So far, we’ve only considered quantitative variables in our models.

We can integrate categorical predictors by constructing artificial variables (known as *dummy variables* or *indicator variables*).

We’ll illustrate here with a binary predictor (e.g. Male/Female).

- Pick one category as the default (say Female).
- Define $z_i = 0$ if observation $i$ is a Female, otherwise $z_i = 1$ (denotes Male).
A simple model with 1 covariate and 1 dummy variable:

\[ Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i \]

where \( Y_i \) and \( x_i \) are continuous variables, and \( \epsilon \sim iid \sim N(0, \sigma^2) \).

– note:

\[ E(Y|x) = \beta_0 + \beta_1 x \] for females

\[ E(Y|x) = (\beta_0 + \beta_2) + \beta_1 x \] for males

– Info from both the males and females were used to estimate \( \beta_1, \sigma^2 \)

– \( \beta_0 \) is interpreted as the female Y-intercept

– \( \beta_2 \) is interpreted as the difference in expected value of Y for identical x units (same x) in the two groups.
• **Example**: Apple yield and tree size.

A botany student wants to model the relationship between apple tree size (diameter) and yield (bushels). She also has information on the pruning method used on the trees (Pyramid or Flattop).

The dependent variable *Bushels* is quantitative, as is the independent variable *Diameter*, but *Pruning* is a categorical or qualitative variable.

Consider the model we previously described:

\[
Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i
\]

where \( z_i = 0 \) if observation \( i \) was a Flattop pruning, otherwise \( z_i = 1 \) (denotes Pyramid pruning).
The data:

```r
> botany.data=read.csv("botany.csv")
> attach(botany.data)

> head(botany.data)
      Diameter Bushels Pruning
1       20    10.56   Pyramid
2       14     6.14   Pyramid
3       16     6.30   Pyramid
4       13     6.38   Pyramid
5       18     8.65   Pyramid
6       17     7.02   Pyramid

> unique(Pruning)
[1] Pyramid Flattop
Levels: Flattop Pyramid

> is.factor(Pruning)
[1] TRUE

> is.numeric(Pruning)
[1] FALSE

> plot(Diameter,Bushels,pch=16)
```
There does appear to be a linear relationship between tree diameter and yield in bushels.
The simple linear regression:

```r
> lm.out=lm(Bushels ~ Diameter)
> summary(lm.out)
```

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -2.1886  | 0.75898    | -2.884  | 0.00988 **|
| Diameter       | 0.62361  | 0.05185    | 12.028  | 4.86e-10 ***|

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.133 on 18 degrees of freedom
Multiple R-Squared: 0.8894, Adjusted R-squared: 0.8832
F-statistic: 144.7 on 1 and 18 DF, p-value: 4.86e-10

```r
> abline(lm.out)
```
Does Pruning Method also make a significant impact on yield?

First, we’ll create the dummy variable:

```r
## Allocate space for the new vector:
> pruning.dummy=rep(0,nrow(botany.data))
> pruning.dummy
[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

## If Pruning equals "Pyramid", code it as 1.
> pruning.dummy[Pruning=="Pyramid"]=1
> pruning.dummy
[1] 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0

> data.frame(Pruning,pruning.dummy)
  Pruning pruning.dummy
 1   Pyramid           1
 2   Pyramid           1
 3   Pyramid           1
 4   Pyramid           1
 5   Pyramid           1
 6   Pyramid           1
 7   Pyramid           1
 8   Pyramid           1
 9   Pyramid           1
10  Pyramid           1
11  Pyramid           1
12  Flattop           0
```
Fit a model with both Diameter and Pruning (as a dummy variable).

> lm.out.2=lm(Bushels ~ Diameter + pruning.dummy)
> summary(lm.out.2)

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -1.90616 | 0.75416    | -2.528  | 0.0217 * |
| Diameter       | 0.63352  | 0.05038    | 12.574  | 4.91e-10 *** |
| pruning.dummy  | -0.76259 | 0.49468    | -1.542  | 0.1416   |

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.092 on 17 degrees of freedom
Multiple R-Squared: 0.9029, Adjusted R-squared: 0.8915
F-statistic: 79.06 on 2 and 17 DF,  p-value: 2.458e-09
Pruning is not a significant regressor given Diameter is already accounted for.

But Diameter is a significant regressor given Pruning is already accounted for.

Create plot distinguishing the Pruning Method:

```r
plot(Diameter,Bushels,type="n") ## Don’t plot the points
points(Diameter[1:11],Bushels[1:11],pch=1,col=1)
points(Diameter[12:20],Bushels[12:20],pch=9,col=4)
legend(8,10,c("Pyramid","Flattop"),col=c(1,4),pch=c(1,9))
```
Plot fitted models for the two Pruning Methods:

## The estimated parameters:

```r
> lm.out.2$coefficients

     (Intercept)  Diameter pruning.dummy
-1.9061600    0.6335207  -0.7625916
```

## Slope is the same for both (it’s the second parameter):

```r
> slope=lm.out.2$coefficients[2]
> slope

Diameter
0.6335207
```

## Intercept is different:

```r
> intercept.Flattop=lm.out.2$coefficients[1]
> intercept.Flattop

(Intercept)
-1.90616

> intercept.Pyramid

(Intercept)
-2.668752
```

## Add the lines on the previous plot:

```r
> abline(intercept.Flattop,slope,col=4)
> abline(intercept.Pyramid,slope,col=1)
```
This shows the fitted lines, but the difference between them ($\beta_2$) was not found to be significantly different than 0, given Diameter is in the model (so there’s really only one line).

$H_0 : \beta_2 = 0$ was not rejected.
This model that we fit with 1 covariate and 1 dummy variable:

\[ Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i \]

where \( Y_i \) and \( x_i \) are continuous variables, 
\( z_i = 0 \) if observation \( i \) is a Flattop method, 
\( z_i = 1 \) if observation \( i \) is a Pyramid method, 
and \( \epsilon \sim iid N(0,\sigma^2) \), 

assumed the same Diameter effect (slope) for both groups, but allowed for different intercepts.
Inclusion of Interaction

Returning to the earlier model with a male/female binary variable.

A slightly more complicated model:

\[ Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \epsilon_i \]

• note:

\[ E(Y|x) = \beta_0 + \beta_1 x \] for females

\[ E(Y|x) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x \] for males

• The \( xz \) interaction term allows for a different slope for each group

• \( xz \) may be called a slope dummy variable

• This model allows for two separate regression lines for each group
• What is the benefit to bringing the data for the two groups together?
  – We can use familiar tests to compare groups (e.g. full vs. reduced in partial $F$-test).
  – If there is constant variance, we’ll have a better estimate for $\sigma^2$ by pooling the data (provided there is a common $\sigma^2$.)
  – More degrees of freedom for error.

• To test if the lines are parallel, test if the interaction term is significant. $H_0 : \beta_3 = 0$.

• If there’s no significant interaction, we can test if there’s two separate parallel lines for the two groups, or if one is sufficient to describe the data. $H_0 : \beta_2 = 0$.

• If there is significant interaction, the effect of covariate $x$ on the response is different for different values of $z$. 
Let’s fit the interaction model to the tree data.

In R, the interaction term between \textit{Diameter} and \textit{pruning.dummy} is shown as \textit{Diameter : pruning.dummy}.

\begin{verbatim}
> lm.out.3=lm(Bushels ~ Diameter + pruning.dummy +
                 Diameter:pruning.dummy)
> summary(lm.out.3)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.00130   0.98989  -2.022  0.0603
Diameter       0.64078   0.06988   9.170 9.05e-08 ***
pruning.dummy -0.53930   1.52761  -0.353  0.7287
Diameter:pruning.dummy -0.01618   0.10434  -0.155  0.8787
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.124 on 16 degrees of freedom
Multiple R-Squared: 0.9031, Adjusted R-squared: 0.8849
F-statistic: 49.69 on 3 and 16 DF, p-value: 2.487e-08
\end{verbatim}

\begin{verbatim}
> lm.out.3$coefficients
            Diameter:pruning.dummy
(Intercept) -2.00129614 0.64077682 -0.53929972 -0.01617838

\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3
\end{verbatim}
## Model allows for different intercepts:

```r
> intercept.Flattop = lm.out.3$coefficients[1]
> intercept.Flattop
(Intercept)
  -2.00129614
```

```r
> intercept.Pyramid = lm.out.3$coefficients[1] +
                      lm.out.2$coefficients[3]
> intercept.Pyramid
(Intercept)
  -2.540596
```

## Model allows for different slopes:

```r
> slope.Flattop = lm.out.3$coefficients[2]
> slope.Flattop
  Diameter
  0.6407768
```

```r
> slope.Pyramid = lm.out.3$coefficients[2] +
                   lm.out.3$coefficients[4]
> slope.Pyramid
  Diameter
  0.6245984
```
The separate fitted line for each group:

```r
c# Don't plot the points
plot(Diameter, Bushels, type="n")
points(Diameter[1:11], Bushels[1:11], pch=1, col=1)
points(Diameter[12:20], Bushels[12:20], pch=9, col=4)
legend(8, 10, c("Pyramid", "Flattop"), col=c(1, 4), pch=c(1, 9))
abline(intercept.Flattop, slope(Flattop), col=4)
abline(intercept.Pyramid, slope.Pyramid, col=1)
```

You can’t see much difference, but the fitted line for Flattop has a slightly steeper slope.
Another statement that fits the same model.

In \texttt{R}, if you use the single term \texttt{Diameter \ast pruning.dummy} in your stated model, then you’re actually including the main effects and the interactions (so this is shorthand for the full model).

\begin{verbatim}
> lm.out.4=lm(Bushels ~ Diameter*pruning.dummy)
> summary(lm.out.4)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.00130   0.98989  -2.022  0.0603 
Diameter      0.64078   0.06988   9.170 9.05e-08 ***
pruning.dummy -0.53930   1.52761  -0.353  0.7287   
Diameter:pruning.dummy -0.01618  0.10434  -0.155  0.8787

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.124 on 16 degrees of freedom
Multiple R-Squared: 0.9031, Adjusted R-squared: 0.8849
F-statistic: 49.69 on 3 and 16 DF, p-value: 2.487e-08

> lm.out.4$coefficients

                     Diameter:
(Intercept)     Diameter pruning.dummy pruning.dummy
   -2.00129614  0.64077682  -0.53929972  -0.01617838
\end{verbatim}
• When interpreting models with dummy variables, it’s important to keep in mind the coding scheme that was used (which variable is associated with 1 and which with 0, for example).

• Numerous covariates can be used in conjunction with dummy variables, though we only showed 1 covariate in this example.

• A model with no interaction terms is said to be additive.

• If an interaction term between two variables is included in the model, then the ‘main effects’ for those variables should also be included. So, if \(xz\) is in the model, so should you include \(x\) and \(z\).
• If there is significant interaction, we do not consider the tests for ‘main effects’. If there is NOT significant interaction, we CAN consider the tests for ‘main effects’.

We look at the interaction significance first, and then proceed from there.
Consider the categorical variable of *Rank* for a professor, which has the categories of Assistant, Associate, and Full.

We can represent this categorical variable with 2 dummy variables:

<table>
<thead>
<tr>
<th>Category</th>
<th>$D_1$</th>
<th>$D_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assistant</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Associate</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Full</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Suppose we wish to predict *Salary* based on *Rank* and the quantitative variable *Grants*, which tells how much grant money a professor has brought in.
The additive model:

\[ Y_i = \beta_0 + \beta_1 x_i + \beta_2 D_{1i} + \beta_3 D_{2i} + \epsilon_i \]

where \( x_i \) is the amount of grant money in dollars for observation \( i \).

The differing models for each Rank:

Assistant: \[ Y_i = (\beta_0 + \beta_2) + \beta_1 x_i + \epsilon_i \]

Associate: \[ Y_i = (\beta_0 + \beta_3) + \beta_1 x_i + \epsilon_i \]

Full: \[ Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \]

There were no interactions terms included, so Grants was assumed to have the same effect on Salary for all Rank categories. This model gives three parallel lines.

The ‘Full’ professor model serves as the baseline model, but this is arbitrarily chosen.
Sometimes there is a natural *baseline* choice.

This choice can make desired comparisons easier.

For example, $H_0 : \beta_3 = 0$, tests if the model for Associate is different than the model for Full.

And, $H_0 : \beta_2 = 0$, tests if the model for Assistant is different than the model for Full.

To compare Assistant to Associate, use $H_0 : \beta_2 - \beta_3 = 0$.

To test if *Rank* has any effect on *Salary* after accounting for *Grants*, use $H_0 : \beta_2 = \beta_3 = 0$. This can be done with Partial F-test.
In general, it takes $m - 1$ dummy variables to represent a categorical variable with $m$ levels.

When interactions are included, it makes interpretation a bit more complex, but interactions are often needed to appropriately model the data.