

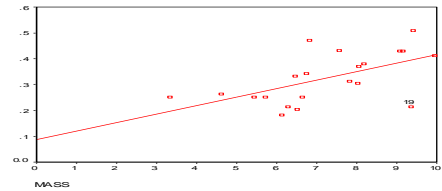
36-309/749
 Experimental Design for Behavioral
 and Social Sciences

Sep. 22, 2015
 Lecture 4: Linear Regression

Simple Regression Example

Male black wheatear birds carry stones to the nest as a form of sexual display. Soler *et al.* wanted to find out whether there is a relationship between the mass (in g) of the stones carried and the immunologic health of the birds as measured by a test of T cell function (in mm).

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
MASS	21	3.33	9.95	7.2043	1.70244
TCELL	21	.18	.51	.3240	.09674
Valid N (listwise)	21				



Example, cont.

Model		Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
		B	Std. Error			Lower Bound	Upper Bound
1	(Constant)	.087	.079	1.112	.280	-.077	.252
	MASS	.033	.011	3.084	.006	.011	.055

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.062	1	.062	9.513	.006 ^b
	Residual	.125	19	.007		
	Total	.187	20			

a. Predictors: (Constant), MASS

b. Dependent Variable: TCELL

Regression: Main Ideas

- **Setting:** Quantitative outcome with a quantitative explanatory variable
- **Model:** (§9.1)
 - Structural (means) model: $E(Y|x) = \beta_0 + \beta_1 x$ (parameters are called "betas" or "coefficients")
 - This defines a linear relationship, on average (linearity assumption).
 - The intercept, β_0 , is the (population) mean of Y when $x = 0$.
 - The slope, β_1 , is the (population) **mean change** in Y when x **increases by 1**.

Model, cont.

- Fixed-x assumption: x 's are measured with no (little) error
- Error model
 - "Errors" (deviations of Y from $\beta_0 + \beta_1 x$) are **Gaussian** ...
 - ... with **constant** variance, σ^2 , called "error variance"
 - ... and the errors are **independent** of each other.
- Alternate model form: $Y_i | X_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ $\varepsilon \sim N(0, \sigma^2)$ iid

5

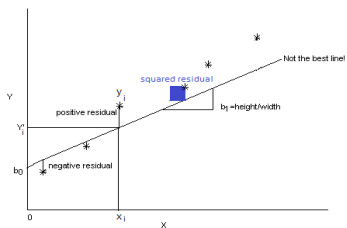
Main Ideas, cont.

- **Simple vs. multiple** regression
- **Null hypothesis:** $H_0: \beta_1=0$ vs. $H_1: \beta_1 \neq 0$.
(Sometimes: $H_0: \beta_0=0$ vs. $H_1: \beta_0 \neq 0$
or $H_0: \beta_1=1$ vs. $H_1: \beta_1 \neq 1$) (§9.2)
- **Interpolation:** good / **extrapolation:** bad

6

Main Ideas, cont.

- The "**least squares principle**" finds the best-fit line by minimizing the sum of squared residuals. (§9.4)



7

Main Ideas, cont.

- **Estimates:** "best" estimates of β_0 and β_1 are called b_0 and b_1 or $\widehat{\beta}_0$ and $\widehat{\beta}_1$ (read "**beta 0 hat**" and "**beta 1 hat**"). (§9.4, 9.5)

- Technical & optional: The estimates of the coefficients are statistics that are unbiased, minimum variance estimates. For linear regression, "least squares" is the same as "maximum likelihood".

8

Main Ideas, cont.

- **Fitted (predicted) values:** $= \hat{Y}_i = Y_i' = \text{fit}_i = \text{pred}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = b_0 + b_1 x_i$. (§9.4)

- **Residual:** $\text{res}_i = r_i = Y_i - \text{fit}_i = Y_i - \hat{Y}_i$
This is an estimate of the true "error". (§9.4, 9.6)

9

Main Ideas, cont.

- **Optional estimation formulas:**

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad b_0 = \bar{y} - b_1 \bar{x}, \quad s^2 = \frac{\sum_{i=1}^n r_i^2}{n-2}$$

- **$Ms_{\text{within}} = Ms_{\text{resid}} = Ms_{\text{error}} = SS_{\text{resid}}/df$** are exactly the same estimate of error variance, σ^2 . (§9.8)

- In general, the df equals (N - #betas).

10

Main ideas, cont.

- The Normality assumption leads to a Normal **sampling distribution of $\hat{\beta}_1$** with mean β_1 and a variance that is a complicated combination of σ^2 and the x values. When we substitute in the estimate of σ^2 and take the square root, we get a standard error (SE) of $\hat{\beta}_1$ that can be used in a t-test (§9.4) of $H_0: \beta_1=0$:

- **95% confidence interval** on β_1 :

$[b_1 - m \text{SE}(b_1), b_1 + m \text{SE}(b_1)]$, where m is approximately 2 and the exact value comes from finding the value of the appropriate t distribution that holds 95% of the distribution. (§9.4)

11

Main ideas, cont.

- **Goodness of fit /model comparison:** (§9.8)

- R^2 (0 to 1) measures the fraction of the variability in the outcome "accounted for" by the explanatory variable(s). It is also r^2 in simple (one x) regression where r is the correlation of x and Y. (See slide 23 below for the formula.)

- Adjusted R^2 corrects for the "more IVs **always** gives a higher R^2 " problem.

- Many consider AIC / BIC, which are "penalized likelihoods", even better for comparing models, e.g., deciding if some additional "x" should be included.

12

Main Ideas, cont.

- **Robustness** of the t-test in regression (§9.7)
 - moderately severe non-normality is OK (CLT)
 - mild to moderate unequal spread are OK (ratio<2)
 - only minimal correlation of errors is OK
 - non-fixed X is OK only if its variability is much less than the variability of Y
 - non-linearity is bad

13

SPSS Output and Interpretation

- SPSS has some mechanisms for trying different sets of explanatory variables (**model selection**) in multiple (>1 x) regression. The default is to include all specified x variables.

Model	Variables Entered	Variables Removed	Method
1	MASS ^b	.	Enter

a. All requested variables entered.
 b. Dependent Variable: TCELL

14

SPSS Output and Interp., cont.

- There is a statistically significant positive **association** between the mass of stones collected and the T cell activity (reject $H_0: \beta_1=0$ or equivalently $\beta_{MASS}=0$; $p=0.006$). *If* the levels of the explanatory variable were randomly assigned, we could say that a **rise** in stone mass **causes** a **rise** in T cell function.
 - Concluding only that “stone mass is statistically significant” is stupid!!!

Model	Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B		
	B	Std. Error			Lower Bound	Upper Bound	
1	(Constant)	.087	.079	1.112	.280	-.077	.252
	MASS	.033	.011	3.084	.006	.011	.055

a. Dependent Variable: TCELL

15

SPSS Output and Interp., cont.

- The three major **causes of an association between variables** X and Y are that changes in X cause changes in Y, changes in Y cause changes in X, and that changes in a third variable causes changes in both X and Y. In addition, type 1 error can result in an **apparent** association between X and Y when there really is none (a type-1 error).
- **Interpretation of slope coefficient:** A rise of 1 gram of mass is associated with an estimated mean rise of 0.033 mm in T cell function. It is also OK (better) to say that, e.g., a rise of 10 gram of mass is associated with an estimated 0.33 mm mean rise in T cell function.

16

SPSS Output and Interp., cont.

- The intercept is **the mean of Y when x is 0**.
 - In this example, since $x=0$ is far from the rest of the data, an **interpretation of the intercept** would be an *unwarranted extrapolation*. If we were to make that extrapolation we would express it as “the *predicted* (estimated) *mean* of T cell function for wheatears that carry *no* stones is 0.087 mm”. Here, the non-significant p-value (and the CI) tell us that we cannot rule out that the mean T cell response for 0 g of stones is 0 mm (cannot reject $H_0; \beta_0=0$).
 - **Only interpret the intercept p-value when $x=0$ makes sense and when we have data at or near $x=0$ and when we care if $E(Y|x=0)$ equals zero or not, i.e. was originally unknown.**

17

SPSS Output and Interp., cont.

- The “CI for B_{MASS} ” has the **interpretation** that “we are **95% confident** that the **mean rise** in T cell function associated with a **one gm increase** in stone mass is between 0.011 and 0.055.”
 - Technically, when the assumptions of the model are met, CI’s constructed by the standard recipe used here will include the true coefficient value 95% of the time (over repeated experiments, whether or not the null hypothesis is true, but only when the assumptions are true).

18

SPSS Output and Interp., cont.

- The **mean squared error (MSE)** is 0.007 which is the best estimate of σ^2 . Our model predicts that for any given x value, the distribution of the outcome for many subjects with that same x value will be normally distributed with mean $\beta_0 + \beta_1 x$ and variance 0.007 and $sd = \sqrt{0.007} = 0.081$ and $2sd = 0.162$.

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.062	1	.062	9.513	.006 ^b
	Residual	.125	19	.007		
	Total	.187	20			

- a. Predictors: (Constant), MASS
b. Dependent Variable: TCELL

19

SPSS Output and Interp., cont.

- The “case-wise diagnostics” tries to detect possible **outliers**. Here it tells us that the subject farthest (vertically) from the best fit line is subject 19, who has actual outcome 0.21 mm, predicted outcome 0.39 mm, and residual -0.18 . So this bird had a T cell measurement 0.18 lower than “expected” by the model. The **standardized residual** divides by the standard deviation of the residuals; -2.24 is not highly unlikely.

Casewise Diagnostics^a

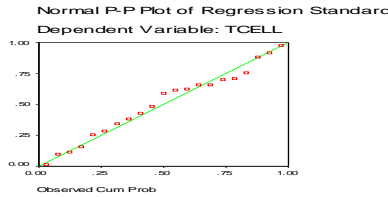
Case Number	Std. Residual	TCELL	Predicted Value	Residual
19	-2.239	.21	.3944	-.1814

- a. Dependent Variable: TCELL

20

SPSS Output and Interp., cont.

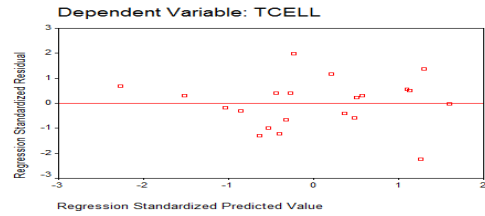
- The “Normal P-P” (similar to the quantile-quantile or **quantile-normal**) plot of the residuals helps test the assumption of a normal distribution of the Y values at each x by combining the residuals from all x values.



21

SPSS Output and Interp., cont.

- The **residual vs. fit plot** of “Standardized Residual” vs “Standardized Predicted Value” helps **test the assumptions of linearity and equal spread**. (Alternatively un-standardized values are used, but they are a bit harder to get in SPSS.) Use the “vertical band” method of interpretation.



22

SPSS Output and Interp., cont.

- The “**R-Squared**” value of 0.334 indicates that 33.4% of the overall variation in the outcome has been “explained” by the explanatory variable(s).
 - $SS_{Total} = \sum (y_i - \bar{y})^2$, $SS_{Error} = \sum (y_i - \hat{y}_i)^2$, $SS_{Explained} = SS_{Total} - SS_{Error}$
 - $R^2 = (SS_{Total} - SS_{Error}) / SS_{Total}$: fraction of the total deviations that is explained by the explanatory variable(s)
 - R^2 values range from 0 to 1 (high in physics, low in psychology)
 - **Adjusted- R^2** corrects for “more is always better” and is more realistic
 - AIC/BIC are even better for model comparison/selection

Model Summary^a

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578 ^a	.334	.299	.08102

a. Predictors: (Constant), MASS
b. Dependent Variable: TCELL

23

Conclusion

There are lots of useful regression results beyond just a p-value!

24