STAT 231

Frank Jin

Contents

Section 1

Introduction to Statistics

Subsection 1.1

Language of Statistics

Definition 1 (Empirical Study) An empirical study is one in which knowledge is gained by observation or by experiment.

Definition 2 (Unit) A unit is an individual person, place, or thing about which we can take some measurement(s).

Definition 3 (Population) A population is a collection of units.

Definition 4 (Process) A process is also a collection of units, but those units are 'produced' over time.

Definition 5 (Variate) A variate is a characteristic of a unit. The types are:

- Continuous variates are those that can be measured - at least in theory - to an infinite degree of accuracy.
- Discrete variates are those that can only take a finite or countably infinite number of values.
- Categorical variates are those where units fall into a (non-numeric) category, such as hair colour or university program.
- Ordinal variates are those where an ordering is implied, but not necessarily through a numeric measure. Examples include 'strongly disagree, disagree, neutral, agree, strongly agree' in surveys.
- Complex variates are more unusual, and include open-ended responses to a survey question, or an image. Analyzing complex variates often requires processing to 'convert' them into one of the other types (e.g., using text analysis to decide if a Tweet is positive, negative, or neutral).

Definition 6 (Attribute) An attribute of a population or process is a function of a variate which is defined for all units in the population or process (eg. average, variability, proportion of something in a population or process of some unit).

Definition 7 (Sample Survey) A sample survey is where information is obtained about a finite population by selecting a 'representative' sample of units from the population and determining the variates of interest for each unit in the sample.

Definition 8 (Observational Study) An observational study is where information about a population or process is collected without any attempt to change one or more variates for the sampled units.

Definition 9 (Experimental Study) An experimental study is one in which the experimenter intervenes and changes or sets the values of one or more variates for the units in the study.

Subsection 1.2

Data Summaries

Definition 10 (Measures of Location) For dataset $\{y_1, ..., y_n\}$, measures of location or central tendacy include:

- Sample mean: $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Sample median: $\hat{m}$ is the middle value when n is odd, and the average of the two middle values when n is even (assuming the dataset is sorted in non-decreasing order).
- The sample mode, or the value of y which appears in the sample with the highest frequency (not necessarily unique).

Definition 11 (Measures of dispersion or variability) For dataset $\{y_1, ..., y_n\}$:

- Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2)$
- Range: $y_{(n)} - y_{(1)}$
- Interquartile range: $q(0.75) - q(0.25)$

Definition 12 (Measures of Shape) For dataset $\{y_1, ..., y_n\}$:

- Sample skewness: $g_1 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^3}{(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2)^{3/2}}$
  - Positive skew: long right tail. Negative skew: long left tail.
- Sample kurtosis: $g_2 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^4}{(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2)^2}$
  - Kurtosis > 3: larger tails. Kurtosis < 3: smaller tails

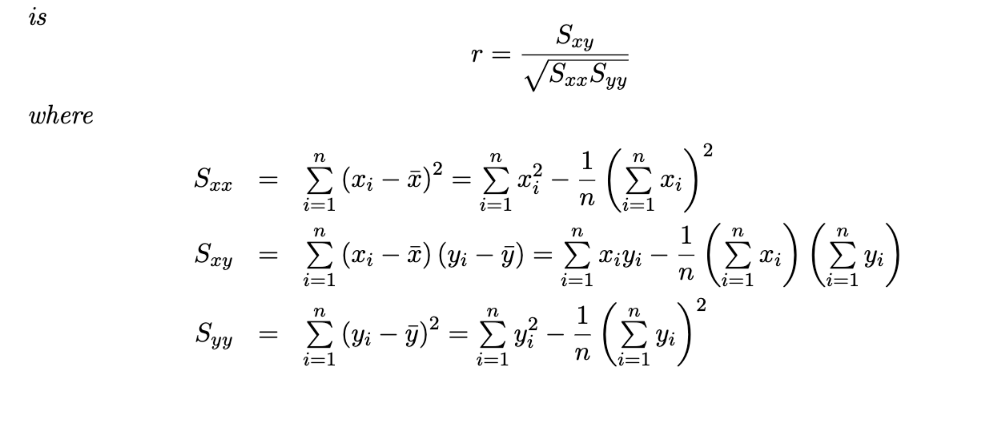
Definition 13 (Quantile) Let $\{y_{(1)}, ..., y_{(n)}\}$ be the order statistic of dataset $\{y_1, ..., y_n\}$. For $0 < p < 1$, the pth sample quantile (also called $100p$th sample percentile) is a value $q(p)$ determined as follows:

- Let $k = (n+1)p$ where n is sample size
- If k is an integer and $1 \leq k \leq n$, then $q(p) = y_{(k)}$
- Otherwise if $1 < k < n$, determine the closest integer j such that $j < k < j+1$ and then $q(p) = \frac{1}{2}(y_{(j)} + y_{(j+1)})$.

$q(0.25)$ is the lower/first quartile, $q(0.5)$ is the median, and $q(0.75)$ is the upper/third quartile.

Definition 14 (5 Number Summary) The 5 number summary of a data set consists of the smallest observation, the lower quartile, the median, the upper quartile and the largest value.

Definition 15 (Sample Correlation) The sample correlation, denoted by r, for data $\{(x_1, y_1), ..., (x_n, y_n)\}$



is

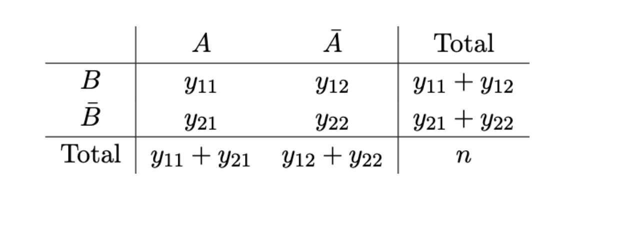
$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

where

$$\begin{aligned}
S_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \\
S_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \\
S_{yy} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2
\end{aligned}$$

The sample correlation, which takes on values between 1 and 1, is a measure of the linear relationship between the two variates x and y. If the value of r is close to 1 then we say that there is a strong positive linear relationship between the two variates while if the value of r is close to 1 then we say that there is a strong negative linear relationship between the two variates. If the value of r is close to 0 then we say that there is no linear relationship between the two variates.

Definition 16 (Relative Risk) Given the following table:



	A	$\bar{A}$	Total
B	y_{11}	y_{12}	$y_{11} + y_{12}$
$\bar{B}$	y_{21}	y_{22}	$y_{21} + y_{22}$
Total	$y_{11} + y_{21}$	$y_{12} + y_{22}$	n

The relative risk of event A in group B as compared to group $\overline{B}$ is

$$\frac{y_{11}/(y_{11} + y_{12})}{y_{21}/(y_{21} + y_{22})}$$

If A and B are independent, then relative risk equals 1. If relative risk equals x, it implies the probability of A given B is x times as likely as the probability of A given $\overline{B}$.

Section 2

Statistical Models and MLE

Subsection 2.1

Point Estimates and Maximum Likelihood Estimation

SUBSECTION 2.2

Likelihood Functions for Continuous Distributions

SUBSECTION 2.3

Invariance Property of Maximum Likelihood Estimate

SUBSECTION 2.4

Qqplots for Gaussian Models

Given data $\{y_1, ..., y_n\}$ and known values μ and σ, we can assess fit of the data to $G(\mu, \sigma)$ using a qqplot. Let $\{y_{(1)}, ..., y_{(n)}\}$ be the order statistic of the data. Let $Q(p)$ the the pth theoretical quantile of $G(\mu, \sigma)$, that is, $P(Y \leq Q(p)) = p$ for $Q \sim G(\mu, \sigma)$. Let $q(p)$ be the pth sample quantile of our observed data. Generally, we want theoretical and sample quantiles to align (eg. $q(0.5) \approx Q(0.5)$). Overall, we plot

$$\left(Q\left(\frac{i+1}{n}\right), q\left(\frac{i+1}{n}\right)\right)$$

for each $i = 1, ..., n$. We use $(i+1)/n$ rather than i/n because $Q(1) = \infty$. If the data is Gaussian, we should see that the plot is relatively straight. If μ and σ are unknown, we can still plot against $G(0, 1)$ to expect a relatively straight line.

- S-shape: indicates distribution is symmetric with skewness close to 0
  - If sample data is above the line on the left and below the line on the right, underlying distribution has lighter tails than Gaussian, so kurtosis should be less than 3.
  - If sample data is below the line on the left and above the line on the right, more observations in tails than expected for Gaussian, so kurtosis is greater than 3.
- U-shape: indicates distribution not symmetric
  - Long right tail: skewness is positive.
  - Long left tail: skewness is negative.

SECTION 3

Planning and Conducting Empirical Studies

SUBSECTION 3.1

Empirical Studies

An empirical study is one which is carried out to learn about a population or process by collecting data. We use the 5 steps:

1. Problem: a clear statement of the study's objectives, usually involving one or more questions
2. Plan: the procedures used to carry out the study including how the data will be collected
3. Data: the physical collection of the data, as described in the Plan

4. Analysis: the analysis of the data collected in light of the Problem and the Plan
5. Conclusion: the conclusions that are drawn about the Problem and their limitations

Subsection 3.2

Steps of PPDAC

We have 3 types of problems:

- Descriptive: the problem is to determine a particular attribute of a population or process
- Causative: the problem is to determine the existence or non-existence of a causal relationship between two variates
- Predictive: the problem is to predict a future value for a variate of a unit to be selected from the process or population

Definition 17 (Target Population) The target population or target process is the collection of units to which the experimenters conducting the empirical study wish the conclusions to apply

Definition 18 (Study Population) The study population or study process is the collection of units available to be included in the study.

The study population is often but not always a subset of the target population.

Definition 19 (Study Error) If the attributes in the study population/process differ from the attributes in the target population/process, then the difference is called study error.

Definition 20 (Sampling Protocol) The sampling protocol is the procedure used to select a sample of units from the study population/process. The number of units sampled is called the sample size.

Definition 21 (Sample Error) If the attributes in the sample differ from the attributes in the study population/process the difference is called sample error.

Definition 22 (Measurement Error) If the measured value and the true value of a variate are not identical the difference is called measurement error.

Section 4

Estimation

Subsection 4.1

Estimators and Sampling Distributions

This chapter focuses on uncertainty of estimations. For example, if we take two independent samples from a $G(0, 1)$ distribution, the first sample's mean will not exactly equal the second one's. This introduces the idea of an estimator.

Definition 23 (Estimator) A (point) estimator $\tilde{\theta}$ is a random variable which is a function $\tilde{\theta} = g(Y_1, ..., Y_n)$ of the random variables $Y_1, ..., Y_n$. The distribution of $\tilde{\theta}$ is called the sampling distribution of the estimator.

Subsection 4.2

Interval Estimation Using Likelihood Function

Definition 24 (Likelihood Interval) A $100p\%$ likelihood interval for θ is the set $\{\theta : R(\theta) \geq p\}$.

A likelihood interval is of the form $[L(\mathbf{y}), U(\mathbf{y})]$ in light of the observed data $\mathbf{y}$. $L(\mathbf{y})$ and $U(\mathbf{y})$ are the two solutions of the equation $R(\theta) - p = 0$ with $L(\mathbf{y}) \leq U(\mathbf{y})$.

Table 4.2

Guidelines for Interpreting Likelihood Intervals

Values of θ inside a 50% likelihood interval are very plausible in light of the observed data.
Values of θ inside a 10% likelihood interval are plausible in light of the observed data.
Values of θ outside a 10% likelihood interval are implausible in light of the observed data.
Values of θ outside a 1% likelihood interval are very implausible in light of the observed data.

Note that a likelihood interval does not guarantee to contain the true value of θ.

Subsection 4.3

Confidence Intervals and Pivotal Quantities

Definition 25 (Confidence Interval) Suppose the interval estimator $[L(\mathbf{Y}), U(\mathbf{Y})]$ has the property that

$$P(\theta \in [L(\mathbf{Y}), U(\mathbf{Y})]) = P(L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y}) = p.$$

Suppose the interval estimate $[L(\mathbf{y}), U(\mathbf{y})]$ is constructed for the parameter θ based on observed data $\mathbf{y}$. The interval estimate $[L(\mathbf{y}), U(\mathbf{y})]$ is called a $100p\%$ confidence interval for θ and p is called the confidence coefficient.

The idea of a confidence interval is grounded in repeated sampling: if we calculated the interval estimate for each sample for a large number of samples of data, we would expect approximately $100p\%$ of those intervals to contain the true value of θ. Given a single sample, we cannot conclude a probability of our interval estimate containing θ.

Definition 26 (Pivotal Quantity) A pivotal quantity $Q = Q(\mathbf{Y}; \theta)$ is a function of the data $\mathbf{Y}$ and the unknown parameter θ such that the distribution of the random variable Q is completely known. That is, statements like $P(Q \leq a)$ and $P(Q \geq b)$ depend on a and b but not on θ or any unknown information.

Confidence interval for μ in $G(\mu, \sigma)$ with known σ

Suppose $\mathbf{Y} = (Y_1, ..., Y_n)$ is a random sample from $G(\mu, \sigma)$. Since

$$Q = Q(\mathbf{Y}; \mu) = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim G(0, 1)$$

is a pivotal quantity, we can find a $100p\%$ confidence interval for μ by finding a value a such that $P(-a \leq Q \leq a) = p$ where $Q \in G(0, 1)$. Equivalently, $P(Q \leq a) = \frac{1+p}{2}$. We can then rearrange $P(-a \leq Q \leq a)$ to arrive at interval $\bar{y} \pm a\frac{\sigma}{\sqrt{n}}$ which is our confidence interval.

Theorem 1 (Asymptotic Gaussian Pivotal Quantities) Suppose $\tilde{\theta}$ is a point estimator for the unknown parameter θ, and that Central Limit Theorem can be used to obtain

$$\frac{\tilde{\theta} - \theta}{g(\theta)/\sqrt{n}} \sim G(0, 1)$$

for large n where $E[\tilde{\theta}] = \theta$ and $\text{Var}[\tilde{\theta}] = (g(\theta)/\sqrt{n})^2$. Then,

$$\frac{\tilde{\theta} - \theta}{g(\tilde{\theta})/\sqrt{n}} \sim G(0, 1)$$

Approximate Confidence Interval for Bin(n, θ)

We know that in a Bin(n, θ) distribution, the MLE of $\tilde{\theta}$ is $\frac{Y}{n}$, $E(\tilde{\theta}) = \theta$, and $sd(\tilde{\theta}) = \sqrt{\frac{\theta(1-\theta)}{n}}$. Then, by the previous theorem and Central Limit Theorem,

$$Q_n = Q_n(Y; \theta) = \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \sim G(0, 1)$$

is an asymptotic Gaussian pivotal quantity for large n. We can then rearrange

$$P\left(-a \leq \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \leq a\right) = p$$

into an $100p\%$ confidence interval $\frac{y}{n} \pm a\sqrt{\frac{y/n(1-y/n)}{n}}$ where a is a value such that $P(Z \leq a) = \frac{1+p}{2}$ where $Z \sim G(0, 1)$.

Interval Width: If we need to solve for a sample size n such that the confidence interval is of width $\leq 2l$, we simply need $2a\sqrt{\frac{y/n(1-y/n)}{n}} \leq 2l$. We can then rearrange for n to arrive at

$$n \geq \left(\frac{a(0.5)}{l}\right)^2$$

as $\hat{\theta}(1 - \hat{\theta})$ is maximized at 0.5 where $\hat{\theta} = y/n$.

Subsection 4.4

Chi-Squared and t Distributions

Definition 27 (Gamma Function) The gamma function is defined as

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$$

for $\alpha > 0$. Properties include:

1. $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$
2. $\Gamma(\alpha) = (\alpha - 1)!$ for $\alpha = 1, 2, \ldots$
3. $\Gamma(1/2) = \sqrt{\pi}$

Definition 28 (Chi-Squared Distribution) The $\chi^2(k)$ distribution is a continuous family of distributions on $(0, \infty)$ with pdf

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2}$$

for $x > 0$ where $k \in \{1, 2, ...\}$ is a parameter of the distribution called the degree of freedom.

For $k = 2$, the pdf is the Exp(2) pdf. For $k > 2$, the pdf is unimodal with maximum at $x = k - 2$. For values $k \geq 30$, the pdf resembles that of $N(k, 2k)$.

Theorem 2 Let $W_1, ..., W_n \sim \chi^2(k_i)$ be independent random variables. Then

$$S = \sum_{i=1}^{n} W_i \sim \chi^2\left(\sum_{i=1}^{n} k_i\right).$$

Theorem 3 If $Z \sim G(0,1)$, then the distribution of $W = Z^2$ is $\chi^2(1)$.

Theorem 4 If $Z_1, ..., Z_n$ are mutually independent $G(0,1)$ random variables, and $S = \sum_{i=1}^{n} Z_i^2$, then $S \sim \chi^2(n)$.

The following are some other useful results:

- If $W \sim \chi^2(1)$, then $P(W \geq w) = 2(1 - P(Z \leq \sqrt{w}))$ where $Z \sim G(0,1)$.
- If $W \sim \chi^2(2)$, then $W \sim \text{Exp}(2)$ and $P(W \geq w) = e^{-w/2}$.

Definition 29 (Student's t distribution) The t distribution has pdf

$$f(t;k) = c_k \left(1 + \frac{t^2}{k}\right)^{-(k+1)/2}$$

for $t \in \mathbb{R}$ and $k = 1, 2, ...$ where constant c_k is given by

$$c_k = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(k/2)}.$$

The parameter k is called the degrees of freedom. The t pdf is similar to $G(0,1)$ in that it is unimodal and symmetric about the origin. For large k, it is more similar to $G(0,1)$. For small k, the t pdf has fatter/taller tails.

Theorem 5 Suppose $Z \sim G(0,1)$ and $U \sim \chi^2(k)$ independently. Then

$$T = \frac{Z}{\sqrt{U/k}}$$

has a student's t distribution with k degrees of freedom.

Subsection 4.5

Likelihood-Based Confidence Intervals

Theorem 6 If $L(\theta)$ is based on $\mathbf{Y} = (Y_1, ..., Y_n)$, a random sample of size n, and if θ is the true value of the scalar parameter, then (under mild mathematical conditions), the distribution of $\Lambda(\theta)$ converges to $\chi^2(1)$ as $n \to \infty$.

This theorem implies that $\Lambda(\theta)$ can be used as an approximate pivotal quantity for sufficiently large n.

Theorem 7 A $100p\%$ likelihood interval is an approximate $100q\%$ confidence interval where $q = 2P(Z \leq \sqrt{-2\log p}) - 1$ and $Z \sim G(0,1)$.

Proof We can write a $100p\%$ likelihood interval as

$$\{\theta : R(\theta) \geq p\} = \left\{\theta : -2\log\left(\frac{L(\theta)}{L(\hat{\theta})} \leq -2\log p\right)\right\}$$

We can then approximate

$$\begin{aligned} P(\Lambda(\theta) \leq -2\log p) &= P\left(-2\log\left(\frac{L(\theta)}{L(\hat{\theta})}\right) \leq -2\log p\right) \\ &\approx P(W \leq -2\log p) \text{ where } W \sim \chi^2(1) \\ &= P(|Z| \leq \sqrt{-2\log p} \text{ where } Z \sim G(0,1) \\ &= 2P(Z \leq \sqrt{-2\log p}) - 1 \end{aligned}$$

□

Theorem 8 If a is a value such that $p = 2P(Z \leq a) - 1$ where $Z \sim G(0,1)$, then the likelihood interval $\{\theta : R(\theta) \geq e^{-a^2/2}\}$ is an approximate $100p\%$ confidence interval.

Proof The confidence interval corresponding to $\{\theta : R(\theta) \geq e^{-a^2/2}\}$ is

$$\begin{aligned} P\left(\frac{L(\theta)}{L(\tilde{\theta})} \geq e^{-a^2/2}\right) &= P\left(-2\log\left(\frac{L(\theta)}{L(\tilde{\theta})}\right) \leq a^2\right) \\ &\approx P(W \leq a^2) \text{ where } W \sim \chi^2(1) \\ &= 2P(Z \leq a) - 1 \text{ where } Z \sim G(0,1) \\ &= p \end{aligned}$$

□

SUBSECTION 4.6

Confidence Intervals for Parameters in $G(\mu, \sigma)$

Theorem 9 Suppose $Y_1, ..., Y_n$ is a random sample from $G(\mu, \sigma)$ with sample mean $\bar{Y}$ and sample variance S^2. Then

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Confidence interval for μ in $G(\mu, \sigma)$ with unknown σ

We can use the pivotal quantity T from the previous theorem. Since the t distribution is symmetric, we can find a such that $P(-a \leq T \leq a)$ for a $100p\%$ confidence interval. This is equivalent to $P(T \leq a) = \frac{1+p}{2}$ where $T \sim t(n-1)$. Rearranging then gives the confidence interval as $\bar{y} \pm as/\sqrt{n}$.

Confidence interval for σ^2 and σ in $G(\mu, \sigma)$

Theorem 10 Suppose $Y_1, ..., Y_n$ is a random sample from the $G(\mu, \sigma)$ distribution with sample variance S^2.

$$U = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left(\frac{Y_i - \bar{Y}}{\sigma}\right)^2 \sim \chi^2(n-1)$$

To construct a $100p\%$ confidence interval for σ^2, we can then find $P(a \leq U \leq b)$ where $P(U \leq a) = P(U \geq b) = \frac{1-p}{2}$ and rearrange to arrive at

$$\left[\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right]$$

Table 4.3
Approximate Confidence Intervals for Named Distributions
based on Asymptotic Gaussian Pivotal Quantities

Named Distribution	Observed Data	Point Estimate $\hat{\theta}$	Point Estimator $\tilde{\theta}$	Asymptotic Gaussian Pivotal Quantity	Approximate $100p\%$ Confidence Interval
Binomial(n, θ)	y	$\frac{y}{n}$	$\frac{Y}{n}$	$\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}}$	$\hat{\theta} \pm a\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$
Poisson(θ)	$y_1, y_2, \ldots, y_n$	$\bar{y}$	$\overline{Y}$	$\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}}{n}}}$	$\hat{\theta} \pm a\sqrt{\frac{\hat{\theta}}{n}}$
Exponential(θ)	$y_1, y_2, \ldots, y_n$	$\bar{y}$	$\overline{Y}$	$\frac{\tilde{\theta}-\theta}{\frac{\tilde{\theta}}{\sqrt{n}}}$	$\hat{\theta} \pm a\frac{\hat{\theta}}{\sqrt{n}}$

Note: The value a is given by $P\,(Z \leq a) = \frac{1+p}{2}$ where $Z \sim G\,(0, 1)$. In R, $a = \texttt{qnorm}\left(\frac{1+p}{2}\right)$

Table 4.4
Confidence/Prediction Intervals for Gaussian and Exponential Models

Model	Unknown Quantity	Pivotal Quantity	100p% Confidence/Prediction Interval
$G(\mu,\sigma)$ σ known	μ	$\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim G(0,1)$	$\bar{y} \pm a\sigma/\sqrt{n}$
$G(\mu,\sigma)$ σ unknown	μ	$\frac{\overline{Y}-\mu}{S/\sqrt{n}} \sim t(n-1)$	$\bar{y} \pm bs/\sqrt{n}$
$G(\mu,\sigma)$ μ unknown σ unknown	Y	$\frac{Y-\overline{Y}}{S\sqrt{1+\frac{1}{n}}} \sim t(n-1)$	100p% Prediction Interval $\bar{y} \pm bs\sqrt{1+\frac{1}{n}}$
$G(\mu,\sigma)$ μ unknown	σ^2	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left[\frac{(n-1)s^2}{d}, \frac{(n-1)s^2}{c}\right]$
$G(\mu,\sigma)$ μ unknown	σ	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left[\sqrt{\frac{(n-1)s^2}{d}}, \sqrt{\frac{(n-1)s^2}{c}}\right]$
Exponential(θ)	θ	$\frac{2n\overline{Y}}{\theta} \sim \chi^2(2n)$	$\left[\frac{2n\bar{y}}{d_1}, \frac{2n\bar{y}}{c_1}\right]$

Notes: (1) The value a is given by $P(Z \leq a) = \frac{1+p}{2}$ where $Z \sim G(0,1)$.
In R, $a =$ `qnorm`$\left(\frac{1+p}{2}\right)$
(2) The value b is given by $P(T \leq b) = \frac{1+p}{2}$ where $T \sim t(n-1)$. In R, $b =$ `qt`$\left(\frac{1+p}{2}, n-1\right)$
(3) The values c and d are given by $P(W \leq c) = \frac{1-p}{2} = P(W > d)$ where $W \sim \chi^2(n-1)$.
In R, $c =$ `qchisq`$\left(\frac{1-p}{2}, n-1\right)$ and $d =$ `qchisq`$\left(\frac{1+p}{2}, n-1\right)$
(4) The values c_1 and d_1 are given by $P(W \leq c_1) = \frac{1-p}{2} = P(W > d_1)$ where $W \sim \chi^2(2n)$.
In R, $c_1 =$ `qchisq`$\left(\frac{1-p}{2}, 2n\right)$ and $d_1 =$ `qchisq`$\left(\frac{1+p}{2}, 2n\right)$

SECTION 5

Hypothesis Testing

SUBSECTION 5.1

Introduction

A test of hypothesis begins by specifying a default hypothesis, and then checking whether the collected data is unlikely under this hypothesis. The default hypothesis is referred to as the null hypothesis H_0, which is tested against an alternative hypothesis H_A.

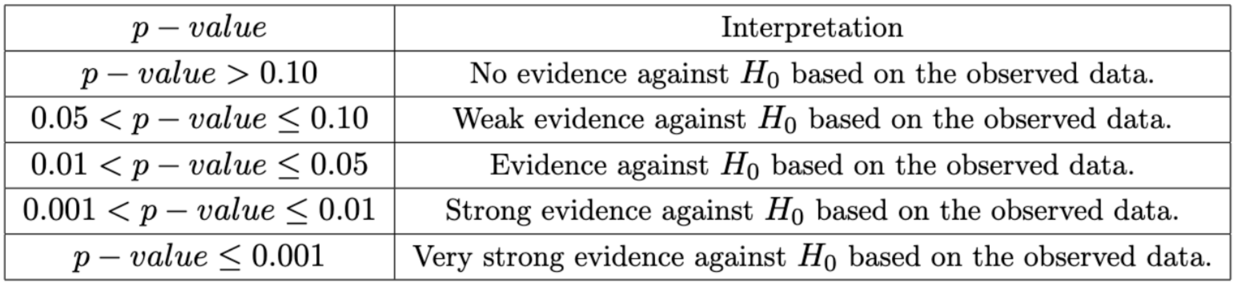
Definition 30 (Test Statistic) A test statistic or discrepancy measure D is a function of the data $\mathbf{Y}$ that is constructed to measure the degree of agreement between the data $\mathbf{Y}$ and the null hypothesis H_0.

We usually define D such that $D = 0$ describes closest alignment between data and H_0.

Definition 31 (p-value) Suppose we use the test statistic $D = D(\mathbf{Y})$ to test the hypothesis H_0. Suppose also that $d = D(\mathbf{y})$ is the observed value of D. The p-value or observed significance level of the test of hypothesis H_0 using test statistic D is

$$p\text{-value} = P(D \geq d; H_0)$$

The following are guidelines for interpreting p-value:



$p - value$	Interpretation
$p - value > 0.10$	No evidence against H_0 based on the observed data.
$0.05 < p - value \leq 0.10$	Weak evidence against H_0 based on the observed data.
$0.01 < p - value \leq 0.05$	Evidence against H_0 based on the observed data.
$0.001 < p - value \leq 0.01$	Strong evidence against H_0 based on the observed data.
$p - value \leq 0.001$	Very strong evidence against H_0 based on the observed data.

SUBSECTION 5.2

Testing $G(\mu,\sigma)$

Test of Hypothesis for μ

Given null hypothesis $H_0 : \mu = \mu_0$, we use test statistic

$$D = \frac{|\bar{Y} - \mu_0|}{S/\sqrt{n}}$$

to obtain the p-value knowing that

$$\frac{\bar{Y} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$$

if $H_0 : \mu = \mu_0$ is true. Let $d = \dfrac{|\bar{y} - \mu_0|}{s/\sqrt{n}}$, then

$$p\text{-value} = P(D \geq d; H_0) = P(|T| \geq d) = 2(1 - P(T \leq d))$$

Theorem 11 Given data $\mathbf{y}$ and model $f(\mathbf{y};\theta)$, suppose we use the same pivotal quantity to construct the (approximate) confidence interval for θ and to test the hypothesis $H_0 : \theta = \theta_0$. Then, the parameter value $\theta = \theta_0$ is an element of the $100q\%$ (approximate) confidence interval for θ if and only if the p-value for testing $H_0 : \theta = \theta_0$ is greater than or equal to $1 - q$.

Example. Suppose $y_1, ..., y_n$ is an observed sample from $G(\mu,\sigma)$. Suppose we test $H_0 : \mu = \mu_0$. Then,

$$\begin{aligned} p\text{-value} \geq 0.05 &\iff P(D \geq d; H_0) \geq 0.05 \\ &\iff P(|T| \geq d) \geq 0.05 \\ &\iff P(|T| \leq d) \leq 0.95 \\ &\iff \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}} \leq a \text{ where } P(|T| \leq a) = 0.95 \\ &\iff \mu_0 \in [\bar{y} - as/\sqrt{n}, \bar{y} + as/\sqrt{n}] \end{aligned}$$

Test of Hypothesis for σ

Given null hypothesis $H_0 : \sigma = \sigma_0$, we use test statistic

$$U = \frac{(n-1)S^2}{\sigma_0^2}$$

to obtain the p-value as such: let $u = (n-1)s^2/\sigma_0^2$ denote the observed value of U from the data. If u is large where $P(U \leq u) > 1/2$, then

$$p\text{-value} = 2P(U \geq u)$$

where $U \sim \chi^2(n-1)$. Otherwise, if $P(U \leq u) < 1/2$, then

$$p\text{-value} = 2P(U \leq u)$$

where $U \sim \chi^2(n-1)$. Note: only one of these two values will be < 1 for any given u, which is the desired p-value.

Subsection 5.3

Likelihood Ratio Test of Hypothesis

Likelihood values are useful in gauging the plausibility of parameters in light of observed data. To test a hypothesis $H_0 : \theta = \theta_0$, we can use the relative likelihood $R(\theta_0) = L(\theta_0)/L(\hat{\theta})$. If $R(\theta_0)$ is close to 1, then our hypothesis is plausible in light of observed data. To convert this into a p-value, we need a sampling distribution of $L(\theta_0)/L(\tilde{\theta})$, so we can use the likelihood ratio statistic

$$\Lambda(\theta_0) = -2\log\left(\frac{L(\theta_0)}{L(\tilde{\theta})}\right).$$

If H_0 is true, then $\Lambda(\theta_0)$ is approximately a $\chi^2(1)$ distribution. Smaller values of $R(\theta_0)$ correspond to larger values of $\Lambda(\theta_0)$ and vice-versa, so large values of $\Lambda(\theta_0)$ are evidence against H_0. To determine the p-value, we calculate the observed value of $\Lambda(\theta_0)$:

$$\lambda(\theta_0) = -2\log\left(\frac{L(\theta_0)}{L(\hat{\theta})}\right) = -2\log R(\theta_0)$$

The approximate p-value is then

$$p\text{-value} \approx P(W \geq \lambda(\theta_0)) = P(|Z| \geq \sqrt{\lambda(\theta_0)}) = 2(1 - P(Z \leq \sqrt{\lambda(\theta_0)}))$$

where $W \sim \chi^2(1)$ and $Z \sim G(0,1)$.

Table 5.2
Hypothesis Tests for Named Distributions
based on Asymptotic Gaussian Pivotal Quantities

Named Distribution	Point Estimate $\hat{\theta}$	Point Estimator $\tilde{\theta}$	Test Statistic for $H_0 : \theta = \theta_0$	Approximate $p - value$ based on Gaussian approximation
Binomial(n,θ)	$\frac{y}{n}$	$\frac{Y}{n}$	$\frac{\lvert\tilde{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}$	$2P\left(Z \geq \frac{\lvert\hat{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}\right)$ $Z \sim G(0,1)$
Poisson(θ)	$\bar{y}$	$\overline{Y}$	$\frac{\lvert\tilde{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0}{n}}}$	$2P\left(Z \geq \frac{\lvert\hat{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0}{n}}}\right)$ $Z \sim G(0,1)$
Exponential(θ)	$\bar{y}$	$\overline{Y}$	$\frac{\lvert\tilde{\theta}-\theta_0\rvert}{\frac{\theta_0}{\sqrt{n}}}$	$2P\left(Z \geq \frac{\lvert\hat{\theta}-\theta_0\rvert}{\frac{\theta_0}{\sqrt{n}}}\right)$ $Z \sim G(0,1)$

Note: To find $2P(Z \geq d)$ where $Z \sim G(0,1)$ in R, use $2 * (1-$ `pnorm`$(d))$

Table 5.3
Hypothesis Tests for Gaussian and Exponential Models

Model	Hypothesis	Test Statistic	Exact $p-value$
$G(\mu,\sigma)$ σ known	$H_0:\mu=\mu_0$	$\frac{\lvert\overline{Y}-\mu_0\rvert}{\sigma/\sqrt{n}}$	$2P\left(Z\geq\frac{\lvert\bar{y}-\mu_0\rvert}{\sigma/\sqrt{n}}\right)$ $Z\sim G(0,1)$
$G(\mu,\sigma)$ σ unknown	$H_0:\mu=\mu_0$	$\frac{\lvert\overline{Y}-\mu_0\rvert}{S/\sqrt{n}}$	$2P\left(T\geq\frac{\lvert\bar{y}-\mu_0\rvert}{s/\sqrt{n}}\right)$ $T\sim t(n-1)$
$G(\mu,\sigma)$ μ unknown	$H_0:\sigma=\sigma_0$	$\frac{(n-1)S^2}{\sigma_0^2}$	$\min(2P\left(W\leq\frac{(n-1)s^2}{\sigma_0^2}\right),$ $2P\left(W\geq\frac{(n-1)s^2}{\sigma_0^2}\right))$ $W\sim\chi^2(n-1)$
Exponential(θ)	$H_0:\theta=\theta_0$	$\frac{2n\bar{Y}}{\theta_0}$	$\min(2P\left(W\leq\frac{2n\bar{y}}{\theta_0}\right),$ $2P\left(W\geq\frac{2n\bar{y}}{\theta_0}\right))$ $W\sim\chi^2(2n)$

Notes:

(1) To find $P(Z\geq d)$ where $Z\sim G(0,1)$ in R, use $1-$ `pnorm`(d)

(2) To find $P(T\geq d)$ where $T\sim t(k)$ in R, use $1-$ `pt`(d,k)

(3) To find $P(W\leq d)$ where $W\sim\chi^2(k)$ in R, use `pchisq`(d,k)

Section 6

Gaussian Response Models

Subsection 6.1

Introduction

Definition 32 (Gaussian Response Model) A Gaussian response model is one for which the distribution of the response variate Y, given the associated vector of covariates $\mathbf{x} = (x_1, ..., x_k)$ for an individual unit, is of the form

$$Y \sim G(\mu(\mathbf{x}), \sigma(\mathbf{x})).$$

If observations are made on n randomly selected units, we write the model as

$$Y_i \sim G(\mu(\mathbf{x_i}), \sigma(\mathbf{x_i}))$$

for $i = 1, ..., n$ independently. We usually assume $\sigma(\mathbf{x}_i) = \sigma$ as a constant. The choice of $\mu(\mathbf{x})$ is guided by past information and current data. We usually assume $\mu(\mathbf{x_i})$ is a linear function of the covariates $\mathbf{x}_i$, where

$$\mu(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

where $\mathbf{x}_i = (x_{i1}, ..., x_{ik})$ is a vector of known covariates associated with unit i and $\beta_1, ..., \beta_k$ are unknown parameters. These models are called linear regression models and the β_j's are regression coefficients. Sometimes, the model is written as $Y_i = \mu(\mathbf{x}_i) + R_i$ where $R_i \sim G(0, \sigma)$ is called a stochastic component.

To introduce model parameter estimation, suppose we have $Y \sim G(\mu, \sigma)$. We can write this model in the form $Y_i \sim \mu + R_i$ where $R_i \sim G(0, \sigma)$ given we have n observations in our sample $Y_1, ..., Y_n$. We can show that $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the "least squares estimator" as $\bar{Y}$ is closer to the data than any other constant:

$$\min_{\mu} \sum_{i=1}^{n} (Y_i - \mu)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

Subsection 6.2

Simple Linear Regression

From the previous section, recall the model $Y_i \sim G(\mu(x_i), \sigma)$ for independent Y_i's where $\mu(x_i) = \alpha + \beta x_i$. Therefore, the unknowin parameters are (α, β, σ) and we can use the likelihood function

$$L(\alpha, \beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right) = \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right)$$

The log-likelihood function is

$$l(\alpha, \beta, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

where $\alpha, \beta \in \mathbb{R}$ and $\sigma > 0$. Taking the partial derivatives

$$\begin{aligned}
\frac{\partial l}{\partial \alpha} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0 \\
\frac{\partial l}{\partial \beta} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)x_i = 0 \\
\frac{\partial l}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = 0
\end{aligned}$$

we get

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} = \frac{S_{xy}}{S_{xx}} \\
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n}(S_{yy} - \hat{\beta}S_{xy})
\end{aligned}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$, and $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

Least Squares

From the previous section, we noticed that to find the fitted line $y = \alpha + \beta x$ using the least squares estimate, we need to minimize the function

$$g(\alpha, \beta) = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2.$$

To do this, we can also take partial derivatives

$$\begin{aligned}
\frac{\partial g}{\partial \alpha} &= \sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0 \\
\frac{\partial g}{\partial \beta} &= \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)x_i = 0
\end{aligned}$$

Upon closer observation, solving this system of equations will yield the same result as from the likelihood function MLE.

Theorem 12 The distribution of estimator $\tilde{\beta}$ is given as

$$\tilde{\beta} \sim G\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right).$$

Proof As previously seen, the maximum likelihood estimator of $\hat{\beta}$ is given as

$$\begin{aligned}\tilde{\beta} &= \frac{1}{S_{xx}}\sum_{i=1}^{n} x_i(Y_i - \bar{Y}) \\ &= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})Y_i \\ &= \sum_{i=1}^{n} a_i Y_i\end{aligned}$$

where $a_i = \frac{(x_i - \bar{x})}{S_{xx}}$. This shows that $\tilde{\beta}$ is a linear combination of Gaussian random variables Y_i and therefore has a Gaussian distribution. Using the identities

$$\sum_{i=1}^{n} a_i = 0, \sum_{i=1}^{n} a_i x_i = 1, \text{ and } \sum_{i=1}^{n} a_i^2 = \frac{1}{S_{xx}},$$

we can find

$$E(\tilde{\beta}) = \sum_{i=1}^{n} a_i E(Y_i) = \sum_{i=1}^{n} a_i(\alpha + \beta x_i) = \beta \sum_{i=1}^{n} a_i x_i = \beta$$

and

$$\text{Var}(\tilde{\beta}) = \sum_{i=1}^{n} a_i^2 \text{Var}(Y_i) = \sigma^2 \sum_{i=1}^{n} a_i^2 = \frac{\sigma^2}{S_{xx}}.$$

□

Definition 33 (Mean Squared Error)

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n-2}(S_{yy} - \hat{\beta}S_{xy})$$

where $\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ is the sum of squared error. We have $E[S_e^2] = \sigma^2$. s_e^2 is not the MLE of σ^2.

Theorem 13 A $100p\%$ confidence interval for $\mu(x)$ is given by

$$\left(\hat{\mu}(x) - as_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}, \hat{\mu}(x) + as_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)$$

where $\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x$ and $P(|T| \leq a) = p$ where $T \sim t(n-2)$.

Proof Consider the maximum likelihood estimator

$$\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x = \bar{Y} + \tilde{\beta}(x - \bar{x})$$

since $\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{x}$. Since

$$\tilde{\beta} = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} Y_i$$

we can rewrite this as

$$\tilde{\mu}(x) = \sum_{i=1}^{n} b_i Y_i$$

where $b_i = \frac{1}{n} + (x - \bar{x})\frac{(x_i - \bar{x})}{S_{xx}}$. We now see that $\tilde{\mu}(x)$ is a linear combination of Gaussian random variables and thus has a Gaussian distribution. Using the identities

$$\sum_{i=1}^{n} b_i = 1, \sum_{i=1}^{n} b_i x_i = x, \text{ and } \sum_{i=1}^{n} b_i^2 = \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}},$$

we can obtain

$$E(\tilde{\mu}(x)) = \sum_{i=1}^{n} b_i(\alpha + \beta x_i) = \alpha + \beta x = \mu(x)$$

and

$$\text{Var}(\tilde{\mu}(x)) = \sum_{i=1}^{n} b_i^2 \text{Var}(Y_i) = \sigma^2 \sum_{i=1}^{n} b_i^2 = \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right).$$

Standardizing, we can show that

$$\frac{\tilde{\mu}(x) - \mu(x)}{\sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim G(0,1)$$

and

$$\frac{\tilde{\mu}(x) - \mu(x)}{S_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

which is a pivotal quantity. We can then rearrange the expression $p = P(-a \leq T \leq a)$ to get the desired result. □

A $100p\%$ confidence interval for α is given by

$$\hat{\alpha} \pm a s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}$$

where $P(|T| \leq a) = p$ and $T \sim t(n-2)$.

PROOF Substitute $\alpha = \mu(0)$ in the previous theorem. □

Theorem 14 A $100p\%$ prediction interval for potential observation $Y = \mu_x + R$ where $R \sim G(0, \sigma)$ is given by

$$\hat{\alpha} + \hat{\beta}x \pm a s_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}$$

where $P(T \leq a) = (1+p)/2$ and $T \sim t_{n-2}$.

PROOF Y is a future observation and thus is also independent of $Y_1, ..., Y_n$. We want to find the distribution of $Y - \tilde{\mu}(x)$: the error in the point estimator of Y. Recall that

$$\tilde{\mu}(x) \sim G\left(\mu(x), \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right).$$

Then,

$$Y - \tilde{\mu}(x) = Y - \mu(x) + \mu(x) - \tilde{\mu}(x) = R + (\mu(x) - \tilde{\mu}(x)).$$

Since R is independent of $\tilde{\mu}(x)$, the above equation is the sum of independent and normally distributed Gaussian random variables and thus also follows a Gaussian distribution. We can see that

$$E(Y - \tilde{\mu}(x)) = E(R) + E(\mu(x)) - E(\tilde{\mu}(x)) = \mu(x) - \mu(x) = 0$$

and

$$\text{Var}(Y - \tilde{\mu}(x)) = \text{Var}(Y) + \text{Var}(\tilde{\mu}(x)) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right).$$

We now see that

$$\frac{Y - \tilde{\mu}(x)}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim G(0,1)$$

and

$$\frac{Y - \tilde{\mu}(x)}{S_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

Rearranging $p = P(-a \leq T \leq a)$ where $T \sim t(n-2)$ gives the result. □

We have explored the case with a single explanatory variate, and now we can look at a more general case. The Gaussian response model can be written as

$$Y_i = G(\mu(\mathbf{x}_i), \sigma)$$

for $i = 1, ..., n$ where $\mathbf{x}_i$ is a vector. We can also write it as

$$Y_i = \mu(\mathbf{x}_i) + R_i$$

where $R_i \sim G(0, \sigma)$ for $1 \leq i \leq n$. We can also say

$$E[Y_i] = \mu(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

where we seek to find $\beta_0, ..., \beta_k$ to minimize

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

We can use R to do this using `mod <- lm(y ~ x1 + x2)`, etc. and taking `summary(mod)`. The `Std. Error` column can be interpreted as follows: we can use $H_0 : \beta_j = 0$ for each parameter using test statistic

$$t_j = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}} = \frac{\texttt{Estimate}}{\texttt{Std. Error}}$$

where $T_j \sim t(n - k - 1)$. In the simple case (one covariate), we have $H_0 : \beta = 0$ where

$$t = \frac{\hat{\beta}}{s_e / \sqrt{S_{xx}}} = \frac{\texttt{Estimate}}{\texttt{Std. Error}}$$

and $T \sim t(n - 2)$.

Table 6.1
Confidence/Prediction Intervals for
Simple Linear Regression Model

Unknown Quantity	Estimate	Estimator	Pivotal Quantity	$100p\%$ Confidence/ Prediction Interval
β	$\hat{\beta} =$ $\frac{S_{xy}}{S_{xx}}$	$\tilde{\beta} =$ $\frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})Y_i}{S_{xx}}$	$\frac{\tilde{\beta}-\beta}{S_e/\sqrt{S_{xx}}}$ $\sim t\,(n-2)$	$\hat{\beta} \pm a s_e/\sqrt{S_{xx}}$
α	$\hat{\alpha} =$ $\bar{y} - \hat{\beta}\bar{x}$	$\tilde{\alpha} =$ $\overline{Y} - \tilde{\beta}\bar{x}$	$\frac{\tilde{\alpha}-\alpha}{S_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}$ $\sim t\,(n-2)$	$\hat{\alpha} \pm a s_e\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}$
$\mu\,(x) =$ $\alpha + \beta x$	$\hat{\mu}\,(x) =$ $\hat{\alpha} + \hat{\beta}x$	$\tilde{\mu}\,(x) =$ $\tilde{\alpha} + \tilde{\beta}x$	$\frac{\tilde{\mu}(x)-\mu(x)}{S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}}$ $\sim t\,(n-2)$	$\hat{\mu}\,(x) \pm a s_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$
σ^2	$s_e^2 =$ $\frac{S_{yy}-\hat{\beta}S_{xy}}{n-2}$	$S_e^2 =$ $\frac{\sum\limits_{i=1}^{n}\left(Y_i-\tilde{\alpha}-\tilde{\beta}x_i\right)^2}{n-2}$	$\frac{(n-2)S_e^2}{\sigma^2}$ $\sim \chi^2\,(n-2)$	$\left[\frac{(n-2)s_e^2}{c}, \frac{(n-2)s_e^2}{b}\right]$
Y	$\hat{Y} =$ $\hat{\alpha} + \hat{\beta}x$		$\frac{Y-\tilde{\mu}(x)}{S_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}}$ $\sim t\,(n-2)$	Prediction Interval $\hat{\mu}\,(x) \pm a s_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$

Notes: The value a is given by $P\,(T \leq a) = \frac{1+p}{2}$ where $T \sim t\,(n-2)$.
The values b and c are given by $P\,(W \leq b) = \frac{1-p}{2} = P\,(W > c)$ where $W \sim \chi^2\,(n-2)$.

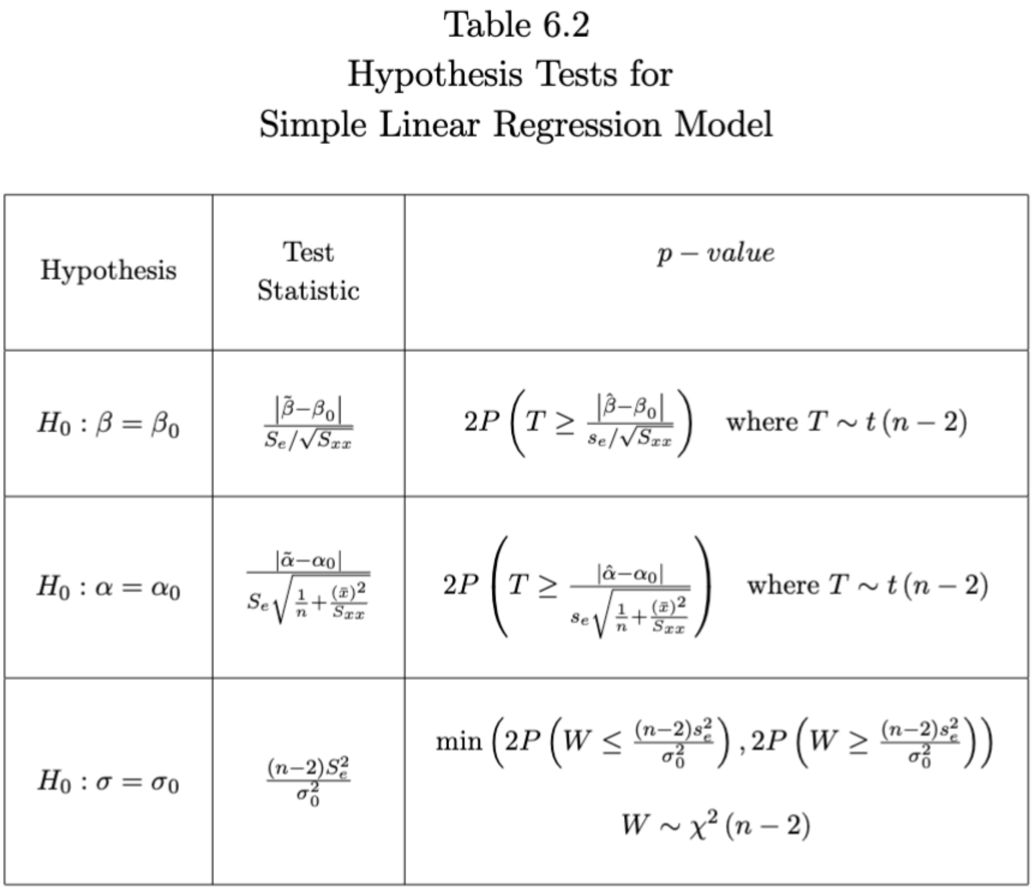


Table 6.2
Hypothesis Tests for
Simple Linear Regression Model

Hypothesis	Test Statistic	$p - value$
$H_0 : \beta = \beta_0$	$\frac{\lvert\tilde{\beta}-\beta_0\rvert}{S_e/\sqrt{S_{xx}}}$	$2P\left(T \geq \frac{\lvert\hat{\beta}-\beta_0\rvert}{s_e/\sqrt{S_{xx}}}\right)$ where $T \sim t\,(n-2)$
$H_0 : \alpha = \alpha_0$	$\frac{\lvert\tilde{\alpha}-\alpha_0\rvert}{S_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}$	$2P\left(T \geq \frac{\lvert\hat{\alpha}-\alpha_0\rvert}{s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}\right)$ where $T \sim t\,(n-2)$
$H_0 : \sigma = \sigma_0$	$\frac{(n-2)S_e^2}{\sigma_0^2}$	$\min\left(2P\left(W \leq \frac{(n-2)s_e^2}{\sigma_0^2}\right), 2P\left(W \geq \frac{(n-2)s_e^2}{\sigma_0^2}\right)\right)$ $W \sim \chi^2\,(n-2)$

Subsection 6.3

Checking the Model

When checking the fit of a linear regression model to data, we can check these assumptions we make:

- Y_i (given covariates x_i) has a Gaussian distribution
- That distribution has standard deviation σ which does not depend on the covariates
- $E(Y_i) = \mu(\mathbf{x}_i)$ is a linear combination of known covariates $\mathbf{x}_i = (x_{i1}, ..., x_{ik})$ and the unknown regression coefficients $\beta_0, \beta_1, ..., \beta_k$

Residual Plots

Let residuals be defined as the difference between observed responses y_i and fitted response $\hat{\mu}_i = \hat{\alpha} + \hat{\beta}x_i$ for each $1 \leq i \leq n$. We can see that the residuals $\hat{r}_i = y_i - \hat{\mu}_i$ behave roughly like a random sample from $G(0, \sigma)$ as

$$0 = \bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = \frac{1}{n}\hat{r}_i.$$

We can also say that the $\hat{r}_i$'s can be thought of as observed values of R_i in

$$Y_i = \mu_i + R_i$$

where $R_i \sim G(0, \sigma)$ for $i = 1, ..., n$ independently. We usually prefer to standardize residuals

$$\hat{r}_i^* = \frac{\hat{r}_i}{s_e} = \frac{y_i - \hat{\mu}_i}{s_e} = \frac{y_i - \hat{\alpha} - \hat{\beta}x_i}{s_e}$$

which roughly follow a random sample from $G(0, 1)$. In this case, the $\hat{r}_i^*$ values should almost entirely lie in the range $(-3, 3)$ as they are approximately $G(0, 1)$. We should expect the data to be scattered randomly in a horizontal band around the line $\hat{r}_i^* = 0$.

In cases where we have multiple covariates (eg. $\mu(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$), we can plot the residual plot as $(\hat{\mu}_i, \hat{r}_i^*)$ and our expectations should still apply. These checks all help us check the second and third assumptions at the beginning of this section.

Qqplot of Residuals

To check the first assumption where Y_i's have Gaussian distribution given covariates x_i, we can use a qqplot of standardized residuals. Since our assumed model

$$\frac{R_i}{\sigma} = \frac{Y_i - \mu_i}{\sigma} \sim G(0, 1),$$

then $\hat{r}_i^*$ should roughly represent a sample from $G(0, 1)$. Therefore, a qqplot of the r_i^* should give approximately a straight line if model assumptions hold.

We have to be careful about making predictions outside the range of our dataset. This is because our observed model may not extrapolate, and we will not be able to check our assumptions.

Subsection 6.4

Comparison of Two Population Means

Let $Y_{11}, ..., Y_{1n_1}$ be an independent random sample from $G(\mu_1, \sigma)$ and $Y_{21}, ..., Y_{2n_2}$ be an independent random sample from $G(\mu_2, \sigma)$. We will derive estimates of μ_1, μ_2, and σ:

$$L(\mu_1, \mu_2, \sigma) = \prod_{j=1}^{2} \prod_{i=1}^{n_j} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(y_{ji} - \mu_j)^2\right)$$

for $\mu_1, \mu_2 \in \mathbb{R}$ and $\sigma > 0$. Maximizing this function gives

$$\begin{aligned}\hat{\mu}_1 &= \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i} = \bar{y}_1 \\ \hat{\mu}_2 &= \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2i} = \bar{y}_2 \\ \hat{\sigma}^2 &= \frac{1}{n_1 + n_2}\left(\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2\right)\end{aligned}$$

Definition 34 (Pooled Estimate of Variance) The pooled estimate of variance is

$$\begin{aligned}s_p^2 &= \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2\right) \\ &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{n_1 + n_2}{n_1 + n_2 - 2}\hat{\sigma}^2\end{aligned}$$

where $s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2$ and $s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2$.

We can show that $E(S_p^2) = \sigma^2$. s_p^2 can also be seen as the weighted average of the two sample variances with weights $w_j = n_j - 1$ for $j = 1, 2$.

Theorem 15 A $100p\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\bar{y}_1 - \bar{y}_2 \pm a s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $P(T \leq a) = (1 + p)/2$ for $T \sim t_{n_1 + n_2 - 2}$.

Proof We can see that $E(\bar{Y}_1 - \bar{Y}_2) = \mu_1 - \mu_2$ and that $\text{Var}(\bar{Y}_1 - \bar{Y}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$. We then have

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim G(0, 1)$$

and

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$

so that

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

Rearranging this pivotal quantity for $P(-a \leq T \leq a)$ where $T \sim t(n_1 + n_2 - 2)$, we arrive at the result. □

Hypothesis Test for $\mu_1 = \mu_2$

To do this, we define $H_0 : \mu_1 - \mu_2 = 0$. Then, we can use test statistic

$$D = \frac{|\bar{Y}_1 - \bar{Y}_2 - 0|}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with

$$p\text{-value} = P(|T| \geq d) = 2(1 - P(T \leq d))$$

where $T \sim t(n_1 + n_2 - 2)$.

Unequal Variances $\sigma_1 \neq \sigma_2$

Another common situation is where we cannot make the assumption that variances are equal. We can use the approximate pivotal quantity

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim G(0, 1).$$

Therefore, a $100p\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\bar{y}_1 - \bar{y}_2 \pm a\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $P(Z \leq a) = (1 + p)/2$ for $Z \sim G(0, 1)$.

Table 6.3
Confidence Intervals for Two Sample Gaussian Model

Model	Parameter	Pivotal Quantity	$100p\%$ Confidence Interval
$G(\mu_1, \sigma_1)$ $G(\mu_2, \sigma_2)$ σ_1, σ_2 known	$\mu_1 - \mu_2$	$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ $\sim G(0,1)$	$\bar{y}_1 - \bar{y}_2 \pm a\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$G(\mu_1, \sigma_1)$ $G(\mu_2, \sigma_2)$ $\sigma_1 = \sigma_2 = \sigma$ σ unknown	$\mu_1 - \mu_2$	$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $\sim t(n_1 + n_2 - 2)$	$\bar{y}_1 - \bar{y}_2 \pm b s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
$G(\mu_1, \sigma)$ $G(\mu_2, \sigma)$ μ_1, μ_2 unknown	σ^2	$\frac{(n_1+n_2-2)S_p^2}{\sigma^2}$ $\sim \chi^2(n_1 + n_2 - 2)$	$\left[\frac{(n_1+n_2-2)s_p^2}{d}, \frac{(n_1+n_2-2)s_p^2}{c}\right]$
$G(\mu_1, \sigma_1)$ $G(\mu_2, \sigma_2)$ $\sigma_1 \neq \sigma_2$ σ_1, σ_2 unknown	$\mu_1 - \mu_2$	asymptotic Gaussian pivotal quantity $\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ for large n_1, n_2	approximate $100p\%$ confidence interval $\bar{y}_1 - \bar{y}_2 \pm a\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Notes:

The value a is given by $P(Z \leq a) = \frac{1+p}{2}$ where $Z \sim G(0,1)$.
The value b is given by $P(T \leq b) = \frac{1+p}{2}$ where $T \sim t(n_1 + n_2 - 2)$.
The values c and d are given by $P(W \leq c) = \frac{1-p}{2} = P(W > d)$ where $W \sim \chi^2(n_1 + n_2 - 2)$.

Table 6.4
Hypothesis Tests for
Two Sample Gaussian Model

Model	Hypothesis	Test Statistic	$p-value$
$G(\mu_1,\sigma_1)$ $G(\mu_2,\sigma_2)$ σ_1, σ_2 known	$H_0:\mu_1=\mu_2$	$\frac{\left\lvert\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)\right\rvert}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$2P\left(Z\geq\frac{\lvert\bar{y}_1-\bar{y}_2-(\mu_1-\mu_2)\rvert}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right)$ $Z\sim G(0,1)$
$G(\mu_1,\sigma)$ $G(\mu_2,\sigma)$ σ unknown	$H_0:\mu_1=\mu_2$	$\frac{\left\lvert\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)\right\rvert}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$2P\left(T\geq\frac{\lvert\bar{y}_1-\bar{y}_2-(\mu_1-\mu_2)\rvert}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)$ $T\sim t(n_1+n_2-2)$
$G(\mu_1,\sigma)$ $G(\mu_2,\sigma)$ μ_1, μ_2 unknown	$H_0:\sigma=\sigma_0$	$\frac{(n_1+n_2-2)S_p^2}{\sigma_0^2}$	$\min(2P\left(W\leq\frac{(n_1+n_2-2)s_p^2}{\sigma_0^2}\right),$ $2P\left(W\geq\frac{(n_1+n_2-2)s_p^2}{\sigma_0^2}\right))$ $W\sim\chi^2(n_1+n_2-2)$
$G(\mu_1,\sigma_1)$ $G(\mu_2,\sigma_2)$ $\sigma_1\neq\sigma_2$ σ_1, σ_2 unknown	$H_0:\mu_1=\mu_2$	$\frac{\left\lvert\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)\right\rvert}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$	approximate $p-value$ $2P\left(Z\geq\frac{\lvert\bar{y}_1-\bar{y}_2-(\mu_1-\mu_2)\rvert}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\right)$ $Z\sim G(0,1)$

Paired Data

Suppose that Y_{1i} and Y_{2i} are not independent and that we "pair" the data for each i assuming they are equal length. We would see that $E(\bar{Y}_1 - \bar{Y}_2) = \mu_1 - \mu_2$ but

$$\mathrm{Var}(\bar{Y}_1 - \bar{Y}_2) = \mathrm{Var}(\bar{Y}_1) + \mathrm{Var}(\bar{Y}_2) - 2\mathrm{Cov}(\bar{Y}_1, \bar{Y}_2) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2\mathrm{Cov}(\bar{Y}_1, \bar{Y}_2).$$

If $\mathrm{Cov}(\bar{Y}_1, \bar{Y}_2) > 0$, then the variance is smaller than for an unpaired experiment. To make inferences about $\mu_1 - \mu_2$, we analyze within pair differences

$$Y_i = Y_i - Y_{1i} - Y_{2i}$$

for $i = 1, ..., n$ by assuming each $Y_i \sim G(\mu_1 - \mu_2, \sigma)$ independently. Then, we can test $H_0 : \mu = \mu_1 - \mu_2 = 0$ using test statistic

$$D = \frac{|\bar{Y} - 0|}{S/\sqrt{n}}$$

with p-value $= 2(1 - P(T \leq d))$ where $T \sim t(n-1)$.

Subsection 6.5

R^2

Another statistic we can use to check model adequacy is the R^2 statistic. The formula is given as

$$R^2 = 1 - \frac{SSE}{S_{yy}} = \frac{S_{yy} - SSE}{S_{yy}} = \frac{\text{Variation explained by regression model}}{\text{Total variation}}$$

where $R^2 = 0$ means the regression explains none of the variation in our response, and $R^2 = 1$ means the regression perfectly explains all variation in our response. $R^2 \in [0, 1]$.

We can also explore the relationship between sample correlation r and R^2. Using the R^2 formula above and $SSE = S_{yy} - \hat{\beta} S_{xy}$, we can find

$$R^2 = \hat{\beta}\frac{S_{xy}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

whereas sample correlation is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

In terms of the R programming language, the `Multiple R-squared` value in `summary(mod)` is what was above. As for `Adjusted R-squared`:

$$\text{Adjusted } R^2 = 1 - \frac{SSE/(n-k-1)}{S_{yy}/(n-1)}$$

where k is the number of explanatory variates in the model. Adding new explanatory variates will always "improve" our R^2 value even if the new variate is unrelated: this is called overfitting. Adjusted R^2 attempts to compensate for this fact.

Another section of the R output is the `F-statistic`. For a simple linear regression model, this hypothesis tests $H_0 : \beta_1 = 0$ where $y = \beta_0 + \beta_1 x_1$. For a multiple linear regression model like $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, it is a hypothesis test for $H_0 : \beta_1 = \beta_2 = 0$.

The value given in `Residual standard error` is equal to s_e, the square root of the mean squared error.

Section 7

Multinomial Models and Goodness of Fit Tests

Subsection 7.1

Multinomial Likelihood Function

Suppose that data arise from the multinomial distribution with joint probability function

$$f(y_1, ..., y_k; \theta_1, ..., \theta_k) = \frac{n!}{y_1!...y_k!}\theta_1^{y_1}...\theta_k^{y_k}$$

where $\sum_{j=1}^k y_j = n$, $0 \leq \theta_j \leq 1$, and $\sum_{j=1}^k \theta_j = 1$. The likelihood function is

$$L(\theta) = \prod_{j=1}^k \theta_j^{y_j}.$$

It can be shown that the MLE of θ is $\hat{\theta}_j = y_j/n$ for each $1 \leq j \leq k$. Note that due to the $\sum_{j=1}^k \theta_j = 1$ constraint, there are only really $k-1$ parameters to estimate rather than k.

Example: Test the hypothesis $H_0 : \theta = \theta_0 = (\frac{1}{k}, ..., \frac{1}{k})$.

We have $L(\theta_0) = \prod_{j=1}^k (\frac{1}{k})^{Y_j}$ and $L(\tilde{\theta}) = \prod_{j=1}^k (\frac{Y_j}{n})^{Y_j}$. Using the likelihood ratio statistic,

$$\Lambda(\theta_0) = -2\log\left(\frac{L(\theta_0)}{L(\tilde{\theta})}\right)$$

alongside

$$\frac{L(\theta_0)}{L(\tilde{\theta})} = \prod_{j=1}^k \left(\frac{n/k}{Y_j}\right)^{Y_j} = \prod_{j=1}^k \left(\frac{E_j}{Y_j}\right)^{Y_j}$$

where Y_j is the frequency observed in category j (which should be n/k if H_0 is true), we write $E_j = n/k$ and simplify

$$\Lambda(\theta_0) = 2\sum_{j=1}^k Y_j \log\left(\frac{Y_j}{E_j}\right)$$

with observed value

$$\lambda(\theta_0) = 2\sum_{j=1}^k y_j \log\left(\frac{y_j}{e_j}\right).$$

When $y_j = e_j$, the statistic isn't affected. When $y_j > e_j$, category j increases the statistic. When $y_j < e_j$, category j decreases the statistic. Categories are not independent in this case: if one category is $y_j > e_j$, another must be $y_j < e_j$.

Theorem 16 If n is large and H_0 is true, then

$$\Lambda(\theta_0) = 2\sum_{j=1}^k Y_j \log\left(\frac{Y_j}{E_j}\right) \sim \chi^2(k-1-p)$$

where k is the number of categories and p is the number of parameters estimated in forming our null hypothesis.

In the above example, $p = 0$. The approximate p-value is therefore

$$p\text{-value} = P(W \geq \lambda(\theta_0))$$

where $W \sim \chi^2(k-1-p)$.

Definition 35 (Pearson Goodness of Fit Statistic) The Pearson goodness of fit statistic is

$$D = \sum_{j=1}^{k} \frac{(Y_j - E_j)^2}{E_j}$$

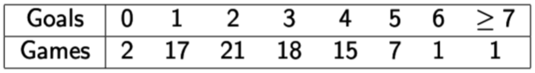
which is larger when expected and actual counts differ more. For large n, $D \sim \chi^2(k-1-p)$ approximately.

We would have p-value $= P(D \geq d)$ where $D \sim \chi^2(k-1-p)$.

Subsection 7.2

Goodness of Fit Tests

We can use goodness of fit tests to test whether or not assuming data follows a specific distribution is a good assumption. Given a frequency table of goals per hockey game and the number of games in which that number of goals were scored, we want to test H_0: the data follows a Poisson(θ) distribution.



Goals	0	1	2	3	4	5	6	≥ 7
Games	2	17	21	18	15	7	1	1

The hypothesis here will be:

$$H_0 : \begin{cases} \theta_j = \frac{\theta^j e^{-\theta}}{j!} & j = 0, ..., 6 \\ \theta_j = \sum_{y=7}^{\infty} \frac{\theta^y e^{-\theta}}{y!} & j = 7 \end{cases}$$

Once we calculate the expected values for each of the categories of goals, we can use the likelihood ratio test statistic in the previous section. To test H_0, we need to test unknown parameter θ so $p = 1$. In this case, we have 8 categories so $k - 1 = 7$. Overall, we use the $\chi^2(6)$ distribution.

Guideline: we often want to know what is "large enough" in a sample. Our guideline is to require $e_j \geq 5$ for all j: that is, the expected counts under the null hypothesis are at least 5 for each category of our dataset.

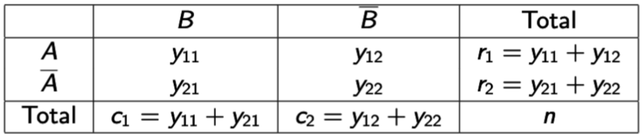
Once we compute the expected values for frequencies for each number of goals, we will see that 6 and 7 goals are less than 5. Here, we can simply collapse the 5, 6, and ≥ 7 categories into a single ≥ 5 category, and update the k value accordingly.

Subsection 7.3

Two-way Contingency Tables

Example:

Consider the following general two-way table:



	B	$\overline{B}$	Total
A	y_{11}	y_{12}	$r_1 = y_{11} + y_{12}$
$\overline{A}$	y_{21}	y_{22}	$r_2 = y_{21} + y_{22}$
Total	$c_1 = y_{11} + y_{21}$	$c_2 = y_{12} + y_{22}$	n

We define the 4 random variables Y_{11}, Y_{12}, Y_{21}, and Y_{22}. A suitable model is

$$Y_{11}, Y_{12}, Y_{21}, Y_{22} \sim \text{Multinomial}(n, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$$

where $\theta_{11} = P(A \cap B)$, etc. We want to test the null hypothesis that A and B are independent: $H_0 : P(A \cap B) = P(A)P(B)$. If we let $P(A) = \alpha$ and $P(B) = \beta$, we can write the null hypothesis as

$$H_0 : \theta_{11} = \alpha\beta.$$

We can also write other θ's such as $\theta_{12} = \alpha(1 - \beta)$, etc. The likelihood function is

$$L(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}) = \theta_{11}^{y_{11}} \theta_{12}^{y_{12}} \theta_{21}^{y_{21}} \theta_{22}^{y_{22}}$$

with MLE $\hat{\theta}_{ij} = \frac{y_{ij}}{n}$ for $i = 1, 2$ and $j = 1, 2$. Then,

$$L(\tilde{\theta}) = \left(\frac{Y_{11}}{n}\right)^{y_{11}} \left(\frac{Y_{12}}{n}\right)^{y_{12}} \left(\frac{Y_{21}}{n}\right)^{y_{21}} \left(\frac{Y_{22}}{n}\right)^{y_{22}}.$$

If H_0 is true, then

$$\begin{aligned} L(\alpha, \beta) &= (\alpha\beta)^{y_{11}} (\alpha(1-\beta))^{y_{12}} ((1-\alpha)\beta)^{y_{21}} ((1-\alpha)(1-\beta))^{y_{22}} \\ &= \alpha^{y_{11}+y_{12}} (1-\alpha)^{y_{21}+y_{22}} \beta^{y_{11}+y_{21}} (1-\beta)^{y_{12}+y_{22}} \end{aligned}$$

with random variable

$$L(\tilde{\alpha}, \tilde{\beta}) = \tilde{\alpha}^{Y_{11}+Y_{12}} (1-\tilde{\alpha})^{Y_{21}+Y_{22}} \tilde{\beta}^{Y_{11}+Y_{21}} (1-\tilde{\beta})^{Y_{12}+Y_{22}}.$$

We can now use the likelihood ratio statistic

$$\Lambda = -2 \log \left(\frac{L(\tilde{\alpha}, \tilde{\beta})}{L(\tilde{\theta})} \right)$$

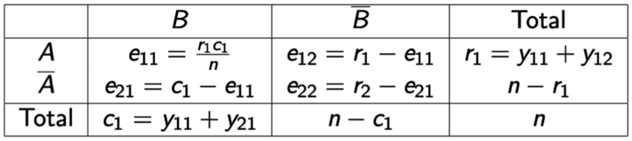
where

$$\tilde{\alpha} = \frac{Y_{11} + Y_{12}}{n} \text{ and } \tilde{\beta} = \frac{Y_{11} + Y_{21}}{n}$$

with the same MLEs. We can finally substitute all the previous steps to solve for Λ. We can also use the simpler

$$\Lambda = 2 \left(Y_{11} \log \left(\frac{Y_{11}}{E_{11}} \right) + Y_{12} \log \left(\frac{Y_{12}}{E_{12}} \right) + Y_{21} \log \left(\frac{Y_{21}}{E_{21}} \right) + Y_{22} \log \left(\frac{Y_{22}}{E_{22}} \right) \right)$$

where $E_{11} = n\tilde{\alpha}\tilde{\beta}$ with $e_{11} = n\hat{\alpha}\hat{\beta}$, etc. We can also use some "tricks" to shorten our time on this calculations:

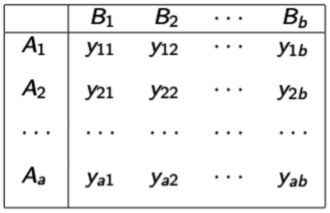


	B	$\overline{B}$	Total
A	$e_{11} = \frac{r_1 c_1}{n}$	$e_{12} = r_1 - e_{11}$	$r_1 = y_{11} + y_{12}$
$\overline{A}$	$e_{21} = c_1 - e_{11}$	$e_{22} = r_2 - e_{21}$	$n - r_1$
Total	$c_1 = y_{11} + y_{21}$	$n - c_1$	n

Now, when we use the chi-squared approximation of $\chi^2(k - 1 - p)$, we have $p = 2$ (α and β) and $k = 4$.

Example:

We now consider a generalized two-way table so that A has a categories $A_1, ..., A_a$ and B has b categories $B_1, ..., B_b$.



	B_1	B_2	$\cdots$	B_b
A_1	y_{11}	y_{12}	$\cdots$	y_{1b}
A_2	y_{21}	y_{22}	$\cdots$	y_{2b}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_a	y_{a1}	y_{a2}	$\cdots$	y_{ab}

Then, we have Y_{ij} to be the random variable representing number of units in category A_i and B_j in a random sample of size n. θ_{ij} is the probability a randomly selected unit is within category A_i and B_j. So,

$$(Y_{11}, ..., Y_{ab}) \sim \text{Multinomial}(n; \theta_{11}, ..., \theta_{ab}).$$

To test whether A and B are independent, we test

$$H_0 : \theta_{ij} = \alpha_i \beta_j$$

for $i = 1, ..., a$ and $j = 1, ..., b$. In general, we have

$$\hat{\alpha}_i = \frac{r_i}{n} \text{ and } \hat{\beta}_j = \frac{c_j}{n}$$

and expected frequency

$$e_{ij} = \frac{r_i c_j}{n} = n \times \frac{r_i}{n} \times \frac{c_j}{n} = n \times \hat{\alpha}_i \times \hat{\beta}_j$$

for each $i = 1, ..., a$ and $j = 1, ..., b$. The likelihood ratio test statistic here is

$$\Lambda = 2 \sum_{i=1}^{a} \sum_{j=1}^{b} Y_{ij} \log \left(\frac{Y_{ij}}{E_{ij}} \right)$$

where for large n, this has an approximate $\chi^2(k - 1 - p)$ distribution. The number of categories here is $k = a \times b$ and $p = (a - 1) + (b - 1)$. p is not equal to $a + b$ because of constraints such as $\sum_{i=1}^{a} \alpha_i = 1$ (so the last parameter isn't actually estimated). Simplifying, we have

$$k - 1 - p = (a - 1)(b - 1).$$

Section 8

Causal Relationships

This is a short chapter that explores statements of the form "X causes Y". Causation is difficult to define due to questions like "does smoking cause lung cancer". Does everyone who smokes get lung cancer?

Definition 36 (Causal Effect - First Definition) Let y be a response variate and let x be an explanatory variate associated with units in a population or process. Then, if all other factors that affect y are held constant, let us change x (or observe different values of x) and

> see if y changes. If it does, then x has a causal effect on y.

Although this can help us measure causation in certain cases, it is not the best definition because even a change in x might not correspond to a change in y for certain instances. So, we consider distributions instead.

Definition 37

> (Causal Effect - Improved Definition) x has a causal effect on Y if, when all other factors that affect Y are held constant, a change in x induces a change in a property of the distribution of Y.

In practice, it is hard to determine all factors affecting Y, let alone keeping them constant. We will look at 6 possible explanations for association between two variates:

1. The explanatory variate is the direct cause of the response variate.

   When you drink water when you are thirsty, you become less thirsty. This is a direct cause. However, even direct causes are not always great measures of relationships: buying a lottery ticket is a direct cause of winning the lottery even though chances are slim.

2. The reponse variate is the direct cause of the explanatory variate.

   For example, let an explanatory variate be gamer toxicity level and a response variate be proportion of matches won. We can argue this relationship both ways: toxicity causes decreased matches won, or decreased matches won causes toxicity.

3. The explanatory variate is a contributing, but not the only, cause of the response variate.

   Many factors cause cancer, but it's not a sole factor that causes it. It could be a necessary contributor to cancer, but still not the sole cause.

4. Both variates are changing over time.

   There is a strong correlation between global average temperature and number of pirates, but they do not have a causal relationship.

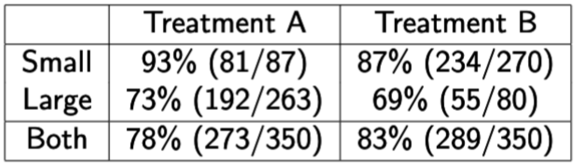
5. The association may be nothing more than coincidence.

   The odds of two people dying of brain cancer in the same office is low. But the low probability has to do with the number of office buildings in the city rather than causation.

6. Both variates mau result from a common cause.

   These are called confounding or lurking variates. For example, hot chocolate and sweater sales increase at the same time. They are not causes of each other, it is due to another variate.

Simpson's Paradox



	Treatment A	Treatment B
Small	93% (81/87)	87% (234/270)
Large	73% (192/263)	69% (55/80)
Both	78% (273/350)	83% (289/350)

In this table, treatment B seems to be a more effective treatment when considering treatment type and success rate. However, the lurking variate of size shows that treatment A is actually better in both groups. This is called an example of Simpson's paradox.

To prevent this, randomization of treatments may help. We see that a lot of patients in the small group get treatment B which skews the numbers for B. However, randomization isn't always possible: for example, in a study on smoking, you can't force people to smoke. Overall, we should consider:

1. The association between the two variates must be observed in many studies of different types among different groups. This reduces the chance that an observed association is due to a defect in one type of study or a peculiarity in one group of subjects

2. The association must continue to hold when the effects of plausible confounding variates are taken into account

3. There must be a plausible scientific explanation for the direct influence of one variate on the other variate, so that a causal link does not depend on the observed association alone

4. There must be a consistent response, that is, one variate always increases (decreases) as the other variate increases