STAT 331 Personal Notes

Marcus Chan
Taught by Peter Balka
UW CS '25

Chapter 1: Introduction

REGRESSION

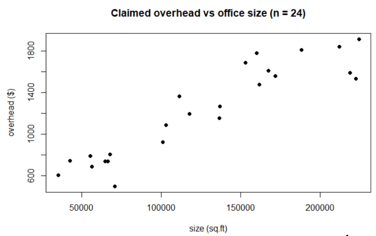
In regression modelling, we attempt to explain or account for variation in a response variate (y) by using a model to describe the relationship between y and one or more explanatory variates ($x_1, x_2, \dots$)

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

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

for some variance σ^2.

2. This yields the normal model

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

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

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

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

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

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

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

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

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

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

* 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

$$\sum e_i = 0$$
$$\sum x_i e_i = 0.$$

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 the 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 $\hat{\beta}_1$

1. We can show for the SLR model that

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

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

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

and so

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

Then

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

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

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

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

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

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

Similarly,

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

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

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

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

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

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

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.