23. Inference for regression

The Practice of Statistics in the Life Sciences
Third Edition
Objectives (PSLS Chapter 23)

Inference for regression

- The regression model
- Confidence interval for the regression slope $\beta$
- Testing the hypothesis of no linear relationship
- Inference for prediction
- Conditions for inference
Most scatterplots are created from *sample* data.

➔ Is the observed *relationship statistically significant* (not entirely explained by chance events due to random sampling)?

➔ What is the *population mean response* $\mu_y$ as a function of the explanatory variable $x$?

$$\mu_y = \alpha + \beta x$$
The least-squares regression line
\[ \hat{y} = a + bx \]
is a mathematical model of the relationship between two quantitative variables:

“sample data = fit + residual”

The regression line is the fit.

For each data point in the sample, the residual is the difference \((y - \hat{y})\).
At the population level, the model becomes

\[ y_i = (\alpha + \beta x_i) + (\varepsilon_i) \]

with residuals \( \varepsilon_i \) independent and Normally distributed \( N(0, \sigma) \).

The population mean response \( \mu_y \) is

\[ \mu_y = \alpha + \beta x \]

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

\( a \) unbiased estimate for intercept \( \alpha \)

\( b \) unbiased estimate for slope \( \beta \)
The regression standard error, $s$, for $n$ sample data points is computed 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 $\sigma$. 

Regression assumes equal variance of $Y$ ($\sigma$ is the same for all values of $x$).
Frog mating call
Scientists recorded the call frequency of 20 gray tree frogs, *Hyla chrysoscelis*, as a function of field air temperature.

**MINITAB**

Regression Analysis:
The regression equation is

\[ \text{Frequency} = -6.19 + 2.33 \text{ Temperature} \]

<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>-6.190</td>
<td>8.243</td>
<td>-0.75</td>
<td>0.462</td>
</tr>
<tr>
<td>Temperature</td>
<td>2.3308</td>
<td>0.3468</td>
<td>6.72</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ S = 2.82159 \quad \text{R-Sq} = 71.5\% \quad \text{R-Sq(adj)} = 69.9\% \]

\[ s: \text{regression standard error, unbiased estimate of } \sigma \]

\[ s = \sqrt{\frac{\sum \text{residual}^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} \]
Confidence interval for the slope $\beta$

Estimating the regression parameter $\beta$ for the slope is a case of one-sample inference with $\sigma$ unknown. Hence we rely on $t$ distributions.

The standard error of the slope $b$ is:

$$SE_b = \frac{s}{\sqrt{\sum (x - \bar{x})^2}}$$

$s$ is the regression standard error

Thus, a level $C$ confidence interval for the slope $\beta$ is:

$$\text{estimate} \pm t^*SE_{\text{estimate}}$$

Thus, $b \pm t^* SE_b$

$t^*$ is $t$ critical for $t(df = n - 2)$ density curve with $C\%$ between $-t^*$ and $+t^*$. 
Frog mating call

Regression Analysis:
The regression equation is
Frequency = -6.19 + 2.33 Temperature

<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>-6.190</td>
<td>8.243</td>
<td>-0.75</td>
<td>0.462</td>
</tr>
<tr>
<td>Temperature</td>
<td>2.3308</td>
<td>0.3468</td>
<td>6.72</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 2.82159   R-Sq = 71.5%   R-Sq(adj) = 69.9%

\[ n = 20, \text{ df } = 18 \]
\[ t^* = 2.101 \]

95% CI for the slope \( \beta \) is:

\[ b \pm t^*SE_b = 2.3308 \pm (2.101)(0.3468) = 2.3308 \pm 0.7286, \text{ or } 1.6022 \text{ to } 3.0594 \]

We are 95% confident that a 1 degree Celsius increase in temperature results in an increase in mating call frequency between 1.60 and 3.06 notes/seconds.
Testing the hypothesis of no relationship

To test for a significant relationship, we ask if the parameter for the slope $\beta$ is equal to zero, using a one-sample $t$ test.

The standard error of the slope $b$ is:

$$SE_b = \frac{S}{\sqrt{\sum (x - \bar{x})^2}}$$

We test the hypotheses $H_0: \beta = 0$ versus a one-sided or two-sided $H_a$.

We compute

$$t = \frac{b}{SE_b}$$

which has the $t\,(n-2)$ distribution to find the $P$-value of the test.
Frog mating call

Regression Analysis:
The regression equation is
Frequency = - 6.19 + 2.33 Temperature

<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>-6.190</td>
<td>8.243</td>
<td>-0.75</td>
<td>0.462</td>
</tr>
<tr>
<td>Temperature</td>
<td>2.3308</td>
<td>0.3468</td>
<td>6.72</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 2.82159   R-Sq = 71.5%   R-Sq(adj) = 69.9%

We test $H_0: \beta = 0$ versus $H_a: \beta \neq 0$

$t = b / SE_b = 2.3308 / 0.3468 = 6.721$, with df = $n - 2 = 18$  

Table C: $t > 3.922 \Rightarrow P < 0.001$ (two-sided test), highly significant.

$\Rightarrow$ There is a significant relationship between temperature and call frequency.
Testing for lack of correlation

The regression slope $b$ and the correlation coefficient $r$ are related and $b = 0 \implies r = 0$. \[ \text{slope } b = r \frac{s_y}{s_x} \]

Similarly, the population parameter for the slope $\beta$ is related to the population correlation coefficient $\rho$, and when $\beta = 0 \implies \rho = 0$.

Thus, testing the hypothesis $H_0: \beta = 0$ is the same as testing the hypothesis of no correlation between $x$ and $y$ in the population from which our data were drawn.
Researchers examined the relationship between straightforwardness (using a personality inventory test score) and brain levels of serotonin transporters (in the dorsal raphe nucleus) in 20 adult subjects. With no obvious explanatory variable, we choose a test for the population correlation $\rho$. We test:

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

With a two-sided test, we get $0.02 < P < 0.01$ from Table E ($P = 0.014$ from software), significant.

**Correlations: serotonin, straightf**  
Pearson correlation of serotonin and straightf = -0.538  
P-Value = **0.014**

---

**TABLE E** Critical values of the correlation $r$

<table>
<thead>
<tr>
<th>n</th>
<th>.20</th>
<th>.10</th>
<th>.05</th>
<th>.025</th>
<th>.02</th>
<th>.01</th>
<th>.005</th>
<th>.0025</th>
<th>.001</th>
<th>.0005</th>
</tr>
</thead>
</table>

$|r| = 0.538$

---

**Regression Analysis: serotonin versus straightf**

<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>4.8544</td>
<td>0.6814</td>
<td>7.12</td>
<td>0.000</td>
</tr>
<tr>
<td>straightf</td>
<td>-0.04114</td>
<td>0.01519</td>
<td>-2.71</td>
<td><strong>0.014</strong></td>
</tr>
</tbody>
</table>
One use of regression is for **prediction within range**: \( \hat{y} = a + bx \).
But this prediction depends on the particular sample drawn.
We need statistical inference to generalize our conclusions.

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 \( \mu_y \).
Confidence interval for $\mu_y$

We may also want to predict the population mean value of $y$, $\mu_y$, for any value of $x$ within the range of data tested.

Using inference, we calculate a **level C confidence interval for the population mean $\mu_y$** of all responses $y$ when $x$ takes the value $x^*$:

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 computed from many different random samples.
A level C prediction interval for a single observation on $y$ when $x$ takes the value $x^*$ is:

$$
\hat{y} \pm t^* \text{SE}_{\hat{y}}
$$

where

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

A level C confidence interval for the mean response $\mu_y$ at a given value $x^*$ of $x$ is:

$$
\hat{y} \pm t^* \text{SE}_{\hat{\mu}}
$$

where

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

Use $t^*$ for a $t$ distribution with $\text{df} = n - 2$. 
Confidence interval for the mean call frequency at 22 degrees C: 
\( \mu_y \) within 43.3 and 46.9 notes/sec

Prediction interval for individual call frequencies at 22 degrees C: 
\( \hat{y} \) within 38.9 and 51.3 notes/sec

**Minitab**
Predicted Values for New Observations: Temperature=22

<table>
<thead>
<tr>
<th>NewObs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95% CI</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45.088</td>
<td>0.863</td>
<td>(43.273, 46.902)</td>
<td>(38.888, 51.287)</td>
</tr>
</tbody>
</table>
Blood alcohol content (bac) as a function of alcohol consumed (number of beers) for a random sample of 16 adults.

Is the relationship linear?

What do you predict for five beers consumed?
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

For any fixed $x$, the responses $y$ follow a Normal distribution with standard deviation $\sigma$. 

$\mu_y = \alpha + \beta x$
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 → $\sigma$ not equal for all values of $x$. 
Frog mating call

The data are a random sample of frogs.

The relationship is clearly linear.

The residuals are roughly Normally distributed.

The spread of the residuals around 0 is fairly homogenous along all values of $x$. 

![Residual plots for frequency](image)