Stat 414 - Day 4B
Categorical predictor variables

Example 1: Pace of Life Recall our pace of life data
Suppose I want to see whether the heart disease appears to differ significantly by region of the country (on average).

(a) How would you suggest answering this question?

ANOVA $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ H_a: at least one μ_i differs

(b) Carry out your analysis and summarize your conclusion.

p-value = .06 => weak evidence that at least μ differs from the other

(c) Could we fit a basic regression model for this relationship? If not, why not? If so, how?

Create K-1 indicator variables
= {1 in category, 0 not in category}

(d) How many parameters are estimated by the model? What are they? How do the ANOVA tables compare?

21.44 = $\bar{y}_{MW} = \beta_0$ = predicted response for reference level (Midwest)

0.5556 = predicted increase in heart comparing NE to midwest $H_0: \beta_1 = 0$

```
## Coefficients:
##                 Estimate Std. Error t value           Pr(>|t|)
## (Intercept)      21.4444     1.6257  13.191 0.0000000000000174 ***
## RegionNortheast   0.5556     2.2990   0.242             0.8106
## RegionSouth      -1.7778     2.2990  -0.773             0.4450
## RegionWest       -5.3333     2.2990  -2.320             0.0269 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.877 on 32 degrees of freedom
```

F tests
$H_0: \beta_1 = \beta_2 = \beta_3 = 0$

(e) What is the difference between "indicator coding" and "effect coding"?

compares to overall mean
19.8 = $\beta_0 = \bar{y}$
1.64 = Region 1 avg above $\bar{y}$

```
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  19.8056     0.8128  24.366 <0.0000000000000002 ***
## Region1       1.6389     1.4079   1.164               0.253
## Region2       2.1944     1.4079   1.559               0.129
## Region3      -0.1389     1.4079  -0.099               0.922
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.877 on 32 degrees of freedom
```

(e) How do you interpret these slope coefficients?

Region 1's mean is 1.6389 larger than average. Region 2's mean is 2.194 above average and Region 3's mean is -.1389 below average.

Computer Problem 4: Market Share (due 8am Friday)

Suppose I have data on market share for 36 consecutive months. I want to decide whether knowing whether a discount promotion is in effect impacts market share.

Fit the regression model to predict MarketShare from Discount.

(a) Interpret the slope and intercept coefficients in context

(b) Are the regression model assumptions met? (Explain in detail what Normality, Linearity, and Equal Variance imply here and comment on each, with supporting evidence.

Perform a weighted regression using the observed sample variances to choose the weights.
(c) Explain what is in the "weights" vector I created and why?

(d) Did the estimated slope coefficient change much?

(e) Do the residuals look any better?

What if we just corrected the standard errors?

(f) The slope t-statistic is more similar to that of the original model 1 or model 2?

model 2, the "correction" is similar