The Linear Model 2: R^2 Strikes Back

Lecture 8

Dr Jennifer Mankin

14 March 2022

UNIVERSITY OF SUSSEX

Looking Ahead (and Behind)

- Previously
  - Samples, distributions, and *t*-tests
  - The take-away paper
- Last week: The Linear Model - Equation of a Line
- This week: The Linear Model - Evaluating the Model

TAP Set Analysis

- View the Set Analysis document on Canvas > Take-away paper Information
    - You **must use the Set Analysis results** to complete your Psychobiology report
- If you don't have exactly these results, DON'T PANIC
    - Marking is based on everything you submitted, not just getting one right answer!

Objectives

After this lecture you will understand:

- The equation for a linear model with one predictor
    - b_0 (the intercept)
    - b_1 (the slope)
- The logic of NHST for b-values
    - Interpreting p and CIs
- How to assess model fit with R^2

General Model Equation

$$outcome = model + error$$

- We can use models to **predict** the outcome for a particular case
- This is always subject to some degree of **error**

The Linear Model

- The linear model predicts the outcome y based on a predictor x
  - General form: $y_i = b_0 + b_1 x_{1i} + e_i$
  - b_0: the **intercept**, or value of y when x is 0
  - b_1: the **slope**, or change in y for every unit change in x
- The slope b_1 represents the relationship between the predictor and the outcome

Today's Example

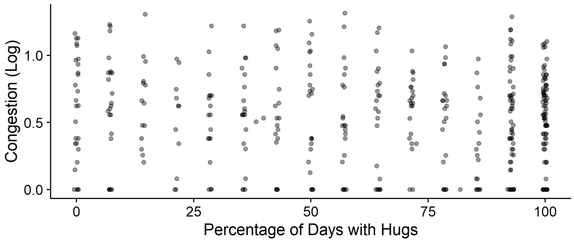
- "Does Hugging Provide Stress-Buffering Social Support? A Study of Susceptibility to Upper Respiratory Infection and Illness" (Cohen et al., 2015)
- Participants completed questionnaires and phone interviews over 14 days
  - Including whether they had been hugged each day
- Then exposed to a cold virus! 🤧
  - Measures of infection: amount of mucus, nasal clearing time
- Does receipt of hugs have a relationship with infection?
  - What kind of relationship might we predict? 🤔

Operationalisation

- Predictor: Percentage of days in which participants were hugged
  - Higher percentage = more hugs
- Outcome: Nasal clearing time
  - A measure of congestion
  - Longer time = more congestion (= worse cold)
- Model: $Congestion_i = b_0 + b_1 \times Hugs_{1i} + e_i$
  - Use the `lm()` function to estimate b_0 and b_1

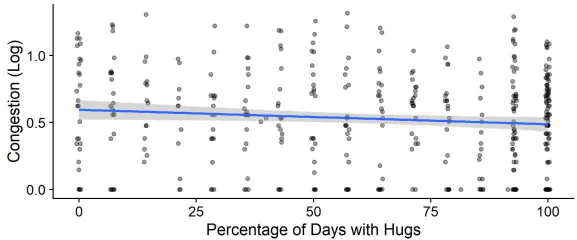
Having a Look

Plot | R code



Having a Look

Plot | R code



- Very slight *negative* relationship
  - How can we interpret this relationship?

Creating the Model

```
##
## Call:
## lm(formula = post_nasal_clear_log ~ pct_hugs, data = cold_hugs)
##
## Coefficients:
## (Intercept)     pct_hugs
##      0.5952      -0.1077
```

- For every *unit increase* in hugs, congestion changes by -0.11
    - Here, "unit increase" = 1%
    - So, congestion goes down by 0.11 for every 1% increase in hugs

Model: $Congestion_i = 0.60 - 0.11 \times Hugs_i$

The Story So Far

- Investigating whether hugs protect against colds
- Linear model shows that more hugs are associated with less congestion (infection)
  - $Congestion_i = 0.60 - 0.11 \times Hugs_i$
- Is this model any good? What do we mean by "good"?
  - Captures a relationship that may in fact exist: significance and CIs of b_1
  - Explains the variance in the outcome: R^2 for the model

NHST for LM

- b_1 quantifies the relationship between the predictor and the outcome
    - The *effect* of interest and the key part of the linear model!
    - Our estimate of the true relationship in the population (the model parameter)
- So...is this value **significant**?

NHST for LM

- Our recipe for significance testing is:
  - Data
  - A test statistic
  - The distribution of that test statistic under the null hypothesis
  - The probability p of finding a test statistic as large as the one we have (or larger) if the null hypothesis is true
- First, we need to sort out the null hypothesis of b_1

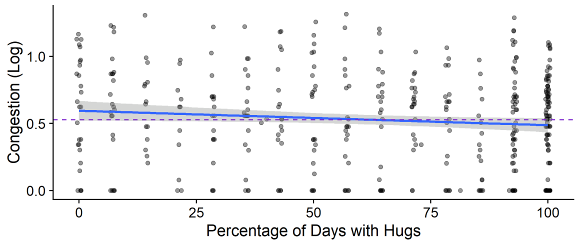
Null Hypothesis of b_1

- b_1 captures the relationship between variables
    - How much the outcome y changes for each unit change in x
    - Null hypothesis: the outcome y **does not change** when x changes
- What would this look like in terms of the linear model? 🤔

Null Hypothesis of b_1

- $Congestion_i = b_0 + 0 \times Hugs_{1i} + e_i$
  - This is the **null** or **intercept-only model**

Plot | R code



Significance of b_1

- $b_1 = 0$ represents the null hypothesis
    - So, the alternative hypothesis is $b_1 \neq 0$
- For our model, does $b_1 = 0$?
    - No! Here $b_1 = -0.11$
- Is our estimate of b_1 different **enough** from 0 to believe that it may actually not be 0 in the population?
    - Compare the estimate of b_1 to the variation in estimates of b_1

Significance of b_1

- Signal-to-noise ratio
  - Signal: the estimate of b_1
  - Noise: the standard error of b_1
- Scale b_1 by its standard error: $\frac{b_1}{SE_{b_1}}$
- What do you get when you divide a normally distributed value by its standard error...???? 🤔

That's the *t*!

Significance of b_1

- $\frac{b_1}{SE_{b_1}} = t$
  - Compare our value of *t* to the *t*-distribution to get *p*, just as we've seen before
  - If *p* is smaller than our chosen alpha level, our predictor is considered to be significant

Term	*b*	SE_b	*t*	*p*
Intercept	0.60	0.04	16.04	< .001
Percentage of Days with Hugs	-0.11	0.05	-2.05	.041

Confidence Intervals for b_1

- Give us the range of likely sample estimates of β_1 from other samples
  - Only if our interval is one of the 95% of intervals that does in fact contain the population value!
  - Review Lecture 2 for more on CIs
- Key info: does the confidence interval cross or include 0?
  - If yes, it's likely that we could have gathered a sample where b_1 was 0

Confidence Intervals for b_1

- What can we conclude from these confidence intervals? 🤔

Term	b	SE_b	t	p	CI_{upper}	CI_{lower}
Intercept	0.60	0.04	16.04	< .001	0.522	0.668
Percentage of Days with Hugs	-0.11	0.05	-2.05	.041	-0.211	-0.005

Interim Summary

- The key element of the linear model is b_1
  - Quantifies the relationship between the predictor and the outcome
  - Null hypothesis: $b_1 = 0$
  - Alternative hypothesis: $b_1 \neq 0$
- Is b_1 different from 0?
  - Significance via t
  - Confidence intervals

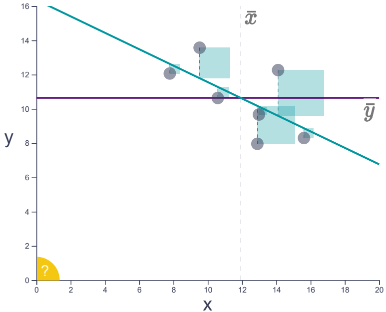
A Good Model

- Captures a relationship that does in fact exist
  - Isn't just noise (random variation)
  - Quantified with significance/CIs
- Is useful for understanding the outcome variable
  - Explains variance in the outcome
  - Quantified with R^2

Explaining Variance

- We want to explain variance, particularly in the outcome
- Goodness of Fit: How well does the model fit the data?
  - Better fit = model is better able to explain the outcome
- So, how do we quantify model fit?

Goodness of Fit with R^2



$$R^2 = 1 - \frac{SS_R}{SS_T} = 1 - \frac{16.48}{26.87} = .39$$

Goodness of Fit with R^2

- $R^2 = \frac{Variance\ explained\ by\ the\ model}{Total\ variance}$
  - Interpret as a percentage of variance explained
  - Applies to our sample only
  - Larger value means better fit
- Adjusted R^2: estimate of R^2 **in the population**
- How does this value look? 🤔

R^2	Adjusted R^2	*F*	*p*
0.01	0.01	4.22	.041

Putting It All Together

```
hugs_lm %>% summary()
```

```
##
## Call:
## lm(formula = post_nasal_clear_log ~ pct_hugs, data = cold_hugs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59522 -0.27880  0.01766  0.25219  0.79262
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  0.59522    0.03710  16.043 <0.0000000000000002 ***
## pct_hugs    -0.10773    0.05243  -2.055              0.0405 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3572 on 403 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.01037,    Adjusted R-squared:  0.007912
## F-statistic: 4.222 on 1 and 403 DF,  p-value: 0.04055
```

Summary

- The linear model (LM) expresses the relationship between at least one predictor, x, and an outcome, $\hat{y}$
  - Linear model equation: $y_i = b_0 + b_1 x_{1i} + e_i$
  - Most important result is the parameter b_1, which expresses the change in y for each unit change in x
- Evaluating the model
  - Is it unlikely that b_1 isn't 0? Significance tests and CIs
  - How well does the model fit the data? R^2 and adjusted R^2

More Strikes??

- Have a look at the impact of pension changes: http://uss-pension-model.com/
- Staff working conditions are your learning conditions!