STAT115: Introduction to Biostatistics

University of Otago
Ōtākou Whakaihu Waka

Lecture 22: Inference with the Linear Regression Model

Outline

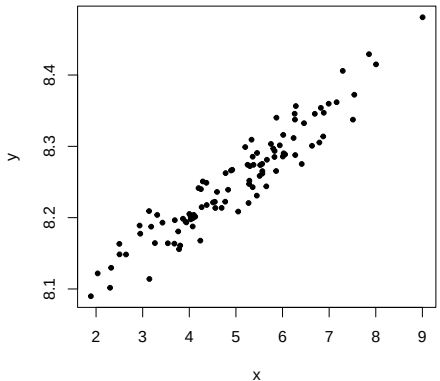
- Previous:
  - Fitting statistical models
  - Checking model assumptions
- Today:
  - Standard error
  - Confidence interval
  - Hypothesis test

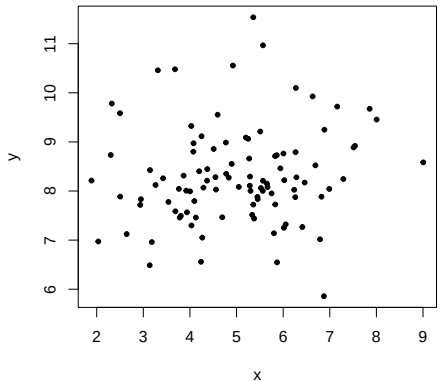
What does a regression model tell us?

- Consider the height of fathers and sons data
- The fitted model is an estimate of the true regression line in population
  - Population may be all male NZ university students (and their fathers)
- We need to assess the precision of the estimated parameters
  - Standard errors of the regression parameters
- Use standard errors to find confidence intervals and conduct hypothesis tests

The importance of the error variance

- Both sets of data come from populations with identical trend: $\mu_y = 8 + 0.05x$.





The importance of error variance

- The linear regresion model is $y = \beta_0 + \beta_1 x + \varepsilon$
  - The error ε is assumed to be normal with mean 0 and variance σ_ε^2
- The larger the error variance (all else equal)
  - The larger the spread of points around the true regression line
  - The more uncertain we are about the fitted regression line
    - That is, the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are less precise
  - To quantify our uncertainty about a fitted model
    - We need to estimate the error variance σ_ε^2

Estimation of the error variance

- The residuals ($\hat{\varepsilon}$) are estimates of the true errors (ε)
- Good estimate of error variance σ_ε^2: sample variance of the residuals $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n$
- We need a few minor technical modifications
- The sample variance of the residuals is $\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2$.
  - The sample mean of the residuals is 0: $\bar{\hat{\varepsilon}} = 0$
  - The correct divisor for simple linear regression is $n-2$ (rather than $n-1$)
- So estimate of error variance is

$$s_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{RSS}{n-2}$$

  - $RSS = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$ is called the residual sum of squares

In R: father/son height data

Input code

- We can get s_ε from the R output (called `Residual standard error`)

```
m_height = lm(son ~ father, data = height)
summary(m_height)
```

In R: father/son height data

Output

```
##
## Call:
## lm(formula = son ~ father, data = height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21.89  -3.89  -0.41   4.59  15.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114.0533     8.4979   13.42  < 2e-16 ***
## father        0.3699     0.0478    7.74  1.9e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.13 on 277 degrees of freedom
## Multiple R-squared:  0.178,        Adjusted R-squared:  0.175
## F-statistic: 59.9 on 1 and 277 DF,  p-value: 1.9e-13
##
```

Standard error of $\hat{\beta}_1$

- In many studies β_1 is the parameter we are most interested in
  - Change in the expected value of y for changing x in the population
- We estimate $\hat{\beta}_1$ from the observed data (sample)
- Measure precision of estimate by standard error $\sigma_{\hat{\beta}_1}$
  - Standard deviation of the sampling distribution of $\hat{\beta}_1$
    - Variation in $\hat{\beta}_1$ if there were many data sets (of the same size) from the population

Standard error of $\hat{\beta}_1$

- The standard error for $\hat{\beta}_1$ is

$$\sigma_{\hat{\beta}_1} = \frac{\sigma_\varepsilon}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

  - The standard error is proportional to the error standard deviation σ_ε
    - As σ_ε^2 increases, the standard error of $\hat{\beta}_1$ also increases
- In principle this tells us about the precision of our estimated slope, $\hat{\beta}_1$
- In practice the formula is useless, since we don't know σ_ε
- We can handle that by estimating σ_ε by s_ε
- In practice, we will then use (estimated) standard error

$$s_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

In R

```
##
## Call:
## lm(formula = son ~ father, data = height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21.89  -3.89  -0.41   4.59  15.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114.0533     8.4979   13.42  < 2e-16 ***
## father        0.3699     0.0478    7.74  1.9e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.13 on 277 degrees of freedom
## Multiple R-squared:  0.178,        Adjusted R-squared:  0.175
## F-statistic: 59.9 on 1 and 277 DF,  p-value: 1.9e-13
##
```

Confidence intervals and hypothesis tests

- The standard error is needed to find confidence intervals and test statistics
- Earlier in semester we have seen that confidence intervals take the form

$$\text{estimate} \pm \text{multiplier} \times \text{std. error}$$

- For testing $H_0 : \beta_1 = \text{null}$ we use the test statistic

$$t = \frac{\text{estimate} - \text{null}}{\text{std. error}}$$

- These continue to apply for a simple linear regression model

Confidence interval for slope

$$\text{estimate} \pm \text{multiplier} \times \text{std. error}$$

- Estimate is $\hat{\beta}_1$
- Multiplier comes from a t-distribution with $\nu = n - 2$ degrees of freedom.
  - Degrees of freedom match denominator in equation $s_\varepsilon^2 = RSS/(n-2)$.
  - So for $100(1-\alpha)\%$ confidence interval, multiplier is $t_{(1-\frac{\alpha}{2},\nu)}$.
- Standard error is

$$s_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Confidence interval 'by hand'

$$\hat{\beta}_1 \pm t_{(1-\frac{\alpha}{2},n-2)} s_{\hat{\beta}_1}$$

$$0.3699 \pm t_{(0.975,277)} \times 0.0478$$

- There are $n = 279$ observations
- From R: `qt(0.975, 277)` $= 1.9686$

$$0.3699 \pm 1.9686 \times 0.0478$$

$$0.3699 \pm 0.0941$$

$$(0.2758, 0.464)$$

- We are 95% confident that the true slope is between 0.2758 and 0.464
  - We estimate that the expected height of a son will increase by between 0.2758 and 0.464 cm for a 1 cm increase in height of father

In R

- We typically find confidence intervals in R using `confint` function
  - It is important to understand how the confidence interval is found

```
confint(m_height)
```

```
##                2.5 %  97.5 %
## (Intercept) 97.3247 130.782
## father       0.2758   0.464
```

- Confidence interval for `father` is identical to that calculated on previous slide
- For a 99% confidence interval

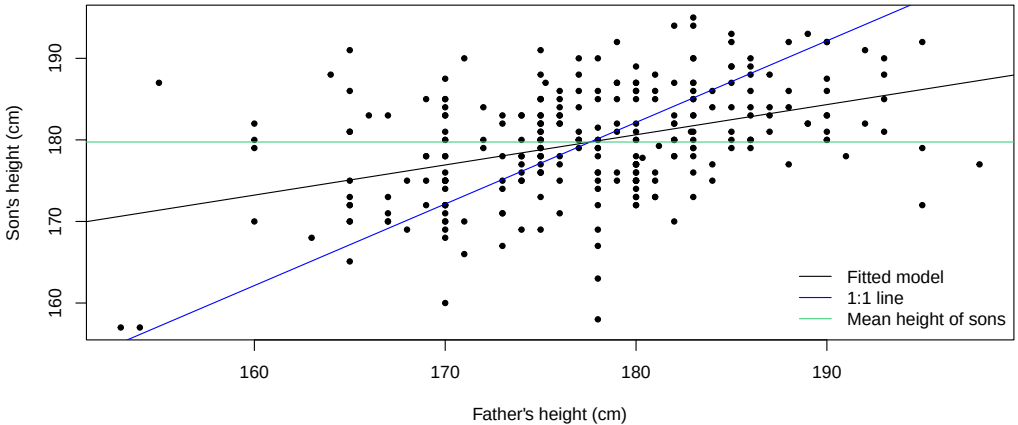
```
confint(m_height, level = 0.99)
```

```
##                0.5 %   99.5 %
## (Intercept) 92.0124 136.0943
## father       0.2459   0.4939
```

Excursion: Regression to the mean

- We might have expected the average height of a son to increase by 1 cm for a 1 cm increase in father's height.
- That it does not, is the origin of the label: regression (to the mean)
  - The son of a short father tends to be short, but on average he is taller than his father
  - The son of a tall father tends to be tall, but on average he is shorter than his father
  - Extreme traits tend to regress to the mean
- 'Regression' introduced by Francis Galton when comparing the heights of parents and children
  - Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, **15**, 246–263

Excursion: Regression to the mean



Hypothesis test for the slope

- Recall: $y = \beta_0 + \beta_1 x + \varepsilon$
  - β_1 describes how the mean response μ_y changes with x at population level
- If $\beta_1 = 0$ then $y = \beta_0 + \varepsilon$
  - $\mu_y = \beta_0$: μ_y does not depend on x
  - Outcome variable is not (linearly) related to the predictor variable
- A hypothesis test about β_1 assesses the hypothesis that two variables are related
  - Null hypothesis: statement of no relationship between x and y
    - $\mathsf{H}_0 : \beta_1 = 0$
  - Alternative hypothesis: relationship exists
    - $\mathsf{H}_A : \beta_1 \neq 0$

The test statistic

- To compute the p-value, we need a test statistic
- The test statistic is

$$t = \frac{\text{estimate} - \text{null}}{\text{std. error}}$$

- The estimate is $\hat{\beta}_1$
- The null value is 0 (previous slide)
- The standard error is $s_{\hat{\beta}_1} = s_\varepsilon / \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - See previous lecture
- So for testing hypothesis about β_1, we use the test statistic

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

Example: PCB in trout

- Concern that polychlorinated biphenyls (PCBs) polluting waterways and accumulating in food chain
  - PCBs used to be commonly found in transformers, capacitors, paints, etc
  - 28 trout collected[1] from Cayuga Lake, NY in 1970
    - Fish were marked and annually stocked (age was known)
    - Each trout was (mechanically) chopped, ground, and mixed before a 5 gm sample taken
    - Chromatography used to find PCB residue in ppm (parts per million)
- Scientific question: is there evidence that (log) PCB residue increases with age?
  - Null hypothesis: $\mathrm{H}_0 : \beta_1 = 0$
  - Alternative hypothesis: $\mathrm{H}_A : \beta_1 \neq 0$
- Treat it as a confirmatory study (specific hypothesis to assess)

[1] Science (1972), **177**, 1191–1192.

Example: PCB in trout

- Import the data into R

```
pcb = read.csv('pcb.csv')
```

- Look at the data

```
head(pcb)
```

```
##   age  logpcb
## 1   1 -0.5108
## 2   1  0.4700
## 3   1 -0.6931
## 4   1  0.1823
## 5   2  0.6931
## 6   2  0.2624
```

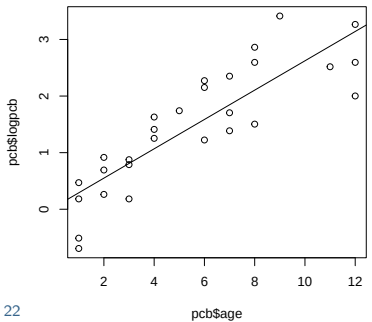
Example: PCB in trout

- Fit simple linear regression

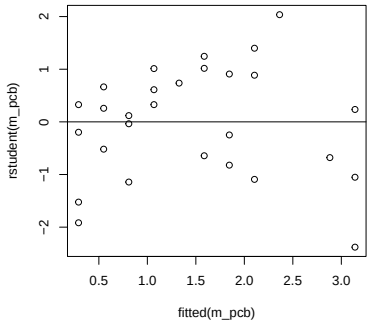
```
m_pcb = lm(logpcb ~ age, data = pcb)
```

- Plot fitted model and residuals: any concerns?

```
plot(pcb$age, pcb$logpcb)
abline(m_pcb)
```




```
plot(fitted(m_pcb), rstudent(m_pcb))
abline(h = 0)
```



R model output

```
summary(m_pcb)

##
## Call:
## lm(formula = logpcb ~ age, data = pcb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1395 -0.3879  0.0957  0.4327  1.0508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0315     0.2014    0.16     0.88
## age           0.2591     0.0308    8.41  6.8e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.567 on 26 degrees of freedom
## Multiple R-squared:  0.731, Adjusted R-squared:  0.721
## F-statistic: 70.8 on 1 and 26 DF,  p-value: 6.78e-09
```

Interpretation

- For a confirmatory study
  - Formal test
- Compare the p-value to α
  - If p-value $< \alpha$: reject H_0
    - Evidence in favour of H_A
  - If p-value $> \alpha$: fail to reject H_0
- For an exploratory study
  - Interpret the p-value as a degree of incompatibility between data and null hypothesis
    - Use α as a guide
    - Try to avoid making a decision between hypotheses
  - Often prefer to use confidence intervals

Interpretation PCB: $\alpha = 0.05$

- The test statistic t is given in column `t value`: `8.41`
- The p-value is given in the column `Pr(>|t|)`: `6.8e-09`
  - These are found assuming the hypothesis: $H_0 : \beta_i = 0$
- p-value $< \alpha$: evidence of incompatibility between the data and null hypothesis
  - Data are incompatible with assumption of no relationship between PCB and age
  - Data are unusual compared to what we would expect if the null hypothesis were correct
- As this is a confirmatory study, we conclude that
  - There is evidence against H_0
  - There is evidence of a relationship between (log) PCB and age of fish (H_A)

Summary

- We want to quantify how precise our estimate is
  - Estimate of error variance
  - Estimate of standard error for $\hat{\beta}_1$
  - Found confidence interval for β_1
  - Hypothesis test for β_1
- Discussed origin of ‘regression’