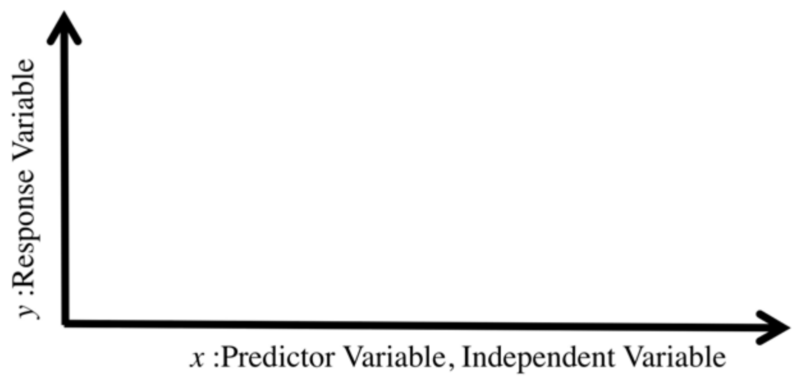
Chapter 10 Notes, Regression and Correlation

Regression analysis allows us to estimate the relationship of a response variable to a set of predictor variables



Let

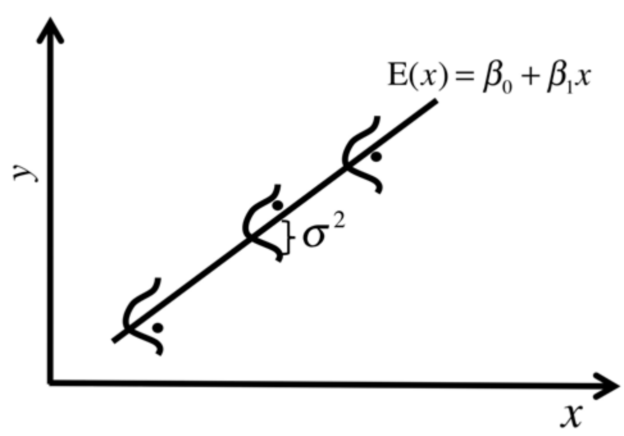
$$\begin{array}{ll} x_1, x_2, \cdots x_n & \text{be settings of } x \text{ chosen by the investigator and} \\ y_1, y_2, \cdots y_n & \text{be the corresponding values of the response.} \end{array}$$

Assume y_i is an observation of rv Y_i (which depends on x_i, where x_i is not random).

We model each Y_i by

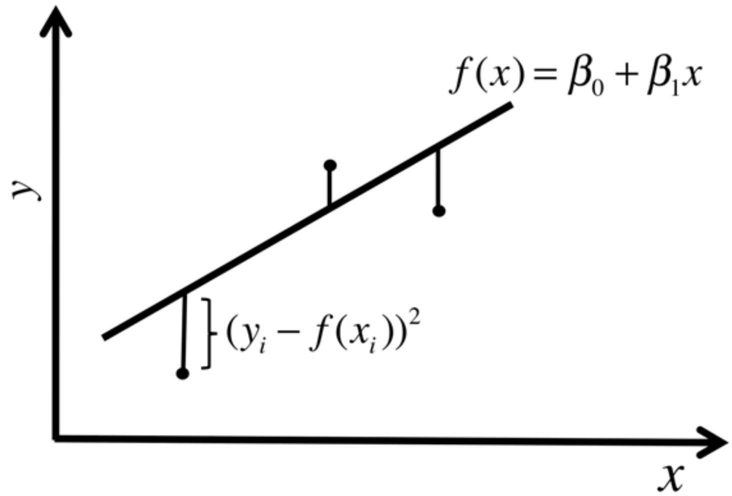
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i is iid noise with $E(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$. We usually assume that ϵ_i is distributed as $N(0, \sigma^2)$, so Y_i is distributed as $N(\beta_0 + \beta_1 x_i, \sigma^2)$.



Note: it is not true for all experiments that Y is related to X this way of course! Always scatterplot to check for a straight line.

For a good fit, choose β_0, β_1 to minimize the sum of squared errors.



Minimize

$$Q = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \quad \leftarrow \quad \text{“least squares”}$$

To minimize Q, set derivatives to 0 and solve for $\beta's$. Call the solutions $\hat{\beta}_0$, and $\hat{\beta}_1$.

$$0 \;=\; \frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right) \tag{1}$$

$$0 \;=\; \frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right). \tag{2}$$

Rewrite equation (1):

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i = 0$$

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \;\; \text{(pull } \beta\text{'s out of the sums)}$$

$$\frac{1}{n} \sum_{i=1}^{n} y_i - \hat{\beta}_0 - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} x_i = 0 \;\; \text{(divide by } n\text{)}$$

$$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

What does this mean about the least square line?

Solve equation (2) for $\hat{\beta}_1$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \hat{\beta}_0 - \sum_{i=1}^{n} x_i^2 \hat{\beta}_1 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \quad \text{(using previous page)}$$

$$\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \quad \text{(using definition of } \bar{x})$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} (\sum_{i=1}^{n} x_i)^2} \quad \text{(using definition of } \bar{y})$$

Consider the expressions (which we'll substitute in later):

$$\tilde{s}_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i y_i \quad \text{(skipping some steps)}$$

$$\tilde{s}_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \sum_{i=1}^{n} x_i^2 \quad \text{(just sub in } x \text{ for } y \text{ in previous eqn)}$$

where $\tilde{s}_{xy}$ is the sample covariance from Chapter 4 times $n - 1$. Look what happened:

$$\hat{\beta}_1 = \frac{\tilde{s}_{xy}}{\tilde{s}_{xx}}.$$

Put it together with the previous result and we get these two little (but important equations):

$$\hat{\beta}_1 = \frac{\tilde{s}_{xy}}{\tilde{s}_{xx}}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Now there is an easy way to find the LS line.

**********Procedure for finding LS line************

Given:

$$x_1, \cdots, x_n$$
$$y_1, \cdots, y_n$$

we compute $\bar{x}, \bar{y}, \tilde{s}_{xy}, \tilde{s}_{xy}$. Then compute

$$\hat{\beta}_1 = \frac{\tilde{s}_{xy}}{\tilde{s}_{xx}}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

And the answer is:

$$y = \hat{\beta}_1 x + \hat{\beta}_0.$$

Then if you want to make predictions you can use this formula - just plug in the x you want to make a prediction for.

Let's examine the goodness of fit. We will define SSE, SST, and SSR. Consider:

$$\text{SSE} = \text{sum of squares error} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where $\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$, these are your model's predictions. Recall $\hat{\beta}_0$ and $\hat{\beta}_1$ were chosen to minimize the sum of squares error (SSE).

The total sum of squares (SST) measures the variation of y's around their mean:

$$\text{SST} = \text{sum of squares total} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \tilde{s}_{yy}.$$

It turns out:

$$\begin{aligned} \text{SST} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \text{SSE} + \text{SSR} \end{aligned}$$

where SSR is called the "regression sum of squares." This is the model's variation around the sample mean.

Consider

$$r^2 = \frac{SSR}{SST} = \frac{\text{model's variation}}{\text{total variation}} = \text{“coefficient of determination.”}$$

It turns out that r^2 is the square of the sample correlation coefficient $r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$. Let's show that. First simplify SSR:

$$\begin{aligned} SSR &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x})\right)^2 \quad \text{note that the } \hat{\beta}_0\text{'s cancel out} \\ &= \hat{\beta}_1^{\,2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \hat{\beta}_1^{\,2}\tilde{s}_{xx}. \end{aligned} \tag{3}$$

And plugging this in,

$$r^2 = \frac{SSR}{SST} = \frac{\hat{\beta}_1^{\,2}\tilde{s}_{xx}}{\tilde{s}_{yy}} = \frac{\tilde{s}_{xy}^2\tilde{s}_{xx}}{\tilde{s}_{xx}^2\tilde{s}_{yy}} = \frac{\tilde{s}_{xy}^2}{\tilde{s}_{xx}\tilde{s}_{yy}} = \frac{s_{xy}^2}{s_{xx}s_{yy}},$$

where we just cancelled a normalizing factor in that last step. So after we take the square root, that shows r^2 really is the square of the sample correlation coefficient.

Back to $SST = SSR + SSE$ and $r^2 = \frac{SSR}{SST}$. If $r^2 = 0.953$, most of the total variation is accounted for by the regression, so the least square fit is a good fit. That is, r^2 tells you how much better a regression line is compared to fitting with a flat line at the sample mean $\bar{y}$.

Note: Compute r using this formula: $\frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$, so you do not get the sign wrong from taking the square root, $r = \pm\sqrt{\frac{SSR}{SST}}$.

To summarize,

- We derived an expression for the LS line

$$y = \hat{\beta}_1 x + \hat{\beta}_0, \text{ where } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

- We showed that $r^2 = \frac{SSR}{SST}$. Its value indicates how much of the total variation is explained by the regression.

One more definition before we do inference. The variance σ^2 measures dispersion of the y_i's around their means $\mu_i = \beta_0 + \beta_1 x_i$. An unbiased estimator of σ^2 turns out to be

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$$

We lose two degrees of freedom from estimating β_0 and β_1, that is why we divide by $n-2$.

Chapter 10.3 Statistical Inference

We want to make inferences on the values of β_0 and β_1. Assume again that we have:

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

where ϵ_i is iid noise and is distributed as $N(0, \sigma^2)$. Then it turns out that $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed with

$$E(\hat{\beta}_0) = \beta_0, \quad SD(\hat{\beta}_0) = \sigma \sqrt{\frac{{}^{n}_{i=1} x_i^2}{n \tilde{S}_{xx}}}$$

$$E(\hat{\beta}_1) = \beta_1, \quad SD(\hat{\beta}_0) = \frac{\sigma}{\sqrt{\tilde{S}_{xx}}}$$

It also turns out that S^2, which is the random variable for $s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$ obeys:

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}.$$

We can do hypothesis tests on β_0 and β_1 using $\hat{\beta}_0$ and $\hat{\beta}_1$ as estimators for the means of β_0 and β_1. We can use

$$SE(\hat{\beta}_0) = s\sqrt{\frac{\sum_{i=1}^n x_i^2}{n\tilde{s}_{xx}}}, \quad SE(\hat{\beta}_1) = \frac{s}{\sqrt{\tilde{s}_{xx}}} \tag{4}$$

as estimators for the SD's. So we can ask for $100(1-\alpha)\%$ CI for β_0 and β_1:

$$\beta_0 \in [\hat{\beta}_0 - t_{n-2,\alpha/2}SE(\hat{\beta}_0), \hat{\beta}_0 + t_{n-2,\alpha/2}SE(\hat{\beta}_0)]$$
$$\beta_1 \in [\hat{\beta}_1 - t_{n-2,\alpha/2}SE(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2,\alpha/2}SE(\hat{\beta}_1)]$$

Hypothesis tests (usually we do not test hypotheses on β_0, just β_1)

$$H_0 : \beta_1 = \beta_1^0$$
$$H_1 : \beta_1 \neq \beta_1^0.$$

Reject H_0 at level-α if

$$|t| = \frac{\hat{\beta}_1 - \beta_1^0}{SE(\hat{\beta}_1)} > t_{n-2,\alpha/2}.$$

***Important: If you choose choose $\beta_1^0 = 0$, you are testing whether there is a linear relationship between x and y. If you reject $\beta_1^0 = 0$, it means y depends on x.

Note that when $\beta_1^0 = 0$, $t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$.

Analysis of Variance (ANOVA)

We're going to do this same test another way. ANOVA is useful for decomposing variability in the y_i's, so you know where the variability is coming from. Recall:

$$SST = SSR + SSE$$

- SST is the total variability ($df = n - 1$ from constraint $\sum_{i=1}^n(\hat{y}_i - \bar{y})^2$),
- SSR is the variability accounted for by regression and
- SSE is the error variability ($df = n - 2$). This leaves one df for SSR.

A sum of squares divided by df is called a "mean square".

- $MSR = \frac{SSR}{1}$ "mean square regression"
- $MSE = \frac{SSE}{n-2} = s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$ "mean square error"

Consider the ratio

$$\begin{aligned} F &= \frac{MSR}{MSE} = \frac{SSR}{s^2} \\ &= \frac{\hat{\beta}_1^2 \tilde{s}_{xx}}{s^2} \quad \text{from (3)} \\ &= \left(\frac{\hat{\beta}_1}{s/\sqrt{\tilde{s}_{xx}}} \right)^2 \\ &= \left(\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \right)^2 = t^2 \quad \text{from (4).} \end{aligned}$$

Hey look, the square of a T_v r.v is an $F_{1,v}$ r.v. Actually that's always true: Consider:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{S^2/\sigma^2}}$$

$$T^2 = \frac{Z^2/1}{S^2/\sigma^2} = F_{1,v}$$

since $Z^2 \sim \chi_1^2$ and $\frac{S^2}{\sigma^2} \sim \frac{\chi_\nu^2}{\nu}$. Therefore we have $t_{n-2,\alpha/2}^2 = f_{1,n-2,\alpha}$.

How come $\alpha/2$ turned into α?

Back to testing:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 = 0 \end{aligned}$$

We'll reject H_0 when $F = \frac{MSR}{MSE} > f_{1,n-2,\alpha}$.

Note: This is just the square of the previous test. We also do it this way because it is a good introduction to multiple regression in Chapter 11.

ANOVA (Analysis of Variance)

ANOVA table - A nice display of the calculations we did.

Source of variation	SS	d.f.	MS	F	p
Regression	SSR	1	$\text{MSR} = \frac{\text{SSR}}{1}$	$F = \frac{\text{MSR}}{\text{MSE}}$	p-value for test
Error	SSE	$n-2$	$\text{MSE} = \frac{\text{SSE}}{n-2}$		
Total	SST	$n-1$			

The pvalue is for the F-test for $H_0 : \beta_1 = 0,\ H_1 : \beta_1 = 0$.

15.075J / ESD.07J Statistical Thinking and Data Analysis
Fall 2011