# Introduction to Statistical Machine Learning with Applications in Econometrics

## Lecture 3: Linear Regression (ISL ch. 3)

Instructor: Ma, Jun

Renmin University of China

October 14, 2021

# Linear regression

- Linear regression is a simple approach to supervised learning. In particular, linear regression is a useful tool for predicting a quantitative response.
- The `Advertising` data has sales as the response ($Y$) and advertising budgets for TV ($X_1$), radio ($X_2$), and newspaper media ($X_3$) as predictors. A statistical model: $Y = f(X) + \epsilon$ with $\epsilon$ being independent of $X = (X_1, X_2, X_3)^\top$.
- Interesting questions:
  - Is there a relationship between advertising budget and sales? (Is $f(x_1, x_2, x_3) = \mathrm{E}[Y \mid X_1 = x_1, X_2 = x_2, X_3 = x_3]$ constant?)
  - How strong is the relationship between advertising budget and sales? (Variance of $\epsilon$?)
  - Which media contribute to sales? (Partial derivatives of $f(x_1, x_2, x_3)$?)
  - How accurately can we predict future sales? (MSE of prediction for an unseen data point.)
  - Is the relationship ($f(x)$) linear?
  - Is there synergy (interaction) among the advertising media? ($\partial f(x_1, x_2, x_3)/\partial x_1$ depends on $(x_2, x_3)$?)

- linear regression model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
    - $\epsilon$ is the error term that is independent of $X$.
    - $\beta_0$ and $(\beta_1, \beta_2, \beta_3)$ are intercept and slopes, which are also called coefficients.
- From the prediction perspective, essentially the model specifies a functional form for $f(X)$ and recovering $f$ reduces to recovering the coefficients.
- From the causal inference perspective, essentially the model assumes that the effects are constant and there is no endogeneity issue.

# Simple linear regression

- Simple linear regression model with a single predictor $X$: $Y = \beta_0 + \beta_1 X + \epsilon$.
- We have the training data:

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n).$$

- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the coefficients, for the unseen data point $(X_0, Y_0)$, we predict $Y_0$ using $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$.
- Let $\hat{Y}_i = b_0 + b_1 X_i$ be the in-sample prediction for $Y_i$ based on the $i$-th value of $X_i$.
- $e_i = Y_i - \hat{Y}_i$ represents the $i$-th residual and we the residual sum of squares (RSS) as

$$\begin{aligned} \text{RSS} &= e_1^2 + e_2^2 + \cdots + e_n^2 \\ &= (Y_1 - b_0 - b_1 X_1)^2 + (Y_2 - b_0 - b_1 X_2)^2 + \\ &\quad \cdots + (Y_n - b_0 - b_1 X_n)^2. \end{aligned}$$

- The least squares approach chooses $b_0$ and $b_1$ to minimize the RSS. The minimizing values can be shown to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$$

and $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$, where $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$ and $\overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i$.

# Assessing the accuracy

- The standard error of an estimator reflects how it varies under repeated sampling:

$$\mathrm{SE}\left(\hat{\beta}_0\right)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}\right] \text{ and } \mathrm{SE}\left(\hat{\beta}_1\right)^2 = \frac{\sigma^2}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2},$$

  where $\sigma^2 = \mathrm{Var}\left[\epsilon\right]$.
- In general, $\sigma^2$ is not known, but can be estimated from the data.
- The estimate of $\sigma$ ($\hat{\sigma}$) is known as the residual standard error:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2},$$

  where the residual sum of squares: $\mathrm{RSS} = \sum_{i=1}^n \left(Y_i - \hat{Y}_i\right)^2$.

- Standard errors

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_0\right)^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}\right] \text{ and } \widehat{\mathrm{SE}}\left(\hat{\beta}_1\right)^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}.$$

  can be used to compute confidence intervals.
- A 95% confidence interval is defined as an interval such that with 95% probability, the interval contains the true unknown value of the parameter.
- Approximately, with 95% probability

$$\left[\hat{\beta}_1 - 2 \cdot \widehat{\mathrm{SE}}\left(\hat{\beta}_1\right), \hat{\beta}_1 + 2 \cdot \widehat{\mathrm{SE}}\left(\hat{\beta}_1\right)\right]$$

  contains $\beta_1$, in a hypothetical scenario where we have repeated samples.

# Hypothesis testing

- Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of

$$H_0\text{: There is no relationship between } X \text{ and } Y$$

against the alternative hypothesis

$$H_a\text{: There is some relationship between } X \text{ and } Y.$$

- This corresponds to testing $H_0 : \beta_1 = 0$ again $H_a : \beta_1 \neq 0$.
- We compute a $t$-statistic, given by

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

which has a $t$-distribution with $n - 2$ degrees of freedom.
- $p$-value: the probability of observing any value equal to $|t|$ or larger.

# Assessing the overall accuracy

- RSE is considered a measure of the lack of (in-sample) fit of the model to the data.
  - If the (in-sample) predictions $\hat{Y}_i$ are very close to the true outcome values $Y_i$, RSE will be small.
  - If $\hat{Y}_i$ is very far from $Y_i$ for one or more observations, then the RSE may be quite large.
- $R^2$: the fraction of variance of $Y$ explained by the model:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}},$$

  where $\text{TSS} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$ is the total sum of squares.
- In simple linear regression, $R^2$ is the square of the sample correlation of $X$ and $Y$:

$$R^2 = \left\{ \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \sqrt{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}} \right\}^2 .$$

# Multiple linear regression

- The multiple linear regression:

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

- We interpret $\beta_j$ as the average effect on $Y$ of a one unit increase in $X_j$, holding all other predictors fixed.
- Our training data:

$$\left\{ \left(Y_1, X_{1,1}, ..., X_{p,1}\right), \left(Y_2, X_{1,2}, ..., X_{p,2}\right), ..., \left(Y_n, X_{1,n}, ..., X_{p,n}\right) \right\}.$$

- Given estimates $b_0, b_1, ..., b_p$, we make in-sample predictions using:

$$\hat{Y}_i = b_0 + b_1 X_{1,i} + b_2 X_{2,i} + \cdots + b_p X_{p,i}.$$

- The values $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ that minimize RSS are the multiple least squares regression coefficient estimates:

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \\ &= \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_{1,i} - b_2 X_{2,i} - \cdots - b_p X_{p,i}\right)^2. \end{aligned}$$

| | 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 |

ISL Table 3.3: simple regression of `sales` on `newspaper`

| | 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 |

ISL Table 3.4: multiple regression

- The `newspaper` simple regression coefficient estimate was significantly non-zero, the multiple regression coefficient estimate for `newspaper` is close to zero, and the corresponding $p$-value is no longer significant.

| | 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 |

ISL Table 3.5

- The sample correlation between `radio` and `newspaper` is 0.35. Markets with high `newspaper` advertising tend to also have high `radio` advertising.
- Suppose that the multiple regression is correct and `newspaper` advertising is not associated with `sales`, but `radio` advertising is associated with `sales`.
- In a simple linear regression, we will observe that higher values of `newspaper` tend to be associated with higher values of `sales`, even though `newspaper` advertising is not directly associated with `sales`.

- Important questions:
  - Is at least one of the predictors $X_1, X_2, ..., X_p$ useful in predicting the response? (Model significance test.)
  - Do all the predictors help to explain $Y$, or is only a subset of the predictors useful? (Model selection will be discussed later in the class.)
  - How well does the model fit the data? (In-sample fit, measured by $R^2$.)
  - Given a set of predictor values, what response value should we predict, and how accurate is our prediction? (MSE of prediction for an unseen data point; is the linear model good enough for our prediction purpose?)

# Model significance test

- Test that none of the regressors explain $Y$ :

$$
\begin{aligned}
H_0 &: \beta_1 = \beta_2 = \ldots = \beta_p = 0 \\
H_a &: \text{at least one } \beta_j \text{ is non-zero.}
\end{aligned}
$$

- Use the $F$-statistic

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

under $H_0$. We expect $F$ to be large if $H_a$ is true.

# Test of subset significance

- Sometimes we want to test that a particular subset of $q$ of the coefficients are zero: $H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots \beta_p = 0$ against $H_a$: $\beta_{p-q+1} \neq 0$ or $\beta_{p-q+1} \neq 0$ or $\cdots$ or $\beta_p \neq 0$.
- We fit a second model that uses all the variables except those last $q$. Suppose that the residual sum of squares for that model is $\mathrm{RSS}_0$.
- Use the $F$-statistic

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p-1)} \sim F_{q,n-p-1}.$$

# Qualitative predictors

- Some predictors are not quantitative but are qualitative, taking a discrete set of values.
- These are also called categorical predictors or factor variables.
- The `Credit` data set records variables for a number of credit card holders.
  - The response is `balance` (average credit card debt for each individual).
  - Quantitative predictors: `age`, `cards` (number of credit cards), `education` (years of education), `income` (in thousands of dollars), `limit` (credit limit), and `rating` (credit rating).
  - Qualitative variables: `gender`, `student` (student status), `status` (marital status), and `ethnicity` (Caucasian, African American (AA) or Asian).

- Example: investigate differences in credit card balance between males and females, ignoring the other variables. We create a new variable

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

- Resulting model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{-th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{-th person is male.} \end{cases}$$

- With more than two levels, we create additional dummy variables. For example, for the ethnicity variable we create two dummy variables. The first could be

$$X_{i1} = \begin{cases} 1 & \text{if } i\text{-th person is Asian} \\ 0 & \text{if } i\text{-th person is not Asian,} \end{cases}$$

  and the second could be

$$X_{i2} = \begin{cases} 1 & \text{if } i\text{-th person is Caucasian} \\ 0 & \text{if } i\text{-th person is not Caucasian.} \end{cases}$$

- Both of these variables can be used:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i \\ &= \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{-th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{-th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{-th person is AA.} \end{cases} \end{aligned}$$

- There will always be one fewer dummy variable than the number of levels. The level with no dummy variable is known as the baseline.

# Interactions

- In our previous analysis of 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.
- The average effect on `sales` of a one-unit increase in `TV` is always $\beta_1$, regardless of the amount spent on `radio`.
- But 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.
- Model takes the form

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

| | 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 |

ISL Table 3.9

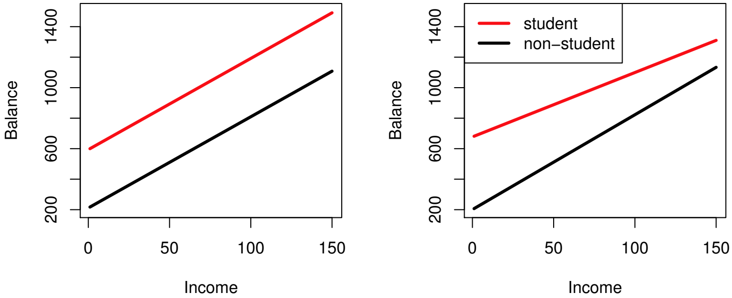
- The results suggest that interactions are important.
- The $p$-value for the interaction term `TV`×`radio` is extremely low, indicating that there is strong evidence for $\beta_3 \neq 0$.

- Consider the `Credit` data set, and suppose that we wish to predict `balance` using `income` (quantitative) and `student` (qualitative). Without an interaction term, the model takes the form

$$
\begin{aligned}
\texttt{balance}_i \approx & \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\
= & \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases}
\end{aligned}
$$

- With interactions, it takes the form

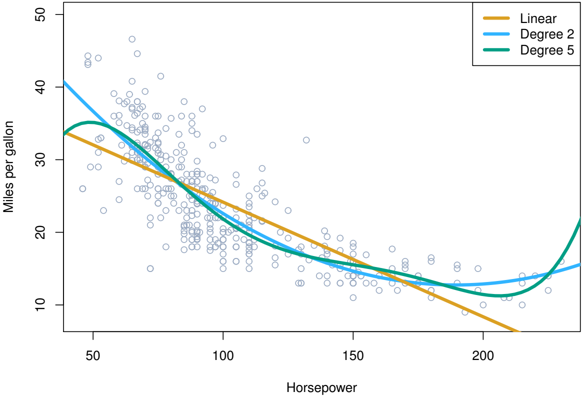
$$
\begin{aligned}
\texttt{balance}_i \approx & \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases} \\
= & \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases}
\end{aligned}
$$




ISL Figure 3.7

- Regression lines have different intercepts, as well as different slopes.

# Non-linear effects of predictors




ISL Figure 3.8

- The `mpg` (gas mileage in miles per gallon) versus `horsepower` is shown for a number of cars in the `Auto` data set.

| | 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 | $< 0.0001$ |

ISL Table 3.10

- It seems clear that this relationship is in fact non-linear. A simple extension to the linear model is to include transformed predictors.
- A nonlinear model

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

may provide a better fit (lower $R^2$).

# Confidence and prediction intervals

- In a linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$ with a single predictor $X$, suppose that for some fixed $x_0$, we wish to construct a confidence interval that covers $y_0 = \beta_0 + \beta_1 x_0$ with 95% probability.
- An estimator of $y_0$ is $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ and

$$\text{SE}\left(\hat{y}_0\right) = \frac{\sigma^2}{n}\left(1 + \frac{\left(\overline{X} - x_0\right)^2}{n^{-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right).$$

- $\widehat{\text{SE}}\left(\hat{y}_0\right)$ replaces $\sigma^2$ with $\hat{\sigma}^2$. An 95% confidence interval for $y_0$:

$$\left[\hat{y}_0 - 2 \cdot \widehat{\text{SE}}\left(\hat{y}_0\right), \hat{y}_0 + 2 \cdot \widehat{\text{SE}}\left(\hat{y}_0\right)\right].$$

- A prediction interval

$$\left[\hat{y}_0 - 2 \cdot \sqrt{\widehat{\text{SE}}\left(\hat{y}_0\right)^2 + \hat{\sigma}^2}, \hat{y}_0 + 2 \cdot \sqrt{\widehat{\text{SE}}\left(\hat{y}_0\right)^2 + \hat{\sigma}^2}\right]$$

  covers $Y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0$ with 95% probability, where $\epsilon_0$ is a new error that is independent of the training data.