STA 35C: Statistical Data Science III

Lecture 5: Multiple Linear Regression & Polynomial Regression

Dogyoon Song

Spring 2025, UC Davis

Agenda

Last time: Simple linear regression

- Model: $Y = \beta_0 + \beta_1 X + \epsilon$
- Least squares: estimate β_0, β_1 by minimizing $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Inference on β_0, β_1: confidence intervals & hypothesis tests using $\mathrm{SE}(\hat{\beta}_i)$
- Model fit: $R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$ where $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

Today: Extending simple linear regression

- What if we have more than one predictor: $X_1, X_2, \ldots$?

  $\rightarrow$ Multiple linear regression
- What if X_1 and X_2 interact, or if Y depends on X^2 instead of X?

  $\rightarrow$ Polynomial regression

Outline

- Multiple linear regression
- Key statistical questions in multiple linear regression
- Accommodating non-linear relationships

Motivation for multiple linear regression

Recall the `Advertising` dataset and the three separate simple linear regression lines:

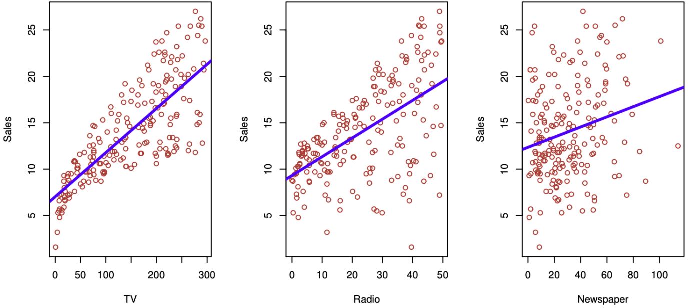

Figure: The `Advertising` data set: `Sales` of a product in 200 different markets against advertising budgets for three media: `TV`, `Radio`, and `Newspaper` [JWHT21, Figure 2.1].

Problem: Each simple linear regression line ignores the other two predictors

Question: Can we extend our analysis to accommodate *all* predictors simultaneously?

Multiple linear regression: Setup

We predict Y using multiple variables $X_1, X_2, \ldots, X_p$, assuming:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

- *Model parameters*: $\beta_0, \beta_1, \ldots, \beta_p$ are fixed, unknown constants
- ϵ is an error term, independent of $X_1, \ldots, X_p$

The coefficient β_j is interpreted as the *average effect* on Y of a unit increase in X_j, *holding all other predictors fixed*

Once we *estimate* $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ from training data, we can predict

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

Visualizing multiple linear regression

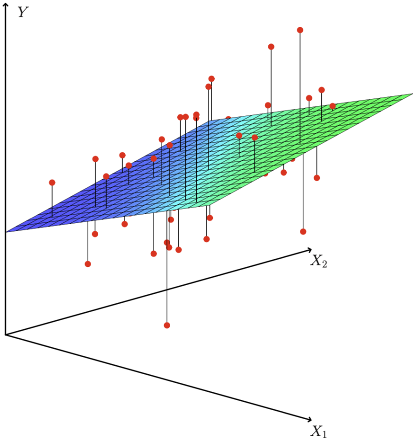

Figure: An illustration of multiple linear regression [JWHT21, Figure 3.4].

Coefficient estimation via least squares

β_0, $\beta_1, \ldots, \beta_p$ are unknown and must be estimated from data $(x_1, y_1), (x_, y_2), \ldots, (x_n, y_n)$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$

Again, we use the **least squares** criterion:

- The *least squares* approach chooses $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ to minimize the RSS:

$$\mathrm{RSS} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 x_i 1 + \hat{\beta}_2 x_i 2 + \cdots \hat{\beta}_p x_i p - y_i\right)^2$$

- The solutions have more complicated forms in this multiple-variable case[1]:
  - $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ where $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

[1]They can be derived by setting the partial derivatives of RSS to zero

Pop-up quiz: Coefficients in multiple lnear regression

Scenario: We fit a multiple linear regression model on the `Advertising` dataset:

$$\texttt{Sales} = \beta_0 + \beta_1\,\texttt{TV} + \beta_2\,\texttt{Radio} + \beta_3\,\texttt{Newspaper} + \varepsilon.$$

Suppose that we obtain:

$$\hat{\beta}_1 = 0.04, \quad \hat{\beta}_2 = 0.18, \quad \hat{\beta}_3 = -0.02.$$

Question: Which statement *best* describes the meaning of $\hat{\beta}_1 = 0.04$ in this model?

Multiple-choice answers:

A) TV advertising alone explains 4% of the variation in `Sales`.

B) For every additional dollar spent on TV, `Sales` increases by 0.04 units, assuming `Radio` and `Newspaper` are both zero.

C) For every additional dollar spent on TV advertising, `Sales` increases by 0.04 units on average, controlling for `Radio` and `Newspaper`.

D) If TV advertising goes up by $100, `Sales` is guaranteed to go up by 4 units, regardless of `Radio` or `Newspaper`.

Some key questions with multiple predictors

When we perform multiple linear regression, we often want to answer questions like:

- Are predictors $X_1, \ldots, X_p$ related to Y (i.e., do they help predict Y)?
- Which subset of $X_1, \ldots, X_p$ is most important?
- How well does the model fit the data?
- Given new predictor values, what response value should we predict and how accurate is that prediction?

Let's address these questions one by one

Hypothesis testing for relationship between Y and each X_j

Recall from simple linear regression that we conduct a hypothesis test using a t-statistic:

- $H_0 : \beta_1 = 0$ (no relationship) vs. $H_1 : \beta_1 \neq 0$ (some relationship)
- We reject H_0 or not, based on the value of

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

In multiple linear regression, we can do the same test to see if Y is related to each X_j

However:

- Would we get the same conclusions from simple vs. multiple regressions?
- What if we want to test whether Y is related to *any* of the X_j's?

Advertising example: Simple vs. multiple regressions

Q: Is newspaper useful in predicting sales?

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

Figure: Separate simple regressions suggest TV, radio, *and* newspaper are all significant [JWHT21, Tables 3.1 & 3.3].

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Figure: Multiple regression suggests newspaper is *not* significant [JWHT21, Table 3.4].

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Figure: Correlation matrix for TV, radio, newspaper, and sales [JWHT21, Table 3.5].

In multiple regression, β_j measures the effect of X_j on Y, *holding all other predictors fixed*

`Advertising` example: Single vs. any predictor

Q: Is *"any"* of `TV`, `radio`, `newspaper` useful in predicting `sales`?

We now test a different, *joint* hypothesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{vs.} \quad H_1 : \text{at least one } \beta_j \neq 0$$

This can be tested using the *F-statistic*:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}$$

Rationale: If H_0 is true,

- $\mathbb{E}[\mathrm{RSS}/(n-p-1)] = \mathbb{E}[\mathrm{TSS} - \mathrm{RSS})/p] = \sigma^2$
- F follows an F-distribution with $(p, n-p-1))$ degrees of freedom

$\Rightarrow$ If H_0 is true, F will likely take a typical value; if F is very large, then we reject H_0

Selecting "important" predictors[2]

Suppose we are confident that *at least some* predictors are related to Y

Variable selection: *"Which subset of predictors is most useful or important?"*

- Naive approach: Try all $2^p - 1$ possible combinations of predictors
  - Evaluate each model by some criterion
  - **Challenge:** Intractable for large p (exponential number of subsets)
- Practical approaches:
  - *Greedy methods:* Forward, backward, or stepwise (mixed) selection
  - *Regularization methods:* Modify the least squares criterion, e.g., LASSO

We will discuss these methods in more detail in future lectures

[2] We will revisit this question later

Evaluating the model fit

The quality of a multiple linear regression fit can be measured by the RSE or the R^2

- **Residual standard error (RSE)**: *"average deviation of Y from the regression line"*

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n-p-1}} \quad \text{where} \quad \mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **The $\mathbf{R^2}$:** "the proportion of variance in Y explained by X"

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}, \quad \text{where} \quad \mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

  - R^2 *always* increases when more predictors are added to the model
  - **"Adjusted"** R^2 compensates for adding predictors:

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{RSS}/(n-p-1)}{\mathrm{TSS}/(n-1)}$$

Pop-up quiz: R^2 vs. adjusted R^2

Scenario: We fit a model on $n = 100$ data points using a single predictor X_1:

$$R^2 = 0.80, \quad R^2_{\mathrm{adj}} = 0.79.$$

After adding a second predictor X_2 (suspected to be mostly noise), we get:

$$R^2 = 0.82, \quad R^2_{\mathrm{adj}} = 0.78.$$

Question: Why did R^2 go up while R^2_{adj} went down?

Multiple-choice answers:

a) There must be a calculation error; if R^2 increases, R^2_{adj} must also increase.
b) X_2 adds a tiny improvement to the fit by chance, raising R^2, but not enough to offset the penalty for extra parameters, so R^2_{adj} drops.
c) Adjusted R^2 always decreases whenever you add predictors, no matter how useful they are.
d) R^2 does not measure model fit at all, whereas R^2_{adj} is the only valid measure of fit.

Confidence intervals and prediction intervals

With $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, we predict $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$

How certain are we about this prediction?

- $\hat{y} = \hat{f}_{\hat{\beta}}(x)$ only *estimates* $f_\beta(x) = \beta_0 + \sum_{j=1}^p \beta_j x_j$.
- $y = f(x) + \epsilon$ has an error term, so additional variability.

Confidence interval for $f(x)$:

- Reflects uncertainty in prediction due to the coefficient estimates
- A 95% CI should contain $f(x)$ with probability 0.95

Prediction interval for y (given x):

- Includes both uncertainty in $\hat{y} = \hat{f}(x)$ and the noise ϵ
- A 95% PI should contain the actual $y = f(x) + \epsilon$ with probability 0.95

Exact formulas are beyond our scope, but in **R**:

```
predict(model, newdata = x0, interval = "confidence", level = 0.95)
```

What if there is a non-linear relationship?

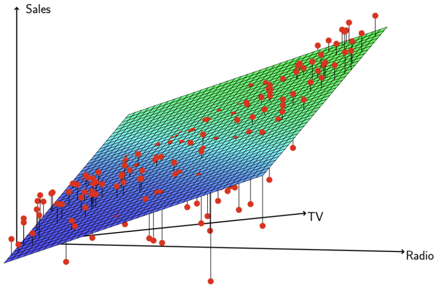

Figure: Pronounced synergy between TV and Radio; positive residuals cluster along the 45-degree line [JWHT21, Figure 3.5].

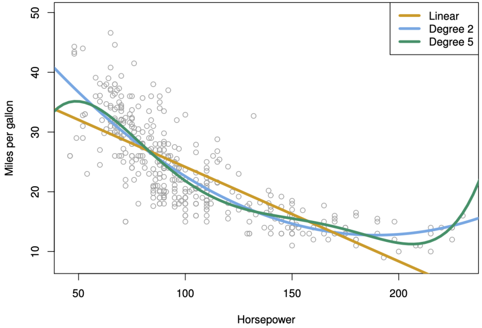

Figure: A non-linear relationship between mpg and horsepower is noticeable [JWHT21, Figure 3.8].

$\rightarrow$ We can add **interaction** terms (TV $\times$ Radio) or **non-linear** terms (horsepower^2) to capture these effects

Polynomial regression

Polynomial regression extends the linear model by including powers of predictors[3]:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_d X^d + \epsilon$$

- Treated as multiple linear regression on transformed predictors $(X, X^2, \ldots, X^d)$
- Although non-linear in X, the model is still linear in the coefficients β_j

Example: Interaction effect (synergy between `TV` and `Radio`)

$$\begin{aligned}\texttt{Sales} &= \beta_0 + \beta_1\texttt{TV} + \beta_2\texttt{Radio} + \beta_3 \underbrace{\texttt{TV} \times \texttt{Radio}}_{\text{interaction term}} + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3\texttt{Radio})\,\texttt{TV} + \beta_2\texttt{Radio} + \epsilon\end{aligned}$$

Example: Quadratic model

$$\texttt{mpg} = \beta_0 + \beta_1\texttt{horsepower} + \beta_2\texttt{horsepower}^2 + \epsilon$$

[3] More generally, $Y = \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|\le d} \beta_{\boldsymbol{\alpha}} \mathbf{X}^{\boldsymbol{\alpha}} + \epsilon$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ and $\mathbf{X}^{\boldsymbol{\alpha}} = X_1^{\alpha_1} X_2^{\alpha_2} \cdots X_p^{\alpha_p}$

Wrap-up

Multiple linear regression assumes a model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

with the parameters typically estimated by least squares

We can address the following questions:

- Is X_j related to Y? $\Rightarrow$ Test $H_0 : \beta_j = 0$
- Is any of $X_1, \ldots, X_p$ related to Y? $\Rightarrow$ Test $H_0 : \beta_1 = \cdots = \beta_p = 0$
- How does Y change per unit change in X_j (others fixed)? $\Rightarrow \hat{\beta}_j$
- How well does the model fit data? $\Rightarrow \mathrm{RSE}, R^2, R^2_{\mathrm{adj}}$
- How certain is our prediction of y? $\Rightarrow$ CI & PI
- What if there is a non-linear relationship? $\Rightarrow$ Add non-linear terms

Next lecture: Dummy variables, pitfalls in linear regression, etc.

References

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
An Introduction to Statistical Learning: with Applications in R, volume 112 of *Springer Texts in Statistics*.
Springer, New York, NY, 2nd edition, 2021.