TEXAS McCombs

Dummy variables

David Puelz

Detecting sex discrimination

You want to relate salary (Y) to gender (X)... how can we do that?

Gender is an example of a **categorical variable**. The gender variable separates our data into 2 groups or categories.

The question we want to answer is: *"how is your salary related to which group you belong to...?"*

Detecting sex discrimination

You want to relate salary (Y) to gender (X)... how can we do that?

Gender is an example of a **categorical variable**. The gender variable separates our data into 2 groups or categories.

The question we want to answer is: *"how is your salary related to which group you belong to...?"*

Could we think about additional examples of categories potentially associated with salary?

- UT education vs. not
- foreign or domestic born citizen
- quarterback vs. wide receiver

Detecting sex discrimination

We can use regression to answer these questions, but first we need to recode the categorical variable into a **dummy variable**:

```
 Gender     Salary  Sex
1      Male  32.00   1
2    Female  39.10   0
3    Female  33.20   0
4    Female  30.60   0
5      Male  29.00   1
... ... ...
208  Female  30.00   0
```

Detecting sex discrimination

Now you can present the following model in court:

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Sex}_i + \epsilon_i$$

How do you interpret β_1? What is your predicted salary for males and females?

Detecting sex discrimination

Now you can present the following model in court:

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Sex}_i + \epsilon_i$$

How do you interpret β_1? What is your predicted salary for males and females?

$$\begin{aligned} E[\text{Salary}|\text{Sex} = 0] &= \beta_0 \\ E[\text{Salary}|\text{Sex} = 1] &= \beta_0 + \beta_1 \end{aligned}$$

β_1 is the male-female difference!

Detecting sex discrimination

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Sex}_i + \epsilon_i$$

```
data = read.table("SalaryData.txt",header=T)
Sex = (data$Gender=="Male")
data$Sex = Sex
fit = lm(Salary~Sex,data)
summary(fit)
```

```
##
## Call:
## lm(formula = Salary ~ Sex, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.805  -6.434  -1.860   4.115  51.495
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2099     0.8945  41.597  < 2e-16 ***
## SexTRUE       8.2955     1.5645   5.302 2.94e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\hat{\beta}_1 = b_1 = 8.29...$ on average, a male makes approximately \$8,300 more than a female in this firm.

Detecting sex discrimination

How can the defense attorney try to counteract the plaintiff's argument?

Perhaps, the observed difference in salaries is related to other variables in the background and NOT to gender discrimination...

Obviously, there are many other factors which we can legitimately use in determining salaries:

- education
- job productivity
- experience

How can we use regression to incorporate additional information?

Detecting sex discrimination

Let's add a measure of experience...

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Sex}_i + \beta_2 \text{Exp}_i + \epsilon_i$$

What does that mean? Write out the model for each gender separately:

$$\begin{aligned} E[\text{Salary}|\text{Sex} = 0, \text{Exp}] &= \beta_0 + \beta_2 \text{Exp} \\ E[\text{Salary}|\text{Sex} = 1, \text{Exp}] &= (\beta_0 + \beta_1) + \beta_2 \text{Exp} \end{aligned}$$

Detecting sex discrimination

Here is our data with this additional variable:

	Exp	Gender	Salary	Sex
1	3	Male	32.00	1
2	14	Female	39.10	0
3	12	Female	33.20	0
4	8	Female	30.60	0
5	3	Male	29.00	1
...	...	...		
208	33	Female	30.00	0

Detecting sex discrimination

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Sex}_i + \beta_2 \text{Exp} + \epsilon_i$$

```
fit = lm(Salary~Sex+Exp,data)
summary(fit)
```

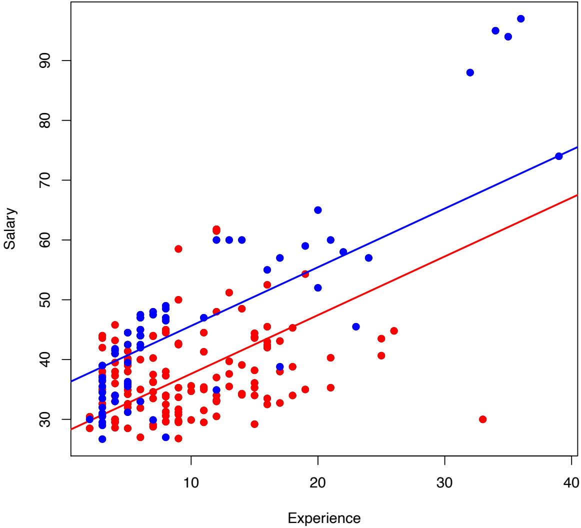
```
##
## Call:
## lm(formula = Salary ~ Sex + Exp, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.1899  -5.7484  -0.6046   4.8129  25.8548
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.81190    1.02789  27.057  < 2e-16 ***
## SexTRUE      8.01189    1.19309   6.715 1.81e-10 ***
## Exp          0.98115    0.08028  12.221  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\rightarrow$ $\text{Salary}_i = 27 + 8\text{Sex}_i + 0.98\text{Exp}_i + \epsilon_i$

Is this good or bad news for the defense?

Detecting sex discrimination

$$\text{Salary}_i = \begin{cases} 27 + 0.98\text{Exp}_i + \epsilon_i & \text{females} \\ 35 + 0.98\text{Exp}_i + \epsilon_i & \text{males} \end{cases}$$



More than two categories

We can use dummy variables in situations in which there are more than two categories. Dummy variables are needed for each category except one, designated as the "base" category.

Why? **Remember that the numerical value of each category has no quantitative meaning!**

House prices revisted

We want to evaluate the difference in house prices in a couple of different neighborhoods.

	Nbhd	SqFt	Price
1	2	1.79	114.3
2	2	2.03	114.2
3	2	1.74	114.8
4	2	1.98	94.7
5	2	2.13	119.8
6	1	1.78	114.6
7	3	1.83	151.6
8	3	2.16	150.7
...	...	...	...

Let's create the dummy variables dn1, dn2 and dn3...

```
   Nbhd  SqFt  Price  dn1 dn2 dn3
1      2 1.79 114.3   0   1   0
2      2 2.03 114.2   0   1   0
3      2 1.74 114.8   0   1   0
4      2 1.98  94.7   0   1   0
5      2 2.13 119.8   0   1   0
6      1 1.78 114.6   1   0   0
7      3 1.83 151.6   0   0   1
8      3 2.16 150.7   0   0   1
... ... ...
```

House prices revisted

$$\text{Price}_i = \beta_0 + \beta_1 \text{dn1}_i + \beta_2 \text{dn2}_i + \beta_3 \text{SqFt}_i + \epsilon_i$$

$$E[\text{Price}|\text{dn1} = 1, \text{SqFt}] \quad = \quad \beta_0 + \beta_1 + \beta_3 \text{SqFt} \quad (\text{Nbhd 1})$$

House prices revisted

$$\text{Price}_i = \beta_0 + \beta_1 \text{dn1}_i + \beta_2 \text{dn2}_i + \beta_3 \text{SqFt}_i + \epsilon_i$$

$$\begin{aligned} E[\text{Price}|\text{dn1} = 1, \text{SqFt}] &= \beta_0 + \beta_1 + \beta_3 \text{SqFt} \quad (\text{Nbhd 1}) \\ E[\text{Price}|\text{dn2} = 1, \text{SqFt}] &= \beta_0 + \beta_2 + \beta_3 \text{SqFt} \quad (\text{Nbhd 2}) \end{aligned}$$

House prices revisted

$$\text{Price}_i = \beta_0 + \beta_1 \text{dn1}_i + \beta_2 \text{dn2}_i + \beta_3 \text{SqFt}_i + \epsilon_i$$

$$\begin{aligned}
E[\text{Price}|\text{dn1} = 1, \text{SqFt}] &= \beta_0 + \beta_1 + \beta_3 \text{SqFt} \quad (\text{Nbhd 1}) \\
E[\text{Price}|\text{dn2} = 1, \text{SqFt}] &= \beta_0 + \beta_2 + \beta_3 \text{SqFt} \quad (\text{Nbhd 2}) \\
E[\text{Price}|\text{dn1} = 0, \text{dn2} = 0, \text{SqFt}] &= \beta_0 + \beta_3 \text{SqFt} \quad (\text{Nbhd 3})
\end{aligned}$$

Model output

```
fit = lm(Price~dn1+dn2+SqFt,data)
summary(fit)
```

```
##
## Call:
## lm(formula = Price ~ dn1 + dn2 + SqFt, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.107 -10.924  -0.305   9.643  38.506
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.776     14.248   4.406 2.25e-05 ***
## dn1          -41.535      3.534 -11.754  < 2e-16 ***
## dn2          -30.967      3.369  -9.192 1.13e-15 ***
## SqFt          46.386      6.746   6.876 2.67e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.26 on 124 degrees of freedom
## Multiple R-squared:  0.6851, Adjusted R-squared:  0.6774
## F-statistic: 89.91 on 3 and 124 DF,  p-value: < 2.2e-16
```

$$\text{Price} = 62.78 - 41.54 * \text{dn1} - 30.97 * \text{dn2} + 46.39 * \text{SqFt} + \epsilon$$

What do these models look like?