

# **STA 35C: Statistical Data Science III**

## **Lecture 6: Multiple Linear Regression & Polynomial Regression**

Dogyoon Song

Spring 2026, UC Davis

# Recap: Simple linear regression

- Linear model assumption:  $Y = \beta_0 + \beta_1 X + \varepsilon$
- Least squares estimation chooses  $(\hat{\beta}_0, \hat{\beta}_1)$  minimizing

$$\text{RSS} = \sum_{i=1}^n (\hat{y}_i - y_i)^2 = \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i)^2$$

- Interpretation:  $Y$  tends to change by  $\beta_1$  on average per unit increase in  $X$ 
  - Associational, nor causal
- Inference:  $(\hat{\beta}_0$  and  $\hat{\beta}_1)$  are random variables
  - Rejecting  $H_0 : \beta_i = 0$  gives evidence of association
- The  $R^2$  is used to assess model fit:

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

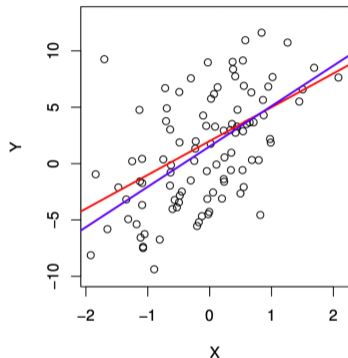


Figure: Least squares coefficient estimates from 100 data points. Red: population regression line, Blue: least squares line [JWHT21, Figure 3.3].

# Agenda

---

## Last time: Simple linear regression

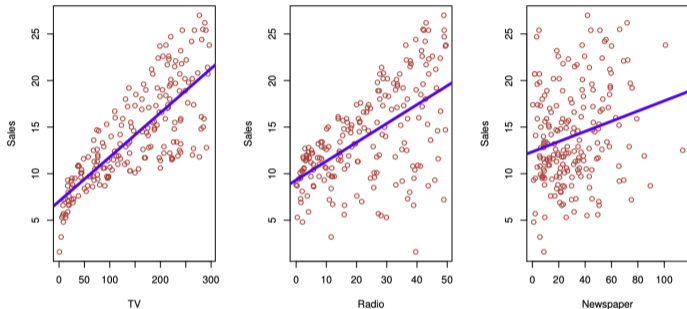
- Model:  $Y = \beta_0 + \beta_1 X + \epsilon$
- Least squares: estimate  $\beta_0, \beta_1$  by minimizing  $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- Inference on  $\beta_0, \beta_1$ : confidence intervals & hypothesis tests using  $SE(\hat{\beta}_i)$
- Model fit:  $R^2 = 1 - \frac{RSS}{TSS}$  where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

## Today: Extending simple linear regression

- Multiple linear regression
  - *"What if we have more than one predictor?"*  $X_1, X_2, \dots$
  - Joint significance, model fit, and prediction intervals
- Polynomial regression
  - *"What if  $X_1$  and  $X_2$  interact, or if  $Y$  depends on  $X^2$  instead of  $X$ ?"*

# 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  $X_1, X_2, \dots, X_p$ , assuming:

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

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

The coefficient  $\beta_j$  is interpreted as the the expected change (=average effect) in  $Y$  associated with a one-unit increase in  $X_j$ , *holding all other predictors fixed*

- This is an associational interpretation, not necessarily a causal one

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

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

# Visualizing multiple linear regression

---

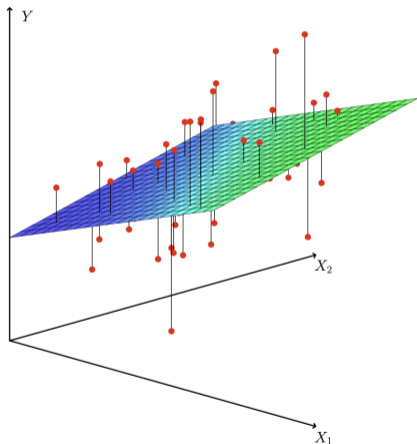


Figure: An illustration of multiple linear regression [JWHT21, Figure 3.4]. With two predictors, the fitted surface is a plane; with many predictors, it is a hyperplane.

## Coefficient estimation via least squares

---

Coefficients  $\beta_0, \beta_1, \dots, \beta_p$  must be estimated from data  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , where  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$

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

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

$$\text{RSS} = \sum_{i=1}^n (\hat{y}_i - y_i)^2 = \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip} - y_i)^2$$

- (Optional) The solutions have more complicated forms in this multiple-variable case<sup>1</sup>:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad \text{where} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

---

<sup>1</sup>This can be derived by setting the partial derivatives of RSS to 0, but is beyond the scope of STA 35C

## Pop-up quiz: Coefficients in multiple linear regression

---

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

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

Suppose that we obtain:  $\hat{\beta}_1 = 0.04$ ,  $\hat{\beta}_2 = 0.18$ ,  $\hat{\beta}_3 = -0.02$ .

**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**.

**Answer: C.**

## Some key questions with multiple predictors

---

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

- Are predictors  $X_1, \dots, X_p$  related to  $Y$  (i.e., do they help predict  $Y$ )?
- Which subset of  $X_1, \dots, X_p$  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 $X_j$

---

For each predictor  $X_j$ , we can test whether it is associated with  $Y$  after accounting for the other predictors:

$$H_0 : \beta_j = 0 \quad (\text{no relationship}) \quad \text{vs.} \quad H_1 : \beta_j \neq 0 \quad (\text{some relationship})$$

using the  $t$ -statistic

$$t = \frac{\hat{\beta}_j}{\widehat{\text{SE}}(\hat{\beta}_j)}$$

- If  $H_0$  is true,  $t$  will likely be close to 0; a very large  $|t|$  is evidence against  $H_0$

This asks whether  $X_j$  contributes to the model *holding the other predictors fixed*

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	<i>t</i> -statistic	<i>p</i> -value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

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

	Coefficient	Std. error	<i>t</i> -statistic	<i>p</i> -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	<i>t</i> -statistic	<i>p</i> -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,  $\beta_j$  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:

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

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

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n - p - 1)} \quad \begin{cases} \text{Reject } H_0 & \text{if } F \text{ is large,} \\ \text{Cannot reject } H_0 & \text{if } F \text{ is “typical”.} \end{cases}$$

**Rationale:** If  $H_0$  is true,

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

$\Rightarrow$  If  $H_0$  is true,  $F$  will likely be close to 1; a very large  $F$  is evidence against  $H_0$

## Selecting “important” predictors<sup>2</sup>

---

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

---

<sup>2</sup>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):** “*typical deviation of  $Y$  from the regression fit*”

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

- **The  $R^2$ :** “the proportion of variance in  $Y$  explained by  $X$ ”

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

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

$$R_{\text{adj}}^2 = 1 - \frac{\text{RSS}/(n - p - 1)}{\text{TSS}/(n - 1)}$$

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

---

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

$$R^2 = 0.80, \quad R_{\text{adj}}^2 = 0.79.$$

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

$$R^2 = 0.82, \quad R_{\text{adj}}^2 = 0.78.$$

**Question:** Why did  $R^2$  go up while  $R_{\text{adj}}^2$  went down?

**Multiple-choice answers:**

- A) There must be a calculation error; if  $R^2$  increases,  $R_{\text{adj}}^2$  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  $R_{\text{adj}}^2$  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  $R_{\text{adj}}^2$  is the only valid measure of fit.

**Answer: B.**

# Confidence intervals and prediction intervals

---

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

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 the mean response  $f(x)$ :**

- Reflects uncertainty in prediction due to estimating coefficients
- Across repeated samples, 95% of CIs would contain the true *mean response* at  $x$

**Prediction interval for a new response  $y$  (at  $x$ ):**

- Accounts for uncertainty in both  $\hat{y} = \hat{f}(x)$  *and* the random noise  $\epsilon$
- Across repeated samples, 95% of PIs would contain a new *observed response*  $y = f(x) + \epsilon$
- **Note:** The PI is always wider than the CI

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

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

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

# What if there is a non-linear relationship?

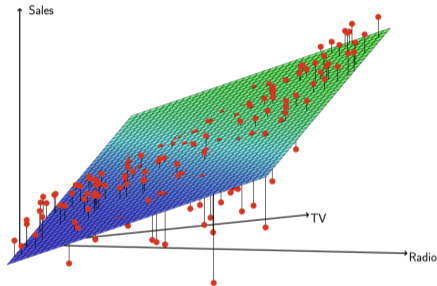


Figure: Pronounced synergy between **TV** and **Radio**; positive residuals cluster along the 45-degree line [JWHT21, Figure 3.5].

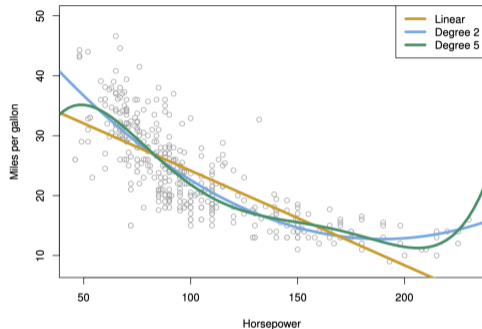


Figure: A non-linear relationship between **mpg** and **horsepower** is noticeable [JWHT21, Figure 3.8].

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

## Polynomial regression: Higher-order and interaction terms

---

**Polynomial regression** extends the linear model by including powers of predictors<sup>3</sup>:

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

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

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

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

**Example:** Quadratic model

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

---

<sup>3</sup>More generally,  $Y = \sum_{\alpha: |\alpha| \leq d} \beta_{\alpha} \mathbf{X}^{\alpha} + \epsilon$  where  $\alpha = (\alpha_1, \dots, \alpha_p)$  and  $\mathbf{X}^{\alpha} = X_1^{\alpha_1} X_2^{\alpha_2} \dots X_p^{\alpha_p}$

## Model flexibility and the risk of overfitting

---

A more flexible model can fit the training data more closely

- Adding predictors or higher-order terms can only decrease the training RSS

But a better fit on the training data does not necessarily mean a better model

- Our actual goal is good performance on new, unseen data
- A model that fits the training set extremely well may perform worse on a test set

**Overfitting:** If a model is too flexible, it can fit noise as well as signal

- Typical symptom: very small training error, but noticeably larger test error
- We will briefly explore this phenomenon in Homework 2, Problem 4
- We will return to this issue repeatedly throughout the course

## Wrap-up

---

Multiple linear regression extends simple linear regression to several predictors:

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

- Each coefficient  $\beta_j$  describes the association between  $X_j$  and the mean response, holding the other predictors fixed
- We can test one predictor at a time with  $t$ -tests, or test whether any predictors matter using the overall  $F$ -test
- Model fit and uncertainty are summarized by RSE,  $R^2$ , adjusted  $R^2$ , confidence intervals, and prediction intervals
- Interaction and polynomial terms make the model more flexible, but extra flexibility increases the risk of overfitting

**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.