STA 235H - Multiple Regression: Interactions, Collinearity, and Residuals

Fall 2021

McCombs School of Business, UT Austin

Let's look at some data

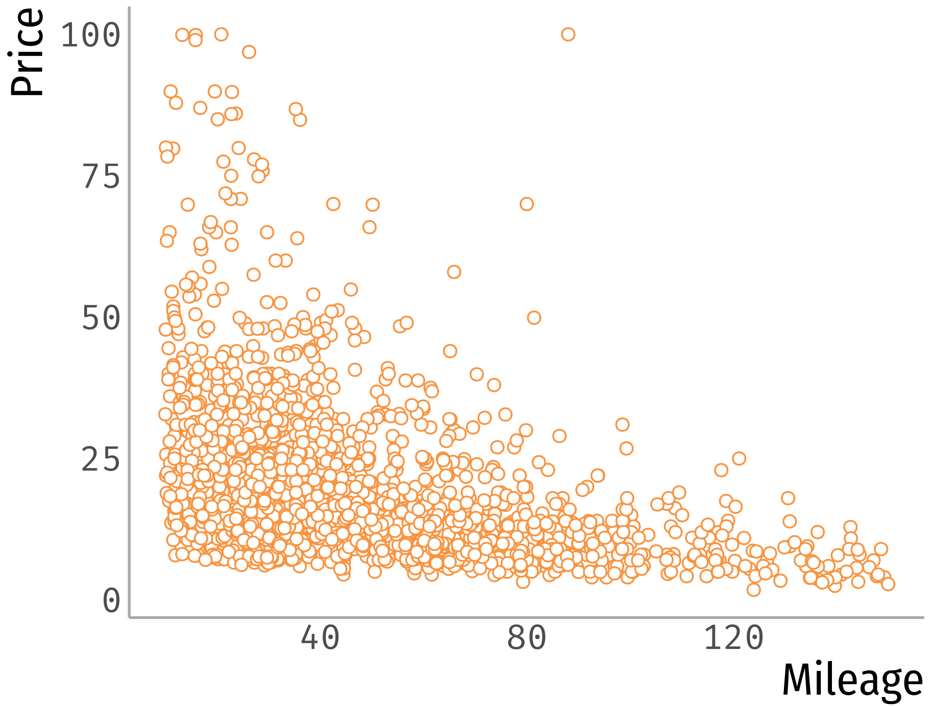
- Used cars in South California (from this week's JITT)

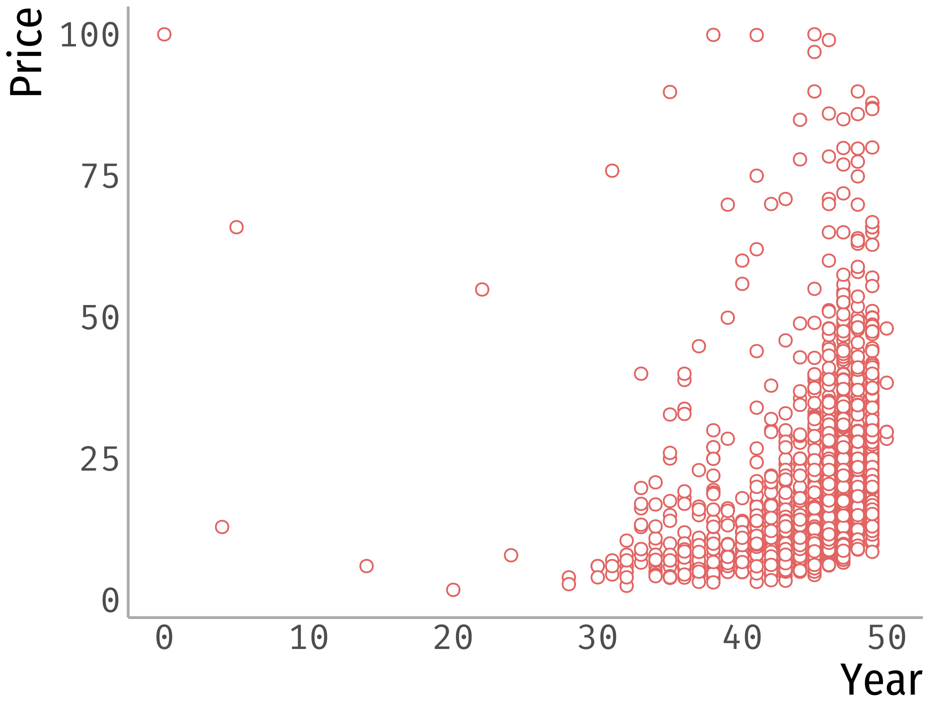
```r
library(vtable)

cars <- read.csv("https://raw.githubusercontent.com/maibennett/sta235/main/exampleSite/content/Classes/Week2/2_OLS_probs/c

names(cars)
```

```
##  [1] "type"      "certified" "body"      "make"      "model"     "trim"
##  [7] "mileage"   "price"     "year"      "dealer"    "city"      "rating"
## [13] "reviews"   "badge"
```

Data source: "Modern Business Analytics" (Taddy, Hendrix, & Harding, 2018)

How do mileage and year affect price?





Let's run a model

```
lm1 <- lm(price ~ year + mileage + rating, data = cars)

summary(lm1)
```

```
##
## Call:
## lm(formula = price ~ year + mileage + rating, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.945  -7.180  -2.465   3.791  72.444
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.78158    4.16486  10.512  < 2e-16 ***
## year        -0.36028    0.08284  -4.349 1.43e-05 ***
## mileage     -0.23406    0.01186 -19.738  < 2e-16 ***
## rating       1.21791    0.15886   7.666 2.69e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.08 on 2086 degrees of freedom
## Multiple R-squared:  0.2098,    Adjusted R-squared:  0.2086
## F-statistic: 184.6 on 3 and 2086 DF,  p-value: < 2.2e-16
```

- Why do you think these partial correlations are both negative for `year` and `mileage`?

Multicollinearity in Regressors

- If two covariates are **highly correlated**, it is difficult to separate the contribution of each!
  - E.g. They move together
- **Be careful** with interpretations
  - The same with extrapolation zones!

Let's start interacting

Luxury vs. non-luxury cars?

Do you think there's a difference between how price changes over time for luxury vs non-luxury cars?

How would you test this?

Let's go to R

Models with interactions

- You include the interaction between two (or more) covariates:

$$\widehat{Price} = \beta_0 + \hat{\beta}_1 Rating + \hat{\beta}_2 Miles + \hat{\beta}_3 Luxury + \hat{\beta}_4 Year + \hat{\beta}_5 Luxury \times Year$$

- $\hat{\beta}_3$ and $\hat{\beta}_4$ are considered the main effects (no interaction)

- The coefficient you are interested in is $\hat{\beta}_5$:
  - Difference in the price change for one additional year between luxury vs non-luxury cars, holding other variables constant.

Let's look at what's left

Residuals in an OLS regression

- **Residuals** are a fundamental part of OLS regression: They represent what is not explained by the covariates.
- When making *probabilistic inference*, we assume (among other things):
  - The distribution of the error term is **normal**
  - The error terms are **iid** (independent and identically distributed).

$$\boxed{\varepsilon_i \sim \mathcal{N}(\mu, \sigma)}$$

Let's look at the residuals

```r
cars <- cars %>% mutate(price_hat = predict(lm(price ~ rating + mileage + year, data = .)),
                        residual = price - price_hat)

ggplot(data = cars, aes(x = residual)) + geom_histogram()
```

Let's look at the variance!

```
ggplot(data = cars, aes(x = price_hat, y = residual)) + geom_point()
```

Let's look at the variance!

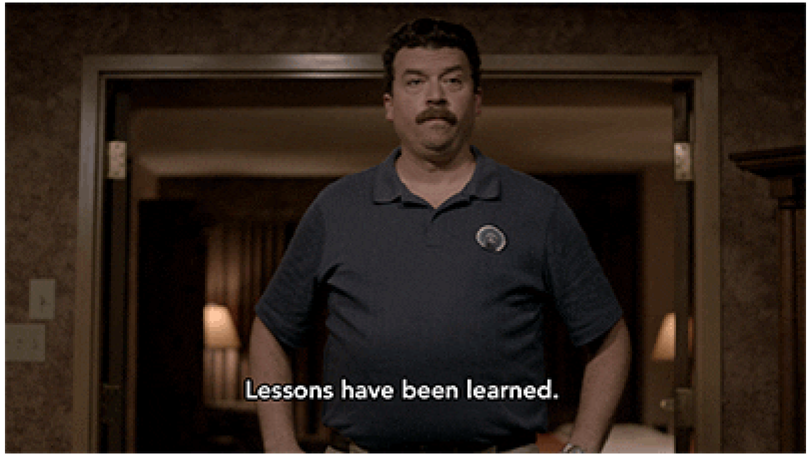
```r
ggplot(data = cars, aes(x = price_hat, y = residual)) + geom_point()
```

Can we fix that?

```
cars <- cars %>% mutate(price_hat = predict(lm(log(price) ~ rating + mileage + year, data = .)),
                        residual = log(price) - price_hat)

ggplot(data = cars, aes(x = price_hat, y = residual)) + geom_point()
```

Takeaway points

- It's important to think about the model we are fitting
    - Does it make sense? *Contextual knowledge is important*
- Interactions terms can capture important heterogeneity
- Always check your assumptions



Next Class

- Start with Causal Inference
- Homework 1 will be posted on Thursday
- JITT 2 will be posted today

References

- Ismay, C. & A. Kim. (2021). “Statistical Inference via Data Science”. Chapter 6 & 10.