Lecture 5: Multiple Linear Regression

CS109A Introduction to Data Science

Pavlos Protopapas and Kevin Rader

Lecture Outline

Simple Regression:

- Predictor variables Standard Errors
- Evaluating Significance of Predictors
- Hypothesis Testing

Multiple Linear Regression:

- Categorical Predictors
- Collinearity
- Hypothesis Testing
- Interaction Terms

Standard Errors

The variances of β_0 and β_1 are also called their ***standard errors***, $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$.

If our data is drawn from a larger set of observations then we can empirically estimate the **standard errors**, $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$ of β_0 and β_1 through bootstrapping.

If we know the variance σ^2 of the noise ϵ, we can compute $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$ analytically, using the formulae below:

$$SE\left(\widehat{\beta}_0\right) = \sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_i \left(x_i - \overline{x}\right)^2}}$$

$$SE\left(\widehat{\beta}_1\right) = \frac{\sigma}{\sqrt{\sum_i \left(x_i - \overline{x}\right)^2}}$$

Standard Errors

In practice, we do not know the theoretical value of σ since we do not know the exact distribution of the noise ϵ. However, if we make the following assumptions,

- the errors $\epsilon_i = y_i - \hat{y}_i$ and $\epsilon_j = y_j - \hat{y}_j$ are uncorrelated, for $i \neq j$,
- each ϵ_i is normally distributed with mean 0 and variance σ^2,

then, we can empirically estimate σ^2, from the data and our regression line:

$$\sigma \approx \sqrt{\frac{n \cdot \mathrm{MSE}}{n-2}} = \sqrt{\frac{\sum_i (y_i - \widehat{y}_i)^2}{n-2}}$$

$$\sigma \approx \sqrt{\sum \frac{(\hat{f}(x) - y_i)^2}{n-2}}$$

Standard Errors

More data: $n \uparrow$ and $\sum_i (x_i - \bar{x})^2 \uparrow \Longrightarrow SE \downarrow$

Largest coverage: $var(x)$ or $\sum_i (x_i - \bar{x})^2 \uparrow \Longrightarrow SE \downarrow$

Better data: $\sigma^2 \downarrow \Rightarrow SE \downarrow$

$$SE\left(\widehat{\beta}_0\right) = \sigma \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_i \left(x_i - \overline{x}\right)^2}}$$

$$SE\left(\widehat{\beta}_1\right) = \frac{\sigma}{\sqrt{\sum_i \left(x_i - \overline{x}\right)^2}}$$

Better model: $(\hat{f} - y_i) \downarrow \Longrightarrow \sigma \downarrow \Longrightarrow SE \downarrow$

$$\sigma \approx \sqrt{\sum \frac{(\hat{f}(x) - y_i)^2}{n-2}}$$

Question: What happens to the $\widehat{\beta_0}$, $\widehat{\beta_1}$ under these scenarios?

Hypothesis Testing

Hypothesis testing is a formal process through which we evaluate the validity of a statistical hypothesis by considering evidence **for** or **against** the hypothesis gathered by ***random sampling of the data***.

Hypothesis Testing

Hypothesis testing is a formal process through which we evaluate the validity of a statistical hypothesis by considering evidence for or against the hypothesis gathered by random sampling of the data.

1. State the hypotheses, typically a **null hypothesis**, H_0 and an **alternative hypothesis**, H_1, that is the negation of the former.
2. Choose a type of analysis, i.e. how to use sample data to evaluate the null hypothesis. Typically this involves choosing a **single test statistic**.
3. **Compute** the test statistic.
4. Use the value of the test statistic to either **reject** or **not reject** the null hypothesis.

Hypothesis testing

1. State Hypothesis:

Null hypothesis:

H_0: There is no relation between X and Y

The alternative:

H_a: There is some relation between X and Y

2: **Choose test statistics**

To test the null hypothesis, we need to determine whether, our estimate for $\hat{\beta}_1$, is sufficiently far from zero that we can be confident that $\hat{\beta}_1$ is non-zero. We use the following test statistic:

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

Hypothesis testing

3. **Compute the statistics** :

   Using the estimated $\hat{\beta}, SE(\beta)$ we calculate the t-statistic.

4. **Reject or not reject the hypothesis:**

If there is really no relationship between X and Y , then we expect that will have a *t-distribution with n-2 degrees* of freedom.

To compute the probability of observing any value equal to $|t|$ or larger, assuming $\hat{\beta}_1 = 0$ is easy. We call this probability the p-value.

a small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance

Multiple Linear Regression

Multiple Linear Regression

If you have to guess someone's height, would you rather be told

- Their weight, only
- Their weight and gender
- Their weight, gender, and income
- Their weight, gender, income, and favorite number

Of course, you'd always want as much data about a person as possible. Even though height and favorite number may not be strongly related, at worst you could just ignore the information on favorite number. We want our models to be able to take in lots of data as they make their predictions.

Response vs. Predictor Variables

X
predictors
features
covariates

Y
outcome
response variable
dependent variable

n **observations**

TV	**radio**	**newspaper**	**sales**
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

p **predictors**

Multilinear Models

In practice, it is unlikely that any response variable Y depends solely on one predictor x. Rather, we expect that is a function of multiple predictors $f(X_1, \ldots, X_J)$. Using the notation we introduced last lecture,

$$Y = y_1, \ldots, y_n, \quad X = X_1, \ldots, X_J \text{ and } X_j = x_{1j}, \ldots, x_{ij}, \ldots, x_{nj}$$

In this case, we can still assume a simple form for f -a multilinear form:

$$Y = f(X_1, \ldots, X_J) + \epsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_J X_J + \epsilon$$

Hence, $\hat{f}$, has the form

$$\hat{Y} = \hat{f}(X_1, \ldots, X_J) + \epsilon = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_J X_J + \epsilon$$

Multiple Linear Regression

Again, to fit this model means to compute $\hat{\beta}_0, \ldots, \hat{\beta}_J$ or to minimize a loss function; we will again choose the MSE as our loss function.

Given a set of observations,

$$\{(x_{1,1}, \ldots, x_{1,J}, y_1), \ldots (x_{n,1}, \ldots, x_{n,J}, y_n)\},$$

the data and the model can be expressed in vector notation,

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_y \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,J} \\ 1 & x_{2,1} & \ldots & x_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,J} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_J \end{pmatrix},$$

Multiple Linear Regression

The model takes a simple algebraic form:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

Thus, the MSE can be expressed in vector notation as

$$\mathrm{MSE}(\beta) = \frac{1}{\mathrm{n}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

Minimizing the MSE using vector calculus yields,

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \mathrm{MSE}(\boldsymbol{\beta}).$$

Collinearity

Collinearity refers to the case in which two or more predictors are correlated (related).

We will re-visit collinearity in the next lectures, but for now we want to examine how does collinearity affects our confidence on the coefficients and consequently on the importance of those coefficients.

First let's look some examples:

Collinearity

Three individual models

TV

Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
6.679	0.478	13.957	2.804e-31	5.735	7.622
0.048	0.0027	17.303	1.802e-41	0.042	0.053

RADIO

Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
9.567	0.553	17.279	2.133e-41	8.475	10.659
0.195	0.020	9.429	1.134e-17	0.154	0.236

NEWS

Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
11.55	0.576	20.036	1.628e-49	10.414	12.688
0.074	0.014	5.134	6.734e-07	0.0456	0.102

One model

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
β_0	2.602	0.332	7.820	3.176e-13	1.945	3.258
β_{TV}	0.046	0.0015	29.887	6.314e-75	0.043	0.049
β_{RADIO}	0.175	0.0094	18.576	4.297e-45	0.156	0.194
β_{NEWS}	0.013	0.028	2.338	0.0203	0.008	0.035

Collinearity

Collinearity refers to the case in which two or more predictors are correlated (related).

We will re-visit collinearity in the next lectures, but for now we want to examine how does collinearity affects our confidence on the coefficients and consequently on the importance of those coefficients.

Assuming uncorrelated noise then we can show:

$$SE(\beta_1) = \sigma^2 (XX^T)^{-1}$$

Finding Significant Predictors: Hypothesis Testing

For checking the significance of linear regression coefficients:

1.we set up our hypotheses H_0:

$$H_0 : \beta_0 = \beta_1 = \ldots = \beta_J = 0 \qquad \textbf{(Null)}$$
$$H_1 : \beta_j \neq 0, \text{ for at least one } j \qquad \textbf{(Alternative)}$$

2. we choose the *F*-stat to evaluate the null hypothesis,

$$F = \frac{\text{explained variance}}{\text{unexplained variance}}$$

Finding Significant Predictors: Hypothesis Testing

3. we can compute the *F*-stat for linear regression models by

$$F = \frac{(\text{TSS} - \text{RSS})/J}{\text{RSS}/(n - J - 1)}, \quad \text{TSS} = \sum_i (y_i - \overline{y}), \text{RSS} = \sum_i (y_i - \widehat{y}_i)$$

4. If $F = 1$ we consider this evidence for H_0; if $F > 1$, we consider this evidence against H_0.

Qualitative Predictors

So far, we have assumed that all variables are quantitative. But in practice, often some predictors are **qualitative**.

Example: The Credit data set contains information about balance, age, cards, education, income, limit , and rating for a number of potential customers.

Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
14.890	3606	283	2	34	11	Male	No	Yes	Caucasian	333
106.02	6645	483	3	82	15	Female	Yes	Yes	Asian	903
104.59	7075	514	4	71	11	Male	No	No	Asian	580
148.92	9504	681	3	36	11	Female	No	No	Asian	964
55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331

Qualitative Predictors

If the predictor takes only two values, then we create an **indicator** or **dummy variable** that takes on two possible numerical values.

For example for the gender, we create a new variable:

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

We then use this variable as a predictor in the regression equation.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male} \end{cases}$$

Qualitative Predictors

Question: What is interpretation of β_0 and β_1?

Qualitative Predictors

Question: What is interpretation of β_0 and β_1?

- β_0 is the average credit card balance among males,
- $\beta_0 + \beta_1$ is the average credit card balance among females,
- and β_1 the average difference in credit card balance between females and males.

Exercise: Calculate β_0 and β_1 for the Credit data.

You should find $\beta_0 \sim \$509, \beta_1 \sim \$19$

More than two levels: One hot encoding

Often, the qualitative predictor takes more than two values (e.g. ethnicity in the credit data).

In this situation, a single dummy variable cannot represent all possible values.

We create additional dummy variable as:

$$x_{i,1} = \begin{cases} 1 & \text{if } i \text{ th person is Asian} \\ 0 & \text{if } i \text{ th person is not Asian} \end{cases}$$

$$x_{i,2} = \begin{cases} 1 & \text{if } i \text{ th person is Caucasian} \\ 0 & \text{if } i \text{ th person is not Caucasian} \end{cases}$$

More than two levels: One hot encoding

We then use these variables as predictors, the regression equation becomes:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i \text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i \text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i \text{th person is AfricanAmerican} \end{cases}$$

Again the interpretation

Beyond linearity

In the Advertising data, we assumed that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media.

If we assume linear model then the average effect on sales of a one-unit increase in TV is always β_1, regardless of the amount spent on radio.

Synergy effect or **interaction effect** states that when an increase on the radio budget affects the effectiveness of the TV spending on sales.

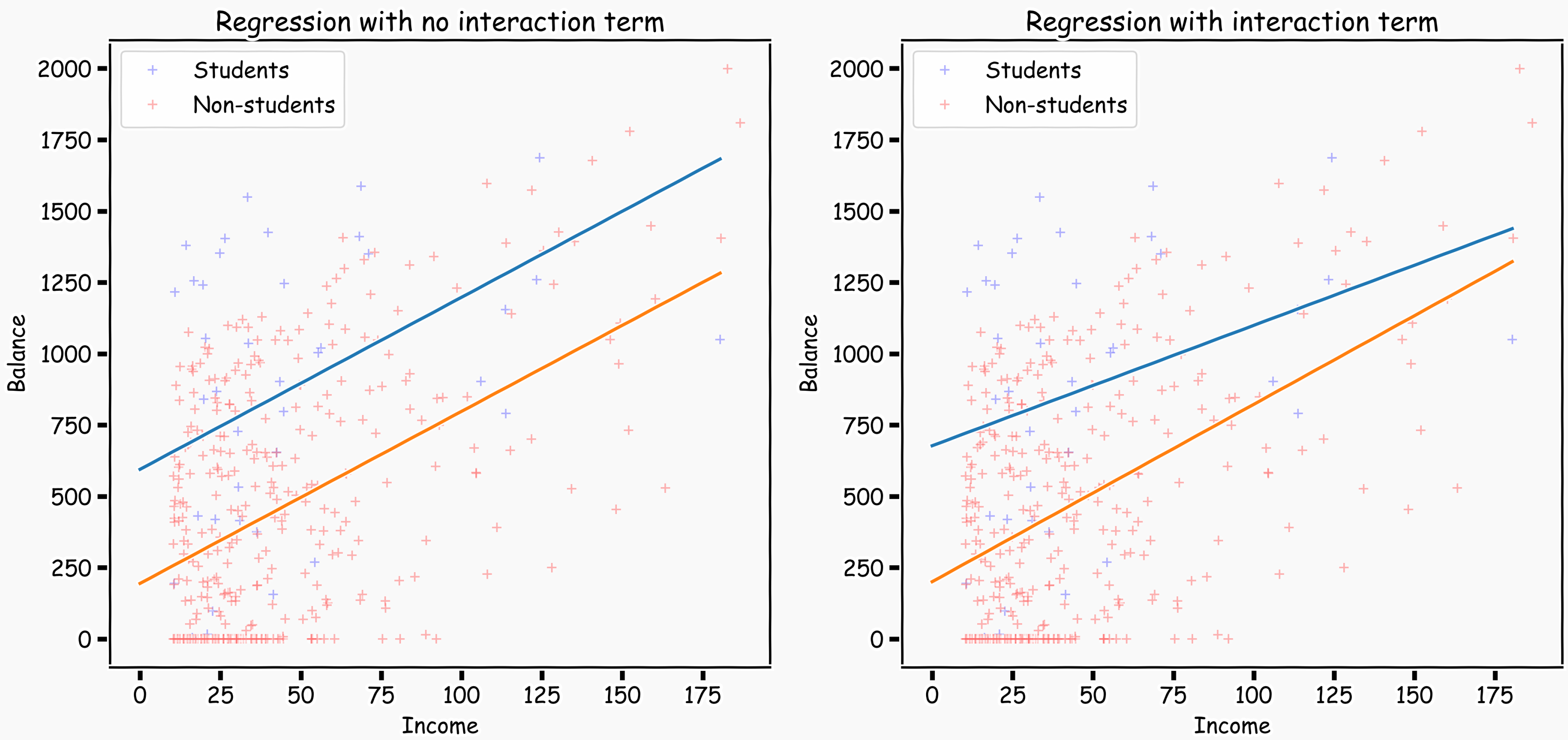
Beyond linearity

We change

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

To

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boxed{\beta_3 X_1 X_2} + \epsilon$$



What does it mean?

Predictors predictors predictors

We have a lot predictors!

Is it a problem?

- Yes: Computational Cost
- Yes: Overfitting

Wait there is more …