

Chapter 2

What We Are Doing in This Course?

여러 가지 통계기법들과 선형회귀모델에 관해 이야기하기에 앞서, 정량적 사회과학연구를 이해하기 위한 몇 가지 핵심적 개념들을 다루고, 이해해보는 시간을 가져보고자 합니다. 먼저, 대체 “함수(function)”란 무엇일까요? 함수는 도대체 왜 유용하며, 왜 사용하는 걸까요? 그리고 이 자료는 선형회귀에 관한 것인데, 그에 앞서 함수에 대한 질문을 하는 것일까요?

함수는 일군의 특성이 다른 일군의 특성과 함께 변화하는(covary) 체계적 변화를 보여주는 방식입니다. 다른 말로 하자면, 하나의 함수는 어떻게 우리가 관측한 일련의 특성들이 서로 연관되어 있는지를 보여준다고 할 수 있습니다. 즉, 우리가 어떤 함수를 가지고 있고, 그와 관련된 일련의 특성들을 관측했다면, 우리는 관측된 특성들이 함수적 관계 속에서 어떠한 다른 특성들과 이어지게 될지를 기대할 수 있게 됩니다. 사회과학 연구에서는 이론적 검토라는 작업이 선행되는데, 선행연구들을 바탕으로 이론적인 관계를 기대하고 경험적 데이터들이 그 기대에 따라 배열되어 나타날 것이라고 주장(가설 수립)합니다. 함수는 그 관계를 간명하게 보여주는 방식입니다. 함수는 우리가 관측한 현상들의 특성들 간 체계적 관계를 논리적으로 보여주는 지도의 역할을 하기 때문에 유용하다고 할 수 있습니다.

선형회귀분석이라는 기법에 본격적으로 들어가기에 앞서 함수의 정의를 묻는 것이 중요한 이유는 선형회귀모델이 둘, 또는 그 이상의 변수들 간의 관계를 서술하고 추론하는 데 함수를 사용하기 때문입니다. 선형회귀모델은 변수들 간의 관계를 하나의 수리적 식(equation)으로 표현하고, 그것이 바로 함수입니다. 사회과학적 연구에서 사용하는 질문들이 어떻게 함수적 관계로 표현될 수 있는지, 그리고 잠정적으로 어떻게 답변되는지¹를 아래의 사례들로 살펴보겠습니다.

- Q1. 소득 수준이 높을수록 사람들이 더 결혼을 할까?
A1. 소득 수준이 높을수록, 결혼으로 예상되는 주거, 양육 등과 같은 미래 지출에 대한 부담이 감소하므로, 개인이 결혼할 확률은 높아질 것이다.
- Q2. 경제적으로 교역을 많이할수록 전쟁을 덜 할까?
A2. 무역은 무역의 당사자들 간 상호연관성을 높이기 때문에, 그 두 당사자들 간 전쟁 확률은 감소할 것이다.
- Q3. 소득 수준이 높은 유권자일수록 보수정당에 투표하는 경향이 나타날까?
A3. 유권자의 소득 수준이 높을수록, 그 유권자는 소득 수준이 낮은 유권자에 비하여 세금이 부과되기를 원하지 않을 것이다². 따라서 소득 수준이 높은 유권자일수록 전형적으로 낮은 세율을 주장하는 보수정당을 선호하는 투표를 할 것이다.
- Q4. 민주주의가 비민주주의보다 더 경제적으로 성장할까?
A4. 민주주의가 비민주주의에 비해 시장체계에 친화적인 사적재산권을 보호할 가능성이 크기 때문에, 민주주의가 비민주주의보다 더 높은 경제성장 수준을 보여줄 것이다.
- Q5. 시간이 경과될수록 정부지출 규모가 지속적으로 증가할까?
A5. 사회는 노령화되고 복합적으로 변화하기 때문에 시간이 지날수록 정부지출은 증가할 것이다. 그러나 사회의 자원 총량은 한정되어 있기 때문에, 일정한 지출 수준에서 그 증가율은 감소할 것이다.
- Q6. SNS 사용 빈도가 높은 유권자일수록 진보정당에 투표할 확률이 높아질까?
A6. 젊은 세대는 상대적으로 SNS를 통해 의사소통을 하며, 일반적으로 젊은 세대가 진보정당을 지지하는 경향이 있으므로, SNS 사용 빈도가 높을수록 이는 진보정당에 대한 투표 증가로 이어질 것이다.

¹여기에서 제시하는 질문과 답변의 예제들은 학술적으로 엄밀하게 검증되거나 검토된 결과가 아닙니다. 다만 함수에 대한 이해를 돋기 위해 수리적 식이 아닌 일반적 표현으로 바꾸어 제시하였을 뿐입니다.

²일반적으로 민주주의 국가에서 세금 제도는 누진적(progressive)이고 소득의 일정 비율로 세금이 책정된다는 가정에 바탕을 두고 있습니다.

- Q7. 구 공산권 국가에서 복지프로그램에 대한 지출 수준이 더 높게 나타날까?
A7. 구 공산권 국가들은 공산주의의 영향으로 복지 프로그램에 대해 상대적으로 더 높은 지출 수준을 보이고, 경로의존성으로 인하여 대중들의 지지를 확보하기 위해 기존의 복지 프로그램에 대한 지출을 유지 및 지속적으로 확대할 것이다.
- Q8. 인터넷 사용 수준이 높을수록, 권위주의 국가가 민주화될 확률이 높을까?
A8. 인터넷은 시민들 간의 의견을 교환하는 데 따르는 정보비용을 감소시키는 역할을 수행하므로, 시민들이 정치적 자유와 경제적 자유라는 공통의 목표를 달성하기 위해 집단행동을 하는 데 드는 비용을 감소시키는 역할을 할 것이다.
- Q9. 복지지출의 증가가 사회의 기대수명을 제고할 것인가?
A9. 높은 수준의 복지지출은 시민들을 잠재적 위험으로부터 구제하는 기능을 수행하므로 평균적인 기대수명은 복지지출이 더 높은 국가에서 증가할 것이다.
- Q10. 아프리카의 선거제도에서 비례성 수준이 높을수록, 국내적 분쟁 발생 확률이 감소할까?
A10. 일반적으로 비례적인 선거제도는 사회 내 갈등을 조정하는 데 이점이 있다고 알려져 있다. 하지만 아프리카와 같이 국민국가의 구획화가 타의에 의하여 이루어진 국가는 한 사회 내에 매우 다른 기원을 가진 다른 종류의 인종 집단들이 혼재하게 된다. 따라서 비례적 선거제도는 서로 다른 사회집단들 간의 필연적인 긴장을 야기하게 되고, 어떤 한 집단이 주도권을 잡지 못하게 만드는 결과로 이어져 오히려 국내 분쟁이 발생할 확률을 높일 수 있다.

위의 10개의 질문과 답변의 예제들은 학계에서 어느 정도 합의된 이론들이 제시된 것도 있고, 아직 논쟁 중인 것들도 있습니다. 하지만 위의 예제들을 통해 살펴볼 수 있는 것은 첫째 우리는 원인과 결과를 상정하고 그 관계를 살펴보고자 하며, 둘째 이때 원인과 결과는 서로 변화하는 관계에 있다는 것입니다. 예를 들어 Q8에서 원인으로 간주되는 인터넷 사용 수준도 변화하는 변수(variable)이며, 민주화 확률(probability)도 변하는 것(variable)입니다. 즉, 우리는 변수와 변수라는 공변하는(covarying) 인자들 간의 관계를 보고자 하는 것입니다. 그리고 선형회귀분석은 이렇게 변화하는 관계의 모습이 선형(linear)일 것이라는 기대를 가지고 있습니다. 우리는 기대하는 선형관계를 수식으로 나타내고 그 도함수(derivative, 혹은 미분)을 취함으로써 변수들 간의 관계를 간략하게 보여주고자 합니다.

그렇다면 도함수 또는 미분이란 무엇일까요? 마찬가지로 도함수가 무엇인지를 수학적 개념을 제시하고 이를 풀어서 설명해보도록 하겠습니다. 나아가 위와 마찬가지로 도함수에 대한 10가지 예제를 제시하고 현실 세계에서 가상의 관계를 제시하고 도함수로 이를 어떻게 해석할 수 있는지를 살펴보도록 하겠습니다.

먼저 우리는 x 에 대한 함수 $f(x)$ 를 도함수를 취해 $f'(x)$ 의 수리적 형태로 보여줄 수 있습니다.

$$f'(x) = \lim_{h \rightarrow \infty} \frac{f(x+h) - f(x)}{h}$$

조금 더 풀어서 설명하자면, 도함수는 원래의 함수로부터 도출된 또 다른 함수로 다른 모든 변수들의 값이 일정할 때, 한 변수의 부분적 변화가 종속변수의 얼마만큼의 변화와 관련이 있는지를 보여줍니다. 따라서 도함수는 변화율을 보여주는 것으로, 직관적으로는 선형관계의 기울기를 보여준다고 할 수 있습니다. 도함수를 가지고 우리는 원함수(original function)가 비선형 관계일 때에도 매 순간의 x 의 변화에 따른 기울기의 변화, 또는 차이를 알 수 있습니다.

- 도함수를 수학적으로 나타내 보겠습니다.

1. 원함수: $f(x) = 2x + 1$
도함수: $f'(x) = 2$
2. 원함수: $f(x) = x^5$
도함수: $f'(x) = 5 \cdot x^4$
3. 원함수: $f(x) = 1 + \frac{2x+7}{x+1}$
도함수: $f'(x) = \frac{-5}{(x+1)^2}$
4. 원함수: $f(x) = x^2 - \frac{4}{x^3}$
도함수: $f'(x) = 2x + \frac{12}{x^4}$
5. 원함수: $f(x) = \sqrt[3]{x} - 6 \cdot \sqrt[3]{x}$
도함수: $f'(x) = \frac{1}{2\sqrt{x}} - \frac{2}{3\sqrt[3]{x^2}}$
6. 원함수: $f(x) = \sqrt[4]{x^3}$
도함수: $f'(x) = \frac{3}{4\sqrt[4]{x}}$
7. 원함수: $f(x) = \frac{1}{x^2}$
도함수: $f'(x) = -\frac{2}{x^3}$
8. 원함수: $f(x) = \frac{x+3}{x+1}$
도함수: $f'(x) = -\frac{2}{(x+1)^2}$

9. 원함수: $f(x) = \sqrt{x}$
 도함수: $f'(x) = \frac{1}{2\sqrt{x}}$
10. 원함수: $f(x) = 2x^3 - 9x^2 + 12x - 3$
 도함수: $f'(x) = 6x^2 - 18x + 12$

- 현실세계의 상황에 가설적으로 이러한 함수적 관계를 적용하여 표현해 보겠습니다.

- 원함수: $f(x) = 2x^3 - 9x^2 + 12x - 3$
 도함수: $f'(x) = 6x^2 - 18x + 12$
- 경제 성장의 경향성에 있어서, 경제성장은 단기에서는 요동치지만 장기에 걸쳐서는 꾸준히 증가하는 증가세를 보입니다.
 위의 함수와 도함수는 경제 성장이 증감하는 경향성과 변동성을 x 가 1, 또는 2일 때를 포착해 보여준다고 할 수 있습니다.

2.1 우리가 사용하는 자료

선형회귀모델을 살펴보기 전에 우리는 분석방법 및 전략을 적용할 대상, 즉 관측된 자료(data)에 대해 이해할 필요가 있습니다. 많이들 헷갈리기도 하는 자료 유형을 각각의 정의를 통해 구분해보도록 하겠습니다. 여기서의 자료의 유형은 R에서 말하는 것과는 조금 다릅니다. R에서의 자료의 유형이란 프로그램 언어로서 엄밀하게 말하면 저장가능한 객체(objects)의 유형을 의미합니다. 그래서 어떻게 저장가능한가에 따라 문자열(string), 숫자형(numeric), 벡터(vector), 매트릭스(matrix), 데이터프레임(dataframe) 등으로 구분하였습니다.

하지만 여기서 말하는 자료란 말 그대로 우리가 관측가능한 자료를 의미합니다. 대개는 시공간적 범주, 그리고 관측된 단위의 동질성과 이질성 등에 따라서 구분하고는 합니다.: 교차사례(cross-sectional), 시계열(time-series), 교차사례 시계열(cross-sectional time-series), 마지막으로 패널(panel) 자료로 그 유형을 구분하여 각각 설명하고자 합니다.

먼저, 교차사례 자료의 경우 획단면 자료라고도 합니다. 간단하게 엑셀을 생각해보면 쉽습니다. 한 연도를 기준으로 어떤 변수에 대해 여러 나라들의 이름을 옆으로 길게 늘어놓는 것을 생각해보면 됩니다. 시간은 특정 시점으로 고정되어 있으니, 그 시점 내에서의 서로 다른 단위들의 차이(variation)을 설명할 수 있는 자료라고 생각할 수 있습니다. 정확히 교차사례 자료는 “동일한 시점에서의 두 개 이상의 공간적 사례들을 가진 자료”라고 할 수 있습니다. 2018년의 OECD 국가들의 국내총생산(GDP) 자료라면 이 기준에 부합할 듯 합니다.

그렇다면 시계열(time-series) 자료란 무엇일까요? 교차사례와는 다르게 종단면 자료라고도 불리는데, 정확하게 교차사례 자료와 대칭적으로 생각하시면 됩니다. 여러 시점에서 동일한 개체에서 수집된 자료입니다. 한국의 1960년부터 2018년까지의 GDP 자료를 엑셀에서 열로 배열해놓으면 이같은 기준을 충족시킬 수 있을 것입니다. 변수로는 아마 연도(year)와 GDP가 있겠네요. 어차피 한국 한 국가만을 대상으로 할테니 국가 변수는 따로 필요 없을 것입니다.

앞의 두 자료 유형은 시간이냐, 공간이냐의 양자택일로 이해하시면 간단합니다. 하지만, 시공간이 결합된 자료의 경우에는 한 가지 차원의 고려가 더 필요합니다. 바로 관측대상-단위의 동질성 또는 이질성이라고 할 수 있겠습니다.

교차사례-시계열 자료(cross-sectional time-series)는 여러 시점에서 동일한 공간적 범주를 대상으로 서로 다른 단위들로부터 수집된 자료를 의미합니다. OECD 국가들의 2000년부터 2018년 사이의 소득 불평등 지표(Gini 계수 등)를 하나의 자료로 만든다면 합동 교차사례 자료라고 할 수 있겠네요. 이 사례에서 자료의 분석단위는 국가-연도(state-year)가 됩니다. 그러나 2000년도부터 2018년도 사이 국가들에서 소득에 대해 응답한 응답자들은 아마 정확하게 동일한 인물들은 아닐 것입니다. 따라서 이 경우에 우리는 단위들이 완전히 동일하다고 단정할 수는 없게 됩니다.

마지막으로 패널 자료는 여러 시점에서 동일한 개체에 대해 수집된 자료를 의미합니다. 설문조사를 예로 들어보겠습니다. 2000년부터 2018년까지 동일한 응답자들을 해마다 추적하여 응답을 수집할 경우, 이는 패널자료라고 할 수 있습니다. 단, 이 경우 연구자는 시간의 흐름에 따라 동일한 표본을 추적하므로 필연적으로는, 의도적으로는 응답이 한 시점에서 누락될 경우 표본의 전체 수는 감소하는 모습을 보이게 됩니다. 따라서 패널 자료를 구축하는 것에는 시간과 자금이 많이 소요되고 시간이 지날수록 공통 표본의 규모가 감소한다는 문제가 있습니다.

2.2 추론(Inference)과 변수(variables), 그리고 측정 수준(척도)

우리는 종종 주어진 자료로 어떠한 결과를 “추론”한다고 합니다. 그렇다면 대체 추론이란 무엇일까요? 추론은 관측가능한 사실로부터 관측불가능한 결과를 도출해내는 일련의 논리적 작업이라고 할 수 있습니다. 만약 우리가 관측가능한 표본들로부터 모집단의

관측불가능한 특성이 어떻게 생겼을 것이라고 추론하고자 한다면, 우리는 기술추론(descriptive inference)을 하게 됩니다. 모집단을 대표적으로 잘 보여준다고 하는 표본이 이렇게 생겼으니 아마도 모집단도 표본처럼 어떠할 것이라고 추론하는 것이지요. 한편, 모집단의 관측불가능한 변수들 간의 인과적 관계를 관측가능한 표본에서 추론하고자 할 때에는 “인과적 추론”(causal inference)을 하게 됩니다. 그리고 사회과학의 본령(本領)은 주로 이 인과적 추론에 놓여있다고 합니다.

자, 이제 변수에 대해 생각해보겠습니다. 흔히들 “그래서 네 연구의 변수가 뭐냐”라는 식으로 질문을 던지고는 합니다. 그런데 변수란 무엇일까요? 왜 우리는 변수를 보는 것일까요? 변수란 특정한 부류의 어떠한 현상, 또는 대상의 공통적인 특성을 포함하는 개념을 의미합니다. 따라서 변수는 개념(concept) 그 자체와는 조금 다르다고 할 수 있습니다. 왜냐하면 변수는 다른 개념들과 비교해 어떠한 개념이 “얼마나” 구별이 되는지를 보여주는 것이기 때문입니다. 따라서 변수는 체계적인 방식으로 측정된, 어떠한 대상의 속성의 “양(amount)”을 보여주는 특정화된 개념(specified concept)이며, 일종의 용기(container)라고 할 수 있습니다.

그렇다면 어떻게 그 정도를 측정하는가에 따라서 변수의 모습은 달라질 수 있다고 쉽게 예상할 수 있습니다. 관측하고자 하는 개체, 현상이 가지고 있는 자연상태의 정보를 기술하는 지표(indicator)로 바꾸는 방법이 바로 측정수준(measurement level)의 문제입니다.

- 연속형(continuous): 만약 측정 수준이 일정한 간격(interval)을 가진 연속된 수로 이루어져 있는 경우를 말합니다. 연속형으로 측정된 변수의 경우에는 가능한 각 값들 간의 특정한 수치적 거리를 특정할 수 있습니다. 예를 들어, 1과 3 간의 거리가 2인 것과 5와 7 간의 거리 2는 동일한 거리라고 말할 수 있는 것이죠. 따라서 어떤 책에서는 연속형을 나누어 이와 같은 일정한 간격, 등간(constant interval)을 가지느냐의 여부로 등간척도를 따로 분류하고는 합니다. 예를 들어, GDP, 인구(명), 연령(년) 등이 연속형으로 간주될 수 있습니다.
- 분류형(categorical): 일련의 분류목(categories)을 가지는 측정 수준을 의미합니다. 분류형 척도 하위로는 명목척도, 순위척도 등을 놓을 수 있습니다.
 - 명목형(nominal): 순위를 매길 수 없이 상호 배타적인 분류목들을 가지는 측정 수준을 의미합니다. 종교, 인종, 혈액형과 같은 경우 어떠한 부류가 다른 부류에 비해 우월하다고 말할 수 없고 서로 공통되지 않은 배타적인 영역을 가집니다.
 - 순위형(ordinal): 한편, 순위형은 분류목들 간의 상대성을 보여줄 수 있는 측정수준을 의미합니다. 문자열로 된 성적(A, A-, B, 등), 소득 구간(상층, 중산층, 하층 등), 그리고 교육 수준(초등학교 졸업, 중학교 졸업, 고등학교 졸업 등) 등이 이에 속하는 대표적인 변수들이라고 할 수 있습니다.
 - 이항(binary): 마지막 측정 수준, 척도는 이항척도인데, 흔히들 이렇게 만들어진 변수를 더미변수(dummy variable)이라고도 합니다. 어떠한 분류목의 특성이 존재하느냐의 여부만을 보여주는 변수로 존재할 경우에는 1, 존재하지 않을 경우는 0으로 표시하며, 이외의 어떠한 정보도 제공하지 못한다는 점에서 “멍청한 변수”라는 의미로 더미변수라는 이름이 붙었습니다. 예를 들어, 냉전 기간을 자료에 입력할 때 특정 연도가 냉전 기간에 속해있으면 1, 아니면 0으로 기입하거나 질병의 감염유무, 성별(남성 = 0, 여성 = 1) 등에서 살펴볼 수 있습니다.

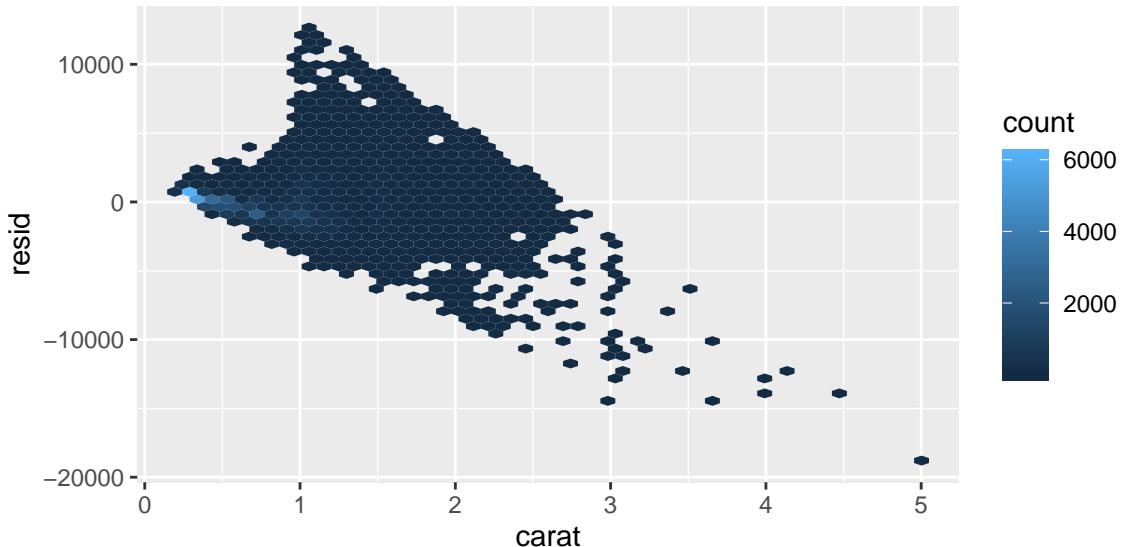
2.3 선형회귀모델 간단히 다시보기

Lv.1.Statistics 자료에서 제6장부터 제8장까지 살펴보았던 선형회귀모델을 간단하게 다시 훑어보도록 하겠습니다. 먼저 diamonds 데이터를 불러오고 다이아몬드의 가격과 캐럿(carat) 간의 단순선형회귀모델을 만들어보겠습니다. 모델의 이름은 mod_diamonds입니다.

```
library(ggplot2)
library(modelr)
library(hexbin)
library(ggpubr)
mod_diamond <- lm(formula=price ~ carat, data=diamonds)
```

자, 선형회귀모델을 아주, 아~주 간단하게 말하면 “우리가 관심을 가지고, 알고 싶어하는 y 의 변화를 어떠한 x 의 변화가 설명할 수 있을 것이라는 기대를 수식으로 표현한 것”이라고 할 수 있습니다. 그리고 현실 속에서 우리가 관측할 수 있는 것의 한계는 명확하므로, 아무리 이론적으로 y 를 설명할 것이라고 기대되는 x 라고 할지라도 완전히 y 를 설명할 수는 없습니다. 따라서 우리의 모델은 x 로 y 를 설명했을 때, 그 예측값(fitted values)과 실제값(actual values; observations) 간의 차이가 최소가 되는(least-squared) 선을 그릴 때 가장 효율적이고 편향되지 않은 결과(BLUE)를 가져다줄 것입니다. 이때, 개별 관측치들과 모델의 예측값 간의 차이를 모집단의 수준에서는 오차(errors), 표본의 수준에서는 잔차(residuals)라고 합니다.

수리적으로는 예측값(\hat{y})과 실제값의 차이(y_i)이므로 잔차는 $y_i - \hat{y}$ 로 구해줄 수 있지만, 여기서는 R의 `residuals()` 함수를 이용하여 간단히 구해주도록 하겠습니다. 그리고 그렇게 구한 잔차와 주요 예측변수, 캐럿 간의 관계를 플롯으로 그려보겠습니다. 왜냐하면 선형회귀모델은 어디까지나 우리가 관측가능하여 종속변수를 설명할 것이라고 기대한 예측변수(들)와 잔차가 서로 독립적이라고 가정하기 때문입니다. KKV (1994)의 표현대로라면, y 의 전체 변화 중 체계적인 것(systematic)들을 예측변수로, 비체계적인 것(non-systematic)들은 잔차로 남겨지는 것이고, 이때 비체계적이란 특정한 경향성이 없는 확률적(probabilistic)이라는 의미를 지닙니다. 따라서 선형회귀모델이 적합하다면 잔차와 설명변수 간에는 어떠한 경향성이 나타나지 않아야 할 것입니다.

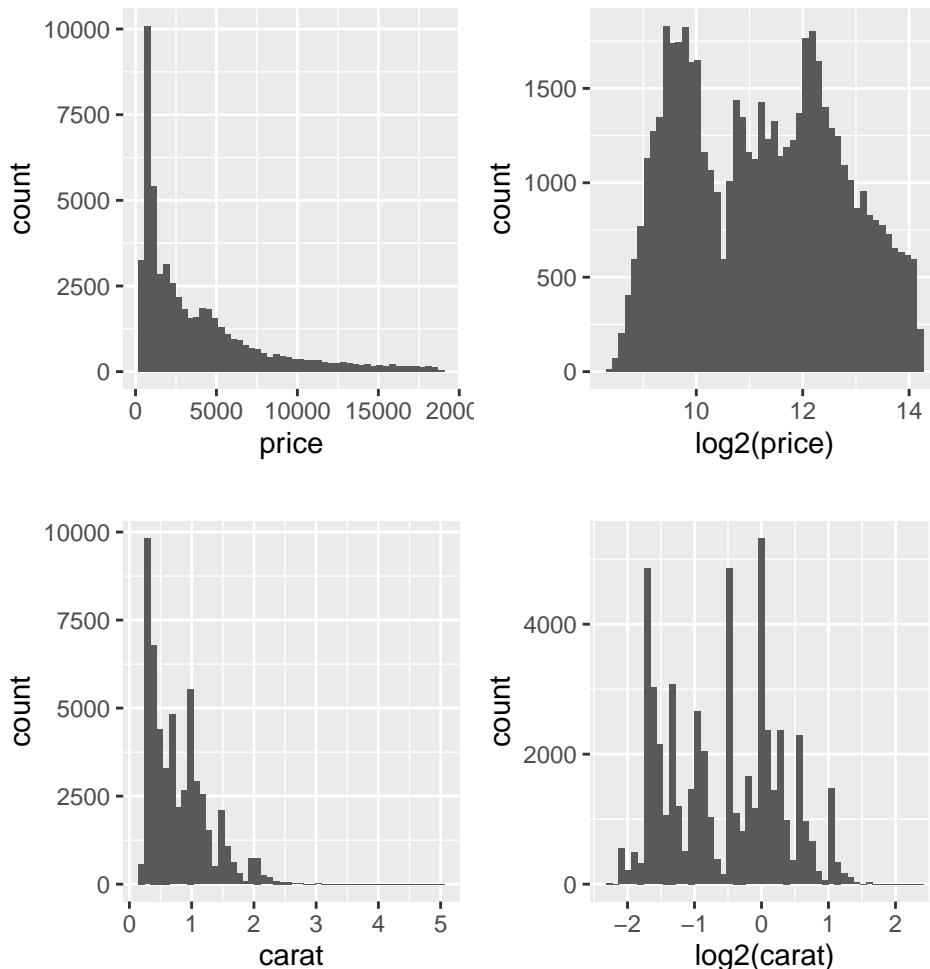


위의 플롯은 우리가 추정한 OLS에 대해 어떠한 정보를 전달해주고 있을까요? 앞서 언급한대로 잔차는 모델의 관측치와 예측치 간의 차이라고 할 수 있습니다. 따라서 여기서의 잔차는 `carat`으로 설명되지 않는 `price`의 변화라고 말할 수 있습니다. 최소자승법, OLS 회귀모델은 모집단의 오차가 일정한 분산(constant variance, homoskedasticity)을 가진다고 가정하고 있습니다. 회귀모델의 가정을 충족시키고, 그 결과가 신뢰할만한 것이기 위해서는 표본 수준에서의 잔차 역시도 동분산에 가까운 결과를 가지고 있어야만 합니다. 즉, 우리가 이 모델로 y 를 적절하게 설명하고 있다고 말하고 싶다면, 잔차는 무작위로(randomly) 분포되어있어야만 합니다.

그러나 위의 플롯은 잔차와 설명변수 간의 일정한 관계, 체계적 경향성(systematic pattern)을 보여주고 있습니다. 이 경우에는 두 가지 가능성을 생각해볼 수 있습니다. 첫째, 우리가 모델에 포함시킨 변수들이 선형 관계를 가지지 않기 때문에, 변수들을 조작(manipulate)³해줄 필요가 있는 경우입니다. x 의 변화와 y 의 변화가 선형으로 연계되어 있지 않기 때문에 잔차의 이질적인 분포를 보여주는 것일 수 있습니다. 잔차의 이분산성(heteroskedasity)은 데이터의 최소값과 최대값 사이의 범위가 클 때에 종종 발견할 수 있습니다. 한 번 종속변수와 예측변수인 `price`, `carat`의 분포를 원자료일 경우랑 로그값을 취해줬을 때를 각각 히스토그램⁴으로 비교해서 살펴보겠습니다.

³이때, 조작(manipulation)이란 변수의 값을 다르게 임의로 입력한다는 것이 아니라 변수의 측정 단위 등을 조정한다는 것을 의미합니다. 같은 변수들 간의 관계라도 그것의 측정 단위 등에 따라서 결과가 다르게 나타날 수 있기 때문입니다.

⁴한 변수의 분포를 살펴볼 때에는 히스토그램으로 보는 것이 좋습니다.



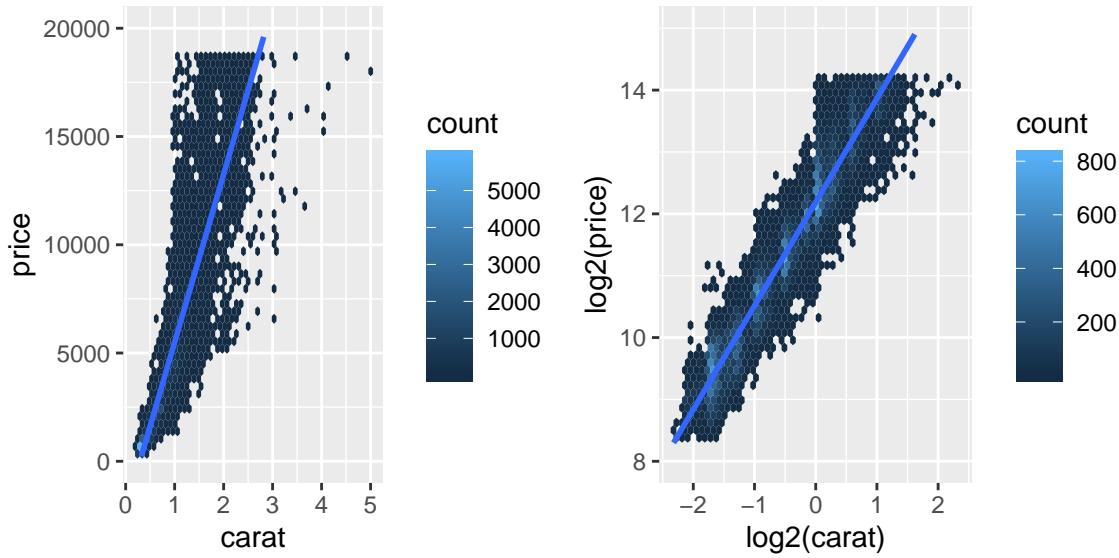
각 변수의 원자료를 보여주는 좌측 패널들에서는 치우친 분포의 양상이 보이는 한 편, 로그값을 취해줬을 경우 상대적으로 정규분포와 같은 종 형태의 분포를 보이는 것을 확인할 수 있습니다. 이같은 결과는 왜 *price*와 *carat* 간의 비선형 관계가 존재할 수 있는지 그 가능성을 엿볼 수 있게 합니다.

둘째, 또 다른 가능성으로는 우리의 모델이 y 를 충분히 설명할 수 있는 체계적 요인을 제외하고 있을 경우, 누락변수의 문제(omitted variable problem)가 있습니다. 만약 체계적으로 y 의 변화에 영향을 미칠 수 있는 변수임에도 불구하고 모형의 예측변수에서 제외되었다면, 그것은 오차 속에서 일종의 교란변수로 x 와 y 모두에 영향을 미쳐 관계 양상을 왜곡시킬 수 있습니다.

그렇다면 이러한 잠재적인 문제들을 어떻게 다루어 우리의 추정을 개선시킬 수 있을까요? 첫 단계로 *price*와 *carat* 간의 잠재적 비선형성의 문제를 먼저 다루어보도록 하겠습니다. 위의 히스토그램에서 살펴본 것과 같이 저는 두 변수 간의 비선형성의 문제가 *price*와 *carat*의 원자료가 매우 치우쳐진 분포를 가지고 있기 때문이라고 의심하고 있습니다. 따라서 두 개의 플롯을 그려보는데, 하나는 원자료 간의 관계를 보여주는 것이고, 다른 하나는 로그값을 취한 변수들 간의 관계를 보여주는 것입니다. `geom_hex()` 옵션을 덧붙이는 이유는 변수들 간의 관계에 더불어 값들이 어디에 중점적으로 모여있는지를 함께 보여줄 수 있기 때문입니다. 색이 옅을수록 관측치들이 모여있다고 보시면 됩니다.

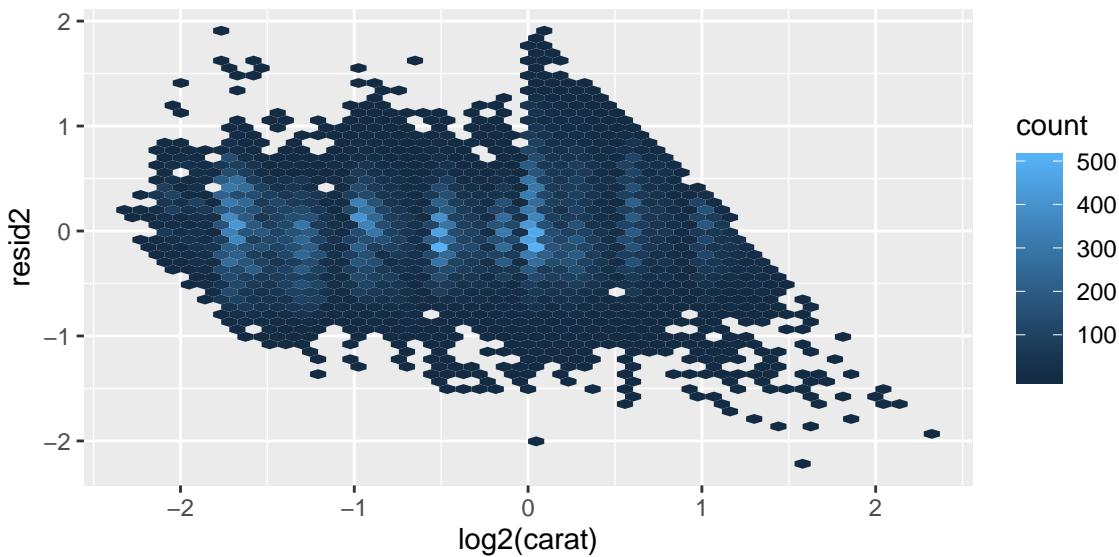
아래의 플롯을 보면, 좌측 패널이 변수들의 원자료들 간의 산포도를 그린 것이며 우측 패널이 로그값을 취한 변수들 간 산포도를 그린 것입니다. 좌측 패널의 경우 일견 선형으로 보이지만⁵ *carat*이 작고 *price*가 작은 수준에 값들이 밀도있게 모여있는 것을 확인할 수 있으며, 가파르게 우상향하는 이차함수의 $x > 0$ 인 곡선의 형태를 보여주고 있는 것을 보여줍니다. 아마 *carat*의 값이 4 이상인 경우에 분포되어 있는 극단적인 관측치들이 아니었다면 더 직관적으로 관계의 비선형성을 확인할 수 있었을 것입니다. 반면, 두 변수에 로그를 취한 경우, 상대적으로 추세선 위에 관측치들이 모여있는(옅은 색) 것을 확인할 수 있고, 선형관계를 보이고 있습니다.

⁵두 플롯을 비교하기 편하도록 동일하게 `geom_smooth(method = "lm")`을 설정하여 직선으로 보일 뿐입니다.

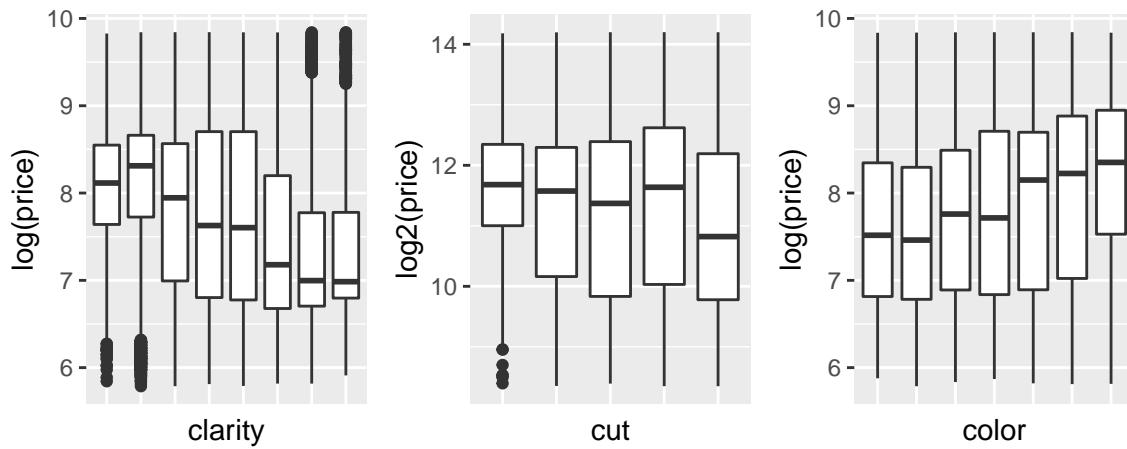


이번에는 종속변수인 `price`와 예측변수인 `carat`에 로그값을 취했을 때, 그 모델 결과로 구해진 잔차와 예측변수 간의 관계를 한번 살펴보겠습니다.

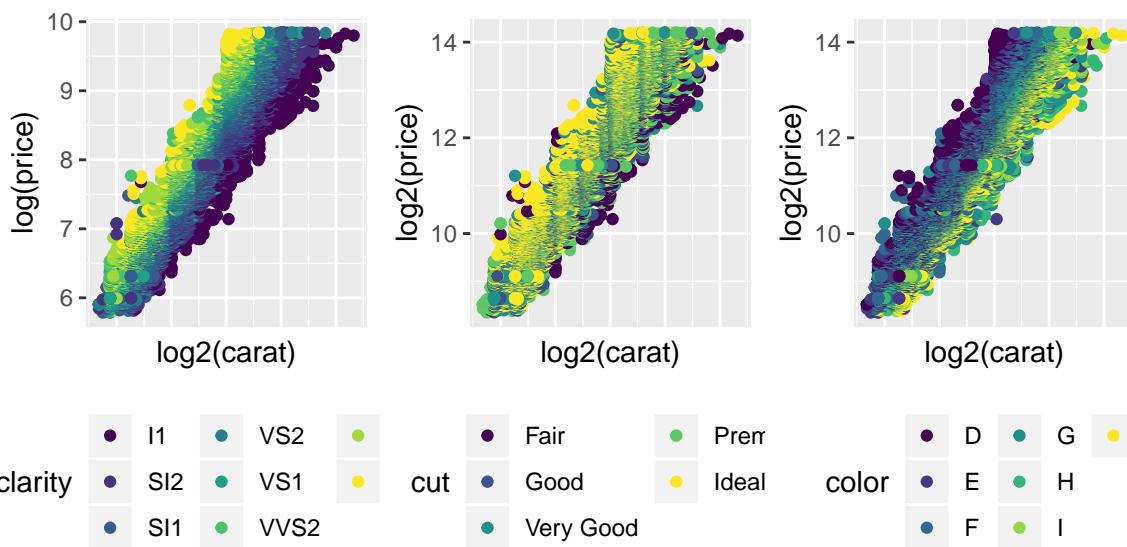
```
mod_diamond2 <- lm(formula= log2(price) ~ log2(carat), data=diamonds)
diamonds$resid2 <- residuals(mod_diamond2)
```



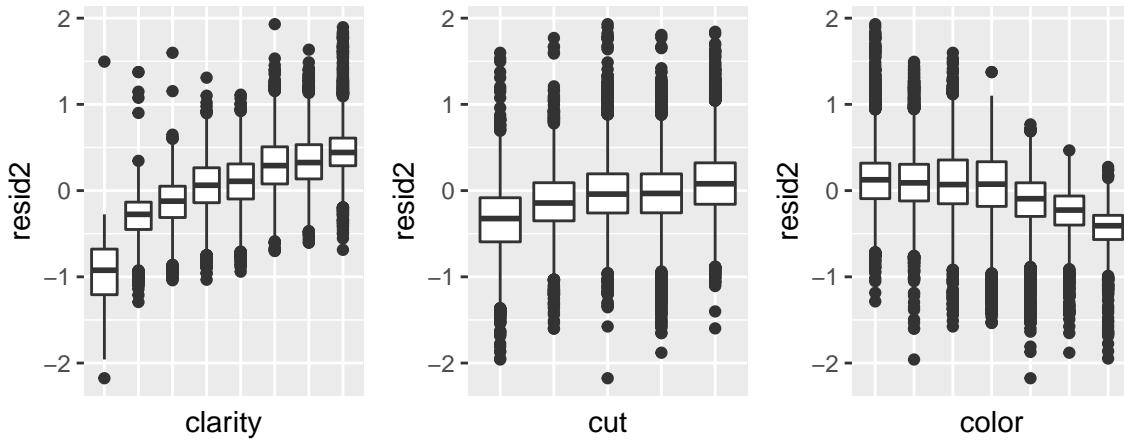
플롯 결과를 보면, 첫 번째 문제점으로 볼 수 있었던 비선형성으로 인한 잔차와 설명변수 간의 상관성의 문제는 로그값을 취함으로써 일부 해결된 것으로 보입니다. 그렇다면 누락변수 문제는 어떻게 다룰 수 있을까요? 저는 `diamonds` 데이터셋에서 `price`와 `carat` 간 양변량 관계에서 `price`에 영향을 미칠 수도 있을 것이라 기대되는 변수들을 모델에 투입해보았습니다. 예를 들어, `clarity`, `cut`, 그리고 `color` 등은 다이아몬드의 질(quality)과 관련된 것으로 `price`에 영향을 미칠 수도 있을 것이라고 기대한 것입니다. 이러한 이론적 기대를 바탕으로 저는 다이아몬드의 질을 보여주는 변수들과 `price` 간의 관계를 보여주는 플롯들을 글보았습니다. 이때, `price`는 로그값을 취한 결과를 사용하였습니다.



플롯의 세 변수와 price 간 관계를 보여주는 각 패널들은 다이아몬드의 질이 가격과 어떤 방식으로든 연관이 되어 있다는 것을 보여줍니다. 그렇다면 각 변수들이 carat과 price의 양변량 관계에 투입될 때 어떠한 양상으로 나타나는지를 살펴보도록 하겠습니다.



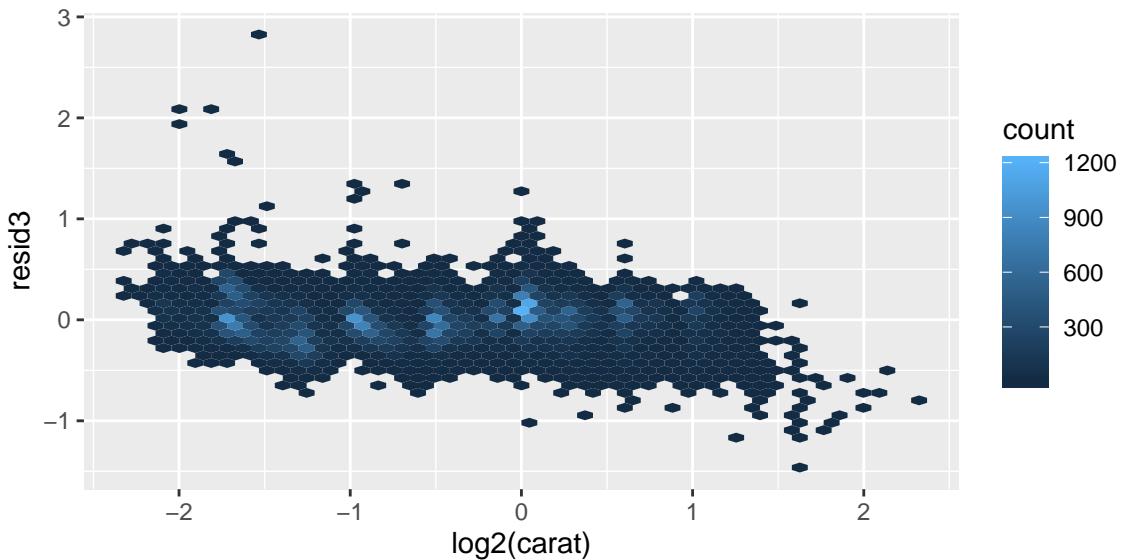
플롯에서 살펴볼 수 있는 이론적 기대와 경험적 관찰을 바탕으로, 저는 price를 설명하기 위한 선형회귀모델에 다이아몬드의 질을 보여주는 clarity, cut, color를 새로운 예측변수로 추가하기로 결정하였습니다. 그렇다면 모델에 새롭게 추가된 각 변수들은 앞의 단순선형모델의 잔차와 어떠한 관계를 맺고 있을까요?



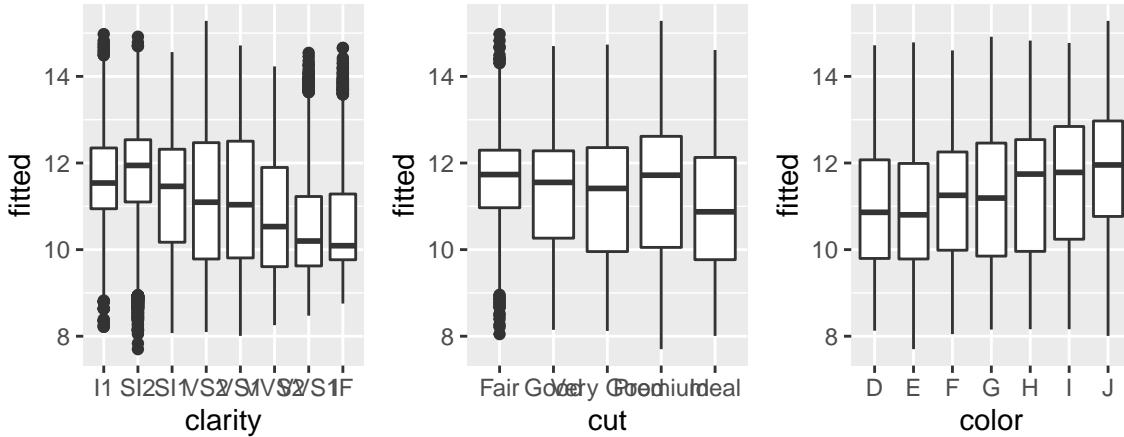
clarity, cut, color 모두가 잔차와 상관이 있는 것 같습니다만. 우리는 단순선형회귀모델의 잔차로부터 체계적 요인—clarity, cut, color를 뽑아내어 price를 설명하기 위한 모델에 투입하기로 결정하였습니다.

```
mod_diamond3 <- lm(formula= log2(price) ~ log2(carat) +
                      clarity + cut + color, data=diamonds)
```

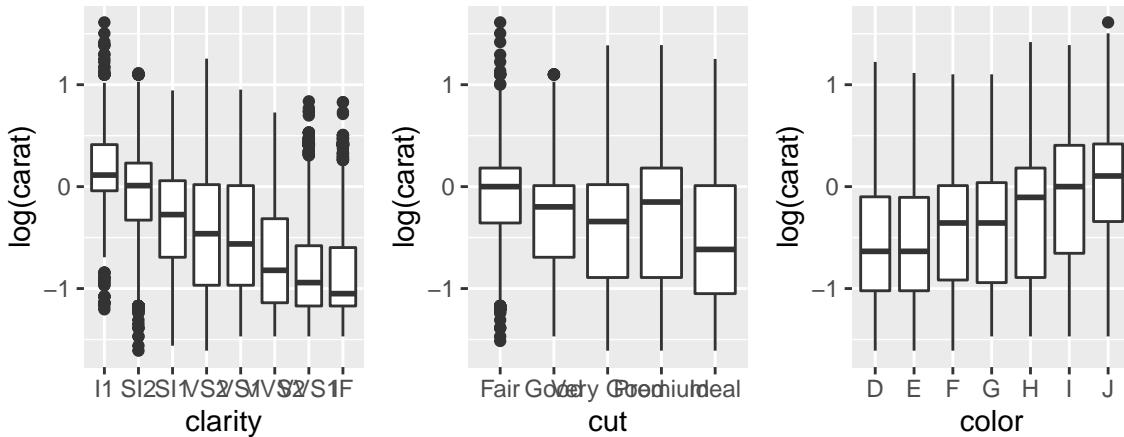
그리고 새롭게 구한 다중선형회귀모델, mod_diamond3를 추정하고 그 결과로 얻은 잔차와 예측변수 carat 간의 관계를 살펴보았습니다.



다중선형회귀모델의 잔차와 로그값을 취한 carat 간에는 뚜렷한 경향성이 존재하는 것 같지는 않습니다. 이번에는 다이아몬드의 질을 보여주는 변수들과 다중선형회귀모델로 예측한 price 값 간의 관계를 플롯으로 보여보겠습니다.



네 개의 예측변수를 가지고 추정한 다중선형회귀모델로 그린 플롯은 상대적으로 단순선형회귀모델에서 충족시키지 못했던 선형회귀모델의 가정에 일부 부합하는 결과로 개선되었음을 확인할 수 있습니다. 먼저, 잔차는 로그값을 취한 carat과 체계적으로 상관하지 않습니다. 이는 잔차가 무작위로 분포되어 있다는 것을 의미합니다. 둘째, 예측된 price의 값과 다이아몬드 질 간의 관계는 관측된 price 값과 다이아몬드의 질 간의 관계와 유사한 양상을 보입니다. 이는 우리의 다중선형회귀모델이 주어진 예측변수들로 price를 잘 예측하고 있다는 것을 의미합니다.



그러나 이 모델은 충분하지 않습니다. 실질적으로 다이아몬드의 질을 보여주는 변수들은 주요 예측변수인 carat과 상관관계가 있음을 확인할 수 있기 때문입니다. 예를 들면, 가장 clarity 수준이 높은 다이아몬드는 극도로 그 크기(carat)가 작은 것으로 나타납니다. 마찬가지로 color도 carat이 작을수록 높은 수준을 보여주고 있습니다. Ideal한 cut 다이아몬드는 전형적으로 작은 크기의 다이아몬드들에서 나타나고 있습니다.

위의 세 박스플롯들은 주요 예측변수 carat과 다른 다이아몬드의 질적 변수들 간의 상관성을 보여주고 있습니다. 즉, 이 네 예측변수를 이용한 다중선형회귀모델은 선형회귀모델의 기본적인 가정 중 하나인 예측변수들 간의 독립성을 심각하게 위배하고 있을 가능성을 시사합니다. 다이아몬드의 가격을 적절하게 예측하기 위해서는 예측변수들 간의 가산적 관계(additive relationship)를 가정하는 선형회귀모델을 새롭게 구축할 필요성이 있습니다.

또한 주요 예측변수와 종속변수에 로그값을 취했기 때문에 변수들 간의 정확한 효과를 직접적으로 해석하는 것은 어렵습니다. 애초에 원자료의 예측변수와 종속변수 간의 관계가 선형이지 않았기 때문에 그들의 관계를 로그값의 결과로 선형으로 해석할 수는 없기 때문입니다. 비록 변수에 로그를 취해서 선형회귀모델의 가정을 충족시켰다고는 하지만, 로그값을 취한 변수들을 사용한 모델의 결과를 통계적으로, 그리고 실질적으로 해석하는 것은 또 다른 문제입니다.