$$y = f(x) + e$$

Chapter 3

Linear Regression

3.1 Simple linear regression

- Model: (Y, X) are random variables that are related via $Y = \beta_0 + \beta_1 X + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$.

  - What are parameters, inputs/predictor/explanatory variable/independent variable, output/response/dependent variable, and error?

Parameters: B1, Bo, Sigma^2 They are fixed, but need to be estiated.

Dependent variable/output/response = Y.

input/predictor/explanatory variable = X

epsilon = error.

- Data: (y_i, x_i) are sampled from the model above so that $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where ϵ_i are iid $N(0, \sigma^2)$.

$$y_i = f(x_i) + e_i$$
$$= \beta_0 + \beta_1 x_i + e_i$$

Example 1. *The sales of a product and amount spent on TV, radio, newspaper ads was measured across 200 different markets.*

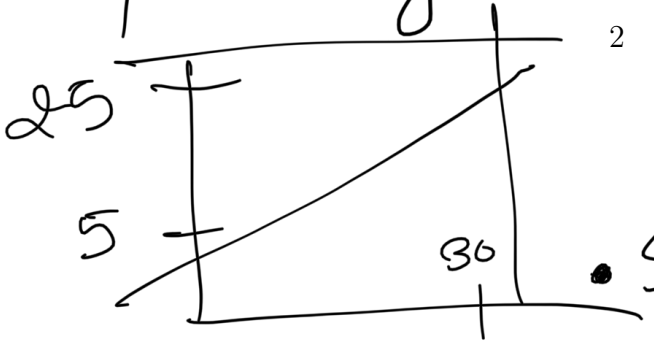
- Identify parameter values, input/predictor/explanatory/independent variable, output/response/dependent variable, and errors.

Code

```
> # Simple linear model
> str(Advertising)
'data.frame':   200 obs. of  4 variables:
 $ TV       : num  230.1 44.5 17.2 151.5 180.8 ...
 $ radio    : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...
 $ newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
 $ sales    : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
> attach(Advertising)
> plot(TV, sales)
> abline(b=25/300, a=5)
```

$Y = 5 + 25/300 X_1 + e$

- Str gives structure of data.
- attach attaches data to R.
- Use data from clipboard
  ↳ read.table("clipboard", header = T)
- Sales → Dependent Variable.
  Explanatory variable → TV, Radio, newspaper.



- Slope = Rise/Run

3.1.1 Least Squares Estimates

- Least Squares Estimates (LSE) of β_0 and β_1 minimize the residual sums of squares

$\hat{y}_i =$ Bo + B1*X

$e_i =$ Yi - (Bo + B1*Xi) (Residual)

RSS = Sum (e^2) = Sum(yi - (Bo + B1*X1))

$$\sum e_i^2 = \sum_i (y_i - (\beta_0 + \beta_1 \cdot x_i))^2$$

- For the simple linear model LSE are

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\hat{\beta}_1 = SS_{xy}/SS_{xx} =$ Sum((X1- Xbar) (yi- ybar))/Sum((Xi- xbar)^2)

- Could you use some calculus to find $\hat{\beta}_1$ and $\hat{\beta}_0$?

$\frac{\partial Q}{\partial \beta_0} = 0$ (I) $\quad \frac{\partial Q}{\partial \beta_1} = 0$ (II)

Solve to get $\hat{\beta}_1$ & $\hat{\beta}_0$.

- Write the model in matrix notation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. What is $\hat{\boldsymbol{\beta}}$?

Y1 = Bo + B1*x1 + e1
y2 = Bo + B1*x2 + e2
.
.
.
yn= Bo + B1*Xn + en

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$n \times 2$ $\quad 2 \times 1$

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

$+ \sum \epsilon_n$

- $\hat{\sigma}^2 = RSS/(n-2) = MSE$ is called the mean squared error
- $\hat{\sigma} = \sqrt{MSE} = RSE$ is called the residual standard error

- Example: Identify $\hat{\beta}_0, \hat{\beta}_1$, (y_1, x_1), $\hat{y}_1$, $\hat{e}_1$, $\mathbf{y}, \mathbf{X}, \hat{\boldsymbol{\beta}}$, RSE and MSE.

Code

```
> # Least Squares Estimates
> TVmodel<-lm(sales~TV)
> coef(TVmodel)
(Intercept)          TV
 7.03259355  0.04753664
> abline(TVmodel, lwd =5)
> e <- resid(TVmodel)
> yhat<-predict(TVmodel)
> cbind(sales, TV, yhat,e)[1:3,]
  sales   TV     yhat      e
1  22.1 230.1 17.970775 4.129225
2  10.4  44.5  9.147974 1.252026
3   9.3  17.2  7.850224 1.449776
> summary(TVmodel)

Call:
lm(formula = sales ~ TV)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

lm(y ~ x)

[Rows, Column]

$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1$

$= 7.03 + 0.48 \times 230.1$

$= 17.97$

$e = y - \hat{y} = 22.01 - 17.97$

$= 4.12$

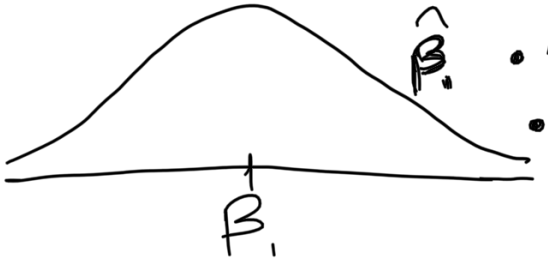
$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_1$

$= 7.03 + 0.48 \cdot 44.5 = 9.14$

$e = |9.14 - 10.40| = 1.25$

3.1.2 Bias and Variance

Question: If we repeat the experiment will we get the same parameter estimates? What are reasonable values of β_0 and β_1 based on the least squares estimates?



- mean follow the parametric distribution
- lm is always unbiased.

- Least squares estimators are unbiased: $E[\hat{\beta}_i] = \beta_i$.
- The variance of the least squares estimators in are the diagonal elements of

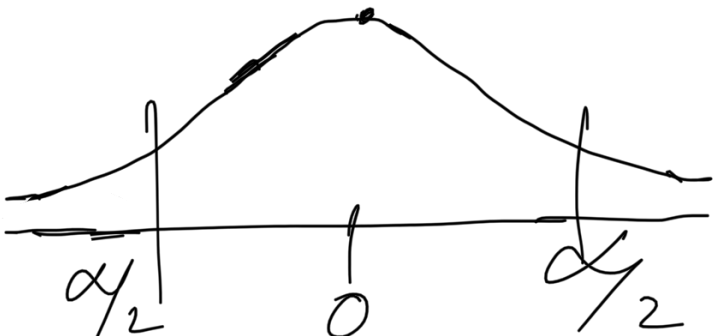
  $Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$ $= \sigma^2 \begin{pmatrix} \times & 0 \\ 0 & \times \end{pmatrix}$ — $Var(\hat{\beta}_0)$, $Var(\hat{\beta}_1)$
- The standard error $SE(\hat{\beta}_i)$ is the standard deviation of $\hat{\beta}_i$. We plug in MSE for σ^2 above.
- You could verify using the matrix formula that

  $SE(\hat{\beta}_1)^2 = \hat{\sigma}^2 / \sum_i (x_i - \bar{x})^2$

  $SE(\hat{\beta}_0)^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right]$

3.1.3 Inference for parameters

- A 95% confidence interval for β_i is $\hat{\beta}_i \pm 2\hat{SE}(\hat{\beta}_i)$ $= \hat{\beta}_1 \pm 1.96\ SE(\hat{\beta}_0)$
- To test the null hypothesis $H_0 : \beta_i = 0$ against $H_a : \beta_i \neq 0$ we
  1. Compute a test statistic $T = \frac{\hat{\beta}_i - 0}{SE(\hat{\beta}_i)}$,
  2. compute a p-value $= \Pr(|T_{n-2}| \geq |T|)$,



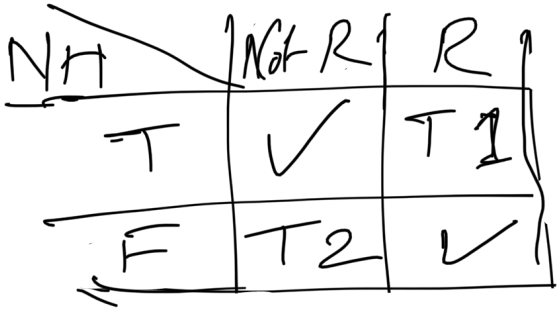
3. reject H_0 if the p-value is small.

- A correct conclusion, type I error, or type II error could be made for a hypothesis test. The power, type I error rate α, and type II error rate describe the properties of the test.

- Identify and interpret the standard error, p-value and confidence interval for β_1.

Code

```
> summary(TVmodel)

Call:
lm(formula = sales ~ TV)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16

> confint(TVmodel)
                2.5 %     97.5 %
(Intercept) 6.12971927 7.93546783
TV          0.04223072 0.05284256
```



β_0 (Intercept) β_1 TV

$SE(\beta)$ $t = \frac{\beta - 0}{SE(\beta)}$

$H_0 : \beta = 0$

$H_A = \beta \neq 0$

we reject H_0 because $p \leq 0.05$. Sales is dependent on TV ads.

We are 95% confident that β_1 is between 0.42 & 0.52. That is the increase in sales per unit increase in TV revenue is between 0.42 - 0.52 with 95% confidence.

(Define β_1 if you use this interpretation in H/W).

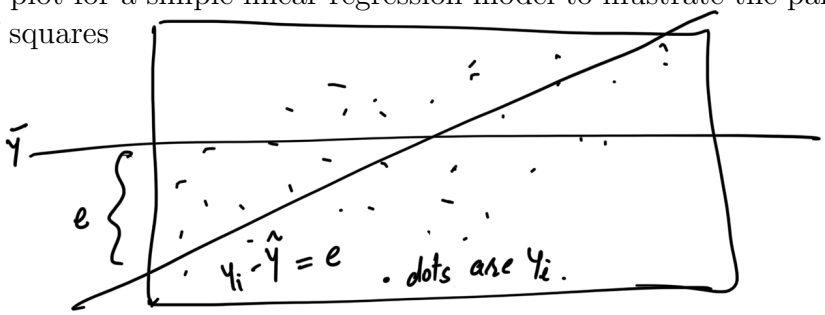
3.1.4 Assessing model accuracy

- In ANOVA we partition the total sums of squares (TSS) into residual sums of squares (RSS) and remaining sums of squares TSS - RSS and organize the results in a table.

$$TSS = \sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 = RSS + (TSS - RSS)$$

	SS	df	F statistic
Model	TSS - RSS	1	$\frac{(TSS - RSS)/df}{RSS/df}$ $(df = 1)$
Residual	RSS	$n-2$	$(df = (n-2)$
Total	TSS	$n-1$	$\sim F_{(1, n-2)}$

- Draw a scatter plot for a simple linear regression model to illustrate the partitioned sums of squares



- How do you interpret the coefficient of determination? $R^2 = \frac{TSS-RSS}{TSS}$

Note this R-squared explain VARIATION, not VARIANCE.

$R^2 = \frac{TSS - RSS}{TSS}$

$\Rightarrow$ The proportion of variation explained by the model. ie. variation explained by TV ad on sales revenue.

- How should the residual standard error $RSE = \sqrt{RSS/(n-2)}$ be interpreted?

$SE = \sqrt{sd} = \sqrt{RSS/(n-2)}$

- Identify and interpret the ANOVA table, RSE to the R^2 and correlation r. Interpret the RSE and R^2.

Code

```
> anova(TVmodel)
Analysis of Variance Table

Response: sales
           Df Sum Sq Mean Sq F value    Pr(>F)
TV          1 3314.6  3314.6  312.14 < 2.2e-16 ***
Residuals 198 2102.5    10.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(TVmodel)
Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
> cor(TV, sales)
[1] 0.7822244
> cor(TV, sales)^2
[1] 0.6118751
```

3.2 Multiple linear regression

(Continuous y).

- Model: (Y, X), where $X^T = (X_1, X_2, ..., X_p)$, are random variables satisfying $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$.

  - What are parameters, inputs/predictor/explanatory variable/independent variable, output/response/dependent variable, and error?

Parameters : $\beta_0, \beta_1, ... \beta_p, \sigma^2$.

Inputs : $X_1, X_2 ... X_p$

Output : Y

Error : e, $e \sim N(0, \sigma^2)$.

- Data: (y_i, x_i), where $x_i^T = (x_{i1}, x_{i2}, ..., x_{ip})$, satisfy $y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i$ where ϵ_i are iid $N(0, \sigma^2)$.

- Can you write $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$?

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix}$$

3.2.1 Least squares estimates

- Least Squares Estimates (LSE) of $\beta_0, \beta_1, ..., \beta_p$ minimize the residual sums of squares

$\hat{y}_i = f(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$

$e_i = y_i - \hat{y}_i = y_i - \left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \right)$

RSS = $\sum_i e^2 = \sum_i \left[y_i - (\beta_0 + \beta_1 x_i + \cdots + \beta_p x_p)\right]^2$

– What is $\hat{\boldsymbol{\beta}}$ in matrix notation?

$\hat{\beta} = (X'X)^{-1}(XY)$

- $\hat{\sigma}^2 = RSS/(n-p-1) = MSE$ is called the mean squared error (MSE)
- $\hat{\sigma} = \sqrt{MSE} = RSE$ is called the residual standard error (RSE)
- The variance of $\hat{\beta}_i$ is the corresponding diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}\hat{\sigma}^2$ and the standard error is its square root.
- A confidence interval for β_i is $\hat{\beta}_i \pm 2\hat{SE}(\hat{\beta}_i)$. 95% Roughly (1.96 to be exact instead of 2).
- A p-value for testing $H_0 : \beta_i = 0$ is by
  1. computing $T_i = \frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)}$,
  2. computing p-value $= \Pr(|T_{n-p-1}| \geq |T_i|)$

$(X'X)^{-1} =$ $\begin{matrix} & \hat{\beta}_0 & \hat{\beta}_1 & \cdots & \hat{\beta}_p \\ \hat{\beta}_0 & & & & \\ \hat{\beta}_1 & & & \text{Covariance} & \\ \vdots & & & & \\ \hat{\beta}_p & & \text{Covariance} & & \end{matrix}$ Variances in diagonal

This is variance-covariance matrix.

- Write out the model being fit below and identify/interpret $\hat{\beta}_j$, $\mathbf{y}, \mathbf{X}, \hat{\boldsymbol{\beta}}$, RSE, MSE, $\hat{SE}(\hat{\beta}_j)$, p-values and confidence intervals.

$sales = \beta_0 + \beta_1 \cdot TV + \beta_2 \cdot radio + \beta_3 \cdot newspaper$

Code

```
> ## Multiple linear regression
> # Least squares estimates
> Advertising[1:3,]
     TV radio newspaper sales
1 230.1  37.8      69.2  22.1
2  44.5  39.3      45.1  10.4
3  17.2  45.9      69.3   9.3
> Adlm<-lm(sales~TV + radio + newspaper, data = Advertising)
> summary(Adlm)

Call:
lm(formula = sales ~ TV + radio + newspaper, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
radio        0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
> X<-as.matrix(cbind(1,Advertising[,1:3]))
> X[1:3,]
     1    TV radio newspaper
[1,] 1 230.1  37.8      69.2
[2,] 1  44.5  39.3      45.1
[3,] 1  17.2  45.9      69.3
> y<-Advertising$sales
```

$y = \underset{\hat{\beta}}{2.94} + \underset{SE(\hat{\beta})}{0.046}\, TV + 0.189\, radio - 0.001\, Newspaper + e.$

RSE $MSE = 1.686^2 =$

- for unit increase in TV ad. revenue increases by 0.046 unit.

```
> solve(t(X)%*%X)%*%t(X)%*%y
                  [,1]
1           2.938889369
TV          0.045764645
radio       0.188530017
newspaper -0.001037493
> solve(t(X)%*%X)*1.686^2
                       1             TV         radio     newspaper
1           0.0973432781 -2.658817e-04 -1.116138e-03 -5.913647e-04
TV         -0.0002658817  1.946868e-06 -4.472992e-07 -3.267848e-07
radio      -0.0011161376 -4.472992e-07  7.419644e-05 -1.781097e-05
newspaper -0.0005913647 -3.267848e-07 -1.781097e-05  3.448878e-05
> sqrt(diag(solve(t(X)%*%X)*1.686^2))
          1          TV       radio   newspaper
0.311998843 0.001395302 0.008613735 0.005872715

> confint(Adlm)
                 2.5 %     97.5 %
(Intercept)  2.32376228 3.55401646
TV           0.04301371 0.04851558
radio        0.17154745 0.20551259
newspaper   -0.01261595 0.01054097
```

---

solve(t(x) %*% x)(t(x) %*% y) ⇒ Gives β coefficients.

solve(t(x) %*% x) × 1.686^2 ⇒ Gives variance covariance matrix.

sqrt(diag(solve(t(x) %*% x) × 1.686^2)) ⇒ Gives standard error.

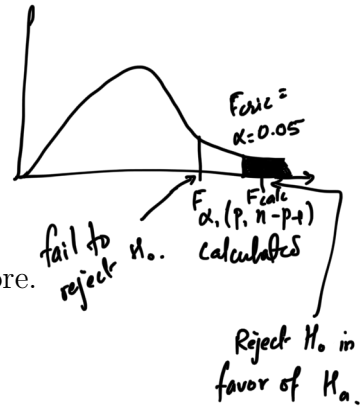
3.2.2 Some typical inferences

Overall ANOVA: Is there a relationship between sales and the predictors?

- The ANOVA table partitions TSS = RSS + (TSS - RSS) as before

	SS	df	MS (mean squares)	F statistic
Model	SST - RSS	÷ p	= MS_{model}	→ $F = \frac{MS_{model}}{MSE}$
Residual	RSS	÷ (n-p-1)	= MSE	
Total	TSS	n-1		

- We can test the overall null $H_0 : \beta_1 = ... = \beta_p = 0$ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$ if β_s are equal, then $Y = f(\varepsilon)$.
  - by comparing $F = \frac{(TSS-RSS)/p}{RSS/(n-p-1)}$ to $F_{\alpha,p,n-p-1}$ or
  - by comparing p-value = $\Pr(F_{p,n-p-1} \geq F)$ to α.



- The R^2 statistic can be interpreted as before or as $R^2 = Cor(y_i, \hat{y}_i)^2$
- The $RSE = \sqrt{RSS/(n-p-1)} = \sqrt{MSE} = \hat{\sigma}$ is interpreted as before.

- Example: Test the overall null hypothesis below and interpret the R^2.

Code

```
> summary(Adlm)

Call:
lm(formula = sales ~ TV + radio + newspaper, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
radio        0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

$Sales = \beta_0 + \beta_1 \cdot TV + \beta_2 \cdot radio + \beta_3 \cdot newspaper + \varepsilon$

↳ $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

H_A : One of them is different.

F-stat = 570.30, p-value $\leq$ 0.001.

so, H_0 is rejected as p-value $\leq$ 0.001.

which means, Atleast one of the βs is different and not zero.

Implication : Because βs are not zero, it is important to include them in the model.

- To test $H_0: \beta_3 = 0$ & $H_a: \beta_3 \neq 0$. (This procedure can be used to test any $\beta = 0$ or $\beta_1 + \beta_2 = 0$, $\beta_1 = \beta_2$, $\beta_1 = 2\beta_2$ etc)

Construct full model : $y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

Reduced model : $y_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ (No X_3 here).

Get RSS from both models. RSS from full model & RSS_r from reduced model & do f-test.

$$F\text{-stat} = \frac{(RSS - RSS_r)/1}{RSS/(n-p-1)}$$

. Then find $F_{critical}$ at df $(1, n-p-1)$ from f-table. If $F_{calc} \geq F_{crit}$, Reject H_0 that $\beta_3 = 0$. So, include in model.

- We can test $H_0 : \beta_3 = 0$ with an F test by

  1. Computing the partial F statistic $F = \frac{(RSS_0 - RSS)/1}{RSS/(n-p-1)}$, where RSS_0 is the residual sum of squares for the reduced model (under the null hypothesis).

  2. Comparing F to $F_{\alpha,1,n-p-1}$ or comparing the p-value $= \Pr(F_{1,n-p-1} \geq F)$ to α.

  3. Using the T test for $H_0 : \beta_3 = 0$ since $T^2_{n-p-1} = F_{1,n-p-1}$

- Compare the F statistic and its p-value to the T statistic -.177 and its p-value 0.86 and interpret the results.

Code

```
> Adlm<-lm(sales~TV + radio + newspaper, data = Advertising)
> Adlm2<-lm(sales ~ TV + radio, data = Advertising)
> anova(Adlm2,Adlm)
Analysis of Variance Table

Model 1: sales ~ TV + radio
Model 2: sales ~ TV + radio + newspaper
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    197 556.91
2    196 556.83  1  0.088717 0.0312 0.8599
> summary(Adlm)$coef
                Estimate  Std. Error    t value     Pr(>|t|)
(Intercept)  2.938889369 0.311908236  9.4222884 1.267295e-17
TV           0.045764645 0.001394897 32.8086244 1.509960e-81
radio        0.188530017 0.008611234 21.8934961 1.505339e-54
newspaper   -0.001037493 0.005871010 -0.1767146 8.599151e-01
```

- we can retain $H_0 : \beta_3 = 0$ because p-value = 0.86 is large. That is, newspaper does not need to be included in the model.
- So, if the p-value is sufficiently large in full model, we can decide to retain or remove that variable based on that p-value. This is easy way to decide model.

- If we want to test the null hypothesis that q of parameters are equal to 0 we can

  1. compute $F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$, where RSS_0 is the residual sum of squares for the reduced model which assumes q of the parameters are 0 (the null hypothesis), and

  2. Compare F to $F_{\alpha,q,n-p-1}$ or compare the p-value $= \Pr(F_{q,n-p-1} \geq F)$ to α.

- Write out the full model and the null hypothesis below and interpret the results

Code

```
> Adlm<-lm(sales~TV + radio + newspaper, data = Advertising)
> Adlm3<-lm(sales~TV, data = Advertising)
> anova(Adlm3,Adlm)
Analysis of Variance Table

Model 1: sales ~ TV
Model 2: sales ~ TV + radio + newspaper
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    198 2102.53
2    196  556.83  2    1545.7 272.04 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Full : $sales = \beta_0 + \beta_1 \cdot TV + \beta_2 \cdot radio + \beta_3 \cdot newspaper.$

Red : $sales = \beta_0 + \beta_1 \cdot TV$

$H_0 : \beta_2 - \beta_3 = 0$ or $H_0 : \beta_2 = \beta_3 = 0$ $H_a : \beta_2 \neq \beta_3 \neq 0$.

Reject H_0 as p_value ≤ 0.001.

Atleast one variable (radio/news/both) should be included in model.

Partial F tests: Which combination of predictors is the correct model?

- In backwards selection, we remove the most insignificant predictors one at a time until all predictors are significant. What is the final model below?

Code

```
> # Backward selection
> # Step 1
> summary(lm(sales~TV + radio + newspaper, data = Advertising))$coefficients
               Estimate  Std. Error    t value     Pr(>|t|)
(Intercept)  2.938889369 0.311908236  9.4222884 1.267295e-17
TV           0.045764645 0.001394897 32.8086244 1.509960e-81
```

Step 1 Remove 'news' as it has biggest p-value.

(arrangement of coeff ?).

```
radio        0.188530017 0.008611234 21.8934961 1.505339e-54
newspaper   -0.001037493 0.005871010 -0.1767146 8.599151e-01
> # Step 2
> summary(lm(sales~TV + radio, data = Advertising))$coefficients
              Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 2.92109991 0.294489678  9.919193 4.565557e-19
TV          0.04575482 0.001390356 32.908708 5.436980e-82
radio       0.18799423 0.008039973 23.382446 9.776972e-59
```

step 2
stop! b/c all variables are significant.

final model:

$sales = \beta_0 + \beta_1 \cdot TV + \beta_2 \cdot Radio$

- In forwards selection, we add predictors one at a time. What is the final model?

Code

```
> # Forward selection
> # Step 1
> summary(lm(sales~ TV, data = Advertising))$coefficients
             Estimate  Std. Error  t value    Pr(>|t|)
(Intercept) 7.03259355 0.457842940 15.36028 1.40630e-35
TV          0.04753664 0.002690607 17.66763 1.46739e-42 ←
> summary(lm(sales~ radio, data = Advertising))$coefficients
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 9.3116381 0.56290050 16.542245 3.561071e-39
radio       0.2024958 0.02041131  9.920765 4.354966e-19 ←
> summary(lm(sales~newspaper, data = Advertising))$coefficients
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 12.3514071 0.62142019 19.876096 4.713507e-49
newspaper    0.0546931 0.01657572  3.299591 1.148196e-03 ←
>
> # Step 2
> summary(lm(sales~TV + radio, data = Advertising))$coefficients
             Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 2.92109991 0.294489678  9.919193 4.565557e-19
TV          0.04575482 0.001390356 32.908708 5.436980e-82
radio       0.18799423 0.008039973 23.382446 9.776972e-59
> summary(lm(sales~TV + newspaper, data = Advertising))$coefficients
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 5.77494797 0.52533779 10.992828 3.145860e-22
TV          0.04690121 0.00258086 18.172707 5.507584e-44
newspaper   0.04421942 0.01017410  4.346276 2.217084e-05
>
> # Step 3
> summary(lm(sales~TV + radio + newspaper, data = Advertising))$coefficients
               Estimate  Std. Error    t value     Pr(>|t|)
(Intercept)  2.938889369 0.311908236  9.4222884 1.267295e-17
TV           0.045764645 0.001394897 32.8086244 1.509960e-81
radio        0.188530017 0.008611234 21.8934961 1.505339e-54
newspaper   -0.001037493 0.005871010 -0.1767146 8.599151e-01 ← N/S.
```

Among these three take model with smallest p. value & build model by adding other variables with that variable in the model.

In step 2, Sales vs radio is taken from step 1 & model expanded.

Step 3: Best model from step 2 is taken & third variable was added but no improvement.

- Mixed selection combines forward and backward. After a forward step, we consider removing variables that are no longer statistically significant. You could get caught in a loop here.

- Information criteria (AIC, BIC, Mallows Cp, adjusted R^2) measures estimate the distance from the true model to a given model and select the best model among all 2^p models. 2^p is a big number.

Interval Estimation $\hat{f}(x) \pm 2\,SE(\hat{f}(x))$ or $\hat{f}(x) \pm t_{\alpha/2}\, SE(\hat{f}(x))$

- For a given value of $x = (x_1, ..., x_p)$ a confidence interval for $f(x) = \beta_0 + \beta_1 x_1 .... \beta_p x_p = (1, x_1, ..., x_p)\boldsymbol{\beta} = x^T\boldsymbol{\beta}$, is

$$\hat{f}(x) \pm 2SE(\hat{f}(x)) \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

  - $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 .... \hat{\beta}_p x_p = x^T\widehat{\boldsymbol{\beta}}$ is an unbiased point estimate for $f(x)$

  - $SE(\hat{f}(x))^2 = \widehat{Var}(x^T\hat{\boldsymbol{\beta}}) = x^T\widehat{Var}(\hat{\boldsymbol{\beta}})x = \hat{\sigma}^2 x^T(\mathbf{X}^T\mathbf{X})^{-1}x$

- For a given value of $x = (x_1, ..., x_p)$ a prediction interval for $y = f(x) + \epsilon = \beta_0 + \beta_1 x_1 .... \beta_p x_p + \epsilon$ is

$$\hat{y} \pm 2SE(\hat{y})$$

$SE = \sqrt{Var}$

$SE(\beta) = \sqrt{Var(\beta)}$

  - $\hat{y} = \hat{f}(x) = x^T\widehat{\boldsymbol{\beta}}$

  - $SE(\hat{y})^2 = Var(\hat{f}(x) + \epsilon) = \hat{\sigma}^2(1 + x^T(\mathbf{X}^T\mathbf{X})^{-1}x)$

- What is the difference between the interpretation and the width of these two intervals?

$SE(\hat{y}) = \hat{\sigma}^2 + \hat{\sigma}^2 x(XX^T)^{-1}x^T$

$Var(\epsilon) + Var(\hat{f}(x))$

↓

Irreducible error.

Note: Confidence interval is for $f(x)$. Prediction interval is for new value of y (range to predict y).

- Identify and interpret the intervals below and describe their width vs. irreducible error.

Code

```
> Admodel<-lm(sales ~ TV + radio, data = Advertising)
> summary(Admodel)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.92110    0.29449   9.919   <2e-16 ***
TV           0.04575    0.00139  32.909   <2e-16 ***
radio        0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16

>
> predict(Admodel, newdata = data.frame(TV = c(5,25), radio = c(10, 15)),
+ interval = "confidence")
       fit      lwr      upr
1 5.029816 4.537977 5.521655
2 6.884884 6.461434 7.308334
>
> predict(Admodel, newdata = data.frame(TV = c(5,25), radio = c(10, 15)),
+ interval = "prediction")
       fit      lwr      upr
1 5.029816 1.677760  8.381872
2 6.884884 3.542178 10.227590
```

Confidence interval for $f(x)$. No reducible error

help(predict.lm).

$5.029 = 2.92 + 0.045 \times 5 + 0.1879 \times 10$.

$\hat{f}(x) \pm 2\sqrt{Var(\hat{f}(x))}$ (no error)

(Has error)

fmla: $\hat{f}(x) \pm 2\sqrt{Var(\hat{f}(x) + \epsilon}$

I am 95% confident that the mean sales among all media market that spent 5 on TV & 10 on radio is between 4.53 to 5.52 CI.

We are 95% confident that a media market that spends 5 on TV & 10 on radio will have sales between 3.54 to 10.22.

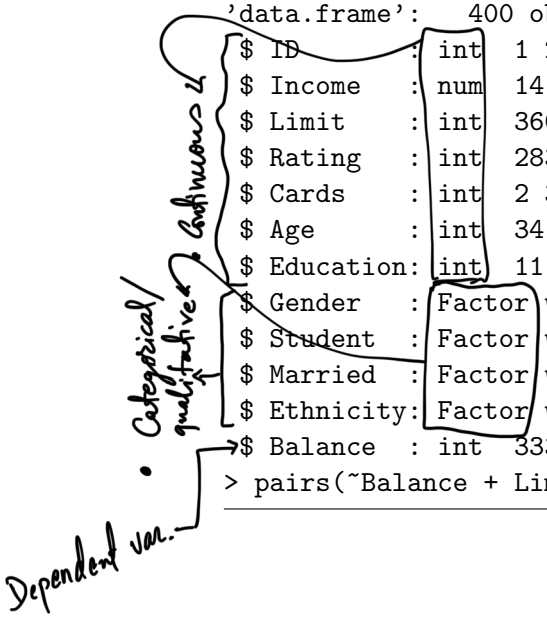
3.2.3 ANOVA: Qualitative Predictors

Example 2. *The Credit data set in the ISLR package contains information on ten thousand customers. The aim here is to predict Y = credit card balance or "balance".*

- Identify categorical / qualitative predictors and continuous / quantitative predictors.

Code

```
> library(ISLR)
> data(Credit)
> str(Credit)
'data.frame':   400 obs. of  12 variables:
 $ ID        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Income    : num  14.9 106 104.6 148.9 55.9 ...
 $ Limit     : int  3606 6645 7075 9504 4897 8047 3388 7114 3300 6819 ...
 $ Rating    : int  283 483 514 681 357 569 259 512 266 491 ...
 $ Cards     : int  2 3 4 3 2 4 2 2 5 3 ...
 $ Age       : int  34 82 71 36 68 77 37 87 66 41 ...
 $ Education: int  11 15 11 11 16 10 12 9 13 19 ...
 $ Gender    : Factor w/ 2 levels " Male","Female": 1 2 1 2 1 1 2 1 2 2 ...
 $ Student   : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 2 ...
 $ Married   : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 1 1 1 2 ...
 $ Ethnicity: Factor w/ 3 levels "African American",..: 3 2 2 2 3 3 1 2 3 1 ...
 $ Balance   : int  333 903 580 964 331 1151 203 872 279 1350 ...
> pairs(~Balance + Limit + Rating + Gender, data = Credit)
```



Qualitative Predictors with 2 levels

Example: Consider predicting Balance with Gender

- Model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - y_i is credit card balance of ith individual (Dependent var)
  - $x_i = I(\text{ i th person female})$ ⇒ 1 if female, 0, otherwise.
  - ϵ_i are iid N(0,σ^2) (Error)
  - Write out the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

- Interpret β_0, β_1, and $\beta_0 + \beta_1$ and $H_0 : \beta_1 = 0$.

β_0 ⇒ mean credit card balance for non-female ie. male.

$\beta_0 + \beta_1$ ⇒ mean cc balance for females.

β_1 ⇒ female mean − male mean.

- Least Squares Estimators are still $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. The formula for MSE, RSE, R^2, hypothesis tests, and confidence and prediction intervals are all the same. The only difference is that predictor(s) take on values 0 or 1 and parameters must be interpreted more carefully.

- Example: Identify and interpret $\hat{\beta}_i$, $\hat{\beta}_1 + \hat{\beta}_2$, R^2, RSE, $\mathbf{X}$, $\hat{\boldsymbol{\beta}}$, the p-values, and confidence interval.

Code

```
> # 2 levels for qualitative predictor
> X<-cbind(1, I(Gender == "Female"))
> Credit[1:3,]
  ID  Income Limit Rating Cards Age Education Gender Student Married Ethnicity
1  1  14.891  3606    283     2  34        11   Male      No     Yes Caucasian
2  2 106.025  6645    483     3  82        15 Female     Yes     Yes     Asian
3  3 104.593  7075    514     4  71        11   Male      No      No     Asian
  Balance
1     333
2     903
3     580
> X[1:3,]
     [,1] [,2]
[1,]    1    0
[2,]    1    1
[3,]    1    0
> y<-Balance
> solve(t(X)%*%X)%*%t(X)%*%y
          [,1]
[1,] 509.80311
[2,]  19.73312
> # Observe R will assign 1 to the second factor value
> BalanceModel<-lm(Balance ~ Gender, data = Credit)
> summary(BalanceModel)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    509.80      33.13  15.389   <2e-16 ***
GenderFemale    19.73      46.05   0.429    0.669
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 460.2 on 398 degrees of freedom
Multiple R-squared:  0.0004611, Adjusted R-squared:  -0.00205
F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
> predict(BalanceModel, newdata = data.frame(Gender = "Female"), interval = "confidence")
       fit      lwr      upr
1 529.5362 466.6493 592.4232
```

only

[attach(credit)]

$\rightarrow \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \text{male cc Balance} \\ \text{female - male cc balance} \end{pmatrix}$

retain $H_0: \beta_1 = 0$ fail to reject H_0.

or Predict

(Confidence interval for mean cc balance among female)

$\text{Var}(\hat{\beta}) = (X^T X)^{-1} \hat{\sigma}^2$

$\hat{f}(x) = x^T \hat{\beta}$

Qualitative Predictors with more than 2 levels (multinomial logistic regression).

Consider predicting Balance with Ethnicity

Code

```
> summary(Ethnicity)
African American            Asian        Caucasian
              99              102              199
```

- Model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$
  - y_i is credit card balance of ith individual
  - $x_{i1} = I(\text{ i th person Asian})$ (1 if Asian, 0 otherwise)
  - $x_{i2} = I(\text{ i th person Caucasian})$ (1 if caucasian, 0 o/w)
  - ϵ_i are iid N(0,σ^2)
  - Write out the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \begin{matrix} \text{Asian} \\ \text{Cauc} \\ \text{Af.Am} \\ \\ \\ \end{matrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \text{A} \end{matrix} \begin{matrix} \\ \\ \\ \\ \text{C} \end{matrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

- Interpret β_0, $\beta_0 + \beta_1$, β_2, $H_0 : \beta_1 = \beta_2 = 0$

β_0 : mean cc balance for Af. amer

$\beta_0 + \beta_1$: mean cc balance for Asian

$\beta_0 + \beta_2$: mean cc balance for Caucasian.

$H_0: \beta_1 = \beta_2 = 0 \Rightarrow$ The credit card balance doesn't differ between among different ethnicities.

- Example: Identify and interpret some of the least squares estimates, the individual and overall p-values, and some confidence intervals.

Code

```
> # 3 levels for qualitative predictor
> BalanceModelEth<-lm(Balance ~ Ethnicity, data = Credit)
> summary(BalanceModelEth)

Call:
lm(formula = Balance ~ Ethnicity, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max
-531.00 -457.08  -63.25  339.25 1480.50

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          531.00      46.32  11.464   <2e-16 ***
EthnicityAsian       -18.69      65.02  -0.287    0.774
EthnicityCaucasian   -12.50      56.68  -0.221    0.826
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 460.9 on 397 degrees of freedom
Multiple R-squared:  0.0002188, Adjusted R-squared:  -0.004818
F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575

> predict(BalanceModelEth,
+ newdata = data.frame(Ethnicity = c("African American","Asian", "Caucasian")),
+  interval = "confidence")
       fit      lwr      upr
1 531.0000 439.9394 622.0606
2 512.3137 422.6023 602.0252
3 518.4975 454.2699 582.7250
```

Intercept is there. So effect model.

There is no enough evidence to suggest the difference in mean credit cards balance between African American and Asian.

$\hat{f}(x) = x^T\hat{\beta}$ form.

- A means model would remove the intercept from the model above and have an indicator for each ethnic group. Write out this model for the data set above

mean model → $Y_{ij} = \mu_i + \varepsilon_{ij}$; μ_i is mean for i^{th} group & Y_{ij} is cc balance for j^{th} individual in i^{th} group.

$\varepsilon_{ij} \rightarrow$ Error.

effect → $Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$ (Isn't this singular?)

AA 1/0 AS 1/0 CA 1/0

- Verify that the confidence intervals for balance within an Ethnicity are the same across models, and determine why the R^2 and some other statistics are different.

Code

```
> BalanceModelEth2<-lm(Balance ~ Ethnicity -1, data = Credit)
> summary(BalanceModelEth2)

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
EthnicityAfrican American   531.00      46.32   11.46   <2e-16 ***
EthnicityAsian              512.31      45.63   11.23   <2e-16 ***
EthnicityCaucasian          518.50      32.67   15.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 460.9 on 397 degrees of freedom
Multiple R-squared:  0.562,     Adjusted R-squared:  0.5587
F-statistic: 169.8 on 3 and 397 DF,  p-value: < 2.2e-16
> confint(BalanceModelEth2)
                               2.5 %   97.5 %
EthnicityAfrican American 439.9394 622.0606
EthnicityAsian            422.6023 602.0252
EthnicityCaucasian        454.2699 582.7250
```

↳ Remove first column of 1s from matrix X. ie. remove intercept.

Mean model & effect model differs in interpretation of β coefficients. But result is same.

← Here, this p-value is testing whether mean cc balance for African american $\beta_1 = 0$.

i.e. $H_0: \beta_1 = 0$
$H_a: \beta_1 \neq 0$.

Summary:

$Y = f(x) + \varepsilon$
$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$
x are dummy variables. 1/0.

LSE : $\hat{\beta} : (X^T Y)(X X^T)^{-1}$

Mean model → without intercept.
Effect model → with intercept.

3.2.4 ANCOVA: Qualitative and Quantitative predictors in the Linear Model

Definition 1. *In analysis of covariance (ANCOVA) some predictors in the regression model are qualitative and some are quantitative.*

- Motivating example: Consider the plot of the Advertising data below.

Code

```
> attach(Advertising)
> median(radio)
[1] 22.9
> plot(TV,sales, pch = as.numeric(I(radio>22.9)))
> legend("topleft", inset = .05, legend = c("radio<22.9","radio>22.0"), pch = c(0,1))
> title("Sales vs. TV vs. radio")
```

- Equal slopes additive model assumes $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

  - y_i is sales, x_{i1} is TV revenue and $x_{i2} = 1$ if the ith market radio > 22.9 and 0 otherwise.

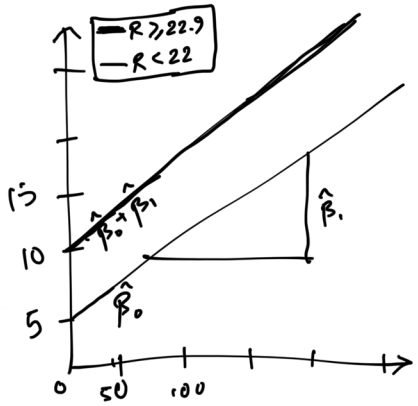
  - What are the two regression equations?

$$f(x) = \begin{cases} \beta_0 + \beta_1 x & \text{if Radio} \leq 22.90 \\ \beta_0 + \beta_2 + \beta_1 x & \text{if radio} > 22.90 \end{cases}$$

β_2: y-intercept

Equal slope model as slope for x is same.

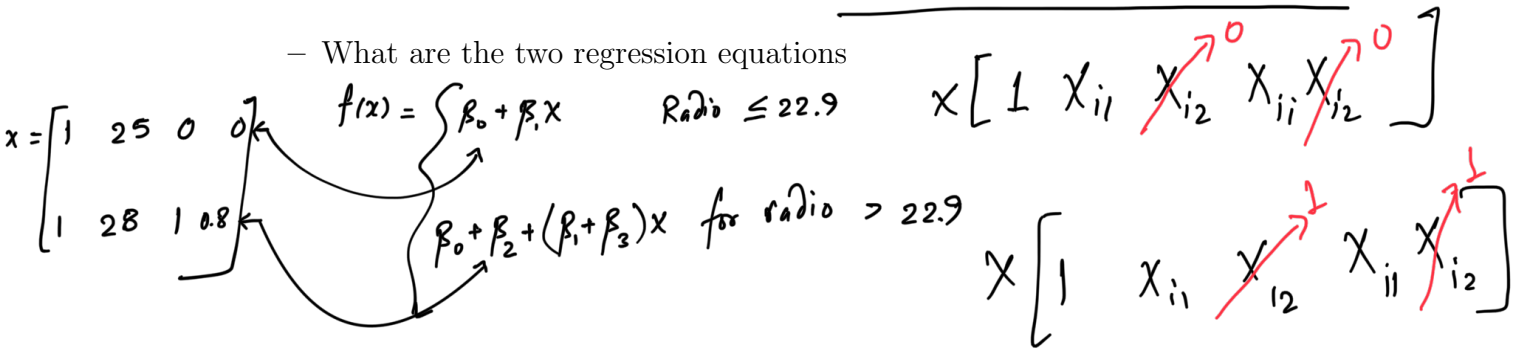
  - Example: Identify each parameter estimate on the plot.



Code

```
> cat.radio<-I(radio>22.9)
> esam<-lm(sales~TV + cat.radio)
> coef(esam)
  (Intercept)            TV cat.radioTRUE
   4.82831666    0.04622553    4.79413260
> abline(a = 4.828, b = .046, lwd =3)
> abline(a = 4.828+4.794, b = .046, lwd =3, lty =2)
```

- Unequal slopes interaction model assumes: $y_i = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i1}x_{i2}+\epsilon_i$ (interaction.)
  - What are the two regression equations



  - How do we interpret β_3. Hint we might interpret a slope here as a return rate for TV advertising.

  - Identify the two regression lines on the plot.

Code

```
> #Unequal slope and intercept
> usim<-lm(sales~TV + cat.radio + TV*cat.radio)
> coef(usim)
     (Intercept)               TV    cat.radioTRUE TV:cat.radioTRUE
      7.10049257       0.03033963       0.63621335       0.02836980
> abline(a = 7.1, b = .03, lwd =3, col =2)
> abline(a = 7.1 + .636, b = .03 + .028, lwd =3, col =2, lty=2)
```

3.2.5 Interactions in the Linear Model

- In ANCOVA we may or may not have interaction between a qualitative and quantitative predictor.
- Consider model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$

  – Where x_{i1} is TV revenue and x_{i2} is Radio revenue for the ith market. What's the difference between x_{i2} here and above?

  – We can interpret β_3 as the increase in effectiveness rate of TV advertising per unit increase in radio advertising.

  – Interpret $\hat{\beta}_3$ below and determine if it should be included in the model?

Code

```
> # Continuous interaction
> cim<-lm(sales~TV + radio + TV*radio)
> summary(cim)

Call:
lm(formula = sales ~ TV + radio + TV * radio)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3366 -0.4028  0.1831  0.5948  1.5246

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
radio       2.886e-02  8.905e-03   3.241   0.0014 **
TV:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,    Adjusted R-squared:  0.9673
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

- In general, we can have p predictors, all pairwise interactions (X_iX_j), three way interactions ($X_iX_jX_k$) and so on in a model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

& So, on.

This become complex.
So we wanna include only imp variables in the model while doing model selection.

3.2.6 Linear models and non-linear relationships

Example 3. *The Auto data set has gas mileage, horsepower, and other information for 392 vehicles. Let consider predicting mpg with horsepower.*

- Identify the response and predictor variable in the plot and write out the final estimated model.

Code

```
> data(Auto)
> names(Auto)
[1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"
[6] "acceleration" "year"         "origin"       "name"
> attach(Auto)
> plot(horsepower, mpg)
> hp.sq<-horsepower^2
> qm.hp<-lm(mpg~horsepower + hp.sq)
> summary(qm.hp)

Call:
lm(formula = mpg ~ horsepower + hp.sq)

Residuals:
     Min       1Q   Median       3Q      Max
-14.7135  -2.5943  -0.0859   2.2868  15.8961

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.9000997  1.8004268   31.60   <2e-16 ***
horsepower  -0.4661896  0.0311246  -14.98   <2e-16 ***
hp.sq        0.0012305  0.0001221   10.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.374 on 389 degrees of freedom
Multiple R-squared:  0.6876,    Adjusted R-squared:  0.686
F-statistic:   428 on 2 and 389 DF,  p-value: < 2.2e-16
> co<-coef(qm.hp)
> curve(co[1] + co[2]*x + co[3]*x^2, add = T, lwd =3)
```

Note: linear model implies linear relationship between parameters (β) but not linearity of variables. Variables can be squared, cubed or logged etc. But parameters should not multiply, divide, exponential forms.

$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$

$\hat{f}(x) = 56.90 - 0.4662x + 0.001231x^2$

3.3 Residual assumptions

- Model $y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i$ where ϵ_i are iid N(0,σ^2)

  $y = X\beta + e$ (Linear relationship)

- The assumptions on the residuals can be broken down into

  (assumptions of OLS)

  - $E[\epsilon_i] = 0$ Expectations of mean zero.
  - $Var(\epsilon_i) = \sigma^2$ (Homoscadesticity) Variance of residuals/error is same/constant & doesn't depend on x.
  - $Cov(\epsilon_i, \epsilon_j) = 0$ (Residuals are uncorrelated).
  - $\epsilon_i \sim N(0, \sigma^2)$ (Error are normally distributed)

- $e \overset{iid}{\sim} N(0, \sigma^2)$
  independent → $Cov(e_i, e_j) = 0$ & Normality.
  identically distributed → $E(e_i) = 0$ & $Var(\epsilon_i) = \sigma^2$

3.3.1 Mean 0 and Nonlinearity

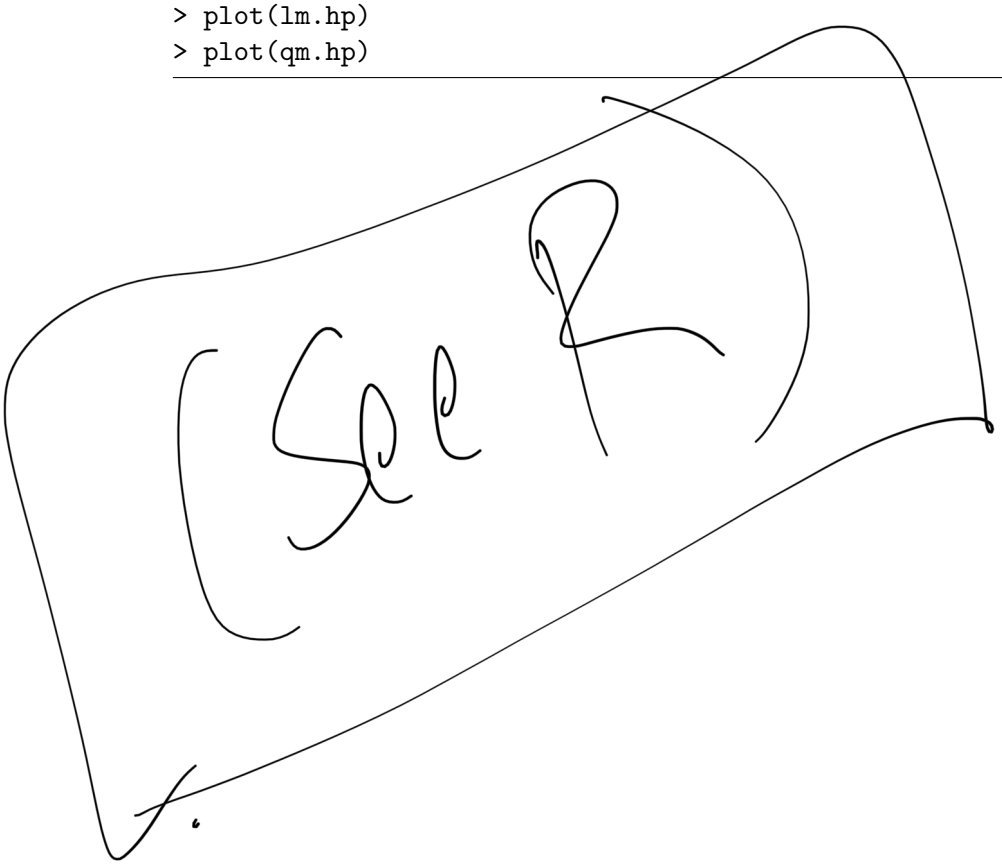
- If $E[\epsilon_i] \neq 0$ for some i then our predictions are biased. This is due to a non-linear relationship that isn't modeled.

- To assess the assumption we get residuals $\hat{e}_i = y_i - \hat{y}_i$ and plot $(\hat{y}_i, \hat{e}_i)$

- How did the linear vs. quadratic model affect the 0 mean assumption and bias.

Code

```
> plot(horsepower, mpg)
> lm.hp<-lm(mpg~horsepower)
> abline(lm.hp)
> par(mfrow = c(2,2))
> plot(lm.hp)
> plot(qm.hp)
```



3.3.2 Correlated errors

- If $Var(\epsilon_i) = \sigma^2$ and $Cov(\epsilon_i, \epsilon_j) = 0$ then $\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_i e_i^2$ is unbiased. Is this true if $Cov(\epsilon_i, \epsilon_j) > 0$?

If errors are correlated, estimate of residual st. err (or variance) is biased.
If $cov(e_i, e_j) > 0 \rightarrow \sigma^2$ will be too small
If $cov(e_i, e_j) < 0 \rightarrow \sigma^2$ will be too large
} → In above fmla.
Imp → $Cov(e_i, e_j) > 0 \Rightarrow$ Underestimating the standard error & will be too small than actual one.
$Cov(e_i, e_j) < 0 \Rightarrow$ Overestimating the SE & SE will be large than actual one.

- How would underestimated σ^2 (for positive correlation), and consequently all standard error estimates, affect the width of our confidence intervals and validity of our hypothesis tests?

Underestimated σ^2 cause (i) Confidence interval will not be wide enough.
(ii) Test statistics will be too far from zero (more +ve)
(iii) P. value will be too small.
& thus, inference won't be valid.

- The following simulates data from $y_i = 0 + \epsilon_i$ with ϵ_is being correlated. Determine if the residuals are positively or negatively correlated, and if the confidence interval is valid.

Code

```
> # correlated errors
> x<-1:50
> y<-cumsum(rnorm(50))
> confint(lm(y~x))
> plot(resid(lm(y~x)), type = "b")
> abline(h=0)
```

3.3.3 Nonconstant variance

- Consider the simulated data below. Is the variance constant? How will this affect prediction intervals?

Code

```
> # Nonconstant Variance
> set.seed(1)
> x<- 1:100/20
> y<-rlnorm(length(x), x)
> plot(x,y)
> abline(lm(y~x))
> plot(lm(y~x))
> z<-log(y)
> plot(x,z)
> lm(z~x)
> plot(lm(z~x))
```

- Typically if the variance is increasing as y_i increases a log or square root transformation works (assuming the data is positive).

- If you perform a log transformation you probably aren't interested in the prediction $\hat{z} = \log(\hat{y}) = \log(\hat{f}(x))$. How would you get $\hat{y}$?

Log:

$$P\left(\hat{a} \leq z \leq \hat{b}\right) = 0.95$$

Unlog:

$$P\left(e^{\hat{a}} \leq y \leq e^{\hat{b}}\right)$$

3.3.4 Outliers

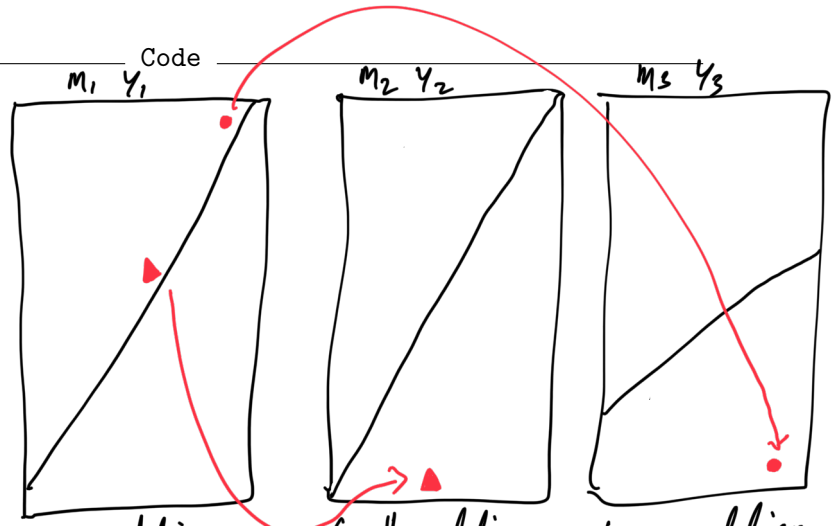
- Typically we check to see that the standardized residuals e_i/RSE are not more than 3 in magnitude.
- The leverage statistic measures how far x_i is from $\bar{x}$. If x_i is far from $\bar{x}$ and e_i is an outlier, the outlier affects the slope of the regression line.

see in residual plot.

- Compare the leverage and standardized residual for each type of outlier and their affect on the analysis. Note Cooks D considers leverage and the value of the standardized residual simultaneously. For details see a regression class.

Code

```
> # Outliers
> set.seed(1)
> x<- 1:10
> y<-x+rnorm(length(x))
> y1<-y
> y1[5]<-0
> y2<-y
> y2[9]<-0
>
> m1<-lm(y~x)
> m2<-lm(y1~x)
> m3<-lm(y2~x)
>
> par(mfrow = c(1,3))
> plot(x,y)
> abline(m1)
> plot(x,y1)
> abline(m2)
> plot(x,y2)
> abline(m3)
```



3.3.5 Normality

- What does the central limit theorem tell us about parameter estimators and normality?
  $\hat{\beta} \sim N$ for large sample.

  for small N,
  $\hat{\beta}/_{SE} \sim T$- distribution

- What's normality of a point estimator got to do with the 95% confidence interval formula below?
  Given $SE(\hat{\theta})$ is estimated consistently estimated and for large sample size, if $\hat{\theta}$ is normally distributed, this is a valid confidence interval. Not necessarily efficient.

  $$\hat{\theta} \pm 2\widehat{SE}(\hat{\theta}))$$

- In practice we plot ordered values of the residuals against their expectations under a normal model. Is normality satisfied below?

Code

```
>hist(rstandard(qm.hp))
>plot(qm.hp)
```

If I have a sample of 100 observations from a normal distribution and look at the smallest value in observation, that value should be close to the expectation of error under normal distribution. So, the smallest sample residual should be close to its smallest theoritical value.

If there is deviation from above condition, the normality assumption might not be satisfied.

3.4 Multicollinearity (Inflates variance / st. error)

If 2 or more predictors are nearly collinear, i.e. $\mathbf{x}_j = c\mathbf{x}_k$ then this is called multicollinearity.

- If a predictor is highly collinear with 1 or more other predictors, it inflates the standard errors.
- The variance inflation factor for $\hat{\beta}_j$ is $VIF(\hat{\beta}_j) = \frac{Var(\hat{\beta}_{j|full})}{Var(\hat{\beta}_{j|reduced})} = \frac{1}{1-R^2_{j|-j}}$
  - where $R^2_{j|-j}$ is the R^2 for the model that regresses X_j on all other predictors.
  - Remark, VIF greater than 5 or 10 (?) is often considered problematic.
- What will happen to $R^2_{j|-j}$ and the VIF if X_j is highly related to the other predictors

car package to get VIF.

$\beta = (X'X)^{-1}(X'y)$

when multicolinearity $(X'X)^{-1}$ calculation become complex & large.

- Example. Identify the correlated predictors, their VIF's, and their impact on the standard errors.

Code

```
> attach(Credit)
> pairs(~Balance + Age + Limit + Rating)
> library(car)
> vif(lm(Balance~ Age + Limit + Rating))
       Age      Limit     Rating
  1.011385 160.592880 160.668301
> vif(lm(Balance~ Age + Limit ))
     Age    Limit
1.010283 1.010283
> summary(lm(Balance ~ Age + Limit+Rating))$coef
                 Estimate  Std. Error    t value     Pr(>|t|)
(Intercept) -259.51751854 55.88219241 -4.6440110 4.656409e-06
Age           -2.34575164  0.66861406 -3.5083792 5.025853e-04
Limit          0.01901346  0.06296388  0.3019741 7.628303e-01
Rating         2.31045954  0.93952549  2.4591771 1.435238e-02
> summary(lm(Balance ~ Age+ Limit))$coef
               Estimate   Std. Error   t value      Pr(>|t|)
(Intercept) -173.410901 43.828387048 -3.956589  9.005366e-05
Age           -2.291486  0.672484540 -3.407492  7.226468e-04
Limit          0.173365  0.005025662 34.495944 1.627198e-121
```

3.5 Summary

- Model $y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i$ or $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
  - ϵ_i are iid N(0,σ^2).
- Y is continuous but X_1, X_2, can be categorical, continuous, interaction terms, nonlinear, ...
- LSE are $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

- We can test hypotheses about parameters and construct interval estimates using LSE and their standard errors or F tests.
  - Which and / or how many predictors are included depends on bias, variance, n and whether our goal is prediction or inference
- Interval estimation
  - Confidence intervals for any $f(x) = x^T\boldsymbol{\beta}$ are

$$x^T\hat{\boldsymbol{\beta}} \pm 2\hat{SE}(x^T\hat{\boldsymbol{\beta}}) = x^T\hat{\boldsymbol{\beta}} \pm 2\sqrt{\hat{\sigma}^2 x^T(\mathbf{X}^T\mathbf{X})^{-1}x}$$

  - Prediction intervals for any $y = f(x) + \epsilon = x^T\boldsymbol{\beta} + \epsilon$ are

$$x^T\hat{\boldsymbol{\beta}} \pm 2\sqrt{\hat{\sigma}^2(1 + x^T(\mathbf{X}^T\mathbf{X})^{-1}x)}$$

  - Choosing the right x above leads to desired confidence and prediction intervals.
- The performance of predictions and inference made with the linear model depends on n and the validity of residual assumptions, which can be verified.
- Multicollinearity can be problematic, especially when we have a lot of predictors.
- Main R pseudo code

Code

```
# Fitting
my.lm<-lm(y~x1 + x2 + ..., data = my.data)

# Summarizing and testing
summary(my.lm)

# Interval estimation
```

```
confint(my.lm)
predict(my.lm, newdata = data.frame(x1= , x2 = ,...), type = )

# Diagnostics
plot(my.lm)
library(car)
vif(my.lm)
```