Again, since $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{(n-2)}$ we can easily show that

$$T = \frac{\frac{Y^* - \hat{Y}^*}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}}}{\sqrt{\frac{(n-2)S^2}{\sigma^2}/(n-2)}} = \boxed{\frac{Y^* - \hat{Y}^*}{S\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}} \sim t_{(n-2)}}$$

Using the above result as a pivotal qty for Y^* ...

$$P\left[-t_{\alpha/2} < \frac{Y^* - \hat{Y}^*}{S\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}} < t_{\alpha/2}\right] = 1-\alpha$$

Isolating Y^* ...

$$P\left[\hat{Y}^* - t_{\alpha/2}\cdot S\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}} < Y^* < \hat{Y}^* + t_{\alpha/2}\cdot S\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}\right] = 1-\alpha$$

$\Rightarrow$ $100(1-\alpha)\%$ Prediction Interval for Y^*:

$$\boxed{\hat{Y}^* \pm t_{\alpha/2}\cdot S\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}}$$

Computer Repair Data:

95% PI for Y^* where $x^* = 9$

$$143.66 \pm 2.179\sqrt{30.363}\sqrt{1+\frac{1}{14}+\frac{(9-6)^2}{114}}$$

$143.66 \pm 12.88 \rightarrow$ (130.78, 156.54) (i.e. This job shouldn't take longer than 157 minutes)

↓ much wider than CI

Y^* for $x^* = 6$ ($\bar{x}$)

$$97.2 \pm 2.179 \sqrt{30.363} \sqrt{1 + \frac{1}{14} + \frac{(6-6)^2}{110}}$$

$97.2 \pm 12.43 \rightarrow$ (84.77, 109.63)

↓ narrower than above, but much wider than CI.

SAS (PI bands, CI bands)

11.8 Correlation

In discussions regarding correlation, we generally assume that X is now a RV too. The usual assumption is that we are sampling from a bivariate normal dist i.e. $(X,Y) \sim BVN(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$

Recall from 5.10 ...

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}$$

Using conditional expectation we established ..

$$E(Y|x) = \mu_y + \rho \cdot \frac{\sigma_y}{\sigma_x}(x-\mu_x)$$

This is the conditional mean our LS line is trying to estimate

$$E(Y|x) = \beta_0 + \beta_1 x \quad , \quad \boxed{\beta_1 = \rho \cdot \frac{\sigma_y}{\sigma_x}} \quad , \quad \boxed{\beta_0 = \mu_y - \beta_1 \cdot \mu_x}$$

note: $H_0: \beta_1 = 0$ is equivalent to testing $H_0: \rho = 0$

The sample correlation coefficient is defined as:

$$\boxed{r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\cdot\sqrt{\sum(y_i-\bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}}$$

Recall the $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \Rightarrow$ $\boxed{\hat{\beta}_1 = r \cdot \sqrt{\frac{S_{yy}}{S_{xx}}}}$

Since we want to test:

$H_0: \rho = 0$ vs. $H_1: \rho \neq 0$

It is equivalent to testing $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

We have already shown that

$$T = \frac{\hat{\beta}_1}{S/\sqrt{S_{xx}}} \sim t_{(n-2)}$$

$$= \frac{r \cdot \sqrt{\frac{S_{yy}}{S_{xx}}}}{S \cdot \sqrt{S_{xx}}} = \frac{r \cdot \sqrt{S_{yy}}}{\sqrt{\frac{SSE}{n-2}}} = \frac{r \cdot \sqrt{S_{yy}}}{\sqrt{\frac{S_{yy} - \hat{\beta}_1 \cdot S_{xy}}{n-2}}}$$

$$= \frac{r \cdot \sqrt{S_{yy}}}{\sqrt{\frac{S_{yy} - r \cdot \sqrt{\frac{S_{yy}}{S_{xx}}} \cdot S_{xy}}{n-2}}} = \frac{r \cdot \sqrt{S_{yy}}}{\sqrt{\frac{S_{yy}\left(1 - \frac{r S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}\right)}{n-2}}}$$

$$= \frac{r \cdot \sqrt{S_{yy}}}{\sqrt{S_{yy}} \cdot \sqrt{\frac{1-r^2}{n-2}}} = \boxed{\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}}$$

So to test $H_0: \rho = 0$ vs. $H_a: \rho \neq 0$ we use

$$\boxed{T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2)}}$$

Computer Repair Data: $n = 14$ $S_{xx} = 114$ $S_{xy} = 1{,}768$ $S_{yy} = 27{,}768.36$

$\bar{x} = 6 \quad \bar{y} = 97.2 \quad \hat{\beta}_0 = 4.16 \quad \hat{\beta}_1 = 15.5 \quad s^2 = 30.363$

$$r = \frac{1,768}{\sqrt{114}\sqrt{27,768.36}} = .9937$$

note: $\hat{\beta}_1 = (.9937)\sqrt{\frac{27,768.36}{114}} = 15.5$

Test $H_0: \rho = 0$ vs. $H_1: \rho \neq 0$ (silly test in the case and in general)

$$t = \frac{.9937\sqrt{14-2}}{\sqrt{1-.9937^2}} = 30.71 \qquad \text{p-value} \approx 0$$

i.e. There is significant correlation (nonzero) between X and Y

note: $T = \frac{\hat{\beta}_1}{s/\sqrt{s_{xx}}} = \frac{15.5}{\sqrt{30.363/114}} = 30$ (same test under BVN assumption)

In practice, we want to test to see if $|\rho|$ is large in magnitude. Since the prob. dist. of r is difficult to obtain, we use an approx. test developed by Fisher

For moderately large samples ...

$\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ is approx. normal with

$$\mu = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \qquad \sigma = \frac{1}{\sqrt{n-3}}$$

$$\Rightarrow z = \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)}{\frac{1}{\sqrt{n-3}}} \sim N(0,1)$$

$H_0: \rho = \rho_0$ $\quad H_a: \rho > \rho_0$ $\quad$ Reject H_0 if $z > z_\alpha$

$H_a: \rho < \rho_0$ $\quad$ Reject H_0 if $z < -z_\alpha$

$H_a: \rho \neq \rho_0$ $\quad$ Reject H_0 if $|z| > z_{\alpha/2}$

Computer Repair Data:

$H_0: \rho = .9$ $\quad H_a: \rho > .9$

$\frac{1}{2}\ln\left(\frac{1+.9937}{1-.9937}\right) = 2.88$ $\quad \frac{1}{2}\ln\left(\frac{1+.9}{1-.9}\right) = 1.47$

$$z = \frac{2.88 - 1.47}{\frac{1}{\sqrt{14-3}}} = 4.67$$

p-value ≈ 0

Conclusion: We are highly confident $\rho > .9$ i.e. very strong relationship between no. of parts and repair time.

Recommendation: Use this test and test for $\rho > .8$ or $\rho < -.7$ or some reasonably large magnitude of ρ. The T-test for $\rho \neq 0$ is virtually worthless!

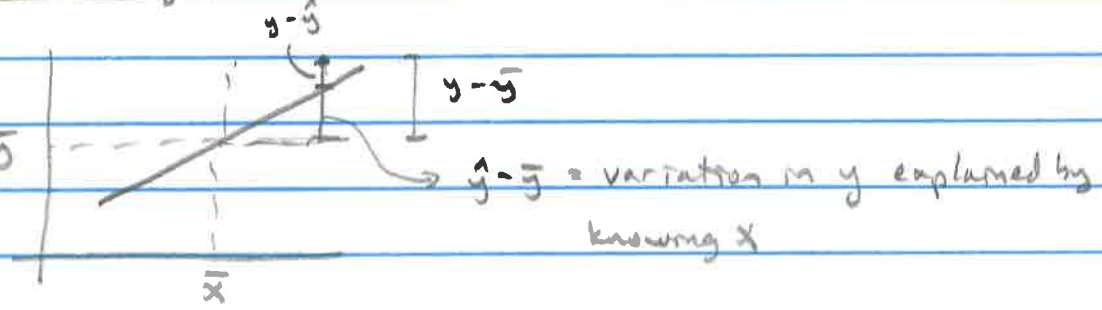
Coefficient of Determination, r^2

$S_{yy} = \sum (y - \bar{y})^2$ = total variation in Y

How much of this is explained by regression on X?

After regression on X, we have new estimates for the avg. value of Y.

$SSE = \sum (Y - \hat{Y})^2$ represents variation in Y that is not explained by X.



Now,

$$r^2 = \left(\frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \right)^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{S_{xy}}{S_{xx}} \cdot \frac{S_{xy}}{S_{yy}}$$

$$= \hat{\beta}_1 \cdot \frac{S_{xy}}{S_{yy}} = \frac{S_{yy} - SSE}{S_{yy}} = \boxed{1 - \frac{SSE}{S_{yy}}}$$

$\Rightarrow r^2$ = proportion of total variation in y, explained by regression on X

note: If $r^2 = 1$, then $SSE = 0$ and all pts fall perfectly on a line

Computer Repair Data: $r^2 = (.9937)^2 = .9874$

$$= 1 - \frac{364.36}{27,768.36} = .987$$

98.7% of all variation in repair times is explained by regression on no. of parts.