Source: The Guardian https://www.theguardian.com/cities/gallery/2016/jun/08/how-far-distance-workers-commute-uk-cities-mapped

Logistic Regression

Who is Willing to Commute Long Distances?

Zi Ye

ENVS225

Exploring the Social World

UNIVERSITY OF LIVERPOOL

So far – Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \epsilon$$

Scale/Continuous variables

Dummy variables

Scale/Continuous variables

UNIVERSITY OF LIVERPOOL

But - Categorical Dependent Variables

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \epsilon$$

Categorical variables

Binominal logistic regression for binary variable: 0 or 1

- Examples:
  - Health outcomes/behaviours
    - eg. smoking, drinking, cancer, heart attack, HIV, etc.
  - Employment outcomes
    - unemployed, employed full-time, employed part-time, self-employed, job satisfaction, etc.
  - Decision making processes
    - Brexit vote, travelling, migration, long-distance commuting, etc.

Learning Outcomes

Aim: Understanding how to estimate and interpret a <u>logistic regression model</u>

Binominal logistic regression for binary variable: 0 or 1

Estimate and interpret a logistic regression model

Assess the model fit

Make predictions using a logistic regression

UNIVERSITY OF LIVERPOOL

What is a Logistic Regression?

Logistic regression

- Used to calculate the probability of a binary event (1 or 0) occurring
- and to deal with issues of classification (1 or 0).

Examples:

- Pass the examination (1: Yes; 0: No)
- Disease (1: Yes; 0: No)
- Online purchase (1: Will buy; 0: No)
- Vote for candidate: (1: Will vote; 0: No)
- etc.

From Probability to Odds

Everything starts with the concept of probability.

Let's say that the probability of success of some event is p =0.8

Then the probability of failure is 1 – p= 0.2

The odds of success are defined as the ratio of the probability of success over the probability of failure.

Odds of success are $\frac{p}{1-p}$ = 0.8/0.2 = 4

We say the odds of success are 4 to 1.

If the probability of success is 0.5, i.e., 50-50 % chance, then the odds of success is 1 to 1.

From probability to odds to log of odds

p Probability: from 0 to 1

$\frac{p}{1-p}$ Odds: from 0 to +∞

$log\left(\frac{p}{1-p}\right)$ Log Odds: from - ∞ to +∞

alternative way of expressing probabilities, but why use the last one?

Linear model $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots$ can ranges over all real numbers $(-\infty,\infty)$

Logistic Regression

Y ← takes values of 1 or 0

p **is the probability of the event occurring: Y=1**

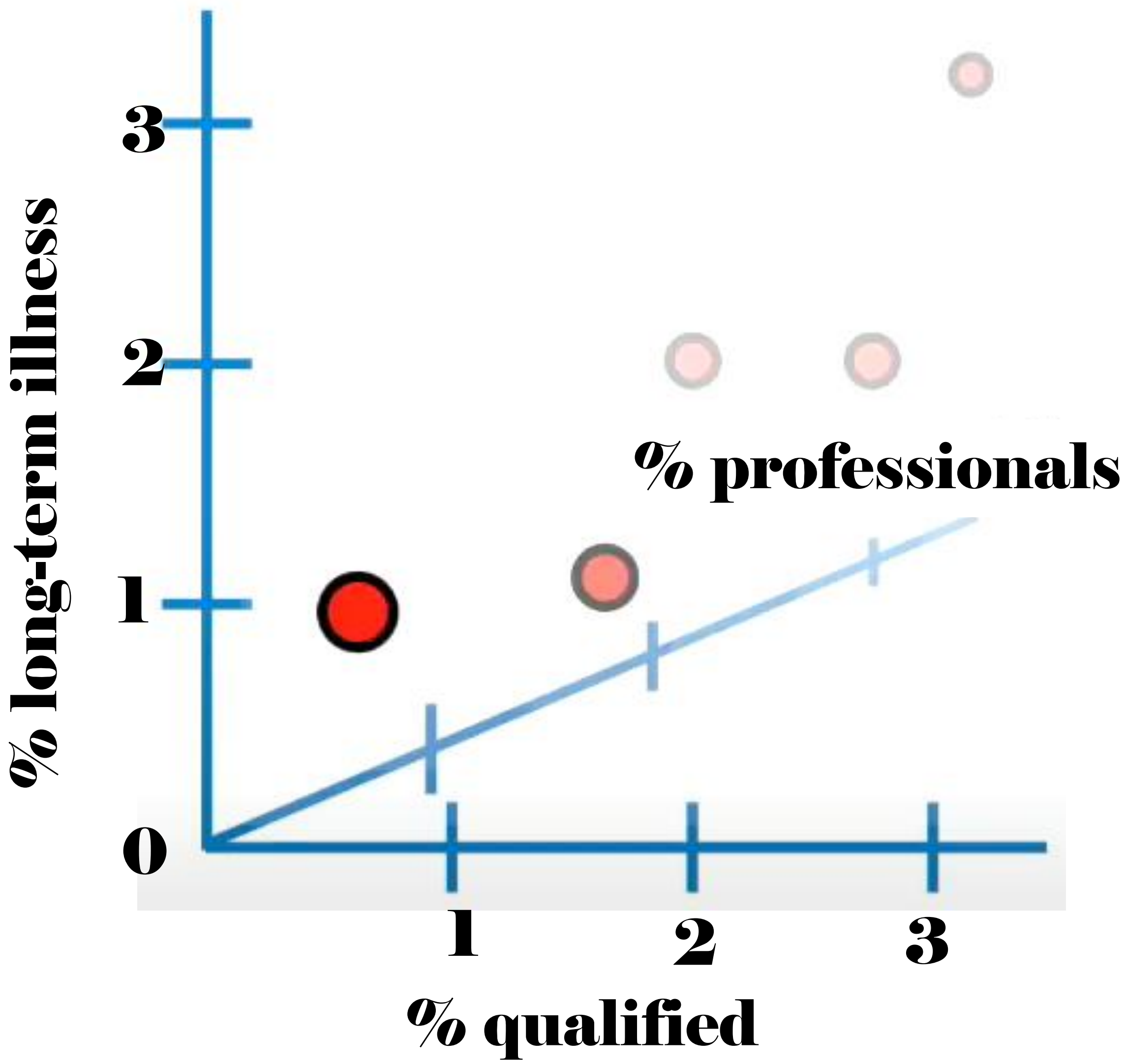
Log odds

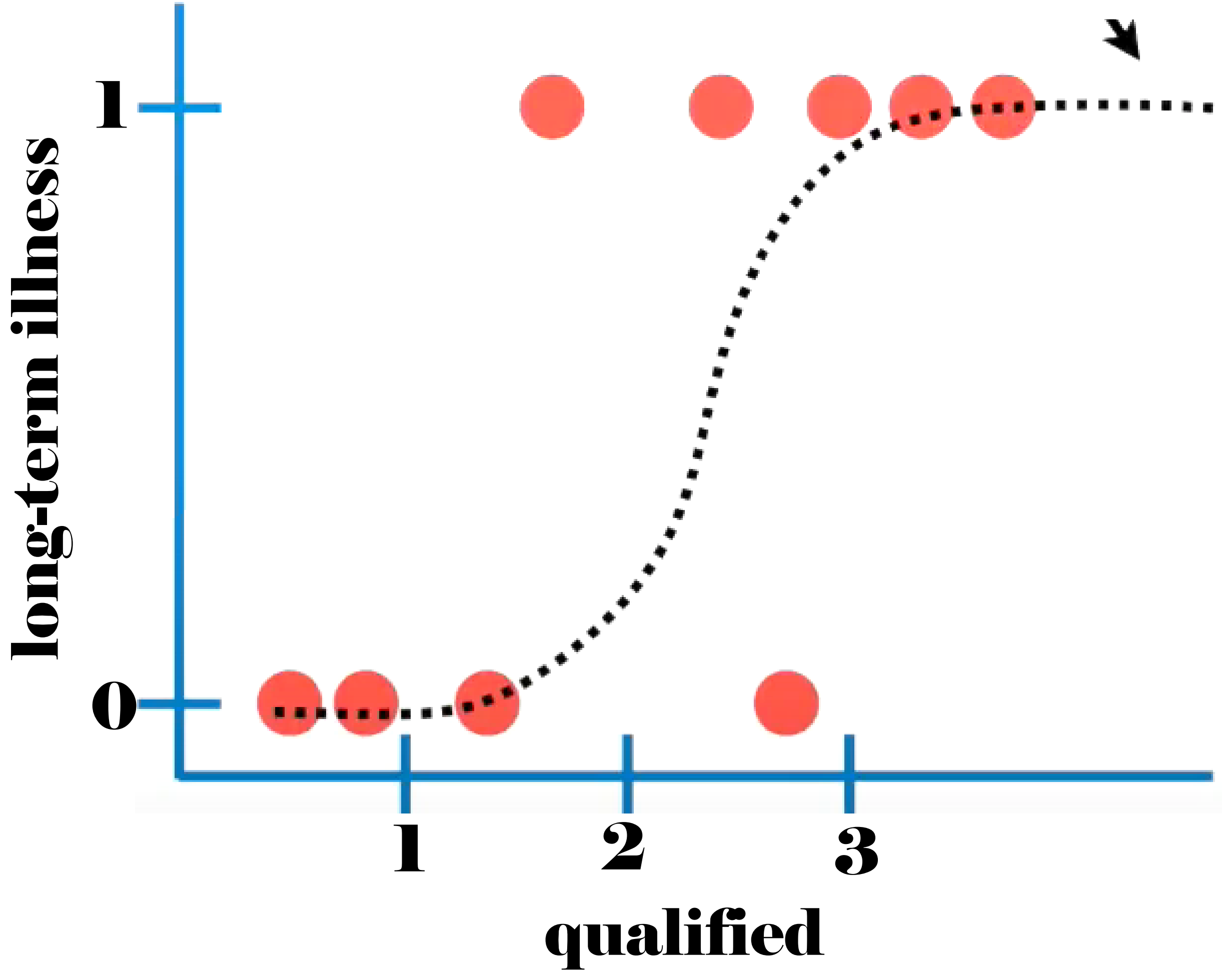
$$\boxed{logit(p) = log\left(\frac{p}{1-p}\right)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

The model estimates the **log-odds** (logit) of the probability p that $Y = 1$:

UNIVERSITY OF LIVERPOOL

Linear vs logistic





What Does a Logistic Regression Do?

- Predict the **probability** of an event happening based on at least one independent variable
  - vs. linear regression - estimates the **average** value

- Dependent variable: **qualitative** (binary) variable **0 or 1,**
  - vs. linear regression - continuous variable

Model the probability of the dependent variable being 1, ensuring that predicted probabilities (*p*) always fall between 0 and 1.

UNIVERSITY OF LIVERPOOL

Also Known as

- binary regression model
- Logit model
- discrete choice model
- probability regression model
- qualitative response regression model

UNIVERSITY OF LIVERPOOL

Interpretation of βs

Intercept

- the log-odds of an event happening (i.e. Y=1) if the value of the explanatory variables Xi is all zero.

Slope

- Each βi estimates the change in the log-odds of Y=1 for a one-unit change in Xi, holding other variables constant.

- The exponentiated coefficient Exp(βi) gives the odds ratio — how much the odds of Y=1 multiply when Xi increases by one unit.

UNIVERSITY OF LIVERPOOL

Interpretation of Exp(βi) Odds Ratio

Calculates the relationship between a variable (e.g., qualifications) and the likelihood of an event occurring (e.g., long-term ill)

- gives the **expected change in the odds** for a unit change in Xi, holding all other variables constant.

- Exp(β) = 1 if β = 0: indicates is equally likely to occur.

- Exp(β) > 1 if β > 0: indicates an event is more likely to occur, or the odds are "exp(bk) times larger"

- Exp(β) < 1 if β < 0: indicates an event is less likely to occur, or the odds are "exp(bk) times smaller"

UNIVERSITY OF LIVERPOOL

Who is Willing to Commute Long Distances?

Dependent Variable (I)

- Defining long-distance commuting using SAR:
  - Distance travelled to work - categorical variable - 12 cat.
  - Select cases reporting kms travelled

1 Less than 2 km
2 2 to <5 km
3 5 to <10 km
4 10 to <20 km
5 20 to <40 km
6 40 to <60km
7 60km or more
8 At home
9 No fixed place
10 Work outside England and Wales but within UK
11 Work outside UK
12 Works at offshore installation (within UK)

Only use these records in the model

UNIVERSITY OF LIVERPOOL

Dependent Variable (2)

- Defining long-distance commuting using SAR:

id	distance travelled to work	long-distance commuting
1	5 to <10km	0
2	20 to <40km	0
3	60km or more	1
4	At home	0
5	60km or more	1

Travel over 60km defines as long-distance commuting

UNIVERSITY OF LIVERPOOL

Explanatory Variables

- Gender:
    - Male (0) & Female (1)

- Socio-Economic Classification:
    - 12 Categories, exclude
        - unemployed,
        - full-time students &
        - not classifiable,
        - child under 15

1	Large employers and higher managers
2	Higher professional occupations
3	Lower managerial and professional occupations
4	Intermediate occupations
5	Small employers and own account workers
6	Lower supervisory and technical occupations
7	Semi-routine occupations
8	Routine occupations
9	Never worked or long-term employed
10	Full-time student
11	Not classifiable
12	Child aged 0-15

UNIVERSITY OF LIVERPOOL

Estimation

Estimation by using R

Long commute 1
Not long commute 0

Males 1
Females 2

```
#create the model
m.glm = glm(New_work_distance~sex + nssec,
            data = sar_df,
            family= "binomial")
# inspect the results
summary(m.glm)
```

Socio-Economic Classification : 1 to 8

UNIVERSITY OF LIVERPOOL

R Output (I)

Y=willing to commute long distance

What is the base category
for Sex and Socio-economic status?

```
Call:
glm(formula = New_work_distance ~ sex + nssec, family = "binomial",
    data = sar_df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.67337    0.05329 -31.401  < 2e-16 ***
sex2        -0.36678    0.04196  -8.742  < 2e-16 ***
nssec1      -0.12881    0.11306  -1.139    0.255
nssec3      -0.38761    0.06467  -5.994 2.05e-09 ***
nssec4      -1.03079    0.08439 -12.214  < 2e-16 ***
nssec5       1.22639    0.06489  18.898  < 2e-16 ***
nssec6      -1.38992    0.10919 -12.730  < 2e-16 ***
nssec7      -1.43909    0.09002 -15.986  < 2e-16 ***
nssec8      -1.48534    0.09646 -15.398  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20441  on 33025  degrees of freedom
Residual deviance: 17968  on 33017  degrees of freedom
AIC: 17986

Number of Fisher Scoring iterations: 6
```

Log odds

As Before, Sig!

If p-value < 0.05;
Statistically significant

R Output (2)

```
# odds ratios
exp(coef(m.glm))
```

Odds Exp(β)

```
(Intercept)        sex2      nssec1      nssec3      nssec4      nssec5
  0.1876138   0.6929649   0.8791416   0.6786766   0.3567267   3.4088847
     nssec6      nssec7      nssec8
  0.2490946   0.2371432   0.2264258
```

Not significant

Exp(β) > 1

```
# confidence intervals
exp(confint(m.glm, level = 0.95))
```

Confidence intervals (CI)

```
Waiting for profiling to be done...

                2.5 %    97.5 %
(Intercept) 0.1688060 0.2080319
sex2        0.6381810 0.7522773
nssec1      0.7017990 1.0935602
nssec3      0.5981911 0.7708192
nssec4      0.3020431 0.4205270
nssec5      3.0037298 3.8739884
nssec6      0.2002766 0.3073830
nssec7      0.1984396 0.2824629
nssec8      0.1869397 0.2729172
```

1 Large employers and higher managers

2 Higher professional occupations (baseline)

3 Lower managerial and professional occupations

4 Intermediate occupations

5 Small employers and own account workers

6 Lower supervisory and technical occupations

7 Semi-routine occupations

8 Routine occupations

Interpretation

Think about what are the expected signs?

Positive/Negative?

Why ?

Interpretation of Socio-economics

Y=willing to commute long distance

```
# odds ratios
exp(coef(m.glm))
```

The predicted change in odds Exp(β) *for a unit increase in the predictor (independent variable)*

Exp(β) = 1: equally likely

Exp(β) < 1: less likely

Exp(β) > 1: more likely

```
(Intercept)        sex2     nssec1     nssec3     nssec4     nssec5
  0.1876138   0.6929649  0.8791416  0.6786766  0.3567267  3.4088847
     nssec6      nssec7     nssec8
  0.2490946   0.2371432  0.2264258
```

Not significant

```
# confidence intervals
exp(confint(m.glm, level = 0.95))
```

```
Waiting for profiling to be done...

                2.5 %    97.5 %
(Intercept) 0.1688060 0.2080319
sex2        0.6381810 0.7522773
nssec1      0.7017990 1.0935602
nssec3      0.5981911 0.7708192
nssec4      0.3020431 0.4205270
nssec5      3.0037298 3.8739884
nssec6      0.2002766 0.3073830
nssec7      0.1984396 0.2824629
nssec8      0.1869397 0.2729172
```

Indicates that the probability of long-distance commuting for those whose socio-economic classification as small employers and own account workers are 3.409 times more likely than the higher prof occupations holding all other variables constant, with a likely range (CI) of between 3.0 to 3.8.

Think about the findings for other socio-economic classification?

UNIVERSITY OF LIVERPOOL

Interpretation of Gender

```r
# odds ratios
exp(coef(m.glm))
```

```
(Intercept)        sex2      nssec1      nssec3      nssec4      nssec5
  0.1876138   0.6929649   0.8791416   0.6786766   0.3567267   3.4088847
     nssec6      nssec7      nssec8
  0.2490946   0.2371432   0.2264258
```

Exp(β) < 1

```r
# confidence intervals
exp(confint(m.glm, level = 0.95))
```

```
Waiting for profiling to be done...

                2.5 %    97.5 %
(Intercept) 0.1688060 0.2080319
sex2        0.6381810 0.7522773
nssec1      0.7017990 1.0935602
nssec3      0.5981911 0.7708192
nssec4      0.3020431 0.4205270
nssec5      3.0037298 3.8739884
nssec6      0.2002766 0.3073830
nssec7      0.1984396 0.2824629
nssec8      0.1869397 0.2729172
```

Indicates that the probability of commuting over long distances for female is 0.693 times less likely than male (the reference group), with the CI between 0.6 to 0.7, holding all other variables constant, or, being females reduces the probability of long-distance commuting by 30.7%(1-0.693).

UNIVERSITY OF LIVERPOOL

More information here: https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/

Assessing the Model

Overall Model Assessment

Pseudo R^2s

```r
# or in better format
pR2(m.glm) %>% round(4) %>% tidy()
```

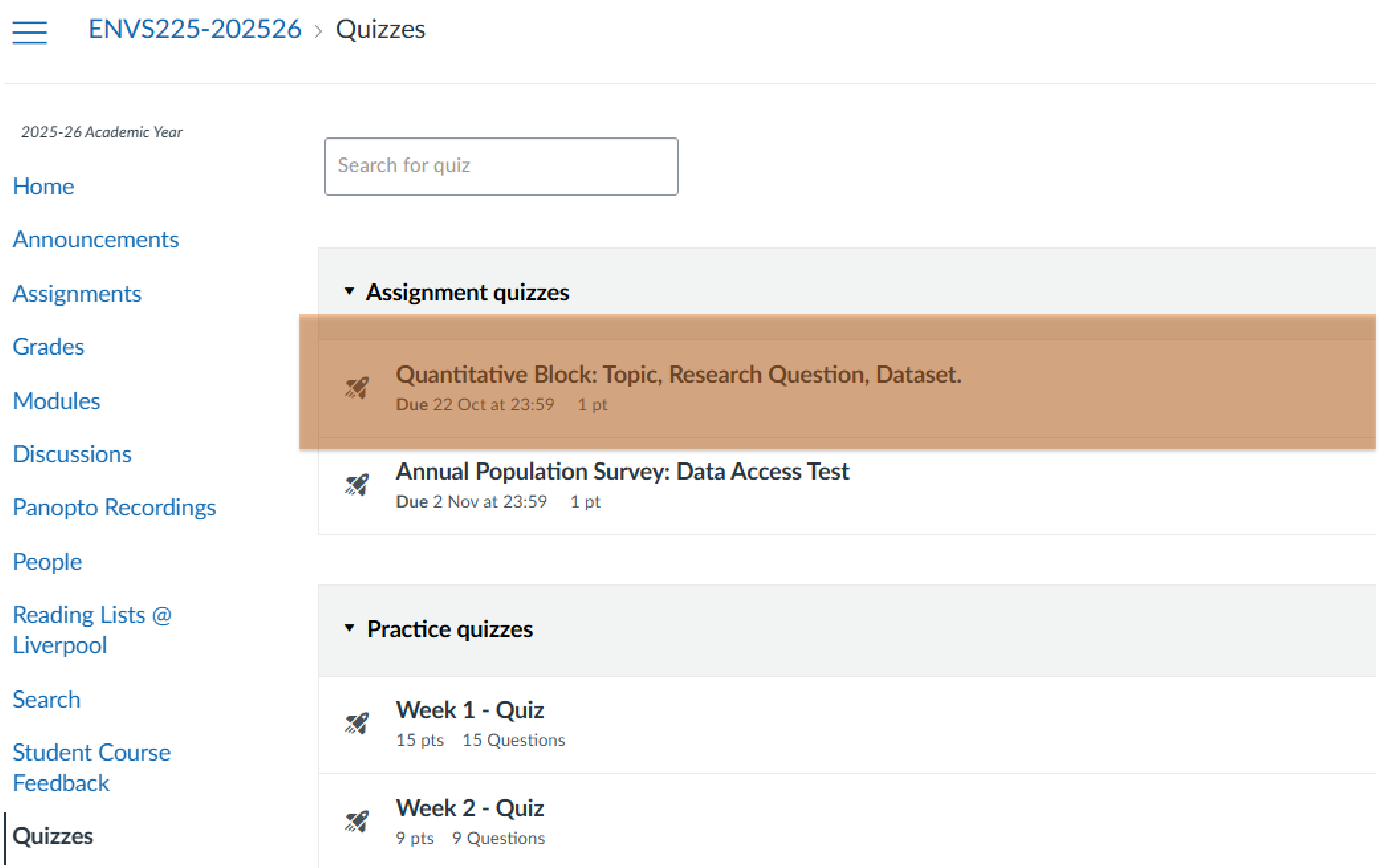
```
fitting null model for pseudo-r2

# A tibble: 6 × 2
  names           x
  <chr>       <dbl>
1 llh        -8984.
2 llhNull   -10220.
3 G2          2473.
4 McFadden   0.121
5 r2ML       0.0721
6 r2CU       0.156
```

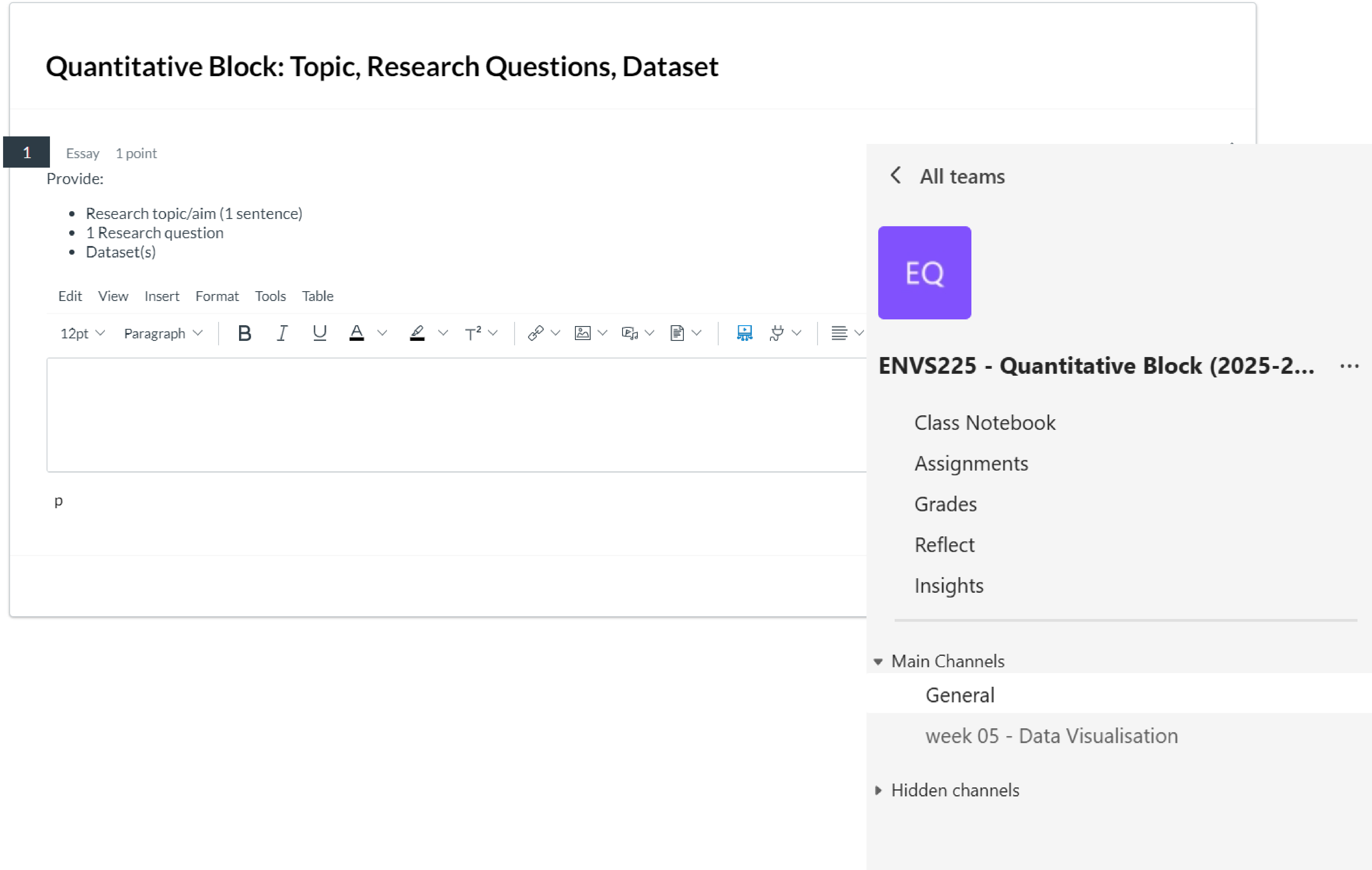
- **llh**: The log-likelihood of the fitted model.
- **llhNull**: The log-likelihood of the null model (without predictors).
- **G2**: The likelihood ratio statistic, showing the model's improvement over the null model.
- **McFadden**: McFadden's pseudo R^2.
- **r2ML**: Maximum likelihood pseudo R^2.
- **r2CU**: Cox & Snell pseudo R^2.

- **Pseudo R^2s** ~ 0.3 are considered fairly good, if using individual-level data
- Note: Pseudo-R^2s **do NOT** have the same meaning that R^2 (i.e.% of explained variance in Y)

Formative assessment support



Formative assessment support



Assessment 1

- Drop-in session
  - This Thursday 15:00-17:00, reservation opened!
  - Week 5 Thursday 11:00-13:00
  - Week 6Wednesday 11:00-13:00
- Formative feedback quiz: 22nd Oct

Submission Date: Week 7 Monday 3rd November

Book drop-in session