DATA 557 - Homework 6

Will Wright

February 19, 2019

DATA 557

Homework Assignment 6

Data set: 'cells.csv'

Summary: a randomized clinical trial of immune cell stimulation in 40 patients

Variables:

id: subject id # dose: drug dose (0, 10, or 100mg) sex: sex (0=female, 1=male) age: age (yrs) count0: pre-treatment cell count count1: post-treatment cell count (the response variable)

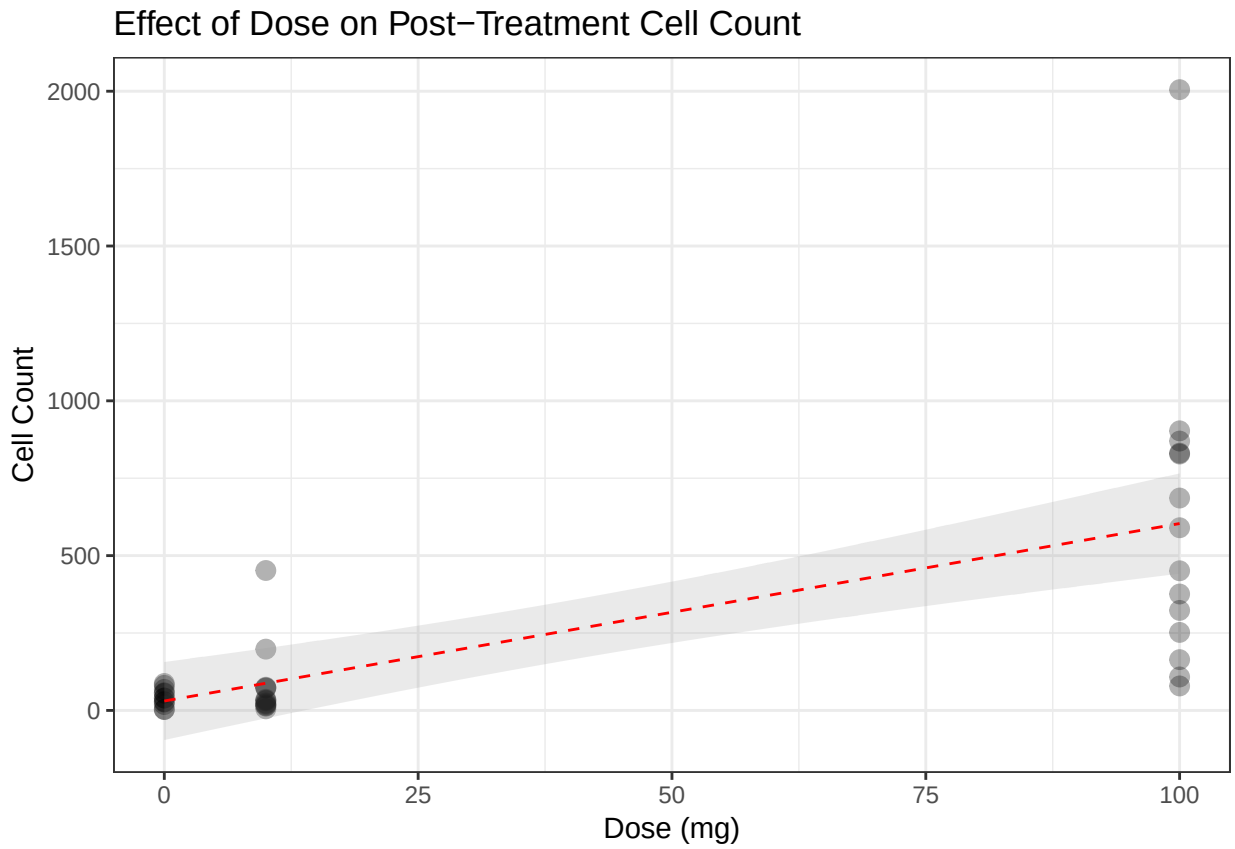
1. Use ANOVA to test for a difference between mean post-treatment cell count between dose groups. Is there evidence for an effect of dose on post-treatment cell count?

```
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## factor(dose)  2 2701378 1350689   14.62 2.09e-05 ***
## Residuals    37 3417120   92355
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, with $p < 0.001$, there is evidence for an effect of dose on post-treatment cell count.

2. Use linear regression to assess the effect of dose on post-treatment cell count? Is there evidence for an effect using regression? Give an interpretation in words of the estimated coefficient for dose. Compare the results using ANOVA and linear regression.

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 30.368713  62.253515 0.4878233 6.284779e-01
## dose         5.731981   1.047054 5.4743884 2.993237e-06
```



With a p-value of <0.001, there is strong evidence for an effect. The coefficient for dose is 5.73, which is the estimate for how much the cell count increases per mg of dose increase. Unlike the ANOVA, here there is stronger significance largely due to the fact that the 100mg dose value is treated as much "further away" in Euclidean space than the 0 and 10 and therefore, the results are more extreme than when treated like a factor. Also, another element of the difference is due to the linear regression testing the hypothesis that the slope is not 0 whereas ANOVA is testing for a difference in means,

3. Add the variable sex to the ANOVA and linear regression models. Describe how the results change when the variable sex is added to the model.

```
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## factor(dose)  2 2701378 1350689  14.274 2.73e-05 ***
## factor(sex)   1   10610   10610   0.112     0.74
## Residuals    36 3406510   94625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

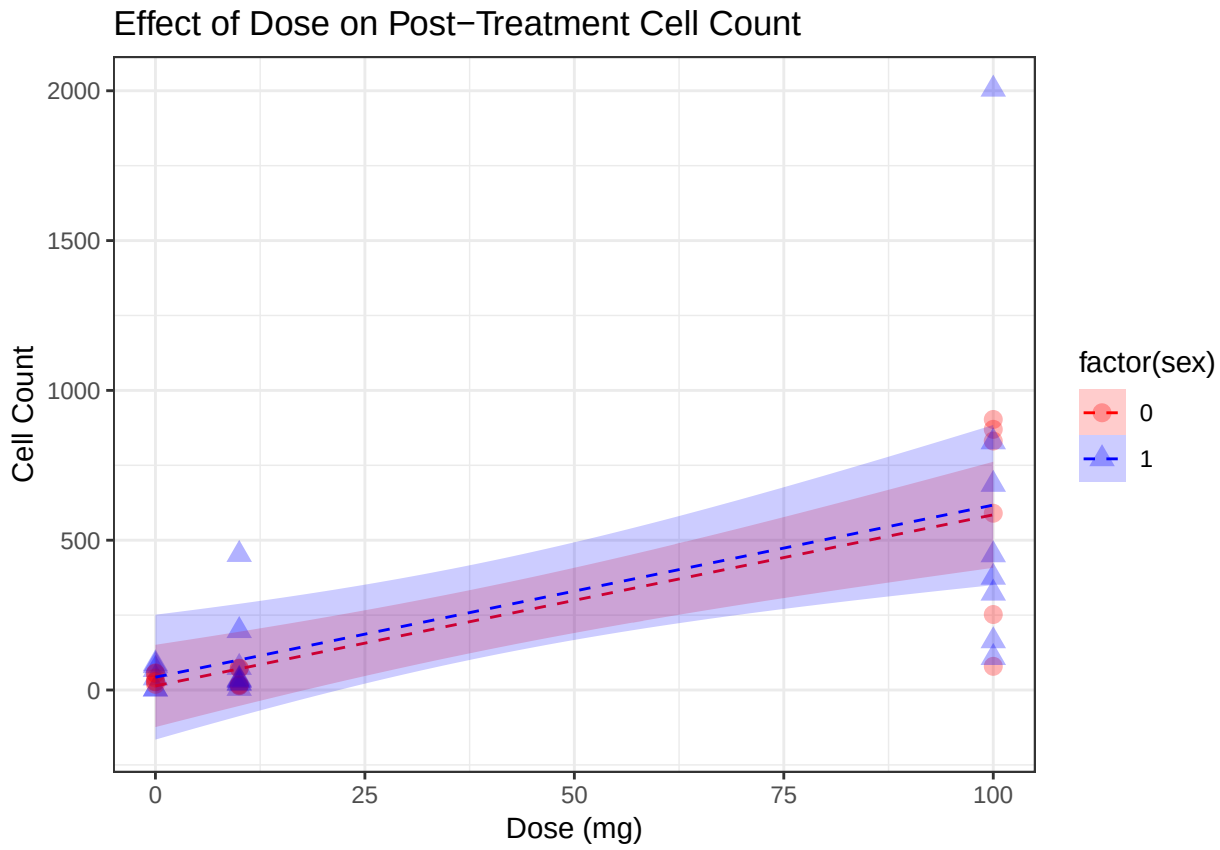
When sex is added (without interaction), the results for dose are still significant, but slightly less so. Sex, however, is not a significant factor.

4. Using ANOVA and linear regression, test for interaction between sex and dose. State the interpretations of the coefficients in the linear regression model with interaction. Give a graphical display of this linear regression model that shows the relationship between dose and response for males and females.

```
##                          Df  Sum Sq Mean Sq F value   Pr(>F)
## factor(dose)              2 2701378 1350689  13.496 4.85e-05 ***
## factor(sex)               1   10610   10610   0.106    0.747
## factor(dose):factor(sex)  2    3875    1938   0.019    0.981
## Residuals                34 3402635  100077
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                       Estimate Std. Error    t value    Pr(>|t|)
## (Intercept)       13.59570957  97.268023 0.13977574 0.889616274
## dose               5.71211221   1.630485 3.50332049 0.001247357
## factor(sex)1      29.25633984 128.978890 0.22683045 0.821838733
## dose:factor(sex)1  0.03216067   2.167533 0.01483745 0.988243786
```



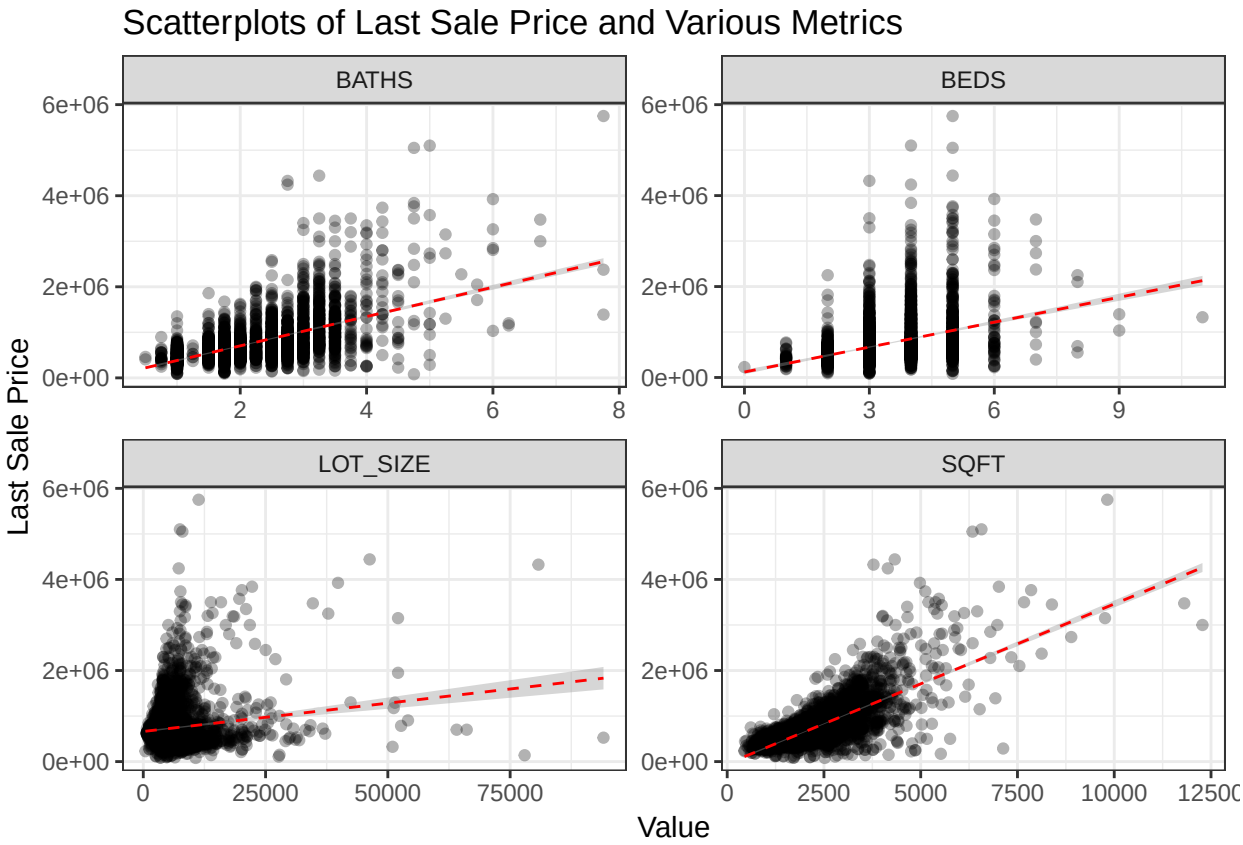
For ANOVA, the results become slightly less significant for than without interaction and both sex and the interaction are not significant. For the linear regression, the results show a less significant effect for dose and no effect for sex or the interaction. This makes sense since there are fewer datapoints in each bucket for gender and this will tend to decrease confidence.

Data set for Questions 5-7: 'Sales.csv'

Variables:

LAST_SALE_PRICE: the sales price of the home SQFT: area of the house (sq. ft.) LOT_SIZE: area of the lot (sq. fit.) BEDS: number of bedrooms BATHS: number of bathrooms

5. Use scatterplots to display the associations between sales price and each of the following predictors: SQFT, LOT_SIZE, BEDS and BATHS. Calculate Pearson correlation coefficients for each plot. Describe the associations in terms of linearity, strength of association and whether the association is positive or negative.



Sqft seems to have the highest degree of linearity with last sale price while lot size seems the most non-linear.

There is a positive correlation between all variables. The strongest correlations are between last sale price and sqft, last sale price and baths, sqft and beds, and beds and baths. The weakest correlations are between last sale price and lot size, lot size and sqft, lot size and beds, and lot size and baths.

6. Use separate linear regression models to assess the association between sales price and each of the four predictor variables (a separate model for each predictor). Interpret the estimated regression coefficients for each model.

```
## (Intercept)        SQFT
##  -47566.522     350.909
```

```
##  (Intercept)     LOT_SIZE
## 661426.19647     12.43952
```

```
## (Intercept)        BEDS
##    119853.3    182697.3
```

```
## (Intercept)       BATHS
##    60140.79   321614.16
```

Sqft's coefficient of 350.9 means that for increase in sqft, price increases by that amount.
Lot size's coefficient of 12.4 means that for increase in lot size, price increases by that amount. This value is perhaps lower than one might expect given the low correlation creating a high intercept. Bed's coefficient of 182,697 means that for increase in number of beds, price increases by that amount. Bath's coefficient of 321,614 means that for increase in number of baths, price increases by that amount.

7. Fit a linear regression model with all 4 predictor variables included. Describe how the estimated coefficients for the predictors change compared to their values in the separate models. Compare the R-squared values for all of the models.

```
##    (Intercept)           SQFT       LOT_SIZE           BEDS          BATHS
##     -10776.748        288.212          5.893      62136.859      -6388.376
##  SQFT:LOT_SIZE      SQFT:BEDS     SQFT:BATHS  LOT_SIZE:BEDS LOT_SIZE:BATHS
##          0.001        -20.069         40.982         -9.047          8.422
##     BEDS:BATHS
##     -15626.105
```

As a single variable, sqft dropped from 351 to 288, lot size dropped from 12.4 to 5.9, beds dropped from 183K to 62K, and now baths converted from a positive 322K to a -6K. The interactions explain these differences. Having fewer beds per sqft increases price, but a property needs more baths per sqft to do the same. The same polarity of interaction applies to lot size for beds and baths. Between beds and baths, having fewer beds per bath is associated with a lower price. Essentially, more baths with fewer, but bigger beds is how to maximize price. I imagine if we took into account the interaction between population density and lot size, we could explain price with more certainty for that factor (with higher densities having a strong effect on price).