Inference for Regression

IPS Chapter 10

- 10.1: Simple Linear Regression
- 10.2: More Detail about Simple Linear Regression
Inference for Regression

10.1 Simple Linear Regression
Objectives

10.1 Simple linear regression

- Statistical model for linear regression
- Estimating the regression parameters
- Confidence interval for regression parameters
- Significance test for the slope
- Confidence interval for $\mu_y$
- Prediction intervals
The data in a scatterplot are a random sample from a population that may exhibit a linear relationship between $x$ and $y$. Different sample $\rightarrow$ different plot.

Now we want to describe the population mean response $\mu_y$ as a function of the explanatory variable $x$: $\mu_y = \beta_0 + \beta_1 x$.

And to assess whether the observed relationship is statistically significant (not entirely explained by chance events due to random sampling).
Statistical model for linear regression

In the population, the linear regression equation is \( \mu_y = \beta_0 + \beta_1 x \).

Sample data then fits the model:

\[
\begin{align*}
\text{Data} &= \text{fit} + \text{residual} \\
y_i &= (\beta_0 + \beta_1 x_i) + (\varepsilon_i)
\end{align*}
\]

where the \( \varepsilon_i \) are independent and Normally distributed \( N(0, \sigma) \).

Linear regression assumes equal variance of \( y \) (\( \sigma \) is the same for all values of \( x \)).
Estimating the parameters

\[ \mu_y = \beta_0 + \beta_1 x \]

The intercept \( \beta_0 \), the slope \( \beta_1 \), and the standard deviation \( \sigma \) of \( y \) are the unknown parameters of the regression model. We rely on the random sample data to provide unbiased estimates of these parameters.

- The value of \( \hat{y} \) from the least-squares regression line is really a prediction of the mean value of \( y \) (\( \mu_y \)) for a given value of \( x \).
- The least-squares regression line (\( \hat{y} = b_0 + b_1 x \)) obtained from sample data is the best estimate of the true population regression line (\( \mu_y = \beta_0 + \beta_1 x \)).

\[ \hat{y} \] unbiased estimate for mean response \( \mu_y \)

\[ b_0 \] unbiased estimate for intercept \( \beta_0 \)

\[ b_1 \] unbiased estimate for slope \( \beta_1 \)
The **population standard deviation** $\sigma$ for $y$ at any given value of $x$ represents the spread of the normal distribution of the $\epsilon_i$ around the mean $\mu_y$.

The **regression standard error**, $s$, for $n$ sample data points is calculated from the residuals $(y_i - \hat{y}_i)$:

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

$s$ is an unbiased estimate of the regression standard deviation $\sigma$. 

$\mu_y = \beta_0 + \beta_1x$
Conditions for inference

- The observations are **independent**.
- The relationship is indeed **linear**.
- The standard deviation of $y, \sigma$, is the same for all values of $x$.
- The response $y$ varies **normally** around its mean.
Using residual plots to check for regression validity

The residuals \((y - \hat{y})\) give useful information about the contribution of individual data points to the overall pattern of scatter.

We view the residuals in a residual plot:

If residuals are scattered randomly around 0 with uniform variation, it indicates that the data fit a linear model, have normally distributed residuals for each value of \(x\), and constant standard deviation \(\sigma\).
Residuals are randomly scattered → **good**!

Curved pattern → the relationship is **not linear**.

Change in variability across plot → **σ not equal** for all values of $x$. 
What is the relationship between the average speed a car is driven and its fuel efficiency?

We plot fuel efficiency (in miles per gallon, MPG) against average speed (in miles per hour, MPH) for a random sample of 60 cars. The relationship is curved.

When speed is log transformed (log of miles per hour, LOGMPH) the new scatterplot shows a positive, **linear** relationship.
Residual plot:
The spread of the residuals is reasonably random—no clear pattern. The relationship is indeed linear.
But we see one low residual \((3.8, -4)\) and one potentially influential point \((2.5, 0.5)\).

Normal quantile plot for residuals:
The plot is fairly straight, supporting the assumption of normally distributed residuals.

⇒ Data okay for inference.
Confidence interval for regression parameters

Estimating the regression parameters $\beta_0, \beta_1$ is a case of one-sample inference with unknown population variance.

* We rely on the $t$ distribution, with $n – 2$ degrees of freedom.

A level $C$ confidence interval for the slope, $\beta_1$, is proportional to the standard error of the least-squares slope:

$$ b_1 \pm t^* \ SE_{b_1} $$

A level $C$ confidence interval for the intercept, $\beta_0$, is proportional to the standard error of the least-squares intercept:

$$ b_0 \pm t^* \ SE_{b_0} $$

$t^*$ is the $t$ critical value for the $t (n – 2)$ distribution with area $C$ between $–t^*$ and $+t^*$. 
Significance test for the slope

We can test the hypothesis $H_0: \beta_1 = 0$ versus a 1 or 2 sided alternative.

We calculate $t = b_1 / SE_{b_1}$ which has the $t(n-2)$ distribution to find the p-value of the test.

$H_a: \beta_1 > 0$ is $P(T \geq t)$

$H_a: \beta_1 < 0$ is $P(T \leq t)$

$H_a: \beta_1 \neq 0$ is $2P(T \geq |t|)$

Note: Software typically provides two-sided p-values.
Testing the hypothesis of no relationship

We may look for evidence of a **significant relationship** between variables $x$ and $y$ in the population from which our data were drawn.

For that, we can test the hypothesis that the regression slope parameter $\beta$ is equal to zero.

$$H_0: \beta_1 = 0 \text{ vs. } H_0: \beta_1 \neq 0$$

Testing $H_0: \beta_1 = 0$ also allows to test the hypothesis of no **correlation** between $x$ and $y$ in the population.

**Note:** A test of hypothesis for $\beta_0$ is irrelevant ($\beta_0$ is often not even achievable).
Using technology

Computer software runs all the computations for regression analysis.

Here is some software output for the car speed/gas efficiency example.

The $t$-test for regression slope is highly significant ($p< 0.001$). There is a significant relationship between average car speed and gas efficiency.
**Regression Statistics**

- Multiple R: 0.946053015
- R Square: 0.895016308
- Adjusted R Square: 0.893206244
- Standard Error: 0.999516364
- Observations: 60

**ANOVA**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>493.9885883</td>
<td>493.9886</td>
<td>494.4668</td>
<td>4.50949E-30</td>
</tr>
<tr>
<td>Residual</td>
<td>58</td>
<td>57.94391174</td>
<td>0.999033</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>59</td>
<td>551.9325</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Coefficients**

- **intercept**: -7.796250129
- **logmph**: 7.874219013

- Standard Error: 1.154944262 for intercept, 0.354110611 for logmph
- t Stat: -6.75033 for intercept, 22.23661 for logmph
- P-value: 7.69E-09 for intercept, 4.51E-30 for logmph
- Lower 95%: -10.10812052 for intercept, 7.16530143 for logmph
- Upper 95%: -5.48437974 for intercept, 8.583047883 for logmph

**Excel**

- P-value for tests of significance
- Confidence intervals

**SAS**

- Root MSE: 0.99952
- R-Square: 0.8950
- Adj R-Sq: 0.8932

| Variable   | DF  | Parameter Estimate | Standard Error | t Value | Pr > |t| 95% Confidence Limits |
|------------|-----|--------------------|----------------|---------|-------|----------------------|
| Intercept  | 1   | -7.79625           | 1.15494        | -6.75   | <.0001| -10.10812052         |
| logmph     | 1   | 7.87422            | 0.35411        | 22.24   | <.0001| 7.16539              | 8.583047883   |
Using inference, we can also calculate a confidence interval for the population mean $\mu_y$ of all responses $y$ when $x$ takes the value $x^*$ (within the range of data tested):

This interval is centered on $\hat{y}$, the unbiased estimate of $\mu_y$.

The true value of the population mean $\mu_y$ at a given value of $x$, will indeed be within our confidence interval in $C\%$ of all intervals calculated from many different random samples.
The level $C$ confidence interval for the mean response $\mu_y$ at a given value $x^*$ of $x$ is:

$$\hat{\mu}_y \pm t_{n-2}^* \text{SE}_\mu$$

$t^*$ is the $t$ critical value for the $t (n – 2)$ distribution with area $C$ between $-t^*$ and $+t^*$.

A separate confidence interval is calculated for $\mu_y$ along all the values that $x$ takes.

Graphically, the series of confidence intervals is shown as a continuous interval on either side of $\hat{y}$. 

95% confidence interval for $\mu_y$
Inference for prediction

One use of regression is for predicting the value of $y$, $\hat{y}$, for any value of $x$ within the range of data tested: $\hat{y} = b_0 + b_1 x$.

But the regression equation depends on the particular sample drawn. More reliable predictions require statistical inference:

To estimate an individual response $y$ for a given value of $x$, we use a prediction interval.

If we randomly sampled many times, there would be many different values of $y$ obtained for a particular $x$ following $N(0, \sigma)$ around the mean response $\mu_y$. 
The **level C prediction interval for a single observation** on $y$ when $x$ takes the value $x^*$ is:

$$\hat{y} \pm t^*_{n-2} SE_{\hat{y}}$$

$t^*$ is the $t$ critical value for the $t (n-2)$ distribution with area $C$ between $-t^*$ and $+t^*$.

The prediction interval represents mainly the error from the normal distribution of the residuals $\varepsilon_i$.

Graphically, the series confidence intervals are shown as a continuous interval on either side of $\hat{y}$.
The **confidence interval** for $\mu_y$ contains with $C\%$ confidence the population mean $\mu_y$ of all responses at a particular value of $x$.

The **prediction interval** contains $C\%$ of all the individual values taken by $y$ at a particular value of $x$.

95% prediction interval for $\hat{y}$
95% confidence interval for $\mu_y$

Estimating $\mu_y$ uses a smaller confidence interval than estimating an individual in the population (sampling distribution narrower than population distribution).
The line graph suggests that 7% to 9% of those diagnosed with the flu died within about a week of diagnosis.

We look at the relationship between the number of deaths in a given week and the number of new diagnosed cases one week earlier.
1918 flu epidemic: Relationship between the number of deaths in a given week and the number of new diagnosed cases one week earlier.

**EXCEL**

Regression Statistics

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.911</td>
</tr>
<tr>
<td>R Square</td>
<td>0.830</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.82</td>
</tr>
<tr>
<td>Standard Error</td>
<td>85.07</td>
</tr>
<tr>
<td>Observations</td>
<td>16.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>St. Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>49.292</td>
<td>29.845</td>
<td>1.652</td>
<td>0.1209</td>
<td>(14.720)</td>
</tr>
<tr>
<td>FluCases0</td>
<td>0.072</td>
<td>0.009</td>
<td>8.263</td>
<td>0.0000</td>
<td>0.053</td>
</tr>
</tbody>
</table>

**P-value for**

\[ H_0: \beta_1 = 0 \]

P-value very small \( \Rightarrow \) reject \( H_0 \) \( \Rightarrow \) \( \beta_1 \) significantly different from 0

There is a **significant relationship** between the number of flu cases and the number of deaths from flu a week later.
CI for mean weekly death count one week after 4,000 flu cases are diagnosed: \( \mu_y \) within about 300–380.

Prediction interval for a weekly death count one week after 4,000 flu cases are diagnosed: \( \hat{y} \) within about 180–500 deaths.
What is this?

A 90% prediction interval for the height (above) and a 90% prediction interval for the weight (below) of male children, ages 3 to 18.
Inference for Regression

10.2 More Detail about Simple Linear Regression
Objectives

10.2 More detail about simple linear regression

- Analysis of variance for regression
- The ANOVA $F$ test
- Calculations for regression inference
- Inference for correlation
Analysis of variance for regression

The regression model is:

\[
\text{Data} = \begin{array}{c}
\text{fit} \\
y_i
\end{array} + \begin{array}{c}
\text{residual} \\
(\varepsilon_i)
\end{array}
\]

where the \( \varepsilon_i \) are independent and normally distributed \( N(0, \sigma) \), and \( \sigma \) is the same for all values of \( x \).

It resembles an ANOVA, which also assumes equal variance, where

\[
\text{SST} = \begin{array}{c}
\text{SS model} \\
\mu_y = \beta_0 + \beta_1 x
\end{array} + \begin{array}{c}
\text{SS error} \\
\text{DFT} = \begin{array}{c}
\text{DF model} \\
\text{DF error}
\end{array} + \begin{array}{c}
\text{DF error}
\end{array}
\]

For any fixed \( x \), the responses \( y \) follow a Normal distribution with standard deviation \( \sigma \).
The ANOVA $F$ test

For a simple linear relationship, the ANOVA tests the hypotheses

$$H_0: \beta_1 = 0 \text{ versus } H_a: \beta_1 \neq 0$$

by comparing MSM (model) to MSE (error): $F = \frac{\text{MSM}}{\text{MSE}}$

When $H_0$ is true, $F$ follows the $F(1, n-2)$ distribution.
The p-value is $P(F \geq f)$.

The ANOVA test and the two-sided t-test for $H_0: \beta_1 = 0$ yield the same p-value.
Software output for regression may provide $t$, $F$, or both, along with the p-value.
## ANOVA table

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares SS</th>
<th>DF</th>
<th>Mean square MS</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>$\sum (\hat{y}_i - \bar{y})^2$</td>
<td>1</td>
<td>SSM/DFM</td>
<td>MSM/MSE</td>
<td>Tail area above F</td>
</tr>
<tr>
<td>Error</td>
<td>$\sum (y_i - \hat{y}_i)^2$</td>
<td>$n-2$</td>
<td>SSE/DFE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$\sum (y_i - \bar{y})^2$</td>
<td>$n-1$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SST = SSM + SSE  
DFT = DFM + DFE

The **standard deviation of the sampling distribution**, $s$, for $n$ sample data points is calculated from the residuals $e_i = y_i - \hat{y}_i$

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

$s$ is an unbiased estimate of the regression standard deviation $\sigma$. 
The coefficient of determination, $r^2$, square of the correlation coefficient, is the percentage of the variance in $y$ (vertical scatter from the regression line) that can be explained by changes in $x$. 

$$r^2 = \frac{\text{variation in } y \text{ caused by } x \text{ (i.e., the regression line)}}{\text{total variation in observed } y \text{ values around the mean}}$$

$$r^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{SSM}{SST}$$
What is the relationship between the average speed a car is driven and its fuel efficiency?

We plot fuel efficiency (in miles per gallon, MPG) against average speed (in miles per hour, MPH) for a random sample of 60 cars. The relationship is curved.

When speed is log transformed (log of miles per hour, LOGMPH) the new scatterplot shows a positive, **linear** relationship.
Using software: SPSS

ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>1</td>
<td>493.989</td>
<td>494.467</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>58</td>
<td>.999</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>551.932</td>
<td>59</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LOGMPH
b. Dependent Variable: MPG

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.946a</td>
<td>.895</td>
<td>.893</td>
<td>.9995</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LOGMPH

Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>-7.796</td>
<td>1.155</td>
</tr>
<tr>
<td>LOGMPH</td>
<td>7.874</td>
<td>.354</td>
</tr>
</tbody>
</table>

Dependent Variable: MPG

ANOVA and t-test give same p-value.
Calculations for regression inference

To estimate the parameters of the regression, we calculate the standard errors for the estimated regression coefficients.

The standard error of the least-squares slope $\beta_1$ is:

$$SE_{b_1} = s \sqrt{\sum (x_i - \bar{x}_i)^2}$$

The standard error of the intercept $\beta_0$ is:

$$SE_{b_0} = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x}_i)^2}}$$
To estimate or predict future responses, we calculate the following standard errors.

The standard error of the mean response $\mu_y$ is:

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

The standard error for predicting an individual response $\hat{y}$ is:

$$\text{SE}_{\hat{y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x - \bar{x})^2}}$$
### 1918 influenza epidemic

<table>
<thead>
<tr>
<th>Date</th>
<th># Cases</th>
<th># Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>week 1</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>week 2</td>
<td>531</td>
<td>0</td>
</tr>
<tr>
<td>week 3</td>
<td>4233</td>
<td>130</td>
</tr>
<tr>
<td>week 4</td>
<td>8682</td>
<td>552</td>
</tr>
<tr>
<td>week 5</td>
<td>7164</td>
<td>738</td>
</tr>
<tr>
<td>week 6</td>
<td>2229</td>
<td>414</td>
</tr>
<tr>
<td>week 7</td>
<td>600</td>
<td>198</td>
</tr>
<tr>
<td>week 8</td>
<td>164</td>
<td>90</td>
</tr>
<tr>
<td>week 9</td>
<td>57</td>
<td>56</td>
</tr>
<tr>
<td>week 10</td>
<td>722</td>
<td>50</td>
</tr>
<tr>
<td>week 11</td>
<td>1517</td>
<td>71</td>
</tr>
<tr>
<td>week 12</td>
<td>1828</td>
<td>137</td>
</tr>
<tr>
<td>week 13</td>
<td>1539</td>
<td>178</td>
</tr>
<tr>
<td>week 14</td>
<td>2416</td>
<td>194</td>
</tr>
<tr>
<td>week 15</td>
<td>3148</td>
<td>290</td>
</tr>
<tr>
<td>week 16</td>
<td>3465</td>
<td>310</td>
</tr>
<tr>
<td>week 17</td>
<td>1440</td>
<td>149</td>
</tr>
</tbody>
</table>

The line graph suggests that about 7% to 8% of those diagnosed with the flu died within about a week of diagnosis. We look at the relationship between the number of deaths in a given week and the number of new diagnosed cases one week earlier.
1918 flu epidemic: Relationship between the number of deaths in a given week and the number of new diagnosed cases one week earlier.

**MINITAB - Regression Analysis:**

**FluDeaths1 versus FluCases0**

The regression equation is

\[ \text{FluDeaths1} = 49.3 + 0.0722 \times \text{FluCases0} \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>49.29</td>
<td>29.85</td>
<td>1.65</td>
<td>0.121</td>
</tr>
<tr>
<td>FluCases</td>
<td>0.072222</td>
<td>0.008741</td>
<td>8.26</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ S = 85.07 \]

\[ s = \sqrt{MSE} \]

\[ r^2 = \frac{SSM}{SST} \]

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>494041</td>
<td>494041</td>
<td>68.27</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>14</td>
<td>101308</td>
<td>7236</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>595349</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P-value for

\[ H_0: \beta_1 = 0; \quad H_a: \beta_1 \neq 0 \]
Inference for correlation

To test for the null hypothesis of no linear association, we have the choice of also using the correlation parameter $\rho$.

- When $x$ is clearly the explanatory variable, this test is equivalent to testing the hypothesis $H_0: \beta = 0$.

- When there is no clear explanatory variable (e.g., arm length vs. leg length), a regression of $x$ on $y$ is not any more legitimate than one of $y$ on $x$. In that case, the correlation test of significance should be used.

- When both $x$ and $y$ are normally distributed $H_0: \rho = 0$ tests for no association of any kind between $x$ and $y$—not just linear associations.

$$b_1 = r \frac{s_y}{s_x}$$
The test of significance for $\rho$ uses the one-sample $t$-test for: $H_0: \rho = 0$.

We compute the $t$ statistics for sample size $n$ and correlation coefficient $r$.

\[ t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \]

The $p$-value is the area under $t$ $(n - 2)$ for values of $T$ as extreme as $t$ or more in the direction of $H_a$:

- $H_a: \rho > 0$ is $P(T \geq t)$
- $H_a: \rho < 0$ is $P(T \leq t)$
- $H_a: \rho \neq 0$ is $2P(T \geq |t|)$
Relationship between average car speed and fuel efficiency

There is a significant correlation \((r\) is not 0) between fuel efficiency (MPG) and the logarithm of average speed (LOGMPH).
Cautions for Regression Inference

1. The observations must be independent.
   Repeated observations on the same cases or individuals.

2. The true relationship must be linear.
   Always plot your data. Look at the scatterplot to check that the overall pattern is roughly linear and that there are no outliers or influential points.

3. The standard deviation of the response about the true line is the same everywhere.
   Look at the scatterplot again. The scatter of the points about the line should be roughly the same over the entire range of the data. This is easier to check on a residual plot.

4. The response varies Normally about the true regression line.
   Make a histogram or stemplot of the residuals and check for skewness or other major departures from Normality.
Alternate Slides

The following slides offer alternate software output data and examples for this presentation.
Using technology

Computer software runs all the computations for regression analysis.
Here is some software output for the car speed/gas efficiency example.

\[ \text{MPG} = -7.79632 + 7.8742447 \text{LOGMPH} \]

| Term     | Estimate  | Std Error | t Ratio | Prob>|t| |
|----------|-----------|-----------|---------|--------|
| Intercept| -7.79632  | 1.154912  | -6.75   | <.0001*|
| LOGMPH   | 7.8742447 | 0.354101  | 22.24   | <.0001*|

The \( t \)-test for regression slope is highly significant (\( p < 0.0001 \)). There is a significant relationship between average car speed and gas efficiency.
MPG = -7.79632 + 7.8742447 LOGMPH

### Summary of Fit

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSquare</td>
<td>0.895022</td>
</tr>
<tr>
<td>RSquare Adj</td>
<td>0.893212</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
<td>0.999489</td>
</tr>
<tr>
<td>Mean of Response</td>
<td>17.725</td>
</tr>
<tr>
<td>Observations (or Sum Wgts)</td>
<td>60</td>
</tr>
</tbody>
</table>

### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>493.99177</td>
<td>493.992</td>
<td>494.4971</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Error</td>
<td>58</td>
<td>57.94073</td>
<td>0.999</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>59</td>
<td>551.93250</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Parameter Estimates

| Term        | Estimate | Std Error | t Ratio | Prob>|t| |
|-------------|----------|-----------|---------|------|
| Intercept   | -7.79632 | 1.154912  | -6.75   | <.0001*|
| LOGMPH      | 7.8742447| 0.354101  | 22.24   | <.0001*|

“intercept”: intercept
“logmph”: slope

P-value for tests of significance
1918 flu epidemic: Relationship between the number of deaths in a given week and the number of new diagnosed cases one week earlier. *JMP*

**Summary of Fit**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSquare</td>
<td>0.830</td>
</tr>
<tr>
<td>RSquare Adj</td>
<td>0.818</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
<td>85.066</td>
</tr>
<tr>
<td>Mean of Response</td>
<td>222.313</td>
</tr>
<tr>
<td>Observations (or Sum Wgts)</td>
<td>16.000</td>
</tr>
</tbody>
</table>

**Parameter Estimates**

| Term             | Estimate | Std Error | t Ratio | Prob>|t| |
|------------------|----------|-----------|---------|------|---|
| Intercept        | 49.292   | 29.8454   | 1.652   | 0.1209|   |
| New cases (-1)   | 0.072    | 0.0087    | 8.263   | <.0001*|   |

$b_1$

P-value for $H_0: \beta_1 = 0$

P-value very small $\Rightarrow$ reject $H_0$ $\Rightarrow$ $\beta_1$ significantly different from 0

There is a **significant relationship** between the number of flu cases and the number of deaths from flu a week later.
Using software: **JMP 6 SE**

### Summary of Fit

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSquare</td>
<td>0.895</td>
</tr>
<tr>
<td>RSquare Adj</td>
<td>0.893</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
<td>0.999</td>
</tr>
<tr>
<td>Mean of Response</td>
<td>17.725</td>
</tr>
<tr>
<td>Observations (or Sum Wgts)</td>
<td>60.000</td>
</tr>
</tbody>
</table>

### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>493.992</td>
<td>493.992</td>
<td>494.4971</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Error</td>
<td>58</td>
<td>57.941</td>
<td>0.999</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>59</td>
<td>551.932</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Parameter Estimates

| Term      | Estimate | Std Error | t Ratio | Prob>|t| |
|-----------|----------|-----------|---------|------|
| Intercept | -7.796   | 1.155     | -6.751  | <.0001* |
| LOGMPH    | 7.874    | 0.354     | 22.237  | <.0001* |

\[ r^2 = \frac{SSM}{SST} = \frac{494}{552} \]

ANOVA and t-test give same p-value.
1918 flu epidemic: Relationship between the number of deaths in a given week and the number of new diagnosed cases one week earlier.

**JMP – Regression Analysis**

**Linear Fit**

Deaths = 49.29 + 0.0722 New cases (−1)

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>SSM 494041</td>
<td>494041</td>
<td>68.27</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Error</td>
<td>14</td>
<td>101308</td>
<td>7236</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>15</td>
<td>SST 595349</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Parameter Estimates**

| Term                  | Estimate | Std Error | t Ratio | Prob>|t| |
|-----------------------|----------|-----------|---------|------|
| Intercept             | 49.2918  | 29.8454   | 1.65    | 0.1209 |
| New cases (−1)        | 0.0722   | 0.0087    | 8.26    | <.0001* |

\[ r^2 = \frac{SSM}{SST} = 0.8298 \]

P-value for \( H_0: \beta_1 = 0; H_a: \beta_1 \neq 0 \)
Relationship between average car speed and fuel efficiency

\[ H_o: \rho = 0 \]
\[ H_a: \rho \neq 0 \]

We had \( n = 60 \) and \( r = 0.946 \)

\[
t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} = 22.225, \text{ df } = 60-2 = 58.
\]

\( p\)-value < 0.0001

There is a significant correlation (\( r \) is not 0) between fuel efficiency (MPG) and the logarithm of average speed (LOGMPH).