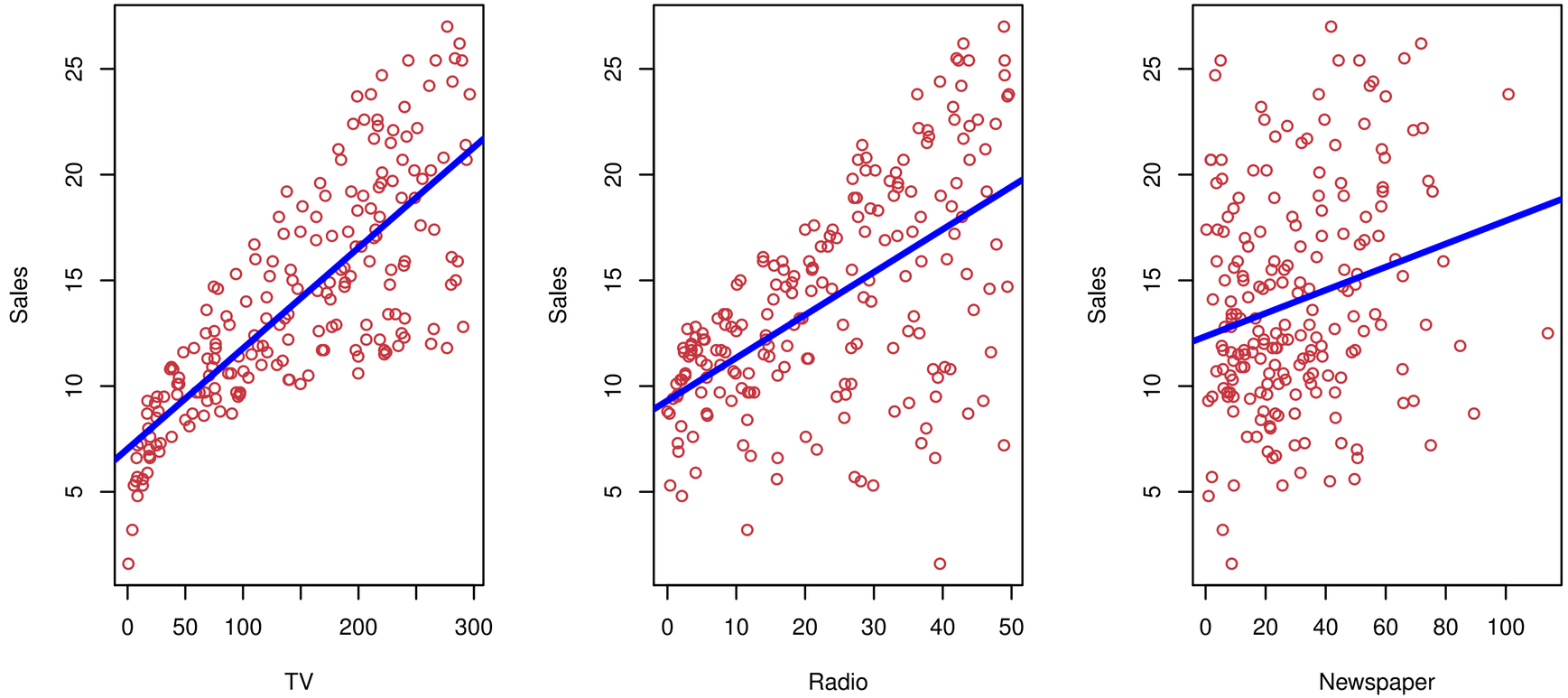
Lecture 5: Multiple Linear Regression

Wednesday, Oct 8

Context & Motivation

Problem this model addresses
Why we care

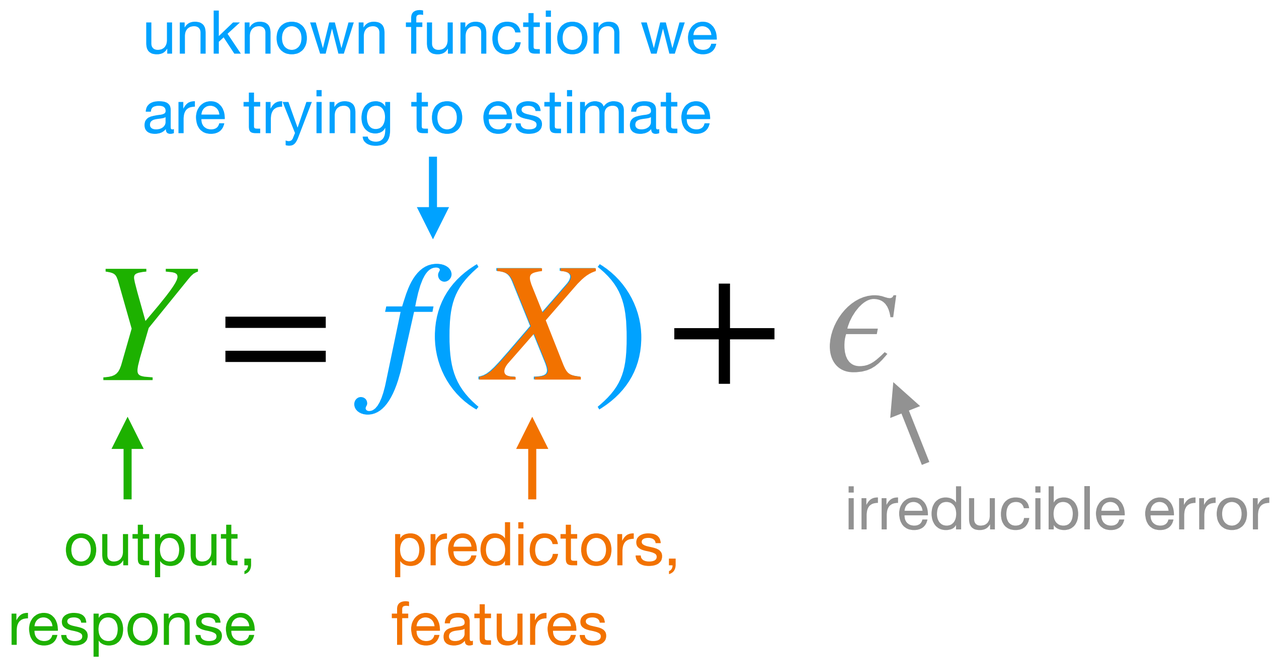
- **What happens when you have many predictors?**
- You could run three separate simple linear regressions…but potential issues:
  1. How to best predict if you have more than one model?
  2. Each model ignores other two potential predictors



Multiple linear regression: understanding the relationship between **many** variables and **one** outcome

Model Definition

Model type
Key equation(s): features, parameters, outputs



Model Definition

Model type
Key equation(s): features, parameters, outputs

unknown *parameters* we are trying to estimate

β_j quantifies the association between the predictor X_j and Y

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

output, response

predictors, features

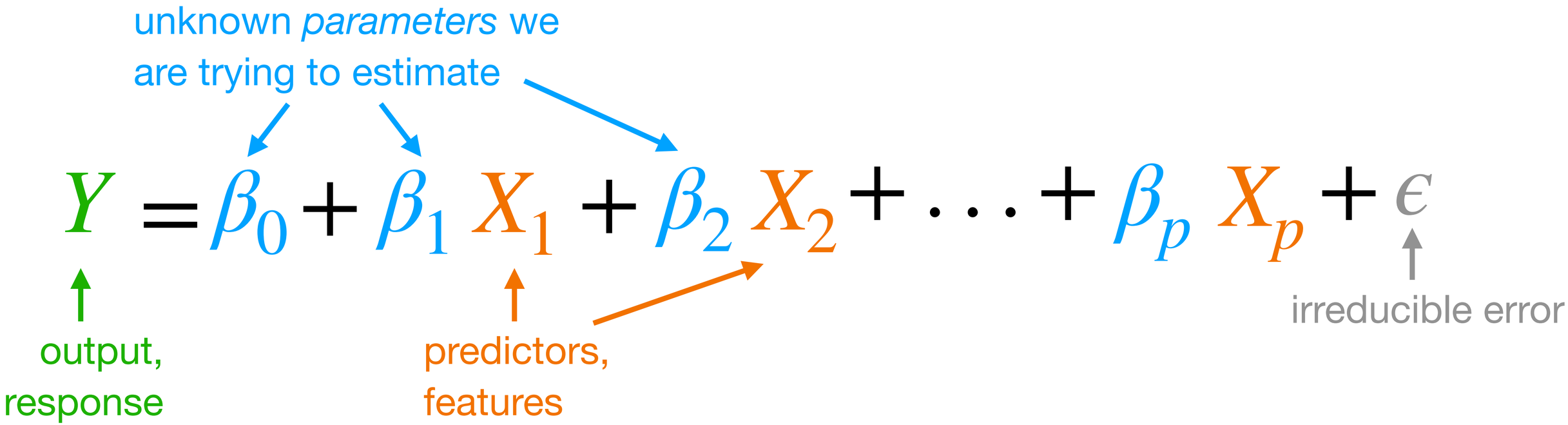
irreducible error

β_j: "slope" or expected change in Y for a 1-unit increase in X_j, holding all other predictors X_{-j} fixed

β_0: "y-intercept," or expected Y when $X_1, \ldots, X_p = 0$

Model Definition

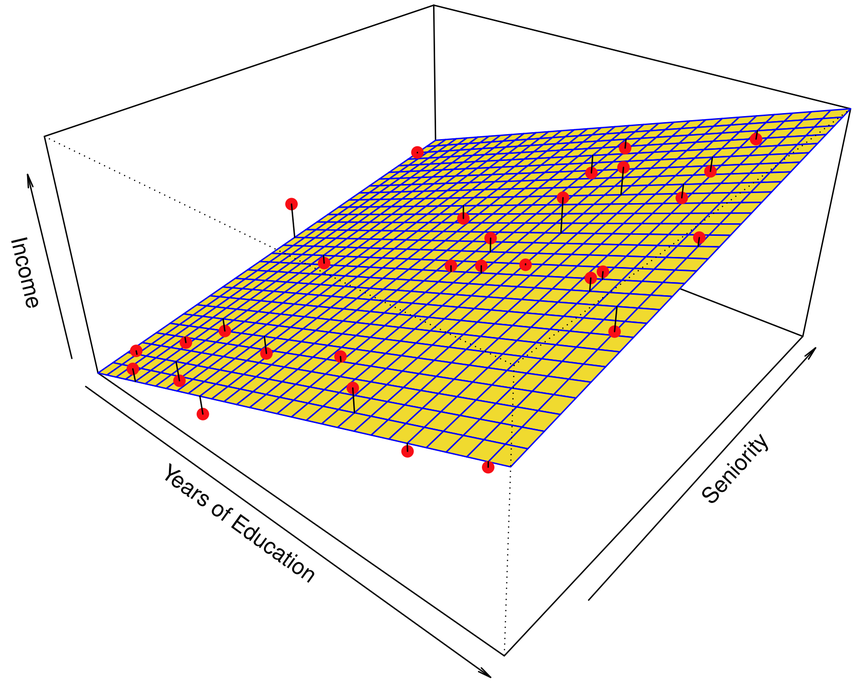
Model type
Key equation(s): features, parameters, outputs

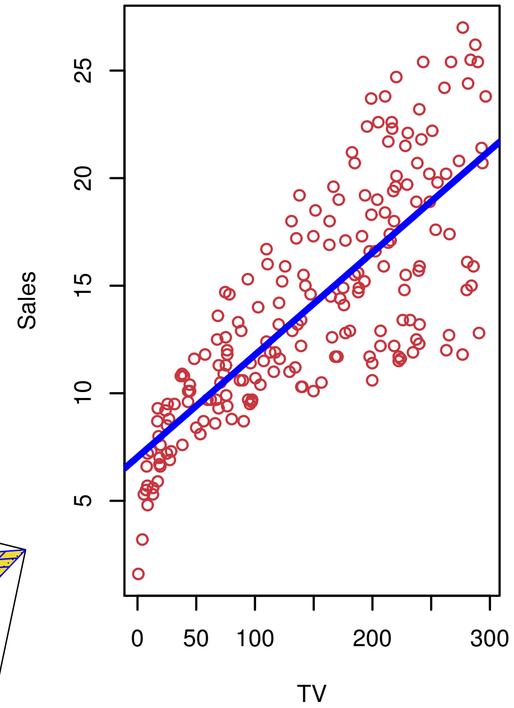


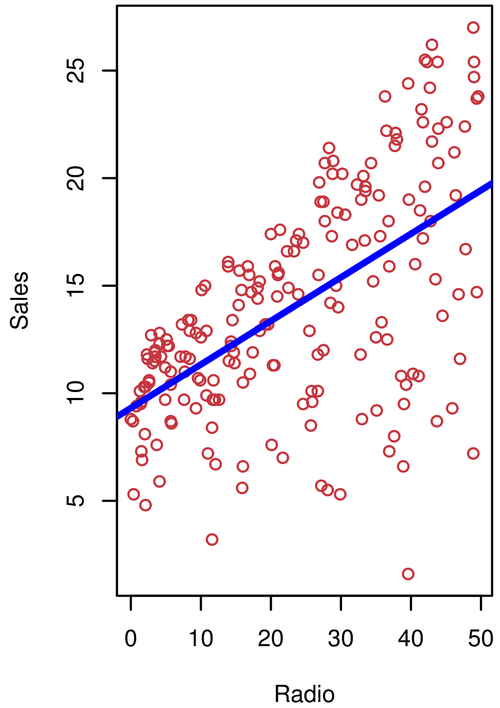
H_0: null hypothesis is that $\beta_1 = \beta_2 = \ldots = \beta_p = 0$; there is *no* relationship between Y and $X_{1:p}$

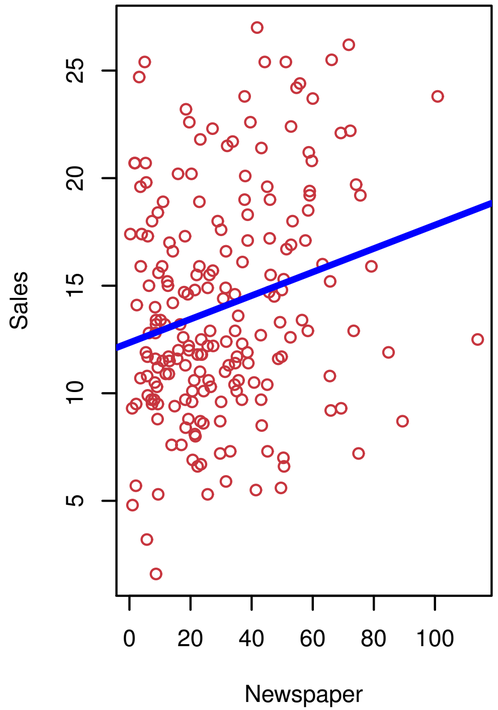
H_A: alternative hypothesis is that *at least one* $\beta_j \neq 0$; there is a *significant relationship* between Y and X_j

Examples







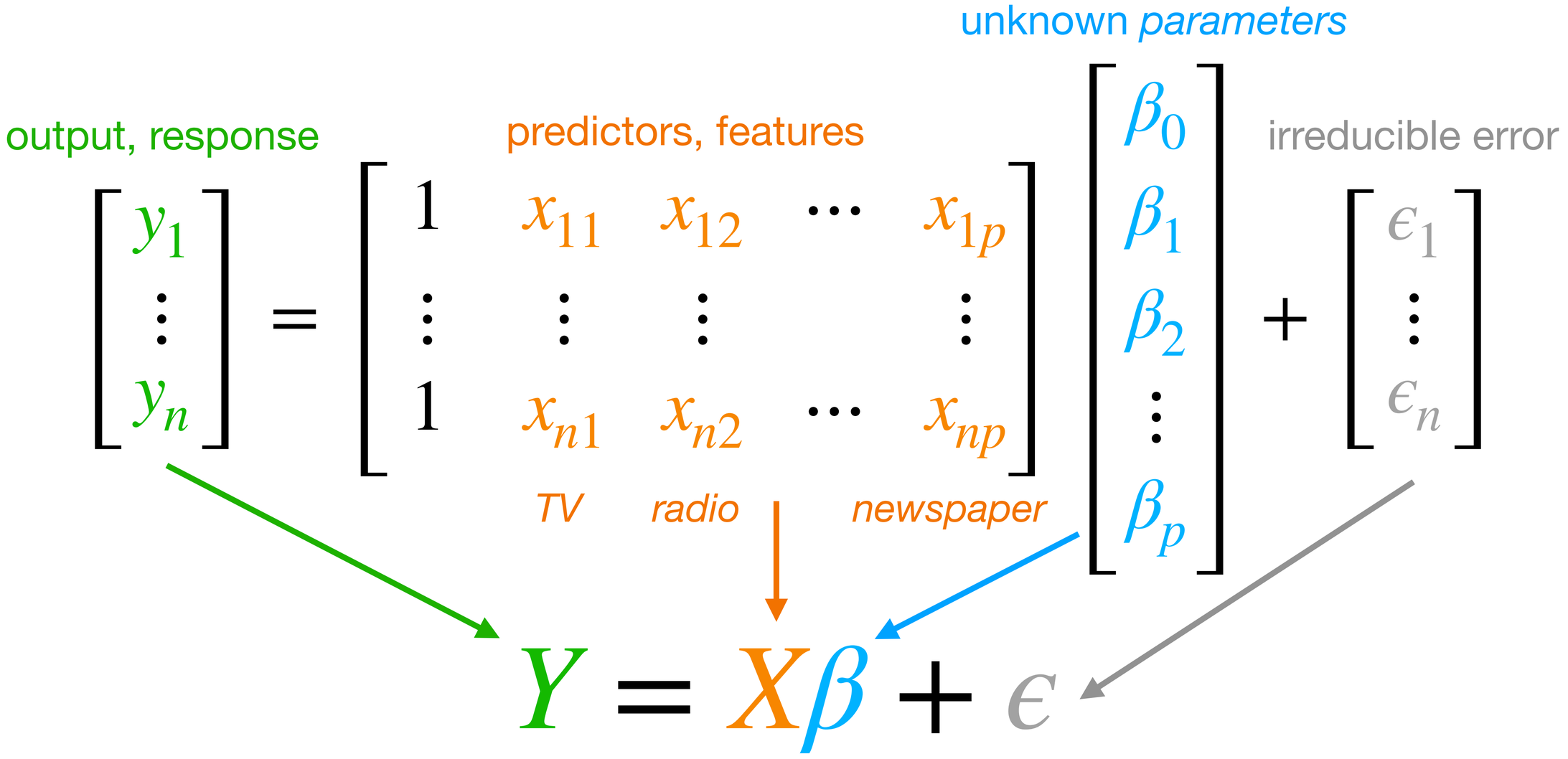


$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper}$$

$$\text{income} = \beta_0 + \beta_1 \times \text{education} + \beta_2 \times \text{seniority}$$

Model Definition

Model type
Key equation(s): features, parameters, outputs



Model Definition

Model type
Key equation(s): features, parameters, outputs

unknown *parameters* we are trying to estimate

$$Y = X\beta + \epsilon$$

output, response

predictors, features

irreducible error

Model specs

Regression	Classification
Supervised	Unsupervised
Parametric	Non-parametric
Prediction	Inference
Flexibility: LOW	Interpretability: HIGH

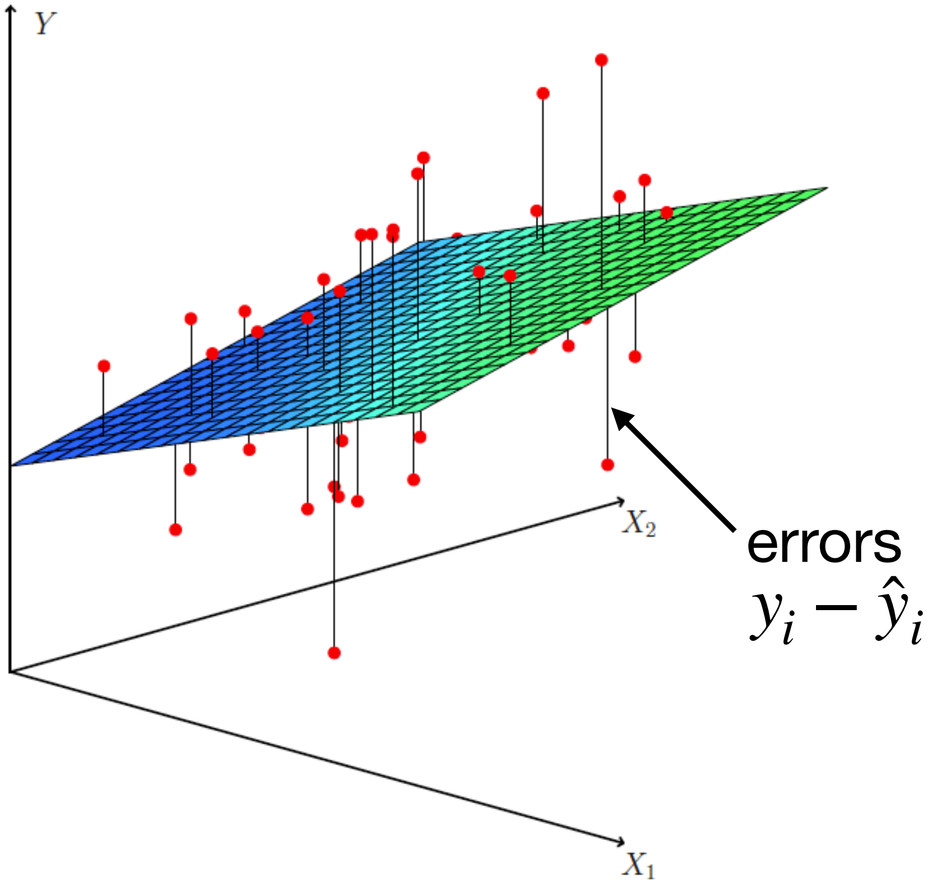
Model Fitting

How to fit the model to data
Parameter estimation
Tools / packages

Ordinary least squares (OLS) method:

Choose $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimizes the **residual sum of squares (RSS):**

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$



Model Fitting

How to fit the model to data
Parameter estimation
Tools / packages

To find the minimum of **RSS**, take partial derivative of RSS w.r.t. each β_j

$$\frac{\partial \text{RSS}}{\partial \beta_j} = \frac{\partial \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{\partial \beta_j} = \sum_{i=1}^{n} \left(x_{ij} - (y_i - \hat{y}_i)\right)^2 = 0$$

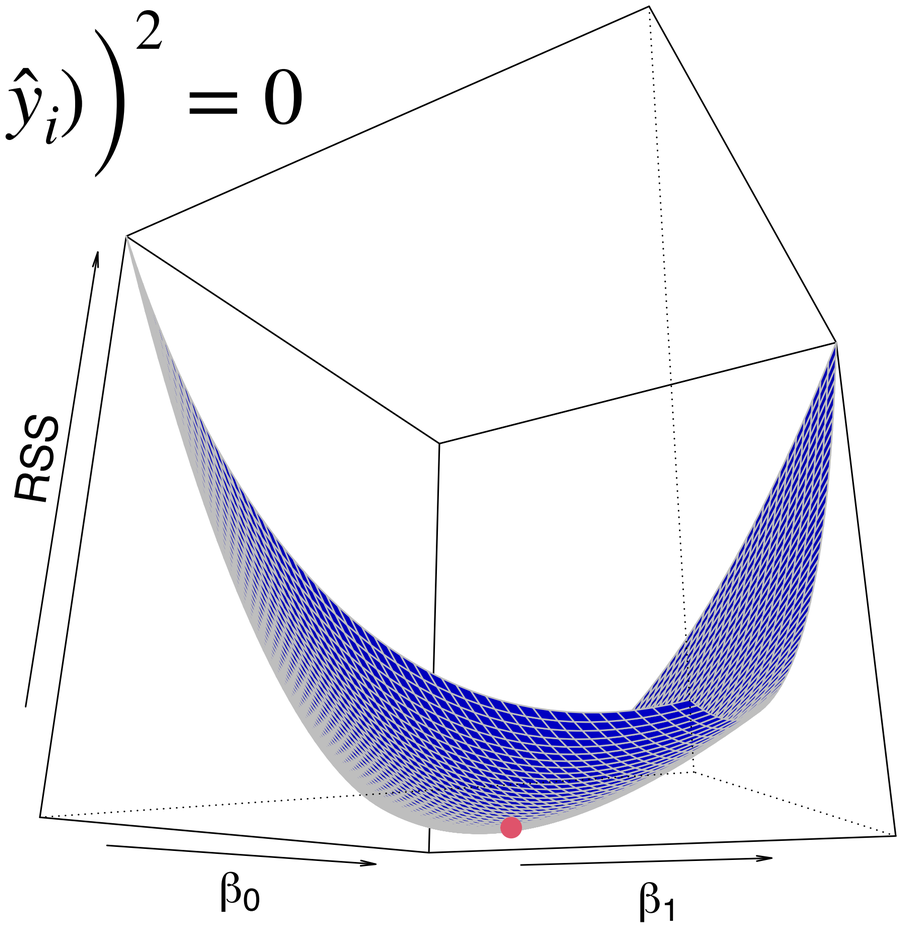
$$= X^T(Y - X\hat{\beta}) = 0 \quad \text{matrix form!}$$

$$= X^TY - (X^TX)\hat{\beta} = 0$$

$$\Rightarrow X^TY = (X^TX)\hat{\beta}$$

$$\Rightarrow \hat{\beta} = (X^TX)^{-1}(X^TY)$$

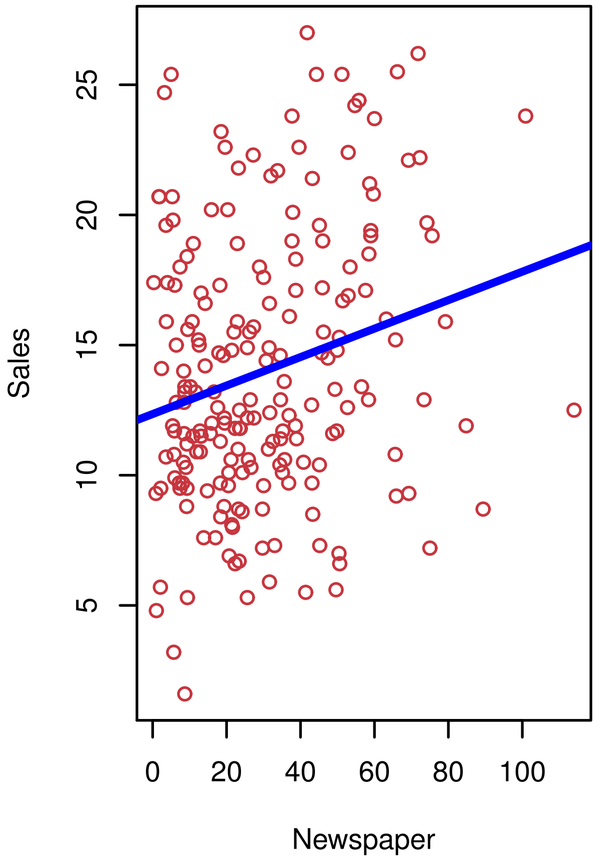
closed-form solution!
(use programs to compute)



Model Assessment

Goodness-of-fit and performance metrics
Model diagnostics

Why SLR ≠ MLR?



	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

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on TV, radio, and newspaper advertising budgets.*

	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

TABLE 3.5. *Correlation matrix for* TV, radio, newspaper, *and* sales *for the* Advertising *data.*

Assumptions

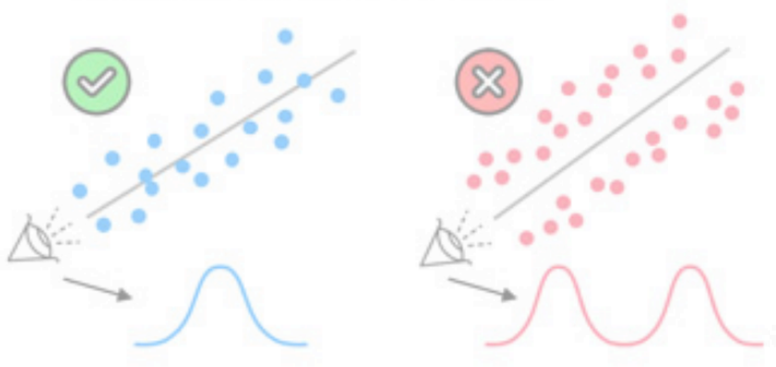
Core assumptions
When to use the model

Linearity:
linear relationship between X and Y

Independent errors:
errors are uncorrelated

Homoscedasticity:
errors have constant variance

Multivariate normality:
errors are normally distributed



No multicollinearity:
predictors are uncorrelated with each other

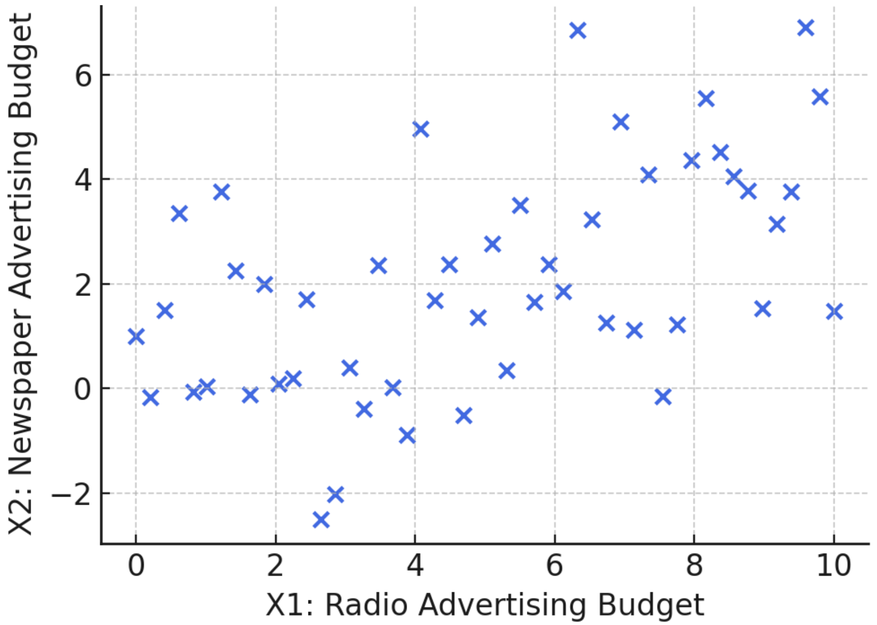
$X_1 \not\sim X_2$

$X_1 \sim X_2$

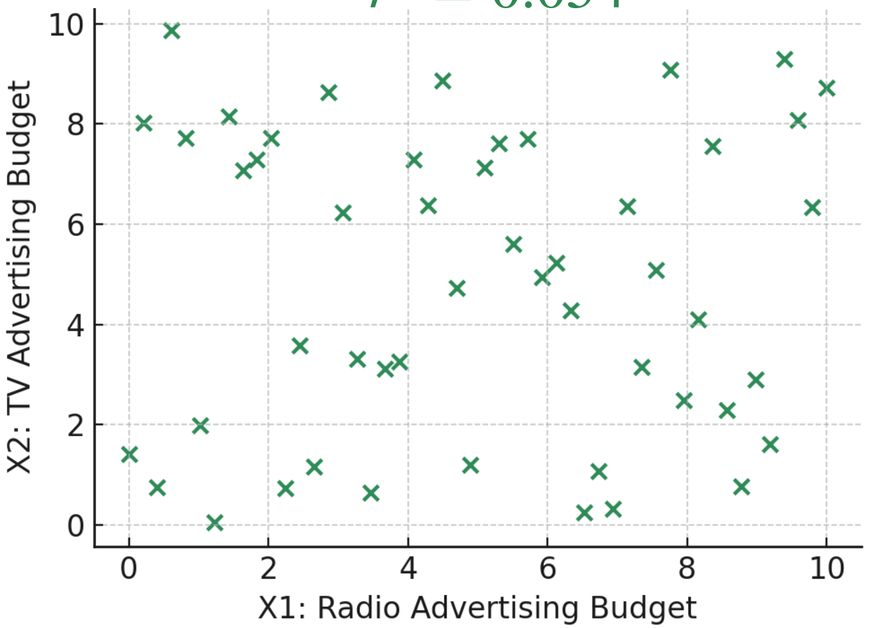
Model Assessment

Goodness-of-fit and performance metrics
Model diagnostics

$r^2 = 0.354$



$r^2 = 0.054$



Takeaway: check for multicollinearity in predictors!

	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

TABLE 3.5. *Correlation matrix for* TV, radio, newspaper, *and* sales *for the* Advertising *data.*

Model Assessment

Goodness-of-fit and performance metrics
Model diagnostics

Is at least one predictor useful in predicting the response?

H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$; *no* relationship between Y and $X_{1:p}$

H_A: *at least one* $\beta_j \neq 0$; *significant relationship* between Y and X_j

This hypothesis test is performed by computing the F-statistic: $F = \dfrac{(TSS - RSS)/p}{RSS/(n-p-1)}$

	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

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

⚠️ For large # of predictors, *p*-values can be <0.05 by chance!

⇒ F-stat adjusts for # of predictors

Model Assessment

Goodness-of-fit and performance metrics
Model diagnostics

Do all the predictors help to explain Y, or is only a subset of them useful?

- Exploring **all possible predictor combos** would take too long → total of 2^p models that contain subsets or combos of p predictors!
- **Variable selection:** process of determining which predictors to include in model

Forward selection

1. Start with null model (β_0 only).
2. Add the variable with the lowest RSS in a SLR.
3. Continue adding variables that give largest RSS improvement.
4. **Stop** when no remaining variable has a *p*-value below some threshold

Backward selection

Implement in HW2!

1. Start with full model (all βs).
2. Remove the variable with the **highest p-value** (least significant).
3. Refit the model and repeat, removing one variable at a time.
4. **Stop** when all remaining variables are significant ($p < 0.05$).

Model Assessment

Goodness-of-fit and performance metrics
Model diagnostics

How well does the model fit the data?

Residual standard error (RSE): $\text{RSE} = \sqrt{\frac{\text{RSS}}{n-p-1}}$ where $\text{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$

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

⇒ Same metrics as for simple linear regression

⚠️ **Caution:** R^2 always increases with more predictors (X's)

- Tiny gains can be meaningless ⇒ use adjusted R^2 instead:

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

Model Assessment

Goodness-of-fit and performance metrics
Model diagnostics

How well does the model fit the data?

RSS, RSE, R^2, Adjusted R^2 ⇒ all used to fit/assess **training data**

What about **test data?** ⇒ model's predictive power

⇒ compute **root mean squared error (RMSE)** on unseen **test set**

$$\text{test RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i^{test} - \hat{y}_i^{test}\right)^2}$$

Model Assessment

Goodness-of-fit and performance metrics
Model diagnostics

Metric	Computed On	Adjusted for Parameters	Used For	Interpretation	Units
RSS	Training data	❌ No	Model fitting	Total squared error between observed and predicted values. Smaller = better fit.	Y^2
RSE	Training data	✅ $n - p - 1$	Model assessment (train)	Typical size of a residual after adjusting for # of predictors.	Y
R^2	Training data	❌ No	Model assessment (train)	Proportion of variance in Y explained by X's. Higher = better fit.	Unitless
Adjusted R^2	Training data	✅ p	Model comparison (train)	Penalizes model complexity; increases only if added predictors significantly improve fit	Unitless
RMSE	Test data	❌ No	Predictive accuracy (test)	Average prediction error on unseen data. Lower = better generalization.	Y

Model Assessment

Goodness-of-fit and performance metrics
Model diagnostics

Takeaways:

- **RSS** is minimized to **fit** the model; all the other metrics are used to **assess** the model.
- **RSE** and **RMSE** are in the same units as Y (interpreted as “average error”).
- R^2 and **Adjusted** R^2 are unitless proportions of explained variance.
- **RSE** and **Adjusted** R^2 correct for the number of parameters.

Assessing train vs. test data:

- Used to assess **training data** → **RSS, RSE,** R^2, **Adjusted** R^2
- Used to assess **test data** → **RMSE**

Applications & Examples

```
import statsmodels.api as sm
X = …, y = …
X = sm.add_constant(X)
model = sm.OLS(y, X)
results = model.fit()
print(results.summary())
```

OLS Regression Results

Dep. Variable:	y	R-squared:	0.701
Model:	OLS	Adj. R-squared:	0.698
Method:	Least Squares	F-statistic:	251.5
Date:	Mon, 06 Oct 2025	Prob (F-statistic):	3.98e-138
Time:	05:12:52	Log-Likelihood:	-4951.0
No. Observations:	543	AIC:	9914.
Df Residuals:	537	BIC:	9940.
Df Model:	5		
Covariance Type:	nonrobust		

95% CI for coefficients

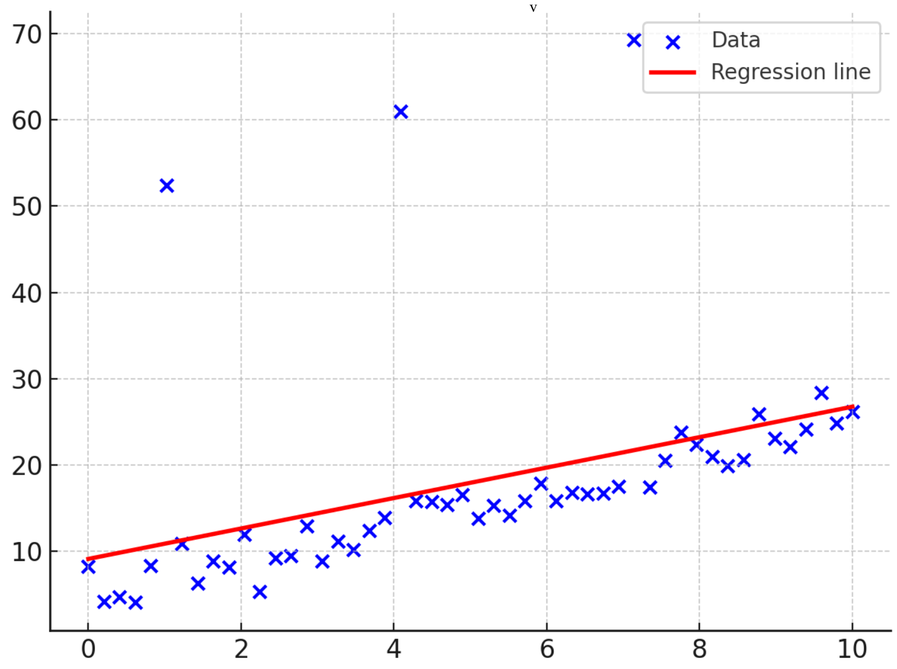
	coef	std err	t	P>|t|	[0.025	0.975]
const	539.5872	750.011	0.719	0.472	-933.729	2012.903
x1	1.2932	0.104	12.379	0.000	1.088	1.498
x2	-120.6634	29.708	-4.062	0.000	-179.021	-62.305
x3	0.2246	0.025	9.030	0.000	0.176	0.273
x4	32.4225	6.807	4.763	0.000	19.051	45.794
x5	76.3294	9.716	7.856	0.000	57.243	95.416

Strengths & Limitations (both SLR & MLR)

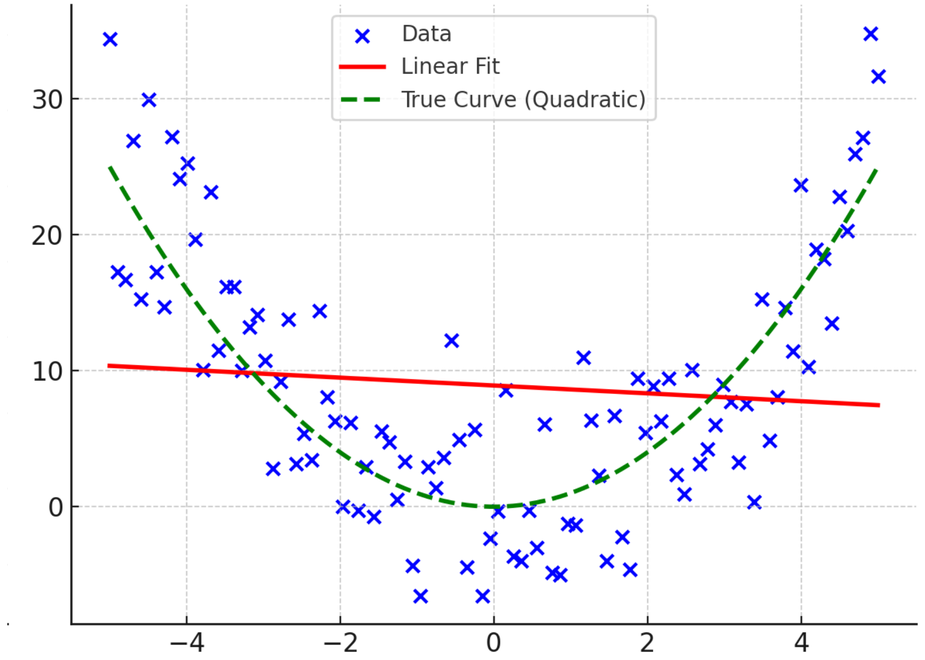
Strengths: simple, fast, interpretable, can use it for inference

Limitations: assumes linearity; sensitive to outliers; limited flexibility; can be misled by omitted predictors / variables

Extreme outliers



Misspecified model or underfitting on training set



Linear regression modeling workflow

1. Define the problem
2. Split data into train / test
3. Fit the model (training data)
4. Evaluate model fit (training data)
5. Evaluate predictive accuracy (test data)
6. Diagnose
   - Overfitting vs. underfitting
   - Variable selection: importance of individual predictors
   - Check model assumptions
7. Interpret model and communicate findings

Workflow guide on course website under “Week 2”!

Exam 1 Details — Monday, October 13th

- **14 MC questions + 2 FR questions** (~4-6 sub-questions/FR, just like quiz)
- 50 mins (9-9:50am) — arrive early!
- Covers Week 1-2 material
  - **To study:** review HW1-2, quizzes, student-submitted exam Q bank
  - **Quiz 2** is “optional,” but will probably help you prep!

Upcoming + Reminders

Updates:

- Project groups have been assigned! If you just joined the class, we will notify you by **Friday**.
- HW1 feedback is released — please take a look on **Gradescope**!
- Quiz 1 feedback will be released **today**

Assignments:

- #FinAid Survey—please complete it!

Friday's topic: ***Other Considerations in Regression***

- **Read:** ISLP Ch. 3.3

Office hours

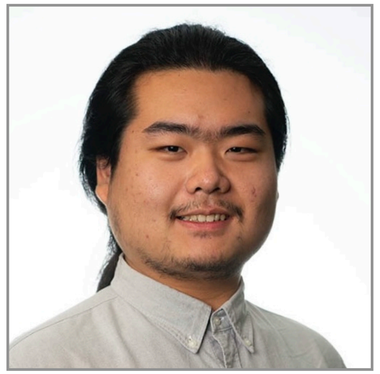
Instructor

Dr. Lucy Lai

CSB 244 or Zoom

- Wed, 3-4:30pm (book / walk-in)
- Thurs, 4:30-5:30pm (book only)
- Fri, 4-6p in WLH 2207 (occasionally)

TA



Jiesen Zhang

During all sections

All office hours can be booked on the course website!

TA/PLAs have office hours during section (last ~30 mins)! Anyone can go!

In-person

Zoom

PLAs

Johny Nguyen

- During W section
- W, 12-1pm (Zoom)

Parinita Saha

- During F, 4pm section
- F, 3-4pm (Zoom)

Exit ticket

Please answer a few questions before you go! 🙂