STP 530
Lecture 7: Regression Models with Categorical Predictors

Yi Zheng, Ph.D.
yi.isabel.zheng@asu.edu

ARIZONA STATE UNIVERSITY
SCHOOL OF MATHEMATICAL AND STATISTICAL SCIENCES

October 17, 2023

Outline

1. One categorical predictor
2. Two categorical predictors without interaction
3. Two categorical predictors with interaction
4. One categorical predictor and one numeric predictor

Example Problem:

Suppose we want to study whether the type of fuel influences the performance of a diesel engine.

How many variables are there in the problem?

1. Response variable Y: Performance score of car engine
2. Predictor X: Type of fuel

Suppose this study involves three fuel types:

- Petroleum-based fuel (F1)
- Coal-based fuel (F2)
- Blended fuel (F3)

A common way of handling categorical variables is to code the categories. For example, we create a column in the dataset named `fuel.type`, and the coding scheme is below:

- Petroleum-based fuel (F1) — `fuel.type = 1`
- Coal-based fuel (F2) — `fuel.type = 2`
- Blended fuel (F3) — `fuel.type = 3`

With the coded variables, technically we can fit a linear regression model:

$$\text{Performance} = \beta_0 + \beta_1(\text{fuel.type}) + \varepsilon$$

What is wrong with this model?

$$\text{Performance} = \beta_0 + \beta_1(\text{fuel.type}) + \varepsilon, \text{ where fuel.type} = \begin{cases} 1 & \text{if F1 fuel} \\ 2 & \text{if F2 fuel} \\ 3 & \text{if F3 fuel} \end{cases}$$

The above model **does not work** because it assumes the increase in the expected engine performance with 1-unit increase in fuel type always equals to β_1, which means —

- The difference in the expected performance between F1 and F2 is the same as that between F2 and F3.
- The difference in the expected performance between F1 and F3 is twice as much as that between F1 and F2 and between F2 and F3.

Those are unreasonable assumptions and illegitimate constraints on the model.

It is legitimate to use that model to compare between any **TWO** fuel types.

For example, if only F1 and F2 are compared, β_1 is the difference in the expected performance between the two fuel types.

$$\text{Performance} = \beta_0 + \beta_1(\text{fuel.type}) + \varepsilon, \text{ where fuel.type} = \begin{cases} 1 & \text{if F1 fuel} \\ 2 & \text{if F2 fuel} \end{cases}$$

If β_1 is positive, which fuel type is expected to yield better performance?

However, to use a single equation to model the expected performance for all three fuel types, a **dummy coding system** is needed:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where

$$X_1 = \begin{cases} 1, & \text{if F1 is used} \\ 0, & \text{if not} \end{cases}$$

$$X_2 = \begin{cases} 1, & \text{if F2 is used} \\ 0, & \text{if not} \end{cases}$$

F3 is the base level where $X_1 = 0$ and $X_2 = 0$.

Note: The textbook refers to X_1 and X_2 as **indicator variables**.

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Fuel Type	X_1	X_2	Expected Performance $E\{Y\}$
Petroleum (F1)			
Coal (F2)			
Blended (F3)			

- $\beta_0 = E\{Y\}_{\mathtt{F3}}$: Expected engine performance score with F3.
- $\beta_1 = E\{Y\}_{\mathtt{F1}} - E\{Y\}_{\mathtt{F3}}$: The difference in the expected engine performance score between F1 and F3.
- $\beta_2 = E\{Y\}_{\mathtt{F2}} - E\{Y\}_{\mathtt{F3}}$: The difference in the expected engine performance score between F2 and F3.
- What is the difference in the expected engine performance score between F1 and F2?

Data in experimental design grids:

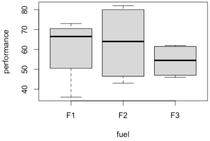
Table 5.8 Performance data for combinations of fuel type and diesel engine brand

		Brand	
		B_1	B_2
FUEL TYPE	F_1	65 73 68	36
	F_2	78 82	50 43
	F_3	48 46	61 62

Data format in R:

```
   performance fuel brand
1           65   F1    B1
2           73   F1    B1
3           68   F1    B1
4           78   F2    B1
5           82   F2    B1
6           48   F3    B1
7           46   F3    B1
8           36   F1    B2
9           50   F2    B2
10          43   F2    B2
11          61   F3    B2
12          62   F3    B2
```

```
> # Import data and inspect
> engine <- read.csv("engine.csv", stringsAsFactors=T)
> engine
   performance fuel brand
1           65   F1    B1
2           73   F1    B1
3           68   F1    B1
4           36   F1    B2
5           78   F2    B1
6           82   F2    B1
7           50   F2    B2
8           43   F2    B2
9           48   F3    B1
10          46   F3    B1
11          61   F3    B2
12          62   F3    B2
>
> # Given the "stringsAsFactors=T" argument in read.csv(), R reads in string
> # variables as a "factor" object.
> str(engine)
'data.frame':	12 obs. of  3 variables:
 $ performance: int  65 73 68 36 78 82 50 43 48 46 ...
 $ fuel       : Factor w/ 3 levels "F1","F2","F3": 1 1 1 1 2 2 2 2 3 3 ...
 $ brand      : Factor w/ 2 levels "B1","B2": 1 1 1 2 1 1 2 2 1 1 ...
>
> # You can view the levels (i.e., categories) in a factor variable by levels()
> levels(engine$fuel)
[1] "F1" "F2" "F3"
>
> #-------------------------------------------------------------------------
> # Visualization
>
> plot(performance ~ fuel, data=engine)
```




```
> # You can fit a linear regression model with a factor-type predictor.
> # When R sees a predictor is a factor object, it will create the dummy
> # codes at the back end.
>
> m1 <- lm(performance ~ fuel, data=engine)
> summary(m1)

Call:
lm(formula = performance ~ fuel, data = engine)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.500  -9.500   5.625   8.938  18.750 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   60.500      7.818   7.738 2.88e-05 ***
fuelF2         2.750     11.057   0.249    0.809    
fuelF3        -6.250     11.057  -0.565    0.586    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.64 on 9 degrees of freedom
Multiple R-squared:  0.07178,	Adjusted R-squared:  -0.1345 
F-statistic: 0.348 on 2 and 9 DF,  p-value: 0.7152

>
> # The model above is different from what we showed in the slides. That
> # is because by default R takes the first level in level() as the base
> # level. See below for the dummy code system it generated for m1.
>
> model.matrix(m1)
   (Intercept) fuelF2 fuelF3
1            1      0      0
2            1      0      0
3            1      0      0
4            1      0      0
5            1      1      0
6            1      1      0
7            1      1      0
8            1      1      0
9            1      0      1
10           1      0      1
11           1      0      1
12           1      0      1
```

```
> # To assign a different category as the base level, run releve().
>
> engine$fuel <- relevel(engine$fuel, ref="F3")
>
> # Check the levels now
>
> levels(engine$fuel)
[1] "F3" "F1" "F2"
>
> # Now the fitted model is what we want.
>
> m1 <- lm(performance ~ fuel, data=engine)
> summary(m1)

Call:
lm(formula = performance ~ fuel, data = engine)

Residuals:
    Min      1Q  Median      3Q     Max
-24.500  -9.500   5.625   8.938  18.750

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   54.250      7.818   6.939 6.77e-05 ***
fuelF1         6.250     11.057   0.565    0.586
fuelF2         9.000     11.057   0.814    0.437
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.64 on 9 degrees of freedom
Multiple R-squared:  0.07178,	Adjusted R-squared:  -0.1345
F-statistic: 0.348 on 2 and 9 DF,  p-value: 0.7152
```

```
> # Check the dummy code system.
>
> model.matrix(m1)
   (Intercept) fuelF1 fuelF2
1            1      1      0
2            1      1      0
3            1      1      0
4            1      1      0
5            1      0      1
6            1      0      1
7            1      0      1
8            1      0      1
9            1      0      0
10           1      0      0
11           1      0      0
12           1      0      0
```

The fitted model

$$\widehat{\text{performance}} = 54.25 + 6.25(\text{fuelF1}) + 9(\text{fuelF2})$$

Interpretation of model coefficients

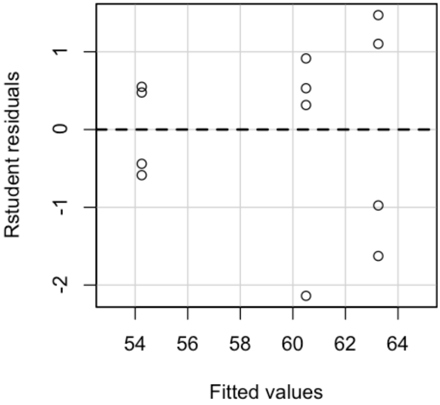
- $b_0 = 54.25$: Expected engine performance score for the base level, F3, blended fuel, is 54.25 points.
- $b_1 = 6.25$: The expected engine performance score for fuel F1 is 6.25 points higher than that for fuel F3.
- $b_2 = 9$: The expected engine performance score for fuel F2 is 9 points higher than that for fuel F3.
- What is the difference in the expected engine performance score between F1 and F2?

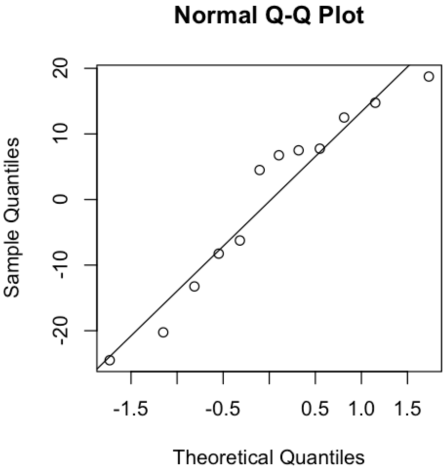
Diagnostics

Before we conduct statistical inferences, we must perform model diagnostics to make sure the underlying assumptions are roughly met.

1. **Linear trend:** This only applies to *numeric* predictors. So we do not need to check this for models including categorical predictors only.
2. **Outliers:** No excessive influences by outliers — Plot studentized residual against the $\hat{Y}$.
3. **Homoskedasticity**: The residuals have constant variance — The same residual plot as above.
4. **Normality**: The error terms are normally distributed — QQ-plot of residuals.
5. **Independence**: The error terms are independent between observations — Check data collection design.

Diagnostic residual plots





Confidence intervals for regression coefficients

The 95% confidence interval for β_1:

$$b_1 \pm t(.975; df = 9) \cdot s\{b_1\} = 6.25 \pm 2.262 * 11.057 = (-18.76, 31.26)$$

Interpretation:

- We are 95% sure that the true population parameter β_1 falls within this interval.
- With a 95% confidence, we estimate that expected engine performance score for F1 falls between 18.76 points lower than F3 and 31.26 points higher than F3.

The 95% confidence interval for β_2:

$$b_2 \pm t(.975; df = 9) \cdot s\{b_2\} = 9 \pm 2.262 * 11.057 = (-16, 34)$$

Confidence interval for $E\{Y_h\}$

We can create the confidence/prediction intervals for Y_h for all possible X values.

```
> # Confidence interval for E{Y}
>
> predict(m1, newdata=data.frame(fuel=c("F1","F2","F3")),
interval="confidence", level=.95)
    fit      lwr      upr
1 60.50 42.81389 78.18611
2 63.25 45.56389 80.93611
3 54.25 36.56389 71.93611
```

Interpretation:

- **F1**: We are 95% sure that the mean engine performance score for all observations with Fuel 1 falls between 42.8 and 78.2.
- **F2**: We are 95% sure that the mean engine performance score for all observations with Fuel 2 falls between 45.6 and 80.9.

Prediction interval for $\hat{Y}_h$

We can create the confidence/prediction intervals for Y_h for all possible X values.

```
> # Prediction interval for Y-hat
>
> predict(m1, newdata=data.frame(fuel=c("F1","F2","F3")),
interval="prediction", level=.95)
    fit      lwr       upr
1 60.50 20.95267 100.04733
2 63.25 23.70267 102.79733
3 54.25 14.70267  93.79733
```

Interpretation:

- **F1**: With a 95% confidence, we predict that the engine performance score for the next observation with Fuel 1 falls between 21 and 100.
- **F2**: With a 95% confidence, we predict that the engine performance score for the next observation with Fuel 2 falls between 24 and 103.

Outline

1. One categorical predictor
2. Two categorical predictors without interaction
3. Two categorical predictors with interaction
4. One categorical predictor and one numeric predictor

Now we add a second predictor in the model: engine brand.

The dataset includes observations from two brands: B1 and B2.

Table 5.8 Performance data for combinations of fuel type and diesel engine brand

		Brand	
		B_1	B_2
FUEL TYPE	F_1	65 73 68	36
	F_2	78 82	50 43
	F_3	48 46	61 62

Data format in R:

```
   performance fuel brand
1           65   F1    B1
2           73   F1    B1
3           68   F1    B1
4           78   F2    B1
5           82   F2    B1
6           48   F3    B1
7           46   F3    B1
8           36   F1    B2
9           50   F2    B2
10          43   F2    B2
11          61   F3    B2
12          62   F3    B2
```

The hypothesized model:

$$E\{Y\} = \beta_0 + \underbrace{\beta_1 X_1 + \beta_2 X_2}_{\text{main effect of fuel type}} + \underbrace{\beta_3 X_3}_{\text{main effect of engine brand}}$$

where

$$X_1 = \begin{cases} 1 & \text{if F1} \\ 0 & \text{otherwise} \end{cases} \quad X_2 = \begin{cases} 1 & \text{if F2} \\ 0 & \text{otherwise} \end{cases} \quad \text{(F3 is base level for fuel type)}$$

$$X_3 = \begin{cases} 1 & \text{if B2} \\ 0 & \text{otherwise} \end{cases} \quad \text{(B1 is base level for engine brand)}$$

$$E\{\text{Performance}\} = \beta_0 + \beta_1(\text{fuelF1}) + \beta_2(\text{fuelF2}) + \beta_3(\text{brandB2})$$

Fuel Type	Engine Brand	X_1 fuelF1	X_2 fuelF2	X_3 brandB2	$E\{Y\}$ Expected performance

The R output for the model is:

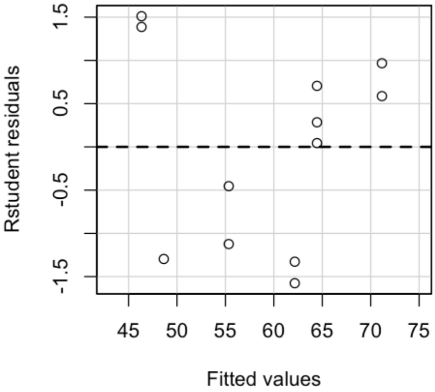
```
> m2 <- lm(performance ~ fuel + brand, data=engine)
> summary(m2)

Call:
lm(formula = performance ~ fuel + brand, data = engine)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.159 -12.415   2.046   9.119  15.659 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   62.159      8.028   7.743 5.52e-05 ***
fuelF1         2.295      9.941   0.231   0.8232    
fuelF2         9.000      9.722   0.926   0.3817    
brandB2      -15.818      8.291  -1.908   0.0928 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.75 on 8 degrees of freedom
Multiple R-squared:  0.362,	Adjusted R-squared:  0.1228 
F-statistic: 1.513 on 3 and 8 DF,  p-value: 0.2838
```

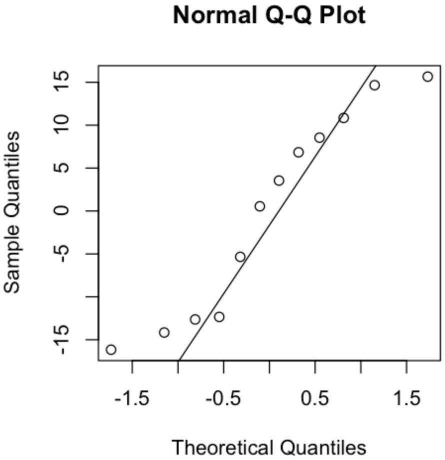
How do you interpret the coefficients of `fuelF1`, `fuelF2`, and `brandB2`?

Interpretation of model coefficients

- $b_0 = 62.159$: The expected engine performance score for the grand base level, which is Fuel 3 Brand 1, is 62.159 points.
- $b_1 = 2.295$: The expected engine performance score for Fuel 1 is 2.295 points higher than that for the base level Fuel 3, holding engine brand constant.
- $b_2 = 9$: The expected engine performance score for Fuel 2 is 9 points higher than that for the base level Fuel 3, holding engine brand constant.
- $b_3 = -15.818$: The expected engine performance score for Brand 2 is 15.818 points lower than that for the base level Brand 1, holding fuel type constant.

Diagnostic residual plots





Additional contribution of Brand to the model

```
> # Statistical difference
>
> anova(m1, m2)
Analysis of Variance Table

Model 1: performance ~ fuel
Model 2: performance ~ fuel + brand
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1      9 2200.5
2      8 1512.4  1    688.09 3.6397 0.09285 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0
>
> # Practical difference
>
> summary(m1)$r.squared
[1] 0.07178009
>
> summary(m2)$r.squared
[1] 0.3620322
```

```
> # Partial R2
> library(rsq)
> rsq.partial(objF=m2, objR=m1)
$adjustment
[1] FALSE

$variables.full
[1] "fuel"  "brand"

$variables.reduced
[1] "fuel"

$partial.rsq
[1] 0.3126975

> # Manually verify the rsq.partial function
> anova(m1)
Analysis of Variance Table

Response: performance
          Df  Sum Sq Mean Sq F value Pr(>F)
fuel       2  170.17  85.083   0.348 0.7152
Residuals  9 2200.50 244.500
> anova(m2)
Analysis of Variance Table

Response: performance
          Df  Sum Sq Mean Sq F value  Pr(>F)
fuel       2  170.17   85.08  0.4501 0.65280
brand      1  688.09  688.09  3.6397 0.09285 .
Residuals  8 1512.41  189.05
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.0
> (2200.5 - 1512.41) / 2200.5
[1] 0.3126971
```

Confidence intervals for regression coefficients

The 95% confidence interval for β_1:

$$b_1 \pm t(.975; df = 8) \cdot s\{b_1\} = 2.295 \pm 2.3 * 9.941 = (-20.6, 25.2)$$

Interpretation:

- We are 95% sure that the true population parameter β_1 falls within this interval.
- **After controlling for engine brand**, we estimate, with a 95% confidence, that the expected engine performance score for Fuel 1 falls between 20.6 points lower than Fuel 3 and 25.2 points higher than Fuel 3.

The 95% confidence interval for β_2:

$$b_2 \pm t(.975; df = 8) \cdot s\{b_2\} = 9 \pm 2.3 * 9.722 = (-13.3, 31.4)$$

Confidence intervals for regression coefficients

The 95% confidence interval for β_3:

$$b_3 \pm t(.975; df = 8) \cdot s\{b_3\} = -15.818 \pm 2.3 * 8.291 = (-34.9, 3.3)$$

Interpretation:

- We are 95% sure that the true population parameter β_3 falls within this interval.
- **After controlling for fuel type**, we estimate, with a 95% confidence, that the expected engine performance score for Brand 2 falls between -34.9 points lower than Brand 1 and 3.3 points higher than Brand 1.

Confidence interval for $E\{Y_h\}$ and Prediction interval for $\hat{Y}_h$

We can create the confidence/prediction intervals for Y_h for all possible X values. Now we have a $3 \times 2 = 6$ combinations of possible X values.

```
> # Create the data frame for the newdata argument
>
> my.newdata <- data.frame(fuel=rep(c("F1","F2","F3"), times=2),
brand=rep(c("B1","B2"), each=3))
>
> my.newdata
  fuel brand
1   F1    B1
2   F2    B1
3   F3    B1
4   F1    B2
5   F2    B2
6   F3    B2
```

Confidence interval for $E\{Y_h\}$

```
> # Confidence interval for E{Y}
>
> predict(m2, newdata=my.newdata, interval="confidence", level=.95)
       fit      lwr      upr
1 64.45455 47.89631 81.01278
2 71.15909 52.64642 89.67176
3 62.15909 43.64642 80.67176
4 48.63636 27.25978 70.01295
5 55.34091 36.82824 73.85358
6 46.34091 27.82824 64.85358
```

Interpretation:

- **F1B1**: We are 95% sure that the mean number of engine performance score for all observations with Fuel 1 and engine Brand 1 falls between 47.9 and 81.0.
- **F3B2**: We are 95% sure that the mean number of engine performance score for all observations with Fuel 2 falls between 27.8 and 64.9.

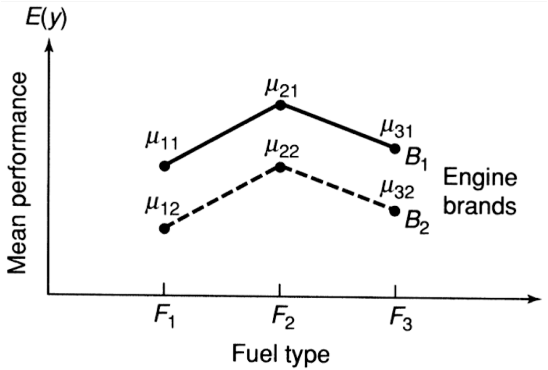
Prediction interval for $\hat{Y}_h$

```
> # Prediction interval for Y-hat
>
> predict(m2, newdata=my.newdata, interval="prediction", level=.95)
       fit       lwr       upr
1 64.45455 28.684672 100.22442
2 71.15909 34.443595 107.87459
3 62.15909 25.443595  98.87459
4 48.63636 10.396760  86.87597
5 55.34091 18.625413  92.05640
6 46.34091  9.625413  83.05640
```

Interpretation:

- **F1B1**: With a 95% confidence, we predict that the engine performance score for the next observation with Fuel 1 and engine Brand 1 falls between 28.7 and 100.2.
- **F3B2**: With a 95% confidence, we predict that the engine performance score for the next observation with Fuel 3 and engine Brand 2 falls between 9.6 and 83.

The fitted model graph



Note this graph represent an additive model *without interaction*, so the lines are parallel.

Outline

The hypothesized model

$$E\{Y\} = \beta_0 + \underbrace{\beta_1 X_1 + \beta_2 X_2}_{\text{main effect of fuel type}} + \underbrace{\beta_3 X_3}_{\text{main effect of engine brand}} + \underbrace{\beta_4 X_1 X_3 + \beta_5 X_2 X_3}_{\text{interaction effect}}$$

where

$$X_1 = \begin{cases} 1 & \text{if F1} \\ 0 & \text{otherwise} \end{cases} \quad X_2 = \begin{cases} 1 & \text{if F2} \\ 0 & \text{otherwise} \end{cases} \quad \text{(F3 is base level for fuel type)}$$

$$X_3 = \begin{cases} 1 & \text{if B2} \\ 0 & \text{otherwise} \end{cases} \quad \text{(B1 is base level for engine brand)}$$

$$E\{\text{Performance}\} = \beta_0 + \beta_1(\text{fuelF1}) + \beta_2(\text{fuelF2}) + \beta_3(\text{brandB2}) + \beta_4(\text{fuelF1})(\text{brandB2}) + \beta_5(\text{fuelF2})(\text{brandB2})$$

Fuel Type	Engine Brand	X_1 fuelF1	X_2 fuelF2	X_3 brandB2	$E\{Y\}$ Expected performance

The R output for the model is:

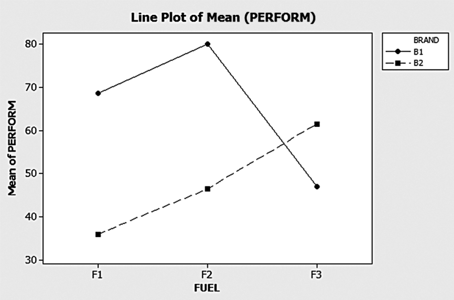
```
> m3 <- lm(performance ~ fuel + brand + fuel:brand, data=engine)
> summary(m3)

Call:
lm(formula = performance ~ fuel + brand + fuel:brand, data = engine)

Residuals:
   Min     1Q Median     3Q    Max 
-3.667 -1.250 -0.250  1.250  4.333 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      47.000      2.375  19.793 1.08e-06 ***
fuelF1           21.667      3.066   7.068 0.000402 ***
fuelF2           33.000      3.358   9.827 6.40e-05 ***
brandB2          14.500      3.358   4.318 0.004995 ** 
fuelF1:brandB2  -47.167      5.130  -9.195 9.33e-05 ***
fuelF2:brandB2  -48.000      4.749 -10.107 5.45e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.358 on 6 degrees of freedom
Multiple R-squared:  0.9715, Adjusted R-squared:  0.9477 
F-statistic: 40.84 on 5 and 6 DF,  p-value: 0.0001477
```

Can you still interpret the coefficients of `fuelF1`, `fuelF2`, and `brandB2` as in the additive model?

$$\widehat{\text{performance}} = 47 + 21.7(\text{fuelF1}) + 33(\text{fuelF2}) + 14.5(\text{brandB2}) - 47.2(\text{fuelF1})(\text{brandB2}) - 48(\text{fuelF2})(\text{brandB2})$$

1. What is the difference between the expected engine performance score of Fuel 1 and Fuel 2 given Brand 1?
2. What is the difference between the expected engine performance score of Fuel 1 and Fuel 2 given Brand 2?
3. What is the difference between the expected engine performance score of Fuel 1 and Fuel 3 given Brand 1?
4. What is the difference between the expected engine performance score of Fuel 1 and Fuel 3 given Brand 2?
5. Does the effect of fuel type differ for the two engine brands?

The fitted model graph



Note when interaction effect is present, the lines are not parallel.

In fact, if we fit an interaction model, the predicted Y value for each crossed condition equals the sample mean of the observed Y values in this condition.

Real data example

Skin cancer has become more concerning in recent years as the understanding of the effect of atmospheric ozone depletion has grown. In this example, we look at the rates of skin cancer incidents by U.S. states and age groups.

The dataset was obtained from the CDC website.

The dataset includes melanoma incidence by age group, state, and year (between 1999 and 2009).

Outline

1. One categorical predictor
2. Two categorical predictors without interaction
3. Two categorical predictors with interaction
4. One categorical predictor and one numeric predictor

Suppose the expected performance of a diesel engine is a function of two variables:

- Fuel type, at levels F1, F2, and F3.
- Engine speed, in revolutions per minute (rpm).

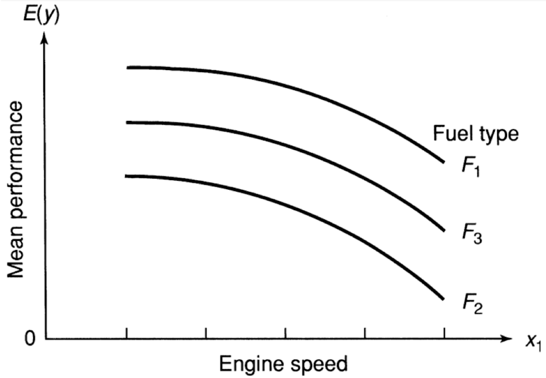
Below is an additive model involving both predictors:

$$E\{y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_3$$

where

$$X_1 = \text{Engine speed (centered)} \quad X_2 = \begin{cases} 1, & \text{if fuel F2} \\ 0, & \text{otherwise} \end{cases} \quad X_3 = \begin{cases} 1, & \text{if fuel F3} \\ 0, & \text{otherwise} \end{cases}$$

The fitted model curve looks like this:



Note because there is no interaction term in the model, the three curves are parallel to each other.