Multiple linear regression

Prof. Maria Tackett

Click for PDF of slides

Review

Vocabulary

- Response variable: Variable whose behavior or variation you are trying to understand.
- Explanatory variables: Other variables that you want to use to explain the variation in the response.
- Predicted value: Output of the **model function**
  - The model function gives the typical value of the response variable *conditioning* on the explanatory variables.
- Residuals: Shows how far each case is from its predicted value
  - **Residual = Observed value - Predicted value**

STA 199

The linear model with a single predictor

- We're interested in the β_0 (population parameter for the intercept) and the β_1 (population parameter for the slope) in the following model:

$$\hat{y} = \beta_0 + \beta_1 \; x$$

STA 199

The linear model with a single predictor

- We're interested in the β_0 (population parameter for the intercept) and the β_1 (population parameter for the slope) in the following model:

$$\hat{y} = \beta_0 + \beta_1 \ x$$

- Unfortunately, we can't get these values
- So we use sample statistics to estimate them:

$$\hat{y} = b_0 + b_1 \ x$$

Least squares regression

The regression line minimizes the sum of squared residuals.

- Residuals: $e_i = y_i - \hat{y}_i$,
- The regression line minimizes $\sum_{i=1}^{n} e_i^2$.
- Equivalently, minimizing $\sum_{i=1}^{n} [y_i - (b_0 + b_1\ x_i)]^2$

Data and Packages

```
library(tidyverse)
library(broom)
```

```
paris_paintings <- read_csv("data/paris_paintings.csv",
                            na = c("n/a", "", "NA"))
```

- Paris Paintings Codebook
- Source: Printed catalogues from 28 auction sales held in Paris 1764 - 1780
- 3,393 paintings, prices, descriptive details, characteristics of the auction and buyer (over 60 variables)

Single numerical predictor

```
m_ht_wd <- lm(Height_in ~ Width_in, data = paris_paintings)
tidy(m_ht_wd)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    3.62    0.254        14.3 8.82e-45
## 2 Width_in       0.781   0.00950      82.1 0.
```

$$\widehat{Height_{in}} = 3.62 + 0.78\ Width_{in}$$

STA 199

Single categorical predictor (2 levels)

```
m_ht_lands <- lm(Height_in ~ factor(landsALL), data = paris_paintings)
tidy(m_ht_lands)
```

```
## # A tibble: 2 x 5
##   term              estimate std.error statistic  p.value
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)          22.7      0.328      69.1 0.
## 2 factor(landsALL)1    -5.65     0.532     -10.6 7.97e-26
```

$$\widehat{Height_{in}} = 22.68 - 5.65\ landsALL$$

STA 199

Single categorical predictor (> 2 levels)

```
m_ht_sch <- lm(Height_in ~ school_pntg, data = paris_paintings)
tidy(m_ht_sch)
```

```
## # A tibble: 7 x 5
##   term              estimate std.error statistic p.value
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)          14.        10.0     1.40  0.162
## 2 school_pntgD/FL       2.33      10.0     0.232 0.816
## 3 school_pntgF         10.2       10.0     1.02  0.309
## 4 school_pntgG          1.65      11.9     0.139 0.889
## 5 school_pntgI         10.3       10.0     1.02  0.306
## 6 school_pntgS         30.4       11.4     2.68  0.00744
## 7 school_pntgX          2.87      10.3     0.279 0.780
```

$$\widehat{Height_{in}} = 14 + 2.33\ sch_{D/FL} + 10.2\ sch_F + 1.65\ sch_G + 10.3\ sch_I + 30.4\ sch_S + 2.87\ sch_X$$

STA 199

The linear model with multiple predictors

STA 199

The linear model with multiple predictors

- Population model:

$$\hat{y} = \beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \cdots + \beta_k\ x_k$$

The linear model with multiple predictors

- Population model:

$$\hat{y} = \beta_0 + \beta_1 \, x_1 + \beta_2 \, x_2 + \cdots + \beta_k \, x_k$$

- Sample model that we use to estimate the population model:

$$\hat{y} = b_0 + b_1 \, x_1 + b_2 \, x_2 + \cdots + b_k \, x_k$$

Data

The data set contains prices for Porsche and Jaguar cars for sale on cars.com.

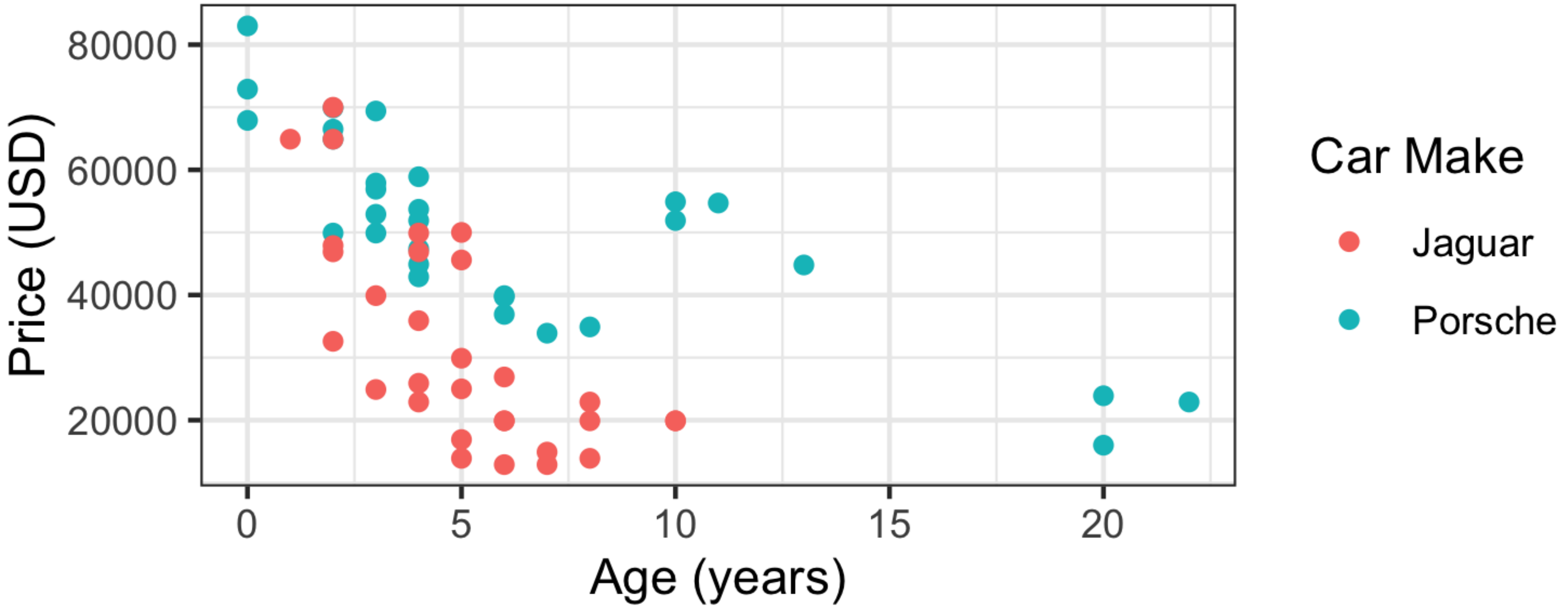
`car`: car make (Jaguar or Porsche)

`price`: price in USD

`age`: age of the car in years

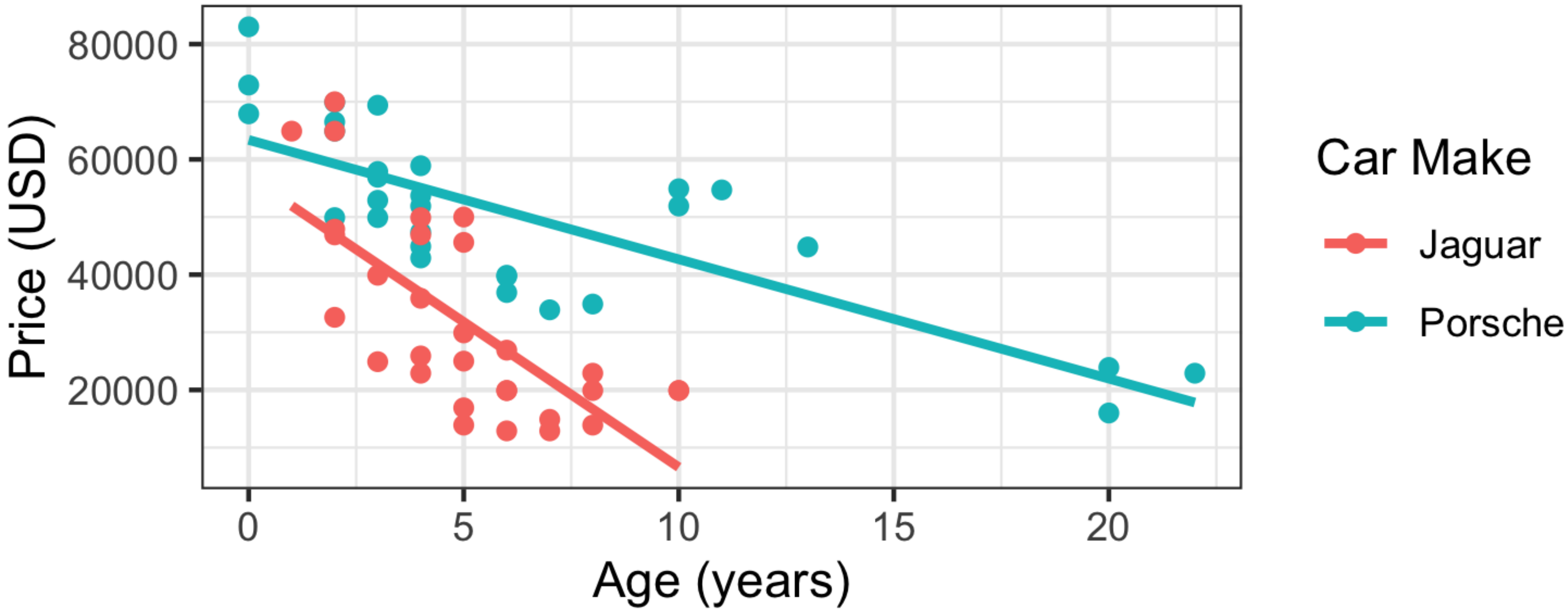
`mileage`: previous miles driven

Price, age, and make



Price vs. age and make

Does the relationship between age and price depend on the make of the car?



STA 199

Modeling with main effects

```r
m_main <- lm(price ~ age + car, data = sports_car_prices)

m_main %>%
  tidy() %>%
  select(term, estimate)
```

```
## # A tibble: 3 x 2
##   term        estimate
##   <chr>          <dbl>
## 1 (Intercept)   44310.
## 2 age           -2487.
## 3 carPorsche    21648.
```

STA 199

Modeling with main effects

```
m_main <- lm(price ~ age + car, data = sports_car_prices)

m_main %>%
  tidy() %>%
  select(term, estimate)
```

```
## # A tibble: 3 x 2
##   term        estimate
##   <chr>          <dbl>
## 1 (Intercept)   44310.
## 2 age           -2487.
## 3 carPorsche    21648.
```

$$\widehat{price} = 44310 - 2487\ age + 21648\ carPorsche$$

STA 199

$$\widehat{price} = 44310 - 2487\ age + 21648\ carPorsche$$

- Plug in 0 for **`carPorsche`** to get the linear model for Jaguars.

STA 199

$$\widehat{price} = 44310 - 2487\ age + 21648\ carPorsche$$

- Plug in 0 for **`carPorsche`** to get the linear model for Jaguars.

$$\begin{aligned}\widehat{price} &= 44310 - 2487\ age + 21648 \times 0 \\ &= 44310 - 2487\ age\end{aligned}$$

STA 199

$$\widehat{price} = 44310 - 2487\ age + 21648\ carPorsche$$

- Plug in 0 for **`carPorsche`** to get the linear model for Jaguars.

$$\begin{aligned}\widehat{price} &= 44310 - 2487\ age + 21648 \times 0 \\ &= 44310 - 2487\ age\end{aligned}$$

- Plug in 1 for **`carPorsche`** to get the linear model for Porsches.

$$\widehat{price} = 44310 - 2487\ age + 21648\ carPorsche$$

- Plug in 0 for **`carPorsche`** to get the linear model for Jaguars.

$$\begin{aligned}\widehat{price} &= 44310 - 2487\ age + 21648 \times 0 \\ &= 44310 - 2487\ age\end{aligned}$$

- Plug in 1 for **`carPorsche`** to get the linear model for Porsches.

$$\begin{aligned}\widehat{price} &= 44310 - 2487\ age + 21648 \times 1 \\ &= 65958 - 2487\ age\end{aligned}$$

Jaguar

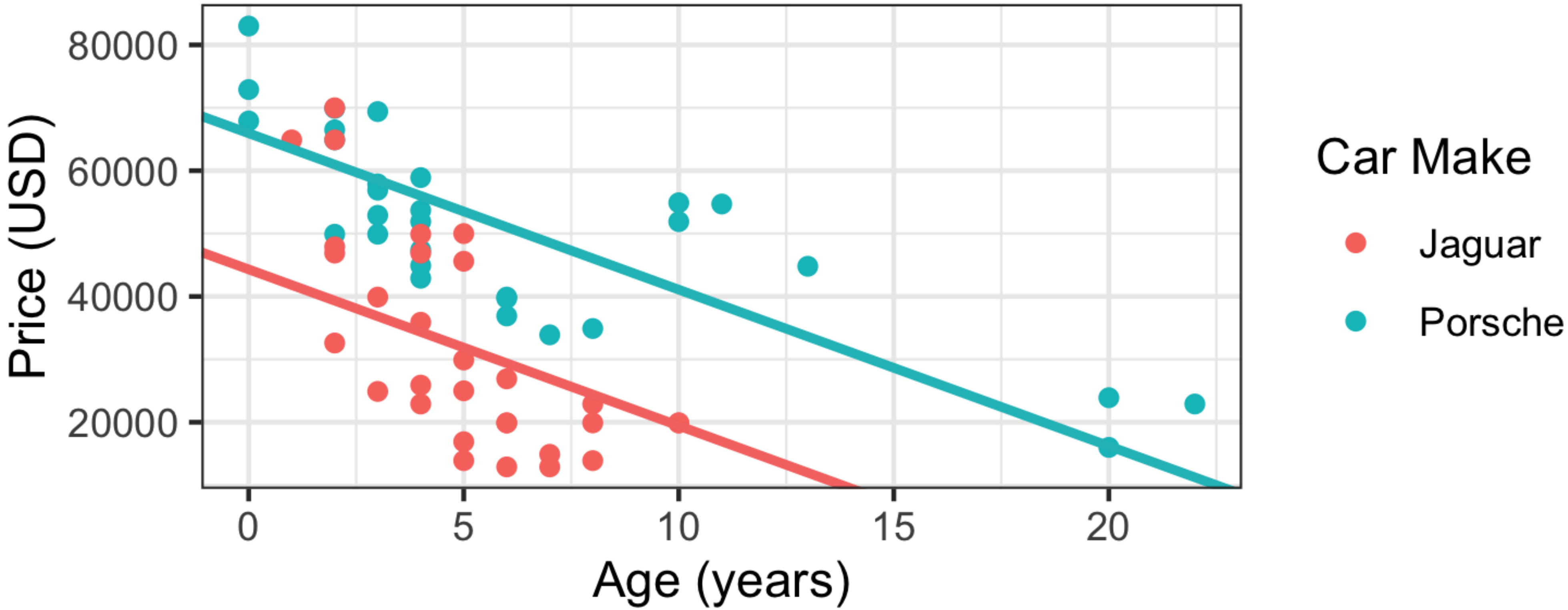
$$
\begin{aligned}
\widehat{price} &= 44310 - 2487\ age + 21648 \times 0 \\
&= 44310 - 2487\ age
\end{aligned}
$$

Porsche

$$
\begin{aligned}
\widehat{price} &= 44310 - 2487\ age + 21648 \times 1 \\
&= 65958 - 2487\ age
\end{aligned}
$$

- Rate of change in price as the age of the car increases does not depend on make of car (same slopes)
- Porsches are consistently more expensive than Jaguars (different intercepts)

Interpretation of main effects



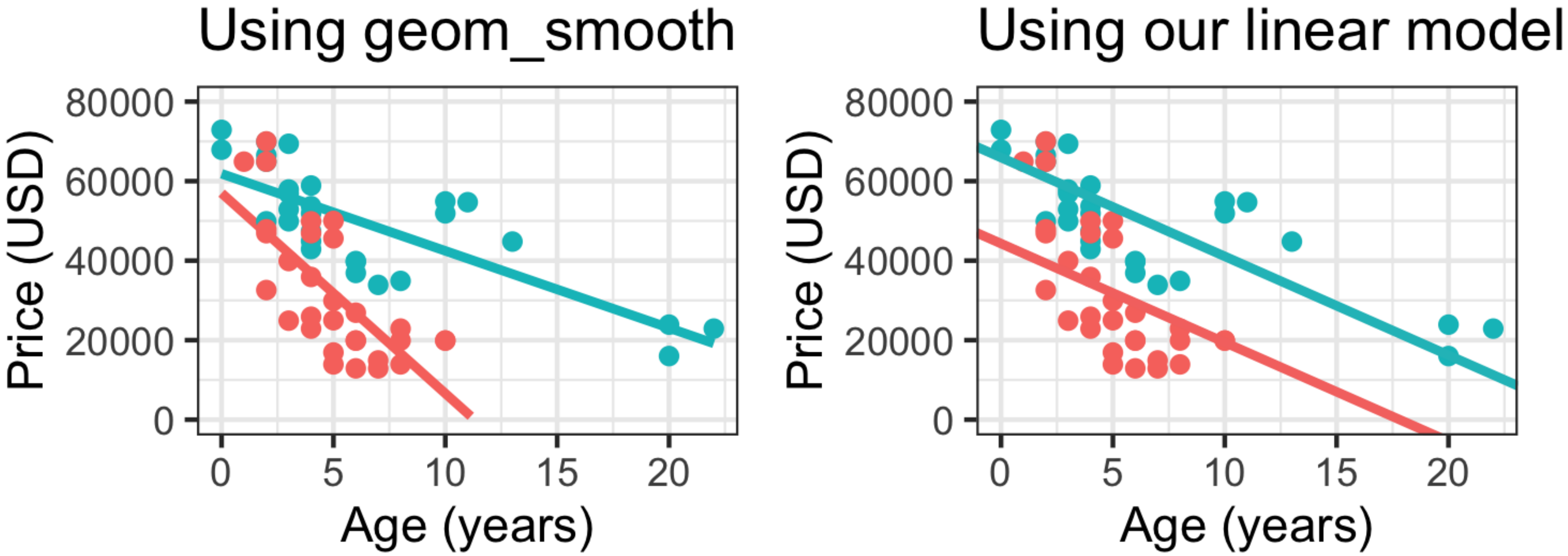
STA 199

Main effects

```
## # A tibble: 3 x 2
##   term          estimate
##   <chr>            <dbl>
## 1 (Intercept)     44310.
## 2 age             -2487.
## 3 carPorsche      21648.
```

- **All else held constant**, for each additional year of a car's age, the price of the car is predicted to decrease, on average, by $2,487.
- **All else held constant**, Porsches are predicted, on average, to have a price that is $21,647 greater than Jaguars.
- Jaguars that are new (age = 0) are predicted, on average, to have a price of $44,309.

Why is our linear regression model different from what we got from **`geom_smooth(method = "lm")`**?



STA 199

What went wrong?

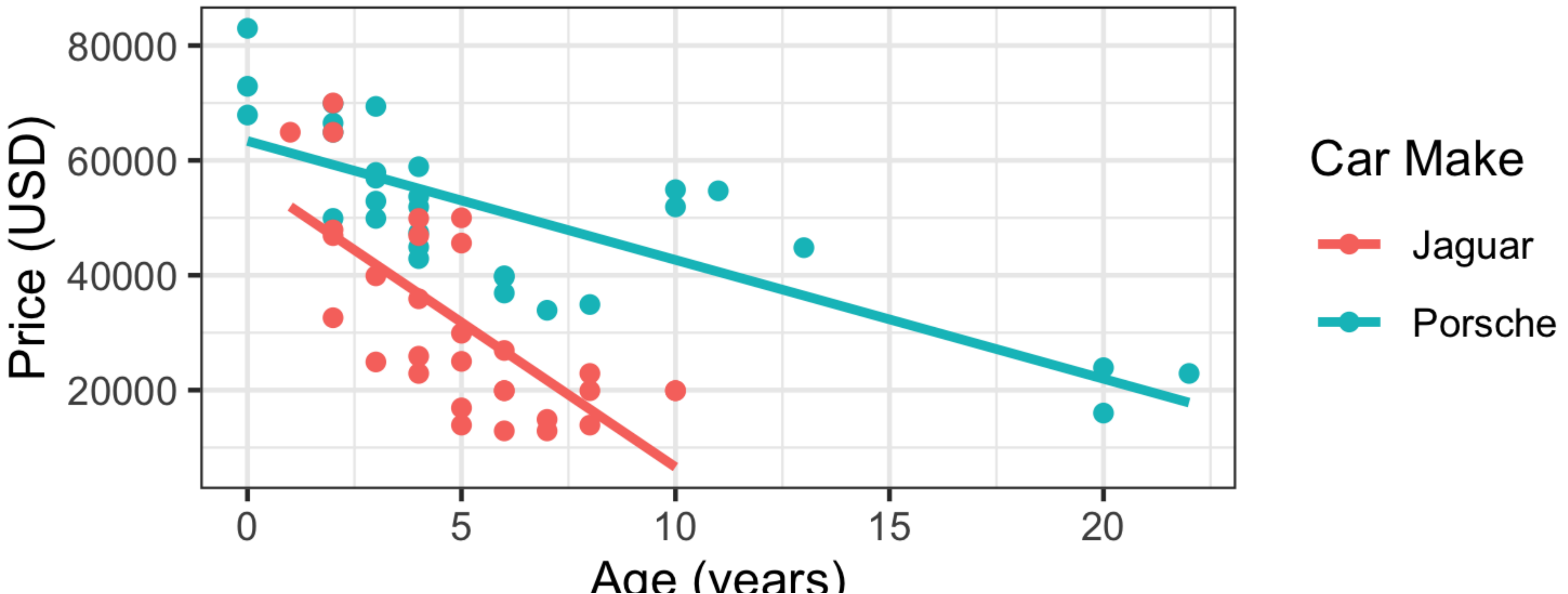
- `car` is the only variable in our model that affects the intercept.
- The model we specified assumes Jaguars and Porsches have the **`same slope`** and **`different intercepts`**.
- What is the most appropriate model for these data?
  - same slope and intercept for Jaguars and Porsches?
  - same slope and different intercept for Jaguars and Porsches?
  - different slope and different intercept for Jaguars and Porsches?

Interacting explanatory variables

- Including an interaction effect in the model allows for different slopes, i.e. nonparallel lines.
- This means that the relationship between an explanatory variable and the response depends on another explanatory variable.
- We can accomplish this by adding an **interaction variable**. This is the product of two explanatory variables.

Price vs. age and car interacting

```r
ggplot(data = sports_car_prices,
       mapping = aes(y = price, x = age, color = car)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Age (years)", y = "Price (USD)", color = "Car Make")
```



STA 199

Modeling with interaction effects

```
m_int <- lm(price ~ age + car + age * car, data = sports_car_prices)
m_int %>%
  tidy() %>%
  select(term, estimate)
```

```
## # A tibble: 4 x 2
##   term            estimate
##   <chr>              <dbl>
## 1 (Intercept)       56988.
## 2 age               -5040.
## 3 carPorsche         6387.
## 4 age:carPorsche     2969.
```

$$\widehat{price} = 56988 - 5040\ age + 6387\ carPorsche + 2969\ age \times carPorsche$$

STA 199

Interpretation of interaction effects

$$\widehat{price} = 56988 - 5040\ age + 6387\ carPorsche + 2969\ age \times carPorsche$$

Interpretation of interaction effects

$$\widehat{price} = 56988 - 5040\ age + 6387\ carPorsche + 2969\ age \times carPorsche$$

- Plug in 0 for **`carPorsche`** to get the linear model for Jaguars.

$$\begin{aligned}\widehat{price} &= 56988 - 5040\ age + 6387\ carPorsche + 2969\ age \times carPorsche \\ &= 56988 - 5040\ age + 6387 \times 0 + 2969\ age \times 0 \\ &= 56988 - 5040\ age\end{aligned}$$

STA 199

Interpretation of interaction effects

$$\widehat{price} = 56988 - 5040\ age + 6387\ carPorsche + 2969\ age \times carPorsche$$

- Plug in 0 for **`carPorsche`** to get the linear model for Jaguars.

$$\begin{aligned}
\widehat{price} &= 56988 - 5040\ age + 6387\ carPorsche + 2969\ age \times carPorsche \\
&= 56988 - 5040\ age + 6387 \times 0 + 2969\ age \times 0 \\
&= 56988 - 5040\ age
\end{aligned}$$

- Plug in 1 for **`carPorsche`** to get the linear model for Porsches.

$$\begin{aligned}
\widehat{price} &= 56988 - 5040\ age + 6387\ carPorsche + 2969\ age \times carPorsche \\
&= 56988 - 5040\ age + 6387 \times 1 + 2969\ age \times 1 \\
&= 63375 - 2071\ age
\end{aligned}$$

Interpretation of interaction effects

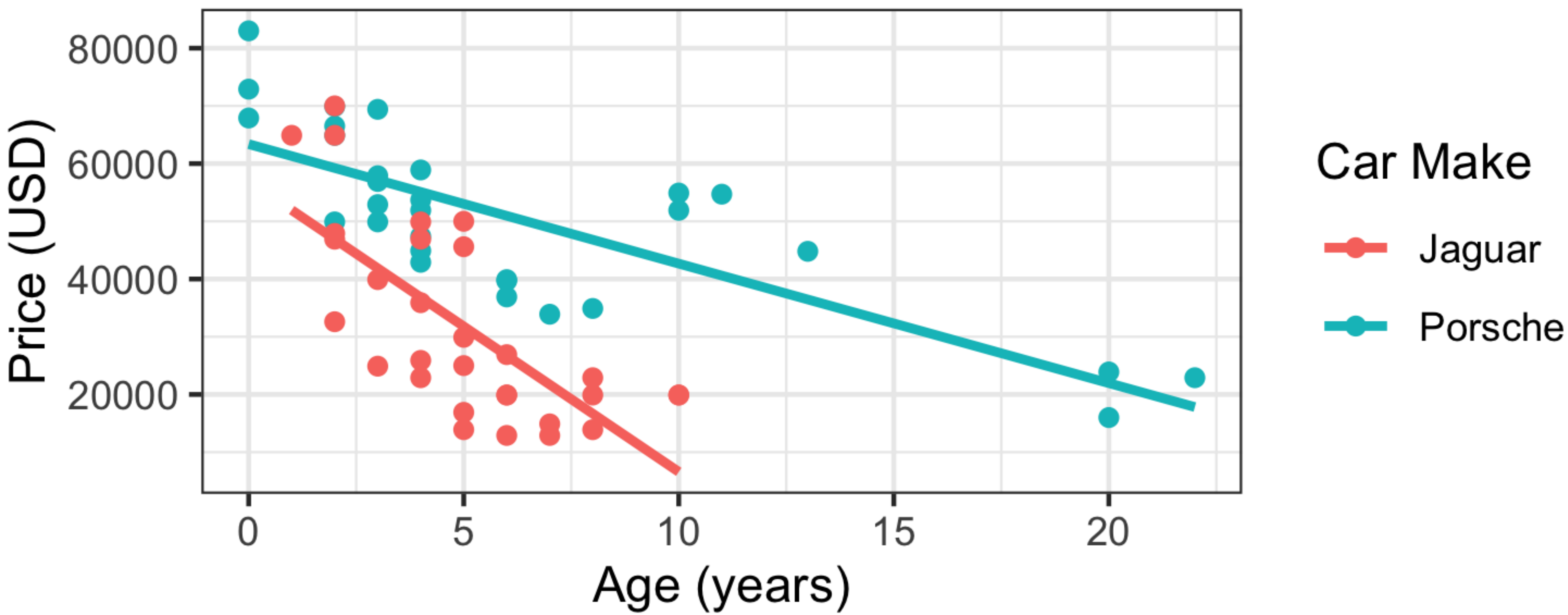
Jaguar

$$\widehat{price} = 56988 - 5040\ age$$

Porsche

$$\widehat{price} = 63375 - 2071\ age$$

- Rate of change in price as the age of the car increases depends on the make of the car (different slopes).
- Porsches are consistently more expensive than Jaguars (different intercepts).


Price (USD)
80000
60000
40000
20000
0
5
10
15
20
Age (years)
Car Make
Jaguar
Porsche

STA 199

Continuous by continuous interactions

- Interpretation becomes trickier
- Slopes conditional on values of explanatory variables

Continuous by continuous interactions

- Interpretation becomes trickier
- Slopes conditional on values of explanatory variables

Third order interactions

- Can you? Yes
- Should you? Probably not if you want to interpret these interactions in context of the data.

Assessing quality of model fit

Assessing the quality of the fit

- The strength of the fit of a linear model is commonly evaluated using R^2.
- It tells us what percentage of the variability in the response variable is explained by the model. The remainder of the variability is unexplained.
- R^2 is sometimes called the coefficient of determination.

What does "explained variability in the response variable" mean?

STA 199

Obtaining R^2 in R

`price` vs. **`age`** and **`make`**

```
glance(m_main)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared  sigma statistic  p.value    df logLik   AIC   BI
##       <dbl>         <dbl>  <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <db
## 1     0.607         0.593 11848.      44.0 2.73e-12     2  -646. 1301. 130
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
glance(m_main)$r.squared
```

```
## [1] 0.6071375
```

Obtaining R^2 in R

`price` vs. **`age`** and **`make`**

```
glance(m_main)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared  sigma statistic  p.value    df logLik   AIC   BI
##       <dbl>         <dbl>  <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <db
## 1     0.607         0.593 11848.      44.0 2.73e-12     2  -646. 1301. 130
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
glance(m_main)$r.squared
```

```
## [1] 0.6071375
```

About 60.7% of the variability in price of used cars can be explained by age and make.

R^2

```r
glance(m_main) %>% #model with main effects
  pull(r.squared)
```

```
## [1] 0.6071375
```

```r
glance(m_int) %>% #model with main effects + interactions
  pull(r.squared)
```

```
## [1] 0.6677881
```

R^2

```r
glance(m_main) %>% #model with main effects
  pull(r.squared)
```

```
## [1] 0.6071375
```

```r
glance(m_int) %>% #model with main effects + interactions
  pull(r.squared)
```

```
## [1] 0.6677881
```

- The model with interactions has a higher R^2.

R^2

```
glance(m_main) %>% #model with main effects
  pull(r.squared)
```

```
## [1] 0.6071375
```

```
glance(m_int) %>% #model with main effects + interactions
  pull(r.squared)
```

```
## [1] 0.6677881
```

- The model with interactions has a higher R^2.
- Using R^2 for model selection in models with multiple explanatory variables is <u>not</u> a good idea as R^2 increases when **any** variable is added to the model.

STA 199

R^2 - first principles

- We can write explained variation using the following ratio of sums of squares:

$$R^2 = 1 - \left(\frac{\text{variability in residuals}}{\text{variability in response}} \right)$$

Why does this expression make sense?

- But remember, adding **any** explanatory variable will always increase R^2

Adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{\text{variability in residuals}}{\text{variability in response}} \times \frac{n-1}{n-k-1} \right)$$

where n is the number of observations and k is the number of predictors in the model.

Adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{\text{variability in residuals}}{\text{variability in response}} \times \frac{n-1}{n-k-1} \right)$$

where n is the number of observations and k is the number of predictors in the model.

- Adjusted R^2 doesn't increase if the new variable does not provide any new information or is completely unrelated.

Adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{\text{variability in residuals}}{\text{variability in response}} \times \frac{n-1}{n-k-1} \right)$$

where n is the number of observations and k is the number of predictors in the model.

- Adjusted R^2 doesn't increase if the new variable does not provide any new information or is completely unrelated.
- This makes adjusted R^2 a preferable metric for model selection in multiple regression models.

Comparing models

```r
glance(m_main)$r.squared
```

```
## [1] 0.6071375
```

```r
glance(m_int)$r.squared
```

```
## [1] 0.6677881
```

```r
glance(m_main)$adj.r.squared
```

```
## [1] 0.5933529
```

```r
glance(m_int)$adj.r.squared
```

```
## [1] 0.649991
```

In pursuit of Occam's Razor

- Occam's Razor states that among competing hypotheses that predict equally well, the one with the fewest assumptions should be selected.
- Model selection follows this principle.
- We only want to add another variable to the model if the addition of that variable brings something valuable in terms of predictive power to the model.
- In other words, we prefer the simplest best model, i.e. parsimonious model.