Clase Virtual 2022-10-12

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

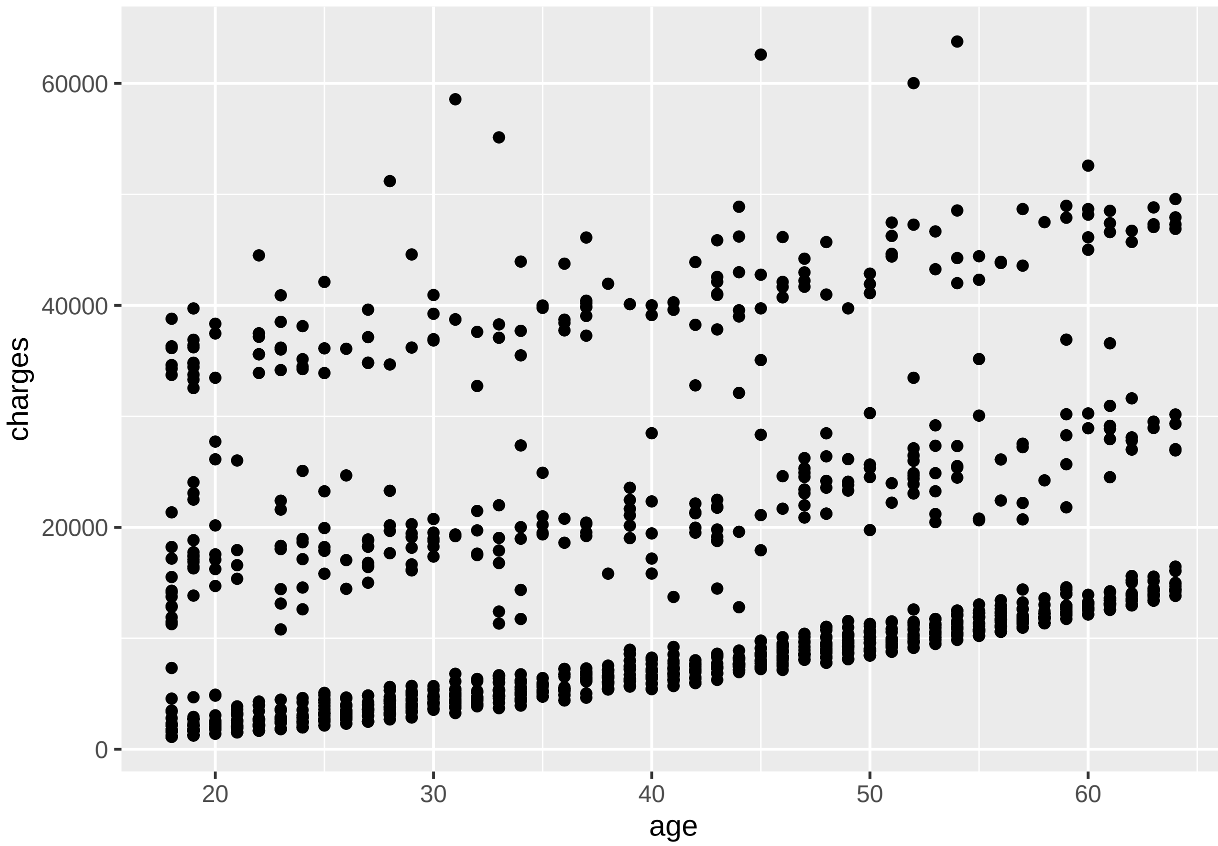
```
library(readxl)
```

Vamos a estudiar la relación entre el costo del seguro médico con la edad de las personas. Usamos el dataset **insurance**,

```
df <- readxl::read_xlsx('../datasets/insurance.xlsx') %>%
  rename(bmi = names(df)[3])
```

Plot de base para el análisis

```
# Plot de base
p0 <- ggplot(df, aes(x=age, y=charges)) + geom_point()
p0
```

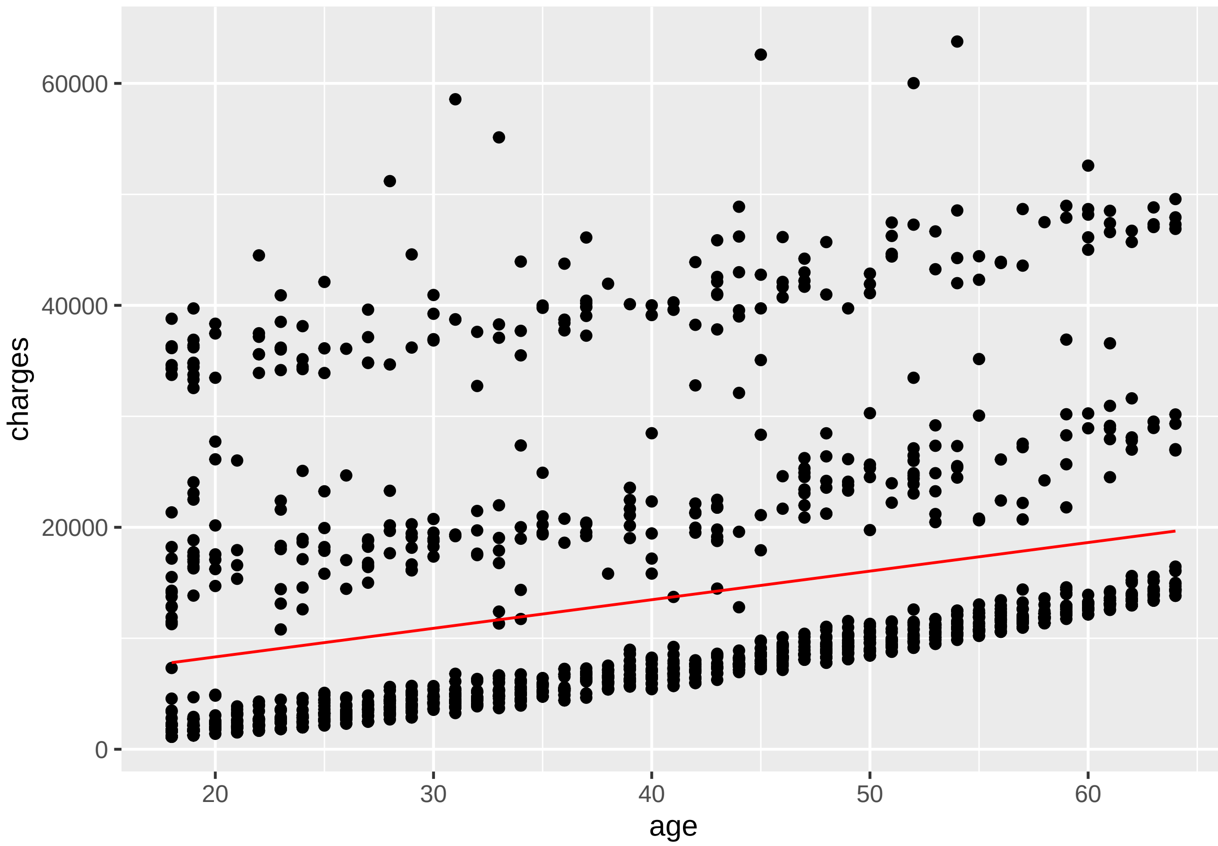


Preparamos la grilla para hacer los modelos (al menos los más simples)

```
grid <- modelr::data_grid(df, age)
```

Modelo 1

```
# Ajusto modelo lineal
mod1 <- lm(charges ~ age, data=df)

# Agrego modelo
grid <- modelr::add_predictions(grid, mod1)

# Agrego modelo
p <- p0 + geom_line(data=grid, mapping=aes(y=pred), color='red')
p
```




```
summary(mod1)
```

```
##
## Call:
## lm(formula = charges ~ age, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8059  -6671  -5939   5440  47829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3165.9      937.1   3.378 0.000751 ***
## age            257.7       22.5  11.453  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11560 on 1336 degrees of freedom
## Multiple R-squared:  0.08941,    Adjusted R-squared:  0.08872
## F-statistic: 131.2 on 1 and 1336 DF,  p-value: < 2.2e-16
```

En este ejercicio, nos interesa particularmente el valor del parámetro que acompaña a `age`, porque queremos estudiar exactamente esto; pero no queremos confundirnos por efectos de otras variables.

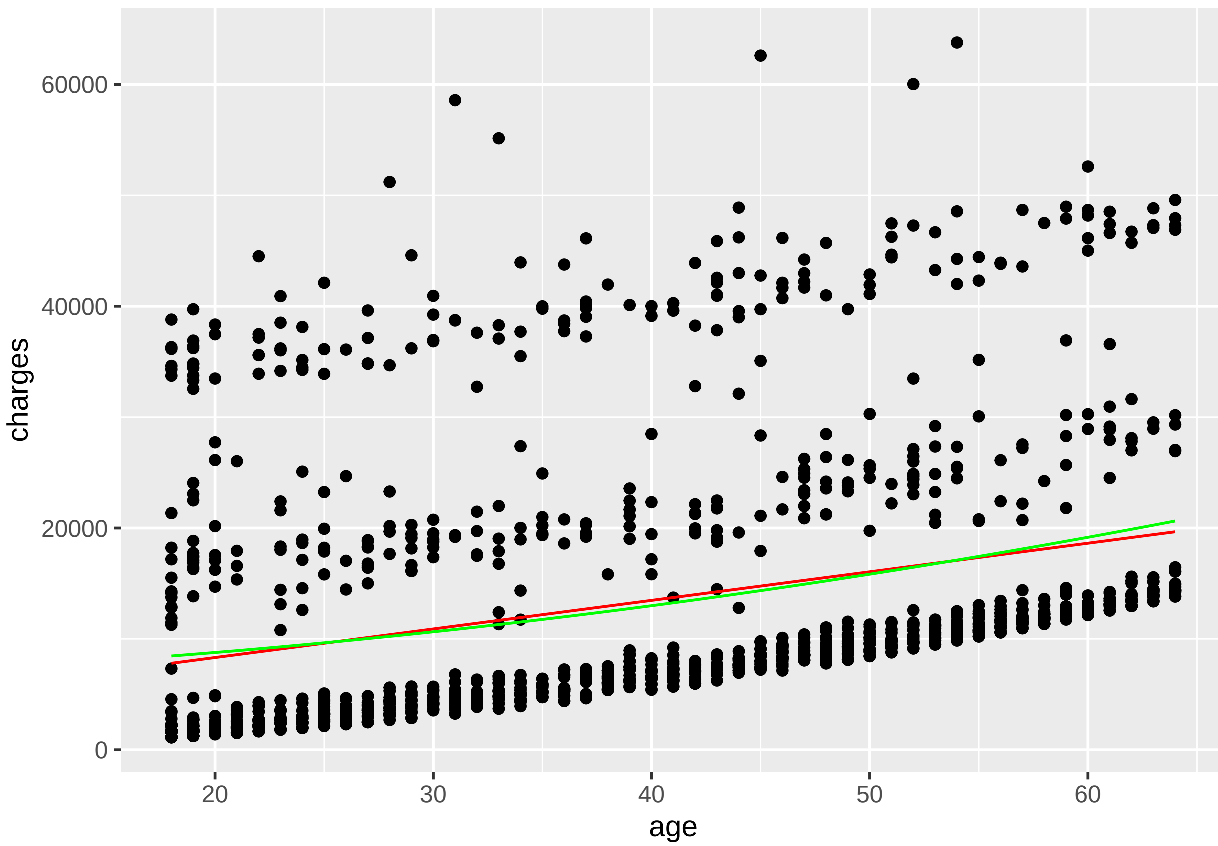
Registremos entonces el valor, 257.7 ± 22.5.

Para estudiar la calidad del modelo podemos ver el desvío de los residuos (Residual standard error), pero también el valor del coeficiente de determinación (R^2, R-squared).

Veamos ahora qué pasa si agregamos un término cuadrático.

Modelo 2

```
# Modelo 2
mod2 <- lm(charges ~ age + I(age^2), data=df)

# Agrego modelo
grid <- modelr::add_predictions(grid, mod2)

# Agrego modelo
p <- p + geom_line(data=grid, mapping=aes(y=pred), color='green')
p
```




```
summary(mod2)
```

```
##
## Call:
## lm(formula = charges ~ age + I(age^2), data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7594  -6640  -5943   5334  48240
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 6508.553   2699.359   2.411    0.016 *
## age           64.573    148.001   0.436    0.663
## I(age^2)       2.439      1.847   1.320    0.187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11560 on 1335 degrees of freedom
## Multiple R-squared:  0.09059,    Adjusted R-squared:  0.08923
## F-statistic:  66.5 on 2 and 1335 DF,  p-value: < 2.2e-16
```

Podemos ver que los errores de los parámetros se vuelven enormes.

De ahora en más, segumos con el modelo lineal en edad.

Modelo con otras variables

```
mod3 <- lm(charges ~ age + smoker, data=df)
```

Escribamos la fórmula para este modelo

$$\text{charges} = w_0 + w_1 * \text{age} + w_2 * \text{smoker}$$

Pero ¿qué significa $w2 * smoker$ cuando smoker es categórica?

En realidad, el modelo son dos curvas:

$$\text{charges} = \begin{cases} w_0 + w_1 * \text{age} & \text{si no sos fumador} \\ w_0' + w_1 * \text{age} & \text{si sos fumador} \end{cases}$$

La única diferencia está en la ordenada al origen, entre w_0 y w_0'.

Pero si vemos la forma en la que está parametrizado el modelo. . . .

```
summary(mod3)
```

```
##
## Call:
## lm(formula = charges ~ age + smoker, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16088.1  -2046.8  -1336.4   -212.7  28760.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2391.63     528.30  -4.527 6.52e-06 ***
## age           274.87      12.46  22.069  < 2e-16 ***
## smokeryes   23855.30     433.49  55.031  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6397 on 1335 degrees of freedom
## Multiple R-squared:  0.7214, Adjusted R-squared:  0.721
## F-statistic:  1728 on 2 and 1335 DF,  p-value: < 2.2e-16
```

encontramos que el parámetro que se ajusta es la *diferencia* entre las ordeandas al origen. Es decir, que la estructura del modelo es:

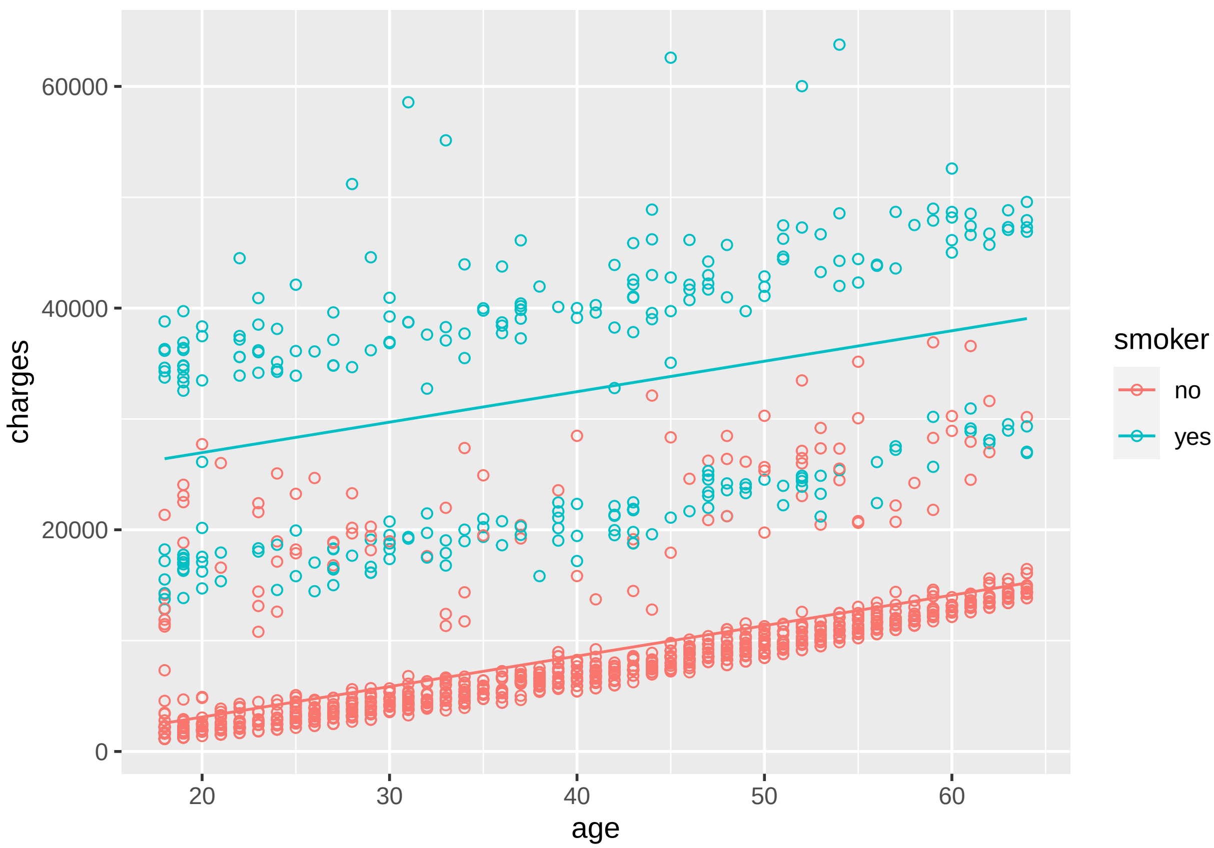
$$\text{charges} = \begin{cases} w_0 + w_1 * \text{age} & \text{si no sos fumador} \\ (w_0 + w_2) + w_1 * \text{age} & \text{si sos fumador} \end{cases}$$

y podemos pensar en el parámetro w_2 como el recargo en el costo del seguro para los fumadores (que aparece como `smokeryes` arriba; y que vale más de $ 23,500 !).

Comparemos la dispersión de los residuos (Residual standard error). Vemos una enorme mejoría en ambos valores (RSE más pequeño y R^2 más alto).

Ahora hagamos el gráfico:

```
# Agrego modelo
grid <- modelr::data_grid(df, age, smoker)

# Agrego modelo
grid <- modelr::add_predictions(grid, mod3)

# Agrego modelo
p <- ggplot(df, aes(x=age, y=charges, color=smoker)) + geom_point(shape=1) +
  geom_line(data=grid, mapping=aes(y=pred, color=smoker))
p
```



Podemos usar la vieja y querida función `plot`, o ver los residuos en función de las predicciones

```
# plot(mod3)
```

```r
df_res <- modelr::add_residuals(df, mod3) %>%
    modelr::add_predictions(mod3)

# Grafico residuos
ggplot(df_res, aes(x=pred, y=resid)) + geom_point(aes(color=smoker), shape=1) +
  geom_hline(yintercept = 0, size=0.25, linetype='dashed') +
  geom_smooth(aes(x=pred, y=resid), method='loess', size=0.5, se = F, color = 'red')
```

```
## `geom_smooth()` using formula 'y ~ x'
```

Este gráfico nos revela que el modelo todavía no capta del todo bien a las personas que pagan mucho seguro. Como se puede ver en el plot de arriba, estos son los fumadores. Tampoco repdoruce un grupo pequeño pero significativo de no fumadores que pagan más de lo predicho.

Pensemos entonces en usar también el BMI

Modelo con dos variables extras

```r
# Hack para que compile el Knit: repetir esta línea
df <- readxl::read_xlsx('../datasets/insurance.xlsx') %>%
  rename(bmi = names(df)[3])

df <- mutate(df, bmi30 = (bmi > 30))

mod4 <- lm(charges ~ age +  smoker + bmi30, data=df)
```

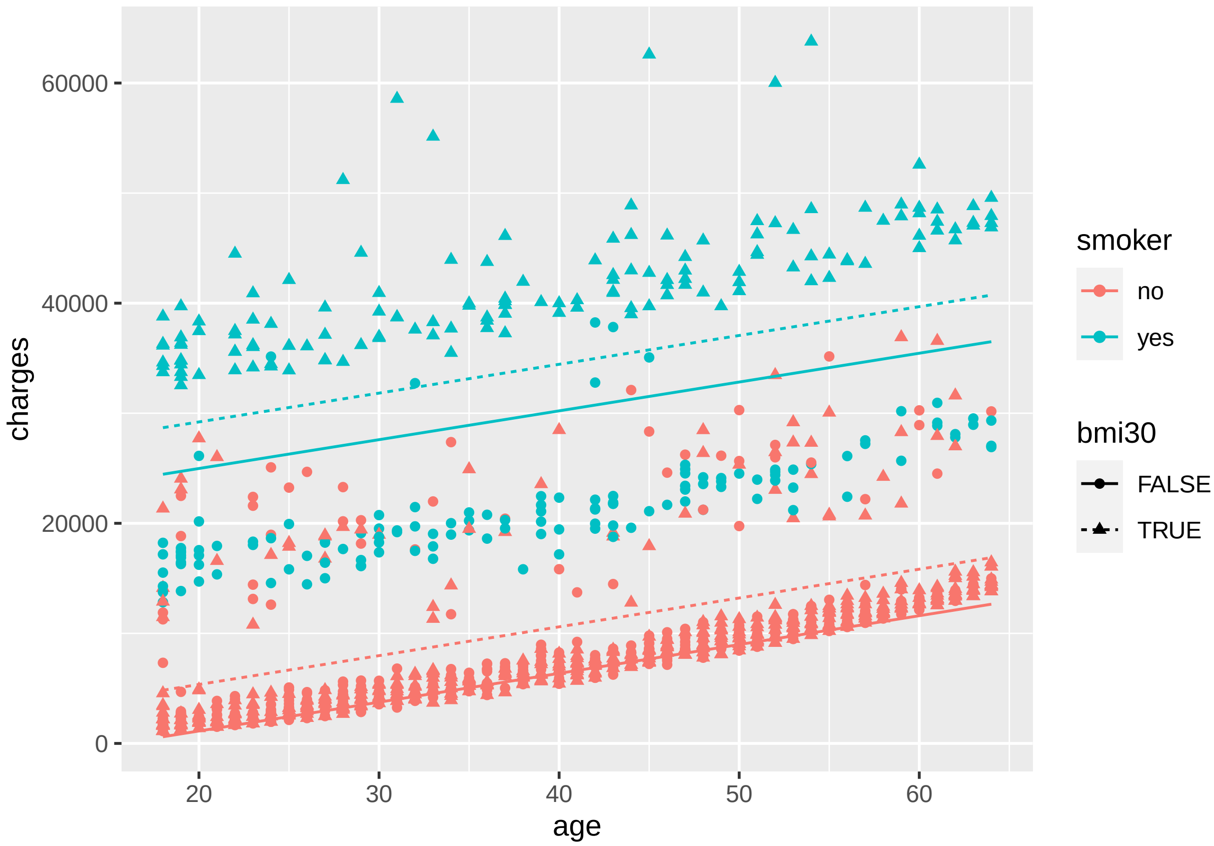
Ahora el modelo son **cuatro curvas**. Pero veamos qué restricciones tenemos:

```
summary(mod4)
```

```
##
## Call:
## lm(formula = charges ~ age + smoker + bmi30, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13872.2  -3654.4   -200.8   1477.7  26838.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4107.59     516.91  -7.946 4.05e-15 ***
## age           261.82      11.81  22.170  < 2e-16 ***
## smokeryes   23851.17     409.48  58.248  < 2e-16 ***
## bmi30TRUE    4229.12     332.11  12.734  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6042 on 1334 degrees of freedom
## Multiple R-squared:  0.7516, Adjusted R-squared:  0.751
## F-statistic:  1345 on 3 and 1334 DF,  p-value: < 2.2e-16
```

Encontramos que el costo extra por BMI alto y el costo extra por ser fumadores son independientes. La penalidad por BMI alto **y** ser fumador solo puede ser la suma de ambos.

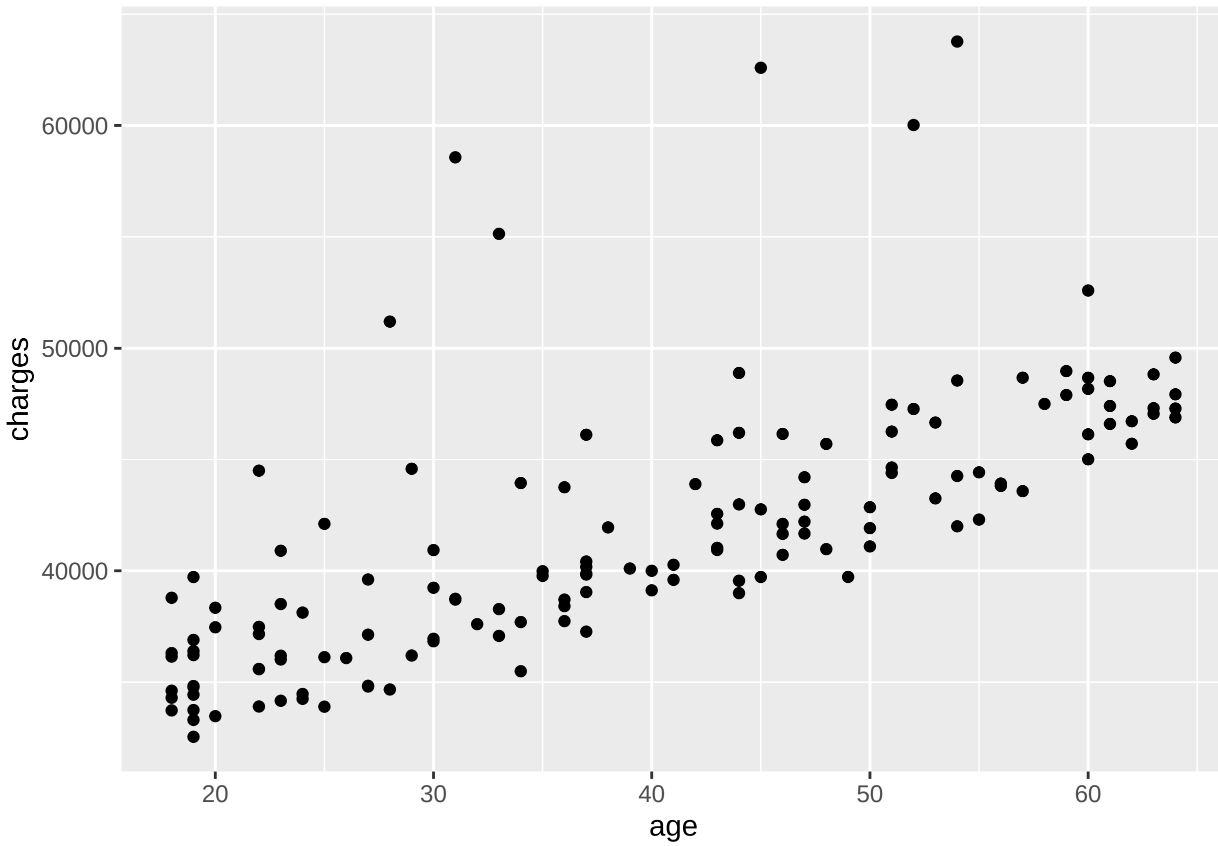
```
# Agrego modelo
grid <- modelr::data_grid(df, age, smoker, bmi30)

# Agrego modelo
grid <- modelr::add_predictions(grid, mod4)

# Agrego modelo
p <- ggplot(df, aes(x=age, y=charges, color=smoker, shape=bmi30)) + geom_point() +
  geom_line(data=grid, mapping=aes(y=pred, color=smoker, linetype=bmi30))
p
```



Vemos que el modelo no parece funcionar. ¿Por qué? Estamos haciendo algo mal?

Veamos solo los datos de un grupo

```
dd <- filter(df, (smoker=='yes') & (bmi30))
ggplot(dd, aes(x=age, y=charges)) + geom_point()
```



Los datos de fumadores con alto BMI están bien localizados en la parte alta del diagrama.

Entonces, ¿por qué la línea de puntos celeste no los encuentra y pasa por abajo? La razón es que el recargo por BMI alto es la misma, tanto para fumadores como para no fumadores. No puede ser de otra forma en un modelo **sin interacciones**. Por lo tanto, si queremos aumentar la diferencia entre la curva celeste sólida y la punteada, necesitamos aumentar también la diferencia entre las curvas rosas, y evidentemente esto no mejor el MSE, y el modelo no lo prefiere.

Si queremos más flexibilidad y recargo por BMI alto que dependa del estado de fumador del sujeto, necesitamos usar un **modelo con interacciones**.

Modelo con interacciones

```
mod5 <- lm(charges ~ age + smoker * bmi30, data=df)

summary(mod5)
```

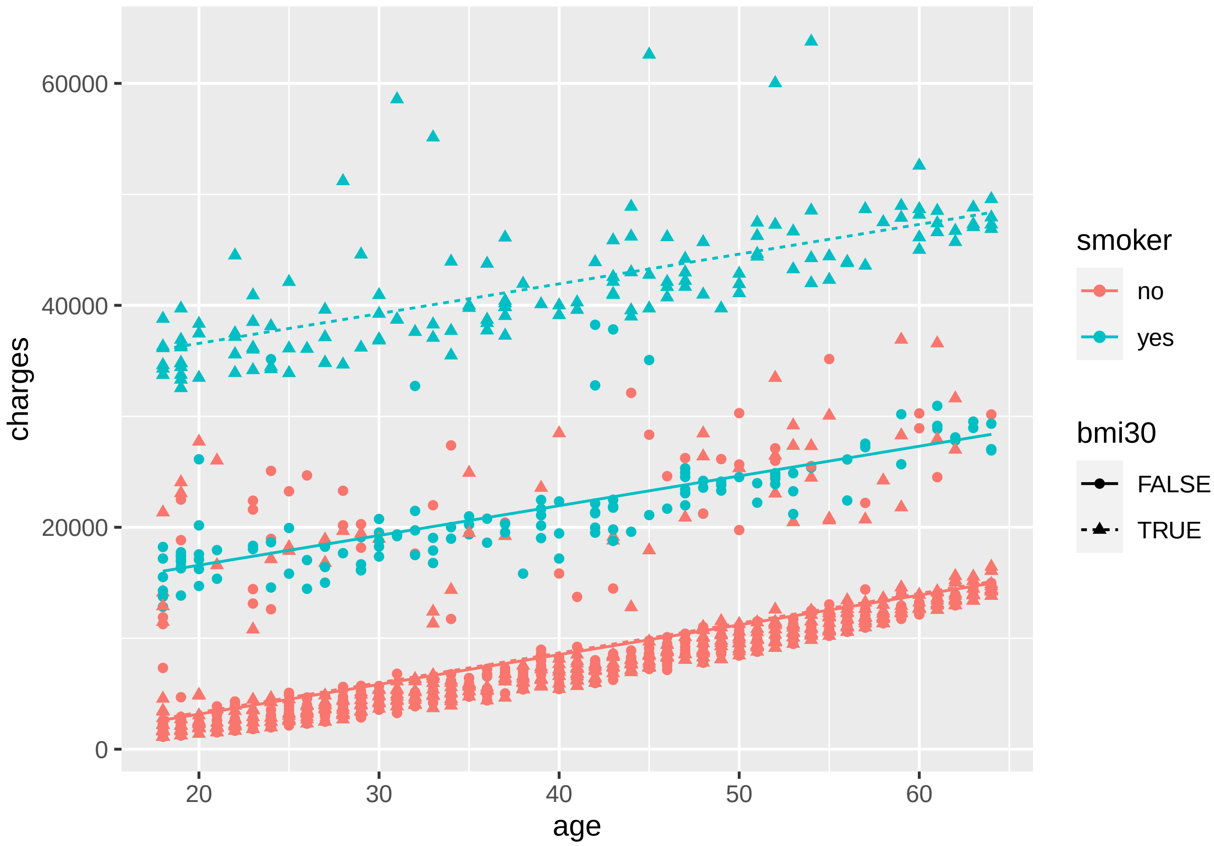
```
##
## Call:
## lm(formula = charges ~ age + smoker * bmi30, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5586  -1944  -1288   -418  24415
##
## Coefficients:
```

```
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         -2200.874    391.831  -5.617 2.36e-08 ***
## age                   268.023      8.852  30.278  < 2e-16 ***
## smokeryes           13422.339    445.503  30.128  < 2e-16 ***
## bmi30TRUE             149.777    279.104   0.537    0.592    
## smokeryes:bmi30TRUE 19840.756    614.472  32.289  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4528 on 1333 degrees of freedom
## Multiple R-squared:  0.8606, Adjusted R-squared:  0.8602 
## F-statistic:  2058 on 4 and 1333 DF,  p-value: < 2.2e-16
```

Vemos que ahora el término que representa el exceso de costo **solo por BMI alto** no es significativo y podríamos sacarlo del modelo.

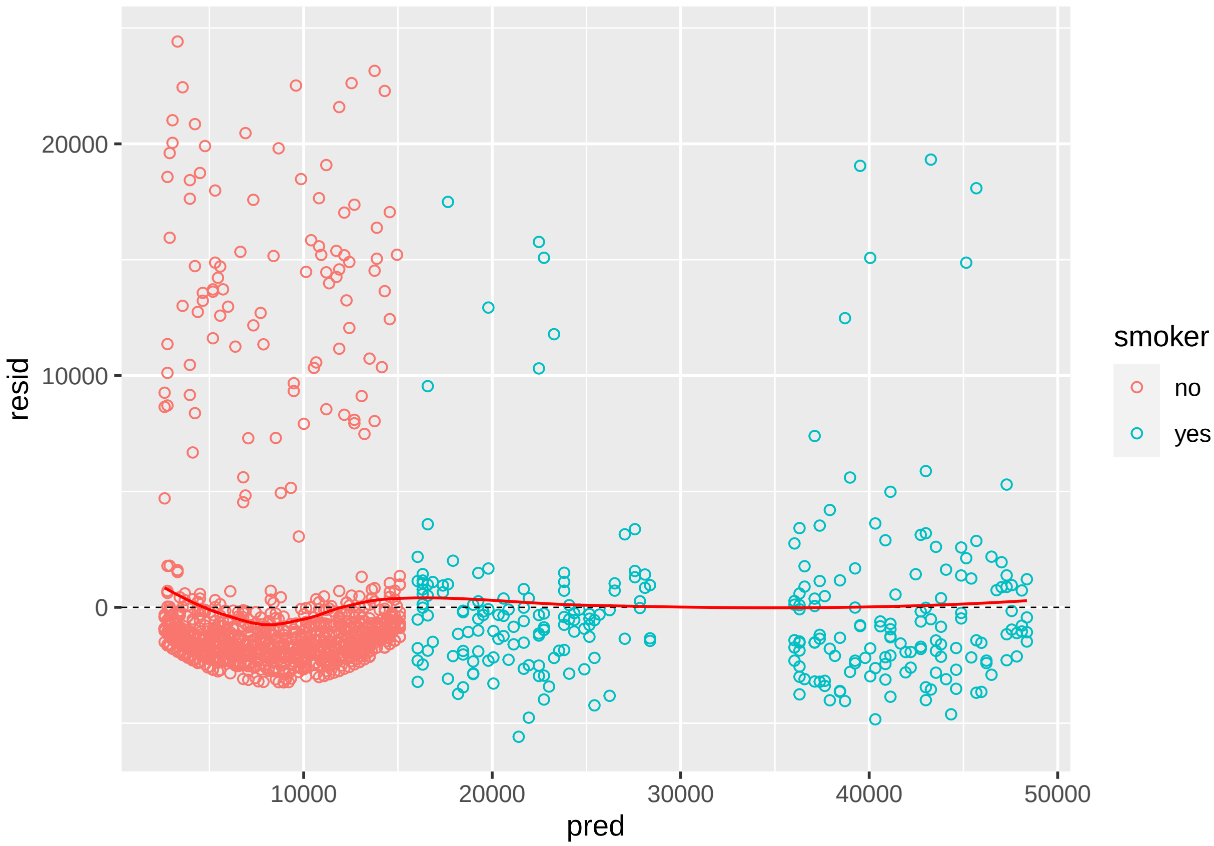
Además, vean cómo mejoró la determinación del coeficiente que acompaña a `age` (recuerden, es nuestro objeto interés!), y cómo cambió el RSE y el R^2.

```r
# Agrego modelo
grid <- modelr::data_grid(df, age, smoker, bmi30)

# Agrego modelo
grid <- modelr::add_predictions(grid, mod5)

# Agrego modelo
p <- ggplot(df, aes(x=age, y=charges, color=smoker, shape=bmi30)) + geom_point() +
  geom_line(data=grid, mapping=aes(y=pred, color=smoker, linetype=bmi30))
p
```




```
# plot(mod5)

df_res <- modelr::add_residuals(df, mod5) %>%
    modelr::add_predictions(mod5)

# Grafico residuos
ggplot(df_res, aes(x=pred, y=resid)) + geom_point(aes(color=smoker), shape=1) +
  geom_hline(yintercept = 0, size=0.25, linetype='dashed') +
  geom_smooth(aes(x=pred, y=resid), method='loess', size=0.5, se = F, color = 'red')
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Vemos que los residuos están mucho mejor, aunque hay varios puntos outliers que habría que mirar con más detalle.

Modelo final

Para terminar, si quisiéramos sacar el parámetro que no era significativo, tenemos que hacer una variable nueva (no encontré cómo hacerlo con fórmulas de R).

```
df <- mutate(df, bmi30smoker = (bmi > 30) & (smoker=='yes'))

mod6 <- lm(charges ~ age + smoker + bmi30smoker, data = df)

summary(mod6)
```

```
##
## Call:
## lm(formula = charges ~ age + smoker + bmi30smoker, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5585.7 -1953.5 -1324.8  -394.4 24493.8
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2138.210    373.926  -5.718 1.33e-08 ***
## age                268.437      8.816  30.449  < 2e-16 ***
```

```
## smokeryes       13343.999    420.794  31.711  < 2e-16 ***
## bmi30smokerTRUE 19990.018    547.769  36.494  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4527 on 1334 degrees of freedom
## Multiple R-squared:  0.8606, Adjusted R-squared:  0.8603
## F-statistic:  2745 on 3 and 1334 DF,  p-value: < 2.2e-16
```

Acá vemos que todos los parámetros son significativos, y que tenemos el menor error el estimador del coeficiente que acompaña a `age`. Además, redujimos la dispersión de los residuos a ~$4000; esto está dominado por los puntos de personas no fumadoras que por alguna razón pagan mucho seguro de salud. ¿Podemos entender esto? ¿Estará la información en el dataset? Queda como un desafío pensar en estas preguntas y explorar posibles respuestas.