Math 455 Class 29 April 7

Last time: How to interpret regression coefficients in models:

$$\log(Y) \sim \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

Note: this is very common: another way to write it is:

$$Y \sim \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon)$$

This is special to $\log(Y)$ models.

⚠ The interpretation that follows will not work if our model is $\sqrt{Y} \sim \beta_0 + \beta_1 X_1 + \cdots + \varepsilon$

Or, worse still $Y^{5.371} \sim \beta_0 + \cdots + \varepsilon$

Once we estimate $\hat{\beta}_0, \hat{\beta}_1$ etc. we have

$$\hat{Y} = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p).$$

This gives an interpretation!

$$\hat{Y} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots (\text{remove } \hat{\beta}_i X_i) + \cdots + \hat{\beta}_p X_p\right) \cdot \exp\left(\hat{\beta}_i X_i\right)$$

Thus the interpretation: if X_i increases by 1 then $\hat{Y}$ changes by a factor of $e^{\hat{\beta}_i}$

Because if $(X_i)_{new} = (X_i)_{old} + 1$

$$\exp\left(\hat{\beta}_i (X_i)_{new}\right) = \exp\left(\hat{\beta}_i \left[(X_i)_{old} + 1\right]\right)$$

$$= \exp\left(\hat{\beta}_i (X_i)_{old}\right) \exp\left(\hat{\beta}_i\right)$$

BUT: usually people write down a different interpretation.

Calculus Review: for small x $\quad e^x \approx 1 + x$

Small means x^2 is so small we can ignore it in our context.

In many cases this allows for a "percentage change" interpretation.

This requires that the "small change" be a plausible level of change in X_i to consider.

If this is true:

$$\exp(\hat{\beta}_i X_i) \approx 1 + \hat{\beta}_i X_i$$

so $100 \cdot \hat{\beta}_i X_i$ is the "percentage change" and $100 \cdot \hat{\beta}_i$ is the percentage change in $\hat{Y}$ for a unit change in X_i.

Standard problem: You see the regression output and then write a sentence interpreting the model.

Next: Chapter 14: "Categorical Variables"

R will give you regression output <u>even if</u> the variables do NOT have numerical values.

How does it do this? There are schemes for replacing the "categorical" with numerical. <u>Interpretation depends on the scheme.</u>

Today: Chapter 14 : Categorical Variables.

Categorical Variables ~~are always~~ Chapter is about translating non-numerical variables (that have values like "red", "yellow", "blue") into numerical variables.

The basic method: "dummy variables"

if we have 3 possible values R, ~~Y~~ W, B.

(W = yellow so as not to confuse with response Y)

Explanatory variable X:

X	D_1	D_2
R	0	1
R	0	1
B	0	0
W	1	0
R	0	1

Introduce 2 "dummy variables" that have values 0 or 1, because X has 3 possible

"levels": R, W, B

Scheme for these variables: arbitrary choice of "Reference Level", let's say B.

Introduce D_1, D_2 where

$$D_1 = \begin{cases} 0 & \text{if color} = R \text{ or } B \\ 1 & \text{if color} = W \end{cases}$$

$$D_2 = \begin{cases} 0 & \text{if color} = B \text{ or } W \\ 1 & \text{if Color} = R \end{cases}$$

Note that if we know D_1 & D_2, we know X, and vice versa.

So D_1 & D_2 contain exactly the same info as X, but D_1, D_2 can be used in a regression.

Q: What if we had 4 levels. (add "green")?

A: Then we need 3 dummy variables.

Q: What if we had another, totally different variable with non-numerical values, like "Toyota", "Ford"?

A: Then we need a separate set of dummy variables for each categorical variable we wish to include.

Q: Wait, that's a lot of variables! And a lot of parameters to estimate.

If we have (say) 7 categorical variables, each of which can take on 5 levels, we need 28 dummy variables, $28 = (5-1) \cdot 7$.

Q: What if we want to include interactions between these?

A: We'd have $\binom{28}{2} = \frac{28 \cdot 27}{2}$ = a lot of interaction terms.

Q: What if we don't have that much data?

A: Choose a model that you have enough data to fit!

Q: Creating these dummy variables and choosing reference levels sounds like a lot of work!

A: R does this automatically!

Problem 12. (3+ 4 + 3 marks) In Problem 3, we expressed history of alcohol use by using the following coding.

Alcohol Use	X_7	X_8
None	0	0
Moderate	1	0
Severe	0	1

(a) Suppose I just build a model of y on X_7 and X_8, then I get the following `lm` output. Interpret the coefficients of X_7 and X_8 in the following output.

```
> lmodInd = lm(y~x7+x8, data = SurgicalUnit)
> summary(lmodInd)

Call:
lm(formula = y ~ x7 + x8, data = SurgicalUnit)

Residuals:
    Min      1Q  Median      3Q     Max
-528.30 -229.81  -43.06  109.82 1296.70

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   599.80      94.99   6.315 6.57e-08 ***
x7             36.51     117.00   0.312  0.75628
x8            446.50     150.19   2.973  0.00449 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 367.9 on 51 degrees of freedom
Multiple R-squared:  0.1753,    Adjusted R-squared:  0.143
F-statistic: 5.421 on 2 and 51 DF,  p-value: 0.007336
```

(b) Now, suppose we express history of alcohol use by using the following coding.

Alcohol Use	X_7	X_8
None	-1	-1
Moderate	1	0
Severe	0	1

Now I build a second model of y on X_7 and X_8, then I get the following `lm` output. Interpret the coefficients of X_7 and X_8 in the following output.

```
> summary(lmodSum)

Call:
lm(formula = y ~ x7 + x8, data = SurgicalUnit1)

Residuals:
    Min      1Q  Median      3Q     Max
-528.30 -229.81  -43.06  109.82 1296.70

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   760.80      55.00  13.833  < 2e-16 ***
x7           -124.49      67.68  -1.839  0.07167 .
x8            285.50      86.81   3.289  0.00183 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 367.9 on 51 degrees of freedom
Multiple R-squared:  0.1753,    Adjusted R-squared:  0.143
F-statistic: 5.421 on 2 and 51 DF,  p-value: 0.007336
```

(c) What do you conclude about the effect of alcohol use on survival time? Does your conclusion change with change in codings at 5% level of significance?

Q: How do we interpret the estimated regression coefficient in the output?

A: Let's look at a past test question (see next page)

$$Y = \beta_0 + \beta_7 X_7 + \beta_8 X_8 + \varepsilon$$

so if patient has "none" history of alcohol use,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_7 \cdot 0 + \hat{\beta}_8 \cdot 0 = \hat{\beta}_0.$$

Thus $\hat{\beta}_0$ is the estimated (fitted) value of Y when the patient has "none" history of alcohol use.

What is the interpretation of $\hat{\beta}_7$?

* If the patient has "moderate" history,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_7 \cdot 1 + \hat{\beta}_8 \cdot 0 = \hat{\beta}_0 + \hat{\beta}_7.$$

Interpretation: $\hat{\beta}_7$ is the predicted (fitted) difference in response between a patient with "moderate"

history and a patient with "none" history.

Similarly: $\hat{\beta}_8$ is the predicted diff in $\hat{Y}$ between a patient with "severe" and "none".

ICBST: $\hat{\beta}_0$ is the average value of Y for patients with "none" history..

$\hat{\beta}_0 + \hat{\beta}_7$ is the avg value of Y for patients with "moderate" history.

$\hat{\beta}_0 + \hat{\beta}_8$ is the avg value of Y for patients with "severe" history

---

Alternate Codings "Non-Dummy Codings"

What do we require of the assignment of values to the dummy variables (X_7 and X_8 in the test question)?

Only that the info in X_7 & X_8 is exactly the info in the categorical variable.

In part (b), and alternate coding is given.

Many alternate codings are possible.

In part (b) we see "sum coding" (see ch. 14).

The coding changes the interpretation of the regression coefficients.

The predicted (fitted) values $\hat{Y}$ do NOT change.

The new interpretation: the value of $\hat{\beta}_7$ is the difference in Y between the average for patients with "moderate" and the "overall average" or "grand mean". (for all patients, not weighted by size of patient groups)

The value of $\hat{\beta}_8$ is the difference in average Y between patients with "severe" and this "grand mean".

Q: What about patients with "none"?

A: The difference in average Y between

patients with "none" and the "grand mean" is $-\hat{\beta}_7 + -\hat{\beta}_8$.

Q: Why don't we introduce a dummy var for every level and avoid this ugly $-\hat{\beta}_7 - \hat{\beta}_8$?

A: If we do this, we have that the sum of the dummy variables is the constant 1.

This means that we have collinearity, which breaks our regression.

Problem 12:

Notice that the fitted values are the same with both codings

In (a)

if "None" is the level of alcohol use:

$$\hat{Y} = 599.80 + 0 \cdot (36.51) + 0 \cdot (446.50)$$

$$= 599.80.$$

if "Moderate", $X_7 = 1$ and $X_8 = 0$

$$\hat{Y} = 599.80 + 1 \cdot (36.51) + 0 \cdot (446.50)$$

$$= 636.51$$

if "Severe", $X_7 = 0$ and $X_8 = 1$

$$\hat{Y} = 599.80 + 0 \cdot (36.51) + 1 \cdot 446.50$$

$$= 1046.3$$

In (b):

if "None" the new X_7 and X_8 are both -1.

$$\hat{Y} = 760.80 + (-1)(-124.49) + (-1)(285.50)$$

$$= 599.79 \quad \text{(same up to rounding)}$$

if "Moderate", $X_7 = 1$, $X_8 = 0$

$$\hat{Y} = 760.80 + 1 \cdot (-124.49) + 0 \cdot (285.50)$$

$$= 636.31 \quad \text{(same again!)}$$

if "Severe", $X_7 = 0$, $X_8 = 1$

$$\hat{Y} = 760.80 + 0 \cdot (-124.49) + 1 \cdot (285.50)$$

$$= 1046.3 \quad \text{(same!)}$$

Changing the coding of categorical variables does not change the fitted values.

We can observe that $2 \cdot 0 - 1 = -1$

$2 \cdot 1 - 1 = 1$

Interpretation:

In (a) the intercept is the average value of survival time for patients whose alcohol use is "None".

The coefficient of X_7 is the difference in $\overset{\text{avg}}{\uparrow}$ survival time for patients whose level of alcohol use is "moderate" vs. patients whose level of alcohol use is "None".

The coefficient of X_8 is the difference in avg survival time between patients whose alcohol use is "severe" and patients whose alcohol use is "None".

Interpretation in (b):

The intercept is the overall average survival time.

The coefficient of X_7 is the difference in avg survival time between a patient with "Moderate" alcohol use and the overall average.

Q: Why is this coeff. negative in (b) and positive in (a)?

A: We are comparing the "Moderate" group to a different population. In (a) we compare to the "None" group and in (b) we compare to everyone. "Everyone" includes the "Severe" group who on average survived much longer.

Finally, the coeff of X_8 is the difference in avg. survival time between the "Severe" group and the average of everyone.