STAT2401 Analysis of Experiments

Lecture Week 12 Dr Darfiana Nur

Aims of this lecture

1. REVISION WEEKS 2-8
2. Statistical modelling, machine learning, & predictive analytics
3. FINAL EXAM INFORMATION

AIM 1 WEEK 2

6 Steps in carrying out a hypothesis test

1. State the hypotheses (Ho and Ha)
2. Calculate the test statistic
3. Sampling distribution of the test statistic
4. Find the p-value based on (3), look at Ha (one sided or two sided)
5. Make a decision based on the p-value

- p-value $\leq \alpha$, reject Ho;
- p-value $> \alpha$, do not reject Ho

6. State your conclusion in the context of your specific setting.

Hypotheses for two independent sample *t*-test

1) $H_0: \mu_A = \mu_B$ $\quad\quad H_A: \mu_A \neq \mu_B$

2) Test statistic:

$$t = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

$$df = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{1}{n_A - 1}\left(\frac{s_A^2}{n_A}\right)^2 + \frac{1}{n_B - 1}\left(\frac{s_B^2}{n_B}\right)^2}$$

3) The sampling distribution of the test statistic: $t_{(df)}$

4) The p-value of the *t*-test is the probability that a random variable having the $t_{(df)}$ distribution exceeds t (in absolute terms)

5) Decision

6) Conclusion

How about ONE-SIDED test for 2 independent sample *t*-test:

[H_A: $\mu_A - \mu_B > 0$ OR H_A: $\mu_A - \mu_B < 0$] - p-value?

How about **the assumptions** for 2 independent sample *t*-test?

Confidence interval (CI)

CI = point estimate ± margin of error (ME)

ME = multiplier × standard error

INTERPRETATIONS
Hypothesis Testing and CI are consistent for the same significance level

CI for 1−sample z−test (proportion) : $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

CI for 1−sample t−test : $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$

CI for paired t−test : $\bar{x}_d \pm t^* \frac{s_d}{\sqrt{n}}$

CI for $(\mu_A - \mu_B)$: $(\bar{x}_A - \bar{x}_B) \pm t^* \sqrt{\frac{{s_A}^2}{n_A} + \frac{{s_B}^2}{n_B}}$

WEEK 3 Aim 2.2 Numerically - The Pearson Sample Correlation coefficient *(r)*

- Measures the direction and strength of the linear relationship between two numerical variables.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{s_{xy}}{s_x s_y}$$

$\bar{x}$ be the sample mean of X;
$\bar{y}$ be the sample mean of Y;
s_x is the sample standard deviation of X;
s_y is the sample standard deviation of Y;
s_{xy} is the sample covariance between X and Y;
n be the number of observations

- R is used to compute this value.
- Note that the formula considers the variation in the *x* variable, in relation to the variation in the *y* variable).

Understanding *correlation*

- Positive *(r)* indicates positive association between the variables

- Negative *(r)* indicates negative association between the variables.

- The correlation *(r)* always falls between -1 and +1.

WEEK 4 SLR – Model and assumptions

- To complete the specification of the model, we assume
  1. $E(\epsilon_i) = 0$, for all i
  2. $\text{var}(\epsilon_i) = \sigma^2$, for all i
  3. ϵ_i and ϵ_j are independent for all $i \neq j$
  4. $\epsilon_i \sim N(0, \sigma^2)$ if we wish to make inferences about the regression model

- The assumptions imply that

$$E(Y \mid X = x) = \beta_0 + \beta_1 x \text{ and}$$
$$\text{var}(Y \mid X = x) = \sigma^2$$

and hence that if we have repeated observations at different values of x, the scatter about the true line will be Normally distributed with constant variance σ^2

Least Squares estimation

- We wish to choose the straight line that minimises Sum of Squares of Error (SSE) $$SSE = \sum (Y_i - \beta_0 - \beta_1 X_i)^2$$
  - To do this we must set the 1st partial derivatives of this formula to 0.

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i) = 0$$

Data = Fit + Error

Y_i = $(\beta_0 + \beta_1 X_i)$ + (ε_i)

$$\frac{\partial SSE}{\partial \beta_1} = -2 \sum X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

After re-arranging some terms we have

$$\sum Y_i = n\beta_0 + \beta_1 \sum X_i$$

$$\sum X_i Y_i = \beta_0 \sum X_i + \beta_1 \sum X_i^2$$

These are called the *normal equations* and must be solved to provide the estimates $\hat{\beta}_0, \hat{\beta}_1$

$\widehat{\beta_0} = b_0$

$\widehat{\beta_1} = b_1$

SLR – Least squares estimation

- Rearranging the equations on the previous slide yields

$$n\beta_0 + \ \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\beta_0 \sum_{i=1}^{n} x_i + \ \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

- These equations are known as the **normal equations**, and solving them yields the least squares estimates of the intercept and slope

Least Squares estimation

THE UNIVERSITY OF WESTERN AUSTRALIA

- We can easily solve these two equations given some data points Y and X.

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \approx \frac{Cov(X,Y)}{Var(X)} \qquad \frac{S_{xy}}{S_{xx}}$$

- It is straightforward to show, using the second order partial derivatives that this point is a minimum for the SSE

$$\widehat{\beta_0} = b_0$$

- Luckily, R calculates these for us!

$$\widehat{\beta_1} = b_1$$

Properties of least squares estimators

AUSTRALIA

- The least squares estimators are **unbiased**

$$E(\hat{\beta}_0) = \beta_0 ; E(\hat{\beta}_1) = \beta_1$$

$\widehat{\beta_0} = b_0$

$\widehat{\beta_1} = b_1$

- What does this mean?

$\hat{\beta}_0, \hat{\beta}_1$ are random variables, they are subject to variation in different samples

- If you take lots of samples and then take the average of the estimates of $\hat{\beta}_0, \hat{\beta}_1$ these will be equal to the true population values β_0, β_1

WEEK 5 Prediction and forecasting I (for μ_Y)

- $\mu_y = E(Y | X)$: the value of the regression line at $X = X_0$
- For any given value of X_0, we know that

$$E(Y \mid X) = \beta_0 + \beta_1 X \quad ; \quad Var(Y \mid X) = \sigma^2$$

- To **predict** the **average** value of Y for a given value of X_0 we use

$$E(Y_i \mid X_0) \approx \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

- To place a confidence interval around this prediction of the mean $E(Y \mid X)$ we need to estimate

$$Var(E(Y \mid X_0)) = Var(\hat{Y}) = Var(\hat{\beta}_0 + \hat{\beta}_1 X_0)$$

Recall that $\hat{\beta}_1 = \sum c_i Y_i ; \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

Meaning that $\hat{Y} = \overline{Y} + \hat{\beta}_1 (X_0 - \overline{X})$

We can obtain that

$$s.e.(\hat{Y} \mid X_0) = \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum (X_i - \overline{X})^2} \right)}$$

This variance term has 2 parts

$$Var(\overline{Y}) = \frac{\sigma^2}{n} ; Var(\hat{\beta}_1 (X_0 - \overline{X})) = \frac{\sigma^2 (X_0 - \overline{X})^2}{\sum (X_i - \overline{X})^2}$$

- To make a confidence interval **we must assume normality (or some other distribution) for the residuals**. Doing so,

$$(\hat{\beta}_0 + \hat{\beta}_1 X_0) \pm t_{n-2;1-\alpha/2} s.e.(\hat{Y} \mid X_0)$$

is a $(1-\alpha)\%$ confidence interval for the mean of future *Y* values corresponding to $X = X_0$.

This is a confidence interval for the regression line at any point X_0

If we know the true value of σ^2 then we can use Normal distribution.

Prediction and forecasting II (for Yhat)

We may wish the confidence interval to cover (1-α)% of **future observations** (not just the mean). We can obtain that

$$s.e.(Y \mid X_0) = \sqrt{\sigma^2\left(1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum (X_i-\bar{X})^2}\right)}$$

Again this variance term has 2 parts.

- $Var(Y \mid X = X_0) = \sigma^2$ $\quad$ $Var(\hat{\beta}_0 + \hat{\beta}_1 X_0) = Var(\hat{Y} \mid X = X_0)$

Firstly the variance for a future observation for a given value *X* and secondly the variance because we have estimated the regression parameters

To make a confidence interval we must assume normality
(or some other distribution) for the residuals.

$$(\hat{\beta}_0 + \hat{\beta}_1 X_0) \pm t_{n-2;1-\alpha/2} s.e.(Y \mid X_0)$$

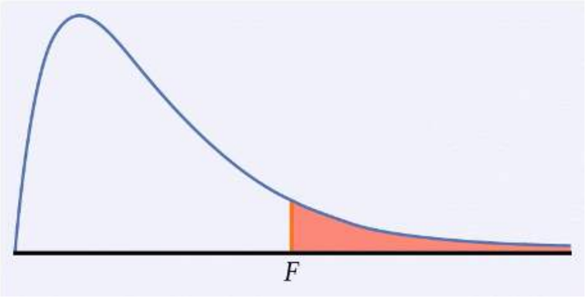
is a $(1-\alpha)\%$ confidence interval for the future *Y* values corresponding to $X = X_0$.

The ANOVA *F* Test

- For a simple linear relationship, the ANOVA tests the hypotheses

$$H_0: \beta_1 = 0 \text{ versus } H_a: \beta_1 \neq 0$$

by comparing MSR (Mean Square Regression) to MSE (Mean Square Error): $F = MSR/MSE$



- When H_0 is true, F follows the $F(1, n - 2)$ distribution. The P-value is $P(F \geq f)$.

- *The ANOVA test and the two-sided* t*-test for* $H_0: \beta_1 = 0$ *yield the same P-value.*

- *Software output for regression may provide* t, F, *or both, along with the P-value.*

The ANOVA Table

Source	Sum of squares SS	DF	Mean square MS	*F*	*P*-value
Regression	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$	1	*MSR=SSM/DFR*	*MSR/MSE*	Tail area above *F*
Error	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	*n* − 2	*MSE=SSE/DFE*		
Total	$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$	*n* − 1			

SST = SSM + SSE *DFT = DFM + DFE* *F=MSM/MSE*

The standard deviation, *s*, of the *n* residuals $e_i = y_i - \hat{y}_i$, $I = 1,\ldots,n$, is calculated from the following quantity:

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{DFE} = MSE$$

s is an approximately unbiased estimate of the regression standard deviation **σ**.

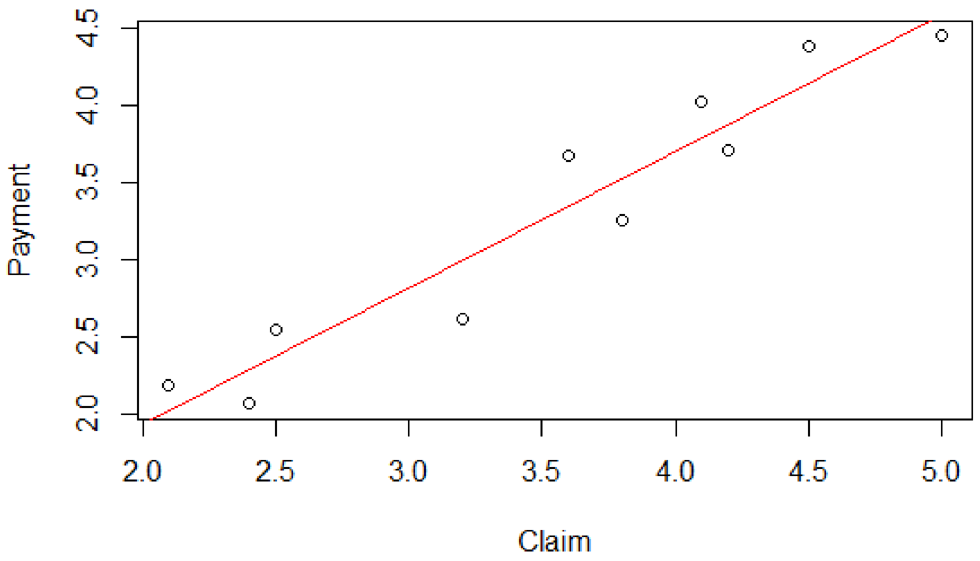
Example 5: Household policies- lm()

A sample of 10 claims and corresponding payments on settlement for household policies is taken from the business of an insurance company.

The amounts, in units of $100, are as follows:

Claim	2.10	2.40	2.50	3.20	3.60	3.80	4.10	4.20	4.50	5.00
Payment	2.18	2.06	2.54	2.61	3.67	3.25	4.02	3.71	4.38	4.45

```
Call:
lm(formula = Payment ~ Claim, data = Insurance)
Residuals:
     Min       1Q   Median       3Q      Max
-0.37702 -0.20571  0.01918  0.22183  0.33006
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.16363    0.34048   0.481    0.644
Claim        0.88231    0.09309   9.478 1.27e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
Residual standard error: 0.2705 on 8 degrees of freedom
Multiple R-squared:  0.9182,  Adjusted R-squared:  0.908
F-statistic: 89.82 on 1 and 8 DF,  p-value: 1.265e-05
```



Example 5: Household policies – anova()

```
Coefficients:
              Estimate  Std. Error   t value
Pr(>|t|)
(Intercept)  0.16363   0.34048     0.481      0.644
Claim         0.88231   0.09309     9.478      1.27e-05
***
---
Signif. codes:  0 '***' 0.001  '**' 0.01  '*' 0.05 '.' 0.1  ' ' 1
```

```
Analysis of Variance Table

Response: Payment
          Df Sum Sq Mean Sq F value    Pr(>F)
Claim      1 6.5734  6.5734  89.824 1.265e-05 ***
Residuals  8 0.5854  0.0732
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Source	Sum of squares SS	DF	Mean square MS	*F*	*P*-value
Regression	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 6.5734$	1	*MSR* *=SSR/DFR=6.5734/1=6.5734*	*MSR/MSE* *=6.5734/0.0732=89.824*	Tail area above *F* *=P(F(1,8) > 89.824)* *=1.265 x 10^(-5)*
Error	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0.5854$	*n* − *2=10-2=8*	*MSE=SSE/DFE* *=0.5854/8=0.0732*		
Total	$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$	*n* − 1=10-1=9			

Example 5: Household policies

T Test

- STEP 1 H0: β1 = 0 vs. Ha: β1 ≠ 0
- STEP 2 Test statistic T=9.478
- STEP 3 The sampling distribution $T \sim t$ df $(n-2)$ that is $T \sim t$ (df=8) given n=10
- STEP 4 The p-value (see Ha):

p-val= P(|t_8| > 9.478)
=2*pt(9.478,8) = 1.27e-05

- STEPS 5 and 6 Decision and Conclusion. As the p-value is very small, we reject the Ho. We conclude that there is a positive relationship between Payment and Claim.

ANOVA or F Test

STEP 1 H0: β1 = 0 vs. Ha: β1 ≠ 0

STEP 2 Test statistic F= 89.824

STEP 3 The sampling distribution

F ~ Fdf(1, $(n-2)$) that is F ~ Fdf(1, 8)

STEP 4 The p-value (see Ha):

p-val= P(Fdf(1, 8)> 89.824)
=pf(89.824,1,8,lower.tail=F)
= `1.265e-05`

STEPS 5 and 6 Decision and Conclusion. As the p-value is very small, we reject the Ho. We conclude that there is a positive relationship between Payment and Claim.

F=89.824 = (9.478)^2=T^2

WEEK 6 Recap: SLR – Model and assumptions

- Simple linear regression model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- To complete the specification of the model, we assume
  1. $E(\epsilon_i) = 0$, for all i (zero means residuals)
  2. $\mathrm{var}(\epsilon_i) = \sigma^2$, for all i (constant variance residuals)
  3. ϵ_i and ϵ_j are independent for all $i \neq j$ (independence residuals)
  4. $\epsilon_i \sim N(0, \sigma^2)$ if we wish to make inferences about the regression model (normality of residuals)

- The assumptions imply that

  $E(Y \mid X = x) = \beta_0 + \beta_1 x$ and

  $\mathrm{var}(Y \mid X = x) = \sigma^2$

  and hence that if we have repeated observations at different values of x, the scatter about the true line will be Normally distributed with constant variance σ^2

Recap: Checking the Conditions for Regression Inference

- You can fit a least-squares line to any set of explanatory-response data when both variables are quantitative. If the scatterplot does not show a roughly linear pattern, the fitted line may be almost useless.
- Before you can trust the results of inference, you must check the conditions for inference one by one.

✓ **The relationship is linear in the population.**
✓ **The response varies Normally about the population regression line.**
✓ **Observations are independent.**
✓ **The standard deviation of the responses is the same for all values of *x*.**

You can check all of the conditions for regression inference by looking at graphs of the residuals or **residual plots.**

Checking model validity

Sheather (2009), p. 50 & 51

1. Determine whether the proposed regression model is a valid model (i.e., determine whether it provides an adequate fit to the data). The main tools we will use to validate regression assumptions are plots of standardized residuals.
2. The plots enable us to assess visually whether the assumptions are being violated and point to what should be done to overcome these violations. Determine which (if any) of the data points have x-values that have an unusually large effect on the estimated regression model (such points are called leverage points).
3. Determine which (if any) of the data points are outliers, that is, points which do not follow the pattern set by the bulk of the data, when one takes into account the given model.

Checking model validity

Sheather (2009), p. 50 & 51

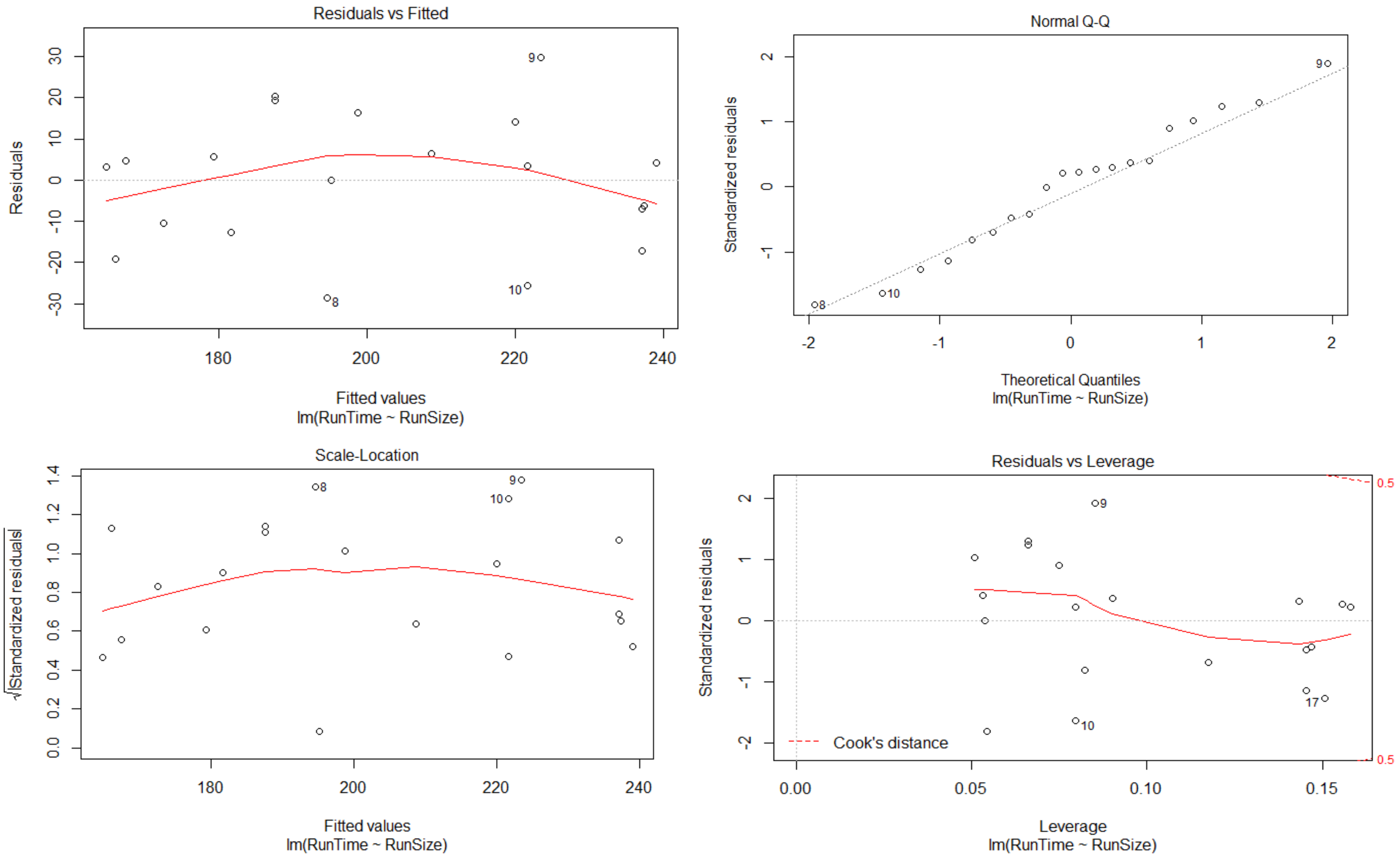
4. If leverage points exist, determine whether each is a bad leverage point. If a bad leverage point exists we shall assess its influence on the fitted model.

5. Examine whether the assumption of constant variance of the errors is reasonable. If not, we shall look at how to overcome this problem.

6. If the data are collected over time, examine whether the data are correlated over time.

7. If the sample size is small or prediction intervals are of interest, examine whether the assumption that the errors are normally distributed is reasonable

Revisiting Example 1: “plot(prod.lm)” in R



- The smoothing red curves to help identifying patterns
- No pattern (random), fairly constant spread (variance),
- Normality is satisfied.
- No leverage points

Influence analysis

- Aims to determine observations that have influential effect on the fitted model
- Potentially influential points become candidate for removal from the model
- Criteria used are
  - The hat matrix elements h_i (we use this one in SLR)
  - The Studentized deleted residuals t_i^*
  - Cook's distance statistic D_i (we use this one in SLR)
- All three criteria are complementary
- Only when all three criteria provide consistent result should an observation be removed

The Hat Matrix Element h_i

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

If $h_i > 4/\mathrm{n}$, X_i is a leverage point

X_i may be considered a candidate for removal from the model if it is a bad leverage point.

Cook's Distance Statistic D_i

$$D_i = \frac{SR_i^2 h_i}{2(1-h_i)}$$

$$SR_i = \frac{e_i}{S_{YX}\sqrt{1-h_i}}$$

If $D_i > 4/(\mathrm{n}-2)$

an observation is considered influential

Use the function `influence.measures` to explore measures of leverage and Cook's distance in R.

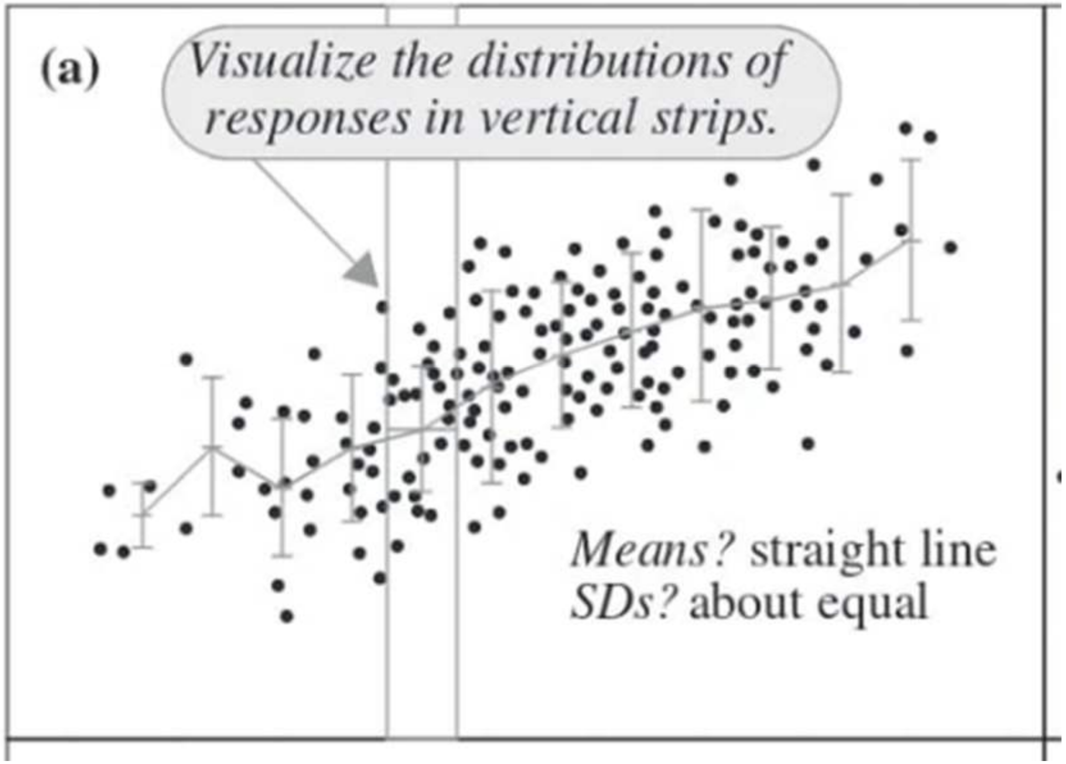
Outliers and leverage

- An **outlier** is a point whose **standardized residual falls outside**
  - the interval from −2 to 2 for small to moderate sample size
  - the interval from −4 to 4 for large sample size
- A **bad leverage point** is a leverage point whose **standardized residual falls outside** the interval from −2 to 2 for small to moderate sample size.
- A **good leverage point** is a leverage point whose standardized residual falls inside the interval from −2 to 2 for small to moderate sample size.

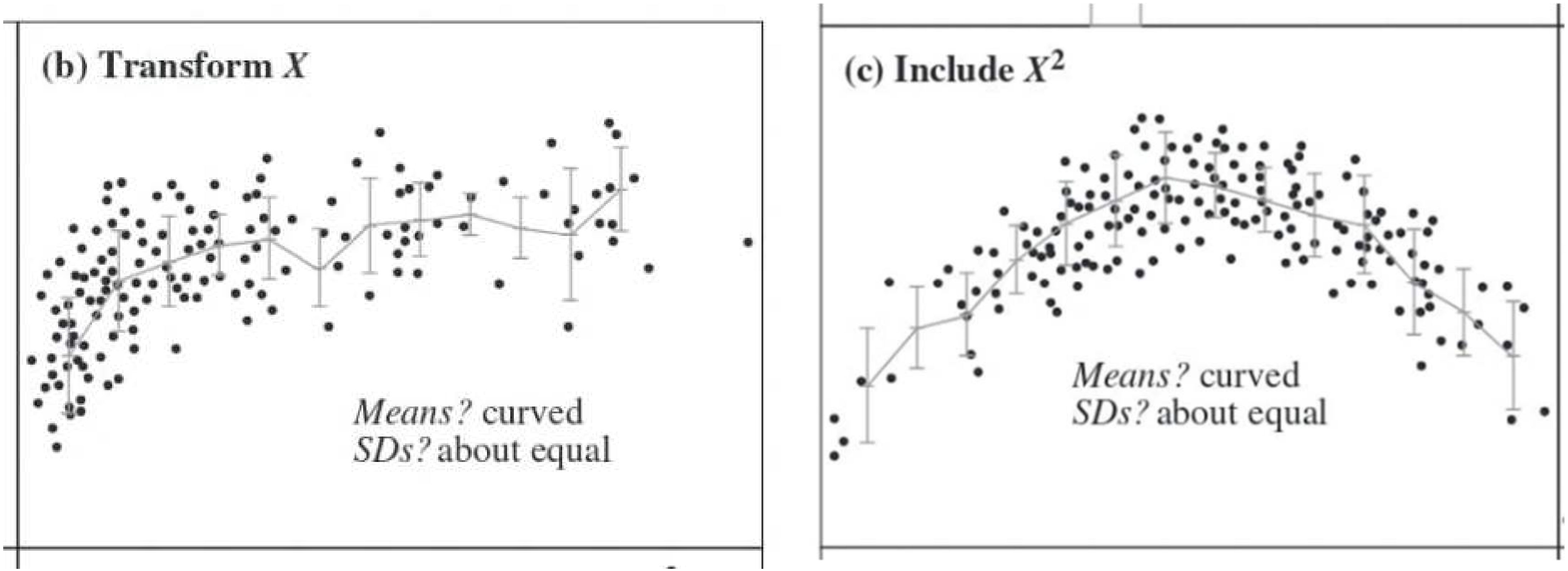
Aim 3 Transformation

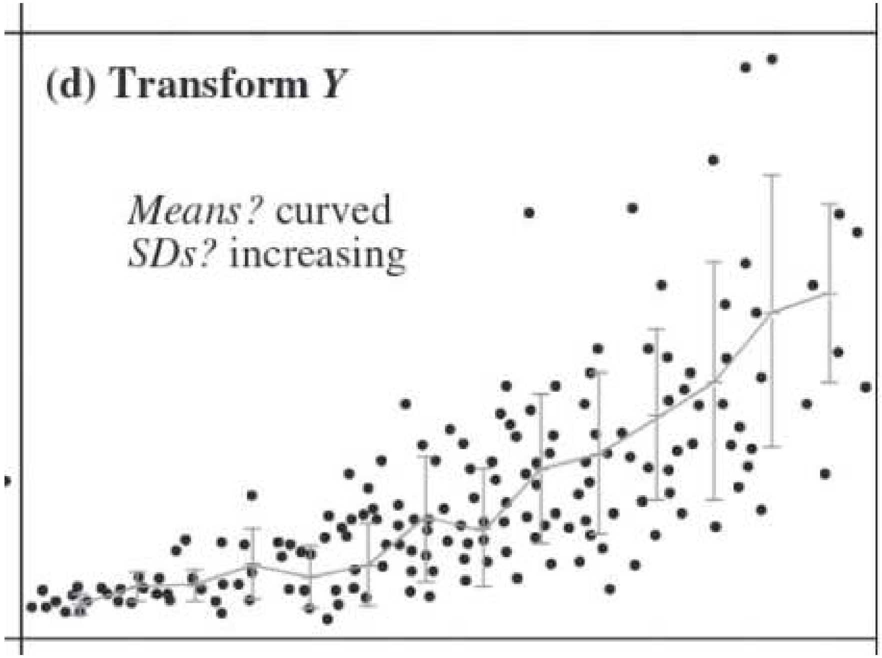
- Transformations can be used to
  - Overcome problems due to nonconstant variance
  - Estimate percentage effects
  - Overcome problems due to nonlinearity

The "ideal" Plot

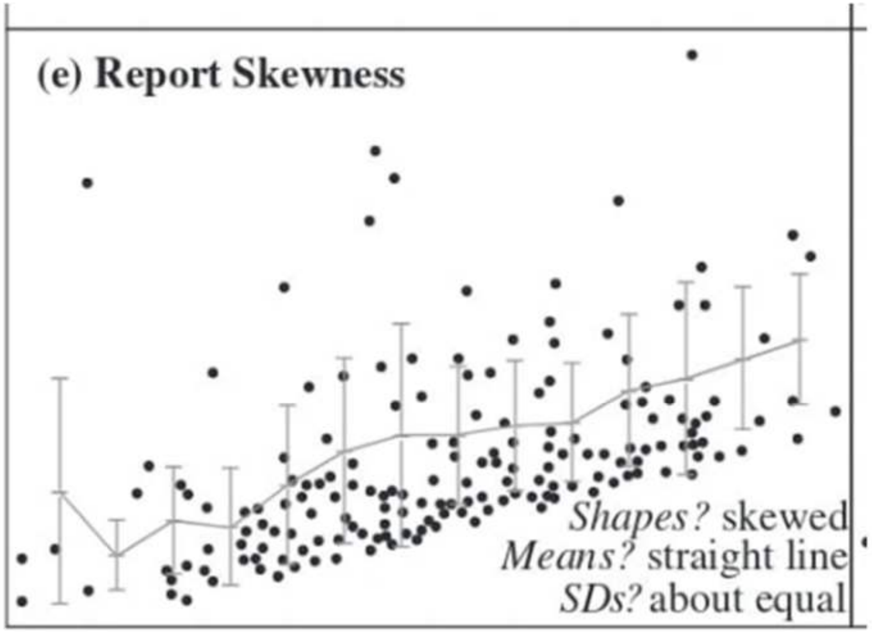


From "The Statistical Sleuth"

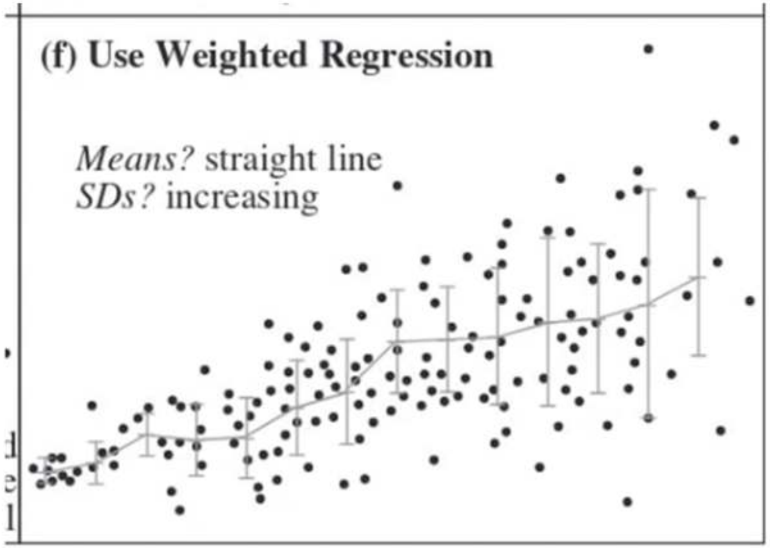

(b) Transform X
Means? curved
SDs? about equal
(c) Include X²
Means? curved
SDs? about equal



(d) Transform *Y*



(e) Report Skewness



(f) Use Weighted Regression

(e) The regression is a straight line, the variability is roughly constant, but the distribution of Y about the regression line is skewed. Remedies are unnecessary, and transformations will create other problems. Use simple linear regression, but report the skewness.

(f) The regression is a straight line but the variability increases as the mean of Y increases. Simple linear regression gives unbiased estimates of the straight line relationship, but better estimates are available using *weighted regression*, as

WEEK 7 The simple linear model in matrix notation is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Often we need sums of squares terms in regression

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}; \boldsymbol{\varepsilon}' = (\varepsilon_1 \quad \varepsilon_2 \quad .. \quad \varepsilon_n)$$

$$\underset{1\text{xn}}{\boldsymbol{\varepsilon}'}\underset{\text{nx1}}{\boldsymbol{\varepsilon}} = \varepsilon_1{}^2 + \varepsilon_2{}^2 + .. + \varepsilon_n{}^2 = \underset{1\text{x1}}{\sum \varepsilon_i^2}$$

For any vector *a*, *a'a* represents the sum of the squares of the elements of *a.* It is a 1x1 scalar number

Properties of least squares estimates: MLR

- $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$ and we have now have p predictors, with
$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$
- The covariance matrix of the LS estimates is
$$\mathrm{var}\left(\widehat{\boldsymbol{\beta}}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$$

and as before, we estimate σ^2 from RSS, i.e.,
$$s^2 = \frac{RSS}{n-p-1} = \frac{1}{n-p-1}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$$
- Hence, for carrying out a t-test for testing $H_0: \beta_i = 0$, we use
$$\frac{\hat{\beta}_i - 0}{\mathrm{se}(\hat{\beta}_i)} \sim t_{n-p-1}$$
- We can obtain $\mathrm{se}(\hat{\beta}_i)$ as the square root of the ith diagonal element of $\mathrm{var}\left(\widehat{\boldsymbol{\beta}}\right)$

Aim 3.1 Parameter estimation: MLR

- If we have p predictors, we can write that $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$, and the least squares estimate is
$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$
- Hence, the fitted values can be written as $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, or $\widehat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, and the matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is known as the 'hat' matrix*
- Residuals are $\hat{\boldsymbol{e}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$, and RSS can be written as
$$RSS = (\boldsymbol{y} - \widehat{\boldsymbol{y}})'(\boldsymbol{y} - \widehat{\boldsymbol{y}})$$
and as before, we estimate σ^2 from RSS, i.e.,
$$s^2 = \frac{RSS}{n-p-1} = \frac{1}{n-p-1}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$$
- Note that the number of degrees of freedom is $n - p - 1$

Confidence Interval for β_j

- Estimating the regression parameters $\beta_0, \ldots, \beta_j, \ldots, \beta_p$ is a case of one-sample inference with unknown population variance.
- We rely on the t distribution, with $\boldsymbol{n - p - 1}$ **degrees of freedom.**

A **level C confidence interval for β_j** is

$$b_j \pm t^* SE_{b_j}$$

where SE_{b_j} is the standard error of b_j and t^* is the t critical for the $t(n - p - 1)$ distribution with area C between $-t^*$ and t^*.

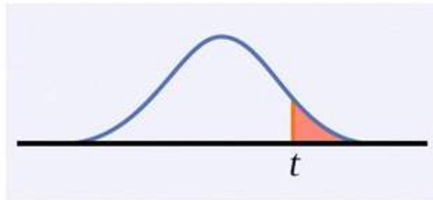
Significance Test for β_j, j=0,1,…,p

- To test the hypothesis H_0: $\beta_j = 0$ versus a one- or two-sided alternative, we calculate the t statistic

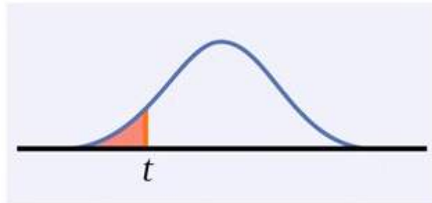
$$t = b_j / SE_{b_j} \sim t\,(n - p - 1)$$

distribution when H_0 is true. The P-value of the test is found in the usual way.

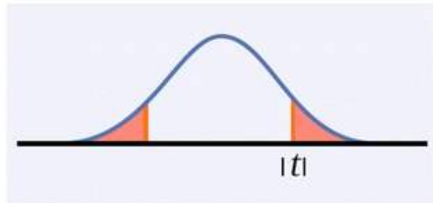
H_a: $\beta_j > 0$ is $P(T \geq t)$



H_a: $\beta_j < 0$ is $P(T \leq t)$



H_a: $\beta_j \neq 0$ is $2P(T \geq |t|)$



Note: *Software typically provides two-sided P-values.*

Example 2: Menu Pricing in a New Italian Restaurant in New York City

- The initial regression model is

Price = – 24.02 + 1.54 Food + 1.91 Decor – 0.003 Service + 2.07 East

At this point we shall leave the variable Service in the model even though its regression coefficient is not statistically significant.

- The variable Décor has the largest effect on Price since its regression coefficient is largest.
- Note that Food, Décor and Service are each measured on the same 0 to 30 scale and so it is meaningful to compare regression coefficients.
- The variable Décor is also the most statistically significant since its *p* -value is the smallest of the three.
- In order that the price achieved for dinner is maximized, the new restaurant should be on the east of Fifth Avenue since the coefficient of the dummy variable is statistically significantly larger than 0.
- It does not seem possible to achieve a price premium for "setting a new standard for high quality service in Manhattan" for Italian restaurants since the regression coefficient of Service is not statistically significantly greater than zero.

```
Call:
lm(formula = Price ~ Food + Decor + East, data = nyc)

Residuals:
    Min      1Q  Median      3Q     Max
-14.0451 -3.8809  0.0389  3.3918 17.7557

Coefficients:
            Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) -24.0269   4.6727     -5.142    7.67e-07 ***
Food          1.5363   0.2632      5.838    2.76e-08 ***
Decor         1.9094   0.1900     10.049    < 2e-16 ***
East          2.0670   0.9318      2.218    0.0279 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.72 on 164 degrees of freedom
Multiple R-squared:  0.6279,  Adjusted R-squared:  0.6211
F-statistic: 92.24 on 3 and 164 DF,  p-value: < 2.2e-16
```

- Dropping the predictor Service from the initial model
- The regression coefficients for the variables in both models are very similar.

WEEK 8 Aim 2 ANOVA in MLR

- $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$
- H_A: at least one of the $\beta_i \neq 0$

Analysis of variance table

Source of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS)	F
Regression	p	SSreg	SSreg/p	$F = \dfrac{\text{SSreg}/p}{\text{RSS}/(n-p-1)}$
Residual	$n-p-1$	RSS	S^2= RSS/ $(n-p-1)$	
Total	$n-1$	SST = SYY		

Partial F-test for comparing models

- ‘small’: model with only `logMiles` $(SSE = RSS = 325216)$
  - $y = \beta_0 + \beta_1 x_1 + \epsilon$ (`q=2 parameters,n=51`)
- ‘big’: model with all four explanatory variables $(SSE = RSS = 193700)$
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$ (add c=3 parameters, $\beta_2, \beta_3, \beta_4$)

$H_0: \beta_2 = \beta_3 = \beta_4 = 0$ against $H_A: \beta_2, \beta_3, \beta_4$ are not all zero

$$F = \frac{(RSS_{\mathrm{small}} - RSS_{\mathrm{big}})/(c\ )}{(RSS_{\mathrm{big}})/(\mathrm{df}_{\mathrm{n-q-c}}\ )} \sim F_{(c,n-q-c)}$$

$$F = \frac{(325216 - 193700)/(3)}{193700/(n-q-c)} = \frac{131516/3}{4210.87} = 10.411 \sim F_{3,46}$$

and $p\left(F_{3,46} > 10.411\right) = 2.4 \times 10^{-5}$

(Reject Ho, we need to add the at least 1 of the 3 variables)

In *R*: partial F-test for comparing models

```
> anova(Fuel.lm0, Fuel.lm1)
Analysis of Variance Table

Model 1: Fuel ~ logMiles
Model 2: Fuel ~ Tax + Dlic + Income + logMiles
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     49 325216
2     46 193700  3    131516 10.411 2.402e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Influential observations

- Single or small groups of observations can strongly influence the fit of a regression model
- ***Influence analysis*** studies changes in a specific part of an analysis under the assumption that the model is correct
  - ‘Easy’ way would be to delete observations from the data one at a time and then study its effects, for example, changes in coefficients $\widehat{\boldsymbol{\beta}}$
  - Observations whose removal causes major changes are called *influential*
- A useful measure of influence is *Cook’s Distance*, D_i, which reflects two aspects: a large residual *and* a large leverage:

$$D_i = \frac{r_i^2}{2}\frac{h_{ii}}{1-h_{ii}},$$

- A useful rule of thumb is that a point is an influential observation if

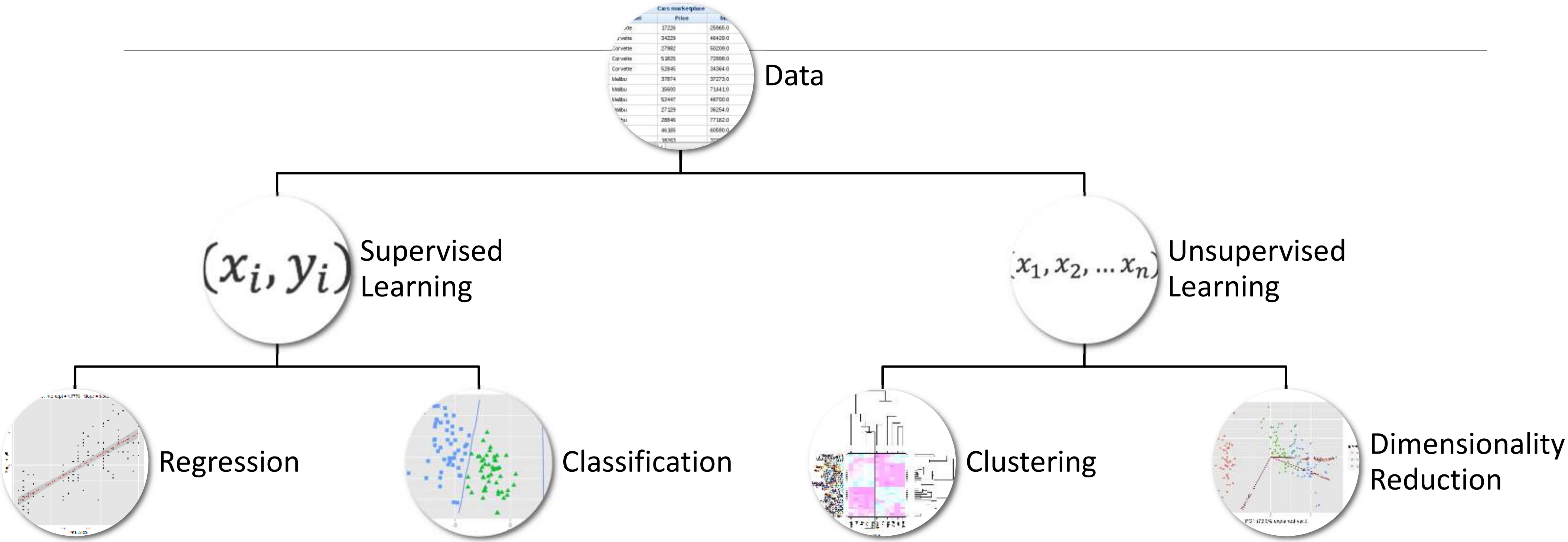
$$D_i > \frac{2(p+1)}{n-(p+1)}$$

Aim 2 Statistical modelling, machine learning, & predictive analytics

One of the main objectives of data science is to be able *predict* the future using statistical models and/or machine learning algorithms

- Predict whether someone has a certain type of cancer based on the over/under-expression of proteins from DNA sequences
- Predict credit risk based on an individual's financial records, demographic data, educational attainment, …
- Predict the frequency and intensity of tropical cyclones by integrating information from climate models and historical data
- Predict the value of a home using historical data, information about amenities in the suburb, and information about the characteristics of the house itself
- Predict the probability that a student will withdraw using information collected passively

Statistical models/machine learning algorithms



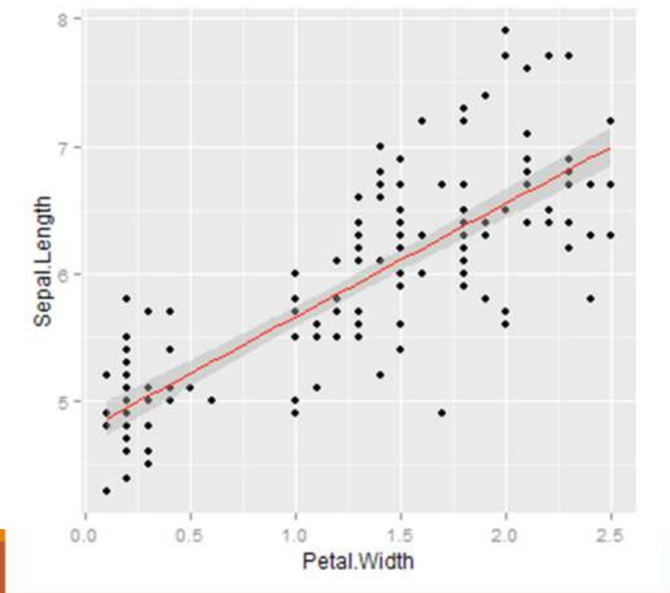
Terminology

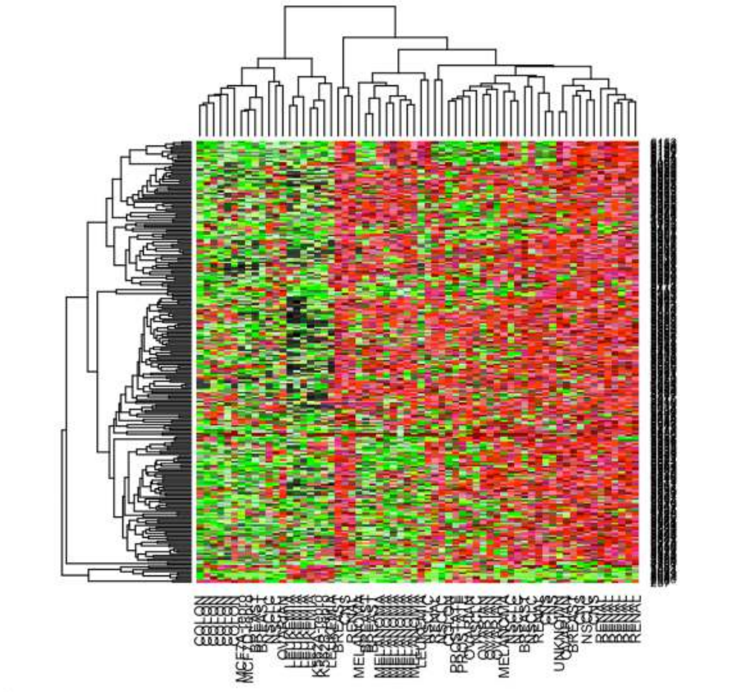
Supervised learning: modelling a specific response variable as a function of some explanatory variables

- Linear and nonlinear regression; neural networks; classification trees

Unsupervised learning: approaches to finding patterns or groupings in data where there is no clear response variable

- Clustering, principal component analysis





Statistical Models/Machine Learning algorithms

Regression
Eg Linear or logistic regression

Distance
Eg K Nearest Neighbour (kNN)

Regularization
Eg Ridge regression, LASSO, LARS

Decision Trees
Eg CART (Classification and Regression Trees), Random Forest

Bayesian
Eg Naïve Bayes, Bayesian Networks

Clustering
Eg k-means clustering; Hierarchical clustering

Association Rule Mining
Eg Context based rule mining, apriori

Artificial Neural Networks
Eg ANN

Deep Learning
Eg DBM, Deep belief networks

Dimensionality reduction
Eg PCA, Linear Discriminant Analysis

Ensemble
Eg Bagging, Boosting

Text mining
Eg Sentiment analysis, Speech recognition

Aim 4 Final Exam Information (DRAFT)

READING THE POLICIES and Check the Timetable

- Please take the time to familiarise yourself with the instructions below to prepare you for your exam experience. These guidelines have been written to ensure adherence with the UWA Assessment Policy.
- Read the information on online exams on the Exams website which includes information to reduce the likelihood of any technical problems during your exam.
- Check your personal published **Exam Timetable** available on **studentConnect**.

Aim 4 Final Exam Information (DRAFT)

THE UNIVERSITY OF WESTERN A

Preparing for a LMS MS Teams Exam – MONITORED

Availability: Item is hidden from students.

Getting Prepared

- To ensure that you correctly sit your online exam, please install the MS Teams app on your computer. MS Teams is free to download - **DO NOT USE MS Teams through a web browser**.
- You need to be logged into MS Teams with your UWA Pheme account - not a 'guest account'
- Connect to a power source and/or have your power cord available and sit near a power source
- Test your webcam (if required) and check it's correctly positioned and working (same with the microphone)
- Have your UWA approved ID ready for inspection as well as your workspace and allowable items
- Headphones and hats are not permitted unless authorised via specific UniAccess requirements
- You must be visible at ALL TIMES during your exam or your work will not be marked

Accessing the MS Teams session

The link to the MS Teams exam supervision session will be published under the folder:

MS Teams session links

It will **appear 45 minutes before your exam start time.** Please patiently wait in the lobby while the exam supervisor checks individual student's ID and workspace. Follow any directions that you are given. You ***may*** be given a **password to access your exam** so you can all start at the same time.

Technical Support

- Inform your exam supervisor if you encounter LMS or IT issues (related to the University system). He/she will try to assist you if possible.
- Ask permission from your exam supervisor if you need to ring the Exams Support phone number **+ (61) 8 6488 1212**

Temporarily leaving the exam for any reason

The exam supervisors will report this to the University as needed.

Final Exam Information

THE UNIVERSITY OF WESTERN AUSTRALIA

1. Materials Permitted:

The prescribed materials are:

- You will only use **lecture slides, computer labs materials and solutions** (RMarkdown or HTML) **available on the LMS STAT2401 under Learning Materials**
- You are permitted to use your own electronic notes **related to the materials that can be accessed from One Drive UWA folder**
- You are permitted to use scrap paper for writing or calculation, which must be prepared prior to the start of the assessment

2. NOT PERMITTED

- You are **not permitted to use internet** other than to access the LMS STAT2401, RStudio/RMarkdown and MS Teams.
- You are **not permitted to open a browser for internet search.**
- **ChatGPT or similar AI tools are strictly prohibited to be used in the examination in any form.**

Academic Integrity. Please read thoroughly the link academic integrity policy

Final Exam Information

THE UNIVERSITY OF WESTERN AUSTRALIA

- **What are available on the LMS during the final examination?**

The LMS restrictions will be applied from Monday 10th June at 7:00am

- **Communication:** Announcement Tab
- **Unit Information**: Unit Outline Tab
- **Unit Materials**: Learning Materials
- **Assessment**: Final Exam 2024 (will be available in Study Week)
  - You must read the information thoroughly
  - The examination can be found within this tab.

Final Exam Information

Preparing for the examination

- **The environment**
  1. Familiarise yourself using the system, accessing LMS, MS Teams, R and R Studio, One Drive folder.
  2. R and RStudio are installed in your laptop, including R packages that you have used in the semester.
  3. Make sure you know how to access your **One Drive folder** as a UWA student.
     1. Create a folder STAT2401 Exam within **One Drive folder.** You will use the folder to save Rmd template for the exam and datasets that will be used for the exam
     2. **Students can access the web version at https://uniwa-my.sharepoint.com/ and sign in with StudentNo@student.uwa.edu.au where StudentNo is their student number.**
     3. **If you are new to One Drive, please read the information within this link https://www.uwa.edu.au/library/help-and-support/student-email-and-collaboration-tools**

Final Exam Information

THE UNIVERSITY OF WESTERN AUSTRALIA

- **The exam (will be available within Final Exam 2024 tab, managed by the university)**
  - This is a 2-hour examination, 1 attempt only.
  - Assessing Lecture and Lab Weeks 2-12 inclusive.
  - The exam is very similar to online tests that you have completed before the examination
  - The questions may be set as *multiple choice*, *matching*, *calculated numeric*, *multiple blanks*.
  - **You are provided with Rmd template for your working that you are strongly recommended to submit under File Response for partial marks.**
  - **The exam is password protected. A password will be provided before the exam starts.**
  - If you have any question or comment about the exam or would like to alert your Unit Coordinator to a perceived error **during the examination,** include a comment in your working in Rmd file, if appropriate, to indicate how you interpreted the question.