MLR: Accuracy and Extensions

Nate Wells

Math 243: Stat Learning

September 14th, 2020

Outline

In today's class, we will. . .

- Quantify model accuracy for linear regression models (both simple and multiple)
- Generalize to include categorical variables and non-linear terms

Section 1

Assessing Model Accuracy

How Strong is a Linear Model?

In an linear model model,

$$Y = f(X) + \epsilon$$

So even if we could perfectly predict f using $\hat{f}$, our model would still have non-zero MSE.

How Strong is a Linear Model?

In an linear model model,

$$Y = f(X) + \epsilon$$

So even if we could perfectly predict f using $\hat{f}$, our model would still have non-zero MSE.

The **Residual Standard Error** (RSE) measures the average size of deviations of the response from the linear regression line, is given by

$$\text{RSE} = \sqrt{\frac{1}{n-1-p}\text{RSS}} = \sqrt{\frac{1}{n-1-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

How Strong is a Linear Model?

In an linear model model,

$$Y = f(X) + \epsilon$$

So even if we could perfectly predict f using $\hat{f}$, our model would still have non-zero MSE.

The **Residual Standard Error** (RSE) measures the average size of deviations of the response from the linear regression line, is given by

$$\mathrm{RSE} = \sqrt{\frac{1}{n-1-p}\mathrm{RSS}} = \sqrt{\frac{1}{n-1-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

It has the property that

$$E(\mathrm{RSE}^2) \approx \mathrm{Var}(\epsilon)$$

Poll 1

Which of the following are most likely to decrease as more and more predictors are added to a linear model (select all that apply)?

a. test MSE
b. training MSE
c. RSS
d. RSE
e. $\mathrm{Var}(\epsilon)$

The R^2 statistic

Large RSE indicates poor model fit, while small RSE indicates good fit. But how do we determine how small is **small**?

The R^2 statistic

Large RSE indicates poor model fit, while small RSE indicates good fit. But how do we determine how small is **small**?

- The answer depends on the units of Y

The R^2 statistic

Large RSE indicates poor model fit, while small RSE indicates good fit. But how do we determine how small is **small**?

- The answer depends on the units of Y

An alternative, standardized measure of goodness of fit is the R^2 statistic:

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \qquad \text{where } \mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The R^2 statistic

Large RSE indicates poor model fit, while small RSE indicates good fit. But how do we determine how small is **small**?

- The answer depends on the units of Y

An alternative, standardized measure of goodness of fit is the R^2 statistic:

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \qquad \text{where } \mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

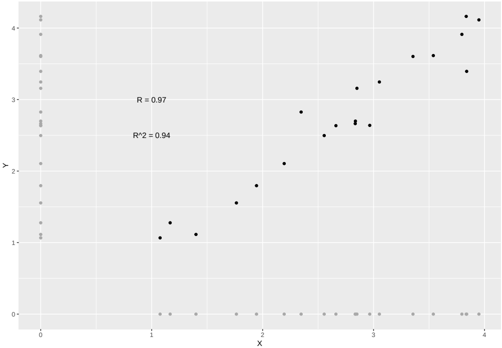
- The value of R^2 is always between 0 and 1, and represents the percentage of variability in values of the response just due to variability in the predictors.

Values of R^2

If $R^2 \approx 1$: nearly all the variability in response is due to variability in the predictor variable.

Values of R^2

If $R^2 \approx 1$: nearly all the variability in response is due to variability in the predictor variable.

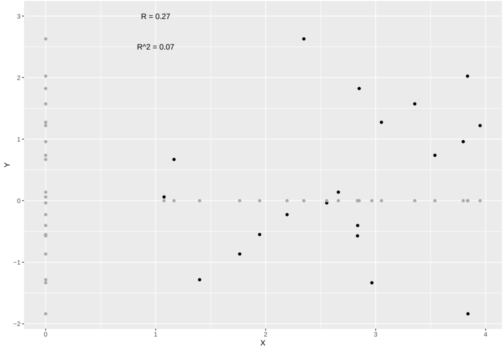


Values of R^2

If $R^2 \approx 0$: almost none of the variability in response is due to variability in the predictor variable.

Values of R^2

If $R^2 \approx 0$: almost none of the variability in response is due to variability in the predictor variable.



Formulas for R^2 in terms of correlation

For SLR,

$$R^2 = [\mathrm{Cor}(X, Y)]^2 = \left[\frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}\right]^2 = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}\right]^2$$

Formulas for R^2 in terms of correlation

For SLR,

$$R^2 = [\mathrm{Cor}(X, Y)]^2 = \left[\frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}\right]^2 = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}\right]^2$$

For MLR,

$$R^2 = \left[\mathrm{Cor}(Y, \hat{Y})\right]^2$$

Formulas for R^2 in terms of correlation

For SLR,

$$R^2 = [\mathrm{Cor}(X, Y)]^2 = \left[\frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}\right]^2 = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}\right]^2$$

For MLR,

$$R^2 = \left[\mathrm{Cor}(Y, \hat{Y})\right]^2$$

We will usually use software to compute R^2.

Model Accuracy in R

```
mod_credit<-lm(Balance ~ Income + Limit , data = Credit)

summary(mod_credit)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit, data = Credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -232.79 -115.45  -48.20   53.36  549.77
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -385.17926   19.46480  -19.79   <2e-16 ***
## Income        -7.66332    0.38507  -19.90   <2e-16 ***
## Limit          0.26432    0.00588   44.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 165.5 on 397 degrees of freedom
## Multiple R-squared:  0.8711, Adjusted R-squared:  0.8705
## F-statistic:  1342 on 2 and 397 DF,  p-value: < 2.2e-16
```

Model Accuracy in R

```r
mod_credit<-lm(Balance ~ Income + Limit , data = Credit)

summary(mod_credit)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit, data = Credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -232.79 -115.45  -48.20   53.36  549.77
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -385.17926   19.46480  -19.79   <2e-16 ***
## Income        -7.66332    0.38507  -19.90   <2e-16 ***
## Limit          0.26432    0.00588   44.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 165.5 on 397 degrees of freedom
## Multiple R-squared:  0.8711, Adjusted R-squared:  0.8705
## F-statistic:  1342 on 2 and 397 DF,  p-value: < 2.2e-16
```

We can use `summary(mod)$r.sq` or `summary(mod)$sigma` to access R^2 and RSE directly.

Adjusted R^2

- It turns out that the samples's R^2 gives a **biased** estimate of the variability in the *population* explained by the model.

Adjusted R^2

- It turns out that the samples's R^2 gives a **biased** estimate of the variability in the *population* explained by the model.
- Instead, we use the adjusted R:

$$R^2_{\text{adjusted}} = 1 - \frac{\text{RSS}}{\text{TSS}} \frac{n-1}{n-p-1}$$

Adjusted R^2

- It turns out that the samples's R^2 gives a **biased** estimate of the variability in the *population* explained by the model.
- Instead, we use the adjusted R:

$$R^2_{\text{adjusted}} = 1 - \frac{\text{RSS}}{\text{TSS}} \frac{n-1}{n-p-1}$$

- This adjusted R^2 is usually a bit smaller than R^2, and the difference decreases as n gets large.

Testing Significance

Suppose we wish to test whether at least one predictor has a significant linear relationship with the response.

Testing Significance

Suppose we wish to test whether at least one predictor has a significant linear relationship with the response.

Why would it be incorrect to conduct p many significant tests comparing each predictor to the response?

The Hypothesis Test

Goal: test whether any predictors are significant.

The Hypothesis Test

Goal: test whether any predictors are significant.

Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad H_a : \text{ at least one of } \beta_i \neq 0$$

The Hypothesis Test

Goal: test whether any predictors are significant.

Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad H_a : \text{ at least one of } \beta_i \neq 0$$

Test statistic:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n - p - 1)}$$

The Hypothesis Test

Goal: test whether any predictors are significant.

Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad H_a : \text{ at least one of } \beta_i \neq 0$$

Test statistic:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}$$

Under the null hypothesis, F is approximately F-distributed with $p, n-p-1$ parameters.

The Hypothesis Test

Goal: test whether any predictors are significant.

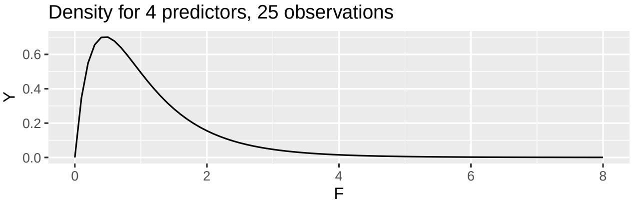
Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad H_a : \text{ at least one of } \beta_i \neq 0$$

Test statistic:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n - p - 1)}$$

Under the null hypothesis, F is approximately F-distributed with $p, n - p - 1$ parameters.



Typical Values of the F statistic

Provided conditions for linear regression are met,

$$E\left[\frac{\mathrm{RSS}}{n-p-1}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

Typical Values of the F statistic

Provided conditions for linear regression are met,

$$E\left[\frac{\mathrm{RSS}}{n-p-1}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

And if H_0 is also true, then

$$E\left[\frac{\mathrm{TSS}-\mathrm{RSS}}{p}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

Typical Values of the F statistic

Provided conditions for linear regression are met,

$$E\left[\frac{\mathrm{RSS}}{n-p-1}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

And if H_0 is also true, then

$$E\left[\frac{\mathrm{TSS}-\mathrm{RSS}}{p}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

Hence, if there is truly no relationship between any of the predictors and the response, then on average,

$$F = \frac{(\mathrm{TSS}-\mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} = 1$$

Typical Values of the F statistic

Provided conditions for linear regression are met,

$$E\left[\frac{\mathrm{RSS}}{n-p-1}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

And if H_0 is also true, then

$$E\left[\frac{\mathrm{TSS}-\mathrm{RSS}}{p}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

Hence, if there is truly no relationship between any of the predictors and the response, then on average,

$$F = \frac{(\mathrm{TSS}-\mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} = 1$$

Moreover, it is unlikely that F is drastically larger than 1.

Poll 2: TSS and RSS

Suppose we have a linear model with 25 observations and 4 predictors. Which of the following provides the best evidence of a relationship between the response and at least 1 of the predictors?

a. $\text{TSS} = 64$, $\text{RSS} = 4$
b. $\text{TSS} = 4$, $\text{RSS} = 16$
c. $\text{TSS} = 48$, $\text{RSS} = 8$
d. $\text{TSS} = 4$, $\text{RSS} = 4$

Improving Model Accuracy

What do we do when model accuracy is low (either high RSE or low R^2)?

Improving Model Accuracy

What do we do when model accuracy is low (either high RSE or low R^2)?

- If some variables are strongly correlated, remove some redundant ones.

Improving Model Accuracy

What do we do when model accuracy is low (either high RSE or low R^2)?

- If some variables are strongly correlated, remove some redundant ones.
  - This process is known as *backwards elimination*.
  - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until accuracy ceases to improve.

Improving Model Accuracy

What do we do when model accuracy is low (either high RSE or low R^2)?

- If some variables are strongly correlated, remove some redundant ones.
  - This process is known as *backwards elimination*.
  - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until accuracy ceases to improve.
- If ϵ is too large, add further variables.

Improving Model Accuracy

What do we do when model accuracy is low (either high RSE or low R^2)?

- If some variables are strongly correlated, remove some redundant ones.
  - This process is known as *backwards elimination*.
  - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until accuracy ceases to improve.
- If ϵ is too large, add further variables.
  - This process is known as *forward selection*.
  - Start with the null model, create p many SLR models (one for each predictor), and select the one with best accuracy. Repeat with this new model, creating $p-1$ two predictor models (one for each remaining predictor). Continue until accuracy ceases to improve.

Improving Model Accuracy

What do we do when model accuracy is low (either high RSE or low R^2)?

- If some variables are strongly correlated, remove some redundant ones.
  - This process is known as *backwards elimination*.
  - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until accuracy ceases to improve.
- If ϵ is too large, add further variables.
  - This process is known as *forward selection*.
  - Start with the null model, create p many SLR models (one for each predictor), and select the one with best accuracy. Repeat with this new model, creating $p-1$ two predictor models (one for each remaining predictor). Continue until accuracy ceases to improve.
- Is it possible that none of these models will have the best possible accuracy among all subsets of predictors?

Improving Model Accuracy

What do we do when model accuracy is low (either high RSE or low R^2)?

- If some variables are strongly correlated, remove some redundant ones.
  - This process is known as *backwards elimination*.
  - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until accuracy ceases to improve.
- If ϵ is too large, add further variables.
  - This process is known as *forward selection*.
  - Start with the null model, create p many SLR models (one for each predictor), and select the one with best accuracy. Repeat with this new model, creating $p-1$ two predictor models (one for each remaining predictor). Continue until accuracy ceases to improve.
- Is it possible that none of these models will have the best possible accuracy among all subsets of predictors?
  - Yes. But we'll cover detailed model selection in Chapter 6.

Section 2

Extending the Linear Model

Qualitative Predictors

Thus far, we have assumed all predictors are quantitative (taking values on a scale).

Qualitative Predictors

Thus far, we have assumed all predictors are quantitative (taking values on a scale).

- It would nice to include qualitative or categorical predictors in our model.

Qualitative Predictors

Thus far, we have assumed all predictors are quantitative (taking values on a scale).

- It would nice to include qualitative or categorical predictors in our model.
- But if we try to include them naively, we immediately run into trouble:

Qualitative Predictors

Thus far, we have assumed all predictors are quantitative (taking values on a scale).

- It would nice to include qualitative or categorical predictors in our model.
- But if we try to include them naively, we immediately run into trouble:

$$\hat{\text{Debt}} = f(X_1, X_2, X_3) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Income} + \hat{\beta}_2 \cdot \text{Limit} + \hat{\beta}_3 \cdot \text{Gender}$$

Qualitative Predictors

Thus far, we have assumed all predictors are quantitative (taking values on a scale).

- It would nice to include qualitative or categorical predictors in our model.
- But if we try to include them naively, we immediately run into trouble:

$$\hat{\text{Debt}} = f(X_1, X_2, X_3) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Income} + \hat{\beta}_2 \cdot \text{Limit} + \hat{\beta}_3 \cdot \text{Gender}$$

Suppose $\hat{\beta}^T = \begin{pmatrix} -400 & -7.5 & .25 & 2.5 \end{pmatrix}$

$$\hat{\text{Debt}} = f(10, 4000, \text{Female}) = -400 - 7.5 \cdot 10 + .25 \cdot 4000 + 2.5 \cdot \text{Female} = ???$$

Coding and Dummy Variables

- For binary categorical variables, we create a new *quantitative* variable by coding the first level as 0 and the second as 1.

Coding and Dummy Variables

- For binary categorical variables, we create a new *quantitative* variable by coding the first level as 0 and the second as 1.
  - For 'Gender', we could code: $1 \leftarrow \text{Female} \quad 0 \leftarrow \text{Male}$

Coding and Dummy Variables

- For binary categorical variables, we create a new *quantitative* variable by coding the first level as 0 and the second as 1.
  - For 'Gender', we could code: $1 \leftarrow \text{Female} \quad 0 \leftarrow \text{Male}$

$$\hat{\text{Debt}} = f(7.5, 4000, \text{Female}) = -400 - 7.5 \cdot 10 + 0.25 \cdot 4000 + 2.5 \cdot 1 = 527.5$$

Coding and Dummy Variables

- For binary categorical variables, we create a new *quantitative* variable by coding the first level as 0 and the second as 1.
    - For 'Gender', we could code: $1 \leftarrow \text{Female} \quad 0 \leftarrow \text{Male}$

$$\hat{\text{Debt}} = f(7.5, 4000, \text{Female}) = -400 - 7.5 \cdot 10 + 0.25 \cdot 4000 + 2.5 \cdot 1 = 527.5$$

- In general, if X_1 is quantitative and X_2 is categorical, the resulting model will be

$$\hat{Y} = f(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = \begin{cases} (\beta_0 + \beta_2) + \beta_1 X_1, & \text{if obs. in 1st level,} \\ \beta_0 + \beta_1 X_1, & \text{if obs. in 2nd level.} \end{cases}$$

Coding and Dummy Variables

- For binary categorical variables, we create a new *quantitative* variable by coding the first level as 0 and the second as 1.
    - For 'Gender', we could code: $1 \leftarrow \text{Female} \quad 0 \leftarrow \text{Male}$

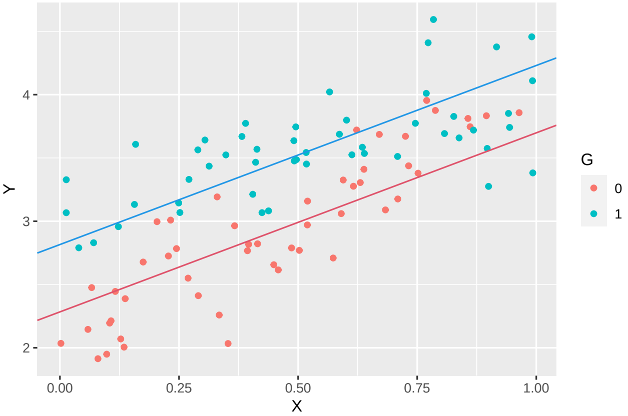
$$\hat{\text{Debt}} = f(7.5, 4000, \text{Female}) = -400 - 7.5 \cdot 10 + 0.25 \cdot 4000 + 2.5 \cdot 1 = 527.5$$

- In general, if X_1 is quantitative and X_2 is categorical, the resulting model will be

$$\hat{Y} = f(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = \begin{cases} (\beta_0 + \beta_2) + \beta_1 X_1, & \text{if obs. in 1st level,} \\ \beta_0 + \beta_1 X_1, & \text{if obs. in 2nd level.} \end{cases}$$

Note that both regression lines have the same slope, but different intercept.

Scatterplot



$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 G = 2.28 + 1.41X + 0.53G$$

The model in R

```
mod_2<- lm(data = my_data, Y ~ X + G)
summary(mod_2)
```

```
##
## Call:
## lm(formula = Y ~ X + G, data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83811 -0.22167 -0.02565  0.21738  0.66865
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.28381    0.06788  33.645  < 2e-16 ***
## X            1.41447    0.11639  12.153  < 2e-16 ***
## G1           0.53199    0.06452   8.246 8.03e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3174 on 97 degrees of freedom
## Multiple R-squared:  0.728,  Adjusted R-squared:  0.7224
## F-statistic: 129.8 on 2 and 97 DF,  p-value: < 2.2e-16
```

Poll 3: MLR Slope Interpretation

The slope on a (binary) categorical variable G tells us (select all that apply)

a. How much we expect the response to change if we increase the value of G from 0 to 1, while holding all else constant.
b. The difference in the average response between observations in the two categories.
c. The value of the response variable if G equals 0.
d. The distance between the two regression lines on the 2d scatterplot

Categorical Variables with more than 2 levels.

We extend to variables with more than 2 levels by creating binary variables for each level.

Categorical Variables with more than 2 levels.

We extend to variables with more than 2 levels by creating binary variables for each level.

In the `Credit` data set, the `Ethnicity` variable takes 3 levels: `African American`, `Asian`, `Caucasion`. (As with `Gender`, the levels here are incomplete)

Categorical Variables with more than 2 levels.

We extend to variables with more than 2 levels by creating binary variables for each level.

In the `Credit` data set, the `Ethnicity` variable takes 3 levels: `African American`, `Asian`, `Caucasion`. (As with `Gender`, the levels here are incomplete)

For categorical variable X_i with levels $j = 1, \ldots, k$, create a dummy variables x_{ij} by

$$x_{ij} = \begin{cases} 1, & \text{obs. in level } j, \\ 0, & \text{obs. not in level } j, \end{cases}$$

Categorical Variables with more than 2 levels.

We extend to variables with more than 2 levels by creating binary variables for each level.

In the `Credit` data set, the `Ethnicity` variable takes 3 levels: `African American`, `Asian`, `Caucasion`. (As with `Gender`, the levels here are incomplete)

For categorical variable X_i with levels $j = 1, \ldots, k$, create a dummy variables x_{ij} by

$$x_{ij} = \begin{cases} 1, & \text{obs. in level } j, \\ 0, & \text{obs. not in level } j, \end{cases}$$

For example,

$$\text{Eth}_{AA} = \begin{cases} 1, & \text{obs. is African American,} \\ 0, & \text{obs. is not African America} \end{cases}$$

$$\text{Eth}_{A} = \begin{cases} 1, & \text{obs. is Asian,} \\ 0, & \text{obs. is not Asian} \end{cases}$$

$$\text{Eth}_{C} = \begin{cases} 1, & \text{obs. is Caucasion,} \\ 0, & \text{obs. is not Caucasion} \end{cases}$$

Categorical Variables with more than 2 levels.

We extend to variables with more than 2 levels by creating binary variables for each level.

In the `Credit` data set, the `Ethnicity` variable takes 3 levels: `African American`, `Asian`, `Caucasion`. (As with `Gender`, the levels here are incomplete)

For categorical variable X_i with levels $j = 1, \dots, k$, create a dummy variables x_{ij} by

$$x_{ij} = \begin{cases} 1, & \text{obs. in level } j, \\ 0, & \text{obs. not in level } j, \end{cases}$$

For example,

$$\text{Eth}_{AA} = \begin{cases} 1, & \text{obs. is African American,} \\ 0, & \text{obs. is not African America} \end{cases}$$

$$\text{Eth}_{A} = \begin{cases} 1, & \text{obs. is Asian,} \\ 0, & \text{obs. is not Asian} \end{cases}$$

$$\text{Eth}_{C} = \begin{cases} 1, & \text{obs. is Caucasion,} \\ 0, & \text{obs. is not Caucasion} \end{cases}$$

- Every observation evaluates to 1 in exactly 1 dummy variable.

Categorical Variables in R

```r
credit_mod <- lm(Balance ~ Limit + Income + Gender + Ethnicity, data = Credit)
summary(credit_mod)$coefficients
```

```
##                        Estimate   Std. Error     t value      Pr(>|t|)
## (Intercept)         -395.7122121 25.890307793 -15.2841834  9.661647e-42
## Limit                  0.2645314  0.005894931  44.8743906 6.014584e-157
## Income                -7.6671626  0.386036409 -19.8612421  2.508448e-61
## GenderFemale           1.9069535 16.599113684   0.1148828  9.085965e-01
## EthnicityAsian        26.8788662 23.412591822   1.1480517  2.516438e-01
## EthnicityCaucasian     3.7623916 20.399222553   0.1844380  8.537648e-01
```

$$\widehat{\text{Balance}} = -395.7 + 0.26 \cdot \text{L} - 7.67 \cdot \text{I} + 1.91 \cdot \text{G}_F + 26.88 \cdot \text{E}_A + 3.76 \cdot \text{E}_C$$