1. 회귀분석이란

$y_i = \alpha + \beta x_i + \epsilon_i$ 여기서 오차항은 평균이 0이고 분산이 σ^2으로 동일합니다.

직관적으로 지금 무엇을 하려고 하는지를 먼저 생각을 합니다.
비료를 주면 어떤 때는 수확량이 많기도 하고 어떤 때는 작기도 한데 평균적인 수확량을 기준으로 많게 수확될 때도 있고 작게 수확될 때도 있는 이유가 저 위의 모형이 설명을 해주는 것입니다.

그럼 이제 그림을 생각합니다. 수업 자료에 있는 파란선 기억나지요?
우리는 비료양에 따른 수확량의 확률분포를 가지고 있고 그 확률분포의 평균이 지나는 선이 $\alpha + \beta x_i$인 것입니다. 우리가 실제로 얻는 데이터는 그 확률분포상의 한 점들이고 그 점들을 가지고 평균선을 추정하자는 것입니다.
어떻게 추정을 할 것인가?
거리를 최소화하는 방법을 쓰자. (최소제곱방법)-이것만 있는가? 여러 가지 다른 방법이 많다. 그래서

$Q(\alpha,\beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$을 최소화하려고 합니다.

어떻게?
α,β 각각에 대해서 편미분을 하여 0이라고 두면 풀면 된다. (α에 대해서 미분할 때 β는 상수라고 생각, β에 대해서 미분할 때 α는 상수라고 생각) --꼭 직접 해봅니다.
s_{xx}, s_{xy}, s_{yy} 등을 쉽게 구할 수 있도록 기출문제를 보면서 연습을 하기 바랍니다.

또한 $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$, $\hat{\beta} = \frac{s_{xy}}{s_{xx}}$를 이용해서 자료를 받으면 회귀계수를 구할 수 있도록 연습합니다.

SST(n-1) =SSR(1) + SSE (n-2) 이 되는 변동성의 합들을 생각해봅니다.
이 식들에 의해 카이제곱 분포가 유도될 것이고 카이제곱 분포의 독립성을 이용해서 f분포를 따르게 되어 검정을 할 수 있게 되는 것입니다.

전체의 변동을 회귀식에 의한 변동과 회귀식이 설명하지 못하는 알 수 없는 오차항의 변동으로 분해를 할 수 있다.

이제 r^2라는 결정계수를 정의합니다.
전체변동에서 회귀변동이 차지하는 부분이 큰지를 보는 것으로 회귀변동이 차지하는 변동이 클수록 회귀선에 의해서 설명을 많이 한다는 이야기가 됩니다. 이 값이 0.9가 나오면 전체 데이터의 90%정도를 회귀선에 의해 설명할 수 있다라는 이야기가 됩니다.
데이터가 직선 주위에 다 놓여있다면 1이 되면 회귀선만 있으면 전체 데이터는 완벽히 설명이 되는 것이지요.

s_{xx},s_{xy},s_{yy}만 구할 수 있으면 SST,SSR,SSE, 결정계수를 다 구할 수 있으니 구하는 연습을 합니다.
----------여기까지 식에 두려워말고 논리적으로 조목 조목 짚어가면서 써보면 충분히 따라올 수 있으리라 믿습니다. ----------

이제 추론을 하러 갑니다.
이제 오차항이 평균 0이고 분산이 동일하고 정규분포를 따른다고 가정합니다.
그러면 Y_i는 평균은 $\alpha+\beta x_i$지만 분산은 오차의 분산인 σ^2이고 역시 오차항에 의해 정규분포를 따르게 됩니다.
위의 합에 대한 분해를 생각하면 합에 대한 분해에는 Y에 대한 식들이고 이 식을 카이제곱분포를 따르도록 만들 수 있습니다.
SST(n-1) =SSR(1) + SSE (n-2)
단!!! 귀무가설인 $\beta=0$이 사실일 때 독립인 카이제곱 분포가 되어 나누면 F분포가 만들어지게 됩니다. $SSR/\sigma^2 \sim X^2(1)$, $SSE/\sigma^2 \sim X^2(n-2)$ 이고 각각은 독립!!!
동일한 분산인 σ^2은 두 개의 카이제곱을 나누면서 없어질 것이고 각 변동합을 자유도로 나눈 것의 비만 남게 되어 이 값이 F 분포를 따르게 됩니다.
즉 F=SSR/1/SSE/(n-2)
이 값은 귀무가설이 참일 때의 분포이고 SSE보다 SSR이 크게 나오면 오차항에 의한 것보다 회귀선에 의해서 더 많이 설명이 된다는 것이므로 회귀계수는 의미가 있게 됩니다. 따라서 F 분포의 윗꼬리 검정에 의해 충분히 큰 증거를 제시한다면 귀무가설인 $\beta=0$을 기각하고 회귀모형은 의미가 있다라는 결론을 얻게 됩니다.
여기까지가 분산분석표라고 주어져 있는 부분에 대한 설명입니다.
그러니 자료를 받으면 회귀계수를 구하고 분산분석표를 완성할 수 있도록 준비하는 것.
충분히 이 글을 보면서 논리적으로 공부할 수 있으니라 생각합니다.

자, 지금까지 우리는 점추정을 했을 뿐이지요. 그러니 이 추정량이 불편성을 만족하는지 이 추정량의 분포는 무엇인지를 알아야 하겠습니다.
그래서 각 추정량에 대한 평균 및 분산을 구합니다.
$\hat{\alpha}$는 뒤에서 나오는 $\hat{\alpha}+\hat{\beta}x$ 에서 x=0인 경우이므로 따로 해볼 필요 없이 $\hat{\beta}$에 대해서 먼저 불편성과 분산 및 분포를 생각합니다.
$E(\hat{\beta})=\beta$, $var(\hat{\beta})=\dfrac{\sigma^2}{s_{xx}}$ 이 된다는 것을 직접 식으로 보여봅니다.
그리고 y의 일차결합으로 $\hat{\beta}$식이 표현되므로 이 또한 정규분포를 따르게 됩니다.
그러나 오차항의 분산을 모르니 추정값을 넣어주게 되고 그래서 t분포를 따르게 됩니다.
오차항의 분산 추정치$\hat{\sigma}= MSE= SSE/(n-2)$ 가 됩니다.

여러분은 자료를 얻으면 회귀계수와 오차항의 추정치를 구할 수 있어야 합니다.

이것을 통해 $\frac{\frac{\hat{\beta}-\beta}{\hat{\sigma}}}{\sqrt{s_{xx}}} \sim t(n-2)$ 를 따르게 됩니다. $SSE/\sigma^2 \sim X^2(n-2)$라고 했으니 표준정규분포를 루트(카이제곱/자유도)으로 나누어 T분포가 된다는 원리에 의해서도 이 분포가 얻어진다는 것을 아시겠지요?

(해보기 바랍니다. $\frac{\frac{\hat{\beta}-\beta}{\sigma}}{\sqrt{s_{xx}}} / \sqrt{(SSE/\sigma^2)/(n-2)} \sim t(n-2)$)

분포가 나왔으니 신뢰구간도 구하고 가설검정도 할 수 있겠지요?

이렇게 해서 β=0인지를 검정한 것이나 F를 이용한 것이나 결과는 동일합니다. $T^2 = F$.

이제 평균 반응값 $\hat{\alpha}+\hat{\beta}x$에 대한 불편성, 분산 및 분포를 생각합니다. 역시 신뢰구간과 가설검정을 할 수 있겠지요? 그래서 내가 추정한 회귀선에 x=10을 넣었을 때 이 값을 근거로 하여 진짜 값 $\alpha+10\beta$ =130인가를 검정하게 되는 것입니다. 또한 $\alpha+10\beta$라는 실제 평균에 대한 신뢰구간을 내가 추정해서 얻은 식 $\hat{\alpha}+10\hat{\beta}$ 값을 이용해서 구하게 되는 것입니다.

이제 추정과 검정을 다 배웠습니다.

회귀분석은 오차항이 평균이 0이고 등분산이라는 가정을 가지고 있으므로 잔차에 의해 그 조건을 만족하는지 확인합니다. 또한 추론을 위해 정규분포를 가정하였으므로 정규분포를 따르는지 확인하기 위하여 잔차를 표준화 시킨 후 관측값이 +-2 이내에서 95%가 얻어지는지 확인을 합니다. 여기서 가정을 만족하지 않게 되면 회귀분석은 의미가 없어지고 변수에 대한 변환 등을 하여 가정을 만족하도록 한 후에 실시해야 합니다.

여러분이 읽은 논문들이 잔차에 대한 언급을 하지 않고 있다면 의심이 가는 결과일 수 있습니다.

중회귀분석은 독립변수의 수가 늘어나서 행렬로 표시하여 추정량을 구하게 되고 SST(n-1)=SSR(k)+SSE(n-k-1)로 자유도가 정해지는 것만 빼고는 동일합니다.

해석은 각 독립변수 앞에 붙어 있는 기울기를 해석할 때 나머지 독립변수를 고정한 상태에서 한 변수가 한단위 변화할 때 y값의 변화량으로 원래 여러분이 해석하는 기울기의 해석을 하면 되겠습니다.

MSE의 불편성 증명

$y_i - \bar{y} = \alpha + \beta x_i + e_i - \alpha - \beta\bar{x} - \bar{e} = \beta(x_i - \bar{x}) + (e_i - \bar{e})$

$Var(y_i - \bar{y}) = var(e_i - \bar{e})$ 이고 e_i는 평균이 0이고 분산이 σ^2이고 독립이고 같은 분포를

갖는 확률변수이다.

우리가 표본분산의 불편성 증명에서 $E(\sum_{i=1}^{n}(x_i-\bar{x})^2)=\sum_{i=1}^{n}E(x_i-\bar{x})^2=(n-1)\sigma^2$ 였으므로 $E(x_i-\bar{x})^2=\frac{(n-1)}{n}\sigma^2$일 것이다. 이것은 평균이 μ이고 분산이 σ^2인 random sample에 대해서 성립하므로 e_i에 대해서도 같은 원리로 성립하게 될 것이다. 따라서 $E(e_i-\bar{e})^2=\frac{(n-1)}{n}\sigma^2$.

$Var(y_i-\bar{y})=var(e_i-\bar{e})=E(e_{i_}\bar{e})^2-[E(e_i-\bar{e})]^2$이고 $E(e_i-\bar{e})=Ee_i-E\bar{e}=0$이므로 $Var(y_i-\bar{y})=var(e_i-\bar{e})=\frac{(n-1)\sigma^2}{n}$이다.

SSE=SST-SSR=$\sum_{i=1}^{n}(y_i-\hat{y_i})^2-\frac{(S_{xy})^2}{S_{xx}}$

E(SSE)=E(SST)-E(SSR)

E(SST)=$E[\sum_{i=1}^{n}(y_i-\bar{y})^2]=\sum_{i=1}^{n}E(y_i-\bar{y})^2$

$E(y_i-\bar{y})^2=Var(y_i-\bar{y})+[E(y_i-\bar{y})]^2=\frac{(n-1)\sigma^2}{n}+(\alpha+\beta x_i-\alpha-\beta\bar{x})^2$

$E(SSR)=E(\frac{(S_{xy})^2}{S_{xx}})$

$E(S_{xy}^2)=Var(S_{xy})+(ES_{xy})^2$

$var(S_{xy})=var[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})]=var[\sum_{i=1}^{n}(x_i-\bar{x})y_i]=\sum_{i=1}^{n}(x_i-\bar{x})^2var(y_i)=S_{xx}\sigma^2$

$E(S_{xy})=E[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})]=\sum_{i=1}^{n}(x_i-\bar{x})E(y_i)=\alpha\sum_{i=1}^{n}(x_i-\bar{x})+\beta\sum_{i=1}^{n}(x_i-\bar{x})x_i=\beta S_{xx}$

$E(SST)=\sum_{i=1}^{n}E(y_i-\bar{y})^2=n\times\frac{(n-1)\sigma^2}{n}+\beta^2\sum_{i=1}^{n}(x-\bar{x})^2=(n-1)\sigma^2+\beta^2S_{xx}$

$E(SSR)=E(\frac{(S_{xy})^2}{S_{xx}})=\sigma^2+\beta^2S_{xx}$

$E(SSE)=E(SST)-E(SSR)=(n-2)\sigma^2$

$E(MSE)=E(\frac{SSE}{n-2})=\sigma^2$