Logistic Regression

Measurement & Evaluation of HCC Systems

Logistic Regression

Today's goal:

Evaluate the effect of multiple variables on a categorical outcome variable

Outline:

- Basic theory: extending regression to logistic regression
- Logistic regression (binary outcome)
- Poisson regression* (count outcome)
- Ordered categorical regression* (Likert scales, etc.)

$x \rightarrow y$

Extending regression

to logistic regression

A quick aside...

Regression with interaction effect:

$$Y_i = a + b_1X_{1i} + b_2X_{2i} + b_3X_{1i}X_{2i} + e_i$$

You can do this with any X!

Just make sure that your variables are **centered**

Centering a factor:

Assign contrasts that sum to zero

Centering a continuous X:

Subtract the mean

A quick aside...

$Y_i = a + b_1X_{1i} + b_2X_{2i} + b_3X_{1i}X_{2i} + e_i$

Interpretation if X_1 is continuous and X_2 is binary:

- b_3 is the additional effect of X_1 in the second group of X_2
- b_3 is the additional difference between the two groups of X_2 with each 1 point increase in X_1

Interpretation if both are continuous:

- b_3 is the additional effect of X_1 with each 1 point increase of X_2 (and vice versa)

Logistic regression

Linear regression:

$Y_i = a + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} + e_i$

What if Y is binary (0 or 1)?

We can try to predict the **probability** of Y=1 — P(Y)

However, this probability is a number between 0 and 1

For linear regression, we want an unbounded linear Y!

Can we find some transformation that allows us to do this?

Yes: $P(Y) = 1 / (1+e^{-U})$

Logistic regression

$P(Y) = 1 / (1+e^{-U})$

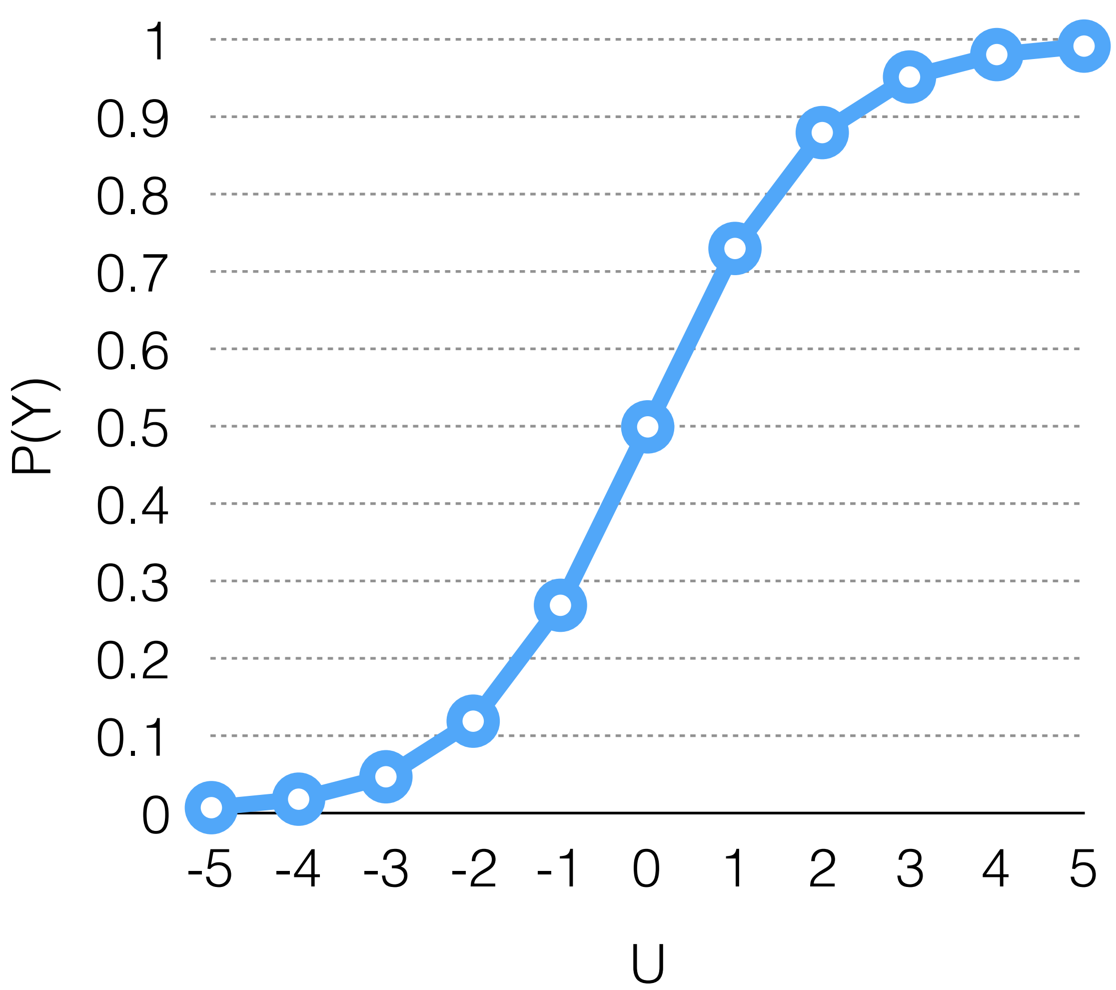
Conversely:

$U = \ln(P(Y)/(1-P(Y)))$

Interpretation:

$P(Y)/(1-P(Y))$ is the **odds** of Y

Therefore, U is the log odds, or **logit** of Y



Logistic regression

Since U is unbounded, we can treat it as our regression outcome:

$U_i = \ln(P(Y_i)/(1-P(Y_i))) = Y_i = a + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} + e_i$

We can always transform it back to $P(Y_i)$ if we want to:

$P(Y_i) = 1 / (1+e^{-(a + b1X1i + b2X2i + \ldots + bkXki + ei)})$

Log-likelihood

How do we assess the fit of a logistic regression?

We calculate the **log-likelihood**, which is a type of residual

Log-likelihood = $\sum(Y_i * \ln(P(Y_i)) + (1 - Y_i) * \ln(1 - P(Y_i)))$

where Y_i is the observed value, and $P(Y_i)$ is the predicted value

Log-likelihood

Log-likelihood = $\sum(Y_i * \ln(P(Y_i)) + (1 - Y_i) * \ln(1 - P(Y_i)))$

If $Y_i = 1$, then this simplifies to $\ln(P(Y_i))$

which is zero when the prediction is correct ($P(Y_i)=1$) but gets a large (negative) value if the prediction is incorrect ($P(Y_i)$ is closer to 0)

If $Y_i = 0$, then this simplifies to $\ln(1 - P(Y_i))$

which is zero when the prediction is correct ($P(Y_i)=0$) but gets a large (negative) value if the prediction is incorrect ($P(Y_i)$ is closer to 1)

Deviance (−2LL)

A more useful measure is deviance (a.k.a. −2LL)

- −2 * log-likelihood

Difference can be used to compare nested models

- Likelihood ratio: $\chi^2 = -2LL_{baseline} - -2LL_{new}$
- Chi-square distribution with $k_{new} - k_{baseline}$ df

In regression we compared against the mean

- In logistic regression we compare against the majority class (either 0 or 1)

Information criteria

Using –2LL to compare non-nested models:

Akaike Information Criterion (AIC):

$$AIC = -2LL + 2k$$

Bayesian Information Criterion (BIC):

$$BIC = -2LL + 2k*\log(N)$$

R^2

We can use –2LL to calculate R^2, but there is some disagreement on how to do this

Hosmer and Lemeshow method:

$$R_L^2 = (-2LL_{baseline} - -2LL_{new}) \,/\, -2LL_{baseline}$$

Cox and Snell method:

$$R_{CS}^2 = 1 - \exp((-2LL_{new} - -2LL_{baseline})/N)$$

Nagelkerke method:

$$R_N^2 = R_{CS}^2 \,/\, (1 - \exp(2LL_{baseline}/N))$$

$x \to y$ Coefficients

In regression, we can test the significance of the b coefficients with a t-test ($t = b/SE_b$)

In logistic regression, this is a z-test

- $z = b/SE_b$ (Wald statistic)

The Wald statistic is prone to inflating type II errors, though

- Better to just do likelihood ratio model comparisons

Coefficients

How to interpret the b coefficients?

- b is the increase in U for each increase of X
- b is the increase in $\ln(P(Y)/(1-P(Y)))$ for each increase in X
- e^b is the ratio of $P(Y)/(1-P(Y))$ for each increase in X
- e^b is the **odds ratio**

Coefficients

Odds ratio examples:

If $e^b > 1$: The odds of Y are e^b times as high for each increase in X

- E.g. $e^b = 3$: The odds of Y are 3 times as high for each increase in X

If $e^b < 1$: The odds of Y are $1/e^b$ times as low for each increase in X

- E.g. $e^b = .333$: The odds of Y are 3 times as low for each increase in X

Coefficients

If e^b = 1.xx: each 1 pt increase in X leads to a xx% increase in the odds of Y

> E.g. e^b = 1.30: The odds of Y are 30% higher for each increase in X

If e^b = 0.xx: each 1pt increase in X leads to a (100-xx)% decrease in the odds of Y

> e^b = 0.70: The odds of Y are 30% lower for each increase in X

Assumptions

Linearity

- In this case, we assume that there is a linear relation between the Xs and the **logit** of Y

Independence

No multicollinearity

Problems

Some times a logistic regression does not converge

- You will get weirdly large standard errors

1. You have no or little data for some combinations of Xs
   - This is especially problematic when Xs are nominal
2. One or a combination of Xs are a perfect predictor of Y
   - The odds ratios are infinite!

Solution:

- Collect more data, or use a simpler model!

Logistic regression

in R

Logistic regression

Dataset “eel.dat”

- Effect of a treatment on constipation

Variables:

- Cured: whether the patient was cured
- Intervention: whether the patient received “No Treatment” or the “Intervention”
- Duration: how long the patient had been constipated

Logistic regression

Relevel the Cured variable so that “Not Cured” becomes the baseline:

eel$Cured <- relevel(eel$Cured, “Not Cured”)

Relevel the Intervention variable so that “No Treatment” becomes the baseline:

eel$Intervention <- relevel(eel$Intervention, “No Treatment”)

Plotting

Plot of the difference in cured percentage between No Treatment and Intervention, with bootstrapped CI:

> ggplot(eel, aes(Intervention, as.numeric(Cured == "Cured"))) + stat_summary(fun.y=mean, geom="bar", fill="white", color="black") + stat_summary(fun.data = mean_cl_boot, geom="errorbar", width=0.2) + ylim(0,1)

Note:

> I'm using as.numeric(Cured == "Cured") to turn this factor into a 0-1 variable…

Plotting

Cured percentage by duration, with bootstrapped CI:

```
ggplot(eel, aes(eel$Duration, as.numeric(Cured ==
“Cured”))) + stat_summary(fun.y=mean, geom=“line”) +
stat_summary(fun.data=mean_cl_boot, geom = “errorbar”,
width=0.2)
```

Split by Intervention:

```
ggplot(eel, aes(eel$Duration, as.numeric(Cured ==
“Cured”), color=Intervention)) + stat_summary(fun.y =
mean, geom=“line”) + stat_summary(fun.data =
mean_cl_boot, geom = “errorbar”, width=0.2)
```

Running a model

Run the model:

```
eel1 <- glm(Cured~Intervention, data=eel, family=binomial)
summary(eel1)
```

This gives us:

- Estimates of the X variable (more on this later)
- Deviance of the baseline model +df
- Deviance of the current model (residual deviance) + df

Model statistics

Model chi-square:

- Likelihood ratio: ratio <- eel1$null.deviance – eel1$deviance
- Degrees of freedom: df <- eel1$df.null – eel1$df.residual
- You can also get these from anova(eel1)
- p-value: 1 - pchisq(ratio, df)

R-square:

- Hosmer-Lemeshow: ratio / eel1$null.deviance
- Cox-Snell: Rcs <- 1-exp(-ratio/113)
- Nagelkerke: Rcs / (1-exp(-eel1$null.deviance/113))

Coefficients

```
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -0.2877     0.2700  -1.065  0.28671
InterventionIntervention      1.2287     0.3998   3.074  0.00212 **
```

Calculate percentages:

With no treatment: $P(Y) = 1/(1+e^{0.2877}) = .429$

With treatment: $P(Y) = 1/(1+e^{0.2877-1.2287}) = .719$

Coefficients

```
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -0.2877     0.2700  -1.065  0.28671
InterventionIntervention       1.2287     0.3998   3.074  0.00212 **
```

The intervention has a significant effect

- The z-score may be underestimated

What does b = 1.23 mean?

- calculate e^b: exp(eel1$coefficients)
- The odds of a treated patient being cured are 3.42 higher than those of a patient who is not treated!

Coefficients

Do the same thing for confidence intervals:

- exp(confint(eel1))
- Note: these are not based on the Wald statistic!
- Does not cross 1, therefore, the intervention is significant

Adding Duration

Run the model:

```
eel2 <- glm(Cured~Intervention+Duration, data=eel, family=binomial)
```

Interpret the results: summary(eel2)

- Duration does not have a significant effect
- Deviance very similar to eel1
- What is the difference? anova(eel1, eel2)
- Significance? 1–pchisq(eel1$deviance–eel2$deviance, eel1$df.residual–eel2$df.residual)

$x \rightarrow y$ Diagnostics

Diagnostics are largely the same as with linear regression:

- You can inspect multicollinearity using VIF
- You can get standardized residuals, Cook's distances, leverage and covariance ratios

Diagnostics

Test for linearity of continuous Xs:

Calculate the interaction of the X with its log:

```
eel$logDurationInt <- eel$Duration*log(eel$Duration)
```

Add this to the model:

```
eel3 <- glm(Cured~Intervention+Duration+logDurationInt,
data=eel, family=binomial)
```

If logDurationInt is significant, then there is non-linearity

Reporting

Use a table like Table 8.2 in Field

- Report not just b and SE_b, but also the odds ratio (and maybe its confidence interval)
- Make sure to report R^2 (Nagelkerke is most accepted), Model χ^2, and p-value

If you test multiple models, present the delta R^2 and results of the χ^2 ratio test

Poisson regression

Something that's not in the book!

Poisson regression

Dataset “awards.dat”

Awards won by high school students

Variables:

id: student id

num_awards: number of awards won

prog: type of high school program the student is in

math: the student’s math score

Plotting

Make sure “General” is the baseline type of school:

```
awards$prog <-relevel(awards$prog, ref=“General”)
```

Histogram by academic program:

```
ggplot(awards,aes(num_awards,fill=prog)) +
geom_histogram(binwidth=0.5, position=“dodge”)
```

$x \to y$ A problem...

Doesn't look very normal!

- This is because num_awards is a count variable!
- Other examples: # of purchases, # of clicks, time*, price*

Can we find some transformation that makes this work?

- Yes: $Y = e^U$

Coefficients

How to interpret the b coefficients?

b is the increase in U for each increase of X

b is the increase in the **log rate** of Y for each increase in X

e^b is the ratio of rate Y for each increase in X

e^b is the **rate ratio**

Why the ratio?

$b = \log(\text{rate}_{x+1}) - \log(\text{rate}_x) = \log(\text{rate}_{x+1} / \text{rate}_x)$

therefore, $e^b = \text{rate}_{x+1} / \text{rate}_x$

Let’s try an lm first

Run the model:

```
alm <- lm(num_awards~prog+math, data=awards)
```

R^2 = 0.277

Coefficients:

- Students in an academic program have 0.48 more awards than students in a general program
- For each 1pt increase in math score, the number of awards increases with 0.048

Let's try an lm first

Residuals:

```
awards$lmresid <- rstandard(alm)
awards$lmresid.large <- (awards$lmresid > 1.96 |
awards$lmresid < -1.96)
awards[awards$lmresid.large,]
```

Some residuals are huge (> 3.29)

Let’s try a glm

Run the model:

aglm <- glm(num_awards~prog+math, data=awards, family=poisson)

R-square:

R^2_{hl}: (aglm$null.deviance–aglm$deviance) / aglm$null.deviance

R^2_{cs}: 1-exp((aglm$deviance–aglm$null.deviance)/200)

R^2_{n}: Rcs / (1-exp(-aglm$null.deviance/200))

Better model fit!

Let's try a glm

Coefficients: exp(aglm$coefficients)

- Students in an academic program have 2.96 times as many awards than students in a general program
- For each 1pt increase in math score, the number of awards increases with 7.27%

Do the same thing for confidence intervals:

- exp(confint(aglm))
- Note: these are not based on the Wald statistic!
- Significant when they do not cross 1

Let’s try a glm

Residuals:

```
awards$glmresid <- rstandard(aglm)
awards$glmresid.large <- (awards$glmresid > 1.96 |
awards$glmresid < -1.96)
awards[awards$glmresid.large,]
```

No huge residuals!

Ordered logistic

Also not in the book!

Ordered logistic

Dataset “consequences.dat”

Consideration of future consequences questionnaire

Variables:

age, gender: participant’s age and gender

Q3: answer to the question “I only act to satisfy immediate concerns, figuring the future will take care of itself.”

answer categories: 1=extremely uncharacteristic, 2=somewhat uncharacteristic, 3=uncertain, 4=somewhat characteristic, 5=extremely characteristic

A problem...

This is ordinal, not interval!

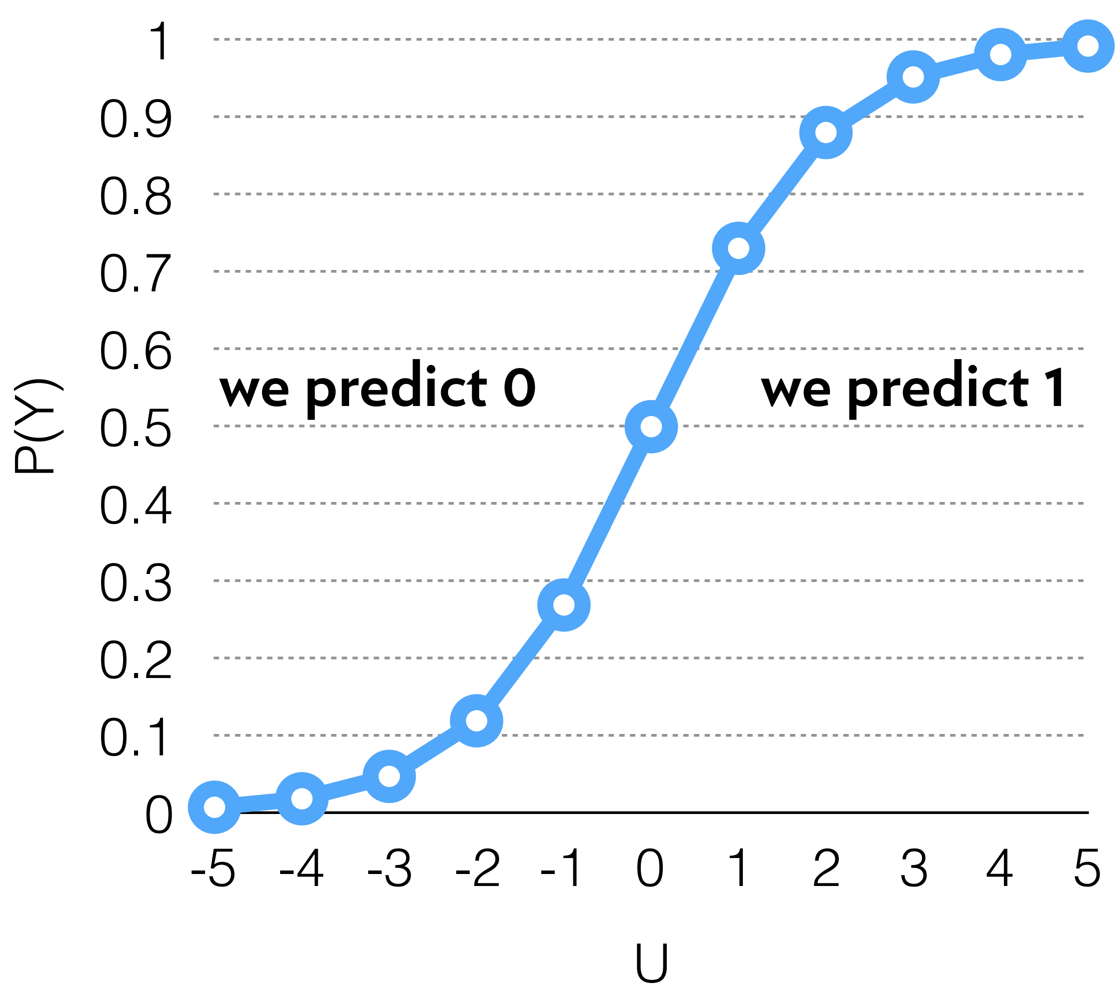
- Is the difference between "extremely uncharacteristic" and "somewhat uncharacteristic" the same as the difference between "uncertain" and "somewhat characteristic"?

Also, not very normally distributed!

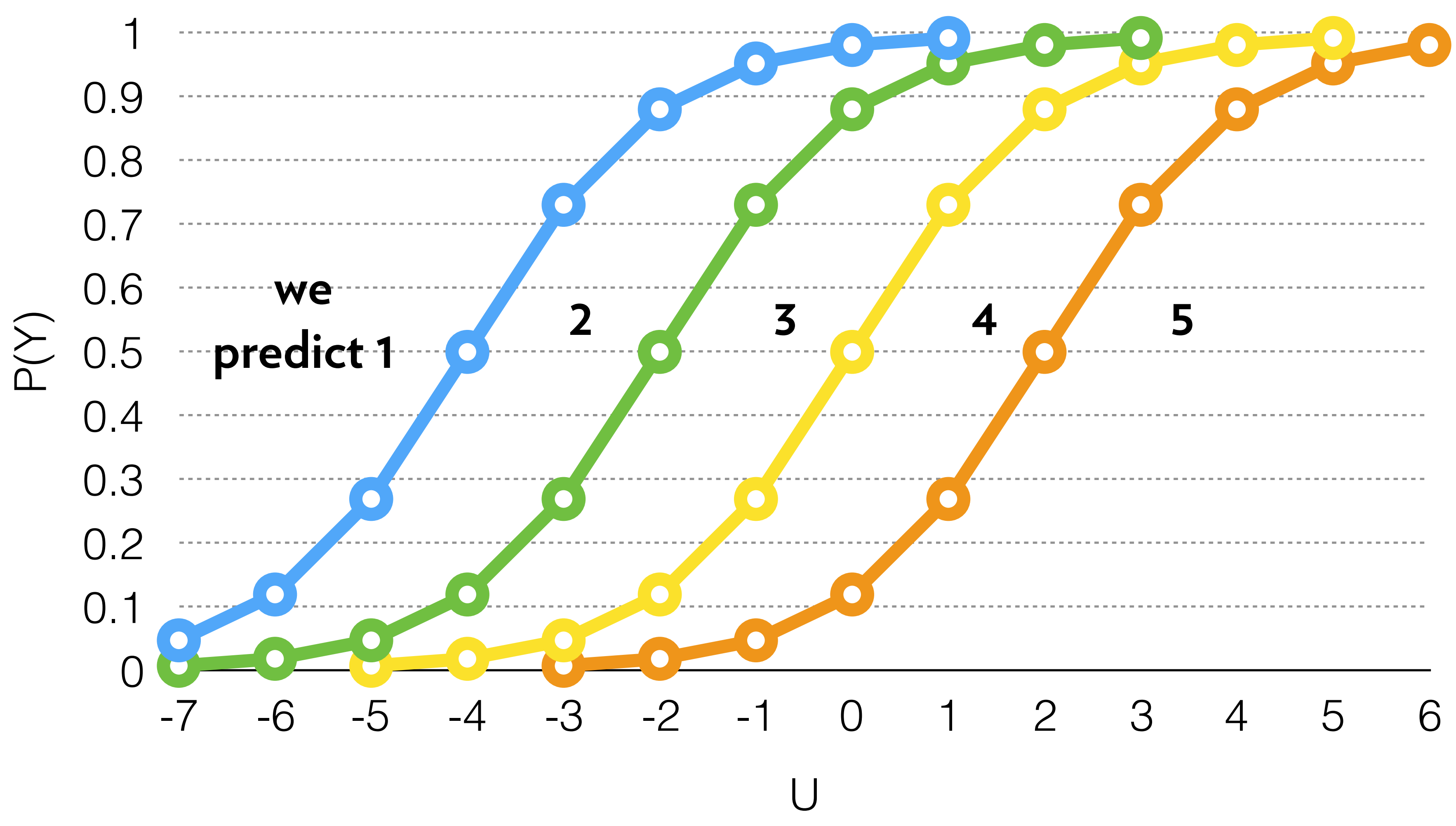
- ggplot(consequences,aes(Q3))+stat_bin(binwidth=1)

How can we solve these problems?

Logistic regression



Ordered logistic



Coefficients

The model estimates intercepts for each threshold

1|2, 2|3, 3|4, 4|5

These thresholds are the **log odds** of any person having **at least** this value

How to interpret the b coefficients?

e^b is the **odds ratio** for a 1pt increase in X

e.g. if the odds ratio is 1.40, then the odds of a higher value by 40% if X is 1 higher

Let's try an lm first

Run the model:

```
clm <- lm(Q3~gender+age, data=consequences)
```

R^2 = 0.030

Coefficients:

- Females score higher on satisfying immediate concerns only (not significant)
- Older individuals score lower on satisfying immediate concerns only (not significant)

Let’s try a polr

Run the model:

```
cplm <- polr(factor(Q3)~gender+age, data=consequences, Hess=T)
```

Run a null model:

```
cplm.null <- polr(factor(Q3)~1,data=consequences, Hess=T)
```

Let's try a polr

R-square:

- R^2_{hl}: (cplm.null$deviance−cplm$deviance) / cplm.null$deviance
- R^2_{cs}: 1-exp((cplm$deviance−cplm.null$deviance)/199)
- R^2_{n}: Rcs / (1-exp(-cplm.null$deviance/199))

The latter two suggest a better model fit!

Let's try a polr

Coefficients: exp(cplm$coefficients)

- Females have a 70% higher likelihood to rate higher on satisfying immediate concerns
- For each 1 year increase in age, the likelihood to rate higher on satisfying immediate concerns decreases by 2.09%

Do the same thing for confidence intervals:

- exp(confint(cplm))
- Note: these are not based on the Wald statistic!
- Significant when they do not cross 1

Robust methods

Also not in the book!

Robust methods

Bootstrapping works the same as linear regression

Alternative method: sandwich estimator of SE (package: "sandwich") — this also works for regular lm!

```
cov.aglm <- vcovHC(aglm, type="HC0")
std.err <- sqrt(diag(cov.aglm))
pval <- 2 * pnorm(abs(coef(aglm)/std.err), lower.tail=F)
LL <- coef(aglm) - 1.96 * std.err
UL <- coef(aglm) + 1.96 * std.err
```

"It is the mark of a truly intelligent person to be moved by statistics."

T H A N K S !

George Bernard Shaw