STA 235H - Review Session I

Fall 2022

McCombs School of Business, UT Austin

Structure

- We will review **different types of regressions**
- I will provide a general framework of **how to think about these regressions**
- We will do one exercise **together**

Lightning round!

Participate!

Even if you make a mistake, everyone can learn from that

Ask questions!

You are here on a Friday afternoon.. take advantage of it :)

Typical regression: Continuous variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Y is a continuous variable
- X is a continuous variable

A one-unit increase in X_1 is associated with an average increase in Y of β_1 units, holding X_2 constant

Typical regression: Example

Let's analyze what factors associate to someone's grade:

$$Grade = \beta_0 + \beta_1 HrsAssign + \beta_2 PreGrade + \beta_3 ClassAtt + \varepsilon$$

Where:

- **Grade**: Grade in an assignment (out of 100 points).
- **HrsAssign**: Hours dedicated to the assignment.
- **PreGrade**: Grade in the previous assignment (out of 100 points).
- **ClassAtt**: Whether someone attended the last class (1) or not (0).

Interpret the coefficient for Hours

Typical regression: Example

```
summary(lm(grade ~ hours + pregrade + classatt, data = d))
```

```
##
## Call:
## lm(formula = grade ~ hours + pregrade + classatt, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.59771 -0.62967  0.07248  0.57736  2.42308
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.22912    1.12686   21.50  < 2e-16 ***
## hours        1.89525    0.02079   91.16  < 2e-16 ***
## pregrade     0.44563    0.01227   36.32  < 2e-16 ***
## classatt     1.04993    0.14461    7.26 8.85e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9254 on 196 degrees of freedom
## Multiple R-squared:  0.9808,    Adjusted R-squared:  0.9805
## F-statistic:  3338 on 3 and 196 DF,  p-value: < 2.2e-16
```

Typical regression: Binary covariate

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Y is a continuous variable
- X is a binary variable

The event of $X_1 = 1$ is associated with an average increase in Y of β_1 units compared to $X_1 = 0$, holding X_2 constant

Typical regression: Example

Let's analyze what factors associate to someone's grade:

$$Grade = \beta_0 + \beta_1 HrsAssign + \beta_2 PreGrade + \beta_3 ClassAtt + \varepsilon$$

Where:

- **Grade**: Grade in an assignment (out of 100 points).
- **HrsAssign**: Hours dedicated to the assignment.
- **PreGrade**: Grade in the previous assignment (out of 100 points).
- **ClassAtt**: Whether someone attended the last class (1) or not (0).

Interpret the coefficient for Class Attendance

Typical regression: Example

- Interpret the coefficient for *ClassAtt*.

```
summary(lm(grade ~ hours + pregrade + classatt, data = d))
```

```
##
## Call:
## lm(formula = grade ~ hours + pregrade + classatt, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.59771 -0.62967  0.07248  0.57736  2.42308
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.22912    1.12686   21.50  < 2e-16 ***
## hours        1.89525    0.02079   91.16  < 2e-16 ***
## pregrade     0.44563    0.01227   36.32  < 2e-16 ***
## classatt     1.04993    0.14461    7.26 8.85e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9254 on 196 degrees of freedom
## Multiple R-squared:  0.9808,    Adjusted R-squared:  0.9805
## F-statistic:  3338 on 3 and 196 DF,  p-value: < 2.2e-16
```

Regressions with logarithms: Log-level

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- The outcome variable is in a logarithm, $\log(Y)$
- X is a continuous variable, not in a logarithm.

A one-unit increase in X_1 is associated with an average increase in Y of $\beta_1 \times 100$ percent, holding X_2 constant

Regressions with logarithms: Example

Let's analyze what characteristics are associated with sales on Etsy:

$$\log(Sales) = \beta_0 + \beta_1 Price + \beta_2 AvgRating + \beta_3 NReviews + \varepsilon$$

Where:

- **log(Sales)**: Weekly sales of a product ($), in a logarithm.
- **Price**: Price of the product ($)
- **AvgRating**: Average rating of the product (in a scale fo 1-5).
- **NReviews**: Number of reviews of a product.

Interpret the coefficient for Number of Reviews

Regressions with logarithms: Example

- Interpret the coefficient for *NReviews*:

```
summary(lm(log(sales) ~ price + avgrating + nreviews, data = d))
```

```
##
## Call:
## lm(formula = log(sales) ~ price + avgrating + nreviews, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54727 -0.03702  0.01991  0.06927  0.19050
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.3319302  0.0405368 156.202  < 2e-16 ***
## price       0.0004073  0.0006945   0.586    0.558
## avgrating   0.0381499  0.0077626   4.915 1.48e-06 ***
## nreviews    0.0190180  0.0003177  59.866  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1119 on 296 degrees of freedom
## Multiple R-squared:  0.9261,    Adjusted R-squared:  0.9253
## F-statistic:  1236 on 3 and 296 DF,  p-value: < 2.2e-16
```

Regressions with logarithms: Your turn

- Interpret the coefficient for *Price*:

```
summary(lm(log(sales) ~ price + avgrating + nreviews, data = d))
```

```
##
## Call:
## lm(formula = log(sales) ~ price + avgrating + nreviews, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54727 -0.03702  0.01991  0.06927  0.19050
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.3319302  0.0405368 156.202  < 2e-16 ***
## price       0.0004073  0.0006945   0.586    0.558
## avgrating   0.0381499  0.0077626   4.915 1.48e-06 ***
## nreviews    0.0190180  0.0003177  59.866  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1119 on 296 degrees of freedom
## Multiple R-squared:  0.9261,    Adjusted R-squared:  0.9253
## F-statistic:  1236 on 3 and 296 DF,  p-value: < 2.2e-16
```

Regressions with Interactions

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 D + \beta_3 X_1 \times D + \beta_4 X_2 + \varepsilon$$

- Y is a continuous variable
- X is a continuous variable
- D is a binary variable

A one-unit increase in X_1 is associated with an average increase in Y of β_1 units for $D = 0$, holding X_2 constant

A one-unit increase in X_1 is associated with an average increase in Y of $\beta_1 + \beta_3$ units for $D = 1$, holding X_2 constant

Why?

Regressions with Interactions: Example

Let's analyze what factors associate to someone's grade:

$$Grade = \beta_0 + \beta_1 PreGrade + \beta_2 HrsAssign + \beta_3 ClassAtt + \beta_4 HrsAssign \times ClassAtt + \varepsilon$$

Where:

- **Grade**: Grade in an assignment (out of 100 points).
- **HrsAssign**: Hours dedicated to the assignment.
- **PreGrade**: Grade in the previous assignment (out of 100 points).
- **ClassAtt**: Whether someone attended the last class (1) or not (0).

Interpret the association between Hours and Grade for students that attended the last class and those that did not

Regressions with Interactions: Example

- Interpret the association the assicuation between *hours* and *grade* for the two groups.

```
summary(lm(grade ~ pregrade + hours*classatt, data = d))
```

```
##
## Call:
## lm(formula = grade ~ pregrade + hours * classatt, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74995 -0.39230 -0.03793  0.43372  1.99286
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    30.42867    0.88578  34.353   <2e-16 ***
## pregrade        0.35736    0.00957  37.341   <2e-16 ***
## hours           1.65246    0.02769  59.669   <2e-16 ***
## classatt        0.12650    0.22545   0.561    0.575
## hours:classatt  0.56198    0.03412  16.471   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7216 on 195 degrees of freedom
## Multiple R-squared:  0.9894,    Adjusted R-squared:  0.9892
## F-statistic:  4540 on 4 and 195 DF,  p-value: < 2.2e-16
```

Quadratic Regressions

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \varepsilon$$

- Y is a continuous variable
- X is a continuous variable and includes a quadratic term

Increasing X_1 in one unit, from x_0 to $x_0 + 1$, is associated with an average increase in Y of $2\beta_2 x_0 + \beta_1$ units, holding X_2 constant

Quadratic Regressions: Example

We think there is a quadratic association between price and sales, so we fit the following model:

$$Sales = \beta_0 + \beta_1 Price + \beta_2 Price^2 + \beta_3 AvgRating + \beta_4 NReviews + \varepsilon$$

Where:

- **Sales**: Weekly sales of a product ($)
- **Price**: Price of the product ($)
- **AvgRating**: Average rating of the product (in a scale fo 1-5).
- **NReviews**: Number of reviews of a product.

Interpret the association between Sales and Price, for an increase in Price from $15 to $16

Quadratic Regressions: Example

- Interpret the association between *sales* and *price* (for an increase in price from 15 to 16 dollars)

```
summary(lm(sales ~ price + I(price^2) + avgrating + nreviews, data = d))
```

```
##
## Call:
## lm(formula = sales ~ price + I(price^2) + avgrating + nreviews,
##     data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -238.849  -52.604    5.526   53.160  228.929
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 106.13919   52.09422   2.037   0.0425 *
## price       -10.33381    2.58888  -3.992 8.29e-05 ***
## I(price^2)    2.50065    0.03527  70.901  < 2e-16 ***
## avgrating    50.24900    5.42595   9.261  < 2e-16 ***
## nreviews     30.03823    0.22195 135.337  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.17 on 295 degrees of freedom
## Multiple R-squared:  0.9981,    Adjusted R-squared:  0.998
## F-statistic: 3.781e+04 on 4 and 295 DF,  p-value: < 2.2e-16
```

Linear Probability Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Y is a binary variable
- X is a continuous variable

Increasing X_1 in one unit is associated with an average increase in the probability of $Y = 1$ of β_1, holding X_2 constant

Increasing X_1 in one unit is associated with an average increase in the probability of $Y = 1$ of $\beta_1 \times 100$ percentage points, holding X_2 constant

Linear Probability Model: Example

Let's analyze what factors associate to getting an A in an assignment:

$$Grade = \beta_0 + \beta_1 HrsAssign + \beta_2 PreGrade + \beta_3 ClassAtt + \varepsilon$$

Where:

- **GradeA**: Binary variable if the grade in the assignment is A (1) or not (0).
- **HrsAssign**: Hours dedicated to the assignment.
- **PreGrade**: Grade in the previous assignment (out of 100 points).
- **ClassAtt**: Whether someone attended the last class (1) or not (0).

Interpret the coefficient for Hours

Linear Probability Model: Example

- Interpret the coefficient for *Hours*.

```
summary(lm(gradeA ~ hours + pregrade + classatt, data = d))
```

```
##
## Call:
## lm(formula = gradeA ~ hours + pregrade + classatt, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18649 -0.05634 -0.01766  0.02702  0.80487
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.516716   0.170502  -3.031  0.00277 **
## hours        0.020050   0.003146   6.374 1.29e-09 ***
## pregrade     0.004937   0.001857   2.659  0.00849 **
## classatt    -0.026360   0.021881  -1.205  0.22976
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.14 on 196 degrees of freedom
## Multiple R-squared:  0.2117,    Adjusted R-squared:  0.1997
## F-statistic: 17.55 on 3 and 196 DF,  p-value: 3.936e-10
```

Putting everything together

Let's interpret what is the return to experience (i.e. association between income and experience) for men and women:

$$Income = \beta_0 + \beta_1 education + \beta_2 exp + \beta_3 exp^2 + \beta_4 female + \beta_5 exp \times female + \beta_6 exp^2 \times female + \varepsilon$$

Putting everything together

```
summary(lm(income ~ education + experience*female + I(experience^2)*female, data = d))
```

```
##
## Call:
## lm(formula = income ~ education + experience * female + I(experience^2) *
##     female, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16033.6  -3393.6   -221.4   3500.0  15994.1
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10865.874   5017.814   2.165  0.03055 *
## education               1955.037     50.306  38.863  < 2e-16 ***
## experience              1713.175    266.341   6.432 1.82e-10 ***
## female                  6662.077   6606.571   1.008  0.31347
## I(experience^2)          -14.771      3.586  -4.119 4.07e-05 ***
## experience:female      -1004.739    382.104  -2.629  0.00866 **
## female:I(experience^2)   -11.941      5.477  -2.180  0.02944 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5147 on 1193 degrees of freedom
## Multiple R-squared:  0.9534,    Adjusted R-squared:  0.9531
## F-statistic:  4066 on 6 and 1193 DF,  p-value: < 2.2e-16
```

Answers

1. A one-hour increase in studying is associated with an average 1.9 point increase in the assignment grade, holding other variables constant.
2. Attending the last class is associated with an average increase in grade of 1.05 points compared to not attending, holding other variables constant.
3. Having one additional review is associated with an average increase in sales of 1.9 percent, holding price and average rating constant.
4. One additional hour dedicated to the assignment is associated to an average increase of 1.65 points in the homework score for students that didn't attend the last class, holding previous grade constant.
5. One additional hour dedicated to the assignment is associated to an average increase of 2.2 points in the homework score for students that attended the last class, holding previous grade constant.
6. Increasing the sales price from 15 to 16 dollars is associated to an average increase in sales of $64.7, holding other variables constant.
7. One additional hour dedicated to the homework is associated to an average increase in the probability of getting an A of 2 percentage points, holding other variables constant.