Point Estimators and Predictors

Once we are convinced that the model is reasonable, we can use the fitted regression equation

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \]

to estimate \( E(y|x) \) and to predict a future value of \( y \) for a given \( x \).
Consider the Salary vs. Height example.

Fitted Line Plot for Salary vs. Height

Salary = -902.2 + 100.4 Height

S = 192.702
R-Sq = 71.4%
R-Sq(adj) = 70.3%
Regression Analysis: Salary versus Height

The regression equation is
Salary = - 902 + 100 Height

<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>-902.2</td>
<td>837.0</td>
<td>-1.08</td>
<td>0.290</td>
</tr>
<tr>
<td>Height</td>
<td>100.36</td>
<td>12.02</td>
<td>8.35</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 192.702    R-Sq = 71.4%    R-Sq(adj) = 70.3%

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>2590433</td>
<td>2590433</td>
<td>69.76</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>28</td>
<td>1039754</td>
<td>37134</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>3630187</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Predicted Values for New Observations

New

<table>
<thead>
<tr>
<th>Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95% CI</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6323.5</td>
<td>45.5</td>
<td>(6230.3, 6416.7)</td>
<td>(5917.9, 6729.1)</td>
</tr>
</tbody>
</table>

Values of Predictors for New Observations

New

<table>
<thead>
<tr>
<th>Obs</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>72.0</td>
</tr>
</tbody>
</table>
The fitted model is $\hat{y} = -902.2 + 100.36 \times x$.

The $p$-value for $\beta_1$ is 0.000, and $R^2$ is 0.714, indicating statistically significant and strong linear association.

Now, consider a (hypothetical) 6-foot-tall graduate.

His salary is random, since it cannot be predicted with certainty based on his height.

The mean (i.e., expected) salary for such a graduate is

$$E(y \mid x) = \beta_0 + \beta_1 (72) ,$$

which we can estimate by the fitted value

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 (72) = -902.2291 + 100.3577 (72) = 6323.5 .$$
If, instead of estimating the mean salary, we wanted to predict the salary *actually obtained* by a given 6-foot-tall graduate, we would use the same value,

\[ \hat{y} = 6323.50. \]

- Thus, the point estimator of \( E(y \mid x) \) and the point forecast of a future \( y \) are numerically identical (both are simply \( \hat{y} \)) but conceptually different, since the "targets" \( E(y \mid x) \) and the future \( y \) are different.

When we ask for *interval* estimators and predictions, however, a numerical difference emerges.
• If we simply want an interval estimator for $E(y \mid x)$, the mean value of $y$ for a given $x$, we should use the confidence interval.

This provides an (exact) 95% confidence interval for $E[Y \mid X = x_0]$ if the $\varepsilon_i$ are normal.

Even if the $\varepsilon_i$ are not normal, the interval will be valid in large sample sizes.

We can obtain this confidence interval from the Minitab output. The interval is denoted by "CI".
Eg: For the Salary vs. Height example, before running the regression, we select "Options". We enter the value 72 in the box marked "Prediction intervals for new observations:". (This will generate a 95% confidence interval and prediction interval).

Minitab gives a fitted value of $6323.50, and a 95% confidence interval for $E(y|x)$ of ($6230.30, $6416.70$).

• If we want to predict a future value of $y$ given a specific value of $x$, we use the prediction interval.

The prediction interval is denoted "PI" by Minitab.

Eg: For the Salary vs. Height example, the 95% prediction interval for the salary of a 6-foot-tall graduate is ($5917.90, $6729.10$).
Interpretation of a Prediction Interval:

If we repeat the experiment of obtaining a regression data set many times, each time forming a 95% prediction interval at $X = x_0$, and waiting to see what the future value of $Y$ is at $X = x_0$, then roughly 95% of the prediction intervals will contain the corresponding actual future value of $Y$.

If we declare that the future value will lie within the given prediction interval, then statements of this kind will be correct 95% of the time, in the long run.
• Both the confidence interval and the prediction interval require that the errors be normally distributed in order to be valid for small samples.

• A prediction interval is similar in spirit to a confidence interval, except that the prediction interval is designed to cover a "moving target", the random future value of $y$.

The confidence interval seeks to cover the fixed target, $E(y|x)$.

• Although both are centered at $\hat{y}$, the prediction interval is wider than the confidence interval, for given $x$.

This makes sense, since the prediction interval must take account of the tendency of $y$ to fluctuate from its mean value, $E(y|x)$. 
The confidence interval simply needs to account for the uncertainty in estimating $E(y|x)$.

• For a given data set, the error in estimating $E(y|x = x_0)$ grows as $x_0$ moves away from $\bar{x}$. Thus, the further $x_0$ is from $\bar{x}$, the wider the confidence intervals and prediction intervals will be.

Why does this make sense?

• If any of the assumptions underlying the model are violated, then the confidence intervals and prediction intervals may be invalid as well. This is one reason why it is so important to check the assumptions by examining residuals, etc.
Fitted Line Plot With CI and PI

Salary = -902.2 + 100.4 Height

<table>
<thead>
<tr>
<th>Regression</th>
<th>95% CI</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>192.702</td>
<td></td>
</tr>
<tr>
<td>R-Sq</td>
<td>71.4%</td>
<td></td>
</tr>
<tr>
<td>R-Sq(adj)</td>
<td>70.3%</td>
<td></td>
</tr>
</tbody>
</table>
Eg: UK Mean Temperatures

UK Mean Temperatures with CI and PI

\[ \text{JUL} = -4.025 + 0.009518 \text{ Year} \]

<table>
<thead>
<tr>
<th>S</th>
<th>1.01247</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-Sq</td>
<td>6.1%</td>
</tr>
<tr>
<td>R-Sq(adj)</td>
<td>5.1%</td>
</tr>
</tbody>
</table>
• The PI refers to the *weather*, while the CI refers to the *climate*.

The 95% PI for the temperature in July 2025 is 
(13.165, 17.333) Degrees C.

This seems very wide, but its width reflects both the uncertainty in the expected temperature (climate) for July, 2025, as well as likely fluctuations (weather) from the expected temperature.

Even the 95% CI for the expected temperature in July 2025 is fairly wide, at (14.702, 15.795) Degrees C. One important reason is that 2025 is far outside the range of the actual historical data used in the regression.
Regression Analysis: JUL versus Year

The regression equation is
JUL = - 4.02 + 0.00952 Year

Predictor      Coef  SE Coef    T      P
Constant     -4.025   7.666  -0.53  0.601
Year         0.009518 0.003911  2.43  0.017

S = 1.01247   R-Sq = 6.1%   R-Sq(adj) = 5.1%

Predicted Values for New Observations

New
Obs  Fit  SE Fit  95% CI  95% PI
1  15.249 0.275  (14.702, 15.795)  (13.165, 17.333)

X denotes a point that is an outlier in the predictors.

Values of Predictors for New Observations
New
Obs  Year
1  2025