Statistical Inference

RANDOM SAMPLE

Definition: Random Sample

$X_1, \ldots, X_n$ is called a random sample from a distribution with pdf $f(x)$ (or pf $P(x)$) if $X_1, \ldots, X_n$ are independent and have identical distribution with pdf $f(x)$ (or pf $P(x)$). It is often denoted "iid" (independent-identically-distributed).

Note

Let $X_1, \ldots, X_n$ be a sample from a distribution with pdf $f(x)$. The joint pdf of $X = (X_1, \ldots, X_n)$ is $f(x_1, \ldots x_n) = f_1(x_1) \cdots f_n(x_n) = f(x_1) \cdots f(x_n)$.

SAMPLE MEAN

Definition: Sample Mean and Sample Variance

Let $X_1, \ldots, X_n$ be a random sample. The sample mean is defined by $\overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + \cdots + X_n}{n}$, and the sample variance is defined by $S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}_n \right)}{n-1}$.

Theorem

If $X_1, \ldots, X_n$ are iid each with a $N(\mu, \sigma^2)$ distribution, then $k_1 X_1 + \cdots + k_n X_n$ has a normal distribution with mean $k_1 \mu + \cdots + k_n \mu = (k_1 + \cdots + k_n)\mu$ and variance $\left(k_1{}^2 + \cdots + k_n{}^2\right)\sigma^2$.
More generally, if $X_1, \ldots, X_n$ are independent and each X_i has a $N\left(\mu_i, \sigma_i{}^2\right)$, then $k_1 X_1 + \cdots + k_n X_n$ has a normal distribution mean $k_1 \mu_1 + \cdots + k_n \mu_n$ and variance $k_1{}^2 \sigma_1{}^2 + \cdots + k_n{}^2 \sigma_n{}^2$.

USEFUL DISTRIBUTION

Theorem

Suppose $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$ distribution. Then $\overline{X}_n$ and $\sum_{i=1}^{n} (X_i - X_n)^2$ are independent, and $\frac{\sum_{i=1}^{n} (X_i - X_n)^2}{\sigma^2}$ has a χ^2_{n-1} distribution.

The t and F Distribution

Two important distributions useful in statistical inference are:

1) If $W \sim N(0,1)$ and $V \sim \chi_r^2$ are independent, then $\dfrac{W}{\sqrt{V/r}}$ has a t-distribution.
2) $U \sim \chi_{r_1}^2$ and $V \sim \chi_{r_2}^2$ are independent, then $\dfrac{U/r_1}{V/r_2}$ has a F-distribution.

THE CENTRAL LIMIT THEOREM

Let $X_1, \ldots, X_n$ is a random sample with finite mean μ and variance $\sigma^2 > 0$. Let $S_n = X_1 + \cdots + X_n$. Then $\dfrac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} = \dfrac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{\mathrm{var}(\overline{X}_n)}} = \dfrac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \xrightarrow[n\to\infty]{} N(0,1)$.

STATISTICAL MODEL

Consider a random sample $X_1, \ldots, X_n$ from a distribution with pdf $f_\theta(x)$. The family $\{f_\theta(x) \mid \theta \in \Omega\}$ (where $f_\theta(x)$ is pdf (or $P_\theta(x)$ a pf), θ is an unknown parameter, Ω is a parameter space) is a statistical model.
We know that the distribution under investigation is in the family, but don't know which one. Based on the sample values $x_1, \ldots, x_n$, we find an estimate for θ; Once we find θ, we know the distribution.

LIKELIHOOD FUNCTION

Definition: The Likelihood Function

Let $X_1, \ldots, X_n$ be a random sample from a distribution with pdf $f_\theta(x)$ (or pf $P_\theta(x)$). The likelihood is defined by $L : \Omega \to \mathbf{R}$ given by $L(\theta \mid x_1 \ldots, x_n) = cf_\theta(x_1, \ldots, x_n) = cf_\theta(x_1) \cdots f_\theta(x_n)$ where $c > 0$ (or $L(\theta \mid x_1 \ldots, x_n) = cP_\theta(x_1, \ldots, x_n) = cP_\theta(x_1) \cdots P_\theta(x_n)$).

Definition: The Maximum Likelihood Estimate (MLE)

The function $\hat{\theta} : S \to \Omega$ is called the maximum likelihood estimator.
$\hat{\theta}(s)$ is called the maximum likelihood estimate of θ if for each $\theta \in \Omega$, $L(\hat{\theta}(s) \mid x_1, \ldots, x_n) \ge L(\theta \mid x_1, \ldots, x_n)$.

The Algorithm

This suggests that in order to obtain the MLE of θ, we maximum the likelihood function. Since a version of the likelihood version with $c = 1$ gives the same maximum value, we use this version. In most cases, this is done by differentiation.

1) Write the likelihood function $L(\theta \mid x_1 \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i)$.

2) Write the log likelihood function defined by $l(\theta \mid x_1 \ldots, x_n) = \ln(L(\theta \mid x_1 \ldots, x_n)) = \sum_{i=1}^{n} \ln f_\theta(x_i)$.
3) Write the score function $S(\theta \mid x_1 \ldots, x_n) = \dfrac{\partial l(\theta \mid x_1 \ldots, x_n)}{\partial \theta}$.
4) Write the score equation $S(\theta \mid x_1 \ldots, x_n) = 0$ and solve for θ.
5) Check that the solution is the global maximum. If it is, then it is the MLE of θ.

Theorem

If $\hat{\theta}(x_1 \ldots, x_n)$ is the MLE in Ω, and $\phi : \Omega \xrightarrow{\text{1-1}} \Omega'$, then the MLE in the new parameterization is $\phi(\hat{\theta}(x_1 \ldots, x_n)) = \hat{\theta}'(x_1 \ldots, x_n)$.

<u>Proof</u>:

$$\begin{aligned} L^*(\hat{\theta}'(x_1, \ldots, x_n) \mid x_1, \ldots, x_n) &= g_{\hat{\theta}'(x_1, \ldots, x_n)}(x_1, \ldots, x_n) = g_{\phi(\hat{\theta}(x_1, \ldots, x_n))}(x_1, \ldots, x_n) \\ &= f_{\hat{\theta}(x_1, \ldots, x_n)}(x_1, \ldots, x_n) = L(\hat{\theta}(x_1, \ldots, x_n) \mid x_1, \ldots, x_n) \\ &\geq L(\theta \mid x_1, \ldots, x_n) = f_\theta(x_1, \ldots, x_n) = g_{\theta'}(x_1, \ldots, x_n) = L^*(\theta' \mid x_1, \ldots, x_n) \end{aligned}$$

Hence for every $\theta' \in \Omega'$, $L^*(\hat{\theta}'(x_1, \ldots, x_n) \mid x_1, \ldots, x_n) \geq L^*(\theta' \mid x_1, \ldots, x_n)$, and so $\hat{\theta}'(x_1, \ldots, x_n)$ is the MLE of the new parameterization.

The Algorithm: The Multidimensional Case

In the multidimensional case, the parameter space is $\Omega = \{(\theta_1, \ldots, \theta_k), k > 1\}$.

1) Write the likelihood function $L((\theta_1, \ldots, \theta_k) \mid x_1 \ldots, x_n)$.
2) Write the log likelihood function defined by $l((\theta_1, \ldots, \theta_k) \mid x_1 \ldots, x_n) = \ln(L((\theta_1, \ldots, \theta_k) \mid x_1 \ldots, x_n))$.
3) Write the score function $S((\theta_1, \ldots, \theta_k) \mid x_1 \ldots, x_n) = \begin{pmatrix} \dfrac{\partial l((\theta_1, \ldots, \theta_k) \mid x_1 \ldots, x_n)}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial l((\theta_1, \ldots, \theta_k) \mid x_1 \ldots, x_n)}{\partial \theta_k} \end{pmatrix}$.
4) Write the score equation $S((\theta_1, \ldots, \theta_k) \mid x_1 \ldots, x_n) = 0 \Rightarrow \begin{pmatrix} \dfrac{\partial l((\theta_1, \ldots, \theta_k) \mid x_1 \ldots, x_n)}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial l((\theta_1, \ldots, \theta_k) \mid x_1 \ldots, x_n)}{\partial \theta_k} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$ and solve.
5) Check that the solutions are the global maximum (the matrix of the second partial derivatives evaluated at $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ must be negative definite, or equivalently, all eigenvalues negative).

Standard Error and Bias

Suppose $\hat{\theta}$ is the MLE; $\hat{\phi}(x_1, \ldots, x_n) = \phi(\hat{\theta}(x_1, \ldots, x_n))$ is the estimate of $\phi(\theta)$. How reliable are the estimates? One measure of accuracy commonly used is MSE (mean squared error).

Definition: Mean Squared Error

The mean squared error is defined as $\mathrm{MSE}_\theta\left(\hat{\phi}\right)=E_\theta\left[\left(\hat{\phi}-\phi(\theta)\right)^2\right]$ for each $\theta\in\Omega$.

Theorem

If $\phi(\theta)\in\mathbf{R}$ and $T:S\to\mathbf{R}$ such that $E_\theta(T)$ exists, then $\mathrm{MSE}_\theta\left(\hat{\phi}\right)=\mathrm{var}_\theta(T)+\left(E_\theta(T)-\phi(\theta)\right)^2$.

Note: When $E_\theta(T)-\phi(\theta)=0\Leftrightarrow E_\theta(T)=\phi(\theta)$, then $\mathrm{MSE}_\theta\left(\hat{\phi}\right)=\mathrm{var}_\theta(T)$ (unbiased).

Definition: Bias

$E_\theta(T)-\phi(\theta)$ is called the bias in the estimate.

Definition: Standard Error

$\mathrm{STD}_\theta(T)=\sqrt{\mathrm{var}_\theta(T)}$ is called the standard error of the estimate.

SUFFICIENCY

The likelihood function for a model and data shows how the data supports the various possible values of the parameters. It is not the actual likelihood but the ratios of the likelihood at different values that are important.

Definition: Statistic

A statistic is a function of one or more random variables that does not depend on the unknown parameters.

Example

$\overline{X}$ is a statistic, but $\dfrac{\overline{X}-\mu}{\sigma}$ is not unless μ and σ are known.

Note

Although a statistic does not depend on unknown parameters, its distribution may.

Definition: Sufficient Statistic

If the statistic T is such that $T(x_1,\ldots,x_n)=T(y_1,\ldots,y_n)\Rightarrow L(\theta\mid x_1,\ldots,x_n)=cL(\theta\mid y_1,\ldots,y_n)$ for some $c>0$ that may depend on $(x_1,\ldots,x_n)$ and $(y_1,\ldots,y_n)$, then T is called a sufficient statistics for the model.

Example

Suppose that $S=\{1,2,3,4\}$, $\Omega=\{\theta_1,\theta_2\}$, and the probability distribution is given by $f_{\theta_1}(x)=\begin{cases}\frac{1}{2}, & x=1\\ \frac{1}{6}, & x=2,3,4\end{cases}$

and $f_{\theta_2}(x)=\frac{1}{4}, x=1,2,3,4$.

If we define T by $T(x)=\begin{cases}0, & x=1\\ 1, & x=2,3,4\end{cases}$, then T is a sufficient statistic. Note that T has only 2 values compared to 4 values in the original model.

Theorem: Factorization Theorem

If the density (or probability) function for a model factors as $f_\theta(x_1,\ldots,x_n)=h(x_1,\ldots,x_n)g_\theta(T(x_1,\ldots,x_n))$ where h and g_θ are non-negative, then T is a sufficient statistic.

Proof: Suppose $T(x_1,\ldots,x_n)=T(y_1,\ldots,y_n)$, then

$$\begin{aligned}L(\theta \mid x_1,\ldots,x_n)&=cf_\theta(x_1,\ldots,x_n)=ch(x_1,\ldots,x_n)g_\theta(T(x_1,\ldots,x_n))\\&=c\frac{h(x_1,\ldots,x_n)g_\theta(T(x_1,\ldots,x_n))}{h(y_1,\ldots,y_n)g_\theta(T(y_1,\ldots,y_n))}h(y_1,\ldots,y_n)g_\theta(T(y_1,\ldots,y_n))\\&=\left(c\frac{h(x_1,\ldots,x_n)}{h(y_1,\ldots,y_n)}\right)h(y_1,\ldots,y_n)g_\theta(T(y_1,\ldots,y_n)),\because T(x_1,\ldots,x_n)=T(y_1,\ldots,y_n)\\&=c'f_\theta(y_1,\ldots,y_n)=L(\theta \mid y_1,\ldots,y_n)\end{aligned}.$$

Hence T is a sufficient statistic.

Definition: Minimal Sufficient Statistic

A minimal sufficient statistic T for a model is any sufficient statistic such that once we have the likelihood function $L(\theta \mid x_1,\ldots,x_n)$ for the model, we can determine $T(x_1,\ldots,x_n)$.

Example

Let $X_1,\ldots X_n$ be a random sample from $N(\mu,\sigma_0^2)$ (σ_0^2 is known). The likelihood function is $L(\mu \mid x_1,\ldots,x_n)=\left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^n e^{-\frac{n}{2\sigma_0^2}(\bar{x}-\mu)^2}e^{-\frac{n-1}{2\sigma_0^2}s^2}$. By the Factorization Theorem, $T(x_1,\ldots,x_n)=\bar{X}$ is a sufficient statistic because any likelihood function is a positive multiple of $e^{-\frac{n}{2\sigma_0^2}(\bar{x}-\mu)^2}$ which is completely determined by its maximum value $\bar{x}$. Hence $\bar{X}$ is a minimal sufficient statistic determined from the likelihood function of the model.

CONFIDENCE INTERVAL

Definition: Quartile

For $p\in[0,1]$, the p^{th} quartile x_p for the distribution with cdf F is the smallest number x_p such that $p\le F(x_p)$. When F is strictly increasing and continuous, then $F^{-1}(p)$ is the unique number x_p such that $p=F(x_p)$ or $x_p=F^{-1}(p)$.

Particular Cases:

1) $F^{-1}(0.5)=x_{0.5}$ is called the median.
2) $F^{-1}(0.25)=x_{0.25}$ is called the first quartile.
3) $F^{-1}(0.75)=x_{0.75}$ is called the third quartile.

Definition: Confidence Interval
An interval $c(x_1,\dots,x_n)=(l(x_1,\dots,x_n),u(x_1,\dots,x_n))$ is an α-confidence-interval for $\phi(\theta)$ if $P_\theta(\phi(\theta)\in c(x_1,\dots,x_n))=P_\theta(l\le\phi(\theta)\le u)\ge\alpha$ for every $\theta\in\Omega$. α is referred to as the confidence level of the interval.

Confidence Interval for μ

1) Let $X_1,\dots X_n$ be a random sample from $N(\mu,\sigma_0^2)$ (σ_0^2 known). Then $\frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}}$ has a $N(0,1)$ distribution. A $100\alpha\%$ confidence interval for μ is $\left(\bar{x}-c\dfrac{\sigma_0}{\sqrt{n}},\bar{x}+c\dfrac{\sigma_0}{\sqrt{n}}\right)$ (it includes the parameter μ with probability α); here $\Phi(c)=\dfrac{\alpha+1}{2}$ or $c=\Phi^{-1}\left(\dfrac{\alpha+1}{2}\right)$, where Φ is the cdf of $N(0,1)$.

2) Let $X_1,\dots X_n$ be a random sample from a distribution (not normal) with finite mean μ and finite variance σ_0^2 (known). By CLT, $\frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}}$ has a limiting $N(0,1)$ distribution. An approximate $100\alpha\%$ confidence interval for μ is $\left(\bar{x}-c\dfrac{\sigma_0}{\sqrt{n}},\bar{x}+c\dfrac{\sigma_0}{\sqrt{n}}\right)$; here $\Phi(c)=\dfrac{\alpha+1}{2}$ or $c=\Phi^{-1}\left(\dfrac{\alpha+1}{2}\right)$, where Φ is the cdf of $N(0,1)$.

3) Let $X_1,\dots X_n$ be a random sample from $N(\mu,\sigma^2)$ (σ^2 unknown). Then $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a t_{n-1} distribution. A $100\alpha\%$ confidence interval for μ is $\left(\bar{x}-c\dfrac{S}{\sqrt{n}},\bar{x}+c\dfrac{S}{\sqrt{n}}\right)$; here $G(c)=\dfrac{\alpha+1}{2}$ or $c=G^{-1}\left(\dfrac{\alpha+1}{2}\right)$, where G is the cdf of t_{n-1}.

4) Let $X_1,\dots X_n$ be a random sample from sample from a distribution (not normal) with finite mean μ and finite variance σ^2 (unknown). By CLT, $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a limiting $N(0,1)$ distribution. An approximate $100\alpha\%$ confidence interval for μ is $\left(\bar{x}-c\dfrac{\sigma_0}{\sqrt{n}},\bar{x}+c\dfrac{\sigma_0}{\sqrt{n}}\right)$; here $\Phi(c)=\dfrac{\alpha+1}{2}$ or $c=\Phi^{-1}\left(\dfrac{\alpha+1}{2}\right)$, where Φ is the cdf of $N(0,1)$.

Note that $\left|\frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}}\right|\le c\Leftrightarrow -c\le\frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}}\le c\Leftrightarrow \bar{X}-c\dfrac{\sigma_0}{\sqrt{n}}\le\mu\le\bar{X}+c\dfrac{\sigma_0}{\sqrt{n}}$, so

$$P\left(\left|\frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}}\right|\le c\right)=\alpha\Leftrightarrow P\left(\bar{X}-c\frac{\sigma_0}{\sqrt{n}}\le\mu\le\bar{X}+c\frac{\sigma_0}{\sqrt{n}}\right)=\alpha .$$

If $Z=\frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}}\sim N(0,1)$, then $P(|Z|\le c)=P(-c\le Z\le c)=\Phi(c)-\Phi(-c)=\Phi(c)-(1-\Phi(c))=2\Phi(c)-1$, so

$$P(|Z|\le c)=\alpha\Leftrightarrow 2\Phi(c)-1=\alpha\Leftrightarrow\Phi(c)=\frac{\alpha+1}{2} .$$

Confidence Interval for *p*

Let $Y \sim \text{binomial}(n, p)$ (p unknown). By CLT, $\dfrac{Y-np}{\sqrt{Y\left(\frac{Y}{n}\right)\left(1-\frac{Y}{n}\right)}}$ has a limiting $N(0,1)$ distribution. An approximate $100\alpha\%$ confidence interval for p is $\left(\dfrac{y}{n}-c\sqrt{\dfrac{\frac{Y}{n}\left(1-\frac{Y}{n}\right)}{n}}, \dfrac{y}{n}+c\sqrt{\dfrac{\frac{Y}{n}\left(1-\frac{Y}{n}\right)}{n}}\right)$; here $\Phi(c)=\dfrac{\alpha+1}{2}$ or $c=\Phi^{-1}\left(\dfrac{\alpha+1}{2}\right)$, where Φ is the cdf of $N(0,1)$.

Confidence Interval for σ^2

Let $X_1,\dots X_n$ be a random sample from $N(\mu_0,\sigma^2)$ (μ_0 known, σ^2 unknown). Then $\sum_{i=1}^{n}\left(\dfrac{X_i-\mu}{\sigma}\right)^2$ has a χ_n^2 distribution. A $100\alpha\%$ confidence interval for σ^2 is $\left(\sum_{i=1}^{n}\dfrac{X_i-\mu}{b}, \sum_{i=1}^{n}\dfrac{X_i-\mu}{a}\right)$; here $P(X<a)=\dfrac{1-\alpha}{2}$ and $P(X>b)=\dfrac{1+\alpha}{2}$, where $X \sim \chi_n^2$.

Let $X_1,\dots X_n$ be a random sample from $N(\mu_0,\sigma^2)$ (μ unknown, σ^2 unknown). Then $\dfrac{(n-1)S^2}{\sigma^2}$ has a χ_{n-1}^2 distribution. A $100\alpha\%$ confidence interval for σ^2 is $\left(\dfrac{(n-1)S^2}{b}, \dfrac{(n-1)S^2}{a}\right)$; here $P(X<a)=\dfrac{1-\alpha}{2}$ and $P(X>b)=\dfrac{1+\alpha}{2}$, where $X \sim \chi_{n-1}^2$.

Testing Statistical Hypotheses

Definition: Statistical Hypothesis

A statistical hypothesis is an assertion about the distribution of one or more random variable(s). If the hypothesis completely determines the distribution, it is called a simply hypothesis. Otherwise, it is called a composite hypothesis.

Example

1) $H_0:\theta=75$ is a simple hypothesis.
2) $H_0:\theta\le 75$, $H_1:\theta>75$ are composite hypothesis.

Definition: Test

A test of statistical hypothesis is a rule such that when the experimental values $(x_1,\dots,x_n)$ have been obtained leads to a decision to accept or reject the hypothesis under consideration.

Definition: Critical Region

Let C be that subset of the sample space which in accordance to the prescribed rule of the test leads to the rejection of the hypothesis under consideration. Then C is called the critical region.

Definition: Power Function

The power function of a test of a statistical hypothesis of H_0 against H_1 is the probability of rejecting the hypothesis under consideration.

Definition: Significance Level

The maximum value of the power function when H_0 is true is called the significance level of the test.

Definition: Best Critical Region

A subset C of the sample space is called a best critical region of size α for testing H_0 against H_1 if for every subset A of S with $P((x_1,\ldots,x_n)\in A \mid H_0)=\alpha$ we have

1) $P((x_1,\ldots,x_n)\in C \mid H_0)=\alpha$.
2) $P((x_1,\ldots,x_n)\in C \mid H_1)\geq P((x_1,\ldots,x_n)\in A \mid H_1)$.

Theorem: Neyman-Pearson Theorem

If we take $C=\left\{(x_1,\ldots,x_n) \mid \dfrac{L(\theta' \mid x_1,\ldots,x_n)}{L(\theta'' \mid x_1,\ldots,x_n)}\leq k, k>0\right\}$, then C is a best critical region of size α for testing $H_0:\theta=\theta'$ against $H_1:\theta=\theta''$.

Basic Tests

***z*-Test**

Let $X_1,\ldots,X_n$ be a random sample from $N(\mu,\sigma_0^2)$ where σ_0^2 is known. When the null hypothesis $H_0:\mu=\mu_0$ is true, then $Z=\dfrac{\overline{X}-\mu}{\frac{\sigma_0}{\sqrt{n}}}\sim N(0,1)$. We reject H_0 if the P-value given by

$$P_{\mu_0}\left(\left|\overline{X}-\mu_0\right|\geq\left|\overline{x}-\mu_0\right|\right)=P_{\mu_0}\left(|Z|\geq\left|\frac{\overline{x}-\mu_0}{\frac{\sigma_0}{\sqrt{n}}}\right|\right)=1-\Phi\left(\left|\frac{\overline{x}-\mu_0}{\frac{\sigma_0}{\sqrt{n}}}\right|\right)+\Phi\left(-\left|\frac{\overline{x}-\mu_0}{\frac{\sigma_0}{\sqrt{n}}}\right|\right)=2\left[1-\Phi\left(\left|\frac{\overline{x}-\mu_0}{\frac{\sigma_0}{\sqrt{n}}}\right|\right)\right]$$ is small.

If the P-value is less than $1-\alpha$, then the results are said to be statistically significant at the $100(1-\alpha)\%$ level.

Bernoulli Model

Let $X_1, \ldots, X_n$ be a random sample from $\text{Bernoulli}(\theta)$ where $\theta \in [0,1]$ is unknown. Suppose we want to test $H_0 : \theta = \theta_0$. When H_0 is true, then $Z = \dfrac{\overline{X} - \theta_0}{\sqrt{\dfrac{\theta_0(1-\theta_0)}{n}}}$ has limiting $N(0,1)$ distribution. Then the approximate P-value is $P\left(|Z| \geq \left|\dfrac{\bar{x} - \theta_0}{\sqrt{\dfrac{\theta_0(1-\theta_0)}{n}}}\right|\right) \approx 2\left[1 - \Phi\left(\left|\dfrac{\bar{x} - \theta_0}{\sqrt{\dfrac{\theta_0(1-\theta_0)}{n}}}\right|\right)\right]$. We reject H_0 when the P-value is small.

Equivalence Between *z*-Test and *z*-Confidence-Intervals

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma_0^2)$ where σ_0^2 is known. The confidence interval $\left(\bar{x} - z_{\frac{1+\alpha}{2}} \dfrac{\sigma_0}{\sqrt{n}}, \bar{x} + z_{\frac{1+\alpha}{2}} \dfrac{\sigma_0}{\sqrt{n}}\right), z_{\frac{1+\alpha}{2}} = \Phi^{-1}\left(\dfrac{1+\alpha}{2}\right)$ includes μ with probability α. If we decide that for any P-value less than $1 - \alpha = 0.05$ we declare the results statistically significant, then we know that the results are statistically significant whenever the 95% confidence interval for μ doesn't contain μ.

t-Test

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. Suppose we want to test $H_0 : \mu = \mu_0$. When H_0 is true, $T = \dfrac{\overline{X} - \mu}{\dfrac{S}{\sqrt{n}}} \sim t_{n-1}$. We reject the null-hypothesis H_0 when the P-value given by

$P_{(\mu,\sigma^2)}\left(|T| \geq \left|\dfrac{\bar{x} - \mu_0}{\dfrac{S}{\sqrt{n}}}\right|\right) = 2\left[1 - G\left(\left|\dfrac{\bar{x} - \mu_0}{\dfrac{S}{\sqrt{n}}}\right|\right)\right]$ is small; here G is the cdf of t_{n-1}.

Sample Size Calculations

We may determine the sample size n so that the margin of error for an $100\alpha\%$ confidence interval for μ does not exceed a prescribed value $\delta > 0$.

THE METHOD OF MOMENTS

Let $X_1, \ldots, X_n$ be a random sample with pdf $f_{(\theta_1,\ldots,\theta_r)}(x), (\theta_1, \ldots, \theta_r) \in \Omega$. The expectation $M_k = E[X^k]$ is the k^{th} moment of the distribution. The sum $m_k = \sum_{i=1}^{n} \dfrac{X_i^k}{n}$ is called the k^{th} moment of the sample.

The Method of Moments

A method of point estimation called the method of moments can be described as follows:

1) Equate M_k to the experimental value m_k.
2) Beginning with $k = 1$ and continuing until there are enough equations to obtain $\theta_1, \ldots, \theta_r$ as functions of $m_1, m_2, \ldots$, that is, $\theta_i = h_i(\theta_1, \ldots, \theta_r), i = 1, \ldots, r$.

Bayesian Inference

Consider a random variable *X* whose distribution depends on $\theta \in \Omega$. We have previously look on θ being some unknown constant. Just as we look on *x* as a possible value of *X*, we now look on θ as a possible value of the random Θ that has a distribution Π on the set Ω.

The Prior Distribution

We shall denote the pdf (or pf) of Θ by π and define $\pi(\theta) = 0$ when $\theta \notin \Omega$. $\pi(\theta)$ is called the prior pdf of Θ. This is because $\pi(\theta)$ is the pdf (or pf) of Θ prior to the observation on *Y*.

The Posterior Distribution

Let $X_1, \ldots, X_n$ be a random sample from the distribution of *X*, and let *Y* be a statistic which is a function of $X_1, \ldots, X_n$. We can find the conditional pdf (or pf) of *Y* given Θ, which we denote by $g(y \mid \theta)$. Thus the joint pdf (or pf) of *Y* and Θ is given by $k(\theta, y) = \pi(\theta)g(y \mid \theta)$. If Θ is continuous, then the marginal pdf of *Y* is given by $m(y) = \int_{-\infty}^{\infty} k(\theta, y)d\theta = \int_{-\infty}^{\infty} \pi(\theta)g(y \mid \theta)d\theta$. If Θ is discreet, then integration would be replaced by summation. In either case the conditional pdf (or conditional pf) of *Y* given Θ is $K(\theta \mid y) = \dfrac{k(\theta, y)}{m(y)} = \dfrac{\pi(\theta)g(y \mid \theta)}{m(y)}, m(y) \neq 0$. $K(\theta \mid y)$ is called the posterior pdf (or pf) of Θ. This is because $K(\theta \mid y)$ is the pdf (or pf) of Θ after the observation on *Y* has been made.

Model Checking

One approach is to choose the discrepancy statistic $D : S \to \mathbf{R}$. If the observed value of *D*, $D(s)$, lies in the region of low probability, then we reject the hypothesis that the model under investigation is true.
In order to compare $D(s)$ with *D*, we need to compute the *P*-value $P(D > D(s))$. This can be done via two methods:

1) This method requires that *D* be ancillary.
2) In this method, we use the conditional distribution of *D* given the value of sufficient statistic *T*. It can be shown that this conditional probability is the same for every parameter θ.

Definition: Ancillary

A statistic whose distribution does not depend on the parameter θ is called ancillary.

Example

We assume $X_1,\dots,X_n$ is a random sample from $N\left(\mu,\sigma_0^2\right)$ (μ unknown, σ_0^2 known). It has been shown that $\overline{X}=\dfrac{X_1+\cdots+X_n}{n}$ is a minimal sufficient statistic.

Consider a sample value $(x_1,\dots,x_n)$ and define $r=r(x_1,\dots,x_n)=(r_1,\dots,r_n)=(x_1-\overline{x},\dots,x_n-\overline{x})$. It can be shown that:

- $R=(R_1,\dots,R_n)=\left(X_1-\overline{X},\dots,X_n-\overline{X}\right)$ has a distribution that is independent of μ. Hence R is ancillary.
- R is independent of $\overline{X}$. So the conditional distribution $R \mid \overline{X}=\overline{x}$ is the same as the distribution of R.

Therefore, the two methods agree in this case.

Also, $R_i \sim N\left(0,\sigma_0^2\left(1-\frac{1}{n}\right)\right), \forall i=1,\dots,n$. Now consider the discrepancy statistic

$D(R)=\dfrac{1}{\sigma_0^2}\sum_{i=1}^{n}R_i^2=\dfrac{1}{\sigma_0^2}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$. We know $D(R)\sim\chi_{n-1}^2$. Compute $P(D(R)>D(s))$ to see if the observed value $D(s)$ is in a region of low probability or not (values close to 0 and 1 indicates tails of the distribution and both cases indicate a region of low probability). If it is, then we reject the model under investigation being true.

Example

We assume $X_1,\dots,X_n$ is a random sample from $N\left(\mu,\sigma^2\right)$ (both μ and σ^2 unknown). It has been shown that $\left(\overline{X},S^2\right)$ is a minimal sufficient statistic.

Consider a sample value $(x_1,\dots,x_n)$ and define $r=r(x_1,\dots,x_n)=(r_1,\dots,r_n)=\left(\dfrac{x_1-\overline{x}}{s},\dots,\dfrac{x_n-\overline{x}}{s}\right)$. It can be shown that:

- $R=(R_1,\dots,R_n)=\left(\dfrac{X_1-\overline{X}}{S},\dots,\dfrac{X_n-\overline{X}}{S}\right)$ has a distribution that is independent of μ and σ^2. Hence R is ancillary.
- R is independent of $\overline{X}$. So the conditional distribution $R \mid \overline{X}=\overline{x}$ is the same as the distribution of R.

Therefore, the two methods agree in this case.

Now consider the discrepancy statistic $D(R)=-\dfrac{1}{n}\sum_{i=1}^{n}\ln\left(\dfrac{R_i^2}{n-1}\right)$. To use this statistic for model checking, we need to know the distribution of the statistic; this can be done via simulation. Then compute $P(D(R)>D(s))$ to see if the observed value $D(s)$ is in a region of low probability or not (values close to 0 and 1 indicates tails of the distribution and both cases indicate a region of low probability). If it is, then we reject the model under investigation being true.

Chi-Squared Goodness of Fit Test

Let $X_1,\dots,X_k$ have a multinomial distribution with parameters $p_1,\dots,p_k$ where $p_k=1-p_1-\cdots-p_{k-1}$ and $X_k=n-X_1-\cdots-X_{k-1}$. Define $Q_{k-1}=\sum_{i=1}^{k}\dfrac{(X_i-np_i)}{np_i}$. It can be shown that as $n\to\infty$, Q_{k-1} has a limiting χ_{n-1}^2 distribution. Hence Q_{k-1} is approximately a χ_{n-1}^2 distribution.

Procedure

1) Let A be the sample space of a random experiment. $A = A_1 \cup A_2 \cup \cdots \cup A_k = \bigcup_{i=1}^{k} A_i$ where $A_i \cap A_k = \varnothing, \forall i \neq j$.
2) Let $p_i = P(A_i), i = 1, \ldots, k$.
3) We repeat the experiment n times.
4) Let X_i denote the number of times the outcome of the random experiment is in $A_i, i = 1, \ldots, k$. Then $X_1, \ldots, X_k$ has a multinomial distribution with parameters $p_1, \ldots, p_k$.
5) We test the simple hypothesis $H_0 : p_1 = p_{1_0}, \ldots, p_k = p_{k_0}$ where $p_{1_0}, \ldots, p_{k_0}$ are specified values against all other alternatives.
6) If H_0 is true, then the statistic $Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_i)}{np_i}$ has an approximate χ^2_{n-1} distribution. We find c so that $P(Q_{k-1} \geq c) = \alpha$, where α is the desired significance level of the test (the significance level of the test is approximately equal to α).
7) We reject H_0 if the observed value of $Q_{k-1} \geq c$.

METHOD OF LEAST SQUARES

Suppose we want to estimate $E(Y)$ based on a sample value $(y_1, \ldots, y_n)$. We select the point $t(y_1, \ldots, y_n)$ in the set of possible values of $E(Y)$ that minimizes $\sum_{i=1}^{n} (y_i - t(y_1, \ldots, y_n))^2$.

The estimate $t(y_1, \ldots, y_n)$ is called the least square estimate of $E(Y)$.

Note

We have
$$\begin{aligned}\sum_{i=1}^{n} (y_i - t(y_1, \ldots, y_n))^2 &= \sum_{i=1}^{n} ((y_i - \bar{y}) + (\bar{y} - t(y_1, \ldots, y_n)))^2 \\ &= \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (\bar{y} - t(y_1, \ldots, y_n))^2 + 2\sum_{i=1}^{n} (y_i - \bar{y})(\bar{y} - t(y_1, \ldots, y_n))\end{aligned}$$
, but

$$\begin{aligned}\sum_{i=1}^{n} (y_i - \bar{y})(\bar{y} - t(y_1, \ldots, y_n)) &= (\bar{y} - t(y_1, \ldots, y_n))\sum_{i=1}^{n} (y_i - \bar{y}) = (\bar{y} - t(y_1, \ldots, y_n))\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y}\right). \\ &= (\bar{y} - t(y_1, \ldots, y_n))(n\bar{y} - n\bar{y}) = 0\end{aligned}$$

So $\sum_{i=1}^{n} (y_i - t(y_1, \ldots, y_n))^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (\bar{y} - t(y_1, \ldots, y_n))^2$. This is minimized when $t(y_1, \ldots, y_n) = \bar{y}$ if $\bar{y}$ is a possible value of $E(Y)$; if not, we choose a possible value of $E(Y)$ that is closest to $\bar{y}$ as the estimate of $t(y_1, \ldots, y_n)$.

Case of Random Vector

1) We have $\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ and $E(\mathbf{Y}) = \begin{pmatrix} E(Y_1) \\ \vdots \\ E(Y_n) \end{pmatrix}$.

2) We observe $(y_1, \ldots, y_n) \in \mathbf{R}^n$ and find $\mathbf{t}(y_1, \ldots, y_n) = \begin{pmatrix} t_1(y_1, \ldots, y_n) \\ \vdots \\ t_n(y_1, \ldots, y_n) \end{pmatrix} \in \{\text{possible values of } E(Y)\}$ that minimizes $\sum_{i=1}^{n} (y_i - t_i(y_1, \ldots, y_n))^2$.

Regression Models: The Simple Linear Regression Model

We study the relation between the response variable Y (dependent) and the predictor variable X (independent). In the Regression Model, we think the change is the through the conditional mean, that is as x changes $E(Y \mid X = x)$ changes.

Definition

In the Simple Regression model, we assume $E(Y \mid X = x) = \beta_1 + \beta_2 x$. β_1 and β_2 are called the regression coefficients.

Definition: Scatter Plot

A scatter plot is a plot of data points $(x_1, y_1), \ldots, (x_n, y_n)$. It shows whether a relation exists between X and Y and the form of the relation.

Least Square Estimate, Predictions, and Standard Errors

Definition: Least Square Estimate

Suppose we observe the independent numbers $(x_1, y_1), \ldots, (x_n, y_n)$. We have $E\left(\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \mid X_1 = x_1, \ldots, X_n = x_n\right) = \begin{pmatrix} \beta_1 + \beta_2 x_1 \\ \vdots \\ \beta_1 + \beta_2 x_n \end{pmatrix}$. The lease square estimate of the conditional mean is the value of $t(y_1, \ldots, y_n)$ in the set of possible values of the conditional mean that minimizes $\sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_1)^2$. The β_1 and β_2 that minimizes this is called the least square estimate of β_1 and β_2.

Theorem

Suppose that $E(Y \mid X = x) = \beta_1 + \beta_2 x$ and we observe the independent values $(x_1, y_1), \ldots, (x_n, y_n)$ for (X, Y). Then the least square estimates of

- β_1 is $b_1 = \bar{y} - b_2 \bar{x}$,

- β_2 is $b_2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, whenever $\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0$.

Note

1) $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i y_i - \bar{x}y_i - x_i\bar{y} + \overline{xy}) = \sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\overline{xy} = \sum_{i=1}^{n} x_i y_i - n\overline{xy}$.

2) $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^{\,2} + \sum_{i=1}^{n}\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^{\,2} + n\bar{x}^2 - 2n\bar{x}^2 = \sum_{i=1}^{n} x_i^{\,2} - n\bar{x}^2$.

So, $b_2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$.

Theorem

If $E(Y \mid X = x) = \beta_1 + \beta_2 x$ and we observe independent values $(x_1, y_1), \ldots, (x_n, y_n)$, then

- $E(B_1 \mid X_1 = x_1, \ldots, X_n = x_n) = \beta_1$, where $B_1 = \bar{Y} - B_2\bar{X}$;
- $E(B_2 \mid X_1 = x_1, \ldots, X_n = x_n) = \beta_2$, where $B_2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

Thus B_1 and B_2 have unbiased property, that is they are unbiased estimators.

Remark

A natural predictor of a future value of Y when $X = x$ is $E(Y \mid X = x) = \beta_1 + \beta_2 x$. Because we do not have the values of β_1 and β_2, we use the estimates b_1 and b_2 for prediction. That is, use the line $b_1 + b_2 x$.

Theorem

If $E(Y \mid X = x) = \beta_1 + \beta_2 x$ and $\text{var}(Y \mid X = x) = \sigma^2, \forall x$, and we observe independent values $(x_1, y_1), \ldots, (x_n, y_n)$, then

- $\text{var}(B_1 \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$,
- $\text{var}(B_2 \mid X_1 = x_1, \ldots, X_n = x_n) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$,
- $\text{cov}(B_1, B_2 \mid X_1 = x_1, \ldots, X_n = x_n) = \dfrac{-\sigma^2\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Corollary

From the theorem above, we obtain $\operatorname{var}(B_1 + B_2 x \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$.

<u>Proof</u>:

$$\begin{aligned}
&\operatorname{var}(B_1 + B_2 x \mid X_1 = x_1, \ldots, X_n = x_n) \\
&= \operatorname{var}(B_1 \mid X_1 = x_1, \ldots, X_n = x_n) + x^2 \operatorname{var}(B_2 x \mid X_1 = x_1, \ldots, X_n = x_n) + 2x \operatorname{cov}(B_1, B_2 \mid X_1 = x_1, \ldots, X_n = x_n) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) + x^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - 2x \frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)
\end{aligned}$$

Theorem

If $E(Y \mid X = x) = \beta_1 + \beta_2 x$ and $\operatorname{var}(Y \mid X = x) = \sigma^2, \forall x$, and we observe independent values $(x_1, y_1), \ldots, (x_n, y_n)$ for (X, Y), then $E(S^2 \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2$ where $S^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - B_1 - B_2 X_i)^2$. Thus $s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - b_1 - b_2 x_i)^2$ is an unbiased estimate of σ^2.

Therefore, the standard error of

- b_1 is $s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$,
- b_2 is $\frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$.

THE ANOVA DECOMPOSITION AND THE F STATISTIC

Lemma

If $(x_1, y_1), \ldots, (x_n, y_n)$ are such that $\sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$, then an useful decomposition of $\sum_{i=1}^n (y_i - \bar{y})^2$ is

$$\sum_{i=1}^n (y_i - \bar{y})^2 = b_2^2 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - b_1 - b_2 x_i)^2 .$$

- $b_2^2 \sum_{i=1}^n (x_i - \bar{x})^2$ is called the regression sum of squares (RSS).
- $\sum_{i=1}^n (y_2 - b_1 - b_2 x_i)^2$ is called the error sum of squares (ESS).

ANOVA (Analysis of Variance Table)

Source	DF (degrees of freedom)	Sum of Squares	Mean Square

X	1	$b_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$	$b_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$
Error	$n-2$	$\sum_{i=1}^{n} (y_2 - b_1 - b_2 x_i)^2$	$\frac{1}{n-2} \sum_{i=1}^{n} (y_2 - b_1 - b_2 x_i)^2 = s^2$
Total	$n-1$	$\sum_{i=1}^{n} (y_i - \bar{y})^2$	

Result

We have $E\left(B_2^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 \mid X_1 = x_1, \ldots, X_n = x_n \right) = \sigma^2 + \beta_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$, which is equal to σ^2 if and only if $\beta_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0$. In other words, $B_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$ is an unbiased estimator of σ^2 if and only if $\beta_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0$.

Proof:

$$E\left(B_2^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 \mid X_1 = x_1, \ldots, X_n = x_n \right) = \left(\sum_{i=1}^{n} (X_i - \bar{X})^2 \right) E\left(B_2^2 \mid X_1 = x_1, \ldots, X_n = x_n\right)$$
$$= \left(\sum_{i=1}^{n} (X_i - \bar{X})^2 \right) \left(\mathrm{var}(B_2 \mid X_1 = x_1, \ldots, X_n = x_n) + \left(E(B_2 \mid X_1 = x_1, \ldots, X_n = x_n)\right)^2 \right).$$
$$= \left(\sum_{i=1}^{n} (X_i - \bar{X})^2 \right) \left(\frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + \beta^2 \right) = \sigma^2 + \beta_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Definition: The F-Statistic

Since $s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2$ is an unbiased estimate of σ^2, hence the F-statistic given by $F = \frac{RSS}{ESS/n-2} = \frac{b_2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{s^2}$ is the ratio of two unbiased estimates of σ^2 when $H_0 : \beta_2 = 0$ (there is a linear effect due to *X*) is true. Therefore reject $H_0 : \beta_2 = 0$ when *F* is large.

THE COEFFICIENT OF DETERMINATION AND CORRELATION

Definition: Coefficient of Determination

Define $R^2 = \dfrac{b_2^2 \sum_{i=0}^n (x_i - \bar{x})^2}{\sum_{i=0}^n (y_i - \bar{y})^2}$ as the coefficient of determination. Since

$\sum_{i=1}^n (y_i - \bar{y})^2 = b_2^2 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - b_1 - b_2 x_i)^2$, we have that

$R^2 = \dfrac{b_2^2 \sum_{i=0}^n (x_i - \bar{x})^2}{b_2^2 \sum_{i=0}^n (x_i - \bar{x})^2 + \sum_{i=0}^n (y_i - b_1 - b_2 \bar{x})^2}$, so $0 \le R^2 \le 1$.

Note: When we use the model to make predictions, a value of R^2 near 1 means a highly accurate prediction, while a value of R^2 near 0 means the predictions will not be very accurate.

Definition: Correlation

We define the correlation coefficient by $\rho_{XY} = \operatorname{corr}(X, Y) = \dfrac{\operatorname{cov}(X, Y)}{\operatorname{SD}(X)\operatorname{SD}(Y)}$. We know that $-1 \le \rho_{XY} \le 1$ and that $\rho_{XY} = \pm 1 \Leftrightarrow Y = a \pm cX$.

Definition: Sample Correlation

We define the sample correlation by $r_{XY} = \dfrac{s_{XY}}{s_X s_Y}$, where $s_{XY} = \dfrac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ is the sample covariance estimating $\operatorname{cov}(X, Y)$, and $s_X = \sqrt{\dfrac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$ and $s_Y = \sqrt{\dfrac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}$ are sample standard deviations of *X* and *Y*. We have that $-1 \le r_{XY} \le 1$ and $r_{XY} = \pm 1 \Leftrightarrow y = a \pm c x_i, \forall i$.

Theorem

If $(x_1, y_1), \ldots, (x_n, y_n)$ are such that $\sum_{i=1}^n (x_i - \bar{x})^2 \ne 0$ and $\sum_{i=1}^n (y_i - \bar{y})^2 \ne 0$, then $R^2 = r_{XY}^2$.

Proof:

$$
\begin{aligned}
r_{XY}^2 &= \frac{\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)\left(\sum_{i=1}^n (y_i - \bar{y})^2\right)} = \frac{b_2^2\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)\left(\sum_{i=1}^n (y_i - \bar{y})^2\right)} = \frac{b_2^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{b_2^2 \sum_{i=1}^n (x_i - \bar{x})^2}{b_2^2 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - b_1 - b_2 x_i)^2} = R^2
\end{aligned}.
$$

CONFIDENCE INTERVALS AND TESTING HYPOTHESES

Theorem

If $Y \mid X = x$ is distributed $N\left(\beta_1 + \beta_2 x, \sigma^2\right)$ and we observe the independent values $(x_1, y_1), \ldots, (x_n, y_n)$ for (X, Y), then the conditional distribution of B_1, B_2, and S^2 given $X_1 = x_1, \ldots, X_n = x_n$ are:

- $B_1 \sim N\left(\beta_1, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right)$,
- $B_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$,
- $B_1 + B_2 x \sim N\left(\beta_1 + \beta_2 x, \sigma^2\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right)$,
- $\dfrac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$ independent of (B_1, B_2).

Corollary

1) $\dfrac{B_1 - \beta_1}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_1 - \bar{x})^2}}} \sim t(n-2)$.
2) $\dfrac{(B_2 - \beta_2)\sqrt{\sum_{i=1}^n (x_1 - \bar{x})^2}}{S} \sim t(n-2)$.
3) $\dfrac{B_1 + B_2 x - \beta_1 - \beta_2 x}{S\sqrt{\frac{\frac{1}{n} + (x - \bar{x})^2}{\sum_{i=1}^n (x_1 - \bar{x})^2}}} \sim t(n-2)$.
4) If F is defined by $F = \dfrac{RSS}{\frac{ESS}{n-2}} = \dfrac{B_2^2 \sum_{i=1}^n (x_i - \bar{x})^2}{S^2}$, then $H_0 : \beta_2 = 0$ is true if and only if $F \sim F(1, n-2)$.

Confidence Intervals

1) An $100\alpha\%$ confidence interval for β_1 is
$$\left(b_1 - t_{\frac{1+\alpha}{2}}(n-2) \cdot s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, b_1 + t_{\frac{1+\alpha}{2}}(n-2) \cdot s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right).$$
2) An $100\alpha\%$ confidence interval for β_2 is
$$\left(b_2 - t_{\frac{1+\alpha}{2}}(n-2) \cdot \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}, b_2 + t_{\frac{1+\alpha}{2}}(n-2) \cdot \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}\right).$$

Testing Hypothesis

We can test the hypothesis $H_0 : \beta_2 = 0$ by computing the P-value $P\left(F \geq \dfrac{b_2^2 \sum_{i=1}^n (x_i - \bar{x})^2}{s^2}\right), F \sim F(1, n-2)$ to see whether or not the observed value lies in a region of low probability.

Notice that we can also test the hypothesis $H_0 : \beta_2 = 0$ by computing the P-value

$$P\left(|T| \geq \frac{\sqrt{b_2 \sum_{i=1}^{n} (x_i - \bar{x})^2}}{s}\right), T \sim t(n-2).$$

It can be shown that these two P-values are equal.

Analysis of Residuals

Model checking is based on the residual $y_i - b_1 - b_2 x_i$ as discussed earlier.

Corollary

1) $E(Y_i - B_1 - B_2 x_i \mid X_1 = x_1, \ldots, X_n = x_n) = 0$.

2) $\operatorname{var}(Y_i - B_1 - B_2 x_i \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right)$.

Definition: Standardized Residual

We define the i^{th} standardized residual by $\dfrac{y_i - b_2 - b_2 x_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}}$.

Note

If the conditional distribution of $Y \mid X_1 = x_1, \ldots, X_n = x_n$ is normal, then $\dfrac{y_i - b_2 - b_2 x_i}{\sigma\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}} \sim N(0,1)$. This is approximately true for $\dfrac{y_i - b_2 - b_2 x_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}}$ for large n.

The Residual and Probability Plots

One approach to model checking is to see if the values of the standardized residuals look like a sample from $N(0,1)$. For this we use the residual and probability plots.

Residual Plot

We defined the residual earlier as $\mathbf{r} = (r_1, \ldots, r_n) = (x_1 - \bar{x}, \ldots, x_n - \bar{x})$ where $(x_1, \ldots, x_n)$ is a sample value from $N(\mu, \sigma_0^2)$. It can be shown that $R = (X_1 - \bar{X}, \ldots, X_n - \bar{X}) \sim N\left(0, \sigma_0^2 \left(1 - \frac{1}{n}\right)\right)$ and is independent of $\bar{X}$.

- We standardize R_i as $R_i^* = \sqrt{\dfrac{n}{\sigma_0^2 (n-1)}}(X_i - \bar{X})$. Then $R_i^* \sim N(0,1)$ has mean 0 and variance 1.

- When σ^2 is unknown, we estimate it by s^2. In this case $R_i^* = \sqrt{\frac{n}{S_0^2(n-1)}}\left(X_i - \overline{X}\right)$ is approximately distributed $N(0,1)$.

The plot of $\left(i, r_i^*\right)$ should not show any discernible pattern. Most of the values should lie in $(-3,3)$. A discernible patter with several extreme values is evidence against the model assumption to be correct. Simulating normal random variables and plotting it gives a good idea of how the plot should look.

Probability Plot

Suppose $X_1, \ldots, X_n$ is a random sample from $N\left(\mu, \sigma^2\right)$; suppose $x_1, \ldots, x_n$ is a sample value. Then, it can be shown that the expectation of the i^{th} order statistic satisfies $E\left(X_{(i)}\right) \approx \mu + \sigma \cdot \Phi^{-1}\left(\frac{i}{n+1}\right)$. If the data x_j corresponds to the order statistic $x_{(i)}$, that is $x_j = x_{(i)}$, then we call $\Phi^{-1}\left(\frac{i}{n+1}\right)$ the normal score of x_j. Then $E\left(X_{(i)}\right) \approx \mu + \sigma \cdot \Phi^{-1}\left(\frac{i}{n+1}\right)$ indicates that if we plot the points $\left(x_i, \Phi^{-1}\left(\frac{i}{n+1}\right)\right)$ they should lie on a line with intercept μ and slope σ. We call such a plot a probability plot.

Regression Models: One Categorical Predictor (One-Way ANOVA)

Now suppose that the predictor X takes a values $1,2,\ldots,a$. Let $\beta_i = E(Y \mid X = i)$ the mean response when X takes the value i. Define $X_i = \begin{cases} 1, X = i \\ 0, X \neq i \end{cases}, i = 1, \ldots, a$. We have $E(Y \mid X_1 = x_1, \ldots, X_a = x_a) = \beta_1 x_1 + \cdots + \beta_a x_a$. Since only one of the $x_i = 1$ and the rest are 0, we obtain the simple linear regression model and hence all the results previously obtained hold.

Least Square Estimate

Now suppose we have n_i values $\left(y_{i1}, \ldots, y_{in_i}\right)$ when $X = i$ and all response values are independent. The least square estimates of β_i are obtained by minimizing $\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(y_{ij} - \beta_i\right)^2$. So the least square estimate of β_i is $b_i = \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$.

It can be shown that $B_i = \overline{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ is an unbiased estimator of β_i, that is $E(B_i) = E\left(\overline{Y}\right) = \beta_i$.

Sample Variance

Assuming $Y \mid X = x$ all have conditional variance σ^2, that is $\operatorname{var}(Y \mid X = i) = \sigma^2, \forall i$. Then $\operatorname{var}\left(\overline{Y}_i \mid X = i\right) = \operatorname{var}\left(\frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \mid X = i\right) = \frac{1}{n_i^2}\left(\operatorname{var}(Y_{i1} \mid X = 1) + \cdots \operatorname{var}\left(Y_{in_i} \mid X = a\right)\right) = \frac{1}{n_i^2}\left(n_i \sigma^2\right) = \frac{\sigma^2}{n_i}$, and the conditional covariance $\operatorname{cov}\left(\overline{Y}_i, \overline{Y}_j\right) = 0$ when $i \neq j$.

We have $s^2 = \frac{1}{N-a}\sum_{i=1}^{a}\sum_{j=i}^{n_i}\left(y_{ij}-\overline{y}_i\right)^2, N = n_1+\cdots+n_a$ is an unbiased estimate of σ^2.

Confidence Interval and Testing

If we assume that $Y \mid X = i \sim N\left(\beta_i, \sigma^2\right)$, then $\overline{Y}_i \sim N\left(\beta_i, \frac{\sigma^2}{n_i}\right)$ independent of $\frac{(N-a)S^2}{\sigma^2} \sim \chi^2(N-a)$.

Now $\frac{\overline{Y}_i-\beta_i}{\sigma/\sqrt{n_i}} \sim N(0,1)$, and hence (by definition) $T = \frac{\frac{\overline{Y}_i-\beta_i}{\sigma/\sqrt{n_i}}}{\sqrt{\frac{(N-a)S^2/\sigma^2}{N-a}}} = \frac{\overline{Y}_i-\beta_i}{S/\sqrt{n_i}} \sim t(N-a)$.

Since $P\left(-a < \frac{\overline{Y}_i-\beta_i}{S/\sqrt{n_i}} < a\right) = \alpha \Leftrightarrow P\left(\overline{Y}_i - a\frac{S}{\sqrt{n_i}} < \beta_i < \overline{Y}_i + a\frac{S}{\sqrt{n_i}}\right) = \alpha$, hence an $100\alpha\%$ confidence interval for β_i is $\left(\overline{y}_i - \frac{s}{\sqrt{n_i}}\cdot t_{\frac{1+\alpha}{2}}(N-a), \overline{y}_i + \frac{s}{\sqrt{n_i}}\cdot t_{\frac{1+\alpha}{2}}(N-a)\right)$.

Also, we can test the null hypothesis $H_0 : \beta_i = \beta_{i_0}$ by computing the *P*-value

$P\left(|T| \geq \left|\frac{\overline{y}_i-\beta_{i_0}}{s/\sqrt{n_i}}\right|\right) = 2\left(1-G\left(\left|\frac{\overline{y}_i-\beta_{i_0}}{s/\sqrt{n_i}}\right|; N-a\right)\right)$ where G is the cdf of the $t(N-a)$ distribution.

Inference About Difference of Means

We now study inference about $\overline{Y}_i - \overline{Y}_j$.

Expectation and Variance

We have $E\left(\overline{Y}_i - \overline{Y}_j\right) = E\left(\overline{Y}_i\right) - E\left(\overline{Y}_j\right) = \beta_i - \beta_j$ and

$\operatorname{var}\left(\overline{Y}_i - \overline{Y}_j\right) = \operatorname{var}\left(\overline{Y}_i\right) + \operatorname{var}\left(\overline{Y}_j\right) = \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_j} = \sigma^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)$ since $\overline{Y}_i$ and $\overline{Y}_j$ are independent.

Confidence Interval

Since $\overline{Y}_i \sim N\left(\beta_i, \frac{\sigma^2}{n_i}\right)$ and $\overline{Y}_j \sim N\left(\beta_j, \frac{\sigma^2}{n_j}\right)$, so $\overline{Y}_i - \overline{Y}_j \sim N\left(\beta_i - \beta_j, \sigma^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)\right)$. We have

$\frac{\left(\overline{Y}_i - \overline{Y}_j\right) - \left(\beta_i - \beta_j\right)}{\sigma/\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}} \sim N(0,1)$ independent of $\frac{(N-a)S^2}{\sigma^2} \sim \chi^2(N-a)$. Thus

$$T = \frac{\frac{\left(\overline{Y}_i - \overline{Y}_j\right) - \left(\beta_i - \beta_j\right)}{\sigma\sqrt{1/n_i+1/n_j}}}{\sqrt{\frac{(N-a)S^2/\sigma^2}{N-a}}} = \frac{\left(\overline{Y}_i - \overline{Y}_j\right) - \left(\beta_i - \beta_j\right)}{S/\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}} \sim t(N-a).$$

Since

$$P\left(-a<\frac{(\bar{Y}_i-\bar{Y}_j)-(\beta_i-\beta_j)}{S\Big/\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}}<a\right)=\alpha \Leftrightarrow P\left((\bar{Y}_i-\bar{Y}_j)-a\frac{S}{\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}}<\beta_i-\beta_j<(\bar{Y}_i-\bar{Y}_j)+a\frac{S}{\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}}\right)=\alpha,$$

hence an $100\alpha\%$ confidence interval for $\beta_i-\beta_j$ is

$$\left((\bar{y}_i-\bar{y}_j)-\frac{s}{\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}}\cdot t_{\frac{1+\alpha}{2}}(N-a),(\bar{y}_i-\bar{y}_j)-\frac{s}{\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}}\cdot t_{\frac{1+\alpha}{2}}(N-a)\right).$$

Also, we can test the null hypothesis $H_0:\beta_i=\beta_j$ by computing the *P*-value

$$P\left(|T|\ge\left|\frac{\bar{y}_i-\bar{y}_j}{s\Big/\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}}\right|\right)=2\left(1-G\left(\left|\frac{\bar{y}_i-\bar{y}_j}{s\Big/\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}}\right|;N-a\right)\right)$$ where *G* is the cdf of the $t(N-a)$ distribution.

Note

If $a=2$, *X* takes on only 2 values. In this case, *T* is called the "two-sample *t*-statistic", the confidence interval is called the "two-sample *t*-confidence-interval", and the *t*-test is called the "two-sample *t*-test". In this case, if we conclude that $\beta_i\ne\beta_j$, then a relationship exists between *X* and *Y*.

In general, when $a\ge 2$, to test the null hypothesis that there is no relationship between the response and the predictor is equivalent to $H_0:\beta_1=\cdots\beta_a=\beta$. If H_0 is true, the least square estimate of β is $\bar{y}$.

ANOVA Table

We have the composition $\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2=\sum_{i=1}^{a}n_i(\bar{y}_i-\bar{y})^2+\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$.

Source	DF (degrees of freedom)	Sum of Squares	Mean Square
X	$a-1$	$\sum_{i=1}^{a}n_i(\bar{y}_i-\bar{y})^2$	$\frac{\sum_{i=1}^{a}n_i(\bar{y}_i-\bar{y})^2}{a-1}$
Error	$N-a$	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$	s^2
Total	$N-1$	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2$	

F-Statistic

To assess H_0 we use the *F*-statistic $\frac{\sum_{i=1}^{a}n_i(\bar{Y}_i-\bar{Y})^2\Big/a-1}{S^2}$. With normality assumption, we have

$F\sim F(a-1,N-a)$ and so we compute $P\left(F>\frac{\sum_{i=1}^{a}n_i(\bar{y}_i-\bar{y})^2\Big/a-1}{s^2}\right)$ and see if the observed value of *F* is in a region of low probability or not.

When $a=2$, this *P*-value is equal to the *P*-value we obtained for the "two-sample *t*-test".

Model Checking

To check the model we look at the standardized residuals given by $\frac{y_{ij} - \overline{y}_i}{s\sqrt{1-\frac{1}{n_i}}}$ and look at the residual plots as before.