가장 먼저 구조를 알면 편하다.

회귀분석을 왜 하는지 알아야 뒤에 나오는 내용들이 뒤섞이지 않고 찬찬히 정리가 된다.
우선 사고의 흐름을 정리하고 각각의 부분을 각론에서 설명하겠다.
회귀분석은 4단계로 나뉜다.

1) 두 변수 사이에 선형적 관계가 있다고 강력히 의심될 때, 회귀분석을 행한다.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$ (모 회귀식)

즉 이론적 뒷받침이 있어야 한다는 얘기다. 그냥 아무거나 끼워맞춰서 회귀분석에 넣는 게 아니라는 뜻이다.
위의 식으로 표현되는 선형관계가 이미 존재한다고 가정한다.

2) 적절한 방식으로 수집된 적절한 데이터를 통해서 두 변수 간 선형관계를 가장 잘 설명해주는 직선을 찾는다.

$$Y_i = b_0 + b_1 X_i + e_i$$ (표본 회귀식)

그런데 실제로 두 변수 간에 관계가 있는지 알 길이 없다. 그래서 늘 그래왔듯이 통계에서는 표본을 추출한다.
표본을 추출할 때에는 올바른 방식으로 해야한다. 뭐가 올바른 거냐 하면 표본 추출 내용만 서술하는 두꺼운 책들이 있다. 그러니 원론에서는 잘 추출되었다고 가정한다.

주어진 표본 자료를 통해서 가장 잘 알맞는 직선, 즉 회귀직선을 긋는다.
도대체 어떻게 가장 잘 들어맞는 직선을 그릴까? 여기서 등장하는 게 바로 최소자승법이다. 아마 여기까지가 교과서 회귀분석 파트에서 다루는 내용일 거다. 뒷 부분은 추정 파트에서 다룰 테고.

3) 우리가 알고 싶은 건 결국 $\beta_0, \beta_1, \varepsilon$ 다. 주어진 표본과 표본을 통해 계산한 b_0, b_1, e를 통해 $\beta 0, \beta 1, \varepsilon$를 추정한다.

우리가 알고 싶은 건 그리스 문자로 표시된 실제 두 변수 간 관계이다. 신뢰구간 파트에서 주구장창 구했던 거랑 같은 내용이다.
참고로 ε는 오차error이고 e는 잔차residual다. 둘을 헷갈리면 안 된다. 전자는 모수parameter이고(모르는 값) 후자는 모수의 추정치estimator이자 표본통계치statistic이다.

4) 추가적으로 회귀분석을 통해 예측 또한 할 수 있다. 예측이란 표본에는 주어져 있지 않은 X값을 대입해 Y값을 어림해보는 걸 의미한다. 이를 extrapolation이라고 한다.

놀랍게도 단순회귀분석의 모든 내용은 이 네 가지 단계 어딘가에 끼워넣어 설명할 수 있다. 아래를 보라.

1단계: 이론의 가정
CLRM(Classical Linear Regression) 이론의 7가지 가정

2단계: 최소자승법과 회귀직선
최소자승법
SST, SSE, SSR
R^2

3단계: 가설검정
β_j에 대한 추정과 가설검정

4단계: 예측
평균치에 대한 예측
개별값에 대한 예측

정확히 내가 지금 구하고 있는 값이 무엇에 해당되며 회귀분석의 어디부분에 아는지 만으로도 많은 의문이 해결될 것이다.

이제부터 각론에 대해서 설명하고 뒤에 다중회귀분석까지 서술하고자 한다.

3단계. 모수 추정 (실제 선형관계 추정)

지금까지 복잡한 계산을 왜 했는지 알아야 한다. 결과적으로 우리가 알고 싶은 것은 실제 두 변수 간의 선형관계이다. 우리는 지금까지 표본데이터를 바탕으로 회귀직선을 긋는 걸 해온거다. 이렇게 해서 구한 b_0, b_1, e, s를 통해서 β_0, β_1, ε, σ 을 알고 싶다는 것이다.

회귀분석을 배울 정도면 기본적인 가설검정의 단계는 전부 알 것이다.
귀무가설을 세우고, z 통계량이나 t 통계량을 구한 뒤, p value를 구해서, 귀무가설을 기각하든 하지 못하든 하는 것이다.

마찬가지다.

여기서도 β에 관한 귀무가설을 세운 뒤, p value를 구해서, 그걸 기각하든 하지 못하든 결론을 내고자 한다.

우리가 가지고 있는 값은 b_0, b_1이다.

1단계에서 서술한 7가지 가정이 모두 만족한다면,

$b_0 \sim N(\beta_0, var(b0))$
$b_1 \sim N(\beta_1, var(b1))$
가 성립한다.

표본 통계량으로 확률변수이다.

그렇다면 b_1, b_0의 분산만 안다면 각 값의 분포를 명확하게 알 수 있다. 이 계산은 수리통계학의 영역이고, 원론에서는 그 계산과정을 생략한다. (왜 분포를 알아야 하나요? 라는 질문이 나올 수 있다. 조금만 생각하면 알 수 있다. 표본평균을 통해 모집단 평균을 추정할 때를 생각해보자. 표본평균의 평균과 분산을 통해 가설검정을 하지 않았나? 그거랑 같은 이치다)

$$\mathrm{var}(b_0) = \frac{\sigma^2 \sum X_i^2}{n\sum (X_i - \overline{X})^2} \Rightarrow s_{b_0} = s\sqrt{\frac{\sum X_i^2}{n\sum (X_i - \overline{X})^2}}$$

$$\mathrm{var}(b_1) = \frac{\sigma^2}{\sum (X_i - \overline{X})^2} \Rightarrow s_{b_1} = s\sqrt{\frac{1}{\sum (X_i - \overline{X})^2}}$$

왼편의 두 식은 오차항의 분산 σ가 알려져 있을 때 쓰는 식이고, 오른편의 두 식은 그걸 모를 때 쓰는 식이다. σ 대신 잔차항의 분산 s를 이용한다. s는 SSE를 n-2로 나눈 것이라고 바로 2단계에서 배웠다.

σ가 알려져 있으면 z통계량을 구한 뒤 정규분포를 이용하고, σ를 모를 때에는 t 분포를 이용해 가설검정을 한다. (t 분포를 이용할 경우 자유도는 n-2로 설정한다.)

그럼 귀무가설로 무엇을 둘 것인지를 알아야겠다.

그때그때 다르나 대부분의 경우

$H_0: \beta_1 = 0$

으로 두고 가설검정을 하는 때가 많다.

이유는 간단하다.

힘들게 데이터를 모아서 회귀직선을 그려서 두 변수 간의 관계인 기울기를 구했고, 그래서 X와 Y간에는 b_1만큼의 관계가 존재한다라고 말하고 싶은데,

어떤 사람이, 그건 우연의 일치 아니야? 실제로 두 변수 간에는 아무런 관계가 없을 수 있잖아? 라고 묻는 것이다.

`b1`은 표본을 통해 구한 통계량이기 때문에, 오히려 저는 `0`은 `b1`이 아닌 `0`일 수도 있다는 의문을 가질 수 있다.

`b1`만큼의

그게 저 귀무가설이 뜻하는 바이고, 그걸 기각할 수 있어야, 자신있게 두 변수 간에는 선형관계가 존재한다고 말할 수 있기 때문이다.

(물론 연구 목적에 따라 다른 귀무가설을 설정할 수도 있다.)

가설검정의 방법은, 그간 해온 거랑 똑같다.

s_{b0}, s_{b1} --> 얘들이 표준오차라는 걸 알 것이다. b0, b1을 표준오차로 각각 나눈 것은 t 통계량이 되고, 그 값이 클수록 p value는 작다. 그래서 귀무가설을 기각할 수 있다.

(t 통계량은 정확히 말하면 b_1의 경우, b_1에 b_1의 기댓값을 뺀 것을 표준오차로 나눈 것이다. 근데 b1의 기댓값은 β_1이고 이를 귀무가설에서 0이라고 가정했으므로 그냥 b_1을 표준오차로 나누기만 하면 된다. b_0의 경우도 마찬가지다.)

앞서 소개했듯이 회귀 분석은 '조건부 평균'을 구하는 것입니다. 때문에 우리가 평균을 구할 때 주의해야 할 점들이 회귀 분석에서도 동일하게 적용됩니다. 그럼 어떤 주의점이 있는지 예시를 통해 살펴 보죠.

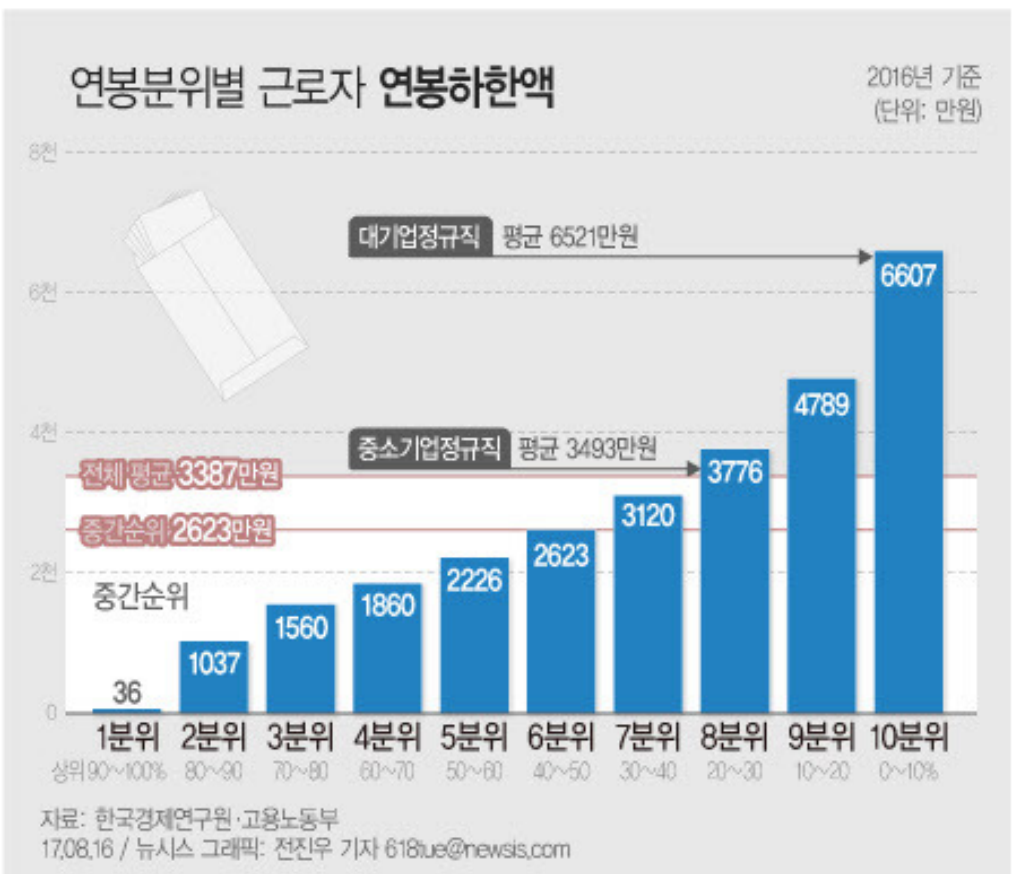
아래 두 데이터 중 평균을 대푯값으로 사용하기 부적절한 것은 무엇일까요?

1) 우리 나라 성인 남성의 키

2) 우리 나라 전체 근로자 연봉

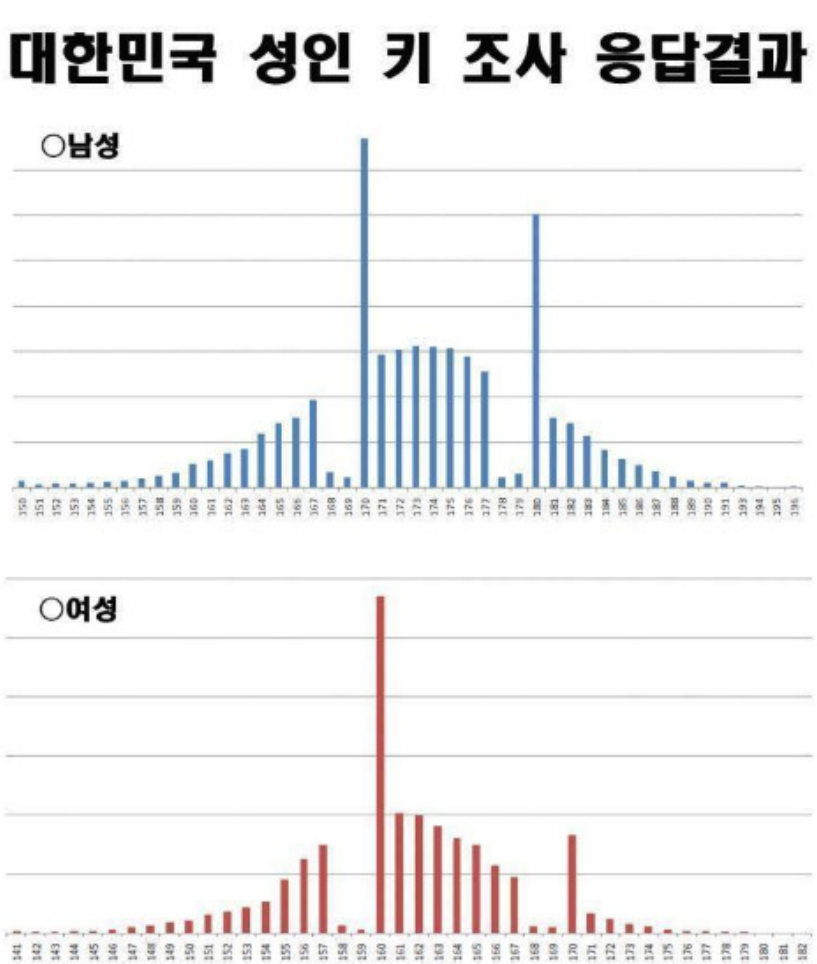
아래 자료는 2016년 기준 우리나라 근로자들의 평균 연봉 및 분위별 연봉하한액을 나타낸 그래프입니다. 아마 뉴스나 신문기사를 통해 아래와 같은 표를 한번쯤 보셨을 겁니다. 아래 자료에 의하면 2016년 기준으로 우리 나라 근로자 평균 연봉은 3387만원이고, 대기업 정규직의 경우 6521만원이라고 합니다.

그런데 그래프를 자세히 보면 중간값이 2623만원입니다. 다시말해, 전체 근로자 중 절반은 연봉이 2623만원이 안된다는 뜻입니다. 반면 평균 연봉 이상을 받은 사람은 전체의 30% 정도에 불과합니다. 아무래도 '평균 연봉 3387만원' 이라는 수치를 접하고선 자괴감에 빠지는 사람들이 꽤 될 것 같습니다. 결국 전체 집단의 연봉을 대표하는 값으로 평균을 사용하는 것은 그리 좋은 선택이 아닌 것 같습니다.



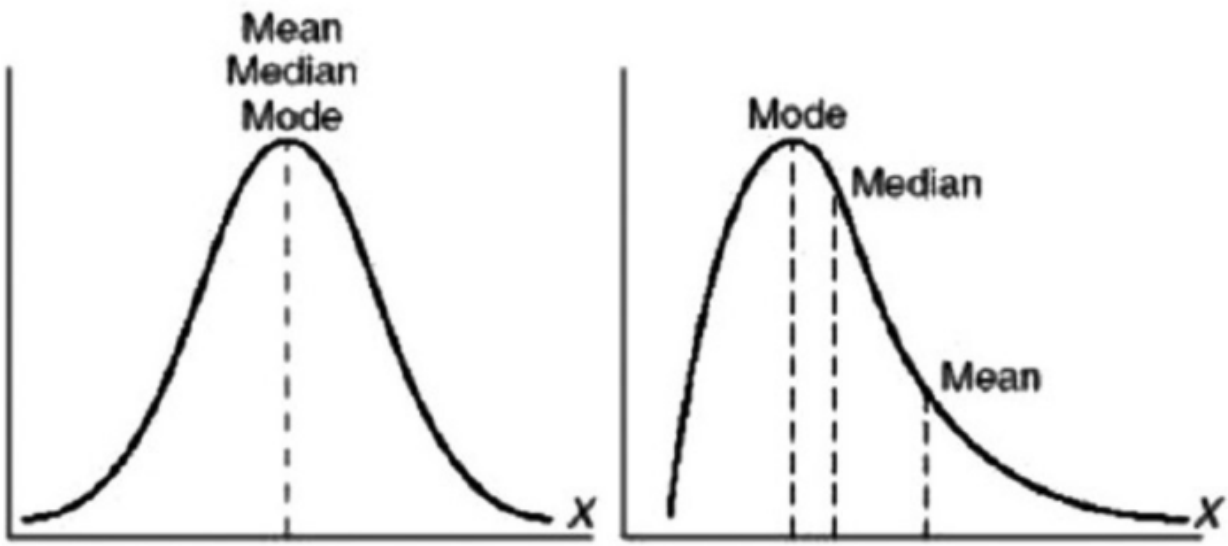
대다수의 직장인들에게 자괴감을 주는 자료의 예 (출처: http://bitly.kr/LUj2nP6r)

반면, 성인 남성의 키 분포는 종형 분포에 가깝기 때문에 대략 평균 키가 전체 집단의 가운데 쯤에 위치합니다. 그러니 이 경우에는 평균이 우리 나라 성인 남성의 대략적인 키를 대표한다고 보는데 (적어도 연봉과 비교했을 때보다는) 크게 문제가 없을 것 같습니다.



물론 모든 자료에는 왜곡이 존재합니다... (출처: https://www.clien.net/service/board/park/7152183)

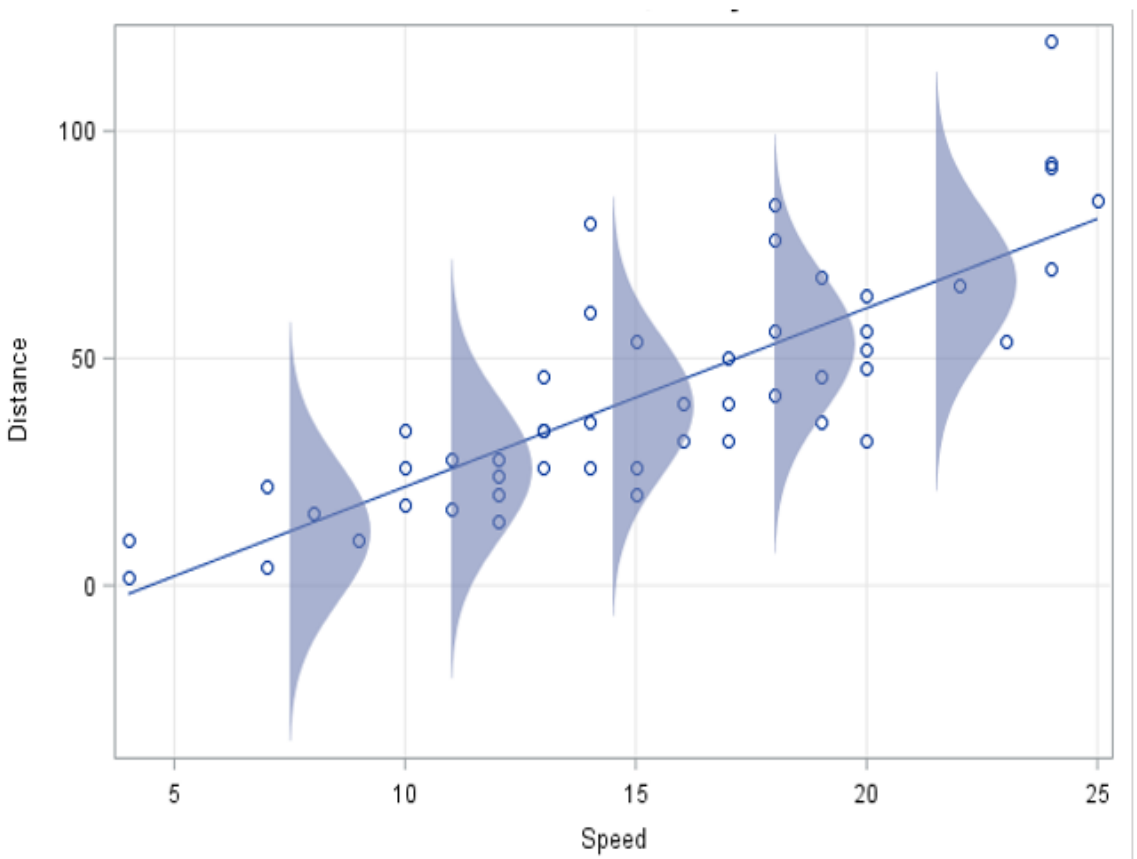
정리하자면, 평균을 전체 집단을 대표하는 값으로 사용하기 좋은 경우는 데이터가 아래 그림의 왼쪽 그래프처럼 좌우 대칭이 되는 종형 분포인 경우입니다. 종형 분포에서는 평균값이 중간값이나 최빈값과 거의 일치하기 때문에 대푯값이 되기에 적절합니다. 반면 오른쪽 그래프처럼 한쪽으로 분포가 쏠리게 되면 이 세 수치가 서로 달라지기 때문에 평균값을 대푯값으로 사용하기에는 그리 적절치 않습니다.



분포의 형태에 따른 평균, 중앙값, 최빈값의 대략적인 위치

회귀 분석 역시 평균을 구하는 기법이기 때문에 회귀 모델이 적절한 대푯값이 되려면 대상 데이터의 분포가 종 모양이어야 합니다. 그런데 회귀 모형은 그냥 전체 집단의 평균이 아니라 조건부 평균이기 때문에 분포를 확인할 때도 전체 분포가 아닌 조건별 분포를 확인해야 합니다.

이를테면 아래 그림 같은 걸 상상하시면 됩니다. 아래 그림에서 Y축에 있는 Distance 는 평균 측정이 되는 타겟 변수이고, X축에 있는 Speed는 Distance의 변화에 영향을 주는 요인 변수입니다. 따라서 회귀 분석을 통해 얻게 되는 회귀 모형이 의미하는 것은 Speed 에 따라 달라지는 Distance의 조건별 평균입니다. 때문에 이 평균이 적절한지 확인하려면 아래 그림처럼 같은 speed 조건을 갖는 distance 값들의 분포가 종형 분포인지를 봐야 하는 것이죠.

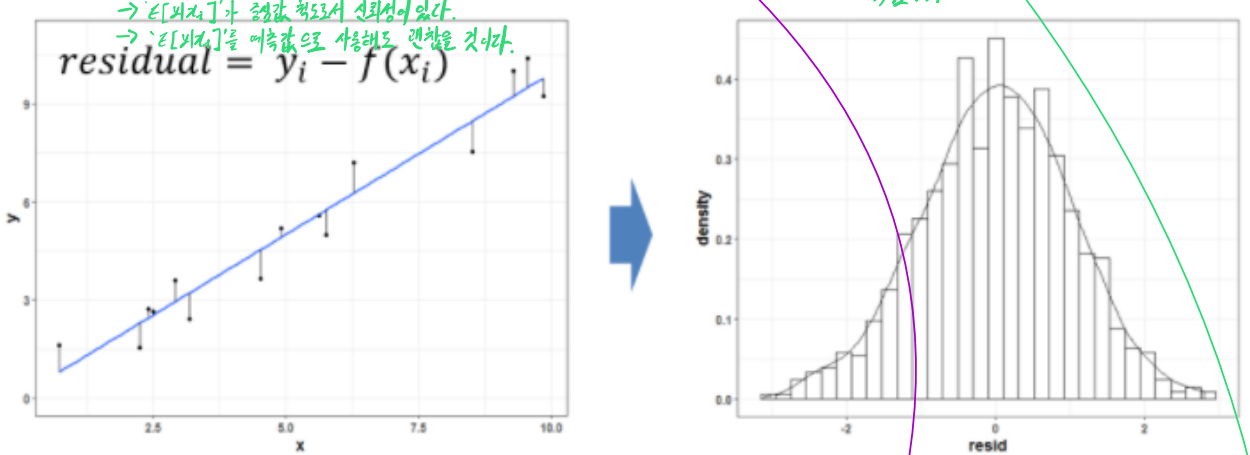


출처: https://blogs.sas.com/content/iml/2015/09/10/plot-distrib-reg-model.html

그렇다면 실제 조건별 데이터의 분포는 어떻게 확인할 수 있을까요? 언뜻 생각하기에는 각 조건에 따라 데이터를 분류해서 분포를 확인하면 될 것 같습니다. 그런데 그렇게 하면 1편에서 예시로 들었던 아파트 평균 가격을 표로 집계할 때와 동일한 문제가 발생합니다. 수천 개의 경우의 수 별로 분포를 그리고 확인해야 할텐데 여간 번거로운 일이 아니죠. 게다가 위 그림에 나온 예시대로라면 나올 수 있는 모든 조건에 비해 데이터 개수가 적어서 분포를 확인하기 힘든 경우가 생깁니다. 가령, 위 자료에서 speed가 15인 데이터는 불과 3개에 불과한데 이걸로 종형 분포인지를 확인하는 건 거의 불가능하겠죠.

다행히 좋은 방법이 있습니다. 아래 그림처럼 일단 먼저 회귀 모형을 만든 후 이 모형을 통해 계산한 평균값(아래 그림에서 f(x))과 실제 관측값 사이의 차이값에 대한 분포를 확인하는 것입니다. 이 차이값의 분포는 중심이 평균이냐 아니면 0이냐의 차이만 있을 뿐 원래 데이터의 조건별 분포와 형태가 비슷하기 때문입니다 (좀 더 엄밀히 얘기하자면, 조건별 분포들이 모두 '정규 분포'인 경우에 그렇습니다). 그러니 이 차이값의 분포가 종형 분포라면 우리가 만든 회귀 모형의 평균값들은 원래 데이터를 잘 대표할 것이라고 생각할 수 있습니다.

↑ 잔차의 분포형태가 정규분포 형태이다.
→ 또한 'y_i'의 분포형태가 정규분포 형태이다.
→ '$E[y_i]$'가 점추정값 역할로서 신뢰성이 있다.
→ '$E[y_i]$'를 예측값으로 사용해도 괜찮을 것이다.



↑ 해석!!!

X 잔차는 '찌꺼기'이다. 찌꺼기가 찌꺼기지 않고 영양가가 남아있으면, 모델에 해당 영양가가 반영되지 않은 것이다.

X 잔차의 분포가 정규분포 형태를 띠는 지에 대한 여부'도 회귀 모델 평가 지표 중 하나임.

ex). 여러 possible 회귀모델을 산출한 뒤, R-square가 가장 높은 모델을 선택함.
- 해당 모델의 잔차 분포는 정규분포 형태가 아님.
- 이때 설명력이 높은 모델(R-square가 가장 높은 모델)을 채택할지, 설명력이 조금 떨어지더라도 안정성이 높은 모델(잔차 분포가 정규분포 모형을 띠는 모델)을 채택할지 분석가가 결정해야 함.

참고로 통계학에서는 실측치와 회귀 모형의 예측치 (조건부 평균) 사이의 차이값을 '잔차 (residual)' 라고 부릅니다. 그리고 종형 분포라는 말 대신 '정규' 분포라는 말을 더 자주 사용합니다 (엄밀히 말하면 종형 분포가 꼭 정규 분포인 것은 아니지만 자세한 이론적인 배경은 복잡하니 일단은 그냥 넘어가겠습니다). 더 나아가 회귀 모형이 원본 데이터를 잘 대표하는 것을 보통 잘 '적합(fit)했다'라고 표현합니다.

따라서 지금까지 설명한 내용을 좀 더 통계학스럽게 표현하자면, 회귀 모형을 데이터에 잘 적합시키려면 잔차가 정규 분포이어야 합니다. 대개 통계학 교재에서는 '**정규성 가정**을 만족해야 한다' 라고 설명하기도 합니다.

이 말을 바꿔서 표현하면 '**잔차가 정규 분포가 되도록 회귀 모형을 만들어야 모형이 데이터에 잘 적합한 것이다**' 라고 얘기할수도 있습니다. 개인적으로는 이 표현이 좀 더 분석가의 능동적인 역할을 강조하는 것 같아 더 마음에 듭니다.

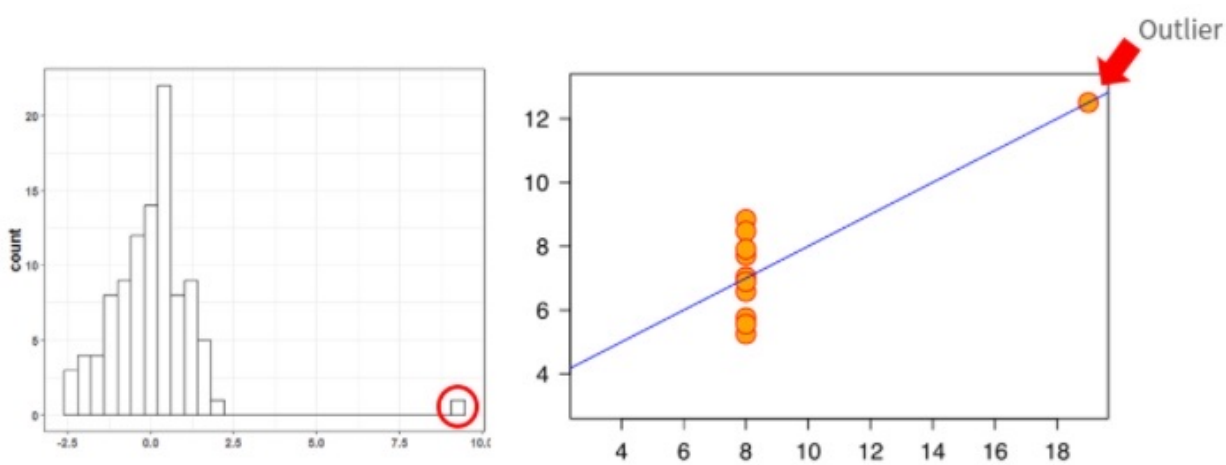
참고로 회귀 분석을 접하는 많은 사람들이 이 '정규성 가정'을 잘못 생각해서 '잔차'의 분포가 아니라 회귀 분석 대상이 되는 관측 데이터의 분포가 정규 분포가 되어야 한다라고 착각하곤 합니다 (솔직히 고백하자면 저 역시 한동안은 이 차이를 잘 몰랐습니다). 그래서 데이터를 분석할 때 전체 아파트 가격이나 연봉 데이터의 분포가 어떤 형태인지를 확인하는 것이죠. 하지만 위에 설명드렸듯이 이건 정규성 가정을 잘못 이해한 것이고, 실제로는 같은 조건 내에서의 데이터 분포 (이것을 조건부 확률 분포라고 부릅니다) 가 정규 분포가 되는지를 확인해야 합니다. 다만 이것을 직접적으로 확인하기가 불가능하기 때문에 잔차의 분포를 대신 확인하는 것이죠.

더 나아가 **조건부 평균이 적절한 값이 되려면 보유하고 있는 각 조건별 데이터들이 특정 조건에 편향되거나 서로 이질적인 특성을 갖고 있어서도 안됩니다.** 예를 들어, 아파트 가격 평균을 구하는데 면적이 작은 아파트들은 모두 강북에 있는 아파트들의 가격 정보만 사용하고 면적이 큰 아파트들은 강남에 있는 아파트들의 가격 정보만 사용한다고 해보죠. 혹은 층별 아파트 가격을 취합하는데 낮은 층 아파트는 모두 오래된 아파트의 가격 정보이고 높은 층은 모두 신축 아파트의 가격 정보라고 하면 어떨까요? 아마 이런 데이터들을 이용해서 조건별 평균을 구할 경우 실제 면적이나 층수에 따른 평균 가격의 변화가 정확히 측정되지 못하고 다른 조건에 의해 편향된 결과가 나올 겁니다.

또 다른 예로 강남에 있는 아파트를 조사할 때는 반포에 있는 특정 아파트 단지의 데이터만 취합하고, 강북에 있는 아파트를 조사할 때는 모든 지역의 아파트 가격을 조사한다면 또 어떨까요? 아마 강남 아파트에 비해 강북 아파트의 가격 편차가 훨씬 클 겁니다. 이런 경우 평균 가격을 구할 수 있다 하더라도 이렇게 구한 평균 가격 정보를 신뢰하기는 힘들겠죠.

즉, ① **신뢰할 수 있는 조건별 평균을 구하려면 데이터들이 특정 조건에 편향되지 않고 독립적으로 공평하게 수집되어야 하고,** ② **각 조건별로 취합된 데이터들의 편차(분산) 역시 서로 비슷해야 합니다.** 통계학에서는 전자를 ① '**잔차의 독립성**' 후자를 ② '**잔차의 등분산성**' 이라고 말합니다. 이 역시 신뢰할 수 있는 회귀 분석 결과를 얻기 위해 만족해야할 중요한 가정입니다. ↑ 설명변수에 대한 잔차의 산점도를 그렸을 때, '등분산성'이 존재해야 함.

마지막으로 평균을 구할 때 주의할 점 중 대표적인 것이 소위 '아웃라이어 (outlier)'라고 부르는 데이터로 인해 평균이 왜곡되는 현상입니다. 이를 테면 어떤 방에 모여있는 사람들의 평균 재산을 조사했더니 5억 쯤 되었는데 이건희씨가 방에 들어오니까 갑자기 평균 재산이 5000억이 되더라...뭐 이런 거죠. 회귀 분석에서도 마찬가지로 아웃라이어는 평균을 크게 왜곡시키기 때문에 이런 자료는 분석 대상에서 제외하는 것이 좋습니다. 우리가 원하는 것은 대다수의 데이터를 대표할 수 있는 평균값이기 때문에 이런 소수의 이질적인 데이터는 예외로 처리하는 것이죠.



다른 데이터와는 동떨어진 특성을 가진 극소수 데이터를 아웃라이어라고 합니다

지금까지 보셨듯이 회귀 분석을 할 때는 이렇게 여러 가지 고려해야할 사항들이 많습니다. 간혹 어떤 자료를 보면 잔차의 정규성, 독립성, 등분산성 같은 조건을 마치 수학적으로 자명한 규칙이라거나 뭔가 심오한 원리인냥 과도한 의미를 부여하는 경우가 있습니다. 하지만 이런 잔차에 대한 가정들을 만족하지 못한다고 해서 회귀 분석이 불가능한 것은 아닙니다. 근로자 연봉 사례에서도 보시다시피 전체 연봉 분포가 크게 평향되어 있더라도 평균 연봉을 구하는데 아무런 문제가 없습니다.

마찬가지로 위에서 소개한 가정들을 만족하지 못하더라도 회귀 분석을 할 수 있습니다. 다만 그렇게 해서 구한 회귀 모델의 신뢰도가 다소 떨어질 뿐이죠. 때문에 **내가 분석한 결과가 얼마나 신뢰할만한 수준인지를 파악하기 위해 위 가정을 잘 만족하는지 확인하는 과정이 필요한 것일 뿐입니다 (이런 확인 과정을 통계학에서는 모델 검정이라고 부릅니다).**

그런데 실제 데이터 분석을 해보면 위 가정을 만족시키기가 대단히 어렵습니다. 좀 더 노골적으로 얘기하면 **실전에서 회귀 분석을 위해 요구하는 가정을 모두 만족하는 회귀 모델은 거의 없습니다.**

여러 가지 이유가 있겠지만 가장 큰 이유는, **현실에서는 확보할 수 있는 데이터에 한계가 있기 때문이죠.** 이건 단순히 데이터의 크기를 말하는 것이 아니라 조건부 평균을 구할 때 사용될 여러 가지 조건 정보를 의미합니다 (이렇게 평균에 영향을 주는 조건들을 독립 변수 혹은 피처라고 부릅니다).

앞서 예로 들었던 아파트 평균 가격을 구하는 문제를 다시 생각해 보면, 가격에 영향을 주는 요인들이 대단히 많기 때문에 이런 요인별 거래 정보를 모두 충분히 확보하는 건 쉽지 않습니다. 가령, 아파트의 면적, 층, 지역 정보 등등의 기본 정보는 알고 있는데 주변 상가나 학군 정보를 모르고 있는 상태라고 가정해보죠. 그러면 분명 동일한 조건의 아파트임에도 불구하고 유명 학원가나 백화점 등의 편의 시설이 주변에 있는 아파트들과 그렇지 않은 아파트들 사이에는 가격 차이가 크게 날 것입니다. 그리고 이로 인해 잔차의 분포가 종 모양으로 깔끔하게 나오지 않고 들쑥날쑥한 모양이 될 가능성이 높습니다. 결국 아무리 회귀 모델을 잘 만들고 싶어도 이렇게 보유한 정보량의 한계로 인해 이론적으로 이상적인 모델이 만들어질 수 없는 것이죠.

이런 경우에는 어쩔 수 없이 지금 보유한 정보 내에서 최선의 결과를 만들 수 밖에 없습니다. 다시 말해, 이론적으로는 정규성, 독립성, 등분산성을 만족해야 하지만 현실에서는 이걸 완벽히 만족하기는 어려우니 적당한 선에서 타협하는 것이죠. 대신 적어도 분석 결과에 어떤 한계점이 있는지는 최대한 정확하게 파악해야 합니다. 또한 이런 한계를 극복하거나 조금이나마 해소하기 위한 여러 가지 기법들도 있습니다. 이런 방법들은 서로 장단점이 있고 대개는 특정 상황에서 적용할 수 있습니다 (현재의 상황에 맞는 적절한 방법을 찾아 사용하는 것은 데이터 분석가에게 필요한 중요한 역량중 하나입니다).

최악의 경우에는 회귀 분석을 하지 말아야 합니다. 즉, 데이터를 확인해 봤을 때 너무 정보가 부족하거나 편향되어 있다고 생각이 들 경우에는 과감하게 회귀 분석을 포기해야 합니다. 자칫 잘못된 정보로 인해 더 큰 혼란을 주거나 혹은 잘못된 의사 결정을 함으로써 오히려 손해를 볼 수도 있기 때문입니다. 이런 판단 역시 데이터 분석가가 해야할 일 중 하나입니다.