

SOSC 5340: Linear regression

Han Zhang

Outline

Logistics

Parametric models

Univariate OLS: estimation

Multivariate OLS: estimation

Multivariate OLS: inference

Bias-variance trade-off

OLS extensions

Today's review

Recommended readings

- Today's topic deals with prediction and OLS regression, one of the simplest prediction model.
- It is often useful to see how other authors present the same content
- More traditional treatment of the topic
 - Wooldridge, *Introductory Econometrics: A Modern Approach*, 2015. Chapters 2 - 8.
- More modern treatment of the topic
 - Aronow and Miller, *Foundations of Agnostic Statistics*, 2019. Chapters 2 - 4.

Expectation of final papers

- Not doing causal inference?
 - I encourage to think about how to do causal inference
 - It makes your study stronger
 - If you are not sure, come to my office hour and let us discuss it
- Collaboration in final paper?
 - Acceptable; no more than 2 people

Parametric Assumptions

- The theorem, “Conditional Expectation as the Best Predictor”, is true in general.
- **Nonparametric** estimator:
 - Estimate $E(Y|X)$ (conditional mean) from the sample directly
 - Pros: no additional assumption
 - More advanced methods in causal inference often rely on this approach
 - Cons: not easy to estimate (e.g., if you have two X , or X is continuous)
 - With an important exception: experiments and causal inference, where X are mostly binary
- **Parametric** methods (e.g., regressions):
 - **explicitly assume functional form** of $E(Y|X = x) = g(X)$
 - E.g., linear regression: $g(X)$ is linear

How do we design the approximating function?

- $E(Y|X)$ is the best predictor of Y given X
- The error of the best predictor $\epsilon = Y - E(Y|X)$ satisfy:
 - $E(\epsilon) = 0$:
 - $E(\epsilon|X)$: the error is mean dependent of X .
 - These two equations are **true**
- If we assume $E(Y|X = x) = g(X)$, we should also assume that the error $\epsilon = Y - g(X)$ satisfy:
 - $E(\epsilon) = 0$
 - $E(\epsilon|X) = 0$
 - It is important to note that now $E(\epsilon) = 0$, and $E(\epsilon|X) = 0$ **are assumptions!**

OLS Assumptions

- Linear regression or Ordinary Least Square regression (OLS)
- Assumption 1: the expected error is 0

$$E(\epsilon) = 0$$

- Assumption 2: **mean independent** between X and the error

$$E(\epsilon|X) = 0$$

OLS Assumptions (cont'd)

- Assumption 3: **linear model**

$$Y = g(X) = \beta_0 + \beta_1 X + \epsilon$$

- β_0 and β_1 are called **parameters**
- $Y = \beta_0 + \beta_1 X + \epsilon$ is called (parametric) **statistical model**

Parameters are unknown constants

- X, Y are **random** variables; their values are generally **unknown**
- We can sample X_1, \dots, X_n from population X ;
 - X_1, \dots, X_n are also random variables
 - Their values are **known**
- Parameters β are **constants**; their values are **unknown**
- [Advanced knowledge]: Bayesian statistics view β as random variables instead of constants

Population Regression Function

- Given the assumptions, it is easy to see that

$$E(Y|X) = E(\beta_0 + \beta_1 X + \epsilon|X) \quad (1)$$

$$= \beta_0 + \beta_1 X + E(\epsilon|X) \quad (2)$$

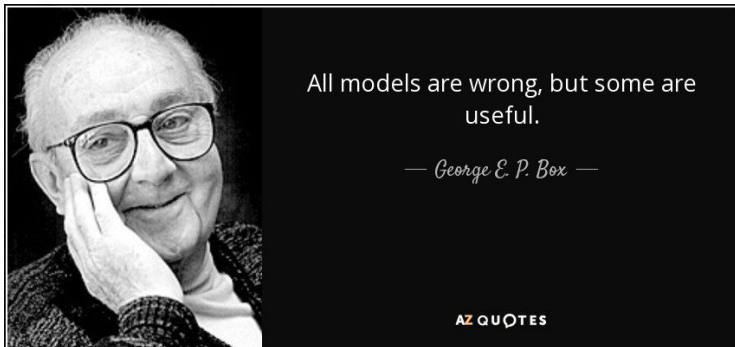
$$= \beta_0 + \beta_1 X \quad (3)$$

- That is, our assumptions lead us to a certain functional form of $E(Y|X) = \beta_0 + \beta_1 X$
- We call $E(Y|X) = \beta_0 + \beta_1 X$ the population regression function
- Note that this is only true when the assumptions are met!

Compare parametric and non-parametric methods

- X is father's height and can take 240 values (every centimeter from 0 to 240); Y is son's height
- Nonparametric: you need a table of 240 cells to represent $E(Y|X = x)$
- Parametric: OLS regression characterizes the relationship between Y and X with 2 parameters: β_0 and β_1
- Parametric model is simpler but need assumptions of the linear relationship

All models are wrong; some are useful



Parametric models can be powerful



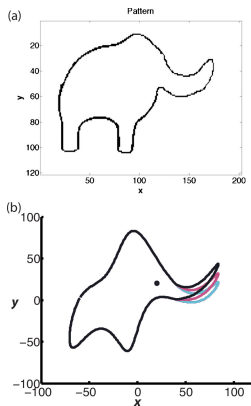
With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.

— *John von Neumann* —

AZ QUOTES

Examples

Jürgen Mayer, Khaled Khairy, and Jonathon Howard, “Drawing an elephant with four complex parameters” *American Journal of Physics*, 78, 648 (2010);



Univariate OLS estimation: set up

- Now we go to the problem of **estimating** population-level conditional expectation $E(Y|X)$ with samples
- Our samples are i.i.d. random samples from X and Y : $(X_1, Y_1), \dots, (X_n, Y_n)$.
- With a assumed statistical model $Y = \beta_0 + \beta_1 X + \epsilon$
- And two assumptions about the error
 1. $E(\epsilon) = 0$
 2. $E(\epsilon|X) = 0$, which implies $E(\epsilon X) = 0$

Univariate OLS estimation (cont'd)

- $E(\epsilon) = 0 \implies E(Y - \beta_0 - \beta_1 X) = 0$
- $E(X\epsilon) = 0 \implies E[X(Y - \beta_0 - \beta_1 X)] = 0$
- Now we **plug-in** sample analog

Population	Sample
$E(Y - \beta_0 - \beta_1 X) = 0$	$\sum_{i=1}^n \frac{1}{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$
$E[X(Y - \beta_0 - \beta_1 X)] = 0$	$\sum_{i=1}^n \frac{1}{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$

Univariate OLS estimation (solution)

- Solving these two equations give:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^N (Y_i - \bar{Y}) (X_i - \bar{X})}{\sum_{i=1}^N (X_i - \bar{X})^2} \quad (4)$$

Univariate OLS estimation: Least Square Perspective

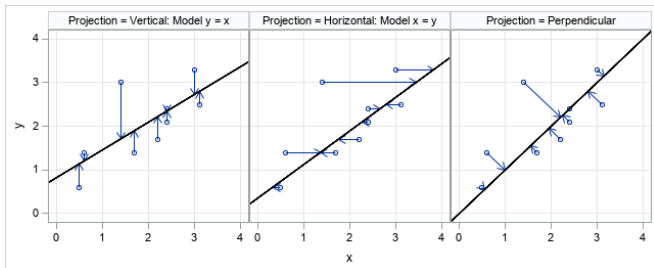
- We have an alternative view of the OLS estimation
- The best predictor of Y , $E(Y|X)$, minimizes Mean Squared Error (MSE), $E[(Y - E(Y|X))^2]$
- Now **with samples**
 - We will try to minimize the sample MSE (also called **empirical MSE**)

$$\text{MSE}_{\text{sample}} = \frac{1}{n} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 \quad (5)$$

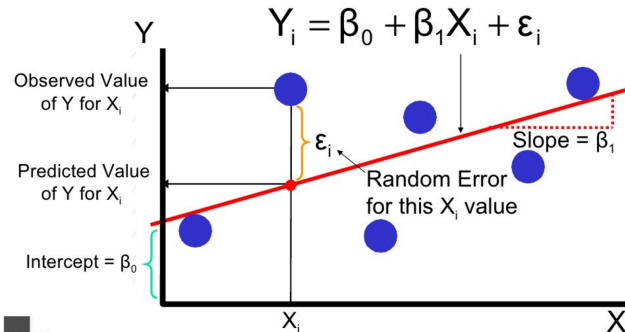
- Note that the sample MSE is almost surely to be larger than the MSE of the best predictor; but this is the best we can do given the assumptions (linear model)
- The problem now is to find $\hat{\beta}_0, \hat{\beta}_1$ that minimize $\text{MSE}_{\text{sample}}$.
 - Use standard calculus; let the derivative of MSE with respect to β to be 0.
- The solutions will be the same

Univariate OLS estimation: geometric perspective

- We have yet another view of the OLS estimation
- Linear regression project observation points vertically onto the “fitted line”
- The left and middle one are linear regressions
- The right one is a special case of a famous machine learning algorithm, “Support Vector Machine” (SVM)



OLS geometry (cont'd)



Multivariate OLS Assumptions (cont'd)

- Now we shift from using a single variable X to predict Y , to use k variables X_1, \dots, X_k to predict Y
- Extend the assumptions of univariate OLS to multivariate OLS case:
- Assumption 1: the expected error is 0

$$E(\epsilon) = 0$$

- Assumption 2: **mean independent** between X and the error

$$E(\epsilon|X_1, \dots, X_k) = 0$$

- This assumption implies that the error and any covariate is uncorrelated, that is, $E(\epsilon X_i)$ for any i

Multivariate OLS Assumptions (cont'd)

- Assumption 3: **linear model**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k + \epsilon$$

- Specific to OLS regression
 - β_0 is called intercept
 - And the rest are also called slope
 - Together, they are called regression coefficients

Multivariate OLS Estimation

- Now going from population to samples
- We sample (Y, X_1, \dots, X_k) for n times
- The sampled data look like

$$\begin{bmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix} \quad (6)$$

- and (Y_1, Y_2, \dots, Y_n)

Multivariate OLS Estimation (cont'd)

- Again, we plug-in the sample analog in place of population equations

Population	Sample
$E(\epsilon) = 0$	$\sum_{i=1}^n \frac{1}{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \dots - \hat{\beta}_k X_{ik}) = 0$
$E(X_1 \epsilon) = 0$	$\sum_{i=1}^n \frac{1}{n} X_{i1} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \dots - \hat{\beta}_k X_{ik}) = 0$
...	...
$E(X_k \epsilon) = 0$	$\sum_{i=1}^n \frac{1}{n} X_{ik} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \dots - \hat{\beta}_k X_{ik}) = 0$

- Now we have $k + 1$ parameters, from β_0 to β_k
- And we have $k + 1$ equations
- Solving these equations will give the solution to OLS regression

Multivariate OLS Estimation (matrix notations)

- It is too complex to write down all these equations every time
- We rewrite the sample data, using matrix notation
- The first column of \mathbf{X} , is added artificially (that is, we just assume $X_0 = 1$)

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}; \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{bmatrix} \quad (7)$$

- The matrix product:

$$\mathbf{X}\beta = \begin{bmatrix} 1 * \beta_0 + X_{11} * \beta_1 + \dots + X_{1k}\beta_k \\ \dots \\ 1 * \beta_0 + X_{n1} * \beta_1 + \dots + X_{nk}\beta_k \end{bmatrix} \quad (8)$$

Multivariate OLS Estimation (matrix version)

- With the notation $X_0 = 1$, our **sample** estimation equations can be written as:

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ X_{k1} & X_{k2} & \dots & X_{kn} \end{bmatrix} \begin{bmatrix} Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \dots - \hat{\beta}_k X_{ik} \\ \vdots \\ Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \dots - \hat{\beta}_k X_{ik} \end{bmatrix} = 0 \quad (9)$$

- Note that \mathbf{X} is transposed here; we write it as \mathbf{X}^T
- And in the matrix term, we write the estimation equations as:

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\beta}) = 0 \quad (10)$$

Multivariate OLS Estimation (matrix algebra)

- Using matrix notation, we can write the estimates of parameters easily:

$$\begin{aligned}
 \mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\beta}) &= 0 \\
 \mathbf{X}^T\mathbf{Y} &= \mathbf{X}^T\mathbf{X}\hat{\beta} \\
 \hat{\beta} &= [\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{Y}
 \end{aligned}
 \tag{11}$$

Multivariate OLS Estimation (unbiasedness)

- Remember, a good estimator has three properties: unbiasedness, consistent, and asymptotically normal
- Does OLS estimator have these good properties?

$$\begin{aligned}
 \hat{\beta} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}, \text{ (then substitute } \mathbf{Y} = \mathbf{X}\beta + \epsilon \text{)} \\
 &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \epsilon \\
 &= \beta + \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \epsilon
 \end{aligned} \tag{12}$$

- β is a constant (unknown)
- It is easy to see that $E(\hat{\beta}) = \beta$, hence $\hat{\beta}$ is **unbiased** estimator of β

Multivariate OLS Estimation (consistency)

- $\hat{\beta} - \beta = \mathbf{X}\epsilon (\mathbf{X}^\top \mathbf{X})^{-1}$
- As n goes to infinity, sample analog of $E(X\epsilon) \rightarrow 0$
- Why?
- Assumption 2
- So $\hat{\beta}$ is consistent

Multivariate OLS in action

```
library(AER)
data(CASchools)
model <- lm(math ~ income + english, data = CASchools)
coef(model)
```

```
(Intercept)      income      english
636.6293146    1.5035886   -0.4005886
```

Multivariate OLS in action

Now we get the matrix estimates

```
X <- model.matrix(model) # this is our X
Y <- CASchools$math
# solve() is to take  $X^{-1}$ 
# %*% is matrix product
# t() is transpose
beta = solve(t(X) %*% X) %*% t(X) %*% Y
beta
```

```
          [,1]
(Intercept) 636.6293146
income      1.5035886
english     -0.4005886
```

Variance of OLS estimators

- We have talked about how to estimate our parameters β
- But how confident are we?
- How can we estimate the variance and confidence intervals?

$$\begin{aligned}
 V(\hat{\beta}) &= V \left(\underbrace{\beta}_{\text{Variance of 0}} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \right) \\
 &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T V(\epsilon) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}
 \end{aligned} \tag{13}$$

- $V(\epsilon)$ is the variance of the **population** error. We have to estimate it!

Assumption 4: homoskedastic error

- We have to add assumptions to estimate $V(\epsilon)$
- **Assumption 4 (homoskedasticity)**: for every sample, they have the same variance of the error $V(\epsilon)$. We also write $V(\epsilon) = \sigma^2$,
- Under Assumption 4: our **estimate** of σ^2 is:
 1. $\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \dots - \hat{\beta}_k X_{ik}$; $\hat{\epsilon}_i$ is called residuals
 2. $\hat{V}(\epsilon) = \hat{\sigma}^2 = \frac{1}{n-k}(\hat{\epsilon}_1^2 + \dots + \hat{\epsilon}_n^2)$ (hint: $n - k$ makes this quantity unbiased)
- That is, the estimate of the variance of the error is the (weighted) sample mean of **residual** squares.
- With the estimate $\hat{\sigma}$, we have the classical standard error of β

$$\hat{V}(\hat{\beta}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \hat{\sigma}^2 \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1} \quad (14)$$

Alternative Assumption 4: heteroskedastic error

- **Alternative Assumption 4 (heteroskedasticity)**: the sample's error can be different (but they are uncorrelated with each other)
- Under Alternative Assumption 4,
 1. $\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \dots - \hat{\beta}_k X_{ik}$
 2. we do not calculate the sample mean; instead:

$$\hat{V}(\epsilon) = \begin{bmatrix} \hat{\epsilon}_{i1}^2 & 0 & 0 \\ 0 & \hat{\epsilon}_{i2}^2 & 0 \\ 0 & 0 & \hat{\epsilon}_{in}^2 \end{bmatrix} \quad (15)$$

- With the estimate of $V(\epsilon)$, we have **heteroskedasticity-robust standard error** of β

$$\hat{V}(\hat{\beta}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \hat{V}(\epsilon) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \quad (16)$$

Homoskedasticity vs heteroskedasticity

- Homoskedastic errors are almost surely to be wrong in reality
- So why we still care about the classical standard error?
- Some math reasons: classical standard errors (and homoskedasticity) make the inference easier, as will be seen later.
- Some historical reasons: it is harder to calculate the robust standard error in the old days (but now with computers, is it very easy)
- In default statistical packages (R, Stata or most other languages), when you run a simple OLS regression, the default standard error you get is still the classical standard error

Assumption 5: normality

- We have derived the estimator for β and its variance
- Now we want to construct confidence intervals
- **Assumption 5:** normal error assumptions (with homoskedasticity)

$$\epsilon \sim N(0, \sigma^2)$$

- Note that Assumption 5 directly implies Assumption 1 (error has mean 0) and Assumption 4 (sample error is the same)
- But not the other way around
- So Assumption 5 is a very strong assumption
- Effectively this also makes the estimator asymptotically normal

Confidence intervals

- With the normality assumption, the α confidence interval of regression coefficients β :

$$\left(\hat{\beta} - z_{\frac{1+\alpha}{2}} \sqrt{\hat{V}(\hat{\beta})}, \hat{\beta} + z_{\frac{1+\alpha}{2}} \sqrt{\hat{V}(\hat{\beta})} \right) \quad (17)$$

- z is the quantile function of a standard normal distribution
 - $\alpha = 0.95$; $z_{0.975} = 1.96$
 - $\alpha = 0.99$; $z_{0.995} = 2.58$

Hypothesis Testing

- Confidence interval can be used to judge the possibility of a particular guess of β
- This is known as **Hypothesis Testing**
- For instance, we think that X_1 is very predictive of outcome Y , so we hypothesize that its coefficient β_1 is not 0.
 - **alternative hypothesis:** $\beta_1 \neq 0$. (this is the hypothesis you truly believe)
 - **null hypothesis:** $\beta_1 = 0$. (this is the “boring” or default hypothesis)

Use confidence interval to perform hypothesis testing (example)

- Our alternative hypothesis is $\beta_1 \neq 0$
- And null hypothesis is $\beta = 0$
- Example: our point estimate of β_1 is $\hat{\beta}_1 = 0.8$, and our estimated 95% confidence interval is $[0.2, 1.4]$
- This indicates 95% of times, our true value of β_1 will be in the range of $[0.2, 1.4]$ over repeated samples
- Therefore, the chance our true β_1 will be 0 (null hypothesis) is small (say 0.05 on average)
- So we reject the null hypothesis and find support for our alternative hypothesis

Use confidence interval to perform hypothesis testing

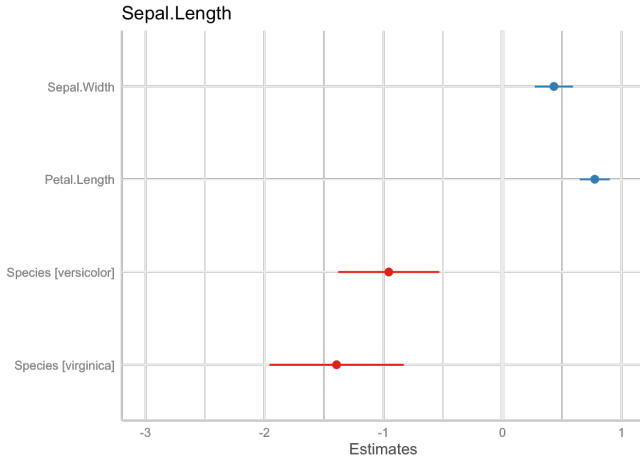
- In general: if we have 95% confidence interval for some quantity θ , $[\theta_{min}, \theta_{max}]$
- And our alternative hypothesis is that $\theta = \theta_{test}$
- Then
 - if θ_{test} does not belong to $[\theta_{min}, \theta_{max}]$, we find support for the null hypothesis
 - if θ_{test} falls into $[\theta_{min}, \theta_{max}]$ or not, we find support for our alternative hypothesis

Illustration

- $\theta_{test} = 0$; we want to test whether regression coefficient differs from 0

```
library(sjPlot); data(iris)
m2 <- lm(Sepal.Length ~ Sepal.Width +
         Petal.Length + Species, data = iris)
plot_model(m2)
```

Illustration



- If 95% confidence interval does not touch 0, we can reject the null and say that coefficients are statistically significantly different from 0

Hypothesis Testing using test statistics

- Alternatively, Hypothesis Testing can be done by calculating **test statistics** of β
- Null Hypothesis: $\beta = 0$
- Alternative Hypothesis: $\beta \neq 0$
- t -statistics is a commonly used test statistics:

$$t = \frac{\hat{\beta} - \beta}{\sqrt{\hat{V}(\beta)}} \quad (18)$$

- If Null Hypothesis is true, we would expect that t is small; otherwise, t should be large
- With homoskedastic error, t follows a student t distribution
- With heteroskedastic error, t distribution is more complex
- Student t distribution depends on data size as well as model; it's not comparably across different data sets
 - E.g., are 10 large enough? 100 large enough?

P-value

- We use something called p-value; it is calculated from the cumulative distribution of the sampling distribution of t
- For example: if p-value is p , it means that the probability of obtaining a t -statistics that equals to t_0 or higher, when sampling from the same population, is approximately p .
- The smaller the p value, the more evidence that we can reject the Null Hypothesis

Confidence interval, t -statistics, and p-value

- Null Hypothesis: $\beta = 0$
- Alternative Hypothesis: $\beta \neq 0$
- The following statements are equivalent
 - The t -statistics is larger than 1.96 (for linear regression only!)
 - The p-value is small or equal to 0.05
 - Estimated 95% confidence interval does not contain 0
 - And each of the three argument can let us to reject the null hypothesis and find support for the alternative hypothesis

Bias-variance trade-off

- To better understand the trade-off between simple (parametric) and complex models, we introduce a concept called bias-variance trade-off
- Previously we have done some math and proved that

$$E[(Y - g(X))^2] > E[(Y - E(Y|X))^2]$$
- Define $e = E(g(X))$ (expectation of any estimate)

Bias-variance trade-off

$$E[(Y - g(X))^2] = E[(Y - e + e - g(X))^2] \quad (19)$$

$$= E[(Y - e)^2 + 2E(Y - e)E(e - g(X)) + (e - g(X))^2] \quad (20)$$

$$= E[(Y - e)^2] + E[(e - g(X))^2] \quad (21)$$

$$= \text{Bias}^2 + \text{var}(g(X)) \quad (22)$$

$$(23)$$

- First is definition of **bias**: $E(\hat{\theta}) - \theta$
 - here our Y is θ (true value) and e is the expectation of estimate $g(X)$
- second is definition of **variance**

Bias-variance trade-off

- Define $e = E(g(X))$ (expectation of any estimate)

$$E[(Y - g(X))^2] = E[(Y - e + e - g(X))^2] \quad (24)$$

$$= E[(Y - e)^2 + 2E(Y - e)E(e - g(X)) + (e - g(X))^2] \quad (25)$$

$$= E[(Y - e)^2] + E[(e - g(X))^2] \quad (26)$$

$$= \text{Bias}^2 + \text{var}(g(X)) \quad (27)$$

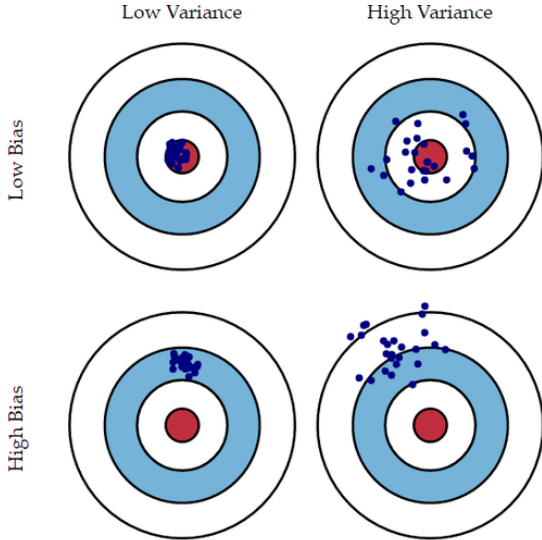
$$(28)$$

- First is definition of **bias**: $E(\hat{\theta}) - \theta$
 - here our Y is θ (true value) and e is the expectation of estimate $g(X)$
 - so the bias is on average how our prediction differs from the truth
- second is definition of **variance**

Bias Variance Trade-off

- It is often true that:
- Simple models (like OLS) have large bias, but small variance
- Complex models (machine learning) have small bias, but estimator variance

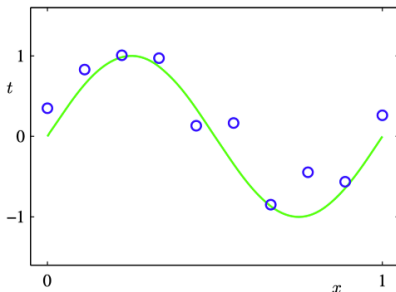
Bias Variance Trade-Off



Bias Variance Trade-off (example)

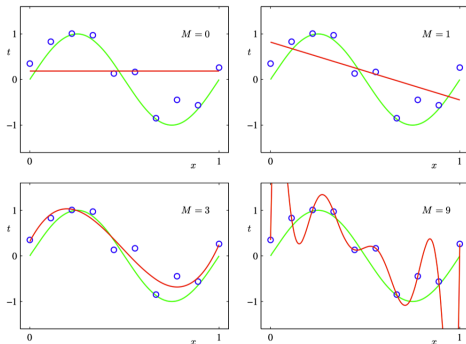
- We have a linear regression with only one variable X , but we add higher order terms
- When M is larger, we get models that fit the data better and better

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + X^M$$



Bias Variance Trade-off (cont'd)

- $M = 1$, OLS gives lots of estimation bias
- $M = 9$, the model fits the data so well, but it is highly sensitive to small changes in observations
- $M = 3$, it achieves a good balance between estimation bias and estimator variance



Bias Variance Trade-off

- Ultimately, if the goal is to improve prediction performance
- We should balance bias and variance, instead of using more and more complex models to reduce bias

Interpret coefficient

- **Marginal effect**: how the change in one variable X_i predicts **average** change in y , holding other variables constant
- Marginal effect is done by taking the partial derivative of Y regarding X_i

$$\frac{\partial Y}{\partial X_i} \quad (29)$$

- In OLS regression, the marginal effect is simple:
- Marginal effect of X_i on Y is its regression coefficient β_i
- Or in other words: one unit change in X_i predicts **on average** β_i change in Y , given the same values for other variables.

Collinearity

- Estimating multivariate OLS with k variables:
 - $k + 1$ equations; $k + 1$ unknown parameters
- But if two variables are colinear, we cannot solve the equation (one equation becomes useless)
- Examples of collinearity:
 - $X_2 = 2X_1$
 - $X_2 = 1 - X_1$
- If X_1 and X_2 have very high correlation: multicollinearity
 - Variance estimate of parameters can be very high
 - Good practice: always check correlation before putting variables into regression

Dummy variable

- Note that we make no assumption about how X and Y should be distributed at all
- So X can be categorical, such as gender (man, female, transgender, etc)
- So naturally, gender looks like (man, female, man, transgender, female) (assume that we have five observations)
- Dummy transformation turns the data into (1 means yes and 0 means no); this is also called zero-one representation or indicator representation
- Each row, one and only one cell is 1; the other cells are 0

observation/gender	man	female	transgender
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1
5	0	1	0

Dummy variable interpretations

- $\text{wage} = \beta_0 + \beta_1 \text{female} + \beta_2 \text{education} + \epsilon$
- Female is a dummy variable for females (i.e., 1 = females and 0 = males), and educ is years of schooling. wage is hourly wage.
- So the reference group is male
- Interpretations:

*β_0 represents the difference in **mean** hourly wage between females and the reference group, male, given the same amount of education.*

Interactions

- Sometimes, effect of X on Y
- We add interaction terms to capture the idea

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \delta X_1 X_2 \quad (30)$$

- Effect of X_1 on Y is the marginal effect, which depends on X_2

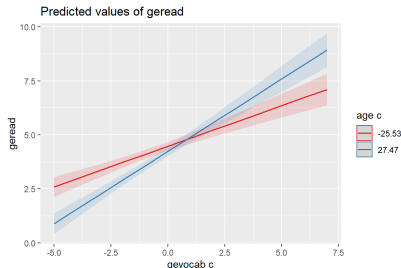
$$\frac{\partial Y}{\partial X_1} = \beta_1 + \delta X_2 \quad (31)$$

- Now, interpreting β_1 needs some caution
 - it is **wrong** to say that one unit increase in X_1 predicts β_1 increase in outcome
 - The amount of increase in outcome also depends on the value of X_2
 - It's much easier to plot the marginal effect

Plotting interaction effect

- E.g., Suppose that we want to test whether more vocabulary predicts higher reading scores differently for age groups

$$\text{reading} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{vocab} + \delta \text{age} \times \text{vocab} + \epsilon$$
- Using R package sjPlot; many other options, such as interactions or interplot



Interactions involving categorical variables

$$\frac{\partial \text{wage}}{\partial \text{education}} = \beta_2 + \delta \text{female}$$

- This means that when $\text{female} = 0$, the marginal effect is β_2
- When $\text{female} = 1$, the marginal effect is $\beta_2 + \delta$
- The marginal effect of education on wage is different, conditional on gender

Wrap Up

- $E(Y|X)$ is the best predictor of Y with MSE
- What are OLS regression's assumptions?
- How to estimate OLS regression coefficient and perform inference?
- How to interpret OLS coefficients