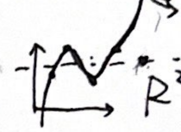
## Important Issues in Multiple Regression

**Oversaturated Models:** Suppose we collect data on a bunch of different variables that could be used as predictor variables. We should **not** just blindly add predictors to the model. Why?

- Bigger models are harder to interpret
  simpler models are better
- Sample size should be at least 5 to 20 times larger than # predictors.

wrong prediction



$R^2 = 100\%$

oversaturated model – When $n = p+1$ (sample size = # parameters)
Eventhough the model predicts perfectly for data set, it's NOT useful for any population!

**Adjusted R²:** One of the shortcomings of $R^2$ is that it **only** increases or stays the same if a new predictor $x_{p+1}$ is added to the model. This is true even if the new predictors are bad. How do we know if the new predictor is actually useful? We look at **adjusted** $R^2$.

$R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1-R^2)$
(You're NOT required to remember this formula)

$R^2_{adj}$ will only go up if new predictor variable is significant while $R^2$ goes up (or stays the same) any time we add a new predictor NO MATTER how bad they are.

**ANOVA vs t tests:** We should always perform ANOVA **first** to see if there are *any* good predictors in our model. If there are not (i.e. p-value > 0.05), then we do not proceed. However, a small ANOVA p-value by itself is not as useful in this case, because it could be the case that only one or a few of the $\beta_i$'s are not 0. We should look at the p-values of the **individual** predictors determined by the t-test for $\beta_i$.

ANOVA – small p-value → at least one predictor is good
t tests – small p-value → this $X_i$ is a good predictor after all other X's taken into account

△ Order of testing (t tests) is important – Test higher-order terms first
e.g. $X^2$ before $X$, $X_1X_2$ before $X_1$ or $X_2$.

**Multicollinearity**: We want the predictors (x's) to be correlated with the response (y). But if several of the predictors are highly correlated with each other, they are not adding anything new to predict y. Each x may be a good predictor by itself, but they should not be used together in the model.

Ex: predict $y = ht$
$x_1$ = length of left arm
$x_2$ = length of right arm
} correlated w/ each other

Together in model we expect:
- ANOVA – small p-value (at least one pred good)
- t test – both large p-vals (neither gives significant information after the other one taken into account in model)

⇒ They should NOT be together in model.

## Categorical Variables in Regression - Dummy Variables

We can use **categorical** variables as predictors in multiple regression. We call the categorical variable a **dummy variable** coded as 0 or 1.

Quant Quant Categorical

**Example**: For example, suppose we wish to predict weight ($y$) using height ($x_1$) and gender ($x_2$). We may code gender with dummy variable:

Dummy variable for gender $X_2 = \begin{cases} 0, & \text{Female} \\ 1, & \text{Male} \end{cases}$ ← Baseline group gets 0 for Dummy Var

**Full Model (No interaction):**

$$\underbrace{y}_{wt} = \alpha + \beta_1 \underbrace{X_1}_{ht} + \beta_2 \underbrace{X_2}_{gender} + \varepsilon$$

F: $X_2 = 0$ $\quad y = \alpha + \beta_1 X_1 + \beta_2(0) + \varepsilon$

$$= \boxed{\alpha + \beta_1 X_1 + \varepsilon}$$ Baseline model F

M: $X_2 = 1$ $\quad y = \alpha + \beta_1 X_1 + \beta_2(1) + \varepsilon$

$$= \alpha + \beta_1 X_1 + \beta_2 + \varepsilon$$

$$= \boxed{(\alpha + \beta_2) + \beta_1 X_1 + \varepsilon}$$ M

**Interpretation of Coefficients in the Full Model:**

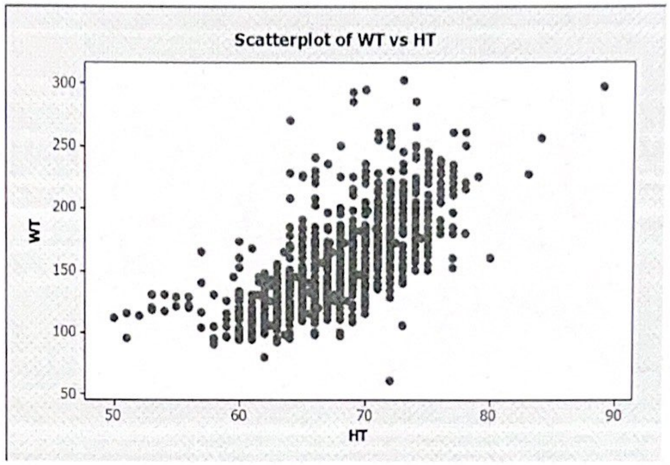
$\alpha$: constant — Intercept for F (Baseline Group)

$\beta_1$: coeff of ht — Slope for both groups

$\beta_2$: coeff of gender — Change in intercept from F to M.

**EXAMPLE – What is the relationship between height and weight for UF students?**
Data on UF students' heights and weights collected by STA3024 students. ~~N=1309~~
Questions about ***some*** data – are these heights correct?

| | HT | WT |
|---|---|---|
| F | 50.0 | 111 |
| F | 51.0 | 115 |
| F | 51.0 | 95 |
| F | 52.0 | 113 |
| F | 53.0 | 118 |
| F | 53.0 | 120 |
| F | 53.0 | 120 |
| F | 53.0 | 130 |
| F | 54.0 | 117 |
| F | 54.0 | 130 |
| F | 55.0 | 121 |
| F | 55.0 | 128 |
| F | 56.0 | 120 |
| F | 56.0 | 122 |
| F | 56.0 | 128 |
| F | 57.0 | 103 |
| F | 57.0 | 116 |
| F | 57.0 | 140 |
| M | 57.0 | 165 |




F 58.0 104 F 58.0 130 F 58.0 90 F 58.0 92 F 58.0 95 F
59.0 104 F 59.0 110 F 59.0 115 F 59.0 125 F 59.0 96 F 59.0
97 F 59.5 145

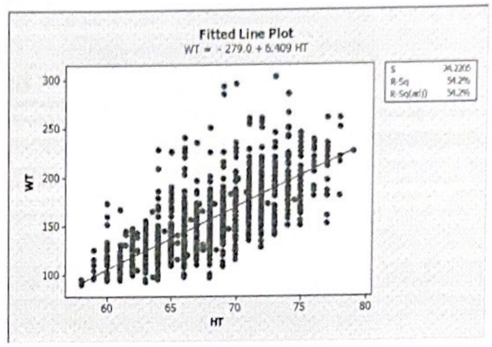
M 80 160 M 83 227 M 83 227 M 84 255 M 89 296

M 72 60 M 73 105
F 64 270

SLR

**Regression Analysis: WT versus HT**

The regression equation is
WT = - 279 + 6.41 HT




| Predictor | Coef | SE Coef | T | P |
|---|---|---|---|---|
| Constant | -279.01 | 11.19 | -24.92 | 0.000 |
| HT | 6.4088 | 0.1649 | 38.86 | 0.000 |

S = 24.2205   R-Sq = 54.2%   R-Sq(adj) = 54.2%

Analysis of Variance

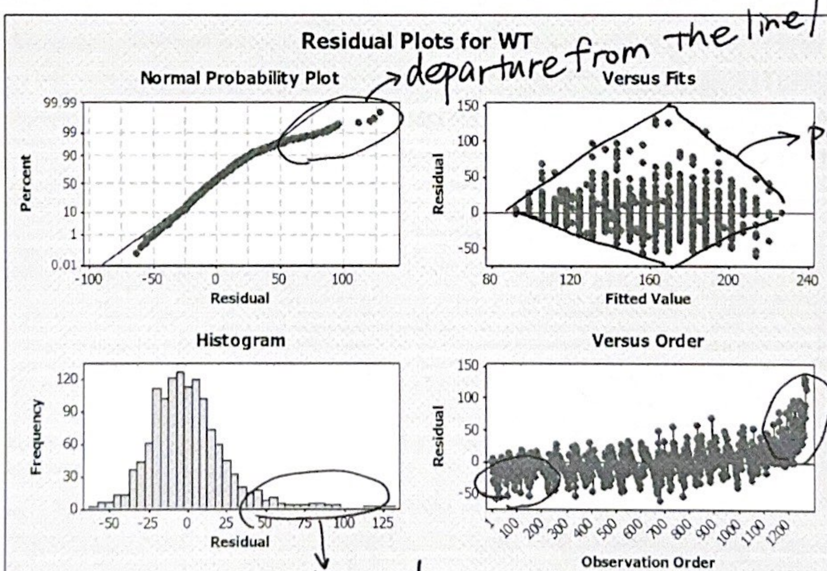
| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 1 | 885986 | 885986 | 1510.29 | 0.000 |
| Residual Error | 1276 | 748543 | 587 | | |
| Total | 1277 | 1634529 | | | |

$n=1278$

$P(X \geq 1510.29)$
where $X \sim F_{1,1276}$

Predicted Values for New Observations

| New Obs | HT | Fit | SE Fit | 95% CI | 95% PI |
|---|---|---|---|---|---|
| 1 | 65 | 137.562 | 0.816 | (135.961, 139.163) | (90.019, 185.106) |
| 2 | 60 | 105.518 | 1.448 | (102.678, 108.359) | (57.917, 153.120) |
| 3 | 76 | 208.059 | 1.519 | (205.080, 211.038) | (160.449, 255.669) |




departure from the line/normality

pattern spotted → hetergeneity of variance

skewed to the right

MLR

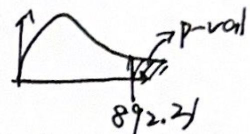
**Regression Analysis: WT versus HT, GENDER_M_1**

| Predictor | Coef | SE Coef | T | P |
|---|---|---|---|---|
| Constant | -164.68 | 14.76 | ~~11.16~~ | ~~0.000~~ |
| HT | 4.5699 | 0.2271 | 20.12 | 0.000 |
| GENDER_M_1 | 20.963 | 1.866 | 11.23 | 0.000 |

S = 23.1134   R-Sq = 58.3%   R-Sq(adj) = 58.3%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 2 | 953389 | 476695 | 892.31 | 0.000 |
| Residual Error | 1275 | 681140 | 534 | | |
| Total | 1277 | 1634529 | | | |




$= P(X \geq 892.31)$

where $X \sim F_{2,1275}$

- **Model:** $WT = \alpha + \beta_1 HT + \beta_2 \underset{(M-1)}{Gender} + \varepsilon$

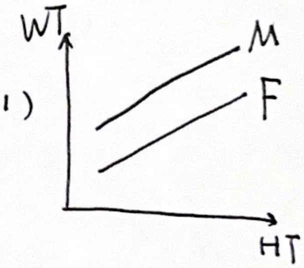
- **Assumptions:** $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

- **Fitted Equations for M and F:** $WT = -164.68 + 4.5699\,HT + 20.963\,Gender$

Baseline F $(X_2 = 0)$ $\boxed{WT = -164.68 + 4.5699\,HT} + \cancel{20.963(0)}$

Male M $(X_2 = 1)$ $WT = -164.68 + 4.5699\,HT + 20.963(1)$

$\boxed{WT = -143.717 + 4.5699\,HT}$




- **ANOVA test**