Saran Cooper

Week 11 Question 4

4. ANOVA and Lack of Fit

a) Show that the pooled two-sample t test is a special case of simple regression on an indicator variable. For example, test whether goalkeepers and other players have the same mean height or weight. Show that the regression output for the t-test of $\beta_1 = 0$ exactly matches the pooled 2-sample t test.

-equal variance ANOVA

a) Show that pooled 2-sample t test is a special case of simple regression.

Group 1 (G1) = goalies

Group 2 (G2) = field players

Test: Are mean height and veight the same across two groups (equal variance)

$H_0: \mu_1 = \mu_2$ $\quad$ $H_a: \mu_1 \neq \mu_2$

Pooled 2-sample Test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{sp^2}{n_1} + \frac{sp^2}{n_2}}}$$

$\bar{x}_1$ and $\bar{x}_2$ are sample group means $\qquad$ sp^2 = pooled sample variance

n_1 and n_2 are sample sizes.

$df = n_1 + n_2 - 2$

$$sp^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

s_1 and s_2 are sample variances

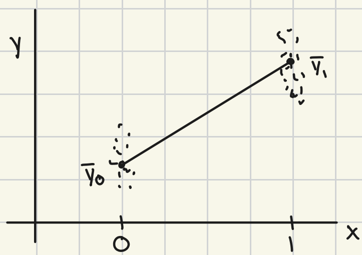
Simple Linear Regression

Y = response variable (height or veight)

$X=0$ goalies

$X=1$ for field players

$\Big\}$ Indicator



$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Y_i = height or veight of player i

X_i = indicator variable

β_0 = intercept, mean height (or veight) for goalies ($x=0$)

β_1 = slope, difference in mean height (or veight) btwn two groups $= \mu_2 - \mu_1$

ε_i = error term

passes through $\bar{Y}_0$ at $x=0$ and $\bar{Y}_1$ at $x=1$

$H_0: \beta_1 = 0$ same as $\mu_2 = \mu_1$ $\quad$ $H_a: \beta_1 \neq 0$

$Y_0 = \beta_0 + \varepsilon_i \rightarrow \beta_0 = E(Y_0) = \mu_0$ $\qquad$ MLE: $\hat{\beta}_0 = \bar{Y}_0$

$Y_1 = \beta_0 + \beta_1 + \varepsilon_i \rightarrow \beta_1 = E(Y_1) - E(Y_0) = \mu_1 - \mu_0$ $\qquad$ MLE: $\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$ $\qquad$ $* \ \Sigma(y_i - \mu)^2 \geq \Sigma(y_i - \bar{y})^2$

Thus $\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$

Comparing t-stats:

$$t_{regression} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{\bar{Y}_1 - \bar{Y}_0}{Sp\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}} \qquad t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{Sp^2\left(\frac{1}{n_0} + \frac{1}{n_1}\right)}}$$

equal variance

*both denominators are the SE of difference between means

$\hat{\beta}_1$ = corresponds directly to difference in sample means

The t-stats are mathematically equivalent, since both estimating for difference btwn groups

$Var(\hat{\beta}_1) = Var(\bar{Y}_1 - \bar{Y}_0)$

$$= \frac{n\sigma^2}{n\Sigma x_i^2 - (\Sigma x_i)^2} = \sigma^2 \frac{n_0+n_1}{(n_0+n_1)n_1 - n_1^2} = \sigma^2 \frac{n_0+n_1}{n_0 n_1} = \sigma^2\left(\frac{1}{n_0} + \frac{1}{n_1}\right)$$

Since pooled the variances are equal and groups are independent.

Part A)

```
soccer_player_data
<-read.table("C:\\Users\\scoop\\OneDrive\\SWAT\\soccerplayer_data.txt", header
= TRUE, sep = ",")
>
>
> # Check the first few rows and structure of the data
> head(soccer_player_data)
  Division Pos GK Weight Height
1        1   F  N    158     71
2        1   M  N    145     71
3        1   M  N    150     67
4        1   D  N    147     68
5        1   F  N    160     68
6        1   M  N    150     68
> str(soccer_player_data)
'data.frame':	1040 obs. of  5 variables:
 $ Division: int  1 1 1 1 1 1 1 1 1 1 ...
 $ Pos     : chr  "F" "M" "M" "D" ...
 $ GK      : chr  "N" "N" "N" "N" ...
 $ Weight  : int  158 145 150 147 160 150 150 160 175 180 ...
 $ Height  : int  71 71 67 68 68 68 69 73 73 73 ...
>
>
> soccer_player_data$Pos <- as.factor(soccer_player_data$Pos)
> soccer_player_data$GK <- as.factor(soccer_player_data$GK)
>
> # Convert GK to a binary variable (1 if "Y", 0 if "N")
> soccer_player_data$GK_binary <- as.integer(soccer_player_data$GK == "Y")
>
> # Perform a pooled 2-sample t-test for height
> t_test_height <- t.test(Height ~ GK, data = soccer_player_data, var.equal =
TRUE)
> print(t_test_height)

	Two Sample t-test

data:  Height by GK
t = -9.382, df = 1038, p-value < 2.2e-16
alternative hypothesis: true difference in means between group N and group Y
is not equal to 0
95 percent confidence interval:
 -2.488077 -1.627333
sample estimates:
mean in group N mean in group Y
       70.71781        72.77551

>
> # Perform a pooled 2-sample t-test for weight
> t_test_weight <- t.test(Weight ~ GK, data = soccer_player_data, var.equal =
TRUE)
> print(t_test_weight)

	Two Sample t-test

data:  Weight by GK
```

```
t = -9.9516, df = 1038, p-value < 2.2e-16
alternative hypothesis: true difference in means between group N and group Y
is not equal to 0
95 percent confidence interval:
 -15.65731 -10.49967
sample estimates:
mean in group N mean in group Y
       164.3841        177.4626

>
> # Simple regression for height
> lm_height <- lm(Height ~ GK, data = soccer_player_data)
> summary(lm_height)

Call:
lm(formula = Height ~ GK, data = soccer_player_data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7178 -1.7178  0.2822  1.2822  7.2822

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.71781    0.08246 857.625   <2e-16 ***
GKY          2.05771    0.21933   9.382   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.464 on 1038 degrees of freedom
Multiple R-squared:  0.07817, Adjusted R-squared:  0.07728
F-statistic: 88.02 on 1 and 1038 DF,  p-value: < 2.2e-16

>
> # Simple regression for weight
> lm_weight <- lm(Weight ~ GK, data = soccer_player_data)
> summary(lm_weight)

Call:
lm(formula = Weight ~ GK, data = soccer_player_data)

Residuals:
    Min      1Q  Median      3Q     Max
-37.463  -9.384   0.616  10.616  51.616

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 164.3841     0.4941 332.699   <2e-16 ***
GKY          13.0785     1.3142   9.952   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.77 on 1038 degrees of freedom
Multiple R-squared:  0.0871,  Adjusted R-squared:  0.08622
F-statistic: 99.03 on 1 and 1038 DF,  p-value: < 2.2e-16
```

Recognize that for both Height and Weight the p-values reflect that we should reject the null hypothesis. This suggests that the mean height and weight of goalkeepers and field players are not the same.

b) Show how to represent a 1-way ANOVA model as a Normal linear model with an intercept and $k-1$ coefficients for indicator variables. Write out the ANOVA table for regression and explain the whole model F test, and how this is the same as the ANOVA F test of equal means in this context. Use the soccer height and weight data as an example.

-ANOVA on Positions

↳GK be baseline

b) 1-way ANOVA, testing whether several group means are equal.

k-groups, testing if mean height (or weight) is the same across k groups

ANOVA:

N individuals in k groups. We have $X_1, X_2, \ldots X_k$ groups.

$Y_i = \mu + a_j + \varepsilon_i$ – 1-Way ANOVA

Y_i = height (or weight) of individual

μ = overall mean height/weight

a_j = effect of the j^{th} group

$\varepsilon_i = \sim N(0, \sigma^2)$ random errors

Represent as Linear Model:

For k groups, we have k-1 indicator variables: $X_1, X_2, \ldots X_{k-1}$

↳ $X_j = 1$ if observation i is in group j, $X_j = 0$ otherwise

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{k-1} X_{(k-1)i} + \varepsilon_i$

β_0 = mean for GK $\beta_1, \beta_2, \ldots \beta_{k-1}$ = deviations of the means of each group from β_0

X_{ji} = indicator variable ε_i = error term

ANOVA table for equal mean / regression

p = # of predictors
As model for linear regression
p = k-1

	DF	SS (group average around average)	MS	F-Stat
Model	p $k-1$	$\sum_{j=1}^{k} (\hat{Y}_j - \bar{Y})^2$	$\frac{SS_{btwn}}{k-1}$	$\frac{MS_{model}}{MS_{error}}$
Error	$N-(p+1)$ $N-k$	$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	$\frac{SS_{Error}}{n-k}$	
Total	$N-1$	$\sum_{i=1}^{n} (Y_i - \bar{Y})^2$		

Hypothesis for F-test:

$\beta_1, \beta_2, \beta_2, \ldots \beta_{k-1} = 0$

$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

H_1: at least one group mean is different

$F = \frac{MS\ regression}{MS\ residual} = \frac{SS_{reg}/k-1}{SS_{res}/N-k}$ reject for large F

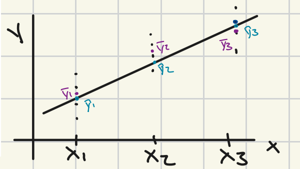
Same as Anova F-test of equal means.

↳ is variation btwn groups sig. larger than variance w/in groups

Part B)

```
# Convert 'Pos' to a factor and set 'GK' as the baseline
(reference level)
> soccer_player_data$Pos <-
factor(soccer_player_data$Pos, levels = c('GK', 'F',
'M','D'))  # 'GK' as baseline
>
> # Perform ANOVA for height based on position
> anova_height <- aov(Height ~ Pos, data =
soccer_player_data)
> summary(anova_height)
              Df Sum Sq Mean Sq F value Pr(>F)
Pos            3    819  273.14   47.03 <2e-16 ***
Residuals   1036   6017    5.81
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # Perform ANOVA for weight based on position
> anova_weight <- aov(Weight ~ Pos, data =
soccer_player_data)
> summary(anova_weight)
              Df Sum Sq Mean Sq F value Pr(>F)
Pos            3  34603   11534   56.03 <2e-16 ***
Residuals   1036 213277     206
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


 >
```

c) Explain how a simple linear regression model becomes an ANOVA model if you treat distinct x values as categories, rather than as a continuous predictor. Note how you need repeated data values with the same x values in order to fit such a model. Describe the Lack of Fit test for regression and how this *saturated* model is used to find an error estimate that does not depend on the linearity assumption. Demonstrate the lack of fit test for several data sets. Note how failing to reject does not imply linearity, or any other regression assumptions.



C) Simple Linear Regression vs ANOVA

$\bar{y}_j$ = average y value for x_j $j=1 \dots k$

SLR: relationship btwn continuous dep. variable y and cont. indep. x.

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim N(0,\sigma^2)$ - Normal linear model

$$SS_{\text{pure error}} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

ANOVA: x as categorical, testing differences between means of groups

$y_i = g(x_i) + \varepsilon_i$ ($g(x_i)$ may be linear)

- you need repeated data points for each value of x. Each category x should have multiple ys w/out it you cannot estimate variability within each group

Lack of Fit Test:

- Used to determine whether a chosen model is inadequate (compares errors of models)

Saturated Model: one parameter for each data point. Has enough parameters to fit the data perfectly, with no residual error. Then error estimate is due to "random noise". Perfect fit.

- compares RSS from chosen model to RSS from saturated. The F-stat determines whether the complexity of saturated model is better fit

Ho: chosen model works Ha: chosen model has lack of fit, more complex is needed

$$SS_{\text{pure error}} = \sum_{j=1}^{k} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 \quad df = n-k \quad \Big\}\ \text{variation within groups at same x}$$

$Y_i = g(x_i) + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

$$SS_{\text{Lack of fit}} = \sum_{j=1}^{k} n_j (\bar{y}_j - \hat{y}_j)^2 \quad \Big\}\ \text{difference in linear model and group means}$$

$$SS_{\text{linear model}} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \hat{y}_j)^2$$

$df = n-2$ $= SS_{\text{pure error}} + SS_{\text{lack of fit}}$

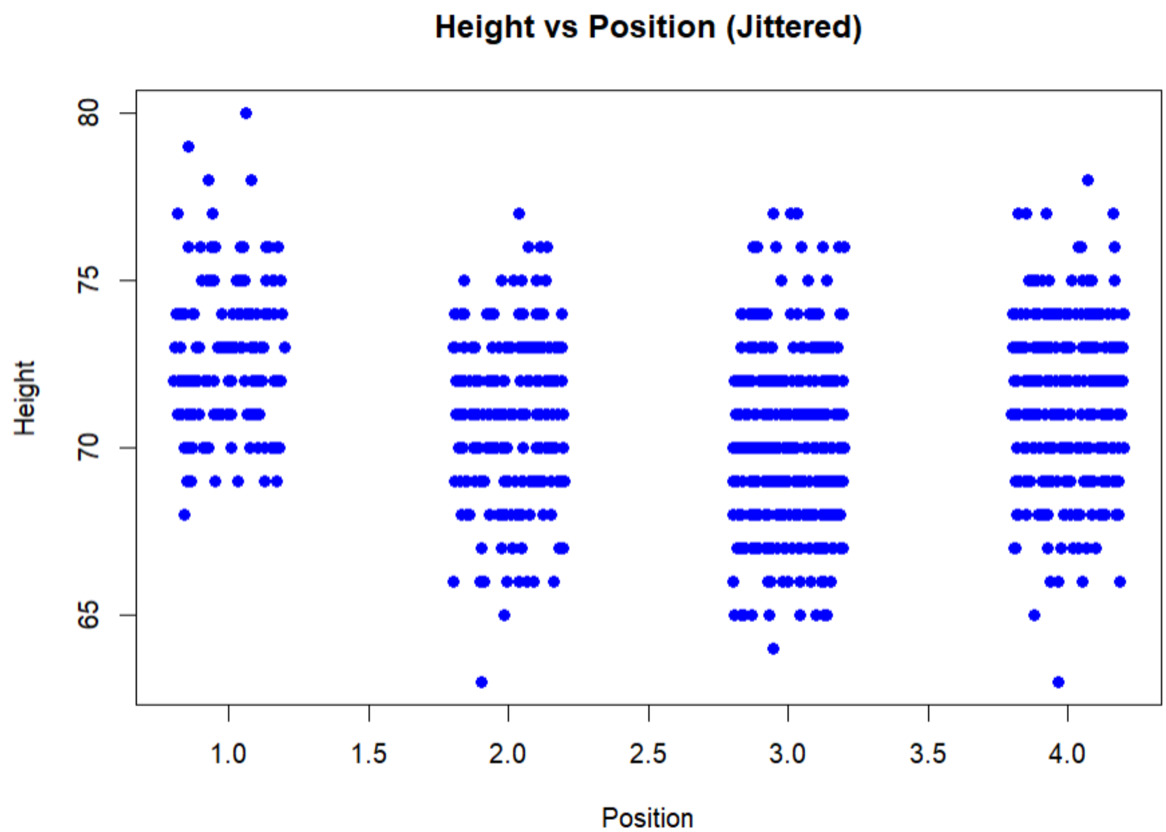
$df = n-k$ $df = k-p$

$$\text{Lack of fit } F = \frac{MS_{LOF}}{MS_{\text{pure error}}} = \frac{SS_{LOF}/k-2}{SS_{\text{pure error}}/n-k}$$

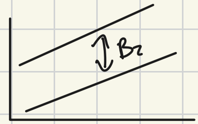
n = total
k = groups
p = parameters in regression

Part C)

```
# 1. Fit a linear regression model (using Pos as the predictor for Height)
> lm_model <- lm(Weight ~ Height, data = soccer_player_data)
> print(anova(lm_model))
Analysis of Variance Table

Response: Weight
            Df Sum Sq Mean Sq F value    Pr(>F)    
Height       1 133284  133284  1207.3 < 2.2e-16 ***
Residuals 1038 114596     110                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # 2. Fit a saturated model (each observation gets its own level in the
model)
> saturated_model <- lm(Weight ~ factor(Height), data = soccer_player_data)
> print(anova(saturated_model))
Analysis of Variance Table

Response: Weight
                 Df Sum Sq Mean Sq F value    Pr(>F)    
factor(Height)   17 135080  7945.9  71.993 < 2.2e-16 ***
Residuals      1022 112799   110.4                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # 3. Perform an ANOVA to compare the two models
> anova_result <- anova(lm_model, saturated_model)
> # Print the ANOVA result to check for lack of fit
> print(anova_result)
Analysis of Variance Table

Model 1: Weight ~ Height
Model 2: Weight ~ factor(Height)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1038 114596                           
2   1022 112799 16    1796.9 1.0175 0.4348
```


Height vs Position (Jittered)
Height
80
75
70
65
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Position

d) Consider the simple ANCOVA model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$ for $i = 1, \ldots, n$, where x_{1i} is height and x_{2i} is an indicator for GK. Give an expression for the mean weight of goal keepers and for the mean weight of other players, as a function of height. Compare the interpretation of β_1 in this model to β_1 in a simple regression of weight on height.



ANCOVA: Analysis of Covariance

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ $i=1,\ldots n$ X_{1i} = height X_{2i} = GK Indicator

Mean veight of GK:

$E[Y_i \mid x_1 = h, x_2 = 1] = \beta_0 + \beta_1 h + \beta_2 = (\beta_0 + \beta_2) + \beta_1 h$

Mean veight of non GK

$E[Y_i \mid x_1 = h, x_2 = 0] = \beta_0 + \beta_1 h$

Compare β_1:

ANCOVA β_1: change in veight for each unit increase in height after adjusting for GK or not

Average change in veight wrt height for goalies and non-goalies

-quantifies the relationship btwn height and veight while holding GK effect constant

Simple Regression β_1: change in veight for each unit increase in height (no other variable)

does not account for how being a goalie could change things

e) Add $\beta_3 x_{3i}$ to the model in part d, with $x_3 = x_1 x_2$. Explain how this generalizes the model to allow different slopes and intercepts in the two groups.

$X_3 = X_1 X_2$

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$

X_3 - allows relationship btwn height and veight to differ depending on gk status

-allows for different slopes and intercepts in GK and non-gk groups.

Non-goalies: $X_2 = 0$

then $X_3 = 0$: $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$ $E[Y_i \mid x_1 = h, x_2 = 0] = \beta_0 + \beta_1 h$

β_1 = slope, how veight changes w/ height, β_0 mean weight when height = 0

Goalies: $(X_2 = 1)$

$X_3 = X_1$; $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 + \beta_3 X_{1i} + \varepsilon_i$ $E(Y_i \mid x_1 = h, x_2 = 1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) h$

$\beta_0 + \beta_2$ = mean veight of GK when heigh is 0

$\beta_1 + \beta_3$ = how veight of GK changes w/ height

-Generalizes because slope changes based on GK status

-captures how relationship btwn height/weight may vary depending on GK status