STAT 331 Personal Notes

Marcus Chan

Taught by Peter Balka

UW CS '25

Chapter 1: Introduction

REGRESSION

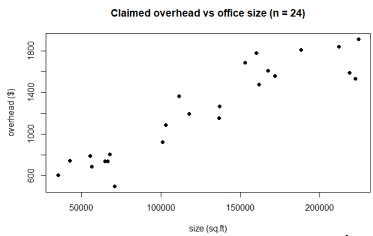
In regression modelling, we attempt to explain or account for variation in a response variate (y) by using a model to describe the relationship between y and one or more explanatory variates $(x_1, x_2, \dots)$

SUMMARIES OF THE DATA

1. A simple LR model involves:
   ① A single explanatory variate; &
   ② A single response variate.

eg Overhead data example:
response (y): claimed overhead (\$)
explanatory (x): office size (sq. ft)

2. We can summarise the data using a scatter-plot.



3. To get a numerical summary of the data, we can use the "sample correlation coefficient".

$$\boxed{r = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2 \sum (y_i-\bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}}$$

Note $-1 \le r \le 1$ and that r is unitless.

4. r tells us the relative strength of the linear relationship.

THE SIMPLE LR MODEL

1. We can describe the observed behavior of the response with a model that includes both
   ① a "deterministic component" that describes the variation in y accounted for by the functional form of the underlying relationship between y & x; &

   eg with the overhead data, the det. comp. is
   $$\mu = \beta_0 + \beta_1 x,$$
   where μ = the mean value of y for a given value of x.

   ② an "error term" ε that describes the random variation in y not accounted for by the underlying relationship with x.

2. Putting these together yields the "simple LR" (or SLR) model:

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

where
① β_0 = the "intercept" parameter
② β_1 = the "slope" parameter
③ i = the index that denotes the observation number.
($x_1, \dots, x_n$ is explanatory data; $y_1, \dots, y_n$ is response data).

*note $\beta_0 + \beta_1 x_i$ is deterministic & ε is random.

THE NORMAL SLR MODEL

1. We typically assume in SLR that

$$\boxed{\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, \dots, n}$$

for some variance σ^2.

2. This yields the normal model

$$\boxed{Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).}$$

3. Assumptions needed to use this model:
   ① the functional form (ie linear) of the relationship between y & x is correctly specified by the deterministic component of the model;
   ② errors follow a normal distribution;
   ③ errors have a constant variance σ^2 (ie "homoskedasticity"); &
   ④ errors are independent.

LEAST SQUARES ESTIMATION OF MODEL PARAMETERS

1. Goal: we want to find values of β_0 & β_1 such that for the data

$$\boxed{\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \varepsilon_n, \end{aligned}}$$

the sum of squares of the errors $\sum_i \varepsilon_i^2$ is minimized.

2. The values of β_0 & β_1 obtained by this procedure (denoted $\hat{\beta}_0$ & $\hat{\beta}_1$) are known as the "least squares estimates" of β_0 & β_1.

3. We show that

$$\boxed{\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}. \end{aligned}}$$

Proof. We wish to minimize

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2.$$

See that

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)].$$

Since we want to minimize S, we can solve

$$\begin{cases} \frac{\partial S}{\partial \beta_0} = 0 \\ \frac{\partial S}{\partial \beta_1} = 0. \end{cases}$$

The resultant solutions for β_0 & β_1 are the desired values as required. ☒

4. In R, we can get these values via

```
> data.slr.lm <- lm(response ~ explanatory).
```

FITTED MODEL

1. For the SLR model, the fitted model is

$$\boxed{\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x,}$$

where $\hat{\mu}$ is the estimated mean value of y given a value of x.

FITTED RESIDUALS

1. The "fitted residual" of the i^{th} observation, e_i, is defined as

$$\boxed{e_i = y_i - \hat{\mu}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).}$$

* ε_i is a random variable in which we impose assumptions;
  e_i is the difference between the observed response & estimated mean response.

2. If we take the partial derivative wrt each parameter and set $= 0$ in our least squares procedure, we get that

$$\boxed{\begin{aligned} \sum e_i &= 0 \\ \sum x_i e_i &= 0. \end{aligned}}$$

3. These constraints allow us to calculate the remaining 2 residuals from $n-2$ observations;
   so we say the fitted model is associated with $n-2$ degrees of freedom.

LEAST SQUARES ESTIMATE OF σ^2: $\hat{\sigma}^2$

In the normal model, we assume

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

In any least squares regression model, we estimate σ^2 by dividing the sum of squares of the residuals by the degrees of freedom.

In particular, this means our estimate for σ^2 is

$$\boxed{\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{n-2}.}$$

* note $E[\hat{\sigma}^2] = \sigma^2$ (ie $\hat{\sigma}^2$ is unbiased).

RESIDUAL STANDARD ERROR: $\hat{\sigma}$

The "residual standard error" is

$$\boxed{\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{\sum e_i^2}{n-2}}}$$

$\hat{\sigma}$ can be interpreted as the estimated std dev of the errors & measures the random variation of the response given a value for x.

The smaller $\hat{\sigma}$ is, the more the variation in y is "explained" by x, and so the better fit the model is.

$\hat{\sigma}$ is part of the summary R output for the fitted model:

```
> summary(audit.lm)
Call:
lm(formula = overhead ~ size)
Residuals:
   Min      1Q Median      3Q     Max
-36639 -12874  -1997    8642   56686
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -27877.06   14172.00  -1.967   0.0619 .
size           126.33      10.88  11.610 7.47e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 23480 on 22 degrees of freedom
Multiple R-squared:  0.8597,    Adjusted R-squared:  0.8533
F-statistic: 134.8 on 1 and 22 DF,  p-value: 7.472e-11
```

INTERPRETATION OF PARAMETER ESTIMATES

We may interpret $\hat{\beta}_1$ as the estimated mean change in the response y associated with a change of one unit in x.

We may interpret $\hat{\beta}_0$ as the estimated mean value of y at $x=0$ only if $x=0$ is a relevant value and is in the range of values we used to fit the model.

* never extrapolate to values of x outside the range used to fit the model.

Lastly, we can interpret $\hat{\sigma}$ as a measure of the variability of the response about the fitted line.

INFERENCE FOR β_1

To investigate whether there is a linear relationship between y & x in the population, we can test the hypothesis "$\beta_1 = 0$."

We can then either use confidence intervals or hypothesis tests to test this.

To do this, we need the least squares estimator of β_1:

$$\boxed{\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2}}$$

DISTRIBUTION OF $\hat{\beta}_1$

1. We can show for the SLR model that

$$\boxed{\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)}$$

Proof. First, note

$$\hat{\beta}_1 = \frac{\sum(x_i-\bar{x})(Y_i-\bar{Y})}{\sum(x_i-\bar{x})^2}$$

$$= \frac{\sum(x_i-\bar{x})Y_i - \bar{Y}\sum(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}$$

$$= \frac{\sum(x_i-\bar{x})Y_i}{\sum(x_i-\bar{x})^2} \qquad \because \sum(x_i-\bar{x})=0$$

$$= \sum c_i Y_i, \qquad c_i = \frac{x_i-\bar{x}}{\sum(x_i-\bar{x})^2}.$$

Then, for the SLR model, $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$.

Since $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, thus

$$Y_i \sim N(\beta_0+\beta_1 x_i, \sigma^2) \quad \& \quad Y_i \text{ are ind.}$$

and so

$$\hat{\beta}_1 = \sum c_i Y_i \sim \text{Normal} \quad \text{(by properties of normal).}$$

Then

$$E(\hat{\beta}_1) = E\left(\sum c_i Y_i\right) = \sum c_i E(Y_i)$$

$$= \sum \frac{x_i-\bar{x}}{\sum(x_i-\bar{x})^2}\cdot(\beta_0+\beta_1 x_i)$$

$$= \frac{\beta_0\sum(x_i-\bar{x}) + \beta_1\sum x_i(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}$$

$$= \frac{\beta_1\sum x_i(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}$$

$$= \frac{\beta_1\sum x_i(x_i-\bar{x}) - \beta_1\bar{x}\underbrace{\sum(x_i-\bar{x})}_{0}}{\sum(x_i-\bar{x})^2}$$

$$= \frac{\beta_1\sum(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2} = \beta_1.$$

Similarly,

$$Var(\hat{\beta}_1) = Var\left(\sum c_i Y_i\right)$$

$$= \sum c_i^2 Var(Y_i) \qquad \because Y_i\text{'s ind.}$$

$$= \sum \frac{(x_i-\bar{x})^2}{\left(\sum(x_i-\bar{x})^2\right)^2}\cdot\sigma^2$$

$$= \frac{\sigma^2}{\sum(x_i-\bar{x})^2} = \frac{\sigma^2}{S_{xx}}.$$

Hence $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$ as required.

2. It follows that

$$\boxed{\frac{\hat{\beta}_1-\beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1-\beta_1}{(\hat{\sigma}/\sqrt{S_{xx}})} \sim t_{n-2}.}$$

(from STAT231/330 result).

This can be used to get t-based CIs & hypothesis tests for β_1.

DISTRIBUTION OF $\hat{\beta}_0$

Similarly, we can show in a SLR model,

$$\boxed{\begin{gathered}\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)\right) \\ \frac{\hat{\beta}_0-\beta_0}{SE(\hat{\beta}_0)} = \frac{\hat{\beta}_0-\beta_0}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}\end{gathered}}$$

CI FOR β_1

1. A $(1-\alpha)100\%$ confidence interval for β_1 is

$$\boxed{\hat{\beta}_1 \pm t_{n-2,\,1-\alpha/2}\, SE(\hat{\beta}_1), \quad SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}.}$$

- $t_{n-2,\,1-\alpha/2}$:= the critical value from a t_{n-2} distribution corresponding to a confidence level of $(1-\alpha)100\%$.

2. "$t_{n-2,\,1-\alpha/2}\, SE(\hat{\beta}_1)$" is called the "margin of error" of the interval.

3. We can use R to calculate this:

```
> summary(data.lm)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27877.06	14172.00	-1.967	0.0619
size	126.33	10.88	11.610	7.47e-11

```
> t <- qt(1-α/2, n-2)
```

↳ The CI is then $(126.33 - t(10.88),\ 126.33 + t(10.88))$.

4. We may interpret the CI as that we are $(1-\alpha)100\%$ confident that for every additional increase of a unit of x, the mean increase of y is between (start of CI) & (end of CI).

5. If "$\beta_1 = 0$" is not in the interval, then we say there is a significant relationship between x & y (at the $(1-\alpha)100\%$ confidence level).

6. Hypothesis test for β_1:

① $H_0: \beta_1 = 0$; $H_A: \beta_1 \neq 0$

② (Assuming H_0) our test statistic is

$$\boxed{t = \frac{\hat{\beta}_1-\beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}}$$

③ p-value is $p = 2P(T>t)$, $T \sim t_{n-2}$

→ In R:

```
> p <- 2 * (1 - pt(t, n-2))
```

④ Check if $p < 0.05$; if yes, reject H_0.

Chapter 2: Multiple Regression

MULTIPLE REGRESSION MODEL

1. If we expand the SLR model to p explanatory variables, we obtain the multiple linear regression model:

$$\boxed{y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i=1,\dots,n}$$

2. This can be expressed as

$$\boxed{y = X\beta + \varepsilon}$$

where $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$, $X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \in \mathbb{R}^{n\times(p+1)}$,

$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1}$ & $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \in \mathbb{R}^n$.

NORMAL MODEL

For the normal model, where we assume $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, we write

$$\boxed{Y = X\beta + \varepsilon, \quad \varepsilon \sim mvN(0, \sigma^2 I)},$$

where $Var(\varepsilon) = \sigma^2 I$ is the covariance matrix of the error random vector ε.

LEAST SQUARES ESTIMATION OF β

1. We wish to minimize

$$\boxed{S(\beta_0, \dots, \beta_p) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]^2}$$

over $\beta_0, \dots, \beta_p$.

2. Taking partial derivatives and setting to 0:

$$\boxed{\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2\sum(y_i - (\hat\beta_0 + \hat\beta_1 x_{i1} + \cdots + \hat\beta_p x_{ip})) = 0 \\ \frac{\partial S}{\partial \beta_1} &= -2\sum x_{i1}(y_i - (\hat\beta_0 + \hat\beta_1 x_{i1} + \cdots + \hat\beta_p x_{ip})) = 0 \\ &\vdots \\ \frac{\partial S}{\partial \beta_p} &= -2\sum x_{ip}(y_i - (\hat\beta_0 + \hat\beta_1 x_{i1} + \cdots + \hat\beta_p x_{ip})) = 0 \end{aligned}}$$

3. This yields the normal equations

$$\boxed{\begin{aligned} n(\hat\beta_0) + \hat\beta_1 \sum x_{i1} + \cdots + \hat\beta_p \sum x_{ip} &= \sum y_i \\ \hat\beta_0 \sum x_{i1} + \hat\beta_1 \sum x_{i1}^2 + \cdots + \hat\beta_p \sum x_{i1}x_{ip} &= \sum x_{i1} y_i \\ &\vdots \\ \hat\beta_0 \sum x_{ip} + \hat\beta_1 \sum x_{i1}x_{ip} + \cdots + \hat\beta_p \sum x_{ip}^2 &= \sum x_{ip} y_i \end{aligned}}$$

4. We can write this as

$$\boxed{(X^T X)\hat\beta = X^T y}$$

and so

$$\boxed{\hat\beta = (X^T X)^{-1}(X^T y)}$$

- note this needs $X^T X$ to be invertible; ie full rank / all columns are linearly independent.

5. Note:

① The fitted line is given by

$$\boxed{\hat\mu = x^T \hat\beta.}$$

② The vector of fitted values is

$$\boxed{\hat\mu = X\hat\beta.}$$

③ The residual vector is

$$\boxed{e = y - \hat\mu}$$

* sum of squares of residuals is $\sum e_i^2 = e^T e$.

THE HAT MATRIX: H

- We can express $\hat{\mu}$ by

$$\boxed{\hat{\mu} = X\hat{\beta} = X(X^TX)^{-1}X^Ty = Hy}$$

where

$$\boxed{H = X(X^TX)^{-1}X^Ty}$$

is the "hat" matrix which maps the vector of response variables to the vector of fitted values.

- Note that
  1. H is symmetric (ie $H^T = H$); &
  2. H is idempotent (ie $H^2 = H$).
- We can express our residual vector e as

$$\boxed{e = y - \hat{\mu} = y - Hy = (I-H)y}$$

LEAST SQUARES ESTIMATION OF σ^2

- The least squares estimate of σ^2 for a p explanatory variable multiple regression model with (p+1) parameters is

$$\boxed{\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-(p+1)}}$$

where $df = n-(p+1)$.

RESIDUAL STANDARD ERROR

- The residual standard error is thus

$$\boxed{\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-(p+1)}}}$$

MLE FOR β

- The MLE for β is equivalent to the least squares estimate; ie the likelihood function

$$\boxed{\begin{aligned} L(\beta_0,\dots,\beta_n \mid y_1,\dots,y_n) &= \prod_{i=1}^{n} f(y_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}}, \quad \mu_i = \beta_0 + \sum_j \beta_j x_{ij} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-\sum(y_i-\mu_i)^2}{2\sigma^2}\right) \end{aligned}}$$

or equivalently the log likelihood function

$$\boxed{\ell(\beta_0,\dots,\beta_n \mid y_1,\dots,y_n) = c - \frac{\sum_i (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2}{2\sigma^2}}$$

is maximized at $\beta = (\beta_0,\dots,\beta_p)$ that minimizes

$$\boxed{\sum \varepsilon_i^2 = \sum (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2.}$$

GAUSS-MARKOV THEOREM & BLUE

- The least squares estimator

$$\boxed{\hat{\beta} = (X^TX)^{-1}X^Ty}$$

is the "best linear unbiased estimator" (BLUE) of β.

- More formally, if we consider the model given by

$$\boxed{Y = X\beta + \varepsilon, \quad E(\varepsilon) = 0, \quad Var(\varepsilon) = \sigma^2 I}$$

then amongst all unbiased linear estimators $\hat{\beta}^* = M^*Y$, the least squares estimator $\hat{\beta} = MY$ has the "smallest" variance; ie

$$\boxed{Var(\hat{\beta}^*) = Var(\hat{\beta}) + \sigma^2(M^*-M)(M^*-M)^T}$$

where $(M^*-M)(M^*-M)^T$ is positive semidefinite.

- A is "positive definite" if $a^TAa \geq 0$ for any vector a.

DISTRIBUTION OF $\hat{\beta}$

1. We show that

$$\boxed{\hat{\beta} \sim MVN(\beta, \sigma^2 (X^TX)^{-1})}$$

where MVN is the multivariate normal distribution.

Proof. First, we have

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim MVN(0, \sigma^2 I).$$

Thus, by properties of MVN,

$$Y \sim MVN(X\beta, \sigma^2 I).$$

Hence $\hat{\beta} = (X^TX)^{-1}X^TY$ also follows a MVN distribution.

Next, see that

$$\begin{aligned} E(\hat{\beta}) &= E((X^TX)^{-1}X^TY) \\ &= (X^TX)^{-1}X^T E[Y] \\ &= (X^TX)^{-1}X^T(X\beta) \\ &= \beta. \end{aligned}$$

Then

$$\begin{aligned} Var(\hat{\beta}) &= Var((X^TX)^{-1}X^TY) \\ &= (X^TX)^{-1}X^T Var(Y) [(X^TX)^{-1}X^T]^T \\ &\quad (Var(AY) = A\,Var(Y)A^T) \\ &= \sigma^2 (X^TX)^{-1}X^T [(X^TX)^{-1}X^T]^T \\ &= \sigma^2 (X^TX)^{-1}X^TX(X^TX)^{-1} \\ &= \sigma^2 (X^TX)^{-1}, \end{aligned}$$

which gives us the desired result.

2. The marginal distribution of $\hat{\beta}_j$ is thus

$$\boxed{\hat{\beta}_j \sim N(\beta_j, \sigma^2 (X^TX)^{-1}_{jj}) \quad \forall j = 0, \dots, p}$$

3. We also have that

$$\boxed{\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim t_{n-(p+1)}, \quad SE(\hat{\beta}_j) = \hat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}}$$

4. Also note that

$$\boxed{Cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 (X^TX)^{-1}_{ij}.}$$

INTERPRETATION OF $\hat{\beta}_j$

$\hat{\beta}_j$ is the estimated mean change in the response associated with a change of one unit of x_j whilst holding all other variables constant.

CIs FOR β_j

1. A $(1-\alpha)100\%$ CI for β_j is

$$\boxed{\hat{\beta}_j \pm t_{n-(p+1),\, 1-\alpha/2}\, SE(\hat{\beta}_j)}$$

2. If $\beta_j = 0$ is not in the CI, then there is a significant linear relationship between y & x_j.

HYPOTHESIS TESTS FOR β_j

Hypothesis test for β_j:

① $H_0: \beta_j = 0$; $H_A: \beta_j \neq 0$

② $t = \dfrac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} = \dfrac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$

③ p-value $= 2P(T > t)$, $T \sim t_{n-(p+1)}$

④ Reject H_0 if $p < 0.05$.

MULTI-COLLINEARITY

1. We say 2 or more explanatory variables exhibit "multicollinearity" if there exist strong linear relationships between them.
2. This
   ① increases the variances (and thus std errors) of the associated parameter estimators; &
   ② leads to wide/imprecise CIs & inaccurate conclusions from hypothesis tests.

VARIANCE INFLATION FACTOR / VIF

1. The "variance inflation factor" is a measure of multicollinearity associated with some explanatory variable x_j.
2. How to calculate VIF for x_j:
   ① Regress x_j onto all other x's; ie fit models for x_j against each other x_i
   ② Then

$$\boxed{VIF_j = \frac{1}{1 - R_j^2},}$$

   where R_j^2 is the coefficient of determination of the model fit with x_j as the response.
   - in R, R_j^2 is the "multiple R squared" parameter.
3. Generally, we remove x_j from the model if $VIF_j > 10 \iff R_j^2 > .90$.
4. In R, we can do

```
> lm(x ~ x1 + x2 + ... + xn)
```

$$\underbrace{x}_{\text{variable we are testing}} \sim \underbrace{x_1 + x_2 + \cdots + x_n}_{\text{other exp. variables}}$$

and check the multiple R-squared value.

CI FOR μ_{new}

1. Idea: We may want to use our fitted model to estimate the mean response of a new unit in the population.
2. In particular,

$$\boxed{\hat{\mu}_{new} = x_{new}^T \hat{\beta}.}$$

3. Then, we show that

$$\boxed{\hat{\mu}_{new} \sim N(x_{new}^T \beta, \ \sigma^2 x_{new}^T (X^T X)^{-1} x_{new})}$$

Proof. Recall

$$\hat{\beta} \sim MVN(\beta, \sigma^2 (X^T X)^{-1}).$$

Then $\hat{\mu}_{new} = x_{new}^T \hat{\beta}$ must also follow a normal distribution.

See that

$$E(\hat{\mu}_{new}) = E(x_{new}^T \hat{\beta}) = x_{new}^T E(\hat{\beta}) = x_{new}^T \beta$$

Then

$$Var(\hat{\mu}_{new}) = Var(x_{new}^T \hat{\beta}) = x_{new}^T Var(\hat{\beta}) x_{new} = \sigma^2 x_{new}^T (X^T X)^{-1} x_{new}.$$

This gives the result, so we're done. ∎

4. Thus, a $(1-\alpha)\%$ confidence interval for μ_{new} is

$$\boxed{\hat{\mu}_{new} \pm t_{n-(p+1),\, 1-\frac{\alpha}{2}} \, \hat{\sigma} \sqrt{x_{new}^T (X^T X)^{-1} x_{new}}}$$

PREDICTION INTERVAL FOR y_{new}

1. Idea: We may also wish to use our fitted model to predict the value of the response of a new unit of the population.
2. Then, note the variance of $\hat{y}_{new}$ is composed of 2 sources of variation:
   ① the variation associated with the parameter estimators; &
   ② the variance σ^2 associated with a random response.
3. Thus our total variation is

$$\boxed{\sigma^2 + \sigma^2 x_{new}^T (X^T X)^{-1} x_{new}.}$$

4. Hence, a $(1-\alpha)100\%$ prediction interval for y_{new} is

$$\boxed{\hat{y}_{new} \pm t_{n-(p+1),\,1-\frac{\alpha}{2}}\ \hat{\sigma}\sqrt{1 + x_{new}^T (X^T X)^{-1} x_{new}}}$$

CONFIDENCE & PREDICTION BANDS FOR THE SLR MODEL

1. For the SLR model,

$$\boxed{x_{new}^T (X^T X)^{-1} x_{new} = \frac{1}{n} + \frac{(x_{new}-\bar{x})^2}{S_{xx}}}$$

2. Thus, the closer x_{new} is to $\bar{x}$, the narrower the CIs & PIs.
3. In general, the closer $\{x_1, \dots, x_p\}$ is to $\{\bar{x}_1, \dots, \bar{x}_p\}$ in the multiple regression model, the narrower the interval.

MODELLING CATEGORICAL EXPLANATORY VARIABLES

1. We can code categorical explanatory variables using indicator variables that take on values of 0 or 1.

   eg $x_1 = \mathbb{I}[A = a_1]$, $x_2 = \mathbb{I}[A = a_2]$

   $\mathbb{I}$ = the indicator function.
2. For a variable with ℓ category levels, we need $\ell - 1$ indicator variables.
3. In particular, if X is a categorical variable that has ℓ distinct values $a_1, \dots, a_\ell$, we can use the model

$$\boxed{\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_{\ell-1} x_{\ell-1},}$$

   where $x_i = \mathbb{I}[X = a_i]$.

   - $(x_1, \dots, x_{\ell-1}) = (0, \dots, 0)$ corresponds with $X = a_\ell$.
4. Then, each $\hat{\beta}_i$ corresponds to the difference in the estimated mean value of the response where $X = a_i$ relative to where $X \neq a_i$.

INFERENCE FOR PARAMETERS ASSOCIATED WITH INDICATOR VARIABLES

1. To test whether there is a difference in $\hat{\mu}$ between data where $x_i = 1$ vs. $x_i = 0$, we can use the following hypothesis test:
   ① $H_0: \beta_i = 0$; $H_A: \beta_i \neq 0$
   ② We get t, p from the `summary` output from the model fit (ie lm); &
   ③ If $p < 0.05$, we reject H_0.
2. To test whether there is a difference in $\hat{\mu}$ between data where $x_i = 1$ vs. $x_j = 1$, we can use the following hypothesis test:
   ① $H_0: \beta_i - \beta_j = 0$, $H_A: \beta_i - \beta_j \neq 0$
   ② We can derive

$$\boxed{\hat{\beta}_i - \hat{\beta}_j \sim N\left(\beta_i - \beta_j,\ \sigma^2\left((X^TX)^{-1}_{ii} + (X^TX)^{-1}_{jj} - 2(X^TX)^{-1}_{ij}\right)\right)}$$

   ③ Thus we get

$$\boxed{t = \frac{(\hat{\beta}_i - \hat{\beta}_j)}{SE(\hat{\beta}_i - \hat{\beta}_j)}, \quad p = 2P(T > t),\ T \sim t_{n-(p+1)}}$$

ORTHOGONAL X MATRIX DESIGNS

1. We may wish to model categorical variables in a way that creates an orthogonal X matrix, thus producing independent parameter estimators.

eg Suppose we had 3 categorical variables X_1, X_2, X_3, each with 2 values 0 & 1.

We can define

$$x_i = \begin{cases} 1, & X_i = 1 \\ -1, & X_i = 0. \end{cases}$$

This yields the "X" matrix

$$X = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

Note the columns of X are orthogonal, and in particular

$$(X^TX) = 8I_4, \quad (X^TX)^{-1} = \frac{1}{8}I_4, \text{ where}$$

I_4 is the 4x4 identity matrix.

Hence $\sigma^2(X^TX)^{-1} = Var(\hat{\beta})$ is diagonal, indicating independent parameter estimators for the normal model.

2. In particular, when comparing one level to another in X for a given factor, it corresponds to an estimated mean change in the response of $2\hat{\beta}_j$.

- since parameter estimators are independent, they are unaffected by inclusion/exclusion of other variables
- so we do not have to account for other variables when interpreting the parameter estimates associated with a given factor.

ANALYSIS OF VARIANCE (ANOVA)

1. We can express the total sum of squares of the observations y_i as

$$\boxed{SS(Tot) = \sum(y_i - \bar{y})^2 = \underbrace{\sum(\hat{\mu}_i - \bar{y})^2}_{SS(Reg)} + \underbrace{\sum(y_i - \hat{\mu}_i)^2}_{SS(Res)}}$$

2. In particular,
   1. The "regression sum of squares" SS(Reg) is the variation explained by the model; &
   2. The "residual sum of squares" SS(Res) is the variation in the response left unexplained by the model.

3. In ANOVA methods of inference, we draw conclusions about the relative fit of models by comparing SS(reg) & SS(res).

4. The greater SS(reg) is compared to SS(res), the better the model fit.

COEFFICIENT OF DETERMINATION / MULTIPLE R-SQUARED: R^2

1. The "coefficient of determination" is

$$\boxed{R^2 = 1 - \frac{SS(Res)}{SS(Tot)}}$$

2. R^2 measures the proportion of the variation in the response explained by the model.

F-TEST FOR MODEL PARAMETERS: $\beta_1 = \cdots = \beta_p = 0$

1. Idea: To test if a relationship exists between the response at least one of the explanatory variates, we can use a F-test.

2. Method:
   1. $H_0: \beta_1 = \cdots = \beta_p = 0; \quad H_A: \exists j \text{ s.t. } \beta_j \neq 0$
   2. Test statistic:

   $$\boxed{F = \frac{SS(Reg)/p}{SS(Res)/(n-p-1)} = \frac{MS(Reg)}{MS(Res)}}$$

   3. Under H_0, F has a F distribution with $p, n-p-1$ degrees of freedom.
   4. The F-test statistic & p-value are provided on the last line of the `summary` output for lm.
   5. We reject H_0 if $p < 0.05$.

ANOVA TABLE

1. We summarise the test $H_0: \beta_1 = \cdots = \beta_p = 0$ in an ANOVA table:

source	df	SS	MS	F	p-value
regression	p	SS(Reg)	$\frac{SS(Reg)}{p}$	$\frac{MS(Reg)}{MS(Res)}$	$P(F_{p,n-(p+1)} > F)$
residual	$n-(p+1)$	SS(Res)	$\frac{SS(Res)}{n-(p+1)}$		
total	$n-1$	SS(Tot)			

2. We can obtain SS(Res) from $\hat{\sigma}$:

$$\hat{\sigma} = \sqrt{\frac{\Sigma e_i^2}{n-(p+1)}} = \sqrt{\frac{SS(Res)}{n-(p+1)}}$$

$$\Rightarrow \quad SS(Res) = (n-(p+1))\,\hat{\sigma}^2$$

ADDITIONAL SUM OF SQUARES

1. Consider the "full" model of

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

and the "reduced" model that reflects the restrictions imposed by $\beta_1 = \cdots = \beta_k = 0$, $k \le p$, ie

$$Y = \beta_0 + \beta_{k+1} x_{k+1} + \cdots + \beta_p x_p + \varepsilon .$$

ie "After accounting for $\beta_1, \dots, \beta_k$, does ≥ 1 of $\beta_{k+1}, \dots, \beta_p$ account for significant variation in y?"

2. To determine which model is better, we can examine

$$SS(Reg)_{full} - SS(Reg)_{red}$$
$$\text{or} \quad SS(Res)_{red} - SS(Reg)_{full}$$

which is the difference in the variation explained by the full & reduced models.

3. We can test $H_0: \beta_1 = \cdots = \beta_k = 0$ with the F test statistic

$$F = \frac{(SS(Res)_{red} - SS(Res)_{full}) \,/\, df_{red} - df_{full}}{SS(Res)_{full} \,/\, df_{full}}$$

where under H_0, $F \sim F_{df_{red} - df_{full},\, df_{full}}$.

4. Our p-value is $P(F \ge F_{test})$, $F \sim F_{df_{red} - df_{full},\, df_{full}}$.

5. In R, we can use the anova function to do this.

```
> anova(audit.red.lm,audit.full.lm)
Analysis of Variance Table

Model 1: overhead ~ col + clients
Model 2: overhead ~ size + age + col + clients
  Res.Df RSS        Df Sum of Sq  F      Pr(>F)
1     21 4954374034
2     19 3901347198  2 1.053e+09  2.5642 0.1033
```

6. Note that our previous ANOVA method to test $H_0: \beta_1 = \cdots = \beta_p = 0$ is just an additional sum of squares test where our reduced model is

$$Y = \beta_0 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

7. Similarly, we can also test $H_0: \beta_j = 0$ using the additional sum of squares test statistic

$$F = \frac{SS(Res)_{red} - SS(Res)_{full}}{SS(Res)_{full} \,/\, df_{full}}$$

where our reduced model is just the full model with x_j omitted & under H_0, $F \sim F_{1,p}$.

(Note $P(|t|_v > |t|) = P(F_{1,v} > t^2)$.)

ADDITIONAL SUM OF SQUARES & CATEGORICAL VARIABLES

1. Suppose we have a model

$$\boxed{Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon}$$

where $x_i = \mathbb{I}[A = a_i]$, where A is a categorical variable.

2. To test $H_0: \beta_1 - \beta_2 = 0$, we can use the reduced model under $H_0: \beta_1 = \beta_2 = \beta^*$; ie

$$\boxed{\begin{aligned} Y &= \beta_0 + \beta^*(x_1 + x_2) + \varepsilon \\ &= \beta_0 + \beta^* x^* + \varepsilon \end{aligned}}$$

where $x^* = \mathbb{I}[A = a_1 \text{ or } A = a_2]$.

3. We can then use the additional sum of squares statistic & proceed as before.

GENERAL LINEAR HYPOTHESIS

1. The "general linear hypothesis" is in the form

$$\boxed{H_0: A\beta = 0}$$

where $A \in \mathbb{R}^{\ell \times (p+1)}$ describes the ℓ linear constraints on the full model as described by H_0.

2. In particular, we can use the additional sum of squares test statistic to test H_0; under H_0,

$$\boxed{F = \frac{(SS(Res)_{red} - SS(Res)_{full}) / (df_{red} - df_{full})}{SS(Res)_{full} / df_{full}} \sim F_{df_{red} - df_{full},\, df_{full}}}$$

eg $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ can be expressed as

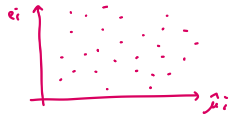
$$H_0: \underbrace{\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}}_{\beta} = \underbrace{\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}}_{0}$$

ASSESSING MODEL ADEQUACY / RESIDUAL ANALYSIS

Idea: We can assess the "model adequacy" (ie the validity of the model assumptions) by examining the fitted residuals $e = y - \hat{\mu}$.

RESIDUAL PLOTS

1. We can plot the residuals e_i against the fitted values $\hat{\mu}_i$.



2. If the model assumptions hold, then e_i and $\hat{\mu}_i$ should be uncorrelated.

3. Thus, if the model assumptions hold, we should see no observable pattern in the plot.

So these are not valid:

(observed values shows quadratic relationship)

(non-constant error variance)

QQ PLOTS

1. In a QQ plot, we plot the ordered residuals $e_{(i)}$ against the expected ordered values $E(z_{(i)})$, $z_i \sim N(0,1)$.

2. We use QQ plots to assess the assumption of normal errors.

3. In particular, if the errors (and hence residuals) are from a normal distribution, then $e_{(i)}$ should be proportional to $E(z_{(i)})$, so the plot should exhibit a straight line relationship.

VARIANCE STABILIZING TRANSFORMATIONS

1. "Variance stabilizing transformations" are transformations to the response and possibly some of the explanatory variates to improve model adequacy / validity of model assumptions.

eg $\log(y)$, $y^{1/2}$, y^{-1}

2. These are useful when we can write

$$\boxed{\sigma^2 = f(\mu).}$$

PROPERTIES OF THE RESIDUALS

1. Recall we can express

$$\boxed{e = (I-H)\varepsilon}$$

where H is the hat matrix ($H = X(X^TX)^{-1}X^T$).

2. Since $\varepsilon \sim N(0, \sigma^2 I)$, it follows that

$$\boxed{\begin{aligned} &e \sim N(0, \sigma^2(I-H)) \\ &\Rightarrow e_i \sim N(0, \sigma^2(1-h_{ii})) \end{aligned}}$$

where h_{ii} is the i^{th} diagonal element of H.

3. Note

① $\boxed{Var(e_i) = \sigma^2(1-h_{ii});}$

- residuals have non-constant variance

② $\boxed{Cov(e_j, e_k) = -\sigma^2 h_{jk}, \quad j \neq k}$

- residuals are not independent
- result of constraint $\sum e_i = 0$ in least squares estimation

STUDENTIZED RESIDUALS

1. For a random variable X, the "studentized" version of X is

$$\boxed{Y = \frac{X-\mu}{\hat{\sigma}}}$$

where μ is the mean and $\hat{\sigma}$ is an estimate of the standard deviation.

2. A "studentized residual" associated with observation i is

$$\boxed{d_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}}$$

which is $\sim N(0,1)$ for large n.

EXTREME RESPONSE VALUES / OUTLIERS

1. An "outlier" is an observation for which $e_i = y_i - \hat{\mu}_i$ is extreme relative to the other residuals.

2. In particular, we can consider an observation an outlier if

$$\boxed{|d_i| > 2.5.}$$

3. Causes of outliers:

① Typos / misrecording of data;

② Values of associated potential explanatory variables not included in the model;

③ Random variability.

LEVERAGE

1. "Leverage" is a measure used to identify those observations whose set of explanatory variables is extreme relative to other observations.

2. In particular, for observation i,

$$\boxed{\text{leverage} = h_{ii} = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum(x_j-\bar{x})^2}}$$

3. Properties:

$$\boxed{\begin{aligned} &\tfrac{1}{n} \leq h_{ii} \leq 1 \\ &\sum h_{ii} = tr(H) = rank(H) = p+1 \end{aligned}}$$

4. Also note

$$\boxed{\hat{\mu} = Hy \Rightarrow \hat{\mu}_i = h_{ii}y_i + \sum_{j \neq i} h_{ij} y_j}$$

and so we can view the leverage as the weight of y_i's contribution to the fitted value $\hat{\mu}_i$.

IDENTIFYING HIGH LEVERAGE CASES

1. Residual plots are not useful in revealing high leverage points.

- as $h_{ii} \to 1$, $Var(e_i) \to 0$
- so high leverage observations' leverage is close to 0.

2. Instead, we can just plot the leverage values.

- in R, we can use 'hatvalues'.

3. We say observation i has high leverage if

$$\boxed{h_{ii} > 2\bar{h} = \frac{2(p+1)}{n}.}$$

INFLUENTIAL OBSERVATIONS

We say an observation is "influential" if its removal from the line fit changes the fitted line (ie parameter estimates) considerably.

IDENTIFYING INFLUENTIAL OBSERVATIONS

1. We can quantify influence using "Cook's distance":

$$\boxed{D_i = \frac{(\hat{\mu}-\mu_{(i)})^T(\hat{\mu}-\mu_{(i)})}{\hat{\sigma}^2(p+1)} = \frac{h_{ii}}{1-h_{ii}} \cdot \frac{d_i^2}{p+1}}$$

where $\hat{\sigma}^2$ is the estimate of the variance from the model fit with the i^{th} observation included.

2. We say an observation is strongly influential if

$$\boxed{D_i \geq 1.}$$

MODEL SELECTION

1. Recall that

$$\boxed{\hat{\sigma} = \sqrt{\frac{SS(Res)}{n-(p+1)}}}$$

2. If we remove variables from the model, both SS(Res) & the df increase.

3. If the increase in SS(Res) is small relative to the degrees of freedom gained, then $\hat{\sigma}$ will decrease, resulting in a more precise model.

4. This is often the case for variates with large associated p-values.

ITERATIVE MODEL SELECTION

Idea: build a reduced model by adding/removing variables one at a time, & refitting at each iteration until no more variables can be added/removed.

BACKWARD ELIMINATION

Idea:

① Fit all p variables;

② Remove variable with largest p-value greater than some threshold (eg $\alpha = 0.1$)

③ Refit with remaining variables;

④ Repeat ①-② until no more variables can be removed.

FORWARD SELECTION

Idea:

① Fit all p single variable SLR models;

② Select variable with smallest p-value $< \alpha$;

③ Fit the $p-1$ 2-variable models that include the variable in ②;

④ Repeat ①-③, continuing to add variables at each iteration until no more variables can be added.

STEPWISE SELECTION

Idea:

① Begin with forward selection;

② Employ both forward selection & backward elimination at each step until no more variables can be added/removed.

ADJUSTED R-SQUARED

1. Motivation: Recall

$$\boxed{R^2 = 1 - \frac{SS(Res)}{SS(Tot)}}$$

Note R^2 will always increase when more variables are added regardless of whether the variables account for a significant amount of the variation of the response.

2. Hence, we cannot use R^2 to compare model subsets with differing amounts of parameters.

3. Thus, we can use the "adjusted R-squared":

$$\boxed{R^2_{adj} = 1 - \frac{SS(Res)/(n-(p+1))}{SS(Tot)/(n-1)}}$$

4. R^2_{adj} will only increase if the variation accounted for by the added variables increases proportionally more than the degrees of freedom through the model decrease parameters.

5. Note we can also write

$$\boxed{R^2_{adj} = 1 - \frac{\hat{\sigma}^2}{SS(Tot)/(n-1)}}$$

So a large R^2_{adj} implies a small $\hat{\sigma}$ and v.v.

MALLOWS' C_p

1. For a k-variable model $(k=1,\dots,p)$, we define

$$\boxed{C_p = \frac{SS(Res)_k}{MS(Res)_p} + 2(k+1) - n}$$

2. Note smaller C_p values relative to the number of variables are associated with more suitable models.

3. A k-variable model is preferred over the full model if

$$\boxed{C_p \le k+1.}$$

MODEL SELECTION IN R

1. To select the best model in R, we can use the `leaps` function, which selects the best model from all possible subsets based on the R^2_{adj} & C_p^2 criteria.

eg

```
> leaps(house[,-9],value,method=c('adjr'),nbest=2,names=names(house[-9]))
   size stories baths rooms   age lotsize basement garage
1   TRUE   FALSE FALSE FALSE FALSE   FALSE    FALSE  FALSE
1  FALSE   FALSE  TRUE FALSE FALSE   FALSE    FALSE  FALSE
2   TRUE   FALSE FALSE FALSE  TRUE   FALSE    FALSE  FALSE
2   TRUE    TRUE FALSE FALSE FALSE   FALSE    FALSE  FALSE
3   TRUE   FALSE FALSE FALSE  TRUE    TRUE    FALSE  FALSE
3   TRUE   FALSE FALSE FALSE  TRUE   FALSE    FALSE   TRUE
4   TRUE   FALSE FALSE FALSE  TRUE    TRUE    FALSE   TRUE
4   TRUE    TRUE FALSE FALSE  TRUE    TRUE    FALSE  FALSE
5   TRUE    TRUE FALSE FALSE  TRUE    TRUE    FALSE   TRUE
5   TRUE   FALSE FALSE  TRUE  TRUE    TRUE    FALSE   TRUE
6   TRUE    TRUE  TRUE FALSE  TRUE    TRUE    FALSE   TRUE
6   TRUE    TRUE FALSE  TRUE  TRUE    TRUE    FALSE   TRUE
7   TRUE    TRUE  TRUE  TRUE  TRUE    TRUE    FALSE   TRUE
7   TRUE    TRUE  TRUE FALSE  TRUE    TRUE     TRUE   TRUE
8   TRUE    TRUE  TRUE  TRUE  TRUE    TRUE     TRUE   TRUE
$adjr2
[1]0.549407 0.317135 0.572833 0.560069 0.618481 0.593374 0.644686 0.6251132
[9]0.651625 0.641082 0.648770 0.648308 0.645221 0.645063 0.641430
```

```
> leaps(house[,-9],value,method=c('Cp'),nbest=1,names=names(house[-9]))
  size stories baths rooms   age lotsize basement garage
1 TRUE   FALSE FALSE FALSE FALSE   FALSE    FALSE  FALSE
2 TRUE   FALSE FALSE FALSE  TRUE   FALSE    FALSE  FALSE
3 TRUE   FALSE FALSE FALSE  TRUE    TRUE    FALSE  FALSE
4 TRUE   FALSE FALSE FALSE  TRUE    TRUE    FALSE   TRUE
5 TRUE    TRUE FALSE FALSE  TRUE    TRUE    FALSE   TRUE
6 TRUE    TRUE  TRUE FALSE  TRUE    TRUE    FALSE   TRUE
7 TRUE    TRUE  TRUE  TRUE  TRUE    TRUE    FALSE   TRUE
8 TRUE    TRUE  TRUE  TRUE  TRUE    TRUE     TRUE   TRUE
$Cp
[1] 27.150656 21.556807 10.144146  4.137270  3.327332  5.096306  7.027277  9.000000
```

2. Note
   1. the full model has a higher R^2 value, as it has more parameters; but
   2. the reduced model has a higher R^2_{adj} value & a lower $\hat{\sigma}$ value.

INTERACTION

1. We say there is "interaction" between x_j, x_k if the effect of x_j on the response depends on the value of x_k.

2. To account for a possible interaction effect, we can include the term $x_j x_k$ in our model.

FITTING LINEAR MODELS TO TIME SERIES DATA

1. "Time series data" is data $\{y_t\}$, where y_t denotes the value of the response at time t $(t = 1, 2, \dots)$.
2. Variation in time series is caused by:
   ① "Seasonal" – repeat regularly over fixed intervals
   ② "Trend" – persistent increase / decrease in μ_t as t increases
   ③ "Cyclical" – oscillations repeating over irregular intervals
   ④ "Irregular/random" – other sources of variation unaccounted by ①-③.

AUTOCORRELATION FUNCTION / ACF

1. The "auto-correlation function" is

$$\boxed{r_k = \frac{\sum_{t=k+1}^{n} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2}}$$

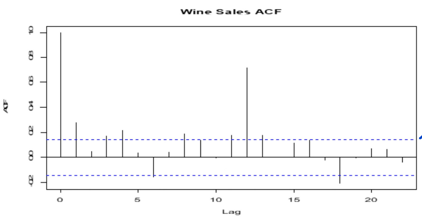
and quantifies the degree of similarity between y_t & the "lagged" version of itself, y_{t-k}.

2. Note
   ① $|r_k| \le 1$ $\forall k$; &
   ② r_k is unitless.

CORRELOGRAM

1. A "correlogram" is a plot of r_k over k (for $k = 1, 2, \dots$).

les)



"significance lines"

2. We can use the `acf` function in R to plot this.

SEASONAL COMPONENT

Idea: We can add model parameters to account for the seasonal variation in the data, which might exacerbate any trends.

TREND COMPONENT

Idea: We can account for the trend component in the model by adding appropriate time variables (eg t, t^2, etc)

Chapter 3: Logistic Regression

Idea: Suppose we want to model a binary response variable

$$Y_i = \begin{cases} 1, & i^{th} \text{ observation is successful} \\ 0, & \text{otherwise} \end{cases}$$

where $P(Y_i=1) = \pi_i$.

We can model $\mu_i = E(Y_i) = \pi_i$ using

$$\pi_i = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$$

$$\Leftrightarrow \log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T\beta$$

where the LHS of the second equation is called the logit function.

In R, we can fit a logistic regression model using

```
> glm(y ~ x1+...+xp, family=binomial(link='logit'))
```

INTERPRETATION OF LOGISTIC REGRESSION PARAMETER ESTIMATES

By manipulating the equation

$$\log\left(\frac{\pi}{1-\pi}\right) = x^T\beta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

we can get

$$e^{\beta_j} = \frac{\pi_{x_j+1} / (1-\pi_{x_j+1})}{\pi/(1-\pi)}$$

where π_{x_j+1} is the value of $E(Y_i)$ after increasing one unit of x_j.

Thus, we can interpret this "odds ratio" as the multiplicative change in the odds of success for a change in one unit of x_j, after accounting for the other variables.

DISTRIBUTION OF $\hat{\beta}_j$

Let $\hat{\beta}_j$ be the MLE of β_j. Then, for sufficiently large n, we can show that

$$\left(\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)}\right)^2 \sim \chi_1^2$$

$$\Leftrightarrow \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim N(0,1)$$

$H_0: \beta_j = 0$ TEST

To test $H_0: \beta_j = 0$, we can use the "Wald test statistic"

$$z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

which is $\sim N(0,1)$ under H_0.

Note the test statistic & p-values associated with this are given in the `summary` output of the glm model in R.

Hence, a $100(1-\alpha)\%$ CI for β_j is

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} SE(\hat{\beta}_j)$$

and a $100(1-\alpha)\%$ CI for the odds ratio is

$$e^{\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} SE(\hat{\beta}_j)}$$

DEVIANCE / D

1. The likelihood function for a logistic regression model with a binary response is

$$\boxed{L(\beta|y) = \prod_{i=1}^{\hat{n}} f(y_i|\pi_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}}$$

with corresponding log-likelihood

$$\boxed{\ell(\beta|y) = \sum_{i=1}^{\hat{n}} (y_i \log \pi_i + (1-y_i) \log(1-\pi_i))}$$

2. The "deviance" of the model is

$$\boxed{D = 2(\ell(\text{saturated}) - \ell(\text{fitted}))}$$

where

① $\ell(\text{fitted})$ is the log-likelihood of the fitted model (with MLE $\hat{\pi}_i$); &

② $\ell(\text{saturated})$ is the log-likelihood of the saturated model, the hypothetical model for which $p+1 = n$ (with MLE $\hat{\pi}_i = y_i$).

3. In particular, the deviance for the binary response logistic regression model is

$$\boxed{D = -2 \sum (y_i \log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) + \log(1-y_i))}$$

MODEL COMPARISON - DEVIANCE

1. To test whether either $x_{k_1}, \dots, x_{k_m}$ are associated with a positive response of y after accounting for the other variates, we can use the test

$$\boxed{H_0 : \beta_{k_1} = \cdots = \beta_{k_m} = 0 .}$$

2. Our full model is

$$\boxed{\beta_0 + \sum_{j=1}^{p} \beta_j x_j}$$

and our reduced model is

$$\boxed{\beta_0 + \sum_{j \neq k_1, \dots, k_m} \beta_j x_j}$$

3. If we assume H_0 is true (ie the reduced model gives a better fit), then we can approximate

$$\boxed{\begin{aligned} D_{red} - D_{full} &= 2 \log\left(\frac{L(\text{fitted})_{full}}{L(\text{fitted})_{red}}\right) \\ &\sim \chi^2_{df_{red} - df_{full}} \end{aligned}}$$

MODEL COMPARISON - AKIAKE INFORMATION CRITERION / AIC

1. The "AIC" is defined as

$$\boxed{AIC = 2(p+1) - 2\ell(\text{fitted})}$$

2. The smaller the AIC, the bette the fit of the model.