richardwu.ca

STAT 331 Course Notes

Applied Linear Models

Peter Balka • Winter 2018 • University of Waterloo

Last Revision: March 4, 2018

Table of Contents

Abstract

These notes are intended as a resource for myself; past, present, or future students of this course, and anyone interested in the material. The goal is to provide an end-to-end resource that covers all material discussed in the course displayed in an organized manner. These notes are my interpretation and transcription of the content covered in lectures. The instructor has not verified or confirmed the accuracy of these notes, and any discrepancies, misunderstandings, typos, etc. as these notes relate to course's content is not the responsibility of the instructor. If you spot any errors or would like to contribute, please contact me directly.

1 January 4, 2018

1.1 Simple linear regression review

In SLRM, there is a single explanatory variate and a response variate.
A good graphical summary for SLRM are **scatterplots**.
A good numerical summary for SLRM is the **correlation coefficient** defined as

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

where $-1 \leq r \leq 1$. If $|r| \approx 1$ then the explanatory/response variates have a strong linear relationship.

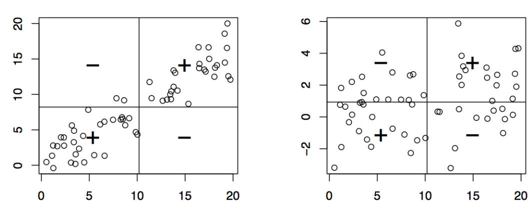
2 January 9, 2018

2.1 Correlation coefficient and covariance

Note: the measure r is also the covariance divided by the standard deviations or

$$r = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Note that the covariance $E[(X - E[X])(Y - E[Y])]$ can be graphically separated by the means $\bar{X}$ and $\bar{Y}$.



One can see that the covariance signage is determined by the sum of the magnitudes in the positive and negative quadrants.

2.2 Simple linear regression (SLR) model

An SLR model can be thought of as a line with covariates x and y where

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \ldots, n$$

where ϵ_i is some error term for each i.

Example 2.1. From the dataset

Overhead	Office Size
218955	1589
224513	1912
$\vdots$	$\vdots$

Thus we have the SLR model

$$\begin{aligned} 218955 &= \beta_0 + \beta_1(1589) + \epsilon_1 \\ 224513 &= \beta_0 + \beta_1(1912) + \epsilon_2 \end{aligned}$$

2.3 Methods of least squares

Find (estimate) the value of β_0, β_1 (denoted by $\hat{\beta}_0, \hat{\beta}_1$, respectively) that minimizes the sum of squares of the errors $\sum_{i=1}^n \epsilon_i^2$. That is: we find values of β_0, β_1 that minimizes the function

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

We take the partial derivatives and set to 0 to find the minimum (assuming convexity)

$$\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2\sum_{i=1}^n y_i - (\beta_0 + \beta_1 x_i) = 0 \\ \frac{\partial S}{\partial \beta_1} &= -2\sum_{i=1}^n x_i[y_i - (\beta_0 + \beta_1 x_i)] = 0 \end{aligned}$$

which yields (the notation changes to estimates of β assuming we can calculate those)

$$\begin{aligned} \sum_{n=1}^n y_i &= n\hat{\beta}_0 + \sum_{n=1}^n x_i \hat{\beta}_1 \\ \sum_{n=1}^n x_i y_i &= \sum_{i=1}^n x_i \hat{\beta}_0 + \sum_{n=1}^n x_i^2 \hat{\beta}_1 \end{aligned}$$

which gives us the estimates

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

where the second equation follows from substituting in the first and re-deriving for $\frac{\partial S}{\partial \beta_1}$.
The corresponding fitted line is

$$\hat{\mu}_{y|X=x} = \hat{\mu} + \hat{\beta}_0 + \hat{\beta}_1 x$$

For the example with overhead above, we'd have

$$\hat{\mu} = -27877.06 + 126.33x$$

2.4 Fitted residuals

These are the difference between the actual values and our fitted value (distinct from the error terms previously)

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

Some **key points** regarding this model

- By estimating two parameters (β_0, β_1), we have imposed two constraints on our residuals (from our partial derivatives)

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

These reduces our number of n independent measures by 2 since we can compute the remaining two residuals from $n-2$ observations. Thus we have $n-2$ **degrees of freedom** (or in general, $n-k$ dfs where k is the number of estimated parameters>).

2.5 Interpretation of estimated parameters $\hat{\beta}_i$

β_1

$$\begin{aligned}
\hat{\mu} &= \hat{\beta}_0 + \hat{\beta}_1 x \\
\hat{\mu_{x+1}} &= \hat{\beta}_0 + \hat{\beta}_1(x+1) \\
&= \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_1 \\
&= \hat{\mu} + \hat{\beta}_1
\end{aligned}$$

thus $\hat{\beta}_1$ can be interpreted as the estimated mean change in the response (y) associated with one unit change of x.

β_0 For $x = 0$, $\hat{\mu} = \hat{\beta}_0$.

However, in the example with overhead, it's evident that when $x = 0$ overhead is negative (-27877.06) which is nonsensical.

Never extrapolate results outside the range of the values of the explanatory variate(s).

3 January 16, 2018

3.1 Invariants for normal SLR models

Recall for the normal SLR model we have

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \dots, n$$

where $\epsilon_i \sim N(0, \sigma^2)$ is some error term for each i.

- $\beta_0 + \beta_1 x_i$ is the **deterministic** and ϵ_i is the **random** components of the model.
- $Var(\epsilon_i) = \sigma^2$ for all i (constant variance)
- ϵ_i, ϵ_j for $i \neq j$ are independent (otherwise we'd need time series)

3.2 Estimate of variance in SLR

Each of our error terms follow a $N(0, \sigma^2)$ distribution. The **unbiased** estimate of σ^2 is

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

The **residual standard error** is $\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-2}}$.

3.3 Unbiased estimator of $\hat{\beta}_1$

The **estimator** of $\hat{\beta}_1$ is a random variable $\hat{\beta}_1$ (usually denoted with a big B) that is similar to the estimate but with r.v. Y_i and $\bar{Y}$

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

Note: $\hat{\beta}_1$ can be expressed as a linear combination of response variables Y_i, $i = 1, 2, \ldots, n$.

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum(x_i - \bar{x})Y_i - \bar{Y}\sum(x_i - \bar{x})}{S_{xx}} \\
&= \frac{\sum(x_i - \bar{x})Y_i}{S_{xx}} && \sum(x_i - \bar{x}) = 0 \\
&= \sum_{i=1}^n c_i Y_i && c_i = \frac{(x_i - \bar{x})}{S_{xx}}
\end{aligned}$$

Remember that

$$\begin{aligned}
&\epsilon_i \sim N(0, \sigma^2) \text{ independent } \Rightarrow Y_i \sim N(\beta_0 + \beta_1 x, \sigma^2) \text{ independent} \\
\Rightarrow &\hat{\beta}_1 \sim \text{ Normal (sum of independent normal r.v.'s)}
\end{aligned}$$

Thus we have

$$\begin{aligned}
E(\hat{\beta}_1) = E(\sum c_i Y_i) &= \sum c_i E(Y_i) \\
&= \sum(\frac{(x_i - \bar{x})}{S_{xx}})(\beta_0 + \beta_1 x_i) \\
&= \frac{\beta_0 \sum(x_i - \bar{x}) + \beta_1 \sum x_i(x_i - \bar{x})}{S_{xx}} \\
&= \frac{\beta_1 \sum x_i(x_i - \bar{x}) - \beta_1 \bar{x}\sum(x_i - \bar{x})}{S_{xx}} && \text{eliminate and introduce 0 term} \\
&= \frac{\beta_1 \sum(x_i - \bar{x})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \\
&= \beta_1
\end{aligned}$$

Since $E(\hat{\beta}_1) = \beta_1$, then $\hat{\beta}_1$ is an unbiased estimator of β_1.

The variance of our estimator $\hat{\beta}_1$ is

$$\begin{aligned}
Var(\hat{\beta}_1) &= Var(\sum c_i Y_i) \\
&= \sum c_i^2 Var(Y_i) && Y_i \text{ independent} \\
&= \sum \frac{\sigma^2 (x_i - \bar{x})^2}{S_{xx}^2} \\
&= \frac{\sigma^2}{S_{xx}} \\
&= \frac{\sigma^2}{\sum (x_i - \bar{x})^2}
\end{aligned}$$

Since our estimator $\hat{\beta}_1$ follows (from above)

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

we have

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0,1)$$

in terms of the sample variance (or estimate $\hat{\sigma}$ we have the **T-distribution**)

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{\sigma}}{\sqrt{S_{xx}}}} \sim t_{n-2}$$

3.4 Identities of distributions

Recall that the distribution of the sample means follows a normal distribution

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

so we have

$$\begin{aligned}
\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} &\sim N(0,1) \\
\frac{\bar{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} &\sim t_{n-1}
\end{aligned}$$

This follows from

$$\begin{aligned}
SD(X) &= \sigma \\
SE(X) &= \hat{\sigma} \\
SD(\bar{X}) &= \frac{\sigma}{\sqrt{n}} \\
SE(\bar{X}) &= \frac{\hat{\sigma}}{\sqrt{n}} \\
\frac{\bar{X}-\mu}{SE(\bar{X})} &\sim t_{n-1} \\
\frac{\hat{\beta}_1-\beta_1}{SE(\hat{\beta}_1)} &\sim t_{n-2}
\end{aligned}$$

4 January 18, 2018

4.1 Inference for β_1 in SLR

"Is there a relationship between overhead and office size (for the population of offices)?"
There is no (linear) relationship $\iff \beta_1 = 0$.
We can statistically check this using two methods

1. Confidence interval for β_1
2. Hypothesis test for β_1 ($H_0 : \beta_1 = 0$)

4.2 Confidence interval for SLR

General example, not necessarily SLR: For a distribution with one parameter μ, we can calculate the $(1-\alpha)100\%$ confidence interval for μ (note: we need only one t value since the T-distribution is symmetric)

$$\begin{aligned}
&\hat{\mu} \pm t_{n-1,1-\frac{\alpha}{2}} \cdot SE(\hat{\mu}) \\
\Rightarrow &\bar{x} \pm t_{n-1,1-\frac{\alpha}{2}} \left(\frac{\hat{\sigma}}{\sqrt{n}}\right)
\end{aligned}$$

where $\bar{x}$ is the sample mean of the distribution.
By a similar line of logic, we can produce confidence intervals for our parameters. The $(1-\alpha)100\%$ confidence interval for β_1 (where the we have $n-2$ degrees of freedom)

$$\hat{\beta}_1 \pm t_{n-2,1-\frac{\alpha}{2}} \cdot SE(\hat{\beta}_1)$$

Example 4.1. The 95% C.I. for β_1 for overhead data is

$$\begin{aligned}
&\hat{\beta}_1 \pm t_{22,0.975} SE(\hat{\beta}_1) \\
=&126.33 \pm 2.074(10.88) \\
=&126.33 \pm 22.57 = (103, 76, 148.90)
\end{aligned}$$

where ± 22.57 is the **margin of error**.

Since $\beta_1 = 0$ is not in the interval, we can conclude that there is a *significant* positive relationship between overhead and office size.

Remark 4.1. An X% confidence interval can be interpreted as: X% of X% confidence intervals established from repeated samples contain the true value.
In other words: they are intervals constructed from a procedure that will contain the population mean for a specified proportion of the time (X% of the time).

4.3 Hypothesis testing for SLR

We form a **null hypthoesis** H_0 and an alternative hypothesis H_1, where we assume H_0 unless there is statistical significance rejecting H_0.
For simple linear regression, we suppose

$$\begin{aligned} H_0 &: \beta_1 = 0 && \text{no relationship} \\ H_1 &: \beta_1 \neq 0 && \text{two-sided alternative} \end{aligned}$$

Our **test statistic** t is the distribution

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

Example 4.2. Under H_0 we have for our example

$$t = \frac{126.33 - 0}{10.88} = 11.61$$

If we look at our t_{22} distribution, we find the total probability of the pdf at $P(t \leq 11.61)$ and $P(t \geq 11.61)$ (the **p-value**).
We see that $P(t_{22} > 2.819) = 0.0005 \Rightarrow P(t_{22} > 11.61) << 0.005$.
Thus the p-value is $2P(t_{22} > 11.61) << 0.01$ (in fact, it is 7.47×10^{-11}), which is lower than **0.05 (the significance level)**, so we reject the null hypothesis.

Remark 4.2. The p-value of a hypothesis test can be interpreted as: the probability that our sample holds (the observed, or more extreme, results) under the null hypothesis. If it is extremely low (past a certain threshold), then we may reject the null hypothesis as not possible.

4.4 Two-sided vs one sided tests

The reason why we took both CDF ends of the T-distribution in the example above is to account for a $\hat{\beta}_1$ equally as extreme but on the negative side. Since we assume all this happens to chance, $\hat{\beta}_1$ could equally be the same magnitude but with a negative sign.

4.5 Confidence interval vs hypothesis testing

Deciding which method to use is problem dependent: usually, hypothesis testing is simpler to interpret for many variates and a confidence interval is only relevant for single variates.
A 95% confidence interval corresponds with a hypothesis test with a 0.05 significance level i.e. we will derive an equivalent conclusion.

4.6 Multiple linear regression (MLR) model

We want to model the following relationship

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

where $\epsilon_i = N(0, \sigma^2)$ and independent.
Note we have p variates and $p+1$ parameters (the bias term) thus we have $n-(p+1)$ degrees of freedom.
In matrix form, this is represented as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

which can be written succinctly as

$$Y = X\beta + \epsilon$$

where $\epsilon = N(\vec{0}, \sigma^2 I)$ or $Var(\epsilon) = \sigma^2 I$ (the **covariance matrix**; note that the covariance between ϵ_i, ϵ_j $i \neq j$ is 0 since they are independent).

5 January 23, 2018

5.1 Least squares estimation of β in MLR

Our residual expression is now, for n observations and p explanatory covariates

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

Taking the partial derivatives with respect to each β_j and finding the minimum

$$\begin{aligned}
\frac{\partial S}{\partial \beta_0} &= -2 \sum [y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}] = 0 \\
\frac{\partial S}{\partial \beta_1} &= -2 \sum x_{i1}[y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}] = 0 \\
&\vdots \\
\frac{\partial S}{\partial \beta_p} &= -2 \sum x_{ip}[y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}] = 0
\end{aligned}$$

which can also be expressed as

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

In matrix form, this is written as

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

yields the best square estimate

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

assuming X is of full rank (i.e. $p+1$ linearly independent columns).

5.2 Estimate of variance in MLR

We can estimate the variance for the error terms (or the variance of the random component in our model) by taking the sum of squared residuals and dividing by the degrees of freedom

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-(p+1)} = \frac{\sum[y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})]^2}{n-(p+1)}$$

The **residual standard error** is the square root of this or

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

5.3 Hat matrix

The **hat matrix** (also known as the *influence matrix*) maps our responses to predicted values.
Given our predicted mean responses

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

The matrix H is the hat matrix

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

Some properties of H are:

H is symmetric ($H = H^T$) Note that

$$\begin{aligned} H^T = [X(X^TX)^{-1}X^T]^T &= X[(X^TX)^{-1}]^TX^T && (AB)^T = B^TA^T \\ &= X[(X^TX)^T]^{-1}X^T && (A^{-1})^T = (A^T)^{-1} \\ &= X(X^TX)^{-1}X^T \\ &= H \end{aligned}$$

H is idempotent ($H = HH$)

$$\begin{aligned} HH &= (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) \\ &= X[(X^TX)^{-1}(X^TX)](X^TX)^{-1}X^T \\ &= X(X^TX)^{-1}X^T \\ &= H \end{aligned}$$

Note that

$$\hat{\mu} = Hy$$

where our residual is

$$\begin{aligned} e &= y - \hat{\mu} \\ &= y - Hy \\ &= (I - H)y \end{aligned}$$

The residuals are a linear combination of our responses.

So we have

$$\begin{aligned} y &= \hat{\mu} + e \\ &= Hy + (I - H)y \end{aligned}$$

where Hy is orthogonal to $(I-H)y$ (that is: $(Hy)^T(I-H)y = 0$ - follows by expansion and the fact that $H^TH = H$). This implies that $\hat{\mu}_i$ and e_i are independent and thus

$$Cov(\hat{\mu}_i, e_i) = 0$$

5.4 Inference for β in MLR

To infer the meaning of the model parameters $(\beta_0, \beta_1, \ldots, \beta_p)$, we note that

$$\epsilon \sim (0, \sigma^2 I) \Rightarrow Y \sim N(XB, \sigma^2 I)$$

since $Y = X\beta + \epsilon$.
The **distribution of** $\hat{\beta}$ is thus

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

so $\hat{\beta} \sim Normal$.
Its model parameters are

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

and for the variance

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

thus $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$. Note that for a specific β_j, its marginal distribution is

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

where $SE(\hat{\beta}_j) = \hat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}$.
One can see that the variance is not constant (some parameters will be estimated with a larger confidence interval) since

$$Var(\hat{\beta}_j) = \sigma^2(X^TX)^{-1}_{jj}$$

where the diagonal entries are not the same.
It is often common for β_j to change as more covariates are added to a multiple linear model. This implies each

explanatory covariate are correlated and thus

$$Cov(\hat{\beta}_j, \hat{\beta}_k) = \sigma^2 (X^T X)^{-1}_{jk} \neq 0$$

The covariance of β_j, β_k $j \neq k$ can all be 0 if all explanatory variates are independent.

Remark 5.1. Covariate x_j, x_k are independent *iff* $Cov(\hat{\beta}_j, \hat{\beta}_k) = 0$.

The β_js can be interpreted as: keeping all other covariates in the model constant, what is the mean response of my covariate x_j? In effect, multiple linear regression corrects for other covariates.

5.5 Confidence interval in MLR

Note: this is a confidence interval for the parameter β_j, not the estimate $\hat{\beta}_j$.
For a $(1-\alpha)100\%$ confidence interval we have

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

Example 5.1. For example, the 95% CI for β_1 (size) in the overhead example is

$$\begin{aligned}&\hat{\beta}_1 \pm t_{18,0.975} SE(\hat{\beta}_1)\\ =&31.26 \pm 2.101(21.47)\\ =&31.26 \pm 45.11 \Rightarrow (-13.85, 76.37)\end{aligned}$$

since the CI encompasses 0, we conclude there is no significant relationship of size with respect to overhead *after accounting for other covariates.*

5.6 Hypothesis testing in MLR

The null hypothesis for testing if a covariate is related to the response is

$$H_0 : \beta_j = 0$$

where we have the test statistic

$$t = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} \sim t_{n-(p+1)}$$

under H_0.

Example 5.2. For β_1 (size), we have

$$t = \frac{31.26}{21.47} = 1.46$$

From the t-distribution table for $n = 18$, we see that this corresponds to a p-value between 0.1 and 0.2 (0.1625 to be exact).
We therefore do not reject H_0 (p-value > 0.05) i.e. there is no significant relationship between overhead nad size, after accounting for the other variates.

6 January 25, 2018

6.1 Scater plot matrix

For a given set of explanatory variates and a response variate, we can plot a matrix of 2D scatter plots of each variate against all the other variates.

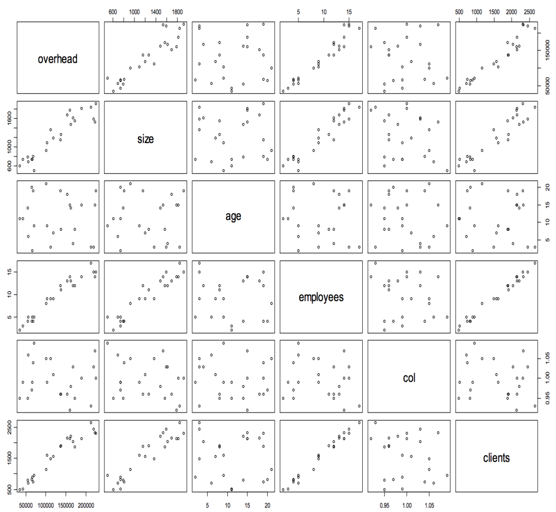


Figure 6.1: Size, employees, and clients are all correlated with overhead. Note however that size, employees, and clients are all correlated with each other therefore it would probably suffice to only include one of these explanatory variates without losing much information in our model.

From this matrix, we can visually see which explanatory variates are correlated to the explanatory variate but also which explanatory variates are correlated with each other.

6.2 Multicollinearity

When strong (linear) relationships are present among two or more explanatory variates, we say the variates exhibit **multicollinearity**.
Intuitively, multicollinearity means some explanatory variates are dependent and it would not be required to have all the extraneous dependent variates in model since they do not introduce much additional explained variance/information.
In fact, multicollinear is **detrimental**: it leads to inflated variances of the associated parameter estimates ($(X^TX)^{-1}$ has inflated diagonal entries, thus $SE(\hat{\beta}_j) = \hat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}$ is inflated), resulting in inaccurate conclusions from hypothesis tests and confidence intervals (which depend on $SE(\hat{\beta}_j)$) (intuitively, our estimate of the impact of one unit change of x_j, $\hat{\beta}_j$, while controlling for the others tend to be less precise since there is some dependency happening when "changing" x_j with another correlated x_k).

6.3 Variance inflation factor (VIF)

To assess whether a variate x_j is a problem in terms of multicollinearity, we can regress x_j onto all other explanatory variates. We can then caclulate the **variance inflation factor** for x_j

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

The VIF$_j$ can be interpreted as the factor by which the variance of $\hat{\beta}_j$ is increased relative to the ideal case in which all explanatory variates are uncorrelated (i.e. columns of X are orthogonal).

Example 6.1. Suppose we do this for $x_j = x_3$: we regress the number of employees on all other explanatory variates (see scatter plot matrix above).
We have $R_3^2 = 0.9855$, thus we have a VIF of $\frac{1}{1-R_3^2} = 68.97$. So the variance is inflated $\approx 69x$ because of multicollinearity (compared to the case where we just have x_3).

As a general rule of thumb: multicollinearity is a serious problem if VIF > 10 (or thereabouts), which corresponds to an $R_j^2 > 0.9$.

7 January 30, 2018

7.1 Maximum likelihood estimation (MLE)

A remark on least squares estimation of β: for a model with *normal errors*, *maximum likelihood estimation (MLE)* and *least squares estimation (LSE)* are equivalent.
The **maximum likelihood estimation** is defined as

$$\begin{aligned} L(\beta_0, \beta_1, \ldots, \beta_p \mid y_1, \ldots, y_n) &= \prod_{i=1}^n P(y_i) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-\mu_i)^2}{2\sigma^2}} \\ &= (2\pi\sigma^2)^{\frac{-n}{2}} e^{\frac{-\sum(y_i-\mu_i)^2}{2\sigma^2}} \qquad \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \end{aligned}$$

Taking the log likelihood function

$$l = \log(L) = c - \frac{\sum[y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}]^2}{2\sigma^2} \qquad = S(\beta_0, \beta_1, \ldots, \beta_p) = \sum \epsilon_i^2 \quad \text{from LSE}$$

7.2 Gauss-Markov theorem

Consider the model given by $Y = X\beta + \epsilon$ where $E(\epsilon) = 0, Var(\epsilon) = \sigma^2 I$. The G-M theorem states that among all unbiased linear estimators $\hat{\beta}^* = M^* Y$, the LSE given by $\hat{\beta} = MY$ (where $M = (X^T X)^{-1} X^T$ in LSE) has the smallest variance.
That is

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

where $(M^* - M)(M^* - M)^T$ is a positive semidefinite matrix (a matrix A is positive semidefinite if $a^T A a \geq 0$ for any vector a).

7.3 Confidence interval for μ_{new}

Example 7.1. Provide an interval in which the mean overhead of a 1000 sq ft office that is 12 years old, has a *col* $= 1.02$ and 1300 clients lies.

Note we can find the **confidence interval for** μ_{new} or a new mean response.

$$\hat{\mu}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new,1} + \ldots + \hat{\beta}_p x_{new,p}$$

which is in vector form: $x_{new}^T\hat{\beta}$ where $x_{new}^T = (1, x_{new,1}, \ldots, x_{new,p})$.
The distribution of $\hat{\mu}_{new}$ can be derived. Recall that

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

so we know that $\hat{\mu}_{new} \sim Normal$. Furthermore

$$\begin{aligned}
E[\hat{\mu}_{new}] &= \mu_{new} = x_{new}^T\beta \\
Var(\hat{\mu}_{new}) &= Var(x_{new}^T\hat{\beta}) \\
&= x_{new}^T Var(\hat{\beta}) x_{new} \\
&= \sigma^2 x_{new}^T (X^TX)^{-1} x_{new}
\end{aligned}$$

Thus we have

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

which has the corresponding pivotal distribution

$$\frac{\hat{\mu}_{new} - \mu_{new}}{\hat{\sigma}\sqrt{x_{new}^T(X^TX)^{-1}x_{new}}} \sim t_{n-(p+1)}$$

Thus the $(1-\alpha)100\%$ CI for μ_{new} is

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

Example 7.2. In the overhead model, we have $x_{new}^T = (1, 1000, 12, 1.02, 1300)$.
So the 95% CI for μ_{new} is

$$(97460.07, 112202.30)$$

where $\hat{\mu}_{new} = 104831.2$ and the margin of error is 7371.1.

Remark 7.1. A confidence interval only establishes an estimate interval *for a population parameter*, but not a particular random variable. We would need to use a **prediction interval** to establish an estimate for Y_{new}.

7.4 Prediction interval for Y_{new}

Example 7.3. An office is 1000 sq ft, 12 years old, with 1300 clients and a *col* = 1.02. Provide an interval for the overhead of **this (particular) office**.

Remark 7.2. This question is different than the previous one since it asks for an interval for a particular office *rather than the mean overhead* of an office of this characteristic in the population.

Consider the prediction error given by $Y_{new} - \hat{\mu}_{new}$. Thus we have

$$\begin{aligned}
Var(Y_{new} - \hat{\mu}_{new}) &= Var(Y_{new}) + Var(\hat{\mu}_{new}) && \text{independence} \\
&= \sigma^2 + \sigma^2 x_{new}^T(X^TX)^{-1}x_{new} \\
&= \sigma^2(1 + x_{new}^T(X^TX)^{-1}x_{new})
\end{aligned}$$

where $Y_{new}, \hat{\mu}_{new}$ are independent since any new observations do not depend on our estimate.

Thus the $(1-\alpha)100\%$ prediction interval for Y_{new} is

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

Example 7.4. In the overhead model, we still have the same x_{new}^T so we get for the 95% prediction interval

$$(73946.20, 135715.70)$$

where $\hat{\mu}_{new} = 104831.2$ (same as CI) and the margin of the error is 30884.5 (much larger than the MoE in the CI).

Note that for the SLR model, the confidence interval and the prediction interval standard errors reduce to

$$\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$$

and

$$\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$$

respectively. Note that the errors are smaller as x_{new} is closer to the mean/centre $\bar{x}$ as we can see in the prediction and confidence bands.

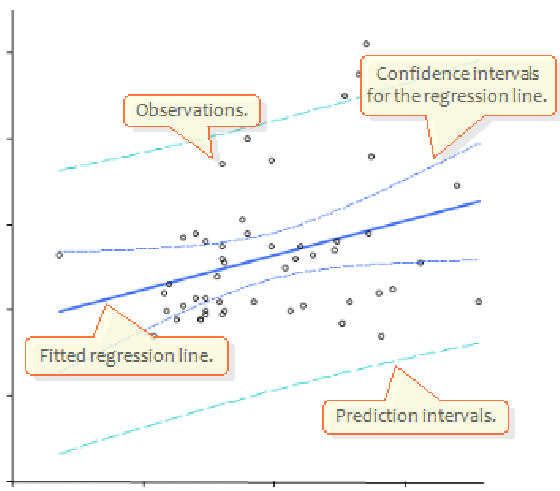


Figure 7.1: The confidence and prediction bands are smaller in closer to the centre of the x's or closer to $\bar{x}$. Furthermore, the prediction bands lie further out from the confidence bands.

8 February 1, 2018

8.1 Modelling categorical variates

Example 8.1. Promotion study: does a wing promotion have any effect on sales? Do different types of promotion affect sales differently?
The sampling protocol is as follows:

- 30 stores randomly selected from population
- 10 stores are randomly assigned to one of three promotion types: promo1, promo2, no promotion (control)

- response variate: change (%) in sales over two-week period of study

One **inappropriate approach may be**:

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

where

$$x_i = \begin{cases} 1 & \text{if ith store uses promo1} \\ 2 & \text{if ith store uses promo2} \\ 3 & \text{if ith store has no promo} \end{cases}$$

There may be no linear relationship depending on the way we assign x_i (e.g. promo 1 has a higher mean response, promo 2 has a lower mean response, no promo has a higher mean response). This model is too **restrictive**.
A more *flexible model*:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

where $x_{i1} = 1$ if ith store uses promo1 (0 otherwise) and $x_{i2} = 1$ if ith store uses promo2.
This is similar to *one-hot encoding* and these are called **indicator or dummy variates**.
Our data might look like

store(i)	x_{i1}	x_{i2}
1	0	0
2	0	0
$\vdots$	$\vdots$	$\vdots$
10	1	0
11	1	0
$\vdots$	$\vdots$	$\vdots$
21	0	1
22	0	1
$\vdots$	$\vdots$	$\vdots$
30	0	1

Suppose we consider adding $x_{i3} = 1$ when the ith store has no promo and 0 otherwise. Then we have our X matrix as

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \end{bmatrix} \tag{8.1}$$

note that $x_{i3} = 1 - (x_{i1} + x_{i2})$ so we have a linear dependent column.
This implies $\text{rank}(X) = 3$ which is not of full rank, thus $X^T X$ is not invertible.

To interpret/inference our parameters, note that the estimate response or estimated change of sales (in %) is given by

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

So for a store that does not have any promos

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1(0) + \hat{\beta}_2(0) = \hat{\beta}_0$$

Similarly for promo 1 stores

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

and for the promo 2 stores

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_2$$

From our data, we may get the regression summary

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.870      1.665   -0.523  0.60552
x1             8.350      2.354    3.547  0.00145 **
x2             2.970      2.354    1.261  0.21792
```

We can't conclude anything about the control case (no promo) and the promo 2 group, but we can conclude that the estimated increase in sales (relative to the control) using promo1 is 8.35% (p-value < 0.05).
More formally, is there a **difference in mean increase in sales** between no promo and promo1 stores? We can assume the null hypothesis $H_0 : \beta_1 = 0$ (no change in promo1 sales) and alternative hypothesis $H_a : \beta_1 \neq 0$.
Thus we have the test statistic

$$\begin{aligned} t &= \frac{\hat{\beta}_1 - \beta_1}{SE\hat{\beta}_1} \\ &= \frac{\hat{\beta}_1 - 0}{SE\hat{\beta}_1} \\ &= 3.547 \end{aligned}$$

We can look up the T-distribution with $30 - 3 = 27$ degrees of freedom to figure out that the p-value is 0.00145. We *reject* H_0: so using promo1 is associated with a significantly higher mean sales than no wing promotion.
A more nuanced question: is there a difference in mean sales between promo1 and promo2? This is not quite clear from our regression summary. Thus we use hypothesis testing with null hypothesis $H_0 : \beta_1 - \beta_2 = 0$. What about our test statistic?
One approach is

$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{SE(\hat{\beta}_1 - \hat{\beta}_2)} \sim t_{27}$$

under H_0. But what is the standard error? We need to take the variance of $\hat{\beta}_1 - \hat{\beta}_2$. Recall that the variances of $\hat{\beta}$ are

$$\begin{aligned} &\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1}) \\ &\hat{\beta}_j \sim N(\beta_j, \sigma^2 (X^T X)^{-1}_{jj}) \\ &\Rightarrow Cov(\hat{\beta}_j, \hat{\beta}_k) = \sigma^2 (X^T X)^{-1}_{jk} \end{aligned}$$

so we have

$$\begin{aligned}
Var(\hat\beta_1 - \hat\beta_2) &= Var(\hat\beta_1) + Var(\hat\beta_2) - 2Cov(\hat\beta_1, \hat\beta_2) \\
&= \sigma^2 (X^TX)^{-1}_{11} + \sigma^2 (X^TX)^{-1}_{22} - 2\sigma^2 (X^TX)^{-1}_{12} \\
&= \sigma^2 [(X^TX)^{-1}_{11} + (X^TX)^{-1}_{22} - 2(X^TX)^{-1}_{12}]
\end{aligned}$$

Thus our standard error is

$$SE(\hat\beta_1 - \hat\beta_2) = \hat\sigma \sqrt{(X^TX)^{-1}_{11} + (X^TX)^{-1}_{22} - 2(X^TX)^{-1}_{12}}$$

Another more general approach is the **F-test (ANOVA)**.

9 February 6, 2018

9.1 X matrices with orthogonal columns

$Y = X\beta + \epsilon$. Suppose we are designing an experiment where the response is the shrinkage of a part (%) during molding process. There are 3 factors: Temp(L,H), Pressure(L,H), Speed(L,H). We thus have 2^3 unique experimental runs

run	T	P	S
1	L	L	L
2	L	L	H
3	L	H	L
4	L	H	H
5	H	L	L
6	H	L	H
7	H	H	L
8	H	H	H

Therefore we have as our model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon$$

where

$$x_{i1} = \begin{cases} 1 & \text{if ith run uses H temp} \\ 0 & \text{otherwise (L temp)} \end{cases}$$
$$x_{i2} = \begin{cases} 1 & \text{if ith run uses H pressure} \\ 0 & \text{otherwise (L pressure)} \end{cases}$$
$$x_{i3} = \begin{cases} 1 & \text{if ith run uses H speed} \\ 0 & \text{otherwise (L speed)} \end{cases}$$

There is an **alternative coding** scheme where we use -1 instead of 0

$$x_{i1} = \begin{cases} 1 & \text{if ith run uses H temp} \\ -1 & \text{otherwise (L temp)} \end{cases}$$
$$x_{i2} = \begin{cases} 1 & \text{if ith run uses H pressure} \\ -1 & \text{otherwise (L pressure)} \end{cases}$$
$$x_{i3} = \begin{cases} 1 & \text{if ith run uses H speed} \\ -1 & \text{otherwise (L speed)} \end{cases}$$

So our X matrix becomes

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

Remark 9.1. All the columns in X are orthogonal.

By noting columns i, j where $i \neq j$ are orthogonal, we can easily see that

$$X^T X = \begin{bmatrix} 8 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix}$$

Taking the inverse, we get

$$(X^T X)^{-1} = \begin{bmatrix} \frac{1}{8} & 0 & 0 & 0 \\ 0 & \frac{1}{8} & 0 & 0 \\ 0 & 0 & \frac{1}{8} & 0 \\ 0 & 0 & 0 & \frac{1}{8} \end{bmatrix}$$

10 February 8, 2018

10.1 Interpretation of parameters for categorical $-1, 1$ variates

Previously we chose $-1, 1$ as values for our indicator variates as opposed to $0, 1$. How do we interpret the parameter estimates now that $x_{ik} \in \{1, -1\}$ (how do we interpret 1 unit of change which is now halved on this scale)?
Taking a look at the previous example we have for two different trials where Tep is toggled between L and H

$$\begin{aligned} \{L, L, L\} &: \hat{\mu} = \hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2 - \hat{\beta}_3 \\ \{H, L, L\} &: \hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 - \hat{\beta}_2 - \hat{\beta}_3 \\ &= (\hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2 - \hat{\beta}_3) + 2\hat{\beta}_1 \end{aligned}$$

So we have $\hat{\mu}_{HLL} = \hat{\mu}_{LLL} + 2\hat{\beta}_1$.

Suppose we get $\hat{\beta}_1 = -0.2875$ in the MLR. Then we have

$$2\hat{\beta}_1 = -0.5750$$

So holding all other factors constant running temperature at a high level is associated with an estimated decrease of 0.5750% in shrinkage compared to running at low temp (although since the p-value is ≥ 0.05, there is no statistical significant relationship).

10.2 Independence of indicator variates

Note from our $X^T X$ matrix, we see that $Cov(\hat{\beta}_j, \hat{\beta}_k) = \sigma^2 (X^T X)^{-1}_{jk} = 0$ for $j \neq k$.
So if we were to take out an explanatory variate, the parameter estimates will not change (no linear dependency between them). **However, it's possible for the p-value to change**.
For example, in a sample dataset we see that the p-value for pressure was 0.0978 (not significant) after correcting for temperature and speed, but was 0.0421 (significant) with just an SLR (just pressure regressed onto shrinkage).
Why is this the case? Well when we removed some variates, we **increased our degrees of freedom**. Algebraically we had

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

and for our standard errors for our parameter we had

$$SE(\hat{\beta}_j) = \hat{\sigma} \cdot \sqrt{(X^T X)^{-1}_{jj}}$$

thus in our reduced model, as the degrees of freedom increases as we take away parameters, our residual standard error decreases hence $SE(\hat{\beta}_j)$ decreases. This is especially true for small samples with low degrees of freedom.

Example 10.1. In our shrinkage vs temperature, pressure and speed we originally had in the "full" model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \quad \epsilon \sim N(0, \sigma^2) \text{ independent}$$

This corresponds to

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{4}} = 3.271 \Rightarrow SE(\hat{\beta}_j) = \hat{\sigma}\sqrt{(X^T X)^{-1}_{jj}} = 1.1563$$

By reducing our model down to only one explanatory variate pressure, we have

$$Y = \beta_0 + \beta_2 x_2 + \epsilon \quad \epsilon \sim N(0, \sigma^2) \text{ independent}$$

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{6}} = 2.733 \Rightarrow SE(\hat{\beta}_j) = 0.9662$$

Note however the p-value does not decrease in general when removing variates. This only happens when the explained variance/variation (see **ANOVA**) of the removed variates is relative less than the increase in the degrees of freedom if that variate was removed. If we had repeated the above experiment many times (to have a larger df) then the p-value will differ less.

10.3 ANOVA and additional sum of squares

Recall that the **sample variance** is given by

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

We want to decompose the sum of squares into two parts: one for the variance we can explain with our model and one for the variance we cannot explain. Doing this algebraically

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum(y_i-\hat{\mu}_i+\hat{\mu}_i-\bar{y})^2 \\ &= \sum(y_i-\hat{\mu}_i)^2+\sum(\hat{\mu}_i-\bar{y})^2+2\sum(y_i-\hat{\mu}_i)(\hat{\mu}_i-\bar{y})\end{aligned}$$

For the last term we have

$$\begin{aligned}\sum(y_i-\hat{\mu}_i)(\hat{\mu}_i-\bar{y}) &= \sum\hat{\mu}_i(y_i-\hat{\mu}_i)-\bar{y}\sum(y_i-\hat{\mu}_i) \\ &= \sum\hat{\mu}_ie_i-\bar{y}\sum e_i \\ &= \hat{\mu}^Te && \sum e_i=0 \\ &= 0 && \hat{\mu},e \text{ is orthogonal}\end{aligned}$$

Thus we have

$$\begin{aligned}&\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum(y_i-\hat{\mu}_i)^2+\sum(\hat{\mu}_i-\bar{y})^2 \\ &SS(Tot)=SS(Res)+SS(Reg)\end{aligned}$$

where $SS(Reg)$ is the variation our regression model accounts for (or **explained sum of squares**, and $SS(Res)$ is the **residual sum of squares**).

10.4 Coefficient of determination R^2

For the SLR case, R^2 is exactly the square of the coefficient of correlation r.
For the MLR case (and in general), it is equivalent to

$$R^2=\frac{SS(Reg)}{SS(Tot)}$$

or the **proportion of variation explained by our model**. Rewriting in terms of $SS(Res)$ or the sum of squares not explained by our model

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

10.5 Testing if any variates are related to response

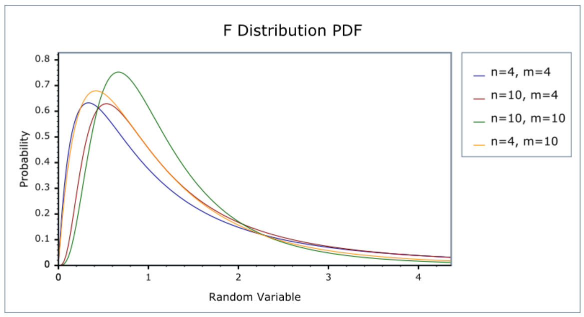
The F-test will check if there is, for example, a relationship between overhead (response variate) and **at least one** of size, age, col, or clients (explanatory variates)? That is: its null hypothesis is (note that β_0 is not included)

$$\begin{aligned}&H_0:\beta_1=\beta_2=\beta_3=\beta_4=0 \\ &H_a:\text{ at least one of }\beta_j\neq 0\quad j=1,2,3,4\end{aligned}$$

What would be the test statistic we use for this? Obviously we cannot use a T distribution and T-statistic like before. We want our statistic to be large when our model explains a lot of variation relative to the variation we cannot explain. We would also like to correct for the degrees of freedom, thus we have

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

where $MS(\cdot)$ is the mean squared error (we will later see that this is a special case of the more general F-statistic). Under $H_0 : F \sim F_{p,n-(p+1)}$. Under H_0, we expect $F = 1$.
Note that $F \geq 0$, so the F-distribution looks something like



Example 10.2. For our overhead study, our $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. We got from our regression summary in R

$$F = \frac{MS(Reg)}{MS(Res)} = 100.5$$

Note that the p-value $= P(F > 100.5) = 1.661 \times 10^{-12}$ (one-tailed!). So we reject H_0 so at least one of size, age, col or clients is significantly related to overhead.

11 February 13, 2018

11.1 ANOVA table in R

A sample output from R for the ANOVA table

```
Response: y
Df Sum Sq  Mean Sq F value Pr(>F)
x          1 0.0049 0.004897  0.0574 0.8112
Residuals 98 8.3654 0.085361
```

The fields correspond to

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

11.2 F-test and ANOVA

After accounting for col index and # of clients, does either size or age account for significant variation in overhead? So in our "full model" we have

$$Y = \beta_0 + (\beta_1 x_1 + \beta_2 x_2) + \beta_3 x_3 + \beta_4 x_4 + \epsilon \quad \epsilon \sim N(0, \sigma^2)$$

in our "reduced" model with just col index and # of clients we have

$$Y = \beta_0 + \beta_3 x_3 + \beta_4 x_4 + \epsilon \quad \epsilon \sim N(0, \sigma^2)$$

(where we take away size and age β_1, β_2). For the reduced model we have $H_0 : \beta_1 = \beta_2 = 0$ and H_a is at least one of $\beta_1, \beta_2 \neq 0$.
We want to see how much more variation our full model explains vs our reduced model.
We can use the **F-statistic**

$$F = \frac{\frac{SS(Res)_{red} - SS(Res)_{full}}{df_{red} - df_{full}}}{\frac{SS(Res)_{full}}{n-(p+1)}}$$

where under $H_0 : F \sim F_{df_{red} - df_{full}, df_{full}}$.
Note that $SS(Res)_{red} - SS(Res)_{full}$ is called the **additional sum of squares** (additional variance explained by full model). Also remark the denominator is simply $MS(Res)_{full} = \hat{\sigma}^2$.

Example 11.1. Recall that

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

So we can calculate it as

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

so we have

$$SS(Res)_{full} = 14330^2(19)$$

and similarly

$$SS(Res)_{red} = 15360^2(21)$$

So our F-statistic value is

$$F = \frac{(15360^2(21) - 14330^2(19))/2}{14330^2} = 2.564$$

So the p-value is the value of $P(F_{2,19} > 2.564)$. From the F-table we see that $P(F_{2,19} > 3.52) = 0.05$ which implies our p-value is > 0.05. So we do not reject H_0. The reduced model is thus preferred: that is age and size together do not account for significant variation in overhead after acconting for col and clients.

In R, we can accomplish this via

```
> anova(audit.red.lm, audit.full.lm)
```

where `audit.red.lm` and `audit.full.lm` are the reduced and full models, respectively. This gives us a p-value of 0.1033 in R for the example.

11.3 F-test special case: testing significance of all parameters of a model

Consider again the case where we wanted to test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ and H_a at least one of $\beta_j \neq 0$ for $j = 1, 2, \ldots, p$. This is simply the F-test but with a 0 parameter reduced model.
Note that if we are testing for the significance of a single model, our "full" model is our model

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

and the "reduced" model is

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

First note that the LSE (least square estimate) of β_0 for the reduced model is

$$\begin{aligned} &\sum (y_i - \hat{\beta}_0)^2 \\ \Rightarrow& -2\sum (y_i - \hat{\beta}_0) = 0 && \text{first-order condition} \\ \Rightarrow& -n\hat{\beta}_0 + \sum y_i = 0 \\ \Rightarrow& \hat{\beta}_0 = \frac{\sum y_i}{n} = \bar{y} = \hat{\mu} \end{aligned}$$

So we end up with

$$\begin{aligned} SS(Res)_{red} &= \sum e_i^2 = \sum (y_i - \hat{\mu}_i)^2 \\ &= \sum (y_i - \hat{\beta}_0)^2 && Y = \beta_0 + \epsilon \Rightarrow \hat{\mu} = \hat{\beta}_0 \\ &= \sum (y_i - \bar{y})^2 \\ &= SS(Tot) \end{aligned}$$

Thus we have

$$\begin{aligned} F &= \frac{(SS(Tot) - SS(Res)_{full})/p}{MS(Res)_{full}} \\ &= \frac{SS(Reg)_{full}/p}{MS(Res)_{full}} \\ &= \frac{MS(Reg)}{MS(Res)} \end{aligned}$$

as we had previously used.

11.4 F-test special case: testing significance of one additional parameter

Remember we previously tested for the significance of a parameter using the T-test where $H_0 : \beta_j = 0$ for some $j = 1, \dots, p$.

Example 11.2. After accounting for size, col, and clients, is age significant related to overhead?
Our full model is thus

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

and our reduced model (with age removed)

$$Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

We thus want to test for $H_0 : \beta_2 = 0$ and $H_0 : \beta_2 \neq 0$. Instead of using the T-statistic

$$t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)}$$

we can instead use the F-statistic

$$F = \frac{(SS(Res)_{red} - SS(Res)_{full})/1}{MS(Res)_{full}}$$

or concretely with our example

$$\begin{aligned} F &= \frac{14210^2(20) - 14330^2(19)}{14330^2} \\ &= 0.666 \end{aligned}$$

Which gives us a p-value > 0.05. We do not reject H_0, so after accounting for size, col, and clients, age is not significantly related to overhead (reduced model is preferred).

Remark 11.1. For the p-values, note that

$$P(F_{1,df} > C) = P(|t_{df}| > C^2) \quad C > 0$$

(our F-statistic value is the square of the t-statistic value). That is: for comparing two models with a difference of one df we have $F = t^2$ (The F-test statistic and t-statistic will yield **identical p-values**).

12 February 15, 2018

12.1 Difference in response from reduced model

In the promo example, we had

$$x_1 = \begin{cases} 1 & \text{if promo1 used} \\ 0 & \text{otherwise} \end{cases}$$
$$x_2 = \begin{cases} 1 & \text{if promo2 used} \\ 0 & \text{otherwise} \end{cases}$$

where our model is $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$.
Is there a difference in sales between promo1 and promo2 stores?
One way is to do the following hypothesis test

$$\begin{aligned} H_0 &: \beta_1 - \beta_2 = 0 \\ H_a &: \beta_1 - \beta_2 \neq 0 \end{aligned}$$

thus we have the T-statistic

$$t = \frac{(\hat\beta_1 - \hat\beta_2) - 0}{SE(\hat\beta_1 - \hat\beta_2)}$$

Another approach: additional sum of squares. We have the full model

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

and reduced model

$$\begin{aligned} Y &= \beta_0 + \beta^* x_1 + \beta^* x_2 + \epsilon & \beta^* = \beta_1 = \beta_2 \\ &= \beta_0 + \beta^*(x_1 + x_2) + \epsilon \\ &= \beta_0 + \beta^* x_3 + \epsilon & x_3 = x_1 + x_2 \end{aligned}$$

where we can interpret x_3 as

$$x_3 = \begin{cases} 1 & \text{if either promotion 1 or 2 used} \\ 0 & \text{otherwise} \end{cases}$$

So we have the F-statistic

$$F = \frac{(SS(Res)_{red} - SS(Res)_{full})/1}{MS(Res)_{full}}$$

where we get $F = 5.2218$ and p-value 0.03038 in our example for a $F_{1,27}$ distribution.
We therefore reject H_0, therefore mean sales is significantly greater for promo1 than for promo2.

12.2 General linear hypothesis

Consider the hypotheses tested so far using additional sum of squares (p is the number of total parameters in our full model)

1. $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$
2. $H_0 : \beta_1 = \beta_2 = 0 \quad (p = 4)$
3. $H_0 : \beta_1 = 0 \quad (p = 3)$
4. $H_0 : \beta_1 - \beta_2 = 0$

The additional sum of squares test can be used to test any set of linear constraints that can be expressed in the form

$$H_0 = A\beta = 0$$

where A is an $l \times (p + 1)$ matrix of l linear constraints.
For the hypothesis (1) we have

$$H_0 : \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & & & & & & \\ 0 & 0 & 0 & \ldots & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where the left matrix is A and the right matrix is β. Similarly for (2)

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

And for (3)

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}$$

and for (4)

$$A = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix}$$

12.3 Residual analysis and model assumptions

Residual analysis lets us **assess our model assumptions**. Recall our model assumptions of $Y = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I)$:

- The "function form" of the relationship is correctly specified (i.e. $\mu = X\beta$)

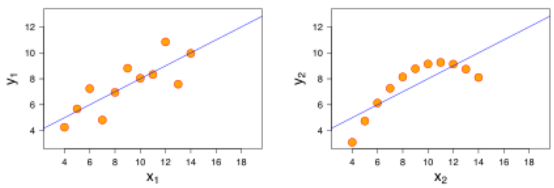


Figure 12.1: The linear function form is correctly specified for the left plot but not for the right plot where although the line fits well, it is not really linear (but rather quadratic).

We want to make sure we're not fitting a linear regression model to data that is actually quadratic or some other nonlinear fit (the significance of our parameters in the regression summary in say R tells us nothing about this: we can get really significant parameter estimates but the data may not actually be linear).

- Errors are normal (specified as $\epsilon \sim N(\cdot)$).

  Remark 12.1. Normality of errors is not too important: although we assumed Y is normal (which depends on ϵ being normal) to derive β and $\hat{\beta}$, since β is the linear combination of Y normal variables it approaches a normal distribution regardless of Y's distribution for large sample sizes by the Central Limit theorem.

- Errors have constant variance i.e. **homoskedastic** (specified as $\sigma_i^2 = c$ some constant c for all β_i).

  It is possible for data to not have non-constant variance that is a function of X

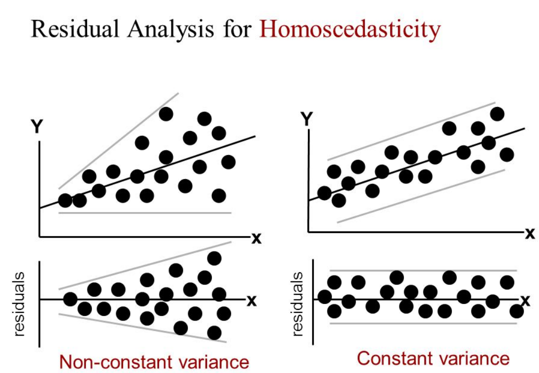


Figure 12.2: Non-constant variance would imply that we have variance as a function of X. For example, if the variance increases linearly as X increases, then we need to transform our data (e.g. square root it or log it) or fit another function to it (e.g. log in GLMs?).

- Errors are independent (specified as variance is $\sigma^2 I$, where the identity matrix has 0 off-diagonal entries).

  This is typically time-series data.

13 February 27, 2018

13.1 Residual plots

One way we can assess our model assumptions is with **residual plots** (where e_i is always on the y-axis):

e_i **vs** $\hat{\mu}_i$ Recall under the model $Y = X\beta + \epsilon$, we have $e^T\hat{\mu} = 0$ (orthogonal vectors) so $Cov(e_i, \hat{\mu}_i) = 0$. Thus they are independent.

A plot of e_i vs $\hat{\mu}_i$ should always reveal no observable pattern or relationship between the residuals and the fitted values. No pattern implies that our **function form** is specified correctly and our **variance is constant**.

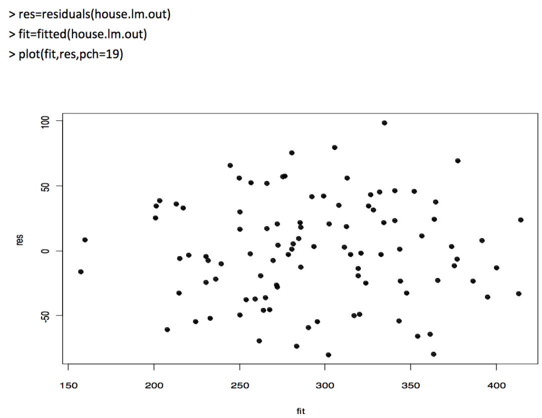


Figure 13.1: Residual plot of e_i vs $\hat{\mu}_i$.

QQ plots These are used to assess assumption of normal errors. It plots ordered (standardized) residuals vs expected quantiles from $N(0, 1)$.

A straight line relationship is an indication that the assumption of normal errors has been reasonably met.

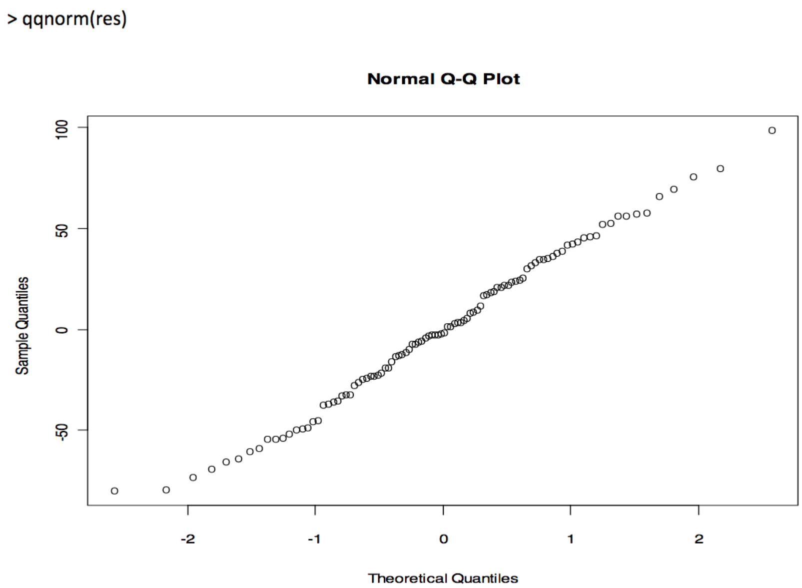


Figure 13.2: The (non-standardized: the sample quantiles are plotted from $[0, 100]$ instead of normalized theoretical $\approx [-3, 3]$ quantiles) residuals are ordered (these are *pm* values) and their CDF are plotted based on their corresponding quantile in the standard normal distribution (e.g. a residual of -2 corresponds to the theoretical standard normal quantile of -1.96).

We do not have a plot to check if the errors are independent (this will be discussed more during time-series).

13.2 Methods to address violated model assumptions

To address perceived violations of model assumptions:

- Transformation of response (and/or one or more explanatory variates). For example, there are **variance-stabilizing transformations** we can use:

  - $\log y$
  - $\sqrt{y}$
  - $\frac{1}{y}$

  (these may also be used for explanatory variates). These can be used if either the errors (variance) is not normal or if the function form is not linear. We can apply each of these transformations and see if it produces a better fit *and produces a "better" fit of our model assumptions.*

  Remark 13.1. After transforming our data, the residual standard error may seemingly decrease (or even increase) dramatically. Note that the RSE is relative to our response values: since the scale of those change based on the transformation we must take care when interpreting the change in the RSE.

  Remark 13.2. We will need to modify our intepretation of our parameter estimates since they are now estimated with respect to the transformed response values, not the original values.

- Inclusion of **higher order terms** (e.g. quadratic (x^2), cubic (x^3) of our explanatory variates (x)). This is sometimes called **polynomial regression**.

This does not violate multicollinearity since these terms are not *linearly* dependent.

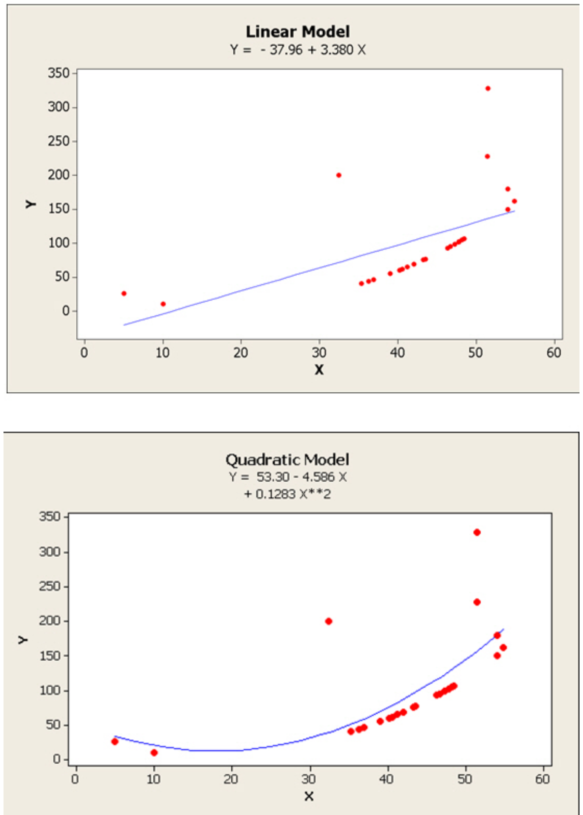


Figure 13.3: We initially have a linear model with a linear explanatory x and response y. We see the relationship between x and y is more quadratic. We then add x^2 as a term to our model which can allow us to fit a quadratic using linear regression with variates x and x^2 (our parameters are then the coefficients A, B, C in the quadratic $y = A + Bx + Cx^2$).

- Inclusion of **interaction terms**.

  When the relationship between an explanatory variate, x_k, and the response depends on the value of another explanatory variate, x_m, we say there is an **interaction** between x_k and x_m (i.e. when other variates have a magnifying/diminishing effect on the relationship between another variate and the response).

  For example, the effect of size on overhead of an office may be magnified if the age of the office is larger (and vice versa).

  May require the inclusion of an interaction term $x_k * x_m$ (where the explicit asterisk $*$ denotes interaction). It is coincidentally often the pairwise product of the variates.

  Remark 13.3. One must be careful introducing too many interaction terms since that will decrease our **degrees of freedom**.

  Example 13.1. In our overhead model, we had

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

if we want an interaction term between age (x_1) and size (x_2) then we use the model

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + +\beta_5(x_1 * x_2) + \epsilon \\
&= \beta_0 + (\beta_1 + \beta_5 x_2)x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon
\end{aligned}
$$

We can then interpret β_5 as the magnifying effect of x_2 on x_1.

14 March 1, 2018

14.1 Fitted residuals e vs errors ϵ

We can derive an expression that relates our fitted residuals (e, a distribution) and our assumed random errors (ϵ) Recall that (below are vectors)

$$
\begin{aligned}
e &= Y - \hat{\mu} \\
&= Y - X\hat{\beta} \\
&= Y - X(X^T X)^{-1} X^T y \\
&= Y - HY \\
&= (I - H)Y \\
&= (I - H)(X\beta + \epsilon) \\
&= X\beta - X(X^T X)^{-1} X^T X\beta + \epsilon - H\epsilon \\
&= X\beta - X\beta + \epsilon - H\epsilon \\
&= (I - H)\epsilon
\end{aligned}
$$

So we end up with

$$
\begin{aligned}
e &= (I - H)Y \\
&= (I - H)\epsilon
\end{aligned}
$$

One might assume $Y = \epsilon$ since

$$
\begin{aligned}
&(I - H)^{-1}(I - H)Y = (I - H)^{-1}(I - H)\epsilon \\
\Rightarrow &Y = \epsilon
\end{aligned}
$$

but this is not necessarily true since $I - H$ need not be invertible (it is invertible iff its rank is n)! Note that

$$
rank(I - H) = rank(H) = tr(H) = rank(X) = p + 1
$$

where the rank of a symmetric matrix is its trace. Since $p + 1 < n$ (usually) $I - H$ is not of full rank so it is not invertible!

14.2 Distribution of fitted residuals e

Since $e = (I - H)\epsilon$ and $\epsilon \sim N$ then $e \sim N$, and we can derive the normal distribution of e

$$
\begin{aligned}
E[e] &= (I - H)E(\epsilon) = 0 \\
Var(e) &= (I - H)Var(e)(I - H)^T \\
&= \sigma^2 (I - H)(I - H)^T \\
&= \sigma^2 (I - H)
\end{aligned}
$$

So we have

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

Assuming our variates are all mutually independent (i.e. $\sigma^2_{jk} = 0$ for $j \neq k$) then we get

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

where all entries of H are strictly positive. This makes sense intuitively: since all residuals must sum to 0 in LSE, some positive residual should coerce other residuals to be negative hence the negative covariance.
Another way to express this

$$e_i = N(0, \sigma^2(I - h_{ii}))$$

14.3 Studentized residuals

Analogous to standardizing with respect to the normal distribution, we can do the same with the T-distribution. As before (remember $\bar{e_i} = 0$), the studentized residuals d_i is defined as

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

where we subtract the mean and divide by the standard error.
This looks very similar to, when $\hat{\beta}_i = N(\beta_i, \sigma^2(X^TX)^{-1}_{ii})$

$$\frac{\hat{\beta}_i - \beta_i}{\sigma^2\sqrt{(X^TX)^{-1}_{ii}}} \sim t_{n-(p+1)}$$

d_i does not exactly follow a T-distribution since e_i and $\hat{\sigma}$ are not independent. Thus it follows a distribution *roughly that of the T-distribution.*

Remark 14.1. There are two types of studentized residuals: **internalized** and **externalized**. They differ in $\hat{\sigma}$ where the internalized uses the biased estimate of σ^2 (whereby sum of squared residuals divided by $n - p$) and externalized removes the ith residual suspected of being improbably large since it may skew the distribution (takes sum of residuals square except ith residual and divide by $n - (p + 1)$).
See the Wikipedia page for more details.

14.4 Outliers

These can be:

extreme values of the response variate How do we define one formally?

An observation is considered an "outlier" if its studentized residual d_i satisfies

$$|d_i| > 2.5$$

or thereabouts.

Why might an observation be an outlier?

1. Typo or human error
2. Missing interacting variate (e.g. type of company may be important to consider overhead of the company: not including such may result in outliers)

extreme values in the explanatory variate space These are extreme values of $(x_1, x_2, \ldots, x_p)$. Suppose we have one such outlier.

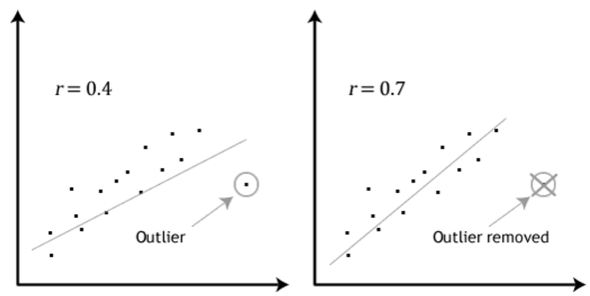


Figure 14.1: An outlier can have a huge effect on the fit and goodness of fit.

Think about the effect of fitting a LSE with and without the outlier. If the outlier lands roughly in the fit without the outlier, then there will be little difference between the two fits (the fit without the outlier will "move away" from where the outlier was before).

If the outlier lands far away from the fit without the outlier, then including it in the fit will skew it dramatically.

We can identify these outliers based on the residual i's h_{ii} (as we derived for its variance in its distribution). h_{ii} is also **called leverage**.