## Interaction Effects in Regression

Interactions –> The effect of one variable is dependent on the value of another.

Libraries

library(tidyverse) #tidyverse set of packages and functions

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.0 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.1 ✔ tibble 3.1.8  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library(tidymodels)

## ── Attaching packages ────────────────────────────────────── tidymodels 1.0.0 ──  
## ✔ broom 1.0.3 ✔ rsample 1.1.1  
## ✔ dials 1.1.0 ✔ tune 1.0.1  
## ✔ infer 1.0.4 ✔ workflows 1.1.2  
## ✔ modeldata 1.0.1 ✔ workflowsets 1.0.0  
## ✔ parsnip 1.0.3 ✔ yardstick 1.1.0  
## ✔ recipes 1.0.4   
## ── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
## ✖ scales::discard() masks purrr::discard()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ recipes::fixed() masks stringr::fixed()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ yardstick::spec() masks readr::spec()  
## ✖ recipes::step() masks stats::step()  
## • Search for functions across packages at https://www.tidymodels.org/find/

library(glmnet) #for Lasso, ridge, and elastic net models

## Loading required package: Matrix  
##   
## Attaching package: 'Matrix'  
##   
## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack  
##   
## Loaded glmnet 4.1-6

library(GGally) #create ggcorr and ggpairs plots

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(ggcorrplot) #create an alternative to ggcorr plots  
library(MASS) #access to forward and backward selection algorithms

##   
## Attaching package: 'MASS'  
##   
## The following object is masked from 'package:dplyr':  
##   
## select

library(leaps) #best subset selection  
library(lmtest) #for the dw test

## Loading required package: zoo  
##   
## Attaching package: 'zoo'  
##   
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library(splines) #for nonlinear fitting

Read in data and examine structure

birth = read\_csv("birth\_smoker.csv")

## Rows: 32 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): smoker  
## dbl (2): weight, gestation  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

summary(birth)

## weight gestation smoker   
## Min. :2420 Min. :34.00 Length:32   
## 1st Qu.:2737 1st Qu.:37.00 Class :character   
## Median :3068 Median :39.00 Mode :character   
## Mean :3020 Mean :38.66   
## 3rd Qu.:3306 3rd Qu.:40.25   
## Max. :3530 Max. :42.00

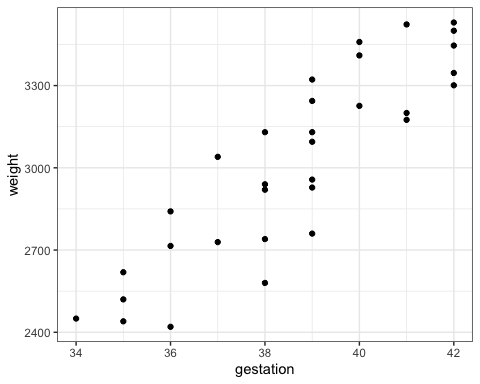
str(birth)

## spc\_tbl\_ [32 × 3] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ weight : num [1:32] 2940 3130 2420 2450 2760 ...  
## $ gestation: num [1:32] 38 38 36 34 39 35 40 42 37 40 ...  
## $ smoker : chr [1:32] "yes" "no" "yes" "no" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. weight = col\_double(),  
## .. gestation = col\_double(),  
## .. smoker = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

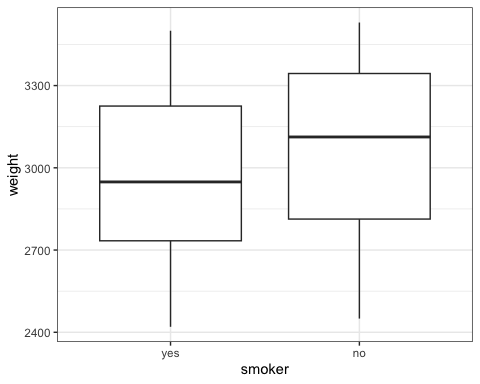
birth = birth %>% mutate(smoker = as\_factor(smoker))

Plots

ggplot(birth,aes(x=gestation,y=weight)) + geom\_point() + theme\_bw()

 Clearly, gestation length affects weight. Longer gestation –> higher weight. How about smoking?

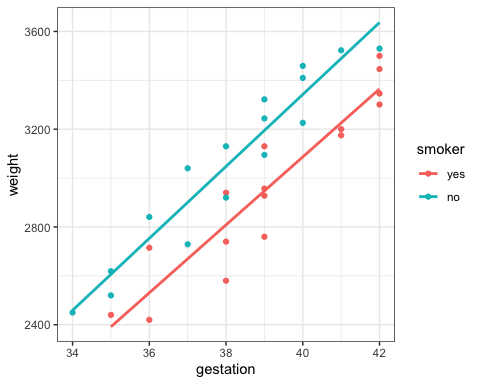
ggplot(birth,aes(x=smoker,y=weight)) + geom\_boxplot() + theme\_bw()

 Clearly, smoking leads to lower birth weights.

Let’s look at one more plot to see how all of these variables interact.

ggplot(birth,aes(x = gestation,y = weight, color = smoker)) +   
 geom\_point() + geom\_smooth(method='lm', se = FALSE)+ theme\_bw()

## `geom\_smooth()` using formula = 'y ~ x'



We have a couple of ways that we can approach this. Let’s look at a model with “gestation” and “smoker” to predict “weight”.

birth\_recipe = recipe(weight ~ gestation + smoker, birth) %>%  
 step\_dummy(all\_nominal())  
  
lm\_model = #give the model type a name   
 linear\_reg() %>% #specify that we are doing linear regression  
 set\_engine("lm") #specify the specify type of linear tool we want to use   
  
lm\_wflow =   
 workflow() %>%   
 add\_model(lm\_model) %>%   
 add\_recipe(birth\_recipe)  
  
lm\_fit = fit(lm\_wflow, birth)

summary(lm\_fit$fit$fit$fit)

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -223.693 -92.063 -9.365 79.663 197.507   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2634.117 358.872 -7.340 4.37e-08 \*\*\*  
## gestation 143.100 9.128 15.677 1.07e-15 \*\*\*  
## smoker\_no 244.544 41.982 5.825 2.58e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 115.5 on 29 degrees of freedom  
## Multiple R-squared: 0.8964, Adjusted R-squared: 0.8892   
## F-statistic: 125.4 on 2 and 29 DF, p-value: 5.289e-15

Both gestation and smoker are significant. Being a smoker has a negative impact on weight.

From our plot above it looks like there might be a slight difference in the slopes of the regression lines by group. Let’s add an interaction effect to the model and see what we get.

birth\_recipe = recipe(weight ~ gestation + smoker, birth) %>%  
 step\_dummy(all\_nominal()) %>%  
 step\_interact(~gestation:starts\_with("smoker\_"))  
  
lm\_model = #give the model type a name   
 linear\_reg() %>% #specify that we are doing linear regression  
 set\_engine("lm") #specify the specify type of linear tool we want to use   
  
lm\_wflow =   
 workflow() %>%   
 add\_model(lm\_model) %>%   
 add\_recipe(birth\_recipe)  
  
lm\_fit2 = fit(lm\_wflow, birth)

summary(lm\_fit2$fit$fit$fit)

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -228.528 -89.560 0.273 83.629 184.529   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2474.564 512.786 -4.826 4.47e-05 \*\*\*  
## gestation 139.029 13.064 10.642 2.39e-11 \*\*\*  
## smoker\_no -71.574 716.950 -0.100 0.921   
## gestation\_x\_smoker\_no 8.178 18.515 0.442 0.662   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 117.2 on 28 degrees of freedom  
## Multiple R-squared: 0.8971, Adjusted R-squared: 0.8861   
## F-statistic: 81.37 on 3 and 28 DF, p-value: 6.144e-14

No apparent interaction (i.e., the slope of the regression lines by smoker group are not significantly different).