BUDT 730
Data, Models and Decisions

Lecture 12

Regression Analysis (4)

Transformation of Variables

Prof. Sujin Kim

Regression Analysis

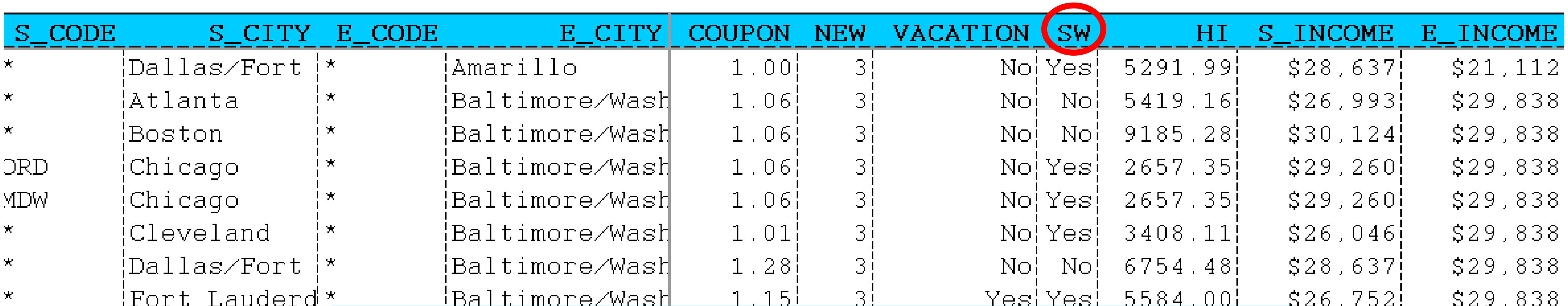
Variable Transformation

Variable Transformations

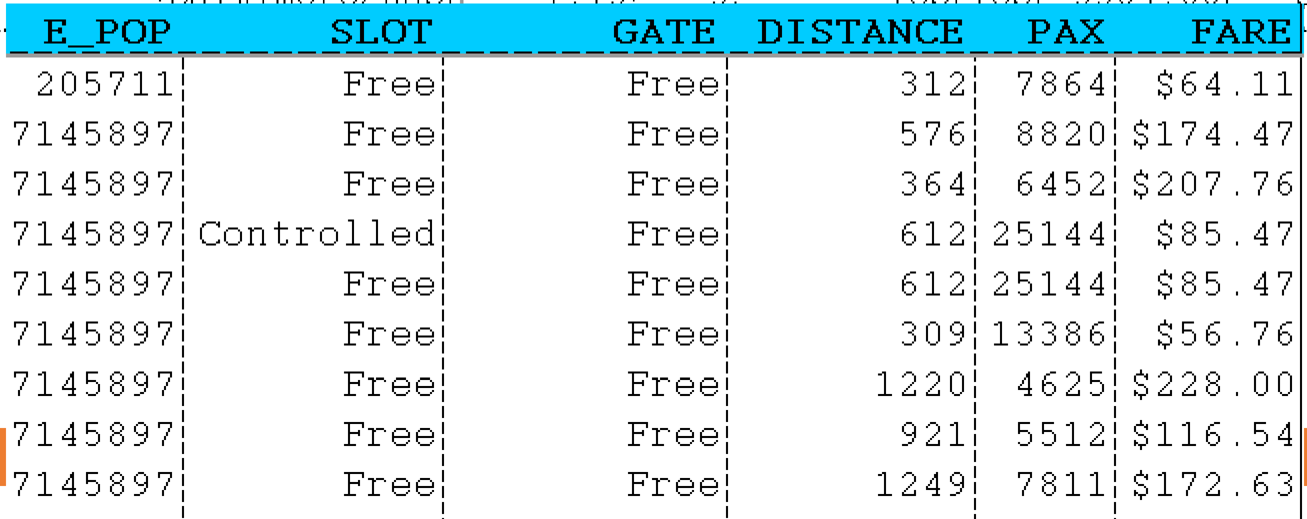
- Several types of independent variables can be used in regression equations:
  - Dummy variables
  - Interaction variables
  - Nonlinear transformations
- We should be selective, and not include too many different types in a particular regression model
  - Only a few might improve the fit!
- Dataset:
  - `Airline data.xlsx`

Example: Southwest Airline Data

- We would like to investigate the effect of Southwest Airlines on Fares.



S_CODE	S_CITY	E_CODE	E_CITY	COUPON	NEW	VACATION	SW	HI	S_INCOME	E_INCOME
*	Dallas/Fort	*	Amarillo	1.00	3	No	Yes	5291.99	$28,637	$21,112
*	Atlanta	*	Baltimore/Wash	1.06	3	No	No	5419.16	$26,993	$29,838
*	Boston	*	Baltimore/Wash	1.06	3	No	No	9185.28	$30,124	$29,838
ORD	Chicago	*	Baltimore/Wash	1.06	3	No	Yes	2657.35	$29,260	$29,838
MDW	Chicago	*	Baltimore/Wash	1.06	3	No	Yes	2657.35	$29,260	$29,838
*	Cleveland	*	Baltimore/Wash	1.01	3	No	Yes	3408.11	$26,046	$29,838
*	Dallas/Fort	*	Baltimore/Wash	1.28	3	No	No	6754.48	$28,637	$29,838
*	Fort Lauderd	*	Baltimore/Wash	1.15	3	Yes	Yes	5584.00	$26,752	$29,838



E_POP	SLOT	GATE	DISTANCE	PAX	FARE
205711	Free	Free	312	7864	$64.11
7145897	Free	Free	576	8820	$174.47
7145897	Free	Free	364	6452	$207.76
7145897	Controlled	Free	612	25144	$85.47
7145897	Free	Free	612	25144	$85.47
7145897	Free	Free	309	13386	$56.76
7145897	Free	Free	1220	4625	$228.00
7145897	Free	Free	921	5512	$116.54
7145897	Free	Free	1249	7811	$172.63

Adding Categorical Variables

- Some independent variables are categorical and are not measured on a quantitative scale

- Therefore, we create one **dummy variable** for each category to indicate whether observations fall into that category

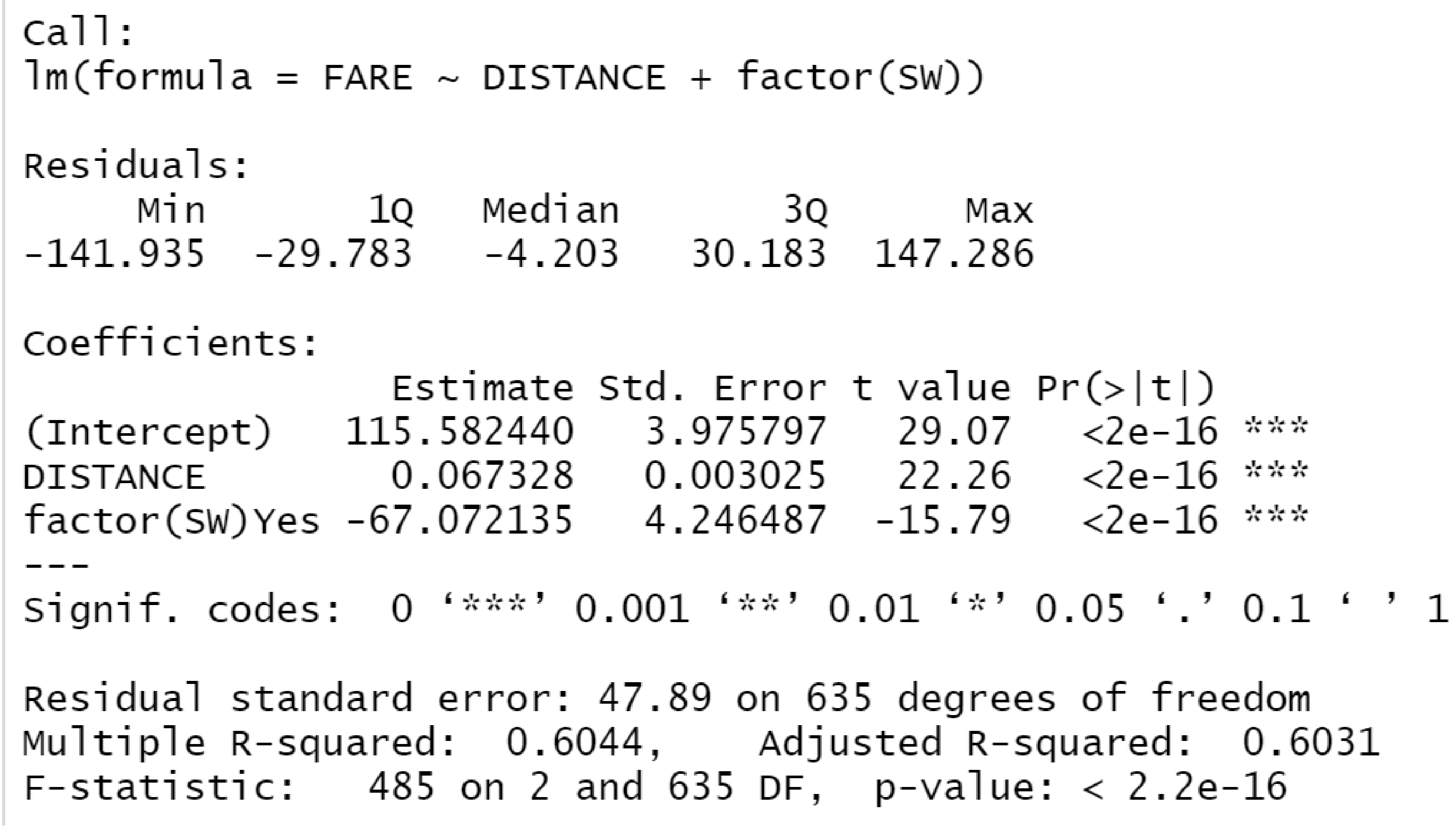
Rules for Using Dummy Variables

- When we add categorical variables to our model, we leave out one of the categories (dummies)
- If there are *m* categories, we include *(m-1)* dummy variables in our model
- Example:
  - We may choose to add *SW=yes* to our model, leaving out *SW=no*

- The category that is left out is called the ***reference*** or ***base* category**
  - In our example, the reference category is *SW=no*

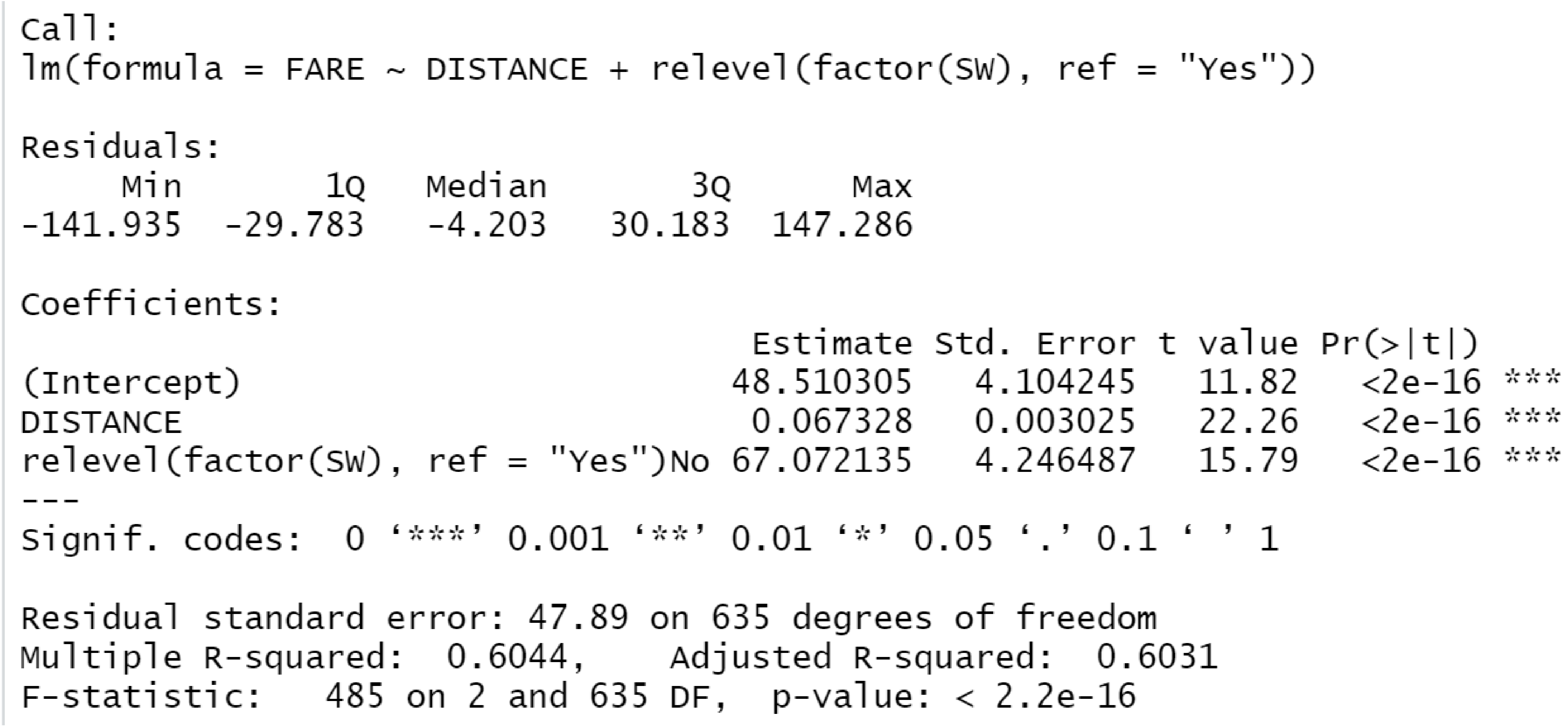
- Any category can be made the reference category.
- In R use **factor** to encode a categorical variables as a set of dummy variables

Adding Southwest to the Model: reference =“No”



```
Call:
lm(formula = FARE ~ DISTANCE + factor(SW))

Residuals:
     Min       1Q   Median       3Q      Max 
-141.935  -29.783   -4.203   30.183  147.286 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   115.582440   3.975797   29.07   <2e-16 ***
DISTANCE        0.067328   0.003025   22.26   <2e-16 ***
factor(SW)Yes -67.072135   4.246487  -15.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.89 on 635 degrees of freedom
Multiple R-squared:  0.6044,	Adjusted R-squared:  0.6031 
F-statistic:   485 on 2 and 635 DF,  p-value: < 2.2e-16
```

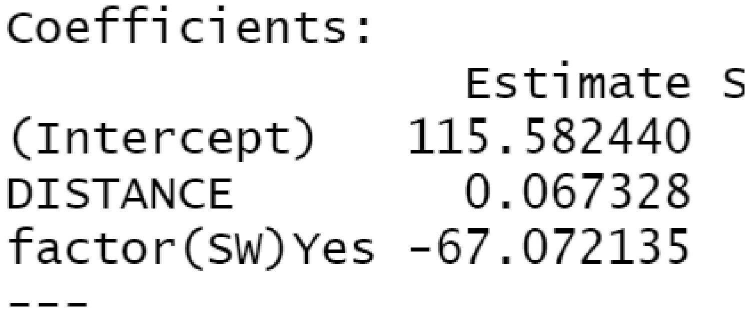
Adding Southwest to the Model: reference =“Yes”



```
Call:
lm(formula = FARE ~ DISTANCE + relevel(factor(SW), ref = "Yes"))

Residuals:
     Min       1Q   Median       3Q      Max 
-141.935  -29.783   -4.203   30.183  147.286 

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                        48.510305   4.104245   11.82   <2e-16 ***
DISTANCE                            0.067328   0.003025   22.26   <2e-16 ***
relevel(factor(SW), ref = "Yes")No 67.072135   4.246487   15.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.89 on 635 degrees of freedom
Multiple R-squared:  0.6044,	Adjusted R-squared:  0.6031 
F-statistic:   485 on 2 and 635 DF,  p-value: < 2.2e-16
```

Adding Southwest to the Model

- Included SW=Yes



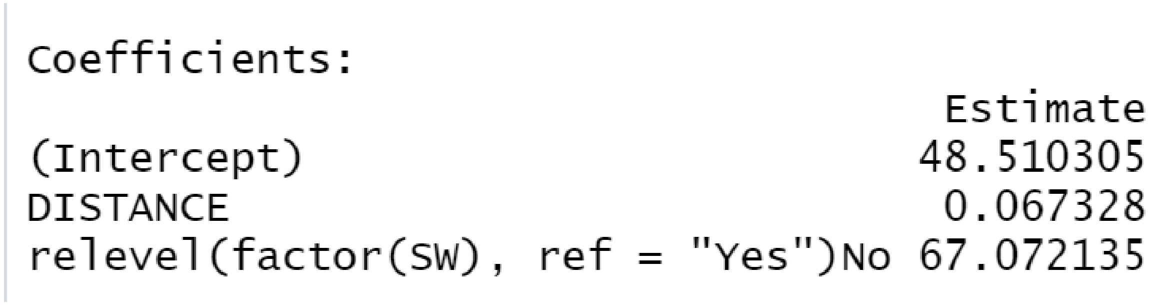
```
Coefficients:
                 Estimate S
(Intercept)    115.582440
DISTANCE         0.067328
factor(SW)Yes  -67.072135
---
```

What is the predicted fare for a route that is 5,000 miles and SW does not fly?

Fare = 115.6 + 0.06733*5,000

= $452.22

- Included SW=No



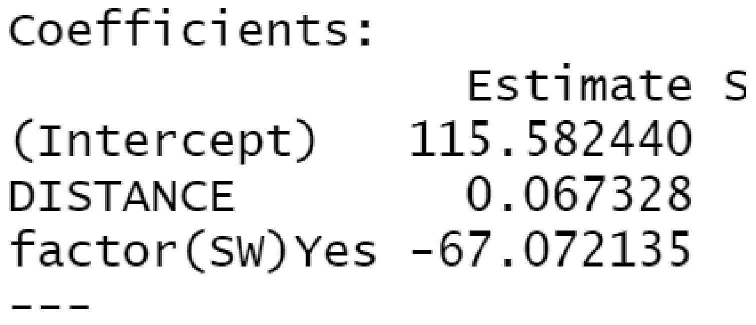
```
Coefficients:
                                        Estimate
(Intercept)                            48.510305
DISTANCE                                0.067328
relevel(factor(SW), ref = "Yes")No     67.072135
```

What is the predicted fare for a route that is 5,000 miles and SW does not fly?

Fare = 48.5 + 0.06733*5,000

+67.07*1 = $452.22

Interpretation of Dummy Variable



```
Coefficients:
                 Estimate S
(Intercept)    115.582440
DISTANCE         0.067328
factor(SW)Yes -67.072135
---
```

- One cannot increase SW by one unit!
- Therefore, the interpretation of coefficients for categorical dummies are always **relative to the base category** that was left out
- Our model:
  - *Fare = a + b1*Distance + b2* (SW=Yes)*
- Interpretation of b2 (= -67)

<u>On average</u>, the average fare is $**67 lower** if SW is present (compared to the route where SW does not present), <u>for routes of the same length.</u>

Interpretation of Dummy Variable

Interpretation of the coefficient of a dummy variable

$$Y = a + b_1(X = 1) + b_2(X = 2) \ldots$$

Suppose that the base variable is $(X = 0)$.

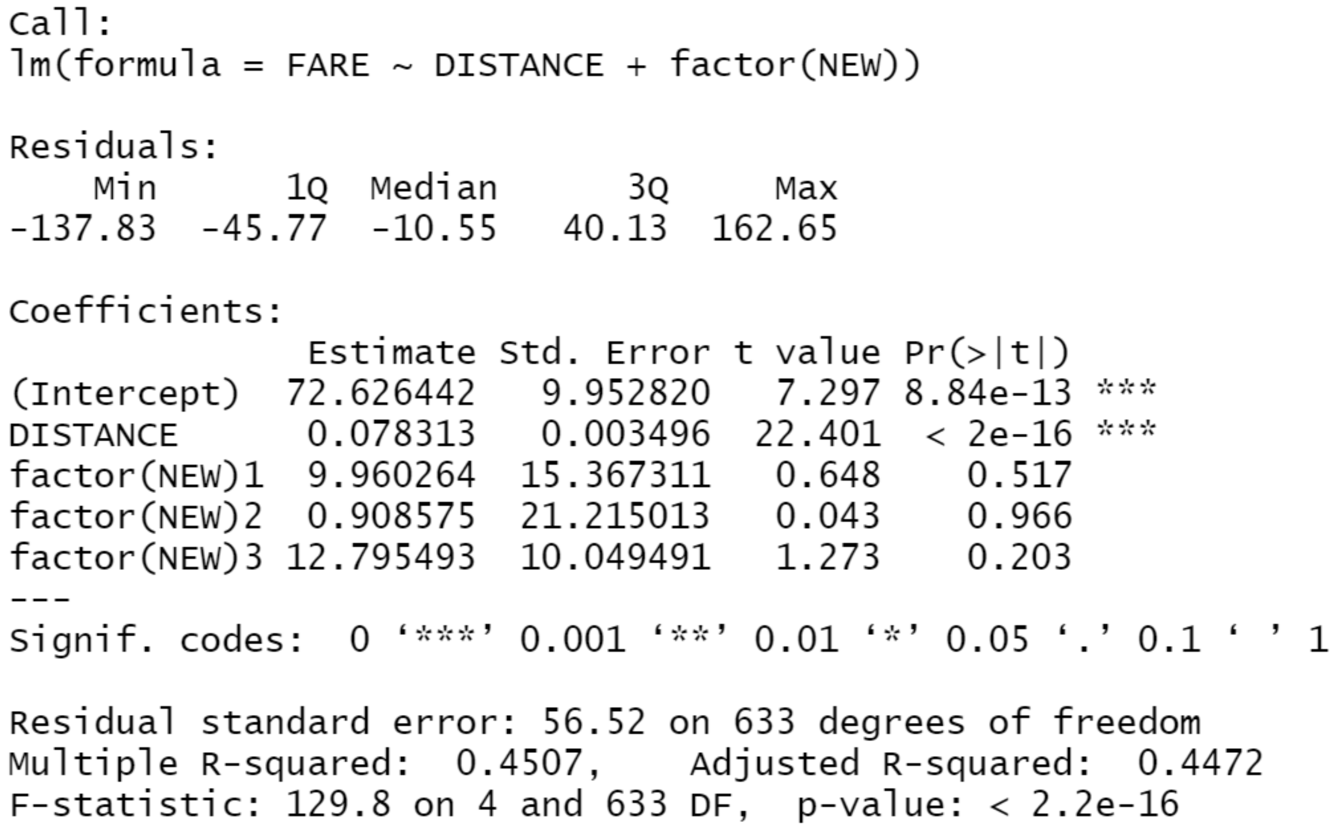
1. On average,

2 the value of Y in category i exceeds the value of Y in category 0 (base category) by b_i units

3. if all else held equal

Adding “NEW” to the Model

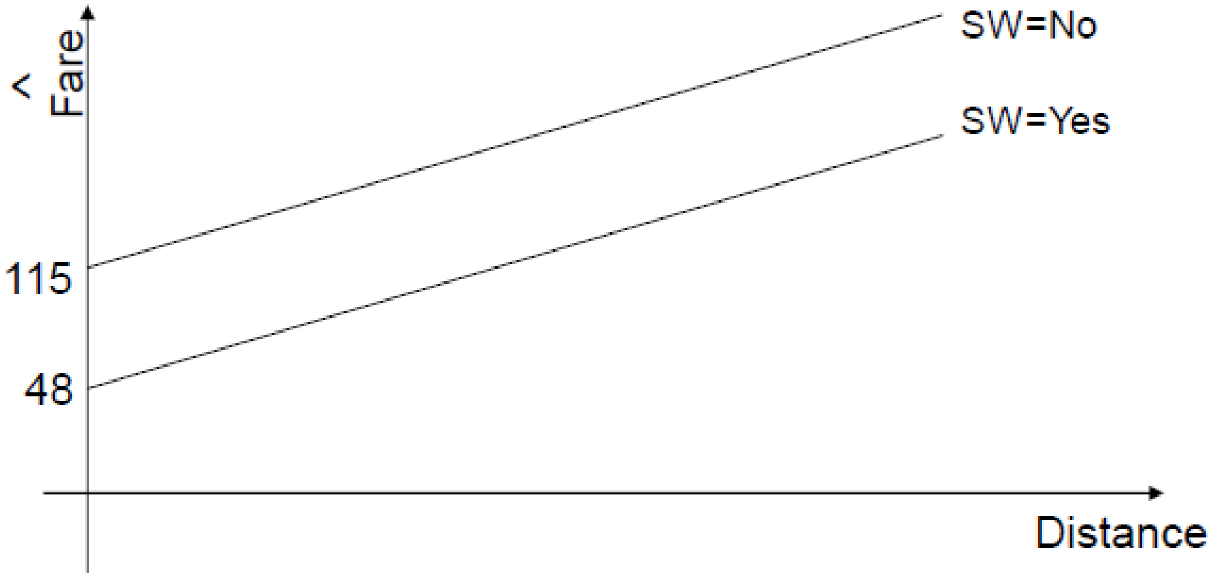
- NEW: number of new carriers entering that route between Q3-96 and Q2-97



```
Call:
lm(formula = FARE ~ DISTANCE + factor(NEW))

Residuals:
    Min      1Q  Median      3Q     Max 
-137.83  -45.77  -10.55   40.13  162.65 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  72.626442   9.952820   7.297 8.84e-13 ***
DISTANCE      0.078313   0.003496  22.401  < 2e-16 ***
factor(NEW)1  9.960264  15.367311   0.648    0.517    
factor(NEW)2  0.908575  21.215013   0.043    0.966    
factor(NEW)3 12.795493  10.049491   1.273    0.203    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 56.52 on 633 degrees of freedom
Multiple R-squared:  0.4507,	Adjusted R-squared:  0.4472 
F-statistic: 129.8 on 4 and 633 DF,  p-value: < 2.2e-16
```

Graphical Interpretation

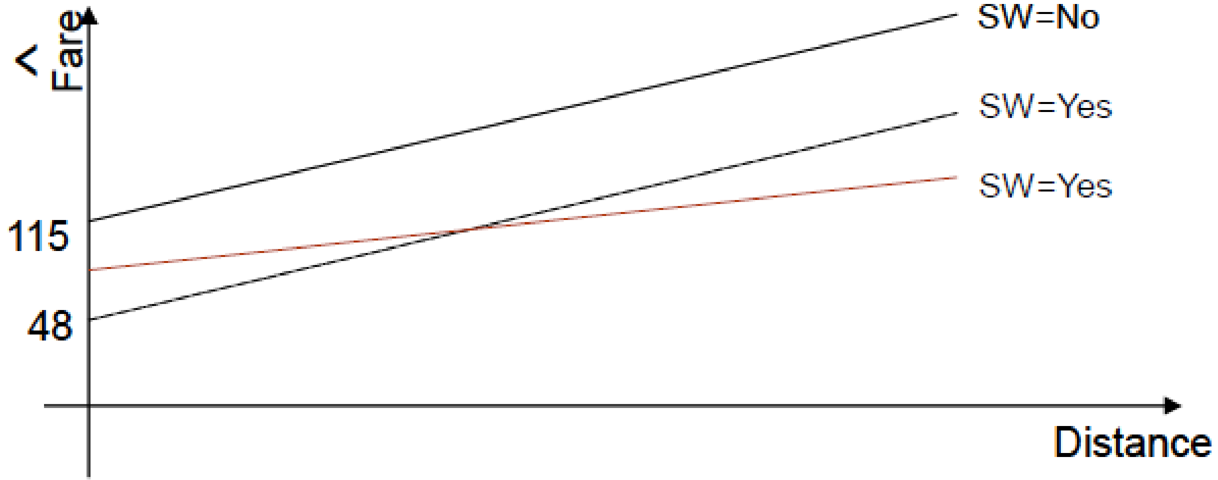
- Fare = 115 + 0.067 * Distance – 67* (SW=Yes)
  - Regression line when SW=no: Fare = 115 + 0.067 * Distance
  - Regression line when SW=yes: Fare = 115 + 0.067 * Distance -67



- After controlling for distance, the fare is more expensive when SW does not fly.

Interaction Terms

- When we include only a dummy variable in a regression equation, we are allowing the intercepts of the two lines to differ, but the lines are be parallel
- We want the rate of change to be different for different groups: To do so we introduce interaction terms

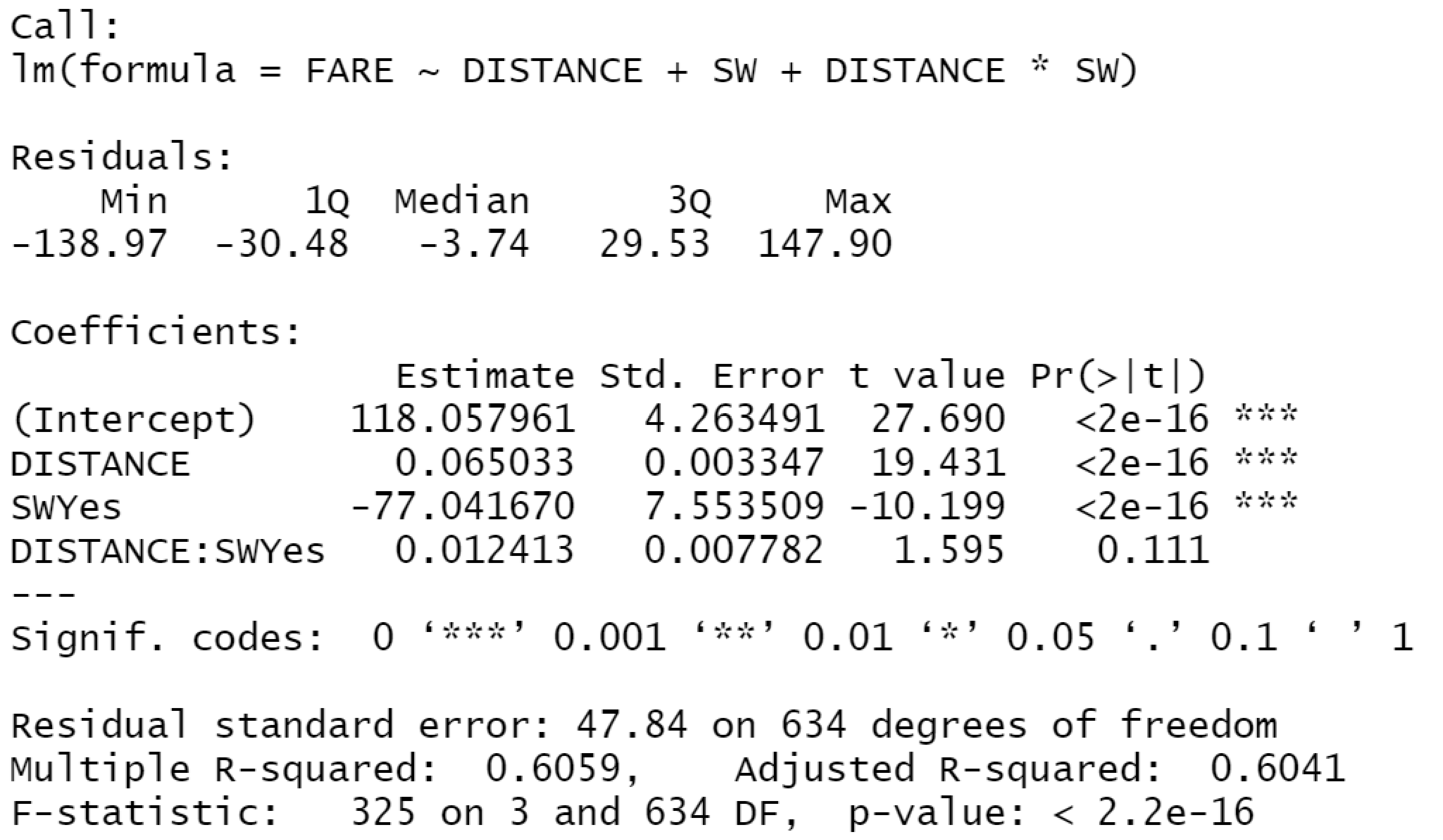


Interaction Terms

- An interaction variable is the product of two independent variables
- Suppose that the amount by which Fare increases for a unit increase in Distance is different for the routes where SW flies and those where SW does not.
- Construct a new variable *(SW=Yes)*Distance*
- This variable is obtained as the **product** between the columns of *(SW=Yes)* and *Distance*

Interaction Terms

- *Fare = a + b1 Distance + b2 (SW=Yes) + b3 Distance*(SW =Yes)*



```
Call:
lm(formula = FARE ~ DISTANCE + SW + DISTANCE * SW)

Residuals:
    Min      1Q  Median      3Q     Max 
-138.97  -30.48   -3.74   29.53  147.90 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    118.057961   4.263491  27.690   <2e-16 ***
DISTANCE         0.065033   0.003347  19.431   <2e-16 ***
SWYes          -77.041670   7.553509 -10.199   <2e-16 ***
DISTANCE:SWYes   0.012413   0.007782   1.595    0.111    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.84 on 634 degrees of freedom
Multiple R-squared:  0.6059,	Adjusted R-squared:  0.6041 
F-statistic:   325 on 3 and 634 DF,  p-value: < 2.2e-16
```

Interpretation of Interaction Coefficients

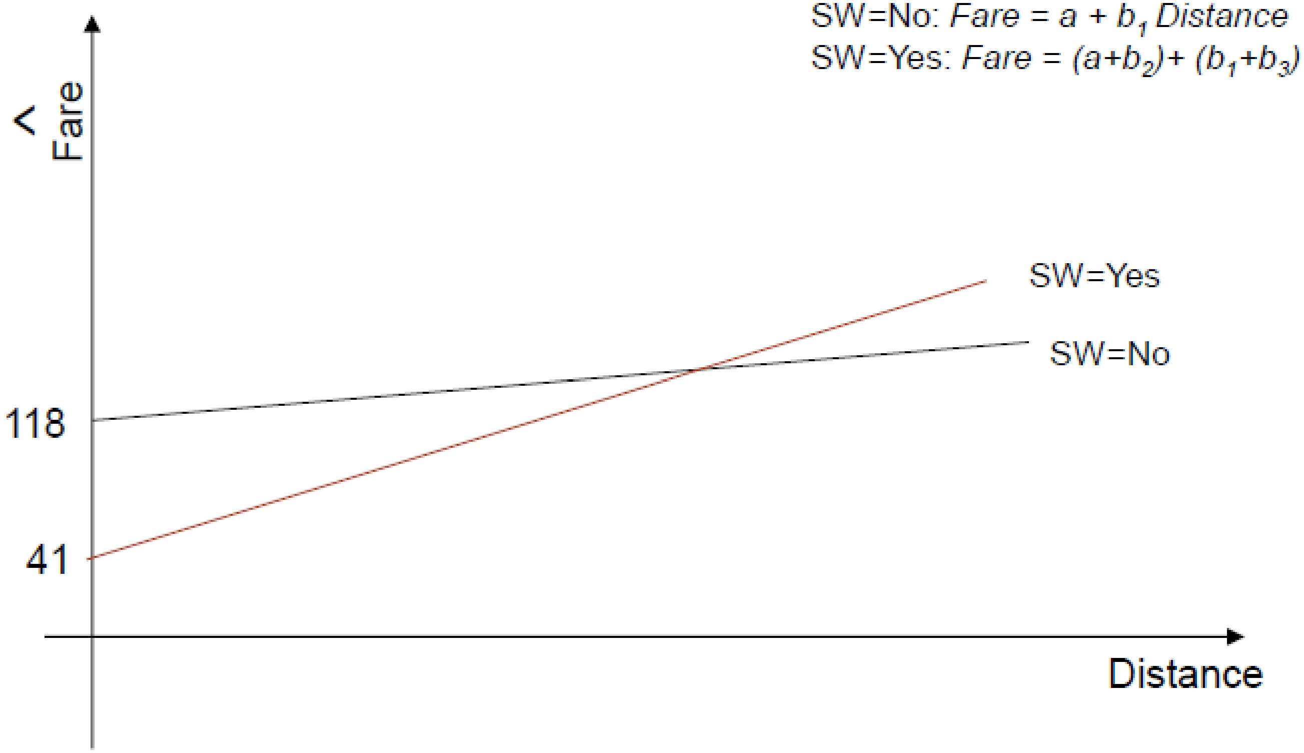
- The model:
  - *Fare = a + b1 Distance + b2 (SW=Yes) + b3 Distance*(SW =Yes)*
- For a route without Southwest (SW=No):
  - *Fare =* ***a*** *+* ***b1*** *Distance*
- For a route with Southwest (SW=Yes):
  - *Fare = a + b1 Distance + b2 (SW=Yes) + b3 Distance*(SW =Yes)*

    *= (**a+b2**) + (**b1+b3**) * Distance*

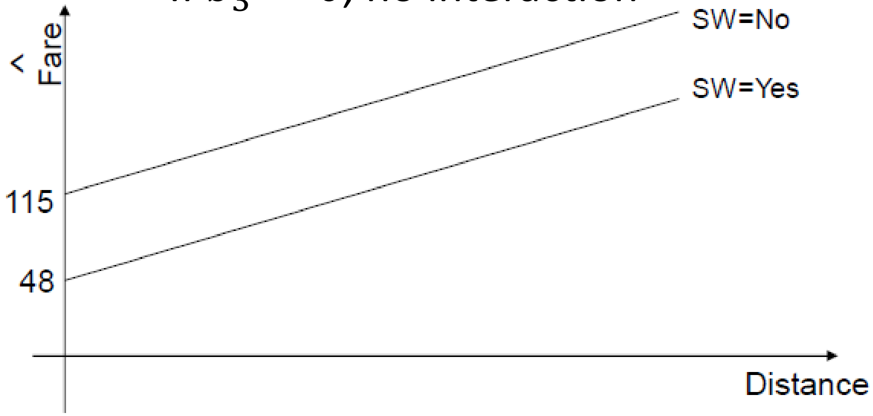
Graphically …

SW=No: $Fare = a + b_1 \, Distance$

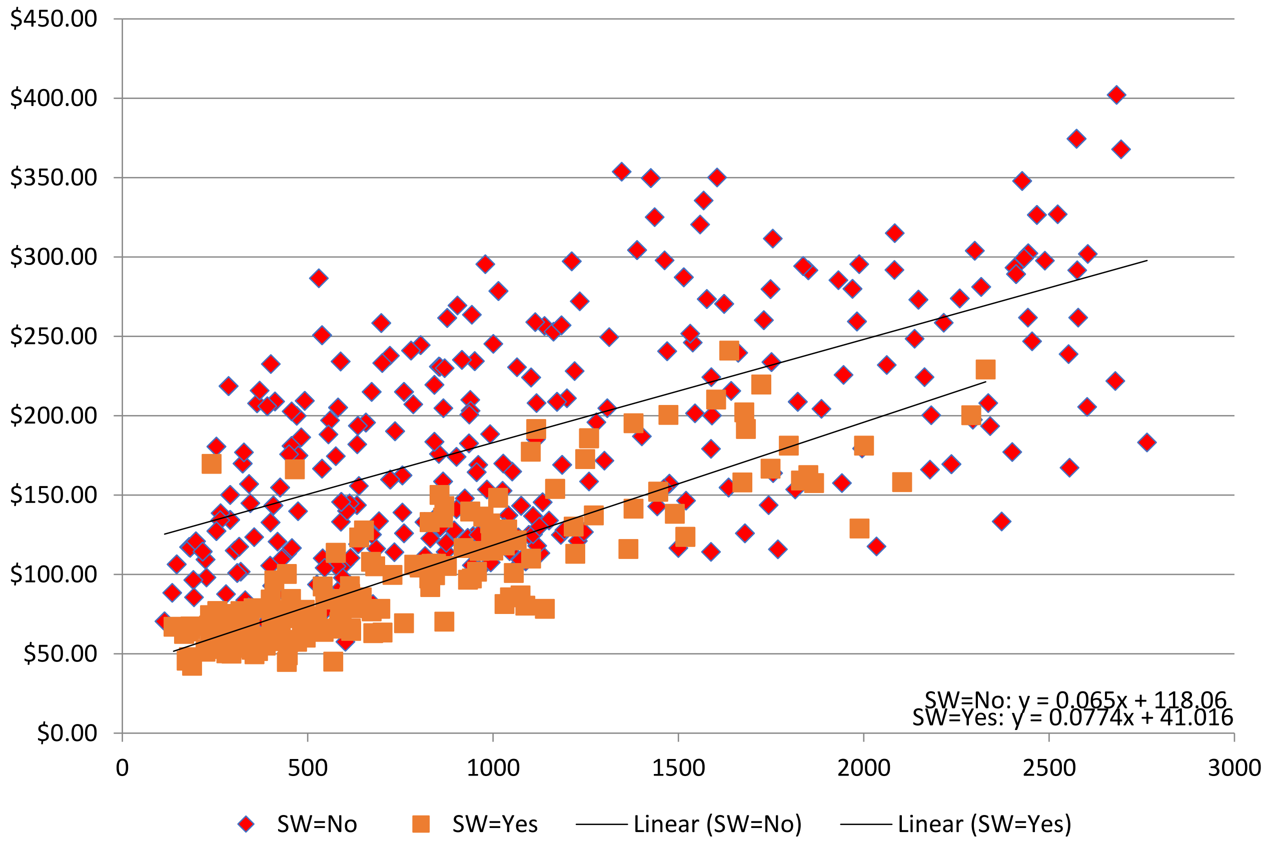
SW=Yes: $Fare = (a+b_2) + (b_1+b_3) * Distance$



If $b_3 = 0$, no interaction



Graphically - Data



Interpretation of Interaction Coefficients

- For a route without Southwest: *Fare = a + b1 Distance*
- For a route with Southwest: *Fare = (a+b2)+ (b1+b3) * Distance*
- Interpretation of the coefficients
    - a = 118, has no economic interpretation
    - b2= -77.0, has no economic interpretation. It is s the change in intercept when Southwest is present.
    - b1=0.065 is the average increase in fare per additional mile on routes where Southwest is not present
    - b3= 0.0124 is the average additional increase in fare per additional mile on routes where Southwest is present, compared to routes where Southwest is not present

Using the Regression Model

- What is the fare on a 5,000 mile route where Southwest is present?
  - *Fare = (a+b2)+ (b1+b3) * Distance*

    *= (118.05-77.041)+ (0.0650+0.0124)*5,000 = $428.24*
- How does it differ from a route where Southwest is not present?
  - *Fare = a + b1 Distance = 118 + 0.065*5,000 = $443.22*

Interaction Terms

- Interactions are a powerful modeling tool :

$$Y = a + b_1 X_1 + b_2 X_2 + c X_1 X_2$$

  - All three variables X_1, X_2 and $X_1 X_2$ must be added to the model

- They can be constructed between:
  - One numerical and one categorical variable
  - Two categorical variables
  - Two numerical variables (but, the interpretation is unclear)
  - … more variables for the adventurous