Multiple Linear Regression 1

Statistical Models and Regression with Multiple Predictors

Grinnell College

Fall 2025

Outline

Where we have been:

- Linear Regression
- making predictions using 'predictor' variables
- quant predictors + categorical predictors

Oh, the place's we'll go

- Statistical "Models"
- Regression with multiple variables
- Assessing Regression quality
- p-values

Statistical Models

A "model" is a simplified representation of some real-world phenomena. Models do not require, but may often use equations or probability concepts to help us explain how something works.

- *simplified* part is key
- reality is complicated, don't get bogged down in details

Example: Crop Growth
Simplistic: More rain (but not too much!) will make crops grow better.

Mathematical Models

Model goals: We may want to see how certain variables are related, calculate likelihood of rare events, or even make predictions.

A **mathematical model** relates things together using equations.

Example: These can be used to understand the relationship between how long it takes an object to fall from a particular height given strength of gravity. $t = \sqrt{\frac{2h}{g}}$ where where t = time to fall, h = height, g = gravity strength.

- ignores air resistance, wind, and strength of gravity changing as the distance from earth's core changes
- still give us very accurate estimates in nearly all situations we encounter in our lives.

Statistical Models

One of the distinctions of "statistical" models from mathematical models in general is that statistical models are not deterministic. Outcomes can change if we re-run a scenario. Much of what you have seen in 209 are various statistical models.

Example: Probability Model
MCAT data was roughly Normal. We could calculate proportions of students have above/below/between some values.

Example: Simple Linear Regression
We were looking at how predictor variables are related to a response.

- Requires a linear relationship
- Requires a quantitative response
- Simplified to only use 1 predictor variable

Theoretical vs Fitted Models

Just like we made distinctions between parameters (for a population) and statistics (for a sample), we are going to do the same thing for regression coefficients.

Theoretical Model: $y_i = \beta_0 + \beta_1 x_i + e_i$ (describes entire population)

Fitted Model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ (estimates of coefficients from sample)

Coefficients in the "theoretical" model describe regression relationships as they truly are. The intercept and slope values we get from our sample are estimates of these and form the "fitted" model

- the e_i terms (residuals) are "random" errors that our regression will never be able to account for
- the fitted model will omit the error terms

Drawbacks of SLR

Simple linear regression as we've seen it is an extremely useful tool for making predictions. But it is often *too* simple for us to describe reality well. We can include more variables in our model, but we need to be careful

- Too few variables could make the model not work well to describe things (over-simplified)
- Too many variables and it will get overly complicated to understand what is really happening

Regression with Multiple Variables

We can are going to extend our regression model to handle multiple explanatory variables. This is our result

$$\text{Theoretical Model:} \quad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i$$

$$\text{Fitted Model:} \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}$$

Notation

- y_i represents response value for observation/case i
- x_{ij} represents a value for a predictor variable for observation i
- β_0 is the intercept
- other β's represent the slopes for other variables
- e_i is the random error term which represents this model is "statistical" and does not make perfect predictions

MLR Coefficients

Theoretical Model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i$

In general, coefficient interpretations are a bit more complicated in MLR.

Intercept (β_0): predicted value for the response variable when all predictor variables are 0

Slopes (other β's) are how much the prediction changes when the corresponding predictor variable changes *with all over variables held constant*

- $\frac{\partial y_i}{\partial x_{ij}} = \beta_j$ (full derivative would need to account for changing one predictor might affect another → MESSY)
- ignores changes to other predictors

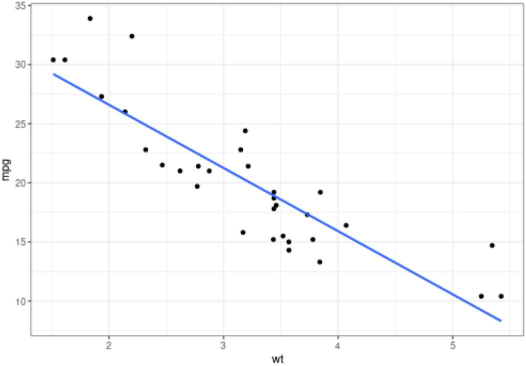
Example Cases

1. $y = \beta_0 + X\beta_1$
2. $y = \beta_0 + \mathbb{1}_A\beta_1$
3. $y = \beta_0 + \mathbb{1}_A\beta_1 + X\beta_2$
4. $y = \beta_0 + \mathbb{1}_A\beta_1 + \mathbb{1}_B\beta_2$
5. $y = \beta_0 + X_1\beta_1 + X_2\beta_2$

1. Simple linear, β_1 shows change in y given change in X
2. Simple categorical, reference variable and group means
3. Continuous and categorical, two regression lines with same slope but different intercept
4. Multiple categorical, combined reference variables
5. Multiple continuous, β_1 shows change in y given change in X_1, *assuming everything else held constant*

Example Case: SLR (Quant Predictor)

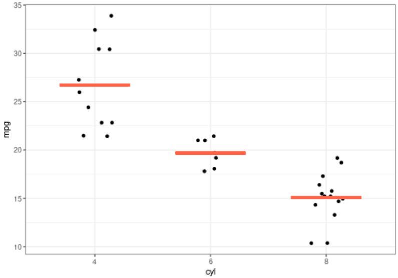
$$\widehat{mpg} = 37.285 - 5.34 \times \text{Weight (tons)}$$



- intercept = 37.285: when weight = 0, predicted mpg = 37.285
- slope = 5.34: when Weight increases by 1 ton, predicted mpg decreases by 5.34 mpg
- not causal! We are just describing overall relationship

Example Case: Multiple Categories

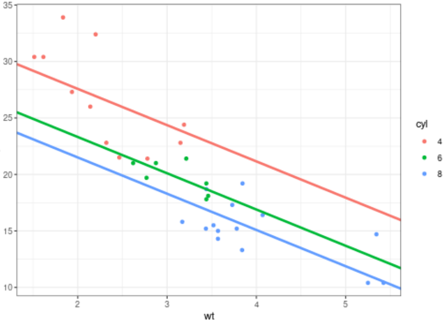
$$\widehat{mpg} = 26.66 - 6.92 \times \mathbb{1}_{\text{6cyl}} - 11.564 \times \mathbb{1}_{\text{8cyl}}$$



- intercept (ref): predicted mpg for 4 cylinder vehicles = 26.66mpg
- predicted mpg for 6 cylinder = 26.66 - 6.92 = 19.74
- predicted mpg for 8 cylinde = 26.66 - 11.564 = 15.096
- slopes adjust intercept to get predictions for other categories

Example Case: Quant+Cat Predictor

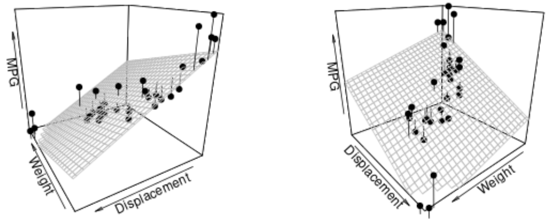
$$\widehat{mpg} = 33.99 - 3.21 \times \text{weight} - 4.26 \times \mathbb{1}_{\text{6cyl}} - 6.07 \times \mathbb{1}_{\text{8cyl}}$$



- intercept: predicted mpg = 33.99 for a zero-weight 4cyl vehicle
- slope: predicted mpg goes down 3.21 for each 1-ton increase in weight
- slopes corresponding to indicators represent vertical shifts for other cylinder categories → tell us which line to use

Example Case: Multiple Quant Predictors

$$\widehat{mpg} = 34.96 - 3.35 \times \text{weight} - 0.017 \times \text{displacement}$$



- intercept: predicted mpg = 34.96 when weight and displacement = 0
- slope 1: predicted mpg decreases by 3.35 when weight increases by 1 ton and displacement is constant
- slope 2: predicted mpg decreases by 0.017 when displacement increases by 1cc and weight is constant
- slope 2 change may seem small, but this is relative to variable magnitude

Summary() in R

The "summary" function gives us back more information than calling a model directly. Let's go back to the SLR using weight to predict mpg.

```
> lm(mpg ~ wt, mtcars) %>% summary()


Coefficients:
            Estimate Std. Error t value         Pr(>|t|)
(Intercept)   37.285      1.878   19.86 < 0.00000000002 ***
wt            -5.344      0.559   -9.56        0.000013 ***


Residual standard error: 3.05 on 30 degrees of freedom
Multiple R-squared:  0.753, Adjusted R-squared:  0.745
F-statistic: 91.4 on 1 and 30 DF,  p-value: 0.00000000129
```

- estimates (of coefficient values)
- std. errors / t-values / p-values?
- Multiple / Adjusted R^2?

F-statistic

In the summary output there is an "F-statistic." What is this doing? It uses the "F" probability distribution (will talk about this more next week) to test the following hypothesis:

H_0: all slope $\beta_j's = 0$ (no linear relationship b/w response & predictors)
H_A: at least one slope $\beta_j \neq 0$ (linear relationship with 1+ predictors)

This is sometimes called the "significance test" for the linear fit. A small p-value indicates the linear model seems to fit better than using only the intercept.

- good starting point for seeing if model works well, not done though
- large p-value says no evidence of linear relationship → predictors don't seem to have linear relationship with response
- just because fit is "significant" doesn't mean it works well!

Other P-values

The other p-values in the summary() output can be used to test if individual slope coefficients equal 0.

H_0: slope $\beta_j = 0$ (no linear relationship b/w response & this predictor)
H_A: slope $\beta_j \neq 0$ (linear relationship with this predictor)

Slope p-values

- Allows us to see whether individual variables are that helpful with prediction
- Use this **ONLY** as a rough guide
- We un into multiple comparison issues if checking lots of slopes → direct p-value interpretation becomes meaningless
- Test-statistic: $\mathrm{T} = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)}$, SE and df complicated

R^2 Stuff

Multiple R^2: When working with multiple variables correlations become more complicated. Multiple R^2 is a measure of how the response is correlated to *multiple* predictors.

- Same interpretation: % of variation in responses explained by the regression model
- $r^2 = R^2$ relationship is no longer valid
- problem: adding more variables can never make R^2 go down $\rightarrow$ adding lots of variables will never make model "worse" in terms of R^2 but it will be needlessly complex

Adjusted R^2 penalizes unhelpful variables

- adj.$R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$
- adding more variables that don't increase R^2 much will actually decrease adj.R^2

Practice

```
> lm(mpg ~ wt, mtcars) %>% summary()


Coefficients:
            Estimate Std. Error t value          Pr(>|t|)
(Intercept)   37.285      1.878   19.86 < 0.00000000002 ***
wt            -5.344      0.559   -9.56        0.000013 ***


Residual standard error: 3.05 on 30 degrees of freedom
Multiple R-squared:  0.753, Adjusted R-squared:  0.745
F-statistic: 91.4 on 1 and 30 DF,  p-value: 0.000000000129
```

- overall regression fit is discernible
- strong evidence of linear relationship b/w mpg and wt
- $R^2 = .753 \rightarrow r \approx .87$ seems like a strong linear relationship

Practice

```
> lm(mpg ~ wt + disp, mtcars) %>% summary()

Coefficients:
            Estimate Std. Error t value         Pr(>|t|)
(Intercept) 34.96055    2.16454   16.15  0.000000049 ***
wt          -3.35083    1.16413   -2.8        0.0074 **
disp        -0.01772    0.00919   -1.93       0.0636 .


Residual standard error: 2.92 on 29 degrees of freedom
Multiple R-squared:  0.781, Adjusted R-squared:  0.766
F-statistic: 51.7 on 2 and 29 DF,  p-value: 0.000000000274
```

- Overall fit still significant
- did adding 'disp' give us a better model?
- check 'disp' slope p-value
- both R^2's went up, but by a lot?
- Maybe the additional complexity is not worth the increased prediction performance

Practice

```
> lm(mpg ~ wt + disp + am,  mtcars) %>% summary()

Coefficients:
            Estimate Std. Error t value       Pr(>|t|)
(Intercept) 34.67591    3.24061   10.70 0.000000000021 ***
wt          -3.27904    1.32751   -2.47          0.020 *
disp        -0.01780    0.00937   -1.90          0.068 .
am1          0.17772    1.48432    0.12          0.906


Residual standard error: 2.97 on 28 degrees of freedom
Multiple R-squared:  0.781, Adjusted R-squared:  0.758
F-statistic: 33.3 on 3 and 28 DF,  p-value: 0.00000000225
```

- overall fit is statistically discernible
- did adding "automatic" help?
- Large "am1" p-value, R^2 did not change, adj.R^2 went down!
- no increase in performance & harder to interpret → bad!