Multiple linear regression

Labor Economics

Instructor: Haoran LEI

Hunan University

Multiple linear regression

Model:

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

We interpret β_j as the **average effect** of a unit increase in X_j on Y, **holding all other predictors fixed** ("ceteris paribus").

Multiple linear regression

Model:

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

We interpret β_j as the **average effect** of a unit increase in X_j on Y, **holding all other predictors fixed** ("ceteris paribus").

In the advertising example:

$$\texttt{sales} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{radio} + \beta_3 \cdot \texttt{newspaper} + \epsilon$$

Interpreting regression coefficients

The ideal scenario is when all predictors are **uncorrelated**:

- An increase in the value of X_1 does not affect the value of X_2
- A trial/experiment design is called a **balanced design** in that case, and each coefficient can be estimated and tested separately

Interpretations such as *"a unit change in X_j is associated with a β_j change in Y, **while all the other variables stay fixed**"*, are possible.

Interpreting regression coefficients

- **Correlations amongst predictors** cause problems:
  1. The variance of all coefficients tends to increase, sometimes dramatically
  2. Interpretations become hazardous — when X_j changes, everything else changes.
- Claims of *causality* should be avoided for *observational data*
  - Indentifying causality is a big topic in economics.
    We'll touch on that topic later in this course.

The woes of (interpreting) regression coefficients

- The regression coefficient β_j estimates the expected change in Y per unit change in X_j, *with all other predictors held fixed*.
- When predictors are correlated, they *change together*!

Two Examples

1. Y = total amount of paper money in your pocket;
   X_1 = # of papers; X_2 = # of 10- and 20- RMB papers.
   By itself, regression coefficients of Y on X_2 will be > 0.
   But how about with X_1 in the model?

Two Examples

1. Y = total amount of paper money in your pocket;
   X_1 = # of papers; X_2 = # of 10- and 20- RMB papers.
   By itself, regression coefficients of Y on X_2 will be > 0.
   But how about with X_1 in the model?

2. Y = number of tackles by a football player in a season;
   W and H are his weight and height. Fitted regression model is
   $\hat{Y} = b_0 + 0.5W - 0.10H$. How do we interpret $\hat{\beta}_2 < 0$?

Interpreting regression coefficients

"The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively." -- Fred Mosteller and John Tukey

- These are said by statisticians. What can we (economists) do to deal with correlations between X_i's?

Interpreting regression coefficients

"The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively." -- Fred Mosteller and John Tukey

Economists Esther and Banerjee won Nobel Prize in 2019 for their usage of **'randomised control trials'** (**RCT**) in economics.

Estimation and Prediction for Multiple Regression

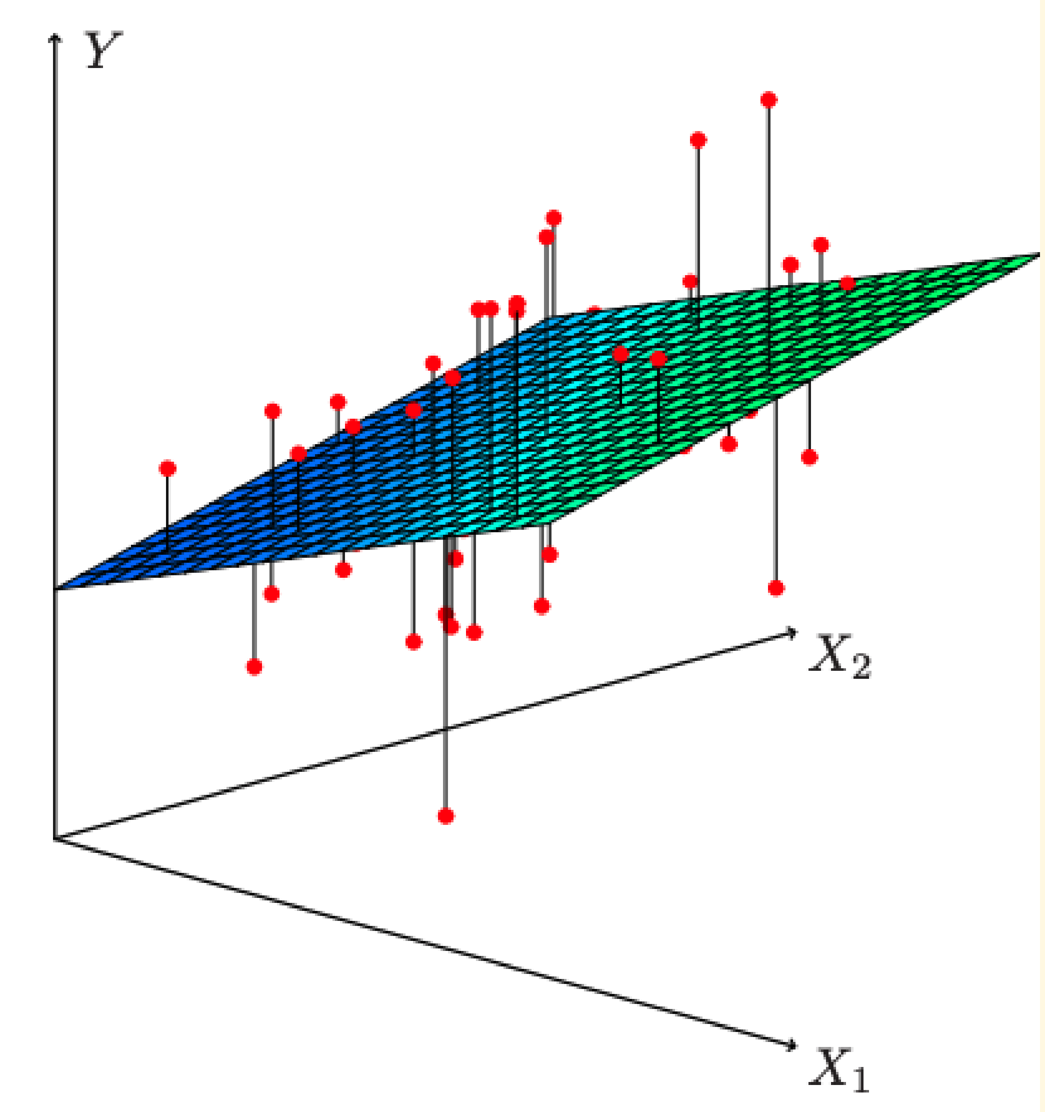
- Given estimates $\hat{\beta}_0$, $\hat{\beta}_1$,..., $\hat{\beta}_p$, we can make predictions using:

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

- LS estimators are obtained by minimizing the RSS:

$$\min_{\hat{\beta}_0, \ldots, \hat{\beta}_p} \text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Example: $p = 2$.



Results for advertising data

	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

Correlations:

	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

Some important questions

1. Is *at least one of the predictors* $X_1, X_2, ..., X_p$ useful in predicting the response?
2. Do *all the predictors* help to explain Y , or only *part of the predictors* useful?
3. How well does the linear model fit the data?

Q1: Is *at least one of the predictors* useful in predicting Y?

- We can use the F-statistic:

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

Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

Q2: Do *all the predictors* help to explain Y , or only *part of the predictors* useful?

- Essentially, this is to **decide on the important variables**.
- The most direct approach is called all subsets or best subsets regression:
  - we compute the least squares fit for all possible subsets;
  - then choose between them based on some criterion that balances training error with model size.
- However, usually we cannot examine all possible models. For example, when $p = 40$, there are $2^p \geq$ a billion models!

Forward selection

- Begin with the *null model* — a model that contains an intercept but no predictors.
- Fit p simple linear regressions and add to the null model the variable that results in the **lowest RSS**.
- Add to that model the variable that results in the **lowest RSS** amongst all two-variable models.
- Continue until **some stopping rule** is satisfied, for example when all remaining variables have a p-value above some threshold

Backward selection

- Start with *all variables* in the model.
- Remove the variable with the **largest p-value** — that is, the variable that is the least statistically significant.
- The new $(p-1)$-variable model is fit, and the variable with the **largest p-value** is removed.
- Continue until a **stopping rule** is reached. For instance, we may stop when all remaining variables have a **significant p-value** defined by some significance threshold.

Model selection

- Forward and Backward selections are two specialized cases of *model selection*.
- There are more systematic criteria for choosing an "optimal" member in the path of models produced by forward or backward stepwise selection.
  - Especially for time-series data.
- These include **Mallow's C_p**, **Akaike information criterion (AIC)**, **Bayesian information criterion (BIC)**, **adjusted R^2** and **Cross-validation (CV)**.

Q3. How well does the model fit the data?

- R^2 fails to be a good judge: adding more predictor variables always increases R^2!
- We need to have **some punishments** for those high R^2 cases with high p.

Q3. How well does the model fit the data?

- R^2 fails to be a good judge: adding more predictor variables always increases R^2!
- We need to have **some punishments** for those high R^2 cases with high p.

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

- where n is # of observations and p is # of predictors.

Other Considerations in the Regression Model

- Some predictors are not **quantitative** but are *qualitative*.
- Also called *categorical/factor* predictors:
  gender, student/martial status, ethnicity, ...

Motivating example: a credit card company (say Bank of China) has the following data about its clients:

- Balance, Age, Cards, Education, Income, Limit, Rating

Qualitative Predictors — continued

Example: investigate differences in credit card balance between males and females, ignoring the other variables.

A *dummy variable* for gender:

$$x_i = \begin{cases} 1 & \text{if i-th person is female} \\ 0 & \text{if i-th person is male} \end{cases}$$

Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

Interpretation?

Results for gender model:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	509.80	33.13	15.389	<0.0001
gender[Female]	19.73	46.05	0.429	0.6690

Qualitative predictors with more than two levels

- With more than two levels, we create additional dummy variables.
- For example, for the ethnicity variable (Asian/African/American) we create two dummy variables:
  - $x_{i1} = 1$ if i-th persion is Asian, or 0 otherwise;
  - $x_{i2} = 1$ if i-th persion is African, or 0 otherwise;
- There will always be one fewer dummy variable than the number of levels. The level with no dummy variable —American in this example — is known as the **baseline**.

Results for ethnicity

	Coefficient	Std. Error	t-stat.	p-value
Intercept	531.00	46.32	11.464	< 0.0001
ethnicity[Asian]	-18.69	65.02	-0.287	0.7740
ethnicity[African]	-12.50	56.68	-0.221	0.8260

Extensions of the Linear Model

Extensions of the Linear Model

- In the ads example, we have assumed that the effect on sales of increasing *one advertising medium* is independent of the amount spent on *the other media*. (**No Interactions**)
- However, suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases.
  - In econ (marketing), this is known as a complementary (synergy) effect
  - in statistics it is referred to as an **interaction effect**.

Modelling interactions — Advertising data

$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon\end{aligned}$$

- Implications: when the expenses on radio ads get higher, the **marginal benefit** of expenses on TV ads get higher!

Modelling interactions — Advertising data

Results as below. **Interpretation?**

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	<0.0001
TV	0.0191	0.002	12.70	<0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	<0.0001

Interpretation

Interactions are important:

- The p-value for the interaction term TV×radio is extremely low, indicating that there is strong evidence for $H_A : \beta_3 \neq 0$.
- The R^2 for the interaction model is 96.8%, compared to only 89.7% for the model that predicts sales using TV and radio without an interaction term.

Interpretation — continued

- This means that (96.8 - 89.7)/(100 - 89.7) = 69% of the variability in sales that remains after fitting the additive model has been explained by the interaction term.
- The coefficient estimates in the table suggest that an increase in TV advertising of $1,000 is associated with increased sales of $(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.
- An increase in radio advertising of $\$1,000$ will be associated with an increase in sales of $(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 29 + 1.1 \times \texttt{TV}$ units.

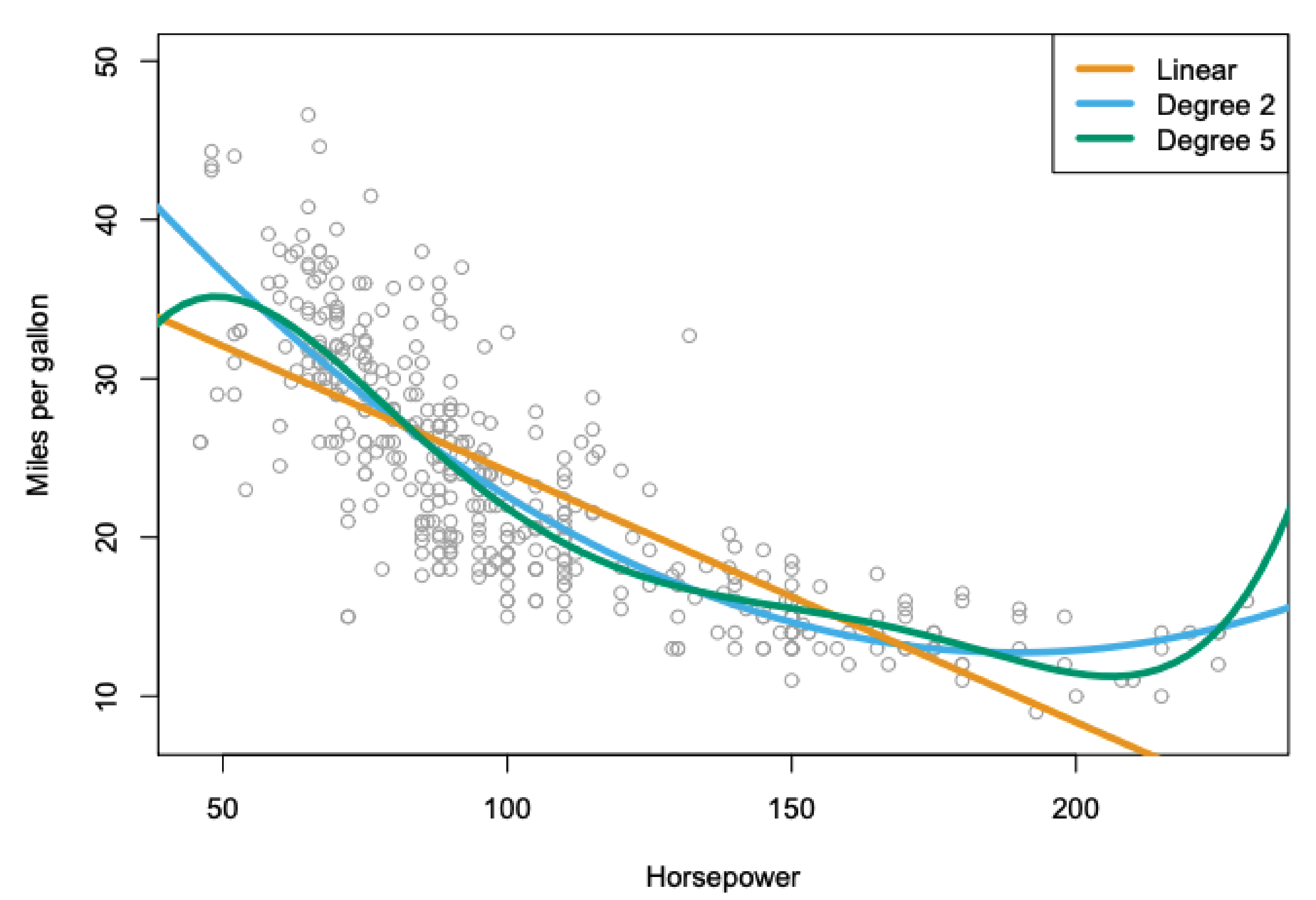
Hierarchy

- Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects (in this case, TV and radio) do not.
- The **hierarchy principle**:
  If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

Hierarchy

- The rationale for this principle is that interactions are hard to interpret in a model without main effects — their meaning is changed.
- Specifically, the interaction terms also contain main effects, if the model has no main effect terms.

Another extension: Incorporating non-linear effects



The figure suggests that

- mpg = β_0 + β_1 × horsepower + β_2 × horsepower^2 + ϵ

may provide a better fit.

The figure suggests that

- mpg = β_0 + β_1 × horsepower + β_2 × horsepower^2 + ϵ

may provide a better fit.

	Coefficient	Std. Error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	<0.0001
horsepower	-0.4662	0.0311	-15.0	<0.0001
horsepower^2	0.0012	0.0001	10.1	<0.0001

Generalizations of the Linear Model

- **Classification problems**: logistic regression, support vector machines
- **Non-linearity**: kernel smoothing, splines and generalized additive models, nearest neighbor methods.
- **Interactions**: Tree-based methods, bagging, random forests and boosting (these also capture non-linearities)
- **Regularized fitting**: Ridge regression and lasso