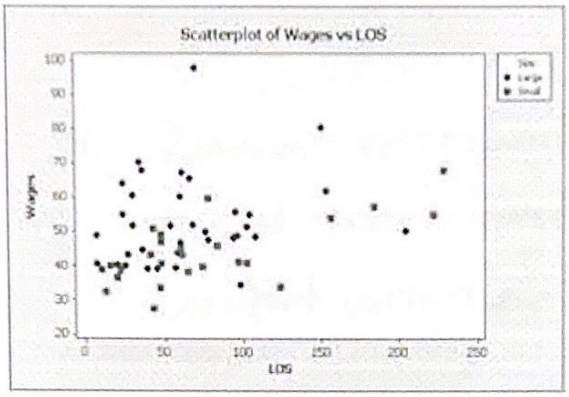
EXAMPLE: Wages vs Length of Service and Size of Company

y X_1 X_2



Coding of size of company: small = 0 large = 1 $\Rightarrow X_2 = \begin{cases} 1, \text{large} \\ 0, \text{small} \end{cases}$

Regression Analysis: Wages versus LOS, size, LOS*size $y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$

The regression equation is
Wages = 35.9 + 0.104 LOS + 13.6 size - 0.0483 LOS*size

Predictor	Coef	SE Coef	T	P
Constant	35.914	3.562	10.08	0.000
LOS	0.10424	0.03632	2.87	0.006
size	13.631	4.910	2.78	0.007
LOS*size	-0.04828	0.05634	-0.86	0.395

S = 10.9612 R-Sq = 26.6% R-Sq(adj) = 22.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	2438.1	812.7	6.76	0.001
Residual Error	56	6728.3	120.1		
Total	59	9166.4			

- Response variable: Wages
- Predictor variables: LOS (Quant) Size (Categ) LOS * Size (Interaction term)
- Identify n= 60 p= 3

- Model $y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$
- Assumptions $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

OR
- Random - errors/wages/employees → Don't know - need ~~story~~ study.
- Normal - errors/wages
- Constant variance - errors/wages

} Can use residual plots (Diagnostics)

- Fitted Equation

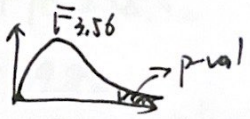
$$\hat{y} = 35.914 + 0.10424 X_1 + 13.631 X_2 - 0.04828 X_1 X_2$$

- ANOVA test

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ H_a: at least one $\beta_i \neq 0$

TS: $F = 6.76$

p-val: $P(X > 6.76) = 0.001$ where $X \sim F_{3,56}$



- t tests

Higher-order terms first.

→ Only the interaction term here

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

TS: $t = -0.86$

p-val: $2P(X > |-0.86|) = 0.395$ where $X \sim t_{56}$ → Fail to rej H_0. NOT SIG

- Interpret?

Do not interpret this model

→ Remove the interaction term and fit a simpler model.

Regression Analysis: Wages versus LOS, size

The regression equation is Wages = 37.5 + 0.0842 LOS + 10.2 size

Predictor	Coef	SE Coef	T	P
Constant	37.466	3.061	12.24	0.000
LOS	0.08417	0.02770	3.04	0.004
size	10.228	2.882	3.55	0.001

S = 10.9357 R-Sq = 25.6% R-Sq(adj) = 23.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	2349.9	1174.9	9.82	0.000
Residual Error	57	6816.6	119.6		
Total	59	9166.4			

- **Model** $y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
- **Assumptions** $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$
- **Fitted Equation** $\hat{y} = 37.466 + 0.08417 X_1 + 10.228 X_2$
- **ANOVA test** $H_0: \beta_1 = \beta_2 = 0$ $TS: F = 9.82$
  H_a: at least one $\beta_i \neq 0$ pval: 0.000 → Rej H_0
  At least one good predictor
- **t tests**

1. $H_0: \beta_1 = 0$ TS: $t = 3.04$
   $H_a: \beta_1 \neq 0$
   p-val: 0.004 (SIG)
   Interpret?

2. $H_0: \beta_2 = 0$ TS: $t = 3.55$
   $H_a: \beta_2 \neq 0$ pval: 0.001 (SIG)

→ Both are good predictors of wages after the other variable is accounted for.

a constant 37.5 y-int for small companies (α)

b_1 LOS 0.08417 slope of both sizes (β_1)

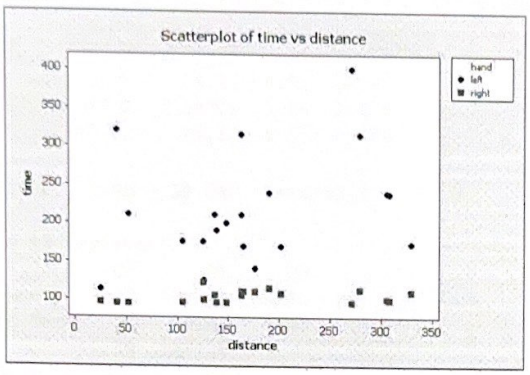
b_2 Size 10.228 change in y-int from small (β_2) to large companies.

① Average starting salary for small companies is $37,466.
(when $X_1 = 0$, zero months of service)

② For each month, we expect the salary to increase $84.17 for both large and small companies.

③ Large companies pay a starting salary $10,228 higher than small companies, on average.

EXAMPLE: Reaction Time in a Computer Game vs Distance to move mouse and Hand used.



y = reaction time

X_1 = distance

X_2 = hand = $\begin{cases} 1, \text{ left} \\ 0, \text{ right} \end{cases}$

All ppl are right-handed in the dataset

Coding of hand: right = 0 left = 1

Regression Analysis: time versus distance, hand, dist*hand

The regression equation is
time = 99.4 + 0.028 distance + 72.2 hand + 0.234 dist*hand

$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$

Predictor	Coef	SE Coef	T	P
Constant	99.36	25.25	3.93	0.000
distance	0.0283	0.1308	0.22	0.830
hand	72.18	35.71	2.02	0.051
dist*hand	0.2336	0.1850	1.26	0.215

S = 50.6067 R-Sq = 59.8% R-Sq(adj) = 56.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	136948	45649	17.82	0.000
Residual Error	36	92198	2561		
Total	39	229146			

Unusual Observations

Obs	distance	time	Fit	SE Fit	Residual	St Resid
25	163	315.00	214.29	11.38	100.71	2.04R
30	271	401.00	242.65	17.19	158.35	3.33R
31	40	320.00	182.09	20.68	137.91	2.99R

R denotes an observation with a large standardized residual.

$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Regression Analysis: time versus distance, hand

The regression equation is time = 79.2 + 0.145 distance + 112 hand

Predictor	Coef	SE Coef	T	P
Constant	79.21	19.72	4.02	0.000
distance	0.14512	0.09324	1.56	0.128
hand	112.50	16.13	6.97	0.000

S = 51.0116 R-Sq = 58.0% R-Sq(adj) = 55.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	132865	66433	25.53	0.000
Residual Error	37	96281	2602		
Total	39	229146			

Unusual Observations

Obs	distance	time	Fit	SE Fit	Residual	St Resid
25	163	315.00	215.39	11.44	99.61	2.00R
30	271	401.00	231.10	14.67	169.90	3.48R
31	40	320.00	197.55	16.80	122.45	2.54R

R denotes an observation with a large standardized residual.

$y = \alpha + \beta_1' X_2 + \varepsilon$

$y = \alpha$. if $X_2 = 0$

$y = \alpha + \beta_1'$. if $X_2 = 1$

Regression Analysis: time versus hand

The regression equation is time = 104 + 112 hand

Predictor	Coef	SE Coef	T	P
Constant	104.25	11.62	8.97	0.000
hand	112.50	16.43	6.85	0.000

S = 51.9573 R-Sq = 55.2% R-Sq(adj) = 54.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	126562	126562	46.88	0.000
Residual Error	38	102583	2700		
Total	39	229146			

Unusual Observations

Obs	hand	time	Fit	SE Fit	Residual	St Resid
30	1.00	401.00	216.75	11.62	184.25	3.64R
31	1.00	320.00	216.75	11.62	103.25	2.04R
32	1.00	113.00	216.75	11.62	-103.75	-2.05R

R denotes an observation with a large standardized residual.

One-way ANOVA: time versus hand

Source	DF	SS	MS	F	P
hand	1	126563	126563	46.88	0.000
Error	38	102584	2700		
Total	39	229146			

S = 51.96 R-Sq = 55.23% R-Sq(adj) = 54.05%

Individual 95% CIs For Mean Based on Pooled StDev

Level	N	Mean	StDev	+---------+---------+---------+---------
0	20	104.25	8.25	(-----*-----)
1	20	216.75	73.01	(-----*-----)
				+---------+---------+---------+---------
				80 120 160 200

Pooled StDev = 51.96

Two-Sample T-Test and CI: time, hand Two-sample T for time

hand	N	Mean	StDev	SE Mean
0	20	104.25	8.25	1.8
1	20	216.8	73.0	16

Difference = mu (0) - mu (1)
Estimate for difference: -112.500
95% CI for difference: (-146.889, -78.111)
T-Test of difference = 0 (vs not =): T-Value = -6.85 P-Value = 0.000 DF= 19

	Parameter	estimator	standard error	sampling distribution
independent Two-group means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	t w/ df between: $\min(n_1-1, n_2-1)$ and n_1+n_2-2.

~~PAIRING~~

Statistical Inference Procedures Covered in STA 2023 and STA 3024

Predictor variables		Categorical response	Quantitative response: Normal-based procedures	Quantitative response: Non-parametric procedures
Categorical predictors	1 group	CI, ST for p	CI, ST for μ	Signed Test for population median
	2 groups	CI, ST for $p_1 - p_2$	CI, ST for $\mu_1 - \mu_2$ (Indep) or μ_d (Dep/matched)	Wilcoxon Rank Sum Test (Indep) or Wilcoxon Signed Rank Test (Matched)
	several groups	χ^2 test for contingency tables	ANOVA	Kruskal Wallis
Quantitative predictors	1 quantitative predictor	Logistic regression	SLR	(Non-parametric Regression)
	several predictors	(Multiple) logistic regression	MLR	
Assumptions		• SRS • enough successes and failures	• SRS • $n \geq 30$ OR original distribution is normal	• SRS (• for K-W, 5 obs min)