Explaining variables with other variables:
Regression and general linear models

Recapitulating

Regression is a method that predicts the value of a numerical from the value of another

Regression and correlation both describe features of a scatter plot, and measure the relationship between two numerical variables.
Correlation treats both variables equally, whereas regression predicts the value of one variable based on the other variable.
Correlation measures the strength of association between the two variables, whereas regression measure how steeply the response variable changes, on average, with the change in the explanatory variable.
**Linear regression**

The equation of the regression line:

\[ Y = a + bX \]

The formula has two coefficients:

- The intercept \( a \) is the value of \( Y \) and \( X \) is zero
- The slope \( b \) is the rate of change in \( Y \) per unit of \( X \)

**Slope**

\[ b = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \]

(Compare with expression for correlation)

** Intercept**

\[ a = \bar{Y} - b\bar{X} \]

**Assumptions**

- At each value of \( X \) there is a population of possible \( Y \)-values whose mean lies on the true regression line.
- At each value of \( X \), the distribution of possible \( Y \)-value is normal.
  
  The variance of \( Y \)-values is the same at all values of \( X \).
  
  At each value of \( X \), the \( Y \)-measurements represent a random sample from the possible \( Y \)-values.
Predicted value

\[ \hat{Y} = a + bX \]

The predicted value of a regression estimates the mean value of Y for all individual/measurements that have given X.

Residuals

\[ Y_i - \hat{Y}_i \]

\[ MS_{\text{residual}} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2} \]

In the sample we have:

\[ Y = a + bX \]

In the population we have the true values:

\[ y = \alpha + \beta x \]

How can we infer the true values of the parameters \( \alpha \) and \( \beta \) from the measured ones \( a \) and \( b \)?

Can we predict the true value of \( y \) as a function of \( x \)?

Interval of confidence for \( \beta \)

When the assumptions of the linear regression are met the sampling distribution then the sampling distribution of \( b \) is a normal distribution with mean \( \beta \) and standard error estimated from the sample by:

\[ SE_b = \frac{MS_{\text{residual}}}{\sqrt{\sum (X_i - \bar{X})^2}} \]

With this we can provide confidence intervals for the true value of \( \beta \)

\[ b - t_{\alpha/2, \nu} SE_b \leq \beta \leq b + t_{\alpha/2, \nu} SE_b \]
Testing hypothesis for slope $\beta$

$$t = \frac{b - b_0}{SE_t}$$

t-distributed with df=n-2

Since we are comparing mean values we could also use ANOVA

ANOVA method for testing zero slope

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of squares</th>
<th>df</th>
<th>Mean Squares</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$SS_{\text{regression}} = \sum(y - \hat{y})^2$</td>
<td>$n - 2$</td>
<td>$\frac{SS_{\text{regression}}}{df_{\text{regression}}}$</td>
<td>$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}}$</td>
</tr>
<tr>
<td>Residual</td>
<td>$SS_{\text{residual}} = \sum(e^2)$</td>
<td>$n - 2$</td>
<td>$\frac{SS_{\text{residual}}}{df_{\text{residual}}}$</td>
<td>$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}}$</td>
</tr>
<tr>
<td>Total</td>
<td>$SS_{\text{total}} = \sum(y^2)$</td>
<td>$n - 1$</td>
<td>$SS_{\text{total}}$</td>
<td>$F = \frac{SS_{\text{regression}}}{SS_{\text{total}}}$</td>
</tr>
</tbody>
</table>

Variance Explained

$$R^2 = \frac{SS_{\text{regression}}}{SS_{\text{total}}}$$

A measure of the quality of the fitting
Confidence bands for the predicted mean measure the precision of the mean Y value for each value of X.

Prediction interval of any one measure of Y measure the precision of the estimate of any individual which as a value of X.

The second is broader denoting the greater uncertainty.

Outliers
Extrapolation
Regression to the mean
Detecting violation of assumptions

Regression to the mean

Results when two variables have a correlation less than 1.

Individuals that are far from the mean for one of the measurements will, on average, lie closer to the mean for the other measurement.

Practical example: before and after drug treatment.
Assumptions

At each value of X there is a population of possible Y-values whose mean lies on the true regression line.

At each value of X, the distribution of possible Y-value is normal.

The variance of Y-values is the same at all values of X.

At each value of X, the Y-measurements represent a random sample from the possible Y-values.