Linear regression with categorical predictors

Poverty vs. region (east, west)

by(poverty$Poverty, poverty$region2,
   function(x) c(mean=mean(x),med=median(x),sd=sd(x),iqr=IQR(x)))

## poverty$region2: east
## mean  med   sd   iqr
## 11.17037 10.30000 3.08543 4.60000
## ---------------------------------------------
## poverty$region2: west
## mean  med   sd   iqr
## 11.55000 10.70000 3.16846 4.00000

summary(lm(Poverty ~ region2, data=poverty))

## Call:  
## lm(formula = Poverty ~ region2, data = poverty)
## ## Residuals:  
##    Min 1Q Median 3Q Max
## -5.5704 -2.2000 -0.8704 2.0398 6.4500
## ## Coefficients:  
##            Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.1704    0.6013 18.576  <2e-16 ***
## region2west 0.3796    0.8766  0.433  0.667
## ---
## Signif. codes:  
##  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1
## ## Residual standard error: 3.125 on 49 degrees of freedom
## Multiple R-squared:  0.003813, Adjusted R-squared:  -0.01652
## F-statistic: 0.1875 on 1 and 49 DF, p-value: 0.6669
Poverty vs. Region (Northeast, Midwest, West, South)

Which region (Northeast, Midwest, West, South) is the reference level?

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 9.50     | 0.87       | 10.94   | 0.00     |
| region4midwest | 0.03   | 1.15       | 0.02    | 0.98     |
| region4west  | 1.79     | 1.13       | 1.59    | 0.12     |
| region4south | 4.16    | 1.07       | 3.87    | 0.00     |

Interpretation:
- Predict 9.50% poverty in Northeast
- Predict 9.53% poverty in Midwest
- Predict 11.29% poverty in West
- Predict 13.66% poverty in South
Linear regression with categorical predictors

Poverty vs. Region (Northeast, Midwest, West, South)

**summary(aov(poverty$Poverty ~ poverty$region4))**

```
## Df Sum Sq Mean Sq F value Pr(>F)
## poverty$region4 3 161.4 53.81 7.933 0.00022 ***
## Residuals 47 318.8 6.78
## ---
## Signif. codes: 
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

---

Multiple predictors in a linear model

Weights of books

<table>
<thead>
<tr>
<th>weight (g)</th>
<th>volume (cm$^3$)</th>
<th>cover</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>800</td>
<td>885</td>
</tr>
<tr>
<td>2</td>
<td>950</td>
<td>1016</td>
</tr>
<tr>
<td>3</td>
<td>1050</td>
<td>1125</td>
</tr>
<tr>
<td>4</td>
<td>350</td>
<td>239</td>
</tr>
<tr>
<td>5</td>
<td>750</td>
<td>701</td>
</tr>
<tr>
<td>6</td>
<td>600</td>
<td>641</td>
</tr>
<tr>
<td>7</td>
<td>1075</td>
<td>1228</td>
</tr>
<tr>
<td>8</td>
<td>250</td>
<td>412</td>
</tr>
<tr>
<td>9</td>
<td>700</td>
<td>953</td>
</tr>
<tr>
<td>10</td>
<td>650</td>
<td>929</td>
</tr>
<tr>
<td>11</td>
<td>975</td>
<td>1492</td>
</tr>
<tr>
<td>12</td>
<td>350</td>
<td>419</td>
</tr>
<tr>
<td>13</td>
<td>950</td>
<td>1010</td>
</tr>
<tr>
<td>14</td>
<td>425</td>
<td>595</td>
</tr>
<tr>
<td>15</td>
<td>725</td>
<td>1034</td>
</tr>
</tbody>
</table>

---

Weights of hard cover and paperback books

Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?
**Multiple predictors in a linear model**

**Modeling weights of books using volume and cover type**

```
book_mlr = lm(weight ~ volume + cover, data = allbacks)
summary(book_mlr)

## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 197.96284  59.19274    3.344  0.005841 **
## volume        0.71795   0.06153  11.669  6.6e-08 ***
## cover:pb     -184.04727  40.49420   -4.545  0.000672 ***
##
## Residual standard error: 78.2 on 12 degrees of freedom
## Multiple R-squared: 0.9275, Adjusted R-squared: 0.9154
## F-statistic: 76.73 on 2 and 12 DF, p-value: 1.455e-07
```

For **hardcover** books: plug in 0 for cover

`\hat{weight} = 197.96 + 0.72 \text{ volume} - 184.05 \times 0`

For **paperback** books: plug in 1 for cover

`\hat{weight} = 197.96 + 0.72 \text{ volume} - 184.05 \times 1`

```
visualising the linear model

![](image)
```

**Interpretation of the regression coefficients**

```
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 197.96   59.19   3.344    0.01
## volume       0.72    0.06  11.669    0.00
## cover:pb     -184.05  40.49  -4.545    0.00

Slope of volume: All else held constant, for each 1 cm\(^3\) increase in volume we would expect weight to increase on average by 0.72 grams.

Slope of cover: All else held constant, the model predicts that paperback books weigh 184 grams less than hardcover books, on average.

Intercept: Hardcover books with no volume are expected on average to weigh 198 grams.

Obviously, the intercept does not make sense in context. It only serves to adjust the height of the line.
```
Multiple predictors in a linear model

Prediction

What is the correct calculation for the predicted weight of a paperback book that has a volume of 600 cm$^3$?

|           | Estimate | Std. Error | t value | Pr(>|t|) |
|-----------|----------|------------|---------|----------|
| (Intercept)| 197.96   | 59.19      | 3.34    | 0.01     |
| volume    | 0.72     | 0.06       | 11.67   | 0.00     |
| cover:pb  | -184.05  | 40.49      | -4.55   | 0.00     |

$\hat{weight} = 197.96 + 0.72 \times \text{volume} - 184.05 \times \text{cover:pb}$

This model assumes that hardcover and paperback books have the same slope for the relationship between their volume and weight. If this isn’t reasonable, then we would include an “interaction” variable in the model.

Example of an interaction

summary(lm(weight ~ volume + cover + volume:cover, data = allbacks))

## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## (Intercept) 161.5865  86.5192  1.87     0.0887
## volume      0.7616    0.0972  7.84     0.0000
## coverpb    -120.2141 115.6590 -1.04     0.3209
## volume:coverpb -0.0757  0.1280 -0.59     0.5661

Regression equations for hardbacks:

$\hat{weight} = 161.58 + 0.76 \times \text{volume} - 120.21 \times 0 - 0.076 \times \text{volume} \times 0$

$= 161.58 + 0.76 \times \text{volume}$

Regression equations for paperbacks:

$\hat{weight} = 161.58 + 0.76 \times \text{volume} - 120.21 \times 1 - 0.076 \times \text{volume} \times 1$

$= 41.37 + 0.686 \times \text{volume}$
### Example of an interaction - Results

![Graph showing interaction between volume (cm³) and weight (g) for hardcover and paperback books.]

### Another look at $R$

For a linear regression we have defined the correlation coefficient to be

$$R = \text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

This definition works fine for the simple linear regression case where $X$ and $Y$ are numeric variables, but does not work well in some of the extensions we will see next week.

A more useful, and equivalent, definition is $R = \text{Cor}(Y, \hat{Y})$, which will work for all regression examples we will see in this class.

### Another look at $R^2$, cont.

**Claim:** $\text{Cor}(X, Y) = \text{Cor}(Y, \hat{Y})$

**Remember:**
- $\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$, $\hat{Y} = b_0 + b_1 X$
- $\text{Var}(aX + b) = a^2 \text{Var}(X)$
- $\text{Cov}(aX + b, Y) = a \text{Cov}(X, Y)$

$$\text{Cor}(Y, \hat{Y}) = \frac{\text{Cov}(Y, \hat{Y})}{\sqrt{\text{Var}(Y) \text{Var}(\hat{Y})}}$$

$$= \frac{\text{Cov}(Y, b_0 + b_1 X)}{\sqrt{\sigma_Y^2 \text{Var}(b_0 + b_1 X)}}$$

$$= \frac{b_1 \text{Cov}(Y, X)}{\sigma_Y \sqrt{b_1^2 \text{Var}(X)}}$$

$$= \frac{b_1 \text{Cov}(Y, X)}{b_1 \sigma_Y \sigma_X}$$

$$= \text{Cor}(X, Y)$$

### Another look at $R^2$

So how can we claim that $R^2$ is a measure of variability “explained” by the model?

Remember, in an ANOVA we can partition total uncertainty into model (group) uncertainty and residual (error) uncertainty.

$$SST = SSG + SSE$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

For a regression we can do the same thing, just replacing $\bar{y}_i$ with $\hat{y}_i$

$$SST = SSR + SSE$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
Another look at $R^2$

After a fair bit of algebra we can show that,

\[ R^2 = \frac{\text{Cor}(Y, \hat{Y})^2}{\text{Var}(Y)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \]

Predicting poverty using % female householder

summary(lm(poverty ~ female_house, data = poverty))

| Estimate   | Std. Error | t value | Pr(>|t|) |
|------------|------------|---------|----------|
| (Intercept)| 3.31       | 1.90    | 1.74     | 0.09     |
| female_house| 0.69       | 0.16    | 4.32     | 0.00     |

\[ R = 0.53 \]
\[ R^2 = 0.53^2 = 0.28 \]

Revisit: Modeling poverty

anova(lm(poverty ~ female_house, data = poverty))

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>female_house</td>
<td>1</td>
<td>132.57</td>
<td>18.68</td>
<td>0.00</td>
</tr>
<tr>
<td>Residuals</td>
<td>49</td>
<td>347.68</td>
<td>7.10</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>480.25</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Another look at $R^2$ - from last week

\[ \text{SS}_{\text{Tot}} = \sum (y - \bar{y})^2 \rightarrow \text{total variability} \]
\[ \text{SS}_{\text{Err}} = \sum e_i^2 = 347.68 \rightarrow \text{unexplained variability} \]
\[ \text{SS}_{\text{Reg}} = \text{SS}_{\text{Total}} - \text{SS}_{\text{Error}} \rightarrow \text{explained variability} \]
\[ \text{SS}_{\text{Reg}} = 480.25 - 347.68 = 132.57 \]
\[ R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57}{480.25} = 0.28 \]
Predicting poverty using % female hh + % cauc

Let’s start by predicting poverty using % female hh + % cauc:

```r
pov_mlr = lm(poverty ~ female_house + cauc, data = poverty)
```

We can then use `summary()` to look at the results:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.58 5.78 -0.45 0.66
female_house 0.89 0.24 3.67 0.00
cauc 0.04 0.04 1.08 0.29
```

We can also look at the `anova()` result:

```
Df Sum Sq Mean Sq F value Pr(>F)
female_house 1 132.57 132.57 18.74 0.00
cauc 1 8.21 8.21 1.16 0.29
Residuals 48 339.47 7.07
Total 50 480.25
```

Let’s calculate the explained variability and total variability to find the $R^2$:

\[
R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57 + 8.21}{480.25} = 0.29
\]

Now let’s calculate the adjusted $R^2$:

\[
R^2_{\text{adj}} = 1 - \left( \frac{SS_{\text{Error}}}{SS_{\text{Total}}} \times \frac{n-1}{n-k} \right)
\]

where $n$ is the number of cases and $k$ is the number of predictors (explanatory variables including the intercept) in the model.

- Because $k$ is never negative, $R^2_{\text{adj}}$ will always be less than or equal to $R^2$.
- $R^2_{\text{adj}}$ applies a penalty for the number of predictors included in the model.
- Therefore, we prefer models with higher $R^2_{\text{adj}}$.

We would like to have some criteria to evaluate if adding an additional variable makes a difference in the explanatory power of the model.

- When any variable is added to the model $R^2$ increases.
- Adjusted $R^2$ is based on $R^2$ but it penalizes the addition of variables.

Let’s calculate the adjusted $R^2$:

```
adjusted_R2 = 1 - (SS_E / SS_T) * (n - 1) / (n - k)
```

where $SS_E$ is the sum of squares for error, $SS_T$ is the total sum of squares, $n$ is the number of cases, and $k$ is the number of predictors.

```
adjusted_R2 = 1 - (339.47 / 480.25) * (51 - 1) / (51 - 3)
adjusted_R2 = 1 - (339.47 / 480.25) * 50 / 48
adjusted_R2 = 1 - 0.74
adjusted_R2 = 0.26
```
We saw that adding the variable cauc to the model only marginally increased adjusted $R^2$, i.e. did not add much useful information to the model. Why?

Two predictor variables are said to be collinear when they are correlated, and this collinearity (also called multicollinearity) complicates model estimation.

Remember: Predictors are also called explanatory or independent variables, so ideally they should be independent of each other.

We don’t like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest model that explains as much as possible - the most parsimonious model.

In addition, inclusion of collinear variables can result in biased estimates of the slope parameters.

While it’s impossible to avoid all collinearity, often experiments are designed to control for correlated predictors.

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are from a survey of adult American women and their children - a subsample from the National Longitudinal Survey of Youth.

cog_full = lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age, data = cognitive)

summary(cog_full)

|              | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------|----------|------------|---------|----------|
| (Intercept)  | 19.59241 | 9.21906    | 2.125   | 0.0341   |
| mom_hs       | 5.09482  | 2.31450    | 2.201   | 0.0282   |
| mom_iq       | 0.56147  | 0.06064    | 9.259   | <2e-16   |
| mom_work     | 2.53718  | 2.35067    | 1.079   | 0.2810   |
| mom_age      | 0.21802  | 0.33074    | 0.659   | 0.5101   |
| Residual standard error: 18.14 on 429 degrees of freedom
| Multiple R-squared: 0.2171, Adjusted R-squared: 0.2098
| F-statistic: 29.74 on 4 and 429 DF, p-value: < 2.2e-16

Inference for MLR

Inference for the model as a whole

Is the model as a whole significant?

\[ H_0: \beta_0 = \beta_1 = \cdots = \beta_k = 0 \]
\[ H_A: \text{At least one of the } \beta_i \neq 0 \]

F-statistic: 29.74 on 4 and 429 DF, p-value: < 2.2e-16

Since p-value < 0.05, the model as a whole is significant.

- The F test yielding a significant result doesn’t mean the model fits the data well, it just means at least one of the βs is non-zero.
- The F test not yielding a significant result doesn’t mean individuals variables included in the model are not good predictors of y, it just means that the combination of these variables doesn’t yield a good model.

Inference for the slope(s)

Is whether or not mom went to high school a significant predictor of kid’s cognitive test score, given all other variables in the model?

\[ H_0: \beta_1 = 0, \text{ when all other variables are included in the model} \]
\[ H_A: \beta_1 \neq 0, \text{ when all other variables are included in the model} \]

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 19.59241 | 9.21906 | 2.125 | 0.0341 |
| mom_hs:yes | 5.09482 | 2.31450 | 2.201 | 0.0282 |
| mom_iq | 0.56147 | 0.06064 | 9.259 | <2e-16 |
| mom_work:yes | 2.53718 | 2.35067 | 1.079 | 0.2810 |
| mom_age | 0.21802 | 0.33074 | 0.659 | 0.5101 |

Residual standard error: 18.14 on 429 degrees of freedom

\[ T = 2.201, \quad df = n - k = 434 - 5 = 429, \quad p-value = 0.0282 \]

Since p-value < 0.05, whether or not mom went to high school is a significant predictor of kid’s test score, given all other variables in the model.

CI Recap from last time

Inference for the slope for a SLR model (only one explanatory variable):

- Hypothesis test:
  \[ T = \frac{b_1 - \text{null value}}{SE_{b_1}} \quad df = n - 2 \]

- Confidence interval:
  \[ b_1 \pm t^*_{df} \times SE_{b_1} \]

The only difference for MLR is that we use \( b_i \) instead of \( b_1 \), and use \( df = n - k \).
Construct a 95% confidence interval for the slope of mom_work.

\[ b_k \pm t^\star SE_{b_k} \]

\[ df = n - k = 434 - 5 = 429 \rightarrow 400 \]

2.54 ± 1.97 \times 2.35

2.54 ± 4.63

\((-2.0895, 7.1695)\)

Interpretation?

Given all variables in the model, which variables are significant predictors of kid’s cognitive test score?

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 19.59241 | 9.21906    | 2.125   | 0.0341   |
| mom_hsyes  | 5.09482  | 2.31450    | 2.201   | 0.0282   |
| mom_iq     | 0.56147  | 0.06064    | 9.259   | <2e-16   |
| mom_workyes| 2.53718  | 2.35067    | 1.079   | 0.2810   |
| mom_age    | 0.21802  | 0.33074    | 0.659   | 0.5101   |