Regression

- Predicts $Y$ from $X$
- Linear regression assumes that the relationship between $X$ and $Y$ can be described by a line

Regression assumes...

- Random sample
- $Y$ is normally distributed with equal variance for all values of $X$

Correlation vs. regression

The parameters of linear regression

$$Y = \alpha + \beta X$$
Estimating a regression line

\[ Y = a + b X \]

Finding the "least squares" regression line

Minimize:

\[ SS_{\text{residual}} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \]

**Nomenclature**

Residual:

\[ Y_i - \hat{Y}_i \]
Best estimate of the slope

\[ b = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \]

(= "Sum of cross products" over "Sum of squares of X")

Remember the shortcuts:

\[ \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \left( \sum X_i Y_i \right) - \frac{\sum X_i \sum Y_i}{n} \]

\[ \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum (X_i)^2 - \frac{\left( \sum X_i \right)^2}{n} \]

Finding a

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

So...

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

Example: Predicting age based on radioactivity in teeth

Many above ground nuclear bomb tests in the ’50s and ’60s may have left a radioactive signal in developing teeth.

Is it possible to predict a person’s age based on dental C\(^{14}\)?

Teeth data:

<table>
<thead>
<tr>
<th>$\Delta^{14}C$</th>
<th>Date of Birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>89</td>
<td>1985.5</td>
</tr>
<tr>
<td>109</td>
<td>1983.5</td>
</tr>
<tr>
<td>91</td>
<td>1990.5</td>
</tr>
<tr>
<td>127</td>
<td>1987.5</td>
</tr>
<tr>
<td>99</td>
<td>1990.5</td>
</tr>
<tr>
<td>110</td>
<td>1984.5</td>
</tr>
<tr>
<td>123</td>
<td>1983.5</td>
</tr>
<tr>
<td>105</td>
<td>1989.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\Delta^{14}C$</th>
<th>Date of Birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>622</td>
<td>1963.5</td>
</tr>
<tr>
<td>262</td>
<td>1971.7</td>
</tr>
<tr>
<td>471</td>
<td>1963.7</td>
</tr>
<tr>
<td>112</td>
<td>1990.5</td>
</tr>
<tr>
<td>285</td>
<td>1975</td>
</tr>
<tr>
<td>439</td>
<td>1970.2</td>
</tr>
<tr>
<td>363</td>
<td>1972.6</td>
</tr>
<tr>
<td>391</td>
<td>1971.8</td>
</tr>
</tbody>
</table>

Let $X$ be the estimated age, and $Y$ be the actual age.

\[
\sum X = 3798, \quad \sum Y = 31674
\]

\[
\sum X^2 = 1340776, \quad \sum (XY) = 7495223
\]

\[
\sum Y^2 = 62704042
\]

\[n = 16\]

\[
\bar{X} = 237.375 \quad \bar{Y} = 1979.63
\]

Calculating $a$

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

\[
a = 1979.63 - (-0.053)237.375 = 1992.2
\]

\[
b = \frac{-23393}{439226} = -0.053
\]
\[ \hat{Y} = 1992.2 - 0.053X \]

Predicting \( Y \) from \( X \)

If a cadaver has a tooth with \( \Delta^{14}C \) content equal to 200, what does the regression line predict its year of birth to be?

\[ \hat{Y} = 1992.2 - 0.053X \]
\[ = 1992.2 - 0.053(200) \]
\[ = 1981.6 \]

\( r^2 \) predicts the amount of variance in \( Y \) explained by the regression line

\( r^2 \) is the “coefficient of determination: it is the square of the correlation coefficient \( r \)
Caution: It is unwise to extrapolate beyond the range of the data.

If we were to extrapolate to ask how many species might be in a pool of 50000m$^2$, we would guess about 20.

More data on fish in desert pools

Log transformed data:

Testing hypotheses about regression

$H_0: \beta = 0$

$H_A: \beta \neq 0$
Sums of squares for regression

\[ SS_{Total} = \sum Y_i^2 - \frac{\left( \sum Y_i \right)^2}{n} \]

\[ SS_{regression} = b \sum (X_i - \bar{X})(Y_i - \bar{Y}) \]

\[ SS_{residual} + SS_{regression} = SS_{Total} \]

With \( n - 2 \) degrees of freedom for the residual

Radioactive teeth: Sums of squares

\[ SS_{Total} = \sum Y_i^2 - \frac{\left( \sum Y_i \right)^2}{n} \]

\[ = 62704042 - \frac{(31674)^2}{16} = 1339.75 \]

\[ SS_{regression} = b \sum (X_i - \bar{X})(Y_i - \bar{Y}) \]

\[ = (-0.053)(-23393) = 1239.8 \]

Teeth: Sums of squares

\[ SS_{residual} = SS_{Total} - SS_{regression} \]

\[ = 1339.75 - 1239.8 \]

\[ = 99.9 \]

\[ df_{residual} = 16 - 2 = 14 \]

Calculating residual mean squares

\[ MS_{residual} = \frac{SS_{residual}}{df_{residual}} \]

\[ MS_{residual} = \frac{99.9}{14} = 7.1 \]
Standard error of a slope

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

$$= \sqrt{\frac{7.1}{439226}} = 0.004$$

**Example:** 95% confidence interval for slope with teeth example

$$b \pm t_{\alpha[2], df} SE_b = b \pm t_{0.05[2], df} SE_b$$

$$= -0.053 \pm 2.14 \times 0.004$$

$$= -0.053 \pm 0.0018$$

$b$ has a $t$ distribution

Confidence interval for a slope: $$b \pm t_{\alpha[2], df} SE_b$$

Hypothesis tests can use $t$: $$t = \frac{b - \beta_0}{SE_b}$$

Confidence bands: confidence intervals for predictions of mean $Y$
Prediction intervals: confidence intervals for predictions of individual $Y$

Hypothesis tests on slopes

$H_0$: $\beta = 0$

$H_A$: $\beta \neq 0$

$$t = \frac{b - \beta_0}{SE_b}$$

$$t = \frac{-0.053 - 0}{0.004} = 13.25$$

$t_{0.0001(2),14} = \pm 5.36$

So we can reject $H_0$, $P<0.0001$

Non-linear relationships

Transformations

Quadratic regression

Splines

Transformations

If $Y = aX^b$ then $\ln Y = \ln a + b \ln X$.

If $Y = ab^X$ then $\ln Y = \ln a + X \ln b$.

If $Y = a + \frac{b}{X}$ then set $X' = \frac{1}{X}$, and calculate $Y = a + bX'$.

All of the equations on the right have the form $Y=a+bX$. 
Non-linear relationship: Number of fish species vs. Size of desert pool

Residual plots help assess assumptions

Original:

Residual plot

Transformed data

Logs:

Residual plot

Polynomial regression

Number of species = 0.046 + 0.185 Biomass - 0.00044 Biomass²
Do not fit a polynomial with too many terms (the sample size should be at least 7 times the number of terms)

Comparing two slopes

Example: Comparing species-area curves for islands to those of mainland populations

Log10(Number of species) = 0.24537 + 0.27554 Log10(Area of "island")

Summary of Fit
RSquare = 0.677098
RSquare Adj = 0.663059
Root Mean Square Error = 0.215751
Mean of Response Observations (or Sum Wgts) = 1.126

Analysis of Variance
Source | DF | Sum of Squares | Mean Square | F Ratio | Prob>F
--- | --- | --- | --- | --- | ---
Model | 1 | 2.2449876 | 2.24499 | 48.2291 | <.0001
Error | 23 | 1.0706124 | 0.04655 | Prob>F
Total | 24 | 3.3156000 |

Parameter Estimates
Term | Estimate | Std Error | t Ratio | Prob>|t|
--- | --- | --- | --- | ---
Intercept | 0.245375 | 0.133946 | 1.83 | 0.0799
Log10(Area of "island") | 0.2755397 | 0.039676 | 6.94 | <.0001

Linear Fit Type of island=M

Log10(Number of species) = 1.59531 + 0.09744 Log10(Area of "island")

Summary of Fit
RSquare = 0.960728
RSquare Adj = 0.95091
Root Mean Square Error = 0.049419
Mean of Response Observations (or Sum Wgts) = 1.925

Analysis of Variance
Source | DF | Sum of Squares | Mean Square | F Ratio | Prob>F
--- | --- | --- | --- | --- | ---
Model | 1 | 0.23898111 | 0.238981 | 97.8540 | 0.0006
Error | 4 | 0.00976889 | 0.002442 |
Total | 5 | 0.24875000 |

Parameter Estimates
Term | Estimate | Std Error | t Ratio | Prob>|t|
--- | --- | --- | --- | ---
Intercept | 1.5953149 | 0.038959 | 40.95 | <.0001
Log10(Area of "island") | 0.0974439 | 0.009851 | 9.89 | 0.0006

Hypotheses

\[ H_0: \beta_M = \beta_I. \]

\[ H_A: \beta_M \neq \beta_I. \]

The error in the difference of two slopes is normally distributed.

\[ t = \frac{(b_1 - b_2) - (\beta_1 - \beta_2)}{SE_{b_1-b_2}} \]

\[ df = n_1 - 2 + n_2 - 2 \]
\[
\left( MS_{error}\right)_p = \frac{\left( SS_{error}\right)_1 + \left( SS_{error}\right)_2}{\left( DF_{error}\right)_1 + \left( DF_{error}\right)_2}
\]

\[
SE_{b1-b2} = \sqrt{\left( \frac{\left( MS_{error}\right)_p}{\sum(X - \bar{X})^2} \right)_1 + \left( \frac{\left( MS_{error}\right)_p}{\sum(X - \bar{X})^2} \right)_2}
\]

**Analysis of covariance (ANCOVA)**

Compares many slopes

\[H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 \ldots\]

\[H_A: \text{At least one of the slopes is different from another.}\]

**Logistic regression**

Tests for relationship between a numerical variable (as the explanatory variable) and a binary variable (as the response).

e.g.: Does the dose of a toxin affect probability of survival?

Does the length of a peacock's tail affect its probability of getting a mate?