Simple Linear Regression and Correlation

- Model for designed experiment:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- $\epsilon_1, \ldots, \epsilon_n$ independent, mean 0, variance σ^2.
- Model for sample of pairs: $(X_i, Y_i), i = 1, \ldots, n$ sample from bivariate population.
- $\mathrm{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$
- So if we define $\epsilon_i = Y_i - \beta_1 X_i - \beta_0$ then
  - The ϵ_i are independent mean 0 constant variance.
  - $\mathrm{E}(\epsilon_i|X_i) = 0$.

Bivariate Normal Populations

- X, Y have a bivariate normal distribution if they have joint density

$$f(x, y) = \frac{1}{\sigma_1 \sigma_2 \sqrt{1-\rho^2}} \exp\left[-q(x, y)/\{2(1-\rho^2)\}\right]$$

where

$$q(x, y) = \frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)}{\sigma_1}\frac{(y-\mu_2)}{\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}$$

- Marginal density of X is $N(\mu_1, \sigma_1^2)$.
- Marginal density of Y is $N(\mu_2, \sigma_1^2)$.

- This is a density if $-1 < \rho < 1$ and σ_1, σ_2 are both positive.
- Covariance of X and Y is

$$\mathrm{E}\left\{(X-\mu_1)(Y-\mu_2)\right\} = \rho\sigma_1\sigma_2$$

- The correlation coefficient is ρ; that is

$$\mathrm{E}\left\{\frac{(X-\mu_1)}{\sigma_1}\frac{(Y-\mu_2)}{\sigma_2}\right\} = \rho$$

- Conditional distribution of Y given $X = x$ is Normal, mean

$$\beta_0 + \beta_1 x = \mu_2 + \rho\sigma_2\frac{x-\mu_1}{\sigma_1}$$

and variance

$$\sigma^2 = (1-\rho)^2\sigma_2^2.$$

Estimation of parameters

- The population means are estimated by sample means:

$$\hat{\mu}_1 = \bar{X} \qquad \hat{\mu}_2 = \bar{Y}$$

- Population SDs are estimated by sample SDs:

$$\hat{\sigma}_1 \equiv s_x = \sqrt{\frac{\sum_i (X_i - \bar{X})^2}{n-1}} \qquad \hat{\sigma}_2 \equiv s_y = \sqrt{\frac{\sum_i (Y_i - \bar{Y})^2}{n-1}}$$

- Population correlation estimated by sample correlation:

$$\hat{\rho} \equiv r = \frac{\frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}}{s_x s_y}$$

Estimation with fixed covariates

- Ordinary least squares estimate of slope β_1 is

$$\hat\beta_1 = r\frac{s_y}{s_x} = \frac{\sum_i (X_i - \bar X)(Y_i - \bar Y)}{\sum_i (Y_i - \bar Y)^2}$$

- Ordinary least squares estimate of intercept β_0 is

$$\hat\beta_0 = \bar Y - \hat\beta_1 \bar X.$$

- Ordinary least squares estimate of σ^2 is residual mean square:

$$\hat\sigma^2 = \sum_i (Y_1 - \hat\beta_0 - \hat\beta_1 X_i)^2/(n-2).$$

- This estimate is unbiased:

$$\mathrm{E}(\hat\sigma^2) = \sigma^2.$$

Relation between the models

- In both models $\mathrm{Var}(\epsilon_i) = \sigma^2$.
- In bivariate normal model

$$\mathrm{Var}(\epsilon_i) = \sigma^2 = \sigma_y^2(1 - \rho^2).$$

Simple linear regression: least squares, inference

- See *Fitting Linear Models* lecture for derivation of least squares formulas.
- The estimates $\hat\beta_1$ and $\hat\beta_2$ are linear combinations of the Y_i. For instance
$$\hat\beta_1 = \sum w_i Y_i$$
where
$$w_i = \frac{x_i - \bar{x}}{\sum_i (x_i - \bar{x})^2}.$$
- So
$$\begin{aligned}\mathrm{E}(\hat\beta_1) &= \sum_i w_i \mathrm{E}(Y_i) = \sum_i w_i(\beta_0 + \beta_1 x_i) \\ &= 0 + \beta_1 \sum_i w_i x_i \\ &= \beta_1\end{aligned}$$

- Notice use of fact that $\sum w_i = 0$ so $\sum w_i \bar{X} = 0$.
- The identity says $\hat{\beta}_1$ is an **unbiased** estimate of β_1.
- We can compute the variance:

$$\begin{aligned}
\mathrm{Var}(\sum_i w_i Y_i) &= \sum_i w_i^2 \mathrm{Var}(Y_i) \\
&= \sigma^2 \frac{\sum (x_i - \bar{x})^2}{\{\sum (x_i - \bar{x})^2\}^2} \\
&= \frac{\sigma^2}{\sum (x_i - \bar{x})^2}
\end{aligned}$$

- The square root of the variance of any estimate is called its **Standard Error**.

Distribution Theory

- Both $\hat{\beta}_1$ and $\hat{\beta}_2$ are linear combinations of the normally distributed Y_i.
- So both have normal distributions.
- So you can form confidence intervals:

$$\hat{\beta}_i \pm t_{n,\alpha/2}\text{Estimated Standard Error}$$

- and test hypotheses using

$$t = \frac{\hat{\beta}_i - \beta_{i,o}}{\text{Estimated Standard Error}}$$

- ESE is theoretical SE with σ estimated.
- Use residual mean square to estimate σ^2.

Output from JMP

```
R Square                   0.534338
Root Mean Square Error 1.96287
Mean of Response        32.44423
Estimates
Term          Estimate     Std Error  t Ratio  Prob>|t|
Intercept    11.098156    1.953928     5.68      <.0001
Distance      0.0481812   0.004389    10.98      <.0001
```

Can form CIs and test hypotheses like $H_o : \beta_1 = 0$.

Output from JMP

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	464.21357	464.214	120.4855
Error	105	404.55022	3.853	Prob > F
C. Total	106	868.76379		<.0001

Notice $F = t^2$, that is $120.4855 = 10.98^2$.
Always happens with 1 df F-test.