

Multiple Linear Regression: Categorical Predictors

Author: Nicholas G Reich, Jeff Goldsmith

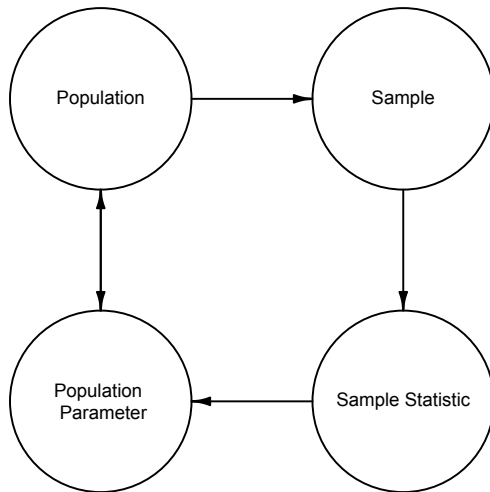
*This material is part of the **statsTeachR** project*

*Made available under the Creative Commons Attribution-ShareAlike 3.0 Unported
License: http://creativecommons.org/licenses/by-sa/3.0/deed.en_US*

Today's Lecture

- Sampling distribution of $\hat{\beta}$
- Confidence intervals
- Hypothesis tests for individual coefficients
- Global tests

Circle of Life



Statistical inference

- We have LSEs $\hat{\beta}_0, \hat{\beta}_1, \dots$; we want to know what this tells us about β_0, β_1, \dots .
- Two basic tools are confidence intervals and hypothesis tests
 - ▶ Confidence intervals provide a plausible range of values for the parameter of interest based on the observed data
 - ▶ Hypothesis tests ask how probable are the data we gathered under a null hypothesis about the data generating distribution

Motivation

How can we draw **inference** about each of these parameters and relationships that our model is encoding?

```
mlr1 <- lm(disease ~ airqual + crowding + nutrition + smoking, d
summary(mlr1)$coef
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	11.86333	2.578819	4.600	1.316e-05
## airqual	0.25788	0.026799	9.623	1.165e-15
## crowding	1.11113	0.102037	10.889	2.404e-18
## nutrition	-0.03278	0.007954	-4.122	8.095e-05
## smoking	4.96093	1.085292	4.571	1.475e-05

Motivation

- Can we say anything about whether the effect of airquality is “significant” after adjusting for other variables?
- Can we say whether adding airquality improves the fit of our model?
- Can we compare this model to a model with only crowding, nutrition and smoking?

Sampling distribution

If our usual assumptions are satisfied and $\epsilon \stackrel{iid}{\sim} N[0, \sigma^2]$ then

$$\hat{\beta} \sim N \left[\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \right].$$

$$\hat{\beta}_j \sim N \left[\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})_{jj}^{-1} \right].$$

- This will be used later for inference.
- Even without Normal errors, asymptotic Normality of LSEs is possible under reasonable assumptions.

Sampling distribution

For real data we have to estimate σ^2 as well as β .

- Recall our estimate of the error variance is

$$\hat{\sigma}^2 = \frac{RSS}{n - p - 1} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n - p - 1}$$

- With Normally distributed errors, it can be shown that

$$(n - p - 1) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p-1}^2$$

Testing procedure

Calculate the probability of the observed data (or more extreme data) under a null hypothesis.

- Often $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$
- Set type I error rate
 $\alpha = P(\text{falsely rejecting a true null hypothesis})$
- Calculate a test statistic assuming the null hypothesis is true
- Compute a p-value =

$$P(\text{As or more extreme test statistic} | H_0)$$

- Reject or fail to reject H_0

Individual coefficients

For individual coefficients

- We can use the test statistic

$$T = \frac{\hat{\beta}_j - \beta_j}{\widehat{se}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2(\mathbf{X}^T \mathbf{X})_{jj}^{-1}}} \sim t_{n-p-1}$$

- For a two-sided test of size α , we reject if

$$|T| > t_{1-\alpha/2, n-p-1}$$

- The p-value gives $P(t_{n-p-1} > T_{obs} | H_0)$

Note that t is a symmetric distribution that converges to a Normal as $n - p - 1$ increases.

Back to the example

```
summary(mlr1)

##
## Call:
## lm(formula = disease ~ airqual + crowding + nutrition + smoking,
##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.130 -2.183 -0.572  1.941 13.326
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.86333    2.57882   4.60 1.3e-05 ***
## airqual      0.25788    0.02680   9.62 1.2e-15 ***
## crowding     1.11113    0.10204  10.89 < 2e-16 ***
## nutrition   -0.03278    0.00795  -4.12 8.1e-05 ***
## smoking      4.96093    1.08529   4.57 1.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.64 on 94 degrees of freedom
## Multiple R-squared:  0.866, Adjusted R-squared:  0.861
## F-statistic: 152 on 4 and 94 DF, p-value: <2e-16
```

Individual coefficients: CIs

Alternatively, we can construct a confidence interval for β_j

- A confidence interval with coverage $(1 - \alpha)$ is given by

$$\beta_j \pm t_{1-\alpha/2, n-p-1} \widehat{se}(\hat{\beta}_j)$$

- Assuming all the standard assumptions hold,

$$(1 - \alpha) = P(LB < \beta_j < UB)$$

Back to the example

```
cbind(coef(mlr1), confint(mlr1))
```

##		2.5 %	97.5 %
## (Intercept)	11.86333	6.74303	16.98364
## airqual	0.25788	0.20467	0.31109
## crowding	1.11113	0.90853	1.31372
## nutrition	-0.03278	-0.04858	-0.01699
## smoking	4.96093	2.80606	7.11580

Inference for linear combinations

Sometimes we are interested in making claims about $c^T \beta$ for some c .

- Define $H_0 : c^T \beta = c^T \beta_0$ or $H_0 : c^T \beta = 0$
- We can use the test statistic

$$T = \frac{c^T \hat{\beta} - c^T \beta}{\widehat{se}(c^T \hat{\beta})} = \frac{c^T \hat{\beta} - c^T \beta}{\sqrt{\hat{\sigma}^2 c^T (\mathbf{X}^T \mathbf{X})^{-1} c}}$$

- This test statistic is asymptotically Normally distributed
- For a two-sided test of size α , we reject if

$$|T| > z_{1-\alpha/2}$$

Inference about multiple coefficients

Our model contains multiple parameters; often we want to perform multiple tests:

$$H_{01} : \beta_1 = 0$$

$$H_{02} : \beta_2 = 0$$

$$\vdots = \vdots$$

$$H_{0k} : \beta_k = 0$$

where each test has a size of α

- For any individual test, $P(\text{reject } H_{0i} | H_{0i}) = \alpha$

Inference about multiple coefficients

What about

$$P(\text{reject at least one } H_{0i} | \text{all } H_{0i} \text{ are true}) = \alpha$$

Family-wise error rate

To calculate the FWER

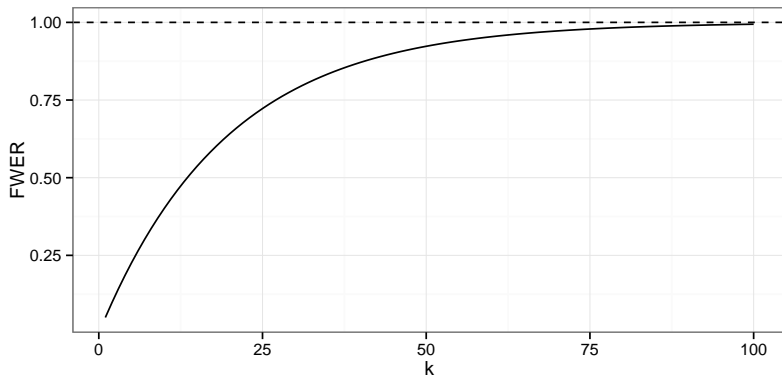
- First note $P(\text{no rejections} | \text{all } H_{0i} \text{ are true}) = (1 - \alpha)^k$
- It follows that

$$\begin{aligned}\text{FWER} &= P(\text{at least one rejection} | \text{all } H_{0i} \text{ are true}) \\ &= 1 - (1 - \alpha)^k\end{aligned}$$

Family-wise error rate

$$\text{FWER} = 1 - (1 - \alpha)^k$$

```
alpha <- 0.05  
k <- 1:100  
FWER <- 1 - (1 - alpha)^k  
qplot(k, FWER, geom = "line") + geom_hline(yintercept = 1, lty = 2)
```



Addressing multiple comparisons

Three general approaches

- Do nothing in a reasonable way
 - ▶ Don't trust scientifically implausible results
 - ▶ Don't over-emphasize isolated findings
- Correct for multiple comparisons
 - ▶ Often, use the Bonferroni correction and use $\alpha_i = \alpha/k$ for each test
 - ▶ Thanks to the Bonferroni inequality, this gives an overall $FWER \leq \alpha$
- Use a global test

Global tests

Compare a smaller “null” model to a larger “alternative” model

- Smaller model must be nested in the larger model
- That is, the smaller model must be a special case of the larger model
- For both models, the RSS gives a general idea about how well the model is fitting
- In particular, something like

$$\frac{RSS_S - RSS_L}{RSS_L}$$

compares the relative RSS of the models

Nested models

- These models are nested:

Smaller = Regression of Y on X_1

Larger = Regression of Y on X_1, X_2, X_3, X_4

- These models are not:

Smaller = Regression of Y on X_2

Larger = Regression of Y on X_1, X_3

Global F tests

- Compute the test statistic

$$F_{obs} = \frac{(RSS_S - RSS_L)/(df_S - df_L)}{RSS_L/df_L}$$

- If H_0 (the null model) is true, then $F_{obs} \sim F_{df_S - df_L, df_L}$
- Note $df_S = n - p_S - 1$ and $df_L = n - p_L - 1$
- We reject the null hypothesis if the p-value is above α , where

$$\text{p-value} = P(F_{df_S - df_L, df_L} > F_{obs})$$

Global F tests

There are a couple of important special cases for the F test

- The null model contains the intercept only
 - ▶ When people say ANOVA, this is often what they mean (although all F tests are based on an analysis of variance)
- The null model and the alternative model differ only by one term
 - ▶ Gives a way of testing for a single coefficient
 - ▶ Turns out to be equivalent to a two-sided t -test: $t_{df_L}^2 \sim F_{1,df_L}$

Lung data: multiple coefficients simultaneously

You can test multiple coefficients simultaneously using the F test

```
mlr_null <- lm(disease ~ nutrition, data = dat)
mlr1 <- lm(disease ~ nutrition + airqual + crowding + smoking, data = dat)
anova(mlr_null, mlr1)

## Analysis of Variance Table
##
## Model 1: disease ~ nutrition
## Model 2: disease ~ nutrition + airqual + crowding + smoking
##   Res.Df  RSS Df Sum of Sq   F Pr(>F)
## 1      97 9193
## 2      94 1248  3      7945 199 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Lung data: single coefficient test

The F test is equivalent to the t test when there's only one parameter of interest

```
mlr_null <- lm(disease ~ nutrition, data = dat)
mlr1 <- lm(disease ~ nutrition + airqual, data = dat)
anova(mlr_null, mlr1)

## Analysis of Variance Table
##
## Model 1: disease ~ nutrition
## Model 2: disease ~ nutrition + airqual
##   Res.Df  RSS Df Sum of Sq   F Pr(>F)
## 1      97 9193
## 2      96 5970  1      3223 51.8 1.3e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(mlr1)$coef

##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   37.6254    2.43946   15.42 9.946e-28
## nutrition     -0.0347    0.01692   -2.05 4.307e-02
## airqual        0.3611    0.05016    7.20 1.347e-10
```

Today's Big Ideas

- Inference for multiple linear regression models