STAT-615 Regression Final Exam

Yunting Chiu

2021-04-25

1. (22 points) (By hand) The following R code was written to study relation between variables y, x1, and x2. Unfortunately, coffee was spilled on the output, and some parts of it became unreadable. Restore the missing parts in the 11 empty boxes.

```
> attach(DATASET)
> summary(DATASET)
     x1                 x2                    y
 No    :10   Min.    :0.04564   Min.    :0.9589
 Yes   :10   1st Qu. :0.45671   1st Qu. :1.5509
             Median  :0.64683   Median  :1.8745
             Mean    :0.65205   Mean    :1.9655
             3rd Qu. :0.82315   3rd Qu. :2.3329
             Max.    :1.12577   Max.    :3.3721

> reg = lm( y ~ x1 + x2, data=DATASET )
> anova(reg)
Analysis of Variance Table

Response: y
           Df  Sum Sq  Mean Sq  F value  Pr(>F)
x1          1  2.0937  .......  .......  .......  ***
x2          1  4.9187  [      ]  [      ]  [      ]  ***
Residuals  17  0.0939  [      ]

> summary(reg)

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   3.40071     0.04405  .......  .......  ***
x1Yes        -0.32623     0.03494  [      ]  [      ]  ***
x2           -1.95094     0.06538  .......  .......  ***

Residual standard error: [      ] on 17 degrees of freedom
F-statistic: [      ] on [      ] and [      ] DF, p-value: 1.11e-16

> predict(reg, data.frame(x1="No", x2=5))
[      ]
```

Final Exam

Monday, April 19, 2021 09:38

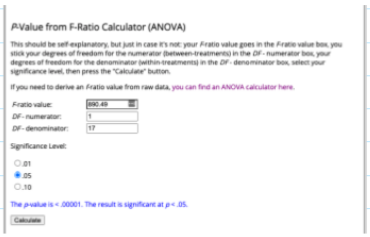
ANOVA Table (n=20)

① MSR(X_2) = SSR(X_2) / df(X_2) = 4.9187 / 1 = 4.9187

MSE = SSE / df(ε) = 0.0939 / 17 = 0.0055223529

F-value(X_2) = MSR(X_2) / MSE = 4.9187 / 0.0055223529 = 890.4995

P-value is < 0.00001 ≒ 0



Coefficients table:

t-value(X_1) = est.(X_1) / Sderr.(X_1) = -0.32623 / 0.03494 = -9.338867

P-value(X_1) = P(-9.338867, df: 17+1+1 = 19)

= P-value is < 0.0001 ≒ 0

Residual standard error = sqrt(SSE / df of residual) = sqrt(0.0939 / 17) = 0.074320445

F-statistic = 634.7766 on 2 (X_1, X_2) and 17 (residuals df) DF, p-value = 1.11e-16

$R^2 = \frac{2.09377 + 4.9187}{2.09377 + 4.9187 + 0.0939} = 0.9867864$ (SSR(X_1), SSR(X_2), SSE)

$F_{stat} = \frac{R^2/p=2}{(1-R^2)/n-(p+1)} = 634.7766$ ($n=20$, -3, $=17$)

predict(reg, data.frame(X_1 = "No", X_2 = 5)) -6.35399

if X_1 = Yes, b_1 = -0.32623

if X_1 = No, b_1 = 0

predict (X_1 = 0, X_2 = 5) ⇒ $y = 3.40071 + 5 \cdot (-1.95094)$

$= -6.35399$

References

- P-Value from F-Ratio Calculator: https://www.socscistatistics.com/pvalues/fdistribution.aspx
- P Value from T Score Calculator: https://www.socscistatistics.com/pvalues/tdistribution.aspx

2. (16 points) Multiple linear regression models are studied for eight different sets of data. Below are

various kinds of residual plots. For each case, give a brief diagnostics of the regression model and suggest a way to improve the model, if necessary.

Heteroscedasticity occurs when the variance for all observations in a data set are not the same. it is a violation of the ordinary least square assumption that $var(yi) = var(ei) = variance2$. In the presence of heteroskedasticity, there are two main consequences on the least squares estimators:

(a) The least squares estimator is still a linear and unbiased estimator, but it is no longer best. That is, there is another estimator with a smaller variance.

(b) The standard errors computed for the least squares estimators are incorrect. This can affect confidence intervals and hypothesis testing that use those standard errors, which could lead to misleading conclusions.

Models with linear pattern and constant variance (no need to improve): 1, 2, 6, 8

Plot 3: Problems: nonlinearity and heterogeneity. However, there is a non-random pattern just not linear (U-shaped). It suggests that this model can be tested on a quadratic model for a better fit.

Plot 4: Non constant variance and it seems like the data is divided. We could transform the data such as taking logs and sqrt and see if there's improvement.

Plot 5: Problems: Nonlinearity and heteroscedasticity. We could transform the variables to improve the model such as taking logs and sqrt.

Plot 7: The "fan out" shape suggests heteroscedasticity. We could transform the data like taking logs and sqrt to improve the model. We could also use a Weighted Least Square model to address this issue.

3. (32 points) (Use R for data analysis) Based on our known "mtcars" dataset, we'll now build a model for predicting variable `qsec`, the time in seconds that it takes a car to drive 1/4 mile from a full stop.

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
#mtcars$cyl <- as.factor(mtcars$cyl)
#mtcars$vs <- as.factor(mtcars$vs)
#mtcars$am <- as.factor(mtcars$am)
# levels(mtcars$am) <- c("automatic", "manual")
#mtcars$gear <- as.factor(mtcars$gear)
#mtcars$carb <- as.factor(mtcars$carb)
```

(a) Select the **optimal** subset of independent variables for the prediction of the `qsec` using your favorite variable selection method. Show all methods used and justify your conclusion.

Exhaustive Search

```
library(leaps)
best <-  regsubsets(qsec ~ ., data = mtcars)
summary(best)
```

```
## Subset selection object
## Call: regsubsets.formula(qsec ~ ., data = mtcars)
## 10 Variables  (and intercept)
##      Forced in Forced out
## mpg      FALSE      FALSE
## cyl      FALSE      FALSE
## disp     FALSE      FALSE
## hp       FALSE      FALSE
## drat     FALSE      FALSE
## wt       FALSE      FALSE
## vs       FALSE      FALSE
## am       FALSE      FALSE
## gear     FALSE      FALSE
## carb     FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          mpg cyl disp hp  drat wt  vs  am  gear carb
## 1  ( 1 ) " " " " " " " "  " " " "  " " "*" " " " "  " "
## 2  ( 1 ) " " "*" " " "  " " " " "  " " " " "*" " "  " "
## 3  ( 1 ) " " " " " " "  "*" " " "  "*" "*" " " " "  " "
## 4  ( 1 ) " " " " " "*"  " " " " "  "*" "*" " " " "  "*"
## 5  ( 1 ) " " " " " "*"  " " " " "  "*" "*" "*" " "  "*"
## 6  ( 1 ) " " "*" "*"  " " " " "  "*" "*" "*" " "  "*"
## 7  ( 1 ) "*" "*" "*"  " " " " "  "*" "*" "*" " "  "*"
## 8  ( 1 ) "*" "*" "*"  " " " " "  "*" "*" "*" "*"  "*"
```

Find out the largest adjusted R squares

- R^2 is not a fair measurement. As the number of parameters increases, so does the R^2.

```
summary(best)$adjr2
```

```
## [1] 0.5394775 0.7317144 0.7916421 0.8196948 0.8227529 0.8317230 0.8351322
## [8] 0.8300857
```

```
which.max(summary(best)$adjr2)
```

```
## [1] 7
```

Find out the smallest Mallows Cp

```
summary(best)$cp
```

```
## [1] 46.688365 16.060664  7.539059  4.317994  4.913419  4.742881  5.390848
## [8]  7.127031
```

```
which.min(summary(best)$cp)
```

```
## [1] 4
```

Find out the smallest BIC (penalized-likelihood criteria)

```
summary(best)$bic
```

```
## [1] -18.93039 -33.83942 -39.58604 -41.91147 -40.20082 -39.65202 -38.14754
## [8] -35.07891
```

```
which.min(summary(best)$bic)
```

```
## [1] 4
```

Sequential Search

- Lower AIC (Akaike information criterion) values indicate a better-fit model

```
reg.null = lm(qsec ~ 1, data = mtcars)
reg.full = lm(qsec ~ ., data = mtcars)

# using algorithm to considers either adding or removing variables at each step:
step(reg.null, scope=list(lower=reg.null, upper=reg.full ), direction = "forward")
```

```
## Start:  AIC=38.14
## qsec ~ 1
##
##        Df Sum of Sq    RSS    AIC
## + vs    1    54.872 44.116 14.275
## + hp    1    49.651 49.338 17.854
## + carb  1    42.631 56.358 22.111
## + cyl   1    34.603 64.385 26.373
## + disp  1    18.619 80.369 33.469
## + mpg   1    17.352 81.636 33.969
## <none>               98.988 38.136
## + am    1     5.230 93.758 38.399
## + gear  1     4.478 94.511 38.655
## + wt    1     3.022 95.966 39.144
## + drat  1     0.823 98.165 39.869
##
## Step:  AIC=14.27
## qsec ~ vs
##
##        Df Sum of Sq    RSS     AIC
## + gear  1   13.8534 30.262  4.2134
## + am    1   12.8533 31.262  5.2538
## + wt    1    8.1318 35.984  9.7548
## + carb  1    7.8975 36.218  9.9625
## + drat  1    6.8737 37.242 10.8545
## + hp    1    5.9853 38.130 11.6088
## <none>               44.116 14.2746
## + disp  1    1.8125 42.303 14.9321
## + mpg   1    1.0151 43.101 15.5297
## + cyl   1    0.0447 44.071 16.2422
##
## Step:  AIC=4.21
## qsec ~ vs + gear
##
```

```
##        Df Sum of Sq    RSS     AIC
## + hp    1    5.3811 24.881 -0.0520
## + cyl   1    5.3612 24.901 -0.0264
## <none>              30.262  4.2134
## + carb  1    1.3073 28.955  4.8002
## + disp  1    1.1573 29.105  4.9655
## + am    1    1.1282 29.134  4.9976
## + wt    1    0.7491 29.513  5.4113
## + mpg   1    0.7020 29.560  5.4623
## + drat  1    0.0034 30.259  6.2097
##
## Step:  AIC=-0.05
## qsec ~ vs + gear + hp
##
##        Df Sum of Sq    RSS     AIC
## + wt    1    7.1658 17.715 -8.9216
## + am    1    3.7916 21.090 -3.3427
## <none>              24.881 -0.0520
## + mpg   1    0.9131 23.968  0.7515
## + disp  1    0.8604 24.021  0.8219
## + drat  1    0.7379 24.143  0.9847
## + cyl   1    0.7075 24.174  1.0249
## + carb  1    0.2762 24.605  1.5908
##
## Step:  AIC=-8.92
## qsec ~ vs + gear + hp + wt
##
##        Df Sum of Sq    RSS      AIC
## + cyl   1   1.74422 15.971 -10.2384
## <none>              17.715  -8.9216
## + disp  1   0.82052 16.895  -8.4392
## + am    1   0.63335 17.082  -8.0866
## + mpg   1   0.54497 17.170  -7.9215
## + carb  1   0.32925 17.386  -7.5220
## + drat  1   0.03047 17.685  -6.9767
##
## Step:  AIC=-10.24
## qsec ~ vs + gear + hp + wt + cyl
##
##        Df Sum of Sq    RSS      AIC
## + am    1   1.35918 14.612 -11.0846
## <none>              15.971 -10.2384
## + disp  1   0.38497 15.586  -9.0192
## + drat  1   0.31580 15.655  -8.8775
## + mpg   1   0.28622 15.685  -8.8170
## + carb  1   0.22227 15.749  -8.6869
##
## Step:  AIC=-11.08
## qsec ~ vs + gear + hp + wt + cyl + am
##
##        Df Sum of Sq    RSS      AIC
## <none>              14.612 -11.0846
## + mpg   1   0.65294 13.959 -10.5474
## + disp  1   0.30408 14.308  -9.7575
```

```
## + carb  1   0.21926 14.393  -9.5684
## + drat  1   0.12833 14.484  -9.3668
```

```
##
## Call:
## lm(formula = qsec ~ vs + gear + hp + wt + cyl + am, data = mtcars)
##
## Coefficients:
## (Intercept)           vs         gear           hp           wt          cyl
##   21.285201     1.267714    -0.387951    -0.009389     0.766989    -0.533956
##          am
##   -0.854222
```

Conclusion

Our final mission is to select the **fewest** predictors in the linear model. In exhaustive search, the adjusted R squares suggests we use 7 variables, the Mallows Cp and BIC suggest 4 variables. In sequential search, we use algorithm to considers either adding or removing variables at each step to final the best model. The lowest AIC = -11.08, suggesting the best model use 6 variables. Therefore, we consider choosing 4 variables' model from Mallows Cp and BIC methods so the best predictors of `qesc` is `disp`, `wt`, `vs`, and `carb`.

According to the coefficient table, all 4 independent variables are completely significant, and the adjusted R-squared is 0.82, meaning that the linear model is a good to predict `qsec`.

```
# the best model
reg <- lm(qsec ~ disp + wt + vs + carb, data = mtcars)
summary(reg)
```

```
##
## Call:
## lm(formula = qsec ~ disp + wt + vs + carb, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0691 -0.3444 -0.0478  0.3004  2.7730
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.584969   0.698748  22.304  < 2e-16 ***
## disp        -0.012291   0.003039  -4.044 0.000393 ***
## wt           1.895087   0.330022   5.742 4.18e-06 ***
## vs           1.469469   0.460193   3.193 0.003560 **
## carb        -0.583230   0.108755  -5.363 1.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7588 on 27 degrees of freedom
## Multiple R-squared:  0.843,  Adjusted R-squared:  0.8197
## F-statistic: 36.23 on 4 and 27 DF,  p-value: 1.733e-10
```

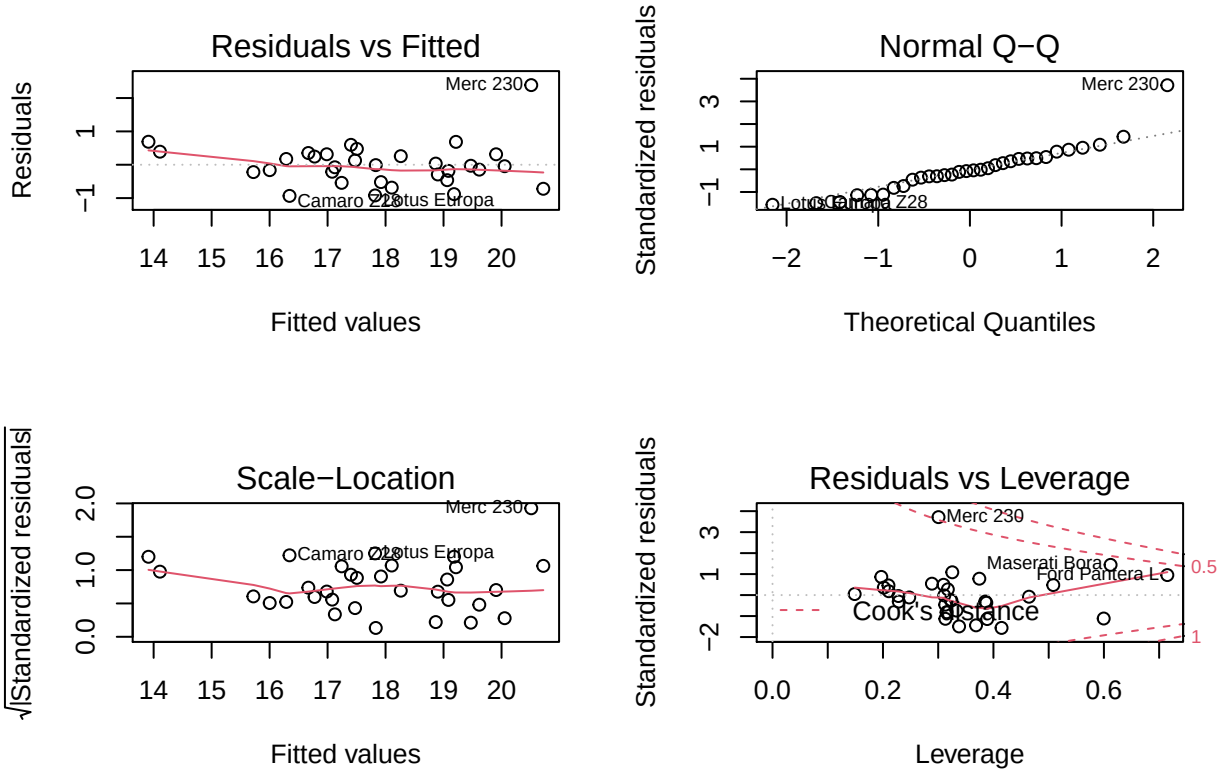
Regression Equation

$$\hat{qsec} = 15.584969 - 0.012291disp + 1.895087wt + 1.469469vs - 0.583230carb$$

(b) Is there any indication of nonlinear relations? Explain how you reach this conclusion.

- **Plot:**: Based on the left top `residuals vs fitted` plot, the red line looks horizontal, meaning that this model may not indicate nonlinear relationships.
- **Table**: Check the inversely related values of Tolerance and VIF. **Tolerance has to be > 0.10 and VIF < 10.** If these stipulated are not fulfilled, multicollinearity is at hand. A correction table reveals `cyl`, `disp`, `hp` and `wt` are highly correlated with each another, meaning that there might be a cause for the multicollinearity problem.

```
library(performance)
mtcarsReg <- lm(qsec ~., data = mtcars)
par(mfrow=c(2,2))
plot(mtcarsReg)
```




```
check_collinearity(mtcarsReg)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##  Term  VIF Increased SE Tolerance
##  drat 3.40         1.84      0.29
##    vs 4.37         2.09      0.23
##    am 4.48         2.12      0.22
##
## Moderate Correlation
##
##  Term  VIF Increased SE Tolerance
##   mpg 7.20         2.68      0.14
##  gear 5.35         2.31      0.19
##  carb 7.44         2.73      0.13
##
```

```
## High Correlation
##
##  Term   VIF Increased SE Tolerance
##   cyl 14.33         3.79      0.07
##  disp 20.03         4.48      0.05
##    hp 10.26         3.20      0.10
##    wt 12.59         3.55      0.08
```

(c) Test significance of interaction between the weight (wt) and transmission (am, 0 = automatic, 1 = manual). (Hint: Use the "best" model according to part (a)).

- Let us take a look on interaction term `wt:am`, the p-value is 0.745148 which is not significant. Compared to model 3a, adding a transmission variable (`am`) is not good to explain the response variable so we consider removing interaction term.

```
# Add interaction into the model
intercationReg <- lm(qsec ~ disp + wt * am + vs + carb, data = mtcars)
summary(intercationReg)$coefficients[7,]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##  0.1596508  0.4857582  0.3286631  0.7451476
```

(d) (**only for 615 students**) Construct a 90% prediction interval for `qsec` of a car that has 6 cylinders (`cyl`) and 150 horsepower (`hp`).

- When cylinders = 6 and horsepower = 150, a 90 % prediction interval for drive 1/4 mile from a full stop (`qsec`) is [15.53595, 20.04211].

```
cylAndHp <- lm(qsec ~ cyl + hp, data = mtcars)
# summary(cylAndHp)

predict(cylAndHp, dplyr::tibble(cyl = 6, hp = 150), interval = "prediction", level = 0.90)
```

```
##        fit      lwr      upr
## 1 17.78903 15.53595 20.04211
```

4. (10 points) (By hand) A student fitted a linear regression function for a class assignment. The student plotted the residuals ei against responses Yi and found positive relation. When the residuals were plotted against the fitted values $\hat{Y}i$, the student found no relation.

(a) How could the differences arise? Which is the more meaningful plot?

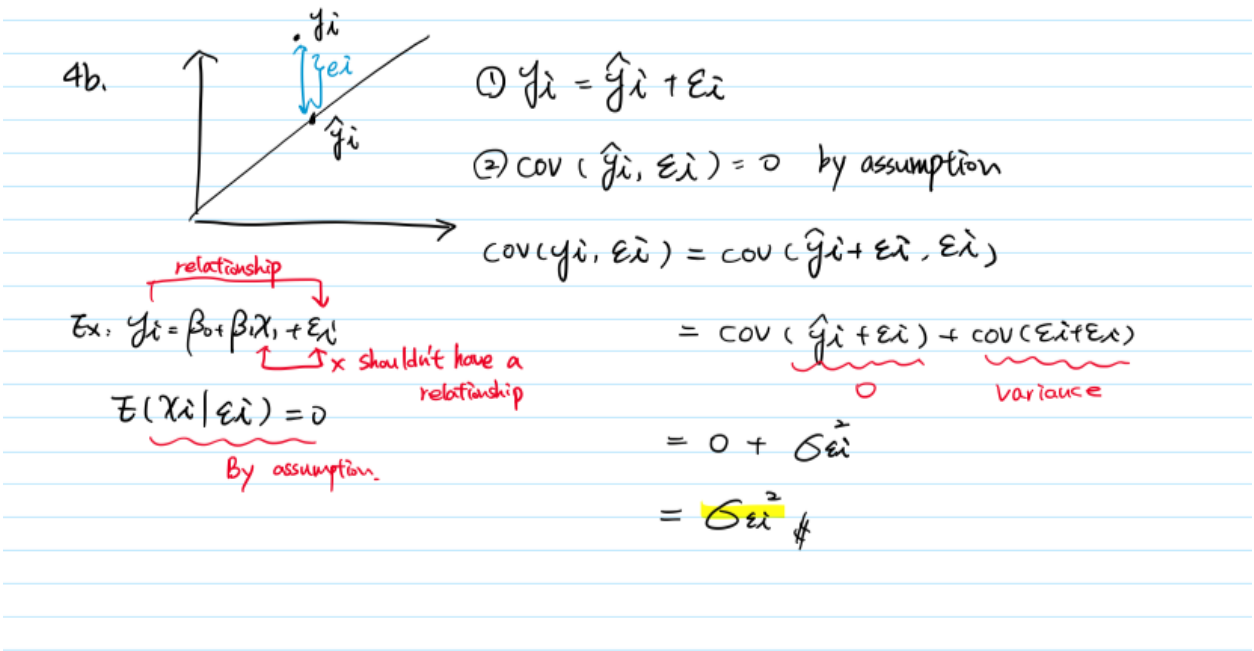
There is a relation between ei against Yi because ei can be found as $Yi - \hat{Y}i$. Plus, ei is dependent of Yi so we can see ei has a positive relation with Yi on the plot. In contract, there is no relation between the ei and $\hat{Y}i$ because as I mentioned that $ei = Yi - \hat{Y}i$ so ei is independent of $\hat{Y}i$. In conclusion, ei vs Yi is the more meaningful plot.

(b) Support your answer by deriving the sample covariance $Cov(ei, Yi)$ and $Cov(ei, \hat{Y}i)$. Feel free to use any Formula and any results that we derived in class.

Recall, the OLS assumptions are below:

- The linear regression model is "linear in parameters."
- There is a random sampling of observations.
- The conditional mean should be zero.
- There is no multi-collinearity (or perfect collinearity).
- Spherical errors: There is homoscedasticity and no autocorrelation
- Optional Assumption: Error terms should be normally distributed.

Then, we start to derive the covariance between ei, Yi and ei, $\hat{Y}i$:



References

- https://www.albert.io/blog/key-assumptions-of-ols-econometrics-review/
- https://stats.stackexchange.com/questions/155587/residual-plots-why-plot-versus-fitted-values-not-observed-y-values

5. (25 points) (Use R for data analysis) An experiment was conducted to evaluate the effect of vitamin C on tooth growth. Sixty guinea pigs received various doses of vitamin C by one of two delivery methods, orange juice or ascorbic acid. Results of this experiment are in dataset "ToothGrowth" which is already loaded in R. You can look at it with commands `attach(ToothGrowth)`, `names(ToothGrowth)`, `summary(ToothGrowth)`, `ToothGrowth`.
   The data set contains the following variables:

- `len`: Tooth length
- `supp`: supplement or delivery method (OJ = orange juice, VC = ascorbic acid)
- `dose`: Dose in milligrams/day

Read the data

Because `supp` is a dummy variable, so we can recode the observations: set `VC (ascorbic acid)` as **1** and set `OJ (orange juice)` as **0**.

```
ToothGrowth %>%
  mutate(supp = dplyr::recode(supp, "VC" = 1, "OJ" = 0)) -> ToothGrowthDummy
```

(a) Fit a linear regression model that can be used to predict the tooth length based on the dose and the delivery method of vitamin C.

```
toothReg <- lm(len ~ ., data = ToothGrowthDummy)
summary(toothReg)
```

```
##
## Call:
```

```
## lm(formula = len ~ ., data = ToothGrowthDummy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.600 -3.700  0.373  2.116  8.800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.2725     1.2824   7.231 1.31e-09 ***
## supp         -3.7000     1.0936  -3.383   0.0013 **
## dose          9.7636     0.8768  11.135 6.31e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.236 on 57 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.6934
## F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

Model Equation

The expected value of tooth length is:

$$\hat{len} = 9.2725 - 3.7000 supp + 9.7636 dose$$

(b) Is delivery method significant?

- The p-value of delivery method (`supp`) is 0.0013, meaning that there is evidence that the `supp` is significant, in addition to X2 (`dose`) that is already in the model. In other words, we can have evidence to reject the null (H0: $\beta_1 = 0$ can be rejected).

```
summary(toothReg)
```

```
##
## Call:
## lm(formula = len ~ ., data = ToothGrowthDummy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.600 -3.700  0.373  2.116  8.800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.2725     1.2824   7.231 1.31e-09 ***
## supp         -3.7000     1.0936  -3.383   0.0013 **
## dose          9.7636     0.8768  11.135 6.31e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.236 on 57 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.6934
## F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

(c) Is there a significant interaction between the dose and the delivery method, at $\alpha = 5\%$ level?

- The interaction term has a p-value of 0.024631, indicating that it is significant.

```r
toothRegInteraction <- lm(len ~ supp * dose, data = ToothGrowthDummy)
summary(toothRegInteraction)
```

```
##
## Call:
## lm(formula = len ~ supp * dose, data = ToothGrowthDummy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2264 -2.8462  0.0504  2.2893  7.9386
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.550      1.581   7.304 1.09e-09 ***
## supp          -8.255      2.236  -3.691 0.000507 ***
## dose           7.811      1.195   6.534 2.03e-08 ***
## supp:dose      3.904      1.691   2.309 0.024631 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.083 on 56 degrees of freedom
## Multiple R-squared:  0.7296, Adjusted R-squared:  0.7151
## F-statistic: 50.36 on 3 and 56 DF,  p-value: 6.521e-16
```

(d) Write two regression equations explicitly, one equation for each delivery method.

- If supp(X1) = 1, meaning that the delivery method is ascorbic acid (VC), the following regression equation is:

$$\hat{len} = 11.550 - 8.255 + 7.811dose + 3.904dose$$

Then:

$$\hat{len} = 3.295 + 11.715dose$$

- If supp(X1) = 0, meaning that the delivery method is orange juice (OJ), the following regression equation is:

$$\hat{len} = 11.550 + 7.811dose$$

(e) What percent of the total variation of the tooth length is explained by this regression?

- We have 72.9 % of the total variation of the tooth length is explained by the predictors.

```r
summary(toothRegInteraction)$r.square
```

```
## [1] 0.7295544
```

```r
# summary(toothRegInteraction)$adj.r
```