

STAT 331



Personal Notes

Marcus Chan

Taught by Peter Balka
UW CS '25



Chapter 1: Introduction

REGRESSION

Q: 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

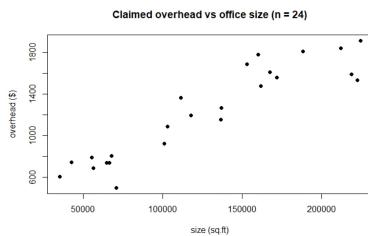
Q: A simple LR model involves:

- ① A single explanatory variate;
- ② A single response variate.

e.g Overhead data example:

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

Q: We can summarise the data using a scatter-plot.



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

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

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

Q: r tells us the relative strength of the linear relationship.

THE SIMPLE LR MODEL

Q: 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 ; &

e.g 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 .

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

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

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

We typically assume in SLR that

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

for some variance σ^2 .

This yields the normal model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim \text{iid } N(0, \sigma^2).$$

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

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

$$y_1 = \beta_0 + \beta_1 x_1 + \epsilon_1$$

⋮

$$y_n = \beta_0 + \beta_1 x_n + \epsilon_n,$$

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

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 .

We show that

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

Proof. we wish to minimize

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

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. \blacksquare

In R, we can get these values via

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

FITTED MODEL

For the SLR model, the fitted model is

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

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

FITTED RESIDUALS

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

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

* e_i is a random variable in which we impose assumptions;

e_i is the difference between the observed response & estimated mean response.

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

$$\sum e_i = 0$$

$$\sum x_i e_i = 0.$$

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 : $\hat{\sigma}^2$

\textcircled{B}_1 In the normal model, we assume

$$\varepsilon_i \sim N(0, \sigma^2).$$

\textcircled{B}_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.

\textcircled{B}_3 In particular, this means our estimate for σ^2 is

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

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

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

\textcircled{B}_1 The "residual standard error" is

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

\textcircled{B}_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 .

\textcircled{B}_3 The smaller $\hat{\sigma}$ is, the more the variation in y is "explained" by x , and so the better fit the model is.

\textcircled{B}_4 $\hat{\sigma}$ 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

\textcircled{B}_1 We may interpret $\hat{\beta}_0$ as the estimated mean change in the response y associated with a change of one unit in x .

\textcircled{B}_2 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.

\textcircled{B}_3 Lastly, we can interpret $\hat{\sigma}$ as a measure of the variability of the response about the fitted line.

INFERENCE FOR β_1

\textcircled{B}_1 To investigate whether there is a linear relationship between y & x in the population, we can test the hypothesis $\beta_1 = 0$.

\textcircled{B}_2 We can then either use confidence intervals or hypothesis tests to test this.

\textcircled{B}_3 To do this, we need the least squares estimator of β_1 :

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

DISTRIBUTION OF $\hat{\beta}_1$

Q: we can show for the SLR model that

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

Proof. First, note

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\ &= \frac{\sum(x_i - \bar{x})y_i - \bar{y}\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \\ &= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} \quad \because \sum(x_i - \bar{x}) = 0 \\ &= \sum c_i y_i, \quad c_i = \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2}.\end{aligned}$$

Then, for the SLR model, $\epsilon_i \sim N(0, \sigma^2)$.

Since $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, 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

$$\begin{aligned}E(\hat{\beta}_1) &= E(\sum c_i y_i) = \sum c_i E(y_i) \\ &= \sum \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2} \cdot (\beta_0 + \beta_1 x_i) \\ &= \frac{\beta_0 \sum(x_i - \bar{x}) + \beta_1 \sum x_i(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \\ &= \frac{\beta_1 \sum x_i(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \\ &= \frac{\beta_1 \sum x_i(x_i - \bar{x}) - \beta_1 \sum \underbrace{x_i}_{0}(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \\ &= \frac{\beta_1 \sum(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2} = \beta_1.\end{aligned}$$

Similarly,

$$\begin{aligned}\text{Var}(\hat{\beta}_1) &= \text{Var}(\sum c_i y_i) \\ &= \sum c_i^2 \text{Var}(y_i) \quad \because y_i \text{ is ind.} \\ &= \sum \frac{(x_i - \bar{x})^2}{(\sum(x_i - \bar{x})^2)^2} \cdot \sigma^2 \\ &= \frac{\sigma^2}{\sum(x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}.\end{aligned}$$

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

Q: It follows that

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

(from STAT231/330 result).

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

DISTRIBUTION OF $\hat{\beta}_0$

Q: Similarly, we can show in a SLR model,

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}))$$

$$\frac{\hat{\beta}_0 - \beta_0}{\text{SE}(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}.$$

CI FOR β_1

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

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

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

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

Q: 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

> $t \leftarrow qt(1-\frac{\alpha}{2}, n-2)$

↳ The CI is then $126.33 - t(0.88)$, $126.33 + t(0.88)$.

Q: 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).

Q: 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.

Q: 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

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

- ③ p-value is $p = 2P(T > t)$, $T \sim t_{n-2}$
 - In R:
> $p \leftarrow 2 * (1 - pt(t, n-2))$
- ④ Check if $p < 0.05$; if yes, reject H_0 .

Chapter 2:

Multiple Regression

MULTIPLE REGRESSION MODEL

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

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

This can be expressed as

$$y = X\beta + \epsilon$$

where $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$, $X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \in \mathbb{R}^{n \times (p+1)}$,

 $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1}$ & $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \in \mathbb{R}^n$.

NORMAL MODEL

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

$$y = X\beta + \epsilon, \quad \epsilon \sim MVN(0, \sigma^2 I),$$

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

LEAST SQUARES ESTIMATION OF β

We wish to minimize

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

over β_0, \dots, β_p . Taking partial derivatives and setting to 0:

$$\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2 \sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip})) = 0 \\ \frac{\partial S}{\partial \beta_1} &= -2 \sum x_{i1} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip})) = 0 \\ &\vdots \\ \frac{\partial S}{\partial \beta_p} &= -2 \sum x_{ip} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip})) = 0 \end{aligned}$$

This yields the normal equations

$$\begin{aligned} n(\hat{\beta}_0) + \hat{\beta}_1 \sum x_{i1} + \dots + \hat{\beta}_p \sum x_{ip} &= \sum y_i \\ \hat{\beta}_0 \sum x_{i1} + \hat{\beta}_1 \sum x_{i1}^2 + \dots + \hat{\beta}_p \sum x_{i1} x_{ip} &= \sum x_{i1} y_i \\ &\vdots \\ \hat{\beta}_0 \sum x_{ip} + \hat{\beta}_1 \sum x_{i1} x_{ip} + \dots + \hat{\beta}_p \sum x_{ip}^2 &= \sum x_{ip} y_i \end{aligned}$$

We can write this as

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

and so

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

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

Note:

① The fitted line is given by

$$\hat{y} = X \hat{\beta}.$$

② The vector of fitted values is

$$\hat{y} = X \hat{\beta}.$$

③ The residual vector is

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

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

THE HAT MATRIX: \hat{H}

$\textcircled{1}$ we can express $\hat{\mu}$ by

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

where

$$H = X(X^T X)^{-1} X^T$$

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

Note that

$\textcircled{1}$ H is symmetric (ie $H^T = H$); &

$\textcircled{2}$ H is idempotent (ie $H^2 = H$).

$\textcircled{3}$ We can express our residual vector e as

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

LEAST SQUARES ESTIMATION OF σ^2

$\textcircled{1}$ The least squares estimate of σ^2 for a p explanatory variable multiple regression model with (p+1) parameters is

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

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

RESIDUAL STANDARD ERROR

$\textcircled{2}$ The residual standard error is thus

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

MLE FOR β

$\textcircled{1}$ The MLE for β is equivalent to the least squares estimate; ie the likelihood function

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

or equivalently the log likelihood function

$$l(\beta_0, \dots, \beta_n | y_1, \dots, y_n) = c - \frac{\sum (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2}{2\sigma^2}$$

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

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

GAUSS-MARKOV THEOREM & BLUE

$\textcircled{1}$ the least squares estimator

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

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

$\textcircled{2}$ More formally, if we consider the model given by

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

then amongst all unbiased linear estimators

$\hat{\beta}^* = M^* y$, the least squares estimator

$\hat{\beta} = My$ has the "smallest" variance;

i.e.

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

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

- A is "positive definite" if $a^T A a > 0$ for any vector a .

DISTRIBUTION OF $\hat{\beta}$

We show that

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

where MVN is the multivariate normal distribution.

Proof: First, we have

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

Thus, by properties of MVN,

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

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

Next, see that

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

Then

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

which gives us the desired result.

Q2: The marginal distribution of $\hat{\beta}_j$ is thus

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

Q3: We also have that

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

Q4: Also note that

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

INTERPRETATION OF $\hat{\beta}_j$

Q5: $\hat{\beta}_j$ 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

Q6: A $(1-\alpha) 100\%$ CI for β_j is

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

Q7: 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

Q8: Hypothesis test for β_j :

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

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

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

④ Reject H_0 if $p < 0.05$.

MULTI-COLLINEARITY

We say 2 or more explanatory variables exhibit "multicollinearity" if there exist strong linear relationships between them.

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

The "variance inflation factor" is a measure of multicollinearity associated with some explanatory variable x_j .

How to calculate VIF for x_j :

- ① Regress x_j onto all other x_i ; ie fit models for x_j against each other x_i ;

② Then

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

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.

Generally, we remove x_j from the model if $VIF_j > 10 \Leftrightarrow R_j^2 > .90$.

In R, we can do

$$> lm(x ~ \underbrace{x_1 + x_2 + \dots + x_n}_{\substack{\text{variable we} \\ \text{are testing}}} + \underbrace{\dots}_{\text{other exp. variables}})$$

and check the multiple R-squared value.

CI FOR μ_{new}

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

In particular,

$$\hat{\mu}_{\text{new}} = \mathbf{x}_{\text{new}}^T \hat{\beta}.$$

Then, we show that

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

Proof. Recall

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

Then $\hat{\mu}_{\text{new}} = \mathbf{x}_{\text{new}}^T \hat{\beta}$ must also follow a normal distribution.
See that

$$\begin{aligned} E(\hat{\mu}_{\text{new}}) &= E(\mathbf{x}_{\text{new}}^T \hat{\beta}) \\ &= \mathbf{x}_{\text{new}}^T E(\hat{\beta}) \\ &= \mathbf{x}_{\text{new}}^T \beta \end{aligned}$$

Then

$$\begin{aligned} \text{Var}(\hat{\mu}_{\text{new}}) &= \text{Var}(\mathbf{x}_{\text{new}}^T \hat{\beta}) \\ &= \mathbf{x}_{\text{new}}^T \text{Var}(\hat{\beta}) \mathbf{x}_{\text{new}} \\ &= \sigma^2 \mathbf{x}_{\text{new}}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_{\text{new}}. \end{aligned}$$

This gives the result, so we're done. \square

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

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

PREDICTION INTERVAL FOR y_{new}

\bullet Idea: We may also wish to use our fitted model to predict the value of the response of a new unit of the population.

\bullet 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.

\bullet Thus our total variation is

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

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

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

CONFIDENCE & PREDICTION BANDS FOR THE SLR MODEL

\bullet For the SLR model,

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

\bullet Thus, the closer x_{new} is to \bar{x} , the narrower the CIs & PIs.

\bullet 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

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

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

\mathbb{I} = the indicator function.

\bullet For a variable with l category levels, we need $l-1$ indicator variables.

\bullet In particular, if X is a categorical variable that has l distinct values a_1, \dots, a_l , we can use the model

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

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

- $(x_1, \dots, x_{l-1}) = (0, \dots, 0)$ corresponds with $X=a_1$.

\bullet Then, each $\hat{\beta}_i$ corresponds to the difference in the estimated mean value of the response where $X=a_i$ relative to where $X=a_1$.

INFERENCE FOR PARAMETERS ASSOCIATED WITH INDICATOR VARIABLES

\bullet 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 .

\bullet 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

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

③ Thus we get

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

ORTHOGONAL X MATRIX DESIGNS

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

e.g. 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

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

Note the columns of X are orthogonal, and in particular

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

I_4 is the 4×4 identity matrix.

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

\therefore 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)

\therefore We can express the total sum of squares of the observations y_i as

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

\therefore 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.

\therefore In ANOVA methods of inference, we draw conclusions about the relative fit of models by comparing SS(Reg) & SS(Res) .

\therefore The greater SS(Reg) is compared to SS(Res) , the better the model fit.

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

\therefore The "coefficient of determination" is

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

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

F-TEST FOR MODEL PARAMETERS:

$$\beta_1 = \dots = \beta_p = 0$$

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

\therefore Method:

$$\textcircled{1} \quad H_0: \beta_1 = \dots = \beta_p = 0; \quad H_A: \exists j \text{ s.t. } \beta_j \neq 0$$

$\textcircled{2}$ Test statistic:

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

$\textcircled{3}$ Under H_0 , F has a F distribution with p, n-p-1 degrees of freedom.

$\textcircled{4}$ The F-test statistic & p-value are provided on the last line of the "summary" output for lm.

$\textcircled{5}$ We reject H_0 if $p < 0.05$.

ANOVA TABLE

We summarise the test $H_0: \beta_1 = \dots = \beta_p = 0$ 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)			

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

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

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

ADDITIONAL SUM OF SQUARES

Consider the "full" model of

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

and the "reduced" model that reflects the restrictions imposed by $\beta_1 = \dots = \beta_p = 0$, ie

$$y = \beta_0 + \beta_{k+1} x_{k+1} + \dots + \beta_p x_p + \epsilon.$$

ie "After accounting for β_1, \dots, β_k , does $\beta_{k+1}, \dots, \beta_p$ account for significant variation in y ?"

To determine which model is better, we can examine

$$SS(Reg)_{full} - SS(Reg)_{red}$$

or

$$SS(Res)_{red} - SS(Res)_{full}$$

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

We can test $H_0: \beta_1 = \dots = \beta_p = 0$ 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 , $F \sim F_{df_{red}-df_{full}, df_{full}}$.

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

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
```

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

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

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

\circlearrowleft Suppose we have a model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

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

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

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

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

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

GENERAL LINEAR HYPOTHESIS

\circlearrowleft The "general linear hypothesis" is in the form

$$H_0: A\beta = 0$$

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

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

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

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

$$H_0: \underbrace{\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}}_A \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}}_{\beta} = \underbrace{\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}}_0$$

ASSESSING MODEL ADEQUACY / RESIDUAL ANALYSIS

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

RESIDUAL PLOTS

\circlearrowleft We can plot the residuals e_i against the fitted values \hat{y}_i :



\circlearrowleft If the model assumptions hold, then e_i and \hat{y}_i should be uncorrelated.

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

QQ PLOTS

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

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

\circlearrowleft 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

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

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

\circlearrowleft These are useful when we can write

$$\sigma^2 = f(\mu).$$

PROPERTIES OF THE RESIDUALS

Recall we can express

$$e = (I - H)e$$

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

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

$$e \sim N(0, \sigma^2 (I - H))$$

$$\Rightarrow e_i \sim N(0, \sigma^2 (1 - h_{ii}))$$

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

Note

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

- residuals have non-constant variance

② $\text{Cov}(e_i, e_j) = -\sigma^2 h_{jk}, j \neq k$

- residuals are not independent

- result of constraint $\sum e_i = 0$ in least squares estimation

STUDENTIZED RESIDUALS

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

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

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

A "studentized residual" associated with observation i is

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

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

EXTREME RESPONSE VALUES / OUTLIERS

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

In particular, we can consider an observation an outlier if

$$|d_i| > 2.5.$$

Causes of outliers:

- ① Typos / misrecording of data;
- ② Values of associated potential explanatory variables not included in the model;
- ③ Random variability.

LEVERAGE

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

In particular, for observation i ,

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

Also note

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

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

Residual plots are not useful in revealing high leverage points.

- as $h_{ii} \rightarrow 1$, $\text{Var}(e_i) \rightarrow 0$

- so high leverage observations' leverage is close to 0.

Instead, we can just plot the leverage values.

- in R, we can use "hatvalues".

We say observation i has high leverage if

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

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

We can quantify influence using "Coo