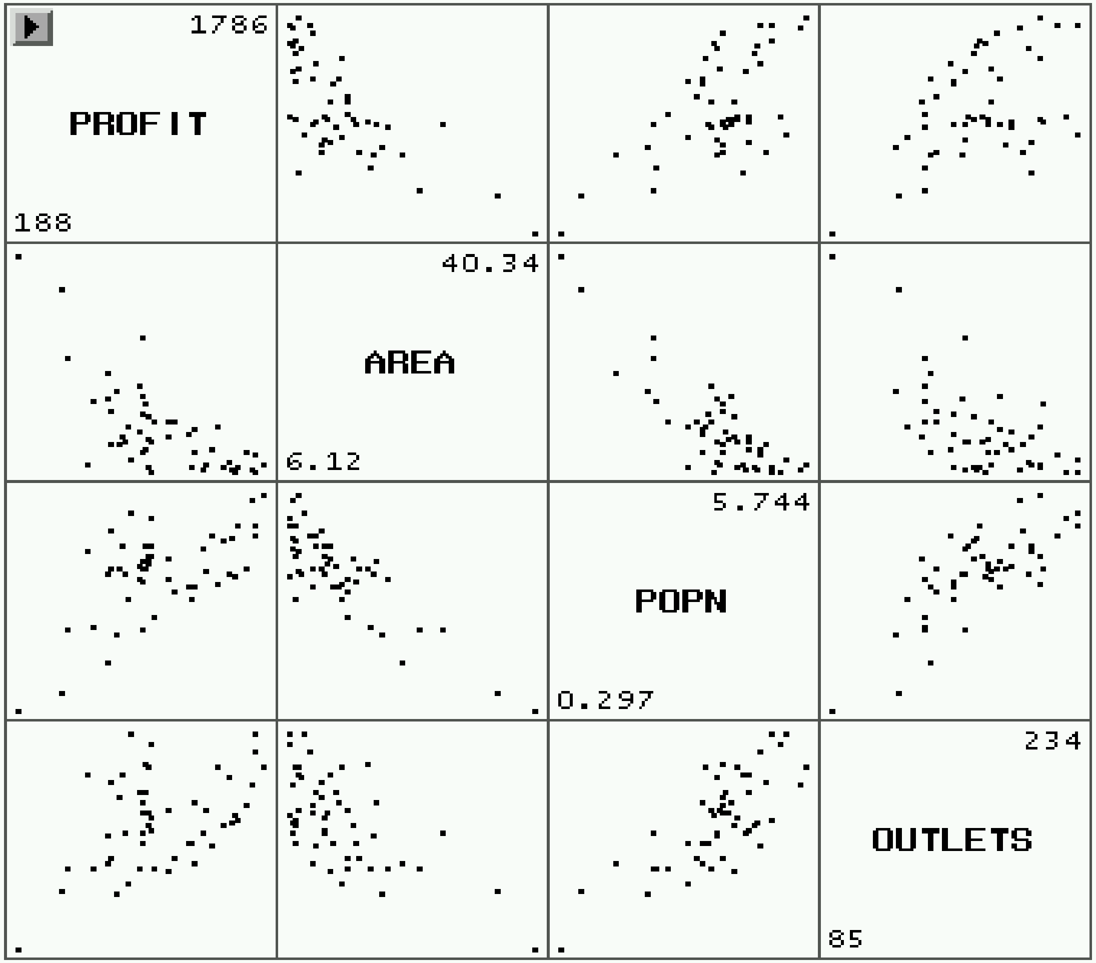
- General Workflow
- Advanced Topics
  - Multicollinearity Problems
  - Dummy Variables (When X is a qualitative variable)
  - Higher-Order Multiple Linear Regressions
  - Interaction Terms
  - Influential Points
- Final Note: Predictions

Multicollinearity Problems

- Multicollinearity refers to the issue that x-variables are strongly correlated.

	Gender	Dept	School
1	M	ITMD	IIT
2	F	ITMS	IIT
3	M	ITMD	IIT
4	M	ITMD	IIT
5	F	ITMS	IIT
6	F	ITMS	IIT

Scatterplot matrix for the 4 quantitative variables.



Which pairs of variables show strong correlation?

Correlation analysis shows some Collinearity

The CORR Procedure

5 Variables: PROFIT AREA POPN OUTLETS COMMIS

Simple Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
PROFIT	51	1120	358.56843	188.00	1786
AREA	51	13.05961	7.03102	6.12	40.34000
POPN	51	3.77822	1.07928	0.297	5.74400
OUTLETS	51	174.21569	30.90651	85.000	234.00000

Pearson Correlation Coefficients, N = 51

Prob > \|r\| under H0: Rho=0

	PROFIT	AREA	POPN	OUTLETS	COMMIS
PROFIT	**1.00000**	**-0.69571**	**0.60172**	**0.46029**	**0.27067**
		<.0001	<.0001	0.0007	0.0547
AREA	**-0.69571**	**1.00000**	**-0.83563**	**-0.63878**	**0.14452**
	<.0001		<.0001	<.0001	0.3116
POPN	**0.60172**	**-0.83563**	**1.00000**	**0.74572**	**-0.31428**
	<.0001	<.0001		<.0001	0.0247
OUTLETS	**0.46029**	**-0.63878**	**0.74572**	**1.00000**	**-0.28831**
	0.0007	<.0001	<.0001		0.0402
COMMIS	**0.27067**	**0.14452**	**-0.31428**	**-0.28831**	**1.00000**
	0.0547	0.3116	0.0247	0.0402	

High correlations among the X variables

Multicollinearity Problems

What to do?

When two X-variables are strongly correlated – there is no need to keep them both in the model! They don't add predictive value to the model.

How do we assess multi-collinearity?

- Pre-processing: Examine the Pearson correlation matrix and the scatter plots for each pair of x-variables. Absolute value of correlation larger than 0.9 or so indicate a serious collinearity problem.
- Post-processing: Build the model first, and then Compute the VIF statistics [suggested!!!]

Variance inflation factor

Tolerance or VIF (variation inflation factor) can be used to assess multivariate multicollinearity.

The value of tolerance for an x-variable is computed by regressing the x-variable on all the others.

If the x-variable is highly correlated with one or more other x-variables, the R^2 value for the regression above is by definition very large.

- **Variance-inflation factor (VIF)** is the variance inflation factor, and is simply the reciprocal of tolerance:

  $VIF=1/(1-R_j^2)$.

  A large value of VIF (larger than 4) is a sign of strong multicollinearity .

Multicollinearity using SAS/R

SAS users

The "tolerance" and "vif" multi-collinearity statistics are computed using the option "vif" or "tol" in the model statement.

```
PROC REG;
MODEL yvar = xvar_1 xvar_2 ... xvar_k / vif tol;
RUN;
```

R users

```
fit = lm(y~xvar1+xvar2)
# Evaluate Collinearity
vif(fit) # variance inflation factors
sqrt(vif(fit)) > 2 # problem?
```

How to identify multicollinearity problem?

How do we assess multi-collinearity?

- Pre-processing: Examine the Pearson correlation matrix and the scatter plots for each pair of x-variables. Correlation values larger than 0.9 or so indicate a serious collinearity problem.
- Post-processing: Compute the VIF statistics

Our suggestions

- Use post-processing and ignore pre-processing
  - We do not know how large correlations can tell a serious problem
  - We do not know which variable to be removed
  - Some variables may be removed after building the model
- How to do by post-processing?
  - Build the model first, then calculate VIF, VIF > 4?
  - If VIF > 4, examine corr of existing x variables in the fitted model

Multiple Linear Regression

- General Workflow
- Advanced Topics
  - Multicollinearity Problems
  - Dummy Variables (When X is a qualitative variable)
  - Higher-Order Multiple Linear Regressions
  - Interaction Terms
  - Influential Points
- Final Note: Predictions

Example – Movie opening ticket sale

A movie producer has two new movie scripts to choose from. He wants to analyze which factors have a strong positive effect on the opening gross revenue of the movies. He collects data on 32 movies released between 1997-1998.

The data are on the variables:

Movie = Title of the movie

Opening = Gross receipts for the weekend after the movie was released (in millions of dollars)

Budget = The total budget for the movie (in millions of dollars)

CHARACTER VARIABLES:

Star = Whether or not the movie has a superstar; VALUE = Star; NoStar

Summer = Whether or not the movie was released in the summer;
VALUE= Summer or NoSummer

ANSWER: Fit a regression model for the gross opening revenue with independent variables chosen among budget, star and summer!

We'll analyze this in class

	Opening	Budget	Star?	Release?
AirForceOne	37.132	85.00	Star	Summer
BatmanandRobin	42.870	110.00	Star	Summer
Bean	2.255	22.00	NoStar	NoSummer
ConAir	24.131	75.00	Star	Summer
Contact	20.584	90.00	Star	Summer
KisstheGirl	13.215	27.00	NoStar	NoSummer
TheLostWorld	92.729	73.00	NoStar	NoSummer
MeninBlack	84.133	90.00	Star	Summer
Metro	18.734	55.00	NoStar	NoSummer
Mimic	7.818	25.00	NoStar	Summer
ThePeacemaker	12.311	50.00	Star	NoSummer
PrivateParts	14.616	20.00	NoStar	NoSummer
TheSaint	16.278	70.00	Star	NoSummer
SoulFood	11.197	7.00	NoStar	NoSummer
....				
Speed2	16.158	110.00	Star	NoSummer
Spawn	21.210	40.00	NoStar	Summer
Volcano	14.581	90.00	NoStar	NoSummer
187	2.912	23.00	NoStar	Summer

How do we include qualitative variables in the regression model?

Each alphanumeric variable is replaced by one or more **dummy variables (that take only 0 or 1 values)**.

For instance:

The variable **Star** is replaced by the *numeric* variable **numstar**.

```
Numstar = 1 if Star = STAR
Numstar = 0 if Star = NOSTAR
```

Analogously for the variable **Summer**:

```
Numsum = 1 if Release= SUMMER
Numsum = 0 if Release = NOSUMMER
```

How do we include qualitative variables in the regression model?

Dummy Variable == Binary Variable

What if a qualitative that has more than 2 values?

Season
Spring
Summer
Fall
Winter
Fall

How do we include qualitative variables in the regression model?

Dummy Variable == Binary Variable

What if a qualitative that has more than 2 values?

Season
Spring
Summer
Fall
Winter
Fall

Spring	Summer	Fall	Winter
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
0	0	1	0

How do we include qualitative variables in the regression model?

Dummy Variable == Binary Variable

What if a qualitative that has more than 2 values?

Season
Spring
Summer
Fall
Winter
Fall

Spring	Summer	Fall
1	0	0
0	1	0
0	0	1
0	0	0
0	0	1

You can convert qualitative variable to multiple dummy variables

Usually N-1 new variables is enough. Not necessary to have N ones

Creating dummy variables in R

METHOD 1

Create dummy variables:

```
numstar= (star == "Star")*1;
numsum= (release == "Summer")*1;
```

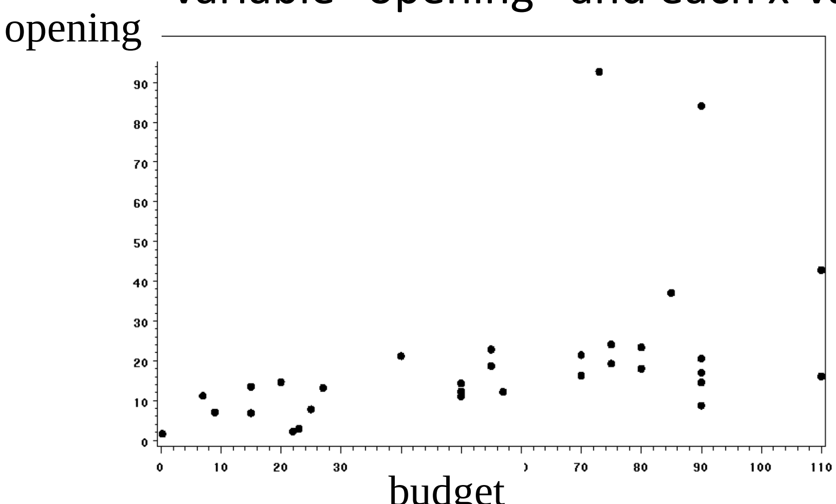
METHOD 2

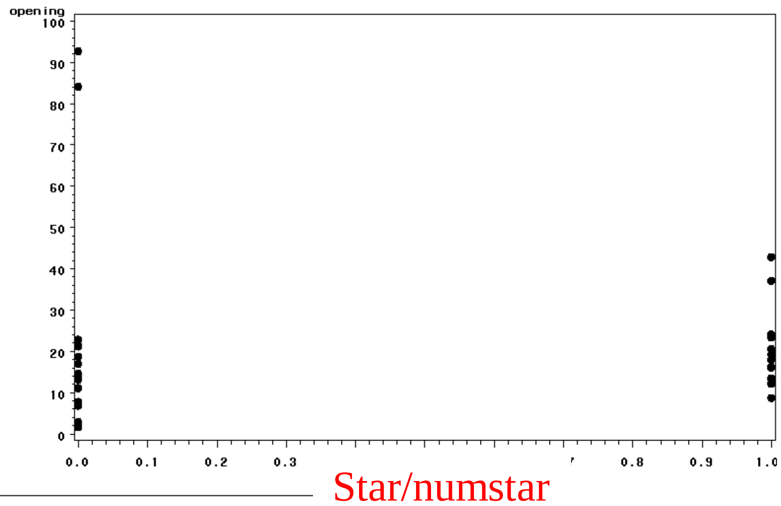
Using the `as.factor()` function to automatically transform the categorical variable in factors or dummy variables to be used in LM() regression model.

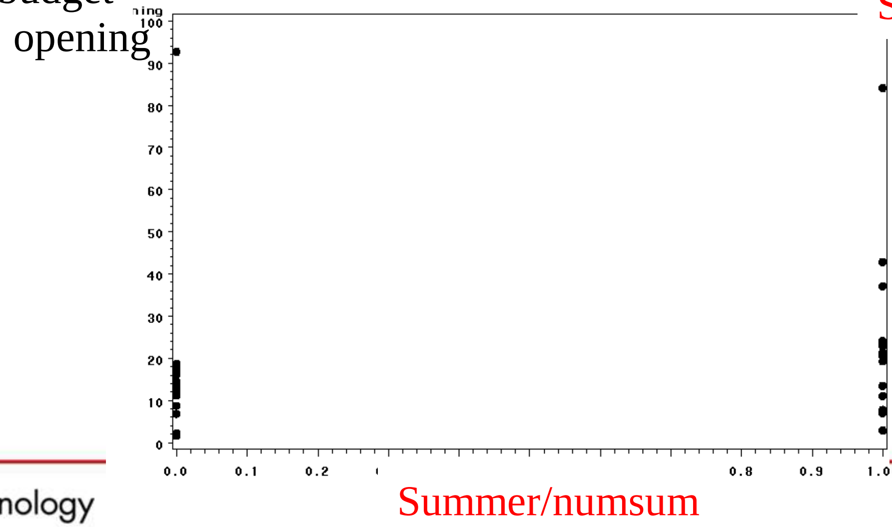
```
fit = lm(y~ xvar1 + xvar2 + as.factor(star)
         +as.factor(release))
summary(fit)
```

Back to our example on the movie data

- **Step 1 : Exploratory data analysis -** examine the scatter plots of the y-variable "opening" and each x-variable.







Correlation matrix

Simple Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
opening	32	20.32619	19.93042	1.64200	92.72900
budget	32	56.19531	32.02662	0.25000	110.00000
numstar	32	0.40625	0.49899	0	1.00000
numsum	32	0.46875	0.50701	0	1.00000

Pearson Correlation Coefficients, N = 32
Prob > |r| under H0: Rho=0

	opening	budget	numstar	numsum
opening	1.00000	0.46839	0.00141	0.1742
		0.0069	0.9939	0.3401
budget	**0.46839**	1.00000	0.51767	0.09748
	0.0069		0.0024	0.5956
numstar	**0.00141**	0.51767	1.00000	0.11555
	0.9939	0.0024		0.5289
numsum	**0.17427**	0.09748	0.11555	1.00000
	0.3401	0.5956	0.5288	

Correlations with dummy variables are hard to interpret

Stronger association between opening revenue and budget money, but the association with star and summer is weak!

Step 2: Fitting the regression model – Find the x-variables that have a significant effect on Y

Start with the **full model**, that includes all the x-variables.

The REG Procedure
Dependent Variable: opening

Analysis of Variance

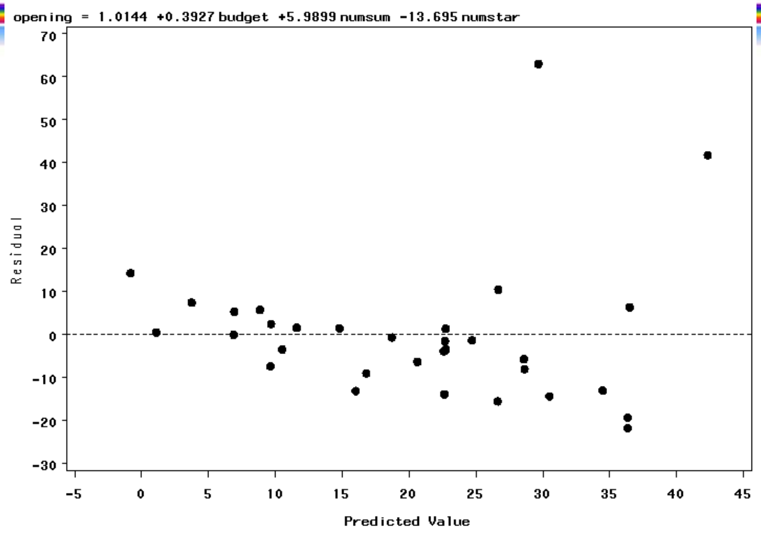
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3960.58314	1320.19438	4.43	0.0115
Error	28	8353.29135	298.33183		
Corrected Total	31	12314			

Root MSE	17.27229	R-Square	0.3216
Dependent Mean	20.32619	Adj R-Sq	0.2490
Coeff Var	84.97553		

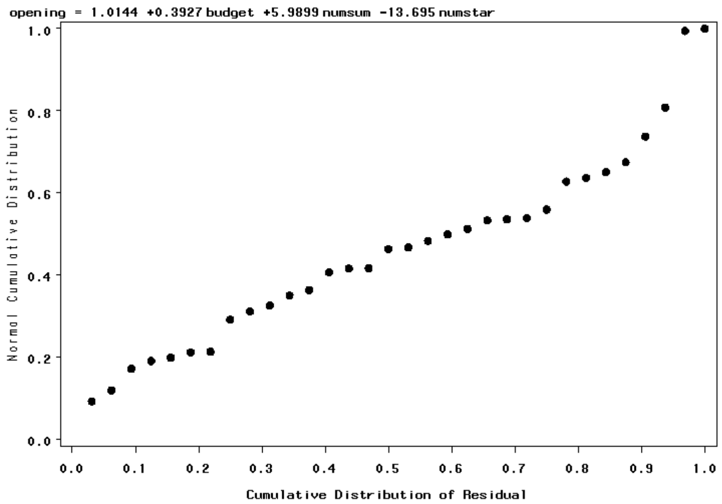
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	**Pr > |t|**
Intercept	1	1.01440	6.68888	0.15	0.8805
budget	1	0.39269	0.11332	3.47	0.0017
numstar	1	-13.69455	7.28767	-1.88	0.0707
numsum	1	5.98995	6.16596	0.97	0.3396

Step 3 – Residual analysis – Plots show some problems!



Residual versus predicted values



Normal probability plots for the model residuals

The residual plots show that the variance is not constant. What can be done?

There are various solutions. The easiest solution is to apply a transformation on the response variable Y to stabilize the variance. Most common transformations are

1. Log(Y) (only if Y not zero)
2. Sqrt(Y) similar to log
3. Square Y = Y^2
4. Cubic Y = Y^3
5. Inverse Y = 1/Y (only for $Y \neq 0$)

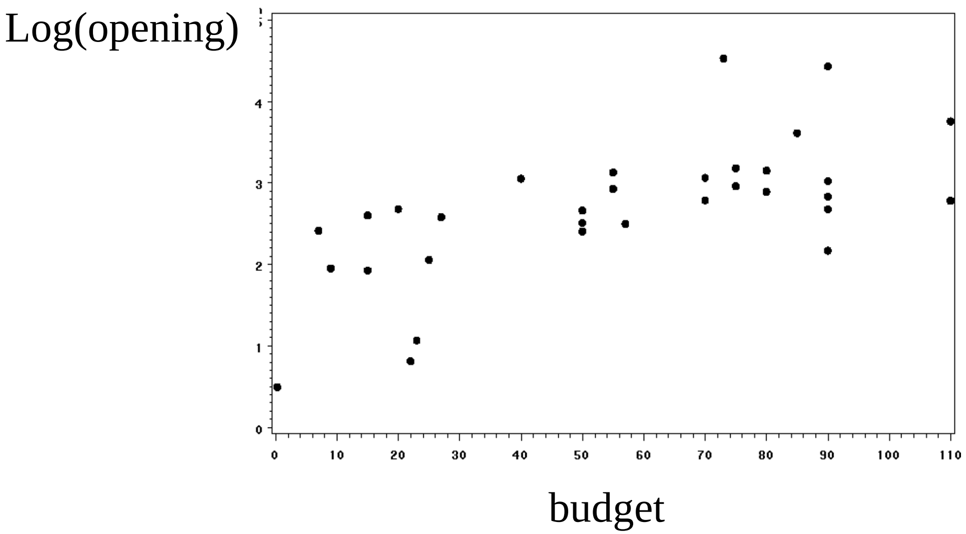
Try them in this order…Fit the regression model on the transformed Y and examine the residual plots to see if the assumptions are now valid!

Perform again the various steps of regression analysis for the new dependent variable log(Y)

Step 1 – Exploratory data analysis

Draw the scatter plots of log(Y) versus each independent variable to check that the transformed variable log(Y) is linearly associated to the x-variables

For instance:



Plot shows that log(Y) and budget are linearly related.

Step 2 - Fit regression model for log(Y) and the x-variables

Start with the full model $\log(y) = \beta_0 + \beta_1 budget + \beta_2 numstar + \beta_3 numsum + e$

The REG Procedure

Dependent Variable: logopen

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	10.93699	3.64566	8.71	0.0003
Error	28	11.72631	0.41880		
Corrected Total	31	22.66330			

Root MSE	0.64715	R-Square	0.4826
Dependent Mean	2.67704	Adj R-Sq	0.4271
Coeff Var	24.17388		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.55414	0.25061	6.20	<.0001
budget	1	0.01919	0.00425	4.52	0.0001
numsum	1	0.32782	0.23102	1.42	0.1669
numstar	1	-0.26913	0.27305	-0.99	0.3327

STEP 2 cont. - Select the x-variables to be included in the model

Examine the results of the t-test for the coefficients of each independent variable. Drop the variable with the largest p-value, because it has the least or no effect on the response variable log(Y).

Rerun the regression analysis without such a variable!

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.55414	0.25061	6.20	<.0001
budget	1	0.01919	0.00425	4.52	0.0001
numsum	1	0.32782	0.23102	1.42	0.1669
numstar	1	-0.26913	0.27305	-0.99	0.3327

Fit the regression model of log(Y) on budget and numsum

The REG Procedure

Dependent Variable: logopen

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	10.53012	5.26506	12.58	0.0001
Error	29	12.13319	0.41839		
Corrected Total	31	22.66330			

Root MSE	0.64683	R-Square	0.4646
Dependent Mean	2.67704	Adj R-Sq	0.4277
Coeff Var	24.16202		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.57344	0.24973	6.30	<.0001
budget	1	0.01705	0.00364	4.68	<.0001
numsum	1	0.31041	0.23023	1.35	0.1880

Are all the regression coefficients significant? If not, eliminate the variable with the largest p-value (>0.05) and rerun the regression analysis for the new model.

Fit the regression model for log(Y) on budget

The REG Procedure

Dependent Variable: logopen

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	9.76959	9.76959	22.73	<.0001
Error	30	12.89372	0.42979		
Corrected Total	31	22.66330			

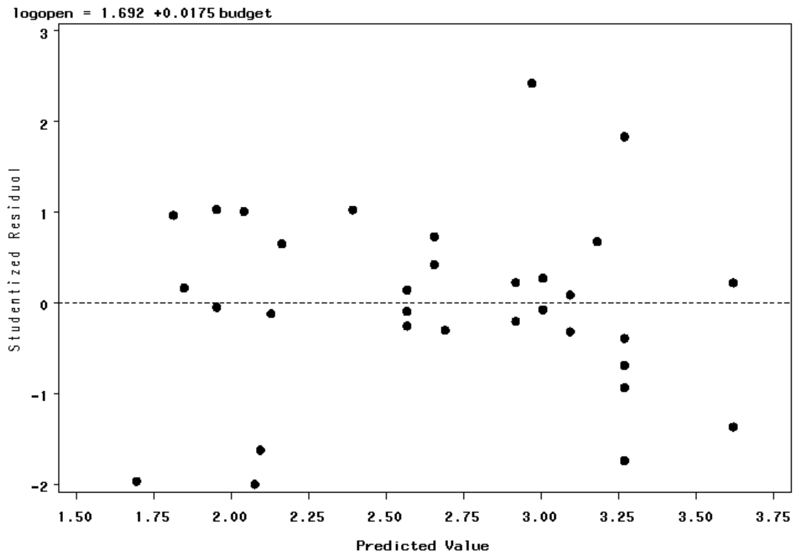
Root MSE	0.65558	R-Square	0.4311
Dependent Mean	2.67704	Adj R-Sq	0.4121
Coeff Var	24.48912		

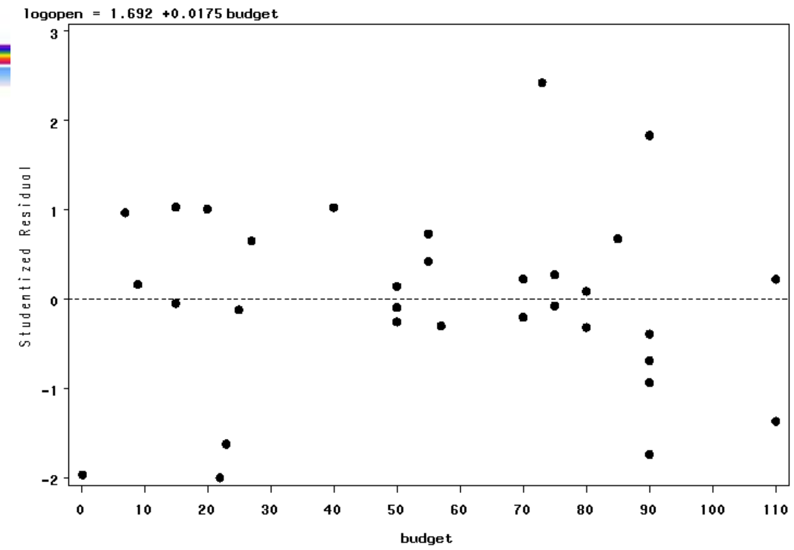
Parameter Estimates

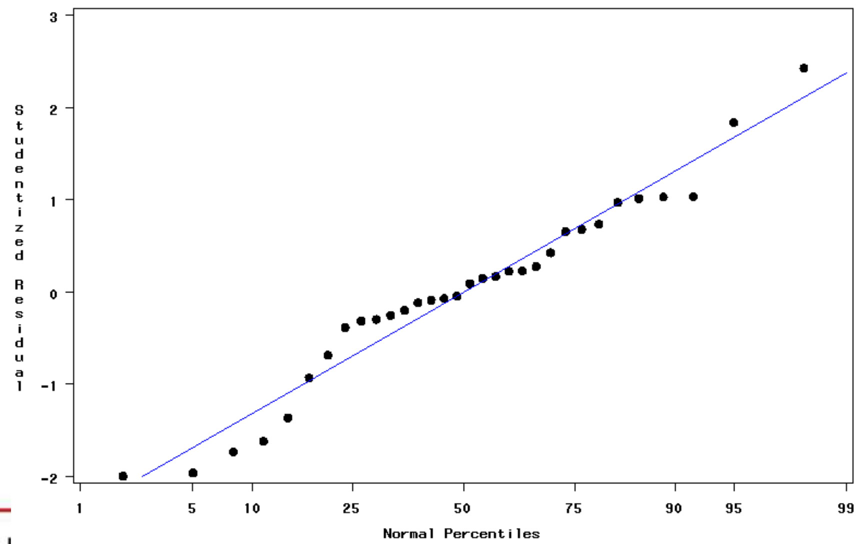
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.69202	0.23689	7.14	<.0001
budget	1	0.01753	0.00368	4.77	<.0001

The t-test for the budget coefficient is significant, indicating that the variable budget has a strong contribution in the explanation of log(Y). The F-test is significant. The value of R^2 indicates that the fitted straight line model explains about 43% of the variation in Y.

Step 3 Diagnostics - The model residual analysis





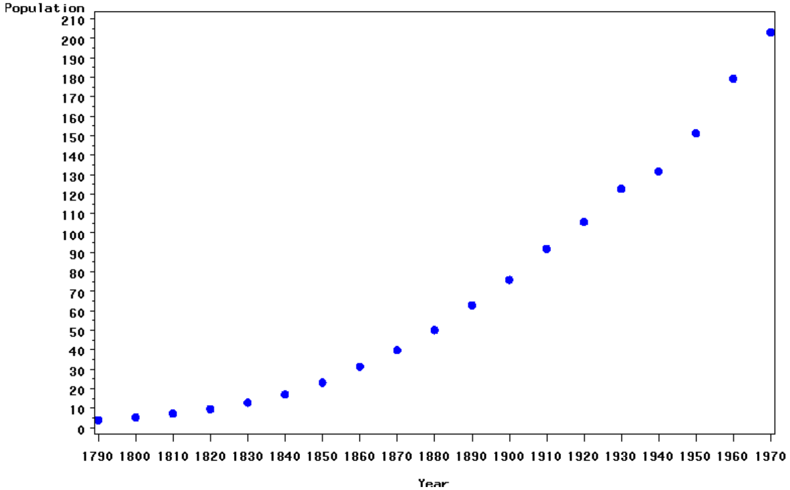


Multiple Linear Regression

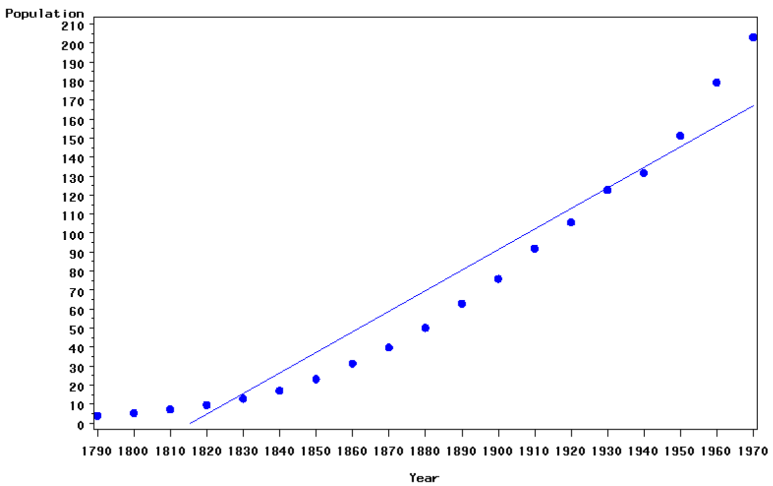
- General Workflow
- Advanced Topics
  - Multicollinearity Problems
  - Dummy Variables (When X is a qualitative variable)
  - Higher-Order Multiple Linear Regressions
  - Interaction Terms
  - Influential Points
- Final Note: Predictions

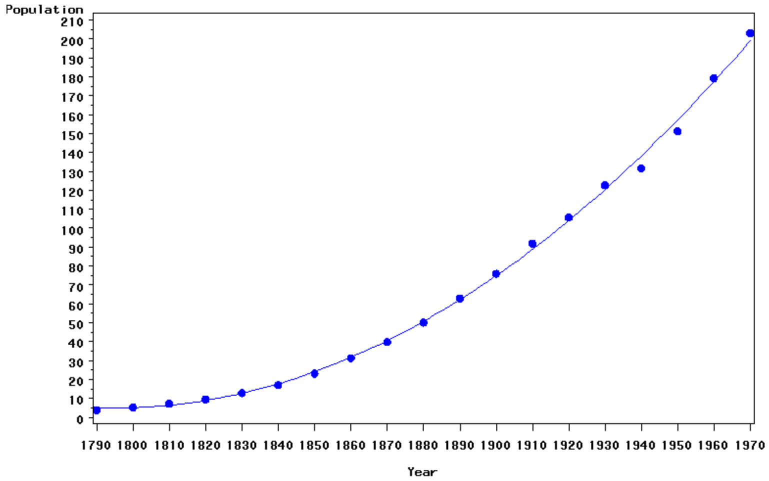
Polynomial regression

Scatter plot shows a quadratic relationship between Y and X.



Line is not a god fit!





Non linear associations

- Detected in scatter plot of response variable Y versus independent variable X: if scatter plot shows a curve, the association is non-linear.
- Use transformation of either X or Y to “straighten out” the curve.
- Typical transformations are
  - Log()
  - Sqrt()
  - Power X^2 or X^3

Polynomial models

- If the association between Y and X is a quadratic or cubic function, *Y* that can be predicted by a polynomial function of *X*.
- Ex: cubic function:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + e$$

- This is simply accomplished by creating additional factors as $X_1 = X$, $X_2 = X^2$, $X_3 = X^3$, etc….

And fitting the cubic regression model

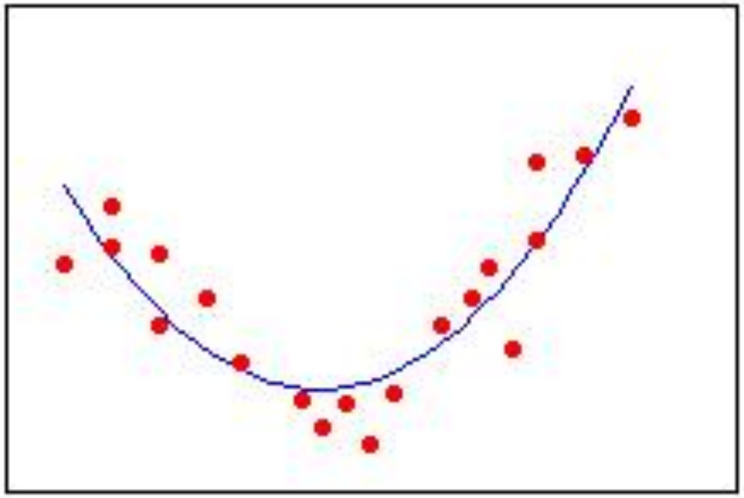
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

Polynomial models

Quadratic

$Y = b_0 + b_1 X + b_{11} X^2$

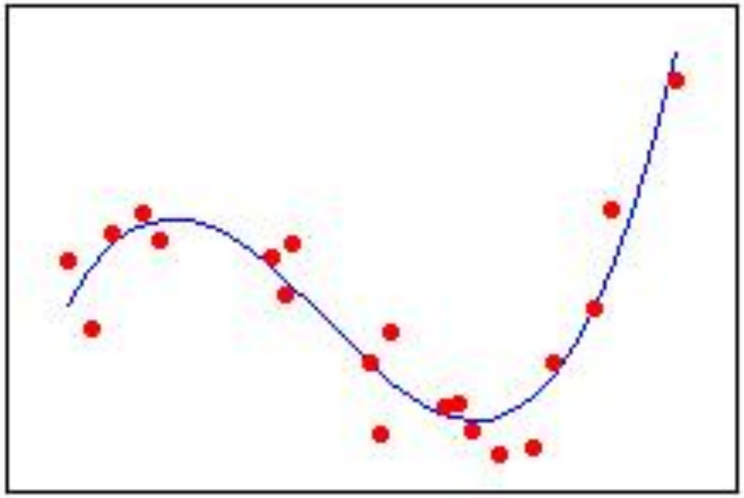
(second order)



Cubic

$Y = b_0 + b_1 X + b_{11} X^2 + b_{111} X^3$

(third order)



Polynomial models

- Based on the shape of the scatter plot, you can make a decision whether there are 2nd order terms or the 3rd order terms
- After that, you can simply create new variables to represent these higher-order terms
- The next steps to build the polynomial model is the same as the way to build multiple linear regression model.
- Special notes: if you are going to add a higher-order term, the lower-order terms should also be added to the model. For example, the 3rd order term is necessary. Therefore, the 1st and 2nd order terms should also be added to build the models

Multiple Linear Regression

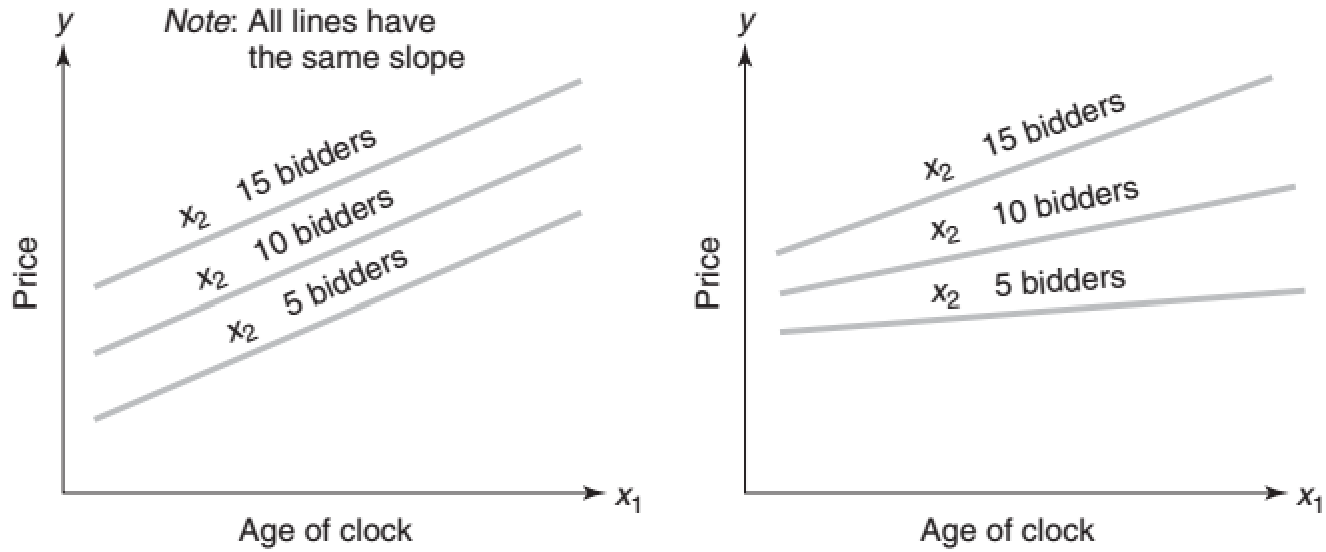
- General Workflow
- Advanced Topics
  - Multicollinearity Problems
  - Dummy Variables (When X is a qualitative variable)
  - Higher-Order Multiple Linear Regressions
  - Interaction Terms => a special case of higher-order
  - Influential Points
- Final Note: Predictions

Interaction models

Assume we build a linear regression model with two independent variables

If we fix the value of x2, and model the relationship between y and x1.

By different x2 values, there should be parallel straight lines



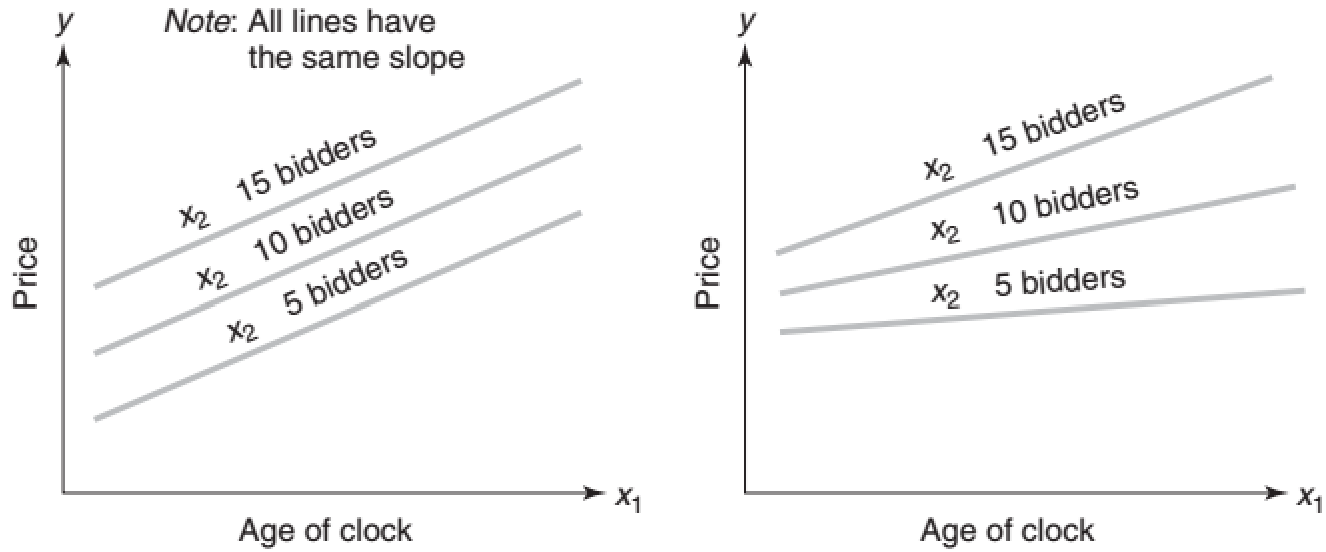
(a) No interaction between x_1 and x_2 (b) Interaction between x_1 and x_2

Interaction models

However, if you can observe straight lines with different slopes, like fig b).

It implies that there should be an interaction term x_1x_2 in your model

This is a special case in higher-order regression models.



(a) No interaction between x_1 and x_2

(b) Interaction between x_1 and x_2

Interaction models

- Modeling changes in response variable Y with quantitative and qualitative variables

Interaction term

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

- Interaction models are useful when associations between Y and X-variables vary with the values of some other variable (slopes are not constant)

- Often used with dummy variables – as association between the response variable Y and a predictor X varies for different levels of the dummy variable

Interaction models

- Once you observe there is an interaction term, you should create a new variable to represent this interaction term.

- You can add the new variable to the multiple linear regression model
  If one of them is dummy variable, you can use the codes below
  fit = lm (y~var1+DAY*var2), where DAY is a dummy variable

- And you can follow the regular steps to build the model

- How to interpret the interaction terms?
  - It is difficult to interpret it if x1 and x2 are two quantitative variables
  - It is relatively easy to interpret it if one of them is a dummy variable
    For example, Male and Female, they may have different impacts on the quantitative variable.

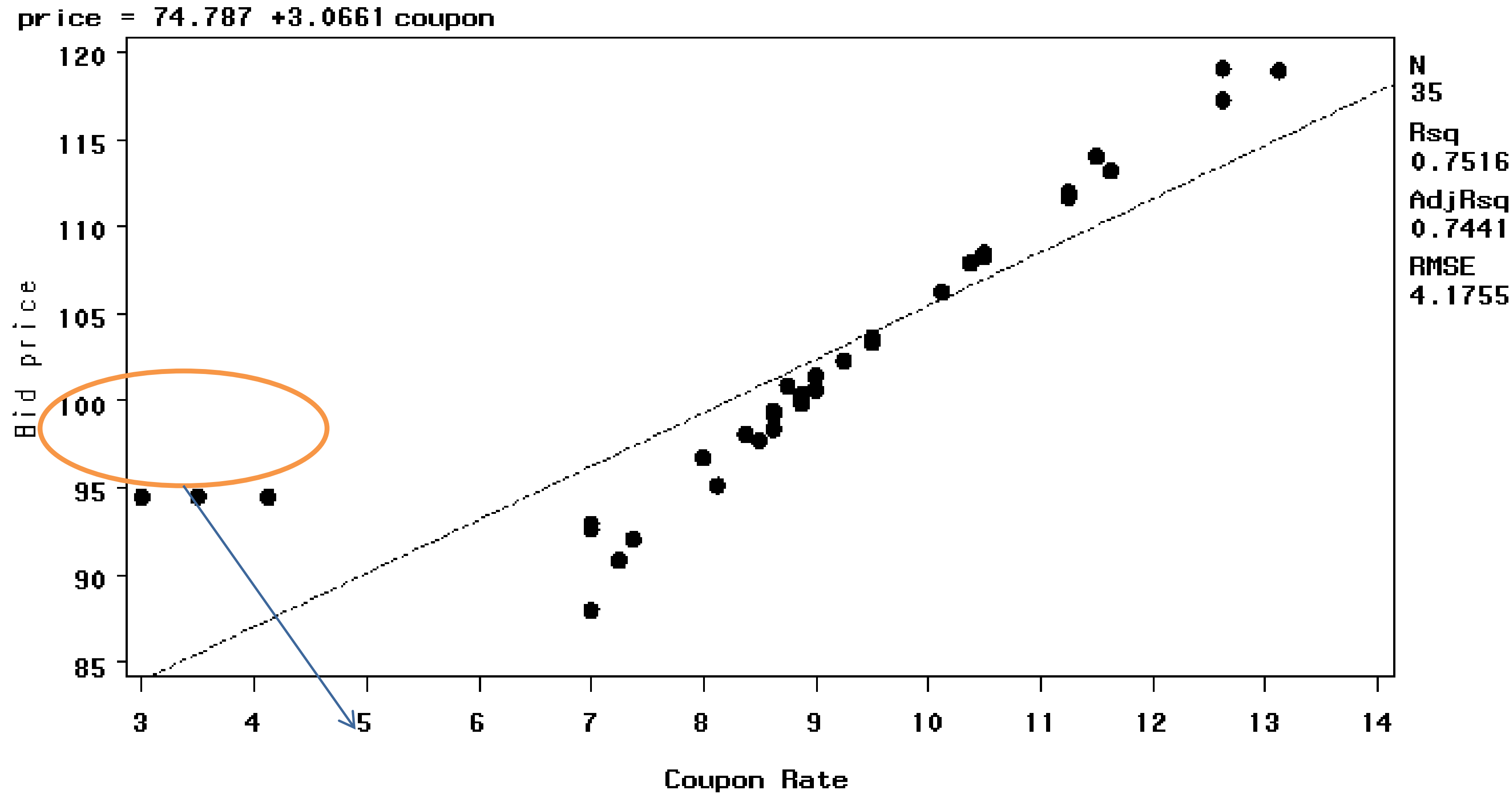
Multiple Linear Regression

- General Workflow
- Advanced Topics
  - Multicollinearity Problems
  - Dummy Variables (When X is a qualitative variable)
  - Higher-Order Multiple Linear Regressions
  - Interaction Terms
  - Influential Points
- Final Note: Predictions

Influential Points

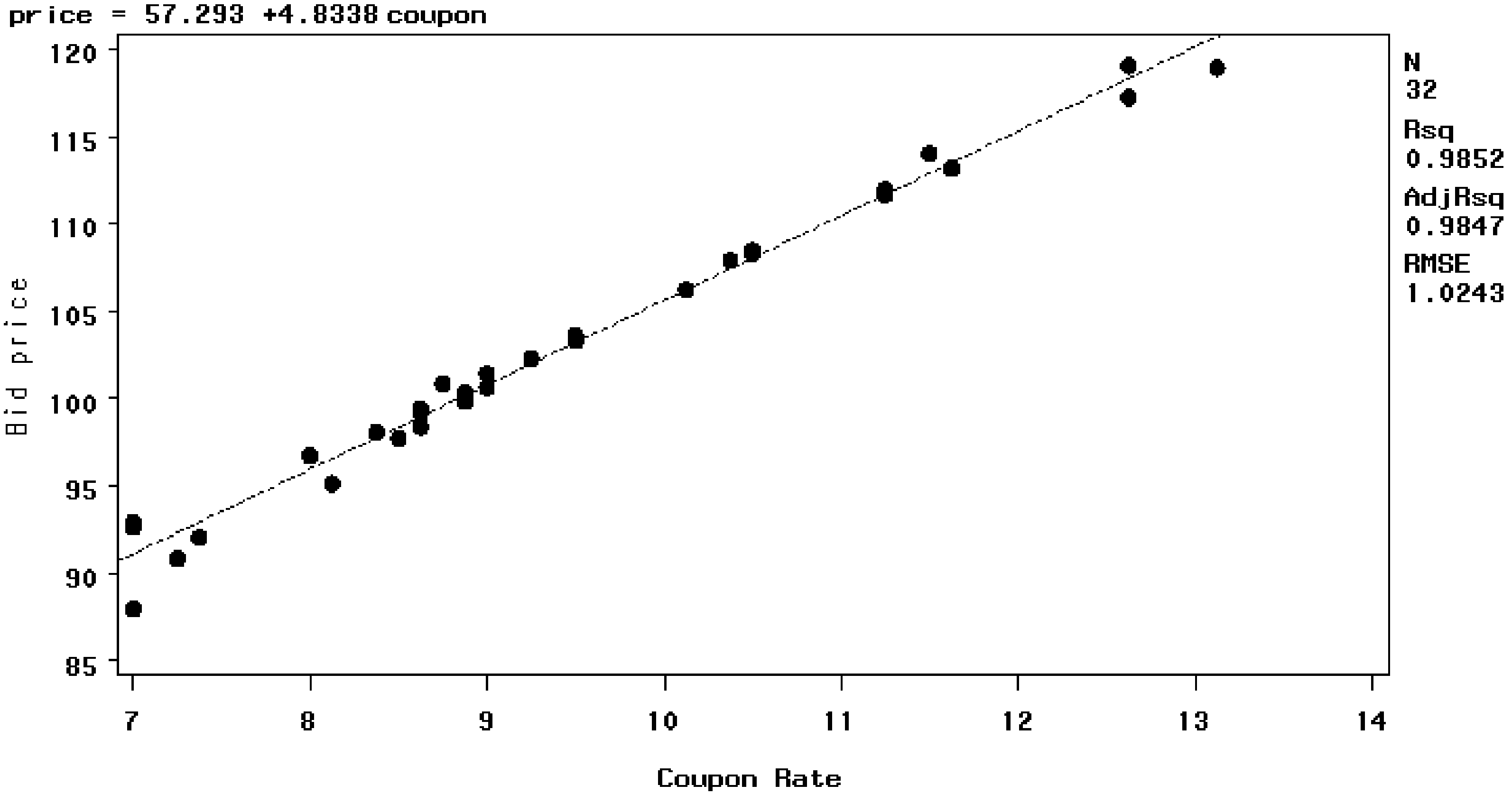
- Influential points are the outliers that affect the fitted model
- Note: not all of the outliers are influential points
- Influential points are observations (typically outliers) that have a strong influence on the fitted model. If removed, the parameter estimates change.

Fitted regression line



Influential Points: They were "flower" bonds with tax advantages, and therefore followed a different model than regular bonds

After removing the influential points



Notice the significant change in the fitted regression line, and the increase in the R^2 value

Outliers vs Influential Points

- Influential points are usually the outliers
- Not all the outliers are influential points
- Outliers can be identified from the data & model
  - From data: outlier detection (a data mining task)
  - From model: residual analysis
- Influential points can only be identified from models
  - "influential": Whether they have impact on the models
  - You need to build models first

Metrics to Identify Influential Points

Function	Description	Rough Cut-off
dffits()	the change in the fitted values (with appropriately scaled)	$\lvert DFFITS \rvert > 2\sqrt{((k+1)/n)}$
dfbetas()	the changes in the**coefficients**(with appropriately scaled)	> 2/sqrt(n)
covratio()	the change in the estimate of OLS covariance matrix	$\lvert covratio-1 \rvert \geq 3*(k+1)/n$
hatvalues()	standardized distance to mean of predictors used to measure the leverage of observation	> 2*(k+1)/n
cooks.distance()	standardized distance change for how far the estimate**vector**	> 4/n

k = Number of x variables

n = Number of records to build the model = the size of your data to build the model

Influential points by R

fit = lm(y~x1+x2+x3)

- Print all of the measures and influential points
  - influence.measure (fit); //influential point measures
  - summary (influence.measure (fit)); //print out only influential observations
- Print measures one by one
  - dfbeta (fit)
  - covratio (fit)
  - dffits (fit)
  - cooks.distance (fit)

Influential Points Identification

```
> mea=influence.measures(m12)
> summary(mea)
Potentially influential observations of
         lm(formula = hours ~ check + cert + cert2 + change + acc) :

   dfb.1_ dfb.chck dfb.cert dfb.crt2 dfb.chng dfb.acc dffit cov.r   cook.d
1  -0.03  -0.04     0.05    -0.05    -0.03     0.02   -0.09  1.57_*  0.00
3   0.06   0.23    -0.26     0.36    -0.04    -0.11    0.52  1.83_*  0.05
4   0.15  -0.34    -0.10     0.03     0.96    -0.18    1.05  1.68_*  0.18
6  -0.18   0.15     0.00    -0.03     0.03     0.15    0.36  1.55_*  0.02
17  0.10  -0.02    -0.17     0.21    -0.07     0.03    0.28  1.85_*  0.01
41  0.25  -0.20     0.32    -0.35    -0.15    -0.21    0.81  0.32_*  0.09
   hat
1   0.24
3   0.40
4   0.44_*
6   0.27
17  0.37
41  0.07
```

Multiple Linear Regression

- General Workflow
- Advanced Topics
  - Multicollinearity Problems
  - Dummy Variables (When X is a qualitative variable)
  - Higher-Order Multiple Linear Regressions
  - Interaction Terms
  - Influential Points
- Final Note: Predictions

A confidence interval for predictions

- Suppose we want to predict a **specific response value Y** at a particular value of the X-variables.

- The **predicted value of Y** for values x^*_1, x^*_2, x^*_3 is computed as

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2^* + \hat{\beta}_3 x_3^*$$

- **Prediction Interval at 95% confidence level:**

$$\hat{y} \pm t_{0.95,n-2} S.E.(\hat{y})$$

$$S.E.(\hat{Y}) = s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Additional term that makes standard error of predictions larger

Prediction and estimations in R

Example of prediction for one data point.
create **new** data frame containing xvalues for prediction

```
new = data.frame(linet=c(7), step=c(6), device=c(3))
```

# use predict() to compute predicted value and standard error
# predict(model_name, new_dataframe, ...)
se.fit=T to compute predicted value

predict(fit, new, se.fit = T)

compute predicted value and prediction interval

predict(fit, new, interval="prediction", level=0.95)

Example of prediction for many data points.

```
linet = c(6, 4, 8)
step = c(6, 3, 1 )
device=c(3, 2, 1)
new <- data.frame(linet, step, device)
```

compute predicted value and standard error

```
predict(fit, new, se.fit = T)
```

compute predicted value and prediction interval

```
predict(fit, new, se.fit = T, interval="prediction", level=0.95)
```

compute average response value and confidence interval

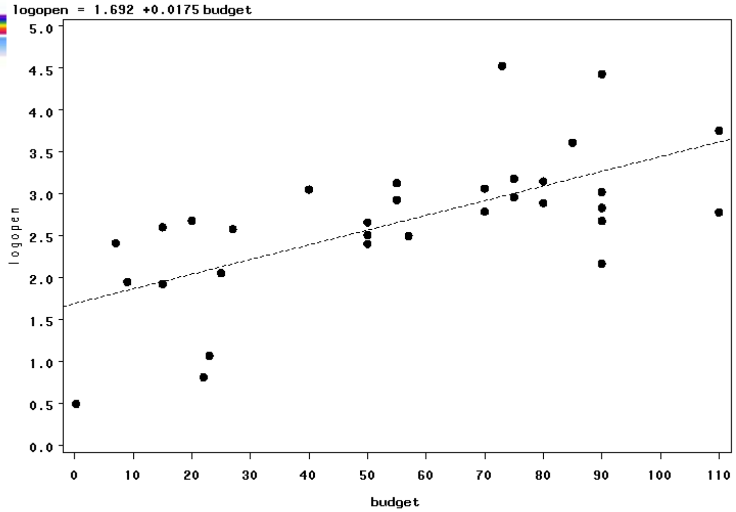
predict(fit, new, se.fit = T, interval="confidence",level=0.95)

Predictions for transformed variables

Data on OPEN = opening revenue for new movies, and BUDGET= cost of the movie. Fitted regression line is

$$\log(open) = 1.692 + 0.0175\, budget$$

Movies with higher budget costs, typically gain more money at their first weekend opening.



Suppose you want to estiamate the average opening revenue for a new movie whose budget was equal to 65 million dollars.

The REG Procedure

Dependent Variable: logopen

Obs	Dep Var logopen	**Predicted Value**	Std Error Mean Predict	95% CL Mean	
		2.8314	0.1203	2.5856	3.0771

Predictions for Original variables

Thus a movie that costs 65 million dollars can expect to gain on **average**

Average Log(Y)= 2.8314 - with 95% C.I. Equal to (2.5856, 3.0771)

Need to transform the dependent variable back to the original value!

Estimated average opening revenue= exp(2.8314)

=16.969 million dollars.

Apply the **same inverse transformation** to the 95% C.I.to obtain an approximate 95% C.I. for the estimated average response.

Thus, the approximate 95% C.I. for the estimated average gross revenues for movies with a budget cost of 65 million dollars is

(exp(2.5856), exp(3.0771))=(13.27, 21.69) million dollars.

Predictions in Linear Regression

- Important Notes
  - Output: predicted value + confidence interval
  - If you applied transformation on the y variable, the predicted value you produce is the predictions based on the transformed y variable. You should convert it back to the original unit
  - For example, $\log(y) = 6 + 2x_1 + 3x_2$
    To get predicted y values, you should use exp() function to be applied on the predicted log(y)

Multiple Linear Regression

- General Workflow
- Advanced Topics
  - Multicollinearity Problems
  - Dummy Variables (When X is a qualitative variable)
  - Higher-Order Multiple Linear Regressions
  - Interaction Terms
  - Influential Points
- Final Note: Predictions

Next Class

- In-Class Practice
  - N-fold Cross validation
  - Advanced Techniques to improve the models
    - Using categorical/dummy variables
    - Examination of multi-collinearity problems
    - Try higher-order terms or interaction terms
    - Improve models by removing influential points