Inference for Regression

- The simple linear regression model
- Estimating regression parameters;
- Confidence intervals and significance tests for regression parameters
- Inference about prediction
- Analysis of variance for regression
- The regression fallacy

Simple Linear Regression

Earlier in the course we discussed how to find the best fitting line for bivariate data. Here, we consider that problem from the perspective of statistical inference.

Suppose we observe pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$

For example,

- x=father's height, y=son's height
- x=midterm score, y=final score
- x=temperature, y=yield

The values of x define different groups of subjects, which we think of as belonging to **subpopulations**: one for each possible value of x.

Let $\mu_{y|x}$ and $\sigma^2_{y|x}$ denote the mean and variance of y in the subpopulation with a certain value of x.

Under the **linear regression model with equal variance**, $\mu_{y|x} = \beta_0 + \beta_1 x$ and $\sigma^2_{y|x} = \sigma^2$, possibly after transformation.

Simple Linear Regression Model

The simple linear regression model states that the response variable y and the explanatory variable x have a linear relationship of the form

$$y = \beta_0 + \beta_1 x + \epsilon$$

where

- β_0 and β_1 are the y-intercept and the slope of the true population regression line.
- $\epsilon \sim N(0, \sigma)$
- The ϵ_i corresponding to the pairs (x_i, y_i) are independent of each other
- Given x, y has mean $\beta_0 + \beta_1 x$ and variance σ^2

  $$E(y|x) = \mu_{y|x} = \beta_0 + \beta_1 x$$

  is called the **population regression line.**

  $$\text{Var}(y|x) = \sigma^2_{y|x} = \sigma^2$$

Estimating the Regression Parameters by Least-Squares

Given a sample of n pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$, we use the *method of least squares* to estimate the unknown parameters β_0, β_1, and σ. This gives us the fitted line

$$\hat{y} = b_0 + b_1 x$$

where

- $b_1 = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} = r\frac{s_y}{s_x}$ is the estimate of β_1
- $b_0 = \bar{y} - b_1\bar{x}$ is the estimate of β_0

Recall that the residual is the difference between the observed value and the predicted value:

$$e_i = y_i - (b_0 + b_1 x_i) = y_i - \hat{y}_i$$

The sample variance of e_i can be used to estimate σ^2:

$$s^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$$

s is called the **regression standard error** and it has $n-2$ degrees of freedom.

Why $n-2$ degrees of freedom?

CIs for Regression Parameters

Under the assumption that $\epsilon \sim N(0, \sigma)$,

$$b_1 \sim N\left(\beta_1, \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}}\right)$$

$$b_0 \sim N\left(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}\right)$$

We don't know σ, so we will use s to estimate it. This leads to t-confidence intervals for β_0 and β_1.

Conditions for regression inference

- The sample is an SRS from the population
- There is a linear relationship in the population

  We check this condition by assessing the linearity of a scatterplot of the sample data.
- The standard deviation of the responses about the population line is the same for all values of the explanatory variable.

  We check this by plotting the residuals and observing whether or not the spread of the observations around the least-squares line is roughly uniform as x varies.
- The response varies Normally about the population regression line.

  We check this condition by observing a *Normal quantile plot* of the residuals.

Note that the last three conditions are statements about the *population* that cannot be verified directly. We use the sample to assess their reasonability.

A **level** $(1-\alpha)$ **confidence interval for** β_0 is given by

$$(b_0 - t^* \,\mathrm{SE}(b_0),\ b_0 + t^* \,\mathrm{SE}(b_0))$$

where t^* is the upper $\alpha/2$ critical value of the t_{n-2} distribution and

$$\mathrm{SE}(b_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$

A **level** $(1-\alpha)$ **confidence interval for** β_1 is given by

$$(b_1 - t^* \,\mathrm{SE}(b_1),\ b_1 + t^* \,\mathrm{SE}(b_1))$$

where t^* is the upper $\alpha/2$ critical value of the t_{n-2} distribution and

$$\mathrm{SE}(b_1) = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$$

Hypothesis Tests for Regression Parameters

To test the hypothesis $H_0 : \beta_1 = a$, we use the test statistic

$$T = \frac{b_1 - a}{\text{SE}(b_1)}$$

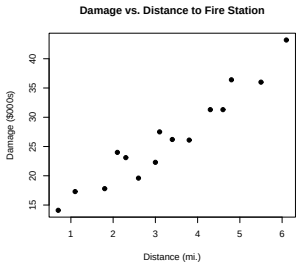
- the p-value for the test statistic is found from the t_{n-2} distribution
- if the regression assumptions are true, testing $H_0 : \beta_1 = 0$ corresponds to testing whether or not there is a linear relationship between y and x

A similar test can be performed for β_0, but it is rarely of interest.

Example: Fire damage and distance to fire station

Suppose a fire insurance company wants to relate the amount of fire damage in major residential fires to the distance between the burning house and the nearest fire station. The study is to be conducted in a large suburb of a major city; a sample of 15 recent fires in this suburb is selected. The amount of damage (in thousands of dollars) and the distance (in miles) between the fire and the nearest fire station are recorded in each fire.

Obs.	Dist.	Damage
1	0.7	14.1
2	1.1	17.3
3	1.8	17.8
4	2.1	24.0
5	2.3	23.1
6	2.6	19.6
7	3.0	22.3
8	3.1	27.5
9	3.4	26.2
10	3.8	26.1
11	4.3	31.3
12	4.6	31.3
13	4.8	36.4
14	5.5	36.0
15	6.1	43.2



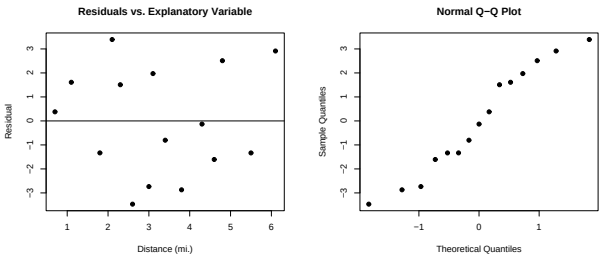
Performing this regression analysis in STATA yields the following results:

```
. regress damage distance

  Source |       SS       df       MS      Number of obs =      15
---------+------------------------------   F(  1,    13) =  156.89
   Model |  841.766403     1  841.766403   Prob > F      =  0.0000
Residual |  69.7509869    13  5.36546053   R-squared     =  0.9235
---------+------------------------------   Adj R-squared =  0.9176
   Total |   911.51739    14   65.108385   Root MSE      =  2.3163

------------------------------------------------------------------------------
  damage |    Coef.   Std. Err.      t    P>|t|   [95% Conf. Interval]
---------+--------------------------------------------------------------------
distance | 4.919331   .3927478   12.525  0.000   4.070851    5.767811
   _cons | 10.27793   1.420278    7.237  0.000   7.209605    13.34625
------------------------------------------------------------------------------
```

The following are a residual plot and a normal quantile plot of the residuals:



Example (cont.)

The fitted line is:

$$\text{damage} = 10.28 + 4.92\,\text{dist}$$

Suppose we want to predict the **mean amount of damage** for fires 2 miles from the nearest fire station. In this case, $x^* = 2$ and our prediction is

$$10.28 + 4.92 \times 2 = 20.12$$

Inference about Prediction

What if we want to predict the amount of damage of a burning house which is 2 miles from the nearest fire station? Still, the prediction is:

$$10.28 + 4.92 \times 2 = 20.12$$

The predicted values are the same, but they have different standard errors. Individual burning houses which are 2 miles away from the fire station don't have the same amount of damage, so the prediction for individual amount of damage has larger standard error than the prediction for mean amount of damage.

CIs for the Mean Response

For a specific value of x, say x^*, the assumption is that y comes from a $N(\mu_{y|x^*}, \sigma)$ distribution, where

$$\mu_{y|x^*} = \beta_0 + \beta_1 x^*$$

Plugging in our estimates of β_0 and β_1, $\mu_{y|x^*}$ is estimated by $\hat{\mu}_{y|x^*} = b_0 + b_1 x^*$, and a level $(1-\alpha)$ confidence interval for the mean response $\mu_{y|x^*}$ is given by

$$\hat{\mu}_{y|x^*} \pm t^* \text{SE}(\hat{\mu}_{y|x^*})$$

where t^* is the upper $\alpha/2$ critical value of the t_{n-2} distribution and

$$\text{SE}(\hat{\mu}_{y|x^*}) = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

How accurate is this estimate?

The error here will be larger than the error for the mean response, $\text{SE}(\hat{\mu}_{y|x^*})$, because there is error in estimating $\mu_{y|x^*}$ as well as error in drawing a value from the normal distribution $N(\mu_{y|x^*}, \sigma)$.

A **level** $(1-\alpha)$ **prediction interval** for a future observation y corresponding to x^* is given by

$$\hat{y} \pm t^* s_{\hat{y}}$$

where t^* is the upper $\alpha/2$ critical value of the t_{n-2} distribution and

$$s_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Prediction Interval for a Future Observation

Suppose we want to predict a specific observation value at $x = x^*$.

At each x^*, $y \sim N(\mu_{y|x^*}, \sigma)$ We want to predict a y drawn from this distribution.

Our best guess is the estimated mean of the distribution,

$$\hat{y} = \hat{\mu}_{y|x^*} = b_0 + b_1 x^*$$

Analysis of Variance for Regression

Analysis of variance is the term for statistical analyses that break down the variation in data into separate pieces that correspond to different sources of variation. In the regression setting, the observed variation in the responses comes from two sources.

- As the explanatory variable x changes, it "pulls" the response with it along the regression line. This is the **variation along the line** or **regression sum of squares**:

$$\text{SS(Regression)} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

- When x is held fixed, y still varies because not all individuals who share a common x have the same response y. This is the **variation about the line** or **residual sum of squares**:

$$\text{SS(Residual)} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The ANOVA Equation

It turns out that SS(Residual) and SS(Regression) together account for *all* the variation in y:

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{SS(Total)}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{SS(Regression)}} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{SS(Residual)}}$$

The degrees of freedom break down in a similar manner:

$$\underbrace{n-1}_{\text{SS(Total)}} = \underbrace{1}_{\text{SS(Regression)}} + \underbrace{n-2}_{\text{SS(Residual)}}$$

Dividing a sum of squares by its degrees of freedom gives a **mean square (MS)**.

$$\text{MS(Residual)} = \frac{\text{SS(Residual)}}{\text{df(Residual)}} = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2} = s^2$$

$$R^2 = \frac{\text{Regression SS}}{\text{Total SS}} = r^2$$

The ANOVA F Statistic

As an alternative test of the hypothesis: $H_0 : \beta_1 = 0$, we use the F statistic:

$$\begin{aligned} F &= \frac{\text{MS(Regression)}}{\text{MS(Residual)}} \\ &= \frac{\text{SS(Regression)}/\text{df(Regression)}}{\text{SS(Residual)}/\text{df(Residual)}} \\ &= \left(\frac{b_1}{\text{SE}_{b_1}}\right)^2 \\ &= t^2 \end{aligned}$$

Under H_0,

$$F \sim F_{1,n-2}$$

where $F_{1,n-2}$ is an F distribution with 1 and $n-2$ degrees of freedom.

The Regression Fallacy

Sir Francis Galton (1822–1911), who was the first to apply regression to biological and psychological data, looked at examples such as the heights of children versus the heights of their parents. He found that the taller-than-average parents tended to have children who were also taller than average, but not as tall as their parents. Galton called this fact "regression toward mediocrity".

As another example, students who score at the bottom on the first exam in a course are likely to do better on the second exam. Is it because they work harder?