Aims of today’s lecture

- To give you an overview of the **modelling cycle**.
- To begin the detailed discussion of **diagnostic procedures**.
The modelling cycle

- We have seen that the regression model describes rather specialised forms of data
  - Data are “planar”,
  - Scatter is uniform over the plane.
- We have looked at some plots that help us decide if the data is suitable for regression modelling
  - pairs
  - reg3d
  - coplot
Residual analysis

- Another approach is to fit the model and examine the residuals.

- If the model is appropriate the residuals have no pattern.

- A pattern in the residuals usually indicates that the model is not appropriate.

- If this is the case we have two options:
  1. Select another form of model e.g. non-linear regression – see other courses;
  2. Transform the data so that the regression model fits the transformed data – see next slide.
The Modelling Cycle

PLOTS and THEORY

Choose Model

Fit Model

Transform

Examine Residuals

Bad fit

Good fit

USE MODEL
What constitutes a bad fit?

- Non-planar data
- Outliers in the data
- Errors depend on covariates (non-constant scatter)
- Errors not independent
- Errors not normally distributed

- First two points can be serious as they affect the meaning and accuracy of the estimated coefficients.

- The others affect mainly standard errors, not estimated coefficients – see subsequent lectures
Detecting non-planar data

- For the next 2 lectures, we look at diagnosing non-planar data and choosing a transformation. We can diagnose non-planar data (nonlinearity) by fitting the model, and
  - plotting residuals versus fitted values, residuals against explanatory variables;
  - fitting additive models
- In each case, a curved plot indicates non-planar data.
Plotting residuals vs. fitted values

data(cherry.df)
cherry.lm <- lm(volume~diameter+height,data=cherry.df)
plot(cherry.lm,which=1)

```r
which=1: selects the plot of residuals vs. fitted values
```
Plotting residuals vs. fitted values

`lm(volume ~ diameter + height)`

Residuals vs Fitted

Fitted values

Residuals

-5 0 5 10

Fitted values

`lm(volume ~ diameter + height)`
Additive models

- These are models of the form

\[ Y = g_1(x_1) + g_2(x_2) + \cdots + g_k(x_k) + \varepsilon \]

where \(g_1, \ldots, g_k\) are transformations.

- Fitted using the `gam` function in R.

- The transformations are estimated by the software.

- Use the function to suggest good transformations.
Example: Cherry trees

library(mgcv)

cherry.gam <- gam(volume~s(diameter)+s(height),
                  data=cherry.df)

plot(cherry.gam,residuals=T,pages=1)
Example: Cherry trees
Fitting polynomials

- To fit a model $y = \beta_0 + \beta_1 x + \beta_2 x^2$, use
  
  $y \sim \text{poly}(x, 2)$

- To fit a model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$, use
  
  $y \sim \text{poly}(x, 3)$

  etc.
The model fitted by $y \sim \text{poly}(x, 2)$ is of the form

$$Y = \beta_0 + \beta_1 p_1(x) + \beta_2 p_2(x)$$

where

- $p_1$: polynomial of degree 1, i.e. of the form $ax + b$
- $p_2$: polynomial of degree 2, i.e. of the form $ax^2 + bx + c$.

$p_1, p_2$ chosen to be uncorrelated (best possible estimation)
Adding a quadratic term (cherry trees)

Call:
  lm(formula = volume ~ poly(diameter, 2) + height, 
      data = cherry.df)
---
Coefficients:

                     Estimate  Std. Error t value  Pr(>|t|)  
(Intercept)       1.565533   6.722180   0.233  0.817603 
poly(diameter, 2)1 80.252233   3.073459  26.111  < 2e-16 *** 
poly(diameter, 2)2 15.399233   2.631567   5.852  3.13e-06 *** 
height             0.376387   0.088229   4.266  0.000218 *** 
---
Residual standard error: 2.625 on 27 degrees of freedom 
Multiple R-squared:  0.9771, Adjusted R-squared:  0.9745 
F-statistic: 383.2 on 3 and 27 DF,  p-value:  < 2.2e-16
Quadratic equation

Call:
`lm(formula = volume ~ diameter + I(diameter^2) + height, 
data = cherry.df)`

---

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -9.92041 | 10.07911   | -0.984  | 0.333729 |
| diameter       | -2.88508 | 1.30985    | -2.203  | 0.036343 *|
| I(diameter^2)  | 0.26862  | 0.04590    | 5.852   | 3.13e-06 ***|
| height         | 0.37639  | 0.08823    | 4.266   | 0.000218 ***|

---

Residual standard error: 2.625 on 27 degrees of freedom
Multiple R-squared: 0.9771, Adjusted R-squared: 0.9745
F-statistic: 383.2 on 3 and 27 DF, p-value: < 2.2e-16
Quadratic equation

\[ \text{volume} \approx -9.92 - 2.89 \times \text{diameter} + 0.27 \times \text{diameter}^2 + 0.38 \times \text{height} \]
Splines

- An alternative to polynomials are splines – these are piecewise cubics, which join smoothly at “knots”.

- Give a more flexible fit to the data.

- Values at one point not affected by values at distant points, unlike polynomials.
Example with 4 knots
Adding a spline term

Call:
lm(formula = volume ~ bs(diameter, df = 6) + height,
    data = cherry.df)

Coefficients:

                          Estimate Std. Error t value  Pr(>|t|) 
(Intercept)            -16.3679   7.4856  -2.187  0.03921 *  
bs(diameter, df = 6)1   0.1941    7.9374   0.024  0.98070
bs(diameter, df = 6)2   5.5744    3.1704   1.758  0.09201 .   
bs(diameter, df = 6)3  10.7976    3.9798   2.713  0.01240 *  
bs(diameter, df = 6)4  31.4053    5.5545   5.654  9.35e-06 *** 
bs(diameter, df = 6)5  42.2665    6.1297   6.895  4.97e-07 *** 
bs(diameter, df = 6)6  58.6454    4.2781  13.708 1.49e-12 ***
height                0.3970    0.1050   3.780  0.00097 ***

---

Residual standard error: 2.8 on 23 degrees of freedom
Multiple R-squared:  0.9778, Adjusted R-squared:  0.971
F-statistic: 144.4 on 7 and 23 DF,  p-value: < 2.2e-16
Difference in fits

![Graph showing volume vs diameter with two fits: splines and polynomial.](graph.png)
Example: Tyre abrasion data

- Data collected in an experiment to study the abrasion resistance of tyres

- Variables are

  - **Hardness**: Hardness of rubber
  - **Tensile**: Tensile strength of rubber
  - **Abrasion Loss**: Amount of rubber worn away in a standard test (response)
Call:
\texttt{lm(formula = abloss \sim hardness + tensile, data = rubber.df)}

---

Coefficients:

|              | Estimate | Std. Error | t value | Pr(>|t|)   |
|--------------|----------|------------|---------|-----------|
| (Intercept)  | 885.161  | 61.751     | 14.334  | 3.84e-14  |
| hardness     | -6.571   | 0.583      | -11.267 | 1.03e-11  |
| tensile      | -1.374   | 0.194      | -7.073  | 1.32e-07  |

---

Residual standard error: 36.49 on 27 degrees of freedom
Multiple R-squared: 0.8402, Adjusted R-squared: 0.8284
F-statistic: 71 on 2 and 27 DF, p-value: 1.767e-11
Tyre abrasion data

- We will use this example to illustrate all the methods we have discussed so far to check if the data are planar, scattered about a flat regression plane i.e.
  - Pairs plot
  - Spinning plot
  - Coplot
  - Residual vs. fitted value plot
  - Fitting GAMs
Spinning

hardness
abloss
tensile
Given: hardness

abloss
Residuals vs. fitted values

lm(abloss ~ hardness + tensile)

Residuals vs Fitted

Fitted values

lm(abloss ~ hardness + tensile)
GAMs
Conclusions

**Pairs plot:** Not very informative

**Spinner:** Hint of a “kink”

**Coplot:** Suggestion of non-linearity, a “kink”

**Residuals vs. fitted values:** weak suggestion that regression is not planar

**GAM plots:** hardness is okay, but strong suggestion that tensile needs transforming

**Suggested transformation:** Looks like a cubic or a $4^{th}$ degree polynomial, so try these (or a spline). With a spline, the $R^2$ increases from 84% to 94%.