Simple Linear Regression

Variance σ^2, S_y^2, S_x^2

measures spread of data vertically (y's) OR horizontally (x's)

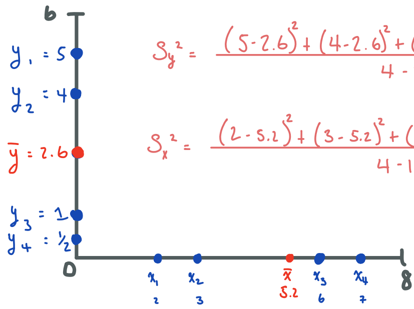
Var of Predictor

$$\boxed{S_x^2 = \frac{\sum^n (x_i - \bar{x})^2}{n-1}}$$

S_x/σ_x is default if unspecified

Var of Outcome

$$\boxed{S_y^2 = \frac{\sum^n (y_i - \bar{y})^2}{n-1}}$$



$$S_y^2 = \frac{(5-2.6)^2 + (4-2.6)^2 + (1-2.6)^2 + (\frac{1}{2}-2.6)^2}{4-1} = \boxed{4.9}$$

$$S_x^2 = \frac{(2-5.2)^2 + (3-5.2)^2 + (6-5.2)^2 + (7-5.2)^2}{4-1} = \boxed{6.32}$$

Standard Deviation σ, S_y, S_x

Same variance but now not in square units. σ^2 is like ft^2 and σ is like ft

$$\boxed{S_y = \sqrt{\frac{\sum^n (y_i - \bar{y})^2}{n-1}}}$$

☆ Same idea for predictors

$\boxed{S_y = \sqrt{4.9}}$

$\boxed{S_x = \sqrt{6.32}}$

Total Sum of Squares TSS, SST

Always refers to outcomes (y's)

$$S_y^2 = \frac{\sum^n (y_i - \bar{y})^2}{\cancel{n-1}} \cdot \cancel{(n-1)} = \text{TSS}$$

$$\boxed{TSS = \sum^n (y_i - \bar{y})^2 = S_y^2 \cdot (n-1)}$$

$\boxed{(5-2.6)^2 + (4-2.6)^2 + (1-2.6)^2 + (\frac{1}{2}-2.6)^2}$ OR $\boxed{4.9 \cdot (4-1)}$

Sum of Square of predictor SSX

like TSS but now with predictors (x's)

$$\boxed{SSX = \sum^n (x_i - \bar{x})^2 = S_x^2 \cdot (n-1)}$$

$\boxed{(2-5.2)^2 + (3-5.2)^2 + (6-5.2)^2 + (7-5.2)^2}$ OR $\boxed{6.32\,(4-1)}$

Covariance S_{xy}

measures spread of data vertically AND horizontally

$(5-2.6) + (4-2.6) + (1-2.6) + (\frac{1}{2}-2.6) = 0.1$

$(2-5.2) + (3-5.2) + (6-5.2) + (7-5.2) = -2.8$

$\frac{(0.1)(-2.8)}{4-1} = \boxed{-0.093}$

$$\boxed{S_{xy} = \frac{\sum^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}}$$

Correlation & Covariance / Standardized covariance

$$\boxed{r_{xy} = \frac{\sum_{i}^{n} (x_i - \bar{x})(y_i - \bar{y})}{S_x S_y (n-1)} = \frac{S_{xy}}{S_x S_y}}$$

$$r_{xy} = \frac{-0.09\bar{3}}{\sqrt{6.32}\sqrt{4.9}} = \boxed{-0.0168}$$

Linear model

$$\hat{y}_i = b_0 + b_1 x_i$$

Slope b_1

$$\boxed{b_1 = \frac{S_{xy}}{S_x^2} = r_{xy} \cdot \frac{S_y}{S_x}}$$

$$b_1 = r_{xy} \cdot \frac{S_y}{S_x} = \frac{S_{xy}}{S_x \cancel{S_y}} \cdot \frac{\cancel{S_y}}{S_x} = \frac{S_{xy}}{S_x^2}$$

$$b_1 = \frac{-0.09\bar{3}}{6.32} = \boxed{-0.0147}$$

Results of simple linear regression

> m1<-lm(gainauthority~group)
> summary(m1)
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	**-2.2535**	0.5510	-4.090	4.58e-05 ***
group	**1.5071**	0.6733	2.238	0.0254 *

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

Residual standard error: 11.55 on 1327 degrees of freedom
(136 observations deleted due to missingness)
Multiple R-squared: 0.003761, Adjusted R-squared: 0.00301
F-statistic: 5.01 on 1 and 1327 DF, p-value: 0.02537

Intercept b_0

$$\boxed{b_0 = \bar{y} - b_1 \bar{x}}$$

$$b_0 = 2.6 - (-0.0147)(5.2) = \boxed{2.676}$$

$$\boxed{\hat{y} = 2.676 + (-0.147)x}$$

Correlation ρ

ρ - "rho" - coefficient of correlation

mesures the relationship btw variables
Variables must be correlated to use one to predict the other

H_0 for bivariate correlation:

- $H_0 : \rho = 0$ No "correlation between x and y"
- $H_a : \rho \neq 0$ Some " "
- $H_a : \rho > 0$ Positive " "
- $H_a : \rho < 0$ Negative " "

T-test of pearson correlation using R
> cor.test(writing,reading)

Pearson's product-moment correlation

data: writing and reading
t = 10.465, df = 198, **p-value < 2.2e-16**
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.4993831 0.6792753
sample estimates:
cor
0.5967765

We reject the null hypothesis that the coefficient of correlation between writing and reading scores in the population is zero. The coefficient of correlation between reading and math is almost 0.6 and it is statistically significant; p =0.000

Correlation Hypothesis t-test

$$\boxed{t = \frac{r}{\sqrt{\frac{(1-r^2)}{N-2}}}}$$

r - sample's corr

Residual e

the error in our prediction
what we CANNOT explain
Actual - Predicted

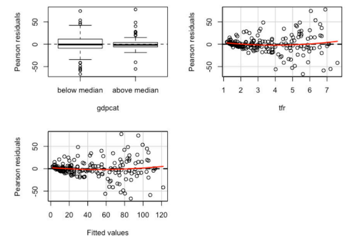
$$e_i = y_i - \hat{y}$$

Since $\hat{y}$ is line of best fit: $\bar{\sum} e_i = \frac{\sum e_i}{n} = 0$

Sum = mean $\bar{e}$ = 0

>Residualplots(m1)

Plot of residuals for the prediction of infant mortality from gdp, total fertility rate, and the combined effect of the two



Variance of Residuals S_e^2

the mean of what our model CANNOT explain

$$S_e^2 = \frac{\sum^n (e_i - \bar{e})^2}{n-2} = \frac{\sum^n e_i^2}{n-2} = \frac{\sum^n (y_i - \hat{y})^2}{n-2} = \frac{\sum^n (y_i - b_0 - b_1 x_i)^2}{n-2}$$

Standard Deviation of Residual S_e

$$S_e^2 = \sqrt{\frac{\sum^n e_i^2}{n-2}} = \sqrt{\frac{\sum^n (y_i - \hat{y})^2}{n-2}} = \sqrt{\frac{\sum^n (y_i - b_0 - b_1 x_i)^2}{n-2}}$$

Sum of Squared Residuals RSS, SS_{res}

the Sum of what our model CANNOT explain

$$RSS = SSE = \sum^n (e_i - \bar{e})^2 = \sum^n e_i^2 = \sum^n (y_i - \hat{y}_i)^2$$

Mean Squared Error MSE

the avg of what our model CANNOT explain

$$MSE = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$$

Standard Error of Slope $S(e)_b$

$$S(e)_b = \frac{S_e}{\sqrt{\sum^n (x_i - \bar{x})^2}} = \frac{S_e}{\sqrt{SSX}}$$

Slope t-test

$$t = \frac{b_1 - B_1}{S(e)_b}$$

B_1 - Pop. Slope
b_1 - Sample Slope

* $H_0 : B_1 = 0$
$H_a : B_1 \neq 0$
$H_a : B_1 > 0$
$H_a : B_1 < 0$

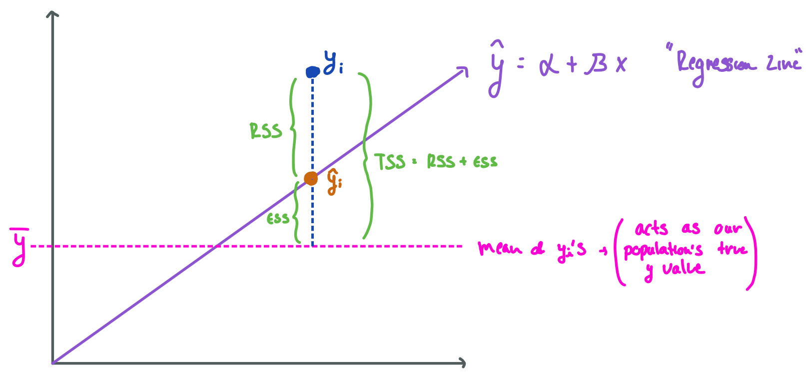
* Recall: CLT

Explained Sum of Squares / Sum of Squared Regression ESS, SS_{reg}

the Sum of what our model CAN explain

$$ESS = SS_{reg} = \sum^n (\hat{y}_i - \bar{y})^2$$

Graphical Interpretation



$$\boxed{SS_{res} = \sum^{n} (y_i - \hat{y})^2}$$

$$\boxed{ESS = \sum^{n} (\hat{y}_i - \bar{y})^2}$$

$$TSS = SS_{res} + ESS$$
$$= \sum^{n} (y_i - \hat{y}_i)^2 + \sum^{n} (\hat{y}_i - \bar{y})^2$$
$$= \sum^{n} (y_i - \cancel{\hat{y}_i} + \cancel{\hat{y}_i} - \bar{y})^2$$

$$\boxed{TSS = \sum^{n} (y_i - \bar{y})^2}$$

R-Squared R^2

Measures how well a Regression model fits data

Proportion of variance our regression model CAN explain

$$\boxed{R^2 = \frac{SS_{reg}}{TSS} = \frac{ESS}{TSS}}$$

Proportion of variance our regression model CANNOT explain

$$\boxed{1 - R^2 = \frac{SS_{res}}{TSS}}$$

Adjusted R^2

adjusts for the # of predictors in model

Increases when a new term improves the model by a stat. sig. amount

⇒ more reliable / conservative measure of goodness of fit in MLR

$$\boxed{\text{Adj } R^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}}$$

$$1 - \underbrace{(1-R^2)}_{\text{Prop. cannot explain}} \underbrace{\frac{(n-1)}{n-p-1}}_{\text{Always } >1}$$

R^2 assumes all Pred explain variation in Response. Adj R^2 tells % of variation explained only by stat sig. Pred

R^2 Change

$$\boxed{R^2_{change} = R^2_{new\ model} - R^2_{old\ model}}$$

Proportion explained by adding new Predictor to our model

F-test of R^2

$$\boxed{F = \frac{SS_{Reg}/k}{SS_{Res}/N-k-1} = \frac{MS_{Reg}}{MS_{Res}} = \frac{Var(Reg)}{Var(Res)} = \frac{R^2/k}{(1-R^2)/N-k-1}}$$

$$\frac{\text{explained var}/k-1}{\text{unexplained var}/n-k}$$

$H_0: R^2 = 0$

$H_a: R^2 \neq 0$

$H_a: R^2 > 0$

$H_a: R^2 < 0$

k - # of predictors

☆ In SLR, $k = 1 \Rightarrow F = t^2$

F Change

$$\boxed{F_{change} = \frac{R^2_{change}/df_{change}}{(1-R^2_{new})/N-k-1}}$$

$$\frac{\text{Prop explain}}{\text{Prop can't explain}}$$

F-test of R-square

```
> m1<-aov(writing~reading, data = hsb2)
> summary(m1)
           Df    Sum Sq    Mean Sq          F value     Pr(>F)
reading    1     6367      6367             109.5       <2e-16 ***
Residuals  198   11511     58
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cohen's f^2

measures effect size for ANOVA and MLR

effect size: the practical/contextual significance of findings

high effect size ⇒ high practical sig

$$\boxed{f^2 = \frac{R^2}{1-R^2}}$$

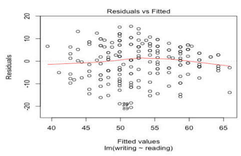
Exploratory Data Analysis EDA

1. Examine Histogram + Scatterplots
2. Check for Outliers, influencial Points, leverage Points
3. Check Normality using QQPlot or expected percentiles in normal model vs. actual percentiles
4. Check homogenicity using plot of Residuals vs. x and Residuals vs. $\hat{y}$

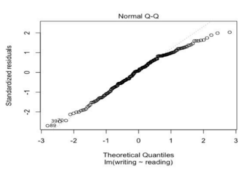
Plot of residuals vs. Y^ (yhat) shows that…

The assumption of equality of the variance of residuals holds. As the value of Y^ increases, the residuals stay the same. In other words, there is no correlation between Y^ and residuals.

Plot thirteen. Plot of residuals vs. Y^ related to the prediction of writing from reading scores

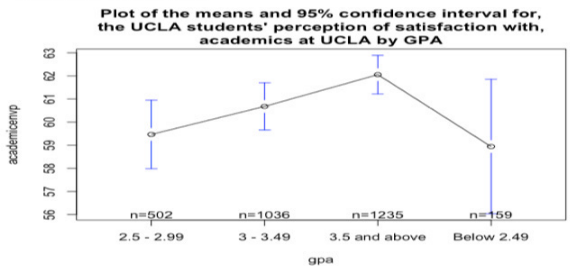


Plot fourteen. qqplot resuled to prediction of writing on reading scores



Based on the qqplot…
The residuals follow the normal model. All the points are on the line indicating that actual and theoretical residuals are close. Toward the two ends, there is some deviance from the line, but it is not too bad.

```
> library(gplots)
> attach(campusclimate)
> gpa <-factor(gpa)
> plotmeans(academicenvp~gpa, main="Plot of the means and 95% confidence interval for,
+the UCLA students' perception of satisfaction with,
+ academics at UCLA by GPA")
```



Statistical vs. Practical Significance

Test	Formula	Standard error
t-test of correlation	$t = \dfrac{r}{\sqrt{\dfrac{(1-r^2)}{N-2}}}$	$\sqrt{\dfrac{(1-r^2)}{N-2}}$
t-test of the slope	$t = \dfrac{b_1 - \beta_1}{SE_b}$	$\dfrac{S_e}{\sqrt{(N-1) * S_X^2}}$
F test of R-squared	$F = \dfrac{MS_{Regression}}{MS_{Residual}}$ $F = \dfrac{SS_{Regression}/k}{\dfrac{SS_{Residual}}{N-K-1}}$ K = number of predictors	$\dfrac{SS_{Residual}}{N-K-1}$

- SE always Dec. as N Inc
  - ↳ as SE Dec. tests Inc.

⇒ Large N's can make deceptive p-value conclusions

P-value could be low but if R^2 is low then conclusion is Statistically Sig. but not Practically Sig.

Power and Effect Size

χ^2, Transformations, Binomial Predictors

Chi-Square χ^2

Criteria:
- Random Sampling
- Large Counts — Condition: expected values > 5
- Categorical variables: qualitative data

```
> qchisq(0.95,1)
[1] 3.841459
```

```
> chisq.test(FIRSTGEN,leaveUCLA)

        Pearson's Chi-squared test with Yates' continuity correction

data:  FIRSTGEN and leaveUCLA
X-squared = 1.7189, df = 1, p-value = 0.1898
```

$$\boxed{\chi^2 = \sum \frac{(\text{Actual} - \text{Expected})^2}{\text{Expected}} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}}$$

$\chi^2 \longrightarrow 0$ as Observed $\rightarrow$ Expected

Goodness of Fit

tests how well the observed data matches our expectations

If actual observations differ enough from expected observations then observations are less likely to have happened by chance

If $\chi^2 >$ Critical Value REJECT H_0

Test for Independence

tests if two categorical variables are independent of one another

If p-value $< \alpha$ REJECT H_0

Test for Homogenous Residuals (homoscedasticity)

an assumption for Regression + ANOVA

↳ variance of Residuals shouldn't INC with fitted values of outcome variable

Breush-Pagan Test / Cook-Weisburg Score Test / Non-Constant Var (NCV)

H_0: constant variance $\quad H_a$: non-constant variance

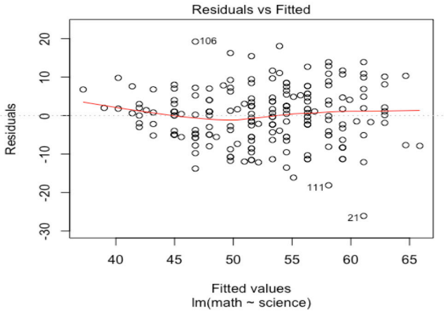
$$\boxed{\chi^2 = \frac{\left(\frac{SSR^*}{2}\right)}{\left(\frac{SSE}{N}\right)^2}}$$

```
> library(car)
> ncvTest((m1))
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 7.122395    Df = 1     p = 0.007612696
```

Based on the ncv Test we reject the null or homoscedasticity of residuals (P = 0.0076<0.05), and conclude that the we are not meeting the assumption of equality of the variance of residuals.

We will now use the Breush Pagan Test to test this assumption.

```
> lmtest::bptest(m1)

        studentized Breusch-Pagan test

data:  m1
BP = 6.3865, df = 1, p-value = 0.0115
```



Plot shows a pattern exists so we suspect heteroscedasticity. Check with χ^2 tests

Transformations

functions applied to skewed data to make normal / symmetric

Tukey - "ladder of powers"

Value of λ	Type of transformation
$\lambda = -1$	inverse transformation
$\lambda = 0$	Log transformation
$\lambda = 1/2$	Square root transformation
$\lambda = 1$	No transformation
$\lambda = 2$	square transformation
$\lambda = 3$	Cubic transformation

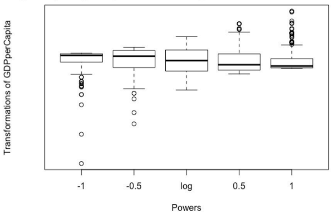
If left/neg skew, move up ladder to achieve symmetry

If Right/pos skew, move down ladder to achieve symmetry

log typically works best

We are going to use the "car" package to find out the most appropriate transformation for the GDP (growth domestic product) data.

```
> library(car)
> symbox(~GDPperCapita, data = undata)
```



Variance Stabalization

- an assumption
- must avoid funnel shape in plot Residuals vs. Fitted values
  - ↳ Square-Root Transformation

Standardize

- transforming vars s.t. mean = 0 translation
  - sd = 1 Dilation
- only for numeric or ordinal vars
- when numeric vars are on different scales/ranges
- allows us to compare coeffs, cannot compare non-standardized
- refers to how many sd's a dep. var will change per 1 sd inc of Ind var ex: $y = 2x_1 + 3x_2$ "for 1 sd inc in x_1 y inc (on avg) by 2 sd"
- Answers Q: which predictor has the highest influence on response? x_2
- Divide anything by its Standard Error it becomes standardized

R statements for creation of standardized coefficients

- install.packages("QuantPsyc")
- library(QuantPsyc)
- lm.beta(m1)

$beta_{science.math}$	$beta_{science.read}$
0.3801172	0.3784138

One-Way ANOVA

Two-Sample test of mean

tests equality of means btw two Independent Populations

ex: Salary btw men + women ceterus perabus

$$\boxed{t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}}$$

Results of two-sample test of the mean

> t.test(gainauthority~group)
data: gainauthority by group
t = -2.1718, df = 806.64, p-value = 0.03016
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.8692372 -0.1449823
sample estimates:

mean in group control	**mean in group experimental**
-2.2534982	**-0.7463884**

Results of ANOVA

> m1=aov(gainauthority~group)
> summary(m1)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	**1**	**668**	**667.8**	**5.01**	**0.0254 ***
Residuals	1327	176879	133.3		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
136 observations deleted due to missingness

Sum of Squares Between SSB

Part we CAN explain.
Diffrence between the mean of two groups

$$\boxed{SSB = \sum^{j} (\bar{y}_j - \bar{y}_a)^2 n_j}$$

j = # of groups
$\bar{y}_j$ - mean of each group
$\bar{y}_a$ - global mean

mean SSB $\quad MSSB = \frac{SSB}{j-1}$

Sum of Squares Within SSW

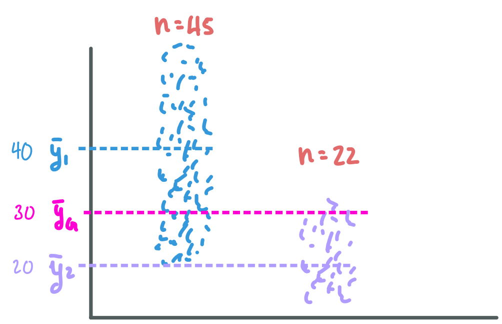
Part we CANNOT explain.
Diffrence between individual observations within each group

$$\boxed{SSW = \sum^{j}\sum^{i} (y_{ij} - \bar{y}_j)^2}$$

y_{ij} - indiv obs i within group j
$\bar{y}_j$ - mean group j

mean SSB $\quad MSSW = \frac{SSW}{N-j}$

$$\boxed{SST = SSB + SSW}$$



$$SSB = \sum^{j} (\bar{y}_j - \bar{y}_a)^2 n_j$$
$$= (\bar{y}_1 - \bar{y}_a)^2 n_1 + (\bar{y}_2 - \bar{y}_a)^2 n_2$$
$$= (40-30)^2 \cdot 45 + (20-30)^2 \cdot 22 = \boxed{6700}$$

$$SSW = \sum (y_{i1} - 40)^2 + \sum (y_{i2} - 20)^2 = \boxed{3200}$$

$$SST = 6700 + 3200 = \boxed{9900}$$

F-Statistic

$$\boxed{F = \frac{MSSB}{MSSW}}$$

$$F = \frac{6700}{3200} = \boxed{2.09}$$

- $F = t^2$ iff $j = 2$ If $j > 2$ then use MANOVA

Homoscedasticity

equality of variance between groups
assumption for ANOVA + Regression

Leven's test, Bartlett's test, Boxplots

Assumption of equality of variance can be tested with Leven's test.

```
> leveneTest(m1)
Levene's Test for Homogeneity of Variance (center = median)
        Df   F value   Pr(>F)
group    3   6.6235    0.0001893 ***
      1869             ---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> bartlett.test(discrimination~edu)

        Bartlett test of homogeneity of variances

data:  discrimination by edu
Bartlett's Chi-squared = 23.248, df = 3, p-value = 3.585e-05
```

Post HOC

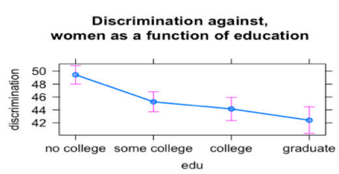
test all possible combinations of outcome variable levels
only needed if levels > 2

Tukey HSD, Bonferroni tests

```
> TukeyHSD(m1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = discrimination ~ edu)

$edu
                            diff        lwr         upr
some college-no college   -4.172143   -6.927233   -1.4170531
college-no college        -5.287388   -8.281651   -2.2931248
graduate-no college       -7.023901  -10.295780   -3.7520223
college-some college      -1.115245   -4.207901    1.9774112
graduate-some college     -2.851758   -6.213915    0.5103992
graduate-college          -1.736513   -5.297298    1.8242722
```

Plot of Means

```
> m1=aov(discrimination~edu)
>library(car)
>library(effects)
>edu<-factor(edu)
> plot(allEffects(m1),ask=FALSE, main="Discrimination against,
+ women as a function of education")
```



Comparing Two Sample Mean, SLR, One-Way ANOVA

1. All should lead to same conclusion
2. Difference of two means = slope
3. Control / base group = intercept
4. $SS_{between}$ analogous to $SS_{Regression}$
5. SS_{within} analogous to $SS_{Residual}$
6. $F = t^2$ iff predictor has 2 levels
7. F-test of R^2 analogous to F-test of ANOVA

Table five. Comparison of the two-sample test of the mean, simple linear regression, and one-way anova in conducting regression analysis for the prediction of a numerical variable from a binary predictor.

Two sample test of the mean	Simple linear regression	One-way ANOVA
$H_0: \mu_1 - \mu_2 = 0$ $H_0: \mu_1 - \mu_2 \# 0$	$H_0: \beta_1 = 0$ $H_0: \beta_1 \# 0$	$H_0: \mu_1 - \mu_2 = 0$ $H_0: \mu_1 - \mu_2 \# 0$
Assumptions: • Normality • Independence • Equality of variance in two populations	**Assumptions** • Normality • Independence • Equality of error variance	**Assumptions:** • Normality • Independence • Equality of variance in two populations
$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$ We assume $\mu_1 - \mu_2 = 0$ $t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$	$t = \frac{b_1 - \beta_1}{\sqrt{\frac{S_e}{SSX}}}$ We assume $H_0: \beta_1 = 0$ $t = \frac{b_1}{\sqrt{\frac{S_e}{SSX}}}$	$F = \frac{SS_{between}/j-1}{SS_{within}/N-j-1}$ $F = \frac{MS_{between}}{MS_{within}}$ $F = t^2$ j = number of groups $SS_{between} = \sum_{j=1}^{2} n_j * (Y_{.j} - \bar{Y}_{..})^2$ $Y_{.j} = mean\ of\ each\ group$ $\bar{Y}_{..} = overall\ mean$ $MS_{with} = \frac{((S_1^2 * (N_1 - 1) + S_2^2 * (N_2 - 1))}{N-j}$ MS_{with} *is like a weighted sum of* variances compare the F formula for ANOVA with the t- formula for two-sample mean and you will see why $F = t^2$

Multiple Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

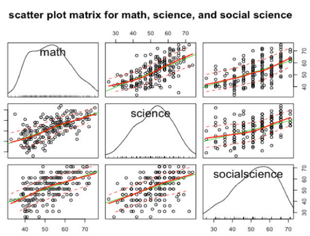
Multicolinearity

- when two+ independent variables are highly correlated with eachother
- Undermines the Stat. significance of an indep. variable

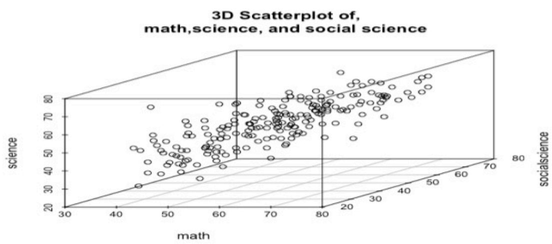
III.1 Scatterplot matrix (see page 125)

```
>library(car)
>scatterplotMatrix(~science+math+socialscience, span=0.7, data =hsb2, main = scatterplot
+matrix for math, science, and social science)
```

We will create the scatterplot matrix for the outcome and predictors to ascertain that the relationship between the outcome and predictors as well as predictors themselves is linear.




```
> library(scatterplot3d)
> attach(hsb2)
> scatterplot3d(math,socialscience,science, main="3D Scatterplot of,
+ math,science, and social science")
```



Variance Inflation Factor VIF

- measures multicolinearity
- If VIF > 5, ∃ multicolinearity

$$\boxed{VIF_j = \frac{1}{1 - R_j^2}}$$

as the j^{th} predictor is correlated with other predictors (multicolinear) $R^2 \to 1 \Rightarrow VIF \uparrow$

R_j^2 - R^2 value when regressing the j^{th} predictor on the remaining predictors

```
> library(car)
> vif(m2)
     math socialscience      reading      writing
 2.085302      1.911425     2.224411     2.002221
```

Resolutions

1. Choose one var and ignore the other
2. Combine vars
   - linear Combination
   - PCA
   - Factor Analysis
3. Regularize Coefficients <u>LASSO or RIDGE</u>

Interactions

- when the effect of one predictor Depends on another Predictor

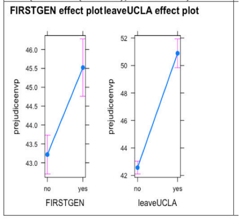
Additive

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Plot of the means for additive model

Plot(allEffects(additive),ask=FALSE)

FIRSTGEN effect plot leaveUCLA effect plot

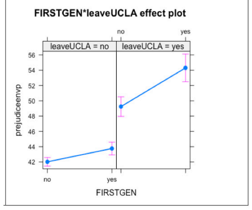


Non-Additive

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Plot of the interaction effect for the non-additive model

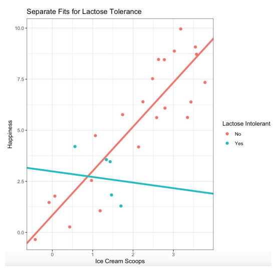
plot(allEffects(nonadditive),ask=FALSE)



• Two-way ANOVA is best when interested in Interactions

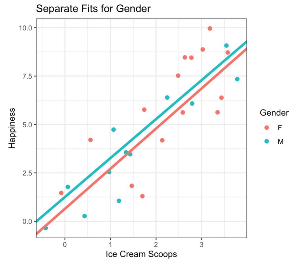
```
m1=aov(prejudiceenvp~FIRSTGEN+leaveUCLA+FIRSTGEN*leaveUCLA)
> summary(m1)
                      Df   Sum Sq    Mean  Sq F value   Pr(>F)
FIRSTGEN              1     6804     6804   26.920     2.2e-07 ***
leaveUCLA             1     51071    51071  202.054    < 2e-16 ***
FIRSTGEN:leaveUCLA    1     1800     1800   7.122      0.00764 **
Residuals             5358  1354269   253
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Disordinal - lack parralelism



Significant Interaction

Ordinal - highly parrallel lines



InSignificant Interaction

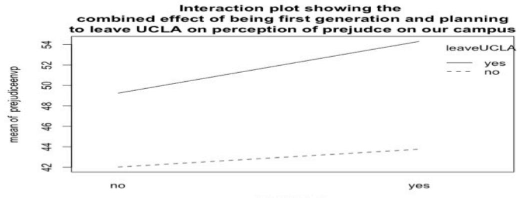
```
Model<-
lm(academicenvp~prejudiceenvp+LowFamilyIncomeIndicator+leaveUCLA+prejudiceenvp*leav
eUCLA+LowFamilyIncomeIndicator*leaveUCLA)
> summary(model)

Coefficients:
                                                   Estimate Std. Error t value   Pr(>|t|)
(Intercept)                                        82.66664   0.90514  91.330    < 2e-16 ***
prejudiceenvp                                      -0.43851   0.01781 -24.628    < 2e-16 ***
LowFamilyIncomeIndicatorNot Low Income             -1.35325   0.59174  -2.287     0.02227 *
leaveUCLAyes                                        -13.84895  2.33874  -5.922   3.57e-09 ***
prejudiceenvp:leaveUCLAyes                         0.11017    0.03894   2.829    0.00469 **
LowFamilyIncomeIndicatorNot Low Income:leaveUCLAyes 0.57981   1.40869   0.412    0.68067
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple R-squared: 0.2574, Adjusted R-squared: 0.2561
F-statistic: 199.8 on 5 and 2882 DF, p-value: < 2.2e-1

Interaction Plot

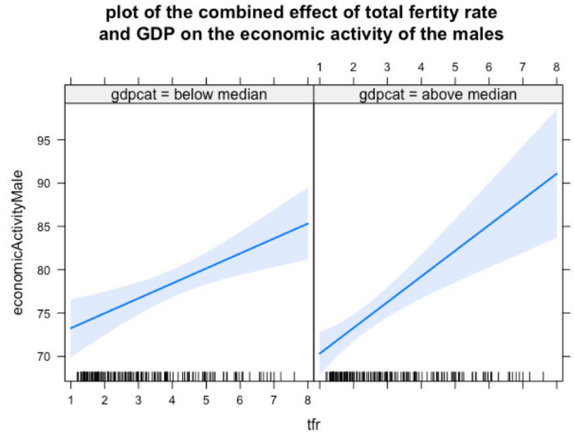
```
>interaction.plot(FIRSTGEN,leaveUCLA,prejudiceenvp,main="Interaction plot showing the
+ combined effect of being first generation and planning
+ to leave UCLA on perception of prejudce on our campus")
```

Plot three



We will now draw the interaction plot.

Library(car)
Library(effects)
Plot(Alleffects(m1), ask=FALSE)



Model Optimization

Leverage Points, Influencial Points, Outliers

Observation x_i far outside the global neighborhood of predictor values $\bar{x}$

$$\boxed{h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SSX}}$$

h_i - leverage Point

$$\boxed{h_i > 4/n \Rightarrow \text{Sig. Leverage}}$$

$$\boxed{|\text{Studentized Residuals}| \begin{cases} > 4, & n \text{ large} \\ > 2, & n \text{ small} \end{cases} \Rightarrow \text{High Influence / Outlier}}$$

$$\boxed{\text{If high leverage AND outlier} \Rightarrow \text{Bad Leverage}}$$

Cook's Distance

Summary of leverage

Calculates influence of each obs on fitted response vals

$$\boxed{D_i = \frac{e_{s_i}^2}{k+1} + \frac{h_i}{1-h_i}}$$

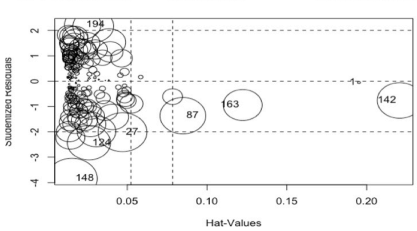
$e_{s_i}^2$ - squared standardized residuals

$$\boxed{e_{s_i}^2 = \frac{e_i}{S_e\sqrt{1-h_i}}}$$

We will also check for leverage and influential points

> influencePlot(m5,id.n=3)

	StudRes	Hat	CookD
1	-0.05171023	0.19519482	0.0001632252
27	-2.01266414	0.04698761	0.0489288427
87	-1.36603806	0.08485239	0.0430052557
124	-2.43280424	0.02598420	0.0382114025
142	-0.76090949	**0.22066653**	0.0411005873
148	**-3.84310683**	0.01533075	0.0526249088
163	-0.94289557	0.12231012	0.0309964549
194	2.22546069	0.02208641	0.0272416692




```
> model<-lm(academicenvp~prejudiceenvp+FIRSTGEN+leaveUCLA)
> library(car)
> outlierTest((model))
```

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:

	rstudent	unadjusted p-value	Bonferonni p
722	**-3.946057**	8.129e-05	**0.24168**

The Bonferroni adjusted p-value is not statistically significant. The largest studentized residual is as large as -3.95. This means we do not have any significant outliers.

Variable Selection

- 2^k possible combinations to select from

Forward Start with most sig. predictor + iteratively add predictor that creates highest R^2 change / lowest P-val

Backward Start with all predictors + iteratively remove predictor with largest P-val

Stepwise Combo of Forward / Backward

Enter All predictors are included regardless of sig. (theory / expert experience)

Blockwise, LASSO, RIDGE

Overfitting Resolutions

Akaike Information Criterion AIC

$$\boxed{AIC = \underbrace{n \cdot \log\left(\frac{RSS}{N}\right)}_{\text{Reward for good fit}} + \underbrace{2k}_{\text{Penalty for complexity}}}$$

as RSS ↑ AIC ↑

⇒ desire small AIC

Dilemma ⇒ weak penalty term (2k) so if $\underbrace{\text{large } k/n}_{\text{small n relative to k}}$ overfitting may occur

Solution ⇒

$$\boxed{AIC_{corrected} = AIC + \underbrace{\frac{2(k+1)}{N-k-1}}_{\text{Larger penalty}}}$$

Bayes Information Criterion BIC

Favors simpler models (small k) compared to AIC

$$\boxed{BIC = n \cdot \log\left(\frac{RSS}{N}\right) + \underbrace{\log\left(\frac{N}{k}\right)}_{\text{Very heavy penalty}}}$$

Sample R commands for the calculation of AIC. AIC corrected, and BIC for subset size one and two.

```
> om1 <- lm(log(Time)~log(Dwgs))
> om2 <- lm(log(Time)~log(Dwgs)+log(Spans))
> #Subset size=1
> n <- length(om1$residuals)
> npar <- length(om1$coefficients) +1
> #Calculate AIC
> extractAIC(om1,k=2)
[1]  2.00000 -94.89754
> #Calculate AICc
> extractAIC(om1,k=2)+2*npar*(npar+1)/(n-npar-1)
[1]  2.585366 -94.312171
> #Calculate BIC
> extractAIC(om1,k=log(n))
[1]  2.00000 -91.28421
> #Subset size=2
> npar <- length(om2$coefficients) +1
> #Calculate AIC
> extractAIC(om2,k=2)
[1]   3.0000 -102.3703
> #Calculate AICc
> extractAIC(om2,k=2)+2*npar*(npar+1)/(n-npar-1)
[1]   4.0000 -101.3703
> #Calculate BIC
> extractAIC(om2,k=log(n))
[1]  3.00000 -96.95036
```

We will now use backward selection based on AIC to build a model

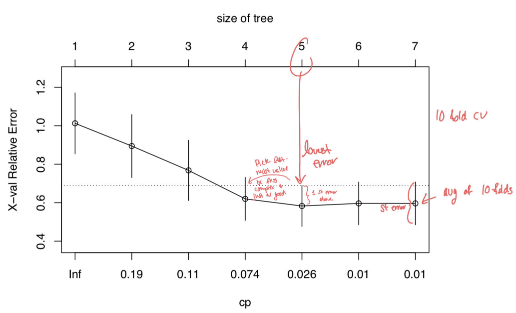
```
> m1 <- lm(log(Time)~log(DArea)+log(CCost)+log(Dwgs)+log(Length)+log(Spans))
> backAIC <- step(m1,direction="backward", data=bridge)
```

	Df	Sum of Sq	RSS	AIC
- log(Length)	1	0.00607	3.8497	-100.640
- log(DArea)	1	0.01278	3.8564	-100.562
<none>			3.8436	-98.711
- log(CCost)	1	0.18162	4.0252	-98.634
- log(Spans)	1	0.26616	4.1098	-97.698
- log(Dwgs)	1	1.45358	5.2972	-86.277

Selection Procedure

Consider data with P Predictors

1. Find best model for each individual Predictor
2. Compare R^2_{adj} for all models of different complexity ($k = 1, 2, 3, \ldots, P$)
3. Pick the k combination with min R^2_{adj} and choose P-1 model (bc it's just as accurate + much less complex than P)



Model Selection

Predictor	Outcome	model	Notes
num	num	SLR, Test of Slope	· Test of Slope if asked for Relationship · SLR if asked for Prediction · SLR if cat has = 2 levels · MLR if ordinal or if >2 levels
cat	num	One-way ANOVA, MLR, SLR	
num + cat	num	multivariate Reg	
cat	cat	Chi-Square	
cat + cat	num	Two-way ANOVA, MLR	
cat + num	num	MLR	

Logistic

- used to classify outcome. Response is num prob. of classification.
- data is fit to linear Reg model then squeezed into sigmoid function to classify.

glm(y ~ . , family = "binomial")

$$f(x) = \frac{1}{1 + e^{-x}}$$

Types
- Binary outcome with 2 levels (yes/no)
- Multinomial outcome with 3+ levels (vegan, not vegan, veggie)
- Ordinal outcome with 3+ ranked levels (1 → 5 rating)

Odds Ratio

	college	HS	
happy	2	3	5
sad	4	5	9
	6	8	14

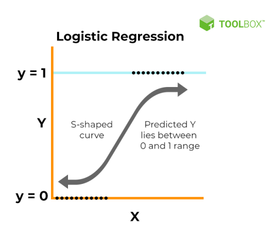
$P(HS) = 8/14$

$P(college) = 6/14$

$$\frac{P(HS)}{P(college)} = \frac{8/\cancel{14}}{6/\cancel{14}} = 8/6 = 4/3 \approx 1.3\bar{3}$$

Log odds

$$\boxed{B_i = \log\left(\frac{p}{1-p}\right)}$$



$$B_i = \log\left(\frac{8/14}{6/14}\right) = \log(4/3)$$

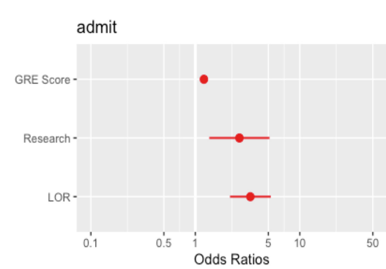
Interpretation should exponentiate coefficients to interpret exp(coef(model))

95% Confidence Interval

Plot of Odds

library(sjPlot)

plot_model(m1)



- All > 1 so all inc Admissions
- GRE doesnt have a clear conf Int bc it has a wide range

Accuracy

Null Deviance - Residual of intercept-only model

Residual Deviance - Residual of model with all predictors

$$\boxed{\text{Pseudo-}R^2 = 1 - \frac{\text{Residual Deviance}}{\text{Null Deviance}}}$$

✱ Caution: Pseudo-R^2 value has NO mathematical meaning so we use confusion matrix to calculate Log Reg model accuracy.

Confusion matrix

a table that describes the performance of a classification model

Actual	Predicted		Colum totals
	No	Yes	
No	60 TRUE NEGATIVE	12 FALSE POSITIVE TYPE I ERROR	72
Yes	8 FALSE NEGATIVE TYPE II ERROR	120 TRUE POSITIVE	128
Row totals	68	132	200

```
m1 <- glm(y~x, data)
p <- predict(m1, newdata)
table(p, y)
```

$$\boxed{\text{Accuracy Rate} = \frac{TP + TN}{\text{Total}}}$$

overall, how accurate is our model?
Gives "Predictive Power"

$$\boxed{\text{Misclassification/Error Rate} = 1 - \text{Accuracy Rate} = \frac{FP + FN}{\text{Total}}}$$

overall, how inaccurate is our model?

$$\boxed{\text{TP (Sensitivity) Rate} = \frac{TP}{\text{Actual Yes}}}$$

when its Actually Yes, how accurate is the prediction?

$$\boxed{\text{FP Rate} = \frac{FP}{\text{Actual NO}}}$$

when its Actually NO, how inaccurate is the prediction?

$$\boxed{\text{Specificity} = 1 - \text{FP Rate} = \frac{TN}{\text{Actual NO}}}$$

When its Actually NO, how accurate is the prediction?

$$\boxed{\text{Precision} = \frac{TP}{\text{Predicted Yes}}}$$

When it Predicts Yes, how accurate is the prediction?

$$\boxed{\text{Prevalence} = \frac{\text{Actual Yes}}{\text{Total}}}$$

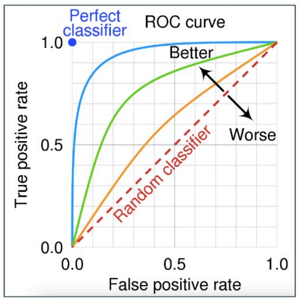
How often does the Yes condition occur in the sample?

ROC Curves

- Reciever Operating Characteristic Curves
- Shows model's performance at all classification thresholds
- Determines the best cutoff for classifying 0/1 Success/Failure
- Accuracy is measured by the Area under the Curve (AUC)

AUC = 0.9 - 1 Excellent fit
0.8-0.9 Good
0.7-0.8 Fair
0.6-0.7 Poor
0.5-0.6 Fail

Sensitivity



Specificity

```
library(ROCR)
roc <- performance(pred, "tpr", "fpr")
plot(roc, colorize = T)
```

Procedure Summary

1. Create Predictions
2. histogram of predictions hist(pred)
3. mode is a good starting threshold *
4. Confusion matrix table(actual y, pred > 0.1)
5. Calculate Accuracy

* use ROC to determine best threshold

multivariate

- multiple numerical outcomes
- If predictors are different types
- If correlations are high btw outcomes. If low we could do MLR multiple times.

```
m1 <- lm(cbind(var1, var2) ~ var3 + var4, data)
```

Correlation table

```
scatterplotMatrix(~ var1 + var2 + var3 | var4, data)
```

MANOVA

`Anova(m1)` where m1 is multivariate model

- to examine the effect of categorical predictors + their combined effect on multiple numerical variables that have collinearity

  * can't handle cat + num predictors ⇒ multivariate reg can

Univariate ANOVA

one categorical predictor w/ 2+ levels with many num

Factorial ANOVA

2+ factors

```
y <- cbind(y1, y2, y3)
m1 <- manova(y ~ x1 * x2, data)
summary(m1, test = "Pillai")
summary.aov(m1)
```

Interaction Effect Plot

```
library(car)
library(effects)
plot(allEffects(m1), ask = F)
```

Ordinal Regression

- Suitable if Outcome is Ranked
- Assumes distances btw ANY two values are equal. diff(1,5) = diff(1,2)

Likter Scale   Strongly Agree → Strongly Disagree

```
library(MASS)
m1 <- polr( y ~ x1 + x2 )
summary (m1)

coef <- coef(summary(m1))
p_value <- pnorm( abs( coef[ , "t value"]) , lower.tail = F) * 2
```

Distributions

z - Stat

t - Stat

F - Stat

χ^2 - Stat