STAT 331 Personal Notes

Marcus Chan

Taught by Peter Balka

UW CS '25

Chapter 1: Introduction

REGRESSION

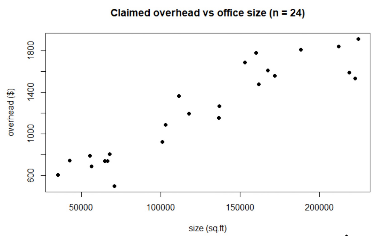
In regression modelling, we attempt to explain or account for variation in a response variate (y) by using a model to describe the relationship between y and one or more explanatory variates ($x_1, x_2, \ldots$)

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.
($x_1, \ldots, x_n$ is explanatory data; $y_1, \ldots, y_n$ 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: $\hat{\sigma}^2$

1. In the normal model, we assume

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

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.
3. 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: $\hat{\sigma}$

1. The "residual standard error" is

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

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

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

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

INFERENCE FOR β_1

1. To investigate whether there is a linear relationship between y & x in the population, we can test the hypothesis "$\beta_1 = 0$."
2. We can then either use confidence intervals or hypothesis tests to test this.
3. 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 $\hat{\beta}_1$

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

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

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

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

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

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,

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

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

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

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

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

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

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 $(1-\alpha)100\%$.

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

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 $(126.33 - t(10.88),\ 126.33 + t(10.88))$.

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).

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).

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

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

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

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

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

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

which gives us the desired result.

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

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

3. We also have that

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

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 $p < 0.05$.