THE UNIVERSITY OF WESTERN AUSTRALIA

STAT2401 Analysis of Experiments

Lecture Week 5 **Dr Darfiana Nur**

Aims of Lecture Week 5

- **Aim 1 Confidence Intervals for the Population Regression Line μ_y**

(Sheather Ch 2.3, Moore et al Ch 10, James ET AL 2023 Ch 2)

- **Aim 2 Prediction Intervals for the Actual Value of Y**

(Sheather Ch 2.4, Moore et al Ch 10, James ET AL 2023 Ch 2)

- **Aim 3 ANOVA** (Sheather Ch 2.5, Moore et al Ch 10, James ET AL 2023 Ch 2)

3.1 The detail

3.2 Coefficient of Determination R^2

Recap 1: Simple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

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

- The residuals have 0 mean and constant variance σ^2
and are independent (follow no pattern)

- To estimate the parameters, we minimise

$$SSE = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Recap 2: Least squares estimates

$$\hat{\beta}_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \sum c_i Y_i$$

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

$$\hat{\sigma}^2 = Var(\hat{\varepsilon}_i) \approx \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

This is called the
Mean square error (MSE)
See ANOVA

These estimates are **unbiased**

$$E(\hat{\beta}_0) = \beta_0 ; E(\hat{\beta}_1) = \beta_1 ; E(\hat{\sigma}^2) = \sigma^2$$

The sampling distributions of slope and intercept estimates

We have also obtained last week that

$$s.e.(\hat{\beta}_0) = \sqrt{\sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right)}; s.e.(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum(X_i - \overline{X})^2}}$$

If we assume the residuals are normally distributed

$$\hat{\beta}_0 \pm t_{n-2;1-\alpha/2} \times s.e.(\hat{\beta}_0) \qquad \frac{\hat{\beta}_0 - \beta_0}{s.e.(\hat{\beta}_0)} \sim t_{n-2}$$

$$\hat{\beta}_1 \pm t_{n-2;1-\alpha/2} \times s.e.(\hat{\beta}_1) \qquad \frac{\hat{\beta}_1 - \beta_1}{s.e.(\hat{\beta}_1)} \sim t_{n-2}$$

Sampling distributions

These are $(1-\alpha)\%$ confidence intervals for the true parameters β_0, β_1

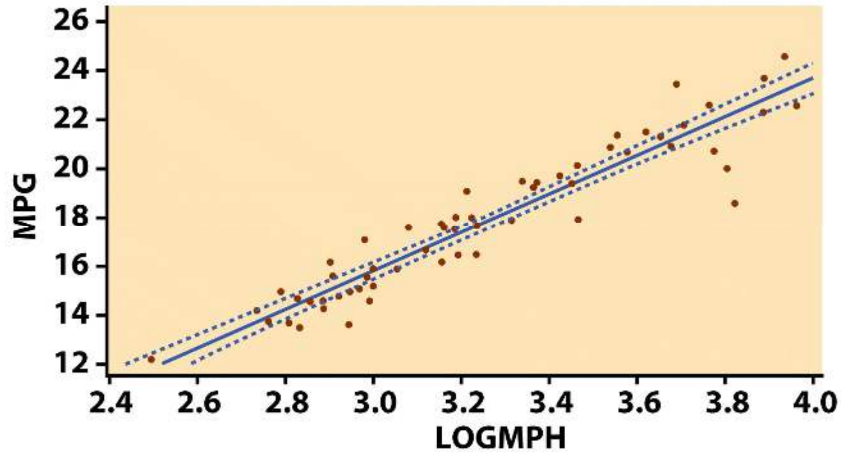
Aim 1 Confidence Intervals for the Population Regression Line μ_y

- We can also calculate a confidence interval for the population mean μ_y of all responses y when x takes the value x^* (within the range of data tested).

- The **level C confidence interval for the mean response μ_y** at a given value x^* of x is: $\hat{\mu}_y \pm t^* SE_{\hat{\mu}}$

where t^* is the value such that the area under the $t(n-2)$ density curve between $-t^*$ and t^* is C.

- A separate confidence interval could be calculated for μ_y along all the values that x takes.
- Graphically, the series of confidence intervals is shown as a continuous curve on either side of $\hat{y}$.



Prediction and forecasting I (for μ_y)

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

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

- 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

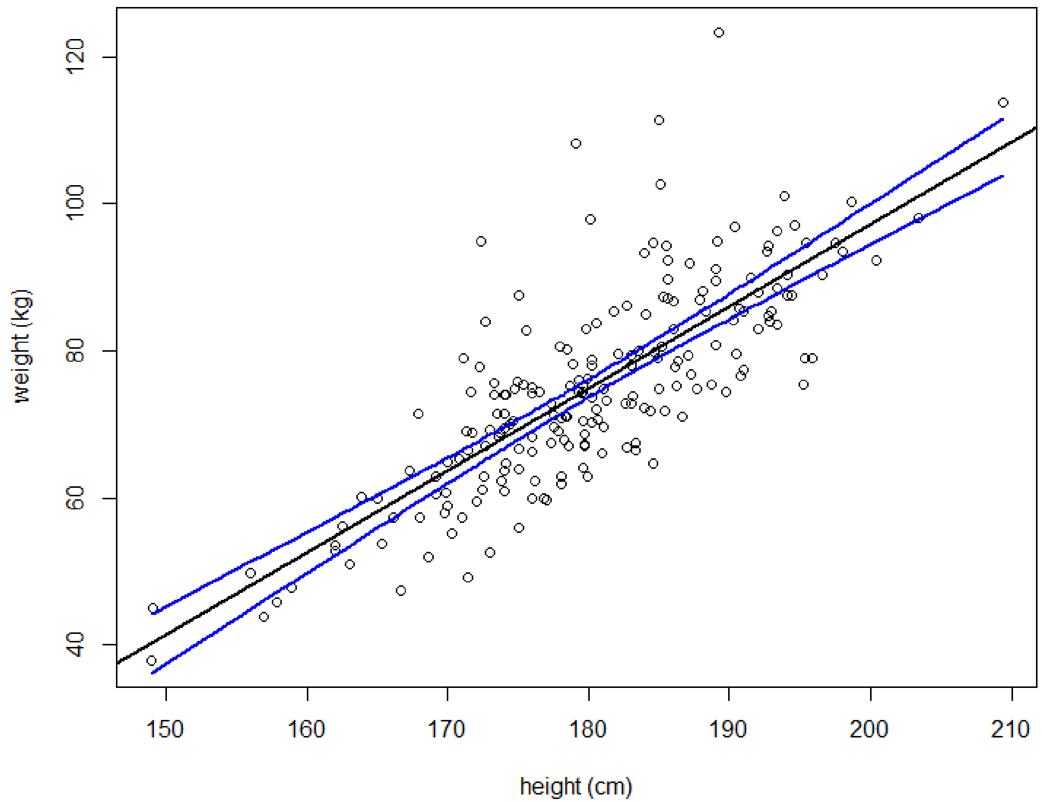
If we know the true value of σ^2 then we can use

Example 1 Confidence intervals for the population regression line: AIS data Using R

```
> ConfMean <-
   predict(ais.lm,
    interval = “confidence”)

> head(ConfMean)
       fit      lwr      upr
1 92.65419 90.34393 94.96445
2 85.72807 84.02709 87.42904
3 72.43438 71.19091 73.67784
4 80.47762 79.12265 81.83258
5 80.03077 78.69750 81.36404
6 68.18933 66.76028 69.61839

> matlines(sort(ais$Ht),
   ConfMean[order(ais$Ht), 2:3],
   lwd = 2, col = "blue",
  lty = 1)
```



AIM 2. Fitted (predicted) values

- Recall equation of a least squares line:

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X$$

- Once we have the equation, we can use it to predict *y* for each actual *x* value in the data.

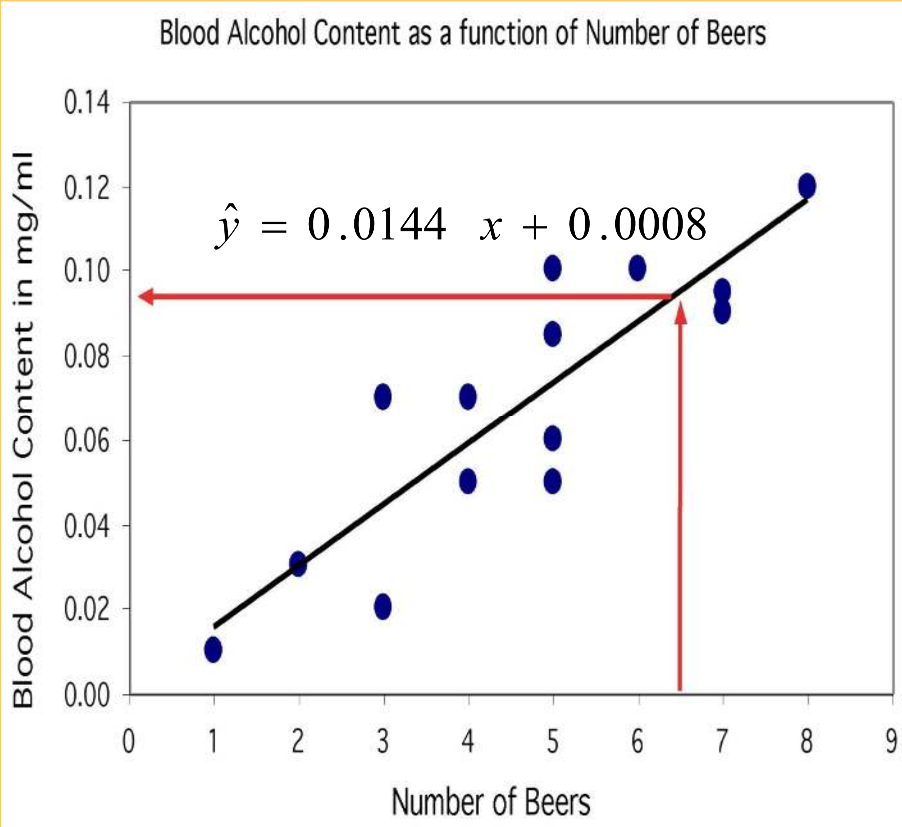
- These are called predicted values.

- The differences between the observed values and the predicted values are called the residuals:

$$\text{residual} = y - \hat{y}$$

Making predictions

The equation of the least-squares regression allows you to predict *y* for any *x* within the range studied.



Example 2

Nobody in the study drank 6.5 beers, but by finding the value of $\hat{y}$ from the regression line for x = 6.5 we would expect a blood alcohol content of 0.094 mg/ml.

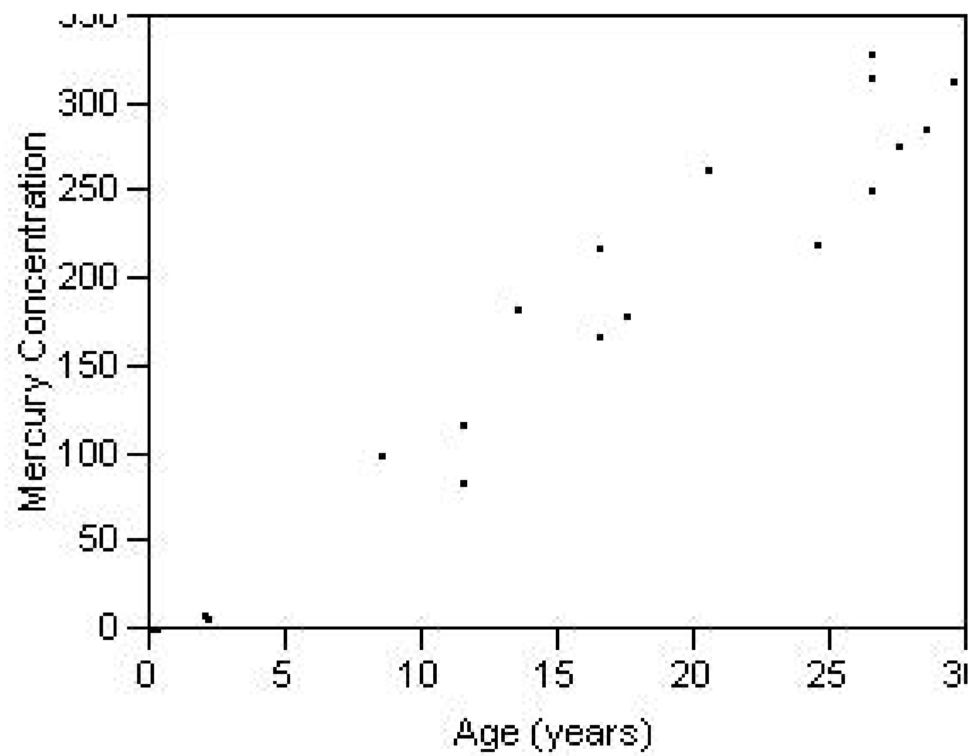
$$\hat{y} = 0.0144*6.5 + 0.0008$$

$$\hat{y} = 0.936 + 0.0008 = 0.0944\text{mg/ml}$$

Example 3:

Dolphin example.

Data for 19 dolphins was collected as part of a marine population study. The data contains the mercury concentration (*y*) in the liver of striped dolphins against the age of the dolphins (*x*).



Mercury Concentration (μg/g)
= -2.65 + 10.90 Age (years)

Example 3 (continued):

Use of least squares (regression) to make predictions

Predict the expected mercury concentration in a dolphin aged 18 years old.

Here $x = 18$.

To predict the value of y, we simply substitute this value of x back into the equation for the regression line:

Use of least squares (regression) to make predictions

$$\begin{aligned} \text{mercury concentration} &= -2.65 + 10.90\text{age} \\ &= -2.65 + 10.90(18) \\ &= 193.55\,\mu\text{g/g} \end{aligned}$$

In other words, we predict the expected mercury concentration in a dolphin aged 18 to be 193.55μg/g.

- **Warning:** Making predictions for a value of ***y*** when the value of ***x*** is outside the range of observed ***x*** values is prone to error and often not accurate.

In-Class Exercise 1:

Use of least-squares (regression) to make predictions

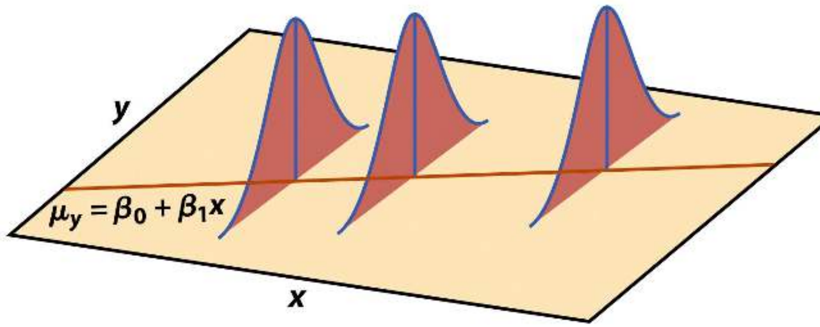
- For example, would it be advisable to use the equation of our regression line to predict the mercury concentration for a dolphin if their age is 40 years old?
- Our *x* values end at approximately 30 years. Beyond this value, the relationship between *y* and *x* might change, so our regression line may no longer be valid.

Prediction Intervals 1

- One use of regression is for **predicting** the value of *y* at some value of *x* within the range of data tested. Reliable predictions require statistical inference.

- To estimate an *individual* response *y* for a given value of *x*, we use a **prediction interval**.

- If we randomly sampled many times, there would be many different values of *y* obtained for a particular *x* following $N(0, \sigma)$ around the mean response μ_y.



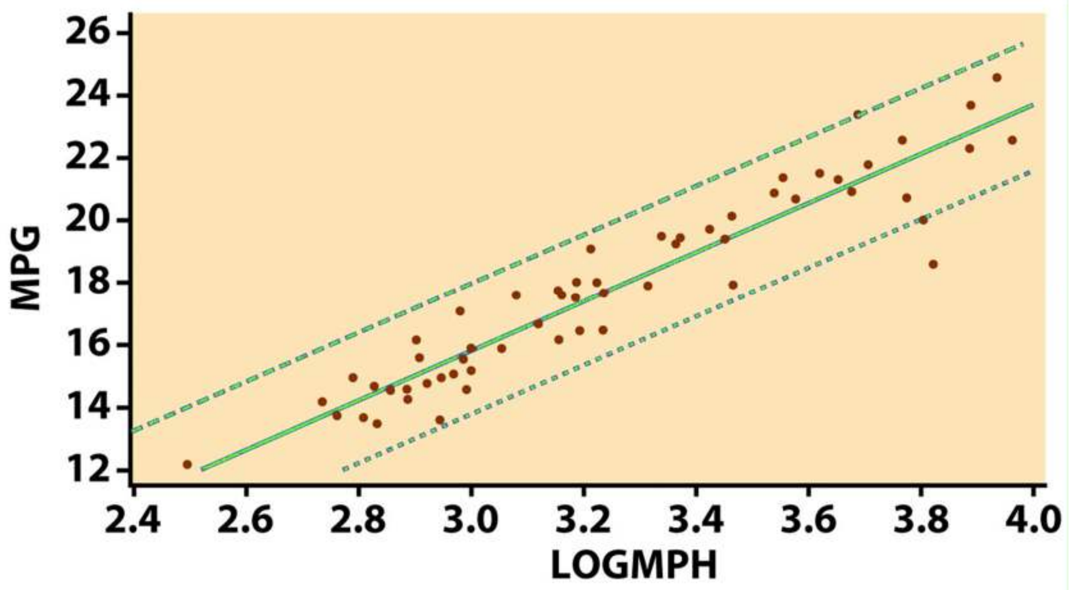
Prediction Intervals 2

The **level *C* prediction interval for a single observation** on *y* when *x* takes the value *x** is

$$\hat{y} \pm t^* SE_{\hat{y}}$$

*t** is the critical value for the ***t* (*n* – 2)** distribution with area C between –*t** and +*t**.

- The prediction interval accounts for error in estimating β_0 and β_1 as well as uncertainty about the value of *y* being predicted.
- Graphically, the series of prediction intervals is shown as a continuous curve on either side of $\hat{y}$.
- These prediction intervals are wider than the corresponding confidence intervals for μ_y.



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.

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

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.

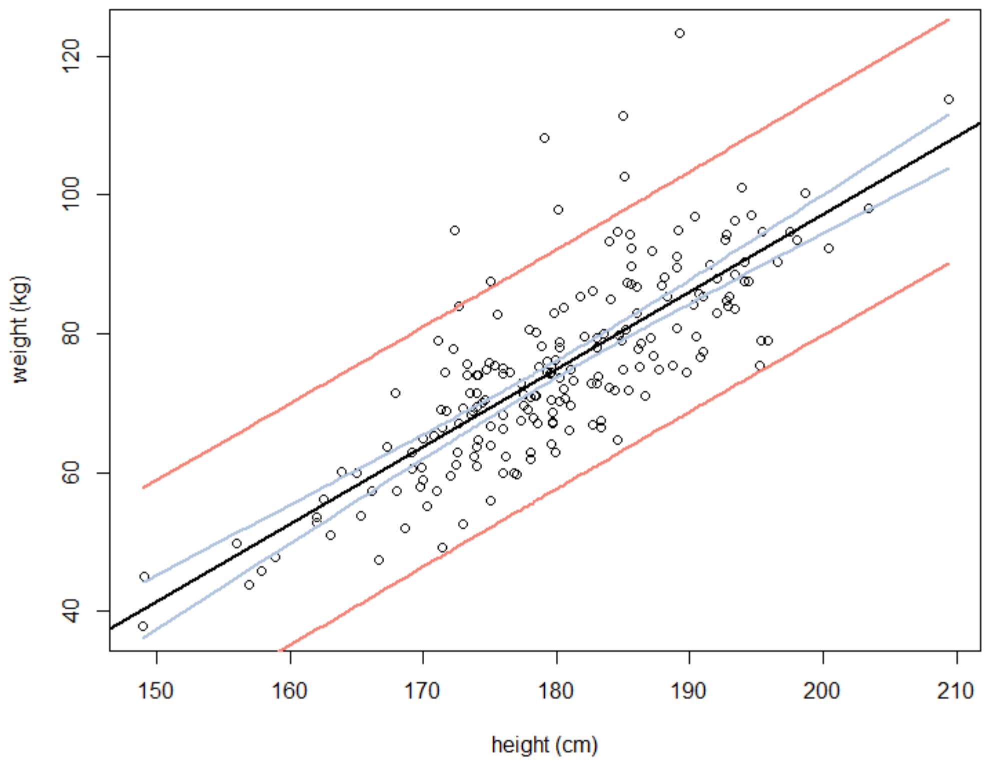
$$(\hat{\beta}_0 + \hat{\beta}_1 X_0) \pm t_{n-2;1-\alpha/2} s.e.(Y \mid X_0)$$

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

Example 4: Prediction intervals AIS data

```
> PredInt <-
   predict(ais.lm,
    interval = "prediction")

> head(PredInt)
       fit      lwr       upr
1 92.65419 75.30416 110.00423
2 85.72807 68.44861 103.00753
3 72.43438 55.19394  89.67481
4 80.47762 63.22878  97.72645
5 80.03077 62.78363  97.27792
6 68.18933 50.93452  85.44415
> matlines(sort(ais$Ht),
   PredInt[order(ais$Ht), 2:3],
   lwd = 2, col = "red",
   lty = 1)
```



Calculations for Regression Inference
A summary so far

To assess variation in the estimates of β_0 and β_1, we calculate the standard errors for the estimated regression coefficients.

The standard error of the slope estimate b_1 is

$$SE_{b1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$$

The standard error of the intercept estimate b_0 is

$$SE_{b0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$

To estimate mean responses or predict future responses, we calculate the following standard errors.

The standard error of the estimate of the mean response μ_y is

$$\mathrm{SE}_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x_i - \overline{x})^2}}$$

The standard error for predicting an individual response y is

$$\mathrm{SE}_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x_i - \overline{x})^2}}$$

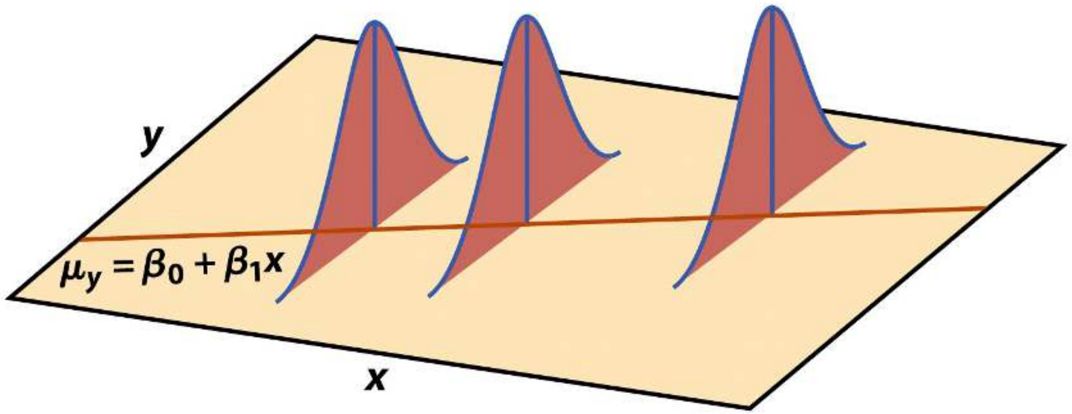
Aim 3 Analysis of Variance (ANOVA) for Regression

- The regression model is

Data = $\boxed{\text{fit}}$ + $\boxed{\text{error}}$

$y_i = \boxed{(\beta_0 + \beta_1 x_i)} + \boxed{(\varepsilon_i)}$

where the ε_i are **independent** and **Normally** distributed $N(0,\sigma)$, and σ is the same for all values of x.



- It resembles an ANOVA, which also assumes equal variance, where

SST = $\boxed{\text{SS Regression}}$ + $\boxed{\text{SS error}}$ and

DFT = $\boxed{\text{DF Regression}}$ + $\boxed{\text{DF error}}$

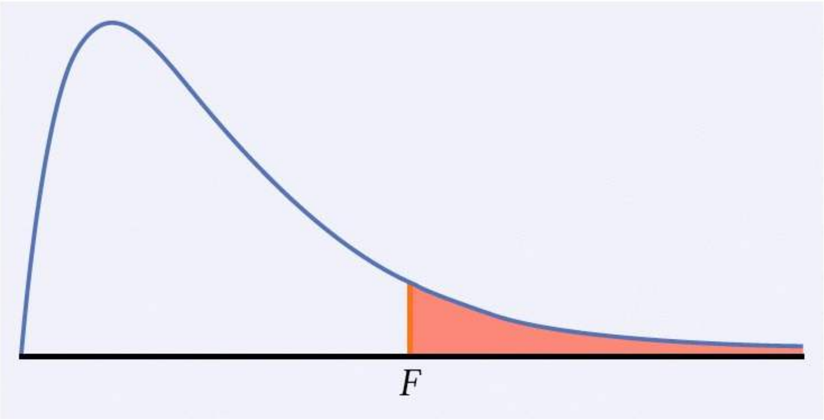
The ANOVA *F* Test

- 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***.

Aim 3.1 Analysis of Variance: The detail

o Before moving on to multiple regression and more complicated models, it is useful to learn about ANOVA.

o ANOVA is used to compare different (but similar) models and choose the best model for a particular set of data.

Consider the two competing models

$$(1)\ \ Y_i = \beta_0 + \varepsilon_i\,; \quad (2)\ \ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$(1)\quad Y_i = \beta_0 + \varepsilon_i\,; \qquad (2)\quad Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Model (1) asserts that the observations Y are unrelated to the variable X and have a constant mean β_0 plus constant variance σ^2

The least squares line for this model minimises

$$SSE = \sum (Y_i - \beta_0)^2 \Rightarrow \hat{\beta}_0 = \bar{Y}$$

The residual sum of squares for this model is

$$\sum (Y_i - \hat{\beta}_0)^2 = \sum (Y_i - \bar{Y})^2 = SST = (n-1)s_Y^2$$

- For model (2) we know that

$$SSE = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum \hat{\varepsilon}_i^2 = (n-2)\hat{\sigma}_\varepsilon^2$$

- We wish to measure the effect that fitting a linear model has in reducing the residual variation over model (1). We can measure this as follows

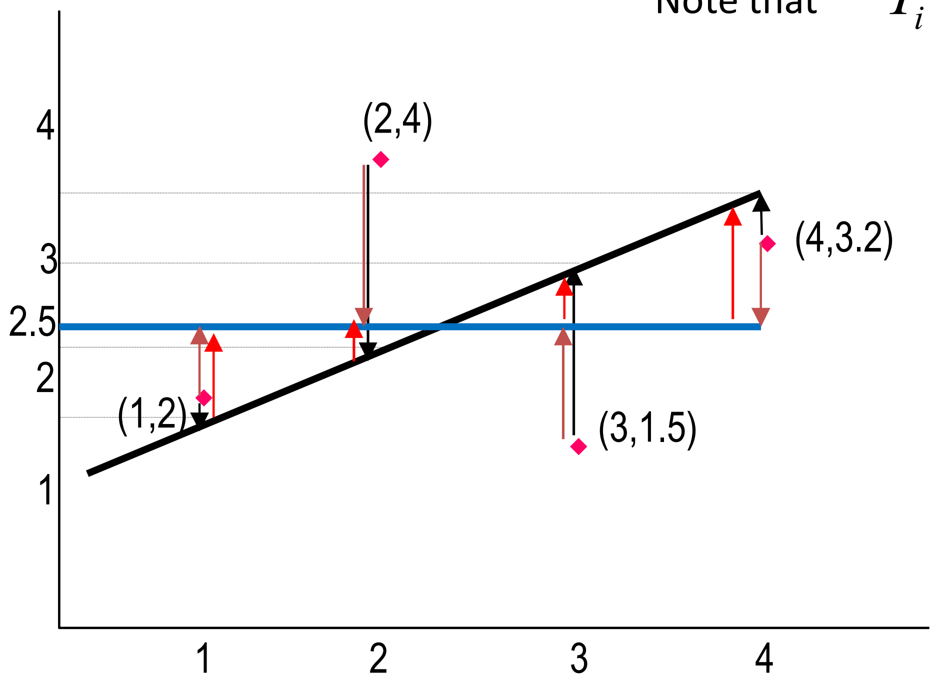
$$SSR = SST - SSE = \sum (Y_i - \overline{Y})^2 - \sum \hat{\varepsilon}_i^2$$

$$= (n-1)s_Y^2 - (n-2)\hat{\sigma}_\varepsilon{}^2$$

- This is called the sum of squares due to regression. It measures how much the residual variation has decreased by fitting a linear model rather than a constant mean to the data.

- The blue arrows represent SST, where a constant mean is assumed
- The black arrows represent SSE, where the linear model is fit
to the data
- The red arrows represent the difference between the blue line
(constant mean) and the black line (linear model)

Note that $Y_i - \overline{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})$



- It is not hard to show (Sheather, 2009) that

$$SSR = SST - SSE = \sum (Y_i - \overline{Y})^2 - \sum (Y_i - \hat{Y})_i^2$$
$$= \sum (\hat{Y}_i - \overline{Y})^2$$

- If *Y* and *X* are linearly related then *SSR* will be large

- If *Y* and *X* are not linearly related then *SSR* will be small

- Degrees of freedom

$$SST = SSR + SSE$$
$$(n-1) = 1 + (n-2)$$

- If we divide a sum of squares by it’s degress of freedom, we obtain a Mean Square (MS)

$$MST = SST/(n-1) = s_Y^2 = \hat{\sigma}_{(1)}^2$$

$$MSE = SSE/(n-2) = \hat{\sigma}_{(2)}^2$$

$$MSR = SSR/1$$

- For the linear model (2) to fit better than the constant mean model (1) we want *MSE* to be much less then *MST*

(otherwise, why bother with the linear model)

- The *F*-statistic compares the *MSR* (=*SSR*/1) to the *MSE.*

$$F = \frac{(SST - SSE)}{MSE} = \frac{SSR/1}{\hat{\sigma}^2_{(2)}}$$

- So, if *SSR* is large then *F* will be large, if *SSR* is ‘small’ then *F* will be ‘small’.

- *F* is a statistic and is subject to random variation as it depends on the parameter estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2, s^2, \overline{Y}$

and the sample data *Y*, all of which are subject to variation.

- *F* has two different degrees of freedom, one for the numerator and one for the denominator

$$F = \frac{SSR/1}{SSE/(n-2)}$$

- This statistic has an *F* distribution with 1 and (n-2) degrees of freedom

- We can use this statistic to test

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

- The p-value in this case is

$$\Pr(F > F_{1-\alpha;1,n-2})$$

We can use R to calculate this for us

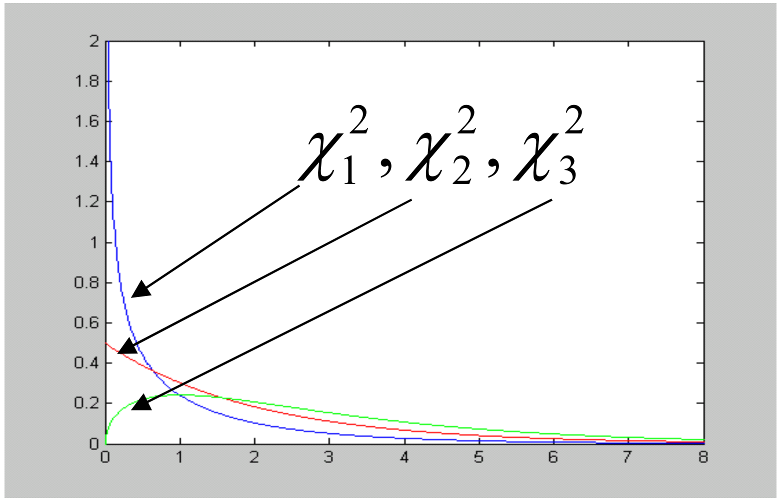
Assumptions for ANOVA

- The statistic will have an F-distribution only if the residuals are **normally distributed.** When using this test we should examine the residuals. If they are highly non-normal then the test is not valid.

- Usually, it is enough if the histogram is roughly mound shaped

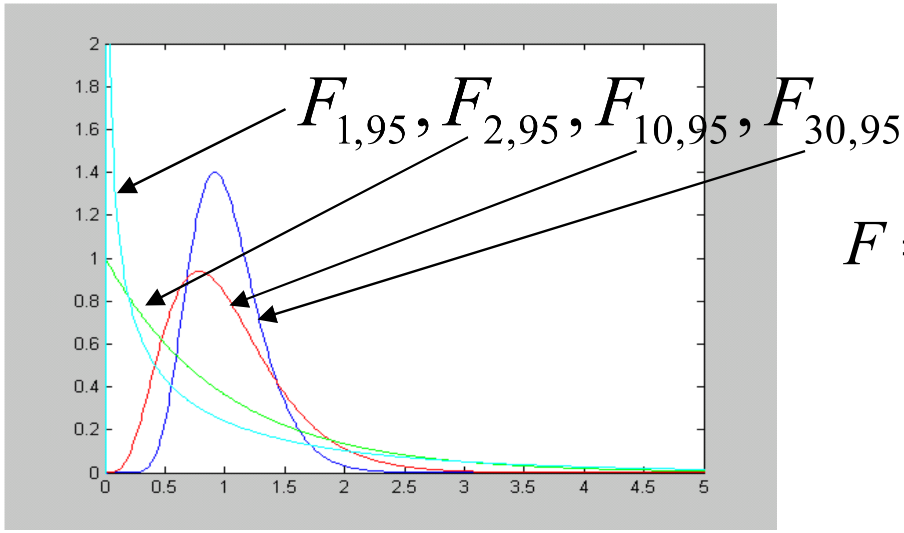
- This means the histogram is roughly symmetric with highest density in the middle and lowest density in the tails

Both *SSE* and *SSR* follow chi-squared distributions with (n-2) and 1 degrees of freedom respectively



A chi-squared variable describes the distribution of a sum of squares of normal data minus its mean, like

$$SSE = \sum (Y_i - \hat{Y}_i)^2 \sim \chi_{n-2}^2$$



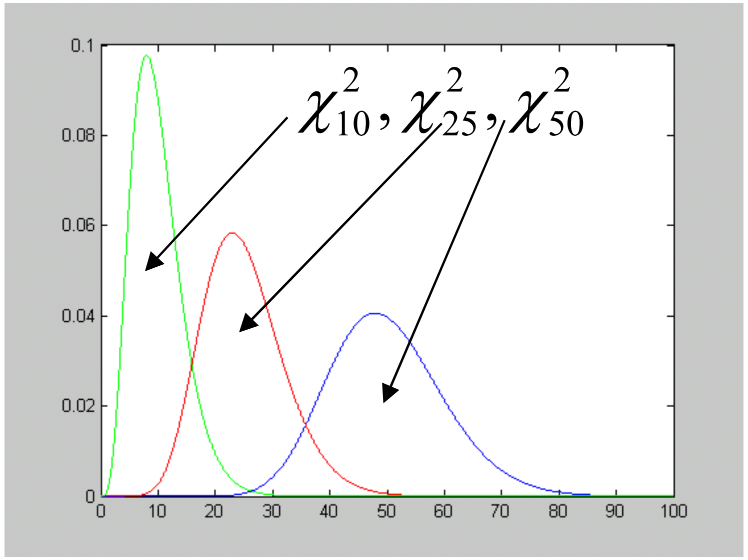
$$F = \frac{SSR/1}{SSE/(n-2)} = \frac{\chi_1^2/1}{\chi_{n-2}^2/(n-2)} \sim F_{1,n-2}$$

Chi-Squared and F distributions

If *Y* has a normal distribution with estimated mean $\hat{Y}$ and variance σ^2 then

$$\frac{\sum (Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi^2_{df}$$

Where *df* is (*n* - no. of parameters estimated) to get $\hat{Y}$



So,
SSE =

$$\sum (Y_i - \hat{Y}_i)^2$$

$$\sim \chi^2_{n-2}$$

$$E(\chi^2_{df}) = df$$

An F distribution is formed by two different chi-squared distributions. Say,

$$U \sim \chi^2_u$$

$$V \sim \chi^2_v$$

Then,

$$F = \frac{U/u}{V/v} \sim F_{u,v}$$

Which is an F distribution with *u* and *v* degrees of freedom.

Usually

$$E(F) \approx 1$$

T-test or F test (ANOVA)

- Although we previously used a t-test

$$T = \frac{\widehat{\beta}_1 - 0}{se(\widehat{\beta}_1)} \sim t_{n-2}$$ to test $H_0\colon \beta_1 = 0$ against $H_A\colon \beta_1 \neq 0$

- We can also use the test statistic

$$F = \frac{\mathrm{SSReg}/1}{\mathrm{RSS}/(n-2)} \sim F_{1,n-2}$$ when H_0 is true

- The two are related by $F = T^2$

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:

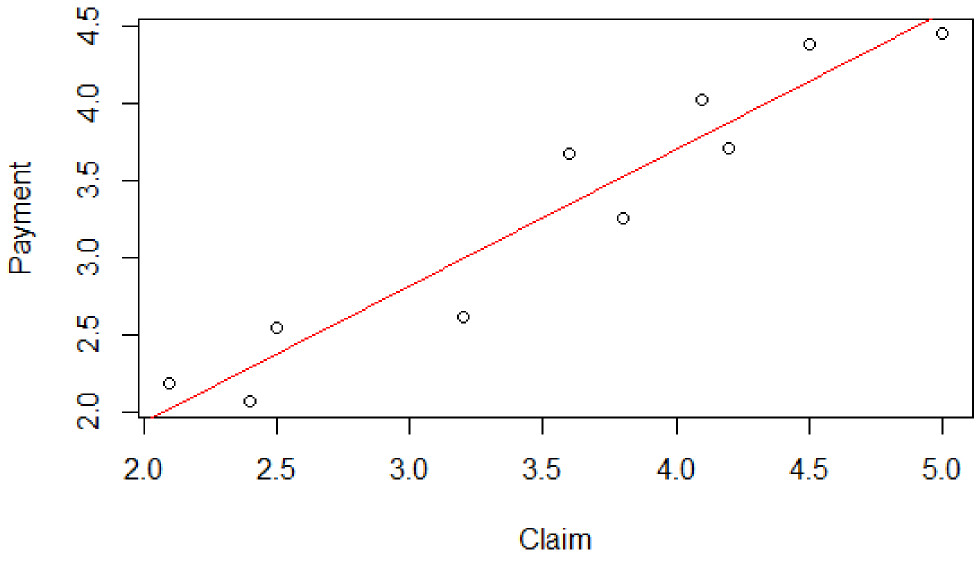
Claim	2.10	2.40	2.50	3.20	3.60	3.80	4.10	4.20	4.50	5.00
Payment	2.18	2.06	2.54	2.61	3.67	3.25	4.02	3.71	4.38	4.45

```
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 '.' 0.1 ' ' 1

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: $\beta 1 = 0$ vs. Ha: $\beta 1 \neq 0$
- 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: $\beta 1 = 0$ vs. Ha: $\beta 1 \neq 0$
- 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

Aim 3.2 R^2 Coefficient of determination

- In the equations

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\text{SST} = \text{SSReg} + \text{SSE}$$

- The proportion of the total variability (SST) explained by the regression is known as R^2, i.e.,

$$R^2 = \frac{\text{SSreg}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

Coefficient of determination R^2

- The coefficient of determination describes the **fraction of the variation in the values of *y* that is explained by the least squares regression line.**

- It is a number between 0 and 1 which we normally convert to a percentage between 0 and 100%.

- The higher the value of the coefficient of determination, the **better the least squares regression line is in explaining the variation in the data**.

- R^2 is the **correlation coefficient (*r*) squared.**

Example 5: Explaining R^2

- The output lm() in slide 38 (see below) implies that 91.82% of the information (variation) in insurance payment is explained by the least squares regression line, suggesting that the model is a good one.

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 6 Simple Linear Regession

Example: Cholesterol Data Set

- Data from 1109 West Australians with measurements pertaining to body mass, cholesterol, blood pressure and other data of medical obsession
- Is `BMI` related to `cholesterol`?

```
> load("cholesterol.RData")
> str(cholesterol)
'data.frame': 1109 obs. of  8 variables:
 $ AGE   : num  32 40 39 37 46 44 51 50 49 49 ...
 $ BMI   : num  24.2 26.3 25.1 28.7 26.3 ...
 $ CHOL  : num  4.7 5.8 5.5 5.6 5.9 5.8 5.4 6 4.9 7.2 ...
 $ DBP   : num  70 70 70 80 80 84 90 80 84 90 ...
 $ HEIGHT: num  175 183 182 183 185 180 170 173 187 178 ...
 $ SEX   : Factor w/ 2 levels "female","male":  2 2 2 2 2 2 2 2 2 2 ...
 $ WAIST : num  82 93 91 98 95 84 82 92 95 89 ...
 $ WEIGHT: num  74 88 83 96 90 76 72 75 98 89 ...
 - attr(*, "variable.labels")=  Named chr  "age" "BMI" "cholesterol" "DBP"  ...
  ..- attr(*, "names")= chr  "AGE" "BMI" "CHOL" "DBP" ...
```

Simple Linear Regession

Example: Cholesterol Data Set

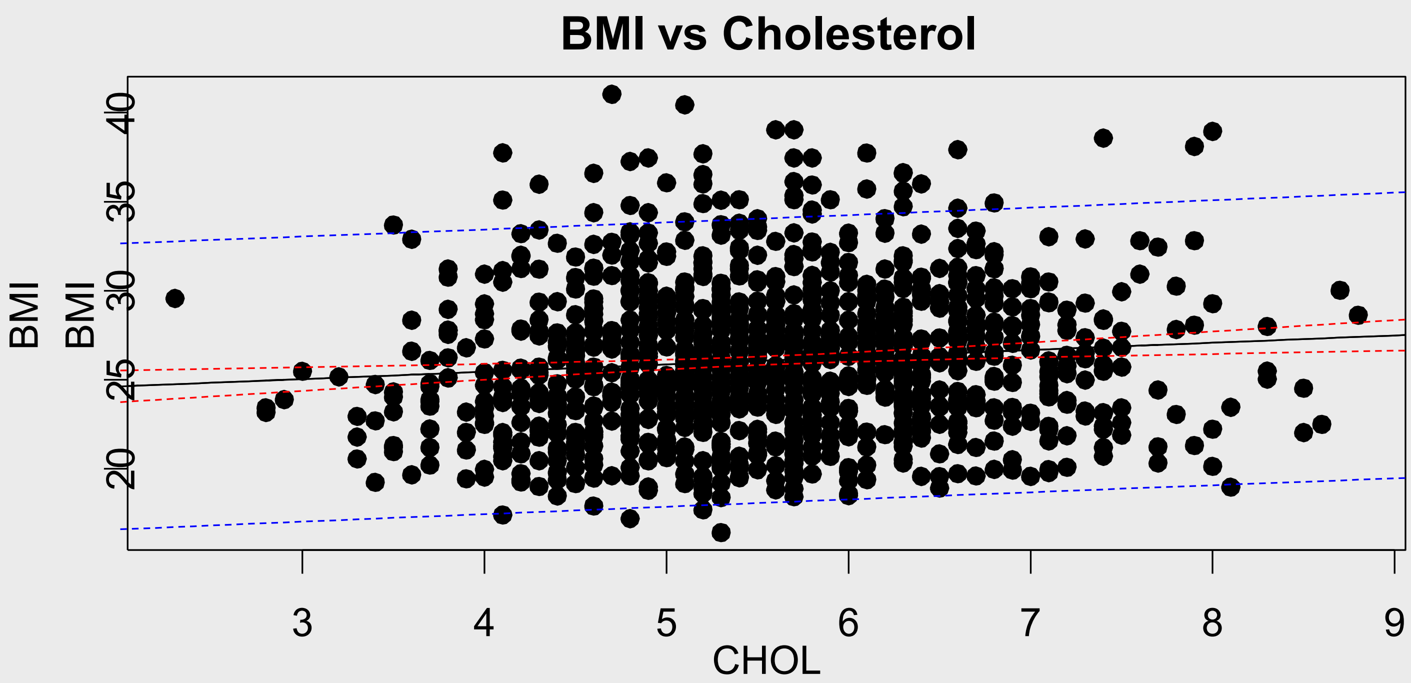
- > cholesterol[1:18,]

	AGE	BMI	CHOL	DBP	HEIGHT	SEX	WAIST	WEIGHT
1	32	24.16	4.7	70	175	male	82	74
2	40	26.28	5.8	70	183	male	93	88
3	39	25.06	5.5	70	182	male	91	83
4	37	28.67	5.6	80	183	male	98	96
5	46	26.30	5.9	80	185	male	95	90
6	44	23.46	5.8	84	180	male	84	76
7	51	24.91	5.4	90	170	male	82	72
8	50	25.06	6.0	80	173	male	92	75
9	49	28.02	4.9	84	187	male	95	98
10	49	28.09	7.2	90	178	male	89	89
11	55	26.51	5.1	100	178	male	105	84
12	52	29.98	4.8	80	178	male	102	95
13	54	29.32	7.1	80	180	male	107	95
14	61	31.26	4.6	80	185	male	109	107
15	59	24.22	6.4	70	170	male	92	70
16	61	23.77	4.7	90	180	male	89	77
17	67	28.73	6.3	80	175	male	98	88
18	67	25.03	8.3	90	181	male	97	82

Simple Linear Regession

Example: Cholesterol Data Set

- The plot



This is a large data set of high variability. The sample size works in favour of the CI (red), but the PI (blue) is left out in the cold.

Simple Linear Regession

Example: Cholesterol Data Set

- The code

```
> C1 = lm(BMI~CHOL,data=cholesterol)
> plot(BMI~CHOL,data=cholesterol,pch=21,bg="black",main="BMI vs Cholesterol" )
> new = data.frame(CHOL=seq(2,10,0.1))
> CIs = predict(C1,new,interval="confidence")
> PIs = predict(C1,new,interval="predict")
> matpoints(new$CHOL,CIs,lty=c(1,2,2),col=c("black","red","red"),type="l")
> matpoints(new$CHOL,PIs,lty=c(1,2,2),col=c("black","blue","blue"),type="l")
```

Simple Linear Regession

Example: Cholesterol Data Set

- The Fitted Model

```
> C1 = lm(BMI~CHOL,data=cholesterol)
> anova(C1) ## Analysis of Variance Table
Analysis of Variance Table

Response: BMI
            Df  Sum Sq Mean Sq F value    Pr(>F)
CHOL         1   184.3 184.333  11.141 0.0008726 ***
Residuals 1107 18316.1  16.546
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Simple Linear Regession

Example: Cholesterol Data Set

- from the output we could identify all the values in the following table

Source of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS)	F-value	Pr(> F)
Regression	1	RegSS	$\frac{\text{RegSS}}{1}$	$\frac{\text{RegSS}/1}{\text{RSS}/(n-2)}$	1-pf (F-value, 1, n-2)
Residual	$n-2$	RSS	$\frac{\text{RSS}}{n-2}$		
Total	$n-1$	TSS	$\frac{\text{TSS}}{n-1}$		

- Here $T = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} \sim \text{T}_{n-2}$ and $F = \frac{\text{RegSS}/1}{\text{RSS}/(n-2)} \sim \text{F}_{1,n-2}$ are related via $F = T^2$

Simple Linear Regession

Example: Cholesterol Data Set

- The Fitted Model

```
> summary(C1)

Call:
lm(formula = BMI ~ CHOL, data = cholesterol)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5632 -2.9024 -0.4118  2.4582 15.3019

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.8078     0.6919  34.409  < 2e-16 ***
CHOL          0.4086     0.1224   3.338 0.000873 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.068 on 1107 degrees of freedom
Multiple R-squared:  0.009964,Adjusted R-squared:  0.009069
F-statistic: 11.14 on 1 and 1107 DF,  p-value: 0.0008726
```

Example: Cholesterol Data Set

- Conclusion

  - ```
    > confint(C1,level=0.95)
                     2.5 %     97.5 %
    (Intercept) 22.4501727 25.1653519
    CHOL         0.1683972  0.6487605
    ```

    As cholesterol (CHOL) increases by 1, mean BMI is estimated to increase by between 0.168 and 0.649
  - Note there is no causal relationship implied or intended to be implied by this statement
    ```

Simple Linear Regession

Example: Cholesterol Data Set

- Conclusion
  - Whether such an increase is of practical significance is not something statistics has an opinion on
  - Clearly in this example, there is a great deal of ”unexplained” variability; 99% in fact
  - There are many other variables involved in determining BMI. Incorporating those variables into the equation to obtain a better explanation is the province of multiple regression
- Disclaimer
  - The interpretations of the examples presented so far are only valid provided the model assumptions are valid
  - Nothing in the numbers presented so far will tell you if this is the case or not