MLR

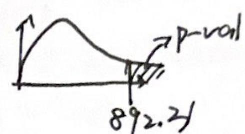
Regression Analysis: WT versus HT, GENDER_M_1

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

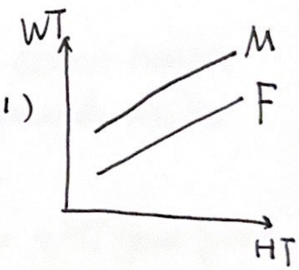
- **Assumptions:** $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

- **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**

$H_0: \beta_1 = \beta_2 = 0$ (neither ht nor gender good pred of wt)

H_a: at least one $\beta_i \neq 0$ ($i = 1, 2$) (at least one of the predictors good)

TS: $F = 892.31$ (from output)

p-val: $P(X \geq 892.31)$ where $X \sim F_{2,1275}$

$= 0.000 \rightarrow$ Rej H_0 at all usual α's

conclusion: Very strong evidence to say that at least one predictor is good.

- t tests

△For Height:

$H_0: \beta_1 = 0 \quad H_a: \beta_1 \neq 0$

$TS: t = \frac{b_1}{s.e._1} = \frac{4.5699}{0.2271} = 20.12$

p-val: $2P(X \geq |20.12|) = 0.000$ where $X \sim t_{1275}$

conclusion: very strong evidence to say that HT good pred of WT after Gender is accounted for in model

△For Gender:

$H_0: \beta_2 = 0 \quad H_a: \beta_2 \neq 0$

$TS: t = \frac{20.963}{1.866} = 11.23$

p-val: $2P(X \geq |11.23|) = 0.000$ where $X \sim t_{1275}$

Conclusion: very strong evidence Gender good pred of WT after HT accounted for in model

- Interpret the coefficients

Parameters	Estimates	~~In the full model~~	
α constant	$a = -164.68$	intercept for F (baseline)	→ Do not interpret b/c "HT=0" tall is impossible (Extrapolation)
β_1 coeff HT	$b_1 = 4.5699$	slope for both genders	→ For each extra inch in ht we predict on avg extra 4.57 lbs for both genders
β_2 coeff Gender	$b_2 = 20.963$	change in intercept from F to M	→ For Males we predict wt 20.963 pounds higher than F at the same height.

- Compute the 95% CI for the true coefficient of gender, β_2

$b_2 \pm t^* s.e._2 \rightarrow t_{1275, 0.025} \approx z_{0.025} = 1.96$

$20.963 \pm (1.96)(1.866)$

$= (17.3, 24.6)$

We are 95% confident that at the same height, male wt is btwn 17.3 and 24.6 pounds more that female wt on average.

- Is this model better than the SLR?

Many criteria!
- Are all predictors good? → Yes, both gender + HT good pred.
- Look at R^2_{adj} → Model with gender + HT has bigger R^2_{adj}

YES

- What if we coded gender the other way?

$X_2 = \begin{cases} 0 ; M \\ 1 ; F \end{cases}$ (Now, M becomes the baseline group)

```
The regression equation is
WT = - 144 + 4.57 HT - 21.0 GENDER_F_1
```

For M: $WT = -144 + 4.57\,HT$

For F: $WT = -144 + 4.57\,HT - 21.0 = -165 + 4.57\,H$

We end up with the same equations for M & F.

Dummy Variables with Interaction

Sometimes there may be interaction between a quantitative term and a categorical term. For example, it is certainly reasonable to expect that there is interaction between height (x_1) and gender (x_2). In this case, we add an **interaction term** to our model.

Full Interaction Model:

$$y = \alpha + \underbrace{\beta_1 X_1}_{\text{ht, quan}} + \underbrace{\beta_2 X_2}_{\text{Gen, cate (dummy)}} + \underbrace{\beta_3 X_1 X_2}_{\text{interaction term}} + \overset{\text{random error}}{\varepsilon} \qquad X_2 = \begin{cases} 0, & F \\ 1, & M \end{cases}$$

Two Equations:

F: $(X_2 = 0)$

$$y = \alpha + \beta_1 X_1 + \beta_2 (0) + \beta_3 X_1 (0) + \varepsilon$$

$$\boxed{y = \alpha + \beta_1 X_1 + \varepsilon} \rightarrow \text{Baseline model}$$

M: $(X_2 = 1)$

$$y = \alpha + \beta_1 X_1 + \beta_2 (1) + \beta_3 X_1 (1) + \varepsilon$$

$$y = \alpha + \beta_1 X_1 + \beta_2 + \beta_3 X_1 + \varepsilon$$

$$\boxed{y = (\alpha + \beta_2) + (\beta_1 + \beta_3) X_1 + \varepsilon}$$

Interaction term btwn a dummy variable and a quantitative variable $X_1 X_2$ allows for different slopes.

Coefficients:

α	constant	intercept for F (Baseline group)
β_1	coeff of ht	Slope for F (Baseline group)
β_2	coeff of gender	change in intercept from F to M
β_3	coeff of interaction	change in slope from F to M

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

The regression equation is
WT = - 128 + 4.00 HT - 56.2 GENDER_M_1 + 1.14 HT*GENDER_M_1

Predictor	Coef	SE Coef	T	P
Constant	-128.05	21.21	-6.04	0.000
HT	4.0039	0.3266	12.26	0.000
GENDER_M_1	-56.16	30.83	-1.82	0.069
HT*GENDER_M_1	1.1382	0.4544	2.50	0.012

S = 23.0840 R-Sq = 58.6% R-Sq(adj) = 58.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	960396	320132	600.77	0.000
Residual Error	1274	678879	533		
Total	1277	1639274			

- **Model:** $y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$ $\quad X_2 = \begin{cases} 0, & F \\ 1, & M \end{cases}$
- **Assumptions:** $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$
- **Tests:**

① ANOVA first. $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ (none of the predictor variables is good)

H_a: at least one of β_i's $\neq 0$ $(i = 1, 2, 3)$

TS: $F = 600.77$ p-val: $P(X \geq 600.77)$ where $X \sim F_{3,1274}$

$= 0.000$

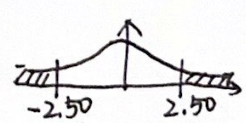
Very strong evidence to say at least one pred good.

② t-tests

α	β_1	β_2	β_3
constant	coeff of ht	coeff of gen	coeff of interaction
X	X	X	higher order term
NEVER			Always test for it first!!

t-test for interaction:

$H_0: \beta_3 = 0$ $\quad H_a: \beta_3 \neq 0$



TS: $t = 2.50$ p-val: $0.012 = 2P(X \geq 2.50)$ where $X \sim t_{1274}$

< 0.05
> 0.01

strong evidence to say that the interaction effect exists
(strong evidence of interaction)

Remark: Test for interaction first:
- SIG, don't even look at p-vals for X_1 or X_2 we include these predictors no matter what
- NOT SIG: Redo the model without interaction