Inference for Simple Linear Regression

Recall: The simple linear regression model relates two quantitative variables X and Y through the linear model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where the ϵ are independent, and normally distributed with mean zero and constant variance σ^2.

Previously, we have investigated ways to "fit" a least squares regression line. This meant to compute estimates for the model parameters

$$b_0 = \hat{\beta}_0, \quad b_1 = \hat{\beta}_1, \quad s^2 = \hat{\sigma}^2$$

so that the sum of squared residuals is minimized (LEAST SQUARES REGRESSION).

The observations on X and Y taken during an experiment are *random variables*. If the experiment were repeated (with different subjects), then the observations would likely be slightly different. Consequently, parameter estimates would be slightly different, too. In this sense, the parameter estimates obtained from the data are random variables. What can we say about their distributions?

The Slope Parameter β_1

The slope parameter β_1 is the true average change in the response if the predictor is increased by one unit. The parameter estimate

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

is a point estimate of this random variable. Thus $\hat{\beta}_1$ is a linear function of the random variables $Y_1, \ldots, Y_n$ (assuming that the values $x_1, \ldots, x_n$ are fixed by the experimenter).

PROPOSITION: What can we say about the distribution of $\hat{\beta}_1$?

Distribution: Since $\hat{\beta}_1$ is a linear combination of the observations y_i, which are normally distributed, the distribution of $\hat{\beta}_1$ is also normal.

Mean: The mean of the point estimate $\hat{\beta}_1$ is the true regression parameter

$$E(\hat{\beta}_1) = \beta_1.$$

Variance: The variance (and standard deviation) of $\hat{\beta}_1$ are

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, \qquad \sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}} \quad \left(= \frac{s}{\sqrt{S_{xx}}} \right)$$

Of course, in reality, we do not know the model parameter σ, so we have to replace it by its estimate $s = \sqrt{\text{MSE}}$.

Test statistic: $\hat{\beta}_1$ has a normal distribution with mean β_1 and (estimated!) standard deviation $s/\sqrt{S_{xx}}$. Thus, the test statistic involving the population parameter β_1 is

$$T = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{xx}}} \sim t(df = n-2)$$

CONFIDENCE INTERVAL: This allows us to formulate a $100(1-\alpha)\%$ confidence interval for the slope parameter:

$$CI = \left[\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot s_{\hat{\beta}_1}\right] = \left[\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot \frac{s}{\sqrt{S_{xx}}}\right]$$

HYPOTHESIS TESTING: We can also formulate hypotheses regarding the slope β_1 and test them using the above distribution assumptions.

Most common null hypothesis: $H_0 : \beta_1 = 0$ (zero slope)

Test statistic: $t = \frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}} \sim t(df = n-2)$.

Regression and ANOVA

There is a large amount of overlap between the basic ideas behind ANOVA and regression. In each case, we are looking at sums of squares and comparing them to get test statistics. The ANOVA sums of squares describe amount of variation due to different factors (and factor combinations) and error. The regression sums of squares describe amount of variation due to regression (SSR) and error (SSE).

ANOVA TABLE FOR SIMPLE LINEAR REGRESSION

Source of Variation	df	Sum of Squares	Mean Square	F
Regression	1	SSR	SSR	$\frac{\text{SSR}}{\text{MSE}}$
Error	$n-2$	SSE	$s^2 = \text{MSE} = \frac{\text{SSE}}{n-2}$	
Total	$n-1$	SST		

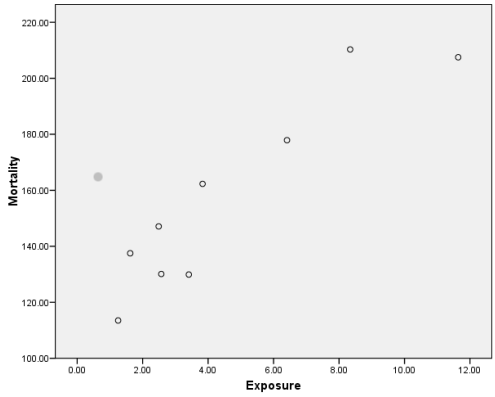
The F-test statistic in the ANOVA table for regression also tests the hypothesis $H_0 : \beta_1 = 0$. Its p-value always gives the exact same result as the one obtained from a t-test for zero slope because it is generally true that $t^2_{\alpha/2,n-2} = F_{\alpha,1,n-2}$.

Remark: Completely analogous to the above case, it may also be of interest to test whether the intercept β_0 is zero. This can also be done with a t-test and the results are simultaneously reported by SPSS in the "Coefficients" table.

Example: Since Word War II, plutonium for use in atomic weapons has been produced at an Atomic Energy Commission facility in Hanford, Washington. One of the major safety problems encountered there has been the storage of radioactive wastes. Over the years, significant quantities of these substances - including strontium 90 and cesium 137 - have leaked from their open-pit storage areas into the nearby Columbia River, which flows along the Washington-Oregon border, and eventually empties into the Pacific Ocean.

To measure the health consequences of this contamination, an index of exposure was calculated for each of the nine Oregon counties having frontage on either the Columbia River or the Pacific Ocean. This particular index was based on several factors, including the county's stream distance from Hanford and the average distance of its population from any water frontage. As a covariate, the cancer mortality rate was determined for each of these same counties.

(a) Look at the scatterplot of mortality rate against exposure. Does this seem like a good candidate for regression?



(b) Interpret the ANOVA table for the regression model

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8309.556	1	8309.556	42.336	.000[a]
	Residual	1373.946	7	196.278		
	Total	9683.502	8			

a. Predictors: (Constant), Exposure

b. Dependent Variable: Mortality

(c) Use $\hat{\beta}_1 = 9.231$, $S_{xx} = 97.5$, $t_{0.025,7} = 2.36$, and the estimate of σ^2 (from the ANOVA table) to compute a 95% confidence interval for the slope β_1.

(d) Read off the 95% confidence intervals for both the slope and the intercept from the COEFFICIENT TABLE produced by SPSS and interpret them in the context of this problem.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	114.716	8.046		14.258	.000	95.691	133.741
	Exposure	9.231	1.419	.926	6.507	.000	5.877	12.586

a. Dependent Variable: Mortality

Inference about Predictions

For a given value x of the predictor X, we know that the response Y is a normal random variable with mean

$$\mu_{Y|x} = \beta_0 + \beta_1 x$$

and variance σ^2. However, in practice, the parameters β_0, β_1, and σ^2 are unknown and have to be replaced by their estimates. Since the estimates depend on the observations (data) the resulting mean estimate $\hat{\mu}_{Y|x}$ is also a random variable. This formulation allows us to predict the behavior of the response Y for a reasonable range of the predictor X.

PROPOSITION: Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ be the predicted value of the response at $X = x^*$. Then $\hat{Y}$ has a normal distribution with:

Mean $E(\hat{Y}) = E(\hat{\mu}_{Y|x=x^*}) = \beta_0 + \beta_1 x^*$, i.e., the estimator of the mean of the response is unbiased.

Variance The variance of $\hat{Y}$ is

$$V(\hat{Y}) = \sigma^2_{\hat{Y}} = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right].$$

To estimate the variance, the error variance σ^2 is replaced by its estimate s^2=MSE.

CONFIDENCE INTERVAL: The above statements allow us to formulate a confidence interval for $\mu_{Y|x=x^*}$:

$$CI_{\mu_{Y|x^*}} = \left[\hat{y} \pm t_{\alpha/2, n-2} s_{\hat{y}} \right]$$

Note that both $\hat{y}$ and $s_{\hat{y}}$ depend on the value of x^*. Therefore the confidence region for the mean of response values $\hat{y}$ as a function of x^* has not a rectangular shape but flares for x-values very different from $\bar{x}$.

Note: If you want to make statements about many different $\hat{y}$-values for different x-values, you have to adjust the confidence level (similarly to the Tukey procedure) in order to keep the overall-confidence level at α.

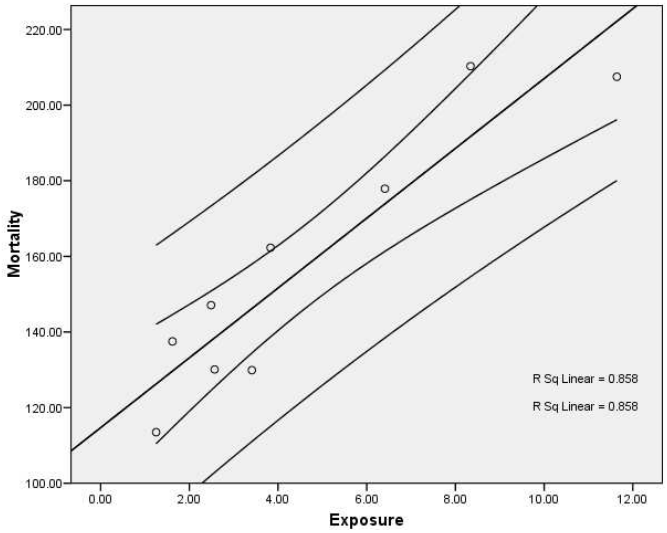
Remark: There is a subtle difference between a confidence interval for the mean $\mu_{Y|x}$ and a PREDICTION INTERVAL for a *future prediction* of Y to be made when $x = x^*$. A confidence interval refers to an unobservable model parameter and its estimation through the data. A prediction interval is an interval in which we expect to see future values of Y $100(1-\alpha)\%$ of the time. There is more uncertainty in predictions than in estimation. Thus, prediction intervals are generally wider than confidence intervals for estimations.

Definition: A $100(1-\alpha)\%$ PREDICTION INTERVAL for a future Y observation to be made at $x = x^*$ is

$$\text{CI} = \left[\hat{y} \pm t_{\alpha/2,n-2}\sqrt{s^2 + s^2_{\hat{y}}}\right] = \left[\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}\right]$$

Example: (cont.)

(e) For the Hanford, Washington cancer rate example, a 95% confidence interval for the expected value $\hat{\mu}_Y$, and a 95% prediction interval as produced by SPSS look like this:



Inference on Correlation

In many cases, the objective of a study is to establish a correlation between two observed random variables. Did the proximity to the reactor in Hanford CAUSE higher mortality rates in the population? Even though statistical studies *cannot* prove (or disprove) causation, they can be used to establish ASSOCIATION of two variables.

Recall: The sample correlation coefficient for observation pairs $(x_1, y_1), \ldots (x_n, y_n)$ is defined as

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Since the data are observations on random variables, the sample correlation coefficient is also a random variable. Thus, it makes sense to perform hypothesis tests for the true (population) correlation ρ of the observed random variables. We will use r as an estimate for ρ.

HYPOTHESIS TEST: Null hypothesis $H_0 : \rho = 0$ (the variables are uncorrelated).
Test Statistic:

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

Example: (cont.)

(f) For the Hanford, Washington cancer example, the sample correlation coefficient between EXPOSURE and MORTALITY is $r = 0.926$. Test the null hypothesis that the random variables are uncorrelated (at level $\alpha = 0.05$).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.926[a]	.858	.838	14.00993

a. Predictors: (Constant), Exposure