POL 212

Regression

Amanda Loehrke

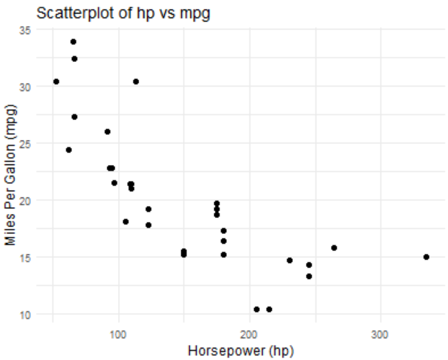
University of California, Davis

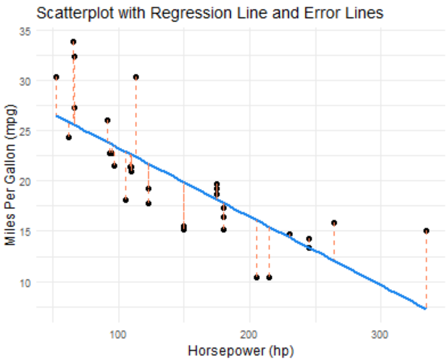
February 19, 2026

Introduction to OLS/Regression

- Used to describe/predict relationships in data
- Ordinary Least Squares (OLS) fits a line to the data to *best* explain how one variable (X) relates to another variable (Y)
- Minimizes the average error from our estimates to our observed data ($Y_i - \hat{Y}_i$)
- This results in a line that best describes the estimated relationship between the treatment and outcome variables

Linear regression





Gauss-Markov assumptions

- When these assumptions are satisfied, we can guarantee that OLS estimators are the Best Linear Unbiased Estimators (BLUE)
  - Linearity in parameters (mean-zero errors)
  - Exogeneity
  - Homoskedasticity
  - Independence

Linearity in parameters

- In our multiple variable regression model:
  $Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i$
  "linear" means linear in the coefficients, β, not in the variables
- $E(\epsilon_i | X_i) = 0$
- When we condition on the regressors, the remaining error averages to 0
- May fail due to model misspecification, omitted nonlinear relationships, group-specific prediction bias

Exogeneity

- The regressors are independent of the error term
  $\text{Cov}(X_i, \epsilon_i) = 0$
- Nothing in the error term is systematically correlated with X
- All variables that jointly affect X and Y are controlled for or irrelevant
- Changes in X are not driven by the same unobserved forces that affect Y
- This is key for unbiasedness
- May be violated with omitted variables, measurement error in X, reverse/simultaneious causation between X and Y

Homoskedasticity

- The variance of the errors is constant across all values of X
  $\text{Var}(\epsilon_i | X_i) = \sigma^2$
- The model is equally uncertain everywhere
- The spread of errors doesn't increase/decrease as X changes
- Ensures standard errors are correctly estimated
- May fail due to outcome scale effects, greater uncertainty for extreme values of X, group-level differences found in nature

Independence

- Errors are uncorrelated across observations
  $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$
- One observation's error tells you nothing about another's
- Each data point contributes independent information
- Commonly violated in time-series data, panel data, spatial dependence and can cause underestimated standard errors

Linear regression

- Used to estimate the linear relationship between variables
- Useful for describing relationships and making predictions
- Helps us to evaluate our theory (it doesn't create it)
- Basic simple linear regression model:

  $Y = \alpha + \beta X + \varepsilon$
- Interpretation: a one-unit increase in X is associated with an average change of β_1 units in Y

Simple linear regression

$$\underbrace{Y}_{\text{Outcome Variable}} = \overbrace{\beta_0}^{\text{Constant/Intercept}} + \underbrace{\beta_1}_{\text{Slope/Coefficient}} \overbrace{X}^{\text{Treatment Variable}} + \underbrace{\varepsilon}_{\text{Error Term}}$$

- Y: dependent, outcome variable
- X: independent, predictor, explanatory variable
- β_0: predicted value of Y when $X = 0$ (intercept) (also α)
- β_1: expected change in Y for a one-unit increase in X (slope)
- ε: error term capturing variation not explained by X
- We assume linearity in the parameters, not in the variables

Linear regression parameters

- β_0: predicted value of Y when $X = 0$ (intercept)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- β_1: expected change in Y for a one-unit increase in X (slope)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Linear regression in R

- Using R, we can create linear models
- The `lm()` function will fit a linear model to our data and specified variables
- The `summary()` function will give the model's output in a useful way:
  - Parameters/coefficients
  - Standard errors
  - t-statistic
  - p-values
  - R^2

```
model <- lm(Y ~ X, data = dataobject)
summary(model)
```

Regression output

- **Estimate**: the regression coefficient showing the expected change in Y associated with a change in X
  - For a **continuous** X variable, a one-unit increase in X is associated with this average change in Y
  - For a **dummy/categorical** variable, the estimate compares that category to the baseline (omitted) category
    - When $X = 1$, Y is this amount higher/lower than when $X = 0$
- **Intercept**: the predicted value of Y for the baseline category or when all X variables equal 0

Regression output, continued

- **Standard error**: the uncertainty in the estimated coefficient
  - How much the estimate would vary across repeated samples
- **t-statistic/t-value**: the ratio of the estimate to its standard error
  - Tests whether the coefficient is significantly different from 0
  - Higher absolute t-value suggests a stronger relationship (can be compared to a t-table)
- **p-value**: the probability of observing a t-statistic as extreme as the one obtained if the true coefficient were 0
  - Used to assess statistical significance of the relationship
  - Lower p-value suggests a stronger relationship (under 0.05 is standard)

Example in R

- Is a car's horsepower associated with gas mileage?
- Two scenarios we may test:

Non-directional	Directional
$H_0 : \beta_{hp} = 0$	$H_0 : \beta_{hp} \geq 0$
$H_1 : \beta_{hp} \neq 0$	$H_1 : \beta_{hp} < 0$

- Treatment = horsepower: `hp`
- Outcome = miles per gallon: `mpg`

```
# Fit a linear model using mtcars:
model1 <- lm(mpg ~ hp, data = mtcars)

# Summarize the model output:
summary(model1)
```

Example output in R

```
> summary(model1)

Call:
lm(formula = mpg ~ hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7121 -2.1122 -0.8854  1.5819  8.2360

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 30.09886    1.63392  18.421 < 0.0000000000000002 ***
hp          -0.06823    0.01012  -6.742          0.000000179 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.863 on 30 degrees of freedom
Multiple R-squared:  0.6024,	Adjusted R-squared:  0.5892
F-statistic: 45.46 on 1 and 30 DF,  p-value: 0.0000001788
```

Example output in R

```
> summary(model1)

Call:
lm(formula = mpg ~ hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7121 -2.1122 -0.8854  1.5819  8.2360 

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)    
(Intercept) 30.09886    1.63392  18.421 < 0.0000000000000002 ***
hp          -0.06823    0.01012  -6.742          0.000000179 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.863 on 30 degrees of freedom
Multiple R-squared:  0.6024,	Adjusted R-squared:  0.5892 
F-statistic: 45.46 on 1 and 30 DF,  p-value: 0.0000001788
```

Interpreting our example

- Intercept: $\beta_0 = 30.1$
  - The predicted mpg is 30.1 when hp is 0
  - Not a lot of practical meaning in this example since a car won't have 0 horsepower
- Coefficient for hp: $\beta_1 = -0.068$
  - Negative relationship between horsepower and miles per gallon
  - As a car's hp increases, its fuel efficiency decreases
  - For every additional horsepower, the mpg decreases by about 0.068

Interpreting our example, continued

- Standard error for hp: 0.010
  - Smaller standard error suggests the estimate is quite precise
- T-value for hp: -6.742
  - Large t-statistic shows the negative relationship between hp and mpg is significantly different from 0
  - Indicates that we would reject the null hypothesis $\beta_{hp} \neq 0$
- P-value for hp: 0.000000179 ***
  - Very low p-value indicates statistic significance
  - Also indicates that we would reject the null hypothesis
- R^2: 0.60
  - The model accounts for 60% of the variation in `mpg`

Interpreting our example, continued

- Our model gives us the regression equation:
  $Y = \beta_0 + \beta_1 X$
  $\hat{mpg} = 30.1 - 0.068(hp)$

- The predicted miles per gallon is 30.1 when horsepower is 0
- For each additional horsepower, miles per gallon decreases by about 0.068
- This reflects a negative relationship between horsepower and fuel efficiency

Multiple variable regression

- What if we want to model more than one independent variable?
- We use multiple variable regression to do so, with the equation:
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \ldots \ + \beta_n X_n$
- This gives us a partial interpretation: "holding other X constant"
- In R code: lm(Y $\sim$ X1 + X2 + X3, dataset)

Interpreting multiple variable regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \dots \ + \beta_n X_n$$

- β_0: the predicted value of Y when **all** independent variables equal zero
- β_1: the expected change in Y associated with a one-unit increase in X_1, holding all other X variables constant
- β_2: the expected change in Y associated with a one-unit increase in X_2, holding all other X variables constant
- β_n: the expected change in Y associated with a one-unit increase in X_n, holding all other X variables constant

Example of multiple variable regression

- What is associated with a car's gas mileage?
- Outcome = miles per gallon: `mpg`
- Predictor = horsepower: `hp`
- Predictor = number of cylinders (4, 6, or 8): `cyl`
- Predictor = transmission type (manual = 1, automatic = 0): `am`

Example, continued

- What might are our research expectations be for each independent variable?
- **Horsepower**
  - As horsepower increases, fuel efficiency should decrease. More powerful engines require more fuel to generate output.
- **Cylinders**
  - Cars with more cylinders are expected to have lower miles per gallon, because additional cylinders increase engine size and fuel consumption.
- **Transmission**
  - Manual transmissions are expected to have higher miles per gallon than automatic transmissions. This data is from 1970s cars, so manual transmissions generally lose less energy through the drivetrain compared to automatic designs.

Example in R

```
model2 <- lm(mpg ~ hp + cyl + am, data = mtcars)
summary(model2)
```

```
Call:
lm(formula = mpg ~ hp + cyl + am, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max 
-4.864 -1.811 -0.158  1.492  6.013 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 30.88834    2.78422  11.094 9.27e-12 ***
hp          -0.03688    0.01452  -2.540  0.01693 *  
cyl         -1.12721    0.63417  -1.777  0.08636 .  
am           3.90428    1.29659   3.011  0.00546 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.807 on 28 degrees of freedom
Multiple R-squared:  0.8041,    Adjusted R-squared:  0.7831 
F-statistic: 38.32 on 3 and 28 DF,  p-value: 4.791e-10
```

Interpreting our example

- Our regression equation:
  $\hat{mpg} = 30.89 - 0.04hp \; - 1.13hp \; + 3.90am$

- Intercept: $\beta_0 = 30.89$
  - The predicted miles per gallon for a car when $\texttt{hp} = 0$, $\texttt{cyl} = 0$, and $\texttt{am} = 0$ (automatic)
  - Less substantive here as the values aren't realistic for a car

Interpretation, continued

- Coefficient for `hp`: $\beta_1 =$ -0.04
  - Holding `cyl` and `am` constant, each 1 additional `hp` is associated with a 0.04 decrease in `mpg`
- Coefficient for `cyl`: $\beta_2 =$ -1.13
  - Holding `hp` and `am` constant, each additional `cyl` (4 to 6 to 8) is associated with a 1.13 decrease in `mpg`
- Coefficient for `am`: $\beta_3 =$ 3.90
  - Holding `hp` and `cyl` constant, manual transmission ($\texttt{am} = 1$) get about 3.9 more `mpg` than automatic cars ($\texttt{am} = 0$)

Performing regression by hand

- Here is a subset of the `mtcars` dataset for horsepower and miles per gallon:

hp	180	150	123	175	245	110	123	110	95
mpg	15.2	15.2	17.8	19.7	14.3	21	19.2	21	22.8

- Using this data, calculate the regression parameters β_0 and β_1:
  - $\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - $\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Regression by hand, continued

- Begin with calculating the mean of hp and mpg:

$$\text{Sum of hp: } 180 + 150 + 123 + 175 + ... + 95 = 1311$$
$$\text{Mean of hp: } \frac{\sum x_i}{n} = \frac{1311}{9} = \boxed{145.67}$$

$$\text{Sum of mpg: } 15.2 + 15.2 + 17.8 + 19.7 + ... + 22.8 = 166.2$$
$$\text{Mean of mpg: } \frac{\sum y_i}{n} = \frac{166.2}{9} = \boxed{18.47}$$

Regression by hand, continued

- Compute the numerator of β_1: $\sum(x_i - \bar{x})(y_i - \bar{y})$

hp (X)	mpg (Y)	$X - \bar{X}$	$Y - \bar{Y}$	Product
180	15.2	34.33	-3.27	-112.16
150	15.2	4.33	-3.27	-14.16
123	17.8	-22.67	-0.67	15.11
175	19.7	29.33	1.23	36.18
245	14.3	99.33	-4.17	-413.89
110	21.0	-35.67	2.53	-90.36
123	19.2	-22.67	0.73	-16.62
110	21.0	-35.67	2.53	-90.36
95	22.8	-50.67	4.33	-219.56

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = \boxed{-905.8}$$

Regression by hand, continued

- Compute the denominator of β_1: $\sum(x_i - \bar{x})^2$

hp (X)	$X - \bar{X}$	$(x_i - \bar{x})^2$
180	34.33	1178.78
150	4.33	18.78
123	-22.67	513.78
175	29.33	860.44
245	99.33	9867.11
110	-35.67	1272.11
123	-22.67	513.78
110	-35.67	1272.11
95	-50.67	2567.11

$$\sum(x_i - \bar{x})^2 = \boxed{18064}$$

Regression by hand, continued

- Calculate $\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{-905.8}{18064} = \boxed{-0.05}$

- Calculate $\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
  $18.47 - (-0.05)(145.67) = \boxed{25.75}$

- The final regression equation:
  $$m\hat{p}g = 25.75 - 0.05(hp)$$

Regression by hand, continued

- To check with R:

```
Call:
lm(formula = mpg ~ hp, data = smallmtcars)

Coefficients:
(Intercept)           hp
   25.77097     -0.05014
```

- The final regression equation:

$$\hat{mpg} = 25.75 - 0.05(hp)$$

- For every one unit increase in horsepower, miles per gallon decrease by about 0.048 units.