INFO 6105
Data Science Engineering Methods and Tools

Lecture 3
Sales of Child Car Seats

Ebrahim Nasrabadi
nasrabadi@northeastern.edu

College of Engineering
Northeastern University

Fall 2019

Uncertainty is everywhere

Uncertainty in input data is a common challenging problem and we need to quantify uncertainty to make better business decisions.

Uncertainty is everywhere

Uncertainty in input data is a common challenging problem and we need to quantify uncertainty to make better business decisions.

Inventory Planning

- Demand is uncertain
- The optimal order quantity depends on the distribution of demand

Uncertainty is everywhere

Uncertainty in input data is a common challenging problem and we need to quantify uncertainty to make better business decisions.

Inventory Planning

- Demand is uncertain
- The optimal order quantity depends on the distribution of demand

Key question How to estimate accurately the distribution of demand?

Uncertainty is everywhere

Uncertainty in input data is a common challenging problem and we need to quantify uncertainty to make better business decisions.

Inventory Planning

- Demand is uncertain
- The optimal order quantity depends on the distribution of demand

Key question How to estimate accurately the distribution of demand? In reality, demand for a product is influenced by various factors, such as

- price
- product quality
- price of related products
- time of the year
- consumer's income,
- growth of population
- climatic conditions
- ...

To predict demand, we analyze the historical data to understand the relationship between sales and factors that influence sales.

Sales of Child Car Seats

Consider a data set containing sales of child car seats at 400 different stores:

- Sales Unit sales (in thousands) at each location
- Price Price company charges for car seats at each site
- CompPrice Price charged by competitor at each location
- Income Community income level (in thousands of dollars)
- Advertising Local advertising budget for company at each location (in thousands of dollars)
- Population Population size in region (in thousands)
- ShelveLoc A factor with levels Bad, Medium, Good indicating the quality of the shelving location for the car seats at each site
- Age Average age of the local population
- Education Education level at each location
- Urban A factor with levels No and Yes to indicate whether the store is in an urban or rural location
- US A factor with levels No and Yes to indicate whether the store is in the US or not

Sales of Child Car Seats

Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urban	US
9.50	138	73	11	276	120	Bad	42	17	Yes	Yes
11.22	111	48	16	260	83	Good	65	10	Yes	Yes
10.06	113	35	10	269	80	Medium	59	12	Yes	Yes
7.40	117	100	4	466	97	Medium	55	14	Yes	Yes
4.15	141	64	3	340	128	Bad	38	13	Yes	No
10.81	124	113	13	501	72	Bad	78	16	No	Yes
6.63	115	105	0	45	108	Medium	71	15	Yes	No
11.85	136	81	15	425	120	Good	67	10	Yes	Yes
6.54	132	110	0	108	124	Medium	76	10	No	No
4.69	132	113	0	131	124	Medium	76	17	No	Yes
9.01	121	78	9	150	100	Bad	26	10	No	Yes
11.96	117	94	4	503	94	Good	50	13	Yes	Yes
3.98	122	35	2	393	136	Medium	62	18	Yes	No
10.96	115	28	11	29	86	Good	53	18	Yes	Yes
11.17	107	117	11	148	118	Good	52	18	Yes	Yes

Sales of Child Car Seats

- We are interested to predict car seat sales on the basis of the other variables.
- We refer to the
  - Sales variable as *target* (also called *response* or *output*) variable
  - Price, ..., US variables as *predictors* (also called *features* or *inputs*).

Sales of Child Car Seats

- We are interested to predict car seat sales on the basis of the other variables.
- We refer to the
  - Sales variable as *target* (also called *response* or *output*) variable
  - Price, ..., US variables as *predictors* (also called *features* or *inputs*).
- Goal:
  - understand the relationship between sales and predictors
  - make predictions for sales

Questions

1. Is there a relationship between predictors and sales?

Questions

1. Is there a relationship between predictors and sales?
2. How strong is the relationship between predictors and sales?

Questions

1. Is there a relationship between predictors and sales?
2. How strong is the relationship between predictors and sales?
3. Which predictors contribute to sales? What predictors matter most?

Questions

1. Is there a relationship between predictors and sales?
2. How strong is the relationship between predictors and sales?
3. Which predictors contribute to sales? What predictors matter most?
4. How accurately can we estimate the effect of each predictor on sales?

Questions

1. Is there a relationship between predictors and sales?
2. How strong is the relationship between predictors and sales?
3. Which predictors contribute to sales? What predictors matter most?
4. How accurately can we estimate the effect of each predictor on sales? What is the degree to which price affects the sales (that is, price sensitivity)? How accurately can we predict this amount?
5. How accurately can we predict future sales?

Questions

1. Is there a relationship between predictors and sales?
2. How strong is the relationship between predictors and sales?
3. Which predictors contribute to sales? What predictors matter most?
4. How accurately can we estimate the effect of each predictor on sales? What is the degree to which price affects the sales (that is, price sensitivity)? How accurately can we predict this amount?
5. How accurately can we predict future sales?

Questions

1. Is there a relationship between predictors and sales?
2. How strong is the relationship between predictors and sales?
3. Which predictors contribute to sales? What predictors matter most?
4. How accurately can we estimate the effect of each predictor on sales? What is the degree to which price affects the sales (that is, price sensitivity)? How accurately can we predict this amount?
5. How accurately can we predict future sales?

Linear regression can be used to answer each of these questions.

Sales of Child Car Seats

Consider a data set containing sales of child car seats at 400 different stores:

- Sales Unit sales (in thousands) at each location
- Price Price company charges for car seats at each site
- CompPrice Price charged by competitor at each location
- Income Community income level (in thousands of dollars)
- Advertising Local advertising budget for company at each location (in thousands of dollars)
- Population Population size in region (in thousands)
- ShelveLoc A factor with levels Bad, Medium, Good indicating the quality of the shelving location for the car seats at each site
- Age Average age of the local population
- Education Education level at each location
- Urban A factor with levels No and Yes to indicate whether the store is in an urban or rural location
- US A factor with levels No and Yes to indicate whether the store is in the US or not

Sales of Child Car Seats

- We are interested to predict car seat sales on the basis of the other variables.
- We refer to the
  - Sales variable as *target* (also called *response* or *output*) variable
  - Price, ..., US variables as *predictors* (also called *features* or *inputs*).

Sales of Child Car Seats

- We are interested to predict car seat sales on the basis of the other variables.
- We refer to the
  - Sales variable as *target* (also called *response* or *output*) variable
  - Price, ..., US variables as *predictors* (also called *features* or *inputs*).
- Goal:
  - understand the relationship between sales and predictors
  - make predictions for sales

Questions

1. Is there a relationship between predictors and sales?

Questions

1. Is there a relationship between predictors and sales?
2. How strong is the relationship between predictors and sales?

Questions

1. Is there a relationship between predictors and sales?
2. How strong is the relationship between predictors and sales?
3. Which predictors contribute to sales? What predictors matter most?

Questions

1. Is there a relationship between predictors and sales?
2. How strong is the relationship between predictors and sales?
3. Which predictors contribute to sales? What predictors matter most?
4. How accurately can we estimate the effect of each predictor on sales?

Questions

1. Is there a relationship between predictors and sales?
2. How strong is the relationship between predictors and sales?
3. Which predictors contribute to sales? What predictors matter most?
4. How accurately can we estimate the effect of each predictor on sales? What is the degree to which price affects the sales (that is, price sensitivity)? How accurately can we predict this amount?
5. How accurately can we predict future sales?

Questions

1. Is there a relationship between predictors and sales?
2. How strong is the relationship between predictors and sales?
3. Which predictors contribute to sales? What predictors matter most?
4. How accurately can we estimate the effect of each predictor on sales? What is the degree to which price affects the sales (that is, price sensitivity)? How accurately can we predict this amount?
5. How accurately can we predict future sales?

Questions

1. Is there a relationship between predictors and sales?
2. How strong is the relationship between predictors and sales?
3. Which predictors contribute to sales? What predictors matter most?
4. How accurately can we estimate the effect of each predictor on sales? What is the degree to which price affects the sales (that is, price sensitivity)? How accurately can we predict this amount?
5. How accurately can we predict future sales?

Linear regression can be used to answer each of these questions.

SUMMARY OUTPUT in R

```
Call:
lm(formula = Sales ~ ., data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8692 -0.6908  0.0211  0.6636  3.4115

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***
CompPrice        0.0928153  0.0041477  22.378  < 2e-16 ***
Income           0.0158028  0.0018451   8.565 2.58e-16 ***
Advertising      0.1230951  0.0111237  11.066  < 2e-16 ***
Population       0.0002079  0.0003705   0.561    0.575
Price           -0.0953579  0.0026711 -35.700  < 2e-16 ***
ShelveLocGood    4.8501827  0.1531100  31.678  < 2e-16 ***
ShelveLocMedium  1.9567148  0.1261056  15.516  < 2e-16 ***
Age             -0.0460452  0.0031817 -14.472  < 2e-16 ***
Education       -0.0211018  0.0197205  -1.070    0.285
UrbanYes         0.1228864  0.1129761   1.088    0.277
USYes           -0.1840928  0.1498423  -1.229    0.220
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.019 on 388 degrees of freedom
Multiple R-squared:  0.8734,    Adjusted R-squared:  0.8698
F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```

SUMMARY OUTPUT in Excel

Regression Statistics	
Multiple R	0.93096584
R Square	0.8666974
Adjusted R Square	0.8632706
Standard Error	1.04427131
Observations	400

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	10	2758.069202	275.8069202	252.9172607	2.1457E-163
Residual	389	424.2054957	1.09050256		
Total	399	3182.274698			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	2.99117944	0.630540793	4.743831753	2.94955E-06	1.751485123
CompPrice	0.09255237	0.004250367	21.77514872	2.5798E-69	0.084195802
Income	0.01615272	0.001889343	8.549386868	2.86354E-16	0.012438122
Advertising	0.12036692	0.011383601	10.57371215	3.87637E-23	0.097985836
Population	0.00029046	0.000379208	0.765975791	0.444155237	-0.00045509
Price	-0.0952477	0.002737379	-34.7952211	1.6832E-121	-0.100629614
Age	-0.0468605	0.00325562	-14.39373328	5.86807E-38	-0.053261339
Education	-0.020948	0.020210801	-1.036476421	0.300623799	-0.060684093
Urban_encoded	0.14120929	0.115711175	1.22036005	0.223067678	-0.086288254
US_encoded	-0.1293475	0.153069318	-0.845025443	0.398616061	-0.430294157
ShelveLoc_encoded	2.41157374	0.078399038	30.76024662	3.2204E-106	2.257434878

- Multiple R. This is the *correlation coefficient*. It tells you how strong the linear relationship is. It is the square root of r squared

Linear Regression in Excel

- Multiple R. This is the *correlation coefficient*. It tells you how strong the linear relationship is. It is the square root of r squared
- R squared. This is r^2, the *Coefficient of Determination*. It tells you the percentage of the response variable variation that is explained by a linear model.

Linear Regression in Excel

- Multiple R. This is the *correlation coefficient*. It tells you how strong the linear relationship is. It is the square root of r squared
- R squared. This is r^2, the *Coefficient of Determination*. It tells you the percentage of the response variable variation that is explained by a linear model.
- Adjusted R square. The adjusted R-square adjusts for the number of predictors in a model. It is used to compare two models with different number of predictors.

Linear Regression in Excel

- Multiple R. This is the *correlation coefficient*. It tells you how strong the linear relationship is. It is the square root of r squared
- R squared. This is r^2, the *Coefficient of Determination*. It tells you the percentage of the response variable variation that is explained by a linear model.
- Adjusted R square. The adjusted R-square adjusts for the number of predictors in a model. It is used to compare two models with different number of predictors.
- Standard Error of the regression: An estimate of the standard deviation of the error term ϵ.

Linear Regression in Excel

- Multiple R. This is the *correlation coefficient*. It tells you how strong the linear relationship is. It is the square root of r squared
- R squared. This is r^2, the *Coefficient of Determination*. It tells you the percentage of the response variable variation that is explained by a linear model.
- Adjusted R square. The adjusted R-square adjusts for the number of predictors in a model. It is used to compare two models with different number of predictors.
- Standard Error of the regression: An estimate of the standard deviation of the error term ϵ.
- Observations. Number of observations in the sample.

Linear Regression in Excel

- Regression Sum of Squares:

$$(\hat{y}_1 - \bar{y})^2 + (\hat{y}_2 - \bar{y})^2 + \ldots + (\hat{y}_n - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 2758$$

Linear Regression in Excel

- Regression Sum of Squares:

$$(\hat{y}_1 - \bar{y})^2 + (\hat{y}_2 - \bar{y})^2 + \ldots + (\hat{y}_n - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 2758$$

- Residual Sum of Squares:

$$(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_n - \hat{y}_n)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 424$$

Linear Regression in Excel

- Regression Sum of Squares:

$$(\hat{y}_1 - \bar{y})^2 + (\hat{y}_2 - \bar{y})^2 + \ldots + (\hat{y}_n - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 2758$$

- Residual Sum of Squares:

$$(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_n - \hat{y}_n)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 424$$

- Total Sum of Squares:

$$(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 3182$$

Linear Regression in Excel

- Regression Sum of Squares:

$$(\hat{y}_1 - \bar{y})^2 + (\hat{y}_2 - \bar{y})^2 + \ldots + (\hat{y}_n - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 2758$$

- Residual Sum of Squares:

$$(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_n - \hat{y}_n)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 424$$

- Total Sum of Squares:

$$(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 3182$$

Linear Regression in Excel

- Regression Sum of Squares:

$$(\hat{y}_1 - \bar{y})^2 + (\hat{y}_2 - \bar{y})^2 + \ldots + (\hat{y}_n - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 2758$$

- Residual Sum of Squares:

$$(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_n - \hat{y}_n)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 424$$

- Total Sum of Squares:

$$(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 3182$$

Note: Total Sum of Squares = Regression Sum of Squares: + Residual Sum of Squares

F-statistic

- F: Overall F test for the null hypothesis:

  H_0 : There is no relationship between predictors and the response

  versus the *alternative hypothesis*

  H_A : There is some relationship between predictors and the response

F-statistic

- F: Overall F test for the null hypothesis:

  H_0 : There is no relationship between predictors and the response

  versus the *alternative hypothesis*

  H_A : There is some relationship between predictors and the response

- F: Overall F test for the null hypothesis:

  H_0 : There is no relationship between predictors and the response

  versus the *alternative hypothesis*

  H_A : There is some relationship between predictors and the response

  Mathematically, this corresponds to testing

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_m = 0$$
$$H_A : \text{ at least one } \beta_j \text{ is non-zero.}$$

F-statistic

- F: Overall F test for the null hypothesis:

  H_0 : There is no relationship between predictors and the response

  versus the *alternative hypothesis*

  H_A : There is some relationship between predictors and the response

  Mathematically, this corresponds to testing

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_m = 0$$
$$H_A : \text{ at least one } \beta_j \text{ is non-zero.}$$

F-statistic

This hypothesis test is performed by computing the F-statistic:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/m}{\mathrm{RSS}/(n - m - 1)}$$

where

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

F-statistic

This hypothesis test is performed by computing the F-statistic:

$$F = \frac{(\text{TSS} - \text{RSS})/m}{\text{RSS}/(n - m - 1)}$$

where

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

If the linear model assumptions are correct, one can show that

$$E\left[\text{RSS}/(n - m - 1)\right] = \sigma^2$$

and that, provided H_0 is true,

$$E\left[(\text{TSS} - \text{RSS})/m\right] = \sigma^2$$

F-statistic

- Hence, when there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1.
- On the other hand, if H_A is true, then $E\left[(\text{TSS} - \text{RSS})/m\right] > \sigma^2$, so we expect F to be greater than 1.

F-statistic

- Hence, when there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1.
- On the other hand, if H_A is true, then $E\left[(\mathrm{TSS} - \mathrm{RSS})/m\right] > \sigma^2$, so we expect F to be greater than 1.

p-value

p-value: the probability of observing any value equal to F-statistic or larger assuming there is no relationship between predictors and the response.

p-value

p-value: the probability of observing any value equal to F-statistic or larger assuming there is no relationship between predictors and the response.

- A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis (data are unlikely with a true null)
- A large p-value (typically > 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis (data are likely with a true null)

Questions

1. Is there a relationship between predictors and sales?
2. How strong is the relationship between predictors and sales?
3. Which predictors contribute to sales? What predictors matter most?
4. How accurately can we estimate the effect of each predictor on sales? What is the degree to which price affects the sales (that is, price sensitivity)? How accurately can we predict this amount?
5. How accurately can we predict future sales?

Questions

1. Is there a relationship between predictors and sales?
2. How strong is the relationship between predictors and sales?
3. Which predictors contribute to sales? What predictors matter most?
4. How accurately can we estimate the effect of each predictor on sales? What is the degree to which price affects the sales (that is, price sensitivity)? How accurately can we predict this amount?
5. How accurately can we predict future sales?

Linear regression can be used to answer each of these questions.

Sales of Child Car Seats

Question Is there a relationship between predictors and sales?

Sales of Child Car Seats

Question Is there a relationship between predictors and sales?

This question can be answered by testing the hypothesis

$$H_0 : \beta_{\text{CompPrice}} = \beta_{\text{Income}} = \ldots = \beta_{\text{US}} = 0$$
$$H_A : \text{ at least one } \beta_j \text{ is non-zero.}$$

Sales of Child Car Seats

Question Is there a relationship between predictors and sales?

This question can be answered by testing the hypothesis

$$H_0 : \beta_{\text{CompPrice}} = \beta_{\text{Income}} = \ldots = \beta_{\text{US}} = 0$$
$$H_A : \text{ at least one } \beta_j \text{ is non-zero.}$$

- The F-statistic can be used to determine whether or not we should reject this null hypothesis.
- The p-value corresponding to the F-statistic is very low, indicating clear evidence of a relationship between predictors and sales.

Sales of Child Car Seats

Question How strong is the relationship between predictors and sales?

Sales of Child Car Seats

Question How strong is the relationship between predictors and sales?

- This question can be answered by R-squared that tells us the percentage of variability in the response that is explained by the predictors.
- The predictors explain almost 93 % of the variance in sales.

Sales of Child Car Seats

Question How accurately can we estimate the effect of each predictor on sales?

Sales of Child Car Seats

Question How accurately can we estimate the effect of each predictor on sales?

This question can be answered by the standard errors of coefficients to construct confidence intervals for each coefficient.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	2.99117944	0.630540793	4.743831753	2.94955E-06	1.751485123	4.23087375	1.95155628	4.03080259
CompPrice	0.09255237	0.004250367	21.77514872	2.5798E-69	0.084195802	0.10090893	0.08554445	0.09956029
Income	0.01615272	0.001889343	8.549386868	2.86354E-16	0.012438122	0.01986733	0.01303761	0.01926783
Advertising	0.12036692	0.011383601	10.57371215	3.87637E-23	0.097985836	0.142748	0.10159786	0.13913597
Population	0.00029046	0.000379208	0.765975791	0.444155237	-0.00045509	0.00103602	-0.0003348	0.0009157
Price	-0.0952477	0.002737379	-34.7952211	1.6832E-121	-0.100629614	-0.0898658	-0.099761	-0.0907344
Age	-0.0468605	0.00325562	-14.39373328	5.86807E-38	-0.053261339	-0.0404597	-0.0522283	-0.0414927
Education	-0.020948	0.020210801	-1.036476421	0.300623799	-0.060684093	0.01878805	-0.0542712	0.01237515
Urban_encoded	0.14120929	0.115711175	1.22036005	0.223067678	-0.086288254	0.36870684	-0.049573	0.33199159
US_encoded	-0.1293475	0.153069318	-0.845025443	0.398616061	-0.430294157	0.17159922	-0.3817251	0.12303019
ShelveLoc_encoded	2.41157374	0.078399038	30.76024662	3.2204E-106	2.257434878	2.56571261	2.28231096	2.54083652

Sales of Child Car Seats

Question Which predictors contribute to sales?

Question Which predictors contribute to sales?

- To answer this question, we can examine the p-values associated with each predictor's t-statistic.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	2.99117944	0.630540793	4.743831753	2.94955E-06	1.751485123	4.23087375	1.95155628	4.03080259
CompPrice	0.09255237	0.004250367	21.77514872	2.5798E-69	0.084195802	0.10090893	0.08554445	0.09956029
Income	0.01615272	0.001889343	8.549386868	2.86354E-16	0.012438122	0.01986733	0.01303761	0.01926783
Advertising	0.12036692	0.011383601	10.57371215	3.87637E-23	0.097985836	0.142748	0.10159786	0.13913597
Population	0.00029046	0.000379208	0.765975791	0.444155237	-0.00045509	0.00103602	-0.0003348	0.0009157
Price	-0.0952477	0.002737379	-34.7952211	1.6832E-121	-0.100629614	-0.0898658	-0.099761	-0.0907344
Age	-0.0468605	0.00325562	-14.39373328	5.86807E-38	-0.053261339	-0.0404597	-0.0522283	-0.0414927
Education	-0.020948	0.020210801	-1.036476421	0.300623799	-0.060684093	0.01878805	-0.0542712	0.01237515
Urban_encoded	0.14120929	0.115711175	1.22036005	0.223067678	-0.086288254	0.36870684	-0.049573	0.33199159
US_encoded	-0.1293475	0.153069318	-0.845025443	0.398616061	-0.430294157	0.17159922	-0.3817251	0.12303019
ShelveLoc_encoded	2.41157374	0.078399038	30.76024662	3.2204E-106	2.257434878	2.56571261	2.28231096	2.54083652

Sales of Child Car Seats

Question Which predictors contribute to sales?

- To answer this question, we can examine the p-values associated with each predictor's t-statistic.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 90.0%	Upper 90.0%
Intercept	2.99117944	0.630540793	4.743831753	2.94955E-06	1.751485123	4.23087375	1.95155628	4.03080259
CompPrice	0.09255237	0.004250367	21.77514872	2.5798E-69	0.084195802	0.10090893	0.08554445	0.09956029
Income	0.01615272	0.001889343	8.549386868	2.86354E-16	0.012438122	0.01986733	0.01303761	0.01926783
Advertising	0.12036692	0.011383601	10.57371215	3.87637E-23	0.097985836	0.142748	0.10159786	0.13913597
Population	0.00029046	0.000379208	0.765975791	0.444155237	-0.00045509	0.00103602	-0.0003348	0.0009157
Price	-0.0952477	0.002737379	-34.7952211	1.6832E-121	-0.100629614	-0.0898658	-0.099761	-0.0907344
Age	-0.0468605	0.00325562	-14.39373328	5.86807E-38	-0.053261339	-0.0404597	-0.0522283	-0.0414927
Education	-0.020948	0.020210801	-1.036476421	0.300623799	-0.060684093	0.01878805	-0.0542712	0.01237515
Urban_encoded	0.14120929	0.115711175	1.22036005	0.223067678	-0.086288254	0.36870684	-0.049573	0.33199159
US_encoded	-0.1293475	0.153069318	-0.845025443	0.398616061	-0.430294157	0.17159922	-0.3817251	0.12303019
ShelveLoc_encoded	2.41157374	0.078399038	30.76024662	3.2204E-106	2.257434878	2.56571261	2.28231096	2.54083652

- The p-values for CompPrice, Income, Advertising, Price, Age, and ShelveLoc are low, but the p-value for Population, Education, Urban, US is not.
- This suggests that only CompPrice, Income, Advertising, Price, Age, and ShelveLoc.

Sales of Child Car Seats

Question How accurately can we predict future sales?

Sales of Child Car Seats

Question How accurately can we predict future sales?

This question can be answered by the assumption that sales is a random variable with normal distribution.

- Sales is random variable:

$$\textbf{Sales} \sim N(13.64192 - 0.05307 \times \textbf{Price}, 2.525987)$$

Sales of Child Car Seats

Question How accurately can we predict future sales?

This question can be answered by the assumption that sales is a random variable with normal distribution.

- Sales is random variable:

$$\textbf{Sales} \sim N(13.64192 - 0.05307 \times \textbf{Price}, 2.525987)$$

- The expected value of **Sales** is

$$E(\textbf{Sales}) = 13.64192 - 0.05307 \times \textbf{Price}.$$

Sales of Child Car Seats

Question How accurately can we predict future sales?

This question can be answered by the assumption that sales is a random variable with normal distribution.

- Sales is random variable:

$$\mathbf{Sales} \sim N(13.64192 - 0.05307 \times \mathbf{Price}, 2.525987)$$

- The expected value of **Sales** is

$$E(\mathbf{Sales}) = 13.64192 - 0.05307 \times \mathbf{Price}.$$

- 95% confidence interval is given by

$$\begin{aligned}&(13.64192 - 0.05307 \times \mathbf{Price} - 2 \times 2.525987,\\ &\ 13.64192 - 0.05307 \times \mathbf{Price} + 2 \times 2.525987)\\ &= (8.589946 - 0.05307 \times \mathbf{Price}, 18.69389 - 0.05307 \times \mathbf{Price})\end{aligned}$$

Sales of Child Car Seats

Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urban	US
9.50	138	73	11	276	120	Bad	42	17	Yes	Yes
11.22	111	48	16	260	83	Good	65	10	Yes	Yes
10.06	113	35	10	269	80	Medium	59	12	Yes	Yes
7.40	117	100	4	466	97	Medium	55	14	Yes	Yes
4.15	141	64	3	340	128	Bad	38	13	Yes	No
10.81	124	113	13	501	72	Bad	78	16	No	Yes
6.63	115	105	0	45	108	Medium	71	15	Yes	No
11.85	136	81	15	425	120	Good	67	10	Yes	Yes
6.54	132	110	0	108	124	Medium	76	10	No	No
4.69	132	113	0	131	124	Medium	76	17	No	Yes
9.01	121	78	9	150	100	Bad	26	10	No	Yes
11.96	117	94	4	503	94	Good	50	13	Yes	Yes
3.98	122	35	2	393	136	Medium	62	18	Yes	No
10.96	115	28	11	29	86	Good	53	18	Yes	Yes
11.17	107	117	11	148	118	Good	52	18	Yes	Yes

Types of Variables

Quantitative: variables whose values representing counts or measurements. A quantitative variable is either

Types of Variables

Quantitative: variables whose values representing counts or measurements. A quantitative variable is either

- Discrete (distinct, separate values): number of bedrooms

Types of Variables

Quantitative: variables whose values representing counts or measurements. A quantitative variable is either

- Discrete (distinct, separate values): number of bedrooms
- Continuos: price, lot size

Types of Variables

Quantitative: variables whose values representing counts or measurements. A quantitative variable is either

- Discrete (distinct, separate values): number of bedrooms
- Continuos: price, lot size

Qualitative: variables whose values can be placed into nonnumeric categories. A qualitative variable is either

Types of Variables

Quantitative: variables whose values representing counts or measurements. A quantitative variable is either

- Discrete (distinct, separate values): number of bedrooms
- Continuos: price, lot size

Qualitative: variables whose values can be placed into nonnumeric categories. A qualitative variable is either

- Binary: Is a US Store?

Types of Variables

Quantitative: variables whose values representing counts or measurements. A quantitative variable is either

- Discrete (distinct, separate values): number of bedrooms
- Continuos: price, lot size

Qualitative: variables whose values can be placed into nonnumeric categories. A qualitative variable is either

- Binary: Is a US Store?
- Nominal: City

Types of Variables

Quantitative: variables whose values representing counts or measurements. A quantitative variable is either

- Discrete (distinct, separate values): number of bedrooms
- Continuos: price, lot size

Qualitative: variables whose values can be placed into nonnumeric categories. A qualitative variable is either

- Binary: Is a US Store?
- Nominal: City
- Ordinal: ShelveLoc (Bad, Medium, Good)

Binary Variables

Question: How do we represent qualitative variables with two levels?

We simply create an indicator or dummy variable that takes on two possible numerical values.

Binary Variables

Question: How do we represent qualitative variables with two levels?

We simply create an indicator or dummy variable that takes on two possible numerical values.

Examples:

- Is a US store?

$$x = \begin{cases} 1 & \text{if Yes} \\ 0 & \text{Otherwise} \end{cases}$$

- Is an Urban store?

$$x = \begin{cases} 1 & \text{if Yes} \\ 0 & \text{Otherwise} \end{cases}$$

Nominal Variables

Question: How do we represent qualitative variables with more than two levels?

When a qualitative predictor has more than two levels, a single dummy variable cannot represent all possible values. In this situation, we can create additional dummy variables.

Nominal Variables

Question: How do we represent qualitative variables with more than two levels?

When a qualitative predictor has more than two levels, a single dummy variable cannot represent all possible values. In this situation, we can create additional dummy variables.

Example:

- CITY with three levels: Seattle, Bellevue, Kirkland

$$x_{\text{Seattle}} = \begin{cases} 1 & \text{if Seattle} \\ 0 & \text{Otherwise} \end{cases} \qquad x_{\text{Bellevue}} = \begin{cases} 1 & \text{if Bellevue} \\ 0 & \text{Otherwise} \end{cases}$$

Nominal Variables

Question: How do we represent qualitative variables with more than two levels?

When a qualitative predictor has more than two levels, a single dummy variable cannot represent all possible values. In this situation, we can create additional dummy variables.

Example:

- CITY with three levels: Seattle, Bellevue, Kirkland

$$x_{\text{Seattle}} = \begin{cases} 1 & \text{if Seattle} \\ 0 & \text{Otherwise} \end{cases} \qquad x_{\text{Bellevue}} = \begin{cases} 1 & \text{if Bellevue} \\ 0 & \text{Otherwise} \end{cases}$$

Note: Only $K - 1$ dummies can (in general) be included, where K is the number of categories of the qualitative variable.