BUDT 730

Data, Models and Decisions

Lecture 14

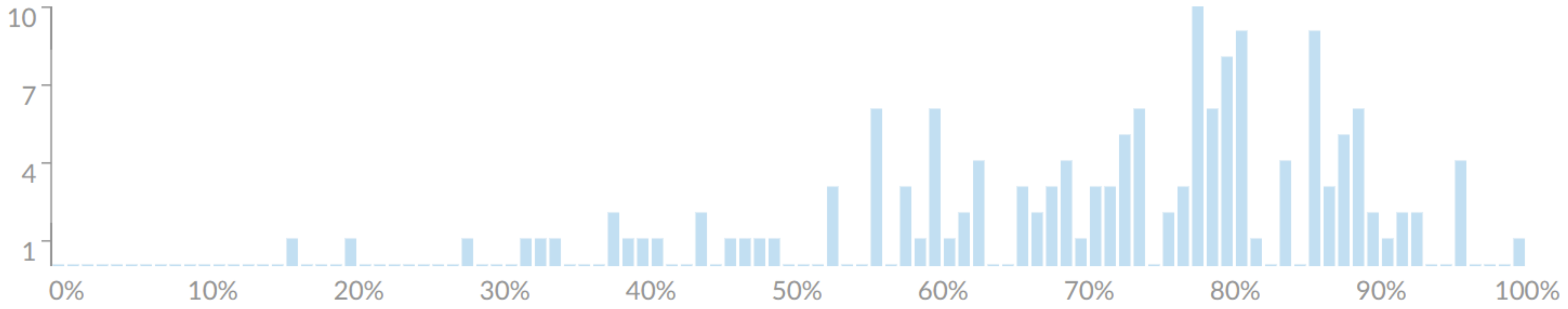
Regression Analysis (6)

Model Validation

Prof. Sujin Kim

Midterm Results

Average Score	High Score	Low Score	Standard Deviation	Average Time
72%	100%	16%	14.77	01:19:32



Review: Nonlinear Transformations Summary

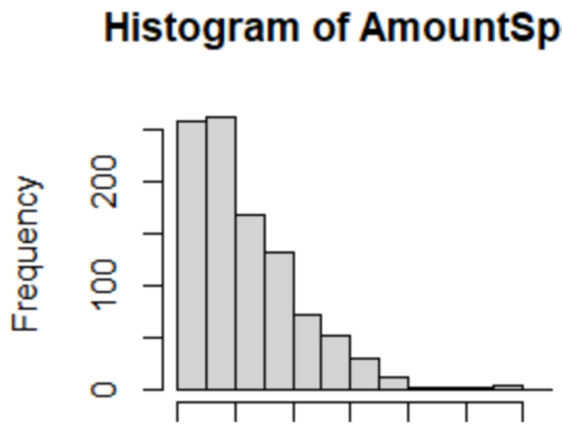
Model	Regression Formula	Interpretation of Model Coefficients
Linear	$Y = a + b X$	Increasing X has a constant effect on Y (b)
Quadratic	$Y = a + b_1 X + b_2 X^2$	$b_1 + 2b_2X$ is the rate of change of Y with respect to X
Log	$Y = a + b \text{ Log}(X)$	When X increases by 1%, Y increases (on average) by b / 100
Exponential	$\text{Log}(Y) = a + bX$	When X increases by one unit, the expected percentage change in Y is approximately b * 100%
Log-Log	$\text{Log}(Y) = a + b \text{ Log}(X)$	When X increases by 1%, Y increases (on average) by b%

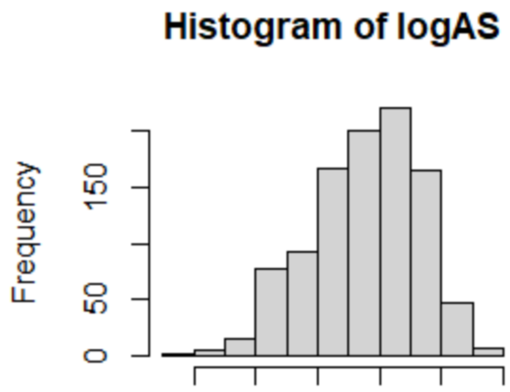
Quiz 10: Catalog_Marketing_Reg.xlsx

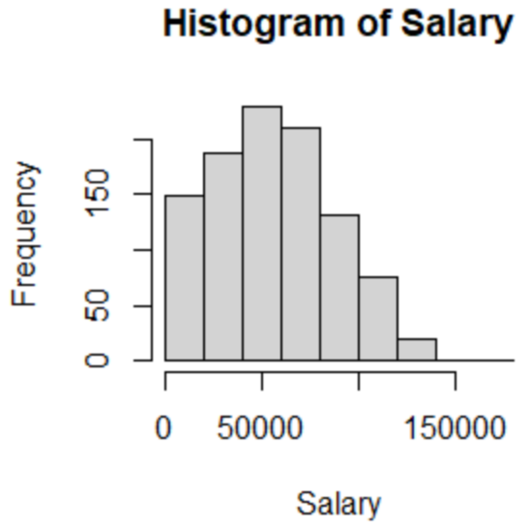
- Build an exponential model: Log(AmountSpent) = Salary + Gender
  - Copy and pate the results
  - Interpret the coefficient of Salary

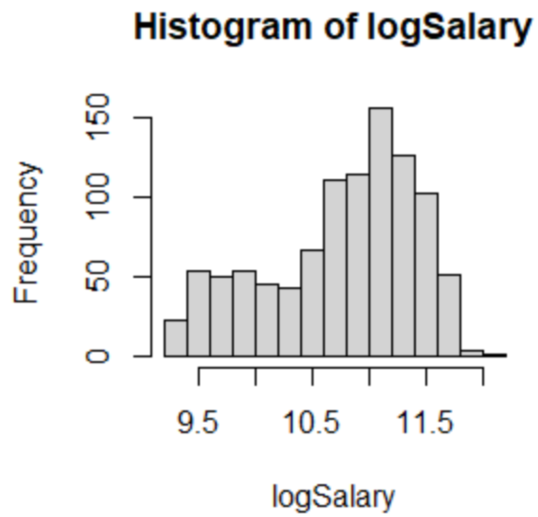
- Build a Log-Log model: Log(AmountSpent) = Log(Salary) + Gender
  - Copy and pate the results
  - Interpret the coefficient of Salary

Practice: Catalog_Marketing_Reg.xlsx





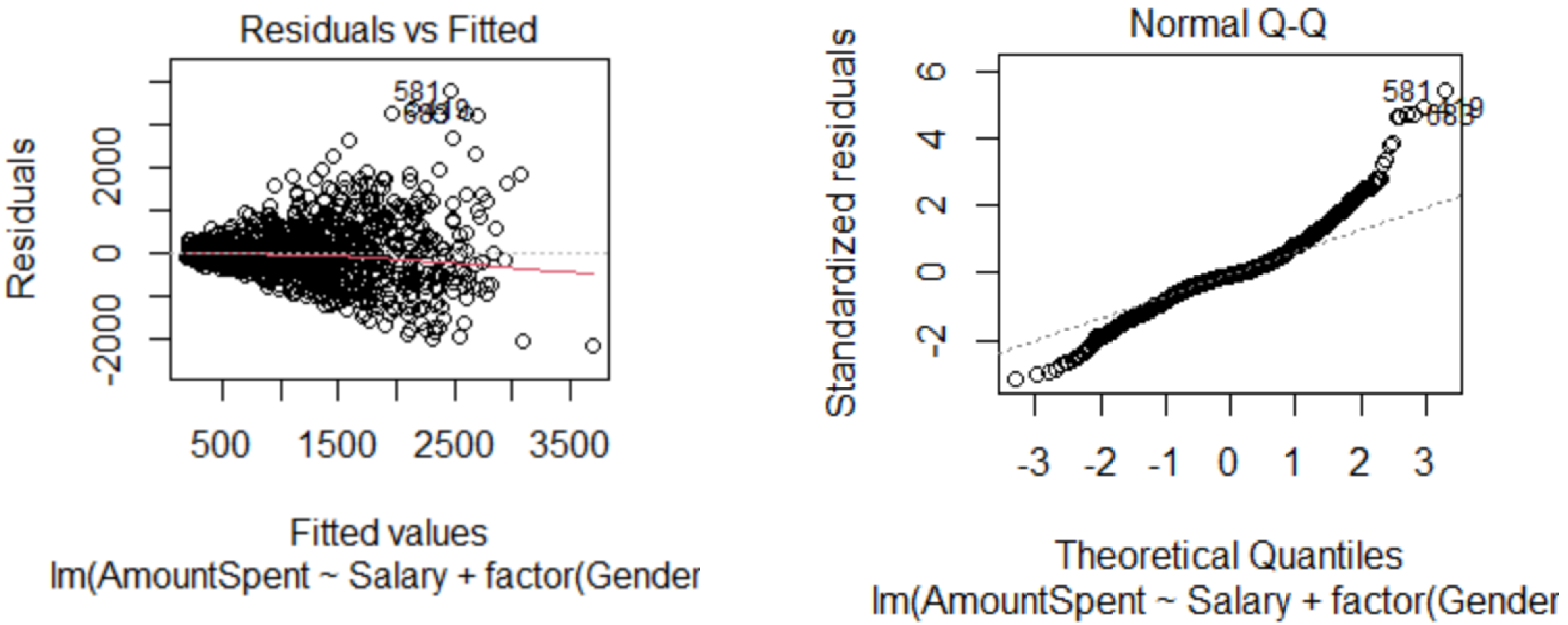




AmountSpent ~ Salary + factor(Gender)

```
Call:
lm(formula = AmountSpent ~ Salary + factor(Gender))

Residuals:
    Min       1Q   Median       3Q      Max 
-2180.9   -323.2   -53.4    282.9   3743.1 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -2.516e+01  4.680e+01  -0.538    0.591    
Salary           2.180e-02  7.357e-04  29.626   <2e-16 ***
factor(Gender)1  3.866e+01  4.503e+01   0.859    0.391    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 687.2 on 997 degrees of freedom
Multiple R-squared:  0.4898,	Adjusted R-squared:  0.4888 
F-statistic: 478.6 on 2 and 997 DF,  p-value: < 2.2e-16
```

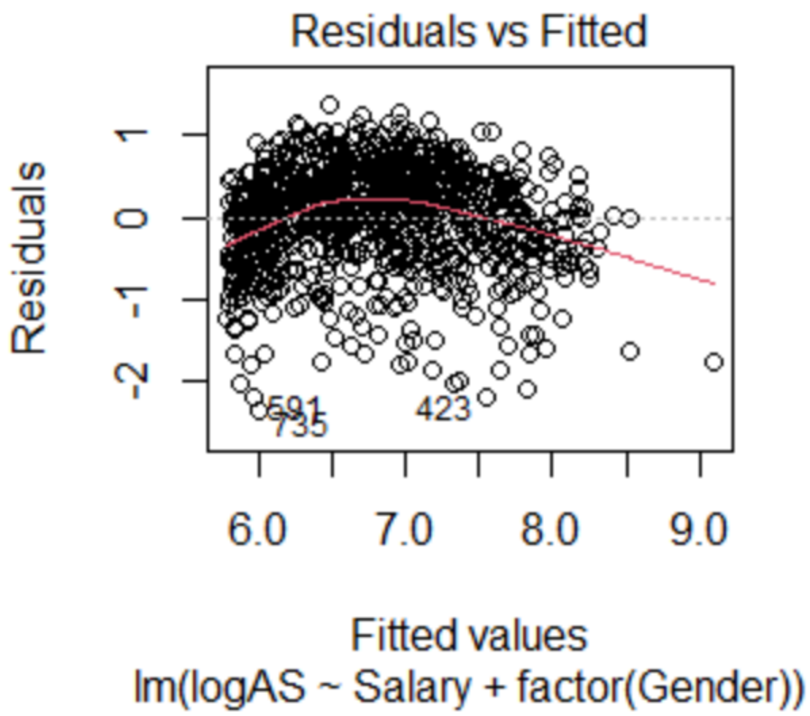
AmountSpent ~ Salary + factor(Gender)

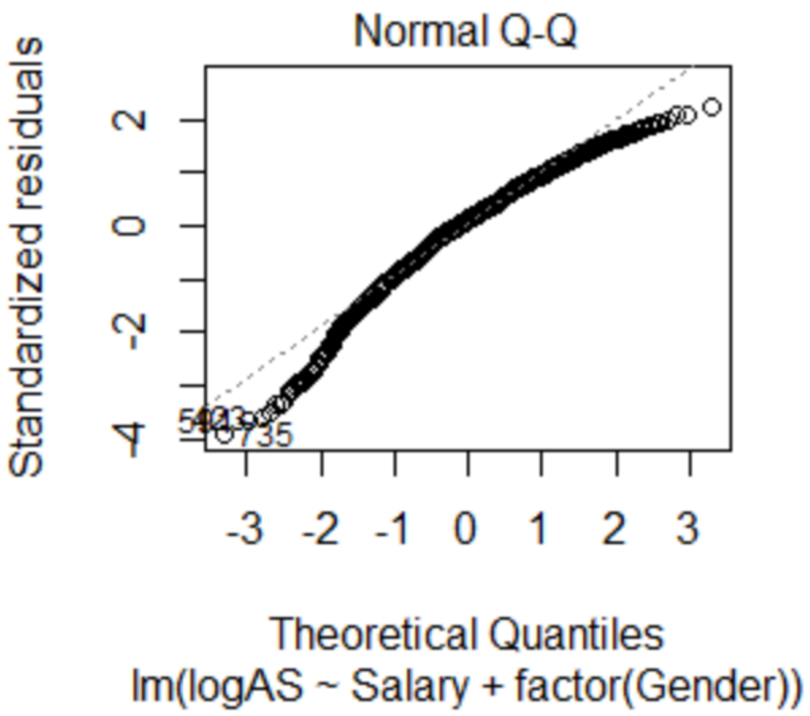


Log(AmountSpent) ~ Salary + factor(Gender)

```
Call:
lm(formula = logAS ~ Salary + factor(Gender))

Residuals:
     Min       1Q   Median       3Q      Max
-2.37490 -0.37032  0.06968  0.42567  1.36035

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.585e+00  4.120e-02 135.570  < 2e-16 ***
Salary          2.009e-05  6.476e-07  31.021  < 2e-16 ***
factor(Gender)1 1.206e-01  3.964e-02   3.042  0.00241 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6049 on 997 degrees of freedom
Multiple R-squared:  0.5236,    Adjusted R-squared:  0.5227
F-statistic: 547.9 on 2 and 997 DF,  p-value: < 2.2e-16
```

Log(AmountSpent) ~ Salary + factor(Gender)

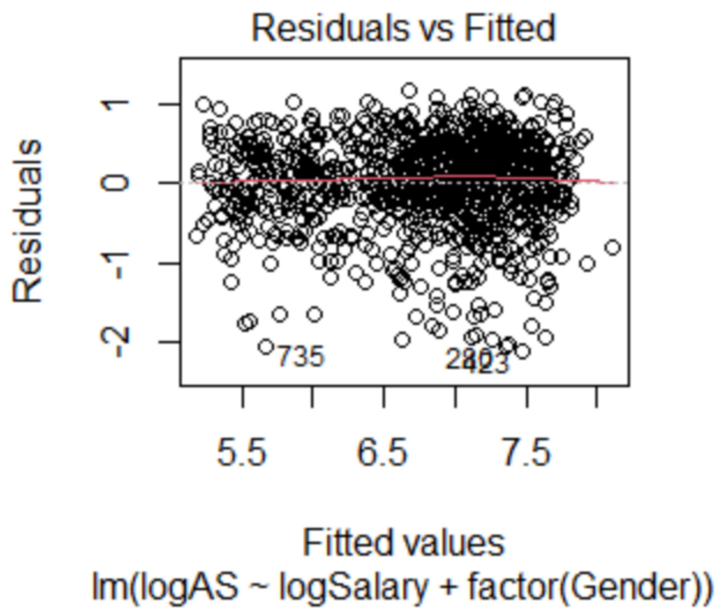


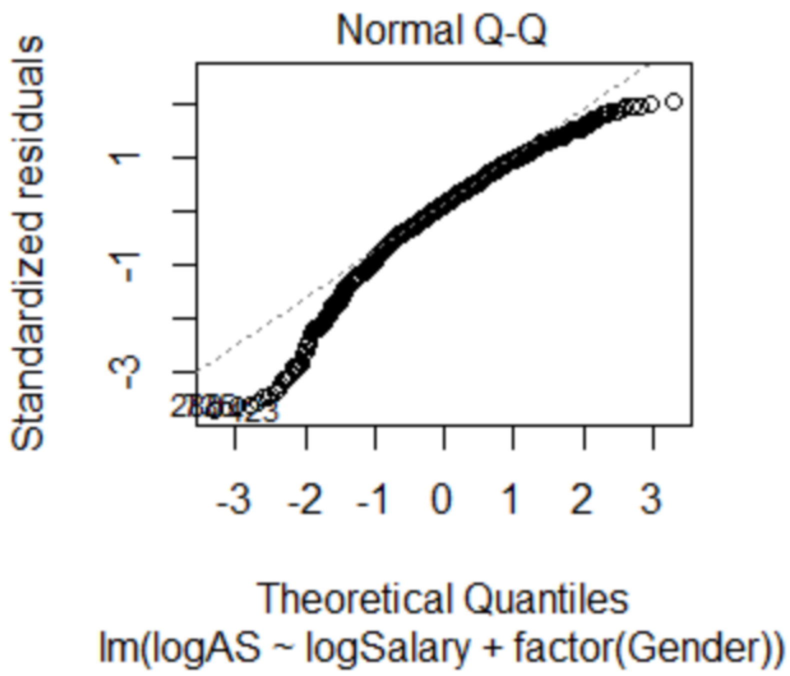


Log(AmountSpent) ~ log(Salary) + factor(Gender)

```
Call:
lm(formula = logAS ~ logSalary + factor(Gender))

Residuals:
     Min       1Q   Median       3Q      Max 
-2.11315 -0.27349  0.06939  0.40125  1.15677 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -4.10001    0.30315 -13.525   <2e-16 ***
logSalary        1.00753    0.02857  35.265   <2e-16 ***
factor(Gender)1  0.08006    0.03722   2.151   0.0317 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5656 on 997 degrees of freedom
Multiple R-squared:  0.5834,	Adjusted R-squared:  0.5826 
F-statistic: 698.2 on 2 and 997 DF,  p-value: < 2.2e-16
```

Log(AmountSpent) ~ log(Salary) + factor(Gender)





Learning Objective

- Model Validation for Explanatory models
- Prediction Models

Explanatory Model and Model Validation

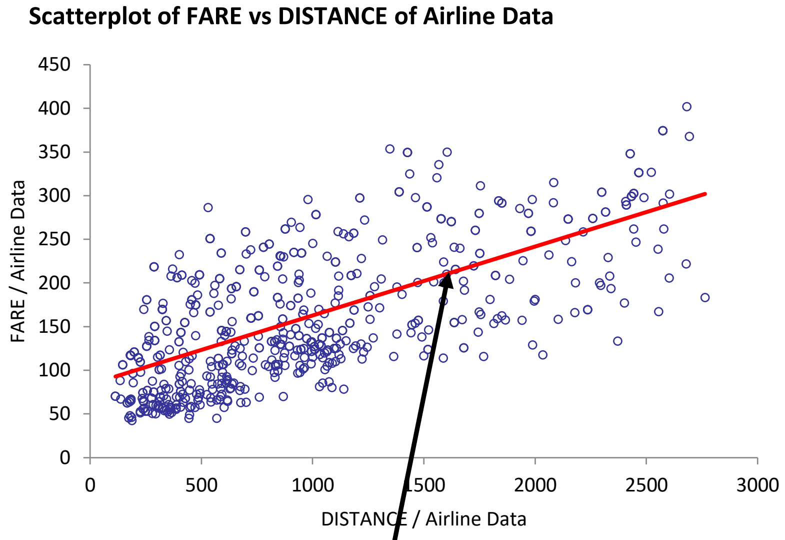
Explanatory regression models

We cared about:

A. Statistical interpretation:

<u>On average</u>, one mile increase in distance is associated with **b** dollars increase in fare.

(all other variables, if any, remain constant)



Fare = a + **b** Distance

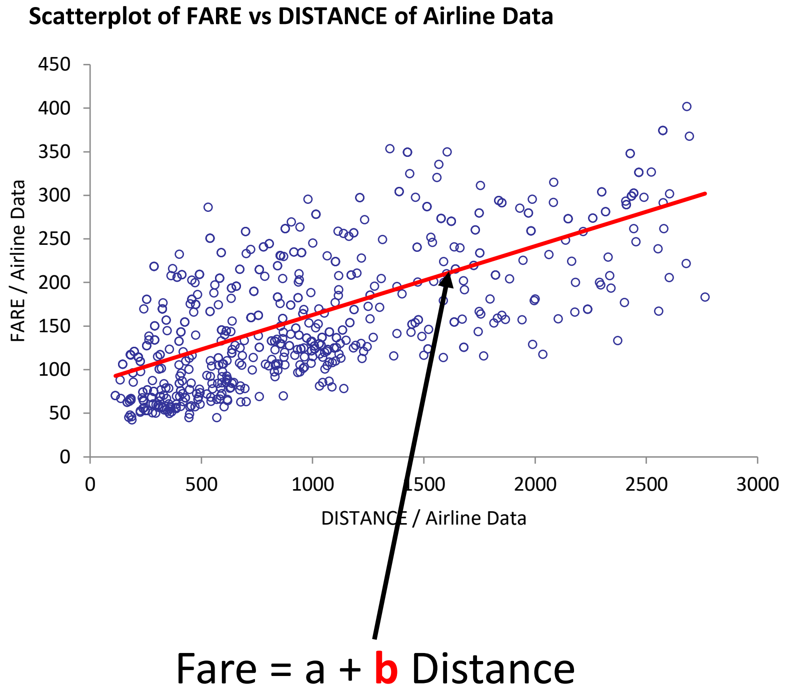
Explanatory regression models

We cared about:

B. Statistical significance:

P-value < α

- You are sure that **b** is not equal to zero
- ‘Distance’ provides meaningful value/information to the model

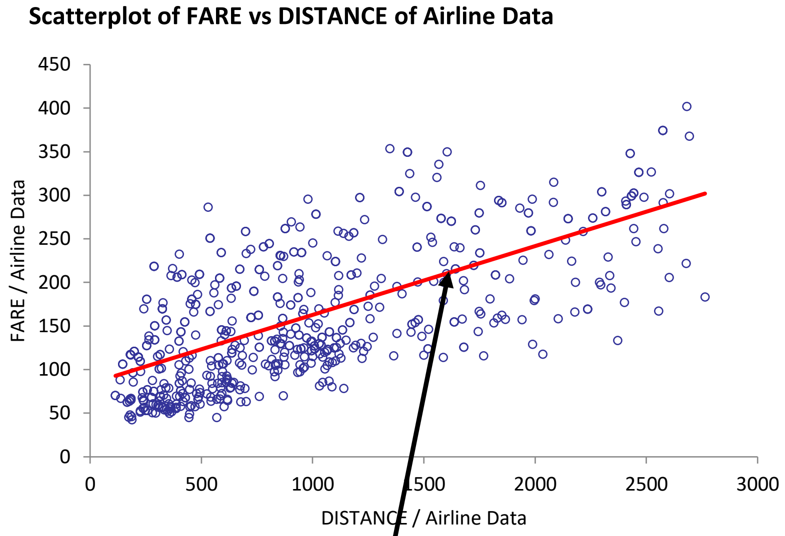


Explanatory regression models

We cared about:

C. Model fit

R^2: the proportion of the variance in the dependent variable that is explained by the independent variable(s)



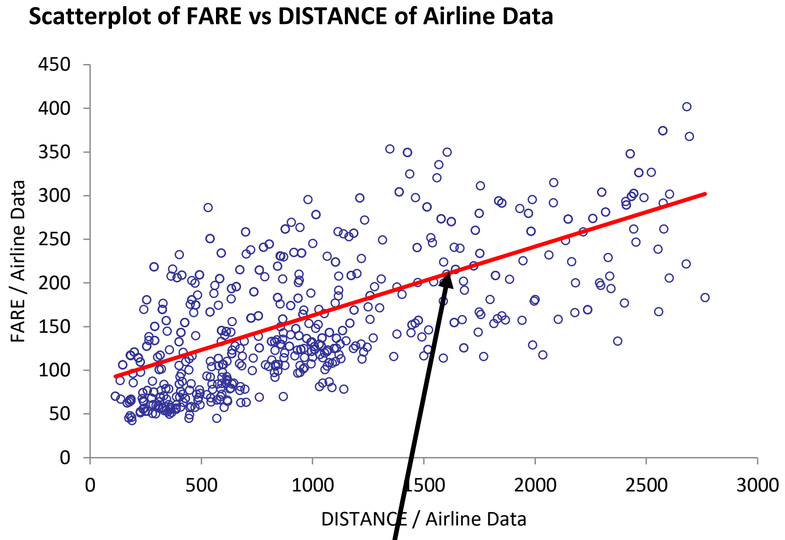
Fare = a + **b** Distance

Explanatory regression models

We cared about:

D. Model validity:

1. Linear relationship between X and Y
2. The variance of the dependent variable is constant (constant error variance)
3. The residual (error term) follows a normal distribution with mean = 0 and residuals are independent
4. Independent variables are independent (no multicollinearity)



Fare = a + **b** Distance

Assumption 1: Linear relationship between Xs and Y

The first assumption is probably the most important:

- For some set of explanatory variables $\mathrm{X} = (X_1, \ldots, X_k)$, there is an exact linear relationship in the population between the *means* of the dependent variable Y and the values of the explanatory variables.
- In other words, there is a **population regression** line that we are estimating from sample data:

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

population intercept population coefficients random error

Note: The mean of Y is

$$\mathrm{E}[Y|X] = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k$$

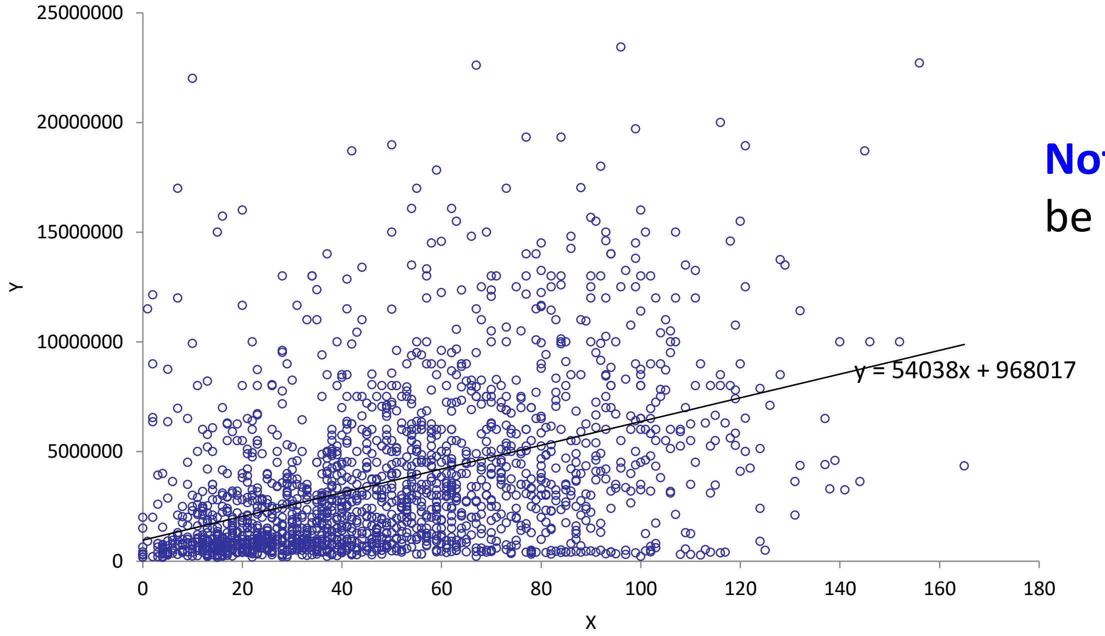
Conditioning on X or fixed X

Assumption 2: Constant Error Variance

- Assumption 2 concerns variation around the population regression line.
  - It states that the variation of the *Y*s about the regression line is the *same*, regardless of the values of the *X*s.
    - The technical term for this property is **homoscedasticity**.
    - A simpler term is **constant error variance**.
  - This assumption is often questionable—the variation in *Y* often increases as *X* increases.
  - **Heteroscedasticity** means that the variability of *Y* values is larger for some *X* values than for others.
    - A simpler term for this is **nonconstant error variance**.

Constant Error Variance

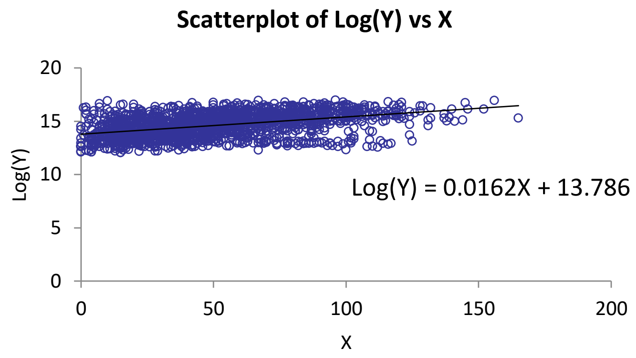
The easiest way to detect nonconstant error variance is through a visual inspection of a scatterplot.



As X increases, the variance of Y increases

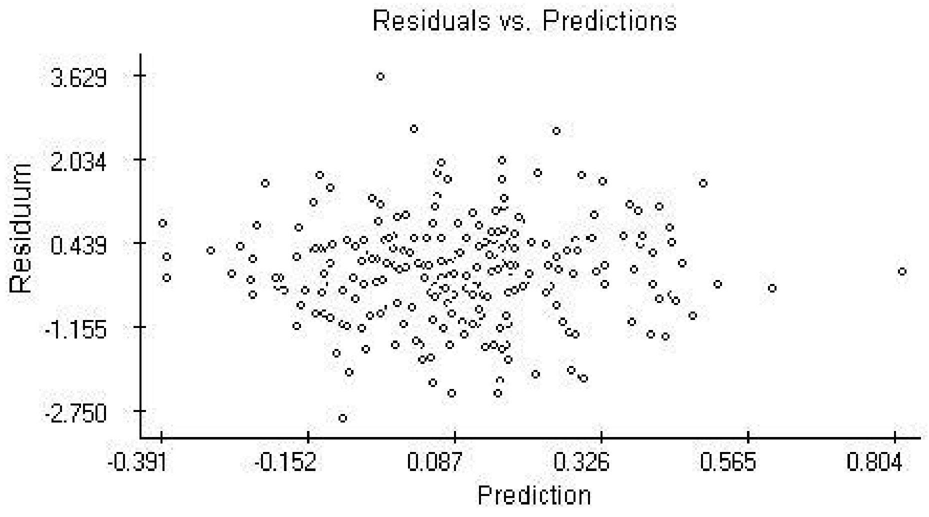
Log transformation can be helpful!

Note: To apply the **log transformation**, all values of data must be **positive**.



Assumption 3: Residuals

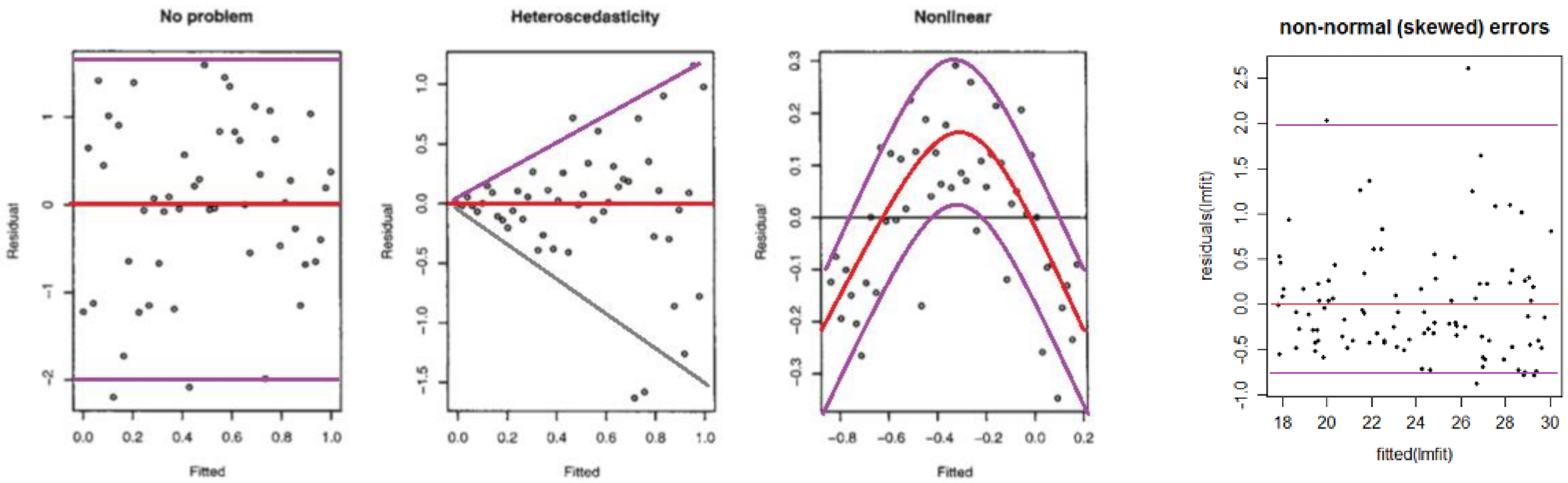
- Assumption 3 is equivalent to stating that the residuals are normally distributed and independent.
- Random residuals, no pattern or trend when plotting residuals



Residuals

- Any deviation is a sign of a problem:
  - Trends
  - Patterns
  - Funnel shape

Log transformation can be helpful!

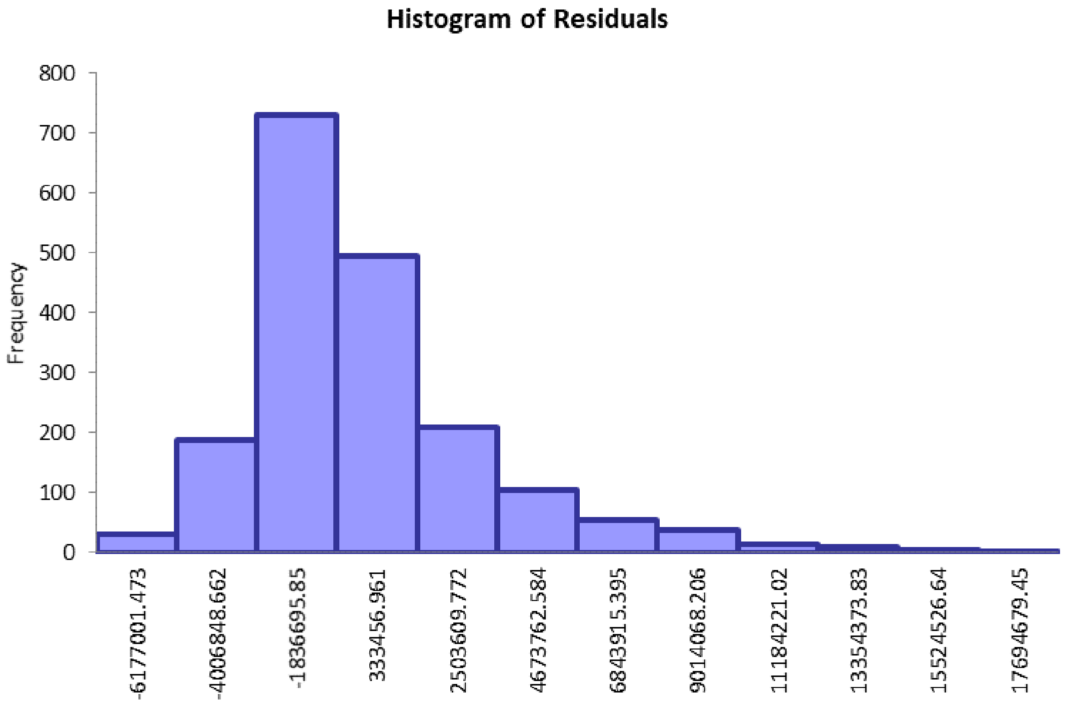


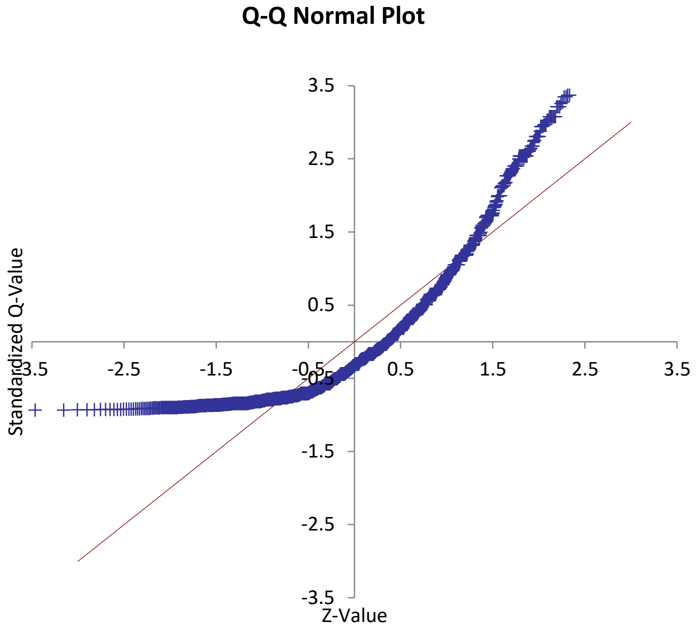
Residual Analysis: Normality

- You can check the normality by forming a histogram or a Q-Q plot of the residuals.
  - The histogram should be approximately symmetric and bell-shaped, and the points of a Q-Q plot should be close to a 45 degree line.
  - If there is an obvious skewness or some other nonnormal property, this indicates a violation of the normality assumption.
- Also, you can conduct a Chi-square goodness of fit test (available in StatTools).

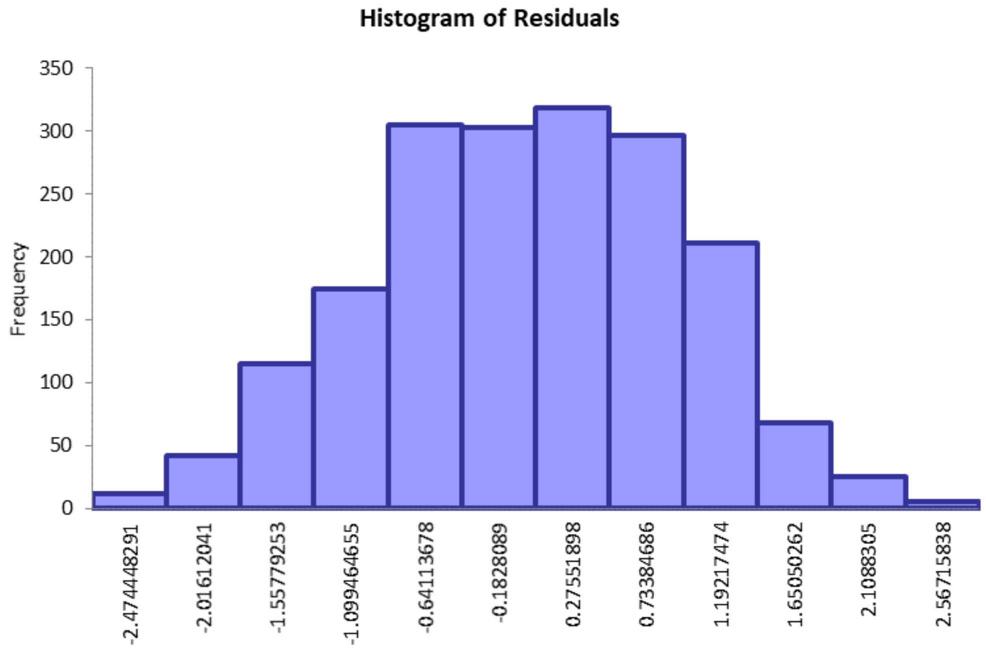
Residual Analysis: Normality Assumption

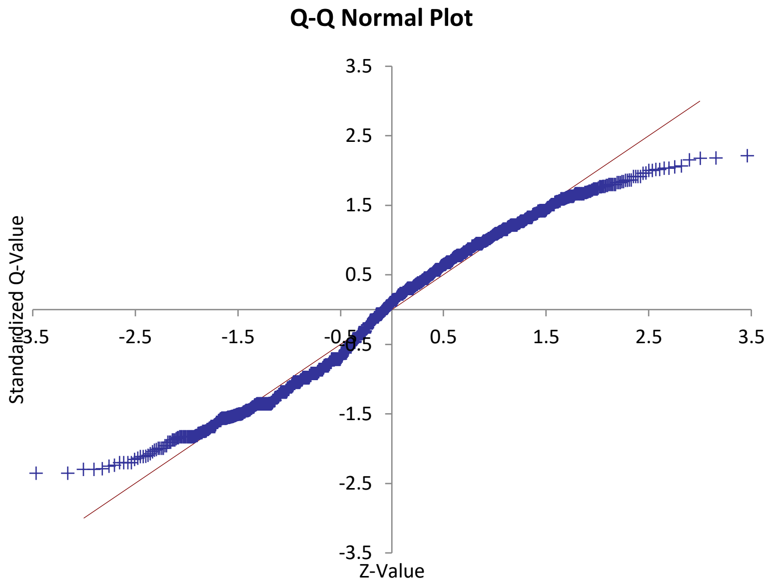
Log transformation can be helpful!





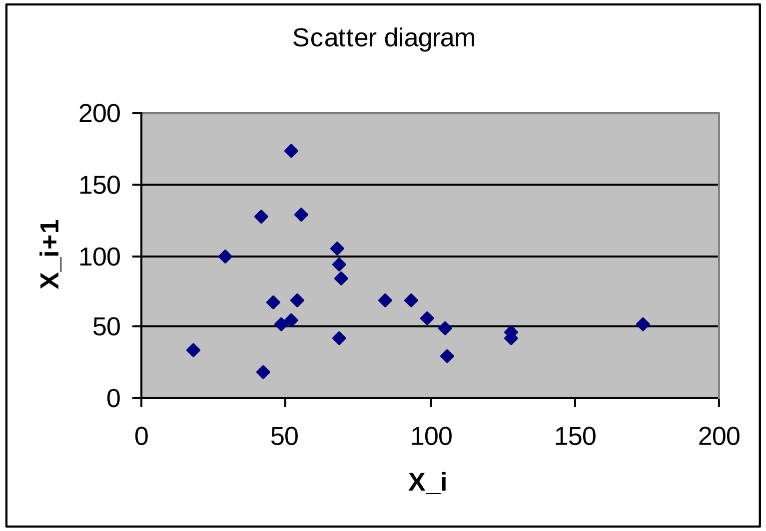
Log Transformation

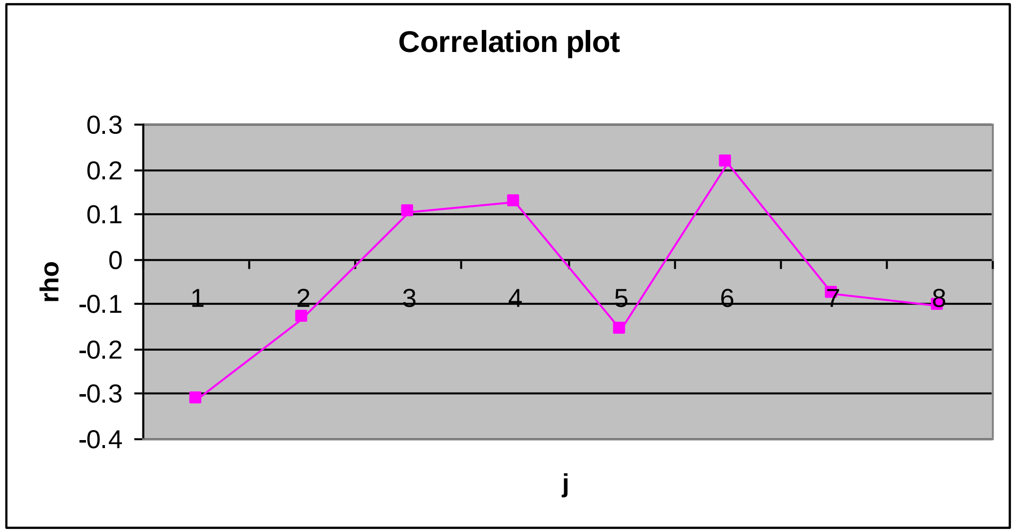




Residual Analysis: Independence

- The independence assumption means that information on some of the errors provides no information on the values of the other errors.
- Use a scatter or a correlation plot





Model Validity

- Assumptions represent idealization of reality, and are never likely to be entirely satisfied for the population in any real study
- If the assumptions are grossly violated, statistical inferences based on these assumptions should be viewed with suspicion

What could go wrong?

1. **Multicollinearity**
2. Omitted Variable Bias (OVB)
3. Outliers
4. Simpson paradox

Multicollinearity

- Linear regression model assumes that all independent variables are independent of each other (Assumption 4)
- Multicollinearity occurs when there is a fairly strong linear relationship among two or more independent variables.
- Multicollinearity can make estimation difficult -it can produce undesirable regression output

Example: Heights vs. Foot Length

We want to explain a person's height by means of foot length.

The dependent variable is Height

The explanatory variables are Right and Left feet lengths.

What happens when we regress Height on both Right and Left ?

Height	Right	Left
77.31	14.49	14.43
67.58	11.96	12.04
70.4	11.21	11.23
64.84	11.74	11.83
77.03	15.06	15.04
79.66	14.24	14.26
72.37	13.19	13.26
73.18	12.89	12.91
77.6	14.76	14.76
71.4	12.40	12.35
72.98	12.63	12.67
69.36	11.81	11.87
74.88	13.63	13.64
67.65	10.96	10.89
78.1	14.73	14.68
72.2	12.83	12.82
67.77	11.78	11.84
73.49	13.78	13.87
69.86	12.86	12.86
77.05	14.48	14.47

Regression Output

```
Call:
lm(formula = Height ~ Right + Left)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.1394 -1.9432  0.1179  2.3544  7.5071 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  29.4430     1.2471  23.609  < 2e-16 ***
Right         3.3535     1.1344   2.956  0.00391 ** 
Left          0.0482     1.1389   0.042  0.96633    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.081 on 97 degrees of freedom
Multiple R-squared:  0.9319,	Adjusted R-squared:  0.9305 
F-statistic: 663.7 on 2 and 97 DF,  p-value: < 2.2e-16
```

- How good is the fit of this regression based on R^2?

Regression Output

- The coefficients of Right and Left are not at all what we might expect.
- The t-value of Left is quite small and the corresponding p-value is quite large – not significant at 5% significance level
- Judging by this, we might conclude that Height and Left are not related.
- The *t*-value and *p*-value for the coefficient of Right are now 2.96 and 0.004

Correlation between Variables

- Very high correlation of 0.997 between Left and Right variables

```
> cor(Height_Foot)
            Left     Right    Height
Left   1.0000000 0.9966079 0.9621697
Right  0.9966079 1.0000000 0.9653519
Height 0.9621697 0.9653519 1.0000000
```

Regression with only Right Foot

```
Call:
lm(formula = Height ~ Right)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.1179 -1.9432  0.1212  2.3427  7.5084 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 29.44652    1.23805   23.79   <2e-16 ***
Right        3.40131    0.09288   36.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.065 on 98 degrees of freedom
Multiple R-squared:  0.9319,    Adjusted R-squared:  0.9312 
F-statistic:  1341 on 1 and 98 DF,  p-value: < 2.2e-16
```

- The R^2 and SE values are roughly same as before
- But, the *t*-value and *p*-value for the coefficient of Right are now 36.62 and <0.0001 - very significant.

Regression with only Left Foot

```
Call:
lm(formula = Height ~ Left)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.7242 -2.1661  0.0314  2.3166  7.5207 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 29.49388    1.29530   22.77   <2e-16 ***
Left         3.40357    0.09735   34.96   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.201 on 98 degrees of freedom
Multiple R-squared:  0.9258,	Adjusted R-squared:  0.925 
F-statistic:  1222 on 1 and 98 DF,  p-value: < 2.2e-16
```

- The R^2 and SE values are again same as before
- But, the *t*-value and *p*-value for the coefficient of Left are 34.96 and <0.0001 - again very significant

Multicollinearity: Conclusion

- The message is that when two variables are very highly correlated, only one of them should be included in the regression equation
- Especially true for estimation/explanation modeling
- Multicollinearity only effects model interpretation, it is **not a significant problem when developing prediction models**

- How to check multicollinearity?
  - Correlation
  - Variance Inflation Factor (VIF):

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

where R^2_j is the coefficient of determination of variable X_j with all other X variables.

The Omitted Variable Bias (OVB)

The good news about multicollinearity is that we can test dependency between the independent variables before we run a regression model.

BUT, what if one relevant independent variable is entirely unobserved?

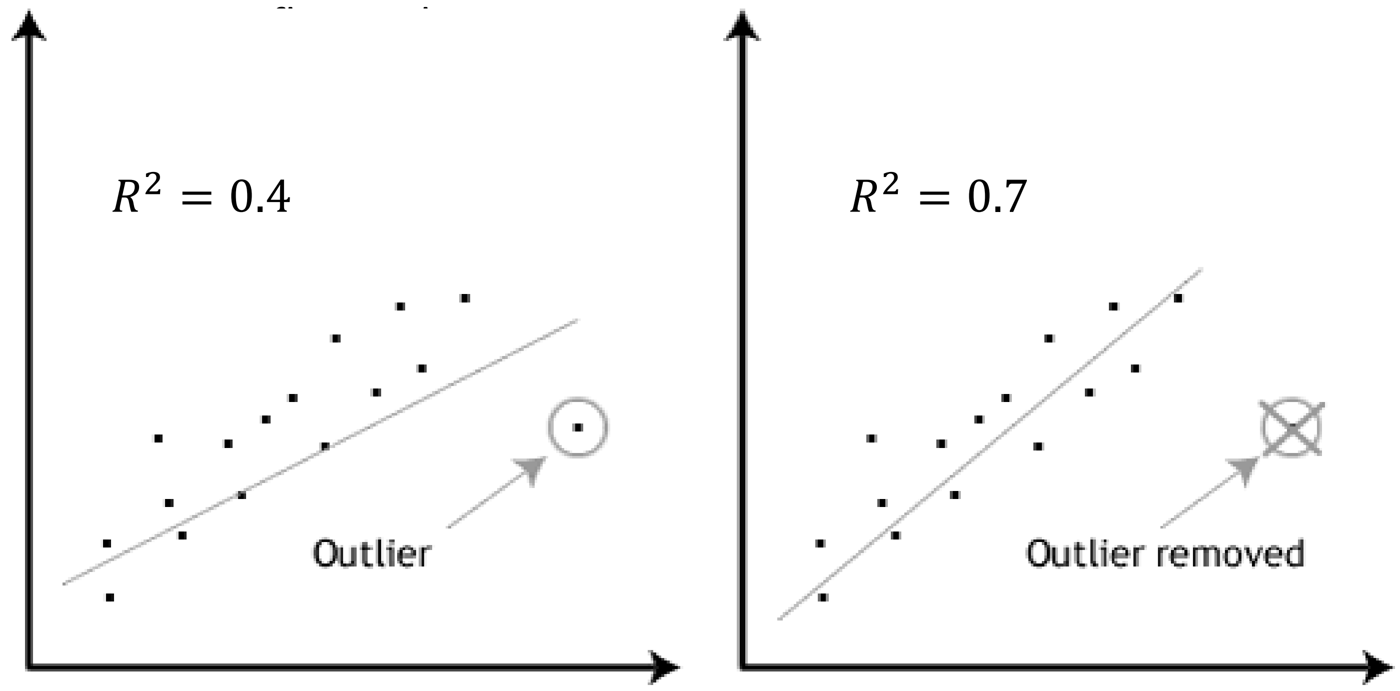
This can cause the **Omitted Variable Bias (OVB)**: a bias that appears in the coefficients of the regression analysis, when we omit an independent variable that is related with both the dependent variable and one or more of the included independent variables.

NOTE! OVB can even affect the sign of the regression coefficient!

Outliers

- Outliers are data points that are substantially different from the rest of the data on the dependent variable
- Outliers could arise due to randomness, data entry error, or some other unknown reason not explained by the regression
- Outliers can distort the regression output
- Coefficient estimate errors will go up
- R^2 will go down
- It is therefore important that we identify and (if warranted) remove outlier data points from the regression

Outlier Examples



Source: https://statistics.laerd.com/stata-tutorials/linear-regression-using-stata.php

Addressing Outliers

Simply finding an outlier does not mean you have to take action, it depends entirely on the situation

- If an outlier is clearly not a member of the population of interest, then it may be best to remove it
- If it is not clear whether outliers are members of the relevant population, run the regression analysis with them and again without them
    - If the results are similar, then it is probably best to report the results with the outliers included
    - Otherwise, you can report both sets of results with an explanation of the outliers
- Outliers can lead to interesting business insights: fraudulent transactions, criminal activity, security breaches, and disease outbreaks

The Simpson paradox

Example: Demand for Umbrellas

A marketing consultant wants to check the impact price of umbrellas on their demand. For that, he collected purchase and price data from several stores across a certain US city.

For simplicity, assume that the number of customers visiting each store is the statistically equal.

The Simpson paradox

- How do we expect the demand Vs. price graph to appear?
- What model will we use to test the relationship between price and demand?

Store	Price	Demand
A	8	1300
B	4	400
C	9	1200
D	3	500
E	7	100
F	10	1100
G	11	1000
H	6	200
I	12	900
J	5	300

The Simpson paradox

- How do we expect the demand Vs. price graph to appear?
- What model will we use to test the relationship between price and demand?

Regression line:

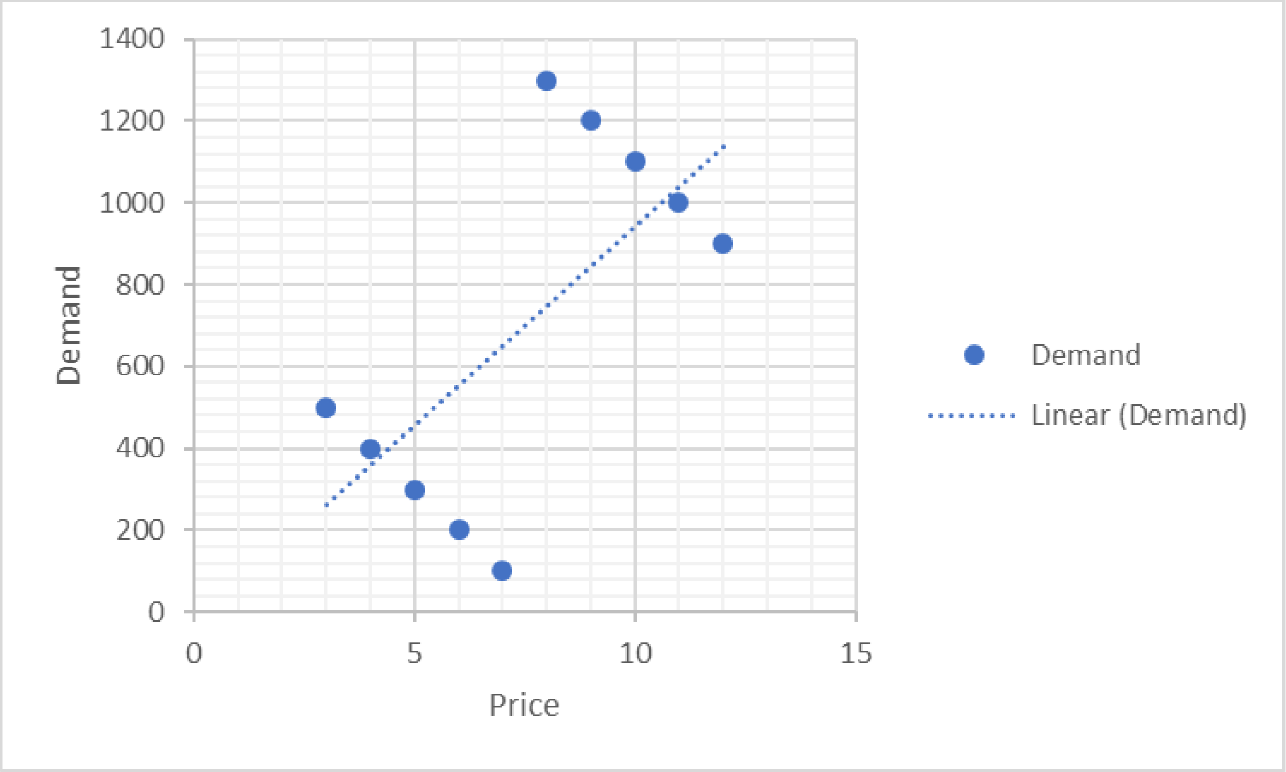
Demand = -27 + 97 Price

Store	Price	Demand
A	8	1300
B	4	400
C	9	1200
D	3	500
E	7	100
F	10	1100
G	11	1000
H	6	200
I	12	900
J	5	300

The Simpson paradox

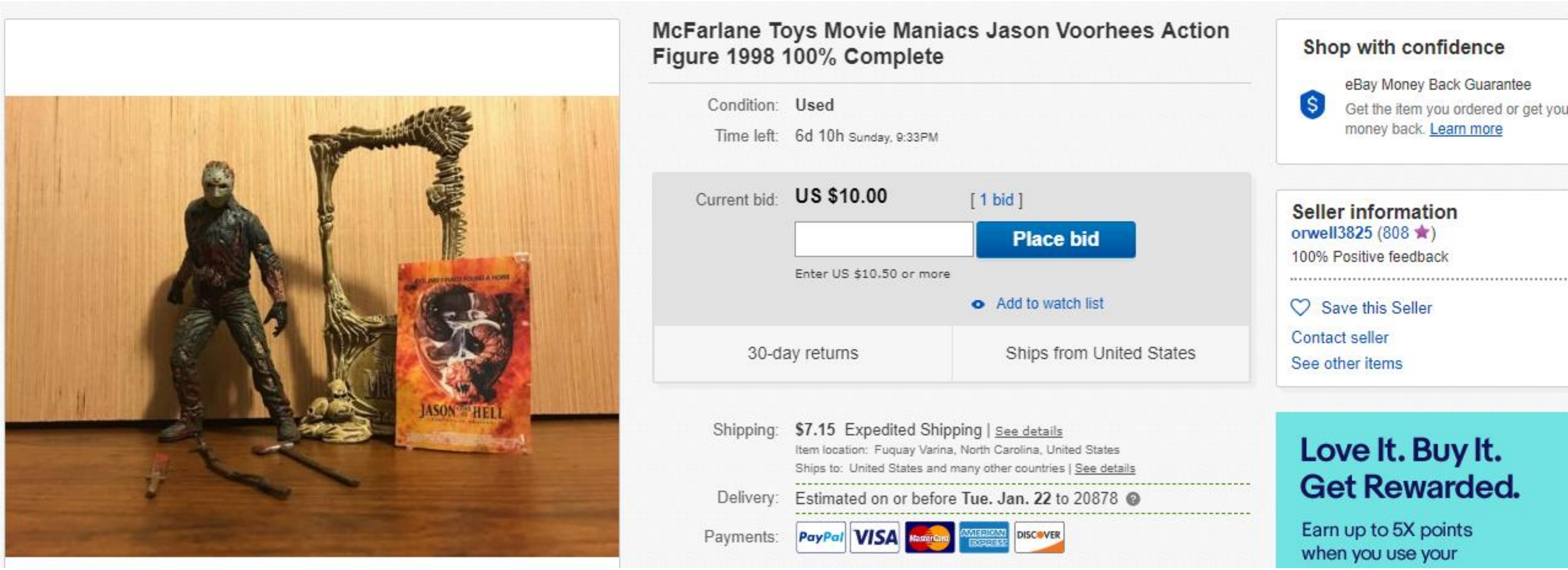
The Simpson paradox is a paradox in which the general trend is the opposite of the trend in sub-populations.

Two different groups or population
-> omitted categorical variables
Ex: Season (or region) - rainy season, dry season



The world of prediction models

Example: eBay Auctions



Example: eBay Auctions

eBay is one of the largest consumer-2-consumer auction website. It enables a global community of sellers and buyers to easily interact and trade.

Many studies use the publicly available historical bid data to learn the bidding behavior and auction outcomes.

Bid data include information about

1) Sellers, such as registration date and feedback score,
2) Buyers, such as shipping address, age and gender, and
3) Auction, such as start date, start price, number of bids and bidders, and close price

Example: eBay Auctions

Two *example* goals:

1) Explanatory task: **Explain** what affects the close price of auction.
2) Predictive task: **Predict** the close price of future auctions.

Example: eBay Auctions

Variable selection: which variables can and cannot by used to address each task? Why?

Bid data include information about
1) **Sellers, such as registration date and feedback score,**
2) **Buyers, such as shipping address, age and gender, and**
3) **Auction, such as start date, start price, number of bids and bidders, and close price**

1) <u>For Explanatory task</u>
2) <u>For Predictive task</u>

Example: eBay Auctions

Variable selection: which variables can and cannot by used to address each task? Why?

Bid data include information about

1) **Sellers, such as registration date and feedback score,**
2) **Buyers, such as shipping address, age and gender, and**
3) **Auction, such as start date, start price, number of bids and bidders, and close price**

1) For Explanatory task: may use everything! Select variables via stepwise selection
2) For Predictive task: Independent variables must proceed dependent and should be available at the time of prediction. Thus, # of bids or bidders are not included

Example: eBay Auctions

Evaluation process:

What does R^2 measure? Is it useful for evaluating prediction quality?

In other words, does high R^2 imply high prediction accuracy?

The problem of overfitting and the concept of data partition

Key question in predictive analytics:

How well will our prediction model perform when we apply it to new data?

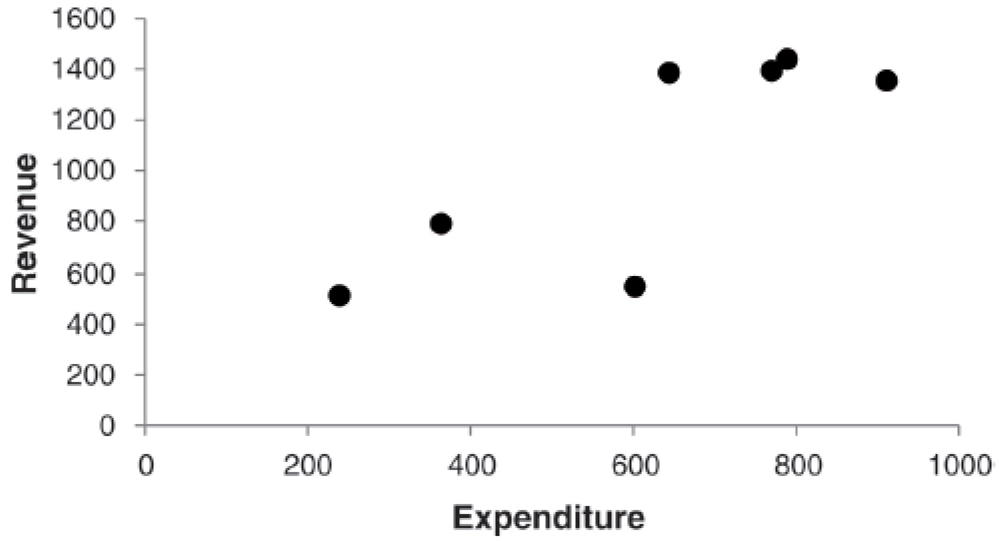
→ *make sure that our model generalizes beyond the dataset that we have at hand*

Overfitting

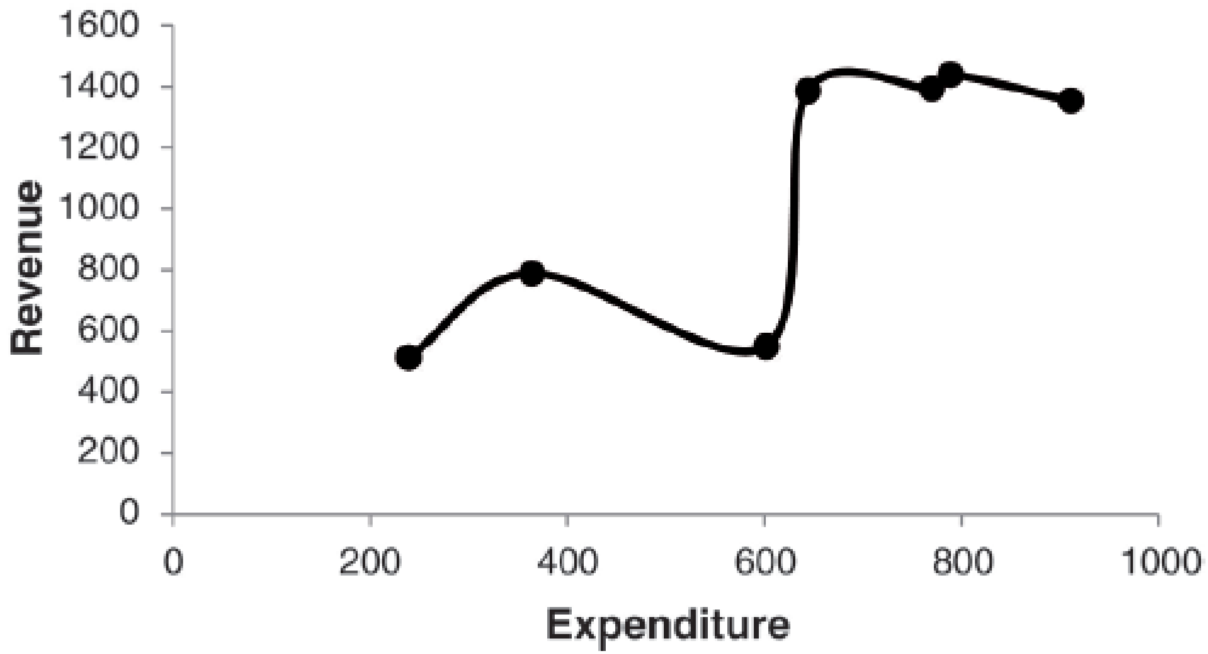
The more variables we include in a model, the greater the risk of **overfitting** the particular data used for modeling.

Example: Sales as a function of advertisement expenditures

Advertising	Sales
239	514
364	789
602	550
644	1386
770	1394
789	1440
911	1354



Overfitting

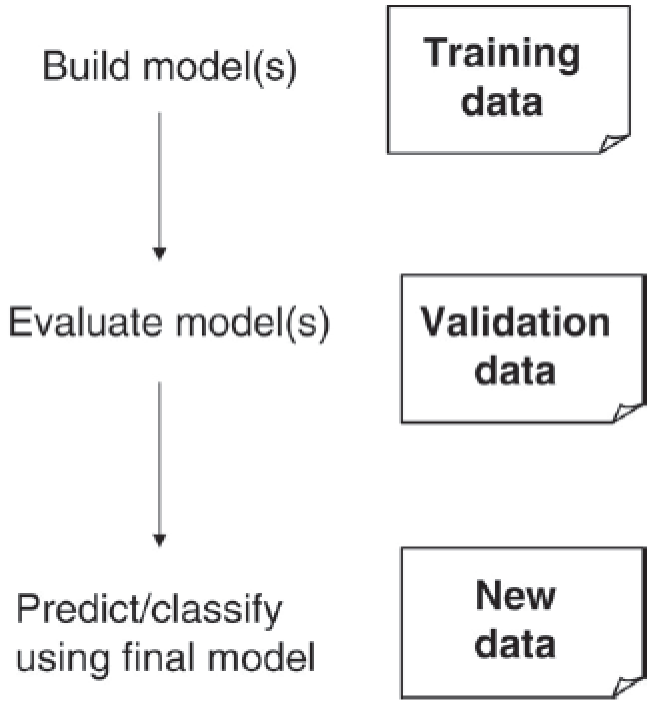


R^2 = ?

No error (residual) model: fit a complex function to Sales ~ Expenditures

Will it be accurate for future sales?

Building a prediction model



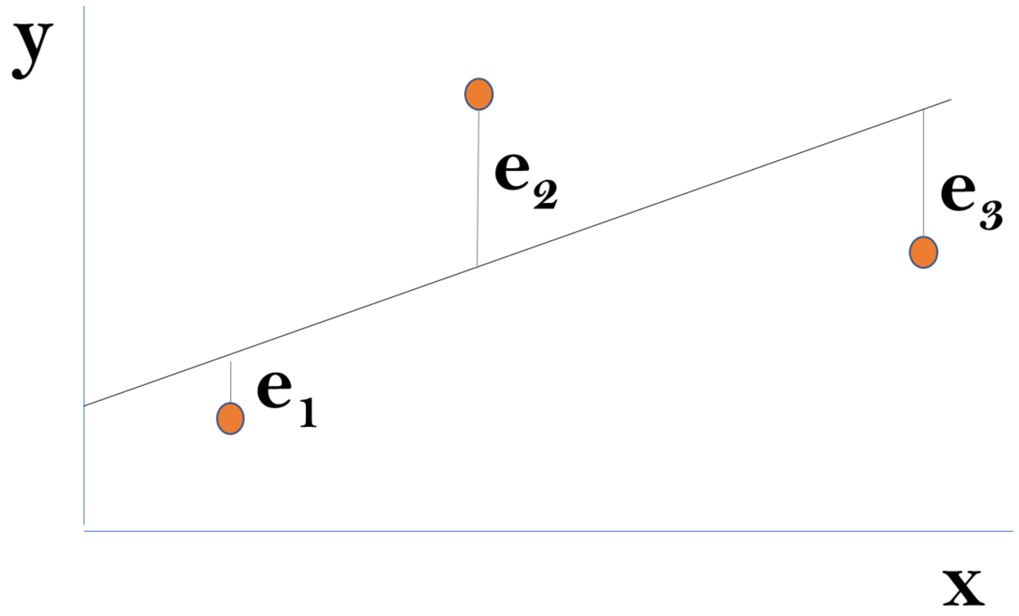
Build model on training.

Evaluate on validation (holdout, test) data.

Deploy on real data.

Evaluation methods

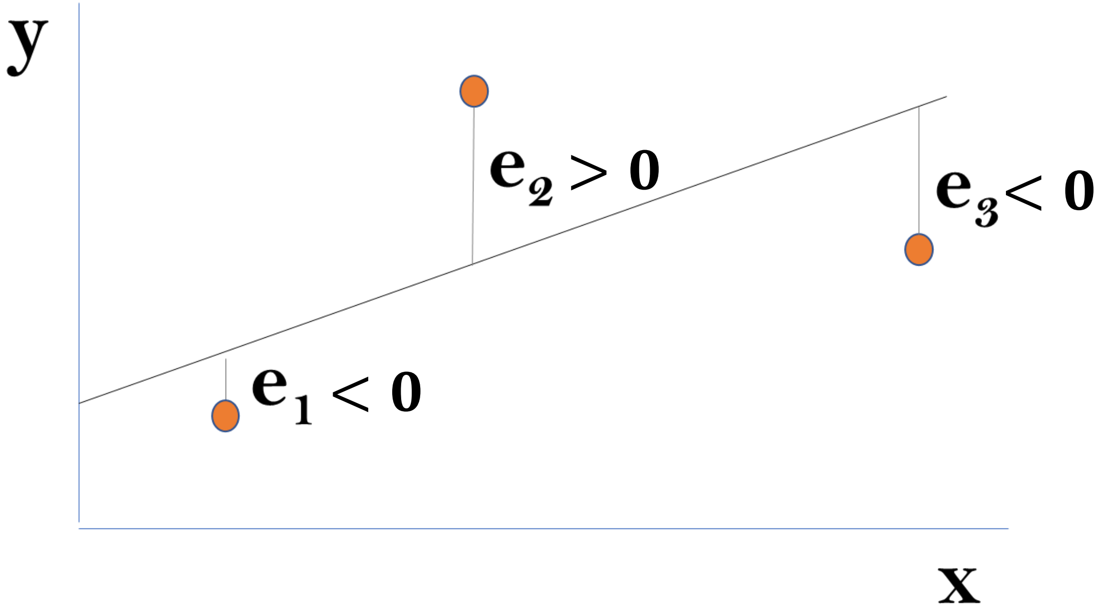
Goal: See how close predictions are to real validation data



$error$: $e_i = Y_i - \hat{Y}_i$ (observed – fitted)

Evaluation methods: ME

- **Mean Forecast Error (MFE)** = $\frac{1}{n}\sum_{i=1}^{n}(e_i)$. MFE shows whether the forecast consistently under- or overestimates demand



Evaluation methods

Mean Absolute Error (MAE) = $\frac{1}{n}\sum_{i=1}^{n}|e_i|$. Gives the magnitude of the average absolute error. On average how much did I miss by?

Root Mean Squared Error (RMSE) = $\sqrt{\frac{1}{n}\sum_{i=1}^{n} {e_i}^2}$. Gives standard error of estimate in linear regression, computed on validation set.

Mean Absolute Percentage Error (MAPE) = $100 \times \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right|$. Gives a percentage score of how predictions deviate (on average) from the actual values.

Example: eBay Auctions

Model selection:

Which model is preferred for a prediction task?

	Model 1	Model 2	Model 3	Model 4
Variables inserted	ALL	Seller data, start price, start date, interaction terms	Seller data, start price, start date	Seller data, start price
Significance	< 0.05	< 0.1	< 0.15	< 0.15
R^2 (Training data)	91%	72%	55%	55%
RMSE (Validation data)	104.31	540.5	210.23	290.41
MAPE (Validation data)	2.1	24.9	5.3	3.8

Example: eBay Auctions

Model selection:

We care about: high prediction accuracy, as measured on **new data** (validation set); that dictates the selection of variables!

We do not care about: statistical interpretation and statistical significance

We less care about: model validity, model fit

Prediction emphasis

- Interpretation is not the goal
- Statistical significance of predictors is not necessarily criterion for retaining predictors
- Residual analysis is not important
- What matters is predictive accuracy
- BUT: any domain knowledge should be included in choice of predictors!