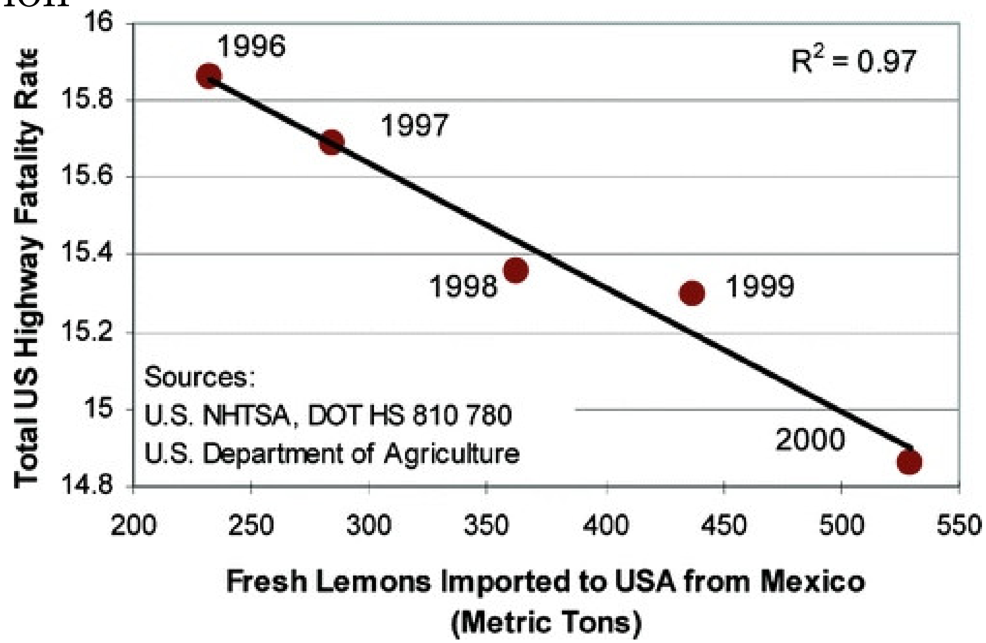
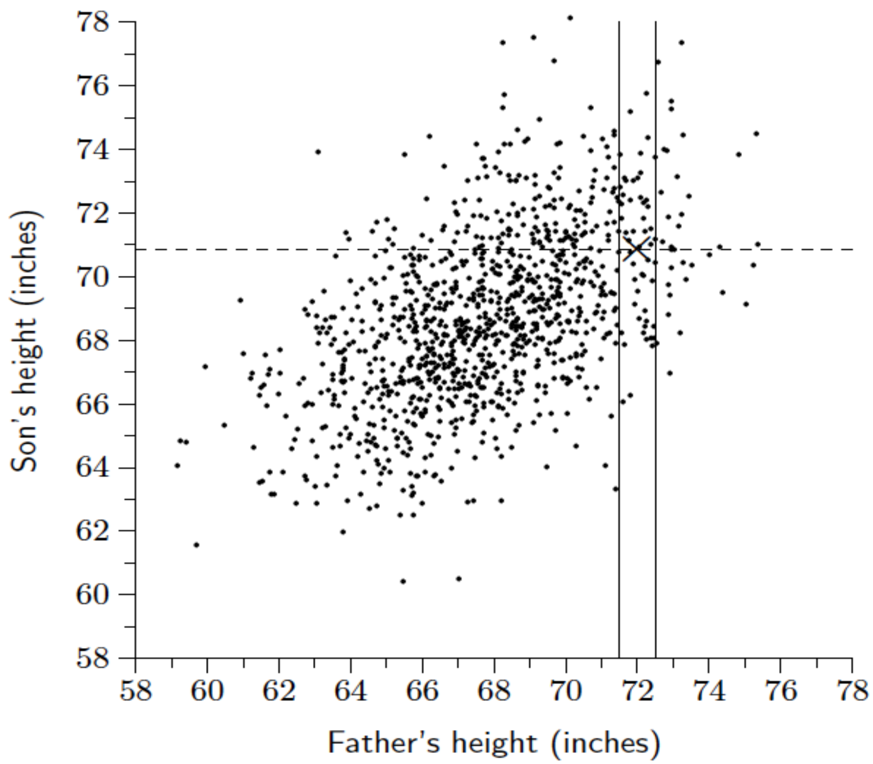
Slides 16: Simple regression model

- Regression to mediocrity
- Least squares estimates
- Coefficient of determination
- Residuals
- Confidence intervals
- Model utility test
- Prediction interval



Correlation does not imply causation

Pearson's father-son data: 1,078 pairs of heights (England, 1900).



Focussing on 6 feet tall fathers, we see that their sons on average are shorter than their fathers. F. Galton called this *regression to mediocrity*.

Question. Which of the previous statistical tools could be applied?

A simple linear regression model is based on the linear relation

$$Y(x) = \beta_0 + \beta_1 x + \epsilon, \qquad \epsilon \sim \mathrm{N}(0, \sigma),$$

where ϵ is the noisy part of the response, that is not explained by the value x of the main explanatory variable. The assumption of *homoscedasticity* requires that σ is independent of the x-value.

For a given collection of x-values $(x_1, \ldots, x_n)$, and a vector $(e_1, \ldots, e_n)$ of independent realisations of ϵ, we get a sample of response values

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n.$$

The likelihood is a function of the 3D parameter $\theta = (\beta_0, \beta_1, \sigma^2)$

$$L(\theta) = \prod_{i=1}^{n} \tfrac{1}{\sqrt{2\pi}\sigma} \exp\Big\{ - \tfrac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \Big\} = C\sigma^{-n} e^{-\frac{S(\beta_0, \beta_1)}{2\sigma^2}},$$

where

$$C = (2\pi)^{-n/2}, \qquad S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Log-likelihood function

$$l(\theta) = \ln L(\theta) = \ln C - n \ln \sigma - \tfrac{S(\beta_0,\beta_1)}{2\sigma^2}$$

Maximisation of $l(\theta)$ over (β_0, β_1) is equivalent to minimisation of the sum of squares $S(\beta_0, \beta_1)$.
Therefore, the MLEs of (β_0, β_1) are called the least squares estimates.

Observe that

$$\begin{aligned}\tfrac{S(\beta_0,\beta_1)}{n} &= \tfrac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \\ &= \beta_0^2 + 2\beta_0\beta_1\bar{x} - 2\beta_0\bar{y} - 2\beta_1\overline{xy} + \beta_1^2\overline{x^2} + \overline{y^2}\end{aligned}$$

with the following set of five sufficient statistics:

$$\bar{x} = \tfrac{x_1+\ldots+x_n}{n}, \quad \bar{y} = \tfrac{y_1+\ldots+y_n}{n}$$

$$\overline{x^2} = \tfrac{x_1^2+\ldots+x_n^2}{n}, \quad \overline{y^2} = \tfrac{y_1^2+\ldots+y_n^2}{n}, \quad \overline{xy} = \tfrac{x_1y_1+\ldots+x_ny_n}{n}$$

To obtain MLEs of $\theta = (\beta_0, \beta_1, \sigma^2)$ compute the derivatives

$$\begin{aligned}
\tfrac{\partial l}{\partial \beta_0} &= -\tfrac{1}{2\sigma^2}\tfrac{\partial S}{\partial \beta_0}, \\
\tfrac{\partial l}{\partial \beta_1} &= -\tfrac{1}{2\sigma^2}\tfrac{\partial S}{\partial \beta_1}, \\
\tfrac{\partial l}{\partial \sigma^2} &= -\tfrac{n}{2\sigma^2} + \tfrac{S(\beta_0,\beta_1)}{2\sigma^4},
\end{aligned}$$

and set them equal to zeros.

Putting $\frac{\partial S}{\partial \beta_0} = 0$ and $\frac{\partial S}{\partial \beta_1} = 0$, we get the so-called normal equations:

$$b_0 + b_1\bar{x} = \bar{y}, \qquad b_0\bar{x} + b_1\overline{x^2} = \overline{xy}.$$

Solving this system of linear equations we get

$$b_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{r s_y}{s_x}, \quad b_0 = \bar{y} - b_1\bar{x},$$

where r is the sample correlation coefficient and

$$s_x^2 = \tfrac{1}{n-1}\sum(x_i - \bar{x})^2, \qquad s_y^2 = \tfrac{1}{n-1}\sum(y_i - \bar{y})^2.$$

The sample correlation coefficient r is an unbiased estimate of ρ

$$r = \frac{s_{xy}}{s_x s_y}, \quad \rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y},$$

where the sample covariance s_{xy} is an unbiased estimate of $\mathrm{Cov}(X,Y)$

$$s_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}), \quad \mathrm{Cov}(X,Y) = \mathrm{E}(X - \mu_x)(Y - \mu_y),$$

provided that $(x_1, \ldots, x_n)$ is a random sample from X-distribution.

As a result, the fitted regression line $y = b_0 + b_1 x$ takes the form

$$y = \bar{y} + r \frac{s_y}{s_x} (x - \bar{x}),$$

Notice that

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

where $\hat{y}_i$ are the predicted responses:

$$\hat{y}_i = b_0 + b_1 x_i, \quad i = 1, \ldots, n.$$

Putting $\frac{\partial l}{\partial \sigma^2} = 0$, we get

$$0 = -\frac{n}{2\sigma^2} + \frac{S(\beta_0,\beta_1)}{2\sigma^4}$$

and replacing (β_0, β_1) with (b_0, b_1), we find the MLE of σ^2 to be

$$\hat{\sigma}^2 = \frac{S(b_0,b_1)}{n},$$

The maximum likelihood estimate of $\hat{\sigma}^2$ is only asymptotically unbiased estimate of σ^2. An unbiased estimate of σ^2 is given by

$$s^2 = \frac{S(b_0,b_1)}{n-2}.$$

The error sum of squares

$$SS_{\mathrm{E}} = S(b_0, b_1) = \sum (y_i - \hat{y}_i)^2 = (n-1)s_y^2(1-r^2)$$

divided by $n-2$ gives a very useful expression

$$s^2 = \frac{n-1}{n-2}\, s_y^2(1-r^2).$$

Question. If $r = 0.5$, what proportion of s_y^2 is explained by s^2?

Residuals are differences between observed and predicted responses

$$\hat{e}_i = y_i - \hat{y}_i = (y_i - \bar{y}) - r\tfrac{s_y}{s_x}(x_i - \bar{x})$$

The residuals $(\hat{e}_1, \ldots, \hat{e}_n)$ are linearly connected via

$$\hat{e}_1 + \ldots + \hat{e}_n = 0, \quad x_1\hat{e}_1 + \ldots + x_n\hat{e}_n = 0, \quad \hat{y}_1\hat{e}_1 + \ldots + \hat{y}_n\hat{e}_n = 0,$$

so we can say that $\hat{e}_i$ are uncorrelated with x_i and $\hat{e}_i$ are uncorrelated with $\hat{y}_i$. The residuals $\hat{e}_i$ are realisations of random variables $\hat{E}_i$ having normal distributions with zero means and

$$\mathrm{Var}(\hat{E}_i) = \sigma^2\Big(1 - \tfrac{\sum_k (x_k - x_i)^2}{n(n-1)s_x^2}\Big), \quad \mathrm{Cov}(\hat{E}_i, \hat{E}_j) = -\sigma^2 \cdot \tfrac{\sum_k (x_k - x_i)(x_k - x_j)}{n(n-1)s_x^2}.$$

Test normality using normal QQ-plot for the standardised residuals

$$\tilde{e}_i = \frac{\hat{e}_i}{s_i}, \quad s_i = s\sqrt{1 - \tfrac{\sum_k (x_k - x_i)^2}{n(n-1)s_x^2}}, \quad i = 1, \ldots, n,$$

where s_i are the estimated standard deviations of $\hat{E}_i$. In some cases, the non-linearity problem can fixed by a log-log transformation of the data.

Using $y_i - \bar{y} = \hat{y}_i - \bar{y} + \hat{e}_i$, we obtain a decomposition

$$SS_{\mathrm{T}} = SS_{\mathrm{R}} + SS_{\mathrm{E}},$$

where

$$SS_{\mathrm{T}} = \sum_i (y_i - \bar{y})^2 = (n-1)s_y^2$$

is the total sum of squares, and

$$SS_{\mathrm{R}} = \sum_i (\hat{y}_i - \bar{y})^2 = (n-1)b_1^2 s_x^2 = (n-1)r^2 s_y^2$$

is the regression sum of squares. Combining these relations, we find that

$$r^2 = \frac{SS_{\mathrm{R}}}{SS_{\mathrm{T}}} = 1 - \frac{SS_{\mathrm{E}}}{SS_{\mathrm{T}}}.$$

Coefficient of determination r^2 is the proportion of variation in the response variable explained by the variation of the predictor.

Observe that r^2 has a more intuitive meaning than the sample correlation coefficient r.

The least squares estimators (b_0, b_1) are unbiased and consistent. Due to the normality assumption we have the following exact distributions

$$B_0 \sim \mathrm{N}(\beta_0, \sigma_0), \quad \sigma_0^2 = \tfrac{\sigma^2 \sum x_i^2}{n(n-1)s_x^2}, \quad s_{b_0}^2 = \tfrac{s^2 \sum x_i^2}{n(n-1)s_x^2}, \quad \tfrac{B_0 - \beta_0}{S_{B_0}} \sim t_{n-2},$$

$$B_1 \sim \mathrm{N}(\beta_1, \sigma_1), \quad \sigma_1^2 = \tfrac{\sigma^2}{(n-1)s_x^2}, \quad s_{b_1}^2 = \tfrac{s^2}{(n-1)s_x^2}, \quad \tfrac{B_1 - \beta_1}{S_{B_1}} \sim t_{n-2}.$$

There is a weak correlation between the two estimators:

$$\mathrm{Cov}(B_0, B_1) = -\tfrac{\sigma^2 \bar{x}}{(n-1)s_x^2}$$

which is negative, if $\bar{x} > 0$, and positive, if $\bar{x} < 0$.

$$\boxed{\text{Exact } 100(1-\alpha)\% \text{ confidence intervals } I_{\beta_i} = b_i \pm t_{n-2}(\tfrac{\alpha}{2}) \cdot s_{b_i}}$$

For $i = 0$ or $i = 1$ and a given value β^*, one would like to the the null hypothesis H_0: $\beta_i = \beta^*$. Use the test statistic

$$t = \frac{b_i - \beta^*}{s_{b_i}},$$

hving the exact null distribution $T \sim t_{n-2}$.

Two important examples of hypothesis testing for the linear regression.

1. Model utility test is built around the null hypothesis

$$H_0 : \beta_1 = 0$$

stating that there is no relationship between the predictor variable x and the response y. The corresponding test statistic, called t-value,

$$t = \frac{b_1}{s_{b_1}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

has an exact t_{n-2} null distribution.

2. Zero-intercept test aims at

$$H_0 : \beta_0 = 0.$$

Compute its t-value

$$t = b_0/s_{b_0},$$

and find whether this value is significant, again using t-distribution with df $= n - 2$.

Given the earlier sample of size n consider a new value x of the predictor variable. We wish to say something on the unobserved response value

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

Its expected value

$$\mu = \beta_0 + \beta_1 x$$

is estimated by

$$\hat{\mu} = b_0 + b_1 x.$$

The standard error of $\hat{\mu}$ is computed as the square root of

$$\mathrm{Var}(\hat{\mu}) = \tfrac{\sigma^2}{n} + \tfrac{\sigma^2}{n-1} \cdot (\tfrac{x-\bar{x}}{s_x})^2.$$

An exact $100(1-\alpha)\%$ confidence interval

$$I_\mu = b_0 + b_1 x \pm t_{n-2}(\tfrac{\alpha}{2}) \cdot s\sqrt{\tfrac{1}{n} + \tfrac{1}{n-1}(\tfrac{x-\bar{x}}{s_x})^2}$$

Question. In what sense $\hat{\mu}$ is a random variable? Is it independent of ϵ that defines the random variable Y?

This I_μ should be compared to the prediction interval for Y

$$I = b_0 + b_1 x \pm t_{n-2}(\tfrac{\alpha}{2}) \cdot s\sqrt{1 + \tfrac{1}{n} + \tfrac{1}{n-1}(\tfrac{x-\bar{x}}{s_x})^2}$$

obtained from

$$\mathrm{Var}(Y - \hat{\mu}) = \mathrm{Var}(\mu + \epsilon - \hat{\mu}) = \sigma^2 + \mathrm{Var}(\hat{\mu}) = \sigma^2(1 + \tfrac{1}{n} + \tfrac{1}{n-1} \cdot (\tfrac{x-\bar{x}}{s_x})^2).$$

Prediction interval I has wider limits than I_μ, since it contains uncertainty due to the noise component ϵ

The further x lies from $\bar{x}$, the wider are the intervals.